Disentangling diatom species complexes: does morphometry suffice?

Saúl Blanco[1], María Borrego-Ramos[1] and Adriana Olenici[2]

[1] Institute of the Environment, Leon, Spain
[2] Faculty of Environmental Sciences and Engineering, Babes-Bolyai University of Cluj-Napoca, Romania

ABSTRACT

Accurate taxonomic resolution in light microscopy analyses of microalgae is essential to achieve high quality, comparable results in both floristic analyses and biomonitoring studies. A number of closely related diatom taxa have been detected to date co-occurring within benthic diatom assemblages, sharing many morphological, morphometrical and ecological characteristics. In this contribution, we analysed the hypothesis that, where a large sample size (number of individuals) is available, common morphometrical parameters (valve length, width and stria density) are sufficient to achieve a correct identification to the species level. We focused on some common diatom taxa belonging to the genus *Gomphonema*. More than 400 valves and frustules were photographed in valve view and measured using Fiji software. Several statistical tools (mixture and discriminant analysis, k-means clustering, classification trees, etc.) were explored to test whether mere morphometry, independently of other valve features, leads to correct identifications, when compared to identifications made by experts. In view of the results obtained, morphometry-based determination in diatom taxonomy is discouraged.

Subjects Bioinformatics, Ecology, Marine Biology, Taxonomy, Freshwater Biology
Keywords Taxonomy, Biometry, Automated identification, Classification, Multivariate statistics

Submitted 28 September 2017
Accepted 22 November 2017
Published 11 December 2017

Corresponding author
Saúl Blanco, sblal@unileon.es

Academic editor
James Reimer

Additional Information and Declarations can be found on page 13

DOI **10.7717/peerj.4159**

Copyright
2017 Blanco et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

INTRODUCTION

Diatoms are unicellular algae inhabiting many different aquatic and terrestrial environments worldwide. To date, ∼10^5 different species have been described (*Mann & Droop, 1996*), with particular ecological preferences, so that there is a clear relationship between diatom communities and the environmental characteristics of their habitats. The reliability of diatom-based biomonitoring methods has long been established, but diatom analyses are also useful in palaeoecology, biotechnology or forensic applications (*Stoermer & Smol, 2001*). However, the main obstacle limiting their use lies in the difficulty of their taxonomic identification, since diagnoses at specific or subspecific levels are often required. This implies important investments in optical equipment and expert training. Currently, the identification and routine counting of diatoms is performed under light/phase contrast optical microscopy, but several tools are being proposed to automate the identification process by means of image analyses (*Buf & Bayer, 2002*; *Kloster, Kauer & Beszteri, 2014*) or DNA metabacoding (e.g., *Vasselon et al., 2017*).

Diatom cell size (length [L], width [W], L/W ratio) and other morphometric parameters (e.g stria density [S]) are commonly used in taxonomic keys aiding identification,

How to cite this article Blanco et al. (2017), Disentangling diatom species complexes: does morphometry suffice? **PeerJ 5:e4159; DOI 10.7717/peerj.4159**

together with shape features and ultrastructural characteristics. Despite the fact that basic morphometric insights are of fundamental importance to diatom taxonomy (*Beszteri, Ács & Medlin, 2005*), the use of quantitative approaches in diatom identification is relatively rare (*Beszteri, 2005*) and taxa identification is mostly based on cell shape, which is considered to be stable on a large scale (*Mou & Stoermer, 1992*). Due to their reproductive cycle, there is a tendency within each species to decrease in cell size after several generations so that natural populations exhibit broad and often skewed size distributions, since they are made up of several age classes (*Spaulding et al., 2012*). This cyclical change in size can vary also between populations (*Pappas & Stoermer, 2003*) and is often accompanied by an allometric change in the L/W ratio (*Schmid, 1994*). Some studies suggest also a dependence of morphometric parameters on habitat features (*Cortese & Gersonde, 2007*).

Diatom taxa are often arranged in groups of morphologically similar forms that share numerous features and with overlapping morphometric ranges (*Potapova & Hamilton, 2007*). A number of "species complexes" have been reported with a high level of diversity (*Ajani et al., 2013*). These complexes may occur in sympatry (*Kulichová & Fialová, 2016*), so that several taxa can be found in a single diatom community. The resolution of morphologically similar taxa is especially difficult for microscopy analyses, and misidentifications can lead to inaccurate environmental inferences. In this paper, we explore the possibility of disentangling one of these groups (species related to *Gomphonema gracile* Ehr. (*Reichardt, 2015*) and *G. parvulum* (Kütz.) Kütz.) by unsupervised classification of individuals based only on their morphometric parameters. We tested a number of classification algorithms and compared their result with identifications made by experts under light microscopy (LM). Our hypothesis is that, provided large training sets are available, morphometric parameters would suffice for unsupervised classification, overriding the need for examining other morphological features.

MATERIALS & METHODS

Field and laboratory routines

A sample of epiphytic diatoms was collected from the surface of reed stems in Lake Villadangos (southeast León, northwest Spain, UTM 30T 272100 4711400) during July 2000. This an anthropic wetland located near Villadangos (León), and has a mean depth of 0.4 m, is 9.4 ha in and has 1.2 km perimeter, characterized by the presence of eutrophic waters. A detailed description of the wetland is available in *Conty (2007)*. The sample was processed and analysed under light microscopy (LM) following standard protocols (*CEN, 2003*; *CEN, 2004*). Large populations of *Gomphonema* taxa dominated the diatom community, with 25 different species that were identified using a microscope (Leica DMRB, DIC 1,000×) according to usual reference works (*Hofmann, Werum & Lange-Bertalot, 2011*). All *Gomphonema* individuals lying in valve view ($N = 523$) were enumerated and photographed (Canon EOS400). Only five species attained large ($N \geq 45$) populations (namely *G. gracile*, *G. auritum* A.Braun, *G. jadwigiae* Lange-Bert. and E.Reichardt, *G. acidoclinatum* Lange-Bert. and E.Reichardt and *G. parvulum*, Fig. 1), which were considered in subsequent analyses ($N = 410$). Morphometric parameters (length, width,

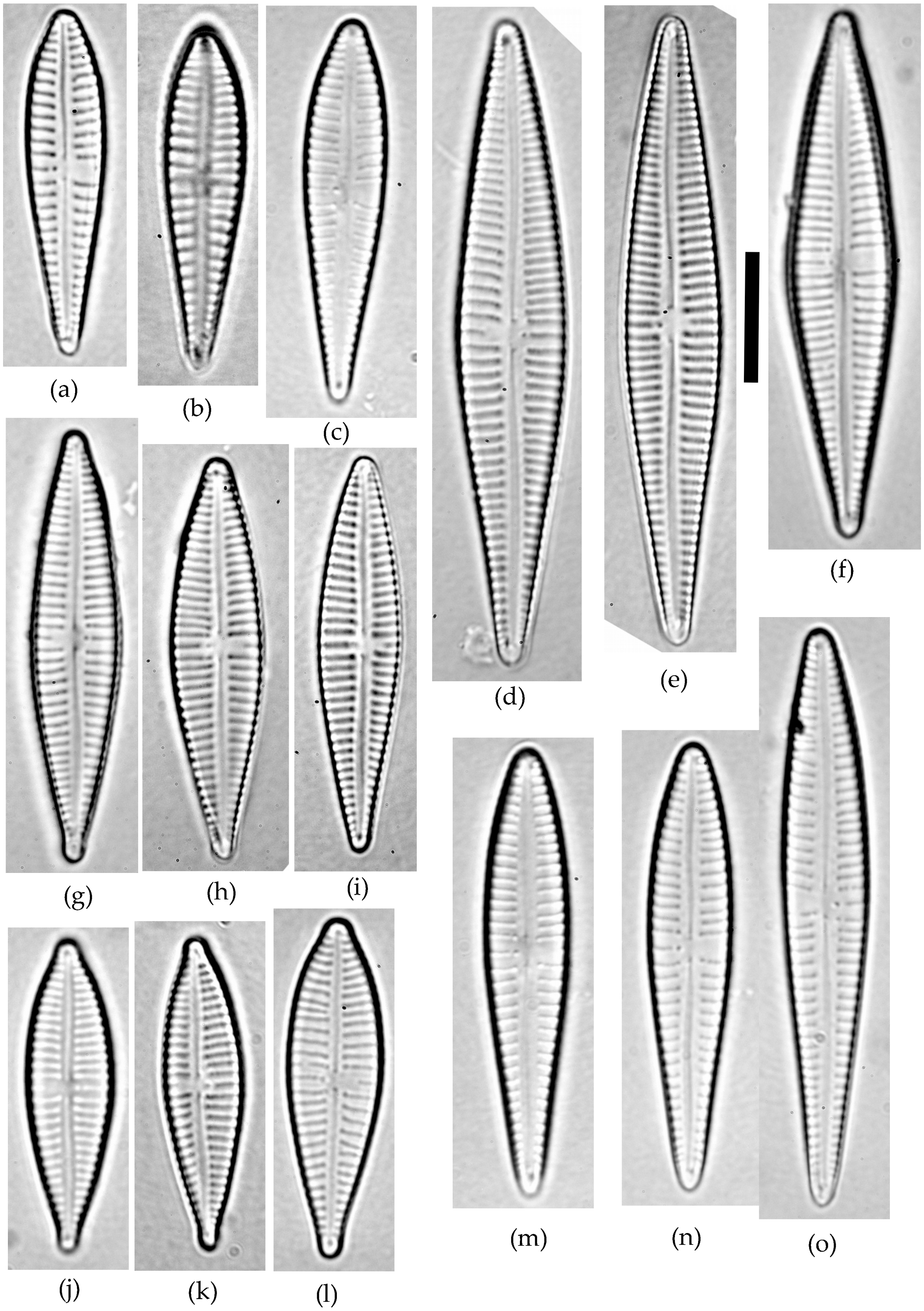

Figure 1 ***Gomphonema*** **species analyzed in the study.** (A–C) *G. jadwigiae*. (D–F) *G. gracile*. (G–I) *G. acidoclinatum*. (J–L) *G. parvulum*. (M–O) *G. auritum*. Scale bar = 10 µm.

Full-size DOI: 10.7717/peerj.4159/fig-1

length-width ratio and stria density, hereafter L, W, L/W and S respectively) were measured in each individual using Fiji software (*Schindelin et al., 2012*). $N = 45$ is the smallest sample size that represents the normal range (90%) of any population following a continuous distribution, with a 0.95 confidence. Original data are publicly available at FigShare (DOI: 10.6084/m9.figshare.4728406).

Statistical analyses

Ten different numerical tools were tested and compared in this study. These algorithms were selected from among other many analogous methods proposed in the body of literature because (a) they are commonly available in statistical software applications, and (b) they have already been used in similar analyses.

1. Mixture analysis (MA): a maximum-likelihood distribution-based approach to test whether a variable distribution fits better in a mixture of (*a priori* defined) *n* normal overlapping distributions. The best-fit model can be chosen based on AIC values. Computations use the EM algorithm (*Dempster, Laird & Rubin, 1977*).

2. Canonical variates analysis (CVA): a discriminant analysis that evaluates quantitatively the distinctiveness of pre-classified groups of objects, estimating the spatial directions that maximize the differences between these groups. The output is a multivariate ordination plot that provides a linear combination of the classification variables having the highest possible multiple correlation with the selected groups. Computational details are available in *Hammer, Harper & Ryan (2001)*.

3. Chi-squared automatic interaction detector (CHAID): a sophisticated segmentation modelling method for analysing large quantities of categorical data (*Kass, 1980*). It consists of a multivariate criterion-based algorithm that divides the test population into a number of distinct groups based on the categories of the most significant attribute. It uses the Chi-square test to determine the best next split at each step.

4. Random forests (RF): a nonparametric ensemble classification method that predicts classes based on the partition of input variables from multiple decision trees. The most reliable predictor is based on the decrease of classification accuracy when values of an attribute in a tree node are permuted randomly (*Breman, 2001*). Random forests produce lower prediction errors than other classification tree algorithms (*Felde et al., 2014*).

5. Boosted classification trees (BCT): a machine learning method which produces a prediction model in the form of an ensemble of decision trees. The algorithm uses a random sample of observations, builds independent sets of boosted trees for each category of the dependent variable, computes the predicted values for the observations in that sample, and fits a regression tree to the residuals, applying a logistic transformation to the predicted values before computing these residuals (*Friedman, 2001*).

6. k-means clustering (KMC): a nonhierarchical clustering method that searches for the partition of a sample into an *a priori* given number of groups so that the within-group sum of squares is minimal (*Hartigan, 1975*).

7. Expectation-maximisation (EM): a similar method that performs clustering by fitting a mixture of *n* different distributions to the data. The algorithm estimates the means and

standard deviations for each cluster so as to maximise the likelihood of the observed input data (*Witten & Frank, 2005*).
8. Support vector machine (SVM): a learning algorithm that uses a hypothesis space of linear functions in a high-dimensional feature space, trained with a learning algorithm from optimization theory. It attempts to minimize the upper bound on the generalization error based on the principle of structural risk minimization (*Hongzong et al., 2007*).
9. Naïve Bayes Classifier (NBC): a simple probabilistic classifier where a single node, which represents a classification variable, is connected to all other nodes that represent predictor variables. The method is based on Bayesian theory with strong independence assumptions that the presence/absence of a particular feature of a class is not related to the presence/absence of any other feature (*Liu et al., 2011*).
10. K-nearest neighbour (KNN): a simple, nonparametric classifier in which the class of each object is determined with respect to the classes assigned to the nearest *k* objects, that is, it is specified according to the most repeated labels of these *k* objects (*Han et al., 2015*).

Computations and graphical outputs were performed with PAST v. 3.14 (*Hammer, Harper & Ryan, 2001*), Statistica v 10 (StatSoft, Tulsa, OK, USA: http://www.statsoft.com) and R (*R Core Team, 2016*) under RStudio (*RStudio Team, 2015*) with the Caret (*Wing et al., 2016*) and Ellipse (*Murdoch & Chow, 2013*) packages.

RESULTS

All analysed morphometric parameters overlapped between the species considered (e.g., L/W, Fig. 2). The lowest variability in terms of L, L/W and S parameters was found in *G. jadwigiae* (CV of 11.9%, 6.2% and 9.9%, respectively), while the least variable population in cell width was that of *G. auritum* (7.5%). Most parameters evaluated showed positive skewness, indicating right-tailed distributions. None of the variables examined in any of the *Gomphonema* species followed Gaussian distributions (Shapiro–Wilk test, $p < 0.05$). As expected, the L/W relationship was monotonic positive for all populations (Fig. 3). Stria density did not correlate with any other morphometric parameter.

MA

The algorithm was set to fit five different populations (Fig. 4), although the lowest AIC values were found forcing 6–7 groups. Since MA is a univariate method, it provided four different classifications according to each parameter tested. On average, only 59.8% of the individuals were correctly classified by the MA method, with the highest success for *G. jadwigiae* (67.0%) and the lowest success for *G. parvulum* (40.0%). The best average results were obtained with W parameter (74.3%), while S had the lowest predictive power (43.8%).

CVA

The resulting ordination of this method is shown in Fig. 5, highlighting cell length as the most important variable for specimen classification. The species most frequently misidentified according to this algorithm was *G. acidoclinatum* (only 33.7% of correct

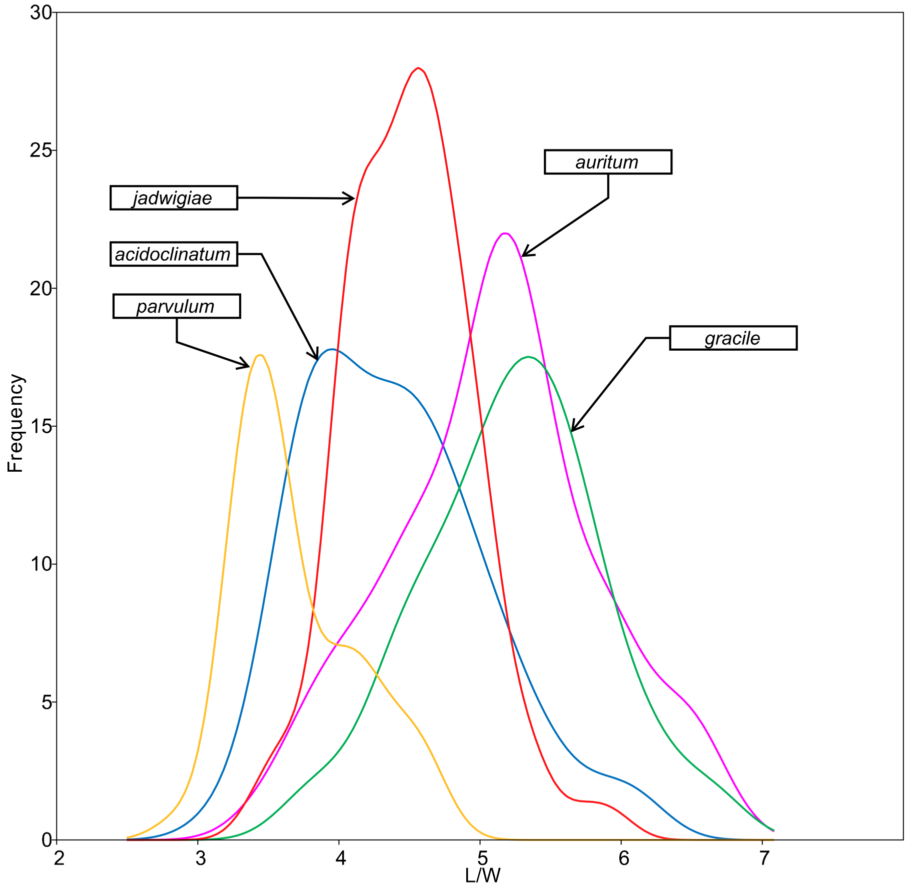

Figure 2 L/W histogram for the analysed Gomphonema populations: *G. jadwigiae* (*n* = 94), *G. acidoclinatum* (*n* = 89), *G. auritum* (*n* = 101), *G. gracile* (*n* = 80) and *G. parvulum* (*n* = 46). Data fitted to kernel distribution estimators.

Full-size DOI: 10.7717/peerj.4159/fig-2

assignations), whereby 29.2% of cases were identified erroneously as *G. parvulum*. On the contrary, *G. gracile* was correctly classified in 68.8% of cases. In total, 55.1% of items were correctly classified.

CHAID

A classification tree using only L and W parameters as classifying variables can be drawn (Fig. 6), but with an estimated classification risk of 44.2 ± 0.02%. According to the resulting confusion matrix (data not shown), as in the case of CVA results, *G. acidoclinatum* was the taxon most often misidentified (in 30.4% of cases), frequently (in 23.6% of cases) as *G. jadwigiae*, whereas *G. gracile* was again correctly classified in 77.5% of cases. Resulting classification used only L and W variables.

RF

Random forests algorithm selected the classification tree shown in Fig. 7 as the most parsimonious among the other 200 models tested. This led to a success percentage in correct assignations of 61.7%. The most important classificatory variable was L. Most

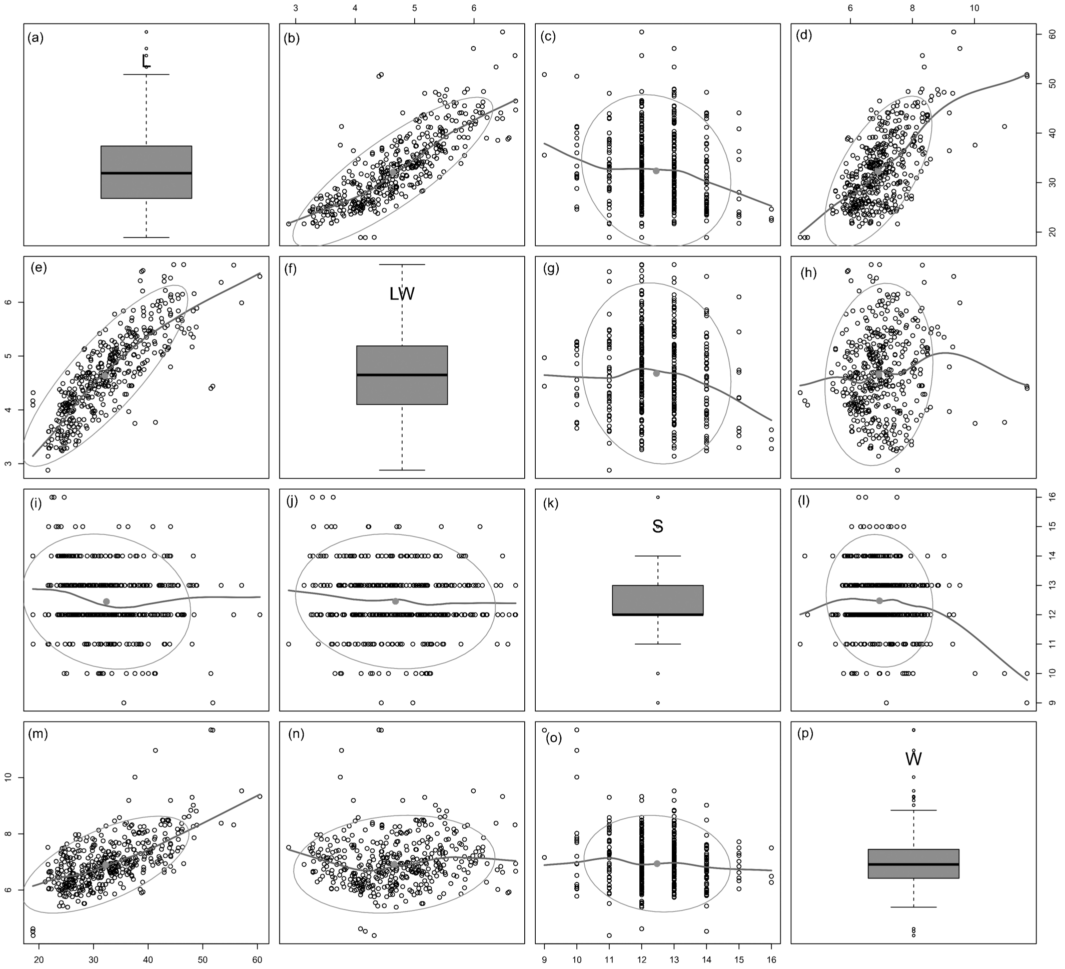

Figure 3 Scatterplot matrix of correlations between analysed morphometric parameters (A–P). Data fitted to LOESS smoothers and 95% confidence ellipses.

Full-size DOI: 10.7717/peerj.4159/fig-3

misidentifications affected *G. acidoclinatum* (50.0%), erroneously considered *G. jadwigiae* in 21.9% of cases. In this algorithm, the taxon achieving maximal correct assignations (68.3%) was *G. jadwigiae*.

BCT

The classificatory success of this method is 67.9%. As in the case of RF, the most important classificatory variable was L. Most frequent misclassifications were observed again in *G. acidoclinatum* (45.2%), frequently (16.1%) assigned to *G. jadwigiae*. *Gomphonema gracile* was correctly identified in 77.1% of cases.

KMC

K-means clustering forced on 5 groups achieved an average of 70.2% of correct classifications. While taxa such as *G. acidoclinatum* or *G. jadwigiae* were always (100%) assigned to unique clusters, the algorithm failed to discriminate *G. auritum* and *G. gracile*,

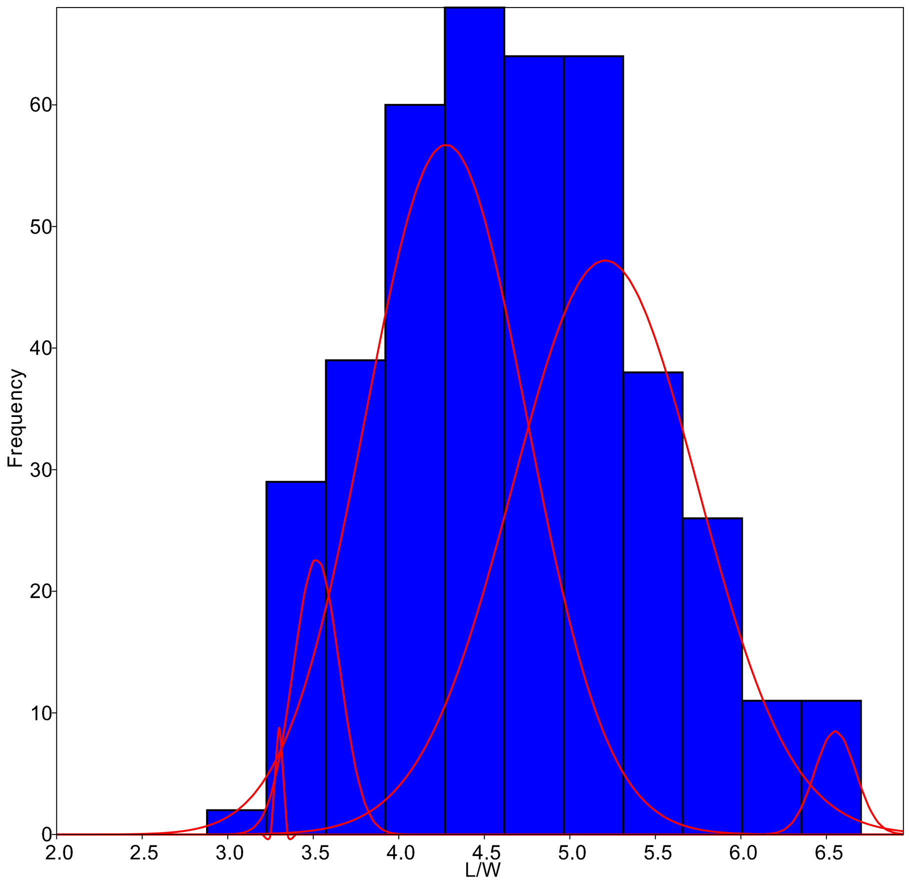

Figure 4 L/W histogram for the whole dataset and MA distributions adjusted to five classes (red curves).

Full-size DOI: 10.7717/peerj.4159/fig-4

which were gathered in the same cluster in 81% of cases. *Gomphonema parvulum* was also misidentified in 48.9% of cases.

EM

The success percentage of this method was only 44.1%. As in the case of KMC, two different class objects (*G. acidoclinatum* and *G. parvulum*) were erroneously gathered together. The identification success ranked from 44.3% (*G. acidoclinatum*) to 71.3% (*G. jadwigiae*).

SVM

A total of 60.2% specimens were correctly classified by this algorithm. The species most often misidentified was *G. parvulum* (50.0%, of which 40.0% were misidentified as *G. jadwigiae*), while 71.4% of *G. gracile* individuals tested were correctly identified.

NBC

A total of 62.1% specimens were correctly classified by this algorithm. The species most often misidentified was *G. acidoclinatum* (71.4%, of which in 23.8% of cases misidentified as *G. parvulum*), while 76.9% of the *G. jadwigae* individuals tested were correctly identified.

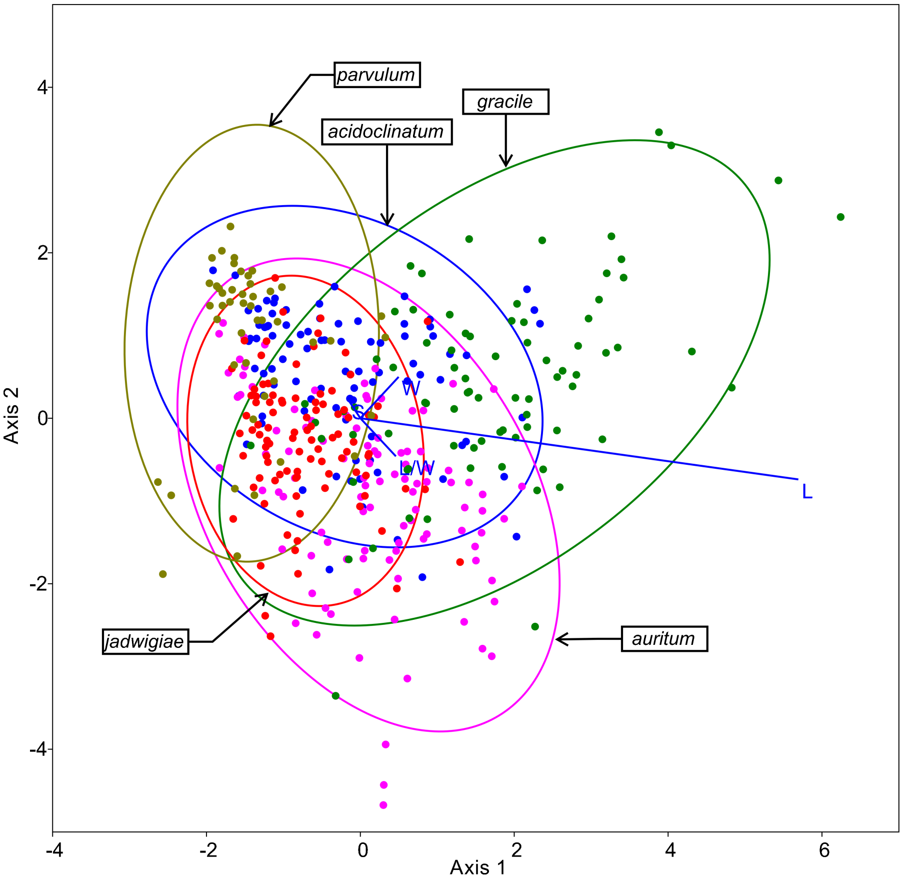

Figure 5 CVA ordination biplot, dots represent individuals and line predictor variables. Points fitted to 95% confidence ellipses.

Full-size DOI: 10.7717/peerj.4159/fig-5

KNN

The predictive power of this method was only 49.5%. The percentage of correct classifications ranked from 61.9% (*G. gracile*) to 30.0% (*G. parvulum*, in 40.0% of cases misidentified as *G. auritum*).

DISCUSSION

During recent decades, quantitative techniques have proposed many different methods to suggest classifications of organisms based on metric characters (*Julius et al., 1998*). Currently, size analysis is a potentially powerful tool for understanding diatom community dynamics and systematic relationships (*Spaulding et al., 2012*). Our study tested different classification algorithms based on metric and meristic parameters that are commonly recorded in diatom taxonomy, and which have proven to be useful to segregate morphologically similar taxa. For instance, *Paull et al. (2008)* demonstrated using linear discriminant analysis that two closely related *Staurosirella* species could be distinguished

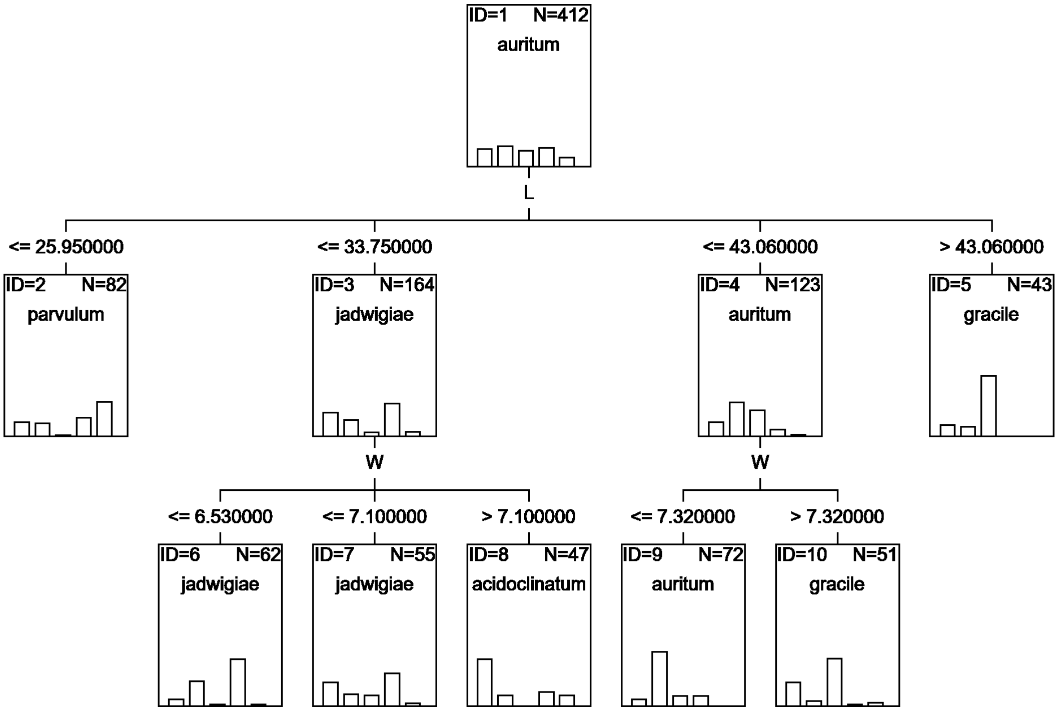

Figure 6 **CHAID classification tree.**

Full-size DOI: 10.7717/peerj.4159/fig-6

with an error of 6% using only cell width and areolae length. In the genus *Kobayasiella*, a single morphometric character can separate all individuals from sister species (*Buczkó, Wojtal & Jahn, 2009*).

Our results show that a majority of methods selected L as the best classificatory parameter, in contrast with previous studies showing S and W as the most stable characters (*Genkal, 2004*). Particularly, S is known to change little with changing valve length, although it may vary due to environmental factors (*Cox, 2010*). On the contrary, S has been found to be inadequate to distinguish diatom species in other works (*Paull et al., 2008*). With respect to L/W parameter, this is the most popular ratio in algal taxonomy, and is commonly regarded as a reliable parameter, despite the fact that it is often size-dependent, compounding variations from several sources (*Theriot, 1988*).

Multivariate methods such as cluster or classificatory algorithms have started to be adopted in diatom taxonomy (*Buczkó, Wojtal & Jahn, 2009*). These techniques are, in contrast to the classical methods, more robust when dealing with complex multivariate data (*Felde et al., 2014*). Both cluster analysis and ordination techniques have similar aims in that they attempt to explore multivariate datasets by reducing their dimensionality and summarising the major patterns of variation within the datasets (*Felde et al., 2014*). In our analyses, cluster methods (KMC, MA, KNN, EM) performed somewhat worse than classification algorithms (average classification risk: 44.1% *vs*. 39.5%, Table 1). Although the computational requirements are much lower in clustering techniques, the assignation of objects to each cluster must be supervised *a posteriori* (in this case, by calculating the

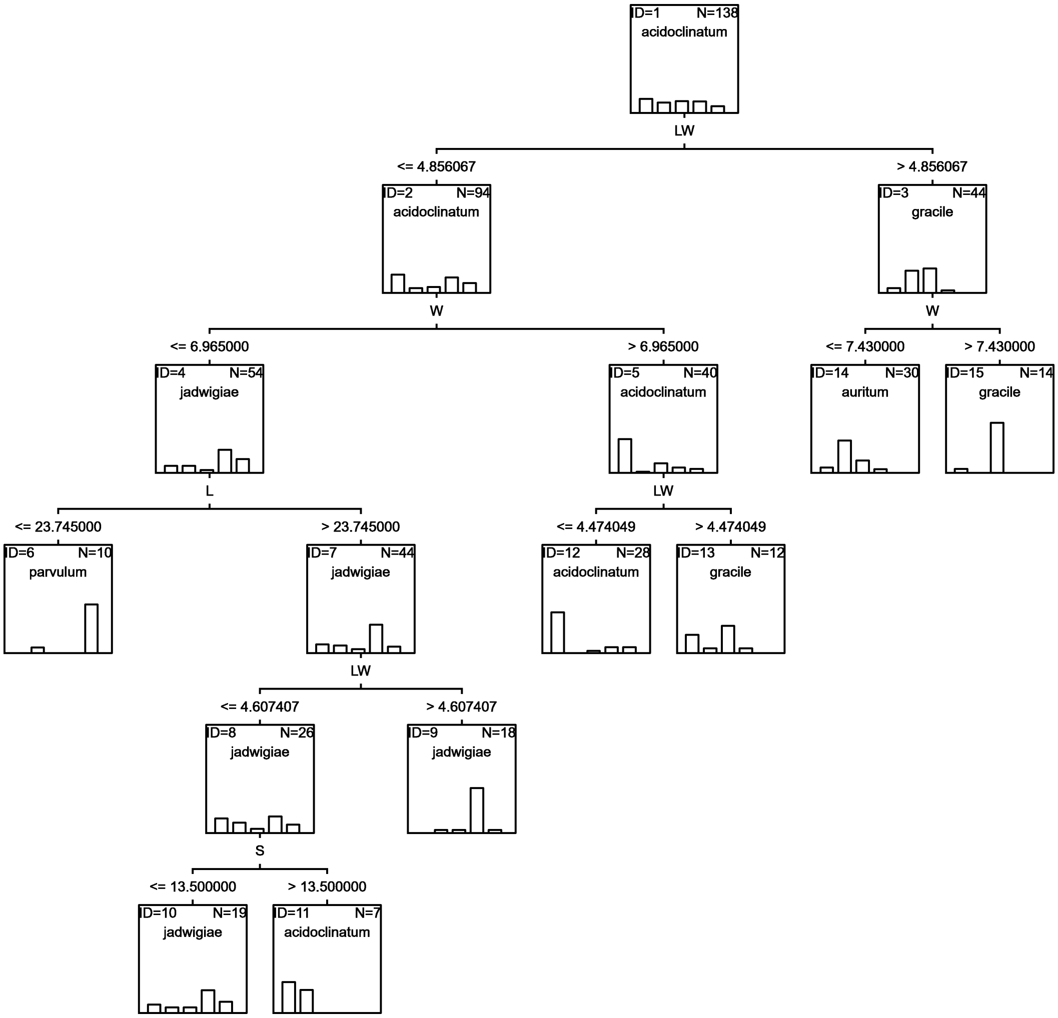

Figure 7 RF classification tree.

Full-size DOI: 10.7717/peerj.4159/fig-7

modal class of each cluster). The best results (70.2% correct identification) were obtained by the relatively simple KMC, but the outcome of this algorithm suggested the presence of only 4 clusters. The Calinsky and silhouette criteria confirm that optimal partitioning of data is obtained for $n = 4$. The confusion matrix shows that the algorithm gathered *G .acidoclinatum* and *G. parvulum* in the same group, and this misidentification accounted for the largest amountof the classification failure observed. The MA method also indicated a different number of taxa in the studied assemblage according to AIC values, although the statistically optimum number determined by AIC may not be particularly useful (*Felde et al., 2014*). Within classification methods, best results were obtained by BCT (67.9%), and this method has proven to provide more accurate results when compared to other tree-based classification techniques (*Bauer & Kohavi, 1999*; *Austin & Lee, 2011*).

Contrary to our expectations, classification success was unrelated to sample size (number of items per class). The taxa that were most oten correctly classified and less

Table 1 Percentage of correct classifications achieved by each method.

Method	Correct classifications (%)
*KMC	70.2
BCT	67.9
NBC	62.1
RF	61.7
SVM	60.2
*MA	59.8
CHAID	55.8
CVA	55.1
*KNN	49.5
*EM	44.1

Notes.

*Cluster methods.

often misclassified (*G. acidoclinatum* and *G. gracile*, respectively) were not the species with the largest or lowest numbers of individuals. Classification risk was also independent of the variability of each species (measured as CV in L, W, L/W and S). This suggests that unsupervised classification based on metric parameters may lead to consistent results independently of the descriptive statistics of the groups involved.

Notwithstanding, the best scores achieved by the methods analysed in this study are far below other automated identification techniques that consider not only morphometry, but also cell shape and structural features, even when tested over similar species. For instance, SHERPA obtained classification accuracies ranging from 98.9% to 100.0% when applied to the *Sellaphora pupula* complex (*Kloster, Kauer & Beszteri, 2014*). Similarly, the system developed in the ADIAC project allowed the identification of 37 species with an accuracy of 97% (*Buf & Bayer, 2002*). Similar success ratios have been reported using a multi-label classification system for diatom image classification developed by *Dimitrovski et al. (2012)*. This shows that the biological relevance of the morphological distinctness of diatoms depends on whether the differences can be explained simply by size differences (*Beszteri, Ács & Medlin, 2005*), and otherwise shape analyses allows the segregation of groups that have different size ranges but vary only subtly in other characters (*Mou & Stoermer, 1992*). When shape group separation is not evident, morphometric measures such as L, W or S may be used (*Kingston & Pappas, 2009*).

CONCLUSIONS

In the light of our results, we cannot recommend the exclusive use of morphometric measurements for unsupervised diatom classification that aims at segregating morphologically similar species. According to our results and the available literature, the combined use of morphometry and morphology seems to best suit this purpose.

Abbreviations

BCT	Boosted classification trees
CHAID	Chi-squared automatic interaction detector

CVA	Canonical variates analysis
EM	Expectation-maximisation
KMC	K-means clustering
KNN	K-nearest neighbour
MA	Mixture analysis
NBC	Naïve Bayes Classifier
RF	Random forests
SVM	Support vector machine

ACKNOWLEDGEMENTS

The authors are grateful to M Doucet and P Del Peso who took the photographs and made the cell measurements.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the Spanish Government under the Aqualitas-retos project (grant number CTM2014-51907-C2-2-R-MINECO). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Spanish Government under the Aqualitas-retos project: CTM2014-51907-C2-2-R-MINECO.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Saúl Blanco conceived and designed the experiments, analyzed the data, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- María Borrego-Ramos and Adriana Olenici performed the experiments, analyzed the data, reviewed drafts of the paper.

Data Availability

The following information was supplied regarding data availability:

Blanco, Saul (2017): Blanco et al Gomphonema.csv. figshare. https://dx.doi.org/10.6084/m9.figshare.4728406.v1.

REFERENCES

Ajani P, Murray S, Hallegraeff G, Lundholm N, Gillings M, Brett S, Armand L. 2013. The diatom genus *Pseudo-nitzschia* (Bacillariophyceae) in New South Wales, Australia: morphotaxonomy, molecular phylogeny, toxicity, and distribution. *Journal of Phycology* **49**:765–785 DOI 10.1111/jpy.12087.

Austin PC, Lee DS. 2011. Boosted classification trees result in minor to modest improvement in the accuracy in classifying cardiovascular outcomes compared to conventional classification trees. *American Journal of Cardiovascular Diseases* **1**:1–15.

Bauer E, Kohavi R. 1999. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. *Machine Learning* **36**:105–139 DOI 10.1023/A:1007515423169.

Beszteri B. 2005. *Morphometric and molecular investigations of species limits in Cyclotella meneghiniana (Bacillariophyceae) and closely related species*. Bremen: Universität Bremen.

Beszteri B, Ács É, Medlin L. 2005. Conventional and geometric morphometric studies of valve ultrastructural variation in two closely related *Cyclotella* species (Bacillariophyta). *European Journal of Phycology* **40**:89–103 DOI 10.1080/09670260500050026.

Breman L. 2001. Random forest. *Machine Learning* **45**:5–32 DOI 10.1023/A:1010933404324.

Buczkó K, Wojtal AZ, Jahn R. 2009. *Kobayasiella* species of the Carpathian region: morphology, taxonomy and description of *K tintinnus* spec. nov. *Diatom Research* **24**:1–21 DOI 10.1080/0269249X.2009.9705780.

Buf H du, Bayer MM. 2002. *Automatic diatom identification*. World Scientific.

CEN. 2003. *Water quality: guidance standard for the routine sampling and pretreatment of benthic diatoms from rivers. EN 13946: 2003*. Geneva: Comité Européen de Normalisation.

CEN. 2004. *Water quality: guidance standard for the identification, enumeration and interpretation of benthic diatom samples from running waters. EN 14407: 2004*. Geneva: Comité Européen de Normalisation.

Conty A. 2007. El bucle microbiano en las lagunas someras esteparias de Castilla y León, importancia ecológica e influencia en la eutrofización [The microbial loop in the shallow steppe lagoons of Castilla y León, ecological importance and influence on eutrophication]. PhD Thesis, University of Leon, Leon.

Cortese G, Gersonde R. 2007. Morphometric variability in the diatom *Fragilariopsis kerguelensis:* implications for Southern Ocean paleoceanography. *Earth and Planetary Science Letters* **257**:526–544 DOI 10.1016/j.epsl.2007.03.021.

Cox EJ. 2010. Morphogenetic information and the selection of taxonomic characters for raphid diatom systematics. *Plant Ecology and Evolution* **143**:271–277 DOI 10.5091/plecevo.2010.403.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM Algorithm. *Journal of the Royal Statistical Society Series B (Methodological)* **39**:1–38.

Dimitrovski I, Kocev D, Loskovska S, Džeroski S. 2012. Hierarchical classification of diatom images using ensembles of predictive clustering trees. *Ecological Informatics* **7**:19–29 DOI 10.1016/j.ecoinf.2011.09.001.

Felde VA, Bjune AE, Grytnes J-A, Birks HJB. 2014. A comparison of novel and traditional numerical methods for the analysis of modern pollen assemblages from major vegetation–landform types. *Review of Palaeobotany and Palynology* **210**:22–36 DOI 10.1016/j.revpalbo.2014.06.003.

Friedman JH. 2001. Greedy function approximation: a gradient boosting machine. *The Annals of Statistics* **29**:1189–1232.

Genkal SI. 2004. Morphological variability and taxonomy of *Diatoma tenue* Ag. (*Bacillariophyta*). *International Journal on Algae* **6**:319–330 DOI 10.1615/InterJAlgae.v6.i4.20.

Hammer Ø, Harper DAT, Ryan PD. 2001. PAST: palaeontological Statistics software package for education and data analysis. *Paleontologia Electronica* **4**:1–9.

Han Y, Virupakshappa K, Pinto E, Oruklu E. 2015. Hardware/Software co-design of a traffic sign recognition system using zynq FPGAs. *Electronics* **4**:1062–1089 DOI 10.3390/electronics4041062.

Hartigan JA. 1975. *Clustering algorithms*. Hoboken: John Wiley & Sons.

Hofmann G, Werum M, Lange-Bertalot H. 2011. *Diatomeen im Süsswasser-Benthos von Mitteleuropa: Bestimmungsflora Kieselalgen für die ökologische Praxis: über 700 der häufigsten Arten und ihre Ökologie*. ARG Gantner.

Hongzong S, Tao W, Xiaojun Y, Huanxiang L, Zhide H, Mancang L, BoTao F. 2007. Support vector machines classification for discriminating coronary heart disease patients from non-coronary heart disease. *West Indian Medical Journal* **56**:451–457.

Julius ML, Estabrook GF, Edlund MB, Stoermer EF. 1998. Recognition of taxonomically significant clusters near the species level, using computationally intense methods, with examples from the *Stephanodiscus niagarae* complex (Bacillariophyceae). *Journal of Phycology* **33**:1049–1054 DOI 10.1111/j.0022-3646.1997.01049.x.

Kass GV. 1980. An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society Series C (Applied Statistics)* **29**:119–127 DOI 10.2307/2986296.

Kingston J, Pappas JL. 2009. Quantitative shape analysis as a diagnostic and prescriptive tool in determining *Fragilariforma* (Bacillariophyta) taxon status. *Nova Hedwigia Beihefte* **135**:103–119.

Kloster M, Kauer G, Beszteri B. 2014. SHERPA: an image segmentation and outline feature extraction tool for diatoms and other objects. *BMC Bioinformatics* **15**:218 DOI 10.1186/1471-2105-15-218.

Kulichová J, Fialová M. 2016. Correspondence between morphology and ecology: morphological variation of the *Frustulia crassinervia-saxonica* species complex (bacillariophyta) reflects the ombro-minerotrophic gradient. *Cryptogamie, Algologie* **37**:15–28 DOI 10.7872/crya/v37.iss1.2016.15.

Liu Z, Zhang Q-M, Lü L, Zhou T. 2011. Link prediction in complex networks: a local naïve Bayes model. *Europhysics Letters* **96**:48007 DOI 10.1209/0295-5075/96/48007.

Mann DG, Droop SJM. 1996. Biodiversity, biogeography and conservation of diatoms. *Hydrobiologia* **336**:19–32 DOI 10.1007/BF00010816.

Mou D, Stoermer EF. 1992. Separating *Tabellaria* (bacillariophyceae) shape groups based on fourier descriptors. *Journal of Phycology* **28**:386–395 DOI 10.1111/j.0022-3646.1992.00386.x.

Murdoch D, Chow ED. 2013. Ellipse: functions for drawing ellipses and ellipse-like confidence regions. *Available at https://cran.r-project.org/package=ellipse*.

Pappas JL, Stoermer EF. 2003. Morphometric comparison of the neotype of *Asterionella formosa* Hassall (Heterokontophyta, Bacillariophyceae) with *Asterionella edlundii* sp. nov. from Lake Hovsgol, Mongolia. *Diatom* **19**:55–65.

Paull TM, Hamilton PB, Gajewski K, LeBlanc M. 2008. Numerical analysis of small Arctic diatoms (Bacillariophyceae) representing the *Staurosira* and *Staurosirella* species complexes. *Phycologia* **47**:213–224 DOI 10.2216/07-17.1.

Potapova M, Hamilton PB. 2007. Morphological and ecological variation within the *Achnanthidium minutissimum* (Bacillariophyceae) species complex. *Journal of Phycology* **43**:561–575 DOI 10.1111/j.1529-8817.2007.00332.x.

R Core Team. 2016. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. *Available at https://r-project.org*.

Reichardt E. 2015. *Gomphonema gracile* Ehrenberg sensu stricto et sensu auct. (Bacillariophyceae): a taxonomic revision. *Nova Hedwigia* **101**:367–393 DOI 10.1127/nova_hedwigia/2015/0275.

RStudio Team. 2015. RStudio: integrated development environment for R. Boston: RStudio, Inc. *Available at https://www.rstudio.com/*.

Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B, Tinevez J-Y, White DJ, Hartenstein V, Eliceiri K, Tomancak P, Cardona A. 2012. Fiji: an open-source platform for biological-image analysis. *Nature Methods* **9**:676–682 DOI 10.1038/nmeth.2019.

Schmid A-MM. 1994. Aspects of morphogenesis and function of diatom cell walls with implications for taxonomy. In: *The protistan cell surface*. Springer, 43–60.

Spaulding SA, Jewson DH, Bixby RJ, Nelson H, McKnight DM. 2012. Automated measurement of diatom size. *Limnology Oceanography: Methods* **10**:882–890 DOI 10.4319/lom.2012.10.882.

Stoermer EF, Smol JP. 2001. *The diatoms: applications for the environmental and earth sciences*. Cambridge University Press.

Theriot E. 1988. An empirically based model of variation in rotational elements in centric diatoms with comments on ratios in phycology. *Journal of Phycology* **24**:400–407.

Vasselon V, Rimet F, Tapolczai K, Bouchez A. 2017. Assessing ecological status with diatoms DNA metabarcoding: scaling-up on a WFD monitoring network (Mayotte island, France). *Ecological Indicators* **82**:1–12 DOI 10.1016/j.ecolind.2017.06.024.

Wing MKCJ, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, Mayer Z, Kenkel BRC, Benesty M, Lescarbeau R. Ziem A, Scrucca L, Tang Y, Candan C, Hunt T. 2016. Caret: classification and regression training. *Available at https://CRAN.R-project.org/package=caret*.

Witten IH, Frank E. 2005. *Data mining: practical machine learning tools and techniques*. Second Edition. Burlington: Morgan Kaufmann.