Limnetica, 40(1): 103-116 (2021). DOI: 10.23818/limn.40.08
 ISSN: 0213-8409

Is a rare diatom relevant for Brazilian reservoirs?

Gisele C. Marquardt[1,*], Saúl Blanco[2] and Carlos E. de M. Bicudo[1]

[1] Instituto de Botânica, Department of Ecology, Av. Miguel Estéfano 3687, 04301-012 São Paulo, SP, Brazil.
[2] Institute of the Environment, La Serna, 58, E-24007 León, Spain.

[*] Corresponding author: giselecmarquardt@gmail.com

Received: 13/08/19 Accepted: 28/05/20

ABSTRACT

Is a rare diatom relevant for Brazilian reservoirs?

Planktonic diatom data sets were compared within different rarity categories to verify their responses using the weighted average (WA) approach. Our hypothesis is to proportionally reach an increased performance for the WA model by reducing the taxa weight according to their categories considering that the use of various cut-off criteria may affect the predictive abilities of different models. The underlying assumption is that WA models are unable to characterize optima and tolerances for low-occurrence taxa; in addition, their overriding may improve the overall model performance. Therefore, we developed forty diatom-training sets for six reservoirs located at two different basins in the Southwest São Paulo, Brazil. 339 diatom taxa were identified and built different models based on their relative abundance values, occurrence frequency, and the data set with no species deletion. The optimum and tolerance per taxon through the WA formula (Zelinka-Marvan weighted averages) was estimated according to their abundance values, in addition to using the pH in the data sets to infer the environmental conditions based on the sample taxonomic composition. First, the procedure using the complete dataset was repeated and subsequently with the down weighting taxa according to their rarity categories. The following procedures were advanced: comparison of predicted and measured pH values via regression analysis, and estimation of the species deletion effects on the predictive ability of the different models in terms of the coefficient of determination values (r^2) of the response curves. In contrast to what is expected, since r^2 values down weighting rare taxa had negative effect on the transference functions, data sets manipulation had significant influence on WA models performance. However, deleting non-abundant taxa had a positive effect on *p* values, thus providing robust reconstructive relationship. This work contributes to an improved understanding on diatom ecology, especially in tropical reservoirs, supporting the development of a diatom biological monitoring protocol for the study area.

Key words: biological indicator, Brazil, ecological optima, ecological tolerance diatoms, Gaussian response curve, pH

RESUMO

Uma diatomácea rara é relevante para os reservatórios brasileiros?

Comparamos conjuntos de dados de diatomáceas planctônicas com diferentes categorias de raridade para verificar suas respostas usando a abordagem da média ponderada (WA). Partimos da hipótese que ao ponderar os táxons de acordo com suas categorias de raridade, um aumento aproximadamente proporcional no desempenho do modelo de WA pode ser obtido. A suposição subjacente é que os modelos de WA são incapazes de caracterizar ótimos e tolerâncias para táxons com baixas ocorrências, e que o desempenho geral do modelo pode ser aprimorado substituindo-os. Para isso, quarenta 'training sets' de diatomáceas foram desenvolvidos para seis reservatórios localizados em duas bacias distintas no sudoeste de São Paulo, Brasil. 339 táxons de diatomáceas foram identificados e diferentes modelos foram construídos com base em suas abundâncias relativas, frequências de ocorrência e o conjunto de dados sem deleção de espécies. Além disso, o ótimo e a tolerância de cada táxon foram estimados com a fórmula WA (médias ponderadas de Zelinka-Marvan) de acordo com sua abundância nas amostras e o pH nos conjuntos de dados utilizados para inferir as condições ambientais com base na composição taxonômica da amostra. Primeiro, repetimos o procedimento com o conjunto de dados completo e depois com os táxons ponderados de acordo com suas categorias de raridade. Os valores de pH previstos e medidos foram comparados através de análise de regressão. Os efeitos da deleção de espécies na capacidade preditiva dos diferentes modelos foram estimados em termos dos valores do coeficiente de determinação (r^2) das curvas de resposta. Manipulação de diferentes conjuntos de dados influenciou

significativamente o desempenho dos modelos WA. Contrário às expectativas, tendo em conta os valores r^2, a ponderação dos táxons raros afetou negativamente as funções de transferência. No entanto, a exclusão de táxons não abundantes teve um efeito positivo nos valores de p*, propiciando uma relação reconstrutiva robusta. O trabalho contribui para uma melhor compreensão da ecologia de diatomáceas, especialmente as de reservatórios tropicais, apoiando o desenvolvimento de um protocolo de monitoramento biológico de diatomáceas para a área de estudo.*

Palavras chave: *Brasil, curva gaussiana, diatomáceas, indicador biológico, ótimo ecológico, pH, tolerância ecológica*

INTRODUCTION

Freshwater ecosystems support unique and complex ecological communities and have a critical role as a resource for humans. For this reason, ecologists are often asked to assess or monitor the "health", "status" or "condition" of these ecosystems (Bailey *et al.*, 2004).

The literature has well-documented reports of algae application in environmental assessment in aquatic habitats, particularly lakes and streams (Dixit & Smol, 1994; Stevenson *et al.*, 1999; Weilhoefer & Pan, 2006). Among algae, diatoms are worldwide used over the past 50 years in water quality monitoring (Round, 1991), for representing good indicators of water conditions because of their short life cycle and narrow ecological tolerance of many taxa (Dixit *et al.*, 1992; Charles & Smol, 1994). Furthermore, their siliceous frustules are usually easily preserved in sediments, and can provide valuable information on past environments (Smol & Glew, 1992; Moser *et al.*, 1996).

Water quality indices based on diatoms are considered to provide more precise data than chemical and zoological assessment methods (Leclercq, 1988; Omar, 2010). However, the choice of a bioindicator group must meet certain criteria, such as the unambiguous identification of each taxon (Cox, 1991; Céspedes-Vargas *et al.*, 2016). Usually, calculations of diatom indicators are based on a weighted-averaging (WA), a model that considers the relative abundances of all taxa in a sample of the site, and the auto ecological parameters of the taxa (Stevenson *et al.*, 1999). Such parameters can be used to predict values of any given environmental variable based on species composition simply by averaging the indicator values of species that are present (Ellenberg, 1979; Ter Braak & Looman, 1986).

However, the excellence of WA-based estimations depends on (1) the shape of the response curves, (2) the definition of each indicator value, and (3) the distribution of the indicator values along the environmental variable (Ter Braak *et al.*, 1986). Nevertheless, the WA approach is mathematically considered very simple and easy to understand. It is also worth mentioning that its quality has proven similar or even superior in relation to other commonly applied methods, and has been used in routine for environmental paleolimnological reconstructions (Hämäläinen, 2000). For example, for rare species (species with low maximum probability of occurrence and/or narrow tolerance), WA proved nearly as efficient as the Gaussian logistic regression (GLR, a form of the generalized linear model that fits a Gaussian-like species response curve to presence-absence data) in most scenarios (Ter Braak *et al.*, 1986).

Despite of representing the vast majority of the species in an assemblage (Gaston, 1994; Kunin & Gaston 1997; Marchant *et al.*, 1997; Hessen & Walseng, 2008; Alahuhta *et al.*, 2014; Gillet *et al.*, 2011; Mouillot *et al.*, 2013), rare species are frequently neglected in statistical analyses, which set relative abundance or occurrence criteria before applying a species-based transfer function for the uncertainty of their optimal (Bellen *et al.*, 2017). Recently, studies linking rarity to bioassessment have demonstrated that the number of diatoms in some rarity categories can be useful indicators of human disturbance in streams and rivers, especially in mountain eco regions (e.g. Potapova & Charles 2004; Gillet *et al.*, 2011).

Freshwater diatoms are very well studied and largely regarded as a good bio-indicator for water quality assessment due to their high diversity, rapid turnover, and sensitivity to numerous envi-

ronmental conditions (e.g. pH, organic and inorganic pollution) (Prygiel & Coste, 1993; van Dam *et al.*, 1994; Foets *et al.*, 2020), and a variety of indices have been developed for this purpose (e.g. Descy, 1979; Coste *in* Cemagref, 1982; Sládeček, 1986; Coste & Ayphassorho, 1991; Lenoir & Coste, 1996; Kelly & Whitton, 1995) (Wu & Kow, 2002). These indices were derived and mainly applied in temperate regions, and there is little information concerning their applicability in the tropics and subtropics (e.g. Wu, 1999; Wu & Kow, 2002; Taylor *et al.*, 2007; Bellinger *et al.*, 2006). Besides, the use of diatoms as indicators of water quality changes has few precedents in South America (e.g. Gómez, 1998; 1999; Gómez & Licursi, 2001).

In Brazil, most studies on diatoms as a bioassessment tool were carried out in the southern region of the country (e.g. Torgan & Aguiar, 1974; Lobo *et al.*, 2004a; 2004b; 2004c; 2004d; Lobo *et al.*, 2006; Hermany *et al.*, 2006; Düpont *et al.*, 2007; Salomoni *et al.*, 2006). Most of them have been carried out in lotic systems and just a few in reservoirs. In the state São Paulo, Bere & Tundisi (2010) applied the WA regression and calibration of benthic diatom assemblages to assess the importance of conductivity and pH in the structuring of benthic diatom communities in streams influenced by urban pollution (São Carlos city, São Paulo state). Specifically for Brazilian reservoirs, diatom research is mostly linked to the AcquaSed project (Base line diagnosis and reconstruction of anthropogenic impacts in the Guarapiranga Reservoir, focusing on water supply sustainability and water quality management in reservoirs of the Upper Tietê and surrounding basins (e.g. Zorzal-Almeida *et al.*, 2017), aiming at further creating an inferential model based on quantitative distribution of diatom species in water and recent sediments. Brazilian reservoirs have a predominant ecological, economic and social role in this regard, and it is essential to conduct integrated studies on such artificial ecosystems, as well as their management perspective (Henry & Nogueira, 1999).

Our study compares diatom datasets within different rarity categories in terms of relative abundance and occurrence frequency values to verify a WA calibration performance of the water pH, whose concentrations are considered a limiting factor for the colonization of aquatic ecosystems by different organisms (Esteves, 2011). This environmental variable is often a major factor influencing the species composition of freshwater diatom assemblages (Round, 1964; Battarbee, 1980; Findlay & Shearer, 1992). In addition, a strong relationship with diatom distribution was demonstrated (e.g. Birks *et al.*, 1990; Dixit *et al.*, 1992; Weckström *et al.*, 1997). Our study also shows a strongly significant correlation of this environmental parameter during stepwise selection in a CCA of diatom assemblages with environmental variables in the study area (Marquardt *et al.*, 2018), and was considered a relevant variable from our dataset, with a relatively long gradient. We used a training set composed of 40 phytoplankton diatom samples from six reservoirs located in Southwest São Paulo (Brazil). We used the correlation (r^2) values of the observed-expected values resulting in the different models tested to assess their relative precision. We expected to obtain more accurate models by down weighting rare taxa in the WA formula (Zelinka-Marvan weighted averages). Since the choice of various cut-off criteria may affect the predictive abilities of different models (Wilson *et al.*, 1996), our hypothesis was that down weighting taxa according to their rarity categories would proportionally increase the WA model performance (e.g. increases in r^2). The underlying assumption is that WA models are unable to characterize optimum and tolerance values for low-occurrence taxa, in addition to an improvement in the overall model performance by overriding them (Payne *et al.*, 2006).

MATERIAL AND METHODS

Study area and field work

The six reservoirs studied are located in two different basins: Ribeira do Iguape/Litoral Sul and Alto Paranapanema. We selected Lamparelli's (2004) Trophic State Index (TSI) based on chlorophyll-*a* (Chl-*a*) and total phosphorous (TP) values as a quantitative measure of the reservoirs trophic state. According to the TSI and current measurements, reservoirs were considered mostly

oligotrophic and mesotrophic (Table 1). Phytoplankton and water were sampled during the austral summer and winter in 2014 with a van Dorn water sampler along the vertical profile from 20 sampling sites distributed along the reservoirs. We measured the pH environmental parameter data for the training set in the field concurrently with the phytoplankton sampling using a multiparameter probe (Horiba U-53). TP analysis followed Standard Methods (APHA, 2005). Chl-*a* corrected for phaeophytin was extracted by using 90 % ethanol (Sartory & Grobbelaar, 1984) (Table 1). Details of the study area and further information on the limnological variables are available in Marquardt *et al.* (2017, 2018).

Diatom sample preparation and analysis: The preparation of diatom samples followed Battarbee *et al.* (2001) over a procedure involving hot digestion with hydrogen peroxide (H_2O_2) and hydrochloric acid (HCl) (37 %). Through a series of dilutions, peroxide and the acid were removed. Subsequently, samples were dried on cover glass, mounted in Naphrax (R.I. = 1.74), and examined on a Zeiss Axio Imager A2 light microscope equipped with DIC and a digital camera AxioCamMR5. A total of 400 diatom valves were counted along random transects at 1000× magnification (Battarbee, 1986) and a minimum sampling efficiency of 90 % (Pappas & Stoermer, 1996). Species abundances were calculated and expressed as a percentage of the total diatom counts per sample. Taxa were identified to the lowest taxonomic level possible based on diatom checklists, specific manuscripts, iconograph (e.g. Krammer, 2000; Metzeltin *et al.*, 2005; Lange-Bertalot *et al.*, 2011), and the on-line catalogue of valid names (California Academy of Sciences site, 2011). Frequent meetings and discussions involving invited diatom taxonomy experts enabled a high level of agreement regarding diatom identification.

Weighted averaging regression and calibration: Diatom taxa derived from data collected as part of the AcquaSed project based on 40 samples referred here as training set.

To assess the effect of excluding rare taxa in WA models, we developed several rarity aggregation scenarios ranging from the complete data set (all species), without any deletion criterion, to three rarity categories established according to the relative abundance values (≥ 1 %, ≥ 2 % or ≥ 5 % of the whole dataset); in addition to other three categories established according to their occurrence frequency values (≥ 1 %, ≥ 2 % or ≥ 5 % of the samples).

Subsequently, we calculated optimal and species tolerance values for pH using the WA approach based on Gaussian response curves of the taxa (Ter Braak & van Dam, 1989), in which each environmental variable value is the weighed/weighed value of each environmental variable based on their abundance in the samples and the pH in the dataset (Lepš & Šmilauer, 2003) (regression step). Although alternative unimodal response curves could be fitted, a Gaussian model represents a compromise between ecological realism and simplicity (Ter Braak & van Dam, 1989; Holden *et al.*, 2008).

To facilitate more direct comparisons of the taxa tolerance to the reservoir pH, tolerance ranges were rescaled within a 0-1 range; these values representing the broadest and narrowest tolerance values, respectively, observed within the dataset:

$$S_{tol} = x - \min / \max - \min$$

Subsequently, we used the computed taxon auto ecological parameters to back-calculate pH values based on the taxonomic composition of the sample with the Zelinka-Marvan WA formula (for each sampling station; calibration step):

$$pH = \sum A.S.V / \sum A.V$$

Where the pH index value corresponds to the mean value of the optimum (S) weighted through abundance (A) and ecological tolerance (V).

Firstly, we repeated the procedure using the complete dataset and later with down weighting taxa according to their rarity categories (see above). After generating all the predicted pH values, we compared them with the measured pH values via regression analysis. The effects of species deletions on the different models predictive ability were assessed in terms of squared correlation (r^2) values of the observed-expected values.

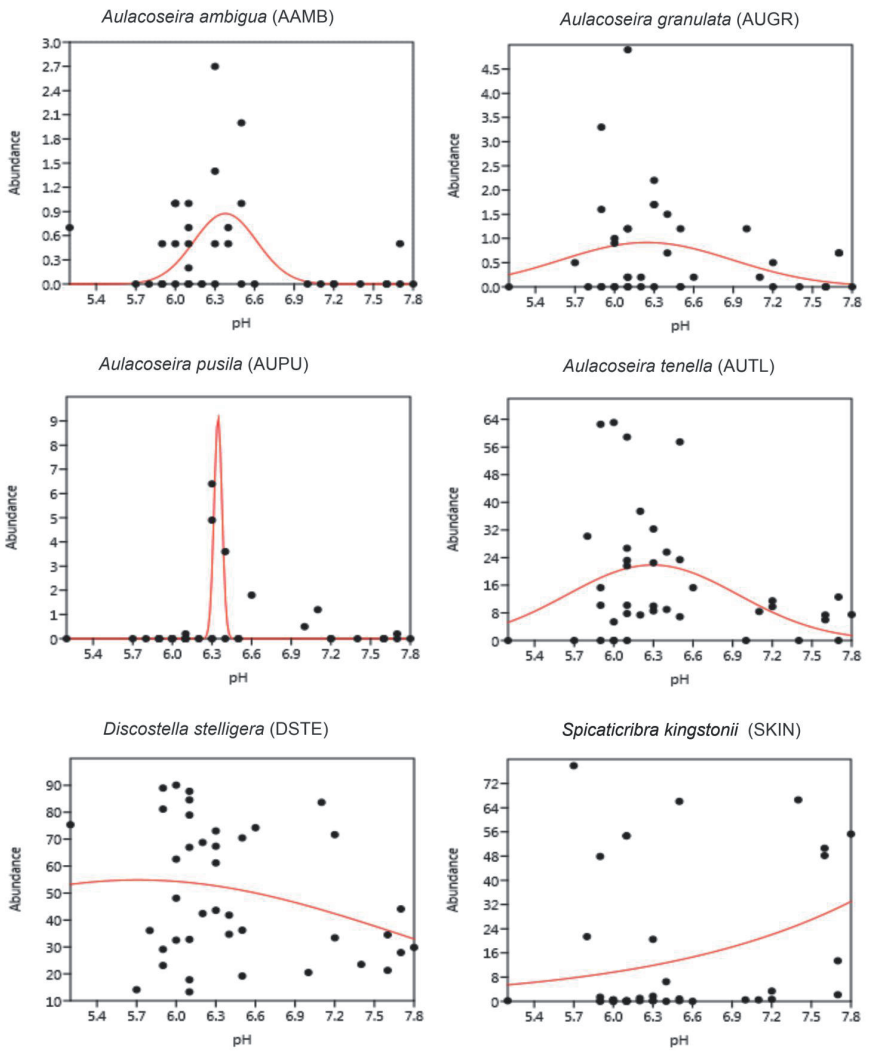

Figure 1. Probability of occurrence of some contrasting species in relation to pH in the six studied reservoirs, as fitted with logistic regression. The curves can be identified by the code near their optimum indicated by dotted lines. The species arranged in order of their optima are: *Aulacoseira ambigua* (AAMB); *A. granulata* (AUGR); *A. pusila* (AUPU); *A. tenella* (AUTL); *Discostella stelligera* (DSTE) e *Spicaticribra kingstonii* (SKIN). *Probabilidade de ocorrência de algumas espécies contrastantes em relação ao pH nos seis reservatórios estudados, adaptados à regressão logística. As curvas podem ser identificadas pelo código próximo ao seu ótimo ideal indicado por linhas pontilhadas.*

All analyses were implemented using PAST version 3.14 (Hammer *et al.*, 2001).

RESULTS AND DISCUSSION

Diatom assemblage's composition

A total 339 species were identified representing 51 diatom genera. Samples were dominated by centric taxa, occurring in almost all sampling sites, especially *Discostella stelligera* (Cleve & Grunow) Houk & Klee, *Aulacoseira tenella* (Nygaard) Simonsen and *Spicaticribra kingstonii* J.R. Johansen, Kociolek & R.L. Lowe. Within the study area, their abundance was correlated with the Secchi disk transparency vector, considered as indicators of oligotrophic conditions (Marquardt *et al.*, 2018). In contrast, *Aulacoseira*

ambigua (AAMB), *A. granulata* (AUGR) and *A. pusila* (AUPU) correlated to the total nitrogen (NT) vector and lowest conductivity and Secchi values (Marquardt *et al.*, 2018).

Species response curves

Most sites were acidic to slightly alkaline with pH values ranging between 5.2 and 7.8 (see table 1). We found that most diatoms exhibited symmetrical, unimodal (bell-shaped) response curves against this variable (Fig. 1), which makes the WA a reliable model to infer pH values.

Considering relative abundance values, the number of taxa per model declined from 339 in the complete species data set to 56 taxa after removing those with less than 1 % relative abundance, remaining 36 taxa with relative abundance over 2 % and just 16 taxa with relative abundance ≥ 5 %. For the occurrence frequency values, the numbers of taxa per model were of 61, 50 and 12, respectively, for selective ranges of ≥ 1 %, ≥ 2 % and ≥ 5 %.

The regression models were marginally significant ($p \leq 0.07$), except for ≥ 2 % ($p = 0.11$) and ≥ 5 % ($p = 0.11$) datasets based on relative abundance, and for the model ≥ 1 % based on occurrence frequency ($p = 0.23$) (Table 2). In contrast, removal of rare taxa from the 40 diatom training set sites according to their occurrence frequency values improved the performance of the WA models significantly, whereas the *p* values decreased from 0.23 (≥ 1 %) to 0.07 (≥ 2 %, ≥ 5 %) (Table 2). However, the model retaining all taxa had the lowest *p* (0.02), and the highest r^2 (0.12) values (Fig. 2A-G; table 2).

Table 1. Means and standard deviation of abiotic variables in the six studied reservoirs. Abbreviations: Secchi (Secchi disk transparency depth), pH, TP (total phosphorus), Chl-*a* (chlorophyll-*a*), TSI (Trophic State Index). FR (Cachoeira do França), FU (Cachoeira da Fumaça), SE (Serraria), JP (Jurupará), SI (Salto do Iporanga), PI (Paineiras). Numbers refer to sample units. *Médias e desvio padrão das variáveis abióticas nos seis reservatórios estudados. Abreviações: Secchi (profundidade de transparência do disco Secchi), pH, TP (fósforo total), Chl-*a *(clorofila-*a*), TSI (Índice de Estado Trófico). FR (Cachoeira do França), FU (Cachoeira da Fumaça), SE (Serraria), JP (Jurupará), SI (Salto do Iporanga), PI (Paineiras). Os números referem-se as unidades amostrais.*

	Secchi	pH	TP	Chl-*a*	TSI (Anual Mean)
FR1	1.7±8.5	7.3±0.9	5.5±2.2	5.4±4.8	Oligotrophic
FR2	1.9±0.2	7.3±1.3	10.0±7.5	4.12±2	Oligotrophic
FR3	1.7±0.005	7.9±0.6	6.6±3.7	20.2±26.5	Oligotrophic
FR4	1.7±0.1	8.6±0.6	6.8±2.5	22.8±29.7	Oligotrophic
FU1	2.6±1.7	6.5±0	5.1±1.6	1.7±0.4	Ultraoligotrophic
FU2	2.8±1.5	6.8±0.7	8.4±2.1	3.4±1.1	Oligotrophic
FU3	2.1±0.3	6.6±0.4	5.8±2.7	3.3±2.3	Oligotrophic
SE1	4.0±1.8	7.4±1.1	8.1±2.9	1.8±0.7	Oligotrophic
SE2	3.7±1.4	7.3±0.6	8.0±1.5	4.6±2.2	Oligotrophic
SE3	4.2±1.4	6.6±0.2	8.5±3.7	7.6±0.7	Oligotrophic
JP1	1.5±0.8	6.4±1.3	19.5±4.2	16.8±1.8	Mesotrophic
JP2	1.9±0.4	6.6±0.1	16.6±0.3	10.1±0.7	Mesotrophic
JP3	2.2±0.2	6.6±0.1	13.5±0.7	7.1±1.2	Mesotrophic
JP4	2.1±0.4	6.8±0.2	13.3±1.9	6.6±3.7	Oligotrophic
SI1	1.5±0.2	8.0±2.1	32.1±3.7	36.7±37.6	Mesotrophic
SI2	1.9±0.4	7.8±2	23.1±1.5	17.4±15.3	Mesotrophic
SI3	1.7±0.6	7.9±2.1	25.6±3.8	36.8±44	Mesotrophic
PI1	0.7±0.7	6.7±0.4	20.3±0.3	9.8±5.7	Mesotrophic
PI2	1.3±0.5	6.9±0.6	16.2±0.4	6.9±2.7	Mesotrophic
PI3	1.3±0.3	7.1±0.5	16.3±2	4.8±3.1	Oligotrophic

Table 2. Comparative performance among predictive models data sets based on relative abundance and occurrence frequency produced by regression analysis of the predicted and measured pH values. (V = ecological tolerance). Percentages refers to the rarity aggregation scenarios based on the data set with no species deletions (All species), relative abundances (≥ 1 %, ≥ 2 %, ≥ 5 %) and occurrence frequencies (≥ 1 %, ≥ 2 %, ≥ 5 %). *Desempenho comparativo entre os conjuntos de dados de modelos preditivos com base na abundância relativa e na frequência de ocorrência produzida pela análise de regressão dos valores de pH previstos e medidos. (V = tolerância ecológica). Porcentagens se referem aos cenários de agregação de raridade com base no conjunto de dados sem deleções de espécies (All species), abundâncias relativas (≥ 1 %, ≥ 2 %, ≥ 5 %) e frequências de ocorrência (≥ 1 %, ≥ 2 %, ≥ 5 %).*

		Relative abundance			Occurrence frequency		
Number of species	339	56	36	16	61	50	12
Correlation	All species	≥ 1 %	≥ 2 %	≥ 5 %	≥ 1 %	≥ 2 %	≥ 5 %
R	0.35	0.29	0.23	0.24	0.19	0.28	0.28
r^2	0.12	0.08	0.05	0.06	0.03	0.08	0.08
RMSE	8.26	5.65	5.2	4.97	3.41	7.88	7.94
V	2.37	1.87	1.50	1.59	1.19	1.84	1.83
p (uncorr.)	0.02	0.06	0.13	0.11	0.23	0.07	0.07
Permutation p	0.02	0.06	0.13	0.11	0.23	0.07	0.07

DISCUSSION

Recently, gradient analytical weighted averaging (WA) regression and calibration modeling (and related techniques) have been developed and successfully applied to historical monitoring of lakes, or used to infer past environmental conditions from the remains of different organisms (Hämäläinen, 2000). In this regard, it is usual to treat data by excluding species occurring for a low number of samples assuming the model inability to characterize the optima and tolerance of low-occurrence species, in addition to a possible improvement of the overall model performance by eliminating them (Payne *et al.*, 2006).

Singular observations (singletons, taxa occurring in only one sample) often occur in ecological series. In nature, singletons result from random fluctuations, migrations or local changes in external forcing. In an aquatic system studied at a fixed location such changes may be derived from temporary movements of water masses. Singletons may also result from improper sampling or inadequate preservation of specimens (Legendre & Legendre, 1998).

It has been observed that taxa deletion in chironomid-based inference models substantially improved the predictive ability of inference models (measured as RMSEP, Martens & Naes, 1989). In this context, the common practice of including taxa with only ≥ 2 % abundance in at least two lakes was one of the deletion criteria that much improved inference models. Similar deletion criteria, such as ≥ 2 % in at least three lakes and ≥ 3 % in at least one lake, produced comparable improvements (≤ 18 % reduction in RMSEP) (Quinlan & Smol, 2001).

Similarly, Payne *et al.* (2006) developed transfer-function models based on different techniques, including weighted averaging, to investigate testate amoebae ecology in southern Alaska. Results showed that the model performance was improved from the selective exclusion of taxa. In relation to previous studies, the relatively poor performance of the model can be explained by the limitations of one-off water-table measurements, the very large environmental gradients covered, and recent climate change in the study area.

Our findings partially disagree with the above-mentioned studies. We found an "all taxa" dataset *p*-value that is below other cut-offs, suggesting the best performance for this model. Therefore, removing rare taxa proved counter-productive and the transfer function models developed from removing rare taxa actually reduced the model performance. This result corroborates those of Birks (1994) and Wilson *et*

al. (1996), in which diatom and pollen pH calibration datasets, as well as in other data sets, the lowest prediction (measured in terms of RMSEboot) always occurs in WA regression and calibration before deleting the taxa on the basis of their effective number of occurrence. Birks (1994) also emphasized that the largest prediction errors occur when only the commonest and

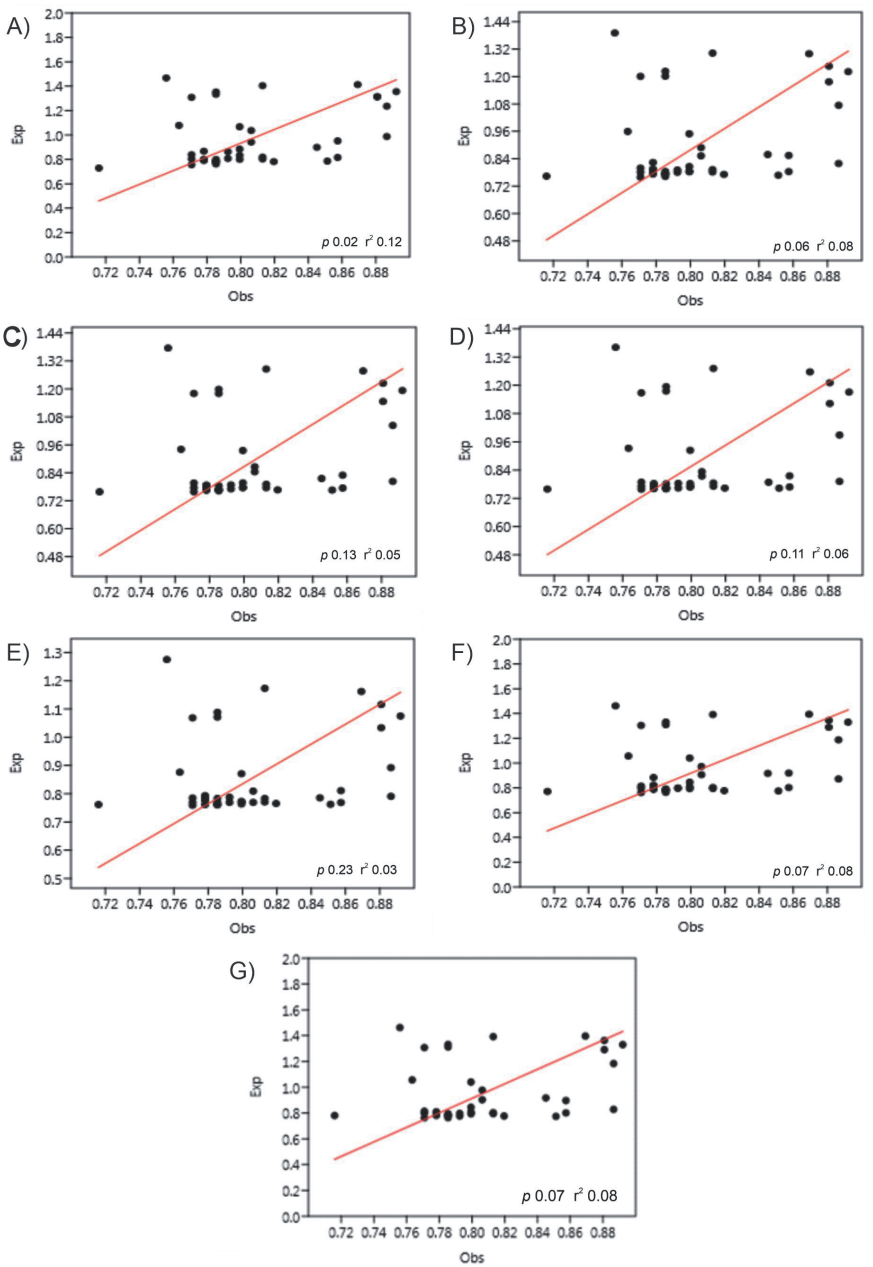

Figure 2. Observed *vs.* expected pH in the different data sets based on the data set with no species deletions (A: all species), relative abundances (B: ≥ 1 %, C: ≥ 2 %, D: ≥ 5 %) and occurrence frequencies (E: ≥ 1 %, F: ≥ 2 %, G: ≥ 5 %). *Modelos de regressão do pH observado* vs. *esperado baseado no conjunto de dados sem deleções de espécies (A: all species), abundância relativa (B: ≥ 1 %, C: ≥ 2 %, D: ≥ 5 %) e frequência de ocorrência (E: ≥ 1 %, F: ≥ 2 %, G: ≥ 5 %).*

numerically most abundant taxa are included in the WA regression and calibration. According to Wilson *et al.* (1994), the effects of species deletions and dataset size on the predictive ability of the models emphasizes the value of training sets with a large number of taxa to develop transfer functions with robust and reliable estimates of species optimum and tolerance values.

Presumably, even the estimated WA optima of very rare taxa (their absence is ignored in WA regression) contribute to some ecological "signals" to the calibration, rather than, as might be expected, having no effect or even having deleterious effects by introducing "noise" into calibration (Birks, 1994).

In contrast, a transfer function training set presents a tendency of nearby sampling locations to floristically resemble one another, more than randomly selected sites with similar species assemblages and environmental conditions. This possibly results in inappropriate model choice and misleading, and in over-optimistic estimates of a transfer function performance (Telford & Birks, 2005).

Our results differ considerably between abundance-based and occurrence-based cut-off criteria. Despite the latter models led to enhanced inference performance, removal of taxa occurring in < 1 % of sampling stations did not improve the model performance. The opposite was observed for the < 1 % model on relative abundances. In this cut-off, rare taxa are probably still restricted to very few samples, and their optimal may be uncertain generating little impact on the predictive ability of the model. In contrast, cut-offs based on relative abundance values (≥ 2 % and ≥ 5 %) may have selected only widespread taxa, considered to have wide tolerance and consequently poor indicators leading to unreliable environmental inferences. The incorporation of abundance data in biological indices may bias accuracy and reduce precision in two ways, e.g. numerically dominant taxa can skew the result in the direction of their indicator scores. Additionally, presence/absence data or strongly transformed abundance values can skew the result in favor of rare taxa by attributing them with weight equal to abundant taxa (Monaghan, 2016).

Finally, it is important to observe a possible season influence that was not assessed in our data. As demonstrated in Winter & Duthie (2000), the quality of inference models built during a study with epilithic diatoms as indicators of stream nutrient concentration was better with the seasonal variation removal from the dataset through mean summer values in relation to the use of full dataset.

CONCLUSION

Our results revealed that bioassessment using the WA modeling is a powerful modeling technique for an accurate assessment of species response to both single and multiple environmental descriptors. However, manipulation of different datasets had significant influence on model performance. Therefore, removing rare taxa proved counter-productive and the transfer function models developed by removing rare taxa actually reduced model performance. Our results proved to have been significantly influenced by the sample size, however, we demonstrated that a model improvement can be reached even at such local scales. Influence of rare taxa on bioassessments still is a subject for much discussion and study, in this context, choosing a cut-off to avoid rare taxa noise is very much subjective. Our work contributes to a better understanding of diatom ecology, especially from tropical reservoirs, and supports the development of accurate biological monitoring protocols based on diatoms for this region.

ACKNOWLEDGEMENTS

This study integrates the AcquaSed project, supported from funds by FAPESP (*Fundação de Amparo à Pesquisa do Estado de São Paulo*, grant number 2009/53898-9) and GCM thesis at the *Instituto de Botânica*, São Paulo, Brazil (FAPESP fellowship number 2013/10314-2). CEMB would like to thank CNPq (*Conselho Nacional de Desenvolvimento Científico e Tecnológico*) for a Research Fellowship (number 305031/2016-3). We deeply appreciated the valuable assistance of personnel from *Votorantim Energia* for their logistic support during the fieldwork. We would also like to thank Prof.

William de Queiróz (*Universidade de Guarulhos, Laboratório de Geoprocessamento*) for the study area illustration.

REFERENCES

ALAHUHTA, J., L. B. JOHNSON, J. OLKER & J. HEINO. 2014. Species sorting determines variation in the community composition of common and rare macrophytes at various spatial extents. *Ecological Complexity*, 20: 61-68. DOI: 10.1016/j.ecocom.2014.08.003

AUSTIN, M. P. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modeling. *Ecological Modelling*, 157(2-3): 101-118. DOI: 10.1016/S0304-3800(02)00205-3

BAILEY, L. L., T. R. SIMONS & K. H. POLLOCK. 2004. Estimating site occupancy and species detection probability parameters for terrestrial salamanders. *Ecological Applications*, 14(3): 692-702. DOI: 10.1890/03-5012

BATTARBEE, R. W. 1980. Diatoms in lake sediments. In: *Paleohydrological changes in the temperate zone in the last 15,000 years, Subproject B. IGCP Project 158 lake and Mire environments*. B.E. Berglung (ed.): 177-225. University of Lund, Sweden.

BATTARBEE, R. W. 1986. Diatom analysis. In: *Handbook of Holocene Paleoecology and Paleohydrology*. B.E. Berglung (ed.): 527-570. Wiley and Sons, London. UK.

BATTARBEE, R. W., V. J. JONES, R. J. FLOWER, N. G. CAMERON, H. BENNION, L. CARVALHO & S. JUGGENS. 2001. Diatoms. In: *Tracking Environmental Change Using Lake Sediments. Vol. 3, Terrestrial, Algal, and Siliceous Indicators*. J.P. Smol, H. J.B. Birks, & W.M. Last (eds.): 155-202. Kluwer Academic Publishers, Dordrecht.

BELLINGER, B. J., C. CHRISTINE. & C. M. O'REILLY. 2006. Benthic diatoms as indicators of eutrophication in tropical streams. *Hydrobiologia*, 573: 75-87. DOI: 10.1007/s10750-006-0262-5

BERE, T. & J. G. TUNDISI. 2010. Epipsammic diatoms in streams influenced by urban pollution, São Carlos-SP, Brazil. *Brazilian Journal of Biology*, 70(4): 921-930. DOI: 10.1590/S1519-69842010000500002

BIRKS, H. J. B., J. M. LINE, S. JUGGINS., A. C. STEVENSON & C. J. F. TER BRAAK. 1990. Diatoms and pH reconstruction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 327: 263-278.

BIRKS, H. J. B. 1994. The importance of pollen and diatom taxonomic precision in quantitative palaeoenvironmental reconstructions. *Review of Palaeobotany and Palynology*, 83(1-3): 107-117. DOI: 10.1016/0034-6667(94)90062-0

CALIFORNIA ACADEMY OF SCIENCES. 2012. *Diatom Collection*. California. URL: http://research.calacademy.org/izg/research/diatom. Accessed 1 May 2017.

CÉSPEDES-VARGAS, E., G. UMAÑA-VILLALOBOS & A. M. SILVA-BENAVIDES. 2016. Tolerancia de diez especies de diatomeas (Bacillariophyceae) a lós factores físico-químicos del agua em El Río Sarapiquí, Costa Rica. *Revista de Biología Tropical*, 64: 105-115.

CHARLES, D. F. & J. P. SMOL. 1994. Long-term chemical changes in lakes: Quantitative inferences from biotic remains in the sediment record. In: *Environmental Chemistry of Lakes and Reservoirs. Advanced in Chemistry Series, no 237*. L. Baker (ed.): 3-31. American Chemical Society, Washington. DC.

COSTE, M., & H. AYPHASSORHO. 1991. Étude de la qualité des eaux du Bassin Artois-Picardie à l'aide des communautés de diatomées benthiques (Application des indices diatomiques). Rapport Cemagref Bordeaux. Agence de l'Eau Artois-Picardie, Douai, 227 pp.

COX, E. J. 1991. What is the basis for using diatoms as monitors of river quality? In: *Use of algae for monitoring rivers*. Whitton, B.A., E. Rott & G. Friedrich (eds.): 33-40. Institut für Botanik, Universitat Innsbruck, Austria.

DESCY J. P. 1979. *A new approach to water quality estimation using diatoms*. Nova Hedwigia, 64: 305-323.

DIXIT, S. S. & J. P. SMOL. 1994. Diatoms as indicators in the Environmental Monitoring and Assessment Program – Surface Waters (EMAP-SW). *Environmental Monitoring and*

Assessment. DOI: 10.1007/BF00577258.

DIXIT, S. S., J. P. SMOL, J. C. KINGSTON & D. F. CHARLES. 1992. Diatoms: powerful indicators of environmental change. *Environmental Science & Technology*, 26: 22-33. DOI: 10.1021/es00025a002

DÜPONT, A., E. A. LOBO, A. B. COSTA & M. SCHUCH. 2007. Avaliação da qualidade da água do Arroio do Couto, Santa Cruz do Sul, RS, Brasil. *Série Biologia* (UNISC), 9: 20-31.

ELLENBERG, H. 1979. Zeigerwerte der Gefäszpflanzen Mitteleuropas. Göttingen: *Scripta Geobotanica*.

ESTEVES, F. A. 2011. *Fundamentos de Limnologia*. Rio de Janeiro: Editora Interciência/FINEP. 3rd ed.

FINDLAY, D. L. & J. A. J. SHEARER. 1992. Relationships between sedimentary diatom assemblages and lakewater pH values in the Experimental Lakes Area. *Paleolimnology*, 7(2): 145-156. DOI: 10.1007/BF00196869

FOETS, J., C. E. WETZEL., A.J. TEULING., & L. PFISTER. 2020. Temporal and spatial variability of terrestrial diatoms at the catchment scale: controls on communities. *PeerJ*, 8:e8296. DOI: 10.7717/peerj.8296

GASTON, K. J. 1994. *Rarity*. Chapman and Hall. London. UK.

GILLET, N. D., Y. PAN, K. M. MANOYLOV, P. STANCHEVA & C. L. WEILHOEFER. 2011. The potential indicator value of rare taxa richness in diatom-based stream bioassessment. *Journal of Phycology*, 47(3): 471-482. DOI: 10.1111/j.1529-8817.2011.00993.x

GÓMEZ, N. 1998. Use of epipelic diatoms for evaluation of water quality in the Matanza-Riachuelo (Argentina), a pampean plain river. *Water Research*, 32: 2029-2034

GÓMEZ, N. 1999. Epipelic diatom from Matanza-Riachuelo river (Argentina), a highly polluted basin from the pampean plain: biotic indices and multivariate analysis. *Aquatic Ecosystem Health & Management*, 2: 301-309.

GÓMEZ, N. & M. LICURSI. 2001. The Pampean Diatom Index (IDP) for assessment of rivers and streams in Argentina. *Aquatic Ecology*, 35: 173-181.

HÄMÄLÄINEN, H. 2000. *Weighted averaging models in contemporary freshwater monitoring with benthic invertebrate assemblages*. Finland.

HAMMER, O, D. A. T. HARPER & P. D. RYAN. 2001. PAST: Paleontological Statistic software package for education and data analysis. *Palaeontologia Electronica*, 4: 1-9.

HENRY, R., & M. G. NOGUEIRA. 1999. A Represa de Jurumirim (São Paulo): primeira síntese sobre o conhecimento limnológico e uma proposta preliminar de manejo ambiental. In: *Ecologia de reservatórios: estrutura, função e aspectos sociais*. R. Henry (ed.): 651-685. FAPESP/FUNDIBIO, Botucatu.

HERMANY, G., E. A. LOBO, A. SCHWARZBOLD & M. A. OLIVEIRA. 2006. Ecology of the epilithic diatom community in a low-order stream system of the Guaíba hydrographical region: subsidies to the environmental monitoring of southern Brazilian aquatic system. *Acta Limnologica Brasiliensia*, 18(1): 9–27.

HESSEN, D. O. & B. WALSENG. 2008. The rarity concept and the commonness of rarity in freshwater zooplankton. *Freshwater Biology*, 53(10): 2026-2035. DOI: 10.1111/j.1365-2427.2008.02026.x

HOLDEN, P., A. MACKAY & G. SIMPSON. 2008. A Bayesian palaeoenvironmental transfer function model for acidified lakes. *Journal of Paleolimnology*, 39: 551-566. DOI: 10.1007/s10933-007-9129-7

KELLY M.G., & B. A. WHITTON. 1995. The trophic diatom index: a new index for monitoring eutrophication in rivers. *Journal of Applied Phycology*, 7: 433–333.

KRAMMER, K. 2000. The genus *Pinnularia*. In: *Diatoms of Europe*. Vol. 1. H. Lange-Bertalot (ed). Koeltz Scientific Books. Germany. 703 pp.

KUNIN, W. E. & K. J. GASTON. 1997. *The biology of rarity*. Chapman and Hall. London. UK.

LANGE-BERTALOT, H., M. BAK, A. WITKOWSKI & N. TAGLIAVENTI. 2011. *Eunotia* and some related genera. In: *Diatoms of Europe, Diatoms of the European Inland waters and comparable habitats*. H. Lange-Bertalot (ed.). Gantner Verlag KG, Ruggell. 747 p.

LECLERCQ, L. 1988. Utilization de trios

indices, chimique, diatomique et biocénotique, pour l'évaluation de la qualité de l'eau de la Joncquiere, riviére calcaire polluée par le village de Doische (Belgique, Prov. Namur). *Mémoires de la Société Royale de Botanique de Belgique*, 10: 26-34.

LEGENDRE, P. & L. LEGENDRE. 1998. *Numerical ecology*. 2nd English edition. Elsevier Scientific Publishing Company, Amsterdam, The Netherlands.

LEPŠ, J. & P. ŠMILAUER. 2003. *Multivariate analysis of ecological data using Canoco*. Cambridge University Press. Cambridge. UK.

LOBO, E. A., V. L. CALLEGARO, G. HERMANY, D. BES, C. E. WETZEL & M. A. OLIVEIRA. 2004a. Use of epilithic diatoms as bioindicators from lotic systems in southern Brazil, with special emphasis on eutrophication. *Acta Limnologica Brasiliensia*, 16(1): 25-40.

LOBO, E. A., V. L. CALLEGARO, G. N. HERMANY & L. ECTOR. 2004b. Review of the use of microalgae in South America for monitoring rivers, with special reference to diatoms. *Vie et Milieu*, 53 (1): 35-45.

LOBO, E. A., D. BES., L. TUDESQUE & L. ECTOR. 2004c. Water quality assessment of the Pardinho River, RS, Brazil, using epilithic diatom assemblages and faecal coliforms as biological indicators. *Vie et Milieu*, 53: 46-53.

LOBO, E. A., V. L. M CALLEGARO, C. E. WETZEL, G. HERMANY & D. BES. 2004d. Water quality study of Condor and Capivara streams, Porto Alegre municipal district, RS, Brazil, using epilithic diatoms biocenoses as bioindicators. *Oceanological and Hydrobiological Studies*, 33(2): 77-93.

LOBO, E. A., S. E. SALOMONI, O. ROCHA, & V. L. CALLEGARO. 2006. Epilithic diatoms as indicatores of water quality in the Gravataí river, Rio Grande do Sul, Brazil. *Hydrobiologia*, 559(1): 233-246. DOI: 10.1007/s10750-005-9012-3

MARCHANT, R., A. HIRST, R. H. NORRIS, R. BUTCHER, L. MIXZELING & D. TILLER. 1997. Classification and ordination of macroinvertebrate assemblages from running waters in Victoria, Australia. *Journal of the North American Benthological Society*, 16(3): 664-681.

MARQUARDT, G. C., C. E. M. BICUDO, T. A. V. LUDWIG, L, ECTOR & C. E. WETZEL. 2018. Diatom assemblages (Bacillariophyta) in six tropical reservoirs from southeast Brazil: species composition and spatial and temporal variation patterns. *Acta Limnologica Brasiliensia*, vol. 30, e201. DOI: 10.1590/s2179-975x6417

MARQUARDT, G. C., S. BLANCO & C. E. M. BICUDO. 2017. Distance decay as a descriptor of the diatom compositional variation in tropical reservoirs. *Marine and Freshwater Research*, 69(1): 105-113. DOI: 10.1071/MF17003

MARTENS, H. & T. NAES. 1989. *Multivariate Calibration*. John Wiley and Sons, New York.

METZELTIN, D., H. LANGE-BERTALOT & F. GARCÍA-RODRIGUEZ. 2005. Diatoms of Uruguay. In: *Iconographia Diatomologica*. H. Lange-Bertalot (ed.). Koeltz Scientific Books, Königstein. 736 pp.

MONAGHAN, K. A. 2016. Four Reasons to Question the Accuracy of a Biotic Index; the Risk of Metric Bias and the Scope to Improve Accuracy. *PLoS ONE*, 11(7): e0158383. DOI: 10.1371/journal.pone.0158383

MOSER, K. A., G. M. MACDONALD & J. P. SMOL. 1996. Applications of freshwater diatoms to geographical research. *Progress in Physical Geography*, 20(1): 21-52.

MOUILLOT, D., D. R. BELLWOOD, C. BARALOTO, J. CHAVE, R. GALZIN, M. HARMELIN-VIVIEN, M. KULBICKI, S. LAVERGNE, S. LAVOREL, N. MOUQUET, C. E. TIMOTHY PAINE, J. RENAUD. & W. THUILLER. 2013. Rare Species Support Vulnerable Functions in High-Diversity Ecosystems. *PLoS Biology*, 11(5): e1001569. DIO: 10.1371/journal.pbio.100156

OMAR, W. M. W. 2010. Perspectives on the Use of Algae as Biological Indicators for Monitoring and Protecting Aquatic Environments, with Special Reference to Malaysian Freshwater Ecosystems. *Tropical Life Sciences Research*, 21(2): 51-67.

PAPPAS, J. L. & E. F. STOERMER. 1996. Quantitative method for determining a representative algal sample count. *Journal of Phycology*, 32(4): 693-696. DOI: 10.1111/j.

0022-3646.1996.00693.x

PAYNE, R., K. KISHABA, J. BLACKFORD & E. MITCHELL. 2006. The ecology of testate amoebae in southcentral Alaskan peatlands: building transfer function models for palaeoenvironmental inference. *Holocene*, 16: 403-414.

POTAPOVA, M. & D. F. CHARLES. 2004. Potential use of rare diatoms as environmental indicators in USA rivers. In: *Proceedings of the 17th International diatom symposium*. M. Poulin (ed.): 281-295. Biopress Ltd., Bristol.

PRYGIEL, J., & M. COSTE. 1993. The assessment of water quality in the Artois-Picardie water basin (France) by the use of diatom indices. *Hydrobiologia*, 269(1): 343-349. DOI: 10.1007/BF00028033

ROUND, F. E. 1964. The diatom sequence in lake deposits: some problems of interpretation. *Verhandlungen des Internationalen Verein Limnologie*, 15: 1012-1020. DOI: 10.1080/03680770.1962.11895641

ROUND, F.E. 1991. Diatoms in river water-monitoring studies. *Journal of Applied Phycology*, 3: 129-145. DOI: 10.1007/BF00003695

SALOMONI, S. E., O. ROCHA, V. L. M. CALLEGARO & E. A. LOBO. 2006. Epilithic diatoms as indicators of water quality in the Gravataí river, Rio Grande do Sul, Brazil. *Hydrobiologia*, 559: 233-246. DOI: 10.1007/s10750-005-9012-3

SLÁDEČEK, V. 1986. Diatoms as indicators of organic pollution. *Acta Hydrochimica et Hydrobiologica*, 14: 555–566.

SMOL, J. P. & J. R. GLEW. 1992. Paleolimnology. In: *Encyclopedia of earth system sciences*. W.A. Nierenberg (ed.): 551–564. Academic Press, San Francisco.US.

STEVENSON, R. J., PAN, Y. & H. VAN DAM. 1999. Assessing environmental conditions in rivers and streams with diatoms. In: *The Diatoms: applications for the environmental and Earth Sciences*. Stoermer, E.F. & J.P. Smol (eds.): 11-40. Cambridge University Press, Cambridge. UK.

TAYLOR, J. C., J. PRYGIEL., A. VOSLOO., P. A. DE LA REY. & L.VAN RENSBURG. 2007. Can diatom-based pollution indices be used for biomonitoring in South Africa? A case study of the Crocodile West and Marico water management area. *Hydrobiologia*, 592: 455-464. DOI: 10.1007/s10750-007-0788-1

TELFORD, R. J. & H. J. B. BIRKS. 2005. The secret assumption of transfer functions. *Quaternary Science Reviews*, 24(20-21): 2173-2179. DOI: 10.1016/j.quascirev.2005.05.001

TER BRAAK, C. J. F. & H. VAN DAM. 1989. Inferring pH from diatoms: a comparison of old and new calibration methods. *Hydrobiologia*, 178(3): 209-223. DOI: 10.1007/BF00006028

TER BRAAK, J. F. & C. W. N. LOOMAN. 1986. Weighted averaging, logistic regression and the Gaussian response model. *Vegetatio*, 65: 3-11. DOI: 10.1007/BF00032121

TORGAN, L. C., L. W. AGUIAR. 1974. Nota preliminar sobre a flora diatomológica do Guaíba, RS. XXV Congresso Nacional de Botânica, Mossoró, Rio Grande do Norte. Anais da Sociedade Botânica do Brasil. Recife, SBB: 141-142.

VAN BELLEN, S., D. MAUQUOY, R. J. PAYNE, T. P. ROLAND, P. D. M. HUGHES, T. J. DALEY, N. J. LOADER, F. A. STREET-PERROTT, E. M. RICE & V. A. PANCOTTO. 2017. An alternative approach to transfer functions? Testing the performance of a functional trait-based model for testate amoebae. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 468: 173-183. DOI: 10.1016/j.palaeo.2016.12.005

VAN DAM, H., A. MERTENS., & J. SINKELDAM. 1994. A coded checklist and ecological indicator values of freshwater diatoms from the Netherlands. *Netherlands Journal of Aquatic Ecology*, 28(1):117-133. DOI: 10.1007/BF02334251

WECKSTRÖM, J., A. KORHOLA., & T. BLOM. 1997. Diatoms as quantitative indicators of pH and water temperature in subarctic Fennoscandian lakes. *Hydrobiologia*, 347(1-3): 171-184. DOI: 10.1023/A:1003091923476

WEILHOEFER, C. L. & Y. PAN. 2006. Diatom-based bioassessment in wetlands: how many samples do we need to adequately characterize the diatom assemblage in a wetland? *Wetlands*, 26(3): 793-802. DOI: 10.1672/0277-5212(2006)26[793:DBIWHM]2.0.CO;2

WINTER, J. G. & H. C. DUTHIE. 2000. Epilithic

diatoms as indicators of stream total N and total P concentration. *Journal of the North American Benthological Society*, 19(1): 32-49. DOI: 10.2307/1468280

WU, J. T. 1999. A generic index of diatom assemblages as bioindicator of pollution in the Keclung River of Taiwan. *Hydrobiologia*, 397: 79-87.

WU, J., & L. KOW. 2002. Applicability of a generic index for diatom assemblages to monitor pollution in the tropical River Tsanwun, Taiwan. *Journal of Applied Phycology*, 14: 63-69.

ZORZAL-ALMEIDA, S., L. M. BINI & D. C. BICUDO. 2017. Beta diversity of diatoms is driven by environmental heterogeneity, spatial extent and productivity. *Hydrobiologia*, 800: 7-16. DOI: 10.1007/s10750-017-3117-3