UNIVERSIDADE FEDERAL DO PARANÁ

DAIANE GRACIELI FALLER

APLICAÇÃO DE MODELOS DE INTELIGÊNCIA ARTIFICIAL NA PREDIÇÃO E ANÁLISE ESPACIAL DA MACROFAUNA BÊNTICA

PONTAL DO PARANÁ
2012

DAIANE GRACIELI FALLER

APLICAÇÃO DE MODELOS DE INTELIGÊNCIA ARTIFICIAL NA PREDIÇÃO E ANÁLISE ESPACIAL DA MACROFAUNA BÊNTICA

Dissertação apresentada ao curso de Pós-Graduação em Sistemas Costeiros e Oceânicos, Centro de Estudos do Mar, Setor de Ciências da Terra, Universidade Federal do Paraná, como requisito parcial para obtenção do grau de Mestre em Sistemas Costeiros e Oceânicos.

Orientador: Maurício Garcia de Camargo

PONTAL DO PARANÁ
2012

Faller, Daiane Gracieli
F194a Aplicação de quatro modelos de inteligência artificial na predição e
análise espacial da macrofauna bêntica. / Daiane Gracieli Faller. – Pontal do Paraná, 2012.
90 f.; 29 cm.

Orientador: Prof. Dr. Maurício Garcia de Camargo.

Dissertação (Mestrado) – Programa de Pós-Graduação em Sistemas Costeiros e Oceânicos, Centro de Estudos do Mar, Setor de Ciências da Terra, Universidade Federal do Paraná.

Bentos - macrofauna. I. Título. II. Maurício Garcia de Camargo
III. Universidade Federal do Paraná.

CDD 574.92

PGSISCO
Curso de Pós-Graduação em Sistemas Costeiros e Oceânicos da UFPR
Centro de Estudos do Mar - Setor Ciências da Terra - UFPR
Avn. Beira-mar, s/n.º - Baln. Pontal do Sul - Pontal do Paraná - Paraná - Brasil
Tel. (41)3511 8644 - Fax (41)3511 8644 - www.cem.ufpr.br/pgsisco - pgsisco@ufpr.br

TERMO DE APROVAÇÃO

Daiane Gracieli Faller

APLICAÇÃO DE MODELOS DE INTELIGÊNCIA ARTIFICIAL NA PREDIÇÃO E ANÁLISE ESPACIAL DA MACROFAUNA BÊNTICA

Dissertação aprovada como requisito parcial para a obtenção do grau de Mestre em Sistemas Costeiros e Oceânicos, da Universidade Federal do Paraná, pela Comissão formada pelos professores:

Dr. Maurício Garcia de Camargo
Orientador e Presidente

Dr. Fabio Teodoro de Souza (UFPR-DHS)
Membro Examinador

Dr. Olavo Correa Pedrolo (UFRGS)
Membro Examinador

Dr. Nelson Francisco Favilla Ebecken (UFRJ)
Membro Examinador

Pontal do Paraná, 23/03/2012.

"APLICAÇÃO DE MODELOS DE INTELIGÊNCIA ARTIFICIAL NA PREDIÇÃO E ANÁLISE ESPACIAL DA MACROFAUNA BÊNTICA"

POR

Daiane Gracieli Faller

Dissertação nº 93 aprovada como requisito parcial do grau de Mestre no Curso de Pós-Graduação em Sistemas Costeiros e Oceânicos da Universidade Federal do Paraná, pela Comissão formada pelos professores:

Dr. Maurício Garcia de Camargo
Orientador e Presidente

Dr. Fabio Teodoro de Souza (UFPR-DHS)
Membro Examinador

Dr. Olavo Correa Pedrolo (UFRGS)
Membro Examinador

Dr. Nelson Francisco Favilla Ebecken (UFRJ)
Membro Examinador

Pontal do Paraná, 23/03/2012.

“Dedico este trabalho:

Aos meus pais, Nilvo e Angelina, pelo amor incondicional, compreensão e dedicação, que fizeram de mim o que sou.

Ao meu irmão, minha inspiração e ídolo, por ser meu apoio nos momentos felizes e nos tristes”.

AGRADECIMENTOS

Inicialmente gostaria de agradecer a Deus pela oportunidade de estudar e trabalhar nesta área tão ampla e desafiadora.

Aos meus pais, Nilvo e Angelina, pelo apoio sempre dedicado a mim. Vocês me apoiaram, incentivaram, ajudaram, compreenderam os longos meses em que estive ausente e acima de tudo abriram mão de muitas coisas para que meu sonho fosse realizado.

Ao meu irmão Nilvanio, por ser minha fonte de inspiração, pelo apoio e paciência que medicou ao longo de meus anos de aprendizado e a sua esposa Dilvete, por compartilhar todos esses momentos.

Ao meu orientador Maurício G. Camargo, pelas inúmeras horas dedicadas, entusiasmo e muita paciência. Obrigado por acreditar em mim e em minha capacidade, seu apoio foi essencial para meu crescimento intelectual e científico.

A banca examinadora deste trabalho Nelson F. F. Ebecken, Fabio T. de Souza e Olavo C. Pedrolo, por dedicarem seu tempo para analisar o presente estudo e por suas contribuições que colaboraram muito no enriquecimento do documento final deste mestrado.

Aos meus colegas e companheiros de LAMEC pelo auxilio nas coletas, triagem e identificação das amostras e a todos os colegas e amigos que aceitaram o desafio de participar de minhas amostragens, regadas a muito bom humor e situações inusitadas, serviram para eu admirar ainda mais cada um de vocês pela amizade que me oferecem.

As minhas amigas Fer, Ju e Daphne pela ajuda que foi primordial para o término deste trabalho e de mais uma etapa da minha vida. Aos amigos não citados aqui, gostaria de dizer que não se preocupem, não foram esquecidos, a lista de nomes seria enorme se eu citasse todos que me ajudaram de alguma maneira. Cada um está em um cantinho especial em meu pensamento!!

Aos laboratórios de Sedimentologia, de Biogeoquímica Marinha e de Geoeoquíca Orgânica e Poluição Marinha. Agradeço os professores responsáveis por estes laboratórios, assim como todo o pessoal técnico e estagiários, por cederem espaço e tempo e pela realização de um expecional trabalho.

Aos barqueiros Abraão e Josias pelas inúmeras vezes que me acompanharam nas excursões aos baixios da Cotinga, e as dicas que foram essenciais para o desenvolvimento do trabalho de campo.

Aos funcionários do Centro de Estudos do Mar da UFPR pelos diversas vezes que “quebraram meu galho”.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro.

Finalmente, um agradecimento geral a cada uma das pessoas, que fazem ou fizeram parte da minha vida, positivamente ou não, pois me ajudaram a modelar e construir a minha personalidade e caráter. A presença constante de vocês seja fisica ou em espírito, ajudou a me fortalecer e ultrapassar barreiras que para muitos poderiam não ser ultrapassáveis. Vocês fizeram este momento ser possível, por isso, só tenho a agradecer.

RESUMO

O objetivo geral deste estudo foi avaliar a aplicação de diferentes técnicas de inteligência artificial (IA) na predição e análise espacial da macrofauna bêntica em relação aos descritores ambientais. Em um primeiro momento, a distribuição da fauna em relação às variáveis ambientais foi analisada através de mapas auto-organizáveis (SOMs), que identificaram cinco grupos ambientais distintos, cada um contendo pontos com características semelhantes. A ocorrência dos táxons foi maior para grupos que registraram maiores valores das variáveis que indicam enriquecimento orgânico. Em seguida, foram desenvolvidos modelos de predição da presença/ausência de 20 táxons, simultânea e individualmente, utilizando *perceptrons* de múltiplas camadas (MLP) com seleção de variáveis ambientais através dos algoritmos genéticos (GA). Para todos os táxons simultaneamente, o melhor modelo foi obtido com apenas oito variáveis. Para os táxons individualmente, somente para *T. lineata* a seleção de variáveis prejudicou a eficiência do modelo, enquanto os demais responderam positivamente à seleção. Através da análise de sensibilidade foi possível identificar que a presença da maioria dos táxons esteve relacionada com variáveis que denotam enriquecimento orgânico no ambiente. Por último, foram desenvolvidos modelos de predição de máquinas de vetor de suporte (SVM) e árvores de classificação (CT). O desempenho das duas técnicas foi comparado e cada técnica foi avaliada antes e após seleção de variáveis com GAs. Os resultados para as duas técnicas foram melhores após a seleção das variáveis ambientais. Os modelos desenvolvidos com CTs, no geral, obtiveram valores superiores para todas as medidas de desempenho. Entre os 20 táxons modelados, 16 obtiveram sucesso para as CTs, enquanto que para as SVMs apenas 10 táxons foram modelados com sucesso.

Palavras-Chave: modelagem da macrofauna bêntica, mapas auto-organizáveis, perceptron de múltiplas camadas, árvore de classificação, máquina de vetor de suporte

ABSTRACT

The aim of this study was to analyze the application of different artificial intelligence techniques in spatial analysis and prediction of benthic macrofauna in relation to environmental variables. Initially, the fauna distribution in relation to environmental variables was identified through self-organizing maps (SOMs). The SOMs identified five different environmental groups, which grouped the points with similar characteristics. The occurrence of taxa was greater for groups which recorded the highest values of the variables indicating organic enrichment. Then, models were developed to predict the presence/absence of 20 taxa, simultaneous and individually, using multilayer perceptrons (MLP) with environmental variables selection using genetic algorithms (GAs). For all taxa simultaneously the best model was obtained with only eight variables. For individual taxa, only for *T. lineata* the variable selection undermined the models efficiency and all other taxa have responded positively to the selection. Through sensitivity analysis it was found that the presence of most taxa was related to variables denoting environmental organic enrichment. Finally, predicting models were developed by support vector machines (SVM) and classification trees (CT). The performances of these techniques were compared and each technique was evaluated before and after variables selection with GAs. The results for both techniques were better after the selection of environmental variables. The models developed with CTs in general had higher values for all performance measures. Among the 20 taxa modeled, 16 were successful to CTs, while for the SVM, only 10 taxa were modeled successfully.

Keywords: benthic macrofauna modeling, self-organizing maps, multilayer perceptron, classification tree, support vector machine

SUMÁRIO

INTRODUÇÃO GERAL

A distribuição e a abundância da macrofauna bêntica são governadas pelas condições ambientais, como a composição sedimentar, disponibilidade de alimento, gradientes de salinidade e nutrientes, além de interações biológicas tais como predação e competição (Ward e Stanford, 1979; Perus e Bonsdorff, 2004). Devido à diversidade taxonômica, comportamento relativamente sedentário, ciclo de vida muitas vezes longo e resposta rápida e/ou contínua às perturbações antrópicas (Rosenberg, 1995), a fauna bêntica é amplamente utilizada na detecção de impactos ambientais por efluentes urbanos e industriais, sendo bons descritores da qualidade ambiental (Hilty e Merenlender, 2000; Rosenberg, 2001; Cooksey e Hyland, 2007).

Com isso, realizar uma modelagem eficaz para determinar as relações entre as espécies e o ambiente em que vivem é uma importante ferramenta em ecologia (Guisan e Zimmermann, 2000; Austin, 2002). Neste contexto, modelos matemáticos e simulações de computador começaram a ser utilizados como meio adequado de obter mais conhecimento sobre estas relações (Cortes, 2000). Técnicas de inteligência artificial (IA), embora mais desenvolvidas em áreas como ciência da computação e engenharia, cada vez mais estão sendo aplicadas em estudos ambientais, principalmente como auxiliares em processos de tomada de decisão (Tan et al., 2005). Entre estas técnicas podem ser citadas as árvores de classificação, redes neurais artificiais, algoritmos genéticos, máquina de vetor de suporte, entre outras, que facilitam o raciocínio ecológico e ambiental.

Estes modelos são capazes de relacionar características ecológicas com os fatores ambientais e assim revelar detalhes dos mecanismos subjacentes responsáveis pela estrutura e organização das comunidades (Austin, 1987; Gogina et al., 2010). Devido à natureza não-linear dos dados ecológicos, técnicas de IA, que possuem grande flexibilidade e capacidade de trabalhar com a não-linearidade de dados ambientais, têm sido incorporadas com eficiência à modelagem ecológica como uma alternativa poderosa para as abordagens tradicionais (Recknagel, 2006; Olden et al., 2006). Quando estas são associadas às técnicas tradicionais de modelagem, geram modelos híbridos eficazes na previsão do funcionamento dos ecossistemas e a sua evolução.

Neste estudo, como objetivo geral, buscou-se examinar a aplicação de diferentes técnicas de IA na predição e análise espacial da macrofauna bêntica em

relação aos descritores ambientais em diferentes graus de distúrbio ambiental. Os objetivos específicos do trabalho estão estruturados em três capítulos, cada qual utilizando uma abordagem distinta na modelagem da fauna bêntica em relação às variáveis ambientais.

No Capitulo I foram aplicados mapas auto-organizáveis (SOM, do inglês *Self-Organizing Maps*) na identificação da distribuição da fauna bêntica em relação às variáveis ambientais em diferentes graus de impacto ambiental. Os SOMs passaram pelo treinamento com as variáveis ambientais para identificar diferentes padrões ambientais e, em um segundo momento, os dados da fauna bêntica foram fornecidos aos mapas previamente treinados buscando identificar as relações entre a fauna e as variáveis ambientais.

O SOM é uma rede neural artificial com aprendizado não-supervisionado, ou seja, a rede não necessita de exemplos rotulados para aprender. O algoritmo aprende a representar (ou agrupar) as entradas submetidas segundo uma medida de qualidade. Estas redes são utilizadas principalmente quando o objetivo é encontrar padrões ou tendências que auxiliem no entendimento dos dados. Entre as características do SOM, está a competição entre os neurônios para determinar o neurônio vencedor, ou seja, o neurônio que irá responder ao máximo à um vetor de entrada. Este neurônio será atualizado durante o treinamento, juntamente com seus vizinhos, para reproduzir o padrão de entrada. O treinamento então ocorre consecutivamente até que todos os possíveis padrões de entrada sejam formados. Portanto, os SOMs são redes competitivas que possuem a habilidade de formar mapeamentos que preservam as mais importantes relações topológicas e/ou métricas dos dados primários (Kohonen, 2001). A rede recebe um número de diferentes padrões de entrada, descobre características significativas nestes padrões e aprende a classificar os dados de entrada em categorias apropriadas. O SOM projeta e visualiza dados de entrada multidimensional em um cenário de duas dimensões, indicando regiões de similaridade.

O Capitulo II objetivou realizar a predição da presença/ausência de 20 táxons da comunidade bêntica em diferentes condições ambientais utilizando as redes perceptron de múltiplas camadas (MLP, do inglês *multi-layer perceptron*) com algoritmo *backpropagation*. Para isso, foram utilizadas três abordagens diferentes: (1) a previsão dos táxons em conjunto com todas as variáveis (19 variáveis) e com as 14, 8 e 4 variáveis mais vezes selecionadas pelos algoritmos genéticos; (2) previsão dos táxons

separadamente antes e após seleção de variáveis com algoritmos genéticos e (3) análise de sensibilidade dos modelos com resultados satisfatórios após a seleção das variáveis.

Entre as redes neurais artificiais, as MLPs com algoritmo *backpropagation* (rede de retropropagação) são algumas das mais populares (Rumelhart et al., 1986; Hagan et al., 1996). Uma rede de retropropagação baseia-se no procedimento 'supervisionado' e pode ser utilizada para o desenvolvimento de modelos de previsão. A rede constrói um modelo baseado em exemplos de dados com saídas conhecidas. O objetivo é que a representação gerada seja capaz de produzir saídas corretas para novas entradas não apresentadas previamente. Em uma MLP com *backpropagation*, os neurônios de uma camada estão ligados à camada seguinte através de sinapses (pesos). Esta rede é composta por três tipos de camada: uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. A camada de entrada contém todas as co-variáveis consideradas e a camada de saída as variáveis de resposta. Já a camadas ocultas auxiliam a rede a trabalhar com a não linearidade dos dados fornecidos para o treinamento.

Por fim, o Capitulo III avaliou o desempenho de máquinas de vetor de suporte (SVM, do inglês *support vector machines*) e árvores de classificação (CT, do inglês *classification trees*) na predição da presença de táxons da macrofrauna bêntica. O desempenho dos modelos desenvolvidos foi avaliado de duas maneiras distintas: (1) comparação entre os modelos de SVM e CT e (2) desempenho de cada modelo antes e depois da seleção de variáveis com os algoritmos genéticos.

A SVM é embasada pela teoria de aprendizado estatístico desenvolvida por Vapnik (1995). A teoria de Vapnik estabelece uma série de princípios que devem ser seguidos na obtenção de classificadores com boa generalização, definida como a sua capacidade de prever corretamente a classe de novos dados do mesmo domínio em que o aprendizado ocorreu. Os SVMs fazem o reconhecimento de padrões e encontram um limite de decisão (vetores de suporte) usando um subconjunto de amostras de treinamento. Os resultados da aplicação dessa técnica são comparáveis e muitas vezes superiores aos obtidos por outros algoritmos de aprendizado, como as redes neurais artificiais (Haykin, 1999; Braga et al., 2000).

As CTs prevêem o valor de uma variável discreta dependente com um conjunto finito de valores (classe) a partir dos valores de um conjunto de atributos independentes (Quinlan, 1986). O procedimento principal para o desenvolvimento de um modelo de CT é dividir os dados da variável alvo baseado na resposta às variáveis de entrada. As

CTs são estruturas hierárquicas, nas quais os nós internos contêm testes sobre os atributos de entrada. Cada ramo de um teste interno corresponde a um resultado do teste e a predição para o valor do atributo de destino é armazenada em uma folha. Cada folha da árvore contém uma previsão para a variável de interesse. Estas redes possuem uma abordagem de modelagem não-paramétrica, que consiste de partições recursivas do espaço multidimensional definidos pelos grupos preditores, que são tão homogêneos quanto possível em termos da resposta (Vayssieres, 2000). O resultado da análise é uma estrutura hierárquica binária com galhos e folhas que contém as regras para prever os novos casos (Dunham, 2002).

REFERÊNCIAS

Austin, M.P., 1987. Models for the analysis of species response to environmental gradients. Vegetatio 69, 35-45.

Austin, M. P., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modeling. Ecological Modelling 157, 101-118.

Cooksey, C., Hyland J., 2007. Sediment quality of the Lower St, Johns River, Florida: An integrative assessment of benthic fauna, sediment-associated stressors, and general habitat characteristics. Marine Pollution Bulletin 54, 9-21.

Cortes, U., 2000. "Artificial intelligence and environmental decision support systems," Applied intelligence, 13(1), 77-91.

Gogina M., Glockzin M., Zettler M.L., 2010. Distribution of benthic macrofaunal communities in the western Baltic Sea with regard to near-bottom environmental parameters. Modelling and prediction, Journal of Marine Systems 80, 57-70.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135, 147–168.

Hilty, J., Merenlender, A., 2000. Faunal indicator taxa selection for monitoring ecosystem health. Biological Conservation 92, 185-197.

Olden, J.D., Poff, N.L., Bledsoe, B.P., 2006. Incorporating ecological knowledge into ecoinformatics: An example of modeling hierarchically structured aquatic communities with neural networks. Ecological Informatics 1, 33-42.

Perus, J., Bonsdorff, E., 2004. Long-term changes in macrozoobenthos in the Åland archipelago, northern Baltic Sea. Journal of Sea Research 52, 45-56.

Recknagel, F., 2006. Ecological Informatics: Scope, Techniques and Applications, Springer, Berlin.

Rosenberg, R., 1995. Benthic marine fauna structured by hydrodynamic processes and food availability, Neth. Journal of Sea Research 34, 303-317.

Rosenberg R., Nilsson H.C., Diaz R.J., 2001. Response of benthic fauna and changing sediment redox profiles over a hypoxic gradient. Estuarine, Coastal and Shelf Science 53, 343-350.

Tan, P.-N., Steinbach, M., Kumar, V., 2005. “Introduction to data mining,” Addison-Wesley, Boston.

Ward J.V., Stanford J.A., 1979. Ecological factors controlling stream zoobenthos with emphasis on thermal modification of regulated streams, In: The ecology of regulated streams (Ward J.V., Stanford J.A., eds), Plenum Press, New York, 35-55.

CAPITULO I

Aplicação de mapas auto-organizáveis na análise do padrão espacial da macrofauna bêntica em relação às variáveis ambientais em planícies de maré

Application of Self-Organizing Maps to analyze the spatial pattern of benthic macrofauna in relation to environmental variables in tidal flats

Revista pretendida: Ecological Modelling (Ecol. Model.), ISSN (0304-3800), Fator de Impacto (JCR, 2010) = 2.438, Qualis CAPES = Estrato A2.

Faller[1], D. G.; Camargo[1], M. G.

[1]Centro de Estudos do Mar, Universidade Federal do Paraná. Av. Beira Mar s/n, Pontal do Paraná, Paraná, Brasil. CEP: 83255-971. Fone: (41) 3511-8600. E-mail: daiafaller@yahoo.com.br

RESUMO

Neste estudo, os mapas auto-organizáveis (SOM) foram utilizados para identificar a distribuição da macrofauna bêntica em relação às variáveis ambientais em diferentes graus de impacto ambiental, identificando características em pequena escala. Para isso, após o treinamento dos SOMs com as variáveis ambientais, os dados da fauna bêntica foram fornecidos aos mapas previamente treinados para identificar estas relações. Após o treinamento, o SOM identificou cinco grupos ambientais distintos dos locais de amostragem (I-V) que agruparam os pontos semelhantes. Os grupos ambientais I e V apresentaram os maiores valores das variáveis que indicam o enriquecimento orgânico do sedimento, como coprostanol, esteróis naturais e nutrientes. Para II e III foram registradas as maiores porcentagens de areia e para IV foi identificada a maior quantidade de sedimentos finos e variáveis que indicam matéria orgânica, principalmente de origem terrestre. A riqueza de táxons foi maior nos grupos ambientais I e V, enquanto que a menor foi registrada para III. Para nove e oito táxons as maiores ocorrências foram registradas nos grupos IV e V, respectivamente. O SOM mostrou-se uma ferramenta analítica eficiente para a extração de informações complexas sobre os dados da fauna e sua distribuição, assim como a sua relação com as variáveis ambientais. Entretanto, a seleção dos táxons foi realizada considerando aqueles com maior ocorrência nos locais de amostragem. Este procedimento provavelmente favoreceu os táxons que habitam uma diversidade maior de condições ambientais, o que subestima a riqueza de espécies nos ambientes menos poluídos do canal.

Palavras-chave: Mapas auto-organizáveis, padrão espacial, fauna bêntica estuarina, padrão ambiental.

ABSTRACT

We applied here SOMs to identify the distribution of benthic macrofauna in relation to environmental variables in different degrees of environmental impact. After the training of the SOMs with environmental variables, the benthic fauna data were provided to the previously trained maps to identify the relationships between them. After training, we identified five environmental groups in the sampling sites (I to V). Groups I and V showed the highest values of the variables that indicates organic enrichment in sediments, such as coprostanol, natural sterols and nutrients. For II and III were recorded higher percentages of sand. In group IV was identified the greatest amount of fine sediments and variables that indicate organic matter, mainly of terrestrial origin. The richness was higher in groups I and V, while the lowest was recorded for III. For nine and eight taxa the major occurrence were recorded in groups IV and V, respectively. The SOM has proved to be an efficient analytical tool for extracting complex information about the fauna and its distribution, as well as the fauna relationship with environmental variables. However, the selection of the taxa was considered only those with highest occurrence in the sampling sites. This procedure may have favored the taxa that inhabit a wider variety of environmental conditions, which may have underestimated the species richness in less polluted environments.

Keywords: self-organizing maps, spatial pattern, benthic macrofauna, environmental patterns.

1. INTRODUÇÃO

A composição da comunidade bêntica depende da estabilidade do ambiente onde está inserida e é potencialmente determinada por diferentes fatores ambientais que atuam em diferentes escalas espaciais e temporais (Stevenson, 1997; Snyder et al., 2002). Características como diversidade taxonômica, comportamento sedentário e ciclo de vida relativamente longo, possibilitam que estes organismos respondam de forma integrada e contínua aos distúrbios naturais ou antrópicos no ambiente (Díaz e Rosenberg, 1995). Com isso, a fauna bêntica tem sido amplamente utilizada como indicadora da qualidade e estado ecológico de ambientes aquáticos (Lenat, 1988; Smith et al., 1999; Wright et al., 2000).

Diferentes táxons respondem de maneiras distintas entre si e muitas vezes não-lineares aos impactos por fatores bióticos (desenvolvimento fisiológico, ciclo de vida, etc.) e abióticos (precipitação, poluição, etc.) (Lek e Guégan, 2000; Ieno et al., 2006). Compreender os padrões de abundância e ocorrência da comunidade é fundamental para a gestão sustentável dos ecossistemas aquáticos. Com isso, diversas técnicas de

modelagem têm sido utilizadas como ferramentas auxiliares para determinar a dinâmica entre a fauna bêntica e as variáveis que influenciam a estrutura da comunidade.

Entre estas técnicas, as redes neurais artificiais (RNAs) têm sido aplicadas na interpretação de fenômenos complexos e não-lineares (Dawson e Wilby, 2001). O crescente uso de RNAs na modelagem de sistemas complexos está relacionado com a capacidade destes modelos de se adaptarem aos dados e por serem menos afetados por *outliers* (Park et al., 2003a). Entre os modelos de RNA, os mapas auto-organizáveis (Kohonen, 1982, 2001) (SOM, do inglês, *self organizing maps*) são amplamente utilizados em ecologia. Os SOMs são redes competitivas não-supervisionadas, que recebem um número de diferentes padrões de entrada, descobrem características significativas nestes padrões e aprendem a classificar os dados de entrada em categorias apropriadas, baseando-se nas semelhanças entre as amostras oferecidas à rede. Estas redes possuem a habilidade de formar mapeamentos que preservam as mais importantes relações topológicas e/ou métricas dos dados primários.

O SOM é uma ferramenta eficiente para mineração de dados não-lineares e tem mostrado particular relevância para detecção de padrões em comunidades biológicas e sua relação com os dados ambientais (Chon et al., 1996, 2000, 2002; Park et al., 2003, 2004, 2006), classificação de assembleias de peixes (Brosse et al., 2001), detecção de padrões de diversidade de macroinvertebrados aquáticos (Cereghino et al., 2001, 2003), identificação de padrões ambientais (Choi et al., 2009) e possíveis fontes de contaminação antrópica (Marengo et al., 2006).

Neste estudo os mapas auto-organizáveis (SOM) foram utilizados para identificar a distribuição da fauna bêntica em relação às variáveis ambientais em diferentes graus de impacto ambiental, identificando características em pequena escala. Para isso, os SOMs passaram pelo treinamento com as variáveis ambientais para identificar diferentes padrões ambientais e, em um segundo momento, os dados da fauna bêntica foram fornecidos aos mapas previamente treinados buscando identificar as relações entre a fauna e as variáveis ambientais.

2. MATERIAL E MÉTODOS

2.1 *Área de Estudo*

O estudo foi realizado em planícies entremarés não-vegetadas do Canal da Cotinga, um subestuário da Baia de Paranaguá, (Complexo Estuarino de Paranaguá - 25°30'S, 48°25'W), (Fig.1). Os rios que fazem parte da margem sul da baia, como o Maciel, Guaraguaçú e Itiberê e que deságuam na Cotinga, recebem grande parte dos efluentes produzidos na cidade e porto de Paranaguá (Lana et al., 2001). Entre os rios da região, o Itiberê pode ser considerado uma das principais fontes pontuais de contaminação na Cotinga, evidenciado por estudos realizados no local que encontraram elevadas concentrações de indicadores orgânicos de poluição, como coliformes fecais na coluna d'água (Kolm et al., 2002) e esteróis fecais no sedimento (Martins et al., 2010). As concentrações de contaminantes no canal são dispersas e diluídas a partir da região interna e mediana em direção à sua desembocadura, formando um gradiente de poluição. Planícies não-vegetadas das duas margens do canal foram selecionadas, totalizando 107 locais de amostragem ao longo da Cotinga.

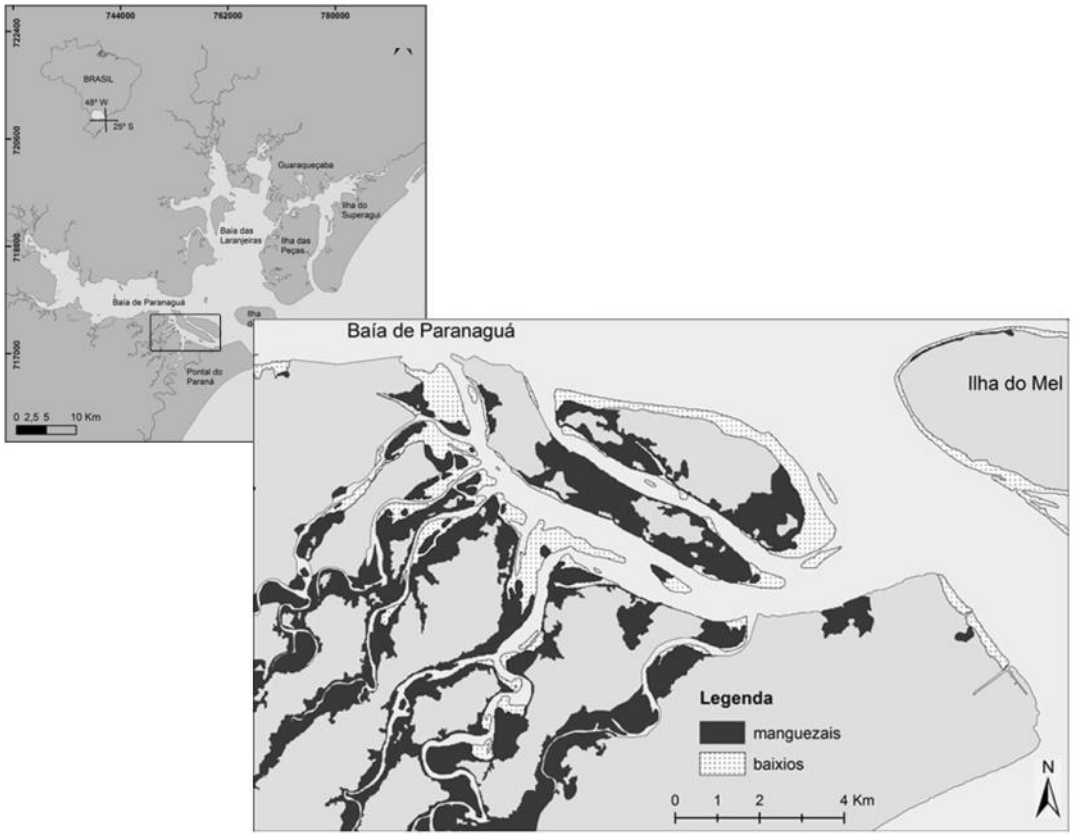


Fig.1. Complexo Estuarino de Paranaguá (CEP) com o Canal da Cotinga no detalhe.

2.2 Conjunto de dados

Em cada local de amostragem foram coletadas duas réplicas com *corers* com diâmetro de 10 cm para a análise da macrofauna bêntica. As amostras biológicas foram fixadas com formol-aldeído (4%), lavadas em peneiras de 0,5 mm, coradas com Rosa de Bengala, preservadas em álcool (70%) e identificadas até o nível de família, totalizando 49 famílias. Entre estas, foram selecionadas as famílias com mais de 15% de ocorrência, que foram identificadas até o menor nível possível. Os 20 táxons com maior ocorrência foram utilizados no desenvolvimento dos modelos do SOM. São eles: *Anomalocardia brasiliana* (35,5%), *Bulla* sp. (64,5%), Capitellidae (72,9%), *Capitella* sp. (23,4%), *Caprella scaura* (14%), *Glycinde multidens* (87,9%), *Heleobia australis* (63,6%), *Laeonereis culveri* (31,8%), *Neanthes succinea* (13,1%), Oligochaeta sp1 (34,6%), Orbiniidae (51,4%), Ostracoda (39,3%), *Polydora* sp. (42,1%), *Prionospio* sp. (44%), *Sigambra* sp. (82,2%), *Sternaspis* sp. (36,4%), *Streblospio benedicti* (61,7%), *Tagelus* sp. (53,3%), *Tellina lineata* (59,8%) e Tubificidae (85%).

Juntamente com a fauna, 19 variáveis ambientais foram utilizadas no desenvolvimento dos modelos (Tabela 1). Em campo, foi realizada a medição da camada redox e amostras de água intersticial foram coletadas para a determinação do pH e salinidade (pHmetro e refratômetro manual, respectivamente). Amostras de sedimento foram coletadas para determinar: fósforo total e nitrogênio total (Grasshoff et al., 1983); carbono orgânico total (Strickland e Parsons, 1972); clorofila-*a* e feoftina (Lorenzen, 1967), porcentagens de cascalho, areia, silte e argila, (Suguio, 1973); carbonato de cálcio, matéria orgânica e esteróis totais (Kawakami e Montone, 2002).

Foram considerados cinco esteróis para identificar diferentes fontes de matéria orgânica: coprostanol e estigmasterol, como indicadores de contaminação por esgotos (Maldonado et al., 2000); colesterol para origem aquática, por estar presente no fito e zooplâncton (Volkman, 1986); brassicasterol para identificar matéria orgânica de origem marinha e estigmasterol para identificar matéria orgânica de origem terrestre (Volkman, 2006). Devido à inviabilidade em estimar a concentração de esteróis em todos os pontos, optou-se por realizar as amostragens em 31 pontos selecionados de acordo com a proximidade à cidade de Paranaguá e à desembocadura dos rios. A extrapolação para os demais pontos foi realizada através da média entre pontos vizinhos.

Tabela 1

Variáveis de entrada utilizadas para o desenvolvimento dos modelos em conjunto com a média, desvio padrão (DP), valores mínimos (Min.) e máximos (Max.).

Variável	Abreviação	Unidade	Média	DP	Min.	Max.
Salinidade	SAL	-	24,63	5,35	2	32
pH	PH	-	7,28	0,27	6,58	7,92
Camada Redox	RED	cm	1,15	1,15	0,1	5,97
Coprostanol	COP	$\mu g.g^{-1}$	0,38	0,46	0	2,04
Epicoprostanol	EPI	$\mu g.g^{-1}$	0	0,01	0	0,08
Colesterol	COL	$\mu g.g^{-1}$	4,37	2,89	0,74	15,5
Brassicasterol	BRA	$\mu g.g^{-1}$	2,2	1,5	0,23	8,22
Estigmasterol	EST	$\mu g.g^{-1}$	2,7	1,62	0,27	10,40
Cascalho	CAS	%	1,72	5,95	0	52,28
Areia	ARE	%	84,98	11,4	33,76	97,89
Silte	SIL	%	10,21	8,9	0	44,56
Argila	ARG	%	3,09	3,01	0	21,31
Carbono	COT	$mg.g^{-1}$	14,55	11,7	0	46,7
Clorofila	CHL	$mg.g^{-1}$	17,39	23,2	0	157,95
Feoftina	FEO	$mg.g^{-1}$	17,88	23,5	0	178,75
Nitrogênio Total	NT	$mg.g^{-1}$	2,09	1,11	0,07	4,41
Fósforo Total	PT	$mg.g^{-1}$	0,03	0,02	0	0,09
Matéria Orgânica	MO	%	4,36	2,17	0,49	13,05
Carbonato de Cálcio	CaCO3	%	4,11	3,98	0,48	32,23

2.3 Mapas auto-organizáveis (SOM)

Os procedimentos necessários para aplicar o SOM podem ser divididos em três etapas distintas: a padronização dos dados, a formação do SOM e a extração da informação após o treinamento.

Primeiramente, os dados passam pelo procedimento de padronização, que evita que algumas variáveis tenham maior impacto que outras e garante que todas as variáveis tenham igual importância na formação do SOM. Com isso, os dados foram escalonados entre 0 e 1, sendo 0 o valor mínimo e 1 o valor máximo dentro de cada variável. Os dados da comunidade ainda foram transformados por logaritmo natural.

Após a padronização, os dados são apresentados à rede para o procedimento iterativo de treinamento para formar o SOM. A estrutura típica de um SOM consiste em duas camadas: uma camada de entrada e uma camada de Kohonen ou saída conectadas por intensidades de conexão w_{ij} (pesos) (Fig. 2).

A camada de entrada é formada por neurônios *j* que recebem informações para cada variável (abundância/riqueza da fauna, salinidade, nutrientes, etc.) a partir da matriz de dados. Cada amostra é representada por um vetor x_i. Quando um vetor x_i é enviado através da rede, cada neurônio *j* da rede calcula a distância $d_j(t)$ entre o vetor de pesos $w_{ij}(t)$ e o vetor de entrada x_i. Os pesos foram inicializados com valores baixos e aleatórios, mudando adaptativamente a cada iteração no tempo *t*. Durante o treinamento os pesos foram calculados usando uma medida de distância (Distância Euclidiana). O cálculo da distância foi:

$$dj(t) = \sum_{i=0}^{n-1} (x_i - wij(t))^2 \qquad (1)$$

A camada de saída é composta por neurônios de saída D_j, que funcionam como locais virtuais retornando um padrão, onde amostras que possuem semelhanças ficam próximas na visualização do mapa. Os neurônios de saída normalmente são organizados em uma grade bidimensional para melhor visualização.

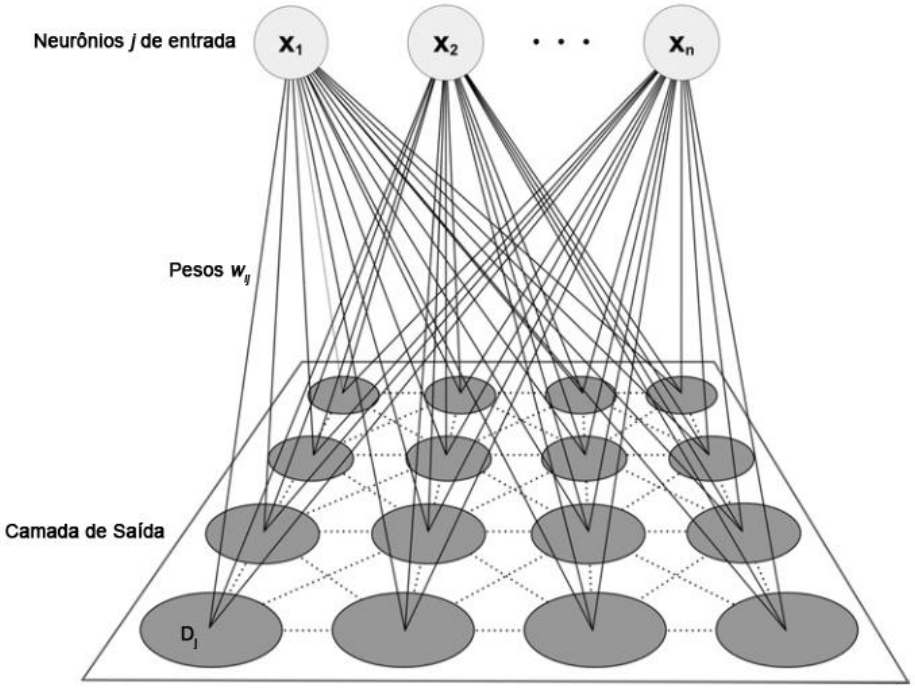
Entre os neurônios da camada *D,* o neurônio que responder ao máximo a um vetor de entrada é escolhido como neurônio vencedor ou melhor unidade de correspondência (BMU, do inglês *best matching unit*). Durante o aprendizado, o neurônio vencedor e seus vizinhos são atualizados para reproduzirem o padrão de entrada, reduzindo ainda mais a distância entre o peso e o vetor de entrada:

$$wij(t+1) = wij(t) + \eta(t)(xi - wij(t))Zj \qquad (2)$$

onde, para Z_j é atribuído 1 aos neurônios vencedores (e seus vizinhos) enquanto que os demais neurônios (não-ativados) recebem 0 e $\eta(t)$ denota o incremento fracionário da correção. O raio da vizinhança normalmente é definido com um valor maior no inicio do processo de formação e então é gradualmente reduzido à medida que se aproxima da convergência. Como função de vizinhança foi utilizada a Função Gaussiana (Kohonen, 2001):

$$N_{j^*}(t) = e^{\frac{\|r_{j^*} - r_j\|^2}{-2\delta^2(t)}} \qquad (3)$$

onde $N_{j^*j}(t)$ é a função de vizinhança da BMU (j^*) na iteração *t*; $\delta(t)$ é o raio da vizinhança na iteração *t*; e $\|r_{j^*} - r_j\|$ é a distância entre os neurônios j^* e j na grade do mapa.



197

198 **Fig.2**. Ilustração do modelo SOM utilizado no estudo. Adaptada de Schmidt et al. (2011).

199

200 O treinamento do SOM foi realizado utilizando diferentes tamanhos de mapa e
201 o tamanho ideal foi escolhido com base no menor erro topográfico. O erro topográfico
202 identifica se a rede foi treinada com sucesso e se os resultados obtidos são confiáveis.
203 Após o treinamento, foi aplicado, para um melhor agrupamento (*clustering*), o método
204 de ligação de Ward com base na Distância Euclidiana (Ward, 1963).

205 A identificação de diferenças significativas para as variáveis ambientais entre
206 os grupos formados pelo SOM foi realizada através do teste não paramétrico de Mann-
207 Whitney, onde os grupos foram comparados par a par. Devido às múltiplas
208 comparações realizadas, o nível critico foi fixado em 1% (α = 0,01), buscando
209 minimizar a ocorrência do erro tipo 1, onde são encontradas diferenças em amostras
210 semelhantes. Para a análise da riqueza e abundância total de táxons em cada
211 agrupamento do SOM, foram utilizados todos os táxons identificados nos locais de
212 amostragem (nível taxonômico de família). Possíveis diferenças entre os grupos foram
213 testadas através do teste não paramétrico de Kruskall-Wallis. As respostas de ocorrência
214 ou não dos 20 táxons submetidos ao SOM foram avaliadas par a par através de modelos
215 lineares generalizados (GLM) da família binomial, usando a função de ligação *logit*
216 (Nelder & Wedderburn, 1972).

Para o desenvolvimento dos SOMs foi utilizado o SOM Toolbox (Vesanto et al., 1999) para o MATLAB (The Mathworks, 2011). O algoritmo do SOM está detalhado em Kohonen (2001) e Chon et al. (1996).

3. RESULTADOS

3.1 Padrão Ambiental

O melhor modelo construído pelo SOM foi composto de 35 neurônios dispostos em sete linhas e cinco colunas (Erro de Quantização: 0,6; Erro Topográfico: 0,009). Após o treinamento com as variáveis ambientais, o SOM identificou cinco grupos ambientais diferentes dos locais de amostragem (I-V) (Fig.3a). Cada grupo corresponde aos pontos de amostragem que possuem características ambientais semelhantes. O dendrograma resultante da análise de agrupamentos, utilizando o método de ligação de Ward e Distância Euclidiana, classificou os neurônios do SOM em dois grandes agrupamentos, o primeiro formado pelos grupos III e IV e o segundo pelos grupos I, II e V (Fig.3b).

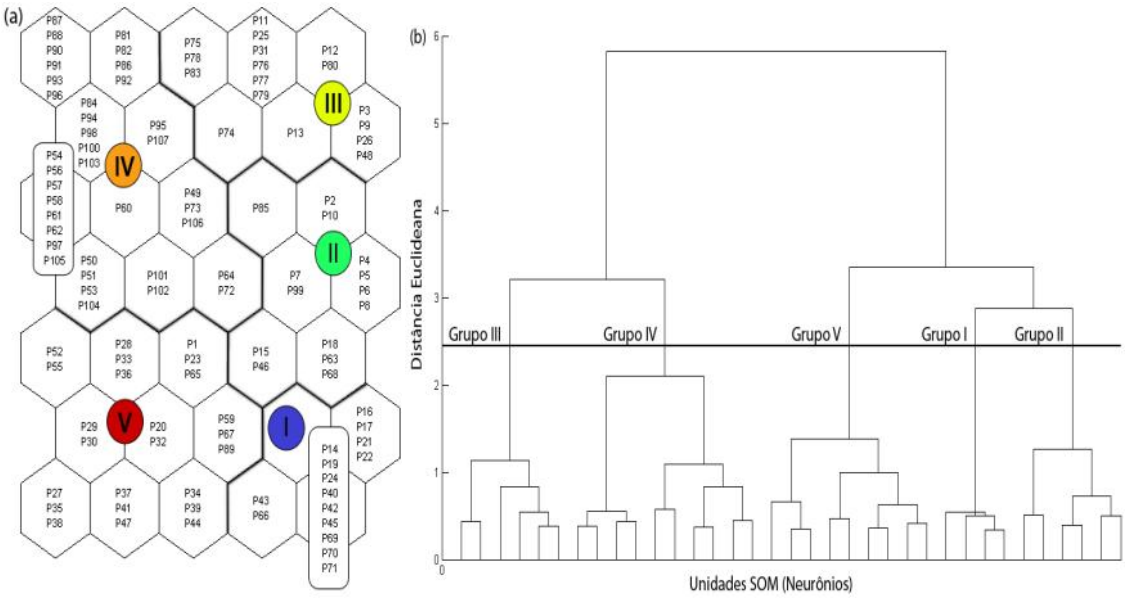


Fig.3- (a) Classificação dos 107 locais de amostragem através do processo de aprendizagem dos mapas auto-organizáveis (SOM). Os pontos foram agrupados em cinco grupos distintos (I-V) relacionados com as características ambientais semelhantes. (b) As unidades SOM foram agrupadas utilizando análise de cluster hierárquica com método de ligação Ward e medida de Distância Euclidiana.

Os grupos ambientais I e V foram formados pelos pontos com maiores concentrações de coprostanol, com valores médios de 0,5 $\mu g.g^{-1}$ e 0,42 $\mu g.g^{-1}$, respectivamente. Além do coprostanol, no grupo I os pontos apresentaram valores mais elevados de clorofila-*a* (27,02 $mg.g^{-1}$), camada redox (1,91 cm), colesterol (7,74 $\mu g.g^{-1}$) e brassicasterol (4,01 $\mu g.g^{-1}$). Já o grupo V, apresentou valores mais elevados dos nutrientes com valores médios de 2,77 $mg.g^{-1}$ para nitrogênio total e 0,04 $mg.g^{-1}$ para fósforo total. Outras variáveis, como clorofila-*a*, carbono orgânico total e camada redox também registraram valores elevados para o grupo (Fig.5). O grupo I foi formado por 15 e o V por 24 pontos, todos localizados na margem sul do Canal da Cotinga, principalmente próximos às desembocaduras dos rios e Baixio dos Papagaios (Fig. 4).

Os grupos II e III foram formados por pontos com altas porcentagens de areia (acima de 90%). As demais variáveis registraram valores medianos para o grupo II, enquanto no grupo III, para a maioria das variáveis, foram registrados os menores valores. O grupo II foi formado por 14 pontos, localizados principalmente próximo à cidade e porto de Paranaguá (Fig. 4). Os 17 pontos que formaram o grupo III estão localizados principalmente próximos à cidade de Paranaguá e na margem norte do Canal da Cotinga, próximo ao Baixio dos Papagaios (Fig. 4).

No grupo IV, foram agrupados 37 pontos semelhantes, caracterizando o grupo mais amplo. Estes pontos estão localizados principalmente na margem norte do canal e entre os rios Guaraguaçú e Maciel (Fig. 4). Estes pontos registraram as maiores porcentagens de sedimento fino (argila com média de 4,01% e silte com 15,31%). Os valores de estigmasterol (3,42 $\mu g.g^{-1}$), pH (7,38), carbonato de cálcio (5,62%), matéria orgânica (5,54%), carbono orgânico total (19,99 $mg.g^{-1}$) e feoftina (25,39 $mg.g^{-1}$) também foram os mais elevados.

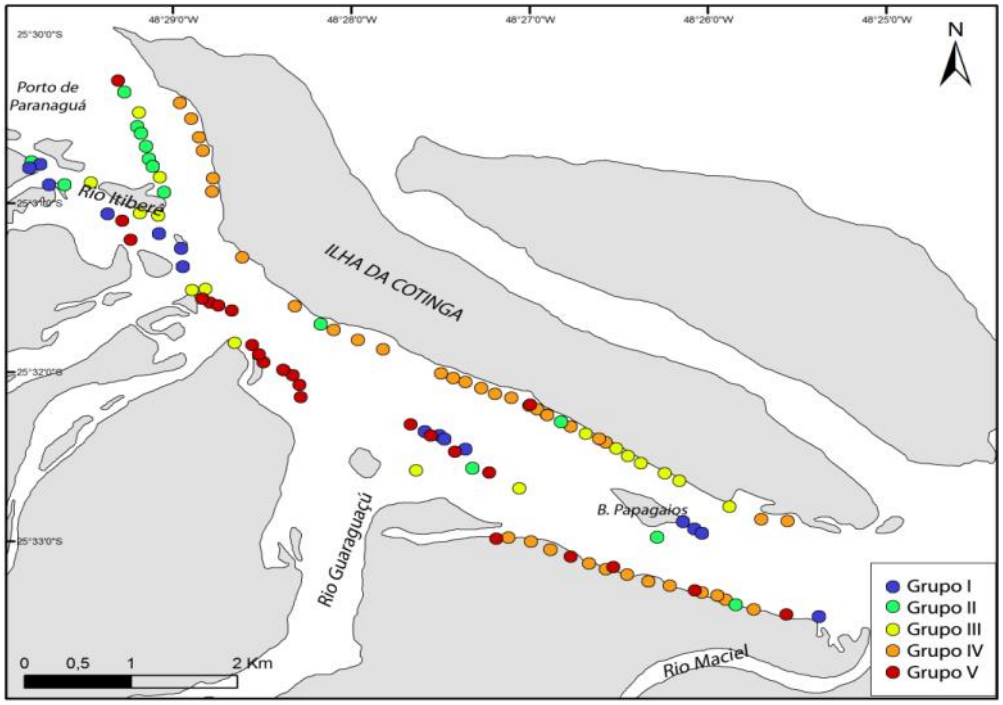


Fig.4 - Distribuição dos grupos de padrão ambiental gerados pelos mapas auto-organizáveis (SOM) em relação aos pontos de amostragem ao longo do Canal da Cotinga.

Entre as variáveis ambientais analisadas, para os grupos par a par, as menores variações foram encontradas para coprostanol, clorofila-*a*, feoftina e cascalho. As concentrações de coprostanol e clorofila-a somente registraram diferenças entre os grupos I e III (Mann-Whitney, $p<0,01$). As concentrações de feoftina diferenciaram para o grupo IV em relação aos grupos I e III (Mann-Whitney, $p<0,01$) e a porcentagem de cascalho para o grupo I em relação aos grupos II e IV (Mann-Whitney, $p<0,01$). O oposto ocorreu com colesterol que apresentou variações significativas para todos os grupos ambientais (Mann-Whitney, $p<0,01$), exceto para o grupo II quando comparado aos grupos IV e V e entre o grupo IV e V (Fig.5). Para os grupos ambientais III e IV o teste para epicoprostanol não foi realizado, pois nos pontos de ambos os grupos os valores de epicoprostanol foram nulos ou abaixo do nível de detecção (Fig. 5).

Os grupos mais semelhantes, avaliado através das diferenças obtidas para as variáveis ambientais, foram II e III com valores significativamente diferentes somente para salinidade, colesterol e nitrogênio total. Para os grupos II e V as diferenças foram encontradas para salinidade, estigmasterol, carbono orgânico total e fósforo total (Fig.5). Em contrapartida, os grupos com diferenças mais acentuadas foram I e IV, que foram semelhantes somente para os valores de coprostanol, clorofila-*a*, nitrogênio total e fósforo total (Fig.5).

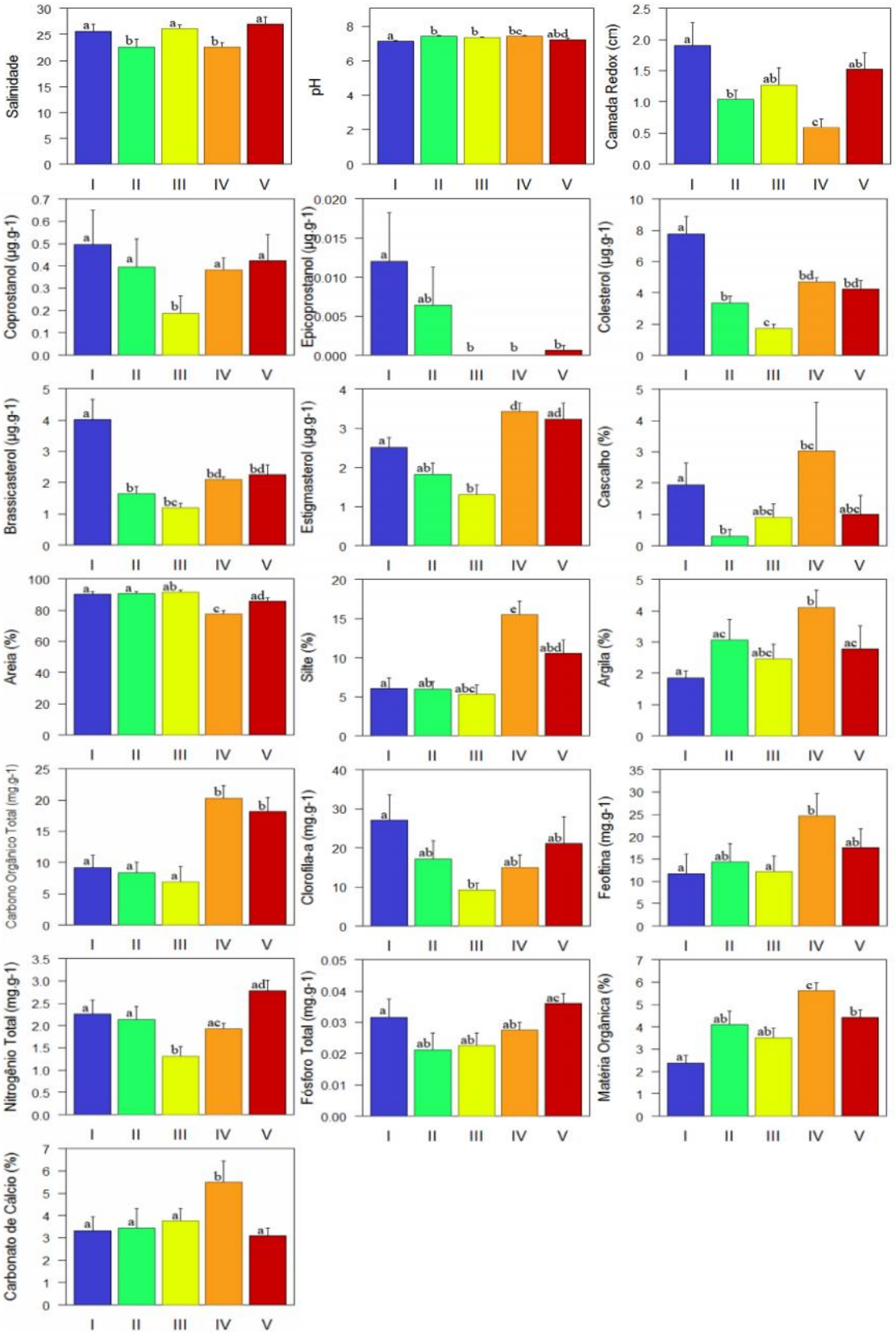


Fig.5 - Valores médios e desvio padrão de cada variável ambiental (ver Tabela 1 para detalhes) para os grupos ambientais (I-V) determinados pelos mapas auto-organizáveis (SOM). Letras diferentes correspondem às diferenças significativas das variáveis dentro e entre os grupos ambientais, de acordo com o teste não-paramétrico de Mann-Whitney.

3.2 Distribuição espacial da macrofauna

Para a determinação da riqueza e abundância de espécies entre os grupos ambientais foram considerados todos os táxons encontrados nos locais de amostragem. A riqueza entre os grupos ambientais foi significativamente diferente (Kruskall-Wallis, $p<0,001$), exceto entre II e IV (Fig.6a). O grupo III apresentou a menor riqueza, com média de 6 táxons por ponto, enquanto a maior foi registrada no o grupo I, com média de 20 táxons por ponto. Para a abundância total, no entanto, não foram encontradas diferenças significativas entre os grupos, com média de 228 a 288 indivíduos por ponto (Fig.6b). Os testes realizados somente com os táxons que foram submetidos aos SOMs apresentaram os mesmos padrões de riqueza e abundância que foi encontrado para todos os táxons.

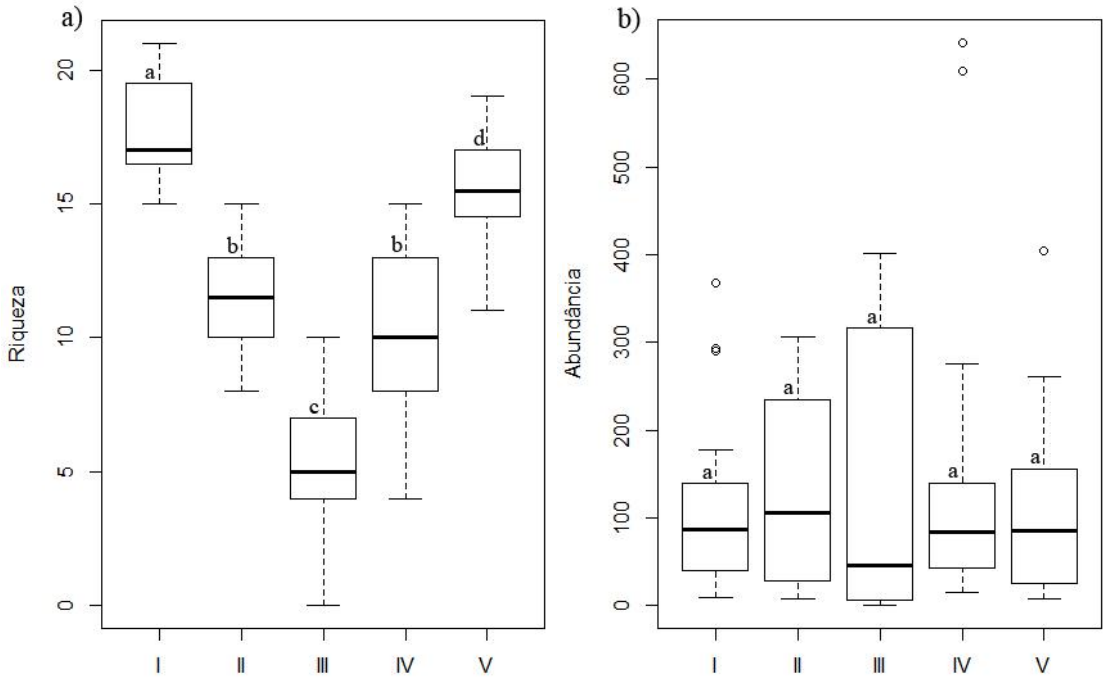


Fig. 6 - Box-plot mostrando as diferenças de riqueza de espécies (a) e abundância total (b) nos diferentes grupos ambientais (I-V) formados pelos mapas auto-organizáveis (SOM). Letras diferentes correspondem às diferenças significativas das variáveis dentro entre os grupos ambientais.

Entre os táxons submetidos ao SOM, os táxons com maior ocorrência nos pontos de amostragem foram *G. multidens* (87,9%), Tubificidae (85%), *Sigambra* sp. (82,2%) e Capitellidae (72,9%) enquanto as menores ocorrências foram encontrados para *N. succinea* (13,1%), *C. scaura* (14%) e *Capitella* sp. (23,4%).

Dentre os 20 táxons analisados, nove foram mais ocorrentes para os pontos formadores do grupo ambiental IV e oito nos pontos do grupo V. Para o grupo I, somente três táxons apresentaram suas maiores ocorrências e para os grupos II e III nenhum táxon registrou maior ocorrência. O táxon *C. scaura* foi ausente nos grupos II, III e IV, enquanto que a ocorrência no grupo V foi de 93,3%. Para os grupos I e II e para o grupo III houve ausência de *Sternaspis* sp. e *N. succinea,* respectivamente (Tabela 3).

Quando analisados os grupos par a par com GLM, foi possível identificar que a composição da fauna foi mais similar para os grupos I e III e entre I e IV, com somente 7 táxons apresentando valores de ocorrência com diferença significativa entre os grupos. As maiores diferenças na composição da fauna foram identificadas entre os grupos I e V com 17 táxons apresentando ocorrência diferenciada para os grupos, entre II e IV, com 16 e entre os grupos I e II com 15 táxons.

Entre os táxons a maior diferença entre os grupos ambientais foi para Capitellidae e *L. culveri*. O táxon Capitellidae foi semelhante entre os grupos I e V e entre II e IV. Para *L. culveri* a semelhança foi encontrada somente para grupo III quando comparado com IV e V. Porém, *N. succinea* registrou porcentagens de ocorrência similares entre os grupos, com exceção dos grupos II e III onde houve diferenças significativas (GLM, $p<0,05$). O mesmo ocorreu para *Sigambra* sp. que foi significativamente diferente somente para o grupo I em relação aos grupos III e IV (Tabela 3).

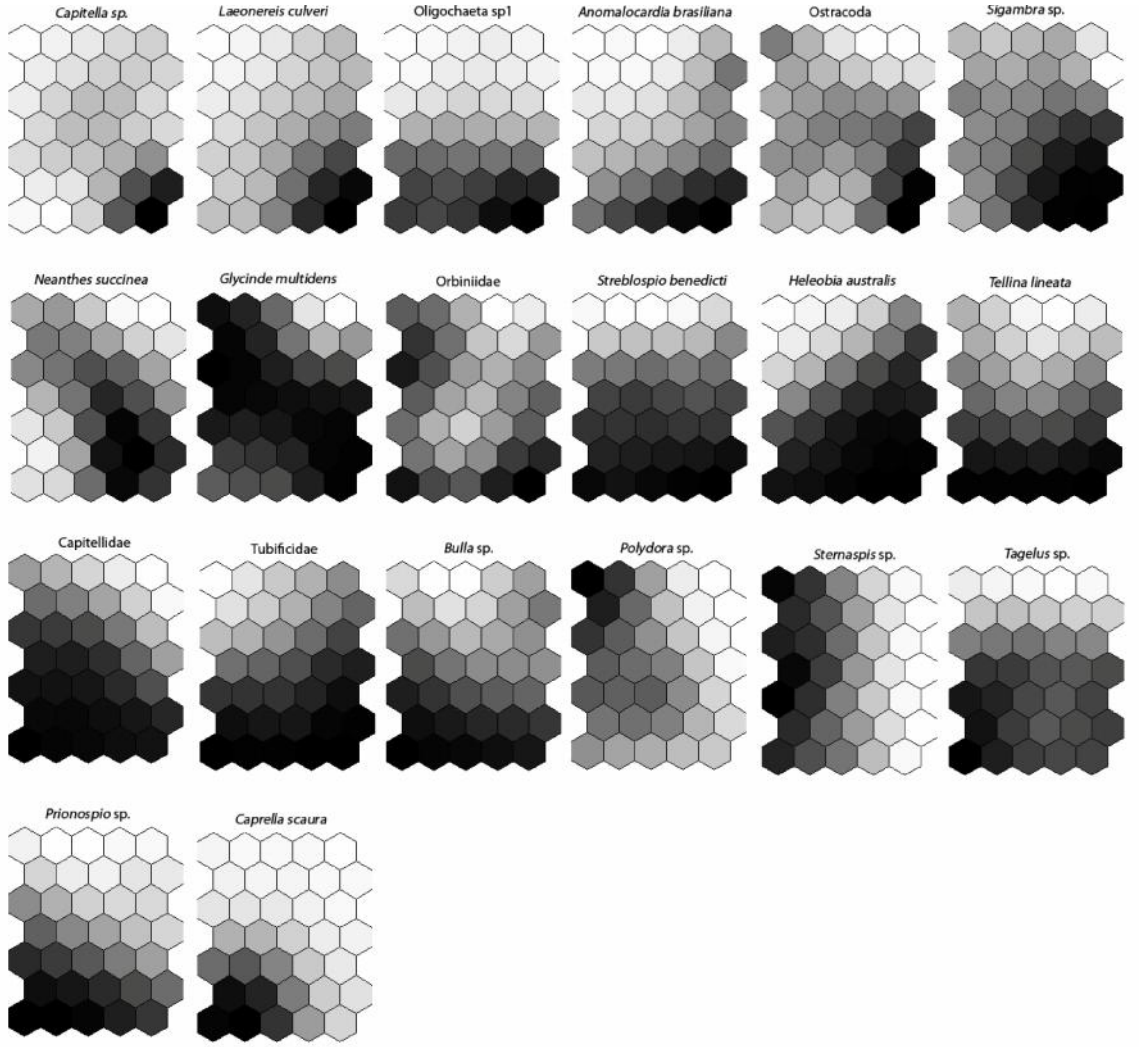


Fig. 7 - Padrões de distribuição dos táxons nos grupos ambientais definidos pelos mapas auto-organizáveis (SOM). Cores escuras representam maior ocorrência dos táxons e claras representam ocorrências baixas ou nulas.

345 **Tabela 3**

346 Porcentagem de Ocorrência (%) dos 20 táxons nos diferentes grupos ambientais (I-V) identificados pelos mapas auto-organizáveis (SOM) e análise da diferença da ocorrência
347 entre os grupos realizada com modelos lineares generalizados (GLM). *p < 0,05; **p < 0,01; ***p < 0,001.

Táxons	Ocorrência (%)	Valores Médios (%)					GLM									
		I	II	III	IV	V	I-II	I-III	I-IV	I-V	II-III	II-IV	II-V	III-IV	III-V	IV-V
Anomalocardia brasiliana	35,5	35,14	13,51	8,11	10,81	32,43	**	***	***	*	0,27	*	0,41	0,50	*	***
Bulla sp.	64,5	18,84	13,04	8,70	26,09	33,33	0,18	**	*	0,32	0,12	0,33	*	0,37	*	***
Caprella scaura	14	6,67	0,00	0,00	0,00	93,33	0,37	0,32	0,13	**	NA	NA	***	NA	***	***
Glycinde multidens	87,9	15,96	14,89	9,57	38,30	21,28	0,97	**	0,55	0,11	**	0,57	0,12	0,06	*	*
Heleobia australis	63,6	20,59	19,12	10,29	16,18	33,82	1,00	**	*	0,76	**	0,07	0,72	0,42	***	***
Oligochaeta sp1	34,6	37,84	10,81	2,70	5,41	43,24	***	**	***	0,06	0,10	*	*	0,96	***	***
Orbiniidae	51,4	21,82	12,73	7,27	38,18	20,00	0,10	**	0,12	*	0,14	0,68	0,82	*	0,15	0,41
Ostracoda	39,3	28,57	19,05	2,38	35,71	14,29	0,20	*	**	***	**	0,30	*	**	0,12	0,22
Sigambra sp.	82,2	17,05	14,77	12,50	32,95	22,73	0,33	*	*	0,11	0,07	0,24	0,42	0,30	0,18	0,64
Sternaspis sp.	36,4	0,00	0,00	2,56	64,10	33,33	1	0,38	*	***	0,40	***	***	*	**	0,30
Tagelus sp.	53,3	19,30	15,79	3,51	29,82	31,58	0,62	***	0,08	0,92	**	0,25	0,50	*	0,09	*
Tellina lineata	59,8	23,44	15,63	3,13	23,44	34,38	*	***	***	0,27	***	*	0,11	*	***	0,08
Tubificidae	85	16,48	15,38	13,19	29,67	25,27	0,96	*	*	0,46	*	*	0,48	0,87	*	*
Capitellidae	72,9	19,23	11,54	3,85	34,62	30,77	*	**	*	0,97	**	0,56	**	***	***	**
Capitella sp.	23,4	44,00	8,00	16,00	24,00	8,00	**	**	0,09	*	0,54	0,88	0,59	0,53	0,19	0,38
Laeonereis culveri	31,8	44,12	23,53	8,82	5,88	17,65	**	**	***	**	*	***	*	0,16	0,59	*
Neanthes succinea	13,1	21,43	21,43	0,00	35,71	21,43	0,95	0,06	0,57	0,55	*	0,50	0,49	0,12	0,14	0,92
Polydora sp.	42,1	8,89	6,67	2,22	55,56	26,67	0,77	0,12	**	0,16	0,22	**	0,09	*	**	0,18
Streblospio benedicti	61,7	22,73	15,15	4,55	27,27	30,30	*	**	***	0,11	**	0,15	0,40	*	*	**
Priosnopio sp.	44	21,28	6,38	4,26	21,28	46,81	*	**	**	*	0,49	0,70	***	0,22	***	***

4. DISCUSSÃO

A aplicação dos mapas auto-organizáveis (SOM) nas variáveis ambientais sugere que os diferentes grupos identificados (Fig.3 e Fig.4) refletem distintos processos ambientais nas planícies entremarés do Canal da Cotinga. As amostras contidas dentro de cada grupo ambiental possuem propriedades semelhantes e podem ser interpretadas como locais onde processos ambientais semelhantes ocorrem. A abordagem de treinamento do SOM com as variáveis ambientais e a possibilidade de introduzir o conjunto de dados biológicos nestes mapas previamente treinados possibilitou analisar as relações entre o conjunto de variáveis e a fauna bêntica.

A distribuição espacial de organismos bênticos ao longo de gradientes estuarinos é fortemente influenciada por gradientes ambientais como salinidade, aporte de água doce e características sedimentares (Underwood et al., 2000). A competição por espaço ou alimento influencia em menor intensidade a distribuição destes organismos (Peterson, 1979; Wilson, 1991). Chapman e Wang (2001) estudaram a contaminação em estuários e relataram que a distribuição da fauna é regida principalmente pela salinidade e secundariamente por fatores como substrato, oxigênio dissolvido e poluição do ambiente. Neste mesmo contexto, Bustamante et al. (2007) demonstraram que no estuário de Bilbao (Espanha) a riqueza de espécies diminui em salinidades inferiores, sendo considerada pelos autores a principal responsável pela distribuição das espécies ao longo do estuário.

Entretanto, ao longo do Canal da Cotinga, a salinidade média foi acima de 22, com poucos pontos apresentando valores inferiores, o que sugere que a distribuição dos táxons no local pode ser influenciada por outros processos que ocorrem no ambiente. Além das variações intrínsecas ao ambiente estuarino, o Canal da Cotinga presumidamente possui um gradiente de contaminação ambiental, com fontes pontuais de contaminação nos principais rios que deságuam no canal, como o Itiberê e o Guaraguaçú (Martins et al., 2010). Estes processos de poluição podem ser responsáveis por alterações nas condições ambientais, além de provocarem mudanças estruturais nas associações bênticas (Elias et al., 2003).

A maior riqueza da fauna foi encontrada para os grupos ambientais I e V, enquanto que o grupo III apresentou a menor riqueza. A diferença entre os pontos formadores destes grupos está associada principalmente aos valores das variáveis que

denotam enriquecimento orgânico. Os grupos I e V, baseados nos valores de concentração de coprostanol, contêm os pontos com a mais elevada contaminação ambiental, atingindo máximos de 2,04 $\mu g.g^{-1}$. Além do coprostanol, estes pontos apresentam valores mais elevados de clorofila-*a,* esteróis naturais (que identificam aporte de matéria orgânica de origem natural) e nutrientes (nitrogênio e fósforo total). Por outro lado, os pontos que formaram o grupo III possuem altos teores de areia e baixas concentrações de variáveis indicadoras de enriquecimento orgânico. Ambientes com enriquecimento orgânico no sedimento favorecem algumas espécies tolerantes ou oportunistas, que se tornam mais abundantes e dominam a comunidade, enquanto espécies menos tolerantes se tornam cada vez mais raras e em muitas situações tendem a desaparecer. A dominância de algumas espécies oportunistas pode ser utilizada como indicadoras positivas de poluição no ambiente, enquanto que a ocorrência de espécies sensíveis indica que o ambiente é prístino (Rygg, 1985a). Entretanto, o nível de contaminação no sedimento encontrado na Cotinga é considerado moderado quando comparado com concentrações observadas em outros sistemas estuarinos brasileiros (Carreira et al., 2004; Cordeiro et al., 2008; Santos et al., 2008; Martins et al., 2008a). Segundo Dauvin et al. (2009), ambientes estuarinos, onde a matéria orgânica se acumula naturalmente, o enriquecimento orgânico ligado às atividades antrópicas, apesar de correlacionadas com alguns organismos mais tolerantes, muitas vezes não está diretamente ligado a distribuição da fauna, enquanto outros fatores ambientais possuem maior influência no processo de distribuição espacial dos organismos.

A macrofauna bêntica responde ao enriquecimento orgânico do substrato através da supressão de espécies sensíveis e dominância de tolerantes e oportunistas (Pearson e Rosenberg, 1978). O enriquecimento orgânico pode ser resultante de processos naturais no ambiente, como aporte de manguezais, o que torna difícil a identificação dos fatores que afetam a distribuição das espécies (Dauvin, 2007; Elliott e Quintino, 2007). Este pode ser o caso da distribuição de *Caprella scaura*, que foi associada aos locais com certo nível de enriquecimento orgânico. No geral, a Ordem Amphipoda está associada a algas e é sensível à poluição por esgoto, mostrando diminuição de abundância e diversidade em estações próximas às áreas poluídas (de-la-Ossa-Carretero et al., 2012). Entretanto, assim como neste estudo, Krapp-Schickel e Vader (1998), no Mar Adriático, identificaram que *Caprella acanthifera* esteve associada a ambientes com baixa energia, ricos em detritos e clorofila-*a*.

A utilização dos SOMs em diversas áreas da modelagem ambiental, principalmente em ecologia, tem mostrado que estes algoritmos são poderosas ferramentas analíticas para a descoberta de padrões entre componentes ecológicos e a dinâmica das comunidades. Entre as áreas mais comuns que os SOMs são aplicados se destacam a análise espacial de dados ambientais (Cereghino et al., 2001; Park et al., 2001; Tran et al., 2003; Kruk et al., 2007), a previsão da ocorrência de espécies de peixes em áreas alteradas (Park et al., 2003) e a distribuição da fauna bêntica (Chon et al., 1996; Park et al., 2003b, 2006).

Alvarez-Guerra et al. (2008) comparou o uso do SOM com análise de cluster hierárquico (HCA) e análise de componentes principais (PCA) na classificação sedimentar, considerando variáveis químicas, físicas e ecotoxicologicas em diferentes sistemas estuarinos e observou que os resultados obtidos entre os métodos foram semelhantes. Porém, segundo os autores e trabalhos correlatos em outras áreas do conhecimento - fontes de carbono (Leflaive et al, 2005), comunidade de diatomáceas (Tison et al,, 2005) e identificação de acido lácteo em bactérias (Piraino et al., 2006) - o SOM fornece uma classificação mais detalhada que é mais eficiente e útil para investigações posteriores e tomadas de decisão, pois é capaz de diferenciar os grupos mais claramente, além da sua visualização ser mais fácil de interpretar.

Os resultados obtidos neste estudo mostraram que o SOM pode ser usado como uma ferramenta analítica e pode identificar relações entre as unidades de amostragem, sendo eficiente na extração de informações complexas sobre os dados da fauna e sua distribuição, assim como a sua relação com as variáveis ambientais. Entretanto, a seleção dos táxons foi realizada considerando apenas aqueles com maior ocorrência nos locais de amostragem e somente os 20 mais frequentes foram utilizados para a modelagem pelo SOM. Este procedimento provavelmente favoreceu os táxons mais tolerantes à poluição, ou seja, que habitam uma diversidade maior de condições ambientais, o que subestima a riqueza de espécies nos ambientes menos poluídos do canal. Um segundo passo seria utilizar todos os táxons identificados neste estudo e determinar a preferência de habitat também dos táxons mais raros.

5. REFERÊNCIAS

Alvarez-Guerra M., González-Piñuela C., Andrés A., Galán B., Viguri J.R., 2008. Assessment of self-organizing map artificial neural networks for the classification of sediment quality. Environment International 34, 782-790.

Brosse, S., Giraudel, J.L., Lek, S., 2001. Utilization of non-supervised neural networks and principal component analysis to study fish assemblages. Ecological Modelling 146, 159-166.

Bustamante, M., Tajadura-Martín, F., Saiz-Salinas, J., 2007. Intertidal macrofaunal communities in an intensely polluted estuary. Environmental Monitoring and Assessment 134, 397-410.

Carreira, R.S., Wagener, A.L.R., Readman, J.W., 2004. Sterols as markers of sewage contamination in a tropical urban estuary (Guanabara Bay, Brazil): space–time variations. Estuarine, Coastal and Shelf Science 60, 587-598.

Céréghino, R., Giraudel, J.L., Compin, A., 2001. Spatial analysis of stream invertebrates distribution in the Adour–Garonne drainage basin (France), using Kohonen self organizing maps. Ecological Modelling 146, 167-180.

Céréghino, R., Park, Y.S., Compin, A., Lek, S., 2003. Predicting the species richness of aquatic insects in streams using a restricted number of environmental variables. Journal of the North American Benthological Society 22, 442-456.

Chapman, P.M., Wang, F., 2001. Assessing sediment contamination in estuaries. Environmental Toxicology and Chemistry, 20,3-22.

Choi K.W., Lee J.H., Kwok K.W., Leung K.M., 2009. Integrated stochastic environmental risk assessment of the harbour area treatment scheme (HATS) in Hong Kong. Environmental Science and Technology 43, 3705-11.

Chon, T.S., Park, Y.S., Moon, K.H., Cha, E.Y., 1996. Patternizing communities by using an artificial neural network. Ecological Modelling 90, 69-78.

Chon, T.S., Park, Y.S., Park, J.H., 2000. Determining temporal pattern of community dynamics by using unsupervised learning algorithms. Ecological Modelling 132, 151-166.

Chon, T.S., Kwak, I.S., Song, M.Y., Park, Y.S., Cho, H.D., Kim, M.J., Cha, E.Y., Lek, S., 2002. Characterizing the effects of water quality on benthic stream macroinvertebrates in South Korea using a self-organizing mapping model. In: Lee, D. (Ed.), Ecology of Korea. Bumwoo Publishing Company, Seoul, Korea.

Cordeiro, L.G.S.M., Carreira, R.S., Wagener, A.L.R., 2008. Geochemistry of fecal sterols in a contaminated estuary in southeastern Brazil. Organic Geochemistry 39, 1097-1103.

Dauvin, J.C., 2007. Paradox of estuarine quality: benthic indicators and indices, consensus or debate for the future. Marine Pollution Bulletin. 55, 271-281.

Dauvin, J.C., Bachelet, G., Barillé, A.I., Blanchet, H., De Montaudoin, X., Lavesque, N., Ruellet, T., 2009. Development of benthic indicators and index approaches in three main estuaries along the French Atlantic coast (Seine Loire and Gironde) for the implementation of the European Water framework Directive (WFD). Marine Ecology 30, 228-240.

Dawson, C.W., Wilby, R.L., 2001. Hydrological modelling using artificial neural networks. Progress in Physical Geography 25 (1), 80-108.

De-la-Ossa-Carretero J.A., Del-Pilar-Ruso Y., Giménez-Casalduero F., Sánchez-Lizaso J.L., Dauvin J.C., 2012. Sensitivity of amphipods to sewage pollution. Estuarine, Coastal and Shelf Science 96, 129-138.

489 Diaz, R.J., Rosenberg, R., 1995. Marine benthic hypoxia: a review of its ecological effects and
490 behavioral responses of marine macrofauna. Oceanography and Marine Biology Annual
491 Review 33, 245-303.
492 Elías, R., Rivero, M. S., Vallarino, E. A., 2003. Sewage impact assessment based on the
493 composition and distribution of Polychaetes associated to intertidal mussel beds of the
494 Southwestern Atlantic shore. Iheringia 93, 309-318.
495 Elliott, M., Quintino, V., 2007. The estuarine quality paradox environmental homeostasis and
496 the difficulty of detecting anthropogenic stress in naturally stressed areas. Marine Pollution
497 Bulletin 54, 640-645.
498 Grasshoff, K., Ehrhardt, M., Kremling, K., 1983. Methods of Seawater Analysis 2 ed, Verlag
499 Chemie: Weinheim.
500 Ieno E.N., Solan M., Batty P., Pierce G.J., 2006. How biodiversity affects ecosystem
501 functioning: roles of infaunal species richness, identity and density in the marine benthos.
502 Marine Ecology Progress Series 311:263-271.
503 Kawakami, S.K., Montone, R.C., 2002. An efficient ethanol-based analytical protocol to
504 quantify fecal steroids in marine sediments, Journal of the Brazilian Chemical Society 13,
505 226-232.
506 Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. Biological
507 Cybernetics 43, 59-69.
508 Kohonen T., 2001. Self-Organizing Maps, 3rd edn. Springer-Verlag, Berlin, Germany.
509 Kolm, H.E., Schoenenberger, M.F., Piemonte, M.R., Souza, P.S.A., Scühli, G.S., Mucciatto,
510 M.B., Mazzuco, R., 2002. Spatial variation of bacteria in surface waters of Paranaguá and
511 Antonina Bays, Paraná, Brazil. Brazilian Archives of Biology and Technology 45, 27-34.
512 Krapp-Schickel, T., Vader W., 1998: What is, and what is not, *Caprella acanthifera* Leach,
513 1814 (Amphipoda, Caprellidea)? Part 1: The acanthifera-group. Journal of Natural History
514 32, 949-967.
515 Kruk, A., Lek, S., Park, Y.S., Penczak, T., 2007. Fish assemblages in the large lowland Narew
516 River system (Poland): application of the self-organizing map algorithm. Ecological
517 Modelling 203 (1-2), 45-61.
518 Lana, P.C., Marone, E., Lopes, R.M., Machado, E.C., 2000. The subtropical estuarine complex
519 of Paranaguá Bay, Brazil, In Coastal Marine Ecosystems of Latin America, Seeliger, U.,
520 Lacerda, L.D., Kjerfve, B. (eds,), Springer Verlag, NY, USA, 467.
521 Leflaive J., Céréghino R., Danger M., Lacroix G., Ten-Hage L., 2005. Assessment of self-
522 organizing maps to analyze sole-carbon source utilization profiles. Journal of
523 Microbiological Methods 62, 89-102.
524 Lek, S., Guégan, J.F., 2000. Artificial neural networks: application to ecology and evolution.
525 Springer, New York.
526 Lenat, D.R., 1988. Water quality assessment of streams using a qualitative collection method
527 for benthic macroinvertebrates. Journal of the North American Benthological Society 7, 222-
528 233.
529 Lorenzen, C.J., 1967. Determination of chlorophyll and phaeopigments: Spectrophotometric
530 equations. Limnology and Oceanography 12, 343-346.
531 Maldonado, C., Venkatesan, M.I., Phillips, C.R., Bayona, J.M., 2000. Distribution of
532 trialkylamines and coprostanol in San Pedro shelf sediments adjacent to a sewage outfall.
533 Marine Pollution Bulletin 40, 680-687.
534 Marengo E., Gennaro M.C., Robotti E., Rossanigo P., Rinaudo C., Roz-Gastaldi M., 2006.
535 Investigation of anthropic effects connected with metal ions concentration, organic matter
536 and grain size in Bormida river sediments. Analytica Chimica Acta 560, 172-83.

537 Martins, C.C., 2008a. Marcadores orgânicos de contaminação por esgotos sanitários em
538 sedimentos superficiais da baía de Santos, São Paulo. Quimica Nova 31(5), 1008-1014.
539 Martins, C.C., Braun, J.A.F., Seyffert, B.H., Machado, E.C., Fillmann, G., 2010. Anthropogenic
540 organic matter inputs indicated by sedimentary fecal steroids in a large South American
541 tropical estuary (Paranaguá estuarine system, Brazil). Marine Pollution Bulletin 60, 2137-
542 2143.
543 Nelder, J.A., Wedderburn, R.W.M., 1972. Generalized linear models. Journal of the Royal
544 Statistical Society: Series A 135, 370-384.
545 Park, Y.S., Kwak, I.S., Chon, T.S., Kim, J.K., Jørgensen, S.E., 2001. Implementation of
546 artificial neural networks in patterning and prediction of exergy in response to temporal
547 dynamics of benthic macroinvertebrate communities in streams. Ecological Modelling 146,
548 143-157.
549 Park, Y.S., Céréghino, R., Compin, A., Lek, S., 2003. Applications of artificial neural networks
550 for patterning and predicting aquatic insect species richness in running waters. Ecological
551 Modelling 160, 265-280.
552 Park, Y.S., Verdonschot, P.F.M., Chon, T.S., Lek, S., 2003a. Patterning and predicting aquatic
553 macroinvertebrate diversities using artificial neural network. Water Research, 37, 1749-
554 1758.
555 Park, Y.S., Chang, J.B., Lek, S., Cao, W.X.S., Brosse S., 2003b. Conservation strategies for
556 endemic fish species threatened by the Three Gorges Dam. Conservation Biology 17, 1748-
557 1758.
558 Park, Y.S., Chon, T.S., Kwak, I.S., Lek, S., 2004. Hierarchical community classification and
559 assessment of aquatic ecosystems using artificial neural networks. Science of the Total
560 Environment 327, 105-122.
561 Park Y.S., Grenouillet G., Esperance B., Lek S., 2006. Stream fish assemblages and basin land
562 cover in a river network. Science of the Total Environment 365, 140-153.
563 Pearson, T.H., Rosenberg, R., 1978. Macrobenthic succession in relation to organic enrichment
564 and pollution of the marine environment. Oceanography and Marine Biology Annual Review
565 16, 229-311.
566 Peterson, C.H., 1979. Predation, competitive exclusion, and diversity in the soft-sediment
567 benthic communities of estuaries and lagoons. In: R.J. Livingston (ed.), Ecological Processes
568 in Coastal and Marine Systems, 233-264. Plenum, New York.
569 Piraino P., Ricciardi A., Salzano G., Zotta T., Parente E., 2006. Use of unsupervised and
570 supervised artificial neural networks for the identification of lactic acid bacteria on the basis
571 of SDS-PAGEpatterns ofwhole cell proteins. Journal of Microbiological Methods 66, 336-
572 46.
573 Rygg, B., 1985a. Effects of sediment copper on benthic fauna. Marine Ecology Progress Series
574 25, 83–89.
575 Santos, E.S., Carreira, R.S., Knoppers, B.A., 2008. Sedimentary sterols as indicators of
576 environmental conditions in southern Guanabara Bay, Brazil. Brazilian Journal of
577 Oceanography 56 (2), 97113.
578 Schmidt A., Hanson C., Kathilankal J., Law B. E., 2011. Classification and assessment of
579 turbulent fluxes above ecosystems in North-America with self-organizing feature map
580 networks, Agricultural and Forest Meteorology 151 508-520.
581 Smith, V.H., Tilman, G.D., Nekola, J.C., 1999. Eutrophication: impacts of excess nutrient
582 inputs on freshwater, marine, and terrestrial ecossystems. Environmental Pollution 100, 179-
583 196.

584 Snyder, E.B., Robinson, C.T., Minshall, G.W., Rushforth, S.R., 2002. Regional patterns in
585 periphyton accrual and diatom assemblages structure in a heterogeneous nutrient landscape.
586 Canadian Journal of Fisheries and Aquatic Sciences 59, 564-577.
587 Stevenson, R.J., 1997. Scale-dependent determinants and consequences of benthic algal
588 heterogeneity. Journal of the North American Benthological Society 16, 248-262.
589 Strickland, J.L.H., Parsons T.R., 1972. A practical handbook of seawater analysis. Bulletin of
590 the Fisheries Research Board of Canada 167, 311.
591 Suguio, K., 1973. Introdução à sedimentologia. São Paulo, Edgard Blucher LTDA: 317.
592 The Mathworks Inc., 2011. MATLAB Version 7.12. The Mathworks, Inc., Massachusetts.
593 Tison J, Park Y.S., Coste M., Wasson J.G., Ector L., Rimet F., 2005. Typology of diatom
594 communities and the influence of hydro-ecoregions: a study on the French hydrosystem
595 scale. Water Research 39, 3177–88.
596 Tran, L.T., Knight, C.G., O'Neill, R.V., Smith, E.R., O'Connell, M., 2003. Self-organizing
597 maps for integrated environmental. Environmental Management 31, 822-835.
598 Underwood, A.J., Chapman, M.G., Connell, S.D., 2000. Observations in ecology: you can't
599 make progress on processes without understanding the patterns. Journal of Experimental
600 Marine Biology and Ecology 250, 97-115.
601 Vesanto J., Himberg J., Alhoniemi E., Parhankangas J., 1999. Self-organising map in Matlab:
602 the SOM Toolbox. In: Proceedings of the Matlab Digital Signal Processing Conference.
603 Espoo, Finland, 35–40.
604 Volkman, J. K., 1986. A review of sterol markers for marine and terrigenous organic matter.
605 Organic Geochemistry 9, 83-100.
606 Volkman, J. K., 2006. Lipids markers for marine organic matter. The Handbook of
607 Environmental Chemistry 2, 27-70.
608 Ward, J. H., 1963. Hierarchical grouping to optimize an objective function. Journal of the
609 American Statistical Association 58, 236.
610 Wilson, Jr. W.H., 1991. Competition and predation in marine soft-sediment communities.
611 Annual Review of Ecology, Evolution, and Systematics 21, 221-241.
612 Wright J.F., Sutcliffe D.W., Furse M.T., 2000. Assessing the biological quality of fresh waters:
613 RIVPACS and other techniques. Freshwater Biological Association, Ambleside, UK.

CAPITULO II

Modelagem preditiva da fauna bêntica em planícies de maré utilizando redes neurais artificiais supervisionadas

Predictive modeling of benthic fauna in tidal flats using supervised artificial neural networks

Revista pretendida: Ecological Modelling (Ecol. Model.), ISSN (0304-3800), Fator de Impacto (JCR, 2010) = 2.438, Qualis CAPES = Estrato A2.

Faller[1], D. G.; Camargo[1], M. G.

[1]Centro de Estudos do Mar, Universidade Federal do Paraná. Av. Beira Mar s/n, Pontal do Paraná, Paraná, Brasil. CEP: 83255-971. Fone: (41) 3511-8600. E-mail: daiafaller@yahoo.com.br

RESUMO

O objetivo deste estudo foi realizar a predição da presença/ausência de 20 táxons da comunidade bêntica em diferentes condições ambientais utilizando perceptron de múltiplas camadas (MLP). Para isso, foram utilizadas três abordagens diferentes: (1) a previsão dos táxons em conjunto com todas as variáveis (19 variáveis) e com as 14, 8 e 4 variáveis mais vezes selecionadas pelos algoritmos genéticos (2) previsão dos táxons separadamente antes e após seleção de variáveis com algoritmos genéticos e (3) análise de sensibilidade dos modelos com resultados satisfatórios após a seleção das variáveis. A eficiência dos modelos foi analisada através das instâncias corretamente classificadas (*CCI*), estatística K (*K*), especificidade e sensibilidade do modelo. Para todos os táxons em conjunto, no geral, o melhor modelo foi obtido com apenas oito variáveis. Para os táxons individualmente, somente para *T. lineata* a seleção de variáveis prejudicou a eficiência do modelo, os demais responderam positivamente à seleção. O procedimento de seleção auxiliou na eliminação do treinamento dos modelos as variáveis irrelevantes para a predição. A análise conjunta de *CCI* e *K* permitiu uma melhor interpretação do desempenho preditivo dos modelos. Através da análise de sensibilidade foi possível construir curvas da probabilidade de ocorrência dos táxons em relação às variáveis ambientais. Foi possível identificar que a presença da maioria dos táxons estava relacionada com variáveis que denotam enriquecimento orgânico do ambiente. Os modelos desenvolvidos neste estudo demonstraram um considerável poder preditivo, principalmente quando a seleção de variáveis ambientais foi utilizada no seu desenvolvimento.

Palavras-chave: seleção de variáveis, perceptron de múltiplas camadas, modelagem preditiva, análise de sensibilidade, macrofauna bêntica

ABSTRACT

The aim of this study was to predict the presence/absence of 20 benthic taxa under different environmental conditions using multilayer perceptron (MLP). For this, we used three different approaches: (1) predicting all taxa simultaneous with all variables (19 variables) and the 14, 8 and 4 variables (2) predicting individual taxa before and after variable selection with genetic algorithms and (3) sensitivity analysis of models with satisfactory results after the selection of variables. The models efficiency was determined by correctly classified instances (*CCI*), K Statistics (*K*), specificity and sensitivity of the model. For all taxa simultaneous, in general, the best model was obtained with only eight variables. For taxa individually, only for *T. lineata* the models efficiency was lower after the variables selection and the others taxa have responded positively to the selection. The selection procedure helped to eliminate irrelevant variables for the taxa prediction. The simultaneous analysis of *CCI* and *K* allowed a better interpretation of predictive models performance. Through the sensitivity analysis was possible to construct curves of taxa occurrence probability in relation to environmental variables. Thus, we found that the presence of most taxa was related to variables denoting environmental organic enrichment. The results obtained by the models developed in this study were found to have considerable predictive power, especially when the selection of environmental variables was applied in the development of the models.

Keywords: variable selection, multilayer perceptron, predictive modeling, sensitivity analysis, benthic macrofauna

1. INTRODUÇÃO

A distribuição espacial e temporal da macrofauna bêntica é potencialmente condicionada por diferentes fatores ambientais como composição do substrato e disponibilidade de alimento (Brown et al., 1995; Benbow et al., 2003; Boyero, 2003). Os padrões resultantes desta interação permitem uma melhor compreensão da estrutura e dinâmica de uma comunidade (Boyero, 2003). Além dos fatores ambientais intrínsecos à dinâmica do sistema estuarino (Baeta et al., 2005), fatores relacionados aos distúrbios ambientais de causa antrópica também agem sobre os organismos bênticos e podem influenciar fortemente a ocorrência e abundância desses organismos (Marshall et al., 2002).

Devido às mudanças nas condições ambientais, a predição da distribuição das espécies bênticas cada vez mais é importante e relevante para a avaliação e conservação ambiental, bem como a gestão da biodiversidade (Manel et al., 1999; Goethals e De Pauw 2001; Austin 2002; D'heygere et al. 2003). Com isso, na última década, o

desenvolvimento de técnicas de modelagem ganhou maior interesse de ecologistas e estas técnicas passaram a ser aplicadas para identificar a relação entre os táxons e o meio onde vivem (Guisan & Zimmermann, 2000). Grande parte destas técnicas é baseada em análise multivariada clássica (Gabriels et al., 2007) . No entanto, técnicas de mineração de dados têm sido cada vez mais utilizadas neste contexto, tais como as redes neurais artificiais (RNAs) (Walley e Fontama 1998; Hoang et al., 2001; Céréghino et al., 2003; Dedecker et al., 2002, 2004; D'heygere et al., 2006; Gabriels et al., 2007; Kim et al., 2008).

As RNAs são ferramentas atraentes para análise e modelagem de dados ecológicos devido à sua capacidade de lidar com dados não-lineares e à sua adaptabilidade aos dados, que permitem a extração das relações importantes entre o ambiente e as espécies, podendo ser usadas para prever a ocorrência dos organismos em condições ambientais alteradas. As redes *perceptron* de múltiplas camadas (MLP, do inglês *multi-layer perceptron*) com algoritmo *backpropagation* estão entre as RNAs mais populares. Uma rede *backpropagation* (retropropagação) é baseada no procedimento supervisionado e pode ser usada para o desenvolvimento de modelos preditivos.

A seleção de variáveis que melhor descreve a relação das espécies com o ambiente é um importante passo para o desenvolvimento de um modelo eficaz. Em muitos casos, um grande número de variáveis de entrada pode fornecer uma descrição mais precisa do problema, porém resultam em modelos mais complexos que requerem maior tempo de processamento e maior quantidade de dados (Maier e Dandy, 2000; Gevrey et al., 2003; Gabriels et al., 2007). Com isso, para melhorar a precisão e o poder preditivo destes modelos, o número de variáveis explicativas utilizado deve ser reduzido a um número razoável (Harrell et al., 1996). Diversas técnicas foram testadas em estudos com o objetivo de selecionar as variáveis de entrada mais explicativas para diferentes táxons, entre elas a eliminação progressiva das variáveis menos importantes (Walley e Fontama, 1998), análise de sensibilidade (Hoang et al., 2001; Park et al., 2007), algoritmos genéticos (Goethals, 2005; D'heygere et al., 2006; Hoang et al., 2010) e seleção de variáveis passo a passo (*stepwise*, Gabriels et al., 2007).

O objetivo deste estudo foi realizar a predição da presença/ausência de 20 táxons da comunidade bêntica em diferentes condições ambientais utilizando redes neurais artificiais supervisionadas denominadas *perceptron* de múltiplas camadas (MLP). Para isso, foram utilizadas três abordagens diferentes: (1) a previsão de todos os

105 táxons simultaneamente com todas as variáveis (19 variáveis) e com as 14, 8 e 4
106 variáveis mais vezes selecionadas pelos algoritmos genéticos; (2) previsão dos táxons
107 individualmente antes e após seleção de variáveis com algoritmos genéticos e (3)
108 análise de sensibilidade para os modelos com resultados satisfatórios após a seleção das
109 variáveis.

110 ## 2. MATERIAL E MÉTODOS

111 ### 2.1 *Área de Estudo*
112

113 O estudo foi realizado em planícies entremarés não-vegetadas do Canal da
114 Cotinga, um subestuário da Baia de Paranaguá, (Complexo Estuarino de Paranaguá -
115 25°30'S, 48°25'W), (Fig.1). Os rios que fazem parte da margem sul da baia, como o
116 Maciel, Guaraguaçú e Itiberê e que deságuam na Cotinga, recebem grande parte dos
117 efluentes produzidos na cidade e porto de Paranaguá (Lana et al., 2001). Entre os rios da
118 região, o Itiberê pode ser considerado uma das principais fontes pontuais de
119 contaminação na Cotinga, evidenciado por estudos realizados no local que encontraram
120 elevadas concentrações de indicadores orgânicos de poluição, como coliformes fecais
121 na coluna d'água (Kolm et al., 2002) e esteróis fecais no sedimento (Martins et al.,
122 2010). As concentrações de contaminantes no canal são dispersas e diluídas a partir da
123 região interna e mediana em direção à sua desembocadura, formando um gradiente de
124 poluição. Planícies não-vegetadas das duas margens do canal foram selecionadas,
125 totalizando 107 locais de amostragem ao longo da Cotinga.
126

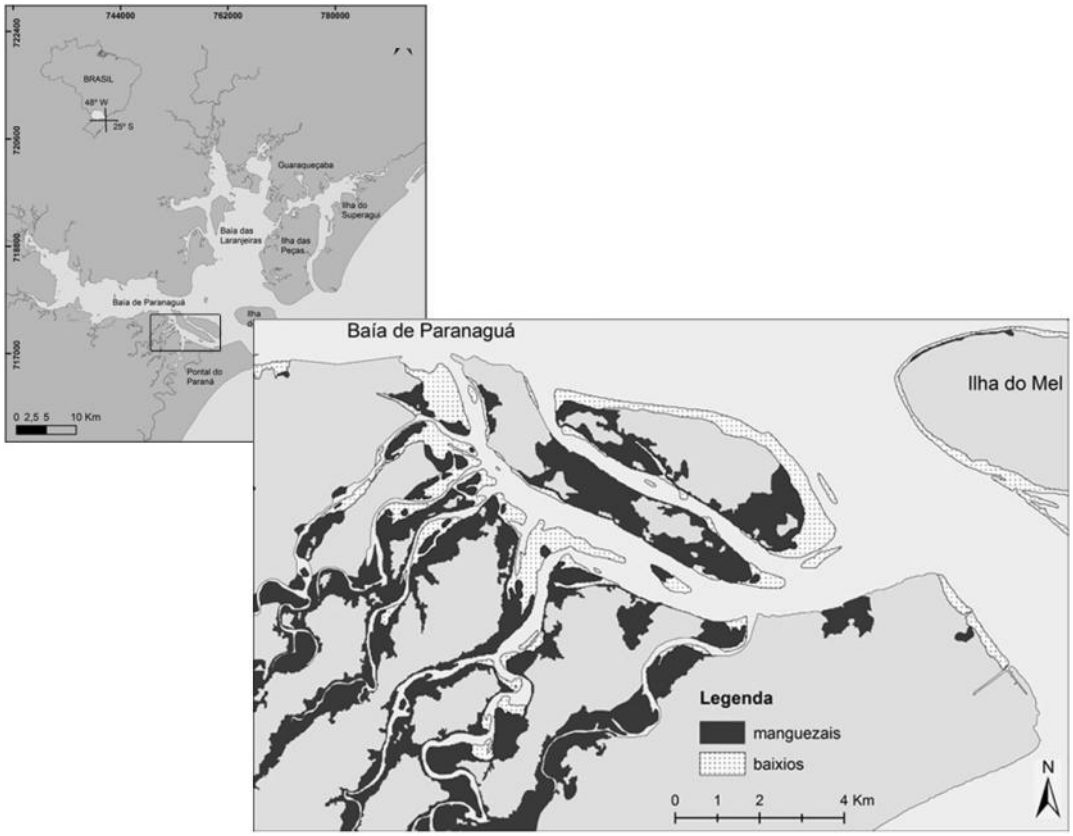


Fig.1. Complexo Estuarino de Paranaguá (CEP) com o Canal da Cotinga no detalhe.

2.4 Conjunto de dados

Em cada local de amostragem foram coletadas duas réplicas com *corers* com diâmetro de 10 cm para a análise da macrofauna bêntica. As amostras biológicas foram fixadas com formol-aldeído (4%), lavadas em peneiras de 0,5 mm, coradas com Rosa de Bengala, preservadas em álcool (70%) e identificadas até o nível de família, totalizando 49 famílias. Entre estas, foram selecionadas as famílias com mais de 15% de ocorrência, que foram identificadas até o menor nível possível. Os 20 táxons com maior ocorrência foram utilizados no desenvolvimento dos modelos do SOM. São eles: *Anomalocardia brasiliana* (35,5%), *Bulla* sp. (64,5%), Capitellidae (72,9%), *Capitella* sp. (23,4%), *Caprella scaura* (14%), *Glycinde multidens* (87,9%), *Heleobia australis* (63,6%), *Laeonereis culveri* (31,8%), *Neanthes succinea* (13,1%), Oligochaeta sp1 (34,6%), Orbiniidae (51,4%), Ostracoda (39,3%), *Polydora* sp. (42,1%), *Prionospio* sp. (44%), *Sigambra* sp. (82,2%), *Sternaspis* sp. (36,4%), *Streblospio benedicti* (61,7%), *Tagelus* sp. (53,3%), *Tellina lineata* (59,8%) e Tubificidae (85%).

Juntamente com a fauna, 19 variáveis ambientais foram utilizadas no desenvolvimento dos modelos (Tabela 1). Em campo, foi realizada a medição da

146 camada redox e amostras de água intersticial foram coletadas para a determinação do
147 pH e salinidade (pHmetro e refratômetro manual, respectivamente). Amostras de
148 sedimento foram coletadas para determinar: fósforo total e nitrogênio total (Grasshoff et
149 al., 1983); carbono orgânico total (Strickland e Parsons, 1972); clorofila-*a* e feoftina
150 (Lorenzen, 1967), porcentagens de cascalho, areia, silte e argila, (Suguio, 1973);
151 carbonato de cálcio, matéria orgânica e esteróis totais (Kawakami e Montone, 2002).

152 Foram considerados cinco esteróis para identificar diferentes fontes de matéria
153 orgânica: coprostanol e estigmasterol, como indicadores de contaminação por esgotos
154 (Maldonado et al., 2000); colesterol para origem aquática, por estar presente no fito e
155 zooplâncton (Volkman, 1986); brassicasterol para identificar matéria orgânica de
156 origem marinha e estigmasterol para identificar matéria orgânica de origem terrestre
157 (Volkman, 2006). Devido à inviabilidade em estimar a concentração de esteróis em
158 todos os pontos, optou-se por realizar as amostragens em 31 pontos selecionados de
159 acordo com a proximidade à cidade de Paranaguá e à desembocadura dos rios. A
160 extrapolação para os demais pontos foi realizada através da média entre pontos
161 vizinhos.

162

163 **Tabela 1**

164 Variáveis de entrada usadas para o desenvolvimento dos modelos em conjunto com a média, desvio
165 padrão (DP), valores mínimos (Min.) e máximos (Max.).

Variável	Abreviação	Unidade	Média	DP	Min.	Max.
Salinidade	SAL	-	24,63	5,35	2	32
pH	PH	-	7,28	0,27	6,58	7,92
Camada Redox	RED	cm	1,15	1,15	0,1	5,97
Coprostanol	COP	$\mu g.g^{-1}$	0,38	0,46	0	2,04
Epicoprostanol	EPI	$\mu g.g^{-1}$	0	0,01	0	0,08
Colesterol	COL	$\mu g.g^{-1}$	4,37	2,89	0,74	15,5
Brassicasterol	BRA	$\mu g.g^{-1}$	2,2	1,5	0,23	8,22
Estigmasterol	EST	$\mu g.g^{-1}$	2,7	1,62	0,27	10,40
Cascalho	CAS	%	1,72	5,95	0	52,28
Areia	ARE	%	84,98	11,4	33,76	97,89
Silte	SIL	%	10,21	8,9	0	44,56
Argila	ARG	%	3,09	3,01	0	21,31
Carbono	COT	$mg.g^{-1}$	14,55	11,7	0	46,7
Clorofila	CHL	$mg.g^{-1}$	17,39	23,2	0	157,95
Feoftina	FEO	$mg.g^{-1}$	17,88	23,5	0	178,75
Nitrogênio Total	NT	$mg.g^{-1}$	2,09	1,11	0,07	4,41
Fósforo Total	PT	$mg.g^{-1}$	0,03	0,02	0	0,09
Matéria Orgânica	MO	%	4,36	2,17	0,49	13,05
Carbonato de Cálcio	CaCO3	%	4,11	3,98	0,48	32,23

2.2 *Perceptron de múltiplas camadas (MLP)*

O treinamento da rede MLP foi baseado nos princípios do algoritmo *backpropagation* (Rumelhart et al., 1986; Lippmann, 1987; Kung, 1993). A rede é composta por três tipos de camadas de neurônios: uma camada de entrada, uma ou mais camadas ocultas (ou intermediárias) e uma camada de saída, cada camada incluindo um ou mais neurônios (nós). Nesta rede, cada nó de uma camada está conectado a todos os nós da camada seguinte, porém não há conexão entre os nós da mesma camada e nem conexões de retorno (*feedback*). As camadas ocultas são utilizadas para auxiliar a rede a lidar com a não-linearidade dos dados.

Com exceção dos neurônios de entrada, que só conectam um valor de entrada com seus valores de peso associados, a entrada líquida de cada neurônio a_j é a soma de todos os valores de entrada x_n multiplicado por seu peso w_{jn}. O termo *bias* z_j pode ser considerado como um peso suplementar e normalmente equivale a 1:

$$a_j = \sum w_{ji} x_i + z_j \quad (1)$$

Os valores de saída y_j podem ser calculados pela alimentação da entrada líquida para a função de transferência do neurônio:

$$y_j = f(a_j) \quad (2)$$

O treinamento da rede consistiu no ajuste dos pesos e *biases* usando o algoritmo *backpropagation*. Para cada vetor de entrada, um vetor de saída foi calculado pela rede e o erro calculado para as saídas foi comparando aos valores reais dos táxons. Este procedimento é repetido até que os erros tornem-se suficientemente pequenos ou um número pré-definido máximo de interações (épocas) seja alcançado. Tanto os pesos quanto os *biases* são utilizados para diminuir o erro e são pré-estabelecidos antes do treinamento da rede. Os valores dos pesos e *biases* foram inicialmente configurados para serem números aleatórios pequenos. Uma descrição mais detalhada do procedimento pode ser encontrada em Lek e Guégan (1999).

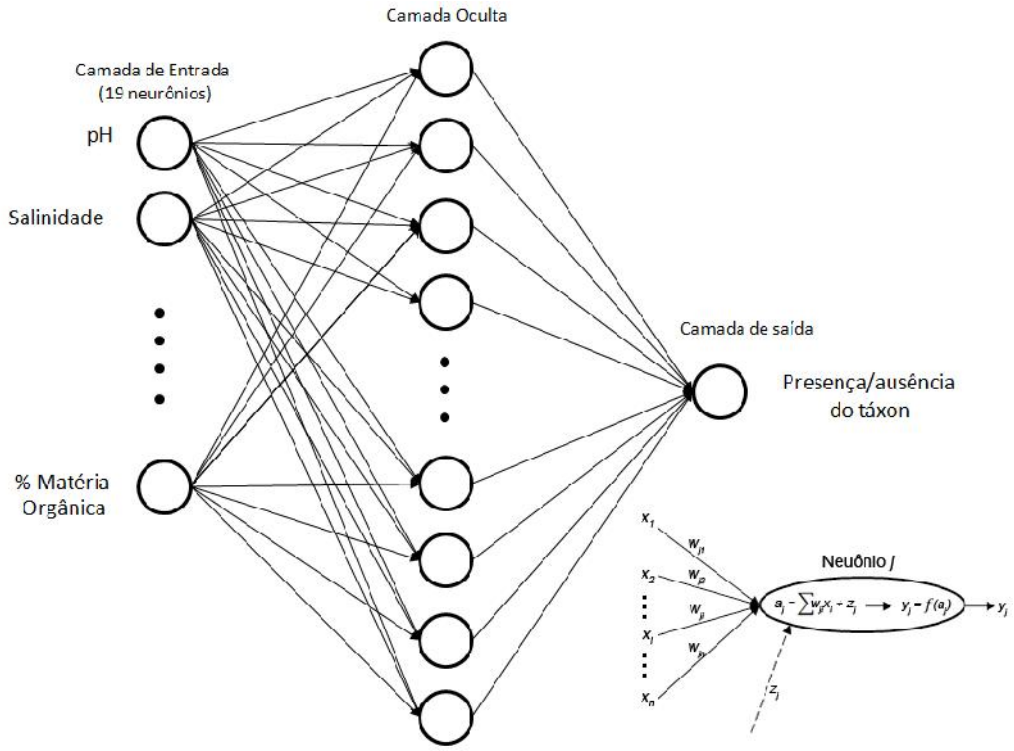


Fig.2. Rede neural MLP esquemática de três camadas com uma camada de entrada, uma camada oculta e uma camada de saída.

Para obter a melhor configuração do modelo preditivo, diferentes arquiteturas de rede foram testadas modificando a quantidade de camadas ocultas e seus neurônios ([5 10], [10 15], [10 10], [10], [15], [25]), com dois tipos de funções sigmóides: a função de transferência logarítmica e a tangencial. Além disso, duas variações do algoritmo *backpropagation* foram usadas neste estudo: os algoritmos de gradiente descendente com aprendizado adaptativo e o de Levenberg-Marquardt (Hagan et al., 1996; Dedecker et al., 2002). Após testes, foi utilizado o algoritmo de gradiente descendente com aprendizado adaptativo e funções de transferência *log*-sigmóide para a camada oculta e para a camada de saída. Todas as redes foram construídas com uma camada oculta com 25 neurônios. Como critérios de parada do treinamento foram estipuladas 1000 épocas ou erro médio quadrado (*mse*) igual a 0,01.

Duas abordagens diferentes foram utilizadas para o treinamento e validação da rede. Primeiramente, foi construída uma rede com 20 neurônios de saída, variando somente a quantidade de neurônios de entrada, com 19, 14, 8 e 4 neurônios (correspondente à quantidade de variáveis ambientais utilizadas no treinamento). Em um segundo momento, foi construída uma rede com somente um neurônio de saída, correspondente a um determinado táxon submetido à rede. Neste caso, a quantidade de

neurônios de entrada também variou, sendo formada por 19 neurônios (todas variáveis de entrada) no treinamento inicial e com as variáveis selecionadas pelos algoritmos genéticos (detalhes em: Tópico 2.4).

Como as variáveis abrangem diferentes faixas de valores e a fim de assegurar que todas as variáveis recebam o mesmo grau de importância durante o processo de treinamento é recomendado que estas passem por uma padronização. Portanto, antes dos dados serem submetidos à rede, todas as variáveis de entrada foram redimensionadas para se restringirem ao intervalo entre -1 e 1 (Dedecker et al., 2004; 2005), enquanto a saída consistiu de 0 (ausência) ou 1 (presença).

2.3 Algoritmos genéticos (GA)

A escolha das variáveis apropriadas em um conjunto de dados é importante porque melhora o desempenho do modelo (Goethals et al., 2007). A seleção das variáveis de entrada para a MLP foi realizada através da combinação da rede com algoritmos genéticos (GA). Portanto, os GAs foram escolhidos na seleção das variáveis devido a capacidade destes algoritmos de proporcionarem uma pesquisa sistemática no universo amostral que seleciona as variáveis que melhor explicam os táxons modelados. Este procedimento auxilia na diminuição da complexidade dos modelos a serem treinados, pois elimina variáveis irrelevantes ao táxon modelado.

O GA é uma técnica baseada na "sobrevivência do mais forte" (Goldberg, 1989) e consiste na busca de soluções ótimas ou próximas do ótimo. O processo de pesquisa é executado em quatro etapas distintas: inicialização, seleção, cruzamento (*crossover*) e mutação (Wong e Tan, 1994). Os GAs são formados por uma população de diferentes soluções concorrentes que evoluem e tendem a convergir em uma solução ideal. Esta solução é representada por um cromossomo, que por sua vez é formado por vários genes. A população inicial é formada aleatoriamente e após sucessivas iterações (gerações), os cromossomos iniciais vão dando lugar aos cromossomos mais fortes obtidos através de suas reproduções, formando novas gerações. Tais gerações podem ser formadas por cruzamento, seleção e mutação.

Existem inúmeras variações de algoritmos genéticos. Neste estudo, porém, os GAs de Goldberg (Goldberg, 1989) foram aplicados para encontrar um melhor conjunto de variáveis de entrada relevante para a previsão da presença/ausência dos táxons da

fauna bêntica. Os cromossomos foram formados por 19 genes, cada um representando uma variável de entrada, com codificação binária. Isso significa que uma determinada variável foi selecionada (representado por '1') ou não (representado por '0').

A taxa de cruzamento (*crossover*) foi fixada com probabilidade de 60%, enquanto que a mutação ocorreu com probabilidade de 3%. A população inicial foi constituída por 20 cromossomos que foram evoluindo ao longo de 45 gerações. Estes parâmetros foram definidos depois de testes preliminares, onde os valores de cruzamento, mutação e gerações foram modificados até encontrar o valor ideal. Através do uso da abordagem *wrapper,* foi possível utilizar os GAs em conjunto com as MLPs. Nesta abordagem, os GAs conduzem a busca por subconjuntos confiáveis no espaço de treinamento utilizando o algoritmo de indução (neste caso, a MLP) como parte da função de avaliação destes subconjuntos (Kohavi e John, 1997).

2.4 Análise de Desempenho dos Modelos

Para as MLPs e GAs a separação dos conjuntos de treinamento e teste foi realizada através da validação cruzada (Dedecker et al., 2004; Goethals et al., 2007) com 10 subconjuntos. Nesta validação, os dados originais foram aleatoriamente divididos em 10 subconjuntos de tamanho aproximadamente igual. Destes, um único conjunto é mantido para validação do modelo e os subconjuntos restantes são usados como dados de treinamento. Este procedimento foi utilizado para estimar o erro dos modelos e assim comparar as performances com e sem a seleção prévia de variáveis.

A eficiência de predição dos modelos foi testada através de quatro indicadores de medida de desempenho: as instâncias corretamente classificadas (*CCI*) (Dedecker et al., 2004), estatística *Kappa* (*K*) de Cohen (Cohen, 1960; Dedecker et al., 2004), especificidade e sensibilidade do modelo (Willems et al., 2008). Para isto, foram identificados através da matriz de confusão (Fielding e Bell, 1997) os casos de verdadeiro positivo (*VP*), falso positivo (*FP*), falso negativo (*FN*) e verdadeiro negativo (*VN*) previstos por cada modelo (Tabela 1).

Tabela 2

Matriz de confusão

		Predito	
		Presente	Ausente
Atual	Presente	*VP*	*FN*

	Ausente	*FP*	*VN*

Neste estudo, para que o desempenho do modelo seja considerado confiável, foram utilizados valores de *CCI* superiores à 70% e *K* superior à 0,3, enquanto que modelos que apresentaram valores inferiores a estes limites foram considerados de desempenho irrelevante (D'heygere et al., 2006). Gabriels et al. (2007) avalia os valores de *K* da seguinte forma: 0-0,2 pobre; 0,2-0,4 razoável; 0,4-0,6 moderada; 0,6-0,8 bom; 0,8-1,0 excelente.

A comparação dos resultados dos modelos para todos os táxons simultaneamente, porém com a modificação da quantidade de variáveis de entrada foi realizada através do teste-*t* independente. Para a comparação entre as respostas dos modelos construídos antes e após a seleção com GAs foram utilizados testes-*t* pareados. Para o desenvolvimento do modelo foi utilizado o NNET Toolbox do MATLAB (The Mathworks, 2011).

2.5 *Análise de Sensibilidade*

A análise de sensibilidade foi realizada para avaliar a contribuição de cada variável de entrada em um determinado táxon, com base no modelo desenvolvido por Lek et al. (1995), denominado de método '*profile*'. Este modelo ilustra diretamente a relação entre as variáveis de entrada e a variável dependente (táxon). A ideia geral do modelo é estudar cada variável de entrada, sucessivamente, enquanto as demais variáveis são bloqueadas em valores fixos. As variáveis utilizadas para construir as curvas de sensibilidade foram aquelas que previamente foram escolhidas pelos GAs para cada táxon analisado. Este procedimento elimina da construção dos modelos da análise de sensibilidade as variáveis irrelevantes para determinado táxon e somente variáveis identificadas como influentes passam pelo treinamento.

Primeiramente, uma variável foi escolhida e em seguida os pontos de amostragens foram divididos em intervalos (escalas) iguais entre os valores mínimos e máximos desta variável. As demais variáveis são definidas primeiro em seus valores mínimos e então a rede é treinada e gera um resultado. Este procedimento é repetido para o primeiro quartil, mediana, terceiro quartil e máximo, gerando cinco resultados preditivos, que são reduzidos à mediana. O mesmo ocorre para todas as demais variáveis selecionadas pelos GAs para o táxon em estudo. Em um segundo momento,

após determinar os valores finais de cada modelo, é possível traçar um perfil da probabilidade de ocorrência do táxon para cada variável nas diferentes escalas estabelecidas. Neste trabalho, os dados foram selecionados em oito escalas diferentes e treinados na rede seguindo o procedimento supracitado (Tabela 3).

Tabela 3

Classes de escalas utilizadas para a construção das curvas da análise de sensibilidade para os táxons que apresentaram valores de predição satisfatórios através dos modelos da rede *perceptron* de múltiplas camadas (MLP).

Classe	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8** Max.
SAL	2,00	5,75	9,50	13,25	17,00	20,75	24,50	28,25
PH	6,58	6,75	6,92	7,08	7,25	7,42	7,59	7,75
RED	0,10	0,83	1,57	2,30	3,03	3,77	4,50	5,23
COP	0,00	0,26	0,51	0,77	1,02	1,28	1,53	1,79
EPI	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07
COL	0,74	2,59	4,43	6,28	8,12	9,97	11,81	13,66
BRA	0,23	1,23	2,23	3,23	4,23	5,22	6,22	7,22
EST	0,27	1,54	2,80	4,07	5,34	6,60	7,87	9,13
CAS	0,00	6,54	13,07	19,61	26,14	32,68	39,21	45,75
ARE	33,76	41,78	49,80	57,81	65,83	73,84	81,86	89,87
SIL	0,00	5,57	11,14	16,71	22,28	27,85	33,42	38,99
ARG	0,00	2,66	5,33	7,99	10,66	13,32	15,98	18,65
COT	0,00	5,84	11,68	17,51	23,35	29,19	35,03	40,87
CHL	0,00	19,74	39,49	59,23	78,98	98,72	118,46	138,21
FEO	0,00	22,34	44,69	67,03	89,38	111,72	134,07	156,41
NT	0,07	0,61	1,15	1,70	2,24	2,78	3,32	3,87
PT	0,00	0,02	0,03	0,04	0,05	0,06	0,07	0,08
MO	0,49	2,06	3,63	5,20	6,77	8,34	9,91	11,48
CaCO3	0,48	4,45	8,42	12,38	16,35	20,32	24,29	28,26

3. RESULTADOS

3.1 Variáveis selecionadas pelos algoritmos genéticos

As variáveis selecionadas pelos algoritmos genéticos (GA) de Goldberg em conjunto com MLP estão representados na Tabela 4. Após aplicação dos GAs, as

maiores reduções de variáveis ocorreram para os táxons *N. succinea* e Orbiniidae onde somente três variáveis foram selecionadas como importantes. Para *N. succinea* foram selecionadas as variáveis brassicasterol, carbono orgânico total e matéria orgânica, enquanto para Orbiniidae foram selecionados estigmasterol, argila e fósforo total.

Os táxons que mantiveram a maior quantidade de variáveis explicativas foram Oligochaeta sp1 e Ostracoda com nove variáveis selecionadas como importantes para cada. Os demais táxons tiveram entre cinco e oito variáveis selecionadas pelos GAs (Tabela 4). Para *Sigambra* sp. não foi encontrado nenhum subconjunto de variáveis explicativas, por isso, análises individuais para o táxon, após a seleção das variáveis, não foram realizadas.

Salinidade e estigmasterol foram as variáveis mais vezes selecionadas, sendo selecionadas para 10 táxons, seguidas de pH e coprostanol selecionadas nove vezes como explicativas. Já argila e feoftina foram as variáveis menos vezes selecionadas, somente identificadas como importante em três e quatro táxons, respectivamente. As demais variáveis foram selecionadas entre cinco e oito vezes.

337 **Tabela 4**

338 Variáveis selecionadas pelos algoritmos genéticos (GA) em conjunto com a rede *perceptron* de múltiplas camadas (MLP) considerando a presença/ausência de cada táxon
339 individualmente.

Táxon	SAL	PH	RED	COP	EPI	COL	BRA	EST	CAS	ARE	SIL	ARG	COT	CHL	FEO	NT	PT	MO	$CaCO_3$
Anomalocardia brasiliana			x	x	x				x		x				x				x
Bulla sp.		x	x	x					x						x		x		
Caprella scaura	x							x	x	x				x		x	x		
Capitelidae	x			x				x		x									x
Capitella sp.		x			x	X					x		x					x	
Glycinde multidens	x		x	x			x	x	x							x			
Heleobia australis		x	x	x						x	x			x		x			x
Laeonereis culveri				x	x	X			x				x	x				x	x
Neanthes succinea							x						x					x	
Oligochaeta sp1	x	x			x	X					x	x				x	x	x	
Orbiniidae								x				x					x		
Ostracoda	x	x	x		x	X		x	x	x									x
Polydora sp		x	x			X		x								x			
Prionospio sp	x	x				X	x	x											x
Streblospio benedicti	x			x				x						x	x		x		
Sigambra sp.	x	x	x	x	x	X	x	x	x	x	x	x	x	x	x	x	x	x	x
Sternaspis sp.					x		x				x		x			x	x		
Tagelus sp	x	x				X	x							x					
Tellina lineata	x			x	x		x	x	x		x								x
Tubificidae							x				x		x		x				x

3.2 Predição da presença/ausência da fauna bêntica

Baseados nos valores de *CCI* e *K,* entre os modelos construídos, para todos os táxons simultaneamente, o melhor modelo foi o construído com as oito variáveis mais vezes selecionadas pelos GAs, com *CCI* de 73% e *K* de 0,46, enquanto que o modelo com quatro variáveis apresentou desempenho preditivo significativamente inferior aos demais modelos (teste-*t* independente, $p<0,05$), com valores de *CCI* e *K* de 67,4% e 0,35, respectivamente (Tabela 5).

A especificidade dos modelos (taxa de predição média negativa) foi significativamente menor (teste-*t* independente, $p<0,05$) para o modelo com quatro variáveis (63,94%) em relação aos demais modelos. A melhor especificidade foi registrada no modelo construído com 14 variáveis (74,32%). Para a sensibilidade (taxa de predição de valores positivos verdadeiros), o melhor resultado foi obtido para o modelo com oito variáveis (72,87%) enquanto que o modelo com 14 variáveis mostrou o menor valor de sensibilidade (67,90%), porém não houve diferenças significativas entre os resultados.

Tabela 5

Indicadores de desempenho dos modelos, da rede *perceptron* de múltiplas camadas (MLP), construídos para a previsão da presença e ausência dos 20 táxons em conjunto utilizando 19 (todas), 14, 8 e 4 variáveis. *CCI* - instâncias corretamente classificadas; *K* - estatística *Kappa* de Cohen.

Medidas de desempenho	19 variáveis	14 variáveis	8 variáveis	4 variáveis
CCI	71,40	71,10	73,00	67,40
K	0,43	0,42	0,46	0,35
Especificidade	71,52	74,32	73,13	63,94
Sensibilidade	71,29	67,90	72,87	71,54

Quando gerados modelos separados para cada táxon, antes da seleção das variáveis, a média geral do *CCI* foi 68,45%, enquanto que após a seleção foi 70,47%, porém não houve diferença significativa entre os valores médios antes e após seleção. Os valores de *CCI* para 10 táxons antes da seleção foram inferiores ao limite. O menor valor de *CCI* foi obtido para Orbiniidae com valor médio de 49,2% (Fig.3). Entre os táxons que superaram 70%, os modelos construidos para *G. multidens* (85,2%),

Oligochaeta sp1 (78,6%), *N. succinea* (77%) e *S. benedicti* (76,4%) registraram os melhores resultados.

Após a seleção das variáveis, 13 táxons superaram 70%, com destaque para *G. multidens* que apresentou valores de *CCI* de 88,6%, assim como N. *succinea*, Oligochaeta sp1 e *S. benedicti* que tiveram valores de *CCI* de 83,8%. Todos os táxons apresentaram *CCI* superiores aos observados antes da seleção, com aumento médio de 6,8%, com exceção de *T. lineata* que respondeu negativamente à seleção de variáveis, com *CCI* 5% inferior. Os táxons com aumento mais significativos (teste-*t* pareado, $p<0,01$) foram *L. culveri* e *Tagelus* sp. com aumento médio de 11,2%, seguidos de *Capitellidade* (9,2%), *Prionospio* sp. (9%) e *H. australis* (8,8%). Os valores de *CCI* para *Capitella* sp., *C. scaura* e *G. multidens* apresentaram os aumentos menos expressivos, com 0,8% para *Capitella* sp. e 3,4% para os outros dois táxons.

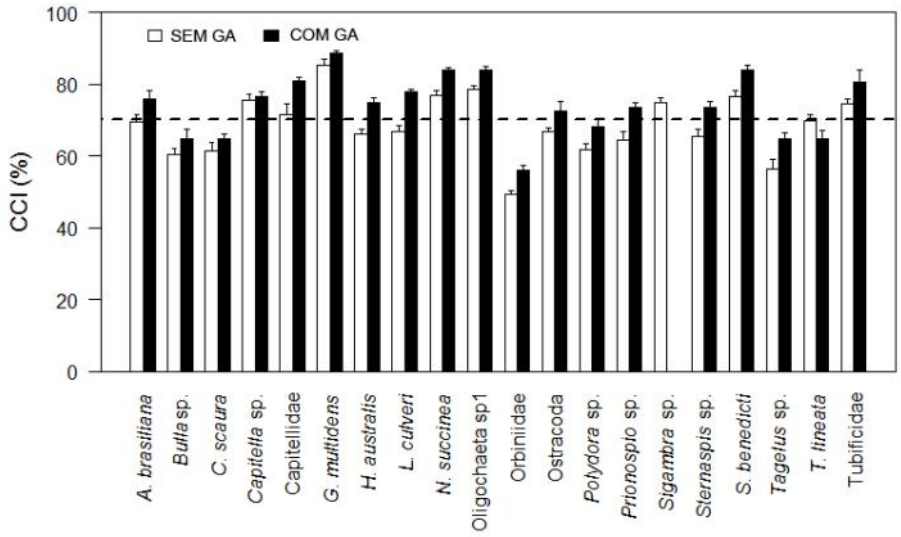


Fig.3 - Instâncias corretamente classificadas (*CCI*) dos modelos das redes *perceptron* de múltiplas camadas (MLP) para os 20 táxons analisados antes (SEM GA) e após seleção das variáveis com algoritmos genéticos (COM GA); Linha pontilhada determina o limite inferior para que os valores do *CCI* dos modelos construídos possam ser considerados confiáveis.

A média geral do *K* antes da seleção das variáveis foi de 0,24, ficando abaixo do limite inferior de 0,3, enquanto que após a seleção das variáveis o valor médio de *K* foi 0,34. Antes da seleção, somente Oligochaeta sp1 (0,49), *S. benedicti* (0,45), *T. lineata* (0,34), *Capitella* sp. (0,3) e *A. brasiliana* (0,3) obtiveram valores de *K* superiores ao limite. Os modelos que apresentaram valores de *K* abaixo de 0,2 podem ser considerados irrelevantes e com baixo poder preditivo (Fig.4). Os menores valores

de *K* foram registrados para Orbiniidae e Tubificidae com valores de 0,04 e 0,1, respectivamente (Fig.4). Com a seleção, todos os modelos desenvolvidos apresentaram aumento nos valores de *K*, com exceção de *T. lineata* e *Capitella* sp., que obtiveram valores inferiores após a seleção. Dos 20 táxons analisados, 12 táxons passaram a apresentar valores de *K* superiores à 0,3 (Fig.4).

Entre os modelos construídos, antes e após seleção, a diferença mais significativa (teste-*t* pareado, $p<0,01$) dos valores de *K* foram obtidos para *L. culveri* e Tubificidae com aumento de 0,2 em ambos, seguido dos táxons Capitellidae, *Tagelus* sp. e *Sternaspis* sp. com aumento de 0,19, 0,17 e 0,16, respectivamente. Para Orbiniidae, mesmo após seleção das variáveis, não foi alcançado valores de *K* superiores à 0,2 e os resultados para o táxon foram considerados irrelevantes.

Quando analisados os valores de *CCI* e *K* em conjunto foi possível identificar que ambas as medidas tiveram o mesmo padrão e os valores de *CCI* foram altos quando os valores de *K* também foram significantes, ou seja, no geral os valores de *CCI* foram superiores a 70% (Fig. 3) quando os valores de *K* (Fig.4) também foram superiores ao limite de 0,3, principalmente após a seleção das variáveis. Porém em alguns táxons, como *N. succinea*, foram registrados valores de *CCI* elevados enquanto os valores de *K* foram inferiores ao limite.

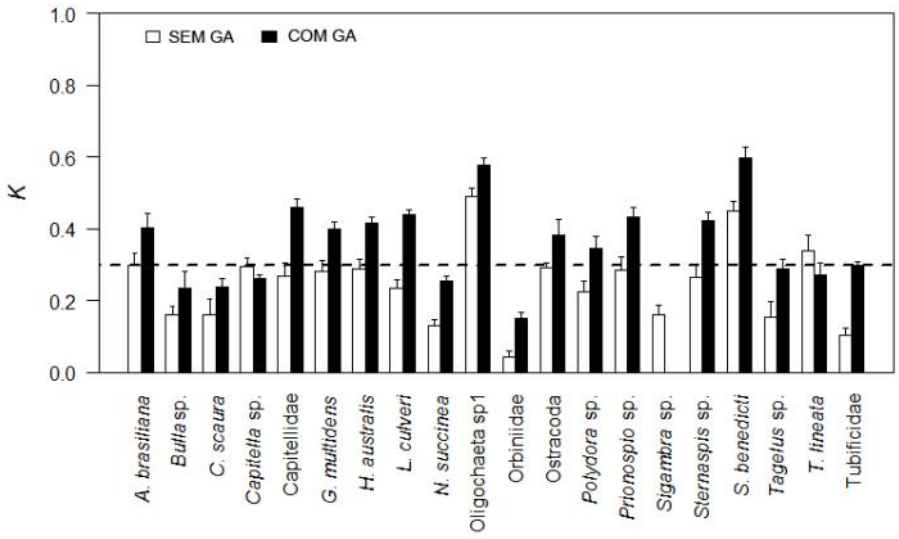


Fig. 4 - *K* de Cohen para os modelos das redes *perceptron* de múltiplas camadas (MLP) para os 20 táxons analisados antes (SEM GA) e depois da seleção das variáveis com algoritmos genéticos (COM GA). Linha pontilhada determina o limite inferior para que os valores de *K* dos modelos construídos possam ser considerados confiáveis.

A sensibilidade, ou , registrou valores médios semelhantes antes e após seleção das variáveis pelos GAs, com valores de 63,5% e 63,8%, respectivamente. O menor valor de sensibilidade antes da seleção foi registrado para *N. succinea* com valores médios de 10,6%, assim como após a seleção com 17,3%. O mesmo ocorreu com os maiores valores de sensibilidade, com *G. multidens* apresentando os maiores valores de sensibilidade antes (92,8%) e após (95,3%) o uso dos GAs na seleção das variáveis (Fig. 5).

Os maiores aumentos dos valores de sensibilidade antes e após o procedimento de seleção de variáveis ambientais foram encontradas para *L. culveri* com aumento médio de 15,4% (teste-*t* pareado, $p<0,01$), seguido de aumento de aproximadamente 11% para Orbiniidae, *Sternaspis* sp. e *Tagelus* sp (teste-*t* pareado, $p<0,05$). Para *Capitella* sp. e *T. lineata* os valores foram inferiores após seleção apresentando diminuição de aproximadamente 7%.

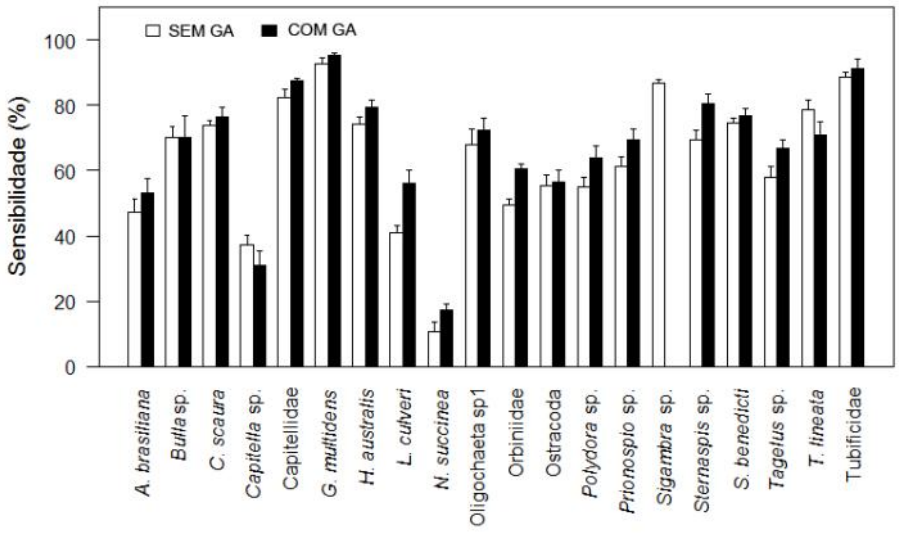


Fig. 5 - Sensibilidade dos modelos das redes *perceptron* de múltiplas camadas (MLP) para os 20 táxons analisados antes (SEM GA) e depois da seleção das variáveis com algoritmos genéticos (COM GA).

A especificidade dos modelos (taxa de predição negativa verdadeira) apresentou valores significativamente maiores (teste-*t* pareado, $p<0,05$) após a seleção das variáveis, com valores médios de 56,7% antes e 67,5% após seleção. Os menores valores de especificidade antes da seleção das variáveis foram encontrados para Tubificidae (6,8%) e *Sigambra* sp. (18,2%). Após seleção, Tubificidae continuou apresentando o menor valor de especificidade com média de 29,3%. Os maiores valores de especificidade antes da seleção foram encontrados para *N. succinea* com 88% e

Capitella sp. (87,8%). Os valores de especificidade atingiram valores médios superiores a 80% para *A. brasiliana*, Oligochaeta sp1 e *Sternaspis* sp.. Após a seleção das variáveis, *N. succinea* continuou sendo o táxon com maiores valores atingindo 94,1%, seguido de Oligochaeta sp1 (90%), *Capitella* sp. (88%), *L. culveri* (87,9%), *Sternaspis* sp. (85,8%) e Ostracoda (82%) (Fig.6).

As variações mais significativas antes e após a seleção das variáveis ambientais (teste-*t* pareado, $p<0,01$) foram para Tubificidae com aumento de 22,5% na taxa de acerto. Para Capitellidae e *S. benedicti* também foi detectado um aumento significativo nos valores de especificidade com valores de 20,2% e 17,9%, respectivamente.

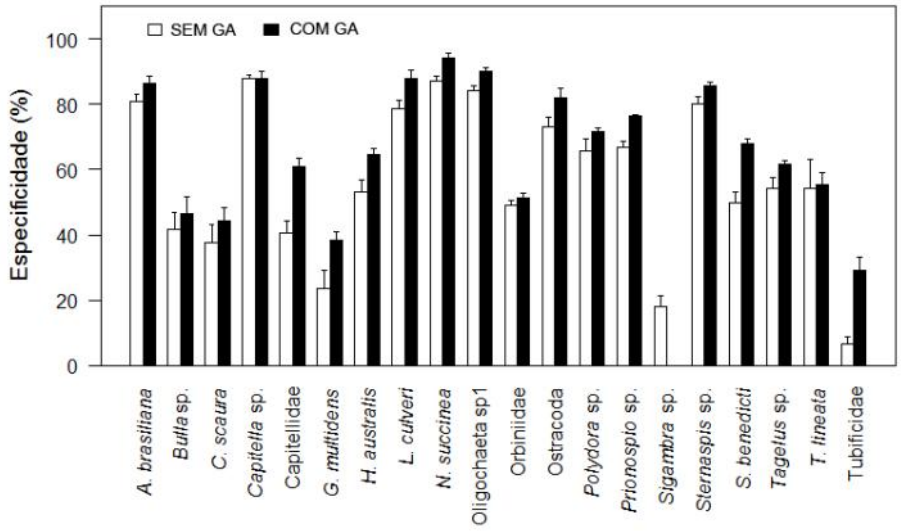


Fig. 6 - Especificidade dos modelos das redes *perceptron* de múltiplas camadas (MLP) para os 20 táxons analisados antes (SEM GA) e depois da seleção das variáveis com algoritmos genéticos (COM GA).

3.3 Análise de sensibilidade

Os modelos dos doze táxons que apresentaram maior potencial preditivo foram submetidos à análise de sensibilidade, utilizando as variáveis selecionadas pelos GAs, para elucidar a direção dos efeitos das variáveis ambientais nas predições da fauna (Fig.7). Entre as variáveis que foram utilizadas na análise de sensibilidade, a porcentagem de carbonato de cálcio e a concentração de epicoprostanol foram selecionadas como influentes na probabilidade de ocorrência de quatro táxons, seguidos de porcentagem de areia e concentração de estigmasterol, detectadas como influentes em três táxons. As demais variáveis foram selecionadas uma ou duas vezes, com

exceção da camada redox e concentração de coprostanol que não foram selecionadas, pois não houve ligação entre a mudança na probabilidade de ocorrência dos táxons e as variações destas variáveis.

A maioria das variáveis mostrou uma relação inversa com a probabilidade de ocorrência da fauna, ou seja, quanto maiores foram os valores das variáveis, menor a probabilidade de ocorrência de determinado táxon (Fig.7). Por exemplo, Tubificidae apresentou sua presença associada às concentrações de feoftina e carbono orgânico total. Para concentrações de feoftina acima da classe 3 (equivalente a 67,03 $mg.g^{-1}$) o modelo não detectou probabilidade de ocorrência do táxon, enquanto que abaixo desse limite a probabilidade foi alta. Já para carbono orgânico total, a probabilidade decresceu acima de valores próximos a 23,4 $mg.g^{-1}$(classe 4).

Entretanto, para areia, salinidade e nitrogênio total houve uma relação direta entre o aumento das escalas e a probabilidade de ocorrência dos táxons. Para areia, quanto maior a porcentagem de areia no sedimento (acima de 49,8%) maior foi a probabilidade de presença de Capitellidae, Ostracoda e *H. australis*. O mesmo ocorreu com o aumento da salinidade para *Prionospio* sp. e Ostracoda (salinidade acima de 6 e 9,5, respectivamente). Para nitrogênio total, essa relação foi detectada para Capitellidae (acima de 1,15 $mg.g^{-1}$). Valores medianos de pH (entre 6,7 e 7,6) para Oligochaeta sp1 e de epicoprostanol (entre 0,3 e 0,6 $\mu g.g^{-1}$) para *A. brasiliana* foram responsáveis pelas maiores probabilidades de ocorrência destes táxons.

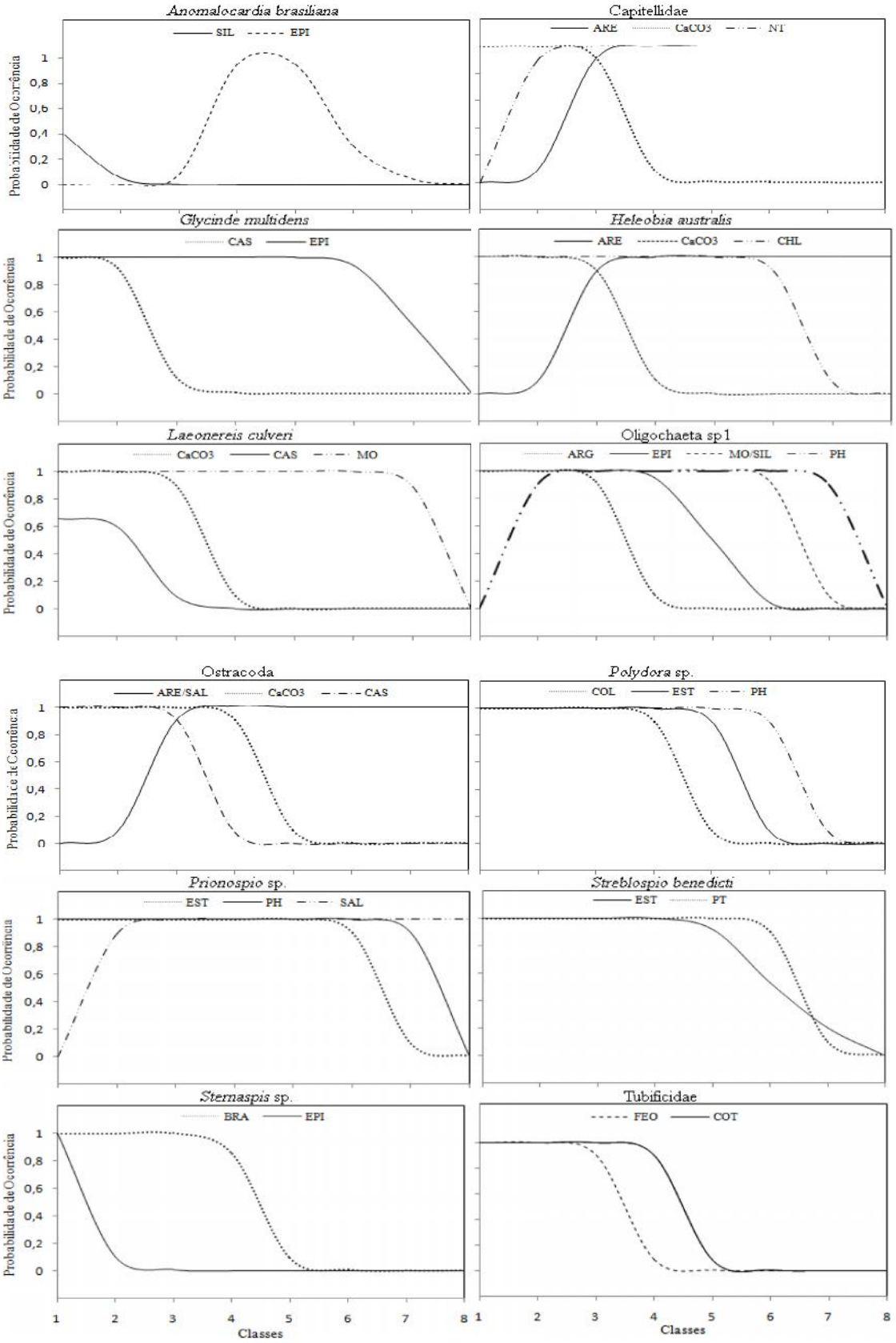


Fig.7 - Análise de Sensibilidade baseada no algoritmo "*profile"* para a probabilidade de ocorrência dos doze táxons que apresentaram valores de predição satisfatórios. Os gráficos foram construídos com as variáveis ambientais mais explicativas. Cada gráfico representa um táxon distinto. As variáveis estão representadas por cada linha e são utilizadas suas respectivas abreviações (ver Tabela 1).

4. DISCUSSÃO

A rede MLP demonstrou ser útil na extração de informações sobre as relações complexas e não-lineares do conjunto de dados, seja quando todos os táxons simultaneamente foram submetidos à rede ou quando os táxons passaram pelo treinamento individualmente. Para todos os táxons em conjunto, os melhores resultados foram obtidos quando algumas variáveis foram excluídas do treinamento das redes. No geral, o melhor modelo foi obtido com apenas oito variáveis ambientais, obtendo valores mais satisfatórios para todas as medidas de desempenho analisadas (*CCI*, *K*, especificidade e sensibilidade). Para os táxons individualmente, esse mesmo comportamento foi identificado em todos os táxons para todas as medidas de desempenho, com exceção de *T. lineata*, que respondeu negativamente à seleção de variáveis.

Com isso, a seleção de variáveis ambientais de entrada mostrou-se uma técnica importante para o pré-processamento dos dados, pois auxilia na eliminação de ruídos da rede e assim, as variáveis irrelevantes para a determinação da presença/ausência dos organismos não precisam ser submetidas ao treinamento. Para *Sigambra* sp. não foi encontrado um subconjunto de variáveis ambientais e entre as possíveis causas deste comportamento do modelo está a atuação de outras variáveis não mensuradas que podem estar agindo na presença ou ausência de *Sigambra* sp.. Segundo Goethals (2005), a seleção das variáveis torna os modelos mais robustos e elimina redundância entre variáveis. A inclusão de variáveis redundantes complica a avaliação da importância relativa de cada variável e a precisão de predição dos modelos melhora quando um conjunto de características ótimas é selecionado (D'heygere et al., 2006; Tirelli et al., 2009). Entre os métodos de seleção de recursos, os algoritmos genéticos (GA) estão entre os procedimentos mais comumente utilizados (Obach et al., 2001; Schleiter et al., 2001;. D'heygere et al., 2006;. Tirelli e Pessani, 2009;. Tirelli et al., 2009).

Os táxons que compõem a fauna bêntica, tendem a responder de maneira variada às mudanças no ambiente, muitas vezes ocorrendo a supressão de táxons mais suscetíveis às mudanças em favorecimento de táxons mais tolerantes (Pearson e Rosenberg, 1978). Porém, como diferentes fontes são responsáveis pelas mudanças no ambiente, como o enriquecimento orgânico, que pode ocorrer devido ao aporte de efluentes urbanos ou pode ser resultante de processos naturais no ambiente, como aporte de manguezais, a tarefa de determinar os fatores que afetam a distribuição da fauna

torna-se complexa (Dauvin, 2007; Elliott e Quintino, 2007). Estudo desenvolvido por Dauvin et al. (2009) exemplifica esta dificuldade em ambientes estuarinos citando que nestes ambientes o acúmulo de matéria orgânica pode ocorrer naturalmente e mudanças drásticas na comunidade bêntica pode ou não estar associadas ao enriquecimento por matéria orgânica de origem antrópica ou natural.

Diversos estudos têm utilizado a análise de sensibilidade (Lek et al., 1996; Guegan et al., 1998; Dedecker et al., 2002) para determinar as contribuições individuais de cada variável de entrada nos organismos previstos. Nesta técnica, uma determinada variável de entrada é fornecida para a rede, variando em toda a sua gama, enquanto as demais são mantidas constantes. A análise das curvas de sensibilidade pode, assim, aumentar a informação sobre os efeitos de vários tipos de impacto sobre um determinado táxon (Marshall et al., 2002). Através da análise de sensibilidade dos modelos da MLP, foram avaliadas as contribuições relativas das variáveis ambientais selecionadas na distribuição dos táxons.

Muitas das variáveis importantes na construção das curvas de sensibilidade estão associadas ao enriquecimento orgânico no ambiente, como é o caso dos nutrientes (fósforo e nitrogênio total), matéria orgânica e esteróis e consequentemente táxons que tiveram a probabilidade de ocorrência associados a estas variáveis podem refletir a tolerância destes ao enriquecimento. Pode ser citado como exemplo o comportamento de *S. benedicti*. Segundo Pearson e Rosenberg (1978), *S. benedicti* é associado principalmente a ambientes poluídos. Os GAs selecionaram principalmente variáveis que demonstram ambientes potencialmente poluídos como coprostanol, fósforo total e clorofila-*a* (Tabela 3), além de estigmasterol que está associado à matéria orgânica de origem terrestre (Volkman, 2006). Nas curvas de sensibilidade a probabilidade de ocorrência de *S. benedicti* esteve associada diretamente às concentrações de fósforo total e estigmasterol.

Quantificar as associações entre a probabilidade de ocorrência de espécies estuarinas e as variáveis abióticas nos permite gerar previsões de distribuição, que podem ser robustas mesmo que os mecanismos ou processos não são conhecidos. De fato, o tipo de análise de sensibilidade realizada neste estudo não permite tirar conclusões diretas sobre os processos que determinam a distribuição dos táxons. No entanto, a modelagem e identificação de possíveis variáveis ambientais condicionantes da ocorrência da fauna são passos críticos em pesquisas ecológicas e gestão de recursos

(Thrush et al., 1999), onde os padrões de distribuição são forte e diretamente ligado a processos físico-químicos.

A abordagem de modelagem aplicada neste estudo, utilizando as redes MLP e outras técnicas auxiliares, foi capaz de predizer as distribuições dos táxons da macrofauna bêntica com um grau relativamente elevado de sucesso. Os resultados obtidos pelos modelos desenvolvidos neste estudo revelaram que estes possuem considerável poder preditivo, principalmente quando a seleção de variáveis ambientais foi utilizada. A inclusão de informações do processo (especialmente sobre os hábitos alimentares das espécies) e informação de história natural, certamente pode melhorar a qualidade e generalidade dos modelos, tornando as previsões mais robustas, permitindo o desenvolvimento de modelos que descrevem a ação e interação de processos operacionais em diferentes escalas (Thrush et al., 1999).

Apesar de, em um primeiro momento, obtermos resultados satisfatórios, mais pesquisas são necessárias para inspecionar e validar as previsões destes modelos, principalmente pela falta de estudos relacionados a ambientes tão dinâmicos como a planícies entremarés, sujeitas às mudanças contínuas no ambiente (Warwick et al., 1990). Além disto, a modelagem de outros táxons mais raros, com habitats mais seletivos, pode auxiliar na obtenção de mais informações sobre a qualidade ecológica destes ambientes. O desenvolvimento de modelos preditivos que fornecem informações sobre as comunidades possuem potencial para contribuir significativamente para a restauração e gestão de sistemas aquáticos. A seleção de características ambientais intrínsecas a cada táxon é essencial para a modelagem a fim de aumentar significativamente os desempenhos dos modelos.

5. REFERÊNCIAS

Austin, M. P., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modeling. Ecological Modelling 157, 101-118.

Baeta, F., Pinheiro, A., Corte-Real, M., Costa, J. L., Raposo de Almeida, P., Cabral, H., Costa, M. J., 2005: Are the fisheries in the Tagus estuary sustainable? Fisheries Research 76, 243-251.

Benbow M.E., Burky A.J., Way C.M. 2003. Life cycle of a torrenticolous Hawaiian chironomid (Telmatogeton torrenticola): stream flow and microhabitat effects. Annales de Limnologie - International Journal of Limnology 39, 103-114.

Boyero L., 2003. The effect of substrate texture on colonization by stream macroinvertebrates. Annales de Limnologie - International Journal of Limnology 39, 211-219.

601 Brown J.H., Mehlman D.W., Stevens G.C., 1995. Spatial variation in abundance. Ecology 76,
602 2028 -2043.
603 Céréghino R., Park Y.S., Compin A., Lek S., 2003. Predicting the species richness of aquatic
604 insects in streams using a limited number of environmental variables. Journal of the North
605 American Benthological Society 22, 442-456.
606 Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological
607 Measurement 20, 37-46.
608 D'heygere, T., Goethals, P.L.M., De Pauw, N., 2003. Genetic algorithms for optimisation of
609 predicitive ecosystem models based on decision trees and neural networks. Ecological
610 Modelling 160, 291-300.
611 D'heygere, T., Goethals, P.L.M., De Pauw, N., 2006. Genetic algorithms for optimisation of
612 predictive ecosystems models based on decision trees and neural networks. Ecological
613 Modelling 195, 20-29.
614 Dauvin, J.C., 2007. Paradox of estuarine quality: benthic indicators and indices, consensus or
615 debate for the future. Marine Pollution Bulletin 55, 271-281.
616 Dauvin, J.C., Bachelet, G., Barillé, A.I., Blanchet, H., De Montaudoin, X., Lavesque, N.,
617 Ruellet, T., 2009. Development of benthic indicators and index approaches in three main
618 estuaries along the French Atlantic coast (Seine Loire and Gironde) for the implementation
619 of the European Water framework Directive (WFD). Marine Ecology 30, 228-240.
620 Dedecker, A.P., Goethals, P.L.M., Gabriels, W., De Pauw, N., 2002. Comparison of artificial
621 neural network (ANN) model development methods for prediction of macroinvertebrates
622 communities in the Zwalm river basin in Flanders, Belgium. Scientific World Journal 2,
623 96-104.
624 Dedecker A.P., Goethals P.L.M., Gabriels W., De Pauw N., 2004. Optimization of Artificial
625 Neural Network (ANN) model design for prediction of macroinvertebrates in the Zwalm
626 river basin (Flanders, Belgium). Ecological Modelling174, 161-173.
627 Dedecker A.P, Goethals P.L.M. and De Pauw N., 2005. Sensitivity and robustness of stream
628 model based on artificial neural networks for the simulation of different management
629 scenarios. In: Lek S., Scardi M., Verdonschot P.F.M., Descy J.P. and Park Y.S. (eds.),
630 Modelling Community Structure in Freshwater Ecosystems, Springer-Verlag, Berlin, 133–
631 146.
632 Elliott, M., Quintino, V., 2007. The estuarine quality paradox environmental homeostasis and
633 the difficulty of detecting anthropogenic stress in naturally stressed areas. Marine Pollution
634 Bulletin 54, 640-645.
635 Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in
636 conservation presence/ absence models. Environmental Conservation, 24, 38-49.
637 Gabriels, W., Goethals, P.L.M., Dedecker, A.P., Lek, S., De Pauw, N., 2007. Analysis of
638 macrobenthic communities in Flanders, Belgium, using a stepwise input variable selection
639 procedure with artificial neural networds. Aquatic Ecology 41, 427-441.
640 Gevrey M., Dimopoulos I., Lek S., 2003. Review and comparison of methods to study the
641 contribution of variables in artificial neural network models. Ecological Modelling 160,
642 249-264
643 Goethals P.L.M., 2005. Data driven development of predictive ecological models for benthic
644 macroinvertebrates in rivers. PhD thesis, Ghent University.
645 Goethals P., De Pauw N, 2001. Development of a concept for integrated river assessment in
646 Flanders, Belgium. Journal of Limnology 60, 7-16.
647 Goethals, P.L.M., Dedecker, A.P., Gabriels, W., Lek, S., De Pauw, N., 2007. Applications of
648 artificial neural networks predicting macroinvertebrates in freshwaters. Aquatic Ecology
649 41, 491-508.
650 Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization and Machine Learning.
651 Addison-Wesley Publishing Company, Reading, MA, 412.
652 Grasshoff, K., Ehrhardt, M., Kremling, K., 1983. Methods of Seawater Analysis 2 ed, Verlag
653 Chemie: Weinheim.

654 Guégan JF, Lek S, Oberdorff T., 1998. Energy availability and habitat heterogeneity predict
655 global riverine fish diversity. Nature 391:382-384.
656 Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology.
657 Ecological Modelling 135, 147-168.
658 Hagan MT, Demuth HB, Beale M., 1996. Neural network design. PWS Publishing Company,
659 Boston.
660 Harrell, F.E., Lee, K.L., Mark, D.B., 1996. Multivariate prognostic models: Issues in developing
661 models, evaluating assumptions and adequacy, and measuring and reducing errors.
662 Statistics in Medicine 15, 361-387.
663 Hoang H, Recknagel F, Marshall J, Choy S. 2001. Predictive modelling of macroinvertebrate
664 assemblages for stream habitat assessments in Queensland (Australia). Ecological
665 Modelling 146:195-206.
666 Hoang, T.U., Lock, K., Mouton, A., Goethals, P.L.M., 2010. Application of classification trees
667 and support vector machines to model the presence of macroinvertebrates inrivers in
668 Vietnam. Ecological Informatics 5, 140-146.
669 Kawakami, S.K., Montone, R.C., 2002. An efficient ethanol-based analytical protocol to
670 quantify fecal steroids in marine sediments, Journal of the Brazilian Chemical Society 13,
671 226-232.
672 Kim, B., Lee, S.E., Song, M.Y., Choi, J.H., Ahn, S.M., Lee, K.S., Cho, E., Chon, T.S., Koh,
673 S.C., 2008. Implementation of artificial neural networks (ANNs) to analysis of inter-taxa
674 communities of benthic microorganisms and macroinvertebrates in a polluted stream.
675 Science of The Total Environment 390, 262-274.
676 Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. Artificial Intelligence
677 Journal 97(1-2), 273-324.
678 Kolm, H.E., Schoenenberger, M.F., Piemonte, M.R., Souza, P.S.A., Scühli, G.S., Mucciatto,
679 M.B., Mazzuco, R., 2002. Spatial variation of bacteria in surface waters of Paranaguá and
680 Antonina Bays, Paraná, Brazil. Brazilian Archives of Biology and Technology 45, 27-34.
681 Kung S., 1993. Digital neural networks. Englewood Cliffs, NJ: Prentice Hall.
682 Lana, P.C., Marone, E., Lopes, R.M., Machado, E. C., 2000. The subtropical estuarine complex
683 of Paranaguá Bay, Brazil, In Coastal Marine Ecosystems of Latin America, Seeliger, U.,
684 Lacerda, L.D., Kjerfve, B. (eds,), Springer Verlag, NY, USA, 467.
685 Lek, S., Guégan, J.F., 1999. Artificial neural networks as a tool in ecological modelling, an
686 introduction. Ecological Modelling 120, 65-73.
687 Lek S., Belaud A., Dimopoulos I., Lauga J., Moreau J., 1995. Improved estimation, using neural
688 networks, of the food consumption of fish populations. Marine & Freshwater Research
689 46:1229-1236.
690 Lek, S., Belaud, A., Baran, P., Dimopoulos, I., Delacoste, M., 1996. Role of some
691 environmental variables in trout abundance models using neural networks. Aquatic Living
692 Resources 9, 23-29.
693 Lippmann R. 1967. An introduction to computing with neural nets. Volume April: IEEE ASSP
694 Magazine; 4-22.
695 Lorenzen, C.J., 1967. Determination of chlorophyll and phaeopigments: Spectrophotometric
696 equations. Limnology and Oceanography 12, 343-346.
697 Maier H.R., Dandy G.C., 2000. Neural networks for the prediction and forecasting of water
698 resources variables: a review of modelling issues and applications. Environmental
699 Modelling & Software 15, 101-124
700 Maldonado, C., Venkatesan, M.I., Phillips, C.R., Bayona, J.M., 2000. Distribution of
701 trialkylamines and coprostanol in San Pedro shelf sediments adjacent to a sewage outfall.
702 Marine Pollution Bulletin 40, 680-687.
703 Manel, S., Dias, J.M., Buckton, S.T., Ormerod, S.J., 1999. Alternative methods for predicting
704 species distribution: an illustration with Himalayan river birds. Journal of Applied Ecology
705 36, 734-747.

706 Marshall J., Hoang H., Choy S., Recknagel F., 2002. Relationships between habitat properties
707 and the occurrence of macroinvertebrates in Queensland streams (Australia) discovered by
708 a sensitivity analysis with artificial neural networks. Verhandlungen des Internationalen
709 Verein Limnologie 28, 1415-1419
710 Martins, C.C., Braun, J.A.F., Seyffert, B.H., Machado, E.C., Fillmann, G., 2010. Anthropogenic
711 organic matter inputs indicated by sedimentary fecal steroids in a large South American
712 tropical estuary (Paranaguá estuarine system, Brazil). Marine Pollution Bulletin 60, 2137-
713 2143.
714 Obach, M., Wagner, R., Werner, H., Schmidt, H.H., 2001. Modelling population dynamics of
715 aquatic insects with arti- ficial neural networks. Ecological Modelling 146: 207-217.
716 Park Y.S., Rabinovich J., Lek S., 2007. Sensitivity analysis and stability patterns of two-species
717 pest models using artificial neural networks. Ecological Modelling 204, 427-438.
718 Pearson, T.H., Rosenberg, R., 1978. Macrobenthic succession in relation to organic enrichment
719 and pollution of the marine environment. Oceanography and Marine Biology Annual Review
720 16, 229-311.
721 Rumelhart, D., Hinton G. E., Williams R. J., 1986. Learning internal representations by error
722 propagation, in parallel Distributed Processing, vol. 1, chap. 8, edited by D. E. Rumelhart
723 and J. L. McCelland, MIT Press, Cambridge, Mass.
724 Schleiter I.M., Obach M., Borchardt D., Werner H., 2001. Bioindication of chemical and
725 hydromorphological habitat characteristics with benthic macro-invertebrates based on
726 artificial neural networks. Aquatic Ecology 35, 147-158
727 Strickland, J.L.H., Parsons T.R., 1972. A practical handbook of seawater analysis. Bulletin of
728 the Fisheries Research Board of Canada 167, 311.
729 Suguio, K., 1973. Introdução à sedimentologia. São Paulo, Edgard Blucher LTDA :317p.
730 The Mathworks Inc., 2011. MATLAB Version 7.12. The Mathworks, Inc., Massachusetts.
731 Thrush SF, Lawrie SM, Hewitt JE, Cummings VJ (1999) The problem of scale: uncertainties
732 and implications for softbottom marine communities and the assessment of human impacts.
733 In: Gray JS, Ambrose WG, Szaniawska A (eds) Biogeochemical cycling and sediment
734 ecology. Kluwer Academic Publishers, Dordrecht, 195-210.
735 Tirelli, T., Pessani, D., 2009. Use of decision tree and artificial neural network approaches to
736 model presence/absence of Telestes muticellus in Piedmont (North-Western Italy). River
737 Research Applications 25 (8), 1001-1012.
738 Tirelli T., Pozzi L., Pessani D., 2009. Use of different approaches to model presence/absence of
739 Salmo marmoratus in Piedmont (Northwestern Italy), Ecological Informatics 4, 234-242.
740 Volkman, J. K., 1986. A review of sterol markers for marine and terrigenous organic matter.
741 Organic Geochemistry 9, 83-100.
742 Volkman, J. K., 2006. Lipids markers for marine organic matter. The Handbook of
743 Environmental Chemistry 2. Part N: 27-70.
744 Walley, W.J., Fontama, V.N., 1998. Neural network predictors of average score per taxon and
745 number of families at unpolluted river sites in Great Britain. Water Research 32, 3-11.
746 Warwick, R. M., Platt, H. M., Clarke, K. R., Agard, J., Gobin, J., 1990. Analysis of
747 macrobenthic and meiobenthic community structure in relation to pollution and disturbance
748 in Hamilton Harbour, Bermuda, Journal of Experimental Marine Biology and Ecology 138,
749 119-142.
750 Willems, W., Goethals, P., Van den Eynde, D., Van Hoey, G., Van Lancker, V., Verfaillie, E.,
751 Vincx, M., Degraer, S., 2008. Where is the worm? Predictive modelling of the habitat
752 preferences of the tube-building polychaete Lanice conchilega, Ecological Modelling 212,
753 74-79.
754 Wong, F., Tan, C., 1994. Hybrid neural, genetic, and fuzzy systems, In G. J. Deboeck (Ed),
755 Trading on the Edge, John Wiley, New York, 243-261.

CAPITULO III

Aplicação de árvores de classificação e máquinas de vetor de suporte na modelagem da presença da macrofauna bêntica em planícies de maré estuarinas

Application of classification trees and support vector machines to model the presence of benthic macrofauna in estuarine tidal flats

Revista pretendida: Ecological Informatics (Ecol. Info.), ISSN (1574-9541), Fator de Impacto (JCR, 2010) = 1.351, Qualis CAPES = Estrato B1.

Faller[1], D. G.; Camargo[1], M. G.

[1]Centro de Estudos do Mar, Universidade Federal do Paraná. Av. Beira Mar s/n, Pontal do Paraná, Paraná, Brasil. CEP: 83255-971. Fone: (41) 3511-8600. E-mail: daiafaller@yahoo.com.br

RESUMO

Este estudo tem como objetivo avaliar o desempenho de máquinas de vetor de suporte (SVM) e árvores de classificação (CT) na predição da presença de táxons da macrofrauna bêntica. O desempenho dos modelos foi avaliado através da comparação dos resultados obtidos para os modelos de SVM e CT e o desempenho de cada técnica foi comparado antes e após a seleção de variáveis ambientais utilizando algoritmos genéticos (GA). A seleção foi realizada através da abordagem *wrapper*. Nesta abordagem os GAs conduzem a busca por subconjuntos de variáveis utilizando as CTs ou SVMs como parte da função de avaliação destes subconjuntos. Para avaliar o desempenho dos modelos foram utilizadas três medidas: as instâncias corretamente classificadas (*CCI*), estatística *Kappa* (*K*) de Cohen e o raiz do erro médio quadrado (*RMSE*). Os resultados tanto para CTs como SVMs foram, no geral, melhores após a seleção das variáveis ambientais. A eficiência dos modelos, baseadas nas medidas de desempenho, apresentaram valores médios superiores para as CTs, antes e após a seleção das variáveis, com exceção dos valores médios de *CCI*, que foram superiores para as SVMs antes de realizar a seleção. Entre os 20 táxons modelados, 16 obtiveram sucesso para as CTs, identificando as relações entre a distribuição dos táxons e as variáveis ambientais, enquanto que para as SVMs, apenas 10 táxons foram modelados com sucesso. Com base nos modelos desenvolvidos, pode-se concluir que, em geral, um desempenho melhor foi detectado para as CTs, com e sem a aplicação dos GAs. A avaliação dos resultados obtidos pelos modelos utilizando mais de uma medida de desempenho, com destaque para o uso em conjunto de *CCI* e *K*, se mostrou a forma mais adequada de avaliar se as previsões foram ou não confiáveis.

Palavras-chave: máquinas de vetor de suporte, árvores de classificação, macrofauna bêntica, modelagem preditiva, algoritmos genéticos, seleção de variáveis

ABSTRACT

The aim of the present study was to analyze the performance of support vector machines (SVM) and classification trees (CT) in predicting the presence of benthic macrofrauna. The models performance was evaluated by comparing the results obtained for SVM and CT models. The performance of each technique was compared before and after the selection of environmental variables using genetic algorithms (GA). The variable selection was done through the *wrapper* approach, which leads to the search for the subset of variables using CT and SVM as part of the evaluation function of these subsets. To evaluate the performance of the models we used three measures: correctly classified instances (CCI), Cohen`s kappa (K) and root mean squared error (RMSE). The results both for CT and SVM techniques were, in general, better after the selection of environmental variables. The efficiency of the models, based on performance measures, showed higher mean values for the CT before and after the selection of variables, except for the mean values of *CCI*, which were superior to SVM before selection occurs. Among the 20 taxa modeled, 16 were successful for the CTs, identifying the relationships between the taxa distribution and environmental variables, whereas for SVMs, only 10 taxa were successfully modeled. Based on the model developed, it can be concluded that, in general, better performance was detected for the CTs models, with and without the Gas application. The evaluation of the results obtained by the models, using more than one performance measure, especially for use in conjunction *CCI* and *K,* was the most appropriate way to assess the relevance of the developed models.

Keywords: support vector machines, classification trees, genetic algorithms, benthic macrofauna, predictive modeling, variables selection

1. INTRODUÇÃO

A composição da fauna bêntica é afetada pela interação de múltiplas variáveis ambientais e biológicas que atuam direta ou indiretamente em diferentes escalas espaciais e temporais (Legendre, 1993; Poff, 1997; Allan, 2004; Culp et al., 2010). A composição destas comunidades pode ser utilizada no biomonitoramento da condição ambiental de um ecossistema, devido à resposta aos distúrbios no ambiente (Statzner et al. 2005; Pollard e Yuan, 2010).

Inúmeros estudos ecológicos aplicaram modelos preditivos, baseados em técnicas de aprendizado de máquina para avaliar, monitorar e gerenciar os recursos naturais, considerando a presença ou ausência de organismos indicadores em um determinado habitat (Goethals et al., 2002; D'heygere et al., 2003, 2006; Dedecker et al., 2005; Dakou et al., 2007; Hoang et al., 2010). Entre as técnicas, as árvores de classificação (CT, do inglês *classification trees*) e máquinas de vetores de suporte (SVM, do inglês *support vector machine*) são dois exemplos de técnicas de modelagem que podem ser utilizadas neste contexto.

Devido à sua transparência e flexibilidade, as CTs vêm ganhando popularidade nos últimos anos, sendo atraentes principalmente pela simplicidade de construção e transparência na classificação de dados em diferentes áreas. Entre os estudos que utilizaram CTs em ecologia, se destacam aqueles de predição de indicadores biológicos, principalmente a fauna bêntica de rios (Edwards et al., 2006; D'heygere et al., 2003, 2006; Dakou et al., 2007). As previsões feitas com as CTs fornecem uma boa indicação do impacto de diferentes distúrbios nas condições ambientais e facilmente podem ser integradas em um sistema ambiental de apoio à decisão (Ghetti e Ravera, 1993; D'heygere et al., 2006; Dakou et al., 2007; Goethals et al., 2007; Hoang et al., 2010).

Já as SVMs são uma abordagem de reconhecimento de padrões (Vapnik, 1998), constituindo uma técnica versátil de aprendizado supervisionado de máquina que podem ser treinadas para aprender relações complexas entre os dados. Nos últimos anos, tem havido um grande interesse em SVMs (Vapnik, 1995; Burges, 1998; Keerthi et al., 2001). Em ecologia, assim como as CTs, as SVMs têm sido aplicadas principalmente em ecologia de rios (Shan et al., 2006; Sanchez-Hernandez et al., 2007a; Ribeiro e Torgo, 2008; Hoang et al., 2010).

Além da escolha da técnica de modelagem, uma das tarefas mais difíceis é a definição do conjunto de variáveis que agem na presença de determinado organismo. No entanto, devido à frequente sobreposição de padrões espaciais, gerados por diferentes processos que atuam em uma ampla gama de escalas, determinar a importância dos processos envolvidos ainda é um desafio para os ecologistas (Legendre 1993; Legendre et al., 2002).

Para melhorar a precisão e o poder preditivo destes modelos, o número de variáveis explicativas utilizado deve ser reduzido a um número razoável (Harrell et al., 1996). Os algoritmos genéticos (GA) estão entre os mais populares utilizados na seleção das variáveis de entrada de uma rede. GAs são algoritmos computacionais de busca e

otimização global, inspirados no princípio de "seleção natural" de Darwin-Wallace. A aplicação de GAs na resolução de problemas complexos auxilia na busca de uma melhor solução para um determinado problema (Holland, 1975; Goldberg, 1989; D'heygere et al., 2006).

Este estudo tem como objetivo avaliar o desempenho de máquinas de vetor de suporte (SVM) e árvores de classificação (CT) na predição da presença de táxons da macrofrauna bêntica. O desempenho dos modelos desenvolvidos foi avaliado de duas maneiras distintas: (1) comparação entre os modelos de SVM e CT e (2) desempenho de cada modelo antes e após a seleção de variáveis com os algoritmos genéticos (GA).

2. MATERIAL E MÉTODOS

2.1 Área de Estudo

O estudo foi realizado em planícies entremarés não-vegetadas do Canal da Cotinga, um subestuário da Baia de Paranaguá, (Complexo Estuarino de Paranaguá - 25°30'S, 48°25'W), (Fig.1). Os rios que fazem parte da margem sul da baia, como o Maciel, Guaraguaçú e Itiberê e que deságuam na Cotinga, recebem grande parte dos efluentes produzidos na cidade e porto de Paranaguá (Lana et al*.,* 2001). Entre os rios da região, o Itiberê pode ser considerado uma das principais fontes pontuais de contaminação na Cotinga, evidenciado por estudos realizados no local que encontraram elevadas concentrações de indicadores orgânicos de poluição, como coliformes fecais na coluna d'água (Kolm et al., 2002) e esteróis fecais no sedimento (Martins et al., 2010). As concentrações de contaminantes no canal são dispersas e diluídas a partir da região interna e mediana em direção à sua desembocadura, formando um gradiente de poluição. Planícies não-vegetadas das duas margens do canal foram selecionadas, totalizando 107 locais de amostragem ao longo da Cotinga.

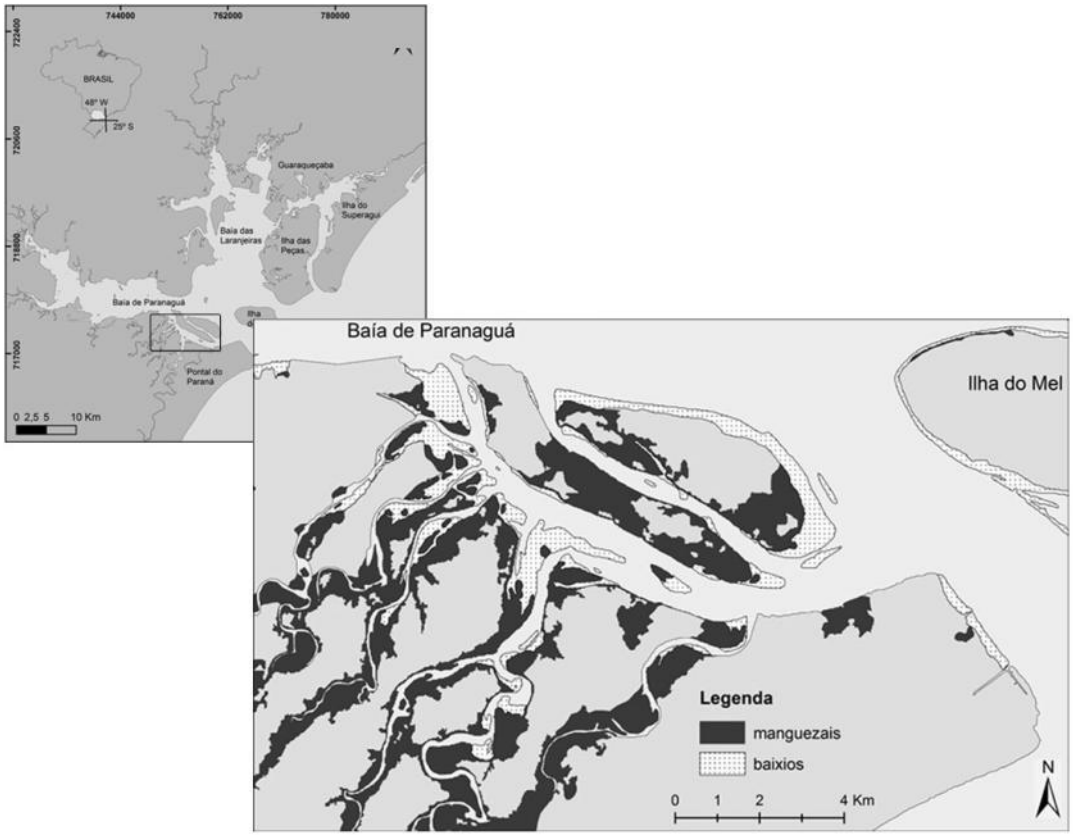


Fig.1. Complexo Estuarino de Paranaguá (CEP) com o Canal da Cotinga no detalhe.

2.5 Conjunto de dados

Em cada local de amostragem foram coletadas duas réplicas com *corers* com diâmetro de 10 cm para a análise da macrofauna bêntica. As amostras biológicas foram fixadas com formol-aldeído (4%), lavadas em peneiras de 0,5 mm, coradas com Rosa de Bengala, preservadas em álcool (70%) e identificadas até o nível de família, totalizando 49 famílias. Entre estas, foram selecionadas as famílias com mais de 15% de ocorrência, que foram identificadas até o menor nível possível. Os 20 táxons com maior ocorrência foram utilizados no desenvolvimento dos modelos do SOM. São eles: *Anomalocardia brasiliana* (35,5%), *Bulla* sp. (64,5%), Capitellidae (72,9%), *Capitella* sp. (23,4%), *Caprella scaura* (14%), *Glycinde multidens* (87,9%), *Heleobia australis* (63,6%), *Laeonereis culveri* (31,8%), *Neanthes succinea* (13,1%), Oligochaeta sp1 (34,6%), Orbiniidae (51,4%), Ostracoda (39,3%), *Polydora* sp. (42,1%), *Prionospio* sp. (44%), *Sigambra* sp. (82,2%), *Sternaspis* sp. (36,4%), *Streblospio benedicti* (61,7%), *Tagelus* sp. (53,3%), *Tellina lineata* (59,8%) e Tubificidae (85%).

Juntamente com a fauna, 19 variáveis ambientais foram utilizadas no desenvolvimento dos modelos (Tabela 1). Em campo, foi realizada a medição da

147 camada redox e amostras de água intersticial foram coletadas para a determinação do
148 pH e salinidade (pHmetro e refratômetro manual, respectivamente). Amostras de
149 sedimento foram coletadas para determinar: fósforo total e nitrogênio total (Grasshoff et
150 al., 1983); carbono orgânico total (Strickland e Parsons, 1972); clorofila-*a* e feoftina
151 (Lorenzen, 1967), porcentagens de cascalho, areia, silte e argila, (Suguio, 1973);
152 carbonato de cálcio, matéria orgânica e esteróis totais (Kawakami e Montone, 2002).

153 Foram considerados cinco esteróis para identificar diferentes fontes de matéria
154 orgânica: coprostanol e estigmasterol, como indicadores de contaminação por esgotos
155 (Maldonado et al., 2000); colesterol para origem aquática, por estar presente no fito e
156 zooplâncton (Volkman, 1986); brassicasterol para identificar matéria orgânica de
157 origem marinha e estigmasterol para identificar matéria orgânica de origem terrestre
158 (Volkman, 2006). Devido à inviabilidade em estimar a concentração de esteróis em
159 todos os pontos, optou-se por realizar as amostragens em 31 pontos selecionados de
160 acordo com a proximidade à cidade de Paranaguá e à desembocadura dos rios. A
161 extrapolação para os demais pontos foi realizada através da média entre pontos
162 vizinhos.

163

164 **Tabela 1**

165 Variáveis de entrada usadas para o desenvolvimento dos modelos em conjunto com a média, desvio
166 padrão (DP), valores mínimos (Min.) e máximos (Max.).

Variável	Abreviação	Unidade	Média	DP	Min.	Max.
Salinidade	SAL	-	24,63	5,35	2	32
pH	PH	-	7,28	0,27	6,58	7,92
Camada Redox	RED	cm	1,15	1,15	0,1	5,97
Coprostanol	COP	$\mu g.g^{-1}$	0,38	0,46	0	2,04
Epicoprostanol	EPI	$\mu g.g^{-1}$	0	0,01	0	0,08
Colesterol	COL	$\mu g.g^{-1}$	4,37	2,89	0,74	15,5
Brassicasterol	BRA	$\mu g.g^{-1}$	2,2	1,5	0,23	8,22
Estigmasterol	EST	$\mu g.g^{-1}$	2,7	1,62	0,27	10,40
Cascalho	CAS	%	1,72	5,95	0	52,28
Areia	ARE	%	84,98	11,4	33,76	97,89
Silte	SIL	%	10,21	8,9	0	44,56
Argila	ARG	%	3,09	3,01	0	21,31
Carbono	COT	$mg.g^{-1}$	14,55	11,7	0	46,7
Clorofila	CHL	$mg.g^{-1}$	17,39	23,2	0	157,95
Feoftina	FEO	$mg.g^{-1}$	17,88	23,5	0	178,75
Nitrogênio Total	NT	$mg.g^{-1}$	2,09	1,11	0,07	4,41
Fósforo Total	PT	$mg.g^{-1}$	0,03	0,02	0	0,09
Matéria Orgânica	MO	%	4,36	2,17	0,49	13,05
Carbonato de Cálcio	CaCO3	%	4,11	3,98	0,48	32,23

2.2 Árvore de Classificação (CT)

As CTs são capazes de prever o valor de uma variável discreta dependente a partir dos valores de um conjunto de atributos independentes contínuos ou discretos (Quinlan, 1986). O procedimento principal para o desenvolvimento de um modelo de CT é a divisão dos dados da variável dependente alvo (20 táxons da fauna bêntica) baseado na sua resposta às variáveis independentes de entrada (19 variáveis ambientais, todas contínuas).

Para a construção das redes foi utilizada a regra de indução *Top-Down* (Quinlan, 1986), na qual a construção da árvore começa com todo o conjunto de dados de treinamento e, em cada etapa, a variável de entrada mais informativa é selecionada como raiz da árvore. Com isso, o conjunto de treinamento atual é divido em subconjuntos de acordo com os valores dos atributos selecionados. Os limites são selecionados e dois ramos são criados, com base no limiar da variável para presença/ausência do táxon (Quinlan, 1986). A construção da árvore encerra quando o critério determinado pelo modelador é satisfeito ou quando todos os exemplos estão na mesma classe (nó). Estes subconjuntos finais formados pelo processo recursivo são chamados de "folhas" da CT e são rotulados como uma classe (Quinlan, 1993).

O algoritmo J48, utilizado na indução das CTs, é uma re-implementação Java para o Software WEKA do algoritmo C4.5 (Hall et al., 2009; Quinlan, 1993; Witten e Frank, 2000). A otimização das árvores foi realizada através da poda (*tree-pruning*) e da mudança do Fator de Confiança (FC) para alterar a intensidade da poda aplicada nas árvores. Esta técnica é utilizada para aumentar a transparência das CTs, reduzindo o seu tamanho e aumentado a precisão de classificação, eliminando os ruídos nos dados (Bratko, 1989). Modelos com diferentes intensidades de poda foram induzidos pela mudança do FC em 0,01, 0,1, 0,25 e 0,5. Os demais parâmetros do algoritmo J48 foram utilizados em sua configuração padrão.

A complexidade da predição foi identificada através da quantidade de folhas da CT, sendo que o acréscimo de folhas nas árvores aumenta a complexidade do modelo e dificulta a tomada de decisão em relação à quais variáveis são mais importantes.

2.3 Máquina de Vetores de Suporte (SVM)

As redes SVM são embasadas pela teoria de aprendizado estatístico desenvolvida por Vapnik (1995). Para a obtenção de classificadores com boa generalização, essa teoria estabelece uma série de princípios que devem ser seguidos. Generalização é definida como a capacidade de um modelo em prever corretamente a classe de novos dados do mesmo domínio em que o aprendizado ocorreu. Dados dois grupos distintos de vetores, o modelo faz o reconhecimento dos padrões e encontra o hiperplano que separa os padrões. Os vetores que definem os limites do hiperplano são chamados de vetores de suporte.

As SVMs foram implementadas com o algoritmo de Platt, onde há uma otimização sequencial mínima (SMO), para a formação de um classificador de vetor de suporte (Keerthi et al., 2001). Esta implementação substitui todos os valores em falta e transforma atributos nominais em binários. A rede consistiu de 19 variáveis de entrada (variáveis ambientais) e uma de saída (presença/ausência do táxon), conectadas através de vetores de peso.

Muitas vezes, na dimensão original dos dados, a tarefa de separação dos padrões não pode ser realizada durante o treinamento dos dados. Para contornar esta dificuldade a SVM mapeia os dados em dimensões mais elevadas, onde a separação não-linear é possibilitada através de uma Função Kernel (Vapinik, 1995; Phan et al., 2005). Para todas as SVMs foi utilizada a Função Kernel Polinomial. Todos os parâmetros da SVM foram mantidos em suas configurações padrões, com exceção dos expoentes da Função Kernel Polinomial. Para obter as melhores performances preditivas, os expoentes variaram de 1 a 5 (Hoang et al., 2010). O modelo que apresentou melhores resultados foi fixado como expoente padrão e foi executado cinco vezes após a randomização para verificar a robustez e reprodutibilidade dos modelos. Para o desenvolvimento dos modelos foi utilizado o Software WEKA (Hall et al., 2009)

2.4 Algoritmos Genéticos (GA)

A escolha das variáveis apropriadas em um conjunto de dados é importante para melhorar o desempenho do modelo (Goethals et al., 2007). A seleção das variáveis de entrada para as CTs e SVMs foi realizada através da combinação das redes com

algoritmos genéticos (GA). Portanto, os GAs foram escolhidos na seleção das variáveis devido a capacidade destes algoritmos de proporcionarem uma pesquisa sistemática no universo amostral que seleciona as variáveis que melhor explicam os táxons modelados. Este procedimento auxilia na diminuição da complexidade dos modelos a serem treinados, pois elimina variáveis irrelevantes ao táxon modelado.

Os GAs são formados por uma população de diferentes soluções concorrentes que evoluem e tendem a convergir em uma solução ideal. Esta solução é representada por um cromossomo, que é formado por vários genes. A população inicial é formada aleatoriamente e após sucessivas iterações (gerações), os cromossomos iniciais dão lugar aos cromossomos mais fortes obtidos através de suas reproduções, formando novas gerações. As gerações podem ser formadas por cruzamento, seleção e mutação.

Existem inúmeras variações de algoritmos genéticos. Porém, neste estudo os GAs de Goldberg (Goldberg, 1989) foram aplicados para encontrar um melhor conjunto de variáveis de entrada relevante para a previsão da presença/ausência dos táxons da fauna bêntica. Os cromossomos foram formados por 19 genes, cada um representando uma variável de entrada, com codificação binária. Isso significa que uma determinada variável foi selecionada (representado por '1') ou não (representado por '0').

A taxa de cruzamento (*crossover*) foi fixada com probabilidade de 60%, enquanto que a mutação ocorreu com probabilidade de 3%. A população inicial foi constituída por 20 cromossomos que foram evoluindo ao longo de 40 gerações. Estes parâmetros foram definidos depois de testes preliminares, onde os valores de cruzamento, mutação e gerações foram modificados até encontrar o valor ideal. Através do uso da abordagem *wrapper,* foi possível utilizar os GAs e os algoritmos de indução (SVM e CT) em conjunto. Nesta abordagem, os GAs conduzem a busca por subconjuntos confiáveis no espaço de treinamento utilizando os algoritmos de indução como parte da função de avaliação destes subconjuntos (Kohavi e John, 1997).

2.5 Treinamento e Validação dos Modelos

Para as CTs, SVMs e GAs, a separação dos conjuntos de treinamento e teste foi realizada através da validação cruzada (Dedecker et al.,2004; Goethals et al., 2007) com 10 subconjuntos. Nesta validação, os dados originais foram aleatoriamente divididos em 10 subconjuntos de tamanho aproximadamente igual. Destes 10, um único conjunto é

mantido para validação do modelo e os subconjuntos restantes são usados como dados de treinamento. Este procedimento foi utilizado para estimar o erro dos modelos e assim comparar as performances com e sem a seleção prévia de variáveis ambientais.

A eficiência de predição dos modelos foi avaliada através de três critérios de medida de desempenho: as instâncias corretamente classificadas (*CCI*), estatística *Kappa* (*K*) de Cohen (Cohen, 1960) e o raiz do erro médio quadrado (*RMSE*). Para isto, foram identificados através da matriz de confusão (Fielding e Bell, 1997) os casos de verdadeiro positivo (*VP*), falso positivo (*FP*), falso negativo (*FN*) e verdadeiro negativo (*VN*) previstos por cada modelo (Tabela 2).

Tabela 2

Matriz de confusão

		Predito	
		Presente	Ausente
Atual	Presente	*VP*	*FN*
	Ausente	*FP*	*VN*

Como primeira medida de desempenho dos modelos foi utilizado o *RMSE*, que é baseado na diferença entre os valores medidos (y_i) no conjunto de teste e os valores previstos ($\hat{y}(x_i)$).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}(x_i))^2} \qquad (1)$$

Os valores de *CCI* são calculados através do percentual de previsões de presença (VP) e de ausência (VN) que o modelo retornou corretamente. Segundo a fórmula:

$$CCI = \frac{VP + VN}{VP + FP + FN + VN} \times 100 \qquad (2)$$

Por último, os modelos foram avaliados quanto aos valores de *K*, que medem a proporção de todos os casos possíveis de presença (VP) ou ausência (VN) previstos corretamente pelo modelo. Gabriels et al. (2007) avalia os valores de *K* da seguinte forma: 0-0,2 pobre; 0,2-0,4 razoável; 0,4-0,6 moderada; 0,6-0,8 bom; 0,8-1,0 excelente.

$$\kappa = \left[(VP + FN) - \left(\frac{(VP + FN)(VP + FP) + (FP + VN)(FN + VN)}{n}\right)\right] \div \left[n - \left(\frac{(VP + FN)(VP + FP) + (FP + VN)(FN + VN)}{n}\right)\right] \qquad (3)$$

Neste estudo, para o desempenho do modelo ser considerado confiável foram utilizados valores de *CCI* superiores à 70% e *K* superior à 0,3, enquanto que modelos

que apresentaram valores inferiores a estes limites foram considerados de desempenho irrelevante (D'heygere et al., 2006).

A comparação entre as respostas dos modelos de predição SVMs e CTs foi realizada através de testes-*t* independentes. Para cada rede, a comparação dos resultados dos modelos antes e após a seleção com GAs foi realizada com testes-*t* pareados.

3. RESULTADOS

3.1 Variáveis selecionadas pelos algoritmos genéticos

As variáveis que foram selecionadas pelos GAs de Goldberg estão representados na Tabela 3 para CTs e na Tabela 4 para SVMs. Houve uma distinção clara entre o número de variáveis que foram selecionadas com os dois sistemas de aprendizagem. A redução do número de variáveis foi consideravelmente menor para SVMs do que para as CTs.

Para as CTs, as maiores reduções ocorreram para *Capitella* sp. onde somente uma variável foi considerada explicativa. As menores reduções foram para *H. australis*, com 11 variáveis, além de *A. brasiliana* e *Polydora* sp., com 10 variáveis. Os demais táxons tiveram entre 2 a 6 variáveis selecionadas pelos GAs. Nas SVMs, os táxons *G. multidens*, *N. succinea*, *Sigambra* sp. tiveram somente uma variável selecionada, enquanto que para *Bulla* sp e Oligochaeta sp1 foram selecionadas 5 e 6 variáveis, respectivamente. Os demais táxons tiveram entre 7 e 12 variáveis selecionadas. Em ambas as redes, para Tubificidae nenhum subconjunto de variáveis ambientais foi selecionado como explicativo, por isso, não foram construidos modelos para o táxon após a seleção de variáveis, portanto os valores das medidas de desempenho (*RMSE*, *CCI*, *K*) não foram calculados.

Os GAs em conjunto com as CTs identificaram que as variáveis epicoprostanol, salinidade e carbonato de cálcio são as mais importantes na construção dos modelos de CTs, pois foram as variáveis mais vezes selecionadas, sendo selecionadas 11, 9 e 8 vezes, respectivamente. Já para as SVMs, foram selecionadas mais vezes a matéria orgânica (13), o nitrogênio total (12) e o pH (11). As demais variáveis foram selecionas oito vezes ou mais.

317 **Tabela 3**

318 As variáveis selecionadas pelos algoritmos genéticos (GAs) para árvores de classificação (CT)

Táxon	SAL	PH	RED	COP	EPI	COL	BRA	EST	CAS	ARE	SIL	ARG	COT	CHL	FEO	NT	PT	MO	CaCO3
Anomalocardia brasiliana		x	x		x	X	x			x	x		x				x		x
Bulla sp.	x				x		x												
Caprella scaura							x	x		x		x				x			x
Capitellidae					x								x				x	x	
Capitella sp.					x														
Glycinde multidens	x			x															
Heleobia australis	x	x	x		x	x			x		x		x			x		x	x
Laoenereis culveri			x		x	x													
Neanthes succinea			x						x		x								
Oligochaeta sp1				x			x				x					x	x		
Orbiniidae									x					x					
Ostracoda	x	x	x		x	x													
Polydora sp				x	x	x	x					x		x	x	x	x		x
Prionospio sp	x	x	x										x						x
Streblospio benedicti	x				x	x		x										x	
Sigambra sp.		x										x							
Sternaspis sp.	x			x				x					x						x
Tagelus sp	x						x										x		x
Tellina lineata					x			x											
Tubificidae	x	X	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

319

320

321 **Tabela 4**

322 As variáveis selecionadas pelos algoritmos genéticos (GAs) para máquinas de vetor de suporte (SVM).

Táxon	SAL	PH	RED	COP	EPI	COL	BRA	EST	CAS	ARE	SIL	ARG	COT	CHL	FEO	NT	PT	MO	CaCO3
Anomalocardia brasiliana			x	x	x	x	x	x		x	x	x		x				x	
Bulla sp.		x					x		x					x	x				
Caprella scaura	x			x	x	x		x						x		x	x		
Capitellidae			x		x	x	x				x	x				x			x
Capitella sp.				x	x	x		x						x				x	x
Glycinde multidens																			x
Heleobia australis	x	x	x	x			x		x				x	x		x	x	x	x
Laoenereis culveri		x	x		x	x	x		x									x	x
Neanthes succinea																		x	
Oligochaeta sp1		x	x	x					x								x		x
Orbiniidae	x					x	x		x							x	x	x	
Ostracoda	x	x	x	x	x	x		x	x							x	x	x	
Polydora sp			x	x				x	x	x		x		x		x	x	x	
Prionospio sp		x		x		x		x		x				x		x	x		x
Streblospio benedicti	x	x					x	x		x				x		x		x	
Sigambra sp.																		x	
Sternaspis sp.		x						x			x		x		x	x		x	
Tagelus sp	x	x		x	x		x							x	x				x
Tellina lineata	x	x	x		x		x	x							x	x			x
Tubificidae	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

3.2 Predição da presença/ausência da fauna bêntica

Para as CTs, o nível do Fator de Confiança ideal para 15 dos 20 táxons foi FC=0,5, com exceção de *N. succinea, Oligochaeta* sp, *Prionospio* sp, *Sternaspis* sp e *Tagelus* sp. Portanto, devido à maior quantidade de táxons responderem melhor a este fator de confiança e buscando minimizar discrepâncias entre os modelos desenvolvidos optou-se por utilizar somente FC=0,5 para todos os táxons. Para as SVMs, foi realizada a otimização do modelo através da aplicação de diferentes expoentes, sendo que, a maioria dos modelos apresentou melhor desempenho com expoente na unidade (1), utilizado então para o desenvolvimento de todos os modelos. Os resultados obtidos para as simulações de ambos os modelos antes e após a seleção das variáveis foram comparados através das três medidas de desempenho: *RMSE*, *CCI* e *K* (Fig.2, 3 e 4, respectivamente).

Para os valores de *RMSE* os modelos que apresentaram os menores valores supostamente foram mais eficientes na predição. Neste caso, os modelos construídos para as CTs apresentaram maiores eficiências preditivas, antes e após a seleção de variáveis ambientais. Apesar dos valores de *RMSE* antes da seleção de variáveis serem inferiores para as CTs ((±0,07) em relação aos modelos desenvolvidos para as SVMs (0,52 ±0,09), as diferenças entre os valores de *RMSE* foram mais significativas (teste-*t* independente; $p<0,01$) após a seleção de variáveis com valores médios de 0,41 (±0,04) e 0,49 (±0,1), para as CTs e SVMs, respectivamente.

Nas CTs, todos os táxons apresentaram valores de *RMSE* significativamente inferiores após a seleção das variáveis (teste-*t* pareado; $p<0,0001$), enquanto que para as SVMs, *T. lineata* apresentou valor significativamente maior (teste-*t* pareado; $p<0,05$) após a seleção de variáveis e Capitellidae, *L. culveri*, Oligochaeta sp1, *Polydora* sp. e *Tagelus* sp., mostraram valores semelhantes (Fig. 2b).

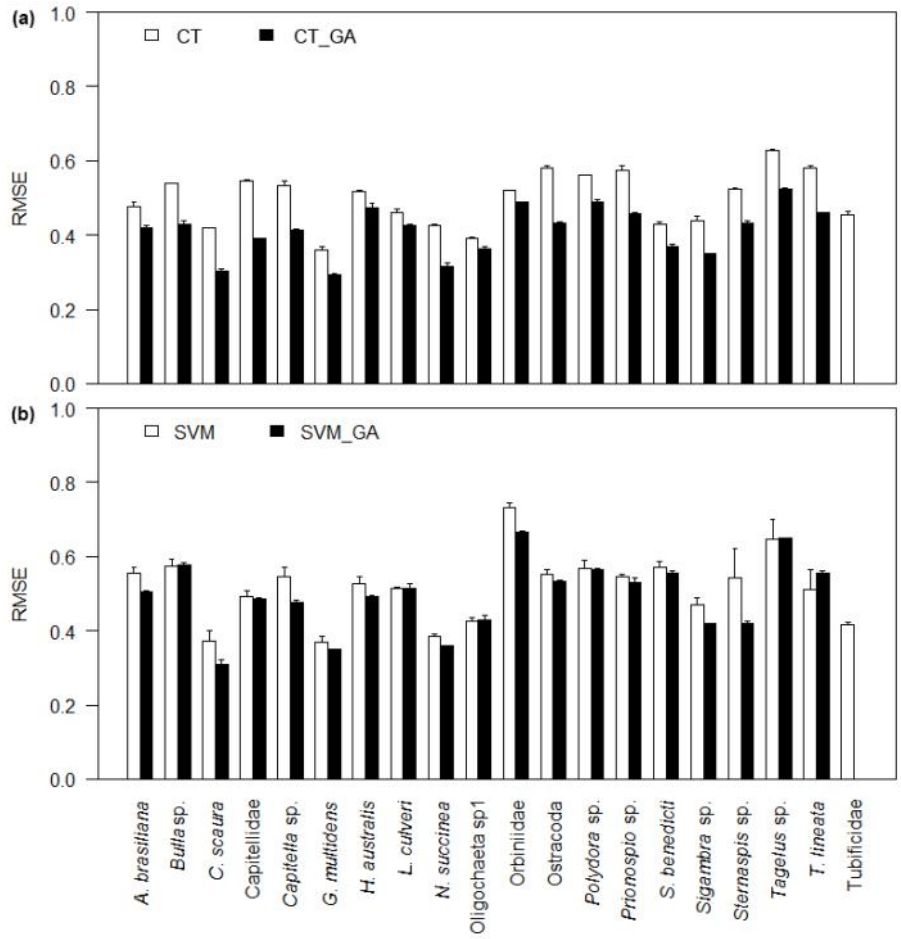


Fig.2 - (a) Raiz do erro médio quadrado (*RMSE*) das árvores de classificação para os 20 táxons analisados antes (legenda: CT) e depois da seleção das variáveis com algoritmos genéticos (CT_GA); (b) *RMSE* para máquina de vetores de suporte para os 20 táxons analisados antes (legenda: SVM) e depois da seleção das variáveis com algoritmos genéticos (SVM_GA). Barras de erro indicam erro padrão da média.

Os valores de *CCI* acima de 70% denotam que os modelos, baseado somente nesta medida, foram eficientes na predição da ocorrência dos táxons. Portanto, os modelos desenvolvidos para Orbiniidae e *Tagelus* sp., para CTs e SVMs não foram satisfatórios e os dados de predição não podem ser considerados relevantes.

Os valores de *CCI* para as SVMs foram em média superiores aos das CTs antes de ocorrer a seleção das variáveis (teste-*t* independente; $p<0,01$), exceto para *A.brasiliana*, *Bulla* sp., *L. culveri*, Oligochaeta sp1, Orbiniidae e *S. benedicti* (Fig. 3a e b). Após a fase de seleção, o *CCI* para as CTs passaram a ser mais elevados em comparação aos valores de *CCI* das SVMs (teste-*t* independente; $p<0,01$), com exceção de *C. scaura*, *H. australis* e *Sternaspis* sp. (Fig. 3a e b).

Analisando os táxons separadamente, todos os modelos desenvolvidos com CTs, exceto para Orbiniidae, registraram aumento significativo (teste-*t* pareado; $p<0,01$)

de *CCI* após a seleção das variáveis, com aumento médio de 7,6% e máximos de 15,7% para *Tagelus* sp e 14,3% para Capitellidae. Para as SVMs, aumento significativo (teste-*t* pareado; $p<0,05$) de eficiência de predição foi identificado somente para *A. brasiliana*, *Tagelus* sp. e Orbiniidae. Nos demais táxons houve aumento, porém os valores antes e após o procedimento de seleção não foram significativamente diferentes. O efeito da seleção de variáveis para SVM foi pouco acentuado, com aumento médio de 2,9%, atingindo máximos para Orbiniidae e *Tagelus* sp., com 9,6% de aumento em ambos. Estes táxons, juntamente com *A. brasiliana* (5,9%), foram os únicos que mostraram um aumento de *CCI* que superou ao das CTs (Fig. 3a e b).

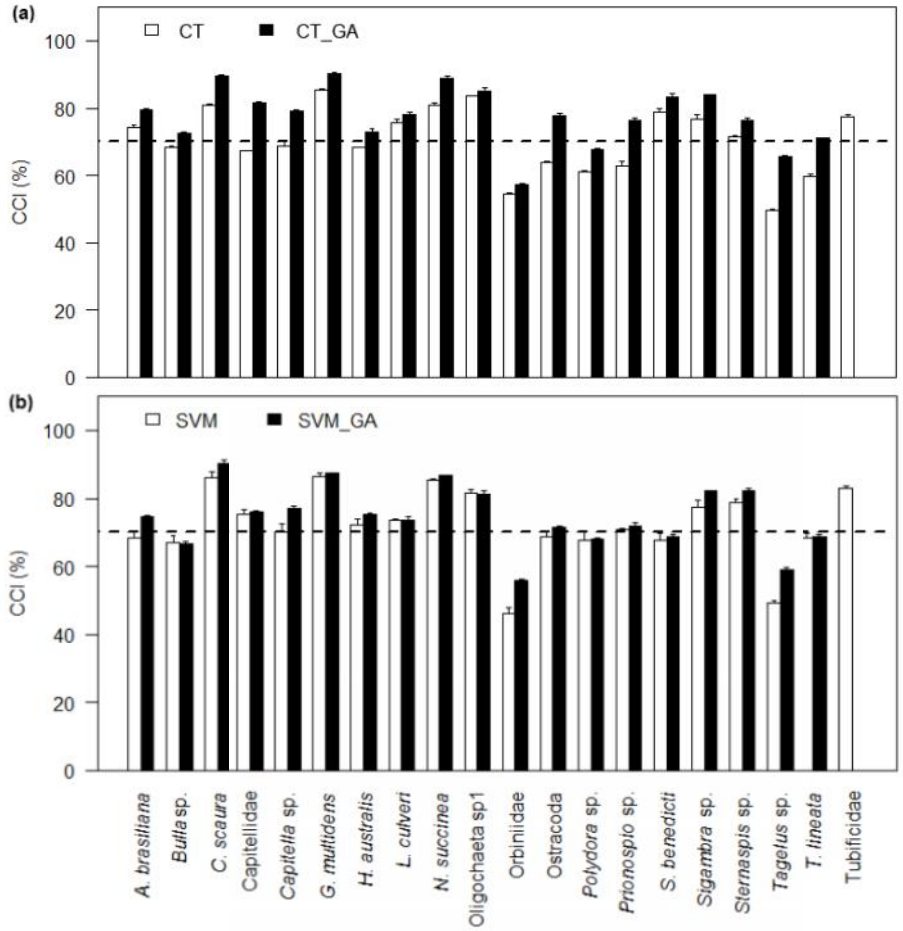


Fig.3 - (a) Instâncias corretamente classificadas (*CCI*) das árvores de classificação para os 20 táxons analisados antes (legenda: CT) e depois da seleção das variáveis com algoritmos genéticos (CT_GA); (b) *CCI* para as máquinas de vetores de suporte para os 20 táxons analisados antes (legenda: SVM) e depois da seleção das variáveis com algoritmos genéticos (SVM_GA). A linha pontilhada determina o limite inferior para que os valores do *CCI* dos modelos construídos possam ser considerados confiáveis.

A última medida de desempenho foi o cálculo do *K* de Cohen e valores acima de 0,3 foram considerados com eficiência moderada, enquanto *K* inferior a 0,2 implicou que os modelos não podem ser considerados relevantes. Baseados nestes limites, os modelos desenvolvidos de CTs para Capitellidae, *Capitella* sp., *N. succinea*, Orbiniidae, *Sigambra* sp., *T. lineata* e Tubificidae e para os modelos de SVM para *Capitella* sp., *G. multidens*, *N. succinea*, Orbiniidae, *Sigambra* sp., *Tagelus* sp. e Tubificidae não foram considerados relevantes antes da seleção das variáveis ambientais com os GAs (Fig. 4a e b). Após a seleção das variáveis, Orbiniidae para CTs e *Tagelus* sp., *Sigambra* sp., Orbiniidae e *Capitella* sp. para SVMs continuaram registrando valores de *K* inferiores a 0,2.

Antes da seleção das variáveis, as SVMs foram estatisticamente superiores às CTs (teste-t independente; $p<0,05$) para *C. scaura*, Capitellidae, Ostracoda, *Polydora* sp., *Prionospio* sp., *Sternaspis* sp. e *T. lineata*. Porém, após a seleção, somente para *C. scaura*, *H. australis* e *Sternaspis* sp. este resultado foi encontrado. O aumento médio do *K* para as SVMs foi muito baixo, com valor de 0,02, enquanto que para as CTs o aumento foi de 0,14. O fato das CTs superarem as SVMs após a seleção das variáveis também foi observado para o *RMSE* e o *CCI*.

Para as CTs, antes da seleção, os táxons *A. brasiliana*, *Bulla* sp., *G. multidens*, *H. australis*, *L. culveri*, Oligochaeta sp1, *S. benedicti* e *Sternaspis* sp. obtiveram valores de *K* acima de 0,3, enquanto que após a seleção das variáveis, todos os táxons mostraram valores superiores a 0,3, exceto *Capitella* sp., Orbiniidae, *Sigambra* sp. e Tubificidae. Os melhores resultados antes e após seleção foi para Oligochaeta sp1 e *S. benedicti*. O maior aumento de desempenho foi identificado para Ostracoda (0,27), Capitellidae (0,28) e *Tagelus* sp. (0,3). Para as SVMs, somente 50% dos táxons apresentaram valores superiores a 0,3 antes e após a seleção das variáveis (Fig. 4b). Os táxons *Bulla* sp., Capitellidae, *G. multidens*, *L. culveri*, *N. succinea* e Oligochaeta sp1 tiveram valores de *K* inferiores após a seleção das variáveis. Os maiores aumentos foram para *A. brasiliana* (0,14), *Tagelus* sp. (0,17) e *Sigambra* sp. (0,18).

Através da análise conjunta de *CCI* e *K* dos modelos de CTs e SVMs, antes e após o procedimento de seleção de variáveis, foi possível identificar que para Orbiniidae e *Tagelus* sp. os resultados de predição foram os menos confiáveis, pois ambos apresentaram valores de *CCI* e *K* abaixo dos limites estipulados (Fig.3 e Fig.4). No entanto, para *Sigambra* sp., para CTs e SVM, o *CCI* encontrado foi alto, enquanto

os valores de *K* foram baixos, o que demonstra que este resultado pode ter sido alcançado com base ao acaso.

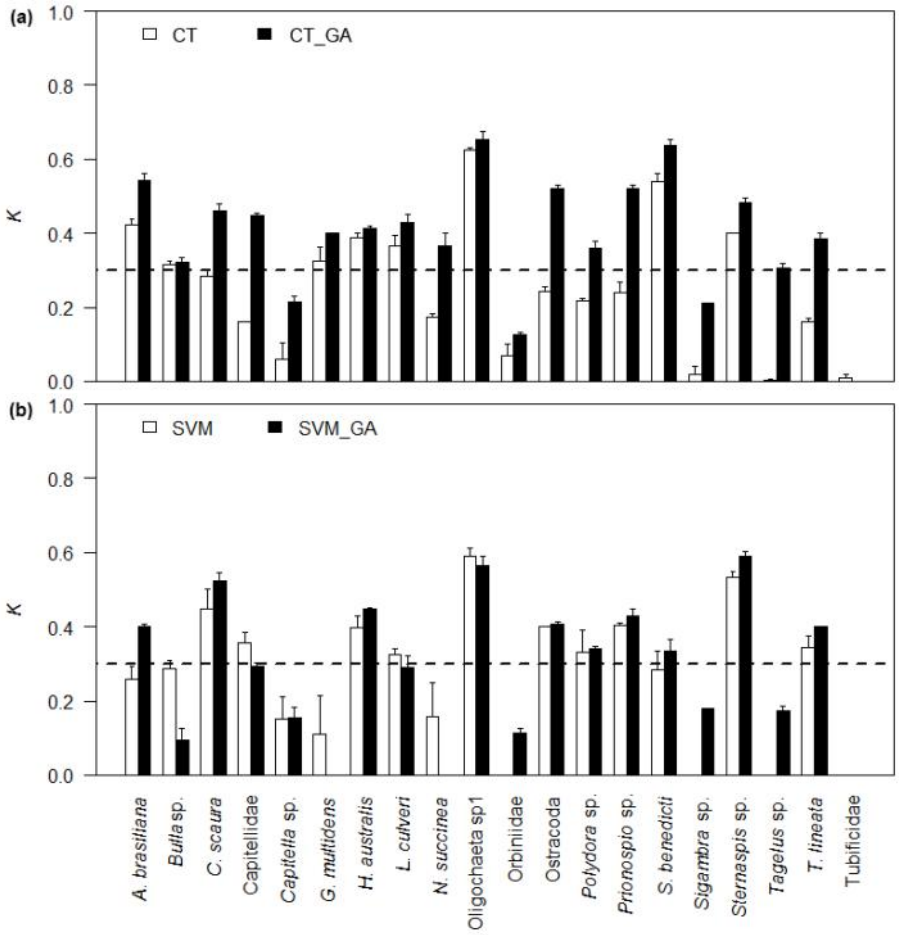


Fig.4 - (a) *K* de Cohen das árvores de classificação (CT) para os 20 táxons analisados antes (legenda: CT) e após a seleção das variáveis com algoritmos genéticos (CT_GA); (b) *K de* Cohen do modelo máquinas de vetores de suporte para os 20 táxons analisados antes (legenda: SVM) e após a seleção das variáveis com algoritmos genéticos (SVM_GA). A linha pontilhada determina o limite inferior a partir do qual os valores de *K* dos modelos construídos passam a ser considerados confiáveis.

Uma vantagem das árvores de classificação é que sua estrutura é transparente e permite mostrar o resultado dos estágios de seleção das variáveis. Por exemplo, o modelo de CT construído para *S. benedicti*, após o uso dos GAs, é apresentado na Fig.6. A partir desta árvore, é possível derivar regras gerais (*SE... ENTÃO...*) para a presença ou ausência de *S. benedicti*. Estas condicionais podem ser deduzidas a partir da árvore, seguindo a rota de uma folha até a raiz. Nesse caso, houve ausência de *S. benedicti* quando a concentração de estigmasterol foi superior a 3,78 $\mu g.g^{-1}$, combinados com

valores de colesterol acima de 2,48 $mg.g^{-1}$. Para valores de estigmasterol abaixo de 3,78 $\mu g.g^{-1}$, a ausência foi em salinidade entre 23 e 20 e matéria orgânica abaixo de 3,56%.

Quase todos os locais com presença de *S.benedicti* foram previstos corretamente (92,4%), enquanto que a ausência foi prevista corretamente em 75,6% (Tabela 5). Os dados previstos incorretamente para presença e ausência foram baseados na variável colesterol. A diferença de folhas utilizadas na construção das árvores de *S. benedicti* foi baixa, com nove folhas e cinco variáveis antes da seleção, enquanto que após a seleção das variáveis, foram sete folhas e quatro variáveis. No entanto, embora o tamanho da árvore mudasse pouco antes e após o procedimento de seleção, o sucesso de previsão antes foi significativamente inferior que após a seleção, com valores médios de *CCI* de 78,5% e *K* de 0, 54 antes e de 83,5% e 0,64 após seleção com os GAs.

Para as SVMs, entretanto, não é possível identificar a atuação das variáveis individualmente sobre o táxon *S. benedicti.* O sucesso de previsão para presença foi de 83% enquanto que a ausência foi prevista corretamente somente em 46,3 % (Tabela 5). A diferença entre os valores de *CCI* antes e após da seleção foi baixa, assumindo valores inferiores a 70% em ambas as fases (67,6 e 68,8% antes e após, respectivamente). Já o valor de *K* antes da seleção foi de 0,28 e após de 0,33.

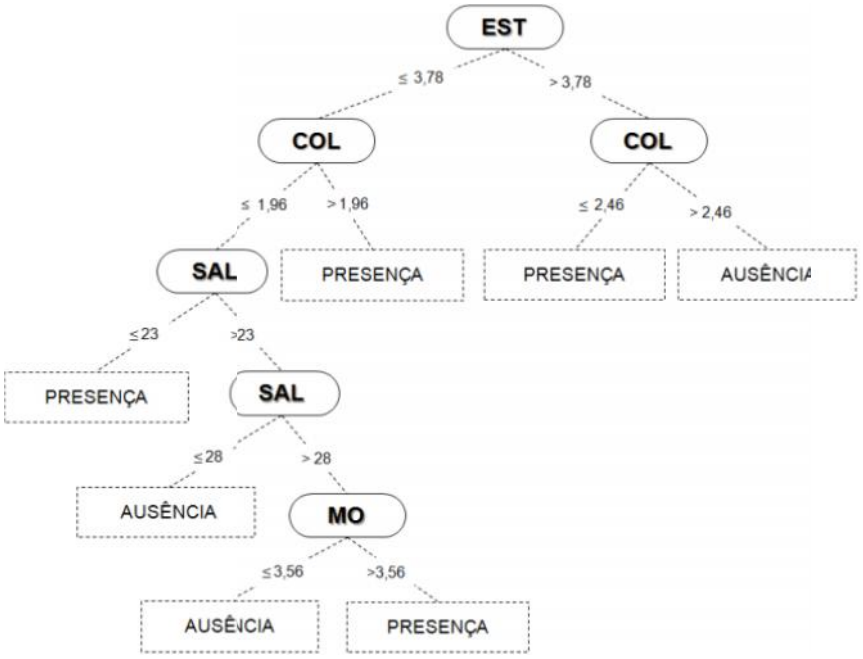


Fig.5 - Árvore de classificação para *S. benedicti* após seleção das variáveis ambientais através da aplicação dos algoritmos genéticos (EST = estigmasterol ($\mu g.g^{-1}$); COL = colesterol ($\mu g.g^{-1}$); SAL = salinidade; MO = matéria orgânica (%))

Tabela 5
Matriz de confusão para *S. benedicti* utilizando as árvores de classificação após seleção de variáveis.

		Predito	
		Presente	Ausente
Atual	Presente	*61*	*5*
	Ausente	*10*	*31*

459

Tabela 6
Matriz de confusão para *S. benedicti* utilizando as máquinas de vetores de suporte após seleção de variáveis.

		Predito	
		Presente	Ausente
Atual	Presente	*55*	*11*
	Ausente	*22*	*19*

460

461 ## 4. DISCUSSÃO

462

463 Entre os 20 táxons modelados, 17 obtiveram sucesso para as árvores de
464 classificação (CTs), identificando as relações entre a distribuição dos táxons e as
465 variáveis ambientais, enquanto que para as máquinas de vetores de suporte (SVMs),
466 apenas 10 táxons foram modelados com sucesso. Devido a estes resultados, é possível
467 identificar que um melhor desempenho foi obtido para os modelos desenvolvidos para
468 as CTs, com e sem a seleção de variáveis com os algoritmos genéticos (GAs). A
469 avaliação dos resultados obtidos para os modelos foi mais satisfatória quando foi
470 realizada a análise conjunta das medidas de desempenho, principalmente para *CCI* e *K*.
471 Segundo Manel et al. (2001), quando se realiza a modelagem de organismos
472 em relação às condicionantes ambientais, inevitavelmente a ocorrência destes
473 organismos varia entre os conjuntos utilizados no treinamento e validação dos modelos,
474 exercendo efeitos importantes sobre algumas medidas de desempenho do modelo. Entre
475 estas medidas, os valores de *CCI* são particularmente sensíveis à frequência de
476 ocorrências dos táxons, enquanto valores de *K*, por exemplo, não são afetados na
477 mesma proporção (Fielding e Bell, 1997; Manel et al., 1999; Dedecker et al., 2002;
478 D'heygere et al., 2006). Por isto, a análise conjunta das medidas de desempenho foi a
479 maneira mais plausível de identificar a eficiência dos modelos desenvolvidos, para as
480 CTs e SVMs.
481 A avaliação da composição das comunidades bênticas fornece informações
482 úteis sobre a qualidade ecológica de determinado local, já que estes organismos são

reconhecidamente sensíveis às perturbações no ambiente (Snyder et al., 2002; Ambelu et al., 2010). No entanto, a previsão da ocorrência destes organismos é dificultada pela complexidade de processos envolvidos na sua distribuição (Guisan & Zimmermann, 2000). A estabilidade dos habitats pode ser afetada por diferentes processos, como o enriquecimento orgânico, atividade de dragagem, aporte de metais pesados e até mesmo a hidrodinâmica local, agindo sobre as associações bênticas (Snyder et al., 2002). Possíveis mudanças no ambiente afetam a estrutura das associações bênticas, podendo ocorrer a supressão de espécies mais sensíveis e o aumento de espécies mais tolerantes as novas condições ambientais.

Além dos processos ambientais supracitados, outro fator que pode afetar os desempenhos preditivos dos modelos é a alta ou a baixa frequência dos organismos no ambiente. Segundo D'heygere et al. (2006) e Dakou et al. (2006a), uma baixa frequência dos táxons nos locais de amostragem dificulta a extração de informações gerais e confiáveis sobre as preferências de habitat destes táxons e neste caso os modelos tendem a identificar mais facilmente a ausência destes organismos. O contrário também pode ocorrer com organismos muito frequentes. Este comportamento foi registrado para *N. succinea,* que ocorre em somente 13% dos locais. Os modelos desenvolvidos para *N. succinea*, no geral foram fortemente afetados pela baixa presença do táxon, identificado através dos altos valores de *CCI* e relativamente baixos valores de *K*. Este comportamento do modelo pode estar refletindo o padrão anormal que o táxon apresentou ao longo da área de estudo. Segundo (Pearson e Rosenberg, 1978), *N. succinea* é associada aos grupos que indicam enriquecimento orgânico por poluição ou fontes naturais, sendo comum em locais onde há dominância dos táxons *S. benedicti*, *Capitella* sp. e *Polydora* sp.. Entretanto, mesmo com a aplicação dos GAs, não foi possível identificar correlação de *N. succinea* com variáveis que denotam enriquecimento orgânico, enquanto *S. benedicti*, *Capitella* sp. e *Polydora* sp. apresentaram forte relação com estas variáveis. Supostamente outros fatores, como disponibilidade de alimento, competição intra e interespecífica e descritores ambientais não mensurados podem estar agindo na supressão de *N. succinea* em locais onde o táxon normalmente seria encontrado.

A aplicação dos GAs mostrou que estes algoritmos podem ser eficazes na detecção e seleção de recursos para a construção de outros modelos. O uso de GAs no pré-processamento dos dados ou de forma híbrida, como no caso deste estudo, é amplamente divulgado por outros estudos e, em todos os casos em que foram aplicados

houve uma melhora significativa da eficiência dos modelos preditivos (Turney, 1995; Cha e Tappert, 2009; Zintzaras e Kowald, 2010). As variáveis que não foram selecionadas podem ser vistas como irrelevantes para o táxon em particular (Witten e Frank, 2000) e na prática, os atributos redundantes podem confundir os sistemas de classificação aumentando a sua complexidade (D'heygere et al., 2003). A eliminação de dados irrelevantes aumenta a facilidade de interpretação das tendências reveladas nos dados, com foco somente nas variáveis mais importantes (Hoang et al., 2010).

Os resultados obtidos neste estudo mostram que as SVMs são mais adequadas para obter uma visão global do efeito das variáveis ambientais porque elas são capazes de avaliar simultaneamente o efeito de todas as variáveis sobre a presença/ausência dos táxons, enquanto que as CTs são mais adequadas para identificar o efeito das variáveis individualmente na ocorrência da fauna, pois consideram apenas uma variável em cada divisão (Hoang et al., 2010). Estudos comparativos entre SVMs e CTs mostraram que, para a fauna bêntica de rios, o desempenho preditivo das SVMs foi melhor que das CTs e a maioria dos táxons apresentou uma melhoria consideravelmente maior de desempenho das SVMs após a seleção de variáveis ambientais (Hoang et al., 2010; Ambelu et al., 2010). No entanto, neste estudo, a CTs foram mais eficazes na predição da fauna bêntica que as SVMs. Algumas condicionantes podem ser responsáveis por este desempenho inferior das SVMs, entre elas, o seu potencial uso pode ser dificultado por um pequeno número de amostras em comparação com um grande número de variáveis (Pai e Hong, 2005). Além disso, as SVMs só recentemente tornaram-se populares na modelagem de dados ecológicos (Ambelu et al., 2010; Hoang et al., 2010; Pino-Mejias et al., 2010; Lock e Goethals, 2012) e estes estudos são voltados principalmente para a modelagem da fauna de rios. Hoang et al. (2010) destaca ainda em seu estudo que apesar de ter obtido resultados superiores para SVM, as CTs são mais facilmente aplicadas por não-especialistas, o que pode ser uma vantagem das CTs.

Apesar de algumas limitações reveladas por este estudo, nossos resultados indicam que, para dados ecológicos que muitas vezes exibem alta variabilidade inerente, as técnicas de inteligência artificial utilizadas podem ajudar no esclarecimento das relações entre a presença ou ausência da fauna bêntica e as variáveis ambientais. Porém, estudos desenvolvidos para estas redes, em destaque para as SVMs, ainda estão concentrados na fauna bêntica de rios, e poucas informações estão disponíveis sobre o uso destas na fauna estuarina onde a dinâmica ambiental é muito mais complexa e mais fatores ambientais estão envolvidos da distribuição da fauna bêntica.

5. REFERÊNCIAS

Allan, J.D., 2004, Landscapes and riverscapes: the influence of land-use on stream ecosystems. Annual Review of Ecology, Evolution, and Systematics 35, 257-284

Ambelu, A., Lock, K., Goethals, P.L.M., 2010. Comparison of modeling techniques to predict macroinvertebrate community composition in rivers of Ethiopia. Ecological Informatics 5, 147-152.

Bratko, I., 1989. Mbachine learning. In: Gilhooly K.J. (ed) Human and machine problem solving. Pelnum Press, New York and London, 265-287.

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2, 121-167.

Cha S.H., Tappert C.C., 2009. A genetic algorithm for constructing compact binary decision trees, Journal of Pattern Recognition Research 4, 1-13.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37-46.

Culp J.M., Cash K.J., Wrona F.J., 2000. Cumulative effects assessment for the Northern River Basins study. Journal of Aquatic Ecosystem Stress and Recovery 8, 87-94.

D'heygere, T., Goethals, P.L.M., De Pauw, N., 2003. Genetic algorithms for optimisation of predicitive ecosystem models based on decision trees and neural networks. Ecological Modelling 160, 291-300.

D'heygere, T., Goethals, P.L.M., De Pauw, N., 2006. Genetic algorithms for optimisation of predictive ecosystems models based on decision trees and neural networks. Ecological Modelling 195, 20-29.

Dakou, E., Goethals, P.L.M., D'heygere, T., Dedecker, A.P., Gabriels, W. and De Pauw, N., 2006a. Development of artificial neural network models predicting macroinvertebrate taxa in the river Axios (Northern Greece), Animal Limnology. 5, 10-17.

Dakou, E., D'heygere, T., Dedecker, A.P., Goethals, P.L.M., Dimitriadou, M.L., De Pauw, N., 2007. Decision tree models for prediction of macroinvertebrate taxa in the river Axios (Northern Greece), Aquat. Ecol.41, 399-411

Dedecker, A.P., Goethals, P.L.M., Gabriels, W., De Pauw, N., 2002. Comparison of artificial neural network (ANN) model development methods for prediction of macroinvertebrates communities in the Zwalm river basin in Flanders, Belgium. Scientific World J. 2, 96-104.

Dedecker A.P., Goethals P.L.M., Gabriels W., De Pauw N., 2004. Optimization of Artificial Neural Network (ANN) model design for prediction of macroinvertebrates in the Zwalm river basin (Flanders, Belgium). Ecological Modelling 174, 161-173.

Dedecker A.P, Goethals P.L.M. and De Pauw N., 2005. Sensitivity and robustness of stream model based on artificial neural networks for the simulation of different management scenarios. In: Lek S., Scardi M., Verdonschot P.F.M., Descy J.P. and Park Y.S. (eds.), Modelling Community Structure in Freshwater Ecosystems, Springer-Verlag, Berlin, 133–146.

Edwards, T.C., Cutler, D.R., Zimmermann, N.E., Geiser, L., Moisen, G.G., 2006. Effects of sample survey design on the accuracy of classification tree models in species distribution models. Ecological Modelling 199, 132-141.

Fielding, A.H. & Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation, 24, 38-49.

Gabriels, W., Goethals, P.L.M., Dedecker, A.P., Lek, S., De Pauw, N., 2007. Analysis of macrobenthic communities in Flanders, Belgium, using a stepwise input variable selection procedure with artificial neural networds. Aquatic Ecology 41, 427-441.

Ghetti, F.P., Ravera, O., 1993. A European perspective on biological monitoring. In: Loeb, S., Spacie, A. (Eds.), Biological Monitoring of Aquatic Systems. Lewis Publishers, Boca Raton.

Goethals, P.L.M., Dedecker, A.P., Gabriels, W., De Pauw, N., 2002. Development and application of predictive river ecosystem models based on classification trees and artificial

604 neural networks, In: Recknagel.F. (Ed,), Ecological Informatics, Understanding Ecology by
605 Biologically-Inspired Computation, Springer-Verlag, Berlin, 203-213.
606 Goethals, P.L.M., Dedecker, A.P., Gabriels, W., Lek, S., De Pauw, N., 2007. Applications of
607 artificial neural networks predicting macroinvertebrates in freshwaters. Aquatic Ecology
608 41, 491-508.
609 Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization and Machine Learning.
610 Addison-Wesley Publishing Company, Reading, MA, pp. 412.
611 Grasshoff, K., Ehrhardt, M., Kremling, K.,1983. Methods of Seawater Analysis 2 ed, Verlag
612 Chemie: Weinheim.
613 Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology.
614 Ecological Modelling 135, 147-168.
615 Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I.H., 2009. The WEKA
616 Data Mining Software: An Update; SIGKDD Explorations 11.
617 Harrell, F.E., Lee, K.L., Mark, D.B., 1996. Multivariate prognostic models: Issues in developing
618 models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat.
619 Med. 15, 361-387.
620 Hoang, T.U., Lock, K., Mouton, A., Goethals, P.L.M., 2010. Application of classification trees
621 and support vector machines to model the presence of macroinvertebrates in rivers in
622 Vietnam. Ecological Informatics 5, 140-146.
623 Holland, J.H., 1975. Adaptation in Natural and Artificial Systems. University of Michigan
624 Press, Ann Arbor, MI.
625 Kawakami, S.K., Montone, R.C., 2002. An efficient ethanol-based analytical protocol to
626 quantify fecal steroids in marine sediments, Journal of the Brazilian Chemical Society 13,
627 226-232.
628 Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K., 2001. Improvements to Platt's
629 SMO algorithm for SVM classifier design. Neural Computation 13, 637-649.
630 Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. Artificial Intelligence
631 Journal 97 (1-2), 273-324.
632 Kolm, H.E., Schoenenberger, M.F., Piemonte, M.R., Souza, P.S.A., Scühli, G.S., Mucciatto,
633 M.B., Mazzuco, R., 2002. Spatial variation of bacteria in surface waters of Paranaguá and
634 Antonina Bays, Paraná, Brazil. Brazilian Archives of Biology and Technology 45, 27-34.
635 Lana, P.C., Marone, E., Lopes, R.M., Machado, E. C., 2000. The subtropical estuarine complex
636 of Paranaguá Bay, Brazil, In Coastal Marine Ecosystems of Latin America, Seeliger, U.,
637 Lacerda, L.D., Kjerfve, B. (eds,), Springer Verlag, NY, USA, 467p.
638 Legendre, P., 1993. Spatial autocorrelation: trouble or new paradigm? Ecology 74: 1659-1673.
639 Legendre, P., Dale, M.R.T., Fortin, M.J., Gurevitch, J., Hohn, M., Myers, D., 2002. The
640 consequences of spatial structure for the design and analysis of ecological field surveys.
641 Ecography 25, 601-615.
642 Lock, K., Goethals, P.L.M., 2012. Habitat suitability modelling for mayflies (Ephemeroptera) in
643 Flanders (Belgium), Ecological Informatics.
644 Lorenzen, C.J., 1967. Determination of chlorophyll and phaeopigments: Spectrophotometric
645 equations. Limnology and Oceanography 12, 343-346.
646 Maldonado, C., Venkatesan, M.I., Phillips, C.R., Bayona, J.M., 2000. Distribution of
647 trialkylamines and coprostanol in San Pedro shelf sediments adjacent to a sewage outfall.
648 Marine Pollution Bulletin 40, 680-687.
649 Manel, S., Dias, J.M., Buckton, S.T., Ormerod, S.J., 1999. Alternative methods for predicting
650 species distribution: an illustration with Himalayan river birds. Journal of Applied Ecology
651 36, 734-747.
652 Manel, S., Williams, H.C., Ormerod, S.J., 2001. Evaluating presence-absence models in
653 ecology: the need to account for prevalence. Journal of Applied Ecology 38, 921-931.
654 Martins, C.C., Braun, J.A.F., Seyffert, B.H., Machado, E.C., Fillmann, G., 2010. Anthropogenic
655 organic matter inputs indicated by sedimentary fecal steroids in a large South American
656 tropical estuary (Paranaguá estuarine system, Brazil). Marine Pollution Bulletin 60, 2137-
657 2143.

658 Pai, P.F., Hong, W.C., 2005. Support vector machines with simulated annealing algorithms in
659 electricity load forecasting, Energy Conversion and Management 46, 2669-2688.
660 Pearson, T.H., Rosenberg, R., 1978. Macrobenthic succession in relation to organic enrichment
661 and pollution of the marine environment. Oceanography and Marine Biology Annual Review
662 16, 229-311.
663 Phan J., Moffitt R., Dale J., Petros J., Young A., Wang M., 2005. Improvement of SVM
664 Algorithm for microarray analysis using intelligent parameter selection, Conf. Proc. IEEE
665 Engineering in Medicine & Biology Society 5. 4838-4841.
666 Pino-Mejias, R., Cubiles-de-la-Vega, M.D., Anaya-Romero, M., Pascual-Acosta, A., Jordan-
667 Lopez, A., Bellinfante-Crocci, N., 2010. Predicting the potential habitat of oaks with data
668 mining models and the R system. Environmental Modelling & Software 25, 826-836.
669 Poff, N.L., 1997. Landscape filters and species traits: towards mechanistic understanding and
670 prediction in stream ecology. Journal of the North American Benthological Society 16,
671 391-409.
672 Pollard A.I., Yuan L.L., 2006. Community response patterns: evaluating benthic invertebrate
673 composition in metal-polluted streams. Ecological Applications, 16, 645-655.
674 Quinlan, J.R., 1986. Induction of decision trees. Machine Learning 1, 81-106.
675 Quinlan, J.R., 1993. C4.5: Program for Machine Learning. Morgan Kaufmann Publishers, San
676 Francisco. 302.
677 Ribeiro R., Torgo L., 2008. A comparative study on predicting algae blooms in Douro River,
678 Portugal. Ecological Modelling 212, 86-91.
679 Sanchez-Hernandez C., Boyd D.S., Foody G.M., 2007a. Mapping specific habitats from
680 remotely sensed imagery: support vector machine and support vector data description
681 based classification of coastal saltmarsh habitats. Ecological Informatics 2, 83-88.
682 Shan Y., Paull D. and McKay R.I., 2006. Machine learning of poorly predictable ecological
683 data. Ecological Modelling 195, 129-138.
684 Snyder, E.B., Robinson, C.T., Minshall, G.W., Rushforth, S.R., 2002. Regional patterns in
685 periphyton accrual and diatom assemblages structure in a heterogeneous nutrient
686 landscape. Canadian Journal of Fisheries and Aquatic Sciences 59, 564-577.
687 Statzner B., Bady P., Dolédec S., Schöll F., 2005. Invertebrate traits for the biomonitoring of
688 large European rivers: an initial assessment of trait patterns in least impacted river reaches.
689 Freshwater Biology 50, 2136-2161.
690 Strickland, J.L.H., Parsons T.R., 1972. A practical handbook of seawater analysis. Bulletin of
691 the Fisheries Research Board of Canada 167, 311.
692 Suguio, K., 1973. Introdução à sedimentologia. São Paulo, Edgard Blucher LTDA :317p.
693 Turney P.D., 1995. Cost-sensitive classification: empirical evaluation of a hybrid genetic
694 decision tree induction algorithm, Journal of Artificial Intelligence Research 2, 369-409.
695 Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York.
696 Vapnik V., 1998. Statistical Learning Theory. Wiley, New York.
697 Verdonschot PFM, Descy JP, Park YS (eds) Modelling community structure in freshwater
698 ecosystems. Springer-Verlag, 133-146.
699 Volkman, J. K., 1986. A review of sterol markers for marine and terrigenous organic matter.
700 Organic Geochemistry 9, 83-100.
701 Volkman, J. K., 2006. Lipids markers for marine organic matter. The Handbook of
702 Environmental Chemistry 2. Part N: 27-70.
703 Witten, I.H., Frank, E., 2000. Data Mining: Practical Machine Learning Tools and Techniques
704 with Java Implementations. Morgan Kaufmann Publishers, San Francisco.
705 Zintzaras E., Kowald A., 2010. Forest classification trees and forest support vector machines
706 algorithms: demonstration using microarray data. Computers in Biology and Medicine 40,
707 519-524.

CONCLUSÃO GERAL

Entre as limitações encontradas neste estudo está a inificiência dos modelos para alguns taxons, que possivelmente estão sendo mais influenciados por outros fatores não mensurados, assim como a pequena quantidade de dados utilizados, que podem acabar prejudicando o desempenho dos modelos, devido ao pequeno número de amostras exemplo para o treinamento das redes. Entretanto, apesar de algumas limitações, reveladas por este estudo, nossos resultados indicam que, para dados ecológicos com alta variabilidade inerente, as técnicas de inteligência artificial utilizadas podem ajudar no esclarecimento das relações entre a presença ou ausência da fauna bêntica e as variáveis ambientais.

No capitulo I, a aplicação dos mapas auto-organizáveis (SOM) nas variáveis ambientais sugere que os diferentes grupos identificados podem estar refletindo distintos processos ambientais nas planícies entremarés. A introdução dos conjuntos de dados biológicos nos mapas SOM previamente treinados com as variáveis ambientais mostrou ser uma abordagem que possibilitou analisar as relações entre as variáveis ambientais e a fauna. Portanto, o SOM pode ser usado como uma ferramenta analítica e pode identificar relações entre as unidades de amostragem, sendo eficiente na extração de informações complexas sobre os dados da fauna e sua distribuição, assim como a sua relação com as variáveis ambientais.

No capitulo II, a rede *perceptron* de múltiplas camadas (MLP) demonstrou ser útil na extração de informações sobre as relações complexas e não-lineares do conjunto de dados, seja quando todos os táxons simultaneamente foram submetidos à rede ou quando os táxons passaram pelo treinamento individualmente. Porém, os modelos mostraram-se susceptíveis às variáveis irrelevantes, ou seja, quando foram desenvolvidos modelos com todas as variáveis os resultados obtidos foram inferiores ao desempenho dos modelos com o uso somente das variáveis selecionadas pelos algoritmos genéticos (GA). Apesar de, em um primeiro momento, obtermos resultados satisfatórios, mais pesquisas são necessárias para inspecionar e validar as previsões destes modelos, principalmente pela falta de estudos relacionados a ambientes tão dinâmicos como a planícies entremarés,

No capitulo III, o desempenho das árvores de classificação (CT) foram melhores que dos modelos desenvolvidos para as máquinas de vetor de suporte (SVM). Porém,

estudos desenvolvidos para estas redes, em destaque para as SVMs, ainda estão concentrados na fauna bêntica de rios e poucas informações estão disponíveis sobre o uso destas na fauna estuarina onde a dinâmica ambiental é muito mais complexa e mais fatores ambientais estão envolvidos da distribuição da fauna bêntica.

Em todos os capítulos foi possível identificar que os modelos desenvolvidos foram afetados pelo procedimento de seleção doa táxons utilizados, ou seja, a seleção dos táxons foi realizada considerando apenas aqueles com maior ocorrência nos locais de amostragem e somente os 20 mais frequentes foram utilizados para a modelagem em todas as redes. Este procedimento provavelmente favoreceu os táxons mais tolerantes à poluição, ou seja, que habitam uma diversidade maior de condições ambientais, o que subestima a riqueza de espécies nos ambientes menos poluídos do canal. Portanto, a modelagem de outros táxons mais raros, com habitats mais seletivos, pode auxiliar na obtenção de mais informações sobre a qualidade ecológica destes ambientes.