UNIVERSIDADE DE LISBOA
FACULDADE DE FARMÁCIA

Previsão da ocorrência de *blooms* de cianobactérias na Albufeira do Roxo

Sara Margarida Nunes Ramos

Trabalho realizado sob a supervisão de

Orientador: Professor Doutor João Pedro Martins de Almeida Lopes

Co-orientador: Mestre Nuno Rafael da Conceição Brôco

Mestrado em Engenharia Farmacêutica

2016

Resumo

A ocorrência de *blooms* de cianobactérias é um fenómeno que acarreta inúmeros problemas para as entidades gestoras de serviços de águas, como a redução da qualidade de água, podendo mesmo ter consequências graves para o Homem devido à presença de cianotoxinas. Estes eventos resultam da interação complexa de vários fatores, como parâmetros de qualidade da água, condições meteorológicas e competição entre espécies, tornando-se a sua previsão um desafio particular e de elevada relevância para as entidades gestoras dos serviços de águas. Este trabalho teve como objetivo a criação de um modelo de previsão antecipada dos *blooms* de cianobactérias, utilizando dados do caso particular da Albufeira do Roxo (Portugal). Os dados correspondem aos parâmetros de monitorização de qualidade de água da Albufeira e a parâmetros meteorológicos, recolhidos entre 2007 e 2015. Construíram-se modelos recorrendo a técnicas lineares (mínimos quadrados parciais) e não lineares (redes neuronais artificiais), para prever a densidade de cianobactérias em termos de um limiar de alerta, que corresponde a 20 000 células/mL. O modelo gerado por uma rede neuronal do tipo *feedforward* com uma camada oculta com quatro nodos foi o que melhor se ajustou aos dados experimentais, apresentando um erro quadrático médio de 6,51x10^7 células/mL. Este modelo, com doze variáveis de entrada (temperatura do ar, velocidade do vento, direção do vento, temperatura da água, pH, condutividade, turvação, cota, azoto amoniacal, dureza, precipitação e radiação), foi testado para os dados de 2016. Apesar do erro de previsão para a densidade de cianobactérias ser elevado, o modelo identificou situações de alerta, cumprindo assim o objetivo proposto.

Palavras-chave: Cianobactérias; Gestão Qualidade da Água; Modelação; Previsão; Redes Neuronais Artificiais

Abstract

Cyanobacteria blooms occurrence is a phenomenon which entails numerous problems for managing bodies of water services, such as reducing the water quality and which may even have serious consequences for humans due to the presence of cyanotoxins. These events result from the complex interaction of several factors such as water quality parameters, the weather conditions and competition between species, making prediction a particular challenge of great relevance for managing bodies of water services. This work aims at developing early prediction models for cyanobacterial blooms, using data from the particular case of Albufeira do Roxo (Portugal). Data correspond to the parameters used to monitor water quality of the reservoir and meteorological data collected between 2007 and 2015. Models based on linear (partial least squares) and non-linear (artificial neural networks) techniques were built to predict cyanobacterial density in terms of an alert value, corresponding to 20 000 cells/mL. The model generated by a feedforward neural network with a four nodes hidden layer, produced the best fit to the experimental data, showing a mean square error of 6.51×10^7 cells/mL. This model, with twelve input variables (air temperature, wind speed, wind direction, water temperature, pH, conductivity, turbidity, water level, ammonium, hardness, precipitation and radiation), was tested with data collected in 2016. Although the prediction error for cyanobacterial density was high, the model successfully identified alert events, thus fulfilling the foreseen objective.

Keywords: Artificial Neural Networks; Cyanobacteria; Forecasting; Modeling; Water Quality Management

Agradecimentos

Gostaria de agradecer a todos os que contribuíram de alguma forma para a realização desta tese. Não sendo viável nomeá-los a todos, há alguns a quem não posso deixar de manifestar o meu sincero agradecimento.

Ao meu orientador, Professor João Almeida Lopes, pela orientação excecional, total apoio, disponibilidade, pelo saber transmitido, pela colaboração na solução dos problemas que surgiram no decorrer deste trabalho e pelo incentivo.

Ao Engenheiro Nuno Brôco, meu co-orientador, pela possibilidade de realização deste estágio na AdP e sugestões.

À Professora Helena Pinheiro pelo apoio, orientação e simpatia.

À Engenheira Olga Martins e restante equipa da AgdA pela disponibilização de dados e das informações solicitadas.

À Cláudia Almeida pelo apoio incansável durante o meu estágio e amizade.

A todos os que prescindiram do seu tempo para tentarem ajudar-me na obtenção de dados/informações úteis para o desenvolvimento deste trabalho, em especial ao Engenheiro José Calmeiro e ao Engenheiro Theo Fernandes.

À Doutora Elisabete Valério, do Instituto Nacional de Saúde Dr. Ricardo Jorge, pela amabilidade de me receber e esclarecer sobre os *blooms* de cianobactérias e apoio bibliográfico.

Às Mafaldas por todo o apoio informático.

Aos meus pais, tia e avó pelo apoio incondicional, força e incentivo, sem eles nada disto seria possível.

À prima Isabelinha e ao primo António pelo incentivo em voltar à Universidade e por todo o apoio.

À "tia"João pelo apoio e amizade.

À Rute e ao André por todo o incentivo e companheirismo ao longo destes dois anos.

A todos, o meu bem hajam!

Glossário

AEs Algoritmos Evolutivos

AgdA Águas Públicas do Alentejo, S.A.

ARH Administração da Região Hidrográfica

ARMA Auto-Regressive Moving Average
(em português: Modelo Autorregressivo e de Médias Móveis)

ARIMA Auto-Regressive Integrated Moving Average
(em português: Modelo Integrado Autorregressivo e de Médias Móveis)

EMAS Empresa Municipal de Água e Saneamento de Beja, E.M.

ANOVA Analysis of Variance
(em português: Análise de Variância)

EPA United States Environmental Protection Agency
(em português: Agência de Proteção Ambiental dos Estados Unidos)

ERSAR Entidade Reguladora dos Serviços de Águas e Resíduos

ETA Estação de Tratamento de Água

GP Genetic Programming
(em português: Programação genética)

IPMA Instituto Português do Mar e da Atmosfera

MLP Multilayer Perceptron
(em português: Perceptrão multicamada)

MSE *Mean Square Error*
(em português: Erro Quadrático Médio)

NARX *Nonlinear autoregressive network with exogenous inputs*
(em português: Rede Autorregressiva com entradas Exógenas Não linear)

N:P Rácio das concentrações de Azoto total e Fósforo total

PC *Principal Component*

(em português: Componente Principal)

PCA *Principal Component Analysis*

(em português: Análise de Componentes Principais)

PLS *Partial least squares*

(em português: Mínimos quadrados parciais)

R^2 Coeficiente de determinação

PSA Plano de Segurança da Água

RAN Reserva Agrícola Nacional

RER *Range Error Ratio*

(em português: Razão amplitude-erro)

RMSE *Root Mean Square Error*

(em português: Raiz do Erro Quadrático Médio)

RNA Rede Neuronal Artificial

SNIRH Sistema Nacional de Informação de Recursos Hídricos

SOM *Self-organization map*

(em português: Mapas auto-organizativos)

ST Série Temporal

ULSBA Unidade Local de Saúde do Baixo Alentejo, EPE

WHO *World Health Organization*

(em português: Organização Mundial de Saúde)

Conteúdo

Lista de Tabelas

Lista de Figuras

Capítulo 1

Motivação

É conhecido que as fluorescências (*blooms*) de cianobactérias são uma ameaça crescente para os ecossistemas aquáticos de todo o mundo. Estes fenómenos têm aumentado como consequência das alterações climáticas e da eutrofização dos sistemas aquáticos, gerada pelo enriquecimento de nutrientes (azoto e compostos ricos em fósforo) consequência do desenvolvimento urbano, agrícola e industrial. Em muitos destes *blooms*, acumulam-se compostos tóxicos (cianotoxinas), metabolitos secundários das cianobactérias. Estes compostos bioativos tóxicos além de diminuírem a qualidade da água podem afetar seriamente a saúde humana e animal [1; 2]. A deteção precoce das espécies que produzem cianotoxinas é uma ferramenta oportuna para iniciar os protocolos de prevenção, controlo e tratamento da água [1].

Segundo Downing *et al.* e Wynne *et al.* (*in* [3]) a previsão de *blooms* é um processo complexo, uma vez que estes fenómenos resultam da combinação de interação de múltiplos fatores, como a química da água, a morfologia do lago e as características da bacia hidrográfica, entre outros. Este processo holístico implica que o sucesso dos modelos de previsão ocorra apenas quando estes são desenvolvidos para um local específico e com elevado número de dados dos parâmetros utilizados no modelo [3]. A natureza imprevisível do aparecimento dos *blooms* de cianobactérias torna a sua previsão desafiante [4]. Foram desenvolvidos vários estudos cujo objetivo passou pela modelação deste fenómeno com o auxílio de metodologias não lineares, como por exemplo as redes neuronais artificiais. A opção por estes métodos prende-se com a sua capacidade em utilizar dados de natureza diversa, aprender a partir de dados históricos e incorporar relações não-lineares, o que os torna eficazes na resolução de problemas muito complexos [5].

O objetivo deste trabalho consiste no desenvolvimento de um modelo que permita a previsão da ocorrência de *blooms* de cianobactérias na Albufeira do Roxo.

Este documento está organizado da seguinte forma:

Capítulo 2 - Descrição das cianobactérias, as consequências da ocorrência de *blooms* destes organismos em origens de água para consumo humano, breve descrição do estado-da-arte, importância

deste trabalho para a entidade gestora de serviços de águas e breve descrição da metodologia utilizada.

Capítulo 3 - Apresentação dos dados utilizados neste trabalho e estratégia adotada para preparação dos mesmos para modelação.

Capítulo 4 - Exposição dos resultados das várias análises realizadas aos dados, bem como dos vários modelos obtidos.

Capítulo 5 - Por fim, neste capítulo, apresentam-se as conclusões deste trabalho, assim como as recomendações para trabalho futuro.

Capítulo 2

Introdução

2.1 O que são cianobactérias?

As cianobactérias, também conhecidas por algas azuis, são um grupo de microrganismos primitivos que partilham características com algas e bactérias - são bactérias gram-negativas contendo clorofila-a que lhes permite realizar a fotossíntese. Acredita-se que estas características lhes confere vantagens relativamente a outros organismos, tendo sido as primeiras plantas a colonizar áreas a descoberto de rocha e solo, e, de acordo com registos fósseis, existem há aproximadamente 3,5 mil milhões de anos [6–8], sendo responsáveis pela oxigenação da atmosfera terrestre [9]. Para além da capacidade de realizar a fotossíntese, algumas espécies de cianobactérias têm a capacidade de regular a sua flutuação através de vacúolos gasosos intracelulares, o que torna possível encontrar estes micro-organismos em diferentes profundidades da massa de água: à superfície, dispersas em diferentes camadas ou no fundo. Há ainda espécies que têm a capacidade de fixar o azoto na forma elementar, dissolvido na água [10].

A designação comum de cianobactérias dever-se-á a uma coloração azul-esverdeada, dada pela presença da ficocianina (pigmento assessório), característica comum às primeiras espécies utilizadas para classificação e identificação destes micro-organismos. No entanto, algumas espécies poderão ter uma coloração vermelha devido à presença de carotenoides e ficoeritrina [7; 10]. Estes micro-organismos podem apresentar diferentes morfologias e formas, como pode ser observado na Figura 2.1. As formas unicelulares são esféricas, ovoides ou cilíndricas, quando surgem em agregados, que são mantidos devido a uma bainha gelatinosa segregada durante o crescimento da colónia, podem tomar a forma de colónias irregulares ou colónias regulares (filamentos), também designados por tricomas, que por sua vez poderão ser retos ou enrolados [7].

As cianobactérias são seres bastante benéficos para o Homem, dado que são importantes produtores primários. O seu valor nutritivo é geralmente elevado e as espécies que fixam azoto contribuem também para a fertilidade do solo e das águas [6]. De acordo com Skulberg *in* [6] o uso de cianobactérias na produção de alimentos e para conversão de energia solar prevê-se promissor. É também estudado o potencial de aplicação dos metabolitos das cianobactérias à farmacologia/medicina. Estes

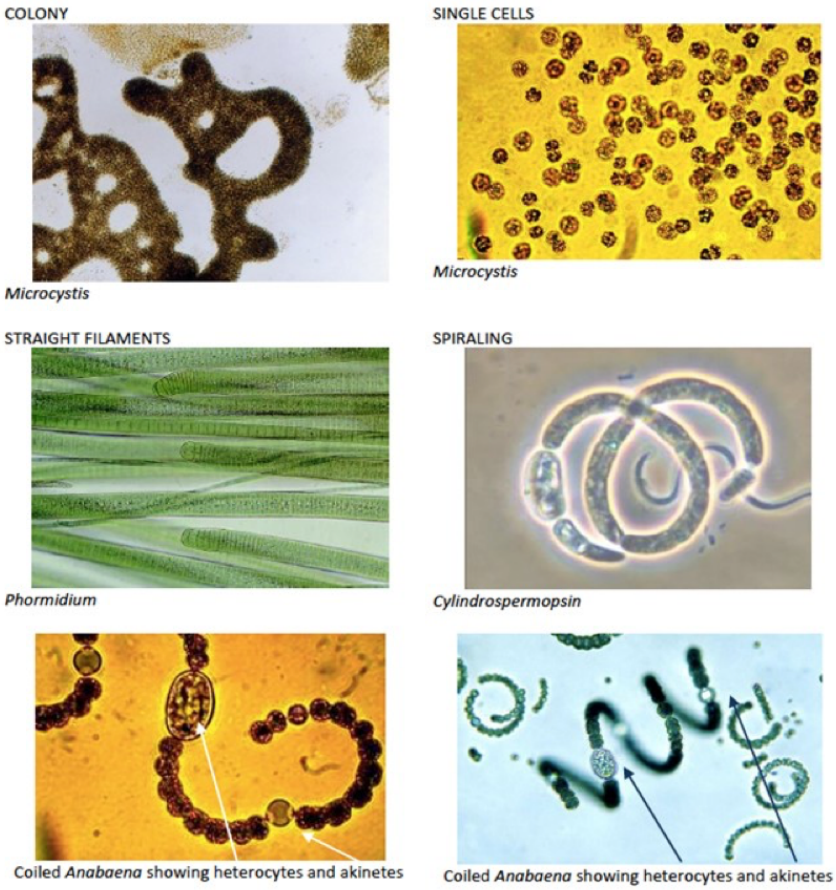

Figura 2.1: Diferentes morfologias de algumas cianobactérias (adaptado de [7])

micro-organismos ocorrem em todas as partes do mundo e em diversos ambientes (solos, água salgada e água doce), com maior frequência em águas calmas e ricas em nutrientes. Em condições ótimas formam *blooms*, tornando-se a espécie dominante, o que pode resultar numa redução da qualidade da água [8; 10; 11].

2.2 *Blooms* de cianobactérias e seus impactes

O crescimento das cianobactérias ocorre devido a uma interação complexa de vários fatores: intensidade de luz, temperatura da água, pH, concentração de dióxido de carbono, disponibilidade de nutrientes (azoto, fósforo, ferro e molibdénio e rácio N:P), características físicas da massa de água (forma e profundidade), estabilidade da coluna de água, caudal de água (rios) ou movimentos horizontais devido a afluências ou vento (reservatórios e lagos), tempo de retenção, estrutura e função do ecossistema aquático. A quantidade de organismos presente nas fontes de água está bastante interligada com a sazonalidade, podendo ocorrer de quantidades ínfimas a um valor excessivo [7; 12]. Quando se atinge rapidamente um valor elevado de células (densidade > 2000 células/mL) e há uma baixa diversidade de espécies estamos na presença de um *bloom* [13; 14].

Como referido anteriormente, a dinâmica dos *blooms* de cianobactérias é fortemente influenciada

tanto pelas condições hidrológicas como pela carga de nutrientes. O enriquecimento de água com nutrientes, que ocorre naturalmente (escorrências, deposição atmosférica de azoto), bem como a degradação da qualidade da água são muitas vezes ampliados pela atividade humana (fertilizantes, esgotos de águas urbanas e industriais) [6; 15]. Na Figura 2.2 é possível verificar a existência de um aumento na concentração de poluentes ao longo dos anos.

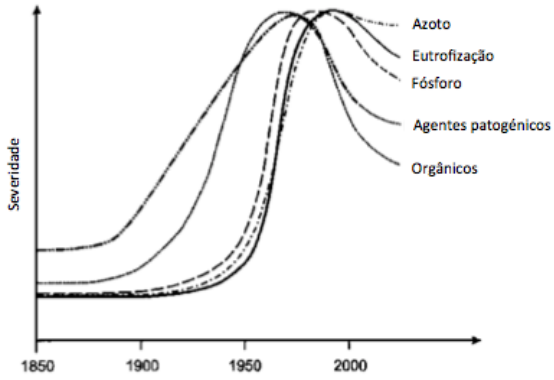

Figura 2.2: Representação esquemática do desenvolvimento da poluição das águas de superfície com patogénicos, consumo de oxigénio por matéria orgânica, fósforo e cianobactérias no Noroeste da Europa e América do Norte (adaptado de [6])

De acordo com os factos supracitados, a ocorrência de *blooms* imprevisíveis poderá ser justificada pela existência de descargas não controladas com origem antropológica [14; 16], bem como pelas condições climáticas. Uma vez que em Portugal o clima é mediterrânico, estes fenómenos ocorrem da primavera ao início do inverno [14]. Na Figura 2.3 é ilustrada a curva de crescimento de fitoplâncton ao longo do ano e respetivo mecanismo de interação com nutrientes, temperatura da água e radiação solar.

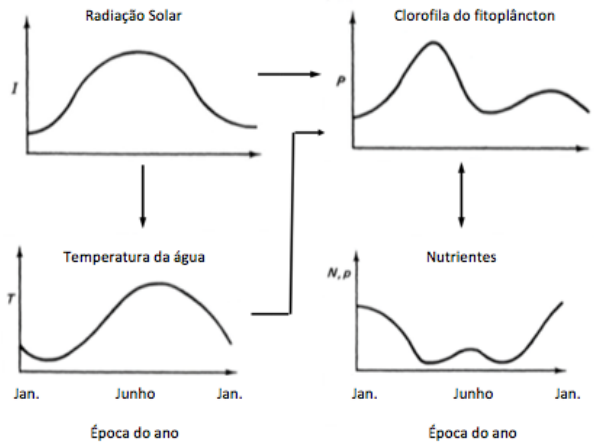

Figura 2.3: Ilustração da curva de crescimento de fitoplâncton (adaptado de [17])

Como é possível observar na Figura 2.3, na primavera há um aumento de fitoplâncton, facto que pode ser justificado por se reunirem todas as condições para o crescimento destes organismos: temperatura da água elevada, radiação forte e concentração de nutrientes suficiente. No verão há um

decaimento destes organismos, que se deve à estratificação do lago em que os nutrientes migram para águas mais profundas, ficando as águas da zona eufótica com concentrações baixas de nutrientes. Uma vez que as algas que cresceram na primavera vão morrendo e não são substituídas, o *bloom* acaba. Com a inversão de outono, que traz para a superfície água carregada de nutrientes, que estavam retidos nas zonas profundas, há um novo aumento de produtividade, embora este aumento seja menos acentuado, devido ao valor da temperatura nesta época do ano. Finalmente, com o decaimento da temperatura e da radiação solar, característico do inverno, volta a haver uma diminuição de biomassa. [17; 18].

De acordo com um estudo realizado por Galvão *et al.* [12] ocorrem *blooms* de cianobactérias na Albufeira do Roxo em diferentes períodos do ano. No intervalo de tempo em que o estudo decorreu verificou-se que os *blooms* mais expressivos surgiram entre julho e dezembro de 2005 (Figura 2.4).

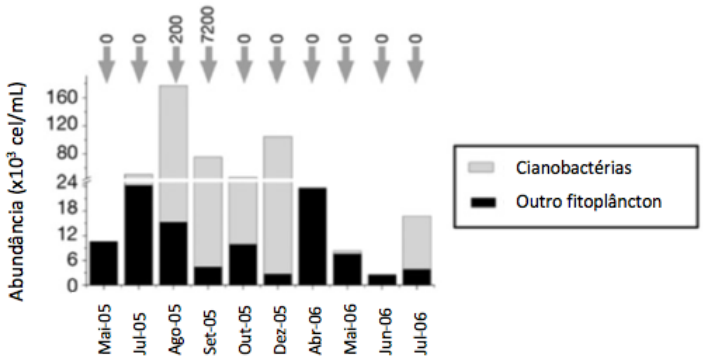

Figura 2.4: Abundância de fitoplâncton e concentrações de microcistinas (valores em cima das setas) [ng/L] (adaptado de [12])

A grande preocupação com a presença de cianobactérias nas origens de água advém da produção de metabolitos por parte destes micro-organismos, que são compostos com sabor e cheiro (em particular a geosmina e o 2-metil-isoborneol), e tóxicos, conhecidos por cianotoxinas [7; 19]. As cianotoxinas pertencem a um grupo diverso de substâncias químicas que possuem mecanismos tóxicos específicos em organismos que não são predadores diretos das cianobactérias, como os mamíferos [11]. As cianotoxinas são constituídas por péptidos cíclicos, que lhes confere elevada estabilidade estrutural em meio aquoso. Deste modo, o consumo de água contendo estas toxinas pode constituir um elevado risco para a saúde pública [6].

O Homem pode ser afetado de diversas formas: ao beber água sem tratamento adequado, ao usar a água para fins recreativos, ao comer animais aquáticos contaminados, uma vez que as toxinas são bioacumuladas nos tecidos de organismos aquáticos, como peixes e crustáceos. Há ainda relatos de contaminação de forma acidental, como, por exemplo, o caso da morte de 52 pessoas, no Brasil, durante sessões de hemodiálise, devido à presença de microcistinas na água utilizada neste procedimento [15; 20; 21]. Os efeitos adversos no Homem e outros animais são diversos, tendo sido identificadas três classes de cianotoxinas de acordo com as implicações na saúde humana: neurotóxicas, hepatotóxicas

e dermatóxicas [6; 15; 19]. Estudos indicam que as espécies hepatotóxicas são as que mais contribuem para a formação de *blooms* em vários lagos e reservatórios portugueses, sendo as microcistinas as cianotoxinas mais comuns [21–23]. De acordo com alguns estudos, as microcistinas são agentes potencialmente cancerígenos para o Homem, mesmo em baixas concentrações, nomeadamente cancro no fígado [24; 25]. Por este motivo, o valor paramétrico estabelecido no Decreto-Lei n^{o} 306/2007, de 27 de Agosto[1], para as microcistinas é de 1 µg/L. Deve-se ter em conta que uma vez que as cianobactérias não se multiplicam no corpo humano estas não são infeciosas [10].

Na Europa, os registos de intoxicações humanas são escassos [11], e Portugal não foge a esta regra. No entanto, segundo Vasconcelos [19], os baixos registos de intoxicações não se devem à inexistência destas toxinas nas fontes de água, mas pelo provável desconhecimento que estes agentes estejam na origem destes problemas. Há, ainda assim, um relato de uma intoxicação humana no Alentejo, entre habitantes que consumiram água proveniente do rio Guadiana que apresentava na altura um *bloom* da espécie *Ap.flos-aquae* [19].

A síntese da problemática dos *blooms* de cianobactérias é ilustrada na Figura 2.5.

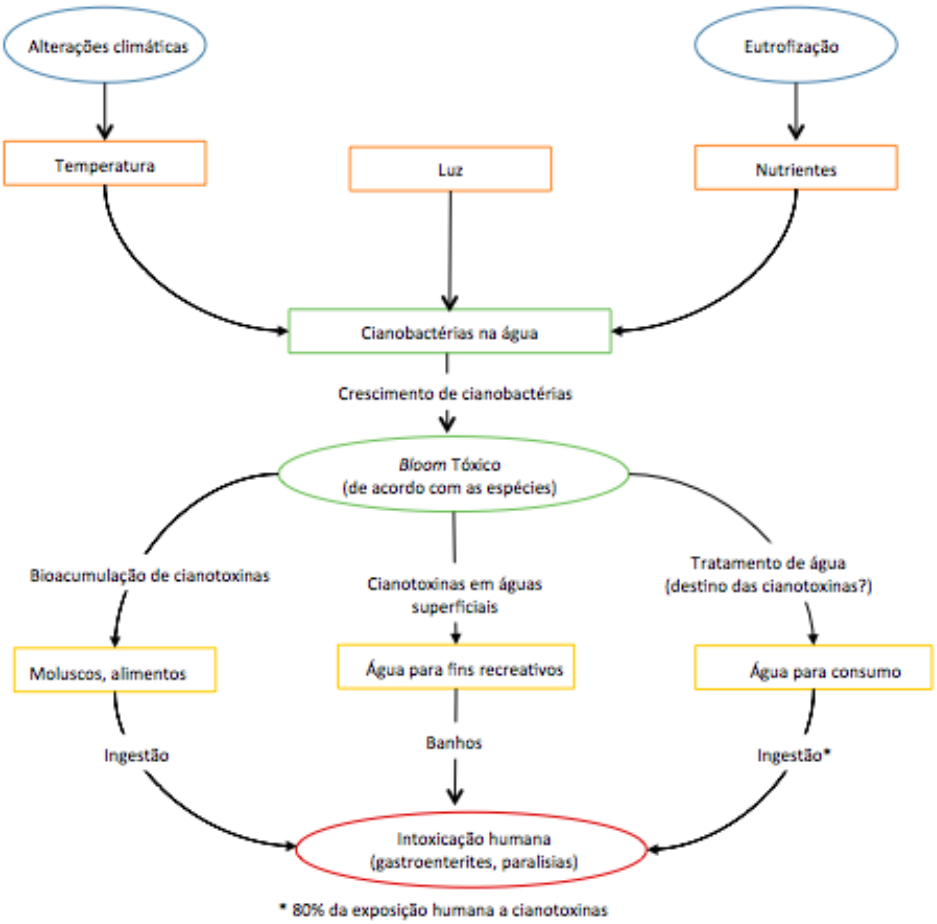

Figura 2.5: Ilustração dos fatores que levam à ocorrência de *blooms* de cianobactérias e respetivas consequências (adaptado de [26])

[1] Decreto-lei que estabelece o regime de qualidade de água destinada ao consumo humano

2.3 Revisão bibliográfica

Como descrito anteriormente, a ocorrência de *blooms* de cianobactérias é um fenómeno que, com o conhecimento atual ainda é imprevisível, mas que acarreta consequências negativas para as massas de água, e, por conseguinte, para o Homem. Por estes factos foram desenvolvidos vários estudos com o objetivo de conhecer e/ou prever estes eventos. Este subcapítulo apresenta uma breve descrição do estado-da-arte.

Existem várias espécies de cianobactérias que se adaptam a diversos habitats. Por este motivo não é possível afirmar genericamente qual a conjugação de fatores responsável pelo aparecimento de *blooms* destes organismos, como nos indica Di Gregorio [27] na sua tese de doutoramento. Ou seja, apesar do crescimento de cianobactérias ser influenciado pelos fatores ambientais, intensidade de luz, nutrientes e hidrologia da bacia, as diferentes espécies de cianobactérias necessitam de conjugações distintas destes parâmetros para se desenvolverem. A autora verificou que apesar de ser conhecido que as cianobactérias preferem elevadas temperaturas de água e elevados níveis de intensidade luminosa, há algumas espécies, incluindo as que produzem toxinas, que são exceção a esta generalização. Segundo a mesma, há evidências experimentais que mostram que a estratificação de temperaturas é, mais do que a temperatura, o fator determinante para a regulação do crescimento de cianobactérias. No que respeita à disponibilidade de nutrientes também não se pode fazer generalizações, pois assume-se que estes organismos necessitam de elevadas concentrações de fósforo e azoto, contudo, segundo a mesma autora, são observados *blooms* de cianobactérias mesmo com baixas concentrações de fósforo dissolvido, ou mesmo em águas com limitações de azoto. Este último porque há espécies que são capazes de fixar o azoto atmosférico.

A capacidade de adaptação e sobrevivência das diferentes espécies de cianobactérias em ambientes diversificados torna a sua previsão num desafio [26], o que instigou a elaboração de diversos estudos cujo objetivo seria modelar para prever este fenómeno. O desconhecimento da relação existente entre os diferentes fatores e o crescimento de cianobactérias fez com que vários autores optassem pela utilização de ferramentas que se mostraram eficazes na produção de funções de aproximação para qualquer tipo de dados - redes neuronais artificiais (RNA) [28]. Os parâmetros de entrada escolhidos, para a construção destes modelos, têm por base os dados de qualidade da água, como nutrientes (p.e. azoto, fósforo e sílica) e parâmetros físico-químicos (p.e. temperatura da água, transparência, oxigénio dissolvido e pH), considerados pelos autores como fatores que influenciam o crescimento das algas e de fácil monitorização por parte das entidades gestoras [29–31]. O conjunto de parâmetros selecionado varia de acordo com os autores. Por exemplo, Yabunaka *et al.* [29] optaram por não incluir a intensidade da luz ao longo da coluna de água, porque os valores recolhidos eram bastante variáveis, quando comparado com os dados de nutrientes e parâmetros químicos, o que iria dificultar a estimativa dos parâmetros do modelo. A inclusão de uma variável relacionada com a irradiação solar disponível ao longo da coluna de água foi também equacionada por Maier *et al.* [30]. No entanto, os autores não incluíram este parâmetro por não haver informação disponível, sendo a irradiação um fenómeno que

resulta da radiação incidente e propriedades óticas da coluna de água. Mas, à semelhança de Yabunaka *et al.* [29], admitiram que este fator estava relacionado com a cor e turvação (*inputs* do modelo). Maier *et al.* [30] referem que fatores como pastagem e competitividade, entre as diversas espécies que formam o fitoplâncton poderão influenciar o crescimento da espécie *Anabaena* (espécie âmbito do estudo), no entanto, não foram utilizados. Os autores do projeto B-Blooms [32], cujo principal objetivo foi o de perceber os *blooms* de cianobactérias que ocorrem na Bélgica, referem que estes podem estar igualmente relacionados com o número de dias sem vento forte, e não só com os parâmetros de qualidade da água. Assim, consideraram que é possível prever estes fenómenos com uma monitorização relativamente fácil utilizando estações meteorológicas e sensores na água, associados a registadores de dados.

À semelhança das entradas (*inputs*), as variáveis de saída (*outputs*) consideradas para a construção de modelos são também diferentes. Há autores que preveem a ocorrência de *blooms* de algas estimando a concentração de clorofila-a [29] ou então criando modelos para diferentes espécies que se encontram nos lagos que foram âmbito do estudo [29; 30].

Sendo o processo de modelação de *blooms* de cianobactérias complexo, a abordagem por parte de alguns autores, a este problema, passou pela utilização de mais do que um tipo de RNA ou pela utilização de duas técnicas de aprendizagem automática (*learning machines*). Oh, H. M. *et al.* [33] utilizaram duas RNAs: mapa auto-organizativo (SOM) e perceptrão multicamada (MLP). A primeira foi utilizada para dividir a comunidade de fitoplâncton em grupos de acordo com a sua composição e a segunda para identificar quais os fatores ambientais que causam a abundância de fitoplâncton em cada grupo. Este estudo permitiu identificar a temperatura da água, o azoto particulado total, a irradiação diária e o azoto total como as variáveis mais importantes para a previsão da abundância de cianobactérias. Por sua vez, Muttil, N. e Chau, K.W. [34] optaram por usar RNA e programação evolucionária para modelar e prever *blooms* de algas costeiros em Tolo Harbour (Hong Kong), utilizando, também dados de monitorização da qualidade da água e meteorológicos como variáveis de entrada dos modelos, sendo a clorofila-a a variável preditiva. Estes autores concluíram que a concentração de clorofila-a pode ser prevista tendo como única entrada valores anteriores dela própria, o que sugere a natureza autorregressiva da dinâmica das algas nas águas costeiras semi-confinadas.

Llewellyn, Chandra T. [3] praticou outra abordagem criando um modelo preditivo através de métodos de agrupamento. A autora conseguiu separar os 50 lagos estudados de acordo com alguns parâmetros de qualidade da água (condutividade, alcalinidade, fósforo total e turvação), no entanto, não conseguiu encontrar um modelo que permitisse prever em qual dos lagos iria ocorrer *blooms* de cianobactérias. Neste caso, ao contrário do esperado, não foi possível encontrar uma relação entre a clorofila-a e concentração de fósforo com a formação de *blooms* de cianobactérias.

A estratégia utilizada por Cha, Y *et al.* [35] para ultrapassar a imprevisibilidade da ocorrência de *blooms* de algas e da dominância das cianobactérias passou pelo desenvolvimento de um modelo Bayseniano. De acordo com os autores, a temperatura é o principal fator para o desencadeamento da presença de cianobactérias nas águas. A temperatura elevada quando combinada com um elevado tempo de residência (baixo caudal de saída), assim como com a estabilidade da coluna de água (quantidade

de sólidos suspensos baixos) criam as condições necessárias para o elevado número de cianobactérias num lago. Os parâmetros selecionados pelos autores como potencialmente relacionados com a ocorrência de *blooms* de cianobactérias foram: temperatura da água, condutividade, profundidade de secchi, sólidos suspensos, oxigénio dissolvido, clorofila-a, fósforo total, fósforo dissolvido total, fosfato, azoto total, azoto dissolvido total, nitrato e azoto amoniacal, precipitação diária, caudal de entrada e caudal de saída. Contrariamente ao esperado, os autores verificaram que os nutrientes, em particular as concentrações de fósforo não apresentaram elevada correlação com a presença/abundância de cianobactérias. De acordo com a análise realizada, é possível prever a quantidade de cianobactérias no Lago Paldang (Coreira do Sul) através de um modelo que tem como variáveis a temperatura da água, o caudal de saída e os sólidos suspensos.

Por sua vez, Ibelings, B. W. *et al.* [36] optaram por utilizar a lógica difusa para previsão da ocorrência/desaparecimento de *blooms* de superfície tendo por base os dados de biomassa de algas, velocidade do vento, irradiação e altura do dia. Esta última foi considerada uma vez que, segundo os autores, estes fenómenos ocorrem num período horário específico, geralmente durante a noite. De acordo com este estudo, o parâmetro meteorológico que mais afeta a formação de *blooms* de superfície é a velocidade do vento. A radiação, por sua vez, aparenta ter um duplo efeito, ou seja, uma elevada insolação promove a estabilidade da coluna de água, que potencia a formação dos *blooms*, para velocidades de vento intermédias. Por outro lado, a elevada irradiação também diminuí a capacidade de flutuação das cianobactérias, reduzindo assim o nível a que as espumas ocorrem. A lógica de difusa foi também utilizada por Lilover, M. e Laanemets, J. [37] para previsão do máximo de biomassa da *Nodularia spumigena* (cianobactéria tóxica) no Golfo da Finlândia. O modelo gerado neste estudo tem em comum com o trabalho anterior as variáveis vento e biomassa da *N. spumigena*. No entanto, estes autores utilizaram, ainda, como entradas a temperatura da camada superficial e concentração de nutrientes (fósforo e azoto) no inverno. De acordo com os resultados obtidos neste estudo, a relação entre excesso de fósforo e o valor de *N. spumigena* não se verifica todos os anos, apesar de estatisticamente haver uma correlação entre estes.

Em Portugal também se realizaram estudos com o objetivo de compreender os *blooms* de cianobactérias e de criar modelos de previsão deste fenómeno em algumas massas de água do território nacional. Na expectativa de compreender a complexidade dos *blooms* de cianobactérias, que ocorrem nas regiões do Alentejo e Algarve, Galvão, H. M. *et al.* [12] sintetizaram, num artigo, três estudos que decorreram nestas regiões. Esta análise permitiu concluir que os diferentes tipos de *blooms* de cianobactérias têm diversas origens, não podendo associar-se sempre este fenómeno ao regime de nutrientes ou a determinadas condições meteorológicas. No entanto, segundo os autores, os *blooms* de verão estão associados às temperaturas elevadas. Para além deste, as estratégias de gestão e a quantidade de água retirada da massa de água também se relacionam com este fenómeno, pois alteram a estratificação da coluna de água. A concentração de sílica influencia, igualmente, a predominância de cianobactérias, uma vez que a depleção de sílica causa a morte das diatomáceas, o que levará ao aumento de clorofitas.

Para além da tentativa de compreensão do evento, foram, ainda, criados modelos de previsão de

blooms de cianobactérias para dois reservatórios portugueses (Torrão e Crestuma), optando-se, em ambos os casos, pela utilização de RNA [5; 38]. No caso do reservatório do Torrão, foram desenvolvidos modelos para previsão da densidade de cianobactérias a três profundidades (superfície, limite da zona eufótica e perto do fundo) e para a concentração total de cianobactérias no reservatório. Os autores escolheram como entradas, dados de qualidade da água, dados meteorológicos e outros, como duração do dia lunar e estratificação de oxigénio. Das trinta variáveis utilizadas apenas seis se revelaram significativas para o desempenho do modelo: amoníaco, fosfato, oxigénio dissolvido, temperatura da água, pH e taxa de evaporação [5]. No estudo realizado para o reservatório de Crestuma os autores optaram por utilizar parâmetros físico-químicos e biológicos para previsão da abundância de cianobactérias. Esta pesquisa permitiu verificar que os modelos construídos com base nas características físico-químicas da água apresentavam melhores previsões quando comparados com os modelos criados com a incorporação da densidade de cianobactérias nas variáveis de entrada [38].

Por sua vez Ribeiro, R. e Torgo, L. [13] desenvolveram um estudo que pretendeu comparar diferentes ferramentas de modelação de *blooms* de algas e identificar o melhor modelo para amostras recolhidas no rio Douro. Nesta investigação foram utilizados dados de amostras de água da barragem de Crestuma-Lever entre 1998 e 2003, como a turvação, temperatura, pH, alcalinidade, condutividade, nitratos, cloratos, sulfatos, sílica, oxidabilidade, oxigénio dissolvido, ferro, ferro dissolvido, sólidos suspensos totais, coliformes fecais, coliformes totais, estreptococos fecais, clostrídios sulfito-redutores, germes totais a 22^{o}C, germes totais a 37^{o}C e *Escherichia coli*. Os autores deste estudo comparam três modelos diferentes: árvores de regressão, RNA e máquinas vetoriais (SVM), tendo concluído que uma variante das SVM revelou ser a ferramenta mais promissora.

Na Tabela 2.1 é apresentada a síntese do estado-de-arte descrito neste subcapítulo.

Tabela 2.1: Síntese do estado-da-arte

Autores	Ref.	Local	Variável preditiva	Entradas	Tipo de modelo	Observações
Ken-ichi Yabunaka, Masaaki Hosomi e Akihiko Murakami	45	Japão (Lago Kasumiguara)	Clorofila-a ou *Microcistis* spp., *Oscillatoria* sp., *Phormidium* spp., *Cyclotella* sp. e *Synedra rumpens*	Fosfato, nitratos, azoto, silício, temperatura da água, OD, pH, transparência, densidade de Rotifera e Diaphanosoma.	RNA retropropagação do erro	
Holger R. Maier, Graeme C. Dandy e Michael D. Burch	46	Australia (Rio Murray)	*Anabaena* spp.	Cor, turvação, temperatura, caudal, fósforo solúvel, fósforo total, óxidos de azoto e ferro total	RNA retropropagação do erro	
Hugh Wilson, Friedrich Recknagel	47	Japão (Lago Kasumiguara e Lago Biwa), Finlândia (Lago Tusulunjarvi), Australia (Rio Darling e Rio Murray) e Coreia do Sul (Lago Soyang)	Várias espécies de algas ou biomassa	Temperatura da água, concentração de fósforo e azoto e penetração da luz (clorofila-a para previsões com um mês de antecedência).	RNA *feedforward*	
Hee-Mock Oh, Chi-Yong Ahn, Jae-Won Lee, Tae-soo Chon, Kyung Hee Choi e Young-Seuk Park	49	Coreia do Sul (Barragem de Daechung)	Clorofila-a ou *Cyanophyceae*	Azoto total, azoto dissolvido total, azoto particulado total, fósforo total, fósforo dissolvido total, fósforo particulado total, temperatura, OD, pH, condutividade, turvação, profundidade de Secchi, precipitação, irradiação.	RNA: SOM e MLP	Parâmetros mais importantes: temperatura da água, azoto particulado total, irradiação e azoto total.
Nitin Muttil e Kwok-Wing Chau	50	Hong Kong (Tolo Harbour)	Clorofila-a	Clorofila-a, azoto inorgânico total, fósforo, OD, profundidade de Secchi, temperatura da água, temperatura, precipitação, radiação solar e velocidade do vento.	RNA e Programação evolucionária	Clorofila-a pode ser prevista apenas com valores anteriores da sua concentração.
Chandra T. Llewellyn	16	Estados Unidos da América (cinquenta lagos a Noroeste de Washington)	Cianobactérias (*Anabaena*, *Aphanizomenon* e *Microcystis*)	OD, temperatura, pH, condutividade, clorofila-a, alcalinidade, turvação, amoníaco, azoto total, nitratos, fósforo total e fósforo solúvel reativo.	Métodos de agrupamento	Não foi possível encontrar um modelo que permitisse identificar a ocorrência de *blooms*.
YoonKyung Cha, Seok Soon Park, Kyunghyun Kim, Myeongseop Byeon e Craig A. Stow	51	Coreia do Sul (Lago Paldang)	Cianobactérias	Temperatura da água, caudal de saída e sólidos suspensos.	Modelo Bayesiano	Temperatura é o principal fator para o desencadeamento de cianobactérias na água.
Bas W. Ibelings, Marijke Vonk, Hans F. J. Los, Diederik T. Van Der Molen e Wolf M. Mooij	42	Holanda	Cianobactérias	Biomassa de algas, velocidade do vento, irradiação e altura do dia.	Lógica difusa	Parâmetro meteorológico com maior influência na formação de *blooms* de superfície é a velocidade do vento.
Madis-Jaak Lilover e Jaan Laanemets	52	Finlândia (Golfo da Finlândia)	Biomassa de *Nodularia spumigena*	Velocidade do vento, biomassa da *N. Spumigena*, temperatura da camada superficial e concentração de fósforo e azoto.	Lógica difusa	
Rita Torres, Elisa Pereira, Vítor Vasconcelos e Luís Oliva Teles	18	Portugal (Barragem do Torrão)	Cianobactérias	Número de dias de rotação da lua, nitritos, nitratos, amoníaco, fosfato, OD, temperatura da água, pH, condutividade, estratificação de oxigénio, precipitação, taxa de evaporação, radiação, descargas.	RNA	Seis variáveis que se revelaram significativas para o modelo: Amoníaco, fosfato, OD, temperatura da água, pH e taxa de evaporação.
Luís Oliva Teles, Vítor Vasconcelos, Elisa Pereira e Martin Saker	53	Portugal (Barragem de Crestuma)	Cianobactérias	Cor, turvação, temperatura da água, pH, alcalinidade, condutividade, oxidabilidade, OD, cloretos, NO_3, SO_4, ferro solúvel, ferro total, SST, radiação, taxa de evaporação, precipitação, descargas, temperatura mínima do ar, temperatura máxima do ar, tempo de retenção, densidade de várias espécies que consitituem o fitoplâncton.	RNA	Modelos com base nas características físico-químicas da água apresentaram melhores resultados que os que incorporavam valores de densidade de cianobactérias.
Rita Ribeiro e Luís Torgo	23	Portugal (Rio Douro)	Cianobactérias	Turvação, temperatura, pH, alcalinidade, condutividade, nitratos, cloratos, sulfatos, sílica, oxidabilidade, OD, ferro, ferro dissolvido, SST, coliformes fecais, estreptococos fecais, clotrídios sulfito-redutores, germes totais a 22^{o}C e 37^{o}C e *Escherichia coli*.	Árvores de regressão, RNA, máquinas vectoriais (SVM)	

2.4 Breve descrição da Albufeira do Roxo

A Barragem e Albufeira do Roxo localizam-se na Região Hidrográfica do Sado e Mira (RH6) como ilustrado na Figura 2.6, sendo uma das origens de água do subsistema Roxo, cuja área de afetação corresponde aos concelhos de Beja, Ferreira do Alentejo e Aljustrel [39]. A capacidade total da Albufeira é de 96,312 hm^3, sendo o volume útil de 89,512 hm^3 e a cota do nível de pleno armazenamento de 136 m. A água é utilizada para irrigação e abastecimento [40].

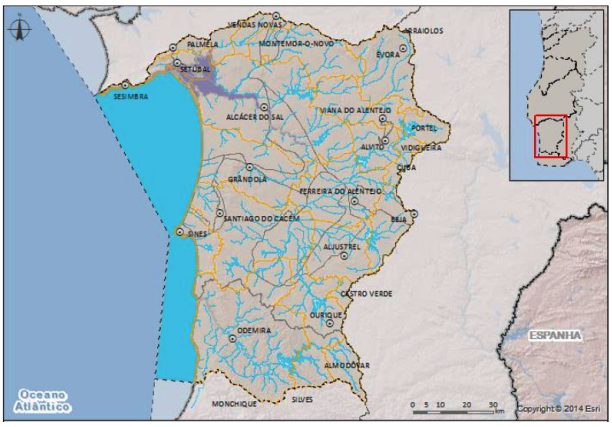

Figura 2.6: Ilustração da Região Hidrográfica do Sado e Mira (adaptado de [39])

Fontes de poluição

De acordo com o Relatório Síntese dos Planos de Gestão das Bacias Hidrográficas integradas na Região Hidrográfica (RH) do Sado e Mira (RH6) e na RH do Guadiana (RH7) a massa de água da Albufeira do Roxo sofre pressões significativas de diversos tipos, como: rejeições pontuais (suinícolas), rejeições de origem difusa (não agrícolas), captações de água (agricultura) e pressões hidromorfológicas [39]. Na Figura 2.7 ilustram-se as localizações das fontes de poluição de origem urbana e das explorações pecuárias.

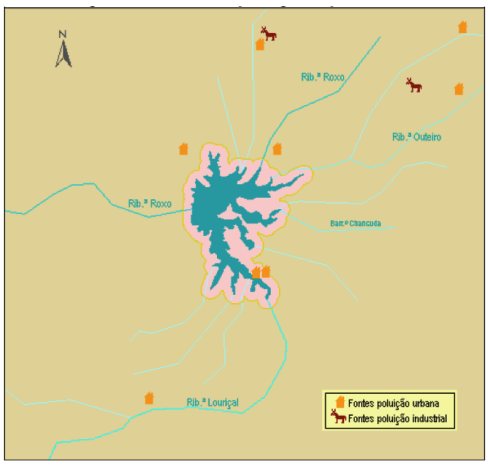

Figura 2.7: Localização das fontes de poluição tópicas da Albufeira do Roxo (adaptado de [41])

As pressões identificadas estarão relacionadas com as concentrações elevadas de fósforo na água da Albufeira, e, por conseguinte, com o facto da Albufeira do Roxo se encontrar Eutrofizada. A qualidade de água da Albufeira do Roxo também não cumpre as Diretivas n.^{o} 75/440/CE (Coli) e n.^{o} 78/659/CEE (OD + NH_3 + $NH_4{}^+$), o que conduz à classificação desta Albufeira como Zona Sensível[1]. Apesar deste cenário, o Estado da massa de água (Estado/Potencial Ecológico e Estado Químico), de acordo com o Plano de Gestão da Bacia Hidrográfica, é Bom, como exposto na Tabela 2.2 [39].

Tabela 2.2: Avaliação do estado da massa de água da Albufeira do Roxo (adaptado de [39])

Tipo de Zona protegida	Estado/Potencial Ecológico	Estado Químico	Estado (2009)	Elementos de Qualidade				
				Fitoplâncton	Qualidade biológica	FQ gerais	Poluentes específicos	Substâncias prioritárias - Estado Químico
Captações; Piscícola; Sensível	Bom	Bom	Bom	Bom	Bom	Bom	Bom	Bom

A presença de elevadas concentrações de nutrientes na água da Albufeira do Roxo contribui para o aumento da biomassa [41], na qual se incluem as cianobactérias. Como é possível verificar na Figura 2.8 a densidades destes organismos na Albufeira do Roxo (assinalado com um quadrado azul) é bastante elevada (3^a da escala).

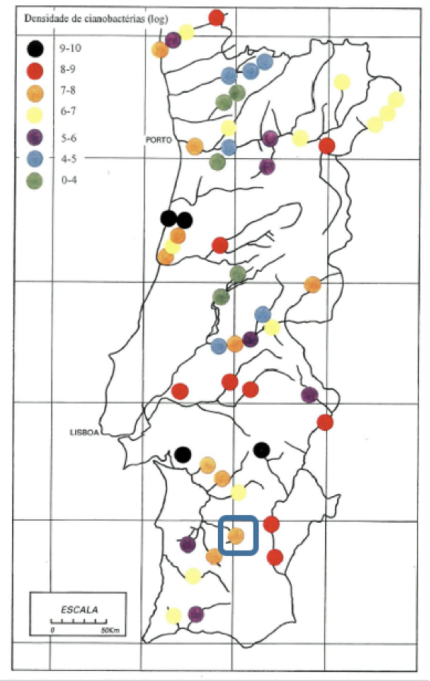

Figura 2.8: Carta de qualidade de água em função da abundância máxima de cianobactérias em águas doces portuguesas (adaptado de [19])

[1] segundo o Decreto-Lei 152/97 são lagoas naturais de água doce, outras extensões de água doce, estuários e águas costeiras que se revelam eutróficos, ou suscetíveis de se tornarem eutróficos, se não forem tomadas medidas de proteção. Bem como, águas doces de superfície destinadas à captação de água potável cujo teor em nitratos possa exceder a concentração de nitrato estabelecida nas disposições pertinentes da Diretiva n^{o}75/440/CEE, de 16 de Julho de 1975 se não forem tomadas medidas de proteção

2.5 Importância deste estudo para a entidade gestora de serviços de águas

É conhecido que cerca de 64% da população portuguesa consome água proveniente de reservas superficiais [42] e que 10% das albufeiras de Portugal Continental são classificadas como zonas sensíveis, sendo 8% devido à eutrofização e/ou elevada concentração de nutrientes [39]. O aumento de nutrientes nas fontes de água devido a ações antropogénicas leva à ocorrência de *blooms* de cianobactérias mais frequentes e imprevisíveis. Aliado a este facto está o fenómeno das alterações climáticas que, segundo a EPA [43], irá contribuir para o agravamento destes *blooms*, uma vez que é previsto um aumento da temperatura, vantagem competitiva para algumas cianobactérias que produzem cianotoxinas e um aumento de fenómenos extremos de precipitação (aumento de transportes de nutrientes) seguidos de períodos de seca (retenção dos nutrientes na albufeira por longos períodos), entre outros.

Apesar de nem todos os *blooms* de cianobactérias serem tóxicos, pois nem todos os géneros de cianobactérias produzem cianotoxinas, deverão ser assim considerados por precaução, dado que mais de 75% dos *blooms* apresentam esta característica [10]. Para além deste perigo, estes fenómenos acarretam outros impactes ecológicos e económicos como o acréscimo da dosagem de reagentes e alteração das características organoléticas da água, entre outros [16; 23; 44; 45]. Por estes motivos o aumento da ocorrência de *blooms* de cianobactérias tem sido encarado com grande preocupação pelas entidades gestoras de água para consumo humano.

No caso concreto da Albufeira do Roxo, a exploração da água para abastecimento está a cargo da AgdA - Águas Públicas do Alentejo, S. A., que pertence ao grupo Águas de Portugal (AdP). O consumo médio anual de água da Albufeira para consumo urbano é de 2,72 hm^3, dos quais são tratados, na Estação de Tratamento de Água (ETA) do Roxo, para consumo humano cerca de 500 m^3/h, que abastecem os concelhos de Aljustrel e Beja. Esta água é recolhida através de uma torre de captação, que permite captar água a três níveis de profundidade, escolhidos em função da qualidade da água da Albufeira (Figura 2.9).

Figura 2.9: Torre de Captação da Albufeira do Roxo

Nos últimos anos a AgdA deparou-se com problemas sérios para o abastecimento de água no

distrito de Beja, nomeadamente cheiro a terra e a mofo na água, que foram associados à presença de cianobactérias na água da Albufeira do Roxo [46]. Em 2015 este problema foi de tal ordem complicado que deu origem a diversas notícias, em meios de comunicação social de âmbito regional e nacional, como é possível ler nos títulos de notícias ilustrados na Figura 2.10.

Figura 2.10: Exemplos de títulos de notícias de agosto de 2015 acerca das consequências da presença de cianobactérias na Albufeira do Roxo

A gravidade do fenómeno levou à criação de um Grupo de Acompanhamento da Qualidade da Água, constituído por diferentes entidades: AgdA, Agência Portuguesa do Ambiente (APA), Câmara Municipal de Aljustrel, Empresa Municipal de Água e Saneamento de Beja, E.M (EMAS), Entidade Reguladora dos Serviços de Água e Resíduos (ERSAR) e Unidade Local de Saúde do Baixo Alentejo, EPE (ULSBA). Este grupo concluiu que seria necessário implementar medidas de adaptação ao nível do tratamento, listadas de seguida, maximizar o uso de água subterrânea para diluição da água proveniente da ETA do Roxo e proceder à reabilitação da ETA [46]:

- utilização de carvão ativado, por forma a tornar o tratamento mais eficaz;
- substituição da areia dos filtros;
- instalação de uma oxidação intermédia e
- aumento da frequência de lavagem e higienização de todos os órgãos, em particular os decantadores.

A adaptação do tratamento realizado nas ETA aquando da captação de água em bacias com cianobactérias potencialmente produtoras de cianotoxinas tem de ser eficaz na remoção destas substâncias e não deve causar a lise das células (usa-se por exemplo, carvão ativado em pó, oxidação por ozono ou cloro, entre outras), uma vez que as toxinas produzidas só serão libertadas quando há decomposição das cianobactérias [19; 45].

Para além das medidas supracitadas ocorreram também transformações ao nível da estrutura de captação, tais como: reabilitação da torre de tomada de água, instalação de novas comportas, alteração do nível de captação e limpeza geral do interior da torre. Estas medidas resultaram num aumento de

custos significativos não apenas relacionados com o investimento de urgência, mas também resultado de um conjunto de medidas operacionais de mitigação dos impactos desta situação.

Por forma a tornar mais célere a resolução de problemas que possam por em causa o abastecimento de água para consumo humano, as empresas do grupo AdP foram incentivadas a elaborar Planos de Segurança da Água (PSA), no qual a entidade gestora de abastecimento de água terá de definir procedimentos de gestão para situações de funcionamento normal, bem como para funcionamento aquando da ocorrência de um "incidente" [47], como no caso de surgimento de um *bloom* de cianobactérias. Uma vez que a monitorização de rotina não é suficiente para que haja um aviso prévio da iminência do surgimento destes fenómenos [36], a criação de um modelo que preveja a ocorrência de *bloom* de cianobactérias com alguns dias de antecedência permite à AgdA cumprir um dos quatro objetivos estratégicos do PSA - Gestão de risco, pois permite o alerta precoce do risco. O conhecimento prévio deste risco permitirá à AgdA executar as adaptações operacionais, descritas anteriormente, em tempo útil, gerir o armazenamento de água para consumo humano, a utilizar em caso de incapacidade de tratamento da água contaminada, e avisar as autoridades competentes (p.e. Entidades Regionais de Saúde) da proximidade da presença destes organismos em águas para fins recreativos, para que sejam tomadas medidas que evitem o contacto com as cianotoxinas [45].

2.6 Metodologias para análise de dados

2.6.1 Análise de variância (ANOVA)

A análise de variância de um fator (ANOVA) é utilizada para perceber a interação entre grupos independentes, comparando a média dos grupos, pelo que a hipótese nula é não existir diferença significativa entre as médias (Ho: $\mu_1 = \mu_2 = \mu_3$) [48].

A execução da ANOVA pressupõe o cumprimento de algumas condições como [48]:

- as variáveis dependentes têm de ter uma distribuição normal,
- tem de existir homogeneidade das variâncias e
- as observações de cada grupo são independentes entre si.

Após a verificação destes pressupostos é calculada a estatística F, que corresponde ao quociente entre a variância entre grupos (SSg) e a variância dentro dos grupos (residual) - SSe (Equação 2.1) [49]:

$$F = \frac{SSg}{SSe} \tag{2.1}$$

Para além da estatística F, é também resultado da ANOVA, como de qualquer teste de hipótese, o valor p (*p-value*) que permite decidir se há evidência suficiente ou não para rejeitar Ho, quando comparado com o nível de confiança. Ou seja, se p-value < α rejeita-se Ho, o que significa que pelo menos um grupo tem média diferente [49]. No entanto, este teste, não permite a identificação de qual(quais)

o(s) grupo(s) que apresenta(m) uma média diferente dos restantes, pois a Ho é uma hipótese *omnibus* (global), pelo que a ANOVA terá de ser precedida por testes, como por exemplo o teste de Turkey (Turkey's HSD). Com este teste, comparam-se as médias dos grupos duas a duas, usando a distribuição q [50].

2.6.2 Análise exploratória (PCA)

A análise de componentes principais (PCA) é uma técnica de análise exploratória que permite analisar a estrutura de correlação dos dados [51]. Esta técnica tem como objetivos extrair a informação mais relevante de um conjunto de dados, reduzir o tamanho do conjunto de dados mantendo apenas a informação importante, simplificar a descrição do conjunto de dados e analisar a estrutura das observações e das variáveis. Ou seja, permite transformar um conjunto de variáveis iniciais correlacionadas num outro conjunto de variáveis não correlacionadas (ortogonais) designadas por Componentes Principais (PC). As PC, resultantes da aplicação do modelo PCA, são uma combinação linear das variáveis originais, por ordem decrescente de importância [52]. A escolha do número de PC pode ser efetuada de acordo com a percentagem da variância explicada por estas. Assim, é possível explicar a variabilidade dos dados com um menor número de variáveis, eliminando a informação redundante e/ou não sistemática (e.g., ruído ou presença de *outliers*) [51].

A transformação dos dados originais pode ser esquematizada como ilustrado na Figura 2.11. Os dados originais (matriz X) são o resultado do produto da matriz dos *scores* (T), pela matriz dos coeficientes ou *loadings* (P), somado à matriz dos resíduos (E), que contém a variabilidade não captada pelos PC selecionados.

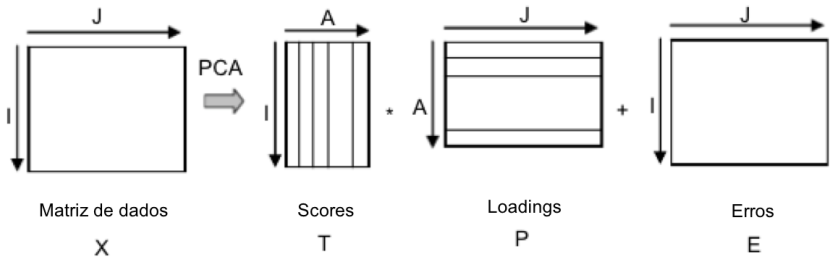

Figura 2.11: Esquema da técnica PCA (adaptado de [53])

A análise da representação gráfica conjunta dos *scores* e *loadings* (*biplot*) permite verificar quais as variáveis responsáveis pelas diferenças observadas entre as amostras.

2.6.3 Modelação linear (PLS)

A técnica dos mínimos quadrados parciais (PLS) é particularmente útil quando se pretende prever um conjunto de variáveis dependentes a partir de um grupo extenso de variáveis independentes. Esta técnica é definida pelas Eqs. 2.2 e 2.3.

$$X = T \times P' + E \tag{2.2}$$

$$Y = U \times Q' + F \tag{2.3}$$

Esta técnica pode ser também descrita como um método de encontrar um conjunto de componentes - variáveis latentes (LV) que executam a decomposição simultânea de X e Y, garantindo a explicação do máximo possível da covariância entre X e Y. De seguida dá-se o passo de regressão em que se tenta prever Y a partir da decomposição de X [54], procurando-se assim encontrar uma relação linear entre os *scores* de X e Y que permita descrever as variáveis de resposta Y (Figura 2.12).

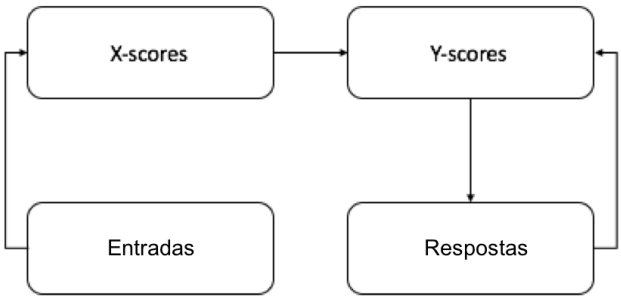

Figura 2.12: Representação esquemática da análise PLS (adaptado de [55])

2.6.4 Modelação não-linear (RNA)

As Redes Neuronais Artificiais (RNAs) são sistemas cujo funcionamento é inspirado na biologia, que tem a capacidade de adquirir conhecimento através da experiência.

O sistema nervoso pode ser visto como um sistema com três estágios, como ilustrado na Figura 2.13. Sendo o sistema central o cérebro (rede neuronal), que recebe informação continuamente e que após a compreender toma decisões acertadas. A transmissão da informação pode ocorrer nos dois sentidos, para a frente (setas pretas) ou de retroalimentação do sistema (setas azuis). Os estímulos do exterior ou do corpo humano são processados pelos recetores em impulsos elétricos por forma a serem percepcionados pela rede neuronal. Por sua vez, os efetores convertem os impulsos elétricos gerados pela rede neuronal em repostas [56].

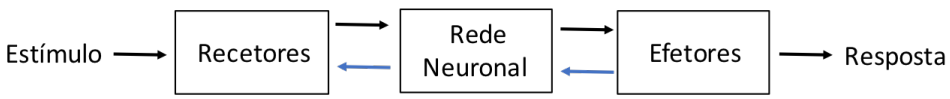

Figura 2.13: Representação esquemática do sistema nervoso (adaptado de [56])

A rede neuronal que constituí o cérebro é uma estrutura bastante eficiente, composta por cerca de 10 mil milhões de neurónios e 60 000 mil milhões de sinapses ou conexões. Estas últimas são

estruturas elementares que realizam a mediação da interação entre neurónios [56]. Na Figura 2.14 é ilustrado o esquema de um neurónio biológico.

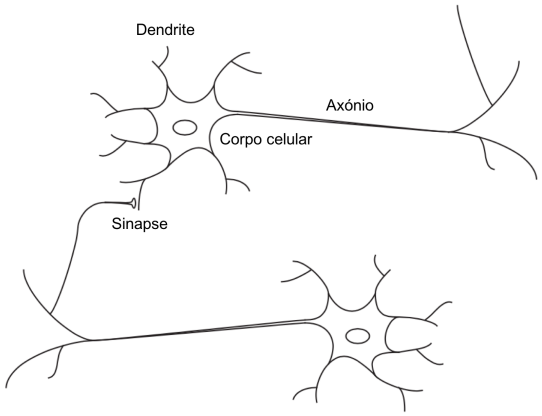

Figura 2.14: Representação do neurónio biológico (adaptado de [57])

Uma rede neuronal artificial pretende simular de forma simplificada o sistema nervoso do cérebro humano, consistindo em sistemas paralelos de processamento, constituídos por unidades de processamento simples (nodos) que calculam funções matemáticas específicas (normalmente não-lineares). Os nodos são dispostos numa ou mais camadas, sendo interligados por um número elevado de conexões. Na maioria dos modelos estas conexões estão associadas a pesos, que armazenam o conhecimento representado no modelo e servem para ponderar a entrada recebida por cada nodo da rede (Figura 2.15).

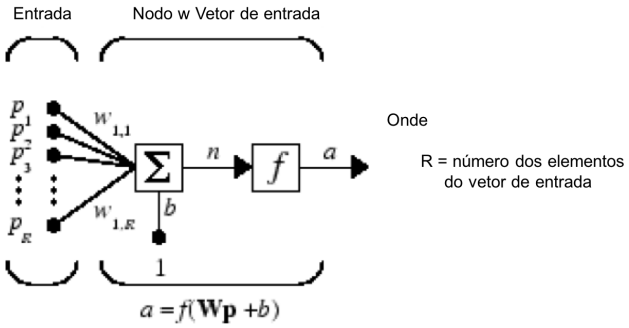

Figura 2.15: Representação esquemática do nodo numa RNA (adaptado de [58])

Num único neurónio, como o representado na Figura 2.15, ocorrem três operações funcionais [58]:

1. função de peso: a entrada (p) é multiplicada pelo peso de ligação (W) resultando no produto Wp
2. função de entrada da rede: ao somatório dos produtos wp (Wp) é adicionado um incremento (*b*), que permite obter uma saída não nula quando todas as entradas (p) forem iguais a zero, convertendo-se na entrada da rede (n)
3. função de ativação: à entrada da rede é aplicada uma função de ativação (f), produzindo a saída (a).

Uma vez que os pesos da ligação são ajustáveis, a ideia principal de uma RNA é o ajuste destes parâmetros até que o comportamento desejado seja atingido, menor diferença entre a saída e o valor desejado. O método mais usado para a estimativa dos erros é a retropropagação (Figura 2.16). Existem três métodos de aprendizagem de RNA: aprendizagem por reforço, aprendizagem supervisionada e aprendizagem não supervisionada. De seguida apresenta-se uma breve descrição destas formas de aprendizagem [59].

1. Aprendizagem por reforço

   Neste método não são fornecidas à rede as saídas corretas para as entradas. A alteração dos pesos das conexões é realizada com a atribuição de prémios ou penalizações de acordo com o facto do *output* ser de acordo com o esperado ou não.

2. Aprendizagem supervisionada

   Nesta forma de aprendizagem os alvos são conhecidos à partida pela rede. Só há alterações dos pesos se a diferença entre os alvos e as saídas da rede forem significativas.

3. Aprendizagem não supervisionada

   Este método é utilizado quando não há conhecimento, *a priori*, dos possíveis classificadores. Desta forma, a rede terá de detetar a existência de regularidades no espaço de entrada, ou seja extrai a estrutura inerente da camada de entrada, sem ter um ajuste direto de um supervisor.

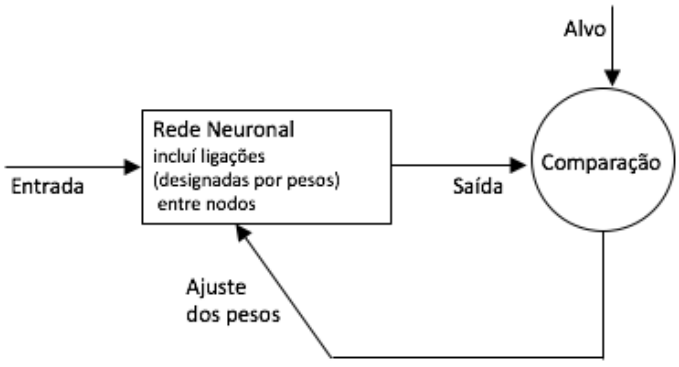

Figura 2.16: Representação esquemática do treino de uma RNA (adaptado de [60])

As RNA pode ter diferentes arquiteturas. Aqui mencionam-se apenas duas.

1. Redes neuronais artificiais *feedforward* (estáticas)

   Esta tipologia de RNA é constituída por uma ou mais camadas escondidas/intermédias, constituída por neurónios (nodos ocultos) cuja função é fazer a ligação entre a camada de entrada e saída. A utilização das camadas intermédias permite à rede modelar dados de complexidade crescente (Figura 2.17) [56].

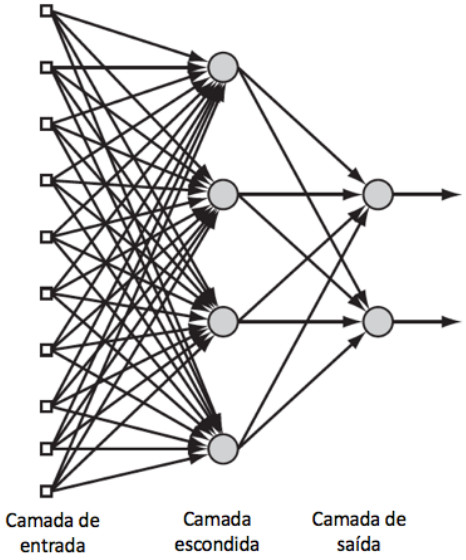

Figura 2.17: Representação esquemática da RNA *feedforward* de múltipla camada (adaptado de [56])

A rede ilustrada na Figura 2.17 designa-se por completamente ligada, uma vez que todos os nodos de cada camada estão ligados a todos os nodos da camada adjacente. Se, no entanto, esta característica não se verificar a rede passa a chamar-se por parcialmente ligada.

2. Redes neuronais artificiais dinâmicas

   As RNA dinâmicas têm memória, ou seja, a resposta dada num determinado momento depende não só da entrada corrente da rede, mas também das entradas anteriores e das saídas ou estados da rede. Por este motivo esta tipologia de rede é mais poderosa que as estáticas, nomeadamente para modelar sistemas dinâmicos. Há dois tipos de categorias de RNA dinâmicas: as que têm apenas ligações *feedforward* e as que têm ligações recorrentes (*feedback*).

   (a) Redes neuronais artificiais recorrentes

   As RNA recorrentes distinguem-se por terem um ou mais laços de retorno (Figura 2.18). As ligações de um neurónio podem ser provenientes de neurónios de camadas anteriores, da própria camada ou até de camadas posteriores. Esta estrutura mais complexa tem um impacto profundo na capacidade de aprendizagem da rede, bem como no seu desempenho. Os laços de retorno envolvem o uso de ramos compostos por atrasos temporais, resultando num comportamento dinâmico não linear [56; 61].

As redes do tipo NARX têm, segundo H. Wang *et al.* [28], uma elevada capacidade de descrição de processos dinâmicos não lineares e complexos, ajustando-se bem para previsão de séries temporais. Este modelo baseia-se no modelo linear ARX, usado na modelação de séries temporais, no qual o valor de saída é obtido com base nos valores anteriores do sinal de saída e dos valores do sinal de entrada antecedentes. De acordo com a forma como a saída é utilizada, as redes são classificadas como de

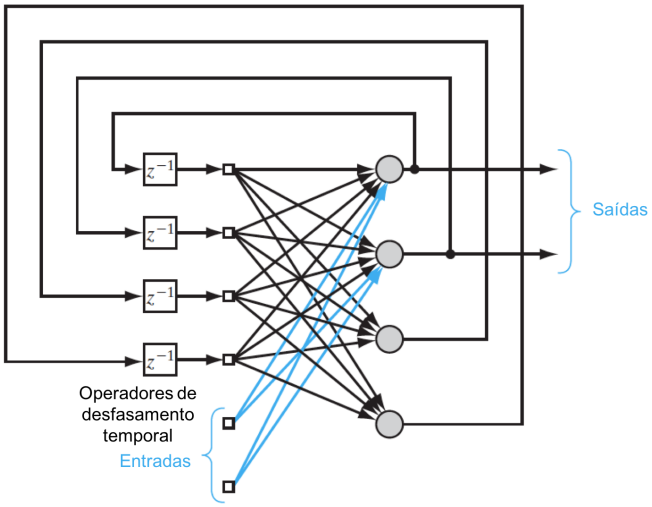

Figura 2.18: Representação esquemática de uma RNA recorrente (adaptado de [56])

arquitetura paralela, em que o valor de resposta realimenta a rede, ou de arquitetura de série-paralela, na qual o alvo da rede é também uma entrada (Figura 2.19) [58].

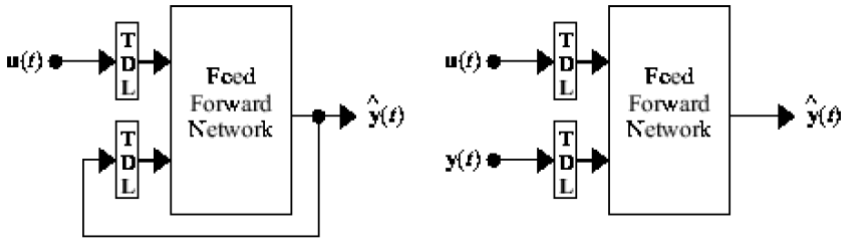

Figura 2.19: Representação esquemática da rede NARX com arquitetura paralela (esquerda) e com arquitetura série-paralela (direita) (adaptado de [58])

2.6.5 Avaliação do desempenho dos modelos

Neste estudo a qualidade dos modelos PLS foi avaliada por três parâmetros, a raiz do erro quadrático médio (RMSE), o R^2 e a razão amplitude-erro (RER).

O valor de RMSE, como o nome indica, resulta da raiz quadrada da média do quadrado da diferença entre os valores experimentais e os valores estimados pelo modelo (Equação 2.4)

$$RMSE = \sqrt{1/N \sum_{i=1}^{N} (Y_{original} - Y_{previsto})^2} \tag{2.4}$$

O R^2 corresponde ao coeficiente de correlação de Pearson, que permite avaliar o poder explicativo do modelo. E, finalmente, o parâmetro RER, relaciona a amplitude da variável de resposta com o respetivo RMSE (Equação 2.5).

$$RER = \frac{(Y^{max} - Y^{min})}{RMSE} \tag{2.5}$$

Para os modelos gerados por RNA optou-se por utilizar o erro quadrático médio (MSE) do conjunto de teste como método de verificação da sua adequabilidade à realidade. Assim, o desempenho do modelo é avaliado pela diferença entre os valores experimentais e os valores estimados para o conjunto de dados de teste (Equação 2.6).

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_{original} - Y_{previsto})^2 \tag{2.6}$$

em que $Y_{original}$ corresponde ao valor experimental, o $Y_{previsto}$ é o valor estimado pelo modelo e N o número de observações.

Capítulo 3

Materiais e Métodos

Neste capítulo são apresentados os dados utilizados no desenvolvimento do estudo, bem como a preparação dos mesmos para as análises efetuadas e para modelação.

3.1 Apresentação dos dados disponíveis

Neste estudo pretendeu-se criar um modelo de previsão da ocorrência de *blooms* de cianobactérias utilizando parâmetros de monitorização do estado de água da Albufeira do Roxo, bem como dados meteorológicos da região onde se insere a Albufeira. Por este motivo, a primeira fase do trabalho passou pela compilação dos dados de qualidade da água provenientes de três entidades responsáveis pela monitorização da qualidade da água da Albufeira do Roxo em três períodos temporais (Tabela 3.1). Foram recolhidos também dados meteorológicos, a partir da base de dados do Sistema Nacional de Informação de Recursos Hídricos - SNIRH (temperatura do ar, velocidade do vento e direção do vento) e outros adquiridos ao Instituto Português do Mar e da Atmosfera - IPMA (precipitação e radiação).

Tabela 3.1: Espaço temporal e entidade fornecedora dos dados

Período temporal	Entidade
2007 - 2010	EMAS
2009 - 2013	ARH
2012 - 2015	AgdA

No final desta tarefa obteve-se uma base de dados com 150 parâmetros de qualidade da água monitorizados em 184 dias, entre 15 de janeiro de 2007 e 4 de novembro de 2015 e cinco variáveis meteorológicas com maior número de dados (de 1 de janeiro de 2007 a 31 de dezembro de 2015). A recolha de amostras para análise da qualidade da água tem, na sua maioria, uma periodicidade mensal no período de 2007 a 2010 e bimestral entre 2010 e 2015, salvo algumas exceções. O aumento de frequência de amostragem estará relacionado com a probabilidade de ocorrência de um *bloom*. De acordo com a literatura é recomendada uma frequência de amostragem de pelo menos uma por semana ou mesmo bissemanal quando se verifica uma densidade de cianobactérias superior a 2000 células/mL [7]. A monitorização dos parâmetros é irregular para os 184 dias. Apenas 12 parâmetros foram monitorizados

em metade dos dias, este facto deverá estar relacionado com as diretrizes do Decreto-Lei n^o 236/98 de 1 de Agosto do Ministério do Ambiente, 1998, que regula a verificação da qualidade das águas doces superficiais destinadas à produção de água para consumo humano, que não obriga a análise de todos os parâmetros.

Um bom modelo deve ser estocástico, ou seja, não deverá ser construído com todas as variáveis disponíveis mas com as indispensáveis para a descrição do processo [62]. Assim, selecionaram-se os parâmetros listados na Tabela 3.2, que de acordo com a literatura consultada (Capítulo 2.3), seriam as variáveis chave para a previsão de *blooms* de cianobactérias.

Tabela 3.2: Parâmetros hipoteticamente relevantes e respetiva notação

Parâmetro	Notação
Radiação	Rad
Velocidade do vento	VV
Direção do vento	DV
Temperatura do ar	Tar
Precipitação	Prec
Cota	Cota
Temperatura da água	Tag
pH	pH
Condutividade	Cond
Cor	Cor
Turvação	Turv
Oxigénio dissolvido	OD
Carência Bioquímica de Oxigénio	CBO_5
Nitrato	NO_3
Nitrito	NO_2
Azoto amoniacal	NH_4
Azoto total	N
Fósforo total	P
Razão azoto total fósforo total	N:P
Manganês	Mn
Ferro	Fe
Fosfato	P_2O_5

Para além dos dados supramencionados, avaliou-se, também, a pertinência do uso de outros parâmetros monitorizados nos últimos anos pela entidade gestora para a criação do modelo. Estes encontram-se listados na Tabela 3.3.

Tabela 3.3: Parâmetros monitorizados pela AgdA para além dos indicados em estudos e respetiva notação

Parâmetro	Notação
Dureza	Dur
Alcalinidade	Alc
Sólidos Suspensos Totais	SST
Carência Química de Oxigénio	CQO
Carbono Orgânico Total	COT
Cálcio	Ca
Magnésio	Mg
Sílica	SiO_2
Fluoretos	F-
Clorofila-a	Cl-a

Na Tabela 3.4 apresenta-se uma breve caracterização dos parâmetros selecionados para análise, bem como o número total de dados disponível de cada uma das variáveis. Ao analisar a Tabela 3.4 deve ter-se em consideração que a maior quantidade de dados dos parâmetros meteorológicos deve-se à diferente proveniência dos mesmos, o que permitiu obter, para algumas desta variáveis, valores para quase todos os dias de 2007 a 2015 (3286 dias).

Tabela 3.4: Algumas estatísticas dos parâmetros considerados para análise

Variáveis	N^o pontos	Mínimo	Máximo	Média	Mediana	Desvio padrão
Rad (kJ/m^2)	3216	1090,6	34 636,3	19 005,5	18 641,5	8869,2
VV (m/s)	2808	0,0	8,6	2,0	1,8	1,2
DirV (^{o})	2812	0	313	142	143	47
Tar (^{o}C)	2456	3,7	30,3	16,5	16,6	5,3
Prec (mm/dia)	3274	0	65,1	1	0	5
Cota (m)	165	124,80	135,98	131,56	132,31	2,92
Tag (^{o}C)	162	9,9	26,2	18,8	19,6	4,6
pH	180	7,10	9,29	8,06	8,07	0,26
Cond [$\mu S/cm(20^o$C)]	178	433	13 330	918	895	194
Cor (mg/L Pt-Co)	92	3,6	47	11	8	7
Turv (NTU)	118	2,2	30,30	7,0	6,0	4,4
OD (mg/L)	89	1,3	14	7	8	2
CBO_5 (mg/L O_2)	79	0	7	3	3	1
NO_3 (mg/L)	84	0	4,98	2	2	1
NO_2 (mg/L)	92	<LQ	0,51	0,04	0,03	0,06
NH_4 (mg/L)	127	<LQ	0,47	0,15	0,11	0,12
N (mg/L)	36	0,43	2,05	0,99	1,00	0,35
P (mg/L)	83	0,019	0,252	0,054	0,040	0,040
NP	24	3,85	34,17	16,78	14,15	9,11
Mn (mg/L)	72	<LQ	0,93	0,13	0,08	0,17
Fe (mg/L)	59	<LQ	0,5	0,1	0,1	0,1
P_2O_5 (mg/L)	51	0	0,227	0,03	0,02	0,04
Dur (mg/L C_2CO_3)	124	172,3	489,03	287,6	260,0	74,3
Alc (mg/L HCO_3)	63	69,36	245	152	151	31
SST (mg/L)	91	2,7	59	9	6	9
CQO (mg/L O_2)	90	12	52	22	21	7
COT (mg/L O_2)	27	6,1	22	10	9	4
Ca (mg/L)	51	24	91,9	63	57	15
Mg (mg/L)	51	15,50	63	40	38	13
SiO_2 (mg/L)	24	0,4	6,4	2,5	2,4	1,6
F- (mg/L)	16	0	0,33	0,22	0,24	0,09
Cl-a ($\mu g/L$)	85	1	66,4	10	7	9
Cianobactérias (n^{o}células/mL)	89	0	463636	23434	6493	53443

Uma vez que é objetivo deste trabalho a previsão de *blooms* de cianobactérias, analisou-se também a variação temporal da densidade destes organismos. Na Figura 3.1 é possível observar a densidade de cianobactérias nos diferentes meses em que este parâmetro foi monitorizado.

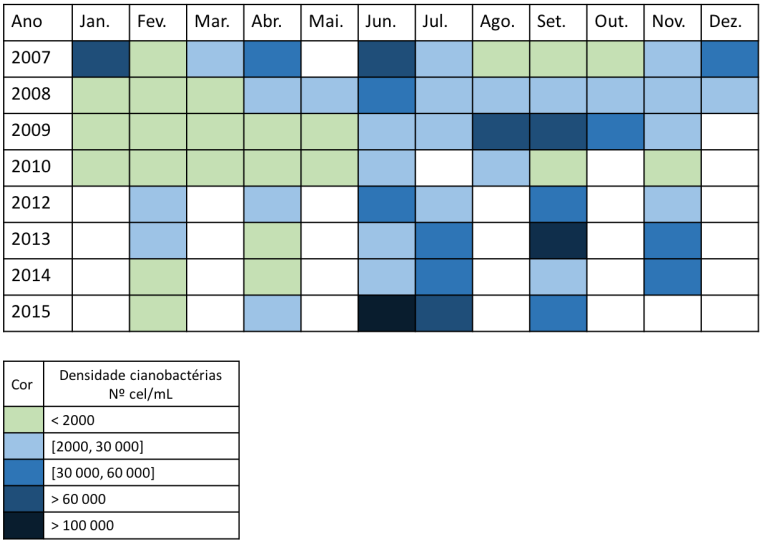

Figura 3.1: Ilustração da densidade de cianobactérias nos diferentes meses

Para perceber a relação dos parâmetros selecionados com a densidade de cianobactérias realizou-se a ANOVA considerando como variável dependente a densidade de cianobactérias. Para este efeito, formaram-se três grupos, tendo por base as recomendações da WHO, para a definição dos níveis de alerta. Estes deverão ser seguidos pelas entidades gestoras de serviços de água para sistematizar a tomada de decisão ao nível da monitorização e gestão, por forma a adequar o tratamento na ETA à presença de cianobactérias, nas origens de água para consumo humano, bem como os níveis recomendados para as águas balneares [6]:

- Grupo 1: < 2000 células/mL (nível de vigilância para águas para consumo humano),
- Grupo 2: [2000, 20 000) células/mL (nível de alerta 1 para águas para consumo humano ao nível de alerta 1 para águas para águas balneares) e,
- Grupo 3: $\geq$ 20 000 células/mL.

3.2 Modelação

3.2.1 Organização dos dados

Ao analisar a densidade de cianobactérias ao longo do tempo (Figura 3.2) verificou-se que no dia 25 de setembro de 2013 foi registado um valor bastante elevado (assinalado a laranja) para este parâmetro. Após a AgdA ter confirmado que era bastante improvável a presença destes organismos nesta ordem de grandeza, a amostra foi considerada como valor atípico (*outlier*).

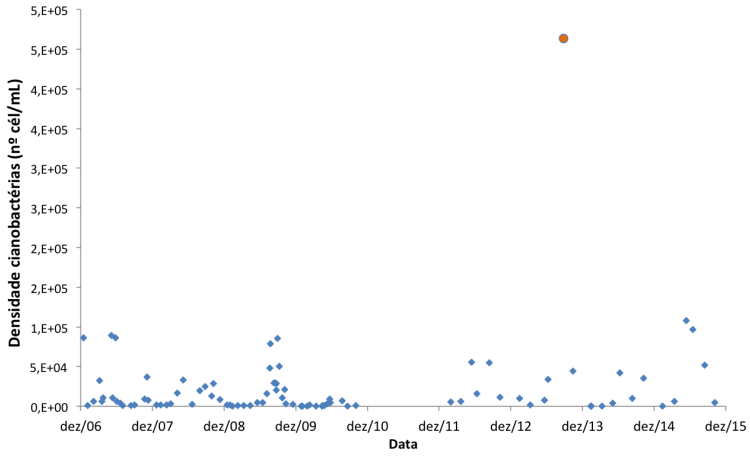

Figura 3.2: Densidade de cianobactérias medidas ao longo do intervalo de amostragem utilizado

Uma vez que a maioria dos dados recolhidos correspondiam a dados de monitorização de qualidade da água e que, por esse motivo, não foram recolhidos especificamente para a realização deste trabalho, a sistematização dos dados, apresentados anteriormente, resultou numa matriz que não pôde ser utilizada diretamente na modelação, dado que os parâmetros não foram todos monitorizados nos diferentes dias. Assim, foi necessário encontrar uma estratégia que permitisse preencher as lacunas nos dados.

O processo de conversão dos dados para modelação começou pela seleção das variáveis que constituíam a matriz original com um número de dados superior a uma percentagem, designada por percX (Figura 3.3).

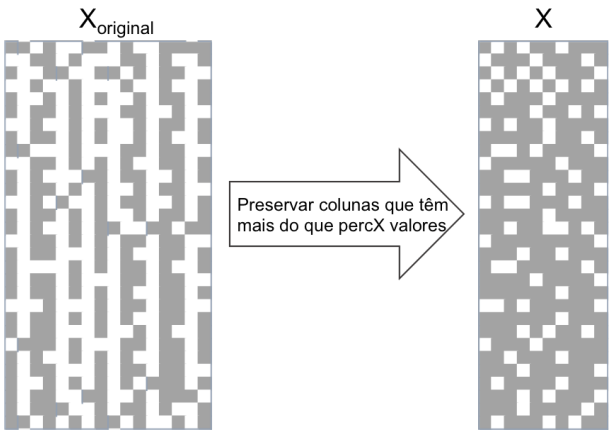

Figura 3.3: Ilustração da construção da matriz X a partir da matriz de dados original (Xoriginal)

Depois de encontrada a matriz X, verificou-se em que instantes de amostragem foi monitorizada a densidade de cianobactérias (valores que constituem a matriz Y). Os instantes que não continham informação foram desprezados. O valor das variáveis monitorizadas nos restantes instantes de amostragem foram então convertido através da obtenção das medianas correspondentes a um determinado período de tempo (delayX) anterior ao dia em que a densidade de cianobactérias é conhecida. Este

processo teve também em consideração o número de dias de previsão (ModelDelay). Este último foi definido, em colaboração com a entidade gestora, como um valor constante de 15 dias, período que se considerou suficiente para poder tomar as medidas de adaptação necessárias para evitar/mitigar as consequências que surgem com estes eventos (Figura 3.4).

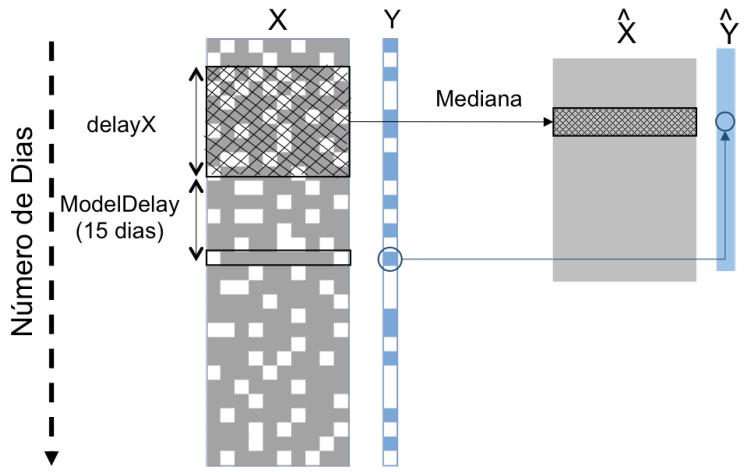

Figura 3.4: Ilustração da construção da matriz usada para modelação

A escolha do valor de percX para a criação das matrizes utilizadas neste estudo foi baseada no número de variáveis/amostras que devem constar nas matrizes, ou seja, se se pretende uma matriz com o maior número de variáveis possível e com um número de amostras também elevado (a utilizar na PCA) ou se o objetivo é construir uma matriz de acordo com o número de amostras (a utilizar na modelação com RNA). Para ajudar nesta seleção, traçou-se o gráfico que ilustra a variação do número de variáveis/amostras de acordo com o valor de percX para valores constantes de ModelDelay de 15 dias e de delayX=60 dias (Figura 3.5)

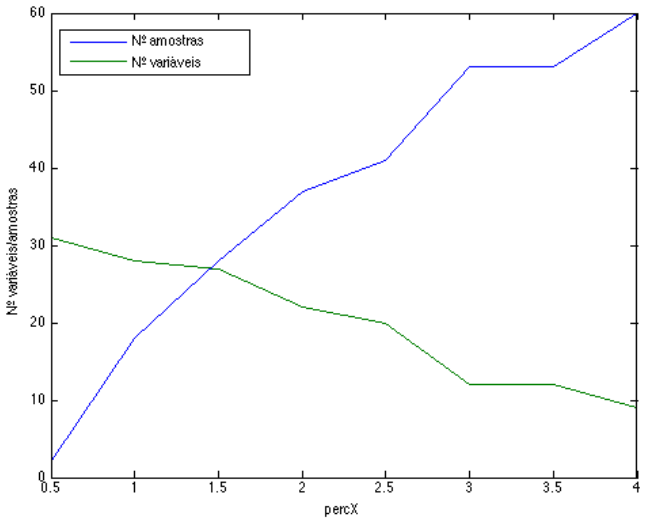

Figura 3.5: Variação do número de variáveis/amostras de acordo com o valor de percX

3.2.2 Análise dos componentes principais

Para a realização da análise de componentes principais era conveniente ter uma matriz de dados com o maior número de parâmetros possível e com um número de amostras também elevado. Por este motivo, construíu-se uma matriz, de acordo com a metodologia descrita no subcapítulo 3.2 considerando um valor de percX de 2% e delayX=60 dias. Este processo resulta numa matriz constituída por 37 amostras e 22 variáveis (Tar, VV, DirV, OD, Tag, pH, Cond, Cor, Turv, Cota, CBO_5, NO_3, NO_2, NH_4, P, Mn, CQO, Cl-a, Dur, SST, Prec e Rad).

3.2.3 Redes neuronais artificiais

Para a criação de RNAs optou-se por gerar uma matriz com maior número de amostras possível. Para este propósito utilizou-se a metodologia descrita no subcapítulo 3.2, optando-se por uma percentagem mínima de dados para as variáveis de entrada (percX) de 3% e um delayX=60. Este processo resultou numa matriz constituída por 53 amostras com 12 variáveis (Tar, VV, DirV, Tag, pH, Cond, Turv, Cota, NH_4, Dur, Prec e Rad).

Nas RNA do tipo NARX o desfasamento dos dados de entrada relativamente ao alvo poderia ser definido durante o treino da rede. No entanto, como os dados utilizados não se apresentavam igualmente espaçados no tempo optou-se por incluir este desfasamento aquando da criação da matriz supracitada.

Treino das RNA

Neste estudo foram utilizadas RNA estáticas (*feedforward*) multicamada e NARX. Em ambas as tipologias optou-se por realizar o treino através do algoritmo de gradientes conjugados, uma vez que segundo Wilson, H. e Recknagel, F. [31] a rede terá, teoricamente, uma performance de treino superior para os dados deste tipo. Para além disso não é necessário definir alguns parâmetros de treino [31].

Uma RNA ideal é aquela que mimetiza a realidade sem que ocorra sobreajustamento de dados. Para garantir estas duas premissas é necessário dividir os dados utilizados para criar a rede em três grupos (enumerados de seguida). Neste estudo usou-se uma proporção de 50/20/30.

- Conjunto de treino (50%), utilizado, como o nome indica, para treino da rede (definição dos pesos e bias)
- Conjunto de validação (20%): cujo erro é monitorizado durante o treino por forma a identificar quando é que a rede entra em sobreajuste. O erro de validação por norma diminui durante a primeira fase de treino e quando a rede começa a sobreajustar-se aos dados este erro começa a crescer. É assim identificada a altura em que se deve parar o treino.
- Conjunto de teste (30%): estes dados não são usados durante o treino, mas servem para avaliar o desempenho da rede [58].

A divisão dos dados supracitada pode ser realizada de diferentes formas de acordo com as funções indicadas na Tabela 3.5.

Tabela 3.5: Funções de divisão dos dados

Função	Algoritmo
dividerand	divisão dos dados é feita aleatoriamente
divideblock	divisão dos dados é feita em blocos de dados contíguos
divideint	divisão dos dados é feita através de uma seleção de dados intercalados
divideind	divisão dos dados é feita através de índices

No decorrer deste trabalho optou-se pela divisão dos dados por índices que correspondiam às amostras, como ilustrado na Figura 3.6. A matriz de dados utilizada para a construção das RNA deste estudo era constituída por 53 amostras. Considerando as proporções de 50/20/30, o conjunto de calibração era constituído pelas amostras correspondentes ao intervalo de índice de 1 a 26, por sua vez, o conjunto de validação correspondia aos índices de 44 a 53 e, finalmente, o conjunto de teste era composto pelas amostras intermédias (27 a 43). Esta disposição dos grupos foi escolhida por permitir avaliar o desempenho da rede com um conjunto de dados (teste) que contem valores de densidade de cianobactérias de diferentes ordens de grandeza (de 11 a 54 719 células/mL).

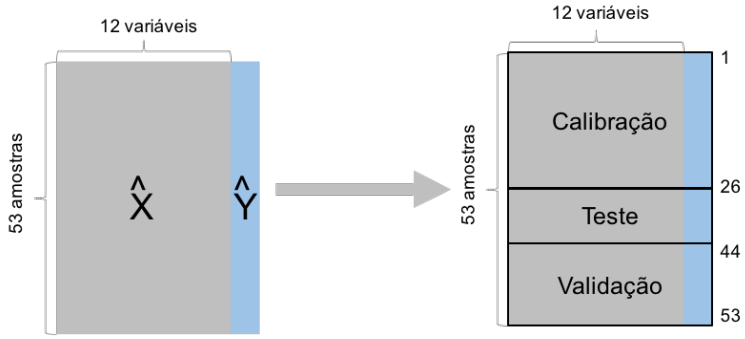

Figura 3.6: Ilustração da divisão dos dados da matriz a utilizar nas RNA

3.3 Definição do limiar de alerta

Uma vez que o modelo gerado neste trabalho será utilizado, pela entidade gestora de serviços de água, para verificar a necessidade de ativar os protocolos de atuação para uma situação da ocorrência de *blooms* de cianobactérias, para além do MSE, os modelos são também avaliados pelo número de falsos positivos/negativos nas previsões dos modelos para os conjuntos de teste, de acordo com a Tabela 3.6.

Tabela 3.6: Matriz de confusão

		Valor experimental	
		Acima do limiar	**Abaixo do limiar**
Valor previsto	**Acima do limiar**	Verdadeiro positivo	Falso positivo
	Abaixo do limiar	Falso negativo	Verdadeiro negativo

O limiar de alerta foi definido de acordo com as recomendações da Organização Mundial de Saúde (WHO). Considerou-se o valor de 20 000 células/mL, por ser o valor limite para a ocorrência de efeitos adversos para a saúde por contacto (águas agrícolas/balneares), e cinco vezes inferior ao limite definido

no nível de alerta II para águas superficiais para abastecimento (número de células de cianobactérias igual ou superior a 100 000 células/mL, com presença de toxinas confirmada por análises químicas ou bioensaios) que descreve uma floração tóxica estabelecida [6].

3.4 *Software* utilizado

Na primeira fase do trabalho, sistematização dos dados, utilizou-se como ferramenta o Microsoft Office Excel Mac 2011. Após definição dos parâmetros a estudar utilizou-se o programa Matlab, nas versões R2009b e R2014a (The Mathworks, Inc), para executar todas as análises e modelos. Para análises de variância foram utilizadas as funções da *toolbox* de estatística versão 9.0. Com as funções incluídas na *toolbox* de PLS (Eigenvector Research, Inc) criaram-se os modelos PCA e PLS. Por sua vez, as RNAs foram geradas pelas funções da *toolbox* de redes neuronais versão 8.2.

Capítulo 4

Resultados e discussão

Neste capítulo são apresentados os resultados obtidos pelas diferentes metodologias utilizadas, nomeadamente: análise de variância, PCA, modelos lineares (PLS) e não lineares (RNAs).

4.1 Análise exploratória univariada

Numa primeira fase do estudo realizou-se uma análise univariada aos dados, verificando como cada um dos parâmetros considerados para análise variava em função da densidade das cianobactérias. O resultado desta análise ilustra-se nas Figura 4.1 e Figura 4.2.

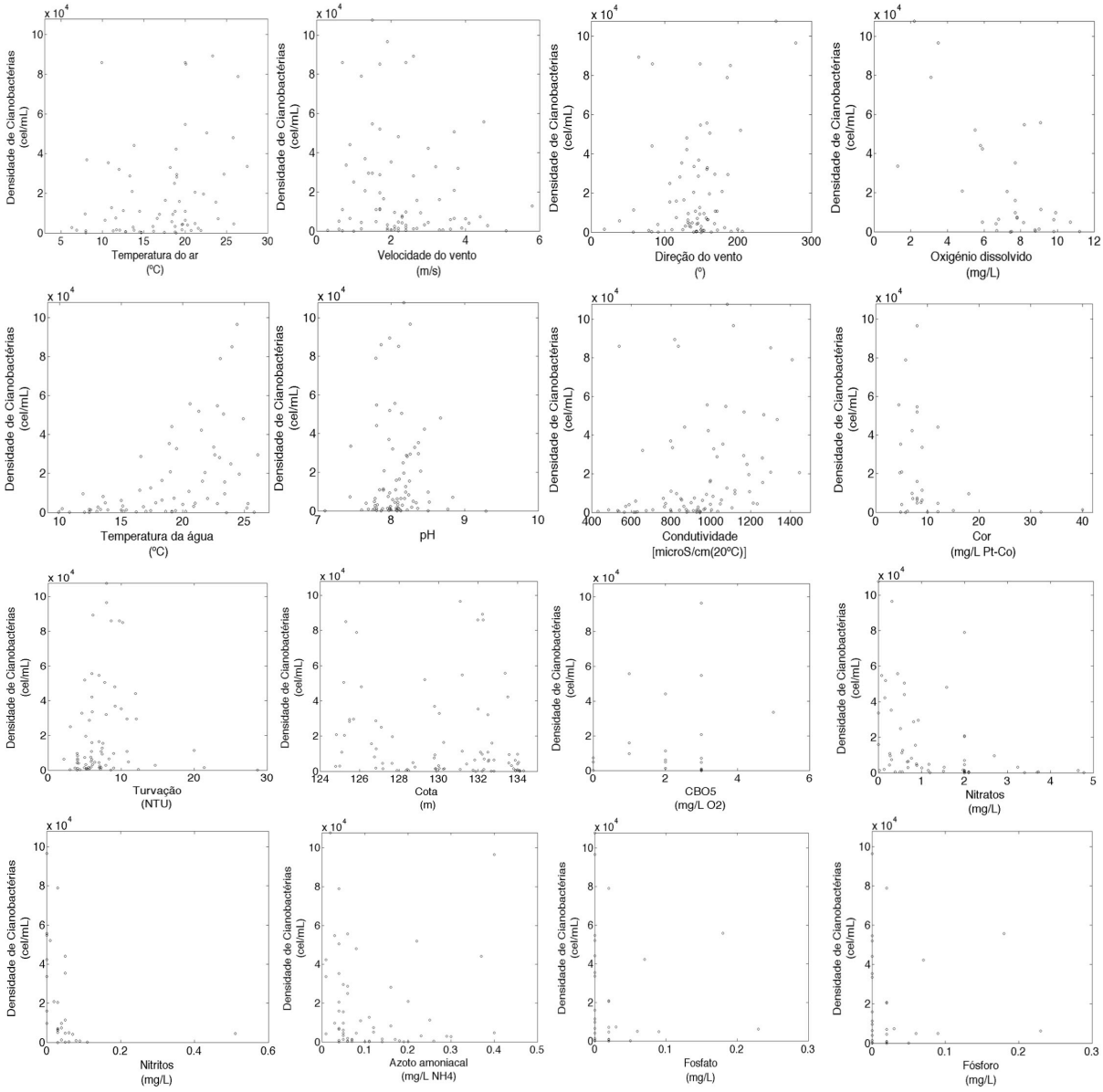

Figura 4.1: Variação de cada um dos parâmetros em relação à densidade de cianobactérias

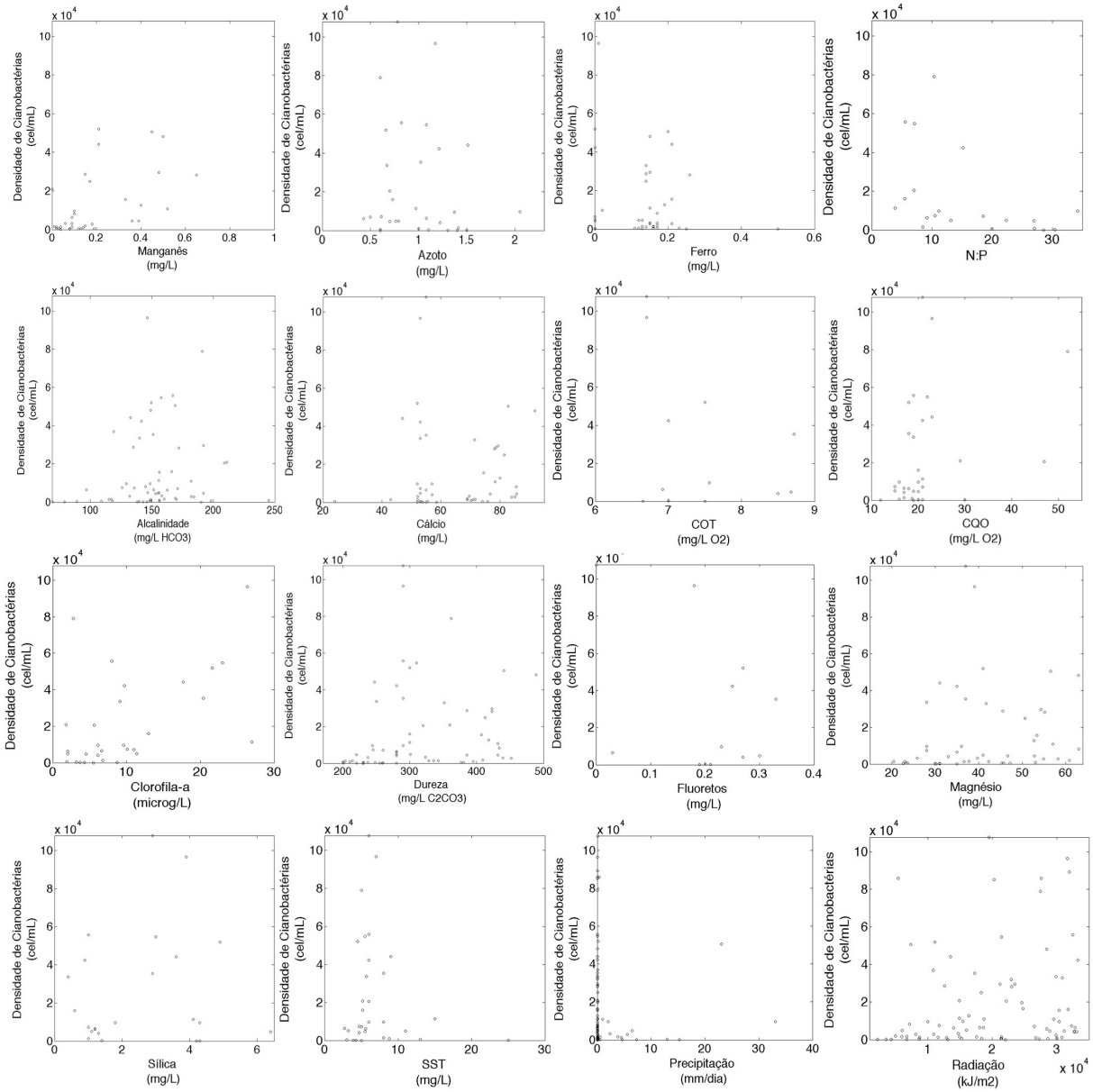

Figura 4.2: Variação de cada um dos parâmetros em relação à densidade de cianobactérias (cont.)

Ainda numa perspetiva de análise univariada dos dados realizou-se a análise de variância (ANOVA - um fator), considerando-se a densidade de cianobactérias como a variável independente e os parâmetros selecionados como as variáveis dependentes, de acordo com os grupos apresentados no subcapítulo 3.1. Nesta análise não se considerou a existência de desfasamento temporal entre o valor de cianobactérias e as restantes variáveis. Na Tabela 4.1 são apresentados os resultados desta análise.

Tabela 4.1: Resultados dos testes de variância para os parâmetros selecionados

Parâmetros	Rad	WV	DirV	Tar	Prec	Cota	Tag	pH	Cond	Cor	Turv	OD	CBO_5	NO_3	NO_2	NH4	N	P	N:P	Mn	Fe	P_2O_5	Dur	Alc	SST	CQO	COT	Ca	Mg	SiO_2	F^-	Cl^-a
Lilliefors	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	0	1	1	1	1	0	1	1	1	1	0	1	0	1
p-value Barlett	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0,1686	-	-	-	-	0,01945	-	-	-	-	0,92534	-	0	-
p-value ANOVA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0,0640	-	-	-	-	-	-	-	-	-	0,0741	-	-	-
p-value Kruskal-Wallis	0,3233	0,0564	0,6072	0,0137	0,1541	0,0437	0,0000	0,5661	0,0001	0,1905	0,0438	0,0027	0,0801	0,0026	0,0012	0,1545	0,4682	0,4938	-	0,0022	0,9654	0,8128	0,0065	0,2994	0,5337	0,0130	0,3456	0,3593	-	0,8207	0,6301	0,0359
Turkey	-	-	-	3≠1 3=2 2=1	-	1=3 2=3	Todos dif.	-	3≠1 3=2 2≠1	-	3=1 2=1	1≠3 1=2 2≠3	-	1≠3 1=2 2=3	1≠3 1=2 2=3	-	-	-	-	3≠1 3=2 3≠1	-	3≠1 3=2 2≠1	-	-	-	3≠1 3=2 2=1	-	-	-	-	-	3≠1 3=2 2=1
Dunn	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Nº dados	78	84	56	73	89	86	70	87	85	33	78	33	22	60	34	62	32	25	20	39	44	32	64	63	33	32	12	12	49	24	12	32
Nº dados G1	20	25	17	21	28	27	23	27	27	8	23	8	6	21	8	22	8	6	5	17	17	8	22	22	8	7	3	3	18	4	3	7
Nº dados G2	31	32	22	29	34	34	25	31	31	12	31	12	9	21	13	21	12	10	10	11	14	12	22	22	12	12	4	4	16	10	4	12
Nº dados G3	27	27	17	23	27	25	22	27	27	13	24	13	7	18	13	19	12	9	5	11	13	12	20	19	13	13	5	5	15	10	5	13

De acordo com o exposto na Tabela 4.1 os parâmetros assinalados a verde, e listados de seguida, estão de certa forma relacionados com a densidade de cianobactérias:

Tar Cond NO_2 CQO Cota Turv Mn Cl-a Tag NO_3 Dur

Uma vez que os modelos gerados neste estudo têm como objetivo prever a ocorrência de *blooms* com 15 dias de antecedência, realizou-se também a ANOVA com desfasamento entre as variáveis independentes e a densidade de cianobactérias, pois os valores anteriores de um ou mais parâmetros poderia ter influenciado o aparecimento de cianobactérias em dias posteriores. Não foi possível utilizar um desfasamento igual à previsão devido à periodicidade das amostragens (descrita na secção 3.1). Por este motivo optou-se por um desfasamento de 30 dias. Como seria de esperar, o número de parâmetros relevantes, listados de seguida, diminuíu (14 para 5). Verificou-se ainda a identificação da possível influência do azoto amoniacal e precipitação, facto que não ocorreu na análise anterior.

Tag Tar Mn NH_4 Prec

De acordo com os resultados apresentados anteriormente, as variáveis Tar, Tag e o Mn estão de alguma forma relacionados com a densidade de cianobactérias com ou sem desfasamento temporal entre os mesmos.

4.2 Análise exploratória multivariada

Para verificar como se relacionam as variáveis estudadas realizou-se uma análise exploratória de dados utilizando a técnica PCA. O mapa de *scores* do modelo PCA encontra-se representado na Figura 4.3. Os pontos estão coloridos de acordo com a densidade de cianobactérias. A azul estão as amostras que correspondem a uma densidade superior (> 20 000 células/mL), a verde a densidade intermédia (2000 a 20 000 células/mL) e, finalmente, a vermelho menor densidade (<2000 células/mL).

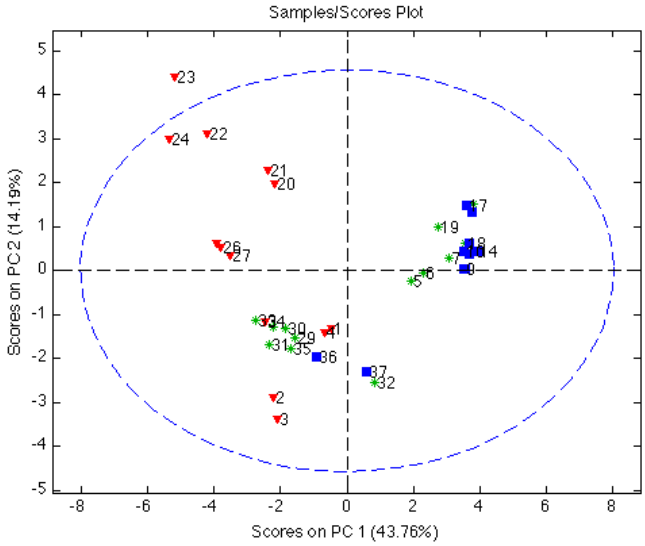

Figura 4.3: Mapa de *scores* relativo ao modelo PCA

De acordo com o ilustrado na Figura 4.3, os dois primeiros componentes principais conseguem

capturar 57,95% de variância. Ao analisar esta figura é possível verificar que as amostras que correspondem a menores valores de densidade de cianobactérias estão mais dispersas que o conjunto relativo à maior densidade, com exceção das amostras 36 e 37. Comparando as contribuições dos *loadings* para estas duas amostras e para a amostra 13, que se encontra agrupada com os restantes pontos de elevada densidade de cianobactérias (Figura 4.4), verifica-se que a amostra 13 tem uma contribuição inferior de clorofila-a e superior de oxigénio dissolvido e ambas de sinal contrário às restantes amostras.

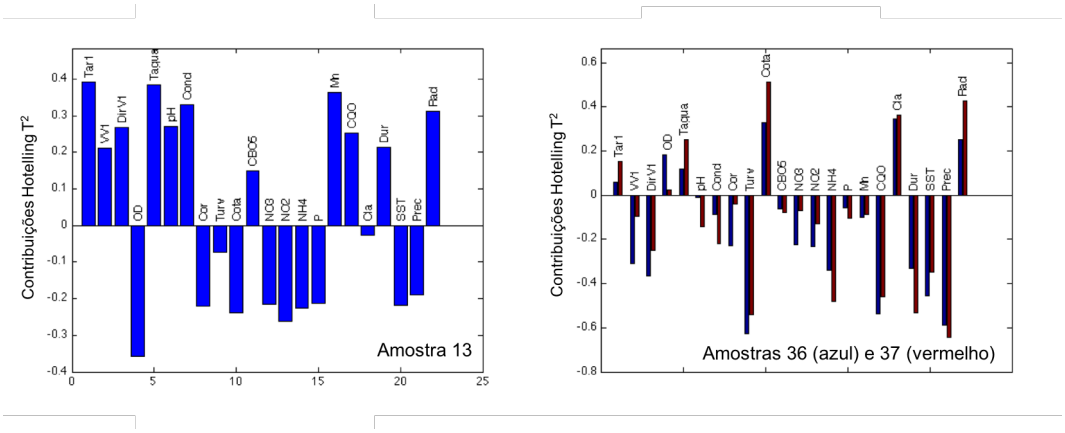

Figura 4.4: Contribuições para a estatística de Hotelling T^2 para as amostras 13 (esquerda) e 36 e 37 (direita)

A técnica PCA permite também identificar quais as variáveis que poderão estar mais relacionadas com a densidade de cianobactérias. Esta análise pode ser realizada, como referido no subcapítulo 2.6.2 pela exploração dos *scores* e *loadings*. Na Figura 4.5 é possível observar o gráfico contendo os *scores* e *laodings* para as duas primeiras componentes principais.

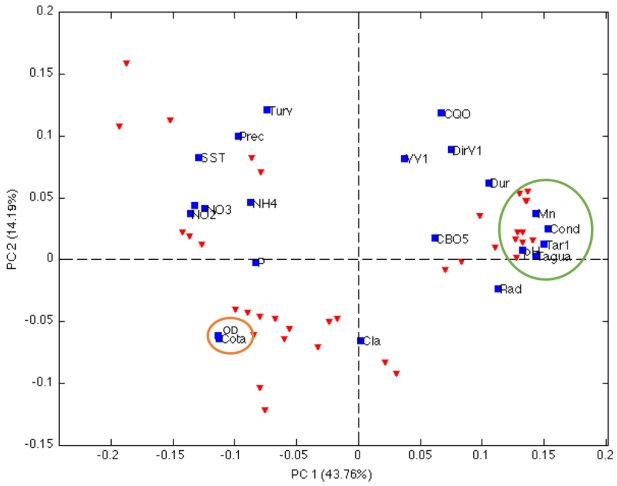

Figura 4.5: *Biplot* contendo os *loadings* e *scores* das duas componentes principais

Da análise da Figura 4.5 resulta que a Cond, o Mn, o pH, a Tag e a Tar estão relacionados positivamente com a densidade de cianobactérias. No quadrante oposto encontram-se as variáveis Cota e

OD, tendo uma relação inversa com estes organismos.

A condutividade estará relacionada positivamente com a densidade de cianobactérias devido à sua ligação à concentração de sais disponíveis na água [63]. Por sua vez o Manganês encontra-se neste grupo provavelmente porque as condições que favorecem o crescimento de cianobactérias também favorece a libertação deste composto dos sedimentos [7]. O pH tem uma correlação positiva com a densidade de cianobactérias, que poderá estar relacionada com o pH ideal para o crescimento de cianobactérias ser de 7,5 a 9,0 [5]. As temperaturas do ar e água influenciam positivamente o crescimento das cianobactérias, uma vez que quanto maior o valor de temperatura maior a taxa de crescimento destes organismos, atingindo um máximo acima dos 25^{o}C [6].

Por sua vez, o nível de oxigénio dissolvido encontra-se no quadrante oposto tendo uma relação negativa com a densidade das cianobactérias, que poderá justificar-se porque o O_2 tem uma relação também inversa com a temperatura, ou seja, quanto maior a temperatura, menor a solubilidade deste gás na água. No mesmo quadrante encontra-se a cota da albufeira, este parâmetro pode influenciar, diretamente ou indiretamente, a densidade de cianobactérias, como por exemplo na relação com a disponibilidade de nutrientes, i.e., quanto maior a cota mais diluídos se encontram os nutrientes [64]. Este parâmetro está também relacionado com a temperatura da água na Albufeira, ou seja, quanto menor a cota maior será a temperatura da água.

4.3 Modelação da densidade de cianobactérias com modelações lineares

A primeira abordagem na modelação dos dados para a previsão do *bloom* de cianobactérias com 15 dias de antecedência passou pela técnica PLS. Esta permite testar modelos mais simples, antes de desenvolver métodos não-lineares mais complexos (RNAs). Esta técnica foi utilizada para três conjuntos de parâmetros: o primeiro com os parâmetros que de acordo com o modelo PCA estão correlacionados com a densidade de cianobactérias, o segundo sem restrição (todos os parâmetros possíveis) e finalmente com os parâmetros cuja monitorização poderia ser realizada pela AgdA com maior frequência (Tabela 4.2). Para o desenvolvimento dos modelos, os dados foram divididos em dois grupos, um para calibração e outro para testar o modelo, numa proporção de 70%/30%. Ou seja, o conjunto de calibração correspondia às primeiras amostras que representavam 70% do total de dados disponível. O modelo foi testado com as restantes amostras (30%). Neste estudo, uma vez que a preparação dos dados para modelação implicava a seleção de vários fatores, optou-se por considerar um número constante de variáveis latentes, não sendo por isso necessário definir um grupo de validação. Para os dados em questão considerou-se que o número ideal de variáveis latentes seria quatro.

A escolha do modelo PLS teve por base dois critérios: a proximidade dos valores de RMSE, R^2 e RER (como descrito no subcapítulo 2.6.5) para os dados de calibração e previsão e a significância das variáveis utilizadas. Assim, parte-se da matriz com os dados de todas as variáveis. Após a geração do modelo verifica-se quais as variáveis com significância para o mesmo e eliminam-se as restantes. De

Tabela 4.2: Variáveis utilizadas na matriz inicial de cada um dos modelos PLS construídos neste estudo

Modelo	Variáveis utilizadas na matriz inicial
I	Variáveis que segundo o PCA estariam mais relacionadas com a densidade de cianobactérias
II	Todas as variáveis
III	Variáveis com maior periodicidade de monitorização

seguida, gera-se um novo modelo com a matriz resultante e volta-se a analisar quais os parâmetros significativos. Este processo é repetido até se encontrar um modelo cujos valores de RMSE, R^2 e RER sejam próximos para os dados de calibração e teste.

Antes de optar pela metodologia supramencionada tentou-se realizar a seleção da conjugação de percX e delayX que permitisse obter um modelo com as características desejadas, criando uma malha de valores de RMSE para os dados de calibração e teste, como ilustrado na Figura 4.6. No entanto, não foi possível utilizar este procedimento uma vez que os dados disponíveis mostraram-se insuficientes para a realização desta análise.

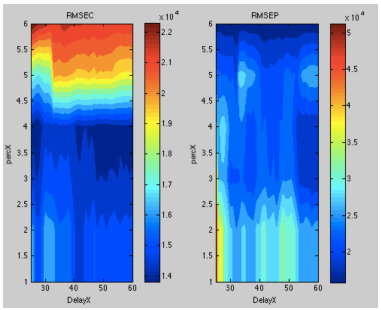

Figura 4.6: Exemplo de resultado da análise dos valores de RMSE de acordo com a conjugação de diferentes percX e delayX

Na Figura 4.7 é apresentado um exemplo do resultado do modelo para os dados de calibração, na qual é possível realizar a análise dos parâmetros utilizados para a seleção do modelo PLS.

Na imagem do canto superior esquerdo da Figura 4.7 é possível identificar quais as amostras utilizadas para calibração. Nos gráficos do canto superior direito e do canto inferior esquerdo são representados os valores experimentais e valores estimados pelo modelo (*predicted*), bem como os respetivos valores de RMSE, RER e R^2. Finalmente, na ilustração do canto inferior direito são representados os coeficientes de regressão e a significância das variáveis utilizados no modelo. Esta é avaliada considerando um nível de confiança. Assim, se a barra de erro, que representa a zona em que há um efeito devido a uma causa aleatória, ultrapassar o limite superior ou inferior dos limites de confiança a variável em questão é significativa ao nível de siginificância definido para o teste (neste trabalho foi sempre 0,05).

4.3.1 Utilização das variáveis com maior correlação linear

Como referido anteriormente, um dos modelos PLS criado neste estudo partiu de uma matriz de dados constituída pelas variáveis que segundo o modelo PCA estariam mais relacionadas com a densidade

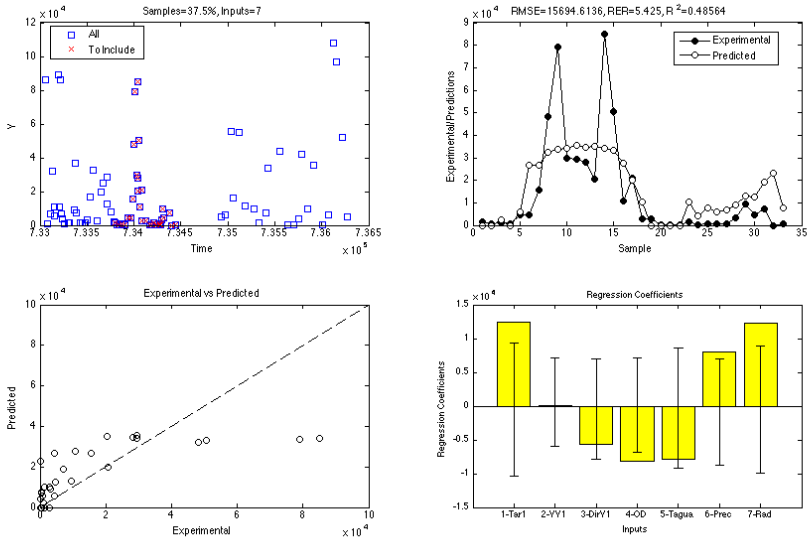

Figura 4.7: Exemplo de resultado do modelo PLS para o conjunto de dados de calibração

de cianobactérias. De acordo com esta premissa, a matriz de partida para este modelo continha as seguintes entradas: Tar, Tag, pH, Cond, OD, Cota e Mn (*vide* Figura 4.5). Após realização do procedimento descrito anteriormente para otimização do modelo PLS selecionou-se o modelo com quatro variáveis de entrada - Tar, Tag, pH e Cond, como ilustrado na Figura 4.8.

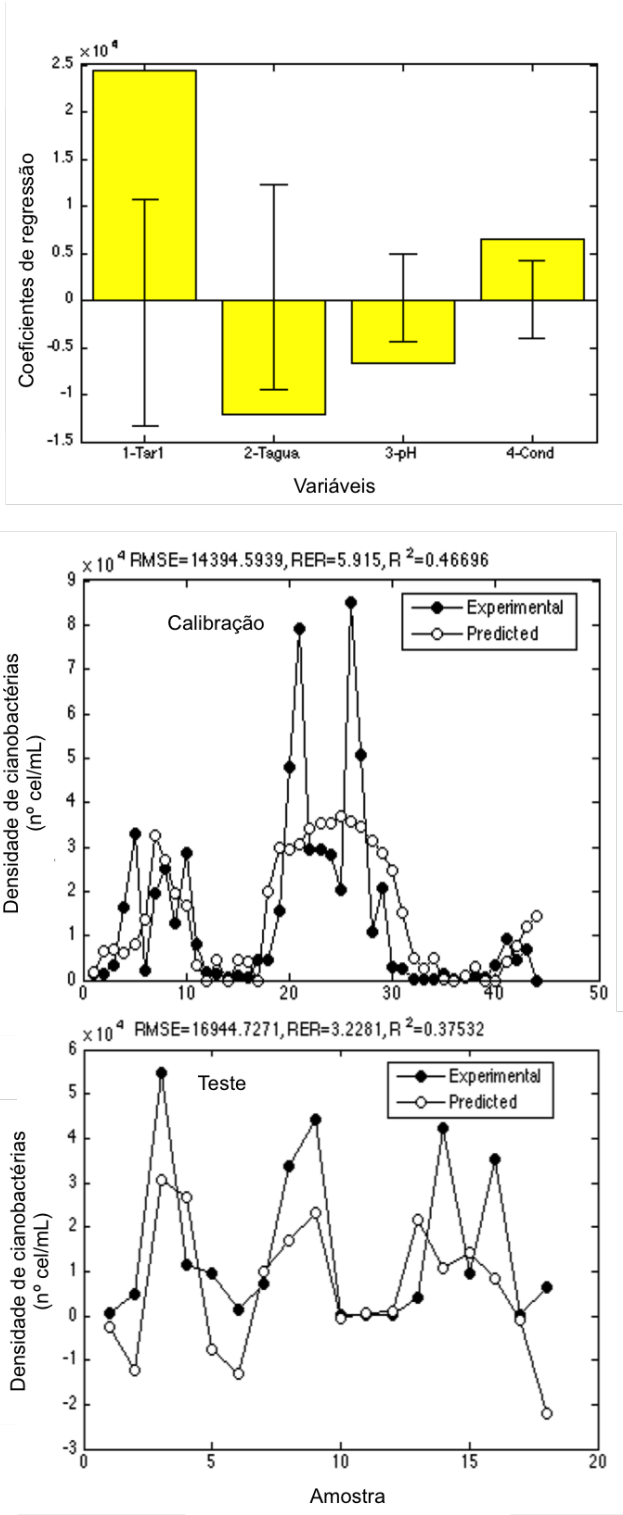

Figura 4.8: Coeficientes de regressão e significância das variáveis (em cima) e resultados da otimização do modelo desenvolvido a partir de uma matriz inicial constituída pelas variáveis que segundo o PCA teriam maior correlação com a densidade de cianobactérias para os grupos de calibração (centro) e de teste (em baixo)

A comparação dos parâmetros de avaliação do modelo permite verificar que o valor de RMSE aumenta para os dados de teste, de 14 395 para 16 945 células/mL. Por sua vez, o valor de RER diminui de 5,9 para 3,2. O mesmo acontece com o valor de R^2 que passa de um valor de 0,47 para 0,38 (Figura 4.9). De notar que as diferenças apresentadas não são significativamente diferentes.

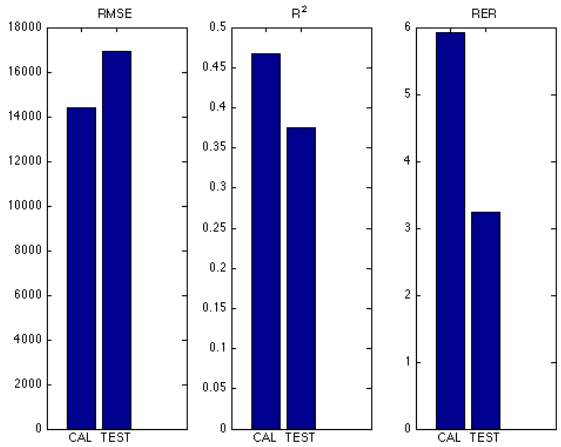

Figura 4.9: Comparação dos parâmetros RMSE, R^2 e RER para os dados de calibração e teste para modelo otimizado desenvolvido a partir de uma matriz inicial constituída pelas variáveis que segundo o PCA estariam mais relacionadas com a densidade de cianobactérias

Ao analisar os dados da Figura 4.8 observa-se que contrariamente ao esperado, a Tar e Tag têm uma contribuição de sinal oposto, também o pH tem contributo de sinal inverso ao resultado do modelo PCA. Colocou-se a hipótese deste facto acontecer por um modelo linear realizar um ajuste quando as variáveis estão correlacionadas, ou seja, dá um maior peso a uma e depois usa a outra variável para contrabalançar, atribuíndo-lhe um sinal contrário. Para verificar a veracidade da hipótese formulada, introduziu-se como variável de entrada o produto da Tar e Tag, o resultado desta análise é apresentado na Figura 4.10.

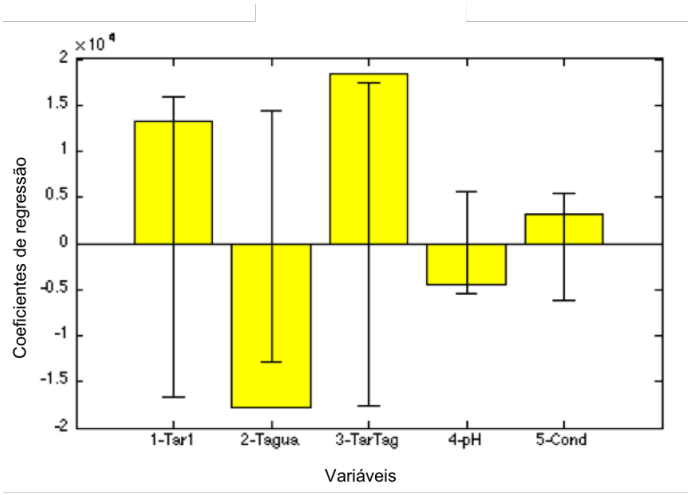

Figura 4.10: Coeficientes de regressão e significância das variáveis para um modelo incluindo o termo de interação Tar e Tag

De acordo com os resultados ilustrados na Figura 4.10, verifica-se que uma vez que as variáveis Tar e Tag não se anulam a hipótese formulada está correta, i. e., as variáveis estão correlacionadas.

A correlação das variáveis pH e Cond foi avaliada pela análise de variância, cujo resultado é apresentado na Tabela 4.3. Como é o possível verificar no resultado do teste Turkey's HSD as médias de todos os grupos são diferentes, pelo que há uma correlação entre as variáveis que estará a ser compensada pelo sinal inverso de contribuição para o modelo linear.

Tabela 4.3: Resultados da ANOVA entre pH e condutividade

Lillietest	Barlett	p-value ANOVA	Turkey	N^o dados	N^o dados G1	N^o dados G2	N^o dados G3
0	0,25063	$3{,}7x10^{-8}$	Todos dif.	47	17	13	17

4.3.2 Utilização de todas as variáveis

Outra tentativa de encontrar um modelo PLS que se adeque à previsão de *blooms* de cianobactérias passou pelo uso de uma matriz de partida sem limitações de variáveis. Os parâmetros de partida foram os seguintes: Tar, VV, DirV, OD, Tag, pH, Cond, Cor, Turv, Cota, CBO_5, NO_3, NO_2, NH_4, P_2O_5, P, Mn, N, Fe, Alc, Ca, CQO, Cl-a. Após a execução dos passos de optimização do modelo selecionaram-se 6 variáveis (Tar, OD, Turv, NO_3, Mn e CQO), cujas significâncias podem ser observadas na Figura 4.11 (esquerda). Nesta mesma figura é possível comparar os valores experimentais com as previsões para os dados de calibração e de teste.

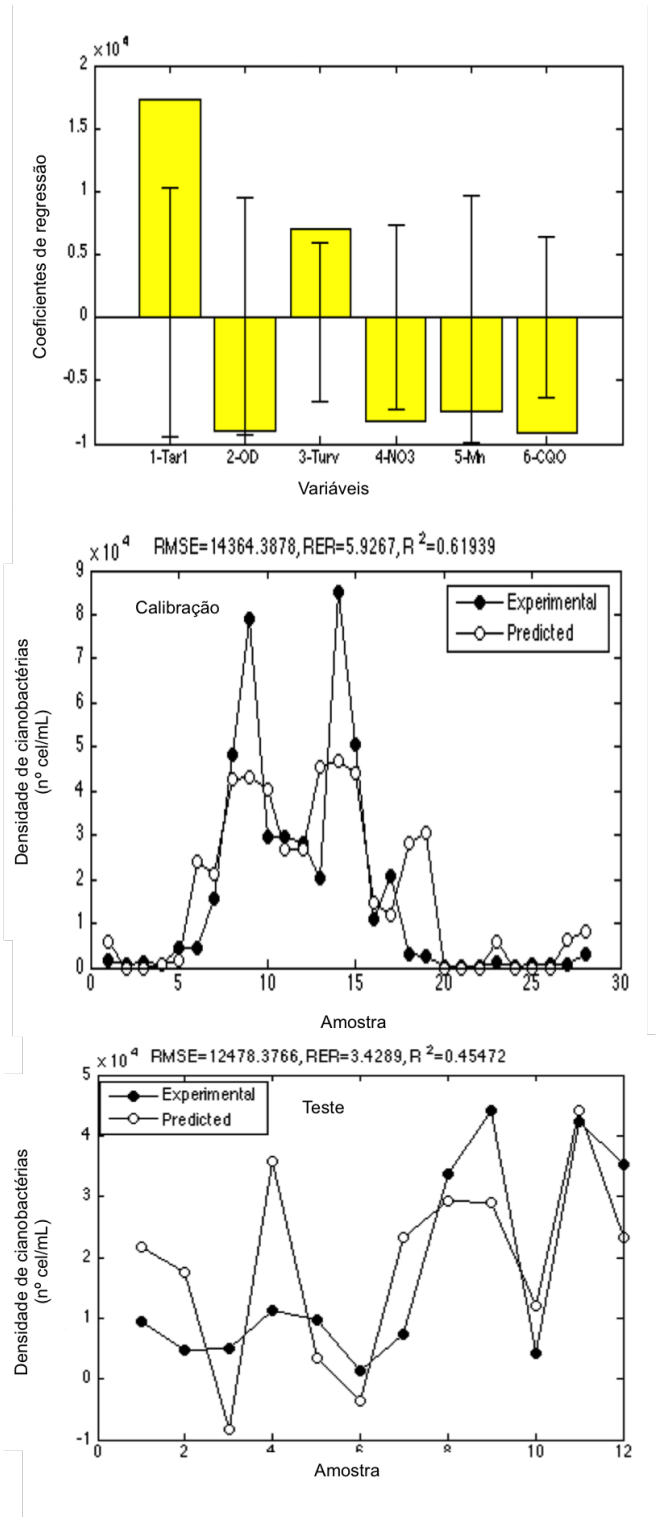

Figura 4.11: Coeficientes de regressão e significância das variáveis (em cima) e resultados da otimização do modelo desenvolvido a partir de uma matriz inicial de dados sem restrições de variáveis para os grupos de calibração (centro) e de teste (em baixo)

De acordo com o ilustrado na Figura 4.12 a aplicação do modelo aos dados de teste resulta numa diminuição do valor de RMSE de cerca de 14 264 células/mL para um valor próximo de 12 478 células/mL. O valor de RER também reduziu passando de 5,9 para 3,4. A mesma tendência foi verificada para o R^2 que alterou de 0,62 para 0,45. Comparando este modelo com o anterior, constata-se que o RMSE de calibração é muito próximo, no entanto o RMSE de teste é inferior para este modelo. Os valores de RER são idênticos para os dois modelos, quer para os dados de calibração como para o teste, por sua vez o R^2 de teste deste modelo é um pouco superior. À semelhança do modelo anterior, as diferenças encontradas entre os dados de calibração e teste não são significativas.

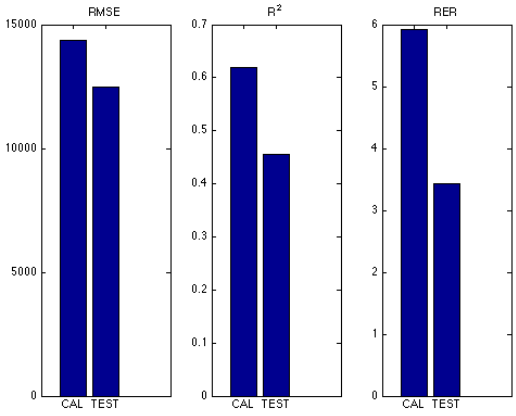

Figura 4.12: Comparação dos parâmetros RMSE, R^2 e RER para os dados de calibração e teste que resultaram da optimização do modelo desenvolvido a partir de uma matriz inicial de dados sem restrições de variáveis

4.3.3 Utilização das variáveis com maior frequência de amostragem

Por fim, criou-se um modelo cujos parâmetros de entrada foram os identificados pela entidade gestora como passíveis de maior periodicidade de monitorização, bem como parâmetros meteorológicos, que a AgdA indicou que, apesar de não serem monitorizados por eles, seriam possíveis de obter em tempo útil, uma vez que se optou por um desfasamento de 60 dias para o cálculo da mediana das entradas. A matriz resultante foi constituída pelas seguintes variáveis: Tar, VV, DirV, OD, Tag, pH, Cond, Turv, Mn, Prec e Rad. Na Figura 4.13 é apresentado o resultado do modelo com melhores resultados.

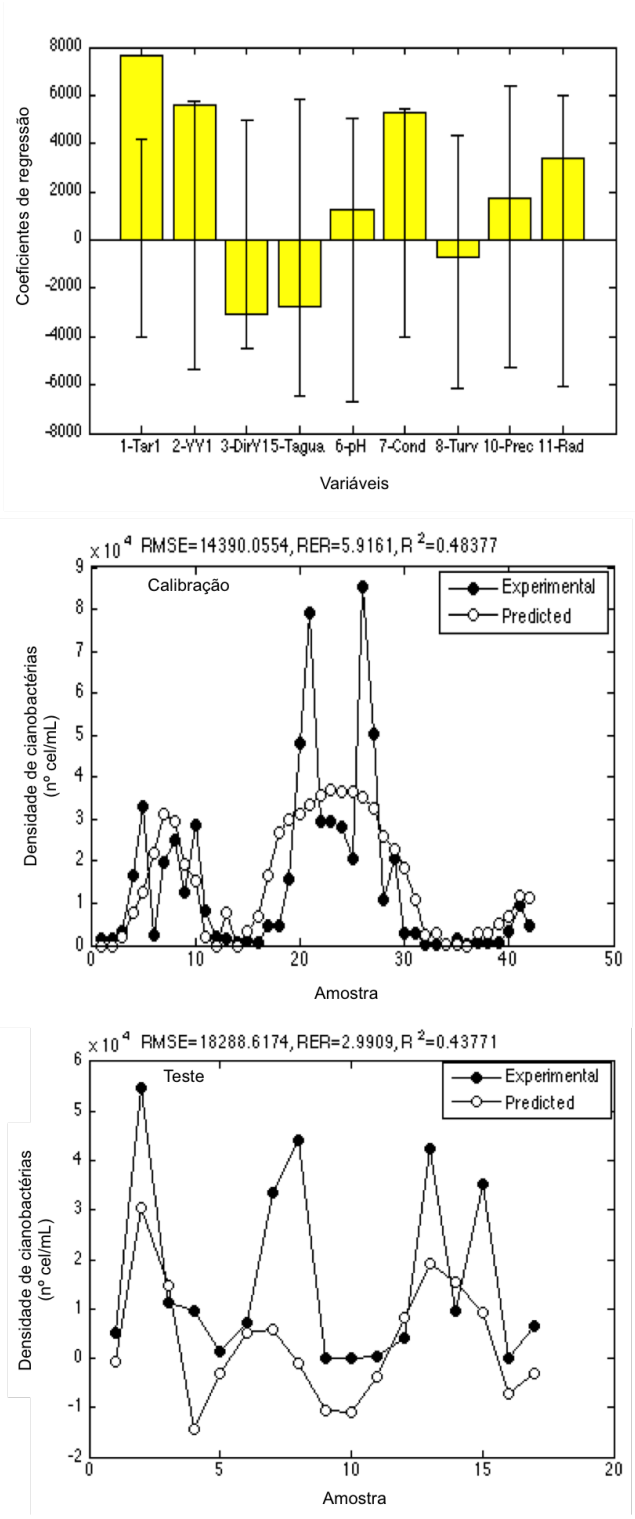

Figura 4.13: Coeficientes de regressão e significância das variáveis (em cima) e resultados da otimização do modelo desenvolvido a partir de uma matriz inicial constituída pelas variáveis com maior periodicidade de monitorização para os grupos de calibração (centro) e de teste (em baixo)

Como é possível verificar na Figura 4.13 o conjunto de parâmetros com melhores resultados é constituído por Tar, VV, DirV, Tag, pH, Cond, Turv, Prec e Rad. Este modelo apresenta um valor de RMSE de calibração de 14 390 células/mL, próximo ao dos restantes modelos, e de teste de 18 289 células/mL, maior valor quando comparado com os modelos anteriores. Os valores de R^2 são iguais ao modelo cuja matriz de partida era constituída pelas variáveis que segundo o PCA estariam mais relacionadas com a densidade de cianobactérias (0,48 e 0,44, para os dados de calibração e teste). Os valores de RER são também muito próximos entre estes dois modelos (5,9 e 3, para os dados de calibração e teste). Na Figura 4.14 são ilustradas estas diferenças dos valores destes parâmetros para os dados de calibração e teste para modelo desenvolvido a partir de uma matriz inicial constituída pelas variáveis com maior periodicidade de monitorização.

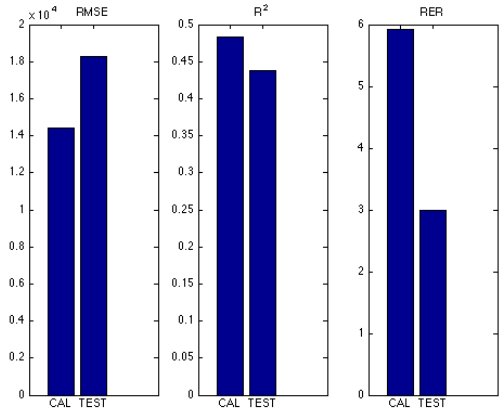

Figura 4.14: Comparação dos parâmetros RMSE, R^2 e RER para os dados de calibração e teste para modelo desenvolvido a partir de uma matriz inicial constituída pelas variáveis com maior periodicidade de monitorização

A Tabela 4.4 contém o resumo dos parâmetros de avaliação dos três modelos lineares selecionados para os diferentes grupos de variáveis de partida.

Tabela 4.4: Síntese dos resultados obtidos para os três modelos PLS criados com diferentes grupos de variáveis de partida

Modelo	RMSEcal	RMSEtest	R^2cal	R^2test	RERcal	RERtest	Nº Variáveis de entrada
Variáveis com maior correlação linear	14 395	16 945	0,47	0,38	5,9	3,2	4
Sem restrição variáveis	14 264	12 478	0,62	0,45	5,9	3,4	6
Variáveis maior periodicidade de monitorização	14 390	18 289	0,48	0,44	5,9	3,0	11

Modelo	Tar	VV	DirV	OD	Tag	pH	Cond	Turv	NO_3	Mn	CQO	Prec	Rad
Variáveis com maior correlação linear	x				x	x	x						
Sem restrição variáveis	x			x				x	x	x	x		
Variáveis maior periodicidade de monitorização	x	x	x	x	x	x	x	x		x		x	x

De acordo com os resultados obtidos verifica-se que os modelos são muito equivalentes em termos de RMSE, RER e R^2. Por este motivo, a AgdA deve optar pelo modelo optimizado a partir de uma matriz com as variáveis com maior correlação linear para previsão dos *blooms* de cianobactérias (primeiro modelo apresentado), uma vez que se obteve uma qualidade idêntica aos restantes com um menor número de variáveis de fácil obtenção e com possibilidade de serem monitorizadas com elevada frequência.

4.4 Modelação da densidade de cianobactérias com modelações não-lineares

De acordo com o descrito anteriormente, a ocorrência de *blooms* de cianobactérias é bastante irregular, o que poderá justificar o erro significativo obtido com os modelos lineares testados. Por este motivo avaliou-se também a potencialidade de modelos não-lineares (RNA) para a previsão deste fenómeno. A escolha de RNA para a modelação destes fenómenos deve-se à sua capacidade de descrever as relações não lineares que existem entre as variáveis que caracterizam os ecossistemas e por sua vez o crescimento de fitoplâncton [45; 65]. Neste subcapítulo são apresentados os resultados de duas tipologias de RNA - as *feedforward* e recorrentes.

4.4.1 RNA do tipo *feedforward*

Como referido no subcapítulo 3.2.3, a matriz criada para a criação de RNA, que permitia a utilização do maior número de amostras, era constituída por 12 variáveis de entrada, nomeadamente Tar, VV, DirV, Tag, pH, Cond, Turv, Cota, NH_4, Dur, Prec e Rad. Na Figura 4.15 apresenta-se uma ilustração da RNA *feedforward* utilizada neste estudo.

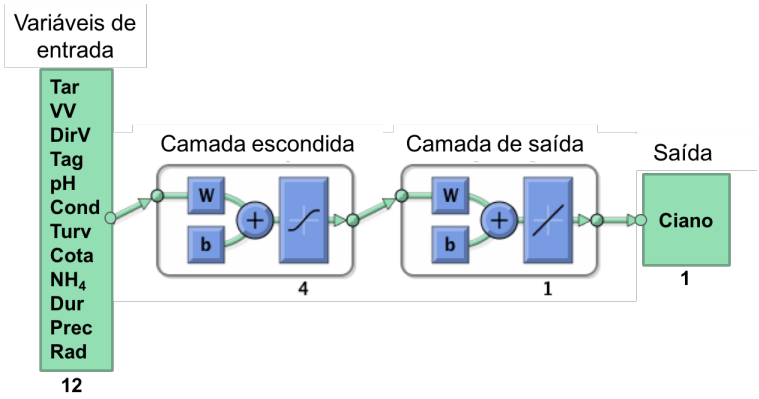

Figura 4.15: Ilustração da RNA *feedforward* utilizada neste estudo

Como é possível observar na Figura 4.15, optou-se por uma rede de uma camada escondida com quatro nodos. A seleção do número de nodos teve por base o valor de erro para o grupo de teste, ou seja, treinaram-se redes com diferentes número de nodos 100 vezes e, no final, selecionou-se a rede cujo erro do grupo de teste era inferior (Tabela 4.5). A opção pela utilização de apenas 1 camada

escondida foi tomada uma vez que, é opinião de vários autores, nomeadamente Hornik *et al.* [66], uma rede neuronal multicamada *feedforward* com uma camada escondida é suficiente para a resolução da maioria dos problemas.

Tabela 4.5: Valor de erro do conjunto de teste de acordo com o número de nodos da RNA

N^o de nodos	Menor erro de 100 treinos (células/mL)
2	1,18x10^8
3	7,61x10^7
4	6,51x10^7
5	3,43x10^8
6	1,91x10^8

O resultado da RNA *feedforward* escolhida (menor erro conjunto de teste = 6,51x10^7) encontra-se ilustrado na Figura 4.16.

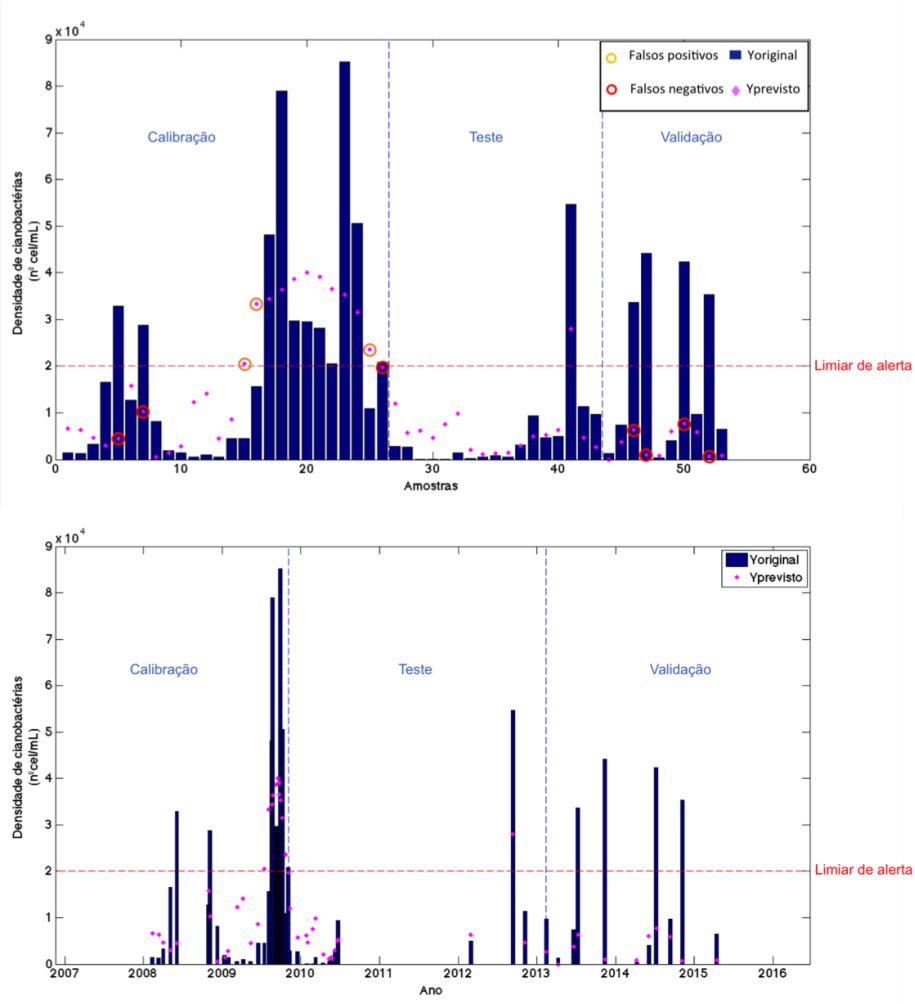

Figura 4.16: Resultados das previsões da RNA *feedforward* igualmente espaçadas (em cima) e espaçados no tempo (em baixo)

Como referido no subcapítulo 3.3, para além do valor de erro, verificou-se também a capacidade do modelo em prever se a densidade de cianobactérias estaria ou não acima das 20 000 células/mL (assinalado com uma linha vermelha a tracejado). Ao analisar a Figura 4.16, é possível verificar que

algumas das previsões desta rede geram falsos negativos - assinalados com círculos vermelhos, e em menor número falsos positivos - assinalados com círculos amarelos. É de assinalar que, nos dados de teste do modelo não ocorreram estes desvios.

4.4.1.1 Análise de sensibilidade para RNA *feedforward*

A análise de sensibilidade dos parâmetros permite verificar como é que a previsão do modelo é influênciada por oscilações de cada uma das variáveis de entrada [67]. Neste estudo, optou-se por verificar o efeito da variação de um parâmetro de cada vez, assumindo que os restantes não se alteram. Para este efeito, fez-se variar o valor do parâmetro a avaliar entre o seu valor máximo e mínimo observado para este conjunto de dados, mantendo as restantes variáveis no seu valor médio. Na Figura 4.17 é possível observar o resultado desta análise para as doze variáveis utilizadas na construção da RNA *feedforward*.

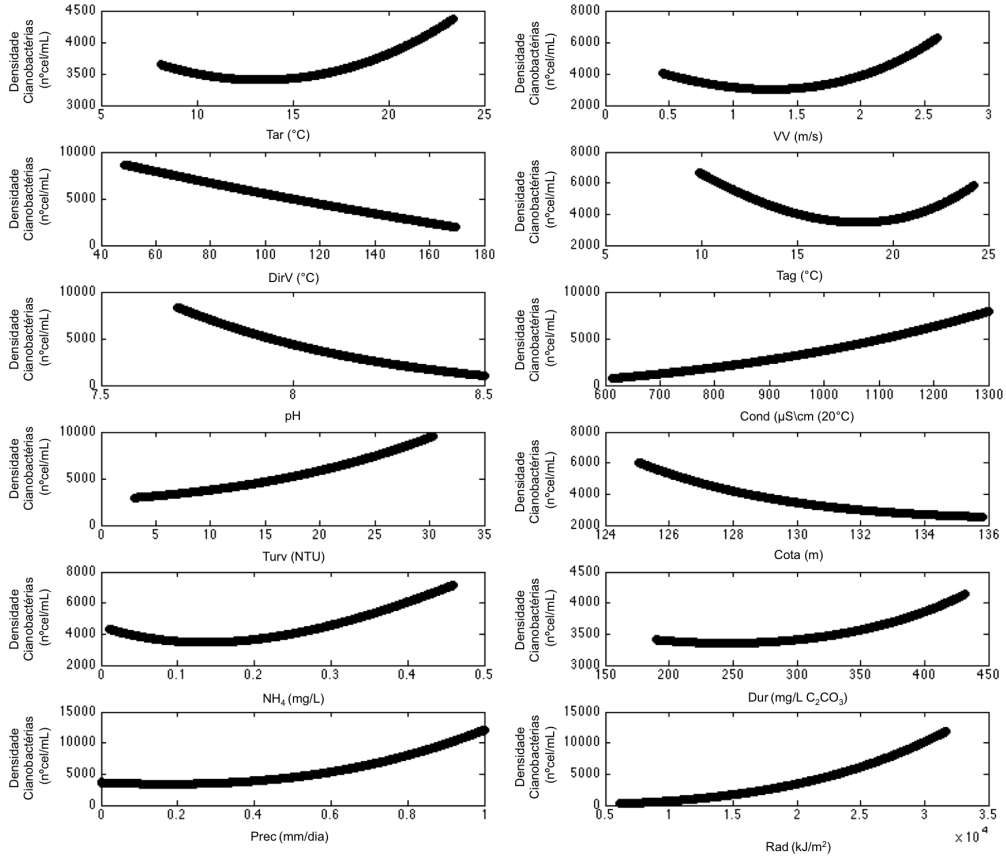

Figura 4.17: Resultado da análise de sensibilidade para a RNA *feedforward*

O coeficiente de sensibilidade de cada variável, que permite quantificar a magnitude da sensibilidade dos diferentes parâmetros, foi calculado através da Equação 4.1.

$$S = \frac{\Delta Y}{\Delta X}, \tag{4.1}$$

em que ΔY corresponde à diferença entre valor máximo e mínimo da previsão e ΔX corresponde à diferença entre valor máximo e mínimo da entrada objeto de análise.

Na Figura 4.18 são ilustrados os valores dos coeficientes de sensibilidade obtidos para as variáveis utilizadas na RNA *feedforward*, na qual é possível observar que variáveis como DirV, pH, Cond, Turv, Prec e Rad têm maior influência para o modelo.

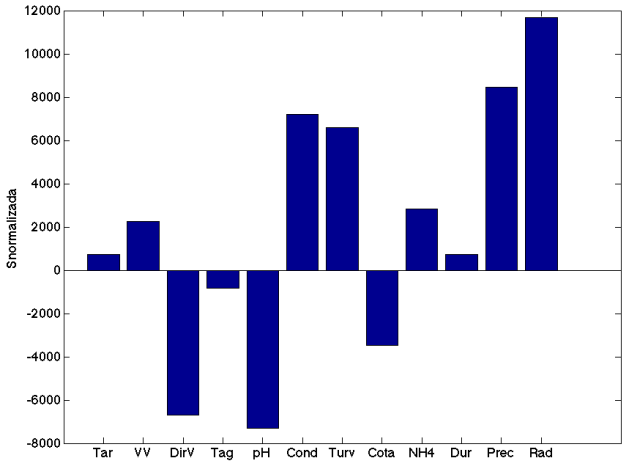

Figura 4.18: Coeficientes de sensibilidade das variáveis utilizadas na RNA *feedforward*

Após a análise de sensibilidade gerou-se uma nova RNA *feedforward* tendo como variáveis de entrada aquelas que revelaram maior valor de sensibilidade - DirV, pH, Cond, Turv, Prec, Rad (Figura 4.18) - para verificar se a qualidade de previsão do modelo alteraria com a diminuição do número de entradas. Optou-se por uma arquitetura de uma camada escondida com dois nodos, cujo resultado é apresentado na Figura 4.19.

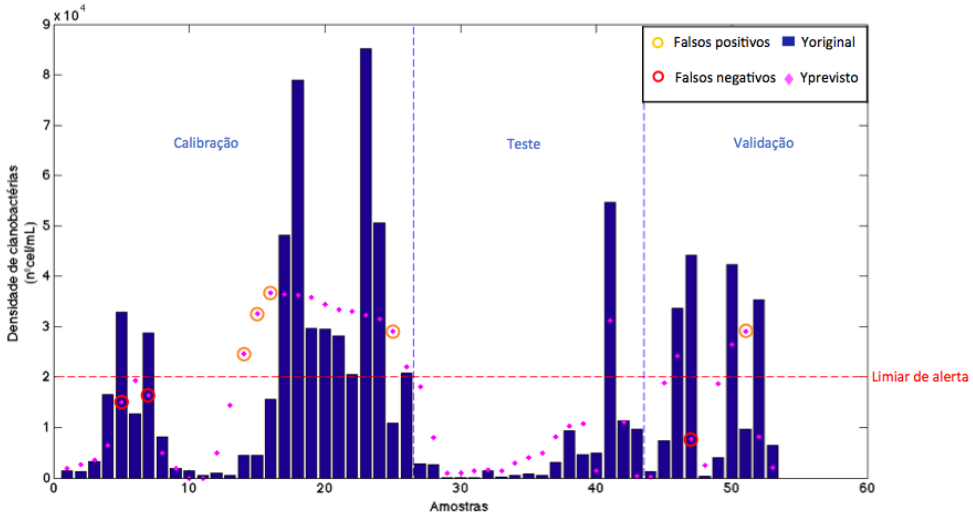

Figura 4.19: Resultados de previsão da RNA *feedforward* com variáveis com maior sensibilidade como entradas

Da análise realizada á Figura 4.19, pode-se concluir que não ocorreram previsões para os dados

de teste que resultaram em falsos negativos/positivos, condição já verificada para a RNA *feedforward* contruída com as 12 variáveis de entrada. Ao comparar o valor de erro para o conjunto de teste obtido para as duas RNA, concluíu-se que a alteração não é significativa, diminuíndo pouco, de 6,51x10^7 para 5,95x10^7 células/mL. Apesar de ter sido obtida uma RNA com menor erro para o conjunto de teste, no decorrer deste trabalho continuou-se a utilizar a RNA *feedforward* com 12 variáveis de entrada para tornar possível a comparação com os restantes modelos.

4.4.1.2 Comparação da sensibilidade das variáveis da RNA *feedforward* e coeficientes do modelo linear

A comparação do desempenho do modelo linear com a qualidade das redes neuronais só faz sentido se as variáveis de entrada utilizadas forem as mesmas. Assim, criou-se um modelo linear com uma matriz de entrada igual à utilizada para as RNA, e compararam-se as previsões obtidas com os valores experimentais, como realizado com os resultados das RNA. Na Figura 4.20 apresenta-se o resultado desta análise. Verifica-se que para o conjunto de teste são previstos três falsos negativos (círculos vermelhos), no entanto, não ocorreram previsões de valores correspondentes de falsos positivos (círculos amarelos). Neste modelo obteve-se um valor de erro do conjunto de teste de 3,20x10^8 células/mL, da mesma ordem de grandeza dos erros das RNA NARX, mas superior ao da RNA *feedforward*.

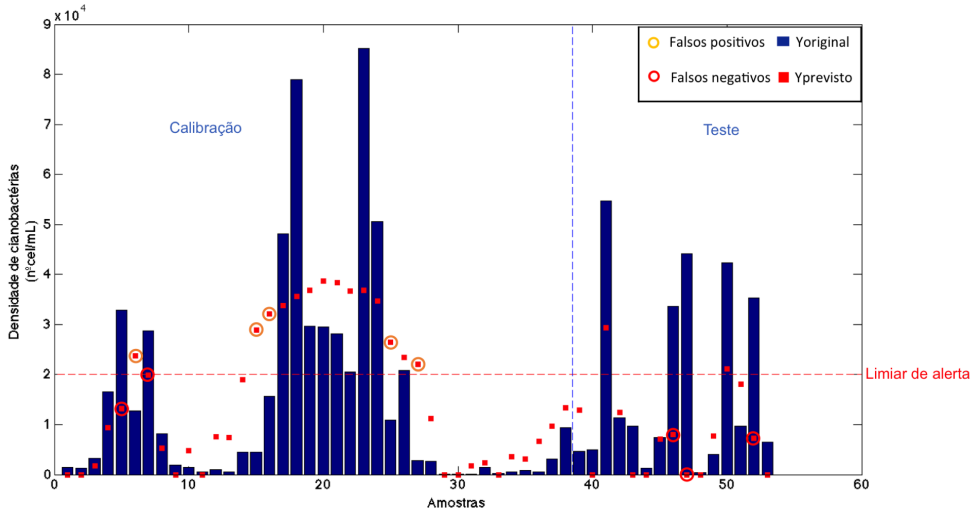

Figura 4.20: Previsões do modelo linear para as variáveis de entrada iguais às das redes neuronais

Uma vez que o modelo linear e a RNA têm como objetivo encontrar uma função que minimize o erro pode-se comparar os coeficientes de regressão com os coeficientes de sensibilidade das variáveis utilizadas. Os valores obtidos para as doze variáveis utilizadas são ilustrados na Figura 4.21. Verifica-se, contrariamente ao esperado, que a ordem de grandeza e sinal dos dois parâmetros não são todos idênticos. As maiores diferenças verificam-se para o pH, pois os sinais são contrários e o valor de sensibilidade é muito superior. A condutividade, turvação, cota, precipitação e radiação, apesar de sinal igual, apresentam um valor de sensibilidade muito superior ao do coeficiente de regressão. Para a temperatura do ar resultou um valor muito superior do coeficiente de regressão, apresentando os dois parâmetros sinal igual.

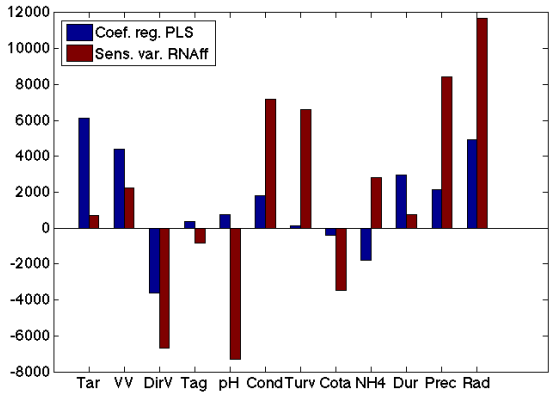

Figura 4.21: Comparação dos coeficientes de regressão do modelo PLS e sensibilidade das variáveis das RNA *feedforward*

4.4.2 RNA do tipo recorrente

Uma vez que o crescimento de algas é um fenómeno dinâmico, complexo e não linear pode ser descrito por uma RNA dinâmica, de acordo com H. Wang *et al.* [28]. Neste trabalho optou-se por avaliar a capacidade preditiva das RNA dinâmicas NARX para estes fenómenos, que, segundo o mesmo autor, têm uma forte capacidade de se adequar a processos não lineares dinâmicos. Na Figura 4.22 ilustram-se os dois tipos de arquitetura utilizadas neste estudo - a arquitetura de série-paralela (open loop) e a arquitetura paralela (closed loop).

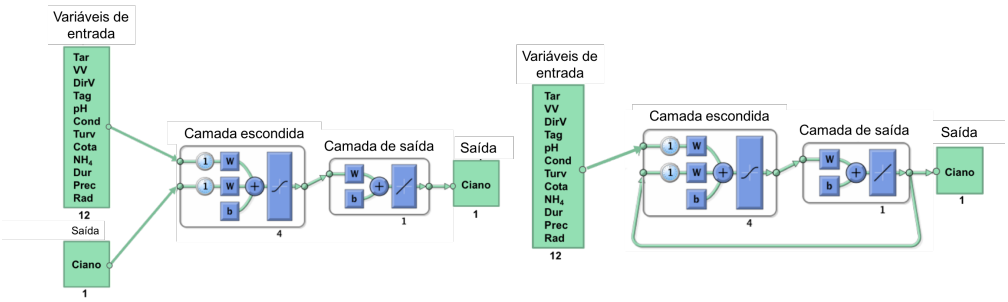

Figura 4.22: Ilustração das arquiteturas das redes NARX utilizadas neste estudo, *open loop* à esquerda e *closed loop* à direita

O número de nodos escolhido para utilizar nesta tipologia de redes foi o mesmo que se utilizou para a rede *feedforward* (4 nodos). Na Figura 4.23 é possível observar o histograma de erros, divididos pelos diferentes conjuntos de dados, da rede NARX *open loop* com 4 nodos.

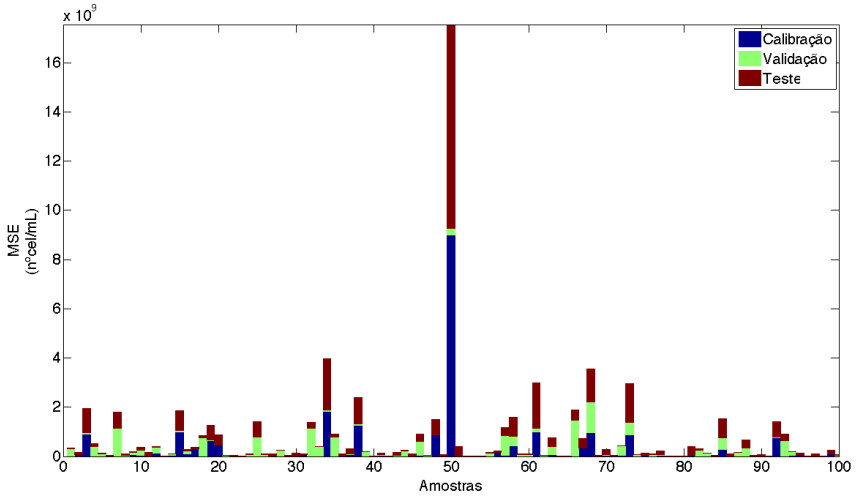

Figura 4.23: Histograma de erros dos diferentes conjuntos de dados utilizados no treino da rede NARX *open loop* com 4 nodos

A escolha da redes NARX foi executada com base no desempenho global da rede NARX *open loop*. Verificou-se que os valores de erro das RNA desta tipologia eram da mesma ordem de grandeza do erro global apresentado pela rede *feedforward*, exposta no subcapítulo 4.4.1 (2,66x10^8 células/mL). Assim, para a rede NARX com arquitetura *open loop* o valor de erro global é de 3,76x10^8 células/mL e para a rede NARX *closed loop* o valor do erro é um pouco menor, 3,55x10^8 células/mL. Na Figura 4.24 apresentam-se os resultados de previsão das duas tipologias de rede NARX utilizadas em comparação com os dados experimentais. Nesta encontram-se ainda identificados os falsos negativos - círculos vermelhos de linha cheia para a rede *open loop* e a tracejado para a rede *closed loop* e falsos positivos - círculos cor-de-laranja de linha cheia para a rede *open loop* e a tracejado para a rede *closed loop*.

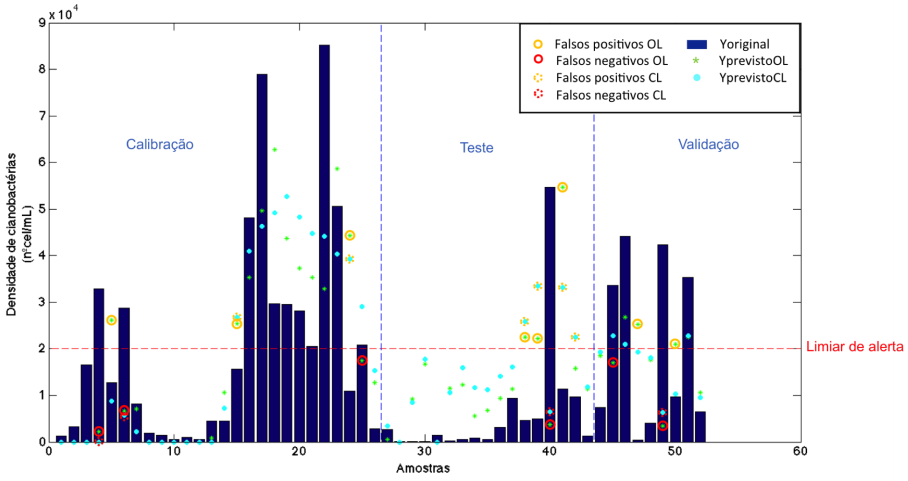

Figura 4.24: Previsões das redes NARX *open loop* e *closed loop*

Como ilustrado na Figura 4.24 ambas as tipologias de RNA NARX preveem falsos negativos e falsos positivos para os mesmos pontos do conjunto de teste, com exceção do ponto 42 em que apenas a rede NARX *closed loop* prevê um falso positivo. É possível verificar também que neste grupo de dados apenas ocorre um falso negativo, no ponto 40, para ambas as tipologias de rede.

Uma vez que os resultados obtidos pelas RNA NARX não eram significativamente melhores que os dos outros modelos, como esperado na literatura consultada, criou-se um modelo utilizando esta tipologia de RNA mas com a divisão dos dados para calibração, validação e teste de forma aleatória, com um rácio de 70/15/15. Os resultados obtidos em termos de erro foram efetivamente melhores do que os obtidos com uma divisão de dados de acordo com o índice das amostras, se bem que da mesma ordem de grandeza, 1,24x10^8 células/mL para a RNA NARX de arquitetura *open loop* e de 1,59x10^8 células/mL para a RNA NARX *closed loop*.

Da análise dos dados registados na Figura 4.25 verifica-se que as previsões geradas por este modelo acompanham de forma mais eficiente o perfil dos dados experimentais. Relativamente ao número de falsos negativos/positivos nas previsões das amostras de teste (assinaladas com T, junto ao eixo das abcissas), verifica-se que há um único ponto (amostra 25) cuja previsão da RNA NARX *open loop* resultou num falso positivo.

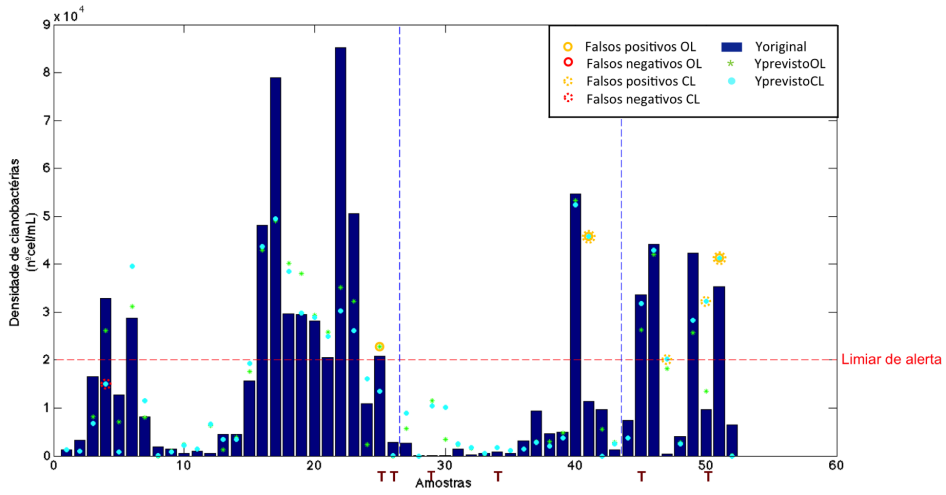

Figura 4.25: Previsões das redes NARX *open loop* e *closed loop* para uma divisão de dados aleatória

Apesar dos melhores resultados obtidos, esta RNA não deverá ser utilizada para previsão de *blooms*, uma vez que em cada simulação as amostras selecionadas para cada um dos grupos (calibração, validação e teste) não seriam as mesmas, pelo que não é possível garantir que a qualidade de previsão da RNA se mantenha.

4.5 Comparação das previsões pelos modelos desenvolvidos

Os modelos gerados neste estudo foram avaliados tendo por base dois parâmetros, o erro do conjunto de teste e o número de previsões que correspondiam a falsos negativos/positivos para o mesmo conjunto de amostras. Na Figura 4.26 ilustram-se os valores experimentais e as previsões obtidas pelos diferentes modelos gerados neste estudo. Nesta são, também, assinalados os falsos negativos/positivos.

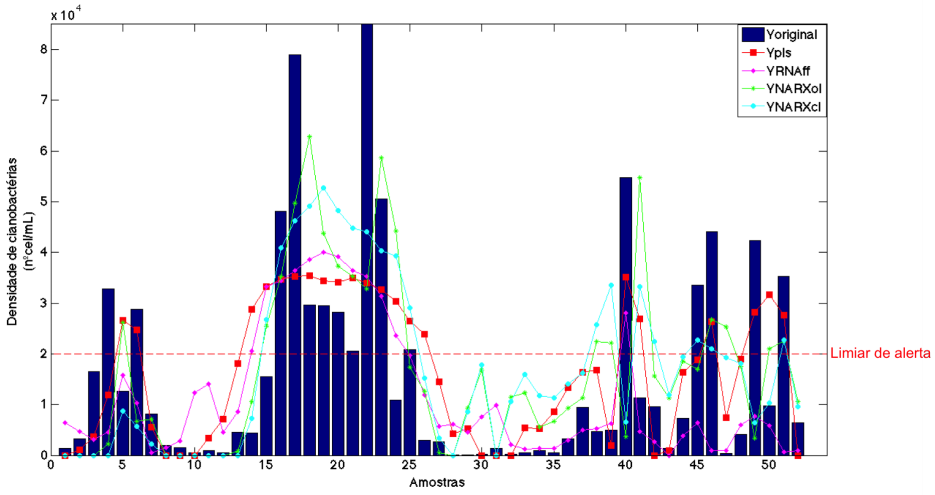

Figura 4.26: Comparação das previsões dos diferentes modelos com os dados experimentais

Ao analisar os dados da Figura 4.26 é possível verificar que em termos de perfil, as redes NARX *open loop* ajustam-se melhor aos valores elevados de densidade de cianobactérias (pontos 17 e 22), este facto poderá estar relacionado pela utilização dos dados experimentais como variável de entrada (Figura 4.22). Para a mesma região constata-se que o modelo linear e a RNA *feedforward* apresentam perfis de previsão semelhantes. Quando comparados os resultados para as amostras do conjunto de teste das redes (ponto 27 a 43) verifica-se qua as redes NARX preveem alguns pontos como falsos positivos/negativos. No entanto, são previstos corretamente, em relação ao limiar de alerta, pela RNA *feedforward* e pelo modelo linear (estes pontos também pertencem ao grupo de teste deste modelo). Na Figura 4.27 apresenta-se, de uma outra forma, a ocorrência de falsos positivos/negativos nas previsões dos vários modelos.

Legenda:

Falso positivo

Falso Negativo

Tipo de Modelo	% Falsos negativo	%Falsos positivo	% falsos positivos teste	% falsos negativos teste
PLS	9%	9%	0%	20%
RNA-FF	13%	6%	0%	0%
RNA-NARX-OL	12%	15%	18%	6%
RNA-NARX-CL	6%	13%	24%	6%

Figura 4.27: Identificação dos falsos positivos e negativos para os diferentes modelos, e percentagens, globais e do conjunto de teste, de falsos negativos

Ao analisar a percentagem global dos falsos negativos, apresentada na Figura 4.27, verifica-se que é semelhante para os diferentes modelos. É possível verificar, também, que o modelo linear gerou maior número de falsos negativos (três) nos dados de teste, o que corresponde a uma percentagem de 20%, e que a rede NARX *closed loop* produziu maior percentagem de falsos positivos (24%). O único modelo para o qual não se observam falsos positivos/negativos no grupo de teste é o que corresponde à RNA *feedforward*.

Na Tabela 4.6 observam-se os valores de erro obtidos pelos diferentes modelos para o conjunto de teste. De acordo com o exposto o modelo gerado pela RNA *feedforward* é o que melhor se ajusta aos dados, uma vez que o seu erro é uma ordem de grandeza inferior quando comparado com os valores obtidos para os restantes modelos. Contrariamente ao esperado, o valor de erro dos modelos criados pelas RNA NARX é superior ao das restantes técnicas. Este facto poderá estar relacionado com a periodicidade de monitorização dos parâmetros de controlo de qualidade da água não ser constante (subcapítulo 3.1). Dentro da tipologia NARX, o modelo produzido pela RNA de arquitetura *closed loop*, imprevisivelmente, apresenta menor valor de erro quando comparada com a RNA NARX *open loop*.

Tabela 4.6: Comparação do erro do conjunto de teste para os diferentes modelos

Modelo	Erro do conjunto de teste (células/mL)
Linear	$3{,}20 \times 10^{8}$
RNA *feedforward*	$6{,}51 \times 10^{7}$
RNA NARX *open loop*	$3{,}51 \times 10^{8}$
RNA NARX *closed loop*	$3{,}21 \times 10^{8}$

Na Figura 4.28 ilustra-se a comparação dos resultados obtidos para os parâmetros de avaliação dos modelos gerados neste estudo, por ordem decrescente de resultado (e.g., o modelo que aparece em primeiro lugar na coluna "MSE para conj. teste" tem menor valor). Considerando que o modelo selecionado irá auxiliar a entidade gestora de serviços de água a prever a ocorrência de um *bloom* de cianobactérias, que poderá estar associados a problemas de qualidade de água, a escolha deve recair no modelo gerado através da RNA *feedforward*, uma vez que, de acordo com os resultados obtidos, apresenta menor valor de falsos negativos para o grupo de teste (conjunto selecionado para avaliar o desempenho dos modelos).

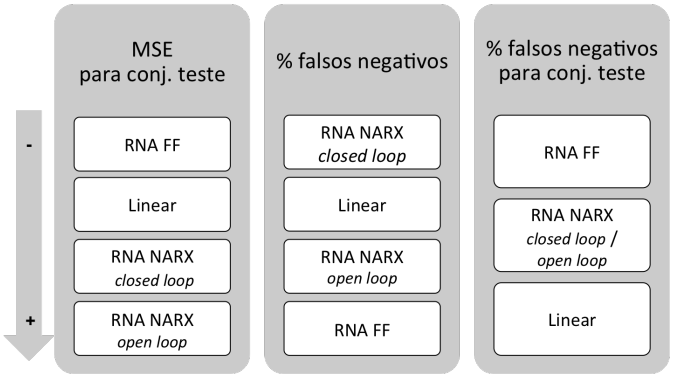

Figura 4.28: Comparação dos diferentes modelos de acordo com os parâmetros escolhidos para avaliar a sua qualidade

4.6 Avaliação do desempenho dos modelos para diferentes horizontes de previsão

Apesar dos modelos apresentados neste trabalho terem como objetivo a previsão de *blooms* de cianobactérias com 15 dias de antecedência, analisou-se também se os erros de modelação, das três tipologias de rede, seriam inferiores para diferentes horizontes de previsão, de 0 a 60 dias. Na Tabela 4.7 apresenta-se o menor valor de erro após 100 treinos das diferentes redes.

Tabela 4.7: Comparação do erro para as três tipologias de redes considerando diferentes horizontes de previsão (x10^8 células/mL)

Dias de previsão	0	5	10	15	20	25	30	35	40	45	50	55	60
RNA *feedforward*	**2,58**	2,71	2,62	2,66	2,87	2,63	3,19	2,94	2,66	2,92	3,36	2,69	2,65
RNA NARX *open loop*	3,40	3,02	3,57	3,76	**2,90**	3,90	3,03	3,43	3,73	3,57	3,68	3,51	3,61
RNA NARX *closed loop*	3,50	**3,09**	9,25	3,55	3,12	5,64	8,5	6,31	5,79	14,2	13,3	6,28	6,82

A partir da análise dos dados apresentados na Tabela 4.7 conclui-se que a ordem de grandeza dos diferentes erros é igual com exceção das previsões da rede NARX *closed loop* para um período de 45 e 50 dias que tem uma ordem de grandeza superior. Para todos os tempos, com exceção dos 30 dias, as RNA *feedforward* apresentam menor erro. Para o tempo de previsão de 30 dias o menor erro corresponde à rede NARX *open loop*. Por sua vez, as redes NARX *closed loop* apresentam erros superiores, com exceção da previsão para 15 dias em que o maior erro corresponde à tipologia NARX *open loop*.

4.7 Validação da metodologia para o ano de 2016

O modelo selecionado, RNA *feedforward*, foi testado utilizando os dados de monitorização de qualidade da água para o ano de 2016. Foram disponibilizadas quatro amostras, correspondentes aos meses de fevereiro, abril, junho e julho. Contrariamente ao esperado a densidade de cianobactérias nos meses de junho e julho atingiu valores da ordem de grandeza do valor desprezado no início deste estudo (subcapítulo 3.2).

Uma vez que a maioria das amostras tem um espaçamento de 60 dias, este teste foi realizado com os valores instantâneos, contrariamente aos dados utilizados para a criação do modelo (medianas dos últimos 60 dias). Apesar da ordem de grandeza de valores e do diferente tratamento dos dados, a RNA reagiu bem aos valores apresentados, prevendo apenas um valor que corresponde a um falso positivo para o mês de fevereiro. Os resultados obtidos apresentam-se na Figura 4.29.

Mês	Valor experimental (células/mL)	Resultado de Alerta (Nível de alerta: 20 000 células/mL)	Adequação do modelo
fevereiro	12	◉	✗
abril	4773	◉	✓
junho	422 466	◉	✓
julho	426 869	◉	✓

Figura 4.29: Resultado das previsões da RNA *feedforward* para valores 2016

Capítulo 5

Conclusões

5.1 Conclusões

Este trabalho teve como objetivo a criação de um modelo de alerta que permitisse à AgdA - Águas Públicas do Alentejo, entidade responsável pela gestão da água da Albufeira do Roxo para consumo humano, prever a ocorrência de *blooms* de cianobactérias com 15 dias de antecedência. Para este fim foram elaborados e testados vários modelos de previsão com ferramentas lineares (PLS) e não-lineares (RNA). Os dados utilizados foram recolhidos no período de 2007 a 2015, correspondendo a parâmetros de monitorização de qualidade da água da Albufeira âmbito do estudo e dados meteorológicos.

Para atingir o objetivo proposto desenvolveu-se uma metodologia que permitiu transformar a informação do controlo operacional, não preparada para ser utilizada em modelação, em informação útil. Ou seja, a AgdA pode utilizar o modelo gerado sem necessidade de alteração dos procedimentos de monitorização de qualidade da água, podendo continuar a recolher amostras, para alimentar o modelo, em intervalos de tempo irregulares e sem que os parâmetros monitorizados sejam os mesmos nos diferentes instantes de amostragem.

Os modelos criados, através de ferramentas lineares e não-lineares, têm a capacidade de prever a tendência do valor da densidade de cianobactérias ao longo do tempo. No entanto, não conseguem mimetizar com grande rigor valores de densidade mais elevados. Assim, embora a estratégia utilizada tenha produzido resultados práticos é de esperar que aumentando a frequência de amostragem das variáveis identificadas como mais relevantes (temperatura do ar, temperatura da água, pH, condutividade, direção do vento, velocidade do vento, cota, azoto amoniacal, dureza, turvação, precipitação e radiação) seja possível melhorar a precisão e exatidão das previsões. É ainda expectável que a relevância aqui atribuída a cada uma das variáveis se altere com o aumento do número de pontos com informação.

Ao comparar o desempenho dos diferentes modelos verificou-se que o modelo gerado por uma rede neuronal do tipo *feedforward* apresentava melhor desempenho, com um menor erro quadrático médio e sem a previsão de falsos negativos ou positivos para o conjunto de dados utilizado para avaliação do modelo.

O objetivo proposto foi atingido, com a criação de um sistema de alerta, considerando um limiar de 20 000 células/mL, que permite à AgdA prever a ocorrência de *blooms* de cianobactérias com a antecedência de 15 dias, para ativar protocolos de atuação definidos para lidar com este tipo de incidente, acrescentando valor aos dados que a entidade já gera de forma rotineira. Uma vez que, na metodologia adotada, não foram utilizados parâmetros específicos da Albufeira do Roxo, poderá explorar-se a possibilidade de replicar esta ferramenta noutras Albufeiras.

5.2 Perspetivas futuras

A construção de um modelo que melhor descreva a realidade depende, profundamente, da qualidade dos dados utilizados. Por este motivo, sugere-se a criação de modelos com base em parâmetros identificados neste estudo com maior relevância (temperatura do ar, temperatura da água, pH, condutividade, direção do vento, velocidade do vento, cota, azoto amoniacal, dureza, turvação, precipitação e radiação), amostrados com maior frequência e em intervalos regulares. A periodicidade deste reforço analítico terá de ser definida pela entidade gestora, uma vez que o custo de amostragem e determinação é insignificante face ao valor acrescentado de uma correta previsão dos valores de contaminação.

E, como neste trabalho um dos critérios de avaliação da qualidade dos modelos passou pela análise do número de previsões que resultaram em falsos negativos ou positivos, seria interessante utilizar como critério de paragem do treino das RNA a taxa de falsos negativos e positivos, em vez do erro quadrático médio.

Sabendo-se que a ocorrência de *blooms* de cianobactérias é um fenómeno imprevisível, e que ocorre a nível nacional, esta ferramenta pode ser um contributo para o trabalho de Gestão de Recursos Hídricos realizada pelas autoridades competentes, em parceria com as entidades gestoras, como ferramenta de suporte à decisão e compreensão do comportamento dos *blooms* de cianobactérias nas Albufeiras.

Bibliografia

[1] Sanz-Alférez S. Distribución de cianobacterias y cianotoxinas. In: Livro de Resumos do 4º Congresso Ibérico de Cianotoxinas. VI Reunião da Rede Ibérica de Cianotoxinas; 08-10 Julho de 2015, Lisboa, Portugal. Instituto Ricardo Jorge; 2015. p. 12.

[2] Moreira C, Mendes R, Matos A, Vasconcelos V, Antunes A. Monitorização de cianotoxinas em águas doces portuguesas: primeira detecção de cilindrospermopsina, anatoxina-a e saxitoxinas. In: Livro de Resumos do 4º Congresso Ibérico de Cianotoxinas. VI Reunião da Rede Ibérica de Cianotoxinas; 08-10 Julho de 2015, Lisboa, Portugal. Instituto Ricardo Jorge; 2015. p. 13.

[3] Llewellyn C. Predicting cyanobacteria blooms in 50 lakes of Northwest Washington [Tese de doutoramento]. Western Washington University; 2010.

[4] Salvador D. Avaliação da expressão de genes envolvidos na síntese de microcistinas em cianobactérias tóxicas sujeitas a diferentes intensidades de luz [Tese de mestrado]. Universidade de Lisboa - Faculdade de Ciências; 2014.

[5] Torres R, Pereira E, Vasconcelos V, Teles LO. Forecasting of cyanobacterial density in Torrao reservoir using artificial neural networks. J Environ Monit. 2011;13(6):1761–1767. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21552584.

[6] Chorus I, Bartram J. Toxic Cyanobacteria in Water: A guide to their public health consequences, monitoring and management. World Health Organization; 1999.

[7] Newcombe G, Coalition GWR. International Guidance Manual for the Management of Toxic Cyanobacteria. Global Water Research Coalition; 2009.

[8] Caeiro JMF. Characterization of cyanobacteria in waters from Portuguese dams. In: Livro de Resumos do 4º Congresso Ibérico de Cianotoxinas. VI Reunião da Rede Ibérica de Cianotoxinas; 08-10 Julho de 2015, Lisboa, Portugal. Instituto Ricardo Jorge; 2015. p. 22.

[9] Berg M, Sutula M. Factors affecting the growth of cyanobacteria with special emphasis on the Sacramento-San Joaquin Delta. Applied Marine Sciences and Southern California Coastal Water Research Project; 2015. Relatório Nº: 869.

[10] WHO. Guidelines for Drinking-water Quality, Fourth Edition. Geneva; 2011.

[11] Dias E, Paulino S, Pereira P. Cyanotoxins: from poisoning to healing - a possible pathway? Limnetica. 2015;34(1):159–172.

[12] Galvão HM, Reis MP, Valério E, Domingues RB, Costa C, Lourenço D, et al. Cyanobacterial blooms in natural waters in southern Portugal: a water management perspective. Aquatic Microbial Ecology. 2008;53:129–140.

[13] Ribeiro R, Torgo L. A comparative study on predicting algae blooms in Douro River, Portugal. Ecological Modelling. 2008;212(1-2):86–91.

[14] Bellém F, Nunes S, Morais M, Fonseca R. Cianobactérias e Toxicidade: Impacte na Saúde Pública em Portugal e no Brasil. Revista Brasileira de Geografia Física. 2012;6:1374–1389.

[15] Norwegian Institute for Water Research. Harmful algal blooms: implications for human health and economic valuation. Oslo: NIVA; 2004. Relatório Nº 4836-2004.

[16] De Figueiredo DR, Reboleira ASSP, Antunes SC, Abrantes N, Azeiteiro U, Gonçalves F, et al. The effect of environmental parameters and cyanobacterial blooms on phytoplankton dynamics of a Portuguese temperate lake. Hydrobiologia. 2006;568(1):145–157.

[17] Monteiro J. Eutrofização. Instituto Superior Técnico; 2004.

[18] Carapeto C. Poluição das águas. Lisboa: Universidade Aberta; 1999.

[19] Vasconcelos V. Toxicologia de cianobactérias - Distribuição de cianobactérias tóxicas e suas toxinas em águas doces portuguesas. Bioacumulação em bivalves [Tese de doutoramento]. Faculdade de Ciências da Universidade do Porto; 1995.

[20] Vasconcelos V. Cyanobacteria Toxins: Diversity and Ecological Effects. Limnetica. 2001;20:45–58.

[21] Vasconcelos V. In: Toxic cyanobacteria in the Mondego basin reservoirs: an overiew. Coimbra: Imprensa da Universidade de Coimbra; 2002. p. 105–114. Available from: `https://digitalis.uc.pt/handle/10316.2/32659`.

[22] Vasconcelos VM. Cyanobacterial toxins in Portugal: effects on aquatic animals and risk for human health. Brazilian Journal of Medical & Biological Research. 1999;32(3):249–254.

[23] Santos CR, Santana FP, Rodrigues AF. Estudo da comunidade de cianobactérias nas lagoas das Sete-Cidades e Furnas (S. Miguel - Açores). Pesquisa de cianotoxinas. In: 6º Congresso da Água. Associação Portuguesa dos Recurso Hídricos; 2002. p. 54–55.

[24] Churro C, Dias E, Paulino S, Alverca E, Pereira P. Importância da monitorização de cianobactérias em albufeiras portuguesas. Boletim Epidemiológico Observações. 2013;4:18–20.

[25] Science European Commission. Link found between ‘ algal blooms ’ and liver disease Assunto 426. European Commission DG Environment News Alert Service. 2015;p. 1.

[26] Merel S, Walker D, Chicana R, Snyder S, Baurès E, Thomas O. State of knowlegde and concerns on cyanobacterial blooms and cyanotoxins. Environment International. 2013;59:303–327.

[27] Gregorio FND. New criteria and methods for cyanobacteria risk assessment and risk management in water for human consumption [Tese de doutoramento]. Università degli Studi di Roma "La Sapienza"; 2014.

[28] Wang H, Yan X, Chen H, Chen C, Guo M. Chlorophyll-a Predicting Model Based on Dynamic Neural Network. Applied Artificial Intelligence. 2015;29:962–978.

[29] Yabunaka KI, Hosomi M, Murakami A. Novel application of a back-propagation artificial neural network model formulated to predict algal bloom. Water Science and Technology. 1997;36(5):89–97. Available from: http://dx.doi.org/10.1016/S0273-1223(97)00464-2.

[30] Maier HR, Dandy GC, Burch MD. Use of artificial neural networks for modelling cyanobacteria *Anabaena* spp. in the River Murray, South Australia. Environmental Modelling Software. 1998;105:257–272.

[31] Wilson H, Recknagel F. Towards a generic artificial neural network model for dynamic predictions of algal abundance in freshwater lakes. Ecological Modelling. 2001;146:69 – 84.

[32] Wilmotte A, Descy JP, Vyverman W. Algal blooms: emerging problem for health and sustainable use of surface waters. Final report BELSPO-project, Brussels. 2008;EV34. Available from: http://orbi.ulg.ac.be/handle/2268/95894.

[33] Oh HM, Ahn CY, Lee JW, Chon TS, Choi KH, Park YS. Community patterning and identification of predominant factors in algal bloom in Daechung Reservoir (Korea) using artificial neural networks. Ecological Modelling. 2007;203(1-2):109–118.

[34] Muttil N, Chau KW. Neural network and genetic programming for modelling coastal algal blooms. International Journal of Environment and Pollution. 2006;28(3/4):223–238.

[35] YooKyung C, Park SS, Kim K, Byeon M, Stow CA. Probabilistic prediction of cyanobacteria abundance in a Korean reservoir using a Bayesian Poisson model. Water Resources Research. 2014;50:2518–2532.

[36] Ibelings BW, Vonk M, Los HFJ, van der Molen DT, Mooij WM. Fuzzy modeling of cyanobacterial surface waterbooms: validation with NOAA-AVHRR satellite images. Ecological Applications. 2003;13(5):1456–1472.

[37] Lilover MJ, Laanemets J. A simple tool for the early prediction of the cyanobacteria Nodularia spumigena bloom biomass in the Gulf of Finland. Oceanologia. 2006;48(SUPPL.):213–229.

[38] Teles LO, Vasconcelos V, Pereira E, Saker M. Time series forecasting of cyanobacteria blooms in the Crestuma Reservoir (Douro River, Portugal) using artificial neural networks. Environ Manage. 2006;38(2):227–237. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16788859.

[39] Agência Portuguesa do Ambiente. Planos de Gestão de Região Hidrográfica. Lisboa: APA; 2012. Available from: http://www.apambiente.pt/index.php?ref=16&subref=7&sub2ref=9&sub3ref=848.

[40] SNIRH. Sistema Nacional de Informação de Recursos Hídricos;. [Internet]. Available from: http://snirh.pt/index.php?idMain=1&idItem=7&albufcode=45 [cited 20 de março de 2016].

[41] Simões J, Barroso S, Roque MA, Marin F, Capitão G, Ferreira IC, et al. Plano de Ordenamento da Albufeira do Roxo. Relatório Síntese. CEDRU/AIA; 2008.

[42] Agência Portuguesa do Ambiente. Inventário Nacional de Sistemas de Abastecimento de Água e de Águas Residuais; 2016. [Internet]. Available from: http://insaar.apambiente.pt/ [cited 15 de julho de 2016].

[43] EPA. Impacts of Climate Change on the Occurrence of Harmful Algal Blooms; 2013. [Internet]. Available from: https://www.epa.gov/nutrientpollution/factsheet-climate-change-and-harmful-algal-blooms [cited 11 de agosto de 2016].

[44] Campinas M, Teixeira M, Lucas H. Previsão da capacidade de remoção de cianobactérias e cianotoxinas na ETA de Alcantarilha. In: 10º Encontro Nacional de Saneamento Básico (ENaSB); 2002. p. 16–19. Available from: http://w3.ualg.pt/{~}mribau/Textos/Previsaocapacidade.pdf.

[45] Bowden G, Dandy G, Maier H. 7. In: Forecasting cyanobacteria (blue-green algae) using artificial neural networks. American Society of Civil Engineers Press; 2005. p. 71–96.

[46] AgdA. Sítio da Águas Públicas do Alentejo; 2015. [Internet]. Available from: http://www.agda.pt/Noticias/nota-informativa.html [cited 24 de março de 2016].

[47] Águas de Portugal. Manual para o Desenvolvimento de Planos de Segurança da Água. Águas de Portugal; 2011.

[48] PennState Eberly College of Science. STAT200;. [Internet]. Available from: https://onlinecourses.science.psu.edu/stat200/node/67 [cited 20 agosto de 2016].

[49] Rutherford A. Introduction ANOVA and ANCOVA a GLM Approach. Daniel B Wright UoS, editor. SAGE Publications Ltd; 2001.

[50] Abdi H, Williams LJ. Turkey's Honestly Significant Difference (HSD) Test. In: Salkind N, editor. Encyclopedia of Research Design. Thonusand Oaks, CA: Sage; 2010. p. 1–5.

[51] Wold S, Esbensen K, Geladi P. Principal Component Analysis. Chemometrics and Intelligent Laboratory Systems. 1987;2:37–52.

[52] Abdi H, Williams LJ. Principal component analysis. WIREs Computational Statistics. 2010;2:433–459.

[53] Böhm K, Smidt E, Tintner J. 2. In: de Freitas LV, de Freitas APBR, editors. Application of Multivariate Data Analyses in Waste Management. InTech; 2013. p. 15–38.

[54] Abdi H. Partial Least Squares (PLS) Regression. In: M LB, Bryman A, T F, editors. Encyclopedia of Social Sciences Research Methods. Thousand Oaks (CA): Sage; 2003. p. 1–7.

[55] StatSoft. Statistica Documentation;. [Internet]. Available from: `http://documentation.statsoft.com/STATISTICAHelp.aspx?path=MSPC/PCAandPLSTechnicalDetails` [cited 29 de julho de 2016].

[56] Haykin S. Neural Networks and Learning Machines. 3rd ed. New York: Pearson Prentice Hall; 2009.

[57] Hagan MT, Demuth HB, Beale MH, De Jesús O. Neural Network Design (2nd Edition). Martin Hagan; 2014. Available from: `https://books.google.pt/books?id=4EW9oQEACAAJ`.

[58] Beale MH, Hagan MH, Demuth HB. Neural Network Toolbox Users Guide. The MathWorks, Inc.; 2016.

[59] Oliveira E. Aprendizagem Automática: Redes Neuronais Computacionais; 2008. Online;.

[60] Beale MH, Hagan MH, Demuth HB. Neural Network Toolbox Getting Started Guide. The MathWorks, Inc.; 2016. Available from: `http://www.mathworks.com/help/nnet/gs/neural-networks-overview.html`.

[61] Martins FG. Simulação e Controlo de Processos Químicos [Tese Doutoramento]. Faculdade de Engenharia da Universidade do Porto; 1997.

[62] Soares A. Geoestatística para as ciências da terra e do ambiente. IST; 2014.

[63] Carneiro RL. Ecofisiologia de *cylindrospermopsis raciborskii* (cyanobacteria): Influências da intensidade e qualidade da luz e da dureza da água sobre o crescimento e a produção de saxitoxinas [Tese Doutoramento]. Universidade Federal do Rio de Janeiro; 2009.

[64] Yang J, Lv H, Yang J, Liu L, Yu X, Chen H. Decline in water level boosts cyanobacteria dominance in subtropical reservoirs. Science of The Total Environment. 2016;557–558:445 – 452. Available from: `http://www.sciencedirect.com/science/article/pii/S0048969716305186`.

[65] Kuo JT, Hsieh MH, Lung WS, She N. Using artificial neural network for reservoir eutrophication prediction. Ecological Modelling. 2007;200:171–177.

[66] Hornik K, Stinchcombe M, White H. Multilayer Feedforward Networks are Universal Approximators. Neural Networks. 1989;2:359–336.

[67] Hamby DM. A Review of techniques for parameter sensitivity analysis of environmental models. Environmental Monitoring and Assessment. 1994;32:135–154.