A comparative study on predicting algae blooms in Douro River, Portugal

Rita Ribeiro\textsuperscript{a,b,\ast}, Lu\'is Torgo\textsuperscript{a,c}

\textsuperscript{a} LIAAD-INESC Porto L.A., R. Ceuta, 118-6o., 4050-190 Porto, Portugal
\textsuperscript{b} FCUP-DCC, University of Porto, R. Campo Alegre, 1021/1035, 4169-007 Porto, Portugal
\textsuperscript{c} FEP, University of Porto, R. Dr. Roberto Frias, 4200-464 Porto, Portugal

\textbf{A R T I C L E I N F O}

Article history:
Published on line 19 November 2007

Keywords:
Multiple regression
Extreme values prediction
Model evaluation
Algae blooms

\textbf{A B S T R A C T}

Algae blooms are ecological events associated with extremely high abundance value of certain algae. These rare events have a strong impact in the river’s ecosystem. In this context, the prediction of such events is of special importance. This paper addresses the problems that result from evaluating and comparing models at the prediction of rare extreme values using standard evaluation statistics. In this context, we describe a new evaluation statistic that we have proposed in Torgo and Ribeiro [Torgo, L., Ribeiro, R., 2006. Predicting rare extreme values. In: Ng, W., Kitsuregawa, M., Li, J., Chang, K. (Eds.), Proceedings of the 10th Pacific-Asia Conference on Knowledge Discover and Data Mining (PAKDD’2006). Springer, pp. 816–820 (number 3918 in LNAI)], which can be used to identify the best models for predicting algae blooms. We apply this new statistic in a comparative study involving several models for predicting the abundance of different groups of phytoplankton in water samples collected in Douro River, Porto, Portugal. Results show that the proposed statistic identifies a variant of a Support Vector Machine as outperforming the other models that were tried in the prediction of algae blooms.

\textcopyright 2007 Elsevier B.V. All rights reserved.

1. Introduction

In Douro River, as in other European rivers, an excessive growth of some algae is registered occasionally. These phenomena, known as algae blooms, degrades water quality as it reduces water clarity and the oxygen levels, causing a massive death of river fish. The early prediction of these blooms is particularly important, especially when potable water is obtained from these rivers, and several authors have addressed this task (e.g. Bobbin and Recknagel, 2001; Ibanez and Conversi, 2002; Lee et al., 2003; Muttil and Lee, 2005).

The data used in our comparative study were obtained from the public company responsible for potable water collection for the metropolitan Porto area, the second largest city of Portugal. We have used the data resulting from the analysis of water samples collected at Crestuma-Lever dam in river Douro between 1998 and 2003. The data include the identification and quantification of different groups of phytoplankton in water samples collected in Douro River, Porto, Portugal. The long term objective of the project behind this paper is to develop reliable models that can forecast micro-
algae blooms in river Douro. Identification of the different groups of phytoplankton in water samples requires intensive and expensive manual labour, while the analysis of most physical–chemical and microbiological parameters can be automated by water probes. Our objective is to avoid the expensive manual identification of phytoplankton groups by using models that are able to accurately anticipate their abundance based on the values of other parameters.

In this paper we describe the results of a comparative experiment between several candidate models. Predicting the abundance level of a certain phytoplankton species, when given the values of a set of other parameters, can be regarded as a multiple regression problem. Several techniques exist for handling this type of problems. They all revolve around the issue of obtaining the model parameters that optimize a certain preference criterion that is estimated on a data sample. The standard error statistics used as preference criteria will tend to favour the model that has good performance on the most frequent cases. However, micro-algae blooms are fortunately rare events that are characterized by extreme values of abundance of those algae. As such, we claim that the use of such preference criteria will introduce a bias into the obtained models that goes against the goals of the application, i.e. predicting rare extreme values of the abundance of micro-algae. Based on these observations, we have developed a different error statistic (Torgo and Ribeiro, 2006) that is able to identify the models that are more “useful” from the perspective of accurately predicting the rare extreme values of the target variable. In our comparative experiments we have used this statistic to compare the models we have tried in our data.

### 2. Problem description

The data used in this study was collected at Crestuma-Lever dam in river Douro and covers the period from 1998 till 2003. As our objective is to predict the algae blooms, we have selected the information concerning certain groups of phytoplankton, physical–chemical and microbiological parameters.

Unfortunately, though this information respects to water samples collected from the same place, their periodicity is far from being homogeneous. This required large pre-processing steps to come up with a data set useful for model construction.

We decided to gather all the data by synchronizing it to the periodicity of the information concerning algae abundance. From all the identified groups of micro-algae, we selected the ones for which we have information along all 6 years: Cyanobacteria, Cryptophyta, Euglenophyta, Chrysophyta, Dinophyta, Chlorophyta and Diatom. The sampling periodicity for these algae was initially weekly, but became biweekly since 2002, as the graph of Fig. 1 shows for the Cyanobacteria case. Therefore, this last biweekly periodicity was established as the periodicity for our data set.

In the original data, the values of physical–chemical and microbiological parameters had higher sampling frequency, so they had to be aggregated. For each parameter, we included the minimum, the median and the maximum value for the 2 weeks respecting each observation in our dataset. We left out from our analysis some physical–chemical and microbiological parameters based on their scarce number of measurements. Still, for the selected parameters, there were some unknown values. We have used an imputation strategy.

### Table 1 - Physical–chemical and microbiological parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Unit</th>
<th>Parameter</th>
<th>Unit</th>
<th>Parameter</th>
<th>Unit</th>
<th>Parameter</th>
<th>Unit</th>
<th>Parameter</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turbidity</td>
<td></td>
<td>Temperature</td>
<td></td>
<td>pH</td>
<td></td>
<td>Alkalinity</td>
<td></td>
<td>Conductivity</td>
<td></td>
</tr>
<tr>
<td>Nitrate</td>
<td></td>
<td>Chlorates</td>
<td></td>
<td>Sulfates</td>
<td></td>
<td>Silica</td>
<td></td>
<td>Oxidability</td>
<td></td>
</tr>
<tr>
<td>Dissolved oxygen</td>
<td></td>
<td></td>
<td></td>
<td>Dissolved iron</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fecal coliforms</td>
<td></td>
<td>Total coliforms</td>
<td></td>
<td>Fecal streptococcus</td>
<td></td>
<td>Sulfite reducer clostridia</td>
<td></td>
<td>Total germs at 22°C</td>
<td></td>
</tr>
<tr>
<td>Total germs at 37°C</td>
<td></td>
<td>Escherichia coli</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
to replace them, namely we used the exponential moving average based on the last 4 observations, which corresponds to, approximately, 2 months. Table 1 summarises the final set of parameters that were used in our study.

In order to capture the temporal oscillations and the existing mutual influence among the algae abundance values (e.g. Lee et al., 2003), we also included, for each observation, the values of the previous 2 weeks of each group of phytoplankton and their total value and diversity index. By including this last information we intended to facilitate the identification of algae blooms that we suppose to be characterized by extremely high total value and a low value of diversity, meaning that a few groups of phytoplankton dominate. For the diversity index we chose to use the Shannon index (e.g. Hill, 1973).

As a result of these pre-processing steps we get to a dataset with 72 variables, 7 target variables (one for each selected group of phytoplankton) and 131 cases. The low number of cases arises from the lack of information about the groups of phytoplankton during the first semester of 1999 and 2002, as we can see in Fig. 1.

3. Multiple regression problem

Our overall objective is to obtain models to predict the values of abundance of each group of phytoplankton. This is a multiple regression problem as we want to estimate the values of abundance of each group of phytoplankton. This is a classification problem (e.g. Provost et al., 1998; Weiss and Provost, 2003a, b). In fact, the classes associated with bloom values will have a very low frequency and thus to make a model more accurate on these classes (that is more interest-

We argue against the use of these error statistics when the goal is to predict accurately a small proportion of values, like the rare extreme values, mainly because they treat equally all the errors made across the target variable range. According to our objective, we should prefer a model which performs especially well at the rare extreme values of the target variable. In this context, more emphasis should be given to the errors made at these values, disregarding the errors made at the “normal” values of the target variable. This is not the case with MSE and MAD. They favour the models that achieve a smaller average error, independently of the true values associated to each error. Given the distribution of the values of target variable, this best score is obtained by models performing well on the most frequent cases. Unfortunately, these are not the relevant cases for our application as blooms are rare.

4. Rare extremes error

According to our objective, the errors made in different ranges of the target variable should not be treated equally. Namely, the errors made at rare extreme values should have more weight than the rest of the errors. A possible approach would be to weight the data set, with larger weights being given to cases with extreme values in the target. However, this would only partially meet our evaluation requirements. When analysing the severity of the errors made by the model, it is not enough to look at the extremeness of the true value and disregard the extremeness of the predicted values. In fact, if a model predicts a bloom for a case with a “normal” value of occurrence of a micro-algal we have a false alarm. False alarms should also be avoided, as they can trigger unnecessary preventive actions. As such, these situations should also have more weight. This means that the weight function should depend on both the true and predicted values. In other words the weight function should penalise both false alarms and missed blooms.

The requirements of our weight function are strongly related to the notion of cost sensitive evaluation (e.g. Turney, 2000) used in classification domains, where the target variable is discrete. In some classification problems we known that making an error in a particular class is more serious than making an error in some other class. A frequently used form of addressing this problem is through a cost matrix which defines all the costs for every combination of predicted and true classes. A possible way of addressing our prediction problem would be to transform it into a classification problem (e.g. Torgo and Gama, 1997) and use one of such cost-sensitive approaches. In particular, we would have to properly select the cut-points on the range of each the micro-algal abundance values, so that we could distinguish a non-bloom abundance value from a bloom abundance value in order to reflect our application objectives. This approach was followed for instance by Bobbin and Recknagel (2001). The resulting classification problem will inevitably lead to an unbalanced classification problem (e.g. Provost et al., 1998; Weiss and Provost, 2003a, b).
ing for our application) would involve using evaluation criteria different from standard prediction accuracy.

We claim that the use of classification approaches in this problem leads to counter-intuitive situations as a result of the necessary discretization process of the abundance values. Suppose that this process uses the cut-point 60 for determining a bloom in some micro-algal. In this context a water sample with a value of 58 would be considered a non-bloom case and two water samples with value 61 or 278 would be regarded as bloom cases. These means that the use of a crisp threshold creates an artificial differentiation between situations that clearly seem more related (the 58 and 61 samples), and considers similar samples that are clearly of different seriousness (the 61 and 278 cases). We think this is clearly counter-intuitive. A possible solution could be to further split the bloom class into more classes to allow and increased differentiation, but that would turn an already unbalanced classification problem into an even harder task.

In this context, we propose the Rare Extremes Error (RExE) statistic presented in Torgo and Ribeiro (2006) and defined by Eq. (3).

\[
\text{RExE} = \frac{1}{n} \sum_{i=1}^{n} w(y_i, \hat{y}_i) \ell(y_i, \hat{y}_i) \tag{3}
\]

where \( w \) is the weight function and \( \ell \) is a loss function that can be some error measure (e.g. the squared error or the absolute deviation).

Inspired by the idea of cost matrix, we proposed a cost surface, \( w(Y, \hat{Y}) \), that can be seen as a continuous version of a cost matrix.

As we are dealing with a regression problem, we are interested in having a continuous cost function that varies smoothly along the range of the target variable and that depends on both predicted and true values. We want this cost function to reflect our preference bias concerning the errors made at rare extreme values.

In order to facilitate the specification of the continuous cost surface, we suggest taking some selected points for which we specify a cost and then use a function approximation method to interpolate the complete surface. In our concrete application we want differentiate the value of the cost surface between extreme and “normal” values.

The problem of appropriately specifying a threshold for a bloom is not trivial (Ibanez and Conversi, 2002). In our application, as we did not have domain knowledge for setting the threshold value defining algae blooms, we have used the distribution properties of the variables. We established the upper adjacent value (adj\(_H\)), as the limit above which the value of abundance of a micro-algal is considered a bloom. This threshold represents the largest value of a micro-algal abundance which is not above the 3rd quartile plus 1.5 of the inter-quartile range. We then specify a set of equally spaced interpolation points between the minimum value and the adj\(_H\). The number of points specified establishes the granularity with which we wish to specify our cost surface.

The next step is to find the costs at these selected points to fill in the cost matrix. These can either be specified by domain experts or, when that sort of knowledge in not available, by the method described in Torgo and Ribeiro (2006). In our study we have followed this latter alternative. We begin by setting to 1, the costs associated to areas of cost surface where no rare extreme values are involved. Then we set the lowest cost of the matrix with a value lying between 0 and 1. This is the cost associated to correct predictions involving extreme values. The rest of the costs were filled in using an arithmetic progression, according to the magnitude of the error made whenever rare extreme values are involved. Lets see an example with Cyanobacteria, where the adj\(_H\) is equal to 204.1 cells/mL. Suppose we set an arithmetic progression of base 6 and ratio 4, and the lower cost to be 0.1. This would result on the cost matrix shown in Table 2.

Once we have the cost matrix completely defined, we can use a function approximation method to obtain a smooth cost surface using the selected points of the cost matrix. We have used the same procedure as in Torgo and Ribeiro (2006), where the loess function implemented in the R statistical software (R Development Core Team, 2005), was used to obtain the surface. This function fits a local polynomial of degree 1 taking the points defined in the cost matrix as samples of points of the cost surface we want to approximate. The result is a smooth cost surface that satisfies the requirement of predicting rare extreme values, as shown in Fig. 2 for the Cyanobacteria case.

\[
\text{Fig. 2 – The cost surface obtained for Cyanobacteria.}
\]
5. Experimental comparisons between different predictive models

The goal of our experiments is to compare different models for predicting the value of abundance of each group of phytoplankton and identify the ones that have the best performance at forecasting micro-algae blooms.

We have compared three different models: regression trees (Breiman et al., 1984), neural networks (Rumelhart et al., 1986) and support vector machines (Vapnik, 1996). This set of models was chosen because they have quite different approaches to regression tasks, thus ensuring some diversity in our analysis. In our experiments, we have used the implementations of these techniques available in R (R Development Core Team, 2005) through the packages: rpart (Therneau and Atkinson, 2006), nnet (Venables and Ripley, 2002) and e1071 (Dimitriadou et al., 2006) which contains an implementation of svm.

Regarding our experimental methodology, we decided to leave out the last 6 months of data for the final testing of the models. We used the remaining 4.5 years of data to estimate the performance of the three different modelling techniques through a Monte Carlo simulation. For each iteration of the Monte Carlo simulation we have randomly selected a date within this 4.5 years period ensuring that there were 2 previous years to be used for training, i.e. 48 biweekly observations, and 9 following months, i.e. 18 biweekly observations, to be used for testing. We have repeated this process 10 times. For each repetition, the predictions made by each model for the 9 months test period were obtained by a sliding-window technique, that moves every two new observations, that is, approximately every month.

The performance of the models was asserted using a normalized version of our rare extremes error statistic (cf. Eq. (4)) to allow better comparisons across the different micro-algae.

\[
\text{NRExE} = \frac{\sum_{i=1}^{n_{\text{test}}} w(y_i, \hat{y}_i) \times |y_i - \hat{y}_i|}{\sum_{i=1}^{n_{\text{test}}} w(y_i, \hat{y}) \times |y_i - \hat{y}|}
\]  

(4)

where \(\hat{y}\) is the training sample median.

In our experiments, we have defined the weighting function \(w\) based on a 4 x 4 cost matrix filled in with costs that follow an arithmetic progression of base 6 and ratio 4 and that establishes 0.1 as the minimum cost. This is a scenario similar to that shown in Table 2, but with different interpolation points according to each micro-algal abundance values.

We have also evaluated the performance of the models using a “standard” error statistic, namely the Normalized Mean Absolute Deviation (NMAD) (c.f. Eq. (5)), to check whether the conclusions are different when using different evaluation metrics.

\[
\text{NMAD} = \frac{\sum_{i=1}^{n_{\text{test}}} |y_i - \hat{y}_i|}{\sum_{i=1}^{n_{\text{test}}} |y_i - \hat{y}|}
\]  

(5)

6. Discussion of the obtained results

We have used the Monte Carlo method to estimate the performance of each modelling technique according to the two error statistics referred above: NMAD and NRExE. For all modelling techniques, we calculated the mean and standard deviation of each statistic in the 10 repetitions. Table 3 shows these results for each group of phytoplankton.

We have asserted the statistical significance of the differences in each group of phytoplankton. For this purpose, we have used the Wilcoxon signed rank test. For each error statistic we have asserted the significance of the differences of the results obtained by nnet and svm against the results of rpart.

We have denoted the results using: (+), (++), (−) and (−−). The positive signs indicate that the difference is positive, i.e. rpart had a smaller error and thus a better performance, while negative signs indicate the opposite. The number of signs reflects the confidence level of the observed difference: one sign represents a 95% confidence level; two signs represent a 99.9% confidence level. The results of these comparisons for each group of phytoplankton are also shown in Table 3, besides the error estimates.

Globally, we can say that the methods that were tried did not achieve good performance. Namely, there are some cases...
where the error estimate is higher than 1, meaning that the respective model is performing worst than simply predicting the median. These high error estimates were somehow expected as the models are biased to the prediction of the most frequent values and, as such, have difficulty in predicting accurately the rare extreme values. Nevertheless, we should remark that we did not explore different parameterizations of these models. We could probably improve the results by carefully adjusting their parameters.

From the observation of Table 3, we can also notice that, according to NMAD, \textit{svm} outperforms the other two techniques for the prediction of values of abundance of almost all the micro-algae. Moreover, if we look at NRExE, that gives more importance to errors on extreme values (blooms), they still have the best performance.

The significance tests carried out reinforce these conclusions. According to NRExE, the \textit{svm} models have not only better error estimates, but also they have, for some groups of phytoplankton, a significantly better performance in terms of rare extreme values prediction. It is interesting to observe that some performance differences that were insignificant from the perspective of the standard evaluation statistic become significant when using our evaluation statistic and vice versa. For instance, in the Cyanobacteria case, the error estimate of \textit{svm} not only decreased according to NRExE, but also became significantly better than the other two techniques. On the contrary, for the Dinophyta case, the apparent good performance of \textit{svm} revealed to be insignificant in the ranking performances of the models for the prediction of the blooms. This is relevant information given the main objectives of this application and stresses the dangers of using an evaluation metric that is not adjusted to these goals.

In summary, we have observed that, according to our proposed evaluation metric, \textit{svm} generally outperforms the other two techniques in terms of predicting rare extreme values of occurrence of almost all the considered groups of phytoplankton.

### 7. Conclusions and future work

This paper presents a comparative study of different prediction models on the difficult task of predicting algae blooms using physical–chemical and microbiological parameters of water samples in river Douro, Portugal. Our study confirms the difficulty of the problem caused mainly by the scarcity of the phenomenon we want to predict accurately. This type of rare events requires specific evaluation criteria. We have used such type of metrics to evaluate the performance of several models. Our results indicate a variant of support vector machines as the most promising model.

Future work will include some adjustments to this evaluation measure so as to make it more interpretable for domain experts in terms of the effectiveness of the models in the identification of the algae blooms. As a long-term objective, we also plan to include the proposed evaluation metric into a modelling technique to better tune the model parameters according to the objectives of this type of applications.

### Acknowledgements

This work was partially supported by FCT project MODAL (POSI/40949/2001) co-financed by POSI and by the European fund FEDER, and by a PhD scholarship of the Portuguese government (SFRH/BD/1711/2004) to R. Ribeiro. We would like to thank Aguas do Douro e Paiva, SA, for providing the data used in this study.

### References


