SCIENTIFIC REPORTS

OPEN

A model predicting the PSP toxic dinoflagellate *Alexandrium minutum* occurrence in the coastal waters of the NW Adriatic Sea

Received: 6 September 2018
Accepted: 15 February 2019
Published online: 12 March 2019

Eleonora Valbi[1,3], Fabio Ricci[1,3], Samuela Capellacci[1,3], Silvia Casabianca[1,3], Michele Scardi[2,3] & Antonella Penna[1,3,4]

Increased anthropic pressure on the coastal zones of the Mediterranean Sea caused an enrichment in nutrients, promoting microalgal proliferation. Among those organisms, some species, such as the dinoflagellate *Alexandrium minutum*, can produce neurotoxins. Toxic blooms can cause serious impacts to human health, marine environment and economic maritime activities at coastal sites. A mathematical model predicting the presence of *A. minutum* in coastal waters of the NW Adriatic Sea was developed using a Random Forest (RF), which is a Machine Learning technique, trained with molecular data of *A. minutum* occurrence obtained by molecular PCR assay. The model is able to correctly predict more than 80% of the instances in the test data set. Our results showed that predictive models may play a useful role in the study of Harmful Algal Blooms (HAB).

Anthropic pressures, highly increased in recent decades, have strong impact along the coasts of the Mediterranean Sea. Among the consequences, there are eutrophication, a nutrient over-enrichment of coastal waters (especially due to the massive use of fertilizers in agriculture), transport of phytoplankton species via ballast-water vessels and translocation of shellfish stocks[1–4]. In particular, eutrophication is increasing due to increased population, increased use of fertilizers both for terrestrial and marine animal farm practices and increased fossil fuel use[5]. These phenomena can favor a fast proliferation of microalgal species, known as algal bloom[6,7]. Further, climate change seems having effects on the frequency and abundance of algal blooms due to the complex of altered environmental factors[8,9].

Some microalgal taxa, such as dinoflagellates, can both originate high density biomass proliferation or blooms and produce a variety of toxin compounds that can accumulate along the trophic web through biomagnification process. Such blooms are known as Harmful Algal Blooms (HABs) and they can cause very serious damages to human health and marine organisms[10]. People can be affected either by breathing aerosols[11–13] or by eating vector species, such as mussels, clams and oysters[14,15], which can accumulate high concentrations of toxins in their digestive glands. HABs can cause also fish kills or hypoxia or anoxia events due to algal biomass proliferation. Therefore, HABs phenomena, in addition to human health, are also concerned with fishing and aquaculture industry[16–18].

In recent years, there has been a significant increase of these HABs phenomena worldwide[19–22], including Mediterranean Sea[23,24]. Therefore, the HAB monitoring programs increased[4]. In the future, the next challenge will be the managing and forecasting of HABs[25]. The mathematical models are shown to be useful tools for this purpose and their use has grown in the last decades. The purpose of these models is to describe[26–29] or to forecast HABs providing a survey[30–32], in order to identify environmental, physical and chemical conditions in which the risk of algal blooms is higher and in which it can concentrate efforts, such as sampling frequency to confirm or discharge the predicted bloom. Methods used to build these models are numerical, mathematical, and statistical ones or artificial intelligence techniques, like Artificial Neural Network (ANN)[33,34] and other Machine Learning

[1]Department of Biomolecular Sciences, University of Urbino, Campus E. Mattei, Via Cà le Suore 2/4, 61029, Urbino (PU), Italy. [2]Department of Biology, University of Rome Tor Vergata, Via della Ricerca Scientifica 1, 00133, Rome, Italy. [3]CoNISMa, Consorzio Interuniversitario per le Scienze del Mare, Pz. Flaminio 9, 00196, Rome, Italy. [4]CNR–IRBIM, Largo Fiera della Pesca 1, 60125, Ancona, Italy. Correspondence and requests for materials should be addressed to A.P. (email: antonella.penna@uniurb.it)

(ML) techniques. Recknagel *et al.*[35] used ANN to predict algal blooms in four freshwater systems. In the northern Adriatic Sea, Volf *et al.*[36] used predictive model for the phytoplankton abundance. Only a few studies used predictive models for HABs in coastal waters: Asnaghi *et al.*[29] used a Quantile Regression Forest to predict the concentration of the toxic benthic dinoflagellate *Ostreopsis* cf. *ovata* in Ligurian Sea (North-western Mediterranean) and Kehoe *et al.*[37] used a Random Forest (RF) to build predictive models of benthic PAR (Photosynthetically Active Radiation) at two sites in Moreton Bay affected by *Lyngbia majuscula* blooms.

In order to develop predictive models, it is crucial to have information about the occurrence of the toxic phytoplankton species. Morphological identification and enumeration of toxic phytoplankton species are usually done by using microscopy methods, which are time-consuming and require taxonomic skills and highly-specialized personnel[38,39]. Moreover, in seawater samples, the target species may be present at very low concentrations, representing only a minor component in the phytoplankton assemblage, and it may risk to remain unnoticed, causing the so-called false negative cases. In addition, morphological identification often stops at genus level failing to discriminate between the various species[40].

Molecular PCR-based techniques have proven to be very useful tools for qualitative identification of microalgal species in coastal waters[41,42]. PCR methods can quickly detect even limited very low abundance of cells[43]. The process is also far more precise, because species-specific ribosomal DNA regions are amplified by using taxon-specific primers. This reduces the risk of inaccuracy, a fundamental condition for the activation of direct analysis that can enable more accurate diagnosis[44–46].

In the Mediterranean Sea, most productive areas, due to the nutrient discharged by numerous rivers, are mainly localized at the mouths of big rivers, among them, the Po River in the northern western Adriatic Sea[47,48]. These riverine discharges can generate eutrophication conditions that may lead to bloom events that can be originated by harmful microalgal species or species complex[49].

The dinoflagellate *Alexandrium minutum* Halim, 1960 is the most widespread toxic species in the western Mediterranean basin[50,51]. This species has been responsible for toxic blooms along the northwestern coast of the Adriatic Sea (Italy) and Ionian Sea, where mussel farms have been contaminated[52,53]. *A. minutum* can produce saxitoxins, GTX1 and 4, that can cause a severe human illness, the Paralytic Shellfish Poisoning (PSP) syndrome[15,54], the most widespread HAB-related shellfish poisoning illness[55]. In the Mediterranean Sea, the increase in the frequency of toxic *A. minutum* outbreaks and the number of areas affected has coincided with the overdevelopment of coastlines, which increasingly offer confined nutrient enriched waters suitable for microalgal proliferation[3,56]. Generally, nutrient rich waters are trigger for its blooming along coastal waters and the physical structure of mass water is critically for the bloom initiation, avoiding cell dispersion and assuring high nutrient levels. In shallow areas, such as coastal shoreline, beaches, bays, *A. minutum* occurs during spring in coincidence with higher temperature, enhanced rainfall and freshwater inputs, which could be related to the supply of macro- and micronutrients, and with stabilization of the water column[23,57]. Furthermore, despite the dinoflagellates' preference for settling in confined environments near shore, *A. minutum* has an enormous natural potential for dispersal because of its capacity to grow and produce resting cysts under a wide range of environmental conditions. This feature can be responsible of toxic bloom dispersion[58,59]. Saxitoxin production in *A. minutum* is difficult to be controlled. It is known that the production of STX in some *A. minutum* strains can be influenced by nutritional conditions. In particular, low levels of phosphorus increase it[60–63]. Moreover, grazer-induced toxin production has been shown in *A. minutum* under nutrient replete conditions[64]. Recently, it was found that *A. minutum* responds to pico- to nanomolar concentrations of copepodamides produced by zooplankton with up to a 20-fold increase in production of paralytic shellfish toxins[65]. The *A. minutum* abundance that can determine the toxic levels dangerous for humans and therefore, representing an alert is not known to date, because many variables can influence the contamination of shellfish filter animals (i.e. environmental parameters, cell concentration in the seawater, cellular toxin content); of course, the conditions of pre-bloom and bloom (10^5–10^6 cells/L) are supposed to be critical for an alert. But, anyway, the presence of *A. minutum* cells in the seawater can represent a potential for a bloom formation, and therefore, it is crucial both to predict and control its occurrence.

Furthermore, in the Adriatic Sea, the *Alexandrium* species that occur frequently are the toxic *A. minutum* together with no PSP producing *A. mediterraneum, A. pseudogonyaulax, A. tamutum* and *A. taylori*[66]. In some cases, light microscopy examination, which is the traditional method used in the monitoring activity, can't identify and distinguish exactly the morpho-type species, due to the similarity of morphology. Therefore, it is important having the tools, such as the molecular techniques to identify properly and rapidly the toxic species from the other no PSP producing *Alexandrium* species, and approach analysis to predict its occurrence.

In this study, we developed a model predicting the occurrence of *A. minutum* in the northern western Adriatic coastal water using a Random Forest (RF) (Breiman, 2001), a Machine Learning ensemble technique that combines many Classification Trees (CT). This technique is particularly effective to develop qualitative predictive models, especially when relationships among variables are unknown.

Methods

Study sites and sampling. A total of 187 surface seawater samples were collected, monthly, from June 2005 to December 2009 along the transects of the Foglia (43°56′0.55N; 12°56′0.18E) and Metauro (43°50′0.54N; 13°05′0.9E) rivers at 500 m and 3000 m (NW Adriatic Sea) from coastland. Seawater samples were collected at 0.5 m depth using polyethylene bottles, and frozen at −20 °C after filtration (0.45 μm nitrocellulose filters, Millipore, USA) until chemical analyses, or fixed with pure ethanol and stored at +4 °C for molecular determinations.

Molecular analysis and PCR assay. Molecular PCR analysis was applied both because *A. minutum* is difficult to distinguish from other species within the same genus, as it is characterized by minute details of its thecal plates[67] and because PCR analysis allows us to be fair more certain about the absence data.

Variables
Day
Distance from coastline (m)
Wind maximum speed (Km h^{-1})
Wind direction
Cloud cover (okta)
Water transparency (m)
Sea surface temperature (°C)
Salinity (PSU)
Dissolved oxygen (mg L^{-1})
Oxygen saturation (% sat.)
Chlorophyll *a* (μg L^{-1})
pH
N-NO_3 (μM L^{-1})
N-NO_2 (μM L^{-1})
N-NH_3 (μM L^{-1})
P-PO_4 (μM L^{-1})
Total P (μM L^{-1})
Si-SiO_2 (μM L^{-1})

Table 1. List of environmental parameters used in the training phase.

For DNA extraction a volume of 100 mL of surface seawater samples, was filtered through a 25 mm diameter Isopore membrane filters with a pore size of 3.0 μm (Merck Millipore, Billerica, MA, USA) under gentle vacuum to avoid cell disruption. The filters were placed in Eppendorf with 1.0 mL of 95% ethanol and stored at +4 °C. Cells were washed out from the filters with ethanol and collected by centrifugation at 12,500 rpm for 10 min at room temperature. Pellets were kept frozen at −80 °C until molecular analyses. Total genomic DNA was purified from pellets, using DNeasy Plant Mini Kit (Qiagen, Valencia, CA). DNA concentration and integrity were evaluated on 0.8% (w/v) agarose gel using serially diluted λ DNA standards (Thermo Fisher Scientific, Hanover Park, IL, USA) and a gel-doc apparatus (Bio-rad, Hercules, CA, USA).

Species-specific primers for the amplification of *A. minutum* ITS–5.8S rDNA region and PCR conditions were reported in Penna *et al.*[41]. The PCR products were resolved on 1.8% (w/v) agarose 1x TAE buffer gel and were visualized by standard ethidium bromide staining under UV light in a gel-doc apparatus (Bio-rad, Hercules, CA, USA).

Chemical-physical analysis. Dissolved oxygen, oxygen saturation, salinity, temperature and pH determinations were performed with a CTD probe (Idronaut mod. Ocean Seven 316). The transparency of the seawater column was approached by Secchi depth. Dissolved inorganic nutrients (N-NO_3, N-NO_2, N-NH_4, P-PO_4 and Si-SiO_2) and chlorophyll "*a*" were performed spectrophotometrically (Shimadzu mod. UV- 1700) on filtered water samples following the methods of Strickland and Parsons[68] and APHA AWWA WPCF[69], respectively. Total phosphorus (TP) was determined on unfiltered water samples according to the method of Valderrama[70].

Modelling procedure. Occurrence data (i.e. presence and absence records based on molecular evidence) were associated not only to oceanographic data, but also to other predictive variables, namely day of the year, distance from coastline and three meteorological variables (wind maximum speed, wind direction and cloud cover). Data about the latter variables were retrieved from SYNOP servers.

At first we associated *A. minutum* occurrence data with all the available predictive variables (Table 1) to train RFs. However, at a later stage we also trained a second RF, using only 12 out of the 18 available predictive variables. The reduced data set excluded information about nutrients to make any future use of the model easier, with no need for water sampling and laboratory analysis to determine nutrients concentrations.

Independently of the number of variables used to predict *A. minutum* occurrence, the available records were divided into two different subsets: one third of them was set aside and *a posteriori* used as test set to validate the model. The remaining data were used as a training set, i.e. to provide the information RFs need to grow.

To assign records to the two subsets (training and test), they were first stratified according to *A. minutum* occurrence (presence or absence). Then each resulting subset was sorted according to the day of the year in which samples were collected, as seasonality is a factor that highly influences the presence of *A. minutum*. Then, in each sequence of three records, one was randomly allocated to the test set and the other two to the training set, thus ensuring the homogeneity of the two subsets.

Using both 18 and 12 predictive variables we tested several RFs, each one with different features given by different combinations of three training parameters. These were: the number of trees in the RF (100, 250, 500 or 1000), the number of variables available at each split (3, 4, 5 or 6) and the minimum number of records in each terminal node, i.e. in each "leaf" (1 to 10).

In RFs the overall output is obtained by collecting the output of each tree for each records. In other words, each tree "votes" for one of the possible states of the target variable and the majority wins. In theory, predicting *A. minutum* presence would need 50% + 1 presence predictions from all the trees in the RF. However, especially

Training set		Predicted values	
		presence	absence
Observed values	presence	43	3
	absence	24	55
	CCI% = 78.4		
	K = 0.58		

Test set		Predicted values	
		presence	absence
Observed values:	presence	21	1
	absence	8	32
	CCI% = 85.5		
	K = 0.70		

Table 2. Confusion matrices for 18-variables new RF, after cut- off optimization (t = 0.310).

when the numbers of presence and absence records are not well balanced, the optimal cut-off value for a successful presence prediction can be different. For instance, a RF could be more accurate if it were allowed to predict *A. minutum* presence even when less than 50% of the trees predict that output. In order to optimize the cut-off value to be used instead of 50%, the ROC (Receiver Operating Characteristic) curve[71] was analyzed to look for the best compromise between true positives and false positives in RF predictions. This way the optimal cut-off value, i.e. the minimum number of presence predictions from the trees in the RF that was needed to issue a presence prediction from the whole RF was found for all the RFs we trained. This procedure was especially necessary because the numbers of presence and absence records were not well balanced in our data set (68 presence and 119 absence records, respectively). As absence records were almost twice as much as those of presence of *A. minutum*, the RF training was slightly biased towards the first case, i.e. to the prediction of absence. Therefore, the optimal cut-off was expected to be smaller than 50% of the votes from the trees, i.e. smaller than 0.5. The ROC curve analysis also provided an AUC (Area Under the Curve) value, that can be regarded as a measure of overall model accuracy. However, in order to select the best model among those we developed with different sets of training parameters, we relied upon the Cohen's K statistics[72].

Results and Discussion

Using all the available predictive variables and different combinations of training parameters (number of trees, number of variables per split and minimum number of records per leaf) we trained 160 RFs. The optimal cut-off value for each RF, i.e. the one that maximized the true positive to false positive ratio, was obtained from the ROC curve analysis. After cut-off optimization, Cohen's K values were calculated for the test set. They ranged from 0.54 to 0.7, with a median value of 0.64 and, as expected, they tended to be inversely proportional to the minimum number of records per leaf. As the best candidate for optimal predictive performance we selected the best model out of the 160 we trained, i.e. we chose the one with the largest K value. The optimal RF model was the one with 100 trees, 3 predictive variables selected at each split and fully-grown trees, with only a single record in each leaf. The latter criterion, by the way, is the default option in the original implementation of the RF[73]. The optimized cut-off value for that RF was 0.31 and K values were 0.58 for the training set and 0.7 for the test set, while the ROC curve analysis returned a 0.895 value for the training set and a 0.88 AUC value for the test set. The K values relative to the test set indicated a substantial[74] to good agreement[75], whereas the AUC testified an excellent performance of the RF model according to Hosmer and Lemeshow[76]. Table 2 showed the confusion matrices for training and test sets as well as K values and the percentage of Correctly Classified Instances (CCI%), which is another index of the accuracy of the model, even though not as robust as Cohen's K in the evaluation of unbalanced data set. CCI% ranged from 78.4 to 85.5, respectively for the training and test set.

Nutrient concentrations are often available in coastal monitoring data, but their acquisition requires the collection of water samples and lab analyses, whereas data about all the other predictive variables can be retrieved from meteorological records or from *in situ* measurements obtained from multiparameter probes. Therefore, we trained more RFs using only 12 predictive variables, i.e. excluding nutrient concentrations. As for the previous RF, we tested several combinations of the training parameters, thus obtaining 160 different RFs. After cut-off optimization K values ranged from 0.51 to 0.7, with a median value of 0.62. As for the RF based on 18 predictive variables, K values were mainly influenced by the minimum number of records in RF leaves, although to a larger extent. The model with the best predictive ability was based on 1000 trees, using only 2 candidate variables at each split and fully-grown trees. The optimized cut-off value for the best RF was 0.361 and K values for training and test set were, respectively, 0.59 and 0.7. While the interpretation of K values was exactly the same as in the RF based on 18 predictive variables, the AUC values were 0.891 for the training set and 0.905 for the test set. AUC value for the test set, in this case, was a bit larger than the value for the training set and it was also a bit larger than the value for the test set of the other model, indicating an outstanding accuracy according to Hosmer & Lemeshow[76]. The confusion matrices for both the training and the test set were shown in Table 3, together with K values and CCI%, which, as in the previous case, were higher for the test set.

Comparing the two RFs, the one based on the full set of predictive variables was less dependent than the other one on the optimization of its training parameters, as shown in Fig. 1, where the central quartiles of the K values were narrower than those for the RF based on 12 predictive variables. Moreover, the median K value was larger (0.64 vs. 0.62) in the first case.

However, while using all the predictive variables allowed obtaining less variability depending on the RF training parameters, the best RF model obtained from the reduced set of predictive variables was as good as the best RF model obtained from the full set of predictive variables, if not marginally better (they're slightly better in the AUC value). Therefore, we have to consider nutrient concentrations as not strictly needed. As obtaining information about nutrients requires additional activities, with larger costs in time and money, we regard the model based

Training set		Predicted values: presence	Predicted values: absence	Test set		Predicted values: presence	Predicted values: absence
Observed values:	presence	39	7	Observed values:	presence	20	2
	absence	18	61		absence	7	33
CCI% = 80.0				CCI% = 85.5			
K = 0.59				K = 0.70			

Table 3. Confusion matrices for 12-variables new RF, after cut- off optimization (t = 0.361).

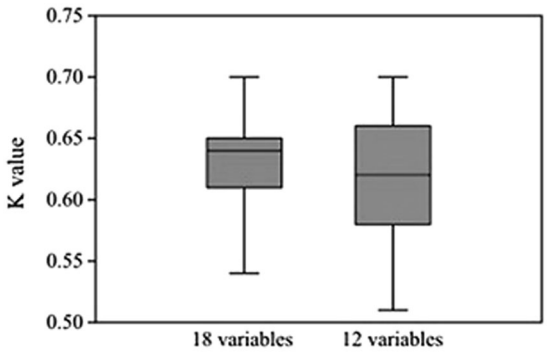

Figure 1. Box plot with K values distribution for all the models tested with different parameters combination. On the left, values of the 18-variables model: minimum value is 0.54, maximum is 0.7. Median value is 0.64. On the right, values of the 12-variables model: minimum value is 0.51, maximum is 0.7 and median value is 0.62.

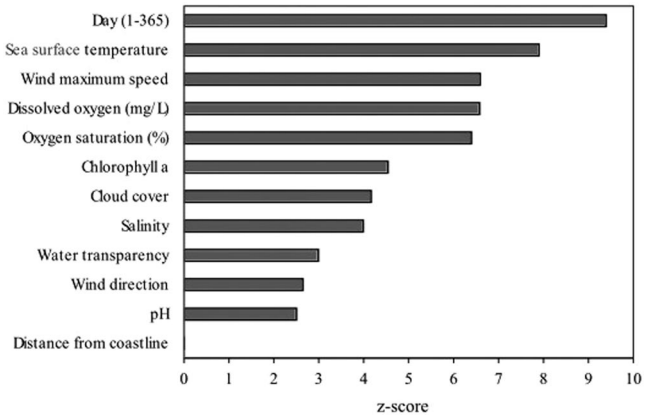

Figure 2. Plot with variable importance for 12-variables RF. The z-scores are obtained by dividing the raw scores by their standard error. All the bars are associated to significant z-scores except the one for distance from coastline, which is non significant and therefore was omitted.

on only 12 predictive variables as the best solution to use for making prediction in the future, not only because of its predictive ability, but also because of practical issues.

While the main drivers of any model can be identified thanks to sensitivity analysis, an interesting property of the RF algorithm is its ability to support an estimate of the relevance of the role played by each predictive variable. Relative importance of the 12 predictive variables used by the reduced RF model was shown in Fig. 2 as z-scores, computed according to the original algorithm proposed by Breiman[73].

As we expected, the day of the year, and therefore the period in which samples were collected, is the variable with the largest importance value, and therefore, the most correlated to *A. minutum* presence. In the studied period, the abundance of *A. minutum* was in the range of 10^3–10^5 cells/L (data not shown). Sea surface temperature (which is obviously not independent of day of the year, i.e. of season) was the second most important predictive variable, followed by wind maximum speed and oxygen concentration and saturation. Interactions between temperature, wind and oxygen concentration were obvious and certainly modulated by seasonal conditions in favoring *A. minutum* presence. The least important variable, according to the z-score obtained from the RF algorithm, was water pH, which was hardly connected, from a theoretical standpoint, to *A. minutum* presence and possibly affected by relatively large measurement errors.

The main goal of HABs management is to provide early warnings to prevent their impacts on public health and economical activities. Microscope identification of target species is a common procedure, although it requires a great deal of taxonomic expertise, in addition to being time consuming and impractical for processing a large number of samples in a monitoring perspective[38,39,77].

Recently, HABs phenomena are increasing in the Mediterranean Sea possibly under the influence of the coastal zone overdevelopment[24]. Climate change and global warming are now the main problems that may increase the risk of reaching critical conditions, especially in the Adriatic Sea. The latter is a very shallow sea and one of the most productive regions in the Mediterranean Sea, with nutrient inputs from riverine discharges[78] and where mussel farms, which play a relevant role in local as well as in Italian mariculture, have already been contaminated[52].

Results obtained in this study suggest that predictive models may be a valid supplementary tool in HABs management. In fact, they could be very useful to gain important information about those events and to identify the particular conditions in which HABs are more likely to occur, thus supporting the implementation of both new research efforts and activities focused on early reaction, whenever the event should occur.

While our models are already able to correctly predict more than 80% of the real-world instances, the RF approach will allow further improvement as soon as more records about *A. minutum* presence or absence will become available. Moreover, while our model was validated only locally, the same procedure can be applied to other sites or to several sites simultaneously. The ultimate goal, obviously, is a general model, trained and validated in a larger region or across the whole Mediterranean basin.

Conclusions

Modelling species distribution, both in space and in time, is usually easier when data about species occurrence are not affected by too many error sources. Undetected occurrences are a very common problem among those that may hinder species distribution models and they are more likely to happen than their positive counterpart, i.e. false occurrences, which may depend on species misidentification. While the first source of error depends on sampling design relative to species distribution, the second source only depends on the taxonomical skills supporting the modeler. As for studies on plankton species or assemblages, using molecular methods for species identification solves both problems, because false negatives and false positives are not likely to occur.

As a consequence, even a relatively small data set can support successful modelling if appropriate methods are selected for species identification. This is certainly the case with our study, because species occurrence data were obtained by molecular PCR analyses, which makes us especially confident about absence records. In fact, the latter can be regarded as real absence rather than as misidentification or undetected presence due to very low density of the target species. Confidence in species detection makes us also confident about the accuracy of our model.

This study was carried out for a single species over a relatively restricted area, but the selected approach can be easily applied elsewhere and at any spatial scale. Moreover, its methodological bases allow an easy application to the prediction of a wide range of different target species and this is the reason why RFs are rapidly becoming one of the most widely applied techniques in species-specific distribution modelling.

Our model allows to correctly classify more than 85% cases of presence or absence of *A. minutum*, with values of the K statistics as high as 0.7 for the test set. This result is certainly adequate for supporting an early warning that can be improved.

While the most common goal of any model is to provide accurate predictions, understanding the underlying ecological relationships is a very common secondary or even alternate objective. In our study, the focus was on the prediction of occurrence, but the importance of the predictive variables was assessed by means of the procedure based on the standardized errors in classification of out-of-bag records obtained from RF training. The assessment of the importance of each predictive variable is obviously based on the available data set only, which can be restricted to a limited number of environmental conditions or to limited sequence of events in a more complex time series. From a purely theoretical viewpoint, however, day of the year, sea surface temperature, wind maximum speed and oxygen concentration and saturation are very likely to be associated to conditions in which *A. minutum* is more frequently found. Needless to say, that association is a fact at local space and time scale and just a hypothesis to be tested at larger scale, as often happens when ecological inferences are based on real data sets.

Our model, however, will certainly play a role in predicting, and possibly better understanding, HABs, although it can only help to identify environmental conditions that might favor HABs, not the actual occurrence of those phenomena. As a matter of fact, we still do not have enough data as to try to understand and possibly modelling what triggers a HAB, but our model is certainly able to point out the conditions that are necessary, although not sufficient, to support that type of event. From this viewpoint, machine learning approaches seem particularly promising because they can be easily updated and optimized as soon as new data become available, thus providing useful support to human experts in HAB risk assessment.

Data Availability

The authors declare the data availability.

References

1. Hamer, J. P., Lucas, I. A. N. & McCollin, T. A. Harmful dinoflagellate resting cysts in ships' ballast tank sediments: potential for introduction into English and Welsh waters. *Phycologia* **40**, 246–255 (2001).
2. Heisler, J. *et al.* Eutrophication and harmful algal blooms: A scientific consensus. *Harmful Algae* **8**, 3–13 (2008).
3. Bravo, I. *et al.* Bloom dynamics and life cycle strategies of two toxic dinoflagellates in a coastal upwelling system (NW Iberian Peninsula). *Deep Sea Res. II* **57**, 222–234 (2010).
4. Anderson, D. M., Cembella, A. D. & Hallegraeff, G. M. Progress in understanding harmful algal blooms (HABs): Paradigm shifts and new technologies for research, monitoring and management. *Ann. Rev. Mar. Sci.* **4**, 143–176 (2012).

5. Glibert, P. M. *et al*. Vulnerability of coastal ecosystems to changes in harmful algal bloom distribution in response to climate change: Projections based on model analysis. *Glob. Chan. Biol.* **20**, 3845–3858 (2014).
6. Smayda, T. J. & Reynolds, C. S. Community assembly in marine phytoplankton: application of recent models to harmful dinoflagellate blooms. *J. Plankton Res.* **23**, 447–461 (2001).
7. Bricker, S. B., Ferreira, J. G. & Simas, T. An integrated methodology for assessment of estuarine trophic status. *Ecol. Model.* **60**, 169–39 (2003).
8. Hallegraeff, G. M. Ocean climate change, phytoplankton community responses, and harmful algal blooms: a formidable predictive challenge. *J. Phycol.* **46**, 220–235 (2010).
9. Fu, F. X., Tatters, A. O. & Hutchins, D. A. Global change and the future of harmful algal blooms in the ocean. *Mar. Ecol. Progr. Ser.* **470**, 207–23 (2012).
10. Hallegraeff, G. M. Harmful algal blooms: a global overview. Manual on Harmful Marine Microalgae. (eds G. M., Hallegraeff, D. M., Anderson & A. D. Cembella) 25–49 (UNESCO, Paris 2003).
11. Gallitelli, M., Ungaro, N., Addante, L. M., Gentiloni Silver, N. & Sabbà, C. Respiratory illness as a reaction to tropical algal blooms occurring in a temperate climate. *J. Am. Med. Assoc.* **293**, 2599–2600 (2005).
12. Casabianca, S. *et al*. Quantification of the toxic dinoflagellate *Ostreopsis spp*. by qPCR assay in marine aerosol. *Environ. Sci. Technol.* **47**, 3788–3795 (2013).
13. Ciminiello, P., Dell'Aversano, C., Dello Iacovo, E., Forino, M. & Tartaglione, L. Liquid chromatography–high-resolution mass spectrometry for palytoxins in mussels. *Anal. Bioanal. Chem.* **407**, 1463–1473 (2015).
14. Deeds, J. R., Landsberg, J. H., Etheridge, S. M., Pitcher, G. C. & Longan, S. W. Non-Traditional vectors for paralytic shellfish poisoning. *Mar. Drugs* **6**, 308–348 (2008).
15. Wiese, M., D'Agostino, P. M., Mihali, T. K., Moffitt, M. C. & Neilan, B. A. Neurotoxic alkaloids: Saxitoxin and its analogs. *Mar. Drugs* **8**, 2185–2211 (2010).
16. Hoagland, P. & Scatasta, S. The economic effects of harmful algal blooms. Ecology of Harmful Algae. (eds E., Graneli & J. T., Turner) 391–401 (Springer-Verlag, Berlin 2006).
17. Morgan, K. L., Larkin, S. L. & Adams, C. M. Firm-level economic effects of HABS: A tool for business loss assessment. *Harmful Algae* **8**, 212–218 (2009).
18. Berdalet, E. *et al*. Marine harmful algal blooms, human health and wellbeing: challenges and opportunities in the 21st century. *J. Mar. Biolo. Ass. UK* **96**, 61–91 (2015).
19. Kudela, R. M. & Gobler, C. J. Harmful dinoflagellate blooms caused by *Cochlodinium* sp.: global expansion and ecological strategies facilitating bloom formation. *Harmful Algae* **14**, 71–86 (2012).
20. Lewitus, A. J. *et al*. Harmful algal blooms along the North American west coast region: history trends, causes, and impacts. *Harmful Algae* **19**, 133–159 (2012).
21. Paerl, H. W. Mitigating harmful cyanobacterial blooms in a human-and climatically-impacted world. *Life* **4**, 988–1012 (2014).
22. Wells, M. L. *et al*. Harmful algal blooms and climate change: learning from the past and present to forecast the future. *Harmful Algae* **49**, 68–93 (2015).
23. Vila, M. *et al*. A comparative study on recurrent blooms of *Alexandrium minutum* in two Mediterranean coastal areas. *Harmful Algae* **4**, 673–695 (2005).
24. Garcés, E. & Camp, J. Habitat changes in the Mediterranean Sea and the consequences for Harmful Algal Blooms formation. Life in the Mediterranean Sea: A Look at Habitat Changes. (ed. Noga Stambler Israel) 519–541 (2012).
25. Kleindinst, J. L. *et al*. Categorizing the severity of paralytic shellfish poisoning outbreaks in the Gulf of Maine for forecasting and management. *Deep-Sea Res. II* **103**, 277–287 (2014).
26. Jeong, K. S., Kim, D. K., Whigham, P. & Joo, G. J. Modelling *Microcystis aeruginosa* bloom dynamics in the Nakdong River by means of evolutionary computation and statistical approach. *Ecol. Model.* **161**, 67–78 (2003).
27. Lee, J. H. W., Huang, Y., Dickman, M. & Jayawardena, A. W. Neural network modeling of coastal algal blooms. *Ecol. Model.* **159**, 179–201 (2003).
28. Wang, J., Tang, D. & Sui, Y. Winter phytoplankton bloom induced by subsurface upwelling and mixed layer entrainment southwest of Luzon Strait. *J. Mar. Syst.* **83**, 141–149 (2010).
29. Asnaghi, V. *et al*. A novel application of an adaptable modeling approach to the management of toxic microalgal bloom events in coastal areas. *Harmful Algae* **63**, 184–192 (2017).
30. Hamilton, G., McVinish, R. & Mengersen, K. Bayesian model averaging for harmful algal bloom prediction. *Ecol. Appl.* **19**, 1805–1814 (2009).
31. Anderson, C. R. *et al*. Predicting potentially toxigenic *Pseudo-nitzschia* blooms in the Chesapeake Bay. *J. Mar. Syst.* **83**, 127–140 (2010).
32. Blauw, A. N., Los, F. J., Huisman, J. & Peperzak, L. Nuisance foam events and *Phaeocystis globosa* blooms in Dutch coastal waters analyzed with fuzzy logic. *J. Mar. Syst.* **83**, 115–126 (2010).
33. Colasanti, R. L. Discussions of the possible use of neural network algorithms in ecological modelling. *Binary* **3**, 13–15 (1991).
34. Edwards, M. & Morse, D. R. The potential for computer-aided identification in biodiversity research. *Trends Ecol. Evol.* **10**, 153–158 (1995).
35. Recknagel, F., French, M., Harkonen, P. & Yabunaka, K. I. Artificial neural network approach for modelling and prediction of algal blooms. *Ecol. Model.* **96**, 11–28 (1997).
36. Volf, G., Atanasova, N., Kompare, B., Precali, R. & Oani, N. Descriptive and prediction models of phytoplankton in the northern Adriatic. *Ecol. Model.* **222**, 2502–2511 (2011).
37. Kehoe, M. *et al*. Random forest algorithm yields accurate quantitative prediction models of benthic light at intertidal sites affected by toxic *Lyngbya majuscula* blooms. *Harmful Algae* **19**, 46–52 (2012).
38. Smayda, T. J. Harmful algal blooms: their ecophysiology and general relevance to phytoplankton blooms in the sea. *Limnol. Oceanogr.* **42**, 1137–1153 (1997).
39. Penna, A. & Galluzzi, L. The quantitative real-time PCR applications in the monitoring of marine harmful algal bloom (HAB) species. *Environ. Sci. Poll. Res.* **20**, 6851–6862 (2013).
40. Godhe, A. *et al*. Intercalibration of classical and molecular techniques for identification of *Alexandrium fundyense* (Dinophyceae) and estimation of cell densities. *Harmful Algae* **6**, 56–72 (2007).
41. Penna, A. *et al*. Monitoring of HAB species in the Mediterranean Sea through molecular methods. *J. Plankton Res.* **29**, 19–38 (2007).
42. Battocchi, C. *et al*. Monitoring toxic microalgae *Ostreopsis* (dinoflagellate) species in coastal waters of the Mediterranean Sea using molecular PCR-based assay combined with light microscopy. *Mar. Pollut. Bull.* **60**, 1074–84 (2010).
43. Perini, F. *et al*. New approach using the real-time PCR method for estimation of the toxic marine dinoflagellate *Ostreopsis* cf. *ovata* in marine environment. *PLoS One* **6**(3), e17699 (2011).
44. Murray, S. A. *et al*. Differential accumulation of paralytic shellfish toxins from *Alexandrium minutum* in the pearl oyster, *Pinctada imbricata*. *Toxicon* **54**, 217–223 (2009).
45. Delaney, J. A., Ulrich, R. M. & Paul, J. H. Detection of the toxic marine diatom *Pseudo-nitzschia multiseries* using the RuBisCO small subunit (rbcS) gene in two real-time RNA amplification formats. *Harmful Algae* **11**, 54–64 (2011).
46. Pugliese, L., Casabianca, S., Perini, F., Andreoni, F. & Penna, A. A high-resolution melting method for the molecular identification of the potentially toxic diatom *Pseudo-nitzschia* spp. in the Mediterranean Sea. *Sci. Rep.* **7**, 4259 (2017).

47. Raicich, F. On the fresh water balance of the Adriatic Sea. *J. Mar. Syst.* **9**, 305–319 (1996).
48. Degobbis, D. *et al*. Long-term changes in the northern Adriatic ecosystem related to anthropogenic eutrophication. *Int. J. Environ. Poll.* **13**, 495e533 (2000).
49. Marić, D. *et al*. Phytoplankton response to climatic and anthropogenic influences in the north-eastern Adriatic during the last four decades. *Estuar. Coast. Shelf Sci.* **115**, 98–112 (2012).
50. Giacobbe, M. G. & Maimone, G. First report of *Alexandrium minutum* Halim in a Mediterranean Lagoon. *Cryptogamie Algol.* **15**, 47–52 (1994).
51. Vila, M., Camp, J., Garcés, E., Masó, M. & Delgado, M. High resolution spatio-temporal detection of potentially harmful dinoflagellates in confined waters of the NW Mediterranean. *J. Plankton Res.* **23**, 497–514 (2001).
52. Honsell, G. *et al*. *Alexandrium minutum* Halim and PSP contamination in the Northern Adriatic Sea (Mediterranean Sea). Harmful and Toxic Algal Blooms. (eds T., Yasumoto, T., Oshima & Y. T., Fukuyo) 77–83 (UNESCO, Paris 1996).
53. Penna, A. *et al*. The *sxt* gene and paralytic shellfish poisoning toxins as markers for the monitoring of toxic *Alexandrium* species blooms. *Environ. Sci. Technol.* **49**, 14230–14238 (2015).
54. Perini, F. *et al*. *SxtA* and *sxtG* gene expression and toxin production in the Mediterranean *Alexandrium minutum* (Dinophyceae). *Mar. Drugs* **12**, 5258–5276 (2014).
55. Anderson, D. M. *et al*. The globally distributed genus *Alexandrium*: Multifaceted roles in marine ecosystems and impacts on human health. *Harmful Algae* **14**, 10–35 (2012).
56. Bravo, I., Vila, M., Maso, M., Ramilo, I. & Figueroa, R. I. *Alexandrium catenella* and *Alexandrium minutum* blooms in the Mediterranean Sea: toward the identification of ecological niches. *Harmful Algae* **7**, 515–522 (2008).
57. Giacobbe, M. G., Oliva, F. D. & Maimone, G. Environmental factors and seasonal occurrence of the dinoflagellate *Alexandrium minutum*, a PSP potential producer in a Mediterranean lagoon. *Estuar. Coast. Shelf Sci.* **42**, 539–549 (1996).
58. Anglés, S., Garcés, E., René, A. & Sampedro, N. Life-cycle alternations in Alexandrium minutum natural populations from the NW Mediterranean Sea. *Harmful Algae* **16**, 1–11 (2012).
59. Anderson, D. M. *et al*. *Alexandrium fundyense* cysts in the Gulf of Maine: Long-term time series of abundance and distribution, and linkages to past and future blooms. *Deep Sea Res. II* **103**, 6–26 (2014).
60. Guisande, C., Frangópulos, M., Maneiro, I., Vergara, A. R. & Riveiro, I. Ecological advantages of toxin production by the dinoflagellate *Alexandrium minutum* under phosphorus limitation. *Mar Ecol Prog Ser* **225**, 169–176 (2002).
61. Lippemeier, S., Frampton, D. M. F., Blackburn, S. I., Geier, S. C. & Negri, A. P. Influence of phosphorus limitation on toxicity and photosynthesis of *Alexandrium minutum* (dinophyceae) monitored by in-line detection of variable chlorophyll fluorescence. *J. Phycol* **38**, 320–331 (2003).
62. Frangópulos, M., Guisande, C., deBlas, E. & Maneiro, I. Toxin production and competitive abilities under phosphorus limitation of *Alexandrium* species. *Harmful Algae* **3**, 131–139 (2004).
63. Touzet, N., Franco, J. M. & Raine, R. Influence of inorganic nutrition on growth and PSP toxin production of *Alexandrium minutum* (Dinophyceae) from Cork Harbour, Ireland. *Toxicon* **50**, 106–119 (2007).
64. Selander, E., Thor, P., Toth, G. B. & Pavia, H. Copepods induce paralytic shellfish toxin production in marine dinoflagellates. *Proc. R. Soc. Lond. Ser. B-Biol. Sci.* **273**, 1673–1680 (2006).
65. Selander, E. *et al*. Predator lipids induce paralytic shellfish toxins in bloom-forming algae. *Proc. Nat. Acad. Sci.* **112**, 6395–6400 (2015).
66. Penna, A. *et al*. Phylogenetic relationships among the Mediterranean *Alexandrium* (Dinophyceae) species based on sequences of 5.8 S gene and Internal Transcript Spacers of the rRNA operon. *Eur. J. Phycol.* **43**, 163–178 (2008).
67. Taylor, F. J. R. & Fukuyo, Y., Larsen, J. Taxonomy of harmful dinoflagellates. Manual of Harmful Microalgae. (eds G. M., Hallegraeff, D. M., Anderson & A. D. Cembella) 283–317 (IOC UNESCO, Paris 1995).
68. Strickland, J. D. H. & Parsons, T. R. A practical handbook of seawater analysis. *J. Fish. Res. Bd.* **167**, 49–89 (1972).
69. American Public Health Association, American Water Works Association, and Water Pollution Control Federation (APHA/AWWA/WPCF). Standard Methods for Water and Wastewater Treatment. (ed.16th APHA) 1067–1072 (Washington 1985).
70. Valderrama, J. C. The simultaneous analysis of total nitrogen and total phosphorus in natural waters. *Mar. Chem.* **10**, 109–122 (1981).
71. Zweig, M. H. & Campbell, G. Receiver-Operating Characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.* **39**, 561–577 (1993).
72. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **20**, 37–46 (1960).
73. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
74. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174 (1977).
75. Fleiss, J. L. Statistical methods for rates and proportions. (Wiley, New York 1981).
76. Hosmer, D. W. & Lemeshow, S. L. Applied Logistic Regression. (Wiley, New York 2000).
77. Sellner, K. G., Doucette, G. J. & Kirkpatrick, G. J. Harmful algal blooms: causes, impacts and detection. *J. Ind. Microbiol. Biotechnol.* **30**, 383–406 (2003).
78. Giani, M. *et al*. Recent changes in the marine ecosystems of the northern Adriatic Sea. *Estuar. Coast. Shelf Sci.* **115**, 1–13 (2012).

Acknowledgements

This research was supported by Regione Marche Project Coastal Monitoring n. 49 of 23/12/2013 of Table C. The monitoring and sampling carried out with Athena Vessel were also funded by the Department of Biomolecular Sciences and University of Urbino "Carlo Bo".

Author Contributions

E.V., M.S., A.P. contributed to the conception and design of the study; E.V. carried out the study. E.V. performed the statistical analyses. F.R. and S.C. carried out the chemical physical analysis. S.C. performed the molecular analyses. All authors were involved in the manuscript preparation and revision approval of the final version of the manuscript.

Additional Information

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.