Ecological Informatics 43 (2018) 12–23

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Using general linear model, Bayesian Networks and Naive Bayes classifier for prediction of *Karenia selliformis* occurrences and blooms

Wafa Feki-Sahnoun[a,*], Hasna Njah[b,c], Asma Hamza[a], Nouha Barraj[d], Mabrouka Mahfoudi[a], Ahmed Rebai[e], Malika Bel Hassen[d]

[a] Institut National des Sciences et Technologies de la Mer, Centre de Sfax, Rue Madagascar, BP 1035, Sfax, CP 3018, Tunisia
[b] Faculté des Sciences Économiques et de Gestion de Sfax, Route de l'Aéroport Km 4, Sfax 3018, Tunisia
[c] Laboratoire de Multimedia, InfoRmation Systems and Advanced Computing Laboratory, Pôle technologique de Sfax, Route de Tunis Km 10 BP 242, CP 3021, Sfax, Tunisia
[d] Institut National des Sciences et Technologies de la Mer (INSTM), 28 rue 2 mars 1934, Salammbô 2025, Tunisia
[e] Centre de Biotechnologie de Sfax, Route Sidi Mansour Km 6, BP 1177, 3018 Sfax, Tunisia

ARTICLE INFO

Keywords:
Karenia selliformis
Naive Bayes classifier
General linear model
Bayesian Network
Hydro-meteorological parameters
the Gulf of Gabès

ABSTRACT

The prediction of the dinoflagellate red tide forming *Karenia selliformis* is a relevant task to aid optimized management decisions in marine coastal water. The objective of the present study is to compare different modeling approaches for prediction of *Karenia selliformis* occurrences and blooms. A set of physical parameters (salinity, temperature and tide amplitude), meteorological constraints (evaporation, air temperature, insolation, rainfall, atmospheric pressure and humidity), sampling months and sampling sites are used. The model prediction included general linear model (GLM), Bayesian Network (BN) and the simplest BN type which is, Naive Bayes classifier (NB). The results showed that three models incriminated high salinity in *Karenia selliformis* blooms and the sampling sites, mainly Boughrara lagoon, in the occurrences. The BN performed better than linear models (NB and GLM) for both *Karenia selliformis* occurrences and blooms prediction. This later is related to the facts that BN considered the inter-independency between predictive variables and that the relationships between the variables and the outcome are often non-linear such us; the transition to bloom situations appeared to be triggered by a salinity threshold. This study is useful in the management of this ecosystem so as to use the best disposal options in the early prediction of the toxic blooms.

1. Introduction

The harmful *Karenia* species are found throughout the world. They have been described as a result of investigations into extensive animal mortalities or human health problems. They have become the most studied species of harmful algae with extensive investigations on the physiology and bloom formation. *Karenia selliformis* (Hansen et al., 2004) has been the most abundant species causing severe harmful blooms in the Gulf of Gabès (Hamza and Abed, 1994). Since the year 1990, *K. selliformis* blooms have occurred annually along the coast, and represented over 64% of the reported blooms in this area (Feki et al., 2008, 2013). It has been argued that their proliferation has been usually related to shellfish toxicity (Ben Naila et al., 2012; Marrouchi et al., 2009; Medhioub et al., 2010). Little is known about the effect of environmental factors on *K. selliformis* occurrences and blooms. Information on optimal growth of this species under controlled temperature and salinity was documented in culture experiments (Medhioub et al., 2009). Tentatively models have been developed to apprehend the effects of physical and meteorological variables on Ks occurrences and blooms in the Gulf of Gabès. A generalized linear mixed-effect model (GLMM) incriminated mainly water temperature and some nutrients in the spatiotemporal occurrences of *Karenia selliformis* (Feki et al., 2013). Bayesian Network approach showed that the bloom can be predicted based on salinity threshold (Feki-Sahnoun et al., 2017). However, the best performance model having the highest goodness of fit on the prediction of *Karenia* blooms and occurrences still needs to be

Abbreviations: Generalized linear mixed-effect model, GLMM; *Karenia selliformis*, *K. selliformis*; Analysis of Variance, ANOVA; General linear model, GLM; Bayesian Network, BN; Naive Bayes classifiers, NB; Tunisian National Meteorological Institute, INM; Akaike Information Criterion, AIC; Bayesian Information Criteria, BIC; SamIam, Sensitivity Analysis, Modeling, Inference And More software; Directed Acyclic Graph, DAG; Conditional Probability Tables, CPTs; Maximum likelihood, ML; Maximum A Posteriori, MAP; Evaporation, Evap; Insolation, Insol; Salinity, Sal; Humidity, Humid; Water temperature, WatT

* *Corresponding author.
E-mail addresses: wafafeki@yahoo.fr (W. Feki-Sahnoun), asma.hamza@instm.rnrt.tn (A. Hamza), barraj.nouha@instm.rnrt.tn (N. Barraj), ahmed.rebai@cbs.rnrt.tn (A. Rebai), belhassen.malika@instm.rnrt.tn (M.B. Hassen).

https://doi.org/10.1016/j.ecoinf.2017.10.017
Received 5 August 2017; Received in revised form 25 October 2017; Accepted 31 October 2017

determined.

To date, a large variety of statistical models are available to analyze a relationship between environmental variables and species distributions (i.e. cross validation criteria, ANOVA) and one of the most popular is the general linear model (GLM) (e.g. McCulloch et al., 2008). Typically, the biological and physical processes, generating this data, are highly complex, resulting in multiple correlations/dependencies between covariates and also between outcome variables. Standard statistical approaches have a limited ability to describe such interdependent multi-factorial relationships. In the last decades, Bayesian Network's (BN) modeling has been widely used in solving environmental problems (Borsuk et al., 2004, 2006; Bromley et al., 2005; Feki-Sahnoun et al., 2017; Pollino et al., 2007; Smith et al., 2007; Zaffalon, 2005) to analyze multi-dimensional data. Among the BN models, the naïve Bayes (NB) model appears to be the most popular (Aguilera et al., 2013; Fytilis and Rizzo, 2013; Markus et al., 2010; Ropero et al., 2014, 2015). Despite its linearity and simplicity, the NB has been found to perform surprisingly well (Friedman et al., 1997) particularly in many complex situations (Boets et al., 2015; Domingos and Pazzani, 1997; Zhang, 2004).

The present study compares results from Bayesian Networks [BN, Bishop, 2006; Pearl, 1985)], Naive Bayes classifier [NB, Friedman et al., 1997] and general linear model [GLM, McCullagh and Nelder, 1989; Nelder and Wedderburn, 1972] in the modeling of Ks occurrences and blooms; which is based on a set of physical and meteorological variables. In contrast to Feki-Sahnoun et al. (2017) who used the same dataset and focused only on the Bayesian Network, this research elaborated upon the comparison between the three models based on a new threshold concentration for bloom and non-bloom conditions established using the Feki-Sahnoun et al. (2017) output results. The choice of these models is due to the fact that GLM analysis is a well-known method that allows predicting a target variable (in this case, *Karenia selliformis*) conditionally on a set of variables by fitting a regression model using least squares (Hastie et al., 2009). Many different models could be fitted by including different subsets of the available observed variables and interaction terms between them. The challenge in implementing this technique is the selection of the best subset of variables, as the inclusion of correlated predictors may result in increased standard errors of the regression coefficients which cause prediction's sensitivity to model changes (Burnham and Anderson, 2002). Whereas GLMs are typically data-driven models techniques (Hastie and Tibshirani, 1990; Madsen and Thyregod, 2011; McCullagh and Nelder, 1989; Nelder and Wedderburn, 1972), BNs can be based on expert knowledge and/or stakeholders (Aguilera et al., 2011; Cain et al., 2003; Chan et al., 2012; Haapasaari and Karjalainen, 2010; Haines-Young, 2011; McVittie et al., 2015; Wang et al., 2009) in order to analyze how a system works, to seek information on the appropriate probabilities, or to explore the usefulness of the decision process (Gonzalez-Redin et al., 2016). BN is an alternative modeling approach applying different criteria for model selection. This method consists of a graphical modeling tool that can be used to construct a predictive model by factorizing the posterior density distribution function assuming a set of stable conditional dependencies (Jensen, 1996; Lauritzen and Spiegelhalter, 1988; Pearl, 1988). Among BN models, Naive Bayes classifier is the simplest; because of the fact that its variables are conditionally independent given the class variable. It has the strength to not only provide a prediction, but also the estimated probability associated with each possible outcome (Fernandes et al., 2010).

The objective of the present study is to compare the efficiency of the linear (GLM, NB) and non-linear (BN) models in order to predict *K. selliformis* occurrences and blooms in the Gulf of Gabès using a set of meteorological (rainfall, atmospheric pressure, insolation, …) and physical (temperature, salinity, tide) variables. By comparing BN with GLM and NB using identical data, the likely impact of such an analytical difference on the inferences was highlighted.

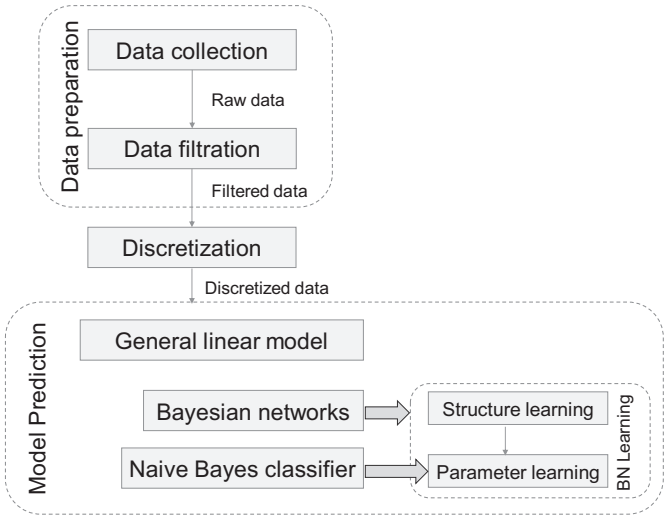

Fig. 1. Framework of the proposed approach describing the three major steps: Data preparation, Discretization and Model Prediction.

2. Material and methods

2.1. Study area

The present study was carried out in the Gulf of Gabès (Tunisia, Eastern Mediterranean Sea), in a wide continental shelf area extending from 9.5 to 12°E in longitude and from 33 to 35.5°N in latitude and sheltering various islands (Kerkennah and Djerba) and lagoons (Boughrara and El Bibane). The Boughrara lagoon is surrounded by solar saltern areas where net evaporation is very high and soil washing is negligible due to limited freshwater supplies coupled with scanty rainfall.

The bathymetry is characterized by a low slope; the 50-m depth is reached at a distance of 100 km (Fig. 1). The tide is semidiurnal, being among the highest in the Mediterranean and having a maximum range of about 2 m generated by resonance phenomena (Sammari et al., 2006). The climate is dry and sunny with strong easterly winds.

Despite being oligotrophic, this Gulf has the peculiarity of being highly productive and accounts for 65% of Tunisian fish production (DGPA, 2005–2009) and is a well-known habitat for marine turtles (Jribi et al., 2008; Lotze and Worm, 2009; Maffucci et al., 2006). Although it has huge importance for the local economy and wildlife conservation, the Gulf of Gabès has been subject to the strong pressures of urbanization, industry, fisheries, tourism and anthropogenic releases (Béjaoui et al., 2004; Ben Brahim et al., 2010; Rekik et al., 2012). It has been the subject of increased anthropogenic interference threatening the entire ecosystem.

2.2. Analysis process

The flowchart in Figure 1was followed in order to conduct a consistent analysis. The analysis (Fig. 1) includes three major steps: Data preparation, Discretization and Model Prediction.

2.2.1. Data preparation

Data were collected in the framework of the National Phytoplankton Monitoring Program in the shellfish harvest areas. This program has been operating since 1995 and covers fifteen sites with weekly measures (Fig. 2). The monitoring was performed on a regular schedule all year round. Each site can include more than one sampling station totalizing thirty-two sampling stations (Fig. 2). The temporal windows of the dataset used for the analysis; that covered the period ranging from 1997 to 2007; are considered as the most consistent in terms of data