Ecological Modelling 354 (2017) 11–19

Contents lists available at ScienceDirect

Ecological Modelling

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

# Effects of species prevalence on the performance of predictive models

Ratha Sor[a,b,c,*], Young-Seuk Park[d], Pieter Boets[b,e], Peter L.M. Goethals[b], Sovan Lek[a]

[a] Université de Toulouse, Laboratoire Evolution & Diversité Biologique, UMR 5174, CNRS – Université Paul Sabatier, 118 route de Narbonne, 31062, Toulouse cédex 4, France
[b] Laboratory of Environmental Toxicology and Aquatic Ecology, Ghent University, Campus Coupure building F, Coupure Links 653, 9000, Ghent, Belgium
[c] Department of Biology, Faculty of Science, Royal University of Phnom Penh, Russian Boulevard, 12000, Phnom Penh, Cambodia
[d] Department of Life and Nanopharmaceutical Sciences and Department of Biology, Kyung Hee University, Seoul, 130-701, Republic of Korea
[e] Provincial Centre of Environmental Research, Godshuizenlaan 95, 9000, Ghent, Belgium

ARTICLE INFO

*Article history:*
Received 14 October 2016
Received in revised form 8 March 2017
Accepted 8 March 2017
Available online 21 March 2017

*Keywords:*
Quadratic effect
Species occurrence
Logistic regression
Random forest
Artificial neural network
Support vector machine
Macroinvertebrates
Habitat suitability
Mekong river

ABSTRACT

Predictive models are useful to support decision making, management and conservation planning. However, the performance of models varies across techniques and is affected by several factors including species prevalence (i.e. the occurrence rate of each species in the total samples). Here, we analysed and compared the performance of four common modelling techniques based on the species prevalence. The occurrence of macroinvertebrates collected at 63 sites along the Lower Mekong Basin was predicted using Logistic Regression, Random Forest, Support Vector Machine and Artificial Neural Network (ANN). Model performance was evaluated using Cohen's Kappa Statistic (Kappa), area under receiver operating characteristic curve (AUC) and error rate. We found a highly significant quadratic effect of species prevalence on the four modelling techniques' performance. Kappa and AUC were less depended on the species prevalence, making them a better measure. The best performance (Kappa and AUC) was reached when predicting species with an intermediate prevalence (e.g. 0.4–0.6). The four modelling techniques significantly yielded different performances ($p < 0.01$), of which ANN performed generally better when using the complete prevalence range (i.e. 0.0–1.0) and the lower prevalence range (i.e. <0.1). However, the four techniques similarly performed when predicting species with a higher prevalence range (i.e. ≥0.3). Our results provide useful insights into the application of modelling techniques in predicting species occurrence and how their performance varies for species with different prevalence ranges. We suggest that the selection of appropriate modelling techniques should carefully take into account the species prevalence, particularly in the case of rare and generalist species.



## 1. Introduction

Various modelling techniques have been widely implemented in different ecological systems, e.g. terrestrial, freshwater lentic and lotic, and marine ecosystems (Guo et al., 2015; Lek et al., 1996; Lencioni et al., 2007; Park et al., 2003; Schröder et al., 2007). The techniques applied are generally used to investigate or to explain the relationship between the occurrence or abundance of studied species and environmental variables or to predict the relationships being measured (Boets et al., 2013; Goethals et al., 2007). The use of modelling techniques to combine both explaining and predicting such relationships is also commonly applied (Call et al., 2016; Roura-Pascual et al., 2009).

The performance of data-driven predictive models is affected by several factors including species prevalence (Brotons et al., 2004; Hernandez et al., 2006; Stokland et al., 2011). In most cases, models predicting species which have unequal occupied and unoccupied samples/sites result in a low performance. With a species having a high prevalence, models tend to become better at predicting the presence of that species, and vice-versa for less occurring species (McPherson and Jetz, 2007). Both cases consequently lead to a low model performance when considering the correct prediction of both the presence and absence of a species. Moreover, it has been demonstrated that species prevalence affects the performance of models in a nonlinear way. For example, Guo et al. (2015) and Manel et al. (2001) reported the nonlinear effect of species prevalence on the performance of models predicting the occurrence of fish and macroinvertebrates. A similar finding has also been revealed for

* Corresponding author at: Université de Toulouse, Laboratoire Evolution & Diversité Biologique, UMR 5174, CNRS – Université Paul Sabatier, 118 route de Narbonne, 31062, Toulouse cédex 4, France.
*E-mail address:* sorsim.ratha@gmail.com (R. Sor).

http://dx.doi.org/10.1016/j.ecolmodel.2017.03.006
0304-3800/

models predicting the distribution of plants and birds (Allouche et al., 2006; McPherson et al., 2004).

Applications of predictive models have provided knowledge and understanding of the ecology and behaviour of studied taxa, which could support decision making, management and conservation planning. For instance, Chen et al. (2015) used different predictive models as an assessment approach to explain and predict the success of invasive species in China. In addition to the increased use of predictive models, an ensemble modelling framework is recommended when aiming to identify important factors influencing model performance (Araújo and New, 2007). With the ensemble modelling approach, some modelling techniques such as Random Forest and Artificial Neural Networks are found to yield a better predictive performance (Grenouillet et al., 2011; Guo et al., 2015; Segurado and Araujo, 2004). However, although there have been studies assessing the performance of predicting models from an ensemble modelling framework, many have not considered analysing the performance based on a complete prevalence range nor comparing the performance based on different prevalence ranges.

The Lower Mekong Basin (LMB) which is known for its high biodiversity (Sodhi et al., 2004) is a breeding ground of numerous endemic, threatened and endangered species of fish, molluscs and crustaceans (Davidson et al., 2006; Zalinge and Van Thuok, 1998). Therefore, it is useful to get more insight into this region based on predictive models which are applicable for different taxonomic groups inhabiting this particular area. To date, the data covering a large spatial scale of the LMB is only available for fish and macroinvertebrates, which were collected by the Mekong River Commission (MRC). The fish data were collected only from the main channel (Poulsen and Viravong, 2001), while macroinvertebrates were collected from both the tributaries and the main channel (Dao et al., 2010). In this study, we used the macroinvertebrate data, sampled over 5 successive years (2004–2008), to build predictive models, which can provide insights on a wide range of keystone species occupying the LMB as well as the neighbouring regions.

The objectives of the present study are to utilize different modelling techniques to 1) predict the occurrence of macroinvertebrate species in the LMB and analyse how the species prevalence (i.e. the occurrence rate of each species in the total samples) affects the behaviour of modelling techniques' performance, and 2) compare the performance of the applied techniques based on the complete prevalence range (i.e. 0.0–1.0), and based on different prevalence ranges (i.e. at a 0.1 interval).

## 2. Methods

### 2.1. Data collection and processing

Benthic macroinvertebrates were sampled at 63 sampling sites along the main channel of the LMB and its tributaries by the MRC. This sampling was carried out once a year in March during the dry season from 2004 to 2008. To obtain as much information as possible on macroinvertebrates inhabiting the main river and the tributaries, the MRC collected samples at three locations from the benthic zone of each sampling site: near the left and right banks, and in the middle of the rivers. At each location, a minimum of three samples (where inter-sample variability is low, e.g. tributaries) to a maximum of five samples (where inter-sample variability is higher, e.g. the main channel and the delta) were collected using a Petersen grab sampler. With the grab which has a sampling area of 0.025 $m^2$, four sub-samples were taken and pooled to give a single sample covering a total area of 0.1 $m^2$. In total, between nine (3 samples × 3 locations) and fifteen (5 samples × 3 locations) pooled samples were collected at each sampling site. Each pooled sample was rinsed using a sieve (0.3 mm mesh size). In the field, samples were sorted and then preserved by adding 10% formaldehyde to obtain a final concentration of about 5%. In the laboratory, the samples were identified to the lowest level possible and counted using a compound microscope (40–1200 magnification) or a dissecting microscope (16–56 magnification). The abundance data of macroinvertebrates per sample (a total area of 0.1 $m^2$) was averaged across all samples (between 9 and 15 samples) collected from each sampling site.

At the sampling site, geographical coordinates and altitude were determined with a GPS (Garmin GPS 12XL). All physical-chemical variables were measured at the three locations where macroinvertebrates were sampled. River width was measured in the field using a Newcon Optik LRB 7 × 50 laser rangefinder, and the river depth was measured using a line metre. Water temperature, dissolved oxygen, pH and water conductivity were measured using a handheld water quality probe (YSI 556MP5). To get a more reliable determination of each variable, the measurement reading was taken at the surface (0.1–0.5 m) and at a depth of 3.5 m or at a maximum depth of the river (wherever less than 3.5 m) and then the average value was recorded for each location. Water transparency was measured with a Secchi disc by lowering it into the water and recording the depth at which it was no longer visible (Dao et al., 2010). The recorded data of each physical-chemical variable was based on the averaged value across the three sampling locations of each site. The distance from the sea was measured by drawing a line from the sea to each locality using GIS-software (ArcGIS version 10.0).

A total of 108 samples were collected from the 63 sampling sites (Fig. 1). Because of unequal sampling efforts (i.e. unequal and different number of samples at each site during the 5-year sampling period) and missing values of environmental variables, we used median values from the collected data to represent each site in our analyses, as suggested by McCluskey and Lalkhen (2007). Therefore, 63 samples remained for the analyses. In total, 299 taxa were obtained from the dataset, of which 131 taxa were insects, 98 were molluscs, 38 were crustaceans and 32 were annelids. The most commonly identified insects belonged to Diptera (37 taxa), Ephemeroptera (32), Odonata (22) and Trichoptera (20). For molluscs, Caenogastropoda (50 taxa), Unionida (18) and Veneroida (15) were represented the most. Most crustaceans belonged to Palaemonidae (10 taxa) and Corophiidae (6 taxa), while most annelids belonged to Naididae (15 taxa) and Nereididae (5 taxa). The detailed information of taxonomic resolution is provided in the Supplementary data Appendix A.

The abundance data of macroinvertebrates from the 63 sites were converted to presence-absence data to analyse how species prevalence (presence/absence) affects the performance of predictive models. Species prevalence was defined as the occurrence rate of each species in the total samples. The species prevalence of a species is an index ranging from 0 to 1, indicating the lowest to highest occurrence rate of that species over all samples. The obtained prevalence values from all macroinvertebrate species formed a complete prevalence range for the present study. For a later analysis, the complete prevalence range was grouped into different ranges based on an interval of 0.1. In other words, the species having a prevalence value between 0.0 and 0.1 were aggregated in a group, and the species having a prevalence between 0.1 and 0.2 were aggregated in another group, and so forth (see Appendix A in Supplementary material).

### 2.2. Predictions

In our predictive models, we used the presence/absence of each species as the response variable. The measured environmental variables used as the input predictors were: altitude, river width, river