Article

Taxonomic identification of hoverfly specimens using neural network and gradient boosting machine techniques

Dunja Popović,**Vuk Popović**, **Nevena Velicković**, **Ante Vujić**
Department of Biology and Ecology, Faculty of Sciences, University of Novi Sad, Novi Sad, Serbia
E-mail: dunja.popovic@dbe.uns.ac.rs, vukmpopovic@gmail.com

Received 10 March 2020; *Accepted 15 April 2020*; *Published 1 September 2020*

Abstract
The correct identification of single specimens on a particular area has great importance in establishing appropriate biodiversity protection programs. Species of the genus *Merodon* Meigen, 1803 (Diptera, Syrphidae) represent important pollinators that are particularly associated with the pollination of wild and cultivated bulbous plants, both wild and cultivated. In order to contribute to a taxonomic issue of separating two cryptic, sibling hoverfly species of *M. avidus* species complex, we programmed and trained specific prediction model that was able to specify to which of two assumed species (*M. avidus* or *M. moenium*) each database specimen belongs. Using two ML techniques (artificial neural network (ANN) and gradient boosting machine (GBM)), we created two separable models, depending on a variable used for a prediction (Model 1 - modelling based on a geographic variable, Model 2 - modelling based on a temporal variable). Moreover, each model was trained and tested with different data sets, resulting in a different predictive accuracy. While ANN modelling showed a higher percent of correct determination when using surrogate information than when using reduced (basic) data set, GBM modelling has given a quite stable result through all three data types. In both ML approaches, comparing Model 1 and Model 2 results showed that prediction based on a temporal variable (day, month and a year of specimen sampling) reached a better predictive performance than a prediction based on a longitude and latitude, on all data sets. This led us to the conclusion that information about the time of sampling was more useful for creating desired determination key with artificial intelligence algorithms than information about longitude and latitude of sampling localities. Therefore, we suggest that time of activity of adult specimens could have been of greater importance in the differentiation of *M. avidus* and *M. moenium* species from a common ancestor. The environmental factors and selective forces connected with the season might have had a more important role in *M. avidus* / *M. moenium* speciation, compared to environmental factors / selective pressures connected with the geographic position of their activity. The demonstrated modelling represents a positive signal in the field of potential implementation of these systems as support in the initial determination *Merodon* specimens. We suggest it´s potential use as technical support in old and partially unreliable databases, in determination of fresh sampled specimens as well as in finding the most efficient sampling strategies.

Keywords *Merodon avidus* complex; neural network; gradient boosting machine; sibling species.

Computational Ecology and Software
ISSN 2220-721X
URL: http://www.iaees.org/publications/journals/ces/online-version.asp
RSS: http://www.iaees.org/publications/journals/ces/rss.xml
E-mail: ces@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

The last 20 years brought the sophistication in the types of statistical models applied in different biological issues, as well as in the development of a wide variety of algorithms particularly suited to prediction. These machine learning (ML) methods include neural nets, ensembles of trees and support vector machines. Those systems are now widely used in ecology, for analysis of morphological relationships (Clarke and Johnston, 1999), population trends (Fewster, 2000), and for predicting the distributions of species (Buckland and Elston 1993).

Artificial neural networks (ANN) represent a highly efficient modern-day tool for creating different models, especially in cases where we are faced with a complex and/or unknown relationship between used data. They are designed for the universal and flexible functional estimation of any type of information (Lek and Guégan, 1999). Model creation consists of two main parts: training phase (presentation of input-outputs examples to networks) and a test phase (testing a prediction, formed on the basis of the knowledge from a test phase). The first biological implementation of ANN was documented in a field of molecular biology and medicine (Lerner et al., 1994; Albiol et al., 1995; Faraggi and Simon, 1995; Lo et al., 1995) with later focus on ecological researches (Lek et al., 1996; Worner and Gevrey, 2006; Zhang and Wei, 2009; Zhang, 2010, 2011).The most popular ANN type today is a backpropagation network (BPN), based on a same-named algorithm, whose architecture is composed of one input layer, one or more "hidden" layers and one output layer.

The search for the best model which would describe already existing ecological patterns or for predicting the new ones, have placed the models based on classification and regression into one of the most important prediction methods in a modern science (De'ath and Fabricius, 2000; Vayssières et al., 2000; De'ath, 2002). Boosted regression trees (BRT) is a ML technique which combines the strengths of two algorithms: regression trees (models that relate a response to their predictors by recursive binary splits) and boosting (an adaptive method for combining many simple models to give improved predictive performance) (Elith et al., 2008). GBM (gradient boosting machine) represents a superior BRT tool, characterized by the ease of use, ease of parallelization and impressive predictive accuracy. Preparation of candidate predictors is simplified because predictor variables can be of any type (numeric, binary, categorical, etc.), model outcomes are unaffected by monotone transformations and different scales of measurement among predictors and irrelevant predictors are seldom selected (Elith et al., 2008). Trees in GBM, and neural networks too can accommodate missing data in predictor variables by using surrogates (Breiman et al., 1984). Our estimation was that this feature could be of great importance for the given research.

The hoverflies stand next to bees in importance as pollinators of wild flowers and crops (Proctor et al., 1996). *Merodon* species are particularly associated with the pollination of bulbous plants, both wild and cultivated (Petanidou, 1991). The genus *Merodon* Meigen, 1803 (Diptera: Syrphidae: Merodontini) has become the largest European hoverfly genus due to several recent studies describing many new taxa (Marcos-García et al., 2007; Popov, 2010; Radenković et al., 2011; Vujic et al., 2007, 2012, 2013a, 2013b, 2015). A considerable number (37 of 57 species in southeastern Europe) of the taxa are morphologically and/or genetically cryptic, with restricted distributional ranges in particular mountain ranges or islands (Vujic et al., 2016). Taxa of the *Merodona vidus* complex have been the subject of many studies in the last decade due to perceived taxonomic difficulties (Milankov et al., 2001, 2009; Ståhls et al., 2009). Species of the *M. avidus* complex are not distinguishable by traditional visual identification of the structures of malegenitalia under a stereomicroscope (Ačanski et al., 2016). The *M. avidus* complexis characterized by a considerable morphological variability, especially in the coloration of the antennae, thorax, abdomen and legs (Popović et al., 2015) and, despite the numerous studies on the subject, difficulties in distinguishing the species of this complex based on morphological characters remain. Spring generations of *Merodona vidus* (Rossi, 1790) are

very similar to those of *M. moenium* (Wiedemann, 1822) based on external morphology and, so, are easily confused using existing diagnostic features (Milankov et al., 2001).Analysis of COI (cytochrome-oksidase I gene) barcodes revealed the presence of separate species *M. megavidus* (Ačanski et al., 2016) and *M. ibericus* (Popović et al., 2015; Ačanski et al., 2016) in the complex. However, all previous phylogenetic researches failed to discriminate taxon *M. avidus* from taxon *M. moenium*, using this DNA marker (Popović et al., 2014, 2015; Ačanski et al., 2016). At the same time, biochemical markers, implemented through the analysis of allozyme variability, successfully confirmed the assumption that *M. avidus*and *M. moenium* represent two different species (Popović et al., 2015). According to the same research, *M. avidus* and *M. moenium* should be declared as two sibling species, whose separation from the rest of the complex occurred the most recently in evolutionary history, which explains the weak resolution of DNA COI marker in their delineation.

Considering the fact that large, decades-old biological databases are often confronted with a loss and/or initial lack of some information, the missing data prediction would have obvious advantages. The loss-of-information problem has also been noticed in a big hoverfly database of Department of biology and ecology, Faculty of Sciences in Novi Sad, which was the main data source for this research. Also, previous researches on this group obtained that a correct differentiation of *M. avidus* and *M. moenium* species from each other represents taxonomically biggest challenge. Those facts brought us to the idea of implementation GBM and ANN advantages on *M. avidus* species complex, by creating and training a model that would be able to perform an accurate taxonomic identification of critical *M. avidus* and *M. moenium* specimens in a database. The aim of this research was to, based on one or more properly chosen variables, develop a model that would be able to correctly determinate species of each specimen, with a great probability. That would demonstrate the potential application of those ML systems in the initial identification of fresh hoverfly samples, in cases of ambiguities connected with their identification.

2 Material and Methods

Although *M. avidus* and *M. moenium* have generally different ecological preferences, those species are only partially geographical isolated. *M. avidus* (generally known as a Mediterranean species) inhabits mainly the Mediterranean basin, but also some continental localities (Fig. 1a). At the other hand, *M. moenium* is generally known as a mountain species, which is widespread in continental parts of Europe but can be also found on few Mediterranean localities (Fig. 1b). Although presence of sympatric localities makes a correct identification of morphologically similar species even more difficult, it is important to emphasize the existence of temporal divergence of these two species, on all continental localities where they occur sympatrically (Popović et al., 2015). Since *M. moenium* occurs as a spring-early summer generation and *M. avidus* as late summer-autumn generation of those localities, it can be assumed that the different temperature preferences prevent their temporal coexistence and result in different seasons of their activity. Therefore, possible importance of temporal factor in a genetic divergence of *M. avidus* and *M. moenium* species can be suggested. This brought us to the decision of including a temporal variable in creating a model for determination of those two species in one sample.

(a)

(b)

Fig. 1 Map of distribution of *M. avidus* (a) and *M. moenium* (b) species.

The material used for testing and presenting an application of species determination modelling in a determination of ambiguous species was taken from a base of Laboratory for Research and Biodiversity Protection, Department of Biology and Ecology, Faculty of Sciences, University of Novi Sad. Basic dataset was created from 1183 specimens of *M. avidus* (Rossi, 1790) and 859 specimens of *M. moenium* (Wiedemann, 1822) from 22 countries (Table 1, Appendix).The specimens were previously identified based on their morphology, sampling locality and a season of their adult activity. In some stages of modelling, we expanded a data source on 3490 *M. avidus* and *M. moenium* entries. The newly added entries consisted of specimens with a lack of some type of information. The information which was missing was imputed again using the ANN. The new entries were added in order to test a model´s ability of prediction in conditions of partially missing data. In order to train a model and achieve a prediction of determination with maximum accuracy, we used available taxonomic information of each entry (name of the species and sex for each specimen), together with one predictor, in each of two model cases. Longitude/latitude of each sampling locality was taken as a predictor for a Model 1, while a time (day, month and year) of sampling was introduced as a predictor in a second case (Model 2). Therefore, we were able to test and compare a model prediction based on a geographic predictor with those based on a temporal predictor and bring conclusions about a potential future implementation of these variables in a determination of specimens for which straight-forward identification could not be performed. This has been done in a frame of each of the two techniques (ANN and GBM), through different types of data preparation.

Since the data preparation represents the most important step in working with all kind of ML algorithms, this phase has been performed with big caution. In determination modelling with GBM method, each of two models was projected on a different data set: raw data set (all 3490 entries), basic data set (created from a raw data set, deleting all the entries for which some kind of information connected to predictors was missing), as well as on a raw data set where the missing data was compensated with surrogate data. In the last case, the unknown information for some specimens is predicted and automatically entered the base, relying on all other available data and the use of an appropriate algorithm. It is important to emphasize that the GBM method allows working with missing data through an implicit data conversion, for which is a neural network not capable. For that reason, species determination modelling with ANN has been done on only two data cases: on basic (reduced) data set (only entries for which all data were already present), as well as on 3490 data entries (filled in with surrogate, automatically predicted, data).

Each of different model variants, defined through data set choice and a choice of a predictor, was implemented into the appropriate artificial intelligence platform. The result, expressed in numerical values-probability of correct prediction (ANN method) or Log Loss value (GBM method), was meant to enable an estimation and comparison of created prediction models.

Both ANN and GBM models were developed using the functionality of R-Studio software (RStudio Team, 2015) and different library sources ("Neuralnet" library for ANN model programming and "gbm" library for GBM programming). Backpropagation algorithm used for neural network training is based on Levenberg-Marquardt approximation. This function, as a network training function, updates weight and bias values according to Levenberg-Marquardt optimization (Sapna et al., 2012). The performance function, for used backpropagation networks, is mean square error function (the average squared error between the network outputs and the target outputs). Afterwards, the model performance is evaluated and verified with the test datasets. Once the ANN prediction models are trained to a satisfactory level, and error rates are acceptable, they are used for prediction on other data.

Neural networks used in this research were standard BPNs, and their architecture was composed of one

input layer, 10 hidden layers and one output layer. Backpropagation process consists of calculating the error contribution of each neuron after processing a batch of data, with each neuron being able to individually introduce the corrections. This is done by changing the values of the weight coefficients on all its inputs, and this change is based on a set error value. The weight of each synapse is connecting a series of tuning neurons and an iterative process to achieve proper tuning (training data) is conducted. In each iteration there is additional fine-tuning, conducted with the backpropagation algorithm, to adjust to the desired level of prediction accuracy. Eventually, the network is tuned and when it is used for prediction it will predict with acceptably low error rates.

For creating a GBM model we used a "gbm" library from CRAN repository. This package contains an implementation of extensions to Freund and Schapire's AdaBoost algorithm and Friedman's gradient boosting machine. It also includes regression methods for least squares, absolute loss, *t*-distribution loss, quantile regression, logistic, multinomial logistic, and many other functionalities. The loss function used in creating the prediction was the Bernoulli distribution. Furthermore, the number of decision trees was set to 2500 trees. The steps taken in the gradient descent of boosting (parameter named – shrinkage) was set to 0.01. Since the higher value of steps means a faster convergence, the shrinkage needs to be balanced with iterations. In the given example, an optimum was achieved with a small shrinkage and an increase the number of iterations. The last parameter "minobsinnode" (the number which specifies the minimum number of observations in the terminal nodes of the trees) was set to 20.

3 Results

For modelling a prediction of taxonomic estimation of specimens, we made two models, each referring to the predictor variable used for a determination. Within an ANN modelling, each of these models was trained on two different data sets, which resulted in a different probability rate, i.e., predictive accuracy. Model 1 (prediction with longitude/latitude as predictor) achieved 65.4% gains on a reduced data set (1183 entries M. *avidus* and *M. moenium*). After data filling with surrogate information, resulting in 3490 entries, the prediction ability of Model 1 was raised on 75.5% (Table 1). Model 2 (prediction with a day, month and year as a predictor) showed better predictive performance by training on the same two data sets. The achieved probability after modelling on a reduced data set was 86.8% and the same value was slightly higher (87.8%) after modelling on the expanded database, with the previous prediction of missing data (Table 1).

Apart from the artificial neural network frame, the idea of species determination modelling was also performed through the gradient boosting machine (GBM) method. As well as in ANN approach, we tested the predictive accuracy of two models by determination of each sample in a database as *M. avidus* or *M. moenium*. Since approaches based on regression trees allow the processing of raw databases, despite eventual lack of some information, the GBM modelling was performed on three data sets: on a raw database (3490 entries including some missing data), as well as on the same two data sets used in ANN modelling (reduced base and the raw base completed with surrogate data). The result of GBM modelling was expressed through Logarithmic Loss (Log Loss) value which measures the performance of a classification model, where the prediction input is a probability value between 0 and 1. Since the lower Log Loss means better predictions, a goal of machine learning models is to minimize this value, which will be increased as the predicted probability diverges from the actual label. Within GBM method, Model 1 (longitude and latitude of sampling locality as a predictor) resulted in the following Log Loss values: 0.894 for a model trained on a raw data set, 0.831 for a model trained on a basis data set (reduced base) and 0.941 for a model trained on a raw data set filled with predicted surrogate data (surrogated data set). Model 2 (day, month and a year of sampling as a predictor), resulted in a better predictive performance in all three cases. Log Loss values were 0.695 for a model trained

on a raw data set, as well as for a model trained on a reduced data set, and 0.694 for a model trained on a surrogated data set (Table 2).

Table 1 Results of artificial neural network modelling. The values show the probability of accurate prediction of determination in each model/data case.

	Probability (%)	
	Basic (reduced) data set (1183)	Surrogated data set (3490)
Model 1 (long/lat)	65,4	75,5
Model 2 (day, month, year)	86,8	87,8

Table 2 Results of gradient boosting machine modelling. The numbers represent Log Loss values. Lower Log Loss value indicates a better prediction.

	Log Loss		
	Raw data set(3490)	Basic (reduced) data set (1183)	Surrogated data set (3490)
Model 1 (long/lat)	0.894	0.831	0.941
Model 2 (day, month, year)	0.695	0.695	0.694

4 Discussion

Determination of morphologically inseparable taxa represents a remaining challenge, which is, within M. *avidus* species complex, the most noticeable in case of *M. avidus*and *M. moenium* species. The species determination modelling with modern artificial intelligence approaches, performed through an appropriate algorithm of ANN and GBM systems, illustrate the possibility of its implementation in every-day taxonomic issues.

Although genetic confirmation of species differentiation represents the most certain method in a taxonomic identification of specimens, this kind of determination demands laboratory conditions and, despite modern techniques of DNA sequencing, a considerable time effort. Moreover, the results based on mitochondrial DNA marker in previous studies explained that these gene loci cannot clearly delineate sibling species *M. avidus* and *M. moenium*, suggesting the necessity of including additional biochemical loci in genetic analysis (Milankov et al., 2009; Popović et al., 2014, 2015; Ačanski et al., 2016). Since only the integrative taxonomy approach can absolutely delimitate those two species, a chance of getting a quick preliminary insight into the taxonomic status of samples could be useful for the researchers and their sampling logistic. The models represented in this paper were trained on data collected from previously determined specimens (based on their morphological, genetic and/or ecological features) and created for delineation of species *M. avidus* and *M. moenium*, on order to support an identification of ambiguous specimens. These models have demonstrated their ability to determine which species each specimen belongs to, based on the chosen predictor. We suggest the potential use of the demonstrated modelling as a support in the initial determination of fresh sampled biological specimens as well as in solving ambiguities in old and partially unreliable databases (in our case ambiguously marked *M. avidus* / *M. moenium* specimens). In the second case, it is important to emphasize that artificial intelligence should not be understood and taken as a definite and completely reliable determination tool. It should be rather accepted as technical support for the taxonomists themselves, in special and complicated

situations of species identification, by helping them in deciding how to mark the most demanding specimens. In future, prediction of species could also contribute to sampling logistic, by helping in finding the most efficient sampling strategies. In other words, the ANN and / or GBM predictions could be included in the estimation of the probability of finding a certain species on a particular locality (defined by its coordinates).

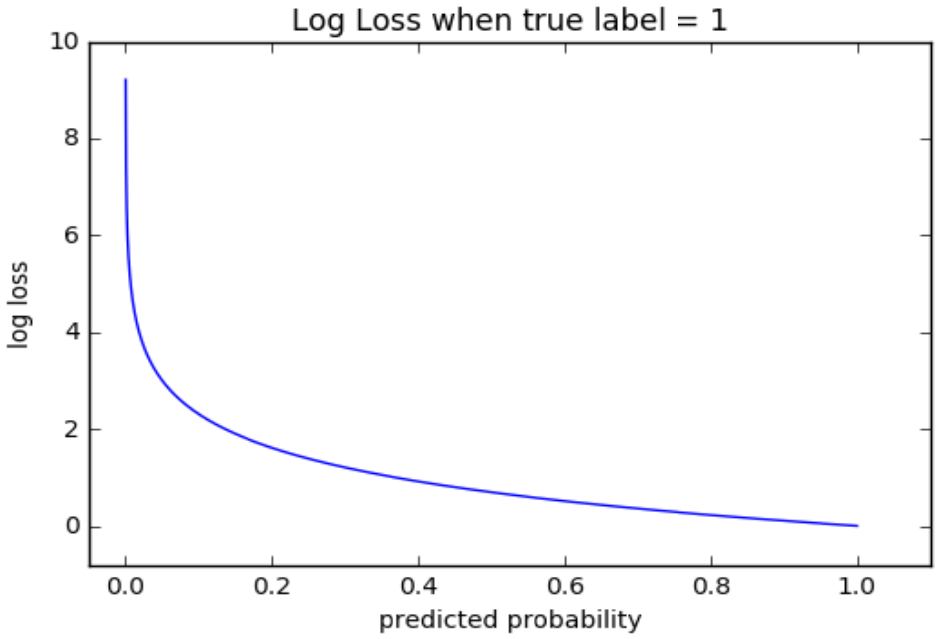

Fig. 2 The range of possible log loss values given a true observation (isDog = 1). As the predicted probability approaches 1, log loss slowly decreases. As the predicted probability decreases, however, the log loss increases rapidly.

Comparing the results of ANN modelling from different data sets showed a higher percent of correct determination in the model created using surrogate information than in the model that was trained on a reduced (basic) dataset. The predictive performance in Model 1 (modelling based on a geographic variable) was raised from 64.5% (basic dataset) on 75.5% (expanded, surrogated database). The same trend was documented in Model 2 (modelling based on a temporal variable), but it was less expressed: the accuracy of determination was raised from 86.8% (basic dataset) on 87.8% (surrogated database). This kind of result encourages the usage of surrogate data, created from the artificial neural network itself, in the prediction of species determination. At the other side, predictive modelling with GBM technique has given a quite stable result through all three data types (raw database, reduced database and database completed with surrogate data). The obtained Log Loss values were in the range from 0.83 to 0.94 (Model 1) and around 0.69 (Model 2). Considering the approximation between Log Loss values and the predicted probability, all GBM results were low enough to match the probability level of over 90% (Fig. 2). Since a lower Log Loss means better predictions, it is obvious that a prediction based on a temporal variable (day, month and a year of specimen sampling) reached a better predictive performance than a prediction based on a longitude and latitude, on all data sets. The same has been noticed in results achieved with ANN models, on both data sets. Therefore, we concluded that information about the time of sampling was more useful for creating desired determination key with artificial intelligence algorithms than information about longitude and latitude of sampling localities. This kind of results suggests that time of activity of adult specimens could have been of greater importance in the differentiation of *M. avidus* and *M. moenium* species from a common ancestor. In other words, we suggest that

environmental factors and selective forces connected with season might have had a more important role in *M. avidus* / *M. moenium* speciation, comparing to those environmental factors/selective pressures that were connected with the geographic position of the area of activity. This assumption is in accordance with the results of the enzyme variability analysis of these two species (Popović et al., 2015). In that study, enzyme variability confirmed the presence of temporal divergence of *M. avidus* and *M. moenium* species at the same locality, indicating that difference in seasonal occurrences of certain populations might lead to genetic differentiation of two sibling species.

5 Conclusion

Creating efficient sampling strategies and appropriate biodiversity protection programs requires an accurate identification of single specimens on a particular area. The creation of a specific computer model with the ability of highly correct identification of *M. avidus* and *M. moenium* specimens confirmed it's usage in described taxonomic issues. The demonstrated modelling represents a positive signal for implementation of those sorts of systems as support in a determination of biological specimens. The results of this research suggest that information about the exact time of activity of adult *Merodon* specimens was more important for their correct identification than information about geographic coordinates of their localities. Therefore, we emphasized the potential importance of environmental factors connected with a season in genetic differentiation of those sibling species.

Appendix: The material used for species determination modeling

Species	Country	Number of specimens
Merodon avidus	Albania	1
	Bosnia and Herzegovina	2
	Bulgarien	10
	Montenegro	108
	France	146
	Greece	422
	Croatia	55
	Italy	185
	Israel	2
	Cyprus	10
	FYR Macedonia	52
	Serbia	182
	Turkey	8
Merodon moenium	Andora	1
	Bosnia and Herzegovina	2
	Montengro	115
	Denmark	4
	France	77
	Greece	29
	Netherland	1

Croatia	24
Italy	99
FYR Macedonia	18
Germany	5
Romania	3
Slovakia	7
Slovenia	2
Serbia	443
Switzerland	25
Sweden	4

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Ačanski J, Vujić A. 2016. Integrative taxonomy in defining species boundaries in *Merodon avidus* complex (Diptera, Syrphidae). European Journal of Taxonomy, 237: 1-25

Albiol J, Campmajo C, Casas C, Poch M. 1995. Biomassestimation in plant cell cultures: a neural network approach. Biotechnology Progrress, 11: 88-92

Breiman L, Friedman JH, Olshen RA, Stone CJ. 1984. Classification and Regression Trees. Wadsworth International Group, Belmont, CA, USA

Buckland ST, Elston DA. 1993. Empirical models for the spatial distribution of wildlife. Journal of Applied Ecology, 30: 478-495

Clarke A, Johnston NM. 1999. Scaling of metabolic rate with body mass and temperature in teleost fish. Journal of Animal Ecology, 68: 893-905

De'ath G. 2002. Multivariate regression trees: a new technique for constrained classification analysis. Ecology, 83:1103-1117

De'ath G. iFabricius KE. 2000. Classification and regression trees: a powerful yet simple technique for the analysis of complex ecological data. Ecology, 81: 3178-3192

Elith J, Leathwick JR, Hastie T. 2008. A working guide to boosted regression trees. Journal of Animal Ecology, 77: 802-813

Faraggi D, Simon R. 1995. A neural network model for survival data. Statistics in Medicine, 14: 73-82

Fewster RM, Buckland ST, Siriwardena GM, Baillie SR, Wilson JD. 2000. Analysis of population trends for farmland birds using generalized additive models. Ecology, 81: 1970-1984

Lek S, Delacoste M, Baran P. 1996. Application of neural networks to modelling nonlinear relationships in ecology. Ecological Modelling, 90: 39-52

Lek S, GoéganJF. 1999. Artificial neural networks as a tool in ecological modelling, an introduction. Ecollogical Modeling, 120(2-3): 65-73

Lerner B, Levinstein M, Rosenberg B, Guterman H, Dinstein I, Romem Y. 1994. Feature selection and chromosomes classification using a multilayer perceptron neural network. IEEE International

Conference On Neural Networks. 3540-3545, Orlando, Florida, USA

Lo JY, Baker JA, Kornguth PJ, Floyd CE. 1995. Application of artificial neural networks to interpretation of mammograms on the basis of the radiologists impression and optimized image features. Radiology, 197: 242-242

Marcos-García MA, Vujić A, Mengual X. 2007. Revision of Iberian species of the genus *Merodon* Meigen, 1803 (Diptera: Syrphidae). European Journal of Entomology, 104: 531-572

Milankov V, Ludoški J, Ståhls G, Stamenković J, Vujić A. 2009. High molecular and phenotypic diversity in the *Merodon avidus* complex (Diptera, Syrphidae): cryptic speciation in a diverse insect taxon. Zoological Journal of the Linnean Society, 155: 819-833

Milankov V, Vujić A, Ludoški J. 2001. Genetic divergence among cryptic taxa of *Merodon avidus* (Rossi, 1790) (Diptera: Syrphidae). International Journal of Dipterological Research, 2: 15-24

Petanidou T. 1991. Pollination Ecology in A Phryganic Ecosystem. PhD Thesis (in Greek, with English summary). Aristotelian University, Thessaloniki, Greece

Popov GV. 2010. *Merodon alexandri* spec. nov. - a new species of hoverfly (Diptera: Syrphidae) from the northern Black Sea Region. Studiadipterologica, 16: 133-151

Popović D, Ačanski J, Đan M, Obreht D, Vujić A, Radenković S. 2015. Sibling species delimitation and nomenclature of the *Merodon avidus* complex (Diptera: Syrphidae). European Journal of Entomology, 112 (4): 790-809

Popović D, Đan M, Šašić L, Šnjegota D, Obreht D, Vujić A. 2014. Usage of different molecular markers in delimitation of cryptic taxa in *Merodon avidus* species complex (Diptera: Syrphidae). Acta Zoologica Bulgarica, 7: 33-38

Proctor M, Yeo P, Lack A. 1996. The Natural History of Pollination. The New Naturalist Series. Harper & Collins Publishers, New York, USA

Radenković S, Vujić A, Ståhls G, Pérez-Bañón C, Rojo S, Petanidou T, Šimić S. 2011. Three new cryptic species of the genus *Merodon*Meigen (Diptera: Syrphidae) from the island of Lesvos (Greece). Zootaxa, 2735: 35-56

RStudio Team. 2015. RStudio: Integrated Development for R. RStudio Inc., Boston, MA, USA. http://www.rstudio.com/

Sapna S, Tamilarasi A, Parvin Kumar M. 2012. Backpropagation learning algorithm based on Levenberg Marquardt algorithm. Computer Science Inform. Technology, 2, 393-398

Ståhls G, Vujić A, Perez-Banon C, Radenković S, Rojo S, Petanidou T. 2009. COI barcodes for identification of *Merodon* hoverflies (Diptera, Syrphidae) of Lesvos Island, Greece. Molecular Ecology Resources, 9: 1431-1438

Vayssières MP, Plant RE, Allen-Diaz BH. 2000. Classification trees: an alternative non-parametric approach for predicting species distributions. Journal of Vegetation Science, 11: 679-694

Vujic A, Perez-Banon C, Radenković S, Ståhls G, Rojo S, Petanidou T, Šimić S. 2007. Two new species of genus *Merodon* Meigen, 1803 (Syrphidae, Diptera) from the island of Lesvos (Greece), in the eastern Mediterranean. Annales de la Societe Entomologique de France. 43: 319-326

Vujic A, Radenković S, Ståhls G, Ačanski J, Stefanović A, Veselić S, Andrić A, Hayat R. 2012. Systematics and taxonomy of the *ruficornis*group of genus *Merodon* Meigen (Diptera: Syrphidae). Systematic Entomology, 37: 578-602

Vujic A, Ståhls G, Ačanski J., Bartsch H, Bygebjerg R, Stefanović A. 2013a. Systematics of Pipizini and taxonomy of European PipizaFallen: molecular and morphological evidence (Diptera: Syrphidae), Zoologica Scripta, 3: 288-305

Vujic A, Radenković S, Likov L, Trifunov S, Nikolić T. 2013b. Three new species of the *Merodon nigritarsis* group (Diptera: Syrphidae) from the Middle East. Zootaxa, 3640: 442-464

Vujic A, Radenković S, Ačanski J., Grković A, Taylor M, Şenol SG, Hayat R. 2015. Revision of the species of the *Merodon*nanusgroup (Diptera: Syrphidae) including three new species. Zootaxa, 4006(3): 439-462

Vujic A, Petanidou T, Tscheulin T, Cardoso P, Radenković S, Ståhls G, Baturan Ţ, Mijatović G, Rojo S, Pérez‐Bañón C, Devalez J. 2016: Biogeographical patterns of the genus *Merodon* Meigen, 1803 (Diptera: Syrphidae) in islands of the eastern Mediterranean and adjacent mainland. Insect Conservation and Diversity, 9: 181-191

Worner SP, Gevrey M. 2006. Modelling global insect pest species assemblages to determine risk of invasion. Journal of Applied Ecology, 43: 858-867

Zhang WJ. 2010. Computational Ecology: Artificial Neural Networks and Their Applications. World Scientific, Singapore

Zhang WJ. 2011. Simulation of arthropod abundance from plant composition. Computational Ecology and Software, 1(1): 37-48

Zhang WJ, Wei W. 2009. Spatial succession modeling of biological communities: a multi-model approach. Environmental Monitoring and Assessment, 158: 213-230