Towards the use of genetic programming in the ecological modelling of mosquito population dynamics

Noname manuscript No.
(will be inserted by the editor)

Towards the Use of Genetic Programming in the Ecological Modelling of Mosquitoes Population Dynamics

Irene Azzali · Leonardo Vanneschi · Andrea Mosca · Luigi Bertolotti · Mario Giacobini

the date of receipt and acceptance should be inserted later

Abstract Predictive algorithms are powerful tools to support infection surveillance plans based on the monitoring of vectors abundance. In this article, we explore the use of genetic programming (GP) to build a predictive model of mosquitoes abundance based on environmental and climatic variables. We claim, in fact, that the heterogeneity and complexity of this kind of dataset demand for algorithms capable of discovering complex relationships among variables. For this reason, we benchmark GP performance with the state of art of machine learning predictive algorithms. In order to provide a real exploitable model of mosquitoes abundance, we train GP and the other algorithms on mosquitoes collections from 2002 to 2005 and we test the predictive ability on 2006 collections. Results reveal that, among the studied methods, GP has the best performance in terms of accuracy and generalization ability. Moreover, the intrinsic feature selection and readability of the solution provided by GP offer the possibility of a biological interpretation of the model which highlights known or new behaviours responsible of mosquitoes abundance. GP reveals therefore to be a promising tool in the field of ecological modelling, opening the way to the use of vector based GP approach (VE-GP) which may be more appropriate and beneficial for the problems in analysis.

Keywords Ecological Modelling · Genetic Programming · Machine Learning · Regression

Irene Azzali (corresponding author) · Luigi Bertolotti · Mario Giacobini
DAMU - Data Analysis and Modeling Unit, Department of Veterinary Sciences, University of Torino, Italy
E-mail: irene.azzali,mario.giacobini,luigi.bertolotti@unito.it

Leonardo Vanneschi
NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal
E-mail: lvannesc@novaims.unl.pt

Andrea Mosca
Istituto per le Piante da Legno e l'Ambiente (IPLA), regional government-owned corporation of Regione Piemonte, Torino, Italy
E-mail: mosca@ipla.org

Acknowledgments. This work was partially supported by Fundação para a Ciência e a Tecnologia (FCT), Portugal, through project BINDER (PTDC/CCI-INF/29168/2017). This study was partially supported by Ministero dell'Istruzione, dell'Universitá e della Ricerca (MIUR) under the programme "Dipartimenti di Eccellenza ex L.232/2016" to the Department of Veterinary Science, University of Turin

1 Introduction

West Nile Virus (WNV) is an infectious disease, transmitted to people by the bite of an infected mosquito called the vector of the disease [9]. The virus causes neurological illnesses and death and no human vaccine is available [19]. It is commonly found in Africa, Middle East, North America and also in Europe. The first outbreak in Italy dates back to 1998, when 14 horses located in Tuscany were confirmed for WNV infection by laboratory analyses [17]. Later, in 2008, besides the largest outbreak, the first human case of WNV neuro-invasive infection in Italy was observed [12]. Since then, a constant circulation of the virus has been highlighted, and a national surveillance plan was established [5]. The aim of the plan is to quantify vector abundance in order to predict the emergence and the amplification of the virus. Therefore, understanding the relationship between vector dynamics and environmental and climatic variables facilitates by far the adaptation of control or eradication strategies. Predictive models of vector spread and abundance are valuable tools to fulfil this goal.

A Generalized Linear Mixed Model (GLMM) was used in [8] to predict *Cx. pipiens* amount, the species most responsible of WNV circulation in Europe [21], in Eastern Piedmont region in Italy. Additionally to the creation of a vector distribution map that identifies high risk area of the region, the model highlighted the most informative environmental determinants of the high abundance of *Cx. pipiens* mosquitoes. Modelling mosquitoes abundance is, however, a hot topic in the whole field of ecological research. In [7] a GLMM was used to evaluate the effect of climatic and ecological factors on the spatio-temporal dynamics of two main species vectors of West African Rift Valley Fever in Senegal. Other methods were used to predict mosquitoes population, including a Poisson regression model in [29] to examine the effect of off-season factors on East Coast vectors species and mathematical matrix models [6] to simulate the abundance of two vectors of the Yellow Fever Virus in Ivory Coast. These approaches undoubtedly showed good performance, providing information that can contribute to the development of more efficient surveillance plans. Nonetheless, they all *a priori* fix the structure of the relationship among ecological variables and mosquitoes abundance which is likely to be more complex due to the heterogeneous and complex data involved. In fact, data regarding mosquitoes abundance, and more in general data in the ecological domain, come from different sources and have different structures (time series, spatial data). This inherent complexity demands for more sophisticated techniques that can provide more realistic distribution models.

In such a framework, Genetic Programming (GP) [14] could potentially represent a promising approach to predict mosquitoes abundance, and deserves to be explored. There are few articles that exploit GP in ecology, specifically marine ecology, and their results look very promising. In [26] GP is used to identify which

environmental variables determine zooplankton abundance, while in [10] a similar approach is used to detect the drivers of planktonic population. As these articles claim, GP has the great capability of generating functions that estimate the target without any assumption on the function structure and on the dependencies among predictors. Although this methodology is different from the classical statistical methods, it still has the advantages of readability and interpretability of the resulting models. These features are of great importance in order to better align the modeling to the decision making process. For this reason, further investigation of GP is recommended in highly non linear problems coming from the ecological field.

In this article, we explore the application of GP in predicting mosquitoes abundance in Piedmont region based on climatic and environmental factors following [8]. A first investigation of GP on the described ecological problem was conducted in [22]. As in [22], also here GP results are compared with those obtained in [8], fitting again the GLMM model since we reprocess some data. The main novelty here is the benchmark of GP with respect to three popular state of art machine learning algorithms on regression problems, namely Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Multilayer Perceptron (MLP). These techniques have already been used in the ecological field [24,32,15] or are recognized as highly effective [30] and, as GP, they have the potential of discovering complex relationships. The comparison has the advantage of highlighting the pros and cons of GP with respect to common machine learning algorithms. Another important element of novelty provided in this paper is the exploration of GP as a concrete device to develop prediction maps. To explain the reason beyond this assertion, we have to define some terms. To create a model, three subsets of the dataset are always adopted: the training set that is used to fit a model, the validation set that is used to evaluate a fitted model and the test set used to assess the performance of a specific fitted model. In [22] the selected GP model used to predict future abundance (test set) was actually the best performing on future abundance among all the executed runs. In real world applications, this approach is not supposed to be used, since we do not have access to future data. Thus, in this work, the model is selected using the performance on validation, while the test set remains untouched during learning.

The paper is structured as follows: in Section 2 we explain the problem in analysis and we introduce the used dataset. In Section 3 we briefly describe all the studied techniques; the comparison between GP and the other methods, as well as the analysis of the best model found by GP, are reported in Section 4. Questions and issues that arise from these results are discussed in Section 5. Finally, Section 6 concludes the paper with suggestions for further research.

2 The dataset

We use the dataset produced by the Casale Monferrato Agreement for mosquitoes control from 2002 to 2006 in the context of the surveillance Piedmont region program. The study area covers the eastern part of Piedmont region, called Casale Monferrato, which offers a suitable environment for the proliferation of the local *Cx. pipiens* species. Mosquitoes were weekly collected from 36 CO_2-baited traps from May to September with a total of 20 collections per year for each trap. We

consider the same predictive variables of [8], selected as the most informative about the abundance of mosquitoes. The variable *TWEEK* is the average Land Surface Temperature (LST) from 8 to 15 days prior to the trapping day derived from the Moderate Resolution Imaging Spectro radiometer (MODIS) satellite (National Aeronautics and Space Administration, NASA [33]). *RAIN* represents the cumulative rainfall from 10 to 17 days prior to trapping, registered by the nearest weather station to the trap [1]. Both these variables catch the effect of climatic conditions on mosquitoes population abundance during a precise window of time. Vegetation changes deemed influential on mosquitoes dynamics are captured by the Normalized Difference Vegetation Index (*NDVI*), derived again from MODIS [33] as a 16 days average. While these variables change in each day of collection, some others are constant and inform about the environment surrounding the trap. *DISTU*, *DISTR*, and *DISTW* estimate the distance of sampling locations to the nearest urban center, rice field and woodland respectively. The area covered by rice field is registered in the variable *RICEA*. To spot the impact of altitude, the elevation of the trap location (*ELEV*) is included among predictors. The last variable involved is *SIN*, a sinusoidal curve with a phase of 1 year. It is an artificial informer of the seasonality of mosquitoes, reflecting the shape of mosquitoes abundances across the year. Its value, thus, has a peak in the first week of August where experts know there is a peak in mosquitoes abundance. The complete dataset used in this work is thus composed of 3600 observations, each one corresponding to a collection of mosquitoes in a precise day and trap. Every observation contains the values of the 9 predictors (*TWEEK*, *RAIN*, *NDVI*,*DISTU*, *DISTR*, *DISTW*, *RICEA*, *ELEV* and *SIN*) and the value of the dependent variable, or target, which is the number of mosquitoes collected.

To perform the experiments, we split the dataset into learning and test set, following the natural order of the years. Collections from 2002 to 2005 are used as the learning set to tune and train the algorithms, while collections of 2006 form the test set, therefore they represent unseen data, used to test the generalization performance of the predictive models. This approach is naturally determined by the real problem at hand: developing predictive models for mosquitoes abundance. Thus, we use data from the past (2002-2005) to train models and we assess their generalization ability by evaluating them on future predictions (2006).

3 GP and the other techniques involved

We use different tools and softwares to run the chosen algorithms, namely MATLAB R2018A, GPLab [28] and R. We do not provide a description of all the parameters, especially for the ones kept at the default value of the respective implementations. We remind that the goal of the paper is to explore GP applicability in ecological modelling giving standard benchmarks (ML techniques) to its results. Only some parameters that need to be manually inserted are tuned according to the following strategy: we propose a value for the parameter and we estimate the performance of the algorithm by running it 60 times on 60 different split of the learning set; in each run the algorithm is trained on 75% of the learning set and validated on the remaining 25%; the median result over the 60 runs on the 25% measures the performance of the algorithm using that particular value of the parameter. In the end, we use the configuration that allowed us to obtain the best

performance. In the following description, we specify which software is used and which parameters are tuned in each case.

3.1 Genetic Programming

We use a tree-based GP [14], where each tree is built combining a set of functions F with a set of terminal symbols T. In our case, T consists of the predictors described in Section 2, plus random constants r between 0 and 1 generated in runtime when building trees. Therefore $T = \{TWEEK, RAIN, NDVI, DISTU, DISTR, DISTW, RICEA, ELEV, SIN, r\}$. The functions set F includes the usual binary addition, subtraction and multiplication operators, plus a protected version of the division, known as `kozadivide`, that returns 1 when the denominator is equal to zero. Thus the set is $F = \{\texttt{plus}, \texttt{minus}, \texttt{times}, \texttt{kozadivide}\}$. Fitness is measured as the Root Mean Squared Error (RMSE) between the predicted values and the real mosquitoes abundances.

The computational tool used for GP experiments is GPLab [28], a public domain GP system implemented in MATLAB. The parameter setting consider is reported in the upper part of Table 1, and it corresponds to the default values provided by GPLab.

3.2 Generalized linear mixed model

The GLMM was the first modelling technique designed to approach our problem [8]. A GLMM is a statistical model that combines the characteristics of Generalized Linear Models (GLM) and mixed models [34]. Therefore, the target is allowed to follow any kind of distribution as a GLM, and fixed and random effects join the predictors to introduce population-average effects and subject-specific effects. We considered the same GLMM as the one selected in [8]. The process of selection consisted in the following steps. Firstly, in [8] a full model was outlined by selecting the environmental variables deemed as the most informative about mosquitoes abundance (*TWEEK*, *RAIN*, *NDVI*, *DISTU*, *DISTR*, *DISTW*, *RICEA*, *ELEV*, *SIN*), and defining two random effects: *RNDtrap*, which represented the spatial difference between traps (the subjects) and *SRNDtrap*, which represented the effect of space location on each trap. The authors then tested 169 GLMMs built with all the variables involved in the previous full model. The best one was chosen by the Deviance Information Criterion (DIC), an approximate model selection method which tries to explicitly balance model complexity with fit to the data [11]. The chosen model was used as the final GLMM model to be trained in order to predict mosquitoes abundance. The resulting expression is:

$$y = I + \beta_1 * RAIN + \beta_2 * TWEEK + \beta_3 * SIN + \beta_4 * ELEV + \beta_5 * DISTU + RNDtrap$$

where I is the intercept representing the fixed effect and y is the abundance of mosquitoes. All the analysis are performed using the R software package [2].

3.3 Random forest

A RF is an ensemble of regression trees that estimates the target by averaging individual tree predictions, in order to regularize the outputs, thus improving the generalization ability. In this work, we use a RF implemented in R [3], that follows the Breiman bagging idea [18] of trees construction. Instead of splitting each node of a tree using the best split among all variables, each node is split using the best among a subset of predictors randomly chosen at that node. This strategy prevents trees of the forest to be correlated since they do not select any more the same strong predictors. The R implementation of RF requires the manual input of the number of trees participating in the forests. Therefore, according to the strategy described above, we tuned this parameter, investigating values from 100 to 700. The selected value was 700, as reported in Table 1.

3.4 Extreme gradient boosting

XGBoost is an implementation of gradient boosted (GB) decision trees, with a difference in modelling details that generally allows XGBoost to obtain better performance [30]. Boosting is a technique where new models (decision trees) are added to correct the errors made by existing models. Specifically, gradient boosting is an approach where new models are created to fit the residuals of prior models and then added together to make the final model. The word "gradient" refers to the use of the gradient descent technique applied to minimize the loss when adding new models. The implementation we choose is the one contained in R [4], which requires the number of rounds (iterations) to be specified by the user. According to the tuning technique described above, we investigate the best number of trees to configure, called the number of rounds, from 1 up to 20. This value and other main parameters of XGBoost in R are listed in Table 1.

3.5 Multilayer perceptron

A multilayer perceptron (MLP) is a multilayer feedforward neural network [27]. It consists of a set of source nodes that constitute the input layer, one or more hidden layers of computational nodes, and an output layer of computational nodes. The input signal propagates through the network in forward direction, on a layer-by-layer basis. The multiple layers are meant to capture more complex relationships among input variables. We adopt the implementation included in the Matlab Neural Network toolbox [20]. Following the strategy previously described, we tuned the number of hidden neurons considering a network with just one hidden layer. We explored all the values between 1 and the number of input variables. All the main parameters used are described in Table 1.

Table 1: Parameters used in the experiments.

GP Parameters	
population size	500
max number of generations	100
initialization	Ramped Half and Half [23]
selection method	Lexicographic parsimony pressure [25]
elitism	best individual kept
crossover rate	0.9
mutation rate	0.1
max tree depth	17
RF Parameters	
number of trees	700
XGBoost Parameters	
η learning rate	0.3
max tree depth	6
number of rounds	7
MLP Parameters	
learning algorithm	Levemberg-Marquardt backpropagation [16]
hidden neurons	1
μ increase factor	0.1
μ decrease factor	10
epochs	1000

4 Experiments and results

4.1 Experiments setup

Our objective is to study GP's ability in predicting mosquitoes abundance during 2006, based on historical data collected from 2002 to 2005, and to show its competitiveness by comparing it with other methods. After a tuning phase of some parameters, RF, XGBoost, MLP and GLMM are trained on the learning set (mosquitoes collections from 2002 to 2005) and then evaluated on the unseen collections of 2006 (test set). Since GP is a population based and stochastic technique, the training phase is conducted differently. We run the algorithm 60 times; in each time the population is trained on a random sample of 75% instances of learning dataset. In each run, the individual of the final population that better perform on the remaining 25% of the learning dataset (called validation set) is selected as the

best predictive model found by GP. The final 60 best models compose the sample population of GP models. Each of the GP models is then evaluated on the test set.

We measure the accuracy of prediction by comparing the RMSE between predicted and real collections on the test data. Statistical significance of the null hypothesis of no difference in performance between GP and each of the other method is based on one sample Wilcoxon signed rank test at $\alpha = 0.0125$, after Bonferroni correction. For further comparison, we measure the overfitting as the difference between test and learning set RMSE.

4.2 Results

Table 2 summarizes the RMSE returned by each techniques on the test set. Regarding GP, we report the median RMSE over the 60 models, choosing the median instead of the mean as a more robust descriptor of outliers, which are likely to be found in stochastic methods. Table 3 presents the *p-value* of the Wilcoxon tests comparing GP with the other methods.

Table 2: Statistics about the RMSE of the different techniques on the test set.

GP	**RF**	**XGBoost**	**MLP**	**GLMM**
83.8 (median)	83.0	87.9	83.7	85.5

Table 3: Statistical significance of the difference in performance between GP and the other methods.

GP vs RF	GP vs XGBoost	GP vs MLP	GP vs GLMM
$p = 3.1 \cdot 10^{-7}$	$p = 1.7 \cdot 10^{-11}$	$p = 0.2$	$p = 1.2 \cdot 10^{-10}$

According to the statistical tests, GP performance differs from all the other methods except MLP. Boxplot in Figure 1 shows that GP is only outperformed by RF. The results on GP performance immediately suggest that the relationship among variables is more complex than the one previously designed with a GLMM in [8]. However, at a first glance, GP does not seem to be the technique that returns the best result. Nonetheless, the quality of predictions should also take into account the quality of learning expressed by overfitting. As mentioned above, we have decided to quantify overfitting by calculating the difference between learning and test median RMSE. For RF and MLP, this measure returns 46.1 and 17.4 respectively. This indicates a severe overfitting for RF. Contrarily, the difference between learning and test RMSE for GP is only 1.8. We hypothesize that RF, and, even though in more reduced form, MLP are not learning the existing relationship between the variables. The boxplot of Figure 1 shows that the same phenomenon appears even more substantially in XGBoost, which is generally considered the top machine learning method nowadays. A possible reason of this fact, strengthened by the RF results, is that regression trees are not suitable for the problem at hand.

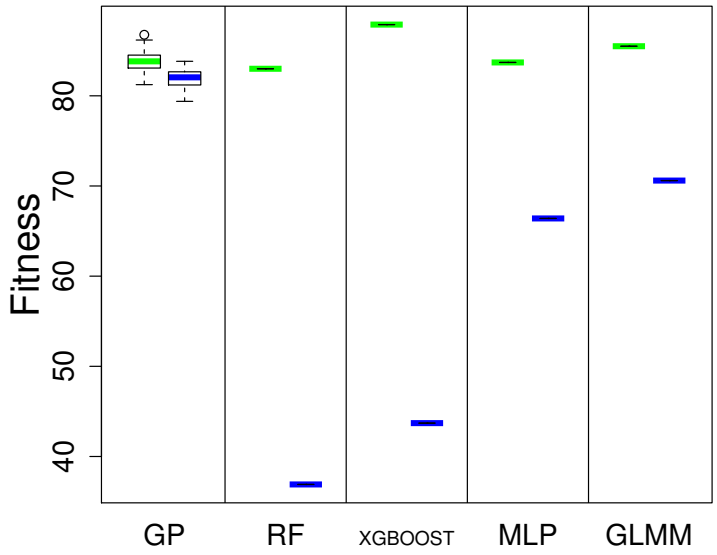

Fig. 1 RMSE on both the learning and test set for the different algorithms. Test set results are plotted in blue, while learning set results are plotted in green.

To corroborate the appropriateness of using GP, compared to the other studied techniques for the problem at hand, we calculate the percentage of GP models in the sample with lower RMSE compared to MLP and RF on the test set. Respectively 22% and 43% of GP models outperform RF and MLP .

4.3 The model for mosquitoes dynamics

Another competitive advantage of GP compared to other machine learning methods, including the ones studied here, is the possibility of reading and interpreting the model. Despite the fact that decision trees are representations easy to understand in logical terms, the process of averaging the results coming from multiple decision trees in a RF muddies the logic. Since model interpretability is a key feature of ecological modelling, GP shows again its benefit in the field.

We select as the predictive model for mosquitoes abundance the best model on its validation set among the 60 best models. The resulting RMSE on the test set is equal to 83.4. Equation (1) represents its expression. The symbols used are the ones classically associated with the primitive functions reported in Table 1; when no symbol is found, a multiplication is occurring.

$$
\begin{aligned}
\#\text{Mosquitoes} = & \frac{DISTW \cdot SIN^3}{DISTU} + \frac{(TWEEK - DISTR + DISTW)SIN^7}{DISTR} \\
& + \frac{DISTW \cdot RICEA \cdot SIN^5}{TWEEK} + 2SIN^5 + \frac{DISTW \cdot SIN^5}{DISTU} \\
& - DISTR \cdot SIN^3 + (TWEEK - DISTR)SIN^8 \\
& + \frac{2SIN^3(2SIN + 1)}{DISTU} + \frac{DISTU \cdot SIN^3}{DISTR} + 2DISTW \cdot SIN^3 \\
& - \frac{DISTW \cdot SIN^2}{DISTR} - DISTR \cdot SIN + DISTW
\end{aligned} \tag{1}
$$

Since there are quite a lot of occurrences of `kozadivision` in the expression, we check whether their result is frequently the constant 1 used to protect the division when the denominator is zero. Luckily, on the test set, the protected version of the division is not too much used, therefore the divisions involved are mainly true division.

The general expression reported in Equation (1) may be hard to interpret, but the analysis of the variables discarded and the general effect of the ones selected may provide meaningful information. To investigate the role of each variable in the prediction we firstly consider, as a subjective measurement, the number of times each variable appears in the model. Table 4 shows these frequencies of occurrence of the single variables in the model.

Table 4: Frequency of each variable in the best model.

Variable	Frequency
TWEEK	3
RAIN	0
NDVI	0
DISTU	4
DISTR	7
DISTW	7
RICEA	1
ELEV	0
SIN	13

The implicit feature selection embedded in GP reveals that *RAIN*, *NDVI* and *ELEV* are not informative about mosquitoes abundance. This assertion contrast with the results of [8], where the GLMM included both *RAIN* and *ELEV*, as shown in Section 3. Interestingly, the most frequent variable is *SIN*. The standardized coefficients in GLMM give a measure of the change in the target (in standard deviations) for every standard deviation change in the predictors. Since the higher standardized coefficient is the one of *SIN* ($\beta_{SIN} = 0.02, \beta_{DISTU} = -0.003, \beta_{ELEV} = -0.007, \beta_{TWEEK} = -0.005, \beta_{RAIN} = 0.003$), we can state that also when using the GLMM this variable is considered as the most informative one. A similar analysis cannot be carried on for the other techniques, since they are mainly black box methods.

Despite the fact that shorter models are often more appealing, GP provides a well performing readable expression where the main effect of variables on the target can be captured. Looking at the main appearance of variables at numerator or denominator, and at the sign, we can make the following observations:

- The LST of the week before trapping, *TWEEK*, has a general positive effect on the abundance of collected mosquitoes. This outcome is in line with the knowledge about mosquitoes development [31]; in fact, the warmer the temperature is, the faster the mosquito larvae will grow and then spread.
- The distance from urban centre, *DISTU*, appears mainly at denominator with positive sign, having a negative effect on the abundance of mosquitoes. This suggests that the closer a place is to an urban area, the more numerous is the presence of mosquitoes. In urban areas in fact there are many breeding sites for mosquitoes such as catch basins and plant pot saucers.
- The distance from the nearest rice field, *DISTR*, has a negative effect too: the closer a place is to a rice field, the more mosquitoes are collected. Moreover, since *RICEA* is at numerator with positive sign, the larger the rice field the more mosquitoes are trapped. The model therefore suggests that rice fields are a suitable habitat for mosquitoes.
- The distance from woodland, *DISTW*, appears as a standalone term, directly influencing the abundance of mosquitoes and thus playing an important role in the prediction. Moreover, it has a general positive effect on the target, suggesting that woodlands, differently from urban and rice fields areas, are not a convenient environment for mosquitoes. This is probably due to the high abundance of various birds that prey this insects.
- *SIN* has a positive effect on the amount of mosquitoes. This is not so surprising, as far as we introduce this variable to reflect peaks in mosquitoes abundance during the hot season.

From GP predictions, we derive the mosquitoes distribution maps during the seasonal peak of abundance. We want in fact to highlight how the GP model can help in identifying high risk area and thereby the planning of surveillance programs. Figure 2 shows the comparison between the median number of mosquitoes collected from the end of June until the first week of August 2006 in each trap and the median predicted values.

(a) Median number of mosquitoes predicted by GP.

(b) Median number of mosquitoes collected.

Fig. 2 Prediction maps for mosquitoes from the end of June until the first week of August 2006. Circles indicates the abundance in each trap. Darkness increase with the number of mosquitoes.

The GP model is able to detect the highest risk area of Casale Monferrato in 2006, having however smoother predictions in space. In fact, while in the real map we find very low abundance traps among dark circles, in the GP map close traps have close abundance values. The fact that nearby traps do not have similar count is suspicious, thus there may be a problem with the data.

5 Discussion

An issue that came out from the results presented so far deserves further attention. The variable *SIN* is an artificial predictor included in the dataset to suggest the period of collection and

therefore how high the abundance of mosquitoes is expected to be. Both GP and GLMM, the first technique that was applied on the problem, strongly rely on this variable, which undoubtedly gives lots of information concerning mosquitoes abundance. However, this predictor is the result of a prior knowledge about the problem and we would rather prefer the algorithms to directly infer from data which combination of environmental and climatic variables determines the fluctuations of mosquitoes abundance.

LST, NDVI and the rainfall are time series predictors whose flow is related to seasons, thus they can be strong predictors of mosquitoes dynamics. Nonetheless, they are not recognized as such by GP (and even GLMM) probably because they are not treated as time series. Time series were managed as independent training cases by all the machine learning techniques studied in this paper, including GP. This may cause a loss of information, which may deteriorate the ability to predict mosquitoes abundance. In this regard, a vector based algorithm such as Vectorial GP (VE-GP) [13] demands an exploration on the problem in analysis. VE-GP is a vectorial approach of genetic programming in which vectors are allowed as terminals. To fully exploit this new representation, VE-GP admits in the primitive functions set new operators such as aggregate functions, parametric or not. Since vectors are appropriate to represent time series, the mosquitoes abundance dataset can definitely be analysed by means of VE-GP. The most informative time windows and aggregations of the predictive time series variable, LST, NDVI and rainfall, may in this way be discovered during the evolutionary process, without any *a priori* assumption. Thus, it may be possible to discard the artificial *SIN* variable.

6 Conclusions

In this paper, we have explored the use of Genetic Programming (GP) in the field of ecological modelling. The problem in analysis was the prediction of mosquitoes abundance during a year in the Italian Piedmont region, in order to control West Nile Virus spread since it is a virus transmitted by mosquitoes. The problem had already been studied using statistical techniques, but we believe that heterogeneous and complex datasets such as the one involved demand for the use of algorithms that can catch more complex relationship among the involved variables. We assess GP performance by means of an experimental comparison with other well known Machine Learning techniques in the field and with Generalized Linear Mixed Model (GLMM), that was the first method ever used to tackle this type of problem.

A first conclusion, based only on an analysis of the Root Mean Squared Error on the test set (2006 abundances), reveals that GP is outperformed by Random Forest. This preliminary result is however misleading since it does not take into account overfitting. It turned out, in fact, that GP is the best approach among the studied ones according to both prediction accuracy and generalization capability. Moreover GP is the only technique that combine to accuracy the readability of the model, which leads to the discovery of patterns in data and provides idea about the domain of investigation. These features are of key importance in ecological modelling to improve the understanding of the problem in analysis. Based on this fact, we investigated the best model provided by GP in order to highlight the more relevant variables for the prediction and their effect on mosquitoes abundance.

A fact that needs to be pointed out is that GP (and also GLMM) gives a substantial importance for the prediction to the artificial *SIN* variable, a sinusoidal curve with a phase of 1 year that suggests mosquitoes abundance dynamics. This *a priori* knowledge is likely to prevent the algorithm to infer the dynamics from other types of information, like the flow of the Normalized Difference Vegetation Index, of the Land Surface Temperature and of rainfalls. However, to avoid the use of *SIN*, the time series involved in the data need to be more effectively exploited. For this reason, our current work is oriented towards the investigation of the problem by means of a novel version of GP called Vectorial GP (VE-GP). The proper time series representation, and the possibility of evolving aggregated values, should allow us to develop a predictive model of the mosquitoes abundance that uses only ecological variables, without the need of any *a priori* knowledge coming from domain experts.

References

1. Arpa Piemonte. http://www.arpa.piemonte.it.

2. INLA. https://inla.r-inla-download.org/R/stable.
3. Random forest, 2002. https://CRAN.R-project.org/doc/Rnews/.
4. XGBoost, R package version 0.82.1, 2019. https://CRAN.R-project.org/package=xgboost.
5. Ministero della Salute. West Nile Disease - Notifica alla Commissione europea e all'OIE - Piano di sorveglianza straordinaria. Gazzetta Ufficiale della Repubblica Italiana, N. 277, 26/11/2008.
6. Shaeffer B., Mondet B., and Touzeau S. Using a climate-dependent model to predict mosquito abundance: Application to *Aedes (Stegomyia) africanus* and *Aedes (Diceromyia)* furcifer (Diptera: Culicidae). *Infection, Genetics and Evolution*, 8:422–432, 2008.
7. Talla C., Diallo D., Dia I., Ba Y., and Ndione J.-A. et al. Statistical modeling of the abundance of vectors of West African Rift Valley Fever in Barkédji, Senegal. *PLoS ONE*, 9(12), 2014.
8. Bisanzio D., Giacobini M., Bertolotti L., Mosca A., Balbo L., Kitron U., and Vazquez-Prokopec G.M. Spatio-temporal patterns of distribution of West Nile virus vectors in eastern Piedmont Region, Italy. *Parasites & Vectors*, 4, 2011.
9. M. S. Diamond. *West Nile Encephalitis Virus Infection*. Springer, 2008.
10. Papworth D.J., Marini S., and Conversi A. A novel, unbiased analysis approach for investigating population dynamics: A case study on *Calanus finmarchicus* and its decline in the North Sea. *PLoS ONE*, 11(7), 2016.
11. Spiegelhalter D.J., Best N., Carlin B.P., and Linde A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 64:1–34, 2002.
12. Monaco F., Lelli R., Teodori L., Pinoni C., Di Gennaro A., Polci A., Calistri P., and Savini G. Re-emergence of West Nile virus in Italy. *Zoonosis public health*, 57:476–486, 2010.
13. Azzali I., Vanneschi L., Silvia S., Bakurov I., and Giacobini M. A vectorial approach to genetic programmig. In *Genetic Programming- 22nd European Conference EUROGP 2019*, Lecture Notes in Computer Science, 2019.
14. Koza J. Genetic programming: On the programming of computers by means of natural selection. 1992.
15. Evans J.S., Murphy M.A., Holden Z.A., and Cushman S.A. Modeling species distribution and change using random forest. *Predictive Species and Habitat Modeling in Landscape Ecology*, pages 139–159, 2011.
16. Levemberg K. A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, 2(2):164–168, 1944.
17. Autorino G. L., Battisti A., Deubel V., Ferrari G., Forletta R., Giovannini A., Lelli R., Murri S., and Scicluna M.T. West Nile virus epidemic in horses, Tuscany region, Italy. *Emerging Infect Dis*, 8(1):372–1378, 2002.
18. Breiman L. Random forests. *Machine Learning*, 45:5–32, 2001.
19. Kramer L.D., Styer L.M., and Ebel G.D. A global perspective on the epidemiology of west nile virus. *Annual review of entomology*, 53:61–81, 2008.
20. The MathWorks. MATLAB Neural Network Toolbox, 2018.
21. Engler O., Savini G., Papa A., Figuerola J., Groschup M.H., and Kampen H. European surveillance for west nile virus in mosquito populations. *Int J Environ Res Public Health*, 10:4869–95, 2013.
22. Gervasi R., Azzali I., Bisanzio D., Mosca A., Bertolotti L., and Giacobini M. A genetic programming approach to predict mosquitoes abundance. In *Genetic Programming*, EuroGP 2019. Lecture Notes in Computer Science, 2019.
23. Poli R., Langdon W. B., and McPhee N. F. *A field guide to genetic programming*. 2008.
24. Wagner R., Obach M., Werner H., and Schmidt H-H. Artificial neural nets and abundance prediction of aquatic insects in small streams. *Ecological Informatics*, 1:423–430, 2006.
25. Luke S. and Panait L. Lexicographic parsimony pressure. In *GECCO 2002: Proceedings of the Genetic and Evolutionary Computation Conference*, pages 829–836, 2002.
26. Marini S. and Conversi A. Understanding zooplankton long term variability through genetic programming. In *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, EvoBIO 2012. Lecture Notes in Computer Science, 2012.
27. S.Haykin S. Neural networks: a comprehensive foundation. 1999.
28. Silva S. GPLAB - A Genetic Programming Toolbox for MATLAB.
29. Walsh A. S., Glass G. E., Lesser C. R., and Curriero F. C. Predicting seasonal abundance of mosquitoes based on off-season meteorological conditions. *Environ Ecol Stat*, 15:279–291, 2008.

30. Chen T. and Guestrin C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, 2016.
31. Ciota A. T., Matacchiero A. C., Kilpatrick A. M., and Kramer L. D. The effect of temperature on life history traits of *Culex* mosquitoes. *J Med Entomol*, 51(1):55–62, 2011.
32. Santosh T. and Ramesh D. Artificial neural network based prediction of malaria abundances using big data: A knowledge capturing approach. *Clinical Epidemiology and Global Health*, 17:121–126, 2019.
33. ORNL DAAC Oak Ridge TennesseeUSA. ORNL DAAC 2018 MODIS and VIIRS land products global subsetting and visualization tool. https://modis.ornl.gov/.
34. Stroup W. W. *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. CLC Press, 2012.