A Genetic Programming Approach to Predict Mosquitoes Abundance

A Genetic Programming Approach to Predict Mosquitoes Abundance

Riccardo Gervasi[1,2][0000−0002−3006−3382], Irene Azzali[1][0000−0002−5808−7044], Donal Bisanzio[3,4][0000−0002−7832−2291], Andrea Mosca[5][0000−0003−2180−0405], Luigi Bertolotti[1][0000−0001−7931−4528], and Mario Giacobini[1][0000−0002−7647−5649]✉

[1] DAMU - Data Analysis and Modeling Unit, Department of Veterinary Sciences, University of Torino, Italy
riccardo.gervasi@edu.unito.it,
{irene.azzali,mario.giacobini,luigi.bertolotti}@unito.it

[2] Department of Management and Production Engineering (DIGEP), Politecnico di Torino, Italy

[3] RTI International, Washington, DC, USA
dbisanzio.epi@gmail.com

[4] Division of Epidemiology and Public Health, School of Medicine, University of Nottingham, Nottingham, UK

[5] Istituto per le Piante da Legno e l'Ambiente (IPLA), regional government-owned corporation of Regione Piemonte, Torino, Italy
mosca@ipla.org

Abstract. In ecology, one of the main interests is to understand species population dynamics and to describe its link with various environmental factors, such as habitat characteristics and climate. It is especially important to study the behaviour of animal species that can hosts pathogens, as they can be potential disease reservoirs and/or vectors. Pathogens of vector borne diseases can only be transmitted from an infected to a susceptible individual by a *vector*. Thus, vector ecology is a crucial factor influencing the transmission dynamics of vector borne diseases and their complexity. The formulation of models able to predict vector abundance are essential tools to implement intervention plans aiming to reduce the spread of vector-borne diseases (e.g. West Nile Virus). The goal of this paper is to explore the possible advantages in using Genetic Programming (GP) in the field of vector ecology. In this study, we present the application of GP to predict the distribution of *Culex pipiens*, a mosquito species vector of West Nile virus (WNV), in Piedmont, Italy. Our modelling approach took into consideration the ecological factors which affect mosquitoes abundance. Our results showed that GP was able to outperform a statistical model that was used to address the same problem in a previous work. Furthermore, GP performed an implicit feature selection, discovered automatically relationships among variables and produced fully explorable models.

Keywords: Genetic Programming · Ecological Modeling · Prediction · West Nile Virus

1 Introduction

West Nile virus (WNV) is a zoonotic virus belonging to the *Flaviviridae* family, genus *Flavivirus*, eventually neuropathogen for birds and mammals including humans [1]. West Nile virus is transmitted by the bite of infected mosquitoes, mainly belonging to the genus *Culex* [2]. The natural cycle (enzoonotic cycle) of the virus involves the pathogen passing from mosquitoes to wild bird species which can be reservoirs. The virus can infect several vertebrate species (mammals, birds, reptiles) and, among mammals, humans and horses are considered dead-end hosts [3] and can show clinical symptoms [4]. In humans, most WNV infections occur asymptomatically: about 20% of infected individuals develop a febrile disease, commonly called *West Nile fever* (WNF). In less than 1% of cases the disease manifests itself in neuro-invasive form (usually encephalitis, meningoencephalitis or flaccid paralysis) [5].

In Italy, the first WNV outbreak occurred in 1998 involving several horses in the Tuscany region [6]. After the first outbreak of 1998, Italian public health authorities implemented a national surveillance plan to detect WNV circulation (in vectors and host) and quantify vector abundance [7]. Given the complexity of the biological cycle of WNV, the control of its circulation requires the integration of surveillance systems in different areas: entomological, veterinary and human. The main objective of integrated surveillance is to detect early, through targeted programmes, the circulation of WNV in birds, insects or mammals on the national territory. The early detection allows to assess the risk of transmission of the disease to humans and to implement all available measures to prevent transmission [8]. *Culex pipiens* is the most common mosquito species belonging to the genus *Culex* in the northern hemisphere. *Cx. pipiens* mosquitoes are the principal vector for WNV in Europe, since they have ornithophilic and anthropophilic subspecies, often inbreeded, and their populations are usually large [9, 10]. For this reason, it is important to keep their abundance under control in order to avoid the outbreak of cases of West Nile Disease (WND).

The main goal of our study is to explore the potential application of genetic programming (GP) for an ecological problem, in particular to create a predictive model for the abundance of *Cx. pipiens* in an eastern area of Piedmont region. The obtained GP models are compared with a statistical model developed to address the same problem in a previous work (Bisanzio et al., 2011) [11].

In the first section we present the available data and how they were collected. Next, we describe the statistical model used to compare the performance of GP and the results obtained using it. The third section contains the results obtained by GP and the comparison with the statistical model via hypothesis tests. Last section concerns conclusions and possible improvements.

2 Data Set

Mosquito data used in this study are based on entomological collections performed during the surveillance Piedmont program started by "Regione Piemonte"

through local Municipalities Agreements, and subsequently carried out by "Istituto per le Piante da Legno e l'Ambiente" (IPLA) from 2002 to 2006. The area of study covered a territory of 987 km^2 in the eastern part of the Piedmont region, where the Municipalities Agreement of Casale Monferrato operated. This territory offers favourable habitats for the reproduction of the local *Cx. pipiens* mosquitoes [11], therefore it is suitable for the introduction and amplification of WNV. From now on when we mention mosquitoes we will refer to the ones of *Cx. pipiens* species. Half of the territory is made up of hills, with an average elevation of $268m$, and the other half of plains where the landscape is mainly composed of mixed agricultural patches (72.2%, mainly in the northeast), rice fields (14.2%, mainly concentrated in the north), deciduous tree forests (8.6%, in the south), urban areas (3.1%) and the river Po with its tributaries (1.9%, in the north). The climate is characterized by cold winters (0.4°C on average) and hot-warm summers (24.0°C on average), with heavy rainfall in spring and autumn (about 600 mm/yr) [12]. To acquire mosquitoes abundance, 36 sampling location were selected on the territory, at a minimum distance of 5 km from each other, in areas suitable for mosquito proliferation (near rice fields, forests, urban suburbs). Captures of mosquitoes were taken at night weekly using CO_2-baited trap from the beginning of May to late September, which is the main period of mosquitoes activity in the area, for a total of 20 collections per year. In addition, various environmental and ecological parameters were collected as influential variables of mosquitoes abundance. The previous work of *Bisanzio et al.* [11] selected among all the variables the ones deemed to be the most effective in predicting the abundance and the spatial distribution of mosquitoes. Table 1 lists the final predictors after processing data as in [11].

Table 1. Description of variables chosen as mosquitoes predictors.

Predictor variable	Description
TWEEK	The average land surface temperature 8-15 days prior to trapping [13].
NDVI	16-days average of normalized difference vegetation index [13].
RAIN	Cumulative rainfall 10-17 days prior to trapping. [12]
ELEV	Elevation of the sample location [13].
DISTU	Distance of the sampling location from the nearest urban area.
DISTR	Distance of the sampling location from the nearest rice field.
DISTW	Distance of the sampling location from the nearest water source.
DISTF	Distance of the sampling location from the nearest forest.
RICEA	Area of the nearest rice field.
SIN	Value of a sinusoidal curve with phase 1 year.

TWEEK and *RAIN* refer to time windows that capture respectively the impact of the temperature and the rain on the abundance of mosquitoes. The parameter *NDVI* reflects environmental changes due to human agricultural activities that influence the distribution of the vector. *ELEV* takes into account altitude differences within the study area. The effect of the surrounding environment is considered through the variables *DISTU*, *DISTR*, *DISTW*, *DISTF* and *RICEA*. *SIN* represents the seasonal pattern of mosquitoes abundance as a positive sinusoidal curve with a peak in period the 12th week of collection (end of July-beginning of August), when mosquitoes are usually more numerous. For our experiments, the years 2002-2006 had been taken into account. The main reasons for this choice are the fact that during this period there was the highest spatial coverage of CO_2baited-traps and that mosquitoes control actions in the region were minimal. Our dataset was composed by the number of captured mosquitoes in each trap for each of the 20 periods, joined with the related predictors values.

3 Classical Statistical Approach

As mentioned in Section 2, mosquitoes abundance in eastern Piedmont had already been investigated in *Bisanzio et al.* [11]. The authors used a spatio-temporal Generalized Linear Mixed Model (GLMM) [14] to study the association between the abundance of *Cx. pipiens* and environmental and ecological variables. They formulated and tested 169 GLMM models and selected the best through the Deviance Information Criterion (DIC) [15]. The parameters of the derived model were tuned according to data using a Bayesian approach based on the Integrated Nested Laplace Approximation (INLA) [16]. For any further detail refer to [11]. In order to investigate the predictive behaviour of a GP approach, we borrowed the selected best model and the INLA technique and we fitted again the statistical model dividing the data in a training and a test dataset. We chose data of weekly collection from 2002 to 2005 as the training set and the 2006 trapping as the test set, which contains unseen data used to validate model performances. This choice follows the natural order of the years, therefore we learn from the past and we test on the future what we have learned. The resulting model is:

$$y = I + \beta_1 * RAIN + \beta_2 * TWEEK + \beta_3 * SIN + \beta_4 * ELEV + \beta_5 * DISTU + RNDtrap \quad (1)$$

where the variables are those described in Section 2, *RNDtrap* is an unstructured spacial random effect to represent unaccountable differences in each trap, and I is the intercept of the model. The estimated parameters according to INLA technique are reported in Table 2.

The GLMM model obtained a training Root Mean Square Error (RMSE) of 70.32851 and a test RMSE of 88.58552. We can notice that:

- Cumulative rain (*RAIN*) does not seem to have a significant effect in predicting abundance, as the credible interval contains zero.

Table 2. Fitted terms of the GLMM applied to predict mosquitoes abundance.

Predictors	Coefficients (Mean)	Credible Intervals
Intercept	1.4831	(0.7400, 2.2303)
RAIN	0.0018	(-0.0006, 0.0043)
TWEEK	-0.1285	(-0.1441, -0.1129)
SIN	7.8351	(7.4870, 8.1846)
ELEV	-0.0069	(-0.0104, -0.0035)
DISTU	-0.1920	(-0.3260, -0.0583)
Random Effects	Mean	St. Deviation
RNDtrap	2.7468	0.7143

- The average temperature of about a week before ($TWEEK$) has a significant negative effect on the amount of mosquitoes. This result, although opposite to that of a previous work [11], is confirmed in a more recent article [17] in which it is pointed out that high temperatures lead to a decrease of *Cx. pipiens* abundance.
- The seasonality variable (SIN) is quite important and has a positive effect on the abundance. As expected, the closer to the period of mosquitoes high activity, the greater the abundance.
- The elevation ($ELEV$) has a negative effect on the expected amount of *Cx. pipiens*. At high altitudes, environmental conditions would not support their proliferation.
- Distance from the nearest urban center ($DISTU$) has a significant negative effect on the amount of *Cx. pipiens*. Thus, the closer to towns, the higher the abundance. This could be explained by the possible presence of stagnant water, which is an important resource for mosquitoes development. Indeed, in an urban environment there is a large number of small breeding sites, mainly underground, particularly adapted for the development of *Cx. pipiens*. The number of potential hosts is large, too.

4 Application of GP

4.1 Experimental Setting

In this work we used a tree-based GP where the individuals are represented in a tree structure. The set of terminal nodes is composed by the 10 variables described in Section 2 and a random constant between 0 and 1 generated during the initialization phase, while the function set is $F = \{\texttt{plus}, \texttt{minus}, \texttt{times}, \texttt{kozadivide}\}$ where `plus`, `minus`, `times` are the classical binary addition, subtraction and multiplication operators and `kozadivide` is the division protected as in [18] that returns 1 if the second argument (i.e. the divisor) is 0. The chosen fitness function was the Root Mean Square Error (RMSE), transforming the problem into a minimization one, since lower values represent better solutions. As the aim of this investigation was to explore the possible application of GP in ecological modeling, we prioritized simplicity considering the parameters default setting

proposed in GPLab [19], the GP toolbox implemented in MATLAB that we used for our experiments. The experimental parameters setting is provided in Table 3.

Table 3. GP parameters used for experiments.

Parameter	Setting
Population size	500
Maximum number of generations	100
Initialization method	Ramped Half-and-Half [18]
Selection method	Lexicographic Parsimony Pressure [20]
Elitism	Survival of the best ('keepbest')
Subtree Crossover rate	0.9
Subtree Mutation rate	0.1
Maximum tree depth	17

Each experiment was repeated 60 times, to ensure a large enough set of simulations in order to investigate the algorithm performance. In each run the algorithm was trained using mosquitoes collection data from 2002 to 2005 and the best model found in each run was tested on mosquitoes abundance prediction of 2006.

4.2 Experimental Results

Figure 1 shows the median training and test RMSEs at each generation over the 60 runs. We chose the median value due to its robustness to outliers which can emerge in such a stochastic algorithm. The evolution of the error reveals the ability of GP in learning the relationship between variables. Moreover, the constant decrease of the test error indicates that no overfitting is occurring, therefore the algorithm is learning with generalization.

At the end of each of the 60 runs of the experiment, we obtained a best solution. The median test RMSE obtained by considering these 60 models is 83.63461, which is less than test RMSE of the GLMM (see Section 5 for the statistical comparison of the performances of the two approaches). After removing possible introns from each best individual, we decided to observe the median frequency with which the variables appear in the terminal nodes of the best solutions. This consideration could give a general idea of which variables play a key role in achieving the best results, while also highlighting variables that could be not so important for explaining the variability of the dependant variable. Table 4 contains the median frequencies obtained for each variable considering the 60 best solutions.

SIN and *DISTF* are the most frequent variables on median, suggesting that these variables could be relevant predictors for mosquito abundance. *RAIN*, *ELEV* and *RICEA* seem to be not so relevant, since their median frequency is 0.

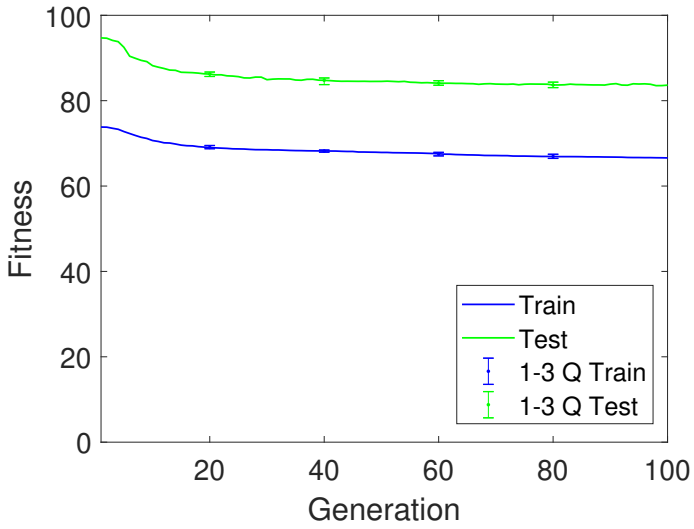

Fig. 1. GP evolution plots. The blue line represents the median RMSE on the training set, while the green line represents the median RMSE on the test set. The bars describe the range between the 1st and 3rd quartiles.

Table 4. Median frequency of each variable in the best 60 solutions.

Variable	Median frequency
X_1 - RAIN	0
X_2 - TWEEK	4
X_3 - SIN	27.5
X_4 - ELEV	0
X_5 - NDVI	1
X_6 - DISTU	4
X_7 - DISTR	7
X_8 - DISTW	4.5
X_9 - DISTF	12
X_{10} - RICEA	0

Among the 60 best solutions, we chose as a prediction model the expression that produced the lowest RMSE on the test set. The selected solution obtained a training RMSE of 65.99388 and a test RMSE of 81.41473. Equation (2) represents the selected model for mosquitoes prediction.

$$
\begin{aligned}
Y = & X_9 + \frac{X_3}{X_6^2} - \frac{X_8}{X_6} - X_3 + X_3 \Bigg\{ X_3 \Bigg[\frac{X_3 X_6^2}{X_7 X_6 + 0.036707\,(X_6 - 1)} + X_3 \Bigg(X_3 \cdot \\
& \Bigg(X_3 \left(X_2 + X_9 - X_7 - \frac{X_2}{X_{10}} \right) - X_7 + \frac{X_3 - X_7}{X_6} - X_3 X_6 + X_3 - 2X_6 \Bigg) + \\
& - \frac{X_9}{X_9 - 2X_7} + \frac{X_3}{X_7} + X_3^3 \left(2X_2 - 2X_7 + \frac{X_3}{X_6^2} \right) - X_7 \Bigg) - X_6 \Bigg] + \frac{X_3}{X_6^2} - 1 \Bigg\}
\end{aligned}
\tag{2}
$$

From Equation (2) it can be noticed that some variables are missing, in particular *RAIN* (X_1), *ELEV* (X_4) and *NDVI* (X_5). This is due to GP's ability to automatically perform an implicit features selection; if solutions with smaller number of variables have a better fitness, they survive into the population, since fitness is the only principle on which individuals are selected. By observing the frequency with which the variables have been selected, it is possible to find out which of them are responsible for good performances with their recurrence.

Table 5. Frequency of each variable in the best model.

Variable	Frequency
X_1 - RAIN	0
X_2 - TWEEK	4
X_3 - SIN	17
X_4 -ELEV	0
X_5 - NDVI	0
X_6 - DISTU	14
X_7 - DISTR	10
X_8 - DISTW	1
X_9 - DISTF	4
X_{10} - RICEA	1

Table 5 shows that *SIN* is the most frequent variable, and the fact that it was a significant predictor also in the previous GLMM gives greater credence to its importance. Although the model (2) found by GP is very complex, it is still possible to interpret it and highlight the effect of some variables:

- *SIN* (X_3) has a general positive effect on the amount of trapped mosquitoes: as expected, as you approach the peak of the mosquito season, the abundance increases. This effect is also confirmed by the previous GLMM.

- *DISTF* (X_9) appears in the model as a standalone term, directly influencing the abundance of mosquitoes. This fact may lead us to believe that it could play an important role in the correct prediction. Moreover, it has a general positive effect on the abundance of *Cx. pipiens*: the further a place is from a forest, the more mosquitoes are expected to be trapped. This could be justified by the fact that in the forest there are few habitats suitable as breeding sites for this species. Moreover, another explanation could be that CO_2-baited traps close to woodlands may have less attraction than those ones placed far, due to abundance of potential hosts for mosquitoes for *Cx. pipiens*, such as various bird species.
- The average temperature of a week before, *TWEEK* (X_2), has a general positive effect on the abundance of collected mosquitoes. Mosquitoes need to accumulate heat so that they can hatch from eggs, and consequently develop.
- The distance from a urban centre, *DISTU* (X_6), appears often negative at numerator or positive at denominator, having basically a negative effect on the abundance of *Cx. pipiens*. Indeed, the closer a place is to an urban area, the more mosquitoes there are. In a urban area there are many breeding sites for mosquitoes (e.g. catch basins and plant pot saucers).
- The distance from a water source, *DISTW* (X_8), has a negative effect on the abundance of *Cx. pipiens*. Water sources (e.g. rivers, lakes) provide higher humidity rates that positively effect the presence and the survival of mosquitoes. Thus, the closer a place is to a water source, the more mosquitoes is likely to be captured.
- The distance from the nearest rice field, *DISTR* (X_7), has in general a negative effect: the closer a place is to a rice field, the more mosquitoes are collected. The dimension of the rice field, *RICEA* (X_{10}), has a positive effect: at the same distance from a rice field, it is expected that more mosquitoes will be trapped in location with the largest rice field. Rice fields are indeed a suitable habitat for mosquitoes.

The study area allows to evaluate the role of environmental variables at fine scale. In general, it can be said that differences in habitat features have a greater impact than differences in climate (temperature or rainfall) when studying a small area.

To validate the ability of the GP model in predicting mosquitoes abundance, we generated a vector suitability map for August 2006, a period that includes the peak of abundance. We represented the median numbers of predicted mosquitoes against the median numbers of truly collected mosquitoes for each trap. Figure 2 shows a good prediction of high populated patterns, information that can be used to identify areas at high risk of exposure to the virus.

5 Comparison with the Statistical Model

To further explore GP ability in prediction we compared the best models found by GP over the 60 runs with the statistical model described in Section 3. We assumed test RMSEs of best solutions followed a normal distribution, since this

(a) Median number of mosquitoes predicted by GP.

(b) Median number of mosquitoes collected.

Fig. 2. Prediction maps for mosquitoes in August 2006. Circles indicates the abundance in each trap. Size and darkness increase with the number of mosquitoes

hypothesis was not rejected by Lilliefors test (p-value: 0.8137). Moreover, looking at the Q-Q plot (Figure 3), the test RMSEs follow pretty well the line and are almost all contained in the grey stip, supporting the normality hypothesis.

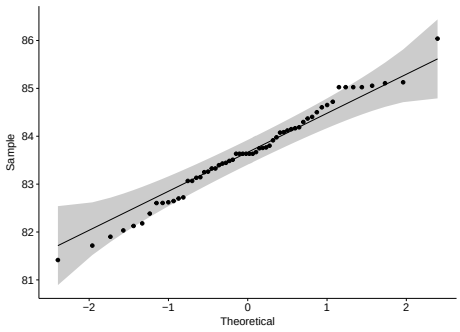

Fig. 3. Q-Q plot for normality of the sample of GP test RMSEs.

We tested whether test RMSEs of the best GP solutions were significantly different from test RMSE of the GLMM, performing a lower-tailed one-sample t-test, with $\alpha = 0.01$. We obtained an extremely significant p-value ($< 2.2 \cdot 10^{-16}$), which indicates a significant difference between the performances of the two methods, therefore GP outperforms the statistical model. Moreover we depicted the best errors on 2006 prediction of both models using a boxplot representation. From Figure 4 emerges that the models selected by GP have better results rather than the GLMM model.

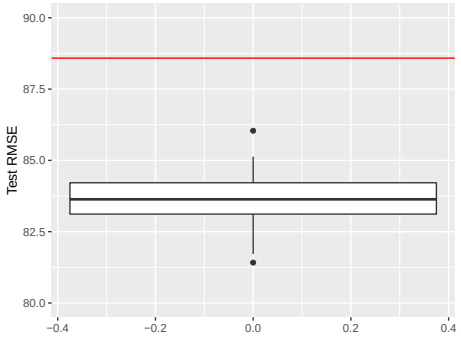

Fig. 4. Boxplot of GP test RMSEs against test RMSE of the GLMM (red line).

To give an idea of how much is the difference we used the *A statistics* (also called *measure of stochastic superiority*). It is a non-parametric effect size measure that quantifies the difference between two populations in terms of the probability that a score sampled at random from the first population will be greater than a score sampled at random from the second [21]. Choosing as first population the 60 test RMSEs of the best GP solutions and as second one the test RMSE of the GLMM, we obtained an A measure equal to 0, indicating that GP greatly outperforms the GLMM. From an ecological point of view, the relevance of the performance difference between the two methods may be marginal, but GP has proved to be consistent in obtaining numerically better results.

Focusing on the GP model with the lowest test RMSE (Equation 2), Figure 5 shows the median predicted values of abundance compared to the median observed values for each of the 20 periods of 2006. We can notice that the GP model was able to detect quite well the moments of greatest mosquito abundance, while the GLMM missed the right peak of abundance.

6 Discussion

Some observations arise from the results achieved by GP and their comparison with the statistical model performance. It is interesting to notice that GP gave some importance to a variable not particularly significant in the GLMM, that is the distance from the nearest woodland (*DISTF*). This result has highlighted how the proximity to forests may be a disturbing element for collection of mosquitoes by means of CO_2-baited traps. Furthermore, this discovery is an example of how certain statistical models could limit the interaction between variables with their predefined structure. The exploration of other types of regression models, with not-only-linear terms, is obviously possible, but we based our work on the state-of-art of the predictive models for the specific problem addressed here.

The *SIN* variable seemed to be quite important in predicting the abundance of *Cx. pipiens* in both methods. This artificial variable was the only one that

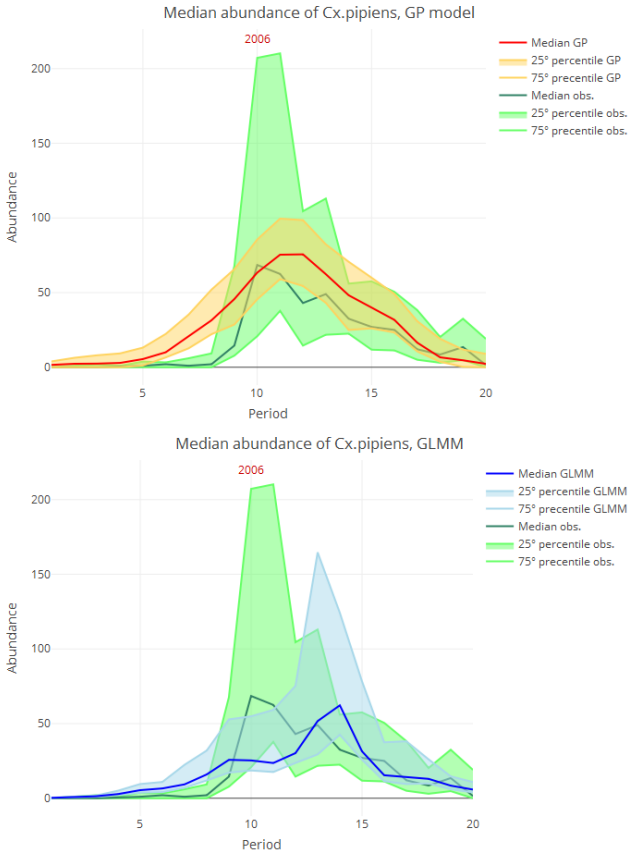

Fig. 5. Predicted median abundance of mosquitoes for 2006 by the GP model (upper figure) and the GLMM (lower figure) compared with median observed values (in green).

allowed to model seasonal peaks and periods of total absence of mosquitoes. We thought it would be interesting if this pattern could be learned by the model directly from ecological variables, such as temperature and rainfall. However, to recognize interesting patterns in these time series variables the algorithm needs to receive them in their entire sequence and not splitted into different fitness cases. This point made us think of a possible application of a vector approach of GP [22] to predict mosquitoes abundance. Using this technique we would have the opportunity to avoid artificial predictors and let GP use only environmental variables to discover the model underlying data.

7 Conclusion

The goal of our work was to explore the possibility of using GP to tackle ecological problems, possibly highlighting the advantages of this technique. In our investigation, we used GP to create a predictive model for the abundance of *Cx. pipiens* in the eastern area of Piedmont region, to detect suitable locations for the entry of WNV. We evaluated GP performances using a statistical model produced in a previous work [11] as threshold. GP was able to outperform the statistical model in predicting the abundance on unseen data, while also finding out interesting relationships among predictors.

In conclusion, GP has proven to be a powerful tool that opens up important perspectives for the creation of forecasting models, particularly in the ecological field. The ability of not requiring a prefixed model structure represent a huge advantage over typical statistical models. In conjunction with the explorability of the model, GP allows to automatically discover rather complex relationships among variables, which is a really attractive feature in the ecological field. Future development could be the application of a vector based GP, as some of the predictors involve time series variables.

References

1. Diamond, M.S. (ed.): West Nile Encephalitis Virus Infection: Viral Pathogenesis and the Host Immune Response. Emerging Infectious Diseases of the 21st Century, Springer-Verlag, New York (2009)
2. Chambers, T., Monath, T.: The Flaviviruses: Detection, Diagnosis and Vaccine Development. Advances in Virus Research, Elsevier Science (2003). https://doi.org/10.1016/S0065-3527(03)61017-1
3. Sfakianos, J.N.: West Nile Virus. Chelsea House Publications (2005)
4. Kramer, L.D., Styer, L.M., Ebel, G.D.: A global perspective on the epidemiology of West Nile virus. Annual Review of Entomology **53**, 61–81 (2008). https://doi.org/10.1146/annurev.ento.53.103106.093258
5. Istituto Superiore di Sanità, http://old.iss.it/
6. Autorino, G.L., Battisti, A., Deubel, V., Ferrari, G., Forletta, R., Giovannini, A., Lelli, R., Murri, S., Scicluna, M.T.: West Nile virus Epidemic in Horses, Tuscany Region, Italy. Emerging Infectious Diseases **8**(12), 1372–1378 (2002). https://doi.org/10.3201/eid0812.020234

7. Ministero della Salute. Piano di sorveglianza nazionale per la encefalomielite di tipo West Nile (West Nile Disease). Gazzetta Ufficiale della Repubblica Italiana, N. 113 (16/05/2002)
8. EpiCentro - Portale di epidemiologia, http://www.epicentro.iss.it/
9. Zeller, H.G., Schuffenecker, I.: West Nile virus: an overview of its spread in Europe and the Mediterranean basin in contrast to its spread in the Americas. European Journal of Clinical Microbiology & Infectious Diseases: Official Publication of the European Society of Clinical Microbiology **23**(3), 147–156 (2004). https://doi.org/10.1007/s10096-003-1085-1
10. Becker, N., Jöst, A., Weitzel, T.: The Culex pipiens Complex in Europe. Journal of the American Mosquito Control Association **28**(4 Suppl), 53–67 (2012). https://doi.org/10.2987/8756-971X-28.4s.53
11. Bisanzio, D., Giacobini, M., Bertolotti, L., Mosca, A., Balbo, L., Kitron, U., Vazquez-Prokopec, G.M.: Spatio-temporal patterns of distribution of West Nile virus vectors in eastern Piedmont Region, Italy. Parasites & Vectors **4**, 230 (2011). https://doi.org/10.1186/1756-3305-4-230
12. Arpa Piemonte, http://www.arpa.piemonte.it
13. NASA MODIS Web, https://modis.gsfc.nasa.gov/
14. Stroup, W.: Generalized Linear Mixed Models: Modern Concepts, Methods and Applications. CRC Press (2012)
15. Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van-der Linde, A.: Bayesian measures of model complexity and fit. Journal of the Royal Statistycal Society **64**(4), 583–639 (2002). https://doi.org/10.1111/1467-9868.02022
16. Rue, H., Martino, S., Chopin, N.: Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **71**(2), 319–392 (2009). https://doi.org/10.1111/j.1467-9868.2008.00700.x
17. Rosà, R., Marini, G., Bolzoni, L., Neteler, M., Metz, M., Delucchi, L., Chadwick, E.A., Balbo, L., Mosca, A., Giacobini, M., Bertolotti, L., Rizzoli, A.: Early warning of West Nile virus mosquito vector: climate and land use models successfully explain phenology and abundance of Culex pipiens mosquitoes in north-western Italy. Parasites & Vectors **7**, 269 (2014). https://doi.org/10.1186/1756-3305-7-269
18. Koza, J.R.: Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge, MA, USA (1992)
19. Silva, S.: GPLAB - A Genetic Programming Toolbox for MATLAB, http://gplab.sourceforge.net/index.html
20. Luke, S., Panait, L.: Lexicographic parsimony pressure. In: Proceedings of the 4th Annual Conference on Genetic and Evolutionary Computation. pp. 829–836. GECCO'02, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2002)
21. Vargha, A., Delaney, H.D.: A Critique and Improvement of the "CL" Common Language Effect Size Statistics of McGraw and Wong. Journal of Educational and Behavioral Statistics **25**(2), 101–132 (2000). https://doi.org/10.2307/1165329
22. A vectorial approach to genetic programming. Submitted to EuroGP 2019
23. Poli, R., Langdon, W., McPhee, N., Koza, J.: A Field Guide to Genetic Programming. Lulu.com (2008). https://doi.org/10.1007/s10710-008-9073-y
24. European Environment Agency, https://www.eea.europa.eu/