PREDICTING *Aedes aegypti* EGGS COUNT USING REMOTE SENSING DATA AND A GENERALIZED LINEAR MODEL

Oladimeji Mudele\textsuperscript{a}, Verónica Andreo\textsuperscript{b,c}, Ximena Porcasi\textsuperscript{b}, Carlos M. Scavuzzo\textsuperscript{b}, Laura Lopez\textsuperscript{d}, Paolo Gamba\textsuperscript{a}

\textsuperscript{a}Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy. oladimeji.mudele01@universitadipavia.it, paolo.gamba@unipv.it

\textsuperscript{b}Instituto de Altos Estudios Espaciales "Mario Gulich". Falda del Cañete, Córdoba, Argentina - veronica.andreo@ig.edu.ar, ximena.porcasi@conae.gov.ar, scavuzzo@conae.gov.ar

\textsuperscript{c}CONICET, Consejo Nacional de Investigaciones Científicas y Técnicas. Buenos Aires, Argentina

\textsuperscript{d}Ministerio de Salud de la Provincia de Córdoba. Córdoba, Argentina - laulop@hotmail.com

\begin{abstract}
Here, we present a method for temporal modeling of the oviposition activity of *Ae. aegypti* mosquitoes based on a weighted generalised linear model (GLM) with explanatory environmental effects extracted from freely available remotely sensing (satellite) images. Our results show potential for operational applications. Experimental results are provided using field collected *Ae. aegypti* eggs count data in Córdoba, Argentina.

\textbf{Index Terms}— Remote sensing, Epidemiology
\end{abstract}

\section{INTRODUCTION}

Remote sensing (RS) data provide information about the Earth's surface at an unprecedented scale. As a result, it has been applied in many domains. One of such domains is landscape epidemiology which focuses on using remotely sensed data information to understand and model the dynamics of environment-dependent disease risk [1, 2, 3]. Traditionally, in order to model health-related risk proxies in urban areas, post-hoc outbreak evaluations were performed. This approach, however, is only effective \textit{a posteriori}. For prevention purposes, monitoring of vector dynamics and disease spread should be performed ahead of outbreak events [1]. Accordingly, there have been multiple attempts in landscape epidemiology towards implementing remote sensing-based systems to predict or forecast, for instance, the oviposition activity and adult population of mosquito species which are known vectors of widespread diseases [3, 4, 5].

Dengue, Zika and Chikungunya which are widespread in over 100 countries in the world are transmitted by the *Aedes aegypti* mosquito species [6]. This mosquito is fully adapted to urban areas and breeds in artificial water containers. Moreover, its oviposition activity, adult development, and disease transmission capabilities are influenced by environmental variables including temperature, humidity, precipitation and vegetation condition [6, 7]. Since these environmental variables have RS-based proxy estimates, it is possible to model the population dynamics of *Aedes aegypti* at scale with RS data; given the right data and modeling technique.

In this study, we apply the weighted GLM technique initially proposed in [8] for modeling and extrapolation of *Aedes aegypti* eggs count (not adult population) based on covariates estimated from RS data. In addition, we apply the method on data from Córdoba, Argentina. As in [8], we compare the results with those obtained by machine learning (ML) methods. In addition, we use the RS variables coefficients from the resulting best model to show how it can be interpreted to support the Argentinean operational risk system.

\section{MODELING}

In this work, the rounded-up mean *Ae. aegypti* eggs count \((Y)^1\) is modeled as a function of environmental variables \((X)\) derived from RS data. For this reason, a Poisson GLM with a logarithm link function is used. It has been shown [9] that this model is suitable to model non-negative discrete non-Gaussian distributed variables. To improve our model quality, we further fit a weighted GLM, labeled GLM-W. The weights \(w\) used to fit GLM-W are obtained as \(w_i = |y_i - \tilde{y}_i|^{-1}\), where \(|y_i - \tilde{y}_i|\) is the \(i\)-th week residual of our initially unweighted GLM. We then apply a step-wise regression [10] based on the Akaike Information Criterion (AIC) [11] to drop uninformative environmental variables from GLM-W. The resulting

\[^{1}\text{Note that a ceiling function is used here to map a real number to an integer.}\]
model is labeled GLM-W∗.

To model the egg counts \( Y \in \mathbb{N} \) as a function of \( X \in \chi \subset \mathbb{R}^p \), we estimate a function, \( m \), such that \( m(x) = \mathbb{E}(Y \mid X = x) \), i.e.

\[
\hat{y}_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} = x_i^T \beta,
\]

where the superscript \( T \) denotes transposition; \( \hat{y}_i \) is the predicted egg count in the \( i \)-th epidemiological week; \( k \) is the number of RS-based environmental variables used to fit the model; \( x_{ij} \) is the value of the \( j \)-th environmental variable in the \( i \)-th week; \( \beta_0 \) is the value of \( \hat{y} \) when all the environmental variables are equal to zero; and \( \beta_j \) is the gradient of \( \hat{y} \) with respect to the \( j \)-th environmental variable. Our observed egg count \( y_i \) is an instance of the random variable \( Y_i \) with mean estimated as \( \hat{y} \) by our model.

The random component of the model is the Poisson distribution for \( y_i \subset Y \). Under this model, the observations are outcomes of a probability process of the form:

\[
\Pr(y_i = y \mid \mu_i) = \frac{e^{-\mu_i} \mu_i^y}{y!},
\]

where \( \mu_i = \mathbb{E}(y_i \mid x_i) \geq 0 \) is the mean and variance of the distribution. The relationship between \( \mu_i \) and the linear predictor is given by a link function, \( g \), such that \( g(\mu_i) = x_i^T \beta \); cf. [12, 9]. As mentioned, a logarithm link function is applied.

For comparison purposes, two ML techniques have been used: Random Forest (RF) and Support Vector Machines (SVM). We assess the quality of all models using the AIC for GLMs only, and the Root Mean Squared Error (RMSE) [13] for all models.

3. MATERIALS

The work refers to the city of Córdoba, the second largest city in Argentina. Entomological data consists of 300 ovitraps distributed in 150 houses over the city (Fig. 1). They were placed in the front yard of each house, usually in shaded places and below or close to bushes or pots with plants. All ovitraps were replaced weekly for two years for a total of 105 weeks from week 40 in 2017 to week 40 in 2019. The average egg count (rounded up) across all 150 traps is obtained for each week as the target variable of this study.

For the purposes of this study, selected environmental variables (temperature, humidity, precipitation and vegetation condition), are represented by proxy features estimated from freely accessible RS products. Specifically, the Enhanced Vegetation Index (EVI), Normalised Difference Water Index (NDWI), and day and nighttime land surface temperature layers of the Moderate Resolution Imaging Spectroradiometer (MODIS) data were used, as in other recent studies [3, 6], as proxies to humidity, temperature and vegetation condition.

Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) data was instead used for precipitation information [14].

To obtain the modeling explanatory variables, the average value for each RS environmental variable across all ovitraps locations (latitude and longitude) is obtained per week. In this way, we work around the differences in spatial resolution among the considered RS products. Variables representing different time intervals were interpolated to obtain weekly values using a third order spline interpolation. As in [1], time-delayed observations with weekly steps up to two weeks are included to account for non-synchronous environmental effects.

To obtain the modeling target variable, the (rounded-up) mean eggs count is obtained for each observed week. 80% of the data (the first 84 weeks) are used to train the models, and the remaining 21 weeks of data are used for model validation.

4. EXPERIMENTAL RESULTS AND DISCUSSION

By including for each of these variable up to two weeks of lag, we reach a total of 15 environmental covariates.

Table 1 shows quality measures for all fitted models. First, we consider GLM, GLM-W, and GLM-W∗. Considering AIC, GLM-W produces an improvement in prediction quality with respect to GLM. This is not the case considering RMSE. It is sufficient, however, to rely on AIC in comparing these two models because it is a measure of the absolute fit corrected for the number of predictors in the model. Also, it can be seen from this table that even with less variables in GLM-W∗ (see Table 2 for selected variables), we are able to obtain a slight improvement in AIC and RMSE values.

Fig. 2 presents line plots comparing observed and fitted values by GLM, GLM-W, and GLM-W∗. Looking at Fig. 2(a), we see that its predicted values and 95% confidence intervals (CI) do not reach the highest observed values which are around weeks 75-85 (corresponding to first week of March to second week of May in 2019). However, as shown in Fig. 2(b), GLM-W slightly improves the predic-
Table 1: Quality measures for GLM and GLM-W. RMSE is the root mean square error.

| Model | AIC   | RMSE  |
|-------|-------|-------|
| GLM   | 572.170 | 5.619 Training |
|       |        | 12.150 Validation |
| GLM-W | 346.932 | 5.951 Training |
|       |        | 12.600 Validation |
| GLM-W*| 342.173 | 6.186 Training |
|       |        | 12.544 Validation |
| RF    | -     | 3.758 Training |
|       |        | 23.392 Validation |
| SVM   | -     | 4.983 Training |
|       |        | 16.126 Validation |

Table 2: Model summary for GLM-W*. 12 out of the 15 variables are retained by AIC-based stepwise regression. β symbolizes the remote sensing data variables as introduced in Eqn. 1. TempN and TempD represent day and nighttime land surface temperature variables. #-1 and #-2 are variables with one and two weeks of lag, respectively.

| β     | Std. Error | p-value |
|-------|------------|---------|
| EVI-2 | 17.3890    | 3.2502  | 8.79 × 10^{-08} |
| NDWI  | 16.0075    | 3.5746  | 7.53 × 10^{-06} |
| NDWI-1| 10.2235    | 1.8757  | 5.02 × 10^{-08} |
| TempN-1| 0.0896    | 0.0333  | 7.20 × 10^{-03} |
| TempN-2| 0.0007    | 0.0213  | 4.50 × 10^{-03} |
| TempN  | 0.0905    | 0.0381  | 1.45 × 10^{-01} |
| Prec-2 | -0.0098   | 0.0031  | 1.80 × 10^{-03} |
| Prec   | -0.0322   | 0.0027  | 1.68 × 10^{-06} |
| Prec-1 | -0.0143   | 0.0034  | 3.16 × 10^{-05} |
| TempD-2| -0.0797   | 0.0260  | 2.20 × 10^{-03} |
| TempD  | -0.0926   | 0.0324  | 4.20 × 10^{-03} |
| EVI    | -31.9761  | 5.8993  | 5.95 × 10^{-08} |
| (Intercept) | 4.5104 | 0.6106  | 1.50 × 10^{-13} |

Table 3: Summary of all the observed and fitted models on validation data

| y     | Min Q1 Median Mean Q3 Max |
|-------|---------------------------|
| GLM   | 0.894 2.126 10.111 14.887 22.346 54.081 |
| GLM-W | 0.749 1.891 9.392 14.546 21.205 55.702 |
| GLM-W*| 0.721 1.892 9.746 14.938 21.319 58.061 |
| SVM   | -1.477 1.348 7.327 10.614 19.487 38.832 |
| RF    | 3.100 4.056 6.250 6.516 7.953 13.359 |

5. CONCLUSIONS

In this work, a Poisson GLM has been used to model and extrapolate Ae. aegypti eggs count based on relevant environmental variables (precipitation, temperature, humidity and vegetation condition) extracted from RS data.

The experiments show that the proposed weighted model performs better than the standard GLM model. It also proved to be robust when qualitatively compared with more complex ML algorithms, such as RF and SVM.

The proposed Poisson model ignores potential temporal autocorrelation of the egg counts data. Future studies can consider ways to include autoregressive components in the model.

References

[1] Oladimeji Mudele, Fábio M Bayer, Lucas FR Zanandrez, Alvaro E Eiras, and Paolo Gamba. Modeling the temporal population distribution of Ae.aegypti mosquito using big earth observation data. IEEE Access, 8:14182–14194, 2020.

[2] Simon I. Hay, Katherine E. Battle, David M. Pigott, David L. Smith, Catherine L. Moyes, Samir Bhatt, John S. Brownstein, Nigel Collier, Monica F. Myers, Dylan B. George, and Peter W. Gething. Global
First 84 weeks are used in model training. The rest are used in validation.

Fig. 2: Observed (points) and predicted (lines) values for GLM, GLM-W and GLM-W* with 95% confidence intervals (CI).

Fig. 3: Comparing GLM-W* with the fitted machine learning models: RF and SVM.