Geomorphometric analysis of wildfire occurrence in a humid tropical protected area: a case study in Southern Pacific Costa Rica

Shu Wei Chou-Chen1,✉, Adolfo Quesada-Román2, and Daniela Vargas-Sanabria3

1 Escuela de Estadística, Universidad de Costa Rica  
2 Physical Geography Lab, School of Geography, University of Costa Rica  
3 Laboratorio de Investigacion e Innovación Tecnológica. Universidad Estatal a Distancia, Costa Rica.

✉ Correspondence: [Shu Wei Chou-Chen <shuwei.chou@ucr.ac.cr>](mailto:shuwei.chou@ucr.ac.cr)

# 1. Introduction

…

# 2. Methods

## 2.1 Data

Data description…

The explanatory variables, FA and TRI, are logarithmically transformed as to address their asymmetry and to avoid undefined values for . Subsequently, all covariates (DAH, logFA, LST, logLSF, SLOPE, logTRI, TWI, WE, and WExpo) are scaled using z-scores to enable the comparison of their standardized contributions.

## 2.2 Statistical models and analysis approach

The goal is to model a count variable (INC), which ranges from 0 to 6 and has a high percentage of zeros (). Therefore, we considered count models, such as Poisson and Negative Binomial regression, as well as the zero-inflated count regression models. There are different ways in R (R Core Team 2023) that implement zero-inflated count models throughout different packages (Zeileis, Kleiber, and Jackman 2008). We performed all the statistical modeling by using a flexible statistical framework called Generalized Additive Models for Location, Scale, and Shape (GAMLSS), since all these models are special cases within this model class. All the statistical analysis in this paper was conducted using the gamlss package (R. A. Rigby and D. M. Stasinopoulos 2005). Although this model allows for non-linear effects from the covariates, we retain the model with linear covariates for simplicity.

Let be the INC of location . The count regression models described above are represented by either one-, two-, or three-parameter GAMLSS models:

where the response variable (INC) is distributed as a two-parameter or three-parameter distribution : the location or mean (), the scale (), and a parameter related to the skewness of the distribution (), as well as the link functions ( for ) for each parameter, and are the regression coefficients of the covariate , measured at location , for each parameter function .

Where the response variable (INC) is distributed as a two-parameter or three-parameter distribution : the location or mean (), the scale (), and a parameter related to the skewness of the distribution (), as well as the link functions ( for ) for each parameter, and are the regression coefficients of the -th covariate , measured at location , for each parameter function .

Regarding to the specification of , the Poisson regression model (PO) includes only one parameter with link function . Then, Negative Binomial model (NB) fits two parameters and , with link functions and .

For the zero-inflated models, a finite mixture distribution of a point mass at zero and a count distribution is specified as follows:

where represents the probability of observing , and is the probability function of a count distribution (either Poisson or Negative Binomial). In this way, zero-inflated models add an additional parameter to the GAMLSS framework.”

That is, the Poisson Zero-Inflated model (ZIP) includes two parameters, with and . Finally, the Zero-Inflated Negative Binomial model (ZINB) fits three parameters, with , , and . Note that the second parameter of ZIP and the third parameter of ZINB model the probability that a given location has zero wildfires, similar to a logistic regression.

Additionally, due to the large number of data points (60,484 locations), we conducted the analysis using a training-testing approach (R. A. Rigby and D. M. Stasinopoulos 2005). In other words, we divided the data into two subsets: a training set comprising 70% (42,213 locations), used for model development and selection, and a testing set comprising 30% (18,271 locations), utilized for out-of-sample validation.

The analysis strategy and the selection of the final model are as follows. First, we considered the models’ mean : PO, NB, ZIP, and ZINB, with full specifications of the location function, while holding other parameters constant. Next, by applying the stepwise algorithm in both directions, we obtained reduced models (PO-red, NB-red, ZIP-red, and ZINB-red). Since the zero-inflated models yielded better results, in the third step, we applied the stepwise algorithm again to add to ZIP-red and both and to ZINB-red, resulting in the final models, ZIP-red-final and ZINB-red-final.”

Finally, the best model is selected based on several metrics and an assessment of model adequacy. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), along with pseudo- (Cox-Snell and Cragg-Uhler), are computed for the training set, while the validation global deviance (VGD) is calculated for the testing set (out-of-sample validation).

# 3. Results

## 3.1 Descriptive analysis

## 3.2

[Table 1](#tbl-ICmodels) shows the goodness of fit measures for the fitted models. Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) evaluate the goodness of fit for the training set, while Validated Globa Deviance (VGD) measures how well the fitted models perform for predicting the out-of-sample data (validation data). All three criterion shows that the best model is ZIP\_red\_final. On the other hand, note that the generalized pseudo r-squared (CoxSmell and Cragg Uhler) show that PO\_red are higher among all models, but the assumptions of this model are not satisfied. On other other hand, all assumptions of ZIP\_red\_final are satisfied and present the best measures (**aquí falta presentar los diagnósticos del modelo y pregunto si para artículos en su área, esto se presenta en el cuerpo del artículo o generalmente no lo toman con tanta importancia y podría ponerlos en anexo**).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Goodness of fit measures for the fitted models.   | Model | AIC | BIC | VGD | R2-CoxSnell | R2-Cragg Uhler | | --- | --- | --- | --- | --- | --- | | PO | 79,265.66 | 79,360.82 | 34,836.93 | 0.2723 | 0.3065 | | NBII | 70,882.70 | 70,986.50 | 31,089.78 | 0.1526 | 0.1812 | | ZIP | 73,309.74 | 73,413.55 | 32,133.96 | 0.0883 | 0.1052 | | ZINBI | 76,092.71 | 76,205.17 | 33,380.19 | 0.0219 | 0.0261 | | PO\_red | 79,269.74 | 79,356.24 | 34,842.41 | 0.2722 | 0.3063 | | NBII\_red | 70,881.29 | 70,976.45 | 31,090.24 | 0.1526 | 0.1812 | | ZIP\_red | 73,308.00 | 73,403.16 | 32,134.62 | 0.0883 | 0.1052 | | ZINBI\_red | 72,903.31 | 72,981.16 | 31,869.63 | 0.0929 | 0.1108 | | ZIP\_red\_final | 69,754.09 | 69,892.50 | 30,790.78 | 0.1621 | 0.1931 | | ZINBI\_red\_final | 69,532.61 | 69,688.32 | 30,566.35 | 0.1629 | 0.1942 | |

[Table 2](#tbl-ZIPmodels) describes the model estimates. The equation for Mu fits the wildfire counts, while the equation for Sigma models the probability of observing zero wildfires in a given centroid. We observe that all these variables are statistically significant, except for WE in the Mu equation.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
Family: c("ZIP", "Poisson Zero Inflated")   
  
Call: gamlss(formula = formula, sigma.formula = sigma.formula,   
 nu.formula = nu.formula, tau.formula = tau.formula,   
 family = "ZIP", data = data, control = control)   
  
Fitting method: RS()   
  
------------------------------------------------------------------  
Mu link function: log  
Mu Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.165197 0.013430 12.301 < 2e-16 \*\*\*  
DAH -0.027644 0.007724 -3.579 0.000346 \*\*\*  
LST -0.405316 0.024243 -16.719 < 2e-16 \*\*\*  
logLSF 0.394362 0.024090 16.371 < 2e-16 \*\*\*  
SLOPE -0.384874 0.026021 -14.791 < 2e-16 \*\*\*  
logTRI -0.001229 0.029799 -0.041 0.967095   
TWI -0.461550 0.023897 -19.314 < 2e-16 \*\*\*  
WE -0.011984 0.013139 -0.912 0.361688   
WExpo -0.199283 0.017533 -11.366 < 2e-16 \*\*\*  
sabanaTRUE 0.471340 0.018106 26.033 < 2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
------------------------------------------------------------------  
Sigma link function: logit  
Sigma Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.84043 0.02063 40.741 <2e-16 \*\*\*  
sabanaTRUE -2.48439 0.06582 -37.747 <2e-16 \*\*\*  
LST -0.31938 0.03054 -10.458 <2e-16 \*\*\*  
WE 0.19459 0.02106 9.240 <2e-16 \*\*\*  
TWI 0.25136 0.02478 10.144 <2e-16 \*\*\*  
WExpo 0.25461 0.03096 8.224 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
------------------------------------------------------------------  
No. of observations in the fit: 42213   
Degrees of Freedom for the fit: 16  
 Residual Deg. of Freedom: 42197   
 at cycle: 18   
   
Global Deviance: 69722.09   
 AIC: 69754.09   
 SBC: 69892.5   
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2: Parameter estimates of the fitted models.   | Covariate | Estimate | Std. Error | t value | Pr(>|t|) | | --- | --- | --- | --- | --- | | Mu |  |  |  |  | | (Intercept) | 0.1652 | 0.0134 | 12.3008 | 0.0000 | | DAH | -0.0276 | 0.0077 | -3.5788 | 0.0003 | | LST | -0.4053 | 0.0242 | -16.7190 | 0.0000 | | logLSF | 0.3944 | 0.0241 | 16.3706 | 0.0000 | | SLOPE | -0.3849 | 0.0260 | -14.7907 | 0.0000 | | logTRI | -0.0012 | 0.0298 | -0.0413 | 0.9671 | | TWI | -0.4616 | 0.0239 | -19.3145 | 0.0000 | | WE | -0.0120 | 0.0131 | -0.9122 | 0.3617 | | WExpo | -0.1993 | 0.0175 | -11.3659 | 0.0000 | | sabanaTRUE | 0.4713 | 0.0181 | 26.0327 | 0.0000 | | Sigma |  |  |  |  | | (Intercept) | 0.8404 | 0.0206 | 40.7412 | 0.0000 | | sabanaTRUE | -2.4844 | 0.0658 | -37.7473 | 0.0000 | | LST | -0.3194 | 0.0305 | -10.4579 | 0.0000 | | WE | 0.1946 | 0.0211 | 9.2396 | 0.0000 | | TWI | 0.2514 | 0.0248 | 10.1444 | 0.0000 | | WExpo | 0.2546 | 0.0310 | 8.2237 | 0.0000 | |

Finall, [Figure 1](#fig-ZIPmusigma) presents the fitted and , which represent wildfire mean of each centroid and the probability of no wildfire, respectively. Those figures are difficult to assess due to the fact that [Figure 1 (a)](#fig-ZIPmu) shows higher fitted wildfire mean for northern parts of the region. However, if we take into account that the model fits a centroid with , meaning this centroid has greater probability to not have wildfire, we can filter out those centroids, and only consider those centroid with , that is, those places will have more than 50% of chance to have more than zero wildfires, then plot the fitted ([Figure 2 (b)](#fig-ZIPmu_sigma)). We can observe that the places with higher wildfire count mean are similar to the observed INC (fig-INC).

Finally, [Figure 1](#fig-ZIPmusigma) presents the fitted and , which represent the mean wildfire count for each centroid and the probability of no wildfire, respectively. These figures can be challenging to interpret, as [Figure 1 (a)](#fig-ZIPmu) shows a higher fitted wildfire mean in the northern parts of the region, which have high probability to not observe wildfire. However, if we consider that the model fits a centroid with , indicating a greater probability of no wildfire, we can filter out those centroids and focus only on those with . This means that these locations have more than a 50% chance of experiencing at least one wildfire. We can then plot the fitted ([Figure 2 (b)](#fig-ZIPmu_sigma)). Notably, the areas with higher mean wildfire counts are similar to the observed INC ([Figure 2 (a)](#fig-INC)).

|  |  |  |  |
| --- | --- | --- | --- |
| |  | | --- | | (a) ZIP mean. | | |  | | --- | | (b) ZIP sigma. | |

Figure 1: The predicted and of the fitted ZIP model.

|  |  |  |  |
| --- | --- | --- | --- |
| |  | | --- | | (a) Wildfire counts. | | |  | | --- | | (b) Predicted fire count mean, conditional on a probability greater than 0.5 of fire. | |

Figure 2: Observed wildfire counts and the model predicted fire count mean, conditional on a probability greater than 0.5 of occurrence of wildfire.

# 4. Conclusiones

…

# 5. Referencias

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.

R. A. Rigby, and D. M. Stasinopoulos. 2005. “Generalized Additive Models for Location, Scale and Shape,(with Discussion).” *Applied Statistics* 54.3: 507–54.

Zeileis, Achim, Christian Kleiber, and Simon Jackman. 2008. “Regression Models for Count Data in r.” *Journal of Statistical Software* 27 (8): 1–25. <https://doi.org/10.18637/jss.v027.i08>.