SCIENTIFIC REPORTS

OPEN

Ordinal regression models for zero-inflated and/or over-dispersed count data

Received: 4 June 2018
Accepted: 23 January 2019
Published online: 28 February 2019

Denis Valle[1], Kok Ben Toh[2], Gabriel Zorello Laporta[3,4] & Qing Zhao[1]

Count data commonly arise in natural sciences but adequately modeling these data is challenging due to zero-inflation and over-dispersion. While multiple parametric modeling approaches have been proposed, unfortunately there is no consensus regarding how to choose the best model. In this article, we propose a ordinal regression model (MN) as a default model for count data given that this model is shown to fit well data that arise from several types of discrete distributions. We extend this model to allow for automatic model selection (MN-MS) and show that the MN-MS model generates superior inference when compared to using the full model or more traditional model selection approaches. The MN-MS model is used to determine how human biting rate of mosquitoes, known to be able to transmit malaria, are influenced by environmental factors in the Peruvian Amazon. The MN-MS model had one of the best fit and out-of-sample predictive skill amongst all models. While *A. darlingi* is strongly associated with highly anthropized landscapes, all the other mosquito species had higher mean biting rates in landscapes with a lower fraction of exposed soil and urban area, revealing a striking shift in species composition. We believe that the MN and MN-MS models are valuable additions to the modelling toolkit employed by environmental modelers and quantitative ecologists.

Count data are ubiquitous in natural sciences[1–8] and other fields[9–13]. The default modeling choice for count data has traditionally been a Poisson regression but it is widely acknowledged that a Poisson likelihood is a poor choice for over-dispersed and/or zero-inflated data and different conclusions may be reached depending on whether zero-inflation and/or over-dispersion are properly accommodated or not[3,8,14]. As a result, considerable research has been devoted to devising alternative statistical modeling approaches to properly accommodate these count data characteristics. A common alternative to the Poisson regression model that accounts for over-dispersion is the negative-binomial [NB] regression model[6,10,11,14,15]. However, other models also exist (e.g., new parameterization of the NB distribution that allows for different quadratic mean-variance relationships[7], the Generalized Poisson distribution[12], and the Quasi-Poisson regression[2]). Similarly, besides the negative-binomial regression model[1,16], various hurdle and mixture models have been proposed in the literature to appropriately deal with zero-inflation (ZI)[3,4,8].

As a result of the large number of potential models for count data and the fact that model choice has important consequences for the derived conclusions, choosing the most appropriate model is critical, even amongst models that properly accommodate over-dispersion and/or zero-inflation[2,3,7,8,14]. Despite substantial research comparing different statistical models using a range of criteria[1,3,12,16,17], several researchers have ultimately concluded that determining the best modeling approach for count data is challenging[2,7].

In this article, we propose a Bayesian ordinal regression model that can flexibly fit count data that arise from various distributions, regardless of zero-inflation and/or over-dispersion, circumventing the need to choose the most appropriate distribution. Furthermore, we extend this model to allow for model selection and parameter estimation within a single coherent modeling framework, enabling researchers to more fully explore the information from covariates (e.g., by accounting for non-linear relationships). We compare the performance of the proposed model to that of other commonly used models using simulations and real data. More specifically, our simulations explore how well the proposed model works for inferential purposes, including how well it (a) fits

[1]School of Forest Resources and Conservation, University of Florida, Gainesville, Florida, United States of America. [2]School of Natural Resources and Environment, University of Florida, Gainesville, Florida, United States of America. [3]Setor de Pos-graduacao, Pesquisa e Inovacao, Faculdade de Medicina do ABC, Santo Andre, Sao Paulo, Brazil. [4]Centro de Engenharia, Modelagem e Ciencias Sociais Aplicadas, Universidade Federal do ABC, Santo Andre, Sao Paulo, Brazil. Correspondence and requests for materials should be addressed to D.V. (email: drvalle@ufl.edu)

data that arise from different distributions, (b) determines which predictors are associated with the response variable (i.e., model selection), and (c) characterizes the (possibly nonlinear) relationship between the response variable and predictor variables. Our case study focuses on determining how land-use/land-cover and precipitation influence malaria risk by modeling mosquito data collected in the Peruvian Amazon. Finally, we end this article with a discussion on important topics for future research.

Methods

Basic model formulation (MN model). A multinomial distribution can approximate any given discrete marginal distribution, with or without zero-inflation and/or over-dispersion. As a result, we rely on the multinomial distribution as the basis of our model and we hypothesize that an ordered multinomial probit model (MN model), also known as an ordinal regression model, can represent a wide range of regression models (i.e., conditional distributions).

Here we described the basic structure of a probit ordinal regression model[18]. We start by ranking the response variable w_i and let $y_i = rank\ (w_i)$, where ties are assigned the same ranking value (i.e., if $w_i = w_k$, then $y_i = y_k$ for $i \neq k$). Therefore, $y_i \in \{1, 2, \ldots, J\}$ where J is the total number of unique w_i values. We assume that:

$$y_i = 1 \text{ if } z_i < b_1$$
$$y_i = j \text{ if } b_{j-1} < z_i < b_j \text{ for j} = 2, \ldots, \text{J} - 1$$
$$y_i = J \text{ if } z_i > b_{J-1}$$

where $b_1, \ldots, b_{J-1}$ are breaks to be estimated and z_i is a continuous latent variable. We further assume that z_i is given by:

$$z_i \sim N(\boldsymbol{x_i}^T\boldsymbol{\beta}, 1)$$

where $\boldsymbol{x_i}^T$ is the design vector and $\boldsymbol{\beta}$ is a vector of regression parameters. For identifiability purposes, we either have to set one of the breaks $b_1, \ldots, b_{J-1}$ to zero or eliminate the intercept from our regression. We opt for the latter because it is not clear which break should be set to zero. Therefore, the design vector $\boldsymbol{x_i}$ does not include a 1 for the intercept.

We use uninformative priors:

$$[b_1, \ldots, b_{J-1}] \sim Unif(-100, 100)I(b_1 < \ldots < b_{J-1})$$
$$\boldsymbol{\beta} \sim N(\mathbf{0}, \boldsymbol{I})$$

Finally, we note that the expected count is given by:

$$E[w_i|\boldsymbol{x_i}] = \sum_{j=1}^{J} u_j \times p(y_i = j|\boldsymbol{x_i}) = \sum_{j=1}^{J} u_j \times [\Phi(b_j - x_i^T\beta) - \Phi(b_{j-1} - x_i^T\beta)]$$

where $\Phi()$ is the cumulative density function of a standard normal distribution, u_j are the ordered unique values of w_i, and $b_0 = -\infty$ and $b_J = \infty$. We rely on this expression for the expected count to create response curves depicting the effect of different covariates. The MN model can be fitted in a straight-forward fashion using standard methods in R, as illustrated in S1 Appendix.

Simultaneously performing model fitting and model selection (MN-MS model). The basic model formulation provided above can be extended to perform model selection and model fitting at the same time (MN-MS model). We start by noticing that the marginal probability associated with a particular model M_k, defined by the subset of covariates k, can be calculated in closed form after integrating out the associated regression parameters $\boldsymbol{\beta_k}$. This is given by:

$$p(M_k|\boldsymbol{z}) \propto \int N(\boldsymbol{z}|\boldsymbol{X_k}\boldsymbol{\beta_k}, \boldsymbol{I})\mathrm{N}(\beta_{\boldsymbol{k}}|0, \mathbf{I})d\boldsymbol{\beta_k}$$
$$\propto exp\left(-\frac{1}{2}[-\boldsymbol{\mu_k}^T\boldsymbol{\Sigma_k^{-1}}\boldsymbol{\mu_k} + \boldsymbol{z}^T\boldsymbol{z}]\right)|\boldsymbol{\Sigma_k}|^{\frac{1}{2}}$$

where $\{\boldsymbol{X_k}^T\boldsymbol{X_k} + \boldsymbol{I}\} = \boldsymbol{\Sigma_k^{-1}}$ and $\boldsymbol{\mu_k} = \boldsymbol{\Sigma_k}\boldsymbol{X_k}^T\boldsymbol{z}$. In these equations, $\boldsymbol{X_k}$ is the design matrix with only the subset of covariates k. Details on this integration can be found in S2 Appendix. Following Denison *et al.*[19], we set the prior for each model M_k as $p(M_k) \propto (P + 1)^{-1}\binom{P}{p_k}^{-1}$, where p_k is the number of covariates in set k. In this prior, each number of covariates 0, …, P (P is the overall number of covariates) is assumed to be equally likely, represented by $\frac{1}{P+1}$. Furthermore, this prior assumes that all models with a given number of covariates p_k are equally likely, represented by $\binom{P}{p_k}^{-1}$, where $\binom{P}{p_k}$ is the number of possible combinations of p_k elements out of P.

Our algorithm explores model space by randomly proposing the birth of a new covariate or the death or swap of an existing covariate. These proposed moves are then accepted or rejected using a standard Metropolis-Hastings acceptance ratio given by:

Reg. model	Mean	Variances	Assumptions	Parameter values
Poisson	Small	—	$w_i \sim Poisson\ (\lambda_i)$	$\beta_0 = \log(1);\ \beta_1 = 0.5$
	Large	—	$w_i \sim Poisson\ (\lambda_i)$	$\beta_0 = \log(5);\ \beta_1 = 0.5$
NB	Small	Small	$w_i \sim Neg\ Binom\ (\mu_i = \lambda_i, n)$	$\beta_0 = \log(1);\ \beta_1 = 0.5;\ n = 1$
	Small	Large	$w_i \sim Neg\ Binom\ (\mu_i = \lambda_i, n)$	$\beta_0 = \log(1);\ \beta_1 = 0.5;\ n = 0.1$
	Large	Small	$w_i \sim Neg\ Binom\ (\mu_i = \lambda_i, n)$	$\beta_0 = \log(5);\ \beta_1 = 0.5;\ n = 1$
	Large	Large	$w_i \sim Neg\ Binom\ (\mu_i = \lambda_i, n)$	$\beta_0 = \log(5);\ \beta_1 = 0.5;\ n = 0.1$
ZIP	Small	—	$q_i \sim Bernoulli\ (\pi_i)$ $w_i \sim Poisson\ (\lambda_i \times q_i)$	$\alpha_0 = \log(3);\ \alpha_1 = 0.5;$ $\beta_0 = \log\left(\frac{4}{3}\right);\ \beta_1 = 0.5$
	Large	—	$q_i \sim Bernoulli\ (\pi_i)$ $w_i \sim Poisson\ (\lambda_i \times q_i)$	$\alpha_0 = \log(3);\ \alpha_1 = 0.5;$ $\beta_0 = \log\left(\frac{20}{3}\right);\ \beta_1 = 0.5$
ZINB	Small	Small	$q_i \sim Bernoulli\ (\pi_i)$ $w_i \sim Neg\ Binom\ (\mu_i = \lambda_i \times q_i, n)$	$\alpha_0 = \log(3);\ \alpha_1 = 0.5;$ $\beta_0 = \log\left(\frac{4}{3}\right);\ \beta_1 = 0.5;\ n = 1$
	Small	Large	$q_i \sim Bernoulli\ (\pi_i)$ $w_i \sim Neg\ Binom\ (\mu_i = \lambda_i \times q_i, n)$	$\alpha_0 = \log(3);\ \alpha_1 = 0.5;$ $\beta_0 = \log\left(\frac{4}{3}\right);\ \beta_1 = 0.5;\ n = 0.1$
	Large	Small	$q_i \sim Bernoulli\ (\pi_i)$ $w_i \sim Neg\ Binom\ (\mu_i = \lambda_i \times q_i, n)$	$\alpha_0 = \log(3);\ \alpha_1 = 0.5;$ $\beta_0 = \log\left(\frac{20}{3}\right);\ \beta_1 = 0.5;\ n = 1$
	Large	Large	$q_i \sim Bernoulli\ (\pi_i)$ $w_i \sim Neg\ Binom\ (\mu_i = \lambda_i \times q_i, n)$	$\alpha_0 = \log(3);\ \alpha_1 = 0.5;$ $\beta_0 = \log\left(\frac{20}{3}\right);\ \beta_1 = 0.5;\ n = 0.1$

Table 1. Assumptions used to simulated data for each model. In these equations, q_i is a latent binary variable, ω_i is the response count variable, x_i is an explanatory variable, $\lambda_i = \exp(\beta_0 + \beta_1 x_i)$, and $\pi_i = \frac{\exp(\alpha_0 + \alpha_1 x_i)}{1 + \exp(\alpha_0 + \alpha_1 x_i)}$. For the negative binomial distribution, $E[w_i] = \mu_i$ and $Var[w_i] = \mu_i + \frac{\mu_i^2}{n}$.

$$\min\left\{1, \frac{p(M_k^*|\boldsymbol{z}, \ldots)p(M_k^*)}{p(M_k|\boldsymbol{z}, \ldots)p(M_k)}\right\} = \min\left\{1, \frac{exp\left(-\frac{1}{2}[-\boldsymbol{\mu}_{\boldsymbol{k}^*}^T \boldsymbol{\Sigma}_{\boldsymbol{k}^*}^{-1} \boldsymbol{\mu}_{\boldsymbol{k}^*} + \boldsymbol{z}^T \boldsymbol{z}]\right)|\boldsymbol{\Sigma}_{\boldsymbol{k}^*}|^{\frac{1}{2}}}{exp\left(-\frac{1}{2}[-\boldsymbol{\mu}_{\boldsymbol{k}}^T \boldsymbol{\Sigma}_{\boldsymbol{k}}^{-1} \boldsymbol{\mu}_{\boldsymbol{k}} + \boldsymbol{z}^T \boldsymbol{z}]\right)|\boldsymbol{\Sigma}_{\boldsymbol{k}}|^{\frac{1}{2}}} \times R\right\}$$

where R is typically equal to 1 and M_k^* and M_k are the proposed and current models, respectively. This model selection procedure is done as part of the MCMC algorithm. A detailed description of this model formulation and associated algorithms can be found in Denison *et al.*[19] and Zhao *et al.*[20]. We provide the derivation of the full conditional distributions used to create our Gibbs sampler in S2 Appendix. The implementation of our algorithm was done in R[21]. All the MN and MN-MS model results reported in this article are based on running our MCMC algorithm for 50,000 iterations and discarding the first half as burn in. The associated code, together with a short tutorial reproducing some of our results for the simulated data, is provided in S3 Appendix. Next, we describe our case study and the three sets of simulations that were performed to compare the performance of the proposed models in fitting data from different discrete distributions, identifying important predictor variables, and modeling nonlinear mean response functions.

Simulation set 1: fitting different discrete distributions. To assess how well our ordinal regression model fits data from a variety of conditional distributions, with and without over-dispersion and/or zero-inflation, we generated 10 simulated datasets for each regression model (from a total of 12 distinct models; see distributional assumptions in Table 1). Each dataset contained 500 observations and the covariate x corresponded to 500 values equally spaced between −2 to 2. Parameter values were chosen to explore a range of possible scenarios. For instance, we simulated data with small and large means ($E[w_i|x_i = 0] = 1$ and $E[w_i|x_i = 0] = 5$, respectively). In addition to small and large means, we experimented with different combinations of small and large variances (n = 1 and n = 1/10, respectively) for the NB and ZINB models. In relation to zero-inflation, we assumed that the proportion of zeroes arising from the Bernoulli mixture component was equal to 0.25 when the covariate x was equal to zero (i.e., $p(q_i = 0|x_i = 0) = \frac{1}{4}$).

We fit our multinomial model with a quadratic specification (i.e., $\beta_1 x_i + \beta_2 x_i^2$) and compare model fit to that of models using the correct distributional assumptions. Because all models were fit under a Bayesian framework, we assess and compare model fit among these models using the posterior distribution of the log-likelihood (LLK), summarized by the median and 95% credible intervals (CI). Two models are judged to fit the data equally well if the 95% CI's for their LLK overlap. If their 95% CI's do not overlap, then the model with the highest LLK is judged to be the best fitting model. The models with the correct distributional assumptions (as described in Table 1) were fit using JAGS[22]. When using JAGS, the number of iterations was set to 10,000 and increased if necessary until all parameters had converged, as assessed by the potential scale reduction factor $\hat{R}$. Values of $\hat{R}$ smaller than 1.1 were assumed to indicate successful convergence.

Simulation set 2: identifying relevant predictors. In our second set of simulations, we aim to examine if the multinomial model with model selection (MN-MS model) can adequately identify the few important

Species	Proportion of zeroes	Maximum number of mosquitoes caught in a 6 hour period
A. darlingi	0.70	109
A. nuneztovari	0.92	24
A. triannulatus	0.60	308
A. benarrochi	0.82	249
A. oswaldoi	0.71	124
A. rangeli	0.86	33

Table 2. Data on mosquito human biting rate is zero-inflated and over-dispersed.

Reg. model	Mean	Variances	MN model fits equally well or has better fit (proportion)
Poisson	Small	—	1.0
	Large	—	1.0
NB	Small	Small	0.9
	Small	Large	1.0
	Large	Small	1.0
	Large	Large	1.0
ZIP	Small	—	0.8
	Large	—	0.0
ZINB	Small	Small	0.9
	Small	Large	1.0
	Large	Small	1.0
	Large	Large	1.0

Table 3. The MN model fits well data generated from a diverse set of conditional distributions despite lack of information on the correct distribution. Numbers correspond to the proportion of datasets (based on 10 datasets) for which the MN model fitted the data equally well or had a better fit when compared to the true model with estimated parameters. Models were judged to fit the data equally well if their 95% credible intervals for the log-likelihood (our measure of goodness-of-fit) overlapped.

predictor variables among a large number of covariates. To this end, we generated data from a Poisson regression model with a large number of covariates:

$$y_i \sim Poisson(\exp(\boldsymbol{x_i}^T \boldsymbol{\beta}))$$

where the design vector $\boldsymbol{x_i}^T$ contains the intercept, 10 covariates and all pairwise interaction terms between these 10 covariates. In total, this model has $(1 + 10 + (10 \times 9/2)) = 56$ regression parameters in the vector $\boldsymbol{\beta}$. We simulate data by assuming that $\boldsymbol{\beta}$ is comprised of zeroes except for the intercept and a given number m (varying from 0 to 10) of randomly chosen elements of $\boldsymbol{\beta}$. These m non-zero elements in $\boldsymbol{\beta}$ were randomly set to 0.5 or to -0.5 and correspond to important predictor variables. We generated 10 datasets for each $m = 0, 1, 2, \ldots, 10$, resulting in a total of 110 datasets with 500 observations per dataset.

According to a recent review, the most common procedure used for model selection in ecological publications is to select covariates based on AIC[23], often within a forward, backward, or stepwise (i.e., combined forward and backward) approach. We compare the performance of this approach in identifying important predictors to that of the MN-MS model. To this end, we performed AIC model selection using the glm() and stepAIC() (from the MASS package) functions in R. The identified best model was subsequently fitted within a Bayesian framework. We compare the results from this best model to that of a Poisson model without any covariate selection and the MN-MS model. These latter models were also fitted within a Bayesian framework and we used the 95% credible intervals (CI) to determine if the method identified the zero and non-zero slope parameters correctly. More specifically, a non-zero coefficient was deemed correctly estimated if its 95% CI did not include zero and had the same sign as the true parameter. On the other hand, a zero coefficient was judged to be correctly estimated if the 95% CI overlapped with zero. Covariates that were excluded by the AIC model selection procedure were deemed to have a slope coefficient of zero.

Simulation set 3: modeling nonlinear response curves. In this set of simulations, we investigate whether the multinomial model with model selection (MN-MS model) can approximate well different non-linear mean response functions in the absence of information on the correct distribution. To this end, we randomly generated 10 datasets, each of which had 500 observations with 6 predictor variables. We assumed that only the first 3 predictor variables influenced the mean response function, based on the following expression:

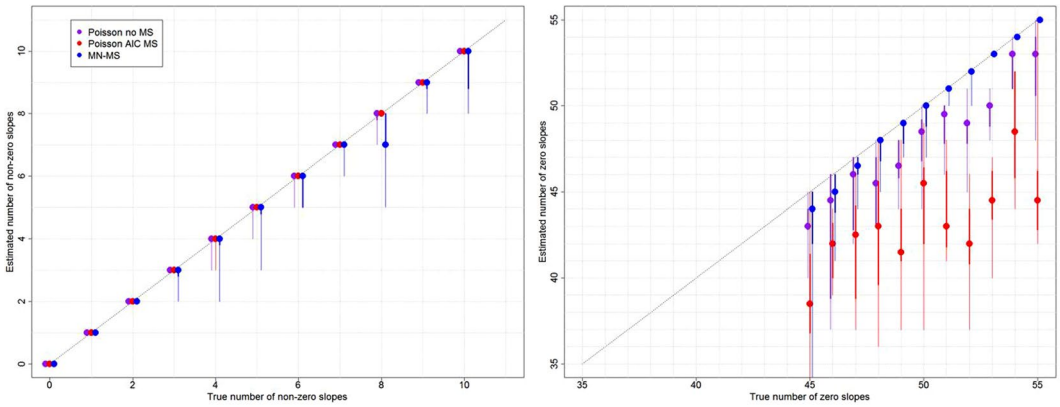

Figure 1. The MN-MS model performs slightly worse than the Poisson regression models in identifying the true non-zero slopes (left panel) but performs substantially better in identifying the true zero slopes (right panel). Results for the Poisson model without model selection (Poisson no MS; purple), with AIC model selection (Poisson AIC MS; red), and the MN model with model selection (MN-MS; blue) are displayed. A 1:1 line was added for reference (dashed diagonal line), where results closer to this line indicate better performance. Circles represent the median, thick lines represent the 20–80% range, while thin lines represent the full range (minimum to maximum) based on 10 datasets. Left panel: The x-axis displays the true number of non-zero slopes used to generate the data while the y-axis reveals how many of these slopes were correctly identified to be non-zero and were estimated with the correct sign. Right panel: The x-axis displays the true number of zero slopes used to generate the data while the y-axis reveals how many of these slopes were correctly identified to be zero.

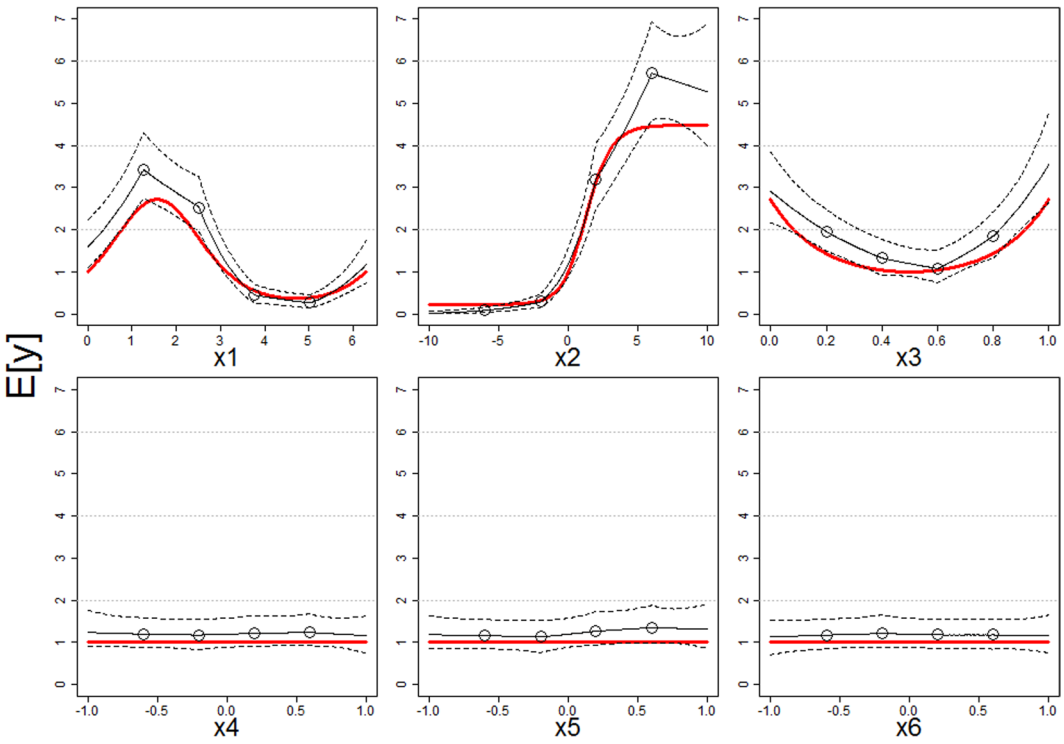

Figure 2. The MN-MS model estimates well non-linear effects of covariates x_1, x_2, and x_3 (top panels) and the absence of effects associated with covariates x_4, x_5, and x_6 (bottom panels). True mean response functions are depicted with red line while the estimated relationship are shown with black lines (continuous and dashed lines are the median and point-wise 95% credible intervals, respectively). Circles show the knot locations for each covariate, *a priori* set to 0.2, 0.4, 0.6, and 0.8 quantile of the corresponding covariate. The displayed response curves are based on one of the 10 simulated datasets and were created by only varying the focal covariate while the other covariates were set to their mean values.

Species	Model fit					
	Poisson	NB	ZINB	ZIP	MN	MN-MS
A. darlingi	−4756	−1283	−1245	−2682	**−1244**	−1245
A. nuneztovari	−407	−318	−311	−316	**−310**	−314
A. triannulatus	−6922	−1616	−1591	−3905	**−1551**	−1552
A. benarrochi	−4131	−775	−770	−1406	**−748**	**−748**
A. oswaldoi	−2533	−1057	−1035	−1682	**−1032**	−1033
A. rangeli	−1147	−587	**−539**	−670	−563	−568

Table 4. The MN and MN-MS generally fit mosquito data better than other competing regression models. The median of the log-likelihood (model fit) is provided for each combination of model and mosquito species. The best model for each species is emphasized in bold. "ZI" stands for zero-inflation.

Species	Predictive performance								
	MN model				MN-MS model				
	Poisson	NB	ZINB	ZIP	Poisson	NB	ZINB	ZIP	MN
A. darlingi	0.86	0.79	0.79	0.64	0.86	0.79	0.79	0.64	0.36
A. nuneztovari	0.86	0.86	0.93	1.00	0.93	0.79	0.93	1.00	0.57
A. triannulatus	0.79	0.71	0.79	0.64	0.79	0.71	0.79	0.64	0.29
A. benarrochi	0.79	0.79	0.86	0.71	0.79	0.86	0.93	0.71	0.79
A. oswaldoi	0.79	0.86	0.71	0.71	0.79	0.79	0.79	0.79	0.64
A. rangeli	0.79	0.93	0.93	0.79	0.71	0.93	0.93	0.79	0.79

Table 5. The MN and MN-MS generally predict out-of-sample mosquito data better than other competing regression models. Numbers indicate the proportion of cross-validation folds (based on 14 folds) in which the MN and MN-MS models had lower MSE scores when compared to each alternative model and for each mosquito species. "ZI" stands for zero-inflation. The last column on the right shows the proportion of cross-validation folds in which the MN-MS model had lower MSE score relative to the MN model.

$$y_i \sim NegBinom(\mu_i = \exp(\sin(x_{1i}) + 3\left[\frac{\exp(x_{2i})}{1+\exp(x_{2i})} - 0.5\right] + 1 - 4x_{3i} + 4x_{3i}^2),\ n = 20)$$

where $E[y_i] = \mu_i$ and $Var[y_i] = \mu_i + \frac{\mu_i^2}{n}$. To approximate this mean response function, we rely on linear splines as our bases functions with four potential inflection points (i.e., knots) for each covariate, *a priori* set to 0.2, 0.4, 0.6, and 0.8 quantiles of the corresponding covariate.

Case study: mosquito data from the Peruvian Amazon. Data on anopheline mosquitoes were collected along the Iquitos-Nauta road, in the Peruvian Amazon, between 2000 and 2001. The original study's goal was to determine how different land-use land-cover (LULC) classes influenced malaria risk. To this end, Vittor *et al.*[9] focused solely on *A. darlingi*, the mosquito species widely regarded as the most important malaria vector in the region, and performed a multinomial regression where biting rates were *a priori* classified as low, medium, or high. Overall, 56 sites (grouped into 14 spatial clusters) were sampled 15 to 16 times between 2000 and 2001. These data are fully described in Vittor *et al.*[9] and a review of how malaria is related to LULC in the Amazon can be found in Tucker-Lima *et al.*[24]. Here we revisit this study but now using the proposed statistical method and using data on the six most common anopheline species in this dataset (i.e., *A. darlingi*, *A. nuneztovari*, *A. triannulatus*, *A. benarrochi*, *A. oswaldoi*, and *A. rangeli*), all of which are known to be able to transmit malaria in the region. We note that adequately modeling these data is challenging because the data are zero-inflated *and* over-dispersed (Table 2).

The covariates in our model consist of precipitation, proportion of forest cover, and proportion of exposed soil/urban area. Precipitation data for each location and month were extracted from the Tropical Rainfall Measuring Mission (TRMM) product 3B43, which provides monthly rainfall estimates with a 0.25×0.25 degree spatial resolution[25]. LULC classification was based on a supervised random forest algorithm applied to a 2000 Landsat image with a 30×30 meter pixel, from which we calculated the proportion of terra-firme forest pixels and exposed soil/urban pixels within a buffer of 500 m around each point. All covariates were standardized to have a mean of zero and variance of one. Similar to the simulation study described above, we model potentially non-linear relationships through the use of linear spline bases, where knots were placed at 0.2, 0.4, 0.6, and 0.8 percentiles of each covariate.

We separately fit data from each of these six mosquito species using the MN and the MN-MS model. To determine how well these models fit and predict these data, we compare the log-likelihood (our measure of model fit) and out-of-sample predictive skill to that of a set of alternative models. As recommend by Roberts, *et al.*[26], because we were primarily interested in spatial covariates (i.e., land use/land cover) and spatial predictions,

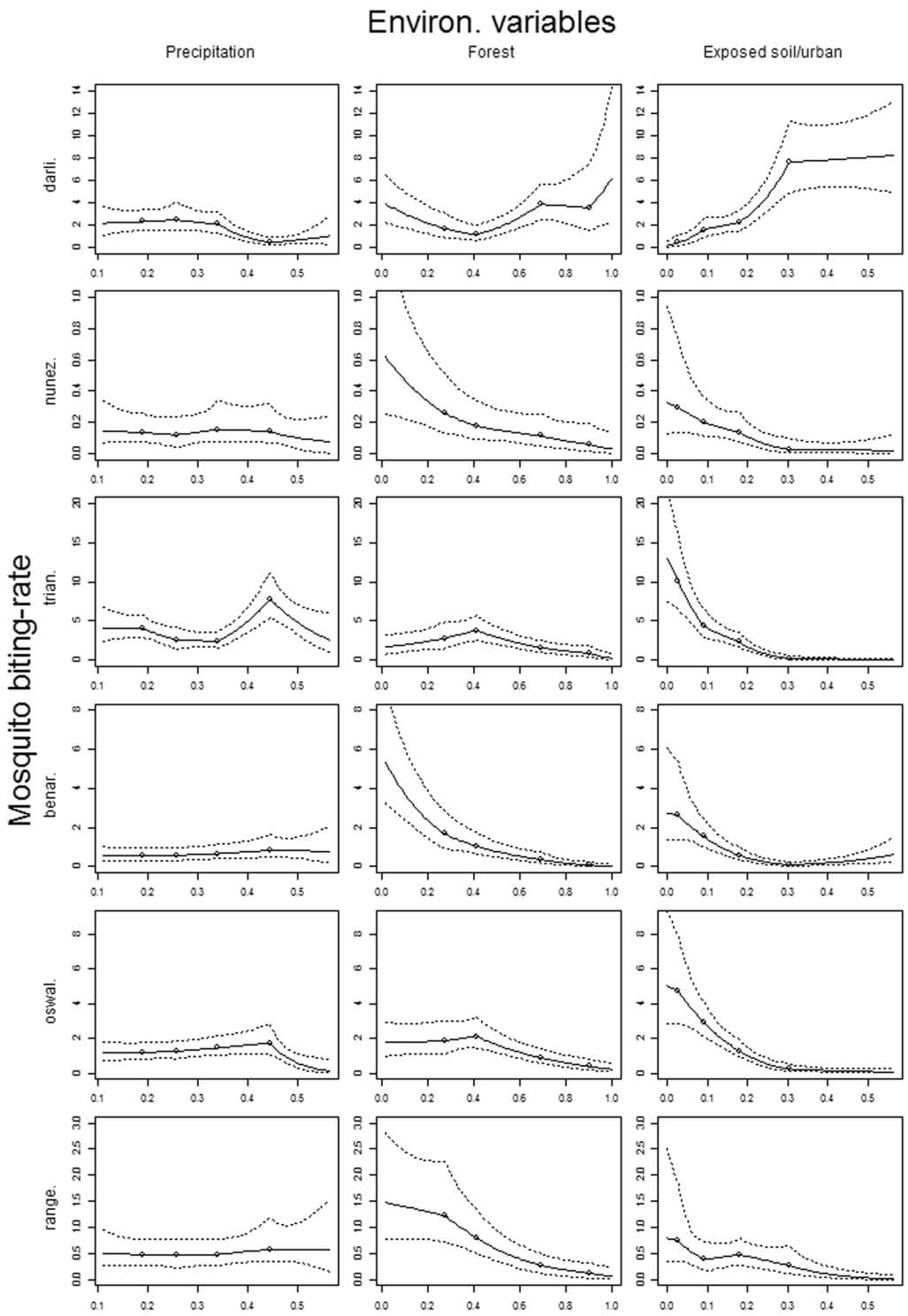

Figure 3. Statistical associations between mosquito biting-rates and environmental covariates based on the MN-MS model. Modeling results from individual mosquito species are shown separately in each row (*A. darlingi* = darli., *A. nuneztovari* = nunez., *A. triannulatus* = trian., *A. benarrochi* = benar., *A. oswaldoi* = oswal., and *A. rangeli* = range.). Continuous and dashed lines represent the median and the 95% credible intervals, respectively. Circles show potential inflection points (i.e., knot locations), *a priori* set to 0.2, 0.4, 0.6 and 0.8 quantiles of the covariate. Left to right panels show the inferred associations between mosquito biting-rate (number of mosquitoes caught per 6-hour period) and precipitation (mm/hr), proportion of forest pixels, and proportion of exposed soil/urban pixels, respectively. Proportion of pixels was calculated within a 500 m buffer of each observation location.

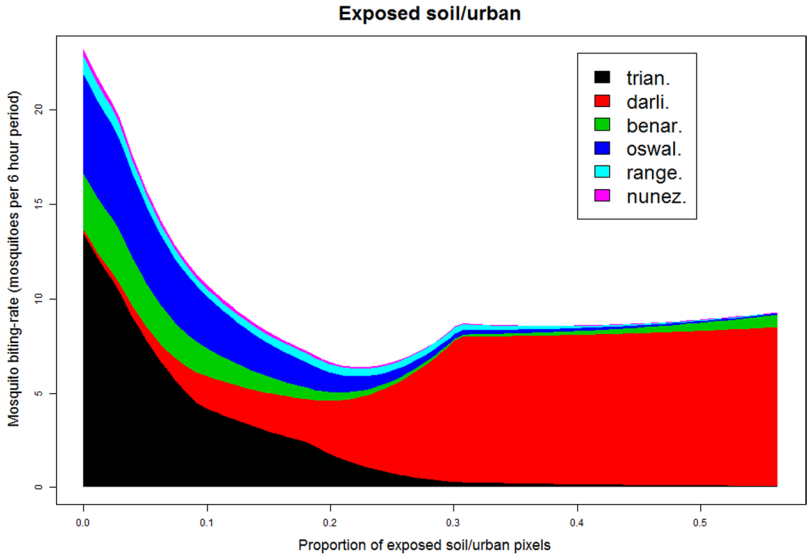

Figure 4. Large shift in species composition in mean mosquito biting-rates associated with changes in the proportion of exposed soil/urban area. Modeling results from individual mosquito species are shown in different colors (*A. darlingi* = darli., *A. nuneztovari* = nunez., *A. triannulatus* = trian., *A. benarrochi* = benar., *A. oswaldoi* = oswal., and *A. rangeli* = range.) as a function of the proportion of exposed soil/urban pixel. Proportion of pixels was calculated within a 500 m buffer of each observation location.

out-of-sample predictive skill was determined through a spatial validation procedure. In this procedure, one spatial cluster of sites was removed for prediction purposes and the rest of the data were used to train the model in each of the 14 validation folds. Out-of-sample predictive performance was evaluated based on mean squared error (MSE). The alternative models were the Poisson, Negative-Binomial (NB), zero-inflated Negative Binomial (ZINB), and zero-inflated Poisson (ZIP) regression models, fitted with JAGS. All models in this comparison had the same set of covariates and spline terms.

Results

Simulation set 1: fitting different distributions. The MN model adequately accounted for over-dispersion and zero-inflation, having similar (based on overlapping 95% credible intervals) or greater goodness-of-fit when compared to that of the true models with estimated parameters (Table 3). The MN model only failed to fit well data originated from the ZIP model with large mean, with a worse fit in all ten simulated datasets. In this case, a comparison of the theoretical and the estimated distributions suggests that the MN model has difficulty representing conditional distributions that are approximately unimodal for small values of the covariate as well as strongly bimodal for large values of the covariate, with little probability mass for numbers in between both modes. Overall, these results highlight the flexibility of the MN model in adequately representing data generated from a wide range of distributions (over-dispersed and/or zero-inflated).

Simulation set 2: identifying relevant predictors. Despite the MN-MS model performing slightly worse in identifying the relevant covariates than the Poisson regression model using all the covariates ("Poisson no MS") and the AIC model selection procedure ("Poisson AIC MS") (left panel in Fig. 1), the MN-MS model performed substantially better than the Poisson models in identifying the slopes that were equal to zero (right panel in Fig. 1). Indeed, the Poisson model using all the covariates ("Poisson no MS") often times identified statistically significant slopes even when the corresponding covariates were independent of the response variable. Surprisingly, the AIC model selection method ("Poisson AIC MS") was the worse approach in this respect, incorrectly identifying a relatively large proportion of "important" covariates. These results are striking because the Poisson models have the advantage of using the correct distributional assumption and yet the MN-MS model performs better overall.

Simulation set 3: modeling nonlinear response curves. We find that the MN-MS model can reliably estimate different non-linear relationships between covariates (e.g., sinusoidal, logistic, and quadratic functions for covariates x_1, x_2, and x_3, respectively; top panels in Fig. 2) and the mean response using linear splines. Importantly, this model can also estimate well the absence of effects (e.g., covariates x_4, x_5, and x_6; bottom panels in Fig. 2). These results suggest that the lack of information on the true distribution and the relationship between covariates and the mean response does not jeopardize the ability of the MN-MS model to infer these non-linear relationships. These results are important because researchers seldom have prior knowledge on the most appropriate distribution and mean response function to use to model their count data.

Case study: mosquito data from the Peruvian Amazon. We find that the MN model was the best fitting model for five of the mosquito species and the second best model for the sixth remaining species (Table 4).

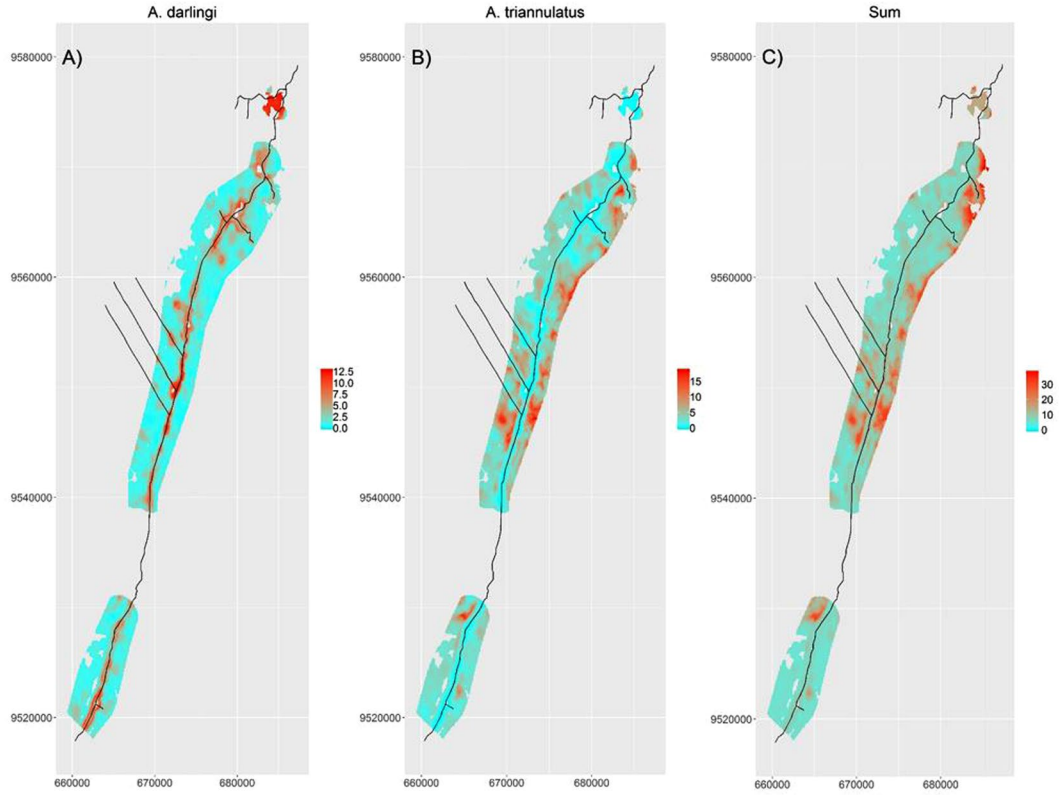

Figure 5. Spatial prediction of mean mosquito biting-rates for the two most common anopheline species and overall biting rate. From left to right, each panel shows the spatial prediction of mean mosquito biting-rate for *A. darlingi* (darli.), *A. triannulatus* (trian.), and the sum of the predicted mean biting-rate of the six anopheline mosquito species (Sum). Axes depict UTM coordinates in meters. The road network is depicted with black lines and covariate extrapolation is avoided by removing all areas for which covariate values were outside the range used to fit the model. Spatial extrapolation is avoided by restricting spatial prediction to within 2.5 km of sampled sites.

The MN-MS model closely followed the model fit metrics of the MN model, being the best model for one mosquito species and the second best model for three other species. Overall, these results suggest that the MN and MN-MS models generally have superior fit to these zero-inflated and over-dispersed mosquito data when compared to other more standard regression models.

The model fit statistics reported in Table 4 can be misleading for the identification of the best model if data are being over-fitted. Because models that over-fit the data have substantially worse out-of-sample predictive performance, we test if these models are over-fitting by comparing the models in Table 4 according to their out-of-sample predictive skill using a spatial block cross-validation procedure. This procedure reveals that both of the proposed models (MN and MN-MS models) tend to consistently have higher out-of-sample predictive skill (i.e., lower MSE values) than the other alternative models across all 6 mosquito species (Table 5). Interestingly, as shown in the right most column of Table 5, the MN-MS model tends to have a better predictive performance when compared to the MN model, with lower MSE for 4 mosquito species.

Using the MN-MS model, we find that the most important factors driving mosquito biting-rates were proportion of forest and exposed soil/urban area whereas precipitation had a comparatively minor role (Fig. 3). In general, we find a negative association between exposed soil/urban area and the biting-rate of all the mosquito species, except for *A. darlingi* which clearly is more common in more heavily disturbed areas (right panels in Fig. 3). Interestingly, three mosquito species (i.e., *A. nuneztovari*, *A. benarrochi*, and *A. rangeli*) also have higher biting-rates in areas with a lower proportion of forest (middle panels in Fig. 3), suggesting that these species thrive in areas that have some vegetation cover but that are not too pristine, such as secondary forest and agricultural lands. The use of linear splines allowed for the detection of several non-linear relationships in the mosquito data. For instance, Fig. 3 reveals that mosquito biting-rates for *A. rangeli* and *A. darlingi* tend to asymptote at intermediate levels of forest and exposed soil/urban area, respectively. Similarly, *A. triannulatus* and *A. oswaldoi* are only strongly influenced by precipitation within a specific range of this covariate.

When results from individual species are put together, they reveal that areas with a lower proportion of exposed soil/urban pixels on average have a substantially higher overall mosquito biting-rate (Fig. 4). Interestingly, there is a pronounced shift in mosquito species composition as the proportion of exposed soil/urban area increases, with *A. darlingi* mosquitoes dominating areas with intermediate or high proportion of exposed soil/urban area. As expected, spatial predictions of mean mosquito-biting rate for *A. darlingi* reveals extremely high biting rates close to the primary road, reiterating the strong association of *A. darlingi* with highly

anthropized sites (Fig. 5A). On the other hand, *A. triannulatus* (the other most common mosquito species in our sample) had a substantially different spatial pattern, being virtually absent from the immediate vicinity of the primary road (Fig. 5B), similar to the spatial pattern that emerges when the predicted mean biting rate for all 6 mosquito species are summed (Fig. 5C).

Discussion

Count data are ubiquitous in multiple fields but these data are often zero-inflated and/or over-dispersed. There are several models that can be used to make inference based on data with these characteristics but determining the best one is challenging and often requires one to *a priori* choose a particular distribution. In this article, we have proposed a new statistical model that relies on a multinomial distribution to fit data from a wide range of different discrete distributions and automatically perform model selection. While ordinal regression models have a long tradition in statistics[18,27], its use to flexibly model count data (rather than ordinal data) and perform model selection is, to our knowledge, a novel idea. We illustrate the features of our model using extensive simulations and apply this model to a case study on environmental drivers of malaria risk.

It is clear that the MN model can fit data from a wide range of conditional distributions, as evidenced by our simulation study. These simulation findings, together with one of the best model fit and out-of-sample predictive skill when applied to the mosquito data, suggest that the MN and MN/MN-MS models might be good default options for drawing inference from count data. While the data generated from the ZIP model with large mean was not well fit, had we chosen the wrong model for these data (i.e., a NB regression model, as suggested by[1]), the fit to these data would be substantially worse than that for the MN model (results not shown). Additional research will be needed to more precisely determine the conditions under which the MN model is likely to fail to fit well and how prevalent these conditions are.

Despite having no prior knowledge of the underlying distribution of the data, the MN-MS model performed very well in variable selection. While the MN-MS model was slightly worse in identifying true explanatory variables, this was greatly outweighed by its superiority in eliminating false predictors, resulting in overall better inference when compared to using a simple Poisson regression with or without AIC model selection. This improved performance is supported by other studies that have compared Bayesian model averaging with simple and stepwise regression methods[20,28]. Finally, our simulation results suggest that the adopted linear spline approach was able to capture a wide range of non-linear patterns. We chose linear splines because they are simple and straight-forward to implement but there is a wide-range of more flexible spline functions that could have been used (e.g., cubic splines, b-splines, and thin-plate splines)[29]. Regardless of the specific type, all spline approaches entail the inclusion of numerous additional "covariates" (i.e., basis functions) into the design matrix, a setting in which our model selection procedure is likely to be particularly effective (e.g.[30]).

In relation to our case study, we build on the original work of Vittor *et al.*[9] in two important aspects. First, we examine multiple malaria vector species rather than just *A. darlingi*. This is important because, despite *A. darlingi* being widely acknowledged to be the main malaria vector in the Amazon region[31], several other anopheline species have been shown to be competent vectors and to be locally important for malaria transmission[32–40]. The second aspect that was improved refers to the statistical modeling approach. Vittor *et al.*[9] relied on a multinomial regression model where biting rates were classified as "low" (0–0.09/hr), "medium" (0.1–0.9/hr) and "high" (1.0–3.8/hr). We have improved on this modeling approach by avoiding the arbitrariness associated with data discretization, and the resulting loss of information, and by allowing for non-linear associations.

Our results suggest that one might arrive at very different conclusions regarding how land-use/land cover (LULC) classes are associated with malaria risk depending on which anopheline species is analyzed. Unlike the other mosquito species, *A. darlingi* seem to thrive in highly anthropized areas, greatly corroborating earlier published results[9,32,41]. Indeed, our model predicts that biting rate for this species concentrates close to roads, particularly in areas with a high proportion of exposed soil/urban area. However, the addition of other mosquito species reveals a different picture in that over-all biting rate is actually higher in areas with lower proportion of exposed soil/urban area. This is partly a result of significant changes in species composition along the urbanity gradient, where *A. triannulatus* dominates areas with less exposed soil/urban area whereas *A. darlingi* is the dominant species at the other side of the spectrum.

We believe that the proposed method will find wide use in natural sciences because it can flexibly fit and predict data with or without zero-inflation and/or over-dispersion while simultaneously identifying the most relevant explanatory variables.

References

1. Warton, D. I. Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. *Environmentrics* **16**, 275–289 (2005).
2. ver Hoef, J. M. & Boveng, P. L. Quasi-Poisson vs. Negative Binomial regression: how should we model overdispersed count data? *Ecology* **88**, 2766–2772 (2007).
3. Potts, J. M. & Elith, J. Comparing species abundance models. *Ecol Modell* **199**, 153–163 (2006).
4. Welsh, A. H., Cunningham, R. B., Donnelly, C. F. & Lindenmayer, D. B. Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecol Modell* **88**, 297–308 (1996).
5. Welsh, A. H., Cunningham, R. B. & Chambers, R. L. Methodology for estimating the abundance of rare animals: seabird nesting on North East Herald Cay. *Biometrics* **56**, 22–30 (2000).
6. White, G. C. & Bennetts, R. E. Analysis of frequency count data using the Negative Binomial distribution. *Ecology* **77**, 2549–2557 (1996).
7. Linden, A. & Mantyniemi, S. Using the negative binomial distribution to model overdispersion in ecological count data. *Ecology* **92**, 1414–1421 (2011).
8. Martin, T. G. *et al.* Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecol. Lett* **8**, 1235–1246 (2005).

9. Vittor, A. *et al*. The effect of deforestation on the human-biting rate of Anopheles darlingi, the primary vector of falciparum malaria in the Peruvian Amazon. *Am J Trop Med Hyg* **74**, 3–11 (2006).
10. Nedelman, J. A negative binomial model for sampling mosquitoes in a malaria survey. *Biometrics* **39**, 1009–1020 (1983).
11. Alexander, N., Moyeed, R. & Stander, J. Spatial modelling of individual-level parasite counts using the negative binomial distribution. *Biostatistics* **1**, 453–463 (2000).
12. Joe, H. & Zhu, R. Generalized Poisson distribution: the property of mixture of Poisson and comparison with Negative Binomial distribution. *Biometrical Journal* **2**, 219–229 (2005).
13. Lord, D., Washington, S. P. & Ivan, J. N. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis and Prevention* **37**, 35–46 (2005).
14. Sileshi, G., Hailu, G. & Nyadzi, G. I. Traditional occupancy-abundance models are inadequate for zero-inflated ecological count data. *Ecol Modell* **220**, 1764–1775 (2009).
15. Shaw, D. J. & Dobson, A. P. Patterns of macroparasite abundance and aggregation in wildlife populations: a quantitative review. *Parasitology* **111**, S111–S133 (1995).
16. Lambert, D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34**, 1–14 (1992).
17. Ghosh, S., Gelfand, A. E., Zhu, K. & Clark, J. The k-ZIG: flexible modeling for zero-inflated counts. *Biometrics* **68**, 878–885 (2012).
18. Agresti, A. *Categorical data analysis*. (John Wiley & Sons, 2003).
19. Denison, D. G. T., Holmes, C. C., Mallick, B. K. & Smith, A. F. M. *Bayesian methods for nonlinear classification and regression*. (Wiley, 2002).
20. Zhao, K., Valle, D., Popescu, S., Zhang, X. & Mallick, B. Hyperspectral remote sensing of plant biochemistry using Bayesian model averaging with variable and band selection. *Remote Sens Environ* **132**, 102–119 (2013).
21. R Core Team. R: A language and environment for statistical computing. (R Foundation for Statistical Computing, Vienna, Austria, 2013).
22. Plummer, M. *JAGS: A program for analysis of Bayesian graphical models using GIbbs sampling*. (2003).
23. Aho, K., Derryberry, D. & Peterson, T. Model selection for ecologists: the worldviews of AIC and BIC. *Ecology* **95**, 631–636 (2014).
24. Tucker-Lima, J., Vittor, A. Y., Rifai, S. & Valle, D. Does deforestation promote or inhibit malaria transmission in the Amazon? A systematic literature review and critical appraisal of current evidence. *Philos Trans R Soc Lond B Biol Sci* (2017).
25. Tropical Rainfall Measuring Mission (TRMM). *TRMM (TMPA/3B43) Rainfall Estimate L3 1 month 0.25 degree × 0.25 degree V7*, https://disc.gsfc.nasa.gov/datasets/TRMM_3B43_V7/summary (Date of access) (2011).
26. Roberts, D. R. *et al*. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40**, 913–929 (2017).
27. McCullagh, P. Regression models for ordinal data. *J R Stat Soc Series B* **42**, 109–142 (1980).
28. Genell, A., Nemes, S., Steineck, G. & Dickman, P. W. Model selection in medical research: a simulation study comparing Bayesian model averaging and stepwise regression. *BMC Medical Research Methodology* **10** (2010).
29. Wood, S. N. *Generalized Additive Models: an introduction with R*. (CRC Press, 2017).
30. Millar, J. *et al*. Detecting risk factors for residual malaria using Bayesian Model Averaging. *Malar J* **17** (2018).
31. Deane, L. M., Causey, O. R. & Deane, M. P. Notas sobre a distribuicao e a biologia dos anofelinos das regioes Nordestina e Amazonica do Brasil. *Revista do Servico Especial de Saude Publica* **4**, 826–965 (1948).
32. Tadei, W. P. & Dutary Thatcher, B. Malaria vectors in the Brazilian amazon: Anopheles of the subgenus Nyssorhynchus. *Rev Inst Med Trop Sao Paulo* **42**, 87–94 (2000).
33. Girod, R. *et al*. Unravelling the relationships between Anopheles darlingi (Diptera: Culicidae) densities, environmental factors and malaria incidence: understanding the variable patterns of malarial transmission in French Guiana (South America). *Ann Trop Med Parasitol* **105**, 107–122, https://doi.org/10.1179/136485911X12899838683322 (2011).
34. Conn, J. *et al*. Emergence of a new neotropical malaria vector facilitated by human migration and changes in land use. *Am J Trop Med Hyg* **66**, 18–22 (2002).
35. Ferreira, R. M. D. A., da Cunha, A. C. & Souto, R. N. P. Distribuicao mensal e atividade noraria de Anopheles (Diptera: Culicidae) em uma area rural da Amazonia Oriental. *Biota Amazonia* **3**, 64–75 (2013).
36. Galardo, A. K. *et al*. Malaria vector incrimination in three rural riverine villages in the Brazilian Amazon. *Am J Trop Med Hyg* **76**, 461–469 (2007).
37. da Silva-Vasconcelos, A. *et al*. Biting indices, host-seeking activity and natural infection rates of anopheline species in Boa Vista, Roraima, Brazil from 1996 to 1998. *Mem Inst Oswaldo Cruz* **97**, 151–161 (2002).
38. Póvoa, M., Wirtz, R., Lacerda, R., Miles, M. & Warhurst, D. Malaria vectors in the municipality of Serra do Navio, State of Amapá, Amazon Region, Brazil. *Mem Inst Oswaldo Cruz* **96**, 179–184 (2001).
39. Schoeler, G. B., Flores-Mendoza, C., Fernandez, R., Davila, J. R. & Zyzak, M. Geographical distribution of Anopheles darlingi in the Amazon Basin region of Peru. *Journal of the American Mosquito Control Association* **19**, 286–296 (2003).
40. Lounibos, P. L. & Conn, J. E. Malaria vector heterogeneity in South America. *Am Entomol* **46**, 238–249 (2000).
41. Turell, M. J. *et al*. Seasonal distribution, biology, and human attraction patterns of mosquitoes (Diptera: Culicidae) in a rural village and adjacent forested site near Iquitos, Peru. *J Med Entomol* **45**, 1165–1172 (2008).

Acknowledgements

We thank Sami Rifai for performing the LULC classification based on the Landsat image and Amy Vittor for providing the mosquito data from Peru. This manuscript has benefited substantially from feedback from Guillaume Blanchet. D.V. and Q.Z. were partly supported by the US National Science Foundation award 1458034. G.Z.P. was supported by the Sao Paulo Research Foundation (FAPESP 2014/09774-1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

D.V. wrote the first draft and performed all the simulations and data analysis. K.B.T., G.Z.L. and Q.Z. provided numerous ideas and edited the manuscript multiple times. All authors have reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-39377-x.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.