Annals of Data Science (2024) 11(3):803–817
https://doi.org/10.1007/s40745-023-00464-6

Count Regression and Machine Learning Techniques for Zero-Inflated Overdispersed Count Data: Application to Ecological Data

Bonelwa Sidumo[1] · Energy Sonono[1] · Isaac Takaidza[1]

Received: 22 October 2021 / Revised: 17 January 2023 / Accepted: 21 March 2023 /
Published online: 13 April 2023

Abstract

The aim of this study is to investigate the overdispersion problem that is rampant in ecological count data. In order to explore this problem, we consider the most commonly used count regression models: the Poisson, the negative binomial, the zero-inflated Poisson and the zero-inflated negative binomial models. The performance of these count regression models is compared with the four proposed machine learning (ML) regression techniques: random forests, support vector machines, $k-$nearest neighbors and artificial neural networks. The mean absolute error was used to compare the performance of count regression models and ML regression models. The results suggest that ML regression models perform better compared to count regression models. The performance shown by ML regression techniques is a motivation for further research in improving methods and applications in ecological studies.

Keywords Count data · Ecology · Machine learning · Overdispersion · Zero-inflation

Energy Sonono and Isaac Takaidza have contributed equally to this work.

✉ Bonelwa Sidumo
Bonelwa.Sidumo@nwu.ac.za

Energy Sonono
Energy.Sonono@nwu.ac.za

Isaac Takaidza
Isaac.Takaidza@nwu.ac.za

[1] School of Mathematical and Statistical Sciences, North-West University, Hendrick Van Eck Blvd, Vanderbijlpark 1911, Gauteng, South Africa

1 Introduction

The aim of this article is to investigate the problem of overdispersion in ecological count data. Overdispersion is an existing and recurring problem that needs attention when dealing with ecological count data. Ignoring overdispersion will cause difficulties in analysis and the decision-making procedures of ecological studies. We approach the problem of overdispersion by using machine learning (ML) regression techniques. To the best of our knowledge an approach to overdispersion in ecological studies using ML techniques has not been extensively researched thus far.

Bolker et al. [1] define overdispersion as the occurrence of more variance in the data than predicted by a statistical model owing to missing observations. The reasons for the existence of missing observations in ecological count data may be due to structural errors (for instance, a bird or fish is not present because the habitat is not suitable), observer error (species are present but cannot be detected) and design error (poor experimental design or sampling surveys are thought to be the reason) [2]. The literature discusses the source of zeros in ecological count data and defines them as either 'true zero counts' or 'false zero counts' [3, 4]. False zero counts occur when species are present at a site during the survey period, but the observer fails to detect them and true zero counts occur when species do not occur at a site because of the ecological process, that is, habitat unsuitability. This study focuses only on false zero counts. In ecology, zero counts do not necessarily mean that there are no species detected during the sampling survey [5]; rather it means that there were no species at that particular sampling time (false zeros). The absence of species results in excess number of zeros termed zero-inflation. The presence of zero-inflation in this study is owing to observer error.

This study will provide an overview of various count regression models: the Poisson, the negative binomial (NB), the zero-inflated Poisson (ZIP) and the zero-inflated negative binomial model (ZINB). The Poisson regression model has been widely used to analyse count data under the assumption of equidispersion, that is, the mean of the response variable is equal to the variance of the response variable [6, 7]. However, as much as this is a naturally occurring property of the Poisson regression model, it is not always true in real life ecological count data as counts may exhibit excess variability. The fact that equidispersion is rarely found in real data has resulted in the development of more general count models which do not assume equidispersion [8]. The NB regression model has been used as an alternative model to the Poisson regression model (see [9, 10]). NB regression models are more flexible than Poisson regression models even though they do not provide exact predictions in certain situations [11]. The next alternative used for modeling count data with excess number of zeros is the ZIP model, which has been applied in many areas of research such as insurance claims [12, 13], education [11, 14], healthcare [15–17] animal ecology [18] and transport [19]. ZIP was found to be inappropriate for data that are both zero-inflated and overdispersed [20]. Furthermore, Minami et al. [20] and Rose et al. [21] propose the ZINB model as another alternative to handle overdispersion. ZINB has been shown to be appropriate to some ecological situations even though the issue of overdispersion still remains [20].

From the empirical evidence provided above, the proposed methods still pose challenges in dealing with overdispersion. There are still numerous false zeros that are being observed or not accounted for. In other words, there is still room to improve the reduction of overdispersion in count data, which this study proposes via a possible new method. Most published studies in ecology sometimes fail to report on overdispersion in respect of their best fitting models [22]. Failing to account for overdispersion can lead to incorrect inferences [23]. There is also limited literature in ecological studies about how overdispersion affects results as researchers would identify predictors as having biologically meaningful effects when, in fact they do not [6].

As a result of the limitations in some statistical methods (for example, Poisson and NB) and the diversity of data, new techniques of data science have been developed. Data science has become an important and growing field as the Internet of Things (IoT) expands worldwide [24]. Encompassing several techniques such as data mining and machine learning, data science solves relevant problems and predicts results by taking into account data quality [25]. The data science techniques have been applied in various research fields such as healthcare [7] and education [26] and these techniques are combined to consolidate statistical analyses.

This study proposes machine learning (ML) regression techniques; random forests (RF), support vector machines (SVM), $k-$nearest neighbors (kNN) and artificial neural networks (ANN) to handle the problem of overdispersion in ecological count data. Lately, ML methods have been cropping up in different areas of science. However, to the best of our knowledge, there is limited empirical evidence showing the use of ML regression techniques in estimating missing observations in population ecological studies. ML models make no distributional assumptions about the response or predictor variables unlike many statistical analysis methods. Some ML techniques accommodate zeros both in the response and in the predictor variables: that is what makes them unique and powerful alternatives to statistical methods. We test the proposed estimation techniques on a real life fisheries count data set, which usually has missing observations due to structural, observer or design errors and we make a comparison to the regression count models to assess whether overdispersion can be reduced further.

The remainder of the study is organized as follows: Sect. 2 outlines the modeling approach relevant in this study; Sect. 3 presents the methods used in this study and in Sect. 4 we present the numerical test of the proposed methodology on the fisheries data set. Section 5 concludes this study and gives recommendations for future work.

2 Modeling Approach

In this section, we outline the modeling approach used in this study. ML regression techniques are proposed as alternative models to the traditional count regression models.

2.1 Count Regression Models

Suppose y_{ij} represents species count data for all sites $i = 1, \dots, n$; for all visits $j = 1, \dots, m$. Let y_{ij} follow a Poisson regression model with a conditional mean μ_{ij}

and a set of predictor variables (x_{ij}). The Poisson regression model is expressed as:

$$P\left(Y = y_{ij} \middle| x_{ij}\right) = \frac{\exp\left(-\mu_{ij}\right) \times \mu_{ij}^{y_{ij}}}{y_{ij}!}$$

where y_{ij} is a non-negative integer [27]. We assume a log link function where $\mu_{ij} = \exp\left(\beta_0, \beta_1 x_{i1}, \ldots, \beta_k x_{ik}\right)$ is a linear combination of predictor variables. By definition, the Poisson regression model cannot model overdispersed count data thus a gamma mixture Poisson is often assumed in this case. The negative binomial (NB) distribution is an alternative to the Poisson when data is overdispersed [8].

Suppose a random variable Y follows a Poisson distribution with a conditional mean μ_{ij} and the parameter λ follows a gamma distribution with mean $\mathrm{E}(\lambda) = \mu_{ij}$ and variance $\mathrm{Var}(\lambda) = \mu_{ij}^2 \theta^{-1}$. Let ϕ denote a dispersion parameter. The joint density of Poisson and gamma distributions leads to the NB regression model

$$P\left(Y = y_{ij} \middle| \mu_{ij}, \theta\right) = \frac{\Gamma\left(y_{ij} + \theta^{-1}\right)}{y_{ij}\Gamma\left(\theta^{-1}\right)} \left(\frac{\mu_{ij}\theta}{1 + \mu_{ij}\theta}\right)^{y_{ij}} \left(\frac{1}{1 + \mu_{ij}\theta}\right)^{\theta^{-1}}$$

where μ_{ij} denote the mean and the dispersion parameter $\phi = 1 + \mu_{ij}\theta$ should be greater than one. When Y follows a NB distribution, its expected value is $\mathrm{E}(Y) = \mu_{ij}$ and the variance is $\mathrm{Var}(Y) = \mu_{ij}(1 + \mu_{ij}\theta)$ [10]. When false zeros are too many to be modeled by a NB regression model, researchers consider the application of zero-inflated models such as zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models which are believed to be adequate.

The zero-inflated models, ZIP or ZINB, are a mixture of two distributions, which models true zeros through the Poisson or NB distribution and a degenerate distribution at zero, which models the true zeros [21]. The ZIP has two components, π which represent 0 observations and $(1 - \pi)$, which represents an observed Poisson random variable [28]. The ZIP regression model for count i at visit j is defined by Rose et al. [21] as:

$$P\left(Y = y_{ij}\right) = \begin{cases} \pi_{ij} + \left(1 - \pi_{ij}\right)\exp\left(-\mu_{ij}\right), & y_{ij} = 0 \\ \left(1 - \pi_{ij}\right)\frac{\exp(-\mu_{ij})\mu_{ij}^{y_{ij}}}{y_{ij}!}, & y_{ij} > 0 \end{cases}$$

where $0 \leq \pi_{ij} < 1$ denotes the probability of zero occurrence and $\mu_{ij} > 0$ with the mean $\mathrm{E}(Y_{ij}) = \mu_{ij}(1 - \pi_{ij})$ and the variance $\mathrm{Var}(Y) = \mu_{ij}(1 - \pi_{ij})(1 + \pi_{ij}\mu_{ij})$. The dispersion parameter for the ZIP regression model is given by $\phi = \mu_{ij}\pi_{ij} + 1$. The ZIP regression model provides a way to model zero-inflation owing to an excess number of zeros: furthermore, the ZINB regression model makes it possible to model both zero-inflation and overdispersion.

The ZINB regression model is defined by Rose et al. [21] and Aráujo et al. [14] as:

$$P\left(Y=y_{ij}\right)=\begin{cases}\pi_{ij}+\left(1-\pi_{ij}\right)\left(\frac{1}{1+\mu_{ij}\theta}\right)^{1/\theta}, & y_{ij}=0\\ \left(1-\pi_{ij}\right)\frac{\Gamma\left(y_{ij}+\theta^{-1}\right)}{\Gamma\left(y_{ij}+1\right)\Gamma\left(\theta^{-1}\right)}\left(\frac{1}{1+\mu_{ij}\theta}\right)^{1/\theta}\left(\frac{\mu_{ij}\theta}{1+\mu_{ij}\theta}\right)^{y_{ij}}, & y_{ij}>0\end{cases}$$

where $\theta > 0$ is a shape parameter which quantifies the amount of overdispersion, π_{ij} denotes the zero-inflation probability, y_{ij} denotes the species count and μ_{ij} the mean of count data. The mean and variance of the ZINB regression model are $\mathrm{E}(Y)=\mu_{ij}(1-\pi_{ij})$ and $\mathrm{Var}(Y)=\mu_{ij}(1-\pi_{ij})(1+\pi_{ij}\mu_{ij}+\theta\mu_{ij})$. The dispersion of the ZINB regression model is given by $\phi=\mu_i\pi_{ij}+\theta\mu_{ij}+1$.

2.2 Overdispersion for Count Data

The Pearson-based dispersion statistic as suggested by Hilbe [8] is to be used to assess overdispersion in traditional count models. If the ratio of the estimator $\hat{\phi}=\frac{D}{n-p}$, where D denotes the deviance and $n-p$ (n is the total number of observations and p is the number of parameters) denotes the degrees of freedom is greater than one, then there is evidence of overdispersion [2, 29]. This study adopts this estimator as a check for overdispersion.

2.3 Machine Learning Regression Techniques

Next, we discuss the machine learning (ML) regression techniques, considered in this study. Our regression problem is of the form: $y_{ij}=f\left(x_{ij}\right)+\varepsilon_{ij}$ in the presence of zero observations where $(\mathbf{X},\mathbf{Y})=\left\{\left(x_{ij};y_{ij}\right)_{i=1}^{N}\right\}$ are data samples with y_{ij} representing the target variable, x_{ij} denoting predictor variables and ε_{ij} the error term.

The aim is to investigate how ML regression techniques can reduce overdispersion for a single species in ecological count data. ML techniques, particularly in fisheries, have not been fully applied. This is due to lack of collaboration between the ML research community and natural scientists, a lack of communication about successful applications of ML in the natural sciences and difficulty in validating ML models [30, 31]. ML techniques have many algorithms and methodologies which are capable to solving real-world problems.

2.3.1 Random Forests

Random forests (RF) is a ML model that is designed to produce accurate predictions that do not overfit the data [32]. The RF model can be written in the form:

$$y(\mathbf{x})=f\left\{\sum_{k=1}^{K}w_k\phi\left(\mathbf{x},\mathbf{v}_k\right)\right\}$$

where $\mathbf{v}_k$ put into code the choice of variable to split on and the function $f()$ depends on whether a regression or classification tree is needed. Each tree is developed using a subset of randomly chosen k features [33].

RF can be used for both regression and classification problems; however, in this study we focus only on regression tasks. RF is commonly demonstrated by building many decision trees from bootstrap samples of a data set. Since individual trees often overfit the training data and result in noisy predictions, averaging is a way to reduce the variance of the model and improve prediction accuracy. In the RF model, there are three tuning parameters of interest: node size, number of trees and the number of predictor variables sampled at each split. The focus in this study is on one tuning parameter which is the number of predictor variables sampled at each split, ($mtry$). The $mtry$ is determined by the total number of predictor variables in the data set and it controls for overfitting [34].

2.3.2 Support Vector Machines

For a given data $D = \{(x_i, y_i)\}_{i=1}^{p} \in \mathbb{R}^n \times \{-1, +1\}$, the aim is to find a function $f(x) = y$ that correctly classifies the patterns of the data, where x_i denote a n-dimensional vector and y_i is its label. The aim of support vector machines (SVM) is to create a hyperplane function that separates the observations into classes. The hyperplane can be defined as:

$$f(x) = (w.x) + b$$

where $w \in \mathbb{R}^n$; $b \in \mathbb{R}$ and the data is then linearly separable, if such a hyperplane exists [35, 36]. SVM also solve non-linear problems by mapping the input vectors to a higher dimensional space using kernel functions $k\left(x_i, x_j\right) = \left\{\varnothing\left(x_i\right) \times \varnothing\left(x_j\right)\right\}$ [37]. Then, the decision function can be written as:

$$f(x) = sign\left\{\sum_{i=1}^{p}(w.x) + b\right\}$$

The computational complexity of the SVM approach is one of its primary features. For the SVM model, the Gaussian radial basis function kernel was used since our count data was non-linear. This particular function relies only on one tuning parameter, which is the cost parameter (C). The cost parameter determines the possible misclassifications of the model. It essentially forces a penalty to the model for making an error. The higher the value of C, the less likely it is that the SVM will misclassify a point.

2.3.3 *k*−nearest Neighbors

The $k-$nearest neighbors (kNN) are determined by calculating the Euclidean distance between the input feature vector x and the data set [37]. The predictor $\hat{f}(x)$ is computed at point x when we first define a neighborhood $N_k(x)$ corresponding to the set of the

$k-$ closest observations to x among learning sample in $\mathbb{R}^d$. The predictor is the average of the output over the kNN,

$$\hat{f}(x) = \frac{1}{k} \sum_{i \in N_k(x)} Y_i$$

where $N_k\ (x)$ is the neighborhood of x defined by the k closest points i in the training sample [5]. To improve accuracy, weights on each neighbor can be added, such as a weight inversely proportional to the Euclidean distance, effectively giving a greater importance to neighbors that are closer.

2.3.4 Artificial Neural Networks

Artificial neural networks (ANNs) are generalizations of linear models inspired by analogies with the biological brain. The architecture of an artificial neural network (ANN) is based on the Multi-Layer Perceptron (MLP) [12]. A perceptron is a simple unit which computes the following function:

$$h(\alpha) = f\ (w_i\,.\,\alpha_i)$$

where the activation function f is a non-linear function of its argument. Examples are the function:

$$f(x) = sign \begin{cases} +1, & \text{if } x \geq 0, \\ -1 & \text{if } x < 0 \end{cases}$$

or the sigmoid function:

$$f(x) = \frac{1}{1 + e^{-x}}.$$

ANNs are interconnected through a feed-forward propagation mechanism, where each neuron recieves input from preceding neurons. The network starts from input layers that are linked to each neuron in the one or many hidden layers that use a backpropagation algorithm to maximize the weights placed at each neuron to improve predictive power. This process is iterative, where the last hidden layer is met by an output layer to produce a predicted response output [38].

3 Materials and Methods

This section discusses the methodologies used in this study; components, as well as data preparation and processing. The performance evaluation measures used to assess the model performance in this study are also discussed.

Table 1 Descriptive statistics for the variables

Variables	Mean	Standard deviation	Median	Minimum	Maximum
Individual count	18.82	32.11	6.00	0.00	282.00
Decimal latitude	0.59	0.61	0.46	− 0.65	1.85
Decimal longitude	28.11	11.87	33.13	0.00	34.02

3.1 Data Description

The data set used here is from "Fish Species Occurrence Records for Uganda Mobilized Observation Archives" sourced from Uganda mobilized unpublished archives, available at https://www.gbif.org. The data set was readily available and accessed in October 2020. The data set chosen consists of 424 observations with 3 variables. This data set presents fish species occurrence records for surveys which were gathered at different periods in most of the aquatic ecosystems of Uganda. The experiment is conducted on a single species, *Lates niloticus*. In the data set, counts of *Lates niloticus* are utilized as the target variable. The fish were captured using fishing gears such as gillnets, hooks and beach seines at specific sites of different waterbodies. The data set had basic useful information for developing occurrences such as GPS co-ordinates (latitude and longitude) on the actual locations or sites where the fish were captured. The latitude and longitude co-ordinates were checked for errors by visualizing them on Google Earth. The predictor variables, latitude and longitude, were used when fitting the models since sampling was done to different sampling locations or sites. Table 1 presents descriptive statistics of the variables for the *Lates niloticus* count data set. We can observe that the large portion of the count data is between 0 and 50 as depicted in Fig. 1. Basically, Fig. 1 shows that there are many zero observations, indicating the presence of overdispersion. Overdispersion in the *Lates niloticus* count data was determined by computing the dispersion of the fitted models as discussed in Sect. 2.2.

3.2 Pre-processing of Data Set

The pre-processing approaches have been used to train ML regression techniques which, might be taken to increase model performance. Models were fit using species observation count data. Prior to model building, the data were scaled to avoid numerical stability in the fitted models. The results were generated using k-fold cross validation (CV) with the hyperparameter optimization performed on every iteration. The data intended for modeling was split into 70% training for building a predictive model and 30% testing for evaluating the model. Splitting of data helps to reduce the possibilities of overfitting on the training set. All the work was conducted using R version 3.6.2 [39] and ML models were trained with caret package using the models' relevant functions.

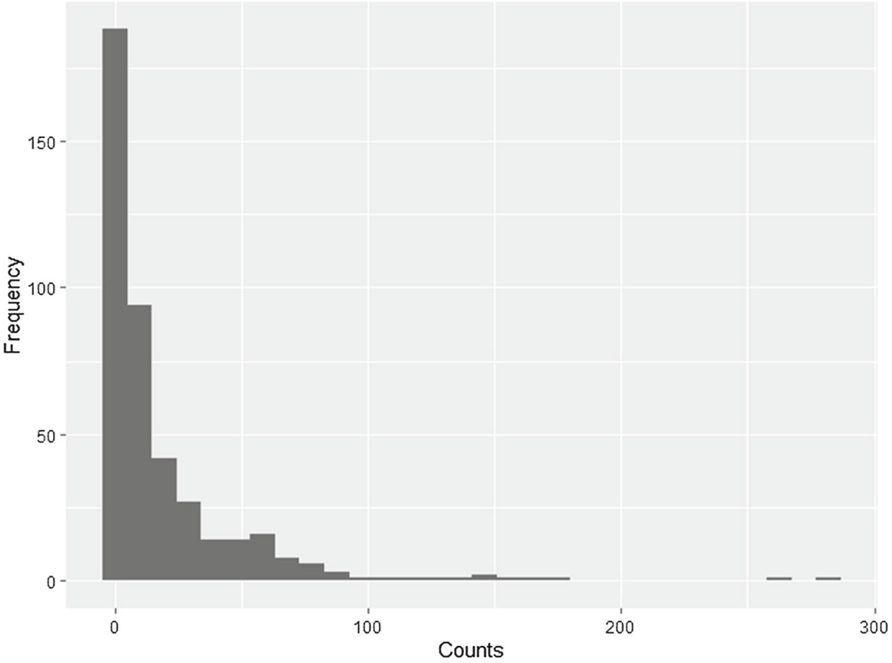

Fig. 1 Graphical representation of count data

Table 2 Mathematical formulas for performance metrics

Metrics	Formulas
Mean squared error (MSE)	$\mathrm{MSE} = \frac{1}{n}\sum_{i=0}^{n}\left(y_i - \hat{y}_i\right)^2$
Root mean squared error (RMSE)	$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=0}^{n}\left(y_i - \hat{y}_i\right)^2}$
Mean absolute error (MAE)	$\mathrm{MAE} = \frac{1}{n}\sum_{i=0}^{n}\left\|y_i - \hat{y}_i\right\|$

3.3 Performance Evaluation Measures

Three common evaluation measures suitable for comparison of both count regression models and ML regression models are mean squared error (MSE), mean absolute error (MAE) or root mean squared error (RMSE). Table 2 presents the formulas for these metrics, where y_i and $\hat{y}_i$ are the observed and predicted values, respectively. The best model, selected on the performance of different evaluation metrics, is the one with the least MSE or RMSE or MAE value. In this study, we consider the MAE as our benchmark for assessing model performance for count regression and ML regression models. This metric was chosen because it is suitable for a data set that has outliers [40, 41].

Table 3 Results for count regression models

Model	Dispersion	MAE
Poisson GLM	47.718	19.709
NB	1.397	19.651
ZIP	–	19.716
ZINB	0.561	19.651

4 Experimental Results and Discussions

In this section, we evaluate the performance of the proposed ML regression models in the presence of zero-inflation and overdispersion for *Lates niloticus* count data. We compare the performance of the proposed ML regression techniques with count regression models and present the empirical insights/findings.

4.1 Count Regression Model Results

Table 3 displays the results of the count regression models. The Poisson regression model was fitted in R using the glm function. After fitting the Poisson regression model, we tested for overdispersion using the dispersion test function from AER package in R software. The Poisson regression model estimates the dispersion parameter as 47.718, which is an indication of overdispersion since it is greater than one. The dispersion parameter in the NB and ZINB models were estimated at 1.397 and 0.561, respectively. Since the dispersion parameter for the ZINB regression model was 0.561, we observe that ZINB works better in comparison to the Poisson regression model when the data is both zero-inflated and overdispersed. All our count regression models exhibit overdispersion regardless of whether they accommodate both zero-inflation and overdispersion. This study confirms that the presence of overdispersion has an impact on the performance of count regression models. The Poisson regression model is the most commonly used model; however, the findings confirm that it is poor in dealing with either zero-inflation or overdispersion.

As highlighted in Sect. 3.3, RMSE and MSE are sensitive to the data set that consists of outliers. Therefore, we concentrate on the comparison of MAE to be able to compare traditional count regression models to ML regression models. Comparison of the MAE results shows that the NB (MAE = 19.651) and ZINB (MAE = 19.651) had the lowest MAE values. In this case, ZIP had the highest MAE of 19.716. Amongst the count regression models, NB and ZINB performed better compared to the Poisson and ZIP models. The reason is the extra dispersion parameter that accommodates overdispersion. The count regression model findings support Minami et al. [20] and Rose et al. [21] suggestion about the ZINB model. The performance metrics of our count regression models improve that of Buyrukoğlu et al. [40].

Table 4 Summary of results on ML regression models

Model	Optimized parameters	MAE
RF	$mtry = 2$	17.525
SVM	$C = 1$	13.670
kNN	$k = 3$	14.850

ANN	Size = 3
	Decay = 1e-04
		19.716

4.2 Machine Learning Regression Models Results

The RF, SVM, kNN and ANN were selected as ML regression models and their results are presented in Table 4. We used observations of single species count data to estimate model performances. The RF regression model showed that $mtry = 2$ resulted in a better MAE value of 17.525. Multiple studies [32, 40] have demonstrated that RF models often perform well in comparison to other methods for ecological modeling. However, the SVM in this study provided better performance when compared to RF the model. For instance, when using the Gaussian radial basis procedure, the optimum SVM model with the smallest MAE value of 13.670 was selected. As SVM are stable algorithms that can deal with large sets of predictors at once, they proved particularly useful in this study. This gives us confidence to say that SVM (owing to flexibility) can catch the information in the data set, even with overdispersed data.

For the kNN model, the number of neighbors (k) is the key parameter. Multiple k-values were used to determine the optimum model. The kNN model with $k = 3$ revealed an optimal model with a lower MAE value of 14.850. The kNN showed a better performance for this study; hence this technique can be used when modeling ecological count data.

For the ANN model, the performance was determined by the number of units within the hidden layers. ANN had the highest MAE value of 19.716 compared to other ML regression models. The highest performance of ANN was observed with three neurons in the hidden layer. The rule of ANNs are that the number of hidden layer neurons should range between 70% and 90% of the size of the input layer and that the number of hidden layer neurons should be less than twice that of neuron numbers in the input layer. ANN is the least performing ML model in this study; hence, we do not recommend it in reducing overdispersion in ecological count data.

4.3 Model Performance Comparison

As stated in Sect. 3.3, MAE was chosen as a benchmark for model performance since it enables us to have a comparison of count regression and ML regression models. Table 5 shows model comparison for both count regression and ML regression models arranged in ascending order for MAE. SVM has the least MAE whereas ANN has the highest MAE. The count regression models presented poor performance in reducing overdispersion. These models presented high MAE values. This is an indi-

Table 5 Model comparison

Model	MAE
SVM	13.670
*k*NN	14.850
RF	17.525
NB	19.651
ZINB	19.651
Poisson GLM	19.709
ANN	19.716

cation that count regression models are not capable enough of modeling zero-inflated overdispersed count data.

The ML results showed that SVM outperforms all other ML regression techniques. In comparison to the count regression models, the ML regression models showed the best performance based on MAE values. This signifies that ML regression models have a capacity to reduce the overdispersion problem. These finding concurs with the previously reported results by Cutler et al. [32]; Kampichler et al. [42] and Olaya-Marín et al. [30] which found that ML regression techniques work better in comparison to count regression models in ecological count data. The comparison of different ML regression techniques should be considered as recommended by Olaya-Marín et al. [30], as this would be helpful in interpreting the quality of the results.

We urge that individual researchers select techniques that are consistent with the specific problem, the nature of the questions being addressed and the availability of the data set. The use of ML techniques in ecological research is motivated by a range of research questions, the type of the available data and the expected outcomes of modeling. This makes ML techniques a very dynamic approach for predictive and exploratory modeling with many user-defined parameters to be considered for each objective [38].

5 Conclusion

The main aims of this study were to investigate the overdispersion problem that is rampant in ecological count data and propose possible methods for reducing it. This study compared the performance of count regression models with ML regression techniques in order to evaluate their capabilities in reducing overdispersion and to assess their conformity in terms of performance metrics through the MAE values. For the traditional count regression models, Poisson GLM, NB, ZIP and ZINB were considered. ZINB performed better compared to the other count regression models.

Four ML regression techniques were implemented: RF, SVM, *k*NN and ANN as well as their performances were compared to traditional count regression models. The results indicated that, generally, the SVM, *k*NN and RF models used in this study provide better performances when compared to the NB, ZINB, Poisson GLM and ANN models. The SVM model performed better in reducing overdispersion in comparison to the other ML regression models. The most prominent finding was that overall, SVM, *k*NN and RF models provide better performances comparing to the traditional count regression models based on MAE values.

ML regression techniques appear to be an attractive route towards tackling the overdispersion problem, particularly in areas where a lack of knowledge exists regarding the implementation of effective models. The comparative study shows the potential of ML regression techniques in reducing overdispersion in ecological count data. The strong performance shown by ML regression techniques should motivate further research, resulting in the improvement of methods and applications in ecological studies. In future, the application of ML regression techniques in ecology will become an increasingly attractive tool for ecologists. Moreover, we conclude that the effect of overdispersion in ecological count data is largely dependent on the dispersion parameter of the fitted model.

Author Contributions All the Authors contributed equally.

Funding Open access funding provided by North-West University. No funding was received for conducting this study.

Data Availability The data that support the findings of this study are available from [www.gbif.org] but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of [National Fisheries Resources Research Institute of Uganda].

Code Availability The code used in this work can be made available upon reasonable request from the corresponding author.

Declarations

Conflict of interest On behalf of all authors, the corresponding author Bonelwa Sidumo (email:Bonelwa.Sidumo@nwu.ac.za) states that there is no conflict of interest.

Ethical Approval The authors follow all the relevant ethical rules.

References

1. Bolker BM, Brooks ME, Clark CJ et al (2009) Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol Evol 24(3):127–135
2. Zuur AF, Ieno EN, Walker NJ et al (2009) Mixed effects models and extensions in ecology with R. Springer, Berlin

3. Martin TG, Wintle BA, Rhodes JR et al (2005) Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecol Lett 8(11):1235–1246. https://doi.org/10.1111/j.1461-0248.2005.00826.x
4. Blasco-Moreno A, Pérez-Casany M, Puig P et al (2019) What does a zero mean? understanding false, random and structural zeros in ecology. Methods Ecol Evol 10(7):949–959. https://doi.org/10.1111/2041-210X.13185
5. Crisci C, Ghattas B, Perera G (2012) A review of supervised machine learning algorithms and their applications to ecological data. Ecol Model 240:113–122. https://doi.org/10.1016/j.ecolmodel.2012.03.001
6. Harrison XA (2014) Using observation-level random effects to model overdispersion in count data in ecology and evolution. PeerJ 2:e616. https://doi.org/10.7717/peerj.616
7. Lee C, Famoye F, Akinsete A (2021) Generalized count data regression models and their applications to health care data. Ann Data Sci 8(2):367–386. https://doi.org/10.1007/s40745-019-00221-8
8. Hilbe JM (2011) Negative binomial regression. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511973420
9. Kassahun W, Neyens T, Faes C et al (2014) A zero-inflated overdispersed hierarchical Poisson model. Stat Model 14(5):439–456. https://doi.org/10.1177/1471082X14524676
10. Lindén A, Mäntyniemi S (2011) Using the negative binomial distribution to model overdispersion in ecological count data. Ecology 92(7):1414–1421. https://doi.org/10.1890/10-1831.1
11. Desjardins CD (2016) Modeling zero-inflated and overdispersed count data: an empirical study of school suspensions. J Exp Educ 84(3):449–472. https://doi.org/10.1080/00220973.2015.1054334
12. Sakthivel K, Rajitha C (2017) A comparative study of zero-inflated, hurdle models with artificial neural network in claim count modeling. Int J Stat Syst 12(2):265–276
13. Zamani H, Ismail N (2013) Score test for testing zero-inflated Poisson regression against zero-inflated generalized Poisson alternatives. J Appl Stat 40(9):2056–2068
14. Aráujo EG, Vasconcelos J, dos Santos DP et al (2021) The zero-inflated negative binomial semiparametric regression model: application to number of failing grades data. Ann Data Sci. https://doi.org/10.1007/s40745-021-00350-z
15. Gupta R, Szczesniak RD, Macaluso M (2015) Modeling repeated count measures with excess zeros in an epidemiological study. Ann Epidemiol 25(8):583–589. https://doi.org/10.1016/j.annepidem.2015.03.011
16. He H, Zhang H, Peng Y et al (2019) A test of inflated zeros for poisson regression models. Stat Methods Med Res 28(4):1157–1169. https://doi.org/10.1177/0962280217749991
17. Bekalo DB, Kebede DT (2021) Zero-inflated models for count data: an application to number of antenatal care service visits. Ann Data Sci 8(4):683–708. https://doi.org/10.1007/s40745-021-00328-x
18. Elliott RJ, Morrell CH (2009) Learning SAS in the computer lab. Cengage Learning, Boston
19. Lord D, Mannering F (2010) The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Trans Res Part A Policy Pract 44(5):291–305. https://doi.org/10.1016/j.tra.2010.02.001
20. Minami M, Lennert-Cody CE, Gao W et al (2007) Modeling shark bycatch: the zero-inflated negative binomial regression model with smoothing. Fish Res 84(2):210–221. https://doi.org/10.1016/j.fishres.2006.10.019
21. Rose CE, Martin SW, Wannemuehler KA et al (2006) On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. J Biopharm Stat 16(4):463–481. https://doi.org/10.1080/10543400600719384
22. Richards SA (2008) Dealing with overdispersed count data in applied ecology. J Appl Ecol 45:218–227. https://doi.org/10.1111/j.1365-2664.2007.01377.x
23. Hilbe JM (2014) Modeling count data. Cambridge University Press, Cambridge
24. Tien JM (2017) Internet of things, real-time decision making, and artificial intelligence. Ann Data Sci 4(2):149–178
25. Waller MA, Fawcett SE (2013) Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. J Bus Logist. https://doi.org/10.1111/jbl.12010
26. Hussain S, Khan MQ (2021) Student-performulator: predicting students' academic performance at secondary and intermediate level using machine learning. Ann Data Sci. https://doi.org/10.1007/s40745-021-00341-0

27. Cheung YB (2002) Zero-inflated models for regression analysis of count data: a study of growth and development. Stat Med 21(10):1461–1469. https://doi.org/10.1002/sim.1088
28. Lambert D (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34(1):1–14
29. Kruppa J, Hothorn L (2021) A comparison study on modeling of clustered and overdispersed count data for multiple comparisons. J Appl Stat 48(16):3220–3232. https://doi.org/10.1080/02664763.2020.1788518
30. Olaya-Marín EJ, Martínez-Capel F, Vezza P (2013) A comparison of artificial neural networks and random forests to predict native fish species richness in Mediterranean rivers. Knowl Manag Aquat Ecosyst 409(7):1–19
31. Thessen AE (2016) Adoption of machine learning techniques in ecology and earth science. One Ecosyst 1:e8621. https://doi.org/10.3897/oneeco.1.e8621
32. Cutler DR, Edwards TC Jr, Beard KH et al (2007) Random forests for classification in ecology. Ecology 88(11):2783–2792. https://doi.org/10.1890/07-0539.1
33. Dastile X, Celik T, Potsane M (2020) Statistical and machine learning models in credit scoring: a systematic literature survey. Appl Soft Comput 91:106–263. https://doi.org/10.1016/j.asoc.2020.106263
34. Fox EW, Hill RA, Leibowitz SG et al (2017) Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. Environ Monit Assess 189(7):1–20. https://doi.org/10.1007/s10661-017-6025-0
35. Yang H, Chan L, King I (2002) Support vector machine regression for volatile stock market prediction. In: International conference on intelligent data engineering and automated learning, Springer, pp 391–396
36. Shi Y, Tian Y, Kou G et al (2011) Optimization based data mining: theory and applications. Springer, Berlin
37. Srinivasa K, Siddesh G, Manisekhar S (2020) Statistical modelling and machine learning principles for bioinformatics techniques, tools, and applications. Springer Nature, Berlin
38. Ghannam RB, Techtmann SM (2021) Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. Comput Struct Biotechnol J 19:1092–1107. https://doi.org/10.1016/j.csbj.2021.01.028
39. Team RC (2020) R: A language and environment for statistical computing
40. Buyrukoğlu G, Buyrukoğlu S, Topalcengiz Z (2021) Comparing regression models with count data to artificial neural network and ensemble models for prediction of generic escherichia coli population in agricultural ponds based on weather station measurements. Microbial Risk Anal. https://doi.org/10.1016/j.mran.2021.100171
41. Do Nascimento RL, Fagundes RAdA, De Souza RM (2022) Statistical learning for predicting school dropout in elementary education: a comparative study. Ann Data Sci 9(4):801–828. https://doi.org/10.1007/s40745-021-00321-4
42. Kampichler C, Wieland R, Calmé S et al (2010) Classification in conservation biology: a comparison of five machine-learning methods. Eco Inform 5(6):441–450. https://doi.org/10.1016/j.ecoinf.2010.06.003