




Research Article

Random Forests in Count Data Modelling: An Analysis of the Influence of Data Features and Overdispersion on Regression Performance

Ciza Arsène Mushagalusa ^{1,2}, Adandé Belarmain Fandohan ^{1,3}
and Romain Glèlè Kakai ¹

¹Laboratoire de Biomathématiques et d'Estimations Forestières, Faculty of Agronomic Sciences, University of Abomey-Calavi, 04 PB 1525, Cotonou, Benin

²Faculty of Agriculture and Environmental Sciences, Université Evangélique en Afrique (UEA), P. O. Box: 3323, Bukavu, Democratic Republic of the Congo

³Unité de Recherche en Foresterie et Conservation des Bioressources, Ecole de Foresterie Tropicale, Université Nationale d'Agriculture, BP 43, Kétou, Benin

Correspondence should be addressed to Ciza Arsène Mushagalusa; shaga.ciza@gmail.com

Received 17 May 2022; Revised 27 October 2022; Accepted 1 November 2022; Published 1 December 2022

Academic Editor: Hyungjun Cho

Copyright © 2022 Ciza Arsène Mushagalusa et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Machine learning algorithms, especially random forests (RFs), have become an integrated part of the modern scientific methodology and represent an efficient alternative to conventional parametric algorithms. This study aimed to assess the influence of data features and overdispersion on RF regression performance. We assessed the effect of types of predictors (100, 75, 50, and 20% continuous, and 100% categorical), the number of predictors ($p = 816$ and 24), and the sample size ($N = 50, 250,$ and 1250) on RF parameter settings. We also compared RF performance to that of classical generalized linear models (Poisson, negative binomial, and zero-inflated Poisson) and the linear model applied to log-transformed data. Two real datasets were analysed to demonstrate the usefulness of RF for overdispersed data modelling. Goodness-of-fit statistics such as root mean square error (RMSE) and biases were used to determine RF accuracy and validity. Results revealed that the number of variables to be randomly selected for each split, the proportion of samples to train the model, the minimal number of samples within each terminal node, and RF regression performance are not influenced by the sample size, number, and type of predictors. However, the ratio of observations to the number of predictors affects the stability of the best RF parameters. RF performs well for all types of covariates and different levels of dispersion. The magnitude of dispersion does not significantly influence RF predictive validity. In contrast, its predictive accuracy is significantly influenced by the magnitude of dispersion in the response variable, conditional on the explanatory variables. RF has performed almost as well as the models of the classical Poisson family in the presence of overdispersion. Given RF's advantages, it is an appropriate statistical alternative for counting data.

1. Introduction

Interest in modelling count data has increased significantly over the past two decades. The Poisson distribution is still the most widely used distribution for modelling count data in many research areas, despite the violation of its well-known property that the mean and variance are equal, conditional on explanatory variables. However, the Poisson distribution is inadequate to model overdispersed count data

[1, 2]. As a result, a multitude of alternative count data models have been developed to address the shortcomings of the Poisson regression in the presence of overdispersion, including quasi-Poisson [3], generalized Poisson [4], negative binomial [5–7], zero-inflated Poisson [8], and zero-truncated Poisson models [9].

Nevertheless, traditional count data models remain limited as they may not be capable of detecting the presence of complex nonlinear interactions between explanatory

variables and outcomes [10]. In addition, traditional statistical models used to predict overdispersed data from selected predictor variables are not suitable for “small- n -large- p ” problems [11].

The random forest (RF) is a good machine learning approach to improving prediction accuracy and model interpretation [12, 13]. RF is a “data-driven statistical method.” It is an ensemble learning approach developed to increase classification accuracy and regression tree prediction by combining many decision trees [14]. The algorithm takes advantage of two powerful techniques: random subspace selection at each split (“classification and regression trees (CART)-split criterion”) [15] and bagging (an abbreviation of “bootstrap aggregating”) of unpruned decision tree learners [16]. The nonlinear nature of RF gives it an edge over linear algorithms [14, 17]. RF algorithms provide variable importance measures for variable selection purposes. Moreover, they involve complex high-order interaction effects and are user-friendly because they have few parameters to set and are less demanding in preprocessing [18]. Over the past 25 years, RF has been successfully applied to a wide range of prediction problems, mainly for classification tasks [19–21]. RF has become an effective data analysis tool that performs well compared to many standard methods [22].

However, in a regression problem, the range of predictions that RF can make is limited by the highest and lowest values in the training dataset. When training and testing data ranges differ, such as when predicting overdispersed response variables, this behaviour becomes troublesome [23]. Overdispersion is one of the important constraints in statistical analysis and may lead a variable to appear as a significant predictor when it is not, by deflating or underestimating the standard errors of the estimates [24]. As a result, it is essential to fully understand the data’s nature. Furthermore, most of the existing machine learning methods for regression tasks are designed for continuous data, and their performance for counted data is not well known [13].

The primary motivation for this study is thus to evaluate the effectiveness of computer-based statistical approaches known as machine learning, such as RF, in the presence of overdispersed response variables. Since the performance of an algorithm depends mainly on the characteristics of datasets to be analysed (type of predictors, number of attributes in the dataset, sample size, noise, and dispersion), high performance requires the identification of the most appropriate algorithm for a given problem and dataset [25]. RFs are considered data type-specific, insensitive to unimportant and noisy predictors, and dependent on the number of informative variables [18, 19, 26].

The RF method has shown comparable or even better prediction performance than other learning methods, such as neural networks, partial least squares regression, and support vector machines [14, 19, 22]. Unfortunately, the impact of overdispersion on the RF performance for different data features has rarely been assessed [27–29]. Thus, determining the impact of data features on RF parameters’ setting and accuracy is an avenue to explore as efforts are needed to improve the accuracy of RF estimates [12].

Therefore, this study aimed to assess the influence of data features on RF regression performance and compare RF regression to count data classic regressions. In addition, datasets with varying magnitudes of overdispersion from two real case studies are also analysed to examine RF performance when the response variable is overdispersed.

2. Materials and Methods

2.1. Statistical Models

2.1.1. Random Forest Technique (RF). RF methods deal with both supervised classification and regression tasks [14]. This research was focused on regression analysis.

(i) *Basic Principles.* Assuming that we are given a training sample $D_n = \{(X_1, Y_1); \dots; (X_n, Y_n)\}$ of i.i.d. $[0, 1]^d \times R$ -valued random variables ($d \geq 2$) with the same distribution as the independent prototype pair D defined by (X, Y) of size n satisfying the $E[Y]^2 < \infty$ condition. The space $[0, 1]^d$ is equipped with the standard Euclidean metric. For a given $x \in [0, 1]^d$, the regression function estimate $m(x)$ can be obtained by

$$m(x) = E[Y|X = x]. \quad (1)$$

Regarding this, the regression function estimate m_n for the dataset D_n is assumed consistent if

$$E[m_n(X) - m(X)]^2 \longrightarrow 0, \quad \text{as } n \longrightarrow \infty. \quad (2)$$

In RF, the final predictor is an average of M independently randomized regression trees. For a given tree j randomly generated, the prediction value at the query point x is denoted $m_n(X; \Theta_j, D_n)$, where $\Theta_1, \Theta_2, \dots, \Theta_M$ are the outputs of a randomized variable Θ (also known as independent random variables).

In practice, the randomized variable Θ is used to resample the training set used in successive individual trees building [18]. The finite forest estimate (combination of the M trees grown) is obtained by

$$m_{M,n}(X; \Theta_1, \dots, \Theta_M, D_n) = \frac{1}{M} \sum_{j=1}^M m_n(X; \Theta_j, D_n). \quad (3)$$

In the randomForest package, the default value of M is 500. Since M may be chosen arbitrarily large (“limited only by available computing resources”), this estimate can be generalized from a modelling point of view, letting M tend to infinity by an infinite forest estimate [15], obtained by

$$m_{\infty,n}(X; D_n) = E_{\Theta} [m_n(X; \Theta, D_n)]. \quad (4)$$

It is noteworthy in practice that the above expectation is evaluated by Monte Carlo, i.e., by generating M (usually large) random trees and taking the average of individual outcomes. This procedure is justified by the law of large numbers [14, 30] which asserts that the almost surely conditional on D_n is expressed as follows:

$$\lim_{M \rightarrow \infty} m_{M,n}(X; \Theta_1, \dots, \Theta_M, D_n) = m_{\infty,n}(X; D_n). \quad (5)$$

(ii) *Algorithm.* The whole process of the RF algorithm can be summarized as follows:

Let M be the number of trees to grow in the forest and m be the number of variables to select at each node. The following steps are considered:

- (i) We resample the training set to create new bootstrap samples.
- (ii) For each bootstrap sample, we create an unpruned tree that is grown until all nodes contain observations no more than the maximal terminal node size defined by the user (a prespecified parameter).
- (iii) At each split node, a randomly selected subset of predictors is uniformly chosen. In regression tasks for p predictors, a value equal to $p/3$ is utilized as the default value in the R package randomForest. Then, the best split using only these predictors is determined or returned. In addition, all observations are not used in tree construction, and a subset a_n is picked from the whole sample D_n . If $a_n = n$, the algorithm runs in a bootstrap mode, whereas $a_n < n$ corresponds to subsampling (with replacement).
- (iv) The RF overall prediction value is the average (in regression) of the results obtained from all M given by individual trees [15].

(iii) *Parameter Tuning.* To date, there are limited studies focusing on RF parameter tuning [31]. According to Probst and Boulesteix [32], the most important parameters in RF methods are as follows:

- (i) Number of variables to possibly split in each node
- (ii) Minimal terminal node size
- (iii) Fraction of observations to sample in each tree
- (iv) Number of trees

2.1.2. Poisson Regression Extensions Used to Predict Count Data. The prediction performance of RF was compared with that of the Poisson model and its extensions. It is suggested that they be used for assessing count data. The following models are utilized in this investigation and are briefly described as follows:

(i) *Linear Model.* It is a linear function of the form:

$$Y = X\beta + \varepsilon; \varepsilon \sim N(0, \sigma_\varepsilon^2), \quad (6)$$

where Y is the vectorized form of the response variable, X is a covariate matrix, β is a vector of coefficients, and ε is a vector of normally distributed errors with 0 mean and a variance of σ_ε^2 [33].

(ii) *Poisson Model (Poisson).* It is a special case of the generalized linear model (GLM) with a log link function. It is the standard GLM for count data [34]. We assume that the response variable (y_i) follows the Poisson distribution with the mean μ_i . Its probability mass function is written as

$$P(Y = y_i; \mu_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} \quad (7)$$

$$= \exp\{[y_i \log \mu_i - \mu_i] - \log(y_i!)\},$$

where the conditional mean is obtained by $\mu_i = E(y_i/x_i) = \exp(x\beta)$ with x being the vector of covariates and β being the vector of unknown parameters [35–37].

(iii) *Negative Binomial Model (Neg.Bin).* In the negative binomial, a parameter of dispersion (K) is introduced to accommodate unobserved heterogeneity in the count data. Its probability density function is written as

$$f \frac{y_i}{x_i} = \frac{\tau(y_i + K)}{\tau(y_i + 1)\tau(K)} \left(\frac{K}{K + \mu_i}\right)^K \left(\frac{\mu_i}{K + \mu_i}\right)^{y_i}, \quad (8)$$

where K is a constant, τ is a function of y_i , $\mu_i = \exp(x\beta)$, x is the vector of covariates, and β is the vector of unknown parameters [38].

(iv) *Quasi-Poisson.* The generalized linear model is used to define the quasi-Poisson model. Let Y be a random variable such that

$$E(Y) = \mu, \quad (9)$$

$$\text{Var}(Y) = \phi \times \mu,$$

where $\text{Var}(Y)$ is Y 's variance, $\mu > 0$, ϕ is the dispersion parameter and $\phi > 0$, and $E(Y)$ is Y 's expectation.

(v) *Zero-Inflated Poisson model (ZIP).* This model is appropriate when there are zeros in excess. Its probability mass function is written as [34, 36]

$$P(Y = y_i | x_i, z_i) = \begin{cases} (w_i z_i + (1 - w_i(z_i))\text{Poisson}(\mu_i; 0 | x_i), & \text{if } y_i = 0, \\ (1 - w_i)\text{Poisson}(\mu_i; y_i | x_i), & \text{if } y_i > 0, \end{cases} \quad (10)$$

where z_i is the vector of covariates, w_i is the probability that the response value is zero, $\text{Poisson}(\mu_i; 0 | x_i) = \exp(-\mu_i)$, and $\text{Poisson}(\mu_i; y_i | x_i) = \exp(-\mu_i) (\mu_i^{y_i} / y_i!)$.

2.2. Simulation Studies

2.2.1. Effect of Dataset Characteristics on RF Algorithm Parameters' Setting and Accuracy. Three sample sizes ($N = 50, 250$, and 1250), three numbers of predictors ($p = 8, 16$, and 24), and five types of predictors (100, 75, 50, and 25% continuous, and 100% categorical) were considered. Schönbrodt and Perugini [39] stated that n should approach 250 for stable estimates. Biau [18] showed that $n > 500$ is sufficiently large, while 50 represents a small population size. Altogether, 45 data types were set. For each case, 1000 datasets were created. Three RF parameters were considered to determine the best combination, while the number of trees was maintained constant (500):

- (i) Number of variables to be randomly selected at each split (mtry): mtry = 2, 3, ..., p , where p = the number of predictors

- (ii) Proportion of samples for model training (sample size) with four levels (55, 63.2, 70, and 80%)
- (iii) Minimal number of samples within each terminal node (node size = 2, 3, 4, ..., 9)

Depending on the number of predictors, the combination of different parameters under study resulted in 224, 480, and 736 different scenarios for each data type. Each scenario was run 1000 times to find optimal parameter combinations. For all datasets, the generated explanatory variables were not correlated and were not informative. The dependent variable was generated randomly from the Poisson distribution with lambda equal to one, which is denoted as $\text{Poisson}(\lambda = 1)$. Categorical predictors were randomly generated by varying the number of levels from 2 to 8. Uncorrelated quantitative covariates were generated from a multivariate normal distribution (MVN) $y \sim N_n(0, 1)$, whose density function in the case of n dimensions is written as

$$f(y_i | \mu, \Sigma) = f \left((y_{i1}, y_{i2}, \dots, y_{in})' | \mu = (\mu_1, \dots, \mu_n)', \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{1n} \\ \cdot & \cdot \\ \sigma_{n1} & \sigma_n^2 \end{bmatrix} \right) \quad (11)$$

$$= \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp \left(-\frac{1}{2} (y_i - \mu)' \Sigma^{-1} (y_i - \mu) \right).$$

The symbol $|\Sigma|$ refers to the determinant of the matrix Σ . The matrix Σ must be positive semidefinite to assure that the most likely point is $\mu = (\mu_1, \mu_2, \dots, \mu_n)$ and that, as y_i moves away from μ in any direction, then the probability of observing y_i declines. The denominator in the formula for the MVN is a normalizing constant, which assures that the distribution integrates to 1.0 [40].

2.2.2. Effect of Overdispersion on RF Parameters' Setting and Accuracy. In order to assess the effect of overdispersion on the setting and performance of RF parameters, negative binomial-distributed (Y_{nb}) and quasi-Poisson-approximated (Y_{qp}) variables were randomly generated. The variance-to-mean relationship of Y_{nb} is quadratic, while the variance-to-mean relationship of Y_{qp} is linear such as

$$\begin{aligned} E(Y_{nb}) &= \mu, \\ E(Y_{qp}) &= \mu, \\ \text{Var}(Y_{nb}) &= \mu + \phi \times \mu^2, \\ \text{Var}(Y_{qp}) &= \mu + \phi \times \mu, \end{aligned} \quad (12)$$

where $\mu > 0$, ϕ is the dispersion parameter, and $\phi > 0$. For this objective, $\mu = 3$ with varying $\phi = 1, 3$, and 5 . Three different samples ($N = 50, 250$, and 1250) and five types of predictors as in the first simulation design were considered. Figure 1 shows the shape and the distribution of randomly generated overdispersed outcomes.

2.2.3. RF Performance Compared to Poisson Model Extensions. Negative binomial-distributed variables (Y_i) were generated with the linear and quadratic variance-to-mean relationships. We let the mean μ vary as a function of two informative covariates $m = 2$ and different levels of the dispersion parameter ϕ . The mean of the outcome Y_i of 1000 observations was obtained according to the model $\log(E(Y_{im} = y | X_{im})) = \log(\mu_i) = \alpha + \sum_{m=1}^2 \beta_m X_{im}$, where β is a collection of parameters ($\beta_1 = 1, \beta_2 = 0.5$) and $\alpha = 0.145$. In addition, three noninformative variables were added from a multivariate normal distribution (MVN) $\sim N_n(0, 1)$.

All datasets were divided into two subsets: training and validation. 70% of the data were randomly selected for model calibration and the remaining 30% for model validation.

Poisson regression, negative binomial, zero-inflated Poisson, and zero-inflated negative binomial were compared

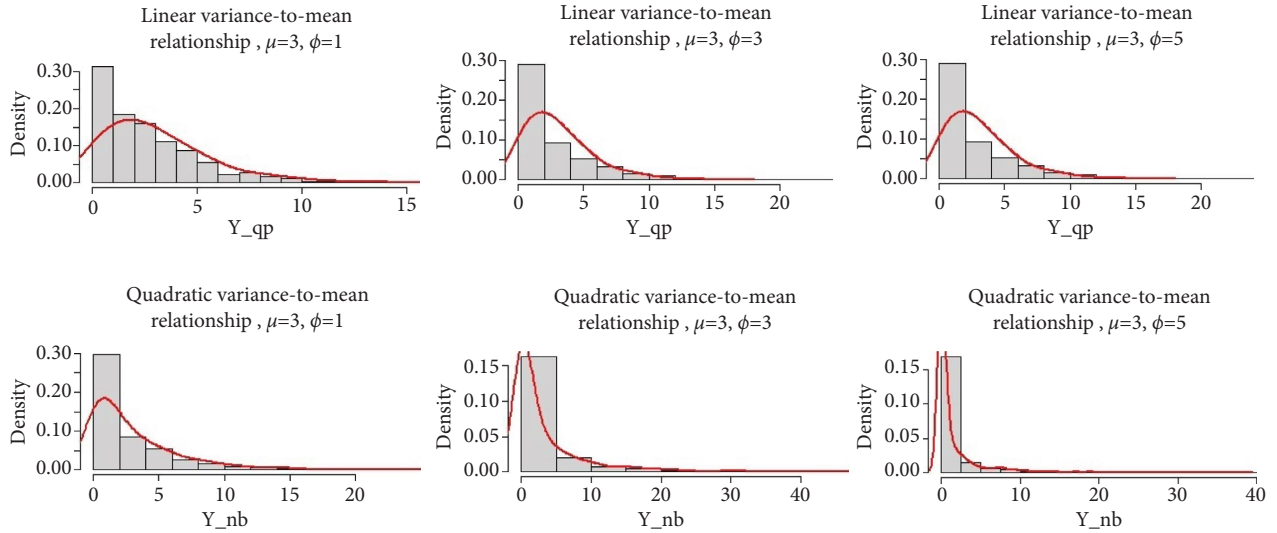


FIGURE 1: Histograms and density plots of some randomly generated overdispersed outcomes.

TABLE 1: Average tick counts on two local goat breeds.

Age (months)	Breed 1							Breed 2		
	Billy 1	Billy 2	Billy 3	Billy 4	Billy 5	Billy 6	Billy 7	Billy 8	Billy 9	Billy 10
1	1	3	4	1	4	3	4	4	2	2
2	3	3	6	1	8	7	6	8	6	6
3	4	3	7	1	13	15	4	10	14	8
4	5	7	16	9	4	7	12	8	6	18
5	16	8	9	6	19	13	19	39	12	17
6	1	3	4	31	16	9	16	16	14	2
7	1	24	11	8	18	10	32	4	37	9
8	17	3	12	9	4	11	12	36	10	18

to the RF regression technique and linear regression applied to the log-transformed response variable $\log(Y + 1)$, as it is considered to be one of the alternatives when dealing with overdispersion or count data [41, 42]. The analysis of variance (ANOVA) test and Tukey's honest significant difference test (Tukey HSD) were conducted to compare the performance of methods.

2.2.4. Criteria for Methods' Accuracy Assessment. The root mean square error (RMSE) and biases were computed for each model. These evaluation criteria are calculated as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (P_i - O_i)^2}, \quad (13)$$

$$\text{Bias} = \frac{1}{n} \sum_{i=1}^n (E(P_i) - O_i),$$

where P_i and O_i are the predicted and observed values in the subunit i , respectively, n is the number of samples (subunits) [43], and $E(P_i)$ is the average of the predicted values [44]. The RMSE represents the overall quality of the prediction. Predictions become increasingly optimal as the RMSE

approaches zero [43, 45]. All analyses were computed using R 4.0.3 software [46].

2.3. Motivating Real Datasets. We used two datasets from real case studies to examine RF performance at two magnitudes of overdispersion caused by outliers or zero inflation. The first example is tick burdens observed on two local goat breeds. Ticks were counted monthly between birth and the age of eight months on ten billy goats belonging to two local breeds [47]. The results are shown in Table 1. The dispersion statistic for this dataset is 3.050, caused by outliers.

The second example is the dataset "NMES1988": Demand for Medical Care in the National Medical Expenditure Survey (NMES) conducted in 1987 and 1988 [48]. This dataset contains 4406 observations on 19 variables. In this example, we examined the relationship between the number of physician office visits (response variable) and health (factor indicating self-perceived health status), age in years (divided by 10), gender, married (factor indicating that the individual is married), income (family income in USD 10000), and insurance (factor indicating that the individual is covered by private insurance) as predictors. The dispersion statistic for this subset is 7.345, caused by outliers and zero inflation.

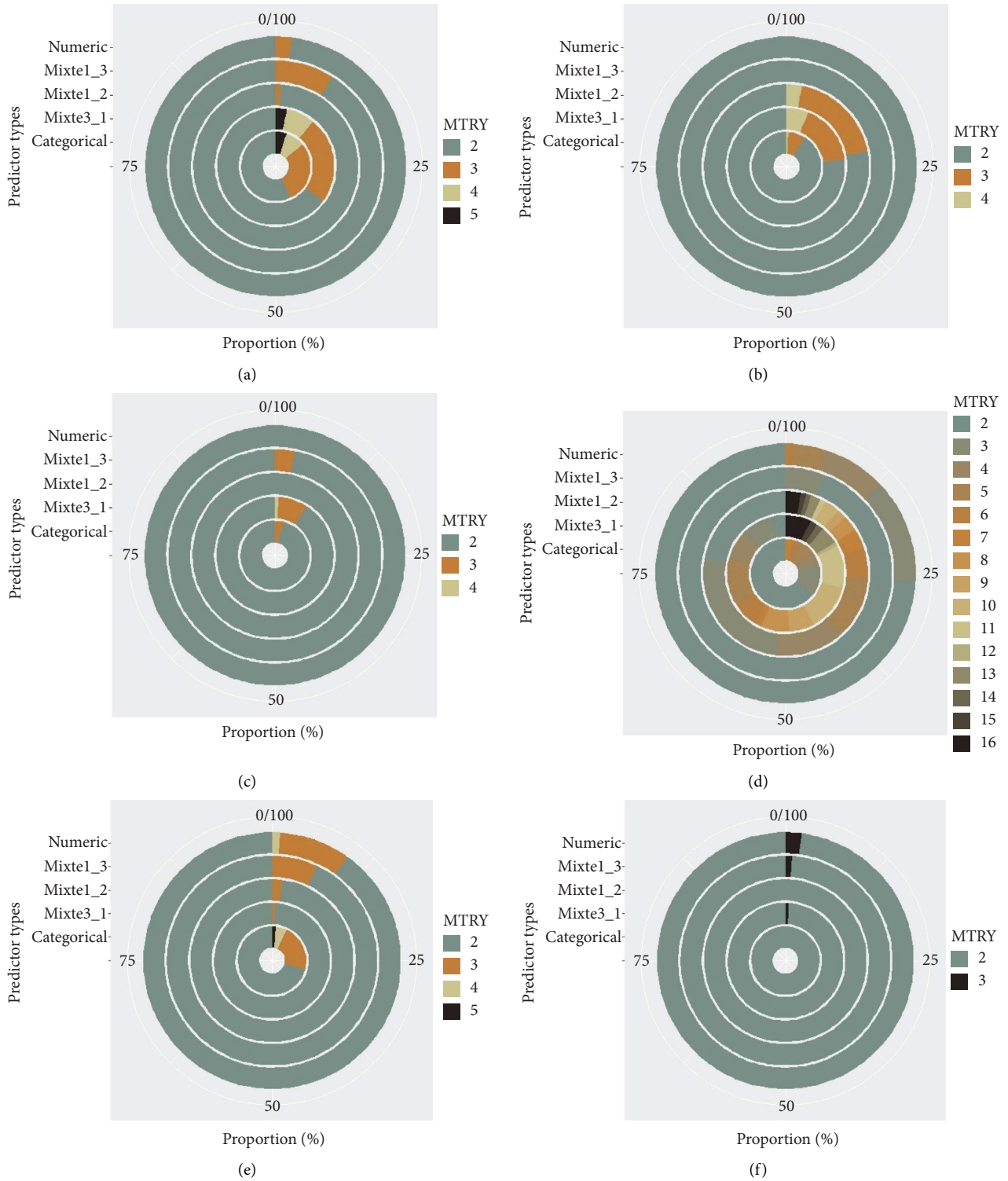


FIGURE 2: Continued.

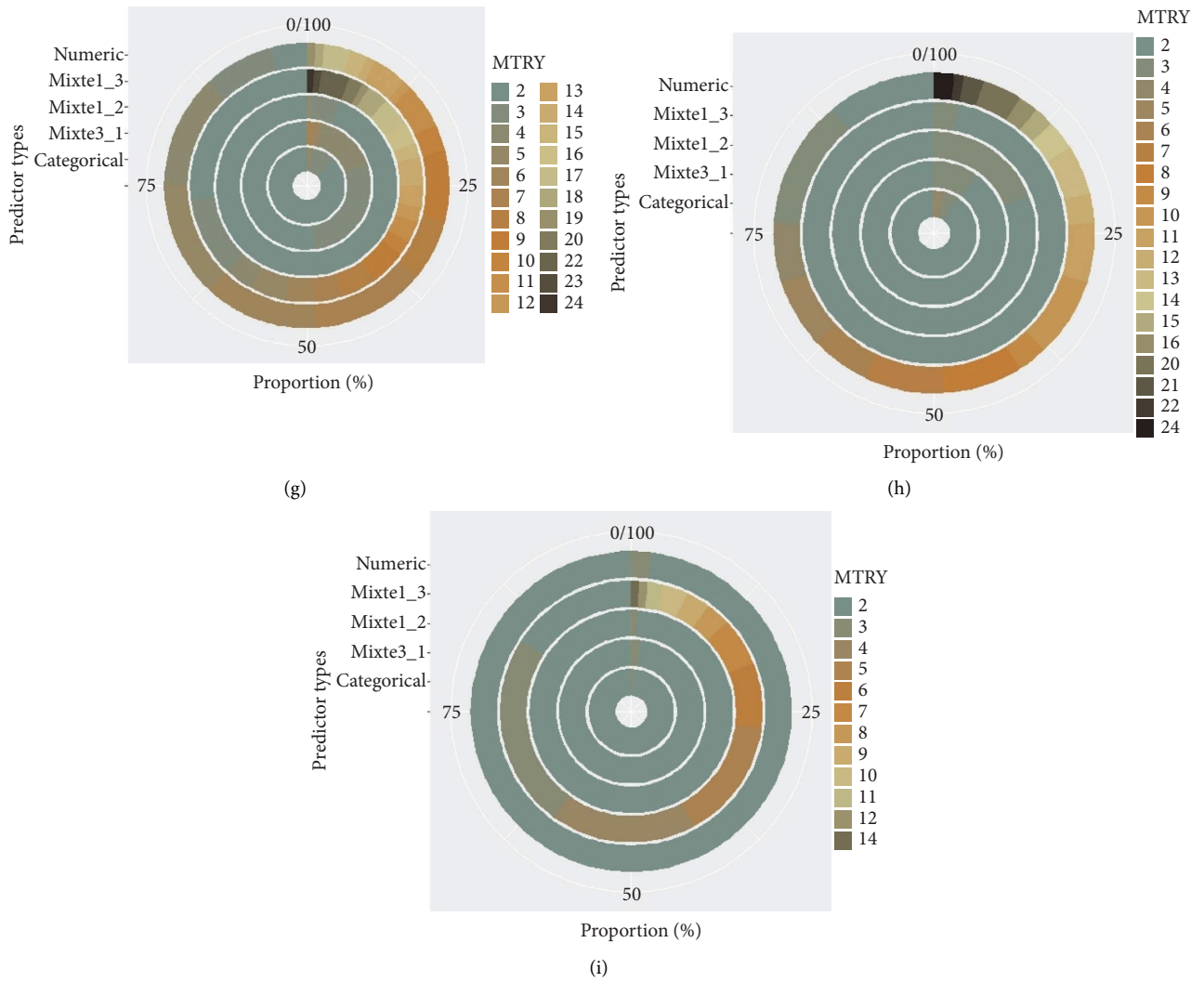


FIGURE 2: Best RF mtry as a function of the population size, number, and type of predictors (Numeric = 100% are quantitative predictors, Mixt1_3 = 25% are quantitative predictors, Mixt1_2 = 50% are quantitative predictors, Mixt3_1 = 75% are quantitative predictors, and Categorical = 100% are categorical predictors). (a) $N = 50, P = 8$, (b) $N = 250, P = 8$, (c) $N = 1250, P = 8$, (d) $N = 50, P = 16$, (e) $N = 250, P = 16$, (f) $N = 1250, P = 16$, (g) $N = 50, P = 24$, (h) $N = 250, P = 24$, and (i) $N = 1250, P = 24$.

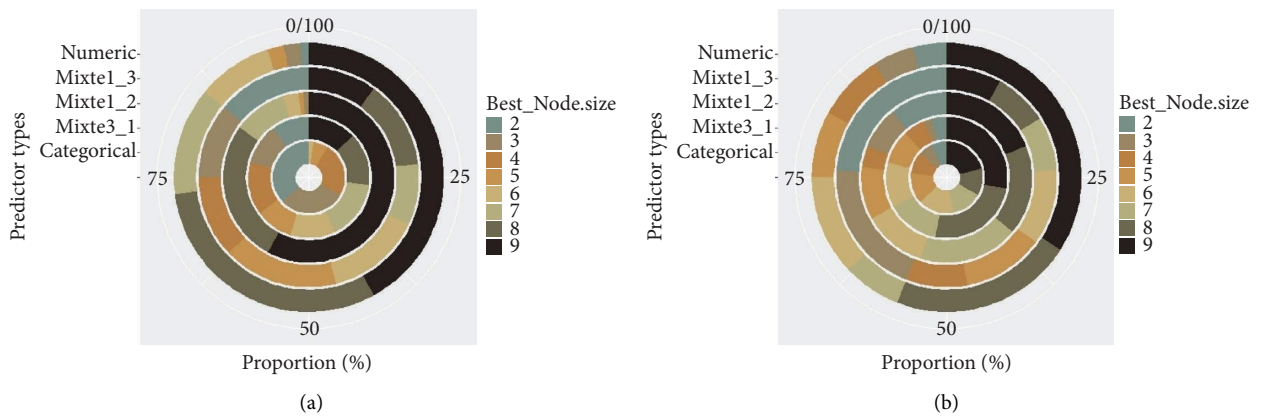


FIGURE 3: Continued.

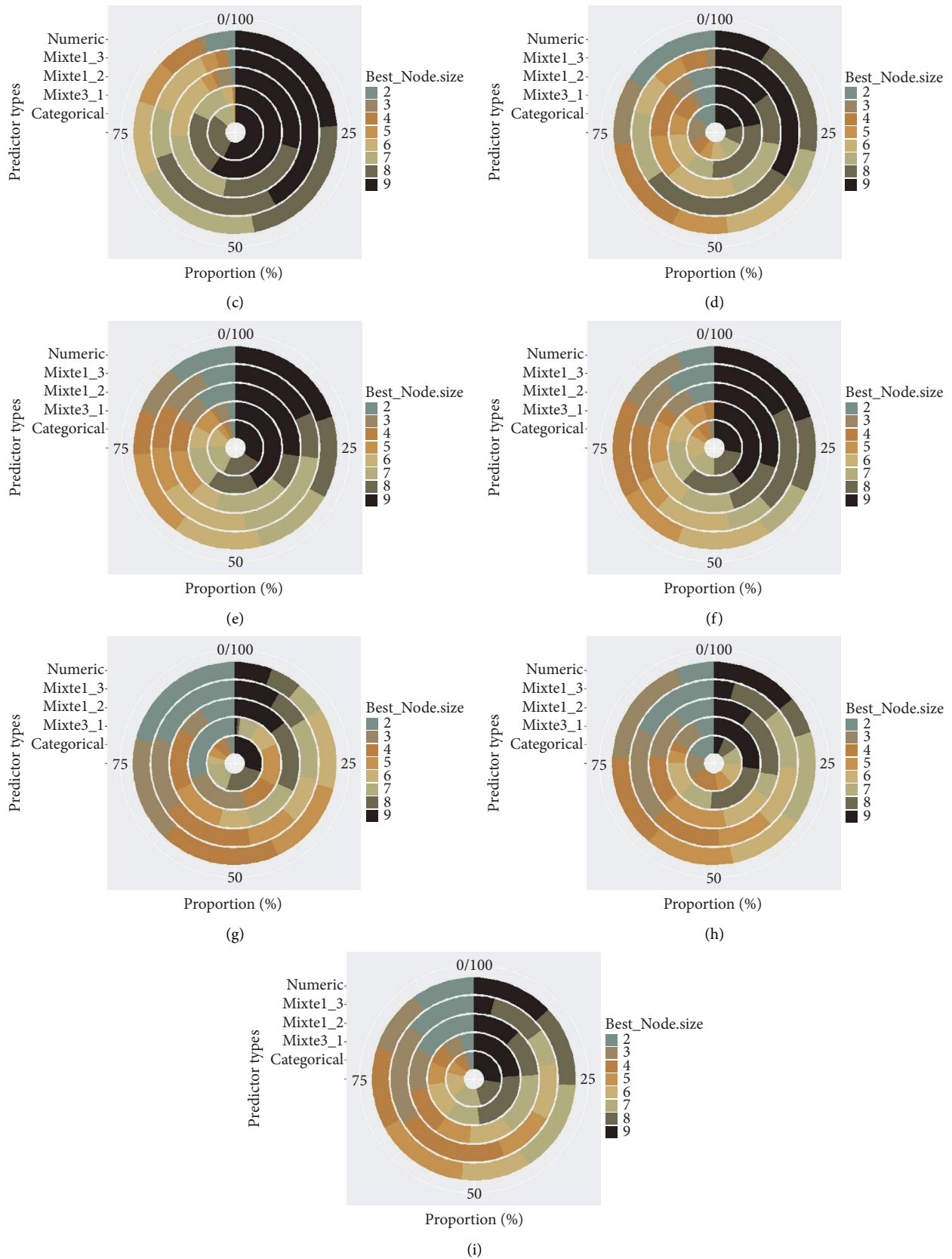


FIGURE 3: Best RF node size as a function of the population size, number, and type of predictors (Numeric = 100% are quantitative predictors, Mixt1_3 = 25% are quantitative predictors, Mixt1_2 = 50% are quantitative predictors, Mixt3_1 = 75% are quantitative predictors, and Categorical = 100% are categorical predictors). (a) $N = 50, P = 8$, (b) $N = 250, P = 8$, (c) $N = 1250, P = 8$, (d) $N = 50, P = 16$, (e) $N = 250, P = 16$, (f) $N = 1250, P = 16$, (g) $N = 50, P = 24$, (h) $N = 250, P = 24$, and (i) $N = 1250, P = 24$.

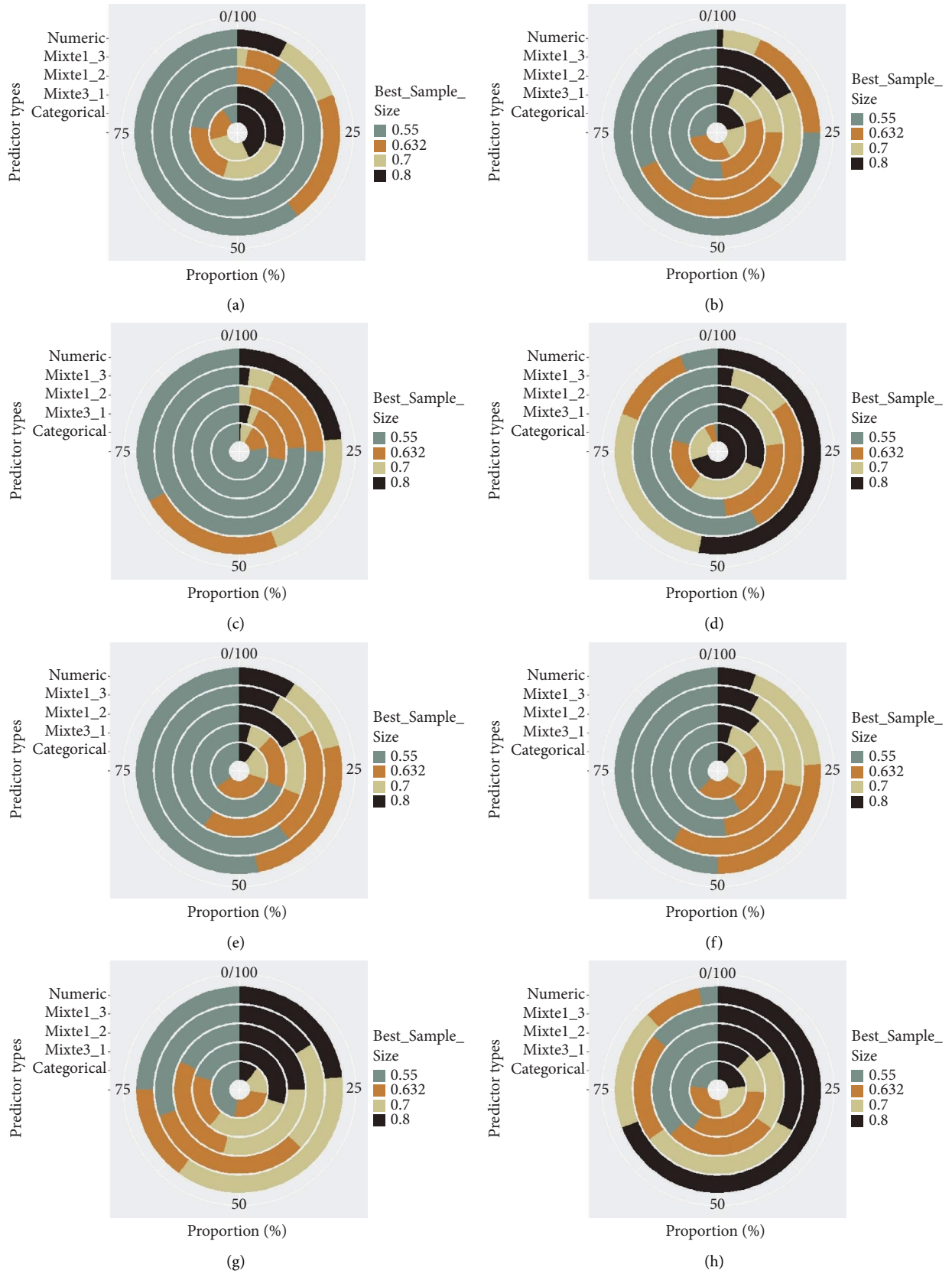


FIGURE 4: Continued.

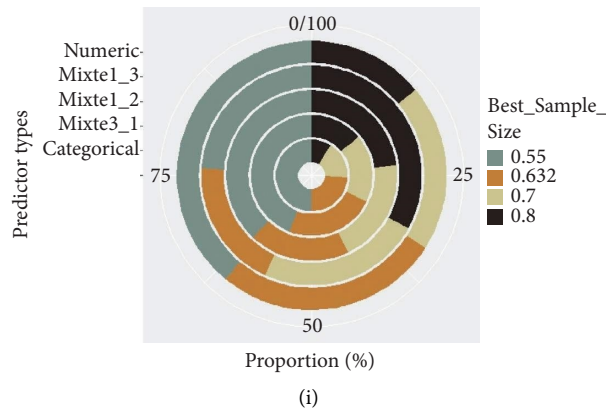


FIGURE 4: Best RF sample size as a function of the population size, number, and type of predictors (Numerical = 100% are quantitative predictors, Mixte1_3 = 25% are quantitative predictors, Mixte1_2 = 50% are quantitative predictors, Mixte3_1 = 75% are quantitative predictors, and Categorical = 100% are categorical predictors). (a) $N = 50, P = 8$, (b) $N = 250, P = 8$, (c) $N = 1250, P = 8$, (d) $N = 50, P = 16$, (e) $N = 250, P = 16$, (f) $N = 1250, P = 16$, (g) $N = 50, P = 24$, (h) $N = 250, P = 24$, and (i) $N = 1250, P = 24$.

3. Results and Discussion

3.1. Results

3.1.1. Effect of Data Features on RF Regression's Parameters' Tuning. The effect of predictor types on RF parameter settings for a response variable following a Poisson distribution with the parameter is $\lambda = 1$.

In most scenarios, as shown in Figure 2, the RF algorithm performs better with the number of variables randomly selected at each split (m_{try}), varying between two and four when there are many observations and a few predictors. Figure 2 shows that the number of variables to be randomly selected at each split is independent of the type of data. Nonetheless, the combination of various types of predictors causes some variability in the m_{try} values that yield the optimal predictive performance. When the sample size is small ($n = 50$) and the number of predictors is large ($p = 24$), there is a considerable fluctuation in m_{try} values (Figure 2(g)).

Figure 3 indicates that when the sample size is large ($n = 1250$) and the number of predictors is small (8), the RF algorithm performs better in the majority of cases with the highest minimum number of observations in each terminal node (between seven and nine). Furthermore, when the fraction of qualitative predictors was high, even for large sample sizes and a small number of predictors, the proportion of scenarios that achieved the optimal predictive performance with high values of the minimum number of observations in each terminal node (best node size \geq) also increased. Small sample sizes ($n = 50$) showed a considerable variation in the number of observations in terminal nodes.

Figure 4 shows that, in most circumstances, the RF method performs better when 55 percent of the whole sample size is used, followed by 63.2 percent. However, it turns out that the variation in this parameter depends on the ratio of the sample size to the number of predictors. The greater the ratio, the better the algorithm performs with fewer observations, whereas the lower the ratio, the more observations are used.

3.1.2. Effect of Overdispersion on the RF Parameter Settings for Various Predictor Types. As shown in Table 2, the RF algorithm works similarly for all types of predictors and magnitudes of dispersion. In most cases, regardless of the number of predictors or the extent of overdispersion, the optimal performance of RF is obtained by setting the number of variables to randomly sample at each split equal to two (mainly a small number of variables between two and four). Overall, for all types of predictors and observation sizes, it is observed that dispersion does not influence RF parameters' tuning.

According to Table 3, when predictors are categorical, for the same number of observations, the RF algorithm tends to use smaller samples to train the model as the number of predictors decreases. Thus, when eight predictors were considered, regardless of their type or level of overdispersion, the best performance was obtained in most cases using 55% of the total number of observations. In most scenarios, this proportion is followed by 63.2%. Furthermore, as the number of predictors increases, the 55% proportion loses dominance. As a result, the likelihood of obtaining the best performance with a large proportion of samples to train the model increases.

However, when all predictors are quantitative continuous or mixed with qualitative predictors, no trend was observed for the best size of the sample to draw and the minimum size of terminal nodes in RF, regardless of the complexity of the response variable.

In most cases, the best performance of RF is obtained with a minimum terminal node size greater than five, the default value used for regression. Table 4 shows that, regardless of the degree of overdispersion of the dependent variable, the RF algorithm performs best with a minimum terminal node size greater than five. For qualitative predictors, the preponderance of the high number of observations per node decreases as the number of predictors increases. However, as the number of predictors increases, the fraction of cases, where the best performance is achieved with a large number of observations per node, decreases. Thus, it turns out that globally, irrespective of the type of

TABLE 3: Effect of overdispersion on the RF optimal size of the sample to draw.

Data types	Variance-to-mean relationship	sample size	N = 250 (%)												N = 1250 (%)											
			p = 8				p = 24				p = 8				p = 24				p = 8				p = 24			
			$\phi = 1$	$\phi = 3$	$\phi = 5$	$\phi = 1$	$\phi = 3$	$\phi = 5$	$\phi = 1$	$\phi = 3$	$\phi = 5$	$\phi = 1$	$\phi = 3$	$\phi = 5$	$\phi = 1$	$\phi = 3$	$\phi = 5$	$\phi = 1$	$\phi = 3$	$\phi = 5$	$\phi = 1$	$\phi = 3$	$\phi = 5$			
Categorical	Linear	0.55	56	46	30	49	31	63	71	59	27	37	39	89	89	74	34	39	43							
		0.632	18	30	18	24	17	24	19	12	26	32	25	27	10	9	22	27	25	25						
		0.7	13	11	14	21	16	23	11	12	6	22	16	21	1	2	4	23	21	16						
	Quadratic	0.8	13	13	12	25	18	22	7	5	9	19	22	13	0	0	0	16	15	16						
		0.55	57	62	57	40	36	37	58	68	62	41	38	37	79	70	86	46	36	43						
		0.632	21	13	21	21	25	30	28	19	23	27	18	25	19	26	10	19	32	26						
25% of predictors are quantitative	Linear	0.7	12	12	10	19	22	13	9	8	12	17	22	20	1	4	0	23	15	17						
		0.8	10	13	12	20	17	20	5	5	3	15	22	18	1	0	4	12	17	14						
		0.55	100	0	0	0	0	100	0	100	100	0	0	100	100	0	100	0	100	0						
	Quadratic	0.632	0	0	100	100	0	0	0	0	0	0	0	0	0	0	0	100	0	0						
		0.7	0	0	0	0	0	0	0	100	0	0	100	0	0	0	0	0	0	0						
		0.8	0	100	0	0	100	100	0	0	0	0	0	0	0	0	0	0	0	0						
50% of predictors are quantitative	Linear	0.55	100	100	100	100	0	100	0	100	100	0	0	100	100	0	100	0	0							
		0.632	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0						
		0.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0						
	Quadratic	0.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0						
		0.55	0	100	0	0	100	100	0	0	0	0	0	100	100	0	0	0	0	0						
		0.632	100	0	100	0	100	0	0	0	0	100	0	0	0	0	0	100	0	0						
75% of predictors are quantitative	Linear	0.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0							
		0.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0						
		0.55	0	100	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0						
	Quadratic	0.632	100	0	100	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0						
		0.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0						
		0.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0						
Quantitative	Linear	0.55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0							
		0.632	100	100	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0						
		0.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0						
	Quadratic	0.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0						
		0.55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0						
		0.632	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0						

predictors, for the same number of observations, the higher the number of predictors, the smaller the minimum size of the terminal nodes. However, it should be noted that the number of predictors is not always inversely proportional to the minimum size of terminal nodes. In some scenarios, the minimum size of terminal nodes does not vary with the number of predictors. In contrast, in others, the best performance is obtained with small minimum sizes of terminal nodes, even when there are few predictors. Thus, when there are a few observations and many predictors, no trend is predictable, and the parameter will depend only on the complexity of the relationship between the data F .

3.1.3. Impact of Overdispersion on the RF Predictive Performance for Different Types of Predictors. There was no difference among the types of covariates considered to be part of this simulation study. Using the best combination of parameters, RF performed similarly for all types of covariates regardless of the level of dispersion in the response variable. However, the performance of the RF for the sample size $n = 50$ varied enormously for different levels of dispersion, whereas the values were stable for large sample sizes (in our case $n = 1250$). RF performance was impacted by the magnitude of overdispersion, but relatively more accurate results were obtained when the population size was large in comparison to a small population size, where the RMSE varied enormously for different levels of dispersion (Figures 5(a) and 5(b)).

3.1.4. Assessment of the RF Performance Compared to Traditional Approaches for Count Data Prediction. In this section, we analyse the performance of RF regression compared to classical log-linear, Poisson, quasi-Poisson, negative binomial, and zero-inflated Poisson models for different sample sizes and overdispersion degrees. Table 5 and Figure 5 show that linear regression applied to the log-transformed response variable has the lowest performance. However, in most scenarios, the biases and RMSE obtained using the RF approach to predict the results of the overdispersed count are not statistically different from the biases and RMSE obtained using traditional Poisson family methods. Furthermore, no statistical difference was observed when the variance-mean relationship was quadratic, with some exceptions for low overdispersion. However, for the linear variance-to-mean relationship, the mean RMSE of RF is higher than that of GLMs, i.e., less accurate with higher variability. As a result, while predicting overdispersed outcomes, the most accurate and precise method will have to be identified case by case, depending on the available data. In some situations, RF will be the best performing method, while in others, it will not. Figures 6(a) and 6(b) present RF compared to the biases of GLM family models and RMSE regression lines as a function of dispersion and indicate the 95% confidence interval (range), represented by the grey zone in which the biases and RMSE would be if the experiment were repeated. Globally, the mean of the RMSE and the biases of RF tend to increase as dispersion increases. Consequently, the slope of the regression of the RMSE in the

function of dispersion and the range of RMSE values are lower for outcomes with linear variance-to-mean relationships than for those with quadratic variance-to-mean relationships.

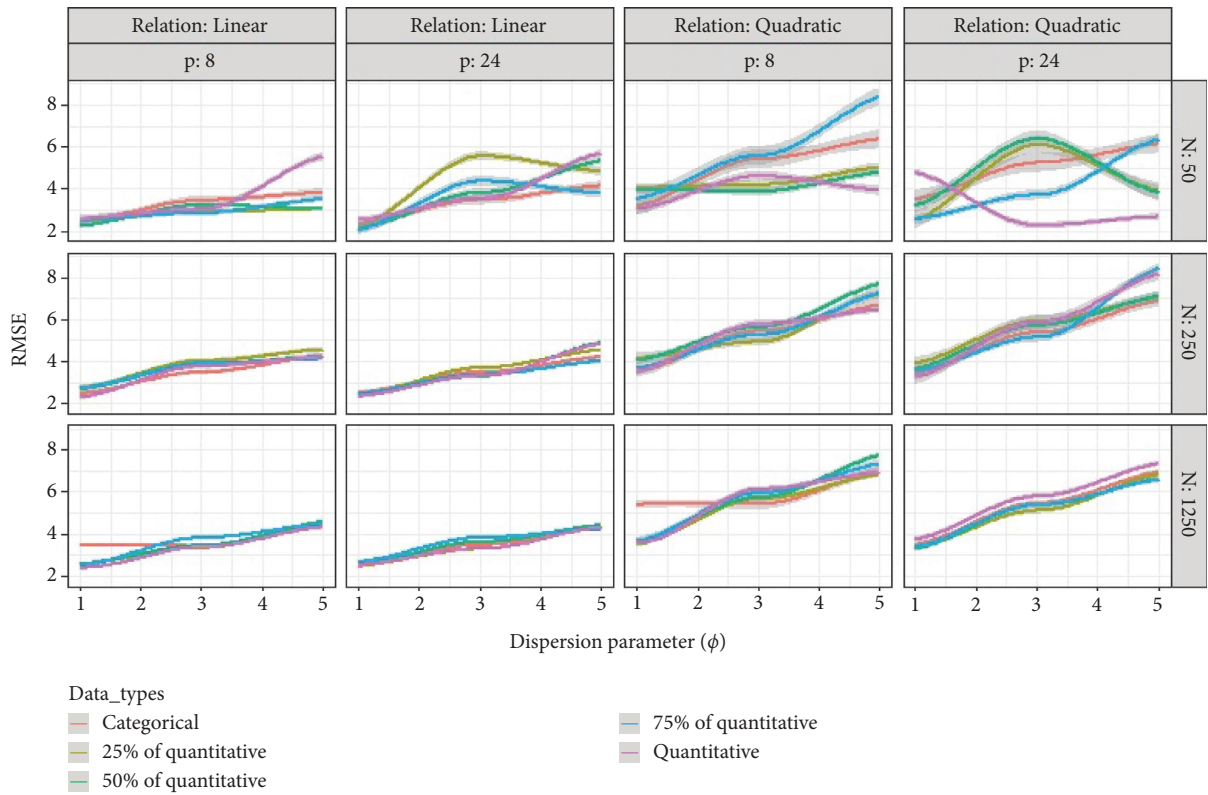
3.1.5. Real Dataset Results. RF yielded the lowest mean of the absolute values of biases (1.51 ± 1.16) compared to linear regression applied to the logarithmic transformation of tick burden (8.24 ± 1.43), negative binomial (1.56 ± 1.19), quasi-Poisson (1.52 ± 1.17), and Poisson (1.52 ± 1.17) regressions for the dataset with overdispersion caused by outliers (Figure 7(a)). However, the difference between the mean absolute values of biases is not significant. Quasi-Poisson and Poisson (7.72 ± 1.32) have the lowest RMSE values, followed by the RF (7.75 ± 1.34) and negative binomial (7.82 ± 1.29) but are not significantly different according to the Tukey HSD test (Figures 7(a) and 7(b)).

For the NMES1988 dataset with overdispersion caused by outliers and zero inflation, quasi-Poisson, Poisson, zero-inflated Poisson, and negative binomial produced the lowest absolute value of biases (0.18 ± 0.14), followed by RF (0.19 ± 0.14). Furthermore, the zero-inflated Poisson model (6.56 ± 0.39) yielded the lowest RMSE, followed by quasi-Poisson (6.57 ± 0.39), negative binomial (6.57 ± 0.39), and RF (6.59 ± 0.4). As for the tick burden dataset, no statistically significant difference was observed between RF and traditional Poisson family models (Figures 7(c) and 7(d)).

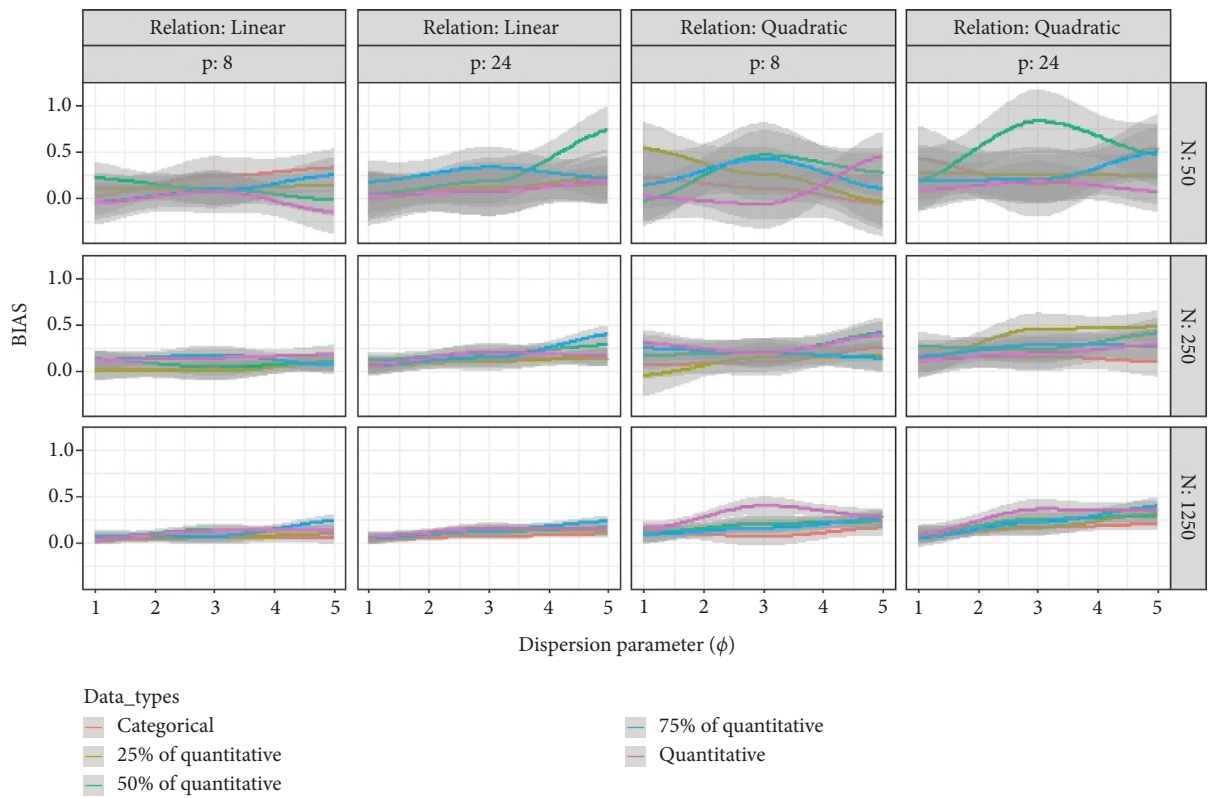
3.2. Discussion

3.2.1. Impact of Data Features on RF Regression. One of the characteristics of many databases that practitioners consider is the complex relationship among data, which typically yields enormous noise. In some situations, this complexity may directly or indirectly result from variables at the hand and affect many statistical procedures. Computer-based statistical methods are more appropriate than traditional statistical models for effectively solving this complex problem and dealing with large-scale issues. However, despite an increasing interest and practical use, computer-based methods have rarely been evaluated [18]. There is a growing trend of replacing traditional statistical models with machine learning in several research areas to improve predictive accuracy. In addition, these techniques are now used to model data that do not fit into the “small- n -large- p ” category.

The RF method works well for all covariates (categorical, continuous, or their mixture) in regression with or without overdispersion in the response variable. No original predictor transformation is required for the predictive task when encoding categorical covariates. Furthermore, RF accuracy is not significantly affected by covariate types. Singh et al. [25] found similar results in classification, suggesting that RF performs equally well with numeric-only or text-only datasets. However, fluctuations in RF parameter settings were observed for exclusively categorical predictors when the number of observations or the relation (ratio) between the number of observations and the number of predictors was low.



(a)



(b)

FIGURE 5: Impact of the response variable overdispersion on RF accuracy (relation is the outcome’s variance-to-mean relationship). (a) RMSE and (b) BIAS.

TABLE 5: Description of methods' predictive performance and Tukey's HSD pairwise comparisons for different levels of overdispersion.

Variance-to-mean relationship	Methods	$\phi = 1$	$\phi = 2$	$\phi = 3$	$\phi = 4$	$\phi = 5$	$\phi = 6$	$\phi = 7$	$\phi = 8$	$\phi = 9$	$\phi = 10$	
		BIAS										
Linear	Log_linear	1.53 ± 0.18b	1.62 ± 0.23c	1.65 ± 0.24c	1.58 ± 0.25b	1.58 ± 0.26 d	1.58 ± 0.26 d	1.56 ± 0.29c	1.56 ± 0.29c	1.70 ± 0.34c	1.67 ± 0.32c	
	Neg.Bin	0.12 ± 0.10a	0.17 ± 0.15ab	0.18 ± 0.13a	0.23 ± 0.17a	0.26 ± 0.22a	0.26 ± 0.22a	0.26 ± 0.22a	0.28 ± 0.22a	0.53 ± 0.44a	0.34 ± 0.26a	
	Poisson	0.12 ± 0.09a	0.15 ± 0.11a	0.18 ± 0.13a	0.20 ± 0.16a	0.22 ± 0.17b	0.22 ± 0.17b	0.22 ± 0.17b	0.22 ± 0.17b	0.27 ± 0.20b	0.27 ± 0.21b	
	Quasi-Poisson	0.12 ± 0.09a	0.15 ± 0.11a	0.18 ± 0.13a	0.20 ± 0.16a	0.22 ± 0.17b	0.22 ± 0.17b	0.22 ± 0.17b	0.22 ± 0.17b	0.27 ± 0.20b	0.27 ± 0.21b	
	RF	0.14 ± 0.11a	0.18 ± 0.13b	0.23 ± 0.16b	0.23 ± 0.18a	0.26 ± 0.20ac	0.26 ± 0.20ac	0.26 ± 0.20ac	0.25 ± 0.19ab	0.29 ± 0.22b	0.30 ± 0.23ab	
	ZIP	0.12 ± 0.09a	0.16 ± 0.12ab	0.19 ± 0.14a	0.21 ± 0.16a	0.23 ± 0.17bc	0.23 ± 0.17bc	0.23 ± 0.17bc	0.24 ± 0.18b	0.27 ± 0.21b	0.27 ± 0.22b	
	Log_linear	1.65 ± 0.28c	1.67 ± 0.57c	1.67 ± 0.32b	1.99 ± 0.66b	1.90 ± 0.58b	1.90 ± 0.58b	1.90 ± 0.58b	1.93 ± 0.66b	1.46 ± 0.50b	2.06 ± 0.72b	
	Neg.Bin	0.23 ± 0.17a	0.41 ± 0.43ab	0.34 ± 0.27a	0.61 ± 0.50a	0.59 ± 0.46a	0.59 ± 0.46a	0.59 ± 0.46a	0.59 ± 0.49a	0.40 ± 0.33a	0.71 ± 0.53a	
	Poisson	0.24 ± 0.18a	0.41 ± 0.42ab	0.34 ± 0.26a	0.63 ± 0.54a	0.60 ± 0.47a	0.60 ± 0.47a	0.60 ± 0.47a	0.57 ± 0.49a	0.38 ± 0.32a	0.73 ± 0.57a	
	Quasi-Poisson	0.24 ± 0.18a	0.41 ± 0.42ab	0.34 ± 0.26a	0.63 ± 0.54a	0.60 ± 0.47a	0.60 ± 0.47a	0.60 ± 0.47a	0.57 ± 0.49a	0.38 ± 0.32a	0.73 ± 0.57a	
Quadratic	RF	0.30 ± 0.22b	0.49 ± 0.51a	0.38 ± 0.28a	0.66 ± 0.50a	0.65 ± 0.49a	0.65 ± 0.49a	0.65 ± 0.49a	0.63 ± 0.51a	0.43 ± 0.35a	0.80 ± 0.61a	
	ZIP	0.24 ± 0.18a	0.40 ± 0.44b	0.33 ± 0.25a	0.60 ± 0.50a	0.58 ± 0.46a	0.58 ± 0.46a	0.58 ± 0.46a	0.56 ± 0.48a	0.39 ± 0.32a	0.71 ± 0.53a	
	Log_linear	3.98 ± 0.56c	4.68 ± 0.89 d	5.19 ± 0.93 d	4.99 ± 0.87c	4.97 ± 0.75 d	5.23 ± 0.82 d	5.23 ± 0.82 d	5.23 ± 1 d	5.54 ± 0.95c	5.95 ± 1.14 d	
	Neg.Bin	2.15 ± 0.2a	2.94 ± 0.67a	3.14 ± 0.3a	3.71 ± 0.51a	3.81 ± 0.47ab	4.18 ± 0.64a	4.18 ± 0.64a	4.26 ± 0.56a	6.02 ± 2.27a	5.29 ± 0.93a	
Linear	Poisson	2.13 ± 0.18a	2.76 ± 0.27b	3.14 ± 0.29a	3.61 ± 0.49a	3.75 ± 0.45a	4.03 ± 0.54b	4.03 ± 0.54b	4.09 ± 0.58b	4.68 ± 0.73b	4.98 ± 0.87b	
	Quasi-Poisson	2.13 ± 0.18a	2.76 ± 0.27b	3.14 ± 0.29a	3.61 ± 0.49a	3.75 ± 0.45a	4.03 ± 0.54b	4.03 ± 0.54b	4.09 ± 0.58b	4.68 ± 0.73b	4.98 ± 0.87b	
	RF	2.42 ± 0.31b	3.24 ± 0.55c	3.88 ± 0.61b	3.86 ± 0.66b	4.03 ± 0.49c	4.33 ± 0.61c	4.33 ± 0.61c	4.43 ± 0.78c	4.86 ± 0.76b	5.19 ± 0.93ac	
	ZIP	2.14 ± 0.21a	2.9 ± 0.38a	3.39 ± 0.51c	3.69 ± 0.61a	3.87 ± 0.49b	4.11 ± 0.58ab	4.11 ± 0.58ab	4.25 ± 0.74a	4.75 ± 0.77b	5.04 ± 0.93bc	
	Log_linear	5.93 ± 1.45c	8.13 ± 6.74a	6.54 ± 1.29c	8.31 ± 2.71a	10.45 ± 4.29a	10.46 ± 4.41a	10.46 ± 4.41a	10.4 ± 5.3a	6.96 ± 2.51a	11.96 ± 6.01a	
	Neg.Bin	4.44 ± 1a	7.19 ± 5.85a	5.93 ± 1.03ab	8.05 ± 2.38a	10.42 ± 4.09a	10.23 ± 3.91a	10.23 ± 3.91a	10.32 ± 4.81a	6.85 ± 2.16a	11.74 ± 5.25a	
	Poisson	4.52 ± 0.98a	7.12 ± 5.37a	5.88 ± 1.07ab	8.1 ± 2.4a	10.59 ± 4.14a	10.26 ± 3.95a	10.26 ± 3.95a	10.15 ± 4.67a	6.82 ± 2.27a	11.89 ± 5.29a	
	Quasi-Poisson	4.52 ± 0.98a	7.12 ± 5.37a	5.88 ± 1.07ab	8.1 ± 2.4a	10.59 ± 4.14a	10.26 ± 3.95a	10.26 ± 3.95a	10.15 ± 4.67a	6.82 ± 2.27a	11.89 ± 5.29a	
	RF	5.1 ± 1.11b	7.68 ± 6.11a	6.06 ± 1.03a	8.15 ± 2.35a	10.62 ± 3.88a	10.49 ± 4.01a	10.49 ± 4.01a	10.46 ± 4.62a	7.03 ± 2.4a	12.08 ± 5.41a	
	ZIP	4.57 ± 1.08a	7.11 ± 5.71a	5.81 ± 1.12b	7.99 ± 2.4a	10.38 ± 4.01a	10.14 ± 3.97a	10.14 ± 3.97a	10.08 ± 4.71a	6.81 ± 2.24a	11.74 ± 5.34a	
Quadratic	Log_linear	3.98 ± 0.56c	4.68 ± 0.89 d	5.19 ± 0.93 d	4.99 ± 0.87c	4.97 ± 0.75 d	5.23 ± 0.82 d	5.23 ± 0.82 d	5.23 ± 1 d	5.54 ± 0.95c	5.95 ± 1.14 d	
	Neg.Bin	2.15 ± 0.2a	2.94 ± 0.67a	3.14 ± 0.3a	3.71 ± 0.51a	3.81 ± 0.47ab	4.18 ± 0.64a	4.18 ± 0.64a	4.26 ± 0.56a	6.02 ± 2.27a	5.29 ± 0.93a	
RMSE												
Linear	Poisson	2.13 ± 0.18a	2.76 ± 0.27b	3.14 ± 0.29a	3.61 ± 0.49a	3.75 ± 0.45a	4.03 ± 0.54b	4.03 ± 0.54b	4.09 ± 0.58b	4.68 ± 0.73b	4.98 ± 0.87b	
	Quasi-Poisson	2.13 ± 0.18a	2.76 ± 0.27b	3.14 ± 0.29a	3.61 ± 0.49a	3.75 ± 0.45a	4.03 ± 0.54b	4.03 ± 0.54b	4.09 ± 0.58b	4.68 ± 0.73b	4.98 ± 0.87b	
	RF	2.42 ± 0.31b	3.24 ± 0.55c	3.88 ± 0.61b	3.86 ± 0.66b	4.03 ± 0.49c	4.33 ± 0.61c	4.33 ± 0.61c	4.43 ± 0.78c	4.86 ± 0.76b	5.19 ± 0.93ac	
	ZIP	2.14 ± 0.21a	2.9 ± 0.38a	3.39 ± 0.51c	3.69 ± 0.61a	3.87 ± 0.49b	4.11 ± 0.58ab	4.11 ± 0.58ab	4.25 ± 0.74a	4.75 ± 0.77b	5.04 ± 0.93bc	
	Log_linear	5.93 ± 1.45c	8.13 ± 6.74a	6.54 ± 1.29c	8.31 ± 2.71a	10.45 ± 4.29a	10.46 ± 4.41a	10.46 ± 4.41a	10.4 ± 5.3a	6.96 ± 2.51a	11.96 ± 6.01a	
	Neg.Bin	4.44 ± 1a	7.19 ± 5.85a	5.93 ± 1.03ab	8.05 ± 2.38a	10.42 ± 4.09a	10.23 ± 3.91a	10.23 ± 3.91a	10.32 ± 4.81a	6.85 ± 2.16a	11.74 ± 5.25a	
	Poisson	4.52 ± 0.98a	7.12 ± 5.37a	5.88 ± 1.07ab	8.1 ± 2.4a	10.59 ± 4.14a	10.26 ± 3.95a	10.26 ± 3.95a	10.15 ± 4.67a	6.82 ± 2.27a	11.89 ± 5.29a	
	Quasi-Poisson	4.52 ± 0.98a	7.12 ± 5.37a	5.88 ± 1.07ab	8.1 ± 2.4a	10.59 ± 4.14a	10.26 ± 3.95a	10.26 ± 3.95a	10.15 ± 4.67a	6.82 ± 2.27a	11.89 ± 5.29a	
	RF	5.1 ± 1.11b	7.68 ± 6.11a	6.06 ± 1.03a	8.15 ± 2.35a	10.62 ± 3.88a	10.49 ± 4.01a	10.49 ± 4.01a	10.46 ± 4.62a	7.03 ± 2.4a	12.08 ± 5.41a	
	ZIP	4.57 ± 1.08a	7.11 ± 5.71a	5.81 ± 1.12b	7.99 ± 2.4a	10.38 ± 4.01a	10.14 ± 3.97a	10.14 ± 3.97a	10.08 ± 4.71a	6.81 ± 2.24a	11.74 ± 5.34a	
Quadratic	Log_linear	3.98 ± 0.56c	4.68 ± 0.89 d	5.19 ± 0.93 d	4.99 ± 0.87c	4.97 ± 0.75 d	5.23 ± 0.82 d	5.23 ± 0.82 d	5.23 ± 1 d	5.54 ± 0.95c	5.95 ± 1.14 d	
	Neg.Bin	2.15 ± 0.2a	2.94 ± 0.67a	3.14 ± 0.3a	3.71 ± 0.51a	3.81 ± 0.47ab	4.18 ± 0.64a	4.18 ± 0.64a	4.26 ± 0.56a	6.02 ± 2.27a	5.29 ± 0.93a	

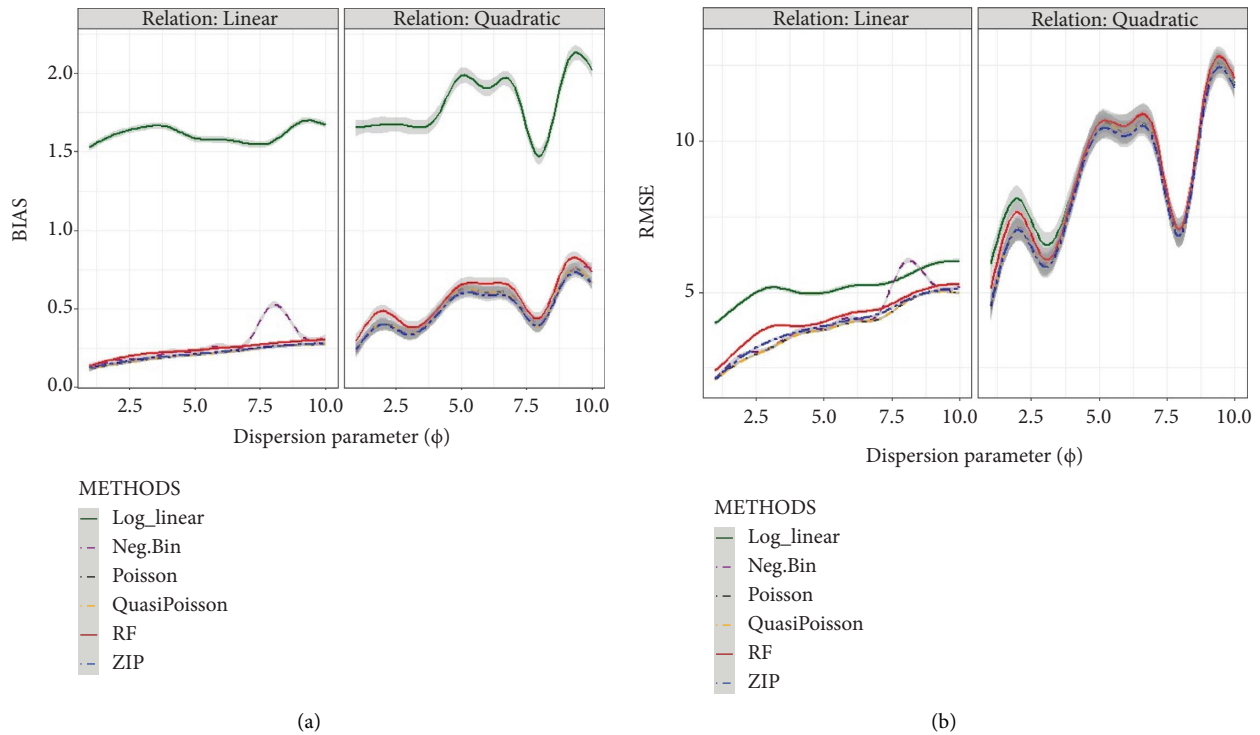


FIGURE 6: Effect of overdispersion on RF predictive performance compared to GLM models (relation is the outcome’s variance-to-mean relationship). (a) |Bias| and (b) RMSE.

Frequently, empirical studies on classification find that the default RF’s parameter values yield good prediction accuracy [19, 49]. However, among all possible combinations, in most scenarios, the best RF performance was not observed with the default parameters in Breiman and Cutler’s random forests for the classification and regression package (randomForest), widely used in R software [44]. For regression tasks, the number of variables to select in each split should equal the number of predictors divided by 3. Our results support the findings of Han and Kim [50], which imply that the default $mtry$ of $p/3$ cannot guarantee the highest level of accuracy in regression. Additionally, Strobl et al. [51] argued that various $mtry$ values should be taken into account.

For all types of covariates, RF performed better when a small number of variables were selected at each split ($mtry = 2$), regardless of the response variable overdispersion, while for some scenarios, $mtry = (p/3)$ was too small. RF prediction accuracy is expected to be higher for smaller values of $mtry$ in general, according to Strobl et al. [51]. Smaller $mtry$ values are preferred if the dataset’s features have similar relevance. If not, greater $mtry$ values are preferred [50]. Furthermore, a small number of predictors in each split can better utilize the available information in many informative predictors with different strengths [52].

However, for a variable to have a chance to appear in a sufficient number of trees, a large number of preselected predictors are required when there are many predictor variables. Therefore, the average variable importance measure will be accurate and not just a random fluctuation based

on enough trials [51]. However, Svetnik et al. [49] claimed that RF performance changes very little over a wide range of $mtry$ values, except for extremes ($mtry = 1$ and $mtry = p$).

RF was relatively insensitive to changes in the minimum number of observations in each terminal node. The minimum node size was adjusted to speed up the computation for large datasets. Our findings suggest that the minimum node size behaviour is influenced by the relationship between the total sample size and the number of predictors. The best RF performance was obtained with many observations in each end node for large n and small p (between seven and nine). When dealing with situations with many observations and a few variables, increasing the terminal node size with the sample size improved the convergence rate of the prediction accuracy [53]. There was no discernible pattern for small n and large p . Although the minimum node size appears to have a minor impact on predictive accuracy, it is necessary to consider this parameter when tuning RF. When determining the ideal combination of the best parameters of RF, the prevalence of specific values can indicate the stability of estimates, avoiding the effect of random fluctuations. According to Lin and Jeon [53], growing large trees, that is, using small observations in each terminal node, generally results in the highest performance for “small- n -large- p ” situations.

We observed that sample size behaviour is also influenced by the relationship between the number of observations and predictors. When the ratio of the number of observations to the number of predictors is high, RF produces the best results with fewer observations (55 percent or 63.2 percent of the total sample size). Buja and Stuetzle [54]

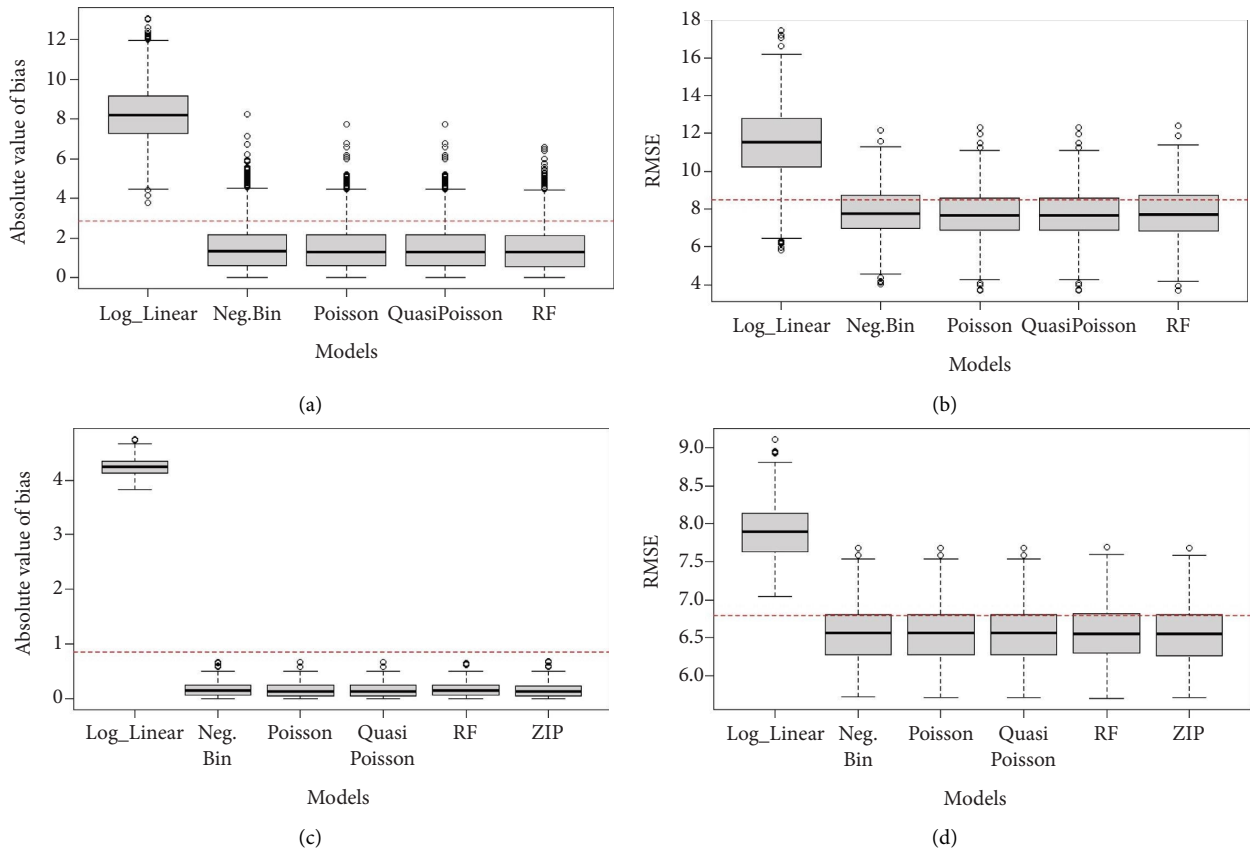


FIGURE 7: Predictive performance of RF compared to GLMs for tick burden and NMES1988 datasets. (a) Tick burden dataset: BIAS, (b) tick burden dataset: RMSE, (c) NMES1988 dataset: BIAS, and (d) NMES1988 dataset: RMSE.

argued that half-sampling might be more suited because bootstrap sampling shares several theoretical traits. Wösthoff [55] claimed that the subsample size is not a critical parameter as the results of sizes 0.5 and the default value (0.632) are nearly identical. For permutation importance estimation in classification, significant predictor variables may not be recognized if the sample size is too small. The variability of the importance measure tends to increase if the subsample size is so large that the out-of-bag sample becomes very small [55].

Considering overdispersion, we realized that, for small sizes (less than 250), the best m try value was randomly selected among possible values. In contrast, for large population sizes, this value was maintained constant (m try = 2) for all types of covariates. According to Svetnik et al. [49], this parameter is the main RF tuning parameter. Furthermore, the sample size influenced the RF accuracy level. For small sample sizes (less than 250), the effect of overdispersion resulted in an erratic trend, challenging to interpret. In contrast, the trend was linear for large population sizes, illustrating the overdispersion effect.

Although RF parameters seemed insensitive to the degree of overdispersion of the response variable, RF accuracy was affected. Thus, despite being an essential step in RF analyses, parameter tuning did not affect the best parameter selection when the response variable was overdispersed. Some authors have recommended data transformation for

RF analyses, especially population estimation. Transforming the response variable allows the RF algorithm to find good splits in data as the relationship between distance-based covariates and the population is, in most cases, more uniformly distributed [56]. Transformation methods may be applied to observed response variables before RF model fitting for higher prediction accuracy. However, O'Hara and Kotze [57] stated that we should not transform count data because, in addition to performing poorly, they can lead to impossible predictions.

3.2.2. RF Performance Compared to Poisson Model Extensions. According to Anderson et al. [33], one of the most critical decisions in the population modelling process is understanding and selecting the appropriate model structure based on the available data. RF regression is an efficient and credible alternative in prediction tasks, as reported by previous studies [12]. The results obtained in this study suggest that RF is as accurate as Poisson, zero-inflated Poisson, and negative binomial regressions when dealing with overdispersed count data. The tick burden dataset confirmed this statement. Novel methods such as RF are more efficient for predicting spatial patterns in species distribution models than more traditional linear models [58]. Furthermore, they consistently outperform more established methods. RF's consistency and predictive performance could be an additional point to explore and

reorient the debate on modelling count data. Some researchers suggest using the Poisson model and its extensions [57]. Others recommend using the linear model when the objective is to test the significance of estimates [41]. On the contrary, others think there is no best solution because it depends on the database in the presence [42]. However, these models have shown similar performance for different levels of overdispersion. Considering GLMs, RF is robust to misspecification of the relationship between the mean and variance [57]. Anderson et al. [33] claimed that traditional regression-based models may not be the most effective model, depending on the available census data (for example, presence of sampling bias, missing data, outliers, and imperfect detection). Given the predictive accuracy demonstrated by RF, its flexibility in dealing with covariates of various natures, and other recognised advantages such as the ability to deal with complex interaction structures, highly correlated variables and the measurement of variable significance, RF is an effective alternative method for predicting count outcomes [52]. RF may be a tool of capital importance for decision and policymakers working in limited data areas or with unreliable census data, especially in developing countries. Furthermore, producing accurate and reliable estimates is a problem of interest for resource allocation, public health, outbreak early detection, food security, conservation policy, environmental planning, and others [33, 56, 59].

4. Conclusion

RF has become a popular analysis tool in many application fields and will probably remain relevant in the future due to its high flexibility and many advantages. This study evaluated the impact of different characteristics of the dataset and the magnitude of overdispersion on the settings of RF parameters and predictive performance through simulated and real-life datasets. We found that the RF algorithm performs relatively well for all types of predictors. Data features do not influence the setting of RF parameters and, consequently, its predictive accuracy. However, the ratio of observations to the number of predictors influences the stability of the optimal RF parameters for a given dataset. Although the degree of overdispersion of the response variable did not significantly influence RF predictive validity, the magnitude of dispersion influenced RF predictive accuracy. The RF algorithm performed almost equally well as standard count data models such as quasi-Poisson, zero-inflated Poisson, and negative binomial models for the outcomes with linear and quadratic variance-to-mean relationships. However, for some scenarios, RF sometimes had a higher bias or RMSE, sometimes a lower bias or RMSE compared to Poisson family models. Furthermore, for the same scenario, depending on the selected subsample, a model may appear to perform better when it does not perform well overall. Thus, the best model choice should be based on data. RF is a reliable method when the assumptions for GLMs are not satisfied. Therefore, despite the sensitivity of RF to the overdispersion of the response variable, it is an efficient method for predicting count outcomes. Thereby, it is a tool of capital importance for decisions and policymakers

working in limited data areas. Therefore, research should focus on decreasing the sensitivity of predictive accuracy to overdispersion.

Data Availability

The data supporting this study are from previously reported studies and datasets, which have been cited. The simulated data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this study.

Acknowledgments

This work was supported by the Regional Universities Forum for Capacity Building in Agriculture through the Graduate Training Assistantship Program supported by the Carnegie Corporation in New York (Grant no. RU/2020/GTA/DRG/019), the DAAD In-Country/In-Region Scholarship Programme FSA/UAC (Grant no. 57546598, 2020), and the International Development Research Centre (IDRC) and the Swedish International Development Cooperation Agency (SIDA) through the Artificial Intelligence for Development (AI4D) Africa Programme, managed by the Africa Center for Technology Studies (ACTS) (Grant no. ACTS/AI4D_2022/109651/006).

References

- [1] E. Altun, "A new model for over-dispersed count data: Poisson quasi-lindley regression model," *Mathematical Sciences*, vol. 13, pp. 241–247, 2019.
- [2] J. Li, A. D. Heap, A. Potter, Z. Huang, and J. J. Daniell, "Can we improve the spatial predictions of seabed sediments? a case study of spatial interpolation of mud content across the southwest australian margin," *Continental Shelf Research*, vol. 31, no. 13, pp. 1365–1376, 2011.
- [3] R. W. M. Wedderburn, "Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method," *Biometrika*, vol. 61, no. 3, pp. 439–447, 1974.
- [4] P. C. Consul, *Generalized Poisson Distributions: Properties and Applications. Statistics, Textbooks and Monographs*, Marcel Dekker, New York, NY, USA, 1989.
- [5] A. L. Bailey, "Credibility procedures: laplace's generalization of bayes' rule and the combination of collateral knowledge with observed data," *Proceedings of the Casualty Actuarial Society*, vol. 37, pp. 7–23, 1950.
- [6] M. Greenwood and G. U. Yule, "An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents," *Journal of the Royal Statistical Society*, vol. 83, no. 2, pp. 255–279, 1920.
- [7] R. Keffer, "An experience rating formula," *Transactions of the Actuarial Society of America*, vol. 30, pp. 130–139, 1929.
- [8] D. Lambert, "Zero-inflated Poisson regression, with an application to defects in manufacturing," *Technometrics*, vol. 34, no. 1, pp. 1–14, 1992.

- [9] M. E. Haque, T. S. Mallick, and W. Bari, "Zero truncated Poisson model: an alternative approach for analyzing count data with excess zeros," *Journal of Statistical Computation and Simulation*, vol. 92, no. 3, pp. 476–487, 2022.
- [10] A. Araldi, "Towards an integrated methodology for model and variable selection using count data: an application to micro-retail distribution in urban studies," *Urban Science*, vol. 4, no. 2, p. 21, 2020.
- [11] Domino, *Analyzing Large P Small N Data-Examples from Microbiome*, Public Library of Science, San Francisco, CA, USA, 2020.
- [12] G. Nicolas, T. P. Robinson, G. R. W. Wint, G. Conchedda, G. Cinardi, and M. Gilbert, "Using random forest to improve the downscaling of global livestock census data," *PLoS One*, vol. 11, no. 3, pp. 01504244–e150516, 2016.
- [13] T. Rahman, H.-E. Huang, A.-S. Tai, W. P. Hsieh, and G. Tseng, "A sparse negative binomial classifier with covariate adjustment for rna-seq data," 2019, <https://www.biorxiv.org/content/10.1101/636340v1>.
- [14] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [15] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, Chapman and Hall/CRC Press, London, UK, 1984.
- [16] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.
- [17] K. Arun and C. Langmead, "Structure-based chemical shift prediction using random non-linear regression," in *Proceedings of the 4th Asia-Pacific Bioinformatics Conference*, pp. 317–326, Taipei, Taiwan, February 2006.
- [18] G. Biau, "Analysis of a random forests model," *Journal of Machine Learning Research*, vol. 13, pp. 1063–1095, 2012.
- [19] R. Díaz-Urriarte and S. Alvarez de Andrés, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, no. 1, p. 3, 2006.
- [20] X. Han, B. Yang, and S. Lee, "Application of random forest algorithm in machine fault diagnosis," *Engineering Asset Management*, Springer, Berlin, Germany, 2006.
- [21] S. Janitzka, E. Celik, and A. L. Boulesteix, "A computationally fast variable importance test for random forests for high-dimensional data," *Advances in Data Analysis and Classification*, vol. 12, no. 4, pp. 885–915, 2018.
- [22] H. Pang, A. Lin, M. Holford et al., "Pathway analysis using random forests classification and regression," *Bioinformatics*, vol. 22, no. 16, pp. 2028–2036, 2006.
- [23] B. Thompson, *A Limitation of Random Forest Regression*, 2019.
- [24] J. M. Hilbe, *Negative Binomial Regression: Modeling*, Cambridge University Press, Cambridge, UK, 2011.
- [25] A. Singh, M. N. Halgamuge, and R. Lakshmanan, "Impact of different data types on classifier performance of random forest, naïve bayes, and k-nearest neighbors algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 12, 2017.
- [26] O. Okun and H. Priisalu, "Random forest for gene expression based cancer classification: overlooked issues," in *Pattern Recognition and Image Analysis. IbPRIA 2007*, J. Martí, J. Benedí, A. M. Mendonça, and J. Serrat, Eds., pp. 483–490, Springer, Berlin, Germany, 2007.
- [27] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for rna-seq data with deseq2," *Genome Biology*, vol. 15, no. 12, p. 550, 2014.
- [28] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "Edger: a bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.
- [29] M. D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of rna-seq data," *Genome Biology*, vol. 11, no. 3, p. 25, 2010.
- [30] E. Scornet, G. Biau, and J. P. Vert, "Consistency of random forests," *Annals of Statistics*, vol. 43, no. 4, pp. 1716–1741, 2015.
- [31] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016.
- [32] P. Probst and A.-L. Boulesteix, "To tune or not to tune the number of trees in random forest?," 2017, <https://arxiv.org/abs/1705.05654>.
- [33] W. Anderson, S. Guikema, B. Zaitchik, and W. Pan, "Methods for estimating population density in data-limited areas: evaluating regression and tree-based models in Peru," *PLoS One*, vol. 9, no. 7, pp. 1–15, 2014.
- [34] P. S. Dewi, Y. S. Dewi, and L. Amaliana, "Zero inflated poisson and geographically weighted zero-inflated Poisson regression model: application to elephantiasis (filariasis) counts data," *Journal of Mathematics and Statistics*, vol. 11, no. 2, pp. 52–60, 2015.
- [35] A. Cameron and K. Trivedi, *Micro Econometrics: Methods and Applications*, Cambridge University Press, Cambridge, UK, 2005.
- [36] Y. Mouatassim and E. H. Ezzahid, "Poisson regression and zero-inflated poisson regression: application to private health insurance data," *European Actuarial Journal*, vol. 2, pp. 187–204, 2012.
- [37] R. Myers, D. Montgomery, G. Vining, and T. Robinson, "Generalized linear models: with applications in engineering and the sciences," *Wiley Series in Probability and Statistics*, Wiley, New York, NY, USA, 2010.
- [38] B. Lokonon and R. Glèlè Kakai, "Effect of overdispersion and sample size on the performance of Poisson model and its extensions in frame of generalized linear models (glms)," *International Journal of Applied Mathematics and Statistics*, vol. 57, pp. 40–53, 2018.
- [39] F. D. Schönbrodt and M. Perugini, "At what sample size do correlations stabilize?" *Journal of Research in Personality*, vol. 47, no. 5, pp. 609–612, 2013.
- [40] A. Gut, "The multivariate normal distribution," *An Intermediate Course in Probability*, vol. 2, pp. 117–145, 2009.
- [41] A. R. Ives, "For testing the significance of regression coefficients, go ahead and log-transform count data," *Methods in Ecology and Evolution*, vol. 6, no. 7, pp. 828–835, 2015.
- [42] D. I. Warton, M. Lyons, J. Stoklosa, and A. R. Ives, "Three points to consider when choosing alm or glm test for count data," *Methods in Ecology and Evolution*, vol. 7, no. 8, pp. 882–890, 2016.
- [43] R. M. Yang, G. L. Zhang, F. Liu et al., "Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem," *Ecological Indicators*, vol. 60, pp. 870–878, 2016.
- [44] A. Liaw and M. Wiener, "Classification and regression by random-forest," *R News*, vol. 2, pp. 18–22, 2002.
- [45] B. A. Walther and J. L. Moore, "The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance," *Ecography*, vol. 28, no. 6, pp. 815–829, 2005.
- [46] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020.

- [47] B. Dirk, *Advanced Epidemiology Course, the Institute of Tropical Medicine (Itm)-Antwerpen*, 2017.
- [48] A. Cameron and P. K. Trivedi, "Regression-based tests for overdispersion in the Poisson model," *Journal of Econometrics*, vol. 46, no. 3, pp. 347–364, 1990.
- [49] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: a classification and regression tool for compound classification and qsar modeling," *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 6, pp. 1947–1958, 2003.
- [50] S. Han and H. Kim, "Optimal feature set size in random forest regression," *Applied Sciences*, vol. 11, no. 8, p. 3428, 2021.
- [51] C. Strobl, A. L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, no. 1, p. 307, 2008.
- [52] A. L. Boulesteix, S. Janitza, J. Kruppa, and I. R. König, "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 6, pp. 493–507, 2012.
- [53] Y. Lin and Y. Jeon, "Random forests and adaptive nearest neighbors," *Journal of the American Statistical Association*, vol. 101, no. 474, pp. 578–590, 2006.
- [54] A. Buja and W. Stuetzle, "Observations on bagging," *Statistica Sinica*, vol. 16, no. 2, pp. 323–351, 2006.
- [55] J. Wösthoff, *Moderne Klassifikationsverfahren in der Biometrie-Einfluss der Stichprobengröße Beim Resampling in Random Forests*, Ludwig Maximilians-Universität München, Munich, Germany, 2008.
- [56] F. R. Stevens, A. E. Gaughan, C. Linard, and A. J. Tatem, "Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data," *PLoS One*, vol. 10, no. 2, pp. 01070422–e107122, 2015.
- [57] R. B. O'Hara and D. J. Kotze, "Do not log-transform count data," *Methods in Ecology and Evolution*, vol. 1, no. 2, pp. 118–122, 2010.
- [58] J. Elith, C. H. Graham, R. P. Anderson et al., "Novel methods improve prediction of species' distributions from occurrence data," *Ecography*, vol. 29, no. 2, pp. 129–151, 2006.
- [59] M. Salvatore, F. Pozzi, E. Ataman, B. Huddleston, and M. Bloise, *Mapping Global Urban and Rural Population Distributions*, Food and Agriculture Organization of the United Nations, Rome, Italy, 2005.