International Journal of Data Science and Analysis
2022; 8(6): 169-181
http://www.sciencepublishinggroup.com/j/ijdsa
doi: 10.11648/j.ijdsa.20220806.11
ISSN: 2575-1883 (Print); ISSN: 2575-1891 (Online)

Modeling the Incidence of Maize Spotted Stem-Borer (*Chilo partellus*) Infestation Under Long-Term Organic and Conventional Farming Systems

Wainaina Stephen[1, 2, *], Anthony Waititu[2], Daisy Salifu[1], Samuel Mwalili[2], Edward Karanja[1], Noah Adamtey[3], Henri Tonnang[1], Felix Matheri[1], Edwin Mwangi[1], David Bautze[3], Chrysantus Tanga[1]

[1]International Centre of Insect Physiology and Ecology (ICIPE), Nairobi, Kenya

[2]Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology (JKUAT), Nairobi, Kenya

[3]Research Institute of Organic Agriculture (Forschungsinstitut für Biologischen Landbau (FiBL)), Frick, Switzerland

Email address:

stephenmangara93@gmail.com (Wainaina Stephen), agwaititu@gmail.com (Anthony Waititu), dsalifu@icipe.org (Daisy Salifu), samuel.mwalili@gmail.com (Samuel Mwalili), ekaranja@icipe.org (Edward Karanja), noah.adamtey@fibl.org (Noah Adamtey), htonnang@icipe.org (Henri Tonnang), fmatheri@icipe.org (Felix Matheri), enderitu@icipe.org (Edwin Mwangi), david.bautze@fibl.org (David Bautze), ctanga@icipe.org (Chrysantus Tanga)

[*]Corresponding author

To cite this article:

Wainaina Stephen, Anthony Waititu, Daisy Salifu, Samuel Mwalili, Edward Karanja, Noah Adamtey, Henri Tonnang, Felix Matheri, Edwin Mwangi, David Bautze, Chrysantus Tanga. Modeling the Incidence of Maize Spotted Stem-Borer (*Chilo partellus*) Infestation Under Long-Term Organic and Conventional Farming Systems. *International Journal of Data Science and Analysis*. Vol. 8, No. 6, 2022, pp. 169-181. doi: 10.11648/j.ijdsa.20220806.11

Received: October 7, 2022; **Accepted**: October 24, 2022; **Published**: November 4, 2022

Abstract: The damage levels of the maize spotted stem borers (Chilo partellus Swinhoe) are estimated at 400,000 metric tons, which is equivalent to 13.5% of farmers' annual maize harvest accounting for US$80 million. Despite the economic importance of the pest, information on the incidence under long-term organic and conventional farming systems is lacking. This study evaluated three different link functions [logit, probit, and complementary log-log – (clog-log)] to reduce prediction errors in overdispersed stem borer incidence data for 12 years in four farming systems. The clog-log link function had the lowest Akaike information criterion (AIC) and Bayesian information criterion (BIC) indexes for the pest incidence model in Thika. Contrarily, probit showed the lowest AIC and BIC in the Chuka incidence data model. The residual diagnostic plots with clog-log demonstrated no patterns against the predicted values. Our findings revealed that clog-log link function provided the best fit in beta-binomial mixed models compared to others. We advocate for the use of clog-log for long-term pest incidence data modelling to obtain biologically realistic projections. Users of mixed models must incorporate explicit consideration of suitable link function discrimination, model fit and model complexity into their decision-making processes if they build biologically realistic models.

Keywords: Maize, Spotted Stem Borer, Pest Incidence, Overdispersion, Binomial Proportions, Beta-Binomial Distribution

1. Introduction

In Africa, maize (*Zea mays* L.) is considered one of the most important staple food crops liked by everyone. In areas of scarcity, the massive shortage has been equated to famine with increased food insecurity challenges. In most cases, this can be attributed to high incidence and damage caused by spotted stem borer (*Chilo partellus,* Swinhoe), leading to reduced maize grain yield loss estimated annually at about 400,000 metric tons, equivalent to US$91 million [1]. Several pest management options have been used to suppress the devastating effects of stem borers on maize but with limited successes. Chemical control methods have proved to be the most effective yet expensive to smallholder farmers and pose risks to humans, livestock, and the environment.

Mitigating the impacts of stem borer attacks requires rapidly identifying incidence in affected farms as they spread further. Understanding the incidence of infestation native species have become increasingly common. Locally established pest species might have explored the range of available environmental conditions and may have uniform and abundant distributions that reflect dispersal limitations. Because these species currently have undergone range expansion at climate equilibrium, describing their incidence thus requires a flexible modelling approach that uses broad-scale, long-term comprehensive data on species occurrences to distinguish informative and uninformative absences. Binomial proportions are encountered in many different fields of study. Modelling of these data has a long history in the statistical domain, including chi-square tests for contingency tables and binomial regression models [2, 3]. Binomial proportions are considered to arise from a number of successes in binomial experiments [4]. It is common to model such data using a generalized linear model (GLM) framework, limiting all specifications to first and second moments [5]. GLMs are distinguished by their membership in the exponential family and their mean-variance relation, except for normal distribution [6]. The GLM framework has been extended to generalized linear mixed models (GLMM) in the case of longitudinal data or repeated measures where the response variable on the experimental unit is recorded over time. The primary feature of GLMM is the addition of subject-specific or random effects in the linear predictors [7].

Response variables that are discrete or grouped binary data (proportion of a number of "successes" out of the total number of units exposed to a particular set of experimental conditions) exhibit overdispersion, indicating that the residual variance is larger than expected under the fitted model. This phenomenon is most common with members of the exponential family, such as binomial and poison distributions with a fixed dispersion parameter $\phi = 1$ associated with the respective distribution variance. Experiments with discrete and bounded outcomes usually have over/underdispersed binomial proportions that appear dispersed and accumulate values at one or both of the distribution scale's edges [8]. Overdispersion is more common than underdispersion. There are two main issues associated with the presence of overdispersion, (i) p-values tend to be too small, increasing the possibility of committing a type I error in which the null hypothesis is rejected when it is true, (ii) confidence intervals are too small leading to overconfidence about the precision of the estimates.

A lot of literature has proposed models that allow the dispersion parameter not to be 1, i.e., $\phi \neq 1$, thus a less restrictive variance-mean relationship [5]. The research of I. Arostegui and others on beta-binomial approach presented the beta-binomial distribution as a suitable fit for some binomial proportions of patient-reported outcomes and found it reliable [8]. The authors demonstrated the appropriateness of the beta-binomial distribution in a cross-sectional framework compared to other exponential family members that are often utilized, such as the binomial and normal distributions. [8] demonstrated that the beta-binomial model adequately accounts for the correlation structure among grouped binary observations while allowing for a flexible relationship between the response's mean and variability.

Within a beta-binomial mixed-effects regression model, the probability of success is related to predictor variables conditional on the random effects; nevertheless, the likelihood of success is not explicitly predicted as a linear combination of the independent variables [9]. A link function is applied to a linear combination of exploratory variables to predict the probability of success conditional on random effects [10]. In other words, a link function connects a nonlinear relationship to a linear relationship to fit a linear model, which is then mapped to the original form by taking the inverse of the link function. The inverse link function can be specified by any monotonically increasing function that transforms values from the range [-∞, ∞] to [0,1]. All inverse link functions are created from known random distributions' cumulative density function (CDF) [11]. For example, the inverse cumulative density functions of the normal, standard logistic, and Gumbel distributions are used to form the probit, logit, and complementary log-log link functions.

Among the most commonly used link functions in beta-binomial mixed effect, models are the probit, logit, and clog-log [12]. The logit link is preferred because of the ease with which parameter estimates can be interpreted using odds ratios. On the other hand, the logit model cannot always be relied upon to provide a satisfactory fit for all beta-binomial mixed-effects regressions, particularly in the case of asymmetric models [13]. When the link function is incorrectly defined, the possibility for significant bias and increased mean squared error increases [14]. Asymmetrical link functions include the complementary log-log link function while probit and logit links are both symmetrical. Symmetric links assume that binomial proportions approach zero at the same rate that they approach one. The asymmetric link assumes that binomial proportions approach zero at a different rate than they approach one. [15] indicated the sensitivity of the inferences if the symmetric link was incorrectly used in the direction of an asymmetric model.

Agriculturalists and researchers may wish to achieve the most reliable estimates in comparison of farming systems' binomial proportions in LTE. One of the factors to consider is the link function in binomial mixed models of their choice to explain different sources of variations. The main objective of this study is to compare the performance of three link functions on binomial proportions from a long-term experiment carried out in two sites in central Kenya highlands. A review of the generalized linear mixed model (GLMM) and beta-binomial mixed effect models for overdispersed binomial proportions is also presented.

2. Related Studies

2.1. Generalized Linear Mixed Models Formulation

Consider $y_i(i = 1,2,3,\ldots,n)$ to denote an observed response. Let X be a design matrix and $x_i{}'$ be the i^th observation in X,β is the fixed effects parameters. Z is the second design matrix and $z_i{}'$ is i^th observation in Z and b is the random effects. Generalized linear mixed models (GLMM) take the following form:

$$g^{-1}(p_i^b) = \eta_i^b = x_i'\beta + z_i'b \tag{1}$$

where g (.) denotes the link function that is generated from the ICDF of the random distribution g. If there is a random effect b GLMM of y_i is an extension of the GLM so that random terms can be included in the linear predictors of conditionally independent observations. This allows overdispersion, correlation, and heterogeneity to be considered.

A GLMM must satisfy the following conditions.

$$f(y_i|b) = exp\left\{\frac{y_i\theta_i - d(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right\} \tag{2}$$

where θ_i is the natural parameter, ϕ is a known overdispersion scalar, $a_i$is a prior weight, and $a_i(\phi)$, $d(\theta_{i,})$ and $c(y_i,\phi)$ are known functions. The conditional mean and variance of a GLMM is give as;

$$E(y_i|b_i) = \mu_i^b = h^{-1}(x_i{}'\beta + z_i{}'b_i) = d'^{(\theta_i)} \tag{3}$$

$$v(y_i|b_i) = \frac{\phi}{a_i}d''^{(\theta)} = \frac{\phi}{a_i}v(\mu_i^b) \tag{4}$$

In this case, $d'(\theta_i)$ is the first derivative, and $d''(\theta_i)$ is the second derivative of $d(\theta_i)$ and a$_i$ is a known prior weight, usually 1.

The random effect $b_1,\ldots,b_N$, are mutually independent with a common underlying distribution G which depends on the unknown parameters α. That is; their probability density function is denoted by $f(b_i)$ with an assumption that means $b_i = 0$ and $var(b_i) = G$.

$$b_i \sim iid(0,G)$$

It is necessary to specify $f(y_{ij}|b_{i,})$ and $f(b_i)$ in GLMM. Based on $f(y_{ij}|b_{i,})$ and $f(b_i)$ the marginal density of Y, $f(y_{ij}|b_i)$ is given by

$$f(y_{ij}|b_i) = \int \prod_{j=1}^{n_i} f\left(y_{ij}|b_i\right)f(b_i)db_i \tag{5}$$

Because the marginal density $f(y_{ij})$ is a normal density, an analytical or closed-form solution for integral eq.5 could be obtained using the linear mixed model (LMM). The likelihood function in the generalized linear mixed model (GLMM) is built using the marginal density $f(y_{ij})$. In general, $f(y_{ij})$ is difficult to compute because $f(y_{ij})\ f(b_i)$ can be a complex function with high-dimension integrals. As a result, approximation methods must be used to solve the integral. Assume there are k groups, and each group has n_i units; $i = 1,2,\ldots,k$. The likelihood function for k groups is then calculated as follows.

$$L(\theta) = f(y_{ij}) = \prod_{i=1}^{k} \int \prod_{j=1}^{n_i} f\left(y_{ij}|b_i\right)f(b_i)db_i \tag{6}$$

To approximate likelihood function eq.6 using Laplace approximation, the equation is rewritten as:

$$\int \prod_{j=1}^{n_i} f\left(y_{ij}|b_i\right)f(b_i)db_i = \int f(\mathrm{y}|\mathrm{b})f(b) \tag{7}$$

$$= \int e^{logf(\mathrm{y}|\mathrm{b})f(b)}d\mathrm{b} = \int e^{g(b)}d\mathrm{b} \tag{8}$$

where $g(b) = logf(\mathrm{y}|\mathrm{b})f(\mathrm{y}|\mathrm{b})$

We want to choose $\hat{b}$ in such a way that $g(b)$ is maximized by satisfying the necessary and sufficient conditions $g'(b) = 0$ and $g''(\hat{b}) < 0$. The following expression gives the second-order Taylor expansion around $\hat{b}$ for $g(b)$:

$$g(b) \approx \tilde{g}(b) = g(\hat{b}) + (b - \hat{b})\, g'(\hat{b}) + \tfrac{1}{2}(b - \hat{b})^2 g''(\hat{b}) \tag{9}$$

$$= g(\hat{b}) - \tfrac{1}{2}(b - \hat{b})^2\left(-g''(\hat{b})\right) \tag{10}$$

This shows that $e^{\tilde{g}(b)}$ is proportional to the normal density (μ_L,σ_L^2) where _L=b and $\mu_L = \hat{b}$ and $\sigma_L^2 = -\frac{1}{g''(\hat{b})}$ is the normal density. So, the Laplace approximation for the likelihood $L(\theta)$ is given by the following formula:

$$(\theta) = \int e^{g(b)}db \approx \int e^{\tilde{g}(b)}db \tag{11}$$

$$= exp\left(g(\hat{b})\right)\int exp\left(-\frac{1}{2\sigma_L^2}(b - \mu_L)^2\right)db = exp\left(g(\hat{b})\right)\sqrt{2\pi\sigma_L^2} \tag{12}$$

Also, be stated that

$$(b|y) = \frac{f(y|b)f(b)}{f(y)} \propto exp\big(g(b)\big) \approx const\ \ exp(b - \mu_L)^2 \tag{13}$$

$$B|Y = y \approx \mathcal{N}(\mu_L,\sigma_L^2).$$

2.2. Beta-Binomial Distribution

Consider the variable $y_i(i = 1,2,3,\ldots,n)$ to represent an observed binary response with a random variable λ that follows a beta distribution with parameters $\alpha,\beta > 0$. The binary responses are assumed to be independently distributed and identical conditional on the random variable λ.

$$y_j|\lambda \sim Ber(\lambda)\ \ iid,\ \ where\ \ \lambda \sim Beta(\alpha,\beta), j = 1,2,3\cdots,n$$

with

$$E(\lambda) = p \quad and \quad Var(\lambda) = \frac{p(1-p)\phi}{(1+\phi)}$$

Where $p = \frac{\alpha}{(\alpha+\beta)}$ and $\phi = \frac{1}{(\alpha_1+\alpha_2)}$.

As a result, the mean and variance of the outcome variable are computed as follows;

$$E(y_j) = E[E(y_j|\lambda)] = p \tag{14}$$

$$Var(y_j) = Var[E(y_j|\lambda)] + E[Var(y_j|\lambda)] = p(1-p), j = 1, \cdots, n. \quad (15)$$

The mean and variance correspond to moments of Bernoulli distribution, but observations are assumed to have a between-subject correlation which is given as;

$$\rho = Corr(y_j, y_k) = \frac{Cov(y_j, y_k)}{\sqrt{Var(y_j)}\sqrt{Var(y_k)}} = \frac{\phi}{1+\phi}, k, j = 1,2,3, \cdots, n \text{ and } j \neq k \quad (16)$$

The parameter ϕ in the eq. 16 is determined as the dispersion parameter. The Sum of all correlated binary outcomes can be defined as:

$$y = \sum_{j=1}^{n} y_j, \quad (17)$$

which is a beta-binomial random variable with three parameters: $m,\ p,\ and\ \phi$. Given a binary response conditional on the occurrence of a random effect, as follows:

$$y|\lambda \sim Bin(m, \lambda) \quad and \quad \lambda \sim Beta(p/\phi, (1-p)/\phi).$$

The beta-binomial distribution's probability mass is given

$$f(y) = \int_0^1 f_{y|\lambda}(y|\lambda) f_\lambda(\lambda) d\lambda \quad (18)$$

$$= \binom{n}{y} \frac{\Gamma\left(\frac{1}{\phi}\right)}{\Gamma\left(\frac{1}{\phi}+m\right)} \frac{\Gamma\left(\frac{p}{\phi}+y\right)}{\Gamma\left(\frac{p}{\phi}\right)} \frac{\Gamma\left(\frac{(1-p)}{\phi}+n-y\right)}{\Gamma\left(\frac{(1-p)}{\phi}\right)} \quad (19)$$

Furthermore, the mean and variance of the beta-binomial distribution is given as;

$$E(y) = np \quad (20)$$

$$Var(y) = np(1-p)\left[1 + (n-1)\frac{\phi}{1+\phi}\right] \quad (21)$$

Thus, the beta-binomial model can be thought of as a binomial distribution with extra variability because of the correlation between the y_j Values.

2.3. The Mixed Effects Beta-Binomial Model

With the addition of random effects in the linear predictor, this section makes the marginal beta-binomial regression more general. Consider the variable $y_i (i = 1,2,3, \ldots, n)$ to represent an observed response conditional on the random effects u . Assuming y is taken to be selected from a beta−binomial distribution and that random effect u is taken derived from a multivariate normal distribution with zero mean and variance−covariance matrix D, then;

$$y_i|u \sim BB(n_i, p_i, \phi) \quad and \quad u \sim N(0, D(\gamma)), \quad i = 1, \cdots, n$$

Let $\theta = (\phi, \gamma)$ denote the parameter vector of mode1's variance or dispersion components. A link function connects the response variable of a beta-binomial distribution to the predictor variables conditional on the random effects. If a logistic link function is used, the model takes the form;

$$\eta_i = log\frac{p_i}{1-p_i} = x_i'\beta + z_i'u, \quad i = 1, \cdots, n \quad (22)$$

And the likelihood function is given as follows;

$$L(\beta, \theta|y) = \int \prod_{i=1}^{n} f(y_i|\beta, \phi, u) f(u|\gamma) du \quad (23)$$

The beta-binomial density function is denoted by $f(y_i|\beta, \phi, u)$ While the multivariate normal density function for the random effects is denoted by $f(u|\gamma)$, the marginal likelihood has no closed form, just as it does in the case of the GLMM, and numerical calculation is almost impossible due to the beta-binomial distribution's complexity. As a result, approximation procedures for estimating the parameters in the model must be developed.

The marginal likelihood of the beta-binomial model can also be expressed in exponential form as;

$$L(\beta, \theta|y) = \int exp\{\sum_{i=1}^{n} l\,ogf(y_i|\beta, \phi, u) + logf(u|\gamma)\} du \quad (24)$$

In this case, the approximation is obtained because the sum of two twice differentiable regular functions is also a twice differentiable regular function.

$$ogL(\beta, \theta|y) \approx l(\beta, \theta|y, \tilde{u}) = h(\beta, \theta|y, \tilde{u}) - \frac{1}{2} log\{det(M)\} \quad (25)$$

and then

$$h(\beta, \theta|y, u) = \sum_{i=1}^{n} l\,ogf(y_i|\beta, \phi, u) + logf(u|\gamma) \quad (26)$$

$$\sum_{i=1}^{n}[\sum_{k=0}^{y_i-1} l\,og(p_i + k\phi) + \sum_{k=0}^{m_i-y_i-1} l\,og(1 - p_i + k\phi) - \sum_{k=0}^{m_i-1} l\,og(1 + k(\phi)) - \frac{1}{2}\log\{det(D)\} - \frac{1}{2}u'D^{-1}u]$$

Eq.26 defines the model's joint log-likelihood. $\tilde{u}$ gives a solution of $\partial h/\partial u = 0$ and − $-\frac{1}{2} log\{det(M)\}$ is the adjusted term with

$$M = \frac{\partial^2 h}{\partial u\, \partial u'}|_{u=\tilde{u}} \quad (27)$$

The marginal log-likelihood approximation is the same as integrating random effects in the first-order Laplace approximation.

2.4. Estimation of the Fixed and Random Effects

When estimating fixed parameters, it is assumed that θ is constant, attempting to maximize the approximated log-likelihood. The log-likelihood is denoted as

$$(\beta|\tilde{u}, \theta) = A(\beta) + h(\beta|\tilde{u}, \theta) \quad (28)$$

Thus, the scoring equation for the model's fixed variables is provided by

$$S(\beta) = \frac{\partial A(\beta)}{\partial \beta} + \frac{\partial h(\beta|y, \tilde{u}, \theta)}{\partial \beta} \quad (29)$$

The beta-binomial mixed-effects model has these characteristics.

$$A(\beta) = -\frac{1}{2} log\{det(M)\} = -\frac{1}{2} log\{det(Z')SWZ - D^{-1}\} \quad (30)$$

$S = diag\left(p_i(1-p_i)\right)$, $W = diag(\omega_i)$, $\omega_i = -v_i p_i(1-p_i) + \xi_i(1-2p_i)$ and

$$\begin{cases} \xi_i = \sum_{k=0}^{y_i-1} \frac{1}{p_i+k\phi} - \sum_{k=0}^{m_i-y_i-1} \frac{1}{1-p_i+k\phi} \\ v_i = \sum_{k=0}^{y_i-1} \frac{1}{(p_i+k\phi)^2} + \sum_{k=0}^{m_i-y_i-1} \frac{1}{(1-p_i+k\phi)^2} \end{cases} \quad (31)$$

all of the preceding formulas were evaluated at $u = \tilde{u}$ for $i = 1, \cdots, n$. It is evident from the weight matrices W and S that A (.) is simply reliant on β. The Weight matrices are assumed to vary slowly or not vary as a function of fixed effects. Next, in eq. 29, the first term is disregarded while the second is maximized to the utmost extent possible to get the most likely estimates of the fixed effects. Eq. 26 defines a joint log-likelihood function, which is responsible for collecting all of the information about β.

As a result, the maximum likelihood estimate of fixed effects can be obtained by maximizing the joint log-likelihood of the fixed and random effects. The log-likelihoods of β and $\mathbf{u}$ may be decomposed into the following scoring equations.

$$\begin{cases} \xi' SX = 0 \\ \xi' SZ - u' D^{-1} = 0 \end{cases} \quad (32)$$

Various numerical algorithms can be used to solve the previous equation iteratively. The delta technique, a variant of the Newton–Raphson method, is presented as a solution to the complexity of the second derivative of the beta-binomial density function.

In this study, we explore different beta-binomial mixed effects models for the analysis of overdispersed binomial proportions on stem borer infestation and identify the best model for the analysis of the data in question and other similar binomial proportions.

3. Methodology

This study utilized data from a long-term experiment (LTE) on cropping systems. The study was carried out in two sites, Chuka and Thika, as part of a long-term farming system comparative trial that began in 2007 [16]. Four treatments were selected to describe farming systems (Conventional and Organic) and input levels (Low and High) in a long-term farming system experiment. Conventional treatments received chemical fertilizers and pesticides, while Organic treatments received only fertilizer materials (compost) and pesticides recommended for organic farming. The input level treatments designated by 'High' were designed to emulate farmers who target distant high-value markets (both local and export) and apply commercial rates of pesticides and fertilizers. In contrast, the 'Low' treatments were intended to emulate typical rates of pesticides and fertilizers used by smallholders, where the produce are consumed by household or sold in the local markets. There were, therefore, two sets of treatments under study per input level, namely: conventional high and organic high system (set 1) and conventional low and low organic system (set 2).

3.1. Link Functions

The main reason to use link functions is to convert the linear combination of covariates ranging from -∞ and +∞ to a probability scale of 0 and 1 [17].

3.1.1. Logit Link Function

A logit link function is used to model the probability of success as a function of explanatory variables [18]. For example, the logit link function is defined as

$$ogit(p) = \ln\left(\frac{p}{1-p}\right) = \mathrm{X}\beta + Zu \quad (33)$$

Where p is the probability of success, X is the fixed effects design matrix, β is the fixed effects regression coefficients, Z is the random effects design matrix, and u is the random effects coefficients.

Using a different arrangement of the variables, eq. 33 may be transformed into the following connection with the variables:

$$p = \frac{exp(\mathrm{X}\beta + Zu)}{1 + exp(\mathrm{X}\beta + Zu)} \quad (34)$$

The logit of p is also referred to as the log odds for success in certain circles. In statistics, the chances of success (i.e., how much bigger the likelihood of success is compared to the probability of failure) are stated as the ratio $\left(\frac{p}{1-p}\right)$.

3.1.2. Complementary Log-Log (Clog-Log)

Complementary log-log (Clog-log) is an asymmetric link function used to represent success probabilities as a function of explanatory factors [3]. The link takes the form:

$$\eta = \log\left(-\log(1-p)\right) = \mathrm{X}\beta + Zu \quad (35)$$

where

$$\mathrm{p} = P(Y = 1 | X = x)$$

So $exp(\eta)$ is not the odds since $\exp(\eta) = -\log(1-\mathrm{p})$

Hence $\exp\left(-\exp(\eta)\right) = 1 - \mathrm{p}$ and $1 - \exp\left(-\exp(\eta)\right) = \mathrm{p}$.

Therefore, if we require an odds ratio for a certain covariate, we can compute one, but the parameters lack a straightforward, clear meaning regarding contribution to log odds. The complementary log-log model's parameters have a tidy explanation in terms of the hazard ratio, that is;

$$e^{\eta} = -\log(1-\mathrm{p}) = -\log(S_x) \quad (36)$$

Where S is the survival function. Now the hazard is

$$h(x) = -\frac{d}{dx}\log(S_x) = \frac{d}{dx}e^{\eta} \quad (37)$$

3.1.3. Probit Link Function

The probit model uses the cumulative distribution function of the standard normal distribution to model the probability of 'success' as a function of explanatory variables [19]. The model is of the form

$$\Phi^{-1}(p) = X\beta + Zu \tag{38}$$

3.2. The Data

Twenty plants per plot were sampled from 2007 to 2013, and 40 from 2015 to 2019, and pest incidence was assessed from the sampled plants. Thus, the response variable (binomial proportions) is the total stemborer incidences for each plot of 20 or 40 sampled plants.

$$\pi_{ijk} = \frac{\sum Number\ of\ plants\ with\ stemborer\ per\ plot}{sampled\ plants} \tag{39}$$

We model the relationship between binomial proportions and the type of farming using a beta-binomial generalized linear mixed model. Three possible models with three-link functions are presented as follows;

$$logit(\pi_{ijk}) = \ln\left(\frac{\pi_{ijk}}{1-\pi_{ijk}}\right) = \eta_{ijk} = E(\pi_{ijk}|b_{ik},) = \beta_0 + \alpha_i + \beta_j + \alpha\beta_{ij} + b_{ik} \tag{40}$$

$$probit(\pi_{ijk}) = \Phi^{-1}(\pi_{ijk}) = \eta_{ijk} = E(\pi_{ijk}|b_{ik},) = \beta_0 + \alpha_i + \beta_j + \alpha\beta_{ij} + b_{ik} \tag{41}$$

$$Clog - log(\pi_{ijk}) = \log\left(-\log(1-\pi_{ijk})\right) = \eta_{ijk} = E(\pi_{ijk}|b_{ik},) = \beta_0 + \alpha_i + \beta_j + \alpha\beta_{ij} + b_{ik}$$

where;

β_0 Denotes the intercept of the model.

α_i Denotes the ith farming system effect.

β_j Denotes the jth Year effect.

$\alpha\beta_{ij}$ Denotes the $ijth$ farming system x Year interaction effect.

$b_{ik} \sim N(0, \sigma_b^2)$ between-subjects (plots) effect.

3.3. Comparison of the Model's Goodness-of-Fit

When comparing models with three distinct link functions, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) were utilized in the study, respectively. The AIC version utilized is minus twice the maximal log-likelihood plus twice the parameter counts. The likelihood function is also used to calculate the BIC, which is linked to the AIC. The best model is the one that has the lowest AIC and BIC.

$$IC = -2L + 2q \tag{42}$$

$$BIC = -2L + q\ln(N) \tag{43}$$

3.4. Residual Diagnostic Plots

To help in spotting departures from uniformity, residual diagnostics plots were created in R using the DHARMa package. The graphic displayed three tests: outlier, overdispersion, and the Kolmogorov-Smirnov test.

3.4.1. DHARMa Residual

DHARMa package [20] creates readily interpretable residuals for mixed models that are standardized between zero and one. The resulting residuals are interpreted as in the case of linear models. A simulation approach is used to produce the standardized residuals in three steps.

1) For each observation, simulate new response data from the fitted model.
2) Calculate the empirical cumulative density function for the simulated observations for each observation, which depicts the potential values (and their probability) for the observed value for the predictor combination, providing the fitted model is valid.
3) After that, the residual is defined as the value of the empirical density function at the observed data value.

The fundamental advantage of this formulation is that if the model is correctly stated, the so-defined residuals always have the same known distribution, regardless of the fit model [21]. The residuals have 0 and 1 as their minimum and maximum values. We should expect asymptotically for a correctly stated model uniform distribution of the scaled residuals. The residuals can be transformed to a normal distribution for a more straightforward interpretation as in linear regression models. The residuals are visualized in two plots;

1) Q-Q-plot, which helps to detect all deviations from the fitted distribution. The plot includes three tests: the Kolmogorov Smirnov (KM) test, the dispersion test, and an outlier test.
2) Residuals plot indicating residuals against the predicted values and simulation outliers.

3.4.2. Kolmogorov-Smirnov Goodness-of-Fit Test

The Kolmogorov-Smirnov test was therefore performed to determine whether binomial proportions were chosen from the beta-binomial distribution in this study. The Kolmogorov-Smirnov (K-S) test is based on the empirical distribution function. The empirical distribution function is defined as follows for N ordered data items $Y_1, Y_2, \cdots, Y_N$:

$$E_N = \frac{n(i)}{N}, \tag{44}$$

where $n(i)$ is the number of points less than Y_i, and which denotes that the points less than Y_i are arranged from least to biggest value. When the value of each ordered data point is greater than zero, this is a step function that grows by $1/N$. The test statistic for the Kolmogorov-Smirnov test is given as:

$$D = max_{1 \le i \le N}\left(F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i)\right) \tag{45}$$

where F is the beta-binomial theoretical cumulative distribution.

All analyses were implemented in R version 4.1.2 [22].

4. Results

This section presents an application of the beta-binomial mixed effect model to overdispersed binomial proportions using logit, probit and clog-log link functions.

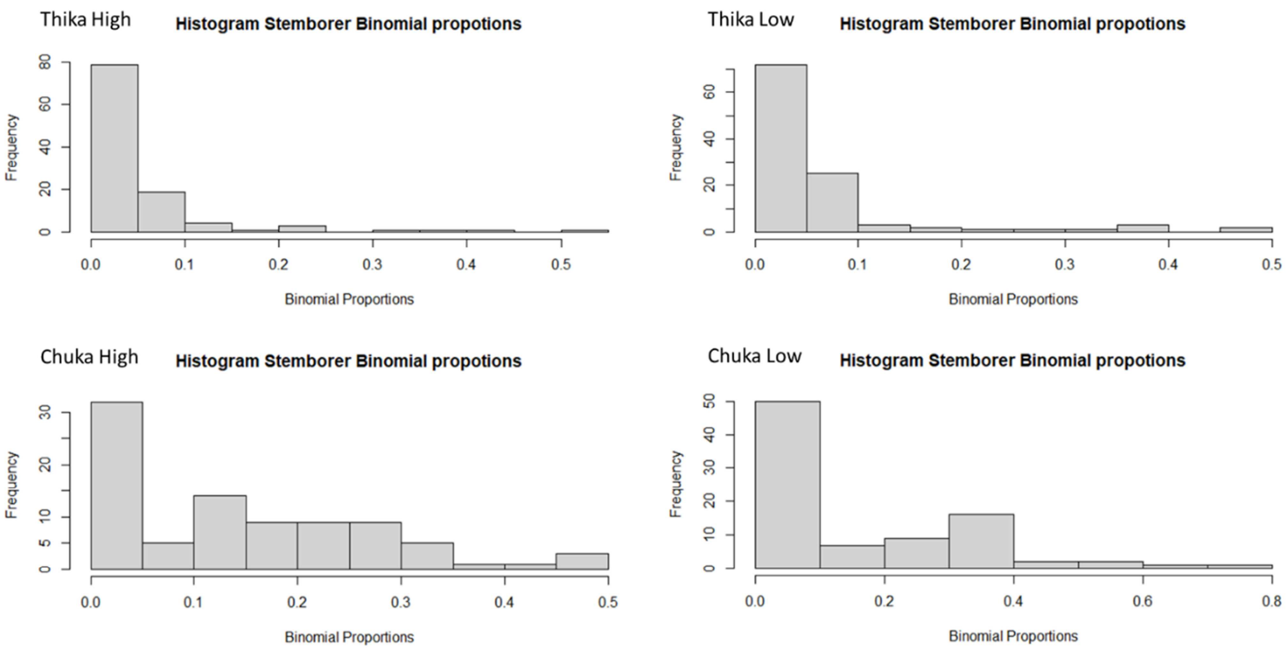

Figure 1. *Distribution of binomial proportions for the two input levels in two sites, Thika and Chuka, for the conventional and organic systems. The first row describes the distribution of proportions for the farming systems used in Thika at two input levels (Low and High). The second row describes the distribution of proportions for the two input levels in Chuka. In a symmetric distribution, the proportions approach one at the same rate; they approach zero, while in asymmetric distribution, the proportions tend to accumulate at the edge of the distribution.*

Table 1. *Overdispersion parameters for the three-link functions binomial generalized linear mixed model.*

Input level	Logit	Probit	Clog-log
Thika High	2.6321	2.8526	2.5729
Thika Low	2.8123	3.128	2.6814
Chuka High	1.2602	1.1919	1.3169
Chuka Low	2.2051	2.1316	2.236

Figure 1 shows that binomial proportions were not symmetrical as they were not approaching zero and one at the same rate. The data showed heavier tails towards zero in all input levels. The proportions surpassed 0.5 for Chuka low input level suggesting the asymmetric distribution of the binomial proportions.

All three-link functions had overdispersion parameters exceeding 1. The parameters appeared very close to each other, and in the presence of a significant overdispersion parameter, no model could be considered better than the other. The majority of the model estimated overdispersion parameter >2.

From Table 2, notably, the year effect was the only statistically significant covariate in all input levels; that is; all three models reduced stemborer infestation over time in both input levels in the two ecological zones,

Table 2. *Parameter estimates and associated p-values (Pr (>|z|) corresponding to the fitted beta-binomial model with different link functions for the binomial proportions on stemborer infestation on maize plants. The first column presents the two input levels in two ecological zones, i.e., Thika and Chuka. The organic system was the reference factor in each; hence the second column shows a comparison of the organic system to the conventional system, year effect and their interaction. Standard errors of the estimates are shown in the brackets.*

		Logit		Probit		Cloglog	
		Estimate (error)	Pr (>|z|)	Estimate (error)	Pr (>|z|)	Estimate (error)	Pr (>|z|)
Thika High	(Intercept)	-1.51 (0.23)	<0.0001	-0.94 (0.12)	<0.0001	-1.58 (0.22)	<0.0001
	(Conventional High)	-0.53 (0.35)	0.128	-0.28 (0.17)	0.109	-0.50 (0.33)	0.129
	year	-0.21 (0.04)	<0.0001	-0.10 (0.02)	<0.0001	-0.20 (0.04)	<0.0001
	Conventional High: year	0.05 (0.06)	0.407	0.03 (0.03)	0.315	0.04 (0.05)	0.420
Thika Low	(Intercept)	-1.65 (0.30)	<0.0001	-1.05 (0.15)	<0.0001	-1.70 (0.28)	<0.0001
	(Conventional Low)	-0.49 (0.41)	0.2306	-0.20 (0.02)	0.313	-0.49 (0.39)	0.2072
	year	-0.19 (0.05)	<0.0001	-0.08 (0.02)	<0.0001	-0.19 (0.05)	<0.0001
	Conventional Low: year	0.01 (0.06)	0.0864	0.04 (0.03)	0.123	0.10 (0.06)	0.0773
Chuka High	(Intercept)	0.07 (0.20)	0.710	-0.01 (0.11)	0.960	-0.24 (0.16)	0.147
	(Conventional High)	-0.03 (0.29)	0.894	-0.05 (0.16)	0.738	-0.01 (0.26)	0.970
	year	-0.37 (0.03)	<0.0001	-0.20 (0.02)	<0.0001	-0.33 (0.03)	<0.0001
	Conventional High: year	-0.08 (0.05)	0.187	-0.03 (0.03)	0.273	-0.08 (0.05)	0.146
Chuka Low	(Intercept)	-0.08 (0.24)	0.737	-0.11 (0.14)	0.454	-0.33 (0.20)	0.0976.
	(Conventional Low)	-0.19 (0.34) 0.8270	0.580	-0.11 (0.20)	0.594	-0.17 (0.29)	0.5548
	year	-0.28 (0.04) 0.7558	<0.0001	-0.15 (0.02)	<0.0001	-0.27 (0.04) 0.7634	<0.0001
	Conventional Low: year	0.004 (0.06)	0.936	0.003 (0.03)	0.924	0.004 (0.05)	0.9407

Table 3. *Model AIC and BIC corresponding to the different link functions of the beta-binomial mixed-effects model.*

	Link function	Thika		Chuka	
		AIC	BIC	AIC	BIC
High Input	Logit	673.96	690.16	534.49	549.35
	Probit	673.98	690.18	525.00	539.86
	Clog-log	673.78	689.98	538.44	553.30
Low Input	Logit	700.81	717.02	650.22	665.08
	Probit	701.47	717.67	648.75	663.61
	Clog-log	700.56	716.76	650.30	665.16

Akaike information criterion (AIC) and Bayesian information criterion (BIC) values were very close in both input levels for the two sites. A beta-binomial mixed-effect model with a complementary log-log link function had the least AIC and BIC values for the Thika site. A model with a probit link function presented the least AIC and BIC values in both input levels in Chuka.

4.1. Thika High Input Systems

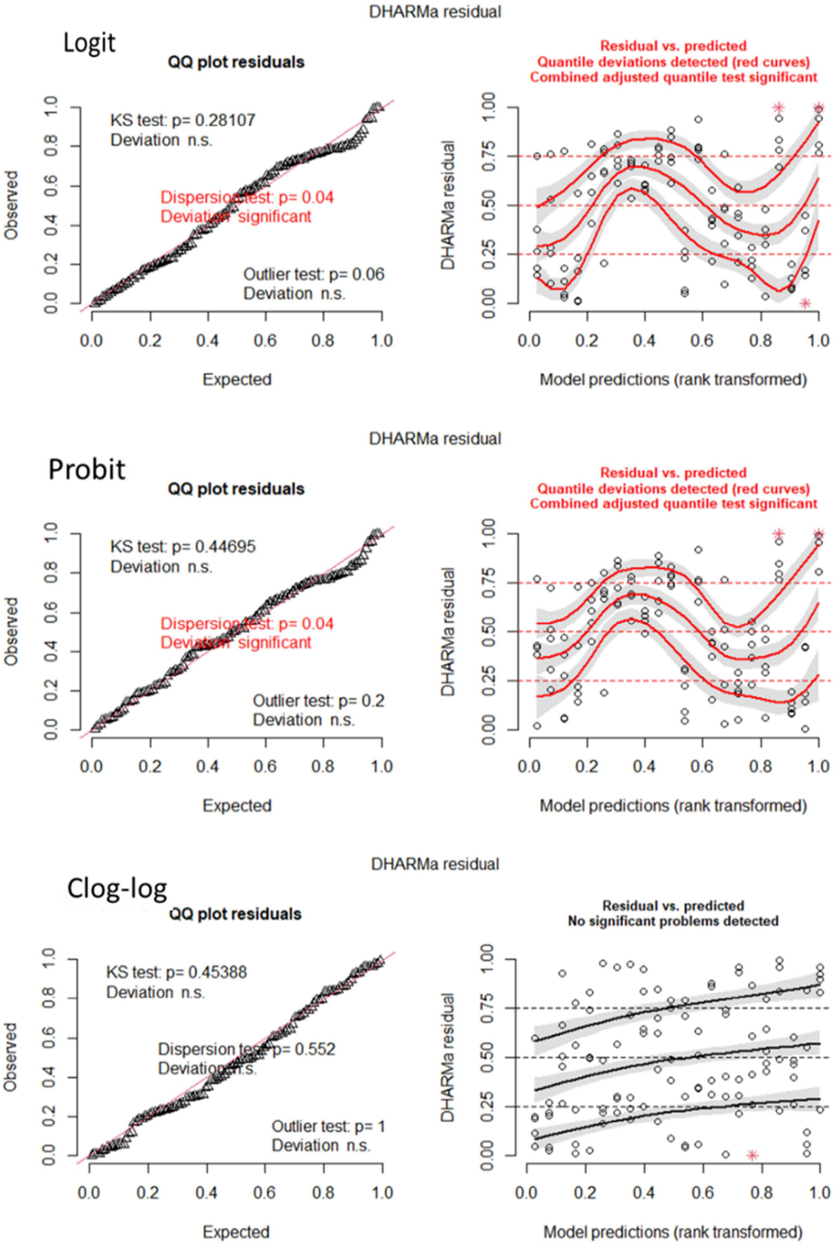

Figure 2. *QQ-plot (left panel) and residual plot (right panel) of three different link functions. QQ-plot detects overall deviations from the expected distribution with added tests for correct distribution (KS test), dispersion and outliers. The residuals plot produces a plot of the residuals against the predicted value. Red curves indicate significant patterns in the residuals for Thika high input level models.*

The plot of residuals for the logit and probit link functions suggests the presence of overdispersion, p = 0.04 in both cases. Kolmogorov Smirnov test showed that Binomial proportions were consistent with the beta-binomial model with an insignificant p-value of 0.2811, 0.4470, and 0.4539 for logit, probit, and Clog-log, respectively (Figure 2). There was also an indication of poor fit by the two link functions (logit and probit), as noted in Dharma residual vs Model predictions plots of the respective link functions in figure 2.

4.2. Thika Low Input Systems

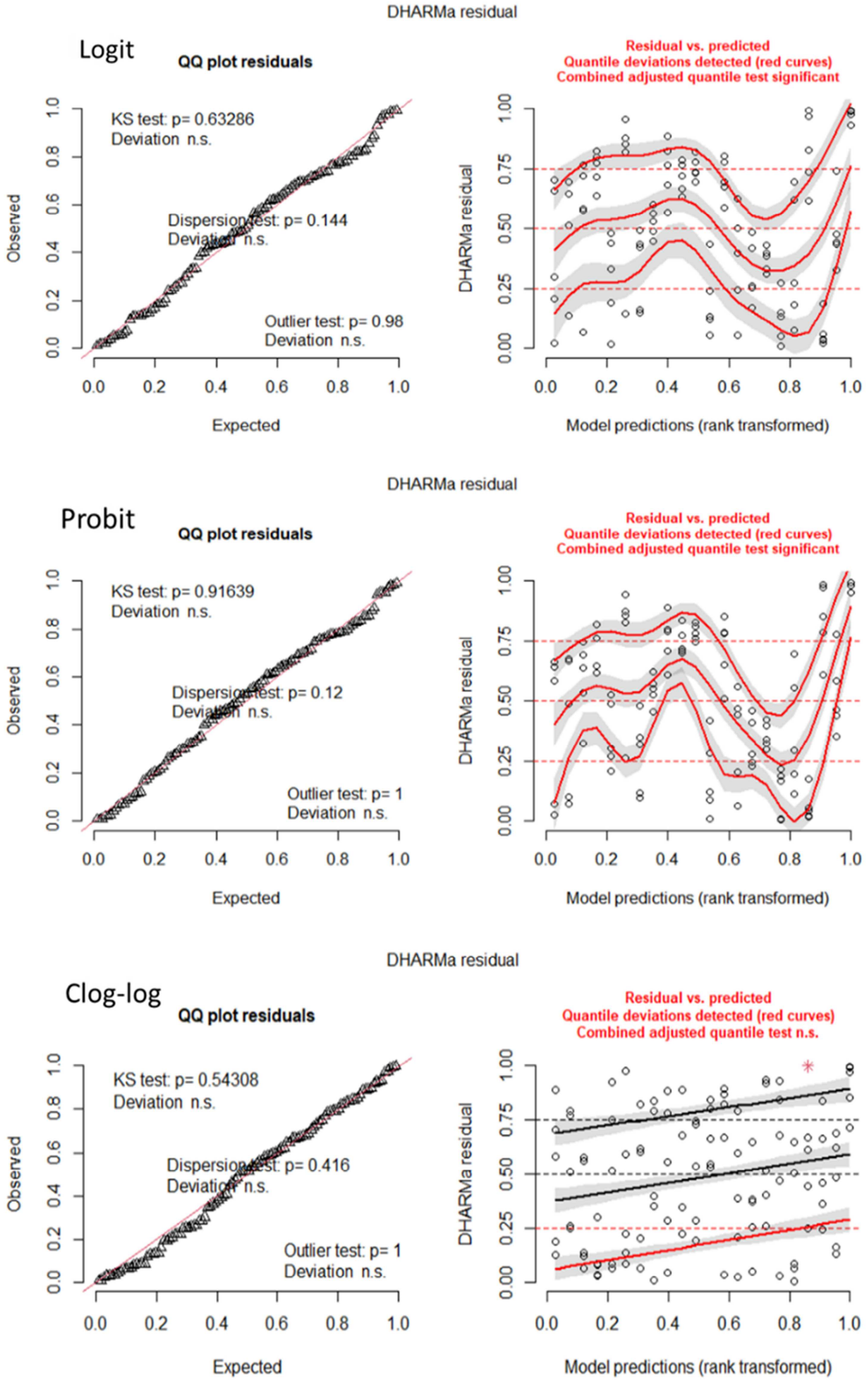

Figure 3. *QQ-plot (left panel) and residual plot (right panel) of three different link functions. QQ-plot detects overall deviations from the expected distribution with added tests for correct distribution (KS test), dispersion and outliers. The residuals plot produces a plot of the residuals against the predicted value. Red curves indicate significant patterns in the residuals for Thika low input level models.*

The three-link functions do not indicate a poor fit to the data. Kolmogorov Smirnov test showed non-significant p-values in the three cases, suggesting that the data was drawn from a beta-binomial model regardless of the link function used. Again, there was no presence of overdispersion and outliers in all link functions. The complementary log-log link function showed a better fit through residual patterns than the logit and probit link functions. Residuals distribution indicated that logit and probit link functions had a poorer fit to the data.

4.3. Chuka High Input Systems

All three tests showed a distributional fit for binomial proportions to the beta-binomial mixed effect model using the three-link functions. Again, complementary log-log presented a better fit than logit and probit link functions based on the residual patterns.

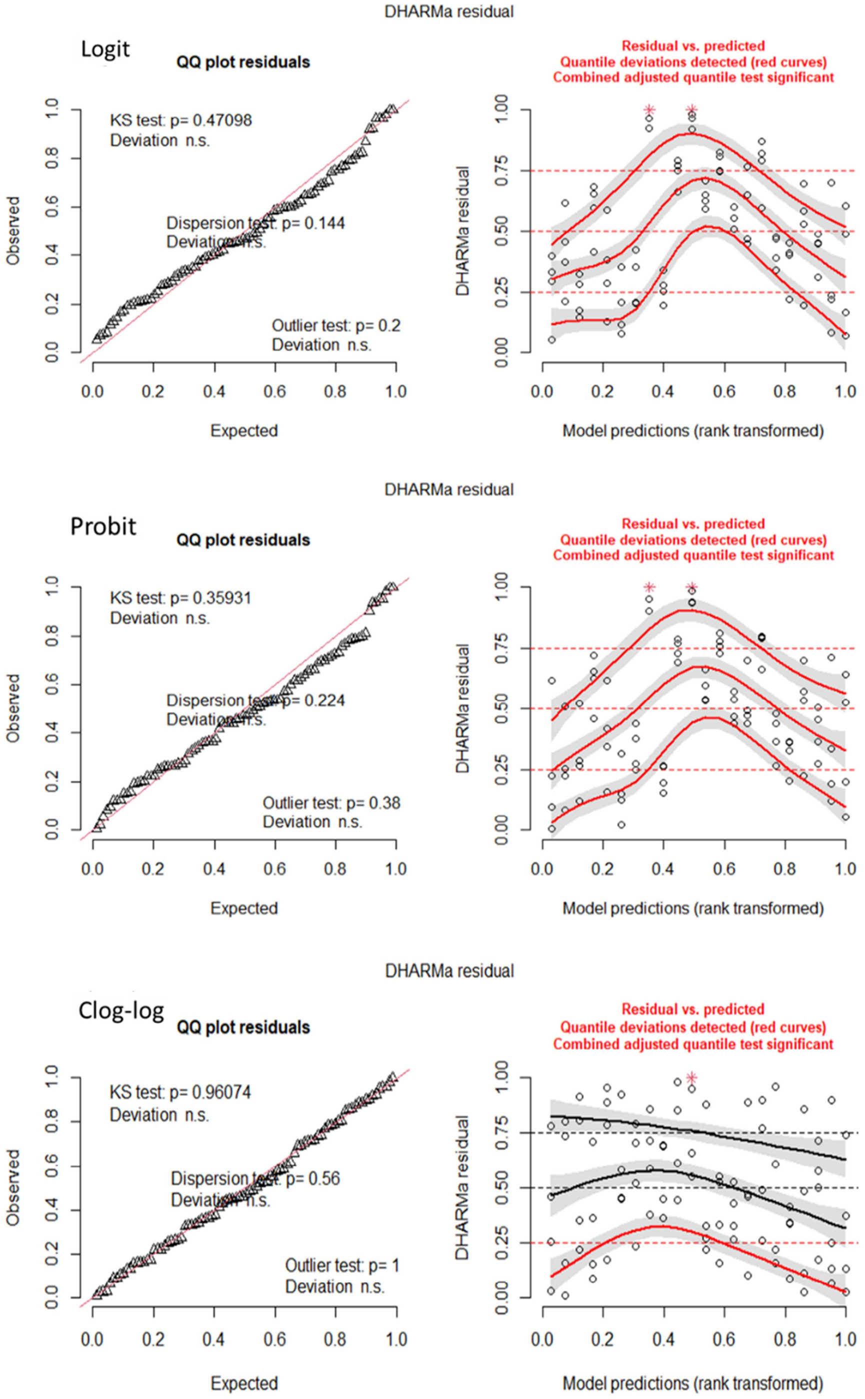

Figure 4. *QQ-plot (left panel) and residual plot (right panel) of three different link functions. QQ-plot detects overall deviations from the expected distribution with added tests for correct distribution (KS test), dispersion and outliers. The residuals plot produces a plot of the residuals against the predicted value. Red curves indicate significant patterns in the residuals for Chuka high input level models.*

4.4. Chuka Low Input Systems

Again, QQ-plot residuals and the attached test showed a distributional fit of binomial proportions to the beta-binomial mixed effect model. Logit and probit showed that there was a significant pattern in the residual. Complementary log-log presented a better fit in residual patterns with significant pattern in upper quartile only.

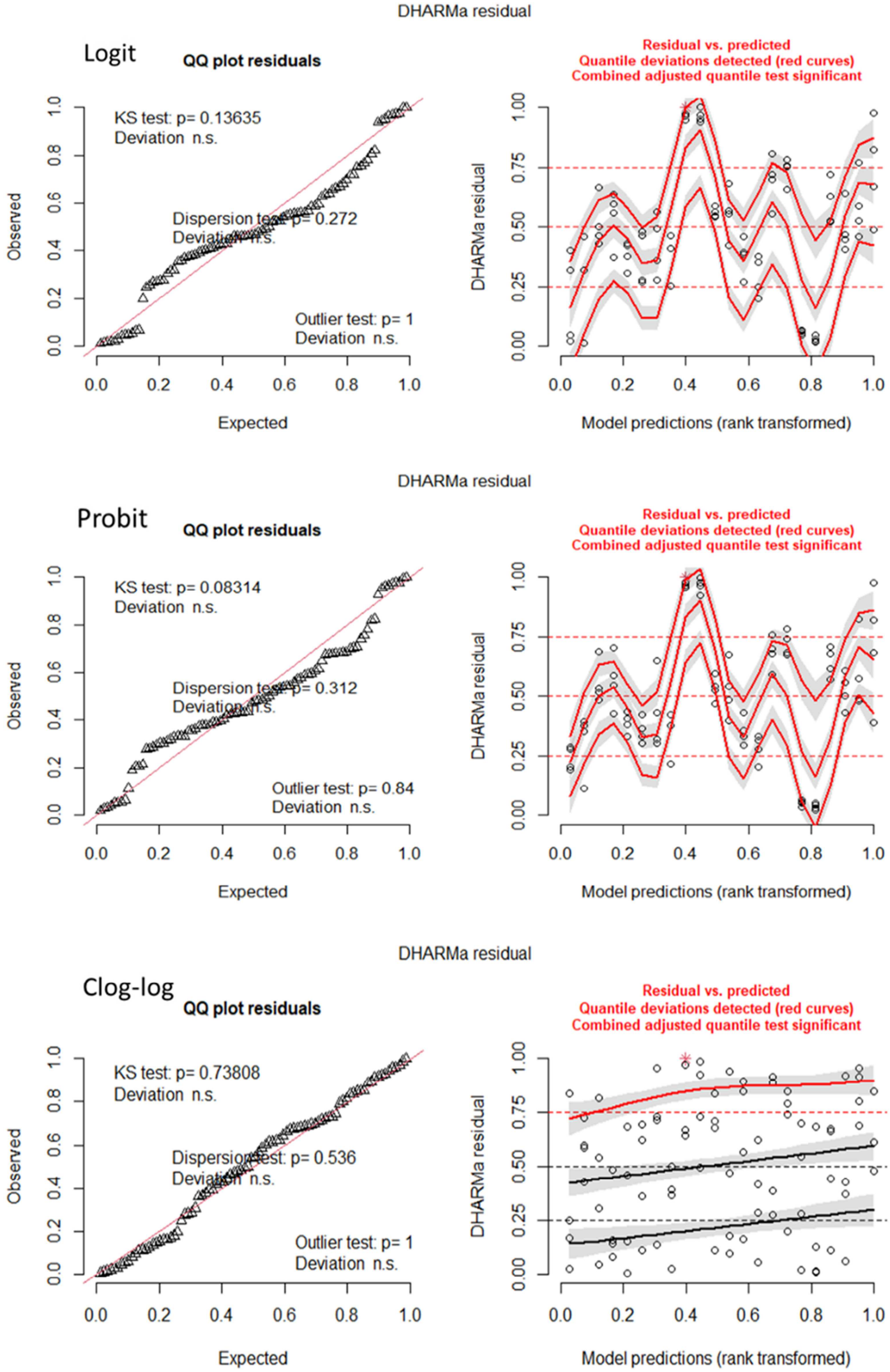

Figure 5. *QQ-plot (left panel) and residual plot (right panel) of three different link functions. QQ-plot detects overall deviations from the expected distribution with added tests for correct distribution (KS test), dispersion and outliers. The residuals plot produces a plot of the residuals against the predicted value. Red curves indicate significant patterns in the residuals for Chuka low input level models.*

5. Discussion

In this study, we compared the performance of three link functions in a beta-binomial mixed effect model for overdispersed proportions on stemborer infestation to identify the best link function for the data. Overdispersion was evident in the data when a binomial generalized linear mixed model (GLMM) was used. This excess dispersion suggests a greater variance of stemborer infestation proportions due to unobserved heterogeneity in the data. For example, unobserved factors that could have affected this may include using different crops in the farming system over the cropping season and pest control intervention. The rule of thumb is that overdispersion parameter should be close to 1 and if greater than 1.10, overdispersion should be modeled [23]. The study considered binomial GLMM inappropriate to model this data as the model would result in biased parameter estimates and underestimated standard errors leading to invalid conclusions [24].

A beta-binomial mixed-effect model was a better fit for the overdispersed stemborer infestation proportions. Many studies have demonstrated the superiority of beta-binomial distribution in modeling this kind of data from different fields. The models was employed to examine the significance of variable protein abundance, and the findings revealed that it performed better than other models on various datasets in terms of false detection rate and true detection rate [25]. The research of O. kush on statistical model using beta-binomial distribuition and bivariate copulas proposed a model that uses beta-binomial distributions for the marginal numbers of the positive and true negatives to overcome the challenges encountered when meta-analyzing data from studies on diagnostic accuracy [26].

Regarding the link functions, the clog-log performed better than probit and logit link with the beta-binomial mixed-effect model. This is not surprising since the data here are asymmetric while the probit and logit links are for symmetric datasets. Using a symmetric link function when the proportions are skewed would be inappropriate [3] and asymmetric link functions provide a suitable alternative in such cases [11]. A beta-binomial mixed-effects model using the clog-log link function, which is asymmetric, fits the overdispersed data better for the Thika site according to AIC values. Furthermore, Q-Q-plots with three embedded tests (Kolmogorov Smirnov test, dispersion test, and outliers test) confirmed that there were residual patterns for all the three link functions. A slight pattern in residuals indicated good reliability that stemborer infestation proportions were adequately modelled using a beta-binomial mixed-effect model with a clog-log link function [27]. Residual analysis is useful to check the fitted model's quality and the underlying assumptions made in the model construction. Using the clog-log link function, the models estimated the year effect as -0.20, -0.19, -0.33 and -0.27 for conventional high vs organic high and conventional low vs organic low systems in Thika and Chuka, respectively. A negative sign indicates a decrement of stemborer infestation rates over the years, probably because of use of biopesticides and intercropping in the farming systems.

6. Conclusion and Recommendations

From the analysis of overdispersed binomial proportions from LTE, it can be concluded that the choice of link function is essential. Based on the results of this study, it was supposed that complementary log-log link functions fit the asymmetric distribution of the response variable better than its counterparts. Using the appropriate link function results in a superior fit for the beta-binomial mixed-effect regression model. Thus, we recommend that complementary log-log link function be considered in modelling binomial proportions with a left-skewed distribution. Navigating these challenges requires developing and selecting link functions using data modifications to understand model behaviour and guard against overdispersed binomial proportions. Link functions must be evaluated holistically with a diverse set of model diagnostics to avoid selecting and relying on non-fitted models that do not properly describe pest incidence over a long time in incipient areas of establishment and spread invasion. Our study suggests that incidence and background data modifications should be implemented when modelling native species to minimize model overfit and spatial biases. Finally, interpretation of these link functions as the description of a species' incidence on long-term farming systems should be done with caution and include additional sources of evidence beyond the current approach used in the present studies. However, users of long-term species incidence data need to incorporate explicit consideration of model discrimination, model fit and model complexity into their decision-making processes if they are to build biologically realistic models.

Acknowledgements

We are grateful to the SysCom Kenya team for facilitating and supporting the study, DMMG unit at icipe and JKUAT team for assistance with data analysis, and language editing of the manuscript. We acknowledge project financial support from Biovision Foundation (Grant No: 1040), Coop Sustainability Fund (Grant No: 1040), Liechtenstein Development Service (LED) (Grant No: 1040) and Swiss Agency for Development and Cooperation (SDC) (Grant No: 1040). We also gratefully acknowledge the support of icipe core funding provided by United Kingdom's Foreign, Commonwealth and Development Office (FCDO); the Swedish International Development Cooperation Agency (Sida); the Swiss Agency for Development and Cooperation (SDC); the Federal Democratic Republic of Ethiopia; and the Government of the Republic of Kenya.

References

[1] S. Mugo, J. Songa, H. Degroote, and D. Hoisington, "Insect Resistant Maize for Africa (IRMA) Project : An overview," Perspect. Evol. Role Priv. Collab. Agric. Res., vol. 522879, no. June 2014, pp. 1–16, 2002.

[2] D. Collett, Modelling binary data. 2002.

[3] R. B. Prasetyo, H. Kuswanto, N. Iriawan, and B. S. S. Ulama, "Binomial Regression Models with a Flexible Generalized Logit Link Function," Symmetry (Basel)., vol. 12, no. 2, p. 221, Feb. 2020, doi: 10.3390/sym12020221.

[4] A. Lott and J. P. Reiter, "Wilson Confidence Intervals for Binomial Proportions With Multiple Imputation for Missing Data," Am. Stat., vol. 74, no. 2, pp. 109–115, 2020, doi: 10.1080/00031305.2018.1473796.

[5] G. Molenberghs, G. Verbeke, C. G. B. Demétrio, and A. M. C. Vieira, "A Family of generalized linear models for repeated measures with normal and conjugate random effects," Stat. Sci., vol. 25, no. 3, pp. 325–347, 2010, doi: 10.1214/10-STS328.

[6] A. M. Malekfar and F. Eskandari, "Risk Assessment of Interval-Valued Variables in Generalized Linear Models," J. Comput. Stat. Model., vol. 1, no. 2, pp. 39–65, 2021.

[7] H. Jeong, E. A. Valdez, J. Y. Ahn, and S. Park, "Generalized Linear Mixed Models for Dependent Compound Risk Models," SSRN Electron. J., vol. 26, no. September, pp. 1318–1342, 2017, doi: 10.2139/ssrn.3045360.

[8] I. Arostegui, V. Núñez-Antón, and J. M. Quintana, "Analysis of the short form-36 (SF-36): the beta-binomial distribution approach," Stat. Med., vol. 26, no. 6, pp. 1318–1342, Mar. 2007, doi: 10.1002/sim.2612.

[9] A. van den Hout and G. Muniz-Terrera, "Joint models for discrete longitudinal outcomes in aging research," J. R. Stat. Soc. Ser. C Appl. Stat., vol. 65, no. 1, pp. 167–186, 2016, doi: 10.1111/rssc.12114.

[10] T. Mathes and O. Kuss, "Beta-binomial models for meta-analysis with binary outcomes: Variations, extensions, and additional insights from econometrics," Res. Methods Med. Heal. Sci., vol. 2, no. 2, pp. 82–89, 2021, doi: 10.1177/2632084321996225.

[11] R. B. Prasetyo, H. Kuswanto, N. Iriawan, and B. S. S. Ulama, "A comparison of some link functions for binomial regression models with application to school drop-out rates in East Java," AIP Conf. Proc., vol. 2194, no. December 2019, 2019, doi: 10.1063/1.5139815.

[12] J. Najera-Zuloaga, D. J. Lee, and I. Arostegui, "A beta-binomial mixed-effects model approach for analyzing longitudinal discrete and bounded outcomes," Biometrical J., vol. 61, no. 3, pp. 600–615, 2019, doi: 10.1002/bimj.201700251.

[13] M. Smithson and J. Verkuilen, "A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables," Psychol. Methods, vol. 11, no. 1, pp. 54–71, 2006, doi: 10.1037/1082-989X.11.1.54.

[14] F. P. Leacy and E. A. Stuart, "On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study," Stat. Med., vol. 33, no. 20, pp. 3488–3508, Sep. 2014, doi: 10.1002/sim.6030.

[15] A. M. Wilson and K. E. Holsinger, "A NEW CLASS OF FLEXIBLE LINK FUNCTIONS WITH APPLICATION TO SPECIES CO- OCCURRENCE IN CAPE FLORISTIC REGION Author (s): Xun Jiang, Dipak K. Dey, Rachel Prunier, Adam M. Wilson and Kent E. Holsinger Source : The Annals of Applied Statistics, Vol.," vol. 7, no. 4, pp. 2180–2204, 2018.

[16] D. Adamtey, N., Musyoka, M. W., Zundel, C., Cobo, J. G., Karanja, E., Fiaboe, K. K. M., & Foster, "Productivity, profitability and partial nutrient balance in maize-based conventional and organic farming systems in Kenya," Agric. Ecosyst. Environ., vol. 235, pp. 61–79, Nov. 2016, doi: 10.1016/j.agee.2016.10.001.

[17] J. C. Douma and J. T. Weedon, "Analyzing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression," Methods Ecol. Evol., vol. 10, no. 9, pp. 1412–1430, 2019, doi: 10.1111/2041-210X.13234.

[18] Y. Li and X. Deng, "An efficient algorithm for Elastic I-optimal design of generalized linear models," Can. J. Stat., vol. 49, no. 2, pp. 438–470, 2021, doi: 10.1002/cjs.11571.

[19] M. Razzaghi, "The probit link function in generalized linear models for data mining applications," J. Mod. Appl. Stat. Methods, vol. 12, no. 1, pp. 164–169, 2013, doi: 10.22237/jmasm/1367381880.

[20] F. Hartig, "Package 'DHARMa'. Version 0.3.0," R Packag., no. Version 0.3.0, 2020.

[21] A. Gelman and J. Hill, Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge: Cambridge University Press, 2006.

[22] R. C. Team, "A language and environment for statistical computing," R Found. Stat. Comput. Vienna, Austria., 2021, [Online]. Available: https://www.r-project.org/.

[23] D. I. Warton and F. K. C. Hui, "The arcsine is asinine: the analysis of proportions in ecology," Ecology, vol. 92, no. 1, pp. 3–10, Jan. 2011, doi: 10.1890/10-0340.1.

[24] D. R. S. Saputro, A. Susanti, and N. B. I. Pratiwi, "The handling of overdispersion on Poisson regression model with the generalized Poisson regression model," in In AIP Conference Proceedings, 2021, p. 020026, doi: 10.1063/5.0040330.

[25] T. V. Pham, S. R. Piersma, M. Warmoes, and C. R. Jimenez, "On the beta-binomial model for analysis of spectral count data in label-free tandem mass spectrometry-based proteomics," Bioinformatics, vol. 26, no. 3, pp. 363–369, Feb. 2010, doi: 10.1093/bioinformatics/btp677.

[26] O. Kuss, A. Hoyer, and A. Solms, "Meta-analysis for diagnostic accuracy studies: a new statistical model using beta-binomial distributions and bivariate copulas," Stat. Med., vol. 33, no. 1, pp. 17–30, Jan. 2014, doi: 10.1002/sim.5909.

[27] R. M. Pires and C. A. R. Diniz, "Bayesian residual analysis for beta-binomial regression models," in AIP Conference Proceedings, 2012, pp. 259–267, doi: 10.1063/1.4759610.