International Journal of Data Science and Analysis
2022; 8(1): 1-10
http://www.sciencepublishinggroup.com/j/ijdsa
doi: 10.11648/j.ijdsa.20220801.11
ISSN: 2575-1883 (Print); ISSN: 2575-1891 (Online)

Analysis of Overdispersed Insect Count Data from an Avocado Plantation in Thika, Kenya

Eric Ali Ibrahim[1, 2, 3], Daisy Salifu[1, *], Samuel Musili Mwalili[3], Thomas Dubois[2], Henri Edouard Zefack Tonnang[1]

[1]Data Management, Modelling, and Geo-Information Unit, International Centre of Insect Physiology and Ecology (Icipe), Nairobi, Kenya

[2]Plant Health Theme, International Centre of Insect Physiology and Ecology (Icipe), Nairobi, Kenya

[3]Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

Email address:

dsalifu@icipe.org (D. Salifu)

[*]Corresponding author

To cite this article:

Eric Ali Ibrahim, Daisy Salifu, Samuel Musili Mwalili, Thomas Dubois, Henri Edouard Zefack Tonnang. Analysis of Overdispersed Insect Count Data from an Avocado Plantation in Thika, Kenya. *International Journal of Data Science and Analysis*. Vol. 8, No. 1, 2022, pp. 1-10. doi: 10.11648/j.ijdsa.20220801.11

Received: December 10, 2021; **Accepted**: January 4, 2022; **Published**: February 16, 2022

Abstract: Avocado (*Persea americana*) farming in East Africa has expanded since recent, contributing significantly toward economic growth and livelihood for small-scale farmers. However, insects attacking avocado fruits reduce fruit quality and size, causing massive losses. Previous studies have identified key avocado insect pests, their temporal population patterns and how landscape vegetation productivity influences their population dynamics. This research analyzed insect count data collected on *Bactrocera dorsalis* and *Ceratitis* spp. in an avocado plantation in Thika, Kenya over a successive period of time, as part of pest management. These data are characterized by overdispersion due to aggregation behaviour of the insects in their habitat and serial correlations since the count data were collected over a successive period of time. Analyzing these data becomes complicated because of overdispersion and the serial correlation in the data. In this study, we explored variants of generalized linear models (GLMs) with a sinusoidal component over time; and with and without timescale decomposition of covariates (weather variables). All GLM variants were fitted assuming the negative binomial distribution to account for overdispersion. Based on the Akaike information criterion (AIC), GLMs with decomposed covariates had lower AIC values than GLMs without decomposed covariates for both *B. dorsalis* and *Ceratitis* spp., and therefore GLMs with a sinusoidal component and decomposed covariates under negative binomial distribution were the best choice for these data. The contribution of the preceding weekly insect pest counts in all models was statistically significant. The study established that both abiotic and biotic factors drive insect pest infestation.

Keywords: Overdispersion, Negative Binomial Distribution, Sinusoidal Component, Time Series Count Data

1. Introduction

Insect pest count data, most of which is non-Gaussian distributed, is increasingly being collected in various research fields [1, 2]. Insect pest counts collected over a successive period of time are characterized by overdispersion and serial correlations, making modeling of such data complicated [3] since time series modeling techniques are adopted for Gaussian distributed data. Commonly used techniques for modeling Gaussian time series data include Box and Jenkins models, which capture the temporal correlation structure of time series [4, 5]. If adopted for the case of time series count data, Box and Jenkins models result in biased estimates, and fail to capture the distribution of count data thus leading to poor models [6]. In addition, when modeling count time series data with a low number of observations, Box and Jenkins models result in inadequate performance. Box and Jenkins models also do not account for overdispersion or excess zeros in the time series count data, which lead to biased estimates in the models if not adequately accounted for. It is therefore important to model count time series data applying adequate statistical approaches that take into account the underlying conditional

distribution, possibilities of excess zeros, overdispersion, positive or negative associations among the counts, and serial dependencies [4, 7, 8].

Efforts to overcome the limitations of Box and Jenkins models in modeling count time series data resulted in the use of the integer-valued autoregressive (INAR) class of Poisson models [9-10]. These models apparently perform better in modeling count time series data but are limited in dealing with seasonality and unobserved heterogeneity [5]. Machine learning algorithms such as artificial neural networks (ANN) and long-short term memory (LSTM) networks have been acknowledged as powerful in modeling count time series data, and do not require consideration of the conditional distribution of the outcome variable [11].

Recently, generalized linear models (GLMs), which are a generalization of the ordinary linear model, have been explored for allowing response variables to assume a non-Gaussian distribution, and for extension to modeling count data recorded over a period of time. GLMs have over the years been applied in modeling cross-sectional count data, allowing the linear model to be related to the response variable via a link function. In such contexts, the dynamics are not of primary concern, but heterogeneity [12]. Preliminary reports on use of GLMs in modeling count data collected over a period of time, where the conditional distribution of the response variable follows Poisson distribution, have revealed that they do so in a parsimonious manner [6].

This paper focuses on aggregated and overdispersed count data that was collected over a period of time, through scouting of pests (*Bactrocera dorsalis* and *Ceratitis* spp.) in avocado (*Persea americana*) fields at Kakuzi PLC (Thika, Kenya). The resulting time series data is characterized by heterogeneity of variance, serial correlations and delayed effects of the predictor variables. In modeling insect count time series data, a study forecasted whitefly and aphid populations using an autoregressive integrated moving average (ARIMA) [13]. ARIMA with exogenous variables (ARIMAX) was used for modelling and forecasting incidence of greenhouse whitefly (*Trialeurodes vaporariorum*) in green houses [11]. The approaches used were for a Gaussian distributed data.

Most researchers adopted GLMs in modeling time series data in the context of epidemiology, where the conditional distribution of the response variable follows a Poisson distribution. Examples include applying GLMs in analyzing non-Gaussian time series data, focusing on incidence of respiratory syncytial virus infection, with Poisson distribution as the underlying conditional distribution of the response variable [14]. On the contrary, insect count data, collected over a period of time, are characterized by overdispersion, hence they are not Poisson-distributed. Studies that used machine learning algorithms in predictive modeling of pest population include prediction of host-parasitoid population using ANN [15], prediction of the severity of *Spodoptera litura* on groundnuts using ANN [16], and forecasting crop attacks by pest using long short-term memory (LSTM) [17].

Studies on avocado farming in Kenya have reported presence and abundance of tephritid fruit flies, especially *B. dorsalis* and various *Ceratitis* spp. (*Ceratitis cosyra*, *Ceratitis capitata* and *Ceratitis rosa*) [18-20]. In addition, climatic factors and avocado plant physiology stages influence insect pest species population densities [16, 19]. Temperature was identified as a key abiotic factor influencing both the distribution and the population dynamics of *B. dorsalis* through affecting their development, survival and reproduction [21]. Varying temperature was reported to have a negative effect on the developmental stages of *B. dorsalis* [22]. Regarding rainfall effect on insect pests, rainfall did not have a significant effect on emergence and survival of the tephritid fruit fly *Anastrepha* spp. attributing yearly population fluctuations to other factors such as fruiting physiology stages of hosts plants [23]. An increase in fruit fly populations was reported during the rainy season and fruiting phenology of hosts, while *Ceratitis* spp. population densities increased during fruiting of hosts plants and relatively dry period [22]. This study explored the variants of GLMs that can be used to analyse overdispersed fruit fly count data to determine the factors that drive insect infestation.

2. Material and Methods

2.1. Available Data

Data comprised of weekly trap counts of *B. dorsalis* and *Ceratitis* spp. collected from two orchards, A and B, in an avocado plantation at Kakuzi PLC. Two traps, one for each taxon, were placed equidistant and at the same elevation within a given orchard. Traps (McPhail traps, Insect Science Company, Tzaneen, South Africa) had been baited with CC EGO Lure (Kenya Biologics, Nairobi, Kenya) and ME Lure (Kenya Biologics) for *B. dorsalis* and *Ceratitis* spp., respectively. We further used the following weather variables: daily total rainfall (mm), daily average temperature (°C) and daily relative humidity (%) as the predictor variables of weekly trap counts. These variables were processed into weekly total rainfall (mm), weekly average temperature (°C) and weekly average relative humidity (%) to align with the weekly measure of insect pest counts.

2.2. Statistical Modeling

GLMs were used to model the data. As described by Nelder and Wedderburn (1972), GLMs contain the following three components [24]. The random component specifies the conditional distribution of the response variable Y_i (for i^{th} of the n independently sampled observations) given the predictor variables in the model. The systemic component includes the linear function of the predictors, and is defined as;

$$\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} \quad (1)$$

and finally, the linearizing link function denoted as g(.), which transforms the expectation of the response variable E(Yi)=μ to k linear predictors and is defined as;

$$g(\mu_i) = \eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} \quad (2)$$

To account for overdispersion, the conditional distribution for this case was considered as a negative binomial

distribution. Negative binomial distribution belongs to the GLM family, and is an extension of Poisson distribution that allows for modeling overdispersed count data, where the Poisson mean is itself a random variable, distributed according to a Gamma distribution. The probability mass function of negative binomial distribution is defined as;

$$f(Y_t = y|X; v, \mu_t) = \frac{\Gamma(v+y)}{\Gamma(v)\Gamma(y+1)} \left(\frac{v}{v+\mu_t}\right)^v \left(\frac{\mu_t}{v+\mu_t}\right)^y \quad (3)$$

with $E(Y_t)=\mu_t$ and $Var(Y_t)=\mu_t + \frac{\mu_t^2}{v}$ where v is the dispersion parameter and μ_t is the mean at time t. Insects counts are often fitted fairly well by the negative binomial distribution [25].

2.2.1. Definition of Models

(i). GLM with Sinusoidal Component

A GLM with sinusoidal component is defined as follows [14];

$$\log \mu = \beta_0 + \alpha Y_{t-1} + \sum_{t=1}^{n}\sum_{k=1}^{m}\sum_{l=0}^{q}\beta_{kl}X_{(t-l)k} + \eta_1 sin\left[\frac{2\pi t}{T}\right] + \eta_2 cos\left[\frac{2\pi t}{T}\right] \quad (4)$$

where Y_t, $t=1,2,....,n$ is the insect pest count recorded per week at corresponding time t. μ_t is the mean count for time t. β_0 is the intercept, α is coefficient of the lagged weekly pest counts by time t-1. $X_{(t-l)k}$ denotes the predictor variable; β_{kl} is the coefficient of the predictor variable with k=1,2..,.m covariates, l=0,1...q is the distributed lags while q is the maximum lag and t=1,...n are the time points. η_1 and η_2 are the coefficients of the sine and cosine functions, respectively, while T is the number of time periods described by one cosine function over the interval [0, 2π].

(ii). GLM with Sinusoidal Component and Decomposed Predictors

The second model, the GLM with sinusoidal component incorporating decomposition, is defined as;

$$\log \mu = \beta_0 + \alpha Y_{t-1} + \sum_{t=1}^{n}\sum_{k=1}^{m}\sum_{s=0}^{r}\sum_{l=0}^{q}\beta_{ksl}X_{(t-l)ks} + n_1 cos\left[\frac{2\pi t}{T}\right] + n_2 sin\left[\frac{2\pi t}{T}\right] \quad (5)$$

where Y_t, t=1, 2,..., n is the insect pest counts recorded per week at corresponding time t. μ_t is the mean for time t. β_0 is the intercept, α is the coefficient of the lagged weekly insect pest species counts by time t-1. $X_{(t-l)ks}$ denotes the decomposed predictor variable; β_{ksl} is the coefficients of the decomposed predictor variables with k=1, 2, …, m covariates, s=1, 2,..., r is the decomposition, l=0,1,..., q is the distributed lags while q is the maximum lag and t=1,...n are the time points. η_1 and η_2 are the coefficients of the sine and cosine functions, respectively, while T is the number of time periods described by one cosine function over the interval [0, 2π]. To decompose the covariates to trend, seasonal and remainder components, moving average (MA) was used, which is a common linear filter defined as;

$$\hat{T}_t = \frac{1}{m}\sum_{j=-k}^{k} X_{t+j} \quad (6)$$

where $\hat{T}_t$ is the estimated trend at time t, m is the order of MA and m=2k+ 1. The estimate of the trend-cycle at time t is obtained by averaging values of the time series within k periods of t. The estimate of seasonal components is;

$$\hat{s}_t = Y_t - \hat{T}_t \quad (7)$$

while the random component is obtained by;

$$\hat{e}_t = Y_t - \hat{T}_t - \hat{s}_t \quad (8)$$

2.2.2. Estimation of Parameters

From the log-likelihood of the negative binomial extended GLM, defined as;

$$l = \sum_{t=1}^{n}\left(y_t \log(\mu_t) + vlog(v) - (v + y_t)\log(v+\mu_t) + log\left(\frac{\Gamma(v+y_t}{\Gamma(v)}\right) - \log(y_t!)\right) \quad (9)$$

the maximum likelihood estimates of β and v, that is $\hat{\beta}$ and $\hat{v}$, in the GLM were estimated simultaneously, by sequential iterations solved by Newton–Raphson iterative scheme, using the iterated weighted least squares (IWLS) algorithm.

2.2.3. Model Evaluation

The models were evaluated using the Akaike information criteria (AIC), which accounts for the likelihood of the observations as well as the number of parameters in the model. AIC is defined as;

$$AIC = -2(\hat{L}) + 2K \quad (10)$$

where K is the number of estimated parameters in the model and $\hat{L}$ is the maximum value of the likelihood function for the model [26].

Further, deviance squared (D^2), proposed by Guisan and Zimmermann [27], was used to determine the amount of deviance accounted for by each GLM. D^2 is obtained by;

$$D^2 = 1 - \frac{null\ deviance}{residual\ deviance} \quad (11)$$

where null deviance is the GLM's null deviance while residual deviance is the GLM's residual deviance.

The models were implemented in R software version 4.0.5 [28], using the "MASS" package [29], "dlnm" package [30] and "lubridate" package [31].

3. Results

3.1. Assessment of the Fitted Generalized Linear Models

We fitted the variants of GLM with sinusoidal components over time, then compared their performances.

3.1.1. Model for B. dorsalis

The results of the GLMs are presented in Table 1. The GLMs were fitted for the *B. dorsalis* counts in orchard A, without and with decomposition of covariates, achieved a D^2 of 0.489 and 0.712 respectively. The corresponding AIC values were 1120.7 and 1032.1, respectively, thus GLM with a sinusoidal component, and incorporating timescale decomposition of covariates, fitted better to *B. dorsalis* count data. The lagged *B dorsalis* counts had a significant effect on weekly *B. dorsalis* counts at 5% significance level. Rainfall and average temperature variables significantly and negatively affected the weekly *B. dorsalis* counts. Avocado plant physiology stages of fruit set and rapid expansion, development and harvesting had positive beta coefficients, implying positive effects on the weekly counts, though statistically insignificant. The results could be due to increased frequency in implementation of pest control strategies, especially during such stages, to ensure that insect counts remained below the set threshold. Relative humidity affected the weekly *B. dorsalis* counts significantly at 10% significance level.

Table 1. *Model parameter estimates for generalized linear models with sinusoidal components and decomposed covariates and without decomposition of covariates for* B. dorsalis *in orchard A.*

Covariate	Orchard A			
	GLM[a]		GLM[b]	
	Estimate (se)		Estimate (se)	
(Intercept)	1.717	(0.832) *	7.437	(1.904) ***
B. dorsalis counts (t-1)	0.006	(0.002) ***	0.006	(0.001) ***
Avocado plant physiology cycle				
Flowering and fruitset	-0.158	(0.517)	-0.910	(0.472).
Fruitset and rapid expansion	1.014	(0.689)	0.271	(0.744)
Fruit development	0.710	(0.671)	0.845	(0.765)
Harvesting	0.018	(0.533)	0.474	(0.461)
Rainfall	-0.487	(0.783)	-	-
Average temperature	-0.918	(1.099)	-	-
Relative humidity	1.592	(0.875).	-	-
Rainfall (seasonality)	-	-	-1.199	(0.640).
Rainfall (trend)	-	-	-7.283	(0.872) ***
Average temperature (seasonality)	-	-	0.165	(1.097)
Average temperature (trend)	-	-	-15.475	(3.567) ***
Relative humidity (seasonality)	-	-	0.656	(1.035)
Relative humidity (trend)	-	-	-3.541	(3.763)
Sinusoidal component				
F(t)	-0.003	(0.003)	0.058	(0.016) ***
$\text{Sine}\left(\frac{2\pi t}{T}\right)$	-0.893	(0.593)	0.901	(1.009)
$\text{Cosine}\left(\frac{2\pi t}{T}\right)$	-0.038	(0.557)	1.288	(0.651) *
$\text{F(t): Sine}\left(\frac{2\pi t}{T}\right)$	0.008	(0.005	0.004	(0.012)
$\text{F(t): Cosine}\left(\frac{2\pi t}{T}\right)$	0.010	(0.007)	0.012	(0.008)
Null deviance	407.98, df=170		670.10, df=170	
Residual deviance	192.48, df=157		174.84, df=154	
AIC	1120.7		1032.1	
Theta	0.629	(0.080)	1.142	(0.157)
2 x log-likelihood	1090.749		-996.111	
D-Squared	0.489		0.712	

[a]Model without covariates decomposition. [b]Model with covariates decomposition. The values in parentheses are standard errors of model parameter estimates. Significance: ***p < 0.001, ** p < 0.01 *p <0.05

Table 2 shows parameter estimates for generalized linear models in orchard B. The GLMs fitted for the *B. dorsalis* counts, without and with decomposition of covariates, achieved a D^2 of 0.620 and 0.633 respectively. The corresponding AICs were 956.62 and 956.35, respectively, implying that there was slight improvement following use of decomposed covariates. The lagged *B. dorsalis* counts and relative humidity significantly affected the weekly *B. dorsalis* counts positively at 5% significant level. Avocado plant physiology stages had no significant effect on the weekly counts, which could be attributed to frequent control of insect pests to ensure that they do not surpass the set threshold.

Table 2. *Model parameter estimates for generalized linear models with sinusoidal components and decomposed covariates and without decomposition of covariates for* B. dorsalis *in orchard B.*

Covariates	Orchard B			
	GLM[a]		GLM[b]	
	Estimate (se)		Estimate (se)	
(Intercept)	2.431	(0.773) **	8.663	(2.993) **
B. dorsalis counts (t-1)	0.008	(0.003) **	0.006	(0.003) *

Covariates	Orchard B			
	GLM[a]		GLM[b]	
	Estimate (se)		Estimate (se)	
Avocado plant physiology cycle				
Flowering and fruitset	-1.332	(0.393) ***	-1.201	(0.441) **
Fruitset and rapid expansion	0.451	(0.324)	0.053	(0.520)
Fruit development	0.365	(0.410)	0.590	(0.459)
Harvesting	0.063	(0.363)	0.722	(0.378).
Rainfall	-0.186	(0.658)	-	-
Average temperature	-0.218	(0.883)	-	-
Relative humidity	2.584	(0.786) **	-	-
Rainfall (seasonality)	-	-	-1.874	(0.662) **
Rainfall (trend)	-	-	-7.281	(12.744)
Average temperature (seasonality)	-	-	1.022	(0.948)
Average temperature (trend)	-	-	12.738	(7.561).
Relative humidity (seasonality)	-	-	0.699	(0.833)
Relative humidity (trend)	-	-	1.015	(8.540)
Sinusoidal component				
F(t)	-0.023	(0.004) ***	-0.117	(0.084)
Sine$\left(\frac{2\pi t}{T}\right)$	-2.422	(0.507) ***	-3.082	(1.322) *
Cosine$\left(\frac{2\pi t}{T}\right)$	-0.944	(0.601)	-7.376	(8.837)
F(t): Sine$\left(\frac{2\pi t}{T}\right)$	0.034	(0.005) ***	0.107	(0.023) ***
F(t): Cosine $\left(\frac{2\pi t}{T}\right)$	0.032	(0.007) ***	0.157	(0.109)
Null deviance	543.13, df=170		586.55, df=170	
Residual deviance	190.36, df=157		195.21, df=154	
AIC	956.62		956.35	
Theta	1.071	(0.167)	1.185	(0.195)
2 x log-likelihood	-926.618		-920.351	
D-Squared	0.620		0.633	

[a]Model without covariates decomposition. [b]Model with covariates decomposition. The values in parentheses are standard errors of model parameter estimates. Significance: ***p < 0.001, ** p < 0.01 *p <0.05

3.1.2. Model for Ceratitis spp.

For orchard A, the fitted GLM without decomposition of covariates achieved a D^2 of 0.814 (AIC=1598) as shown in Table 3. However, using decomposed variables, the GLM had a D^2 of 0.838 with corresponding AIC of 1578.2. At 5% significance level, the lagged counts of *Ceratitis* spp. had statistically significant effect on weekly *Ceratitis* spp. counts. Avocado plant physiology stages of fruit set and rapid expansion, development and harvesting (with reference to dormant stage) had a positive effect on the weekly *Ceratitis* spp. counts, though statistically insignificant.

Table 3. *Model parameter estimates for generalized linear models with sinusoidal components and decomposed covariates and without decomposition of covariates for* Ceratitis *spp. in orchard A.*

	Orchard A			
	GLM[a]		GLM[b]	
	Estimated (se)		Estimated (se)	
(Intercept)	2.120	(0.430) ***	0.019	(1.010)
Ceratitis spp. counts (t-1)	0.002	(0.000) ***	0.001	(0.000) ***
Avocado plant physiology cycle				
Flowering and fruitset	-0.132	(0.219)	-0.009	(0.282)
Fruitset and rapid expansion	0.106	(0.247)	0.350	(0.388)
Fruit development	0.448	(0.330)	0.338	(0.366)
Harvesting	0.398	(0.269)	0.124	(0.253)
Average rainfall	0.079	(0.427)	-	-
Average temperature	-0.567	(0.563)	-	-
Relative humidity	0.119	(0.479)	-	-
Rainfall (seasonality)	-	-	-0.197	(0.394)
Rainfall (trend)	-	-	0.276	(1.419)
Average temperature (seasonality)	-	-	0.100	(0.586)
Average temperature (trend)	-	-	3.981	(1.625) *
Relative humidity (seasonality)	-	-	-1.149	(0.626).
Relative humidity (trend)	-	-	4.865	(2.950).
Sinusoidal component				
F(t)	0.016	(0.001) ***	-0.001	(0.007)
Sine$\left(\frac{2\pi t}{T}\right)$	-0.405	(0.243).	-0.315	(0.949)
Cosine$\left(\frac{2\pi t}{T}\right)$	0.670	(0.292) *	1.176	(0.656).
F(t):sine$\left(\frac{2\pi t}{T}\right)$	-0.005	(0.003).	0.002	(0.009)
F(t): Cosine $\left(\frac{2\pi t}{T}\right)$	-0.002	(0.003)	-0.008	(0.008)

	Orchard A			
	GLM[a]		GLM[b]	
	Estimated (se)		Estimated (se)	
Null deviance	1112.98, df=170		1320.85, df=170	
Residual deviance	191.69, df=157		194.18, df=154	
AIC	1598		1578.2	
Theta	2.258	(0.266)	2.714	(0.334)
2 x log-likelihood	-1568.013		-1542.158	
D-Squared	0.814		0.838	

[a]Model without covariates decomposition. [b]Model with covariates decomposition. The values in parentheses are standard errors of model parameter estimates. Significance: ***p < 0.001, ** p < 0.01 *p <0.05.

In orchard B, the fitted GLM without decomposition had a D^2 of 0.761 (AIC=1681.4). However, the GLM with decomposed variables achieved a D^2 of 0.776 with corresponding AIC of 1673. The previous weekly counts of *Ceratitis* spp., avocado plant stages of fruits development and harvesting, and relative humidity had a statistically significant, positive relationship with the weekly *Ceratitis* spp. counts at 5% significance level. Orchard B bordered mango (*Mangifera indica*) farms and forests with fig trees, which acted as alternative host for the fruit fly. Hence, this could be a potential explanation for variations in the models in orchard B, despite controls. Table 4 presents the results.

Table 4. *Model parameter estimates for generalized linear models with sinusoidal components and decomposed covariates and without decomposition of covariates for* Ceratitis *spp. in orchard B.*

	Orchard B			
	GLM[a]		GLM[b]	
	Estimate (se)		Estimate (se)	
(Intercept)	3.282	(0.619) ***	3.307	(1.983).
Ceratitis spp. counts (t-1)	0.002	(0.000) ***	0.001	(0.000) ***
Avocado plant physiology cycle				
Flowering and fruitset	-0.111	(0.246)	-0.010	(0.245)
Fruitset and rapid expansion	-0.272	(0.282)	-0.079	(0.354)
Fruit development	0.955	(0.384) *	1.149	(0.442) **
Harvesting	1.028	(0.298) ***	0.841	(0.294) **
Average rainfall	-0.343	(0.460)	-	-
Average temperature	0.553	(0.608)	-	-
Relative humidity	1.095	(0.496) *	-	-
Rainfall (seasonality)	-	-	0.345	(0.445)
Rainfall (trend)	-	-	0.706	(0.501)
Average temperature (seasonality)	-	-	0.239	(0.662)
Average temperature (trend)	-	-	0.896	(4.472)
Relative humidity (seasonality)	-	-	-1.258	(0.627) *
Relative humidity (trend)	-	-	1.366	(2.515)
Sinusoidal component				
F(t)	-0.009	(0.005).	-0.013	(0.014)
Sine$\left(\frac{2\pi t}{T}\right)$	-3.491	(0.660) ***	-3.487	(1.565) *
Cosine$\left(\frac{2\pi t}{T}\right)$	-1.345	(0.440) **	-1.303	(0.526) *
F(t):sine $\left(\frac{2\pi t}{T}\right)$	0.047	(0.008) ***	0.045	(0.014) **
F(t): Cosine $\left(\frac{2\pi t}{T}\right)$	0.019	(0.005) ***	0.020	(0.007) **
Null deviance	872.19, df=170		955.81, df=170	
Residual deviance	192.41, df=157		193.66, df=154	
AIC	1681.4		1673	
Theta	1.869	(0.215)	2.061	(0.243)
2 x log-likelihood	-1651.368		-1636.952	
D-Squared	0.761		0.776	

[a]Model without covariates decomposition. [b]Model with covariates decomposition. The values in parentheses are standard errors of model parameter estimates. Significance: ***p < 0.001, ** p < 0.01 *p <0.05

3.2. Analysis of Deviance (Type III Tests)

To determine the unique contribution of each predictor variable while controlling for other variables, type III tests (likelihood chi-square statistic) results were obtained for each model, and presented in Tables 5 and 6. In all the models, the contribution of preceding weekly counts of insect pest species were statistically significant.

Table 5. *Type III tests (likelihood chi-square statistic) for predictor variables for generalized linear model without covariate decomposition and with covariate composition for* B. dorsalis *trap counts in orchard A and B, respectively.*

	Orchard A		Orchard B		Df
	LR Chisq		LR Chisq		
	GLM[a]	GLM[b]	GLM[a]	GLM[b]	
B. dorsalis counts (t-1)	15.965***	17.705***	8.628**	4.1548 *	1
Plant physiology stages	5.897	6.925	29.381***	18.2862**	4
Rainfall	0.284	-	0.077	-	1
Average temperature	0.651	-	0.050	-	1
Relative humidity	3.045.	-	9.811**	-	1
Rainfall (seasonality)	-	3.405.	-	8.3686**	1
Rainfall (trend)	-	67.443***	-	0.2939	1
Average temperature (Seasonality)	-	0.018	-	1.1135	1
Average temperature (trend)	-	19.731***	-	2.744.	1
Relative humidity (seasonality)	-	0.341		0.7255	1
Relative humidity (trend)	-	0.799	-	0.0123	1
Sinusoidal component					
F(t)	1.863	13.730***	28.259***	1.8343	1
$\text{Sine}\left(\frac{2\pi t}{T}\right)$	1.987	0.684	23.062***	4.8222*	1
$\text{Cosine}\left(\frac{2\pi t}{T}\right)$	0.004	3.757.	2.335	0.6426	1
$\text{F(t):sine}\left(\frac{2\pi t}{T}\right)$	2.318	0.071	40.745***	21.5216***	1
$\text{F(t): Cosine }\left(\frac{2\pi t}{T}\right)$	1.924	2.100	18.344***	1.9093	1

[a]Model without covariates decomposition. [b]Model with covariates decomposition. LR Chisq – likelihood ratio chi-square, *B. dorsalis* counts (t-1) is the lagged counts of *B. dorsalis*. Significance: ***p < 0.001, ** p < 0.01 *p <0.05.

Table 6. *Type III tests (likelihood chi-square statistic) for predictor variables for generalized linear model without covariate decomposition and with covariate composition for* Ceratitis *spp. trap counts in orchard A and B respectively.*

	Orchard A		Orchard B		Df
	LR Chisq		LR Chisq		
	GLM[a]	GLM[b]	GLM[a]	GLM[b]	
Ceratitis spp. counts (t-1)	61.401***	27.779***	38.109***	27.634***	1
Plant physiology stages	3.948	1.317	30.337***	17.659**	4
Rainfall	0.031	-	0.490	-	1
Average temperature	0.948	-	0.774	-	1
Relative humidity	0.060	-	4.318*	-	1
Rainfall (seasonality)	-	0.251	-	0.594	1
Rainfall (trend)	-	0.036	-	1.742	1
Average temperature (seasonality)	-	0.027	-	0.116	1
Average temperature (trend)	-	5.375*	-	0.039	1
Relative humidity (seasonality)	-	3.249.	-	4.133*	1
Relative humidity (trend)	-	2.430	-	0.284	1
Sinusoidal component					
F(t)	122.024***	0.037	3.226.	0.786	1
$\text{Sine}\left(\frac{2\pi t}{T}\right)$	2.605	0.114	24.219 ***	4.722*	1
$\text{Cosine}\left(\frac{2\pi t}{T}\right)$	4.607*	2.959.	9.037 **	6.309*	1
$\text{F(t):sine}\left(\frac{2\pi t}{T}\right)$	2.717.	0.043	31.227 ***	9.687**	1
$\text{F(t): Cosine }\left(\frac{2\pi t}{T}\right)$	0.379	0.927	14.794***	8.539**	1

[a]Model without covariates decomposition. [b]Model with covariates decomposition. LR Chisq – likelihood ratio chi-square, *Ceratitis* spp. counts (t-1) is the lagged counts of *Ceratitis* spp. Significance: ***p < 0.001, ** p < 0.01 *p <0.05.

Table 7. *The AIC and deviance for the GLM models with and without covariate decomposition.*

Pests	Orchards	AIC		Deviance explained (%)	
		GLM[a]	GLM[b]	GLM[a]	GLM[b]
Ceratitis spp.	A	1598.000	1578.200	0.814	0.838
	B	1681.400	1673.000	0.761	0.776
B. dorsalis	A	1120.700	1032.100	0.489	0.712
	B	956.620	956.350	0.620	0.633

[a]Model without covariates decomposition. [b]Model with covariates decomposition.

3.3. Model Evaluation

To compare the fitting of the GLMs with and without timescale decomposition of covariates to the data, we used AIC and D^2. The GLMs with decomposed covariates had a better fit to the data than GLMs without decomposed covariables as indicated by lower AICs and higher D^2 (Table 7).

4. Discussion

GLMs with sinusoidal component, and without and with decomposition of covariates perform differently when fitted to the overdispersed insect count data. The GLM with decomposition of covariates fit the data better, attaining lower AIC and higher D^2 compared to GLMs with sinusoidal component, and without decomposition of covariates. This is because decomposition of the covariates allows accounting for the seasonality and long-term trend effects of the covariates in the resulting GLM. Presence of seasonality and trend in a time series influences the model results [32]. Without decomposition of covariates, the seasonality and long-term trend effects of the covariates are not accounted for in the model.

Inclusion of sinusoidal component into the GLM improves the model fitting, by allowing for the accounting of the time varying effects. Harmonic regression improves goodness-of-fit when modeling time series data exhibiting seasonal patterns [33]. The significance of sinusoidal component in the model could be because insect pest population varies with time, implying that seasonal force plays a significant role in influencing the observed insect pest population dynamics. Sinusoidal component therefore allows for accounting for the seasonal variations of abiotic factors within a model [34, 35]. In addition, presence of a sinusoidal component when modeling count time series using GLM allowed for avoidance of confounding by season and long-term trend [36].

The negative binomial models estimated the parameter of dispersion (theta) for the data within the region theta ≈ 2. Small values of theta (theta < 2) indicate aggregation, but high values (theta > 10) implies randomness, and thereby resulting in the distribution being indistinguishable from Poisson [37]. The estimated theta by the negative binomial model confirms that the insect count data herein are overdispersed and hence the negative binomial distribution is appropriate. When using time series regression for counts, accounting for both autocorrelation and overdispersion contribute to a better model fit, adding that GLM using Poisson as the distribution of counts often fail to adequately control for the overdispersion [38].

The weekly insect counts are significantly influenced by the counts recorded in the preceding week. In time series data, counts are usually correlated, contrary to the regression assumption that observations must be independent and identically distributed [38]. In addition, the insect counts increase during the fruiting stages of avocado plants. The populations of fruit fly tend to increase during rainy seasons and fruiting stage of hosts [39]. This is not the case with our data because insect population could not build up due to control measures that were undertaken when counts exceeded a set threshold. The contribution of temperature in the model imply that temperature negatively influences the expected counts of *B. dorsalis*. High temperature has a negative effect on the developmental stages of *B. dorsalis* [22].

Despite the performance of the GLMs, a foreseeable limitation is that sinusoidal components and decomposition may not be included into the GLMs if the data collected over a successive period of time does not exhibit seasonal variations and trend. In addition, the smoothing function may not perfectly fit such data especially when the amplitudes and frequencies change unsystematically. Further, abrupt changes in insect infestation over time because of control measures implemented in the orchards may have interfered with the seasonal pattern of the data.

5. Conclusion

A negative binomial model with sinusoidal components and decomposed covariates performed better in analyzing overdispersed insect population dynamics data compared to negative binomial model with sinusoidal components and undecomposed covariates. Thus, negative binomial model with sinusoidal components and decomposed covariates is recommended for modeling of population dynamics of *B. dorsalis* and *Ceratitis* spp or any other similar insect count data.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

This work received financial support from the German Federal Ministry for Economic Cooperation and Development (BMZ) commissioned and administered through the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) Fund for International Agricultural Research (FIA), grant number 17.7860.4–001; the Norwegian Agency for Development Cooperation, the Section for Research, Innovation, and Higher Education, grant number RAF-3058 KEN-18/0005; 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. The views expressed herein do not necessarily reflect the official opinion of the donors.

References

[1] T. Liboschik, "Modeling Count Time Series following Generalized Linear Models," 2016, [Online]. Available: https://eldorado.tu-dortmund.de/bitstream/2003/35144/1/Dissertation.pdf.

[2] K. Fokianos, "Some recent progress in count time series, statistics," *A J. Theor. Appl. Stat.*, vol. 45, no. 1, pp. 49–58, 2011.

[3] Y. Shapovalova, N. Baştürk, and M. Eichler, "Multivariate count data models for time series forecasting," *Entropy*, vol. 23, no. 6, pp. 1–23, 2021.

[4] R. C. Jung, M. Kukuk, and R. Liesenfeld, "Time series of count data: modeling, estimation and diagnostics," *Comput. Stat. Data Anal.*, vol. 51, no. 4, pp. 2350–2364, 2006, doi: 10.1016/j.csda.2006.08.001.

[5] M. A. Quddus, "Time series count data models: An empirical application to traffic accidents," *Accid. Anal. Prev.*, vol. 40, no. 5, pp. 1732–1741, 2008.

[6] N. Bosowski, V. Ingle, and D. Manolakis, "Generalized Linear Models for count time series," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing-Proceedings*, 2017, pp. 4272–4276.

[7] V. Serhiyenko, "Dynamic modeling of multivariate counts-Fitting, diagnostics, and applications," *Dr. Diss.*, vol. 858, 2015.

[8] C. W. S. Chen and S. Lee, "Generalized Poisson autoregressive models for time series of counts," *Comput. Stat. Data Anal.*, vol. 99, no. xxxx, pp. 51–67, 2016, doi: 10.1016/j.csda.2016.01.009.

[9] N. Alzahrani, P. Neal, S. E. F. Spencer, T. J. McKinley, and P. Touloupou, "Model selection for time series of count data," *Comput. Stat. Data Anal.*, vol. 122, pp. 33–44, 2018, doi: 10.1016/j.csda.2018.01.002.

[10] F. Lu and D. Wang, "A new estimation for INAR (1) process with Poisson distribution," *Comput. Stat.*, vol. 2021, 2021, doi: 10.1007/s00180-021-01157-5.

[11] L. Y. Chiu, D. J. Arcega Rustia, C. Y. Lu, and T. Te Lin, "Modelling and forecasting of greenhouse whitefly incidence ssing time-series and ARIMAX analysis," *IFAC-PapersOnLine*, vol. 52, no. 30, pp. 196–201, 2019.

[12] J. Hinde and C. G. B. Demétrio, "Overdispersion: Models and estimation," *Comput. Stat. Data Anal.*, vol. 27, no. 2, pp. 151–170, 1998, doi: 10.1016/S0167-9473(98)00007-3.

[13] P. Arya, R. K. Paul, A. Kumar, K. N. Singh, N. Sivaramne, and P. Chaudhary, "Predicting pest population using weather variables : An arimax time series framework," *Int. J. Agric. Stat. Sci.*, vol. 11, no. 2, pp. 381–386, 2015.

[14] R. Nyoka, J. Omony, S. M. Mwalili, T. N. O. Achia, A. Gichangi, and H. Mwambi, "Effect of climate on incidence of respiratory syncytial virus infections in a refugee camp in Kenya: A non-Gaussian time-series analysis," *PLoS One*, vol. 12, no. 6, pp. 1–14, 2017.

[15] H. E. Z. Tonnang, L. V. Nedorezov, J. O. Owino, H. Ochanda, and B. Löhr, "Host-parasitoid population density prediction using artificial neural networks: Diamondback moth and its natural enemies," *Agric. For. Entomol.*, vol. 12, no. 3, pp. 233–242, 2010, doi: https://doi.org/10.1111/j.1461-9563.2009.00466.x.

[16] H. et al. Vennila, S; Singh, G; Jha, G K; Rao, M S; Panwar, "Artificial neural network techniques for predicting severity of Spodoptera litura (Fabricius) on groundnut," *J. Environ. Biol.*, vol. 38, pp. 1–6, 2017, doi: 10.22438/jeb/38/3/MS-163.

[17] T. Wahyono, Y. Heryadi, H. Soeparno, and B. S. Abbas, "Enhanced lstm multivariate time series forecasting for crop pest attack prediction," *ICIC Express Lett.*, vol. 14, no. 10, pp. 943–949, 2020.

[18] J. J. Odanga *et al.*, "Spatial distribution of bactrocera dorsalis and thaumatotibia leucotreta in smallholder avocado orchards along altitudinal gradient of taita hills and mount kilimanjaro," *Insects*, vol. 9, no. 2, pp. 1–11, 2018, doi: 10.3390/insects9020071.

[19] N. K. Toukem, A. A. Yusuf, T. Dubois, E. M. Abdel-Rahman, M. S. Adan, and S. A. Mohamed, "Landscape vegetation productivity influences population dynamics of key pests in small avocado farms in Kenya," *Insects*, vol. 11, no. 7, pp. 1–14, 2020, doi: 10.3390/insects11070424.

[20] J. J. Odanga, S. Mohamed, R. Nyankanga, F. Olubayo, T. Johansson, and S. Ekesi, "Temporal population patterns of oriental fruit flies and false codling moths within small-holder avocado orchards in Southeastern Kenya and Northeastern Tanzania," *Int. J. Fruit Sci.*, vol. 20, no. 2, pp. 542–556, 2020, doi: 10.1080/15538362.2020.1746728.

[21] K. S. Choi, A. C. Samayoa, S. Y. Hwang, Y. B. Huang, and J. J. Ahn, "Thermal effect on the fecundity and longevity of Bactrocera dorsalis adults and their improved oviposition model," *PLoS One*, vol. 15, no. 7, pp. 3–6, 2020.

[22] R. Ma, A. Verghese, R. R. Pv, and S. Kandakoor, "Effect of climate change on biology of oriental fruit fly, Bactrocera dorsalis hendel (Diptera: Tephritidae)," vol. 8, no. May, pp. 935–940, 2020.

[23] P. Montoya, S. Flores, and J. Toledo, "Effect of rainfall and soil moisture on survival of adults and immature stages of Anastrepha ludens and A. obliqua (Diptera: Tephritidae) under semi-field conditions," *Florida Entomol.*, vol. 91, no. 4, pp. 643–650, 2008.

[24] J. A. Nelder and R. W. M. Wedderburn, "Generalized Linear Models," *J. R. Stat. Soc. Ser. A*, vol. 135, no. 3, pp. 370–384, 1972, doi: 10.2307/2344614.

[25] F. J. Anscombe, "The statistical analysis of insect counts based on the Negative Binomial distribution," *Int. Biometric Soc.*, vol. 5, no. 2, pp. 165–173, 1949, doi: 10.2307/3001918.

[26] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proceedings of the 2nd International Symposium on Information Theory*, 1973, pp. 267–281.

[27] A. Guisan and N. E. Zimmermann, "Predictive habitat distribution models in ecology," *Ecol. Modell.*, vol. 135, no. 2–3, pp. 147–186, 2000, doi: 10.1016/S0304-3800(00)00354-9.

[28] R. CoreTeam, "R: A Language and Environment for Statistical Computing," 2020. https://www.r-project.org/ (accessed Aug. 16, 2021).

[29] B. D. Venables, W. N. & Ripley, "Modern applied Statistics with S-PLUS," *J. R. Stat. Soc. Ser. D (The Stat.*, vol. 52, no. 4, pp. 704–705, 2002.

[30] A. Gasparrini, "Distributed lag linear and non-linear models in R: The package dlnm," *J. Stat. Softw.*, vol. 43, no. 8, pp. 2–20, 2011.

[31] J. Law and D. Mitarotonda, "Package ' lubridate ' R topics documented:," 2021. https://cran.r-project.org/web/packages/lubridate/lubridate.pdf.

[32] R. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 2nd ed. Melbourne, Australia: Melbourne, Australia: OTexts, 2018.

[33] K. Ramanathan *et al.*, "Assessing seasonality variation with harmonic regression: Accommodations for sharp peaks," *Int. J. Environ. Res. Public Health*, vol. 17, no. 4, pp. 1–14, 2020.

[34] D. A. Ewing, C. A. Cobbold, B. V. Purse, M. A. Nunn, and S. M. White, "Modelling the effect of temperature on the seasonal population dynamics of temperate mosquitoes," *J. Theor. Biol.*, vol. 400, pp. 65–79, 2016.

[35] C. A. Johnson *et al.*, "Effects of temperature and resource variation on insect population dynamics: the bordered plant bug as a case study," *Funct. Ecol.*, vol. 30, no. 7, pp. 1122–1131, 2016.

[36] C. Imai, B. Armstrong, Z. Chalabi, P. Mangtani, and M. Hashizume, "Time series regression model for infectious disease and weather," *Environ. Res.*, vol. 142, pp. 319–327, 2015.

[37] M. H. Stephens, *A Primer of Ecology with R*, vol. 32, no. Book Review 3. New York, United States: Springer, 2010.

[38] A. Tobías and M. Saez, "Time-series regression models to study the short-term effects of environmental factors on health," 11, 2004.

[39] H. Mze *et al.*, "Invasion by Bactrocera dorsalis and niche partitioning among tephritid species in Comoros," *Bull. Entomol. Res.*, vol. 106, no. 6, pp. 749–758, 2016.