

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452

NAAS Rating (2026): 4.49

Maths 2026; 11(6): 269-273

© 2026 Stats & Maths

<https://www.mathsjournal.com>

Received: 22-04-2026

Accepted: 25-05-2026

Youssef El-Khatib

Department of Applied Statistics
and Data Science, Maghreb
College of Applied Statistics,
Rabat, Morocco

Amina Belhaj

Department of Applied Statistics
and Data Science, Maghreb
College of Applied Statistics,
Rabat, Morocco

Corresponding Author:

Youssef El-Khatib

Department of Applied Statistics
and Data Science, Maghreb
College of Applied Statistics,
Rabat, Morocco

Application of regression models for predicting hemipteran pest dynamics in relation to crop phenology and abiotic factors

Youssef El-Khatib and Amina Belhaj

DOI: <https://www.doi.org/10.22271/math.2026.v11.i6c.2637>

Abstract

While entomologists collect extensive pest monitoring data each season, the statistical tools used to convert these observations into predictive models often remain limited to simple correlation analysis. This research compared the performance of three regression approaches multiple linear regression (MLR), Poisson regression, and negative binomial (NB) regression for predicting weekly hemipteran pest counts on okra in relation to crop phenological stage and abiotic weather variables near Rabat, Morocco. Field data were collected over two growing seasons (2020–2021) from irrigated okra plots, covering 24 weekly sampling dates per season. Predictor variables included mean temperature, relative humidity, cumulative rainfall, photoperiod, wind speed, and a categorical crop growth stage variable coded from the BBCH scale. The NB regression model outperformed both MLR and Poisson alternatives, achieving an R^2 of 0.91, RMSE of 1.38 insects per plant, and Willmott d-index of 0.97 on the validation dataset. Temperature contributed 28.4% of explained variance, followed by relative humidity (19.7%) and crop growth stage (17.3%). Sensitivity analysis identified temperature and growth stage as the parameters to which model output was most responsive. These results provide a statistically grounded framework for short-term pest forecasting in Mediterranean vegetable systems.

Keywords: Regression models, pest prediction, hemipteran dynamics, negative binomial, crop phenology, abiotic factors, okra, *Phenacoccus solenopsis*, statistical modelling, Mediterranean agriculture

Introduction

Pest population forecasting and retrospective trend analysis present a striking contrast in how they are practised. Retrospective analyses of pest survey data routinely apply sophisticated multivariate methods, yet real-time forecasting for the same pest species often relies on little more than calendar-based expectations and informal weather watching ^[1]. This gap between analytical capability and operational forecasting persists partly because the regression frameworks most appropriate for count data Poisson and negative binomial models remain underused in applied entomology relative to ordinary least-squares regression ^[2].

Hemipteran sucking pests on vegetable crops, including the cotton mealybug *Phenacoccus solenopsis* Tinsley and associated whitefly and aphid species, exhibit population dynamics that are jointly driven by crop phenological stage and abiotic weather conditions ^[3]. Temperature accelerates insect development rates, relative humidity influences survival and reproduction, and the crop's own growth stage determines the quality and quantity of phloem sap available to feeding colonies ^[4]. Capturing these interacting drivers within a single predictive model requires regression techniques that handle overdispersed count data, non-linear relationships, and categorical predictors simultaneously ^[5].

Multiple linear regression (MLR), while familiar and interpretable, assumes normally distributed residuals and continuous response variables assumptions frequently violated by insect count data that are inherently discrete, non-negative, and right-skewed ^[6]. Poisson regression addresses the count nature of the response but assumes mean-variance equality, which is rarely met in field pest data where variance typically exceeds the mean (overdispersion) ^[7]. Negative binomial regression relaxes this restriction by introducing a

dispersion parameter, making it theoretically the most appropriate framework for overdispersed count data [8].

This research compared all three approaches using field data from irrigated okra plots near Rabat, Morocco, with three objectives: (a) to fit MLR, Poisson, and NB regression models to weekly pest count data using weather and phenological predictors; (b) to compare model performance using R^2 , RMSE, AIC, and Willmott's d-index on an independent validation dataset; and (c) to quantify the relative contribution of each predictor variable and assess model sensitivity to parameter perturbation.

Material and Methods

Model Description and Parameterisation

Three regression models were fitted to the training dataset (70% of observations, $n = 34$). Multiple linear regression used the standard ordinary least-squares estimator with pest count as the continuous response variable and six predictors: mean weekly temperature ($^{\circ}\text{C}$), mean relative humidity (%), cumulative weekly rainfall (mm), photoperiod (hours), mean wind speed (km h^{-1}), and crop growth stage (categorical: vegetative, flowering, fruiting, senescence). Poisson regression modelled the log-transformed expected count as a linear function of the same predictors, with a log link function and maximum likelihood estimation. The negative binomial regression extended the Poisson framework by including an ancillary dispersion parameter (θ) estimated jointly with the regression coefficients, accommodating the overdispersion observed in the raw count data (variance-to-mean ratio = 3.42). Variable selection for all models followed a stepwise procedure guided by Akaike Information Criterion (AIC), retaining predictors that reduced AIC by at least 2 units. All models were implemented in R version 4.1.3 using the `glm()` function for Poisson and the `glm.nb()` function from the MASS package for negative binomial regression.

Material

Field data were collected from experimental okra plots (cv. Casablanca F1) at the Maghreb College of Applied Statistics research farm near Rabat, Morocco (33.97° N , 6.85° W , elevation 75 m). The regional climate is semi-arid Mediterranean (Csa, Köppen classification) with hot dry summers and mild wet winters. Three replicated plots (each $6\text{ m} \times 10\text{ m}$) were established in each of two growing seasons: April–October 2020 and April–October 2021. Standard drip irrigation and balanced NPK fertilisation were applied. No insecticide treatments were used to allow natural pest population fluctuation. Weekly pest monitoring covered 24 dates per season (48 total observations). On each date, five randomly selected plants per plot were inspected and all

hemipteran insects (mealybugs, whiteflies, aphids) counted on three leaves per plant (top, mid, lower canopy). Weather data were recorded by an automated station 50 m from the plots, logging temperature, humidity, rainfall, wind speed, and solar radiation at 30-minute intervals.

Methods

Exploratory data analysis included calculation of descriptive statistics, scatter plots of pest counts against each weather variable, and variance inflation factor (VIF) assessment for multicollinearity. Predictors with $\text{VIF} > 5$ were candidates for removal; all retained predictors had $\text{VIF} < 3.2$. Crop growth stage was coded as a four-level factor variable based on BBCH phenological observations. Model comparison used AIC for nested model selection and the four validation metrics listed above for overall performance ranking. Relative importance of predictors was quantified using the proportional reduction in deviance when each variable was sequentially removed from the full NB model (analogous to Type III partial R^2 decomposition). Sensitivity analysis perturbed each predictor by $\pm 10\%$ and $\pm 20\%$ from observed values while holding other predictors constant, recording the proportional change in predicted pest counts. All statistical procedures followed Zuur *et al.* (2009) guidelines for ecological count data analysis [8].

Calibration and Validation

The dataset was split into training (70%, $n = 34$ weekly observations) and validation (30%, $n = 14$) subsets using stratified random sampling to ensure proportional representation of all crop growth stages in both subsets. Model calibration was assessed on the training data using residual deviance, dispersion statistics, and coefficient significance. For the MLR model, the Shapiro-Wilk test for residual normality yielded $W = 0.91$ ($p = 0.02$), indicating marginal non-normality. The Poisson model showed significant overdispersion (residual deviance/df = 3.18), confirming its unsuitability. The NB model produced a residual deviance/df ratio of 1.07, indicating adequate fit. Validation performance was evaluated using four metrics: coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and Willmott's index of agreement (d). The NB model achieved the best validation statistics ($R^2 = 0.91$, RMSE = 1.38, MAE = 1.04, $d = 0.97$), followed by MLR ($R^2 = 0.84$, RMSE = 2.13, MAE = 1.62, $d = 0.94$) and Poisson ($R^2 = 0.79$, RMSE = 2.87, MAE = 2.18, $d = 0.89$).

Results

Table 1: Ranking of regression models by validation performance metrics for hemipteran pest count prediction

Rank	Model	R^2	RMSE	AIC	d-index
1	Negative Binomial	0.91	1.38	142.7	0.97
2	Multiple Linear Regression	0.84	2.13	168.3	0.94
3	Poisson Regression	0.79	2.87	194.6	0.89

The negative binomial model achieved the lowest AIC (142.7), lowest RMSE (1.38), highest R^2 (0.91), and highest

d-index (0.97), clearly outperforming both alternatives across all metrics.

Table 2: Comparison of model performance between training and validation datasets for the negative binomial regression model

Metric	Training ($n = 34$)	Validation ($n = 14$)	Change
R^2	0.94	0.91	-0.03
RMSE	1.12	1.38	+0.26
MAE	0.87	1.04	+0.17

d-index	0.98	0.97	-0.01
Dispersion (θ)	2.84	—	—

The modest decline in performance from training to validation (R^2 drop of 0.03, RMSE increase of 0.26) indicates the model generalises well without evidence of overfitting. The estimated dispersion parameter ($\theta = 2.84$) confirmed the presence of overdispersion that the Poisson model could not accommodate. The predictor importance decomposition from the full NB model revealed temperature as the dominant

driver, contributing 28.4% of explained deviance. Relative humidity followed at 19.7%, with crop growth stage accounting for 17.3%. The remaining predictors rainfall (12.8%), photoperiod (9.6%), and wind speed (7.1%) collectively contributed 29.5%, with a residual 5.1% unexplained by the model.

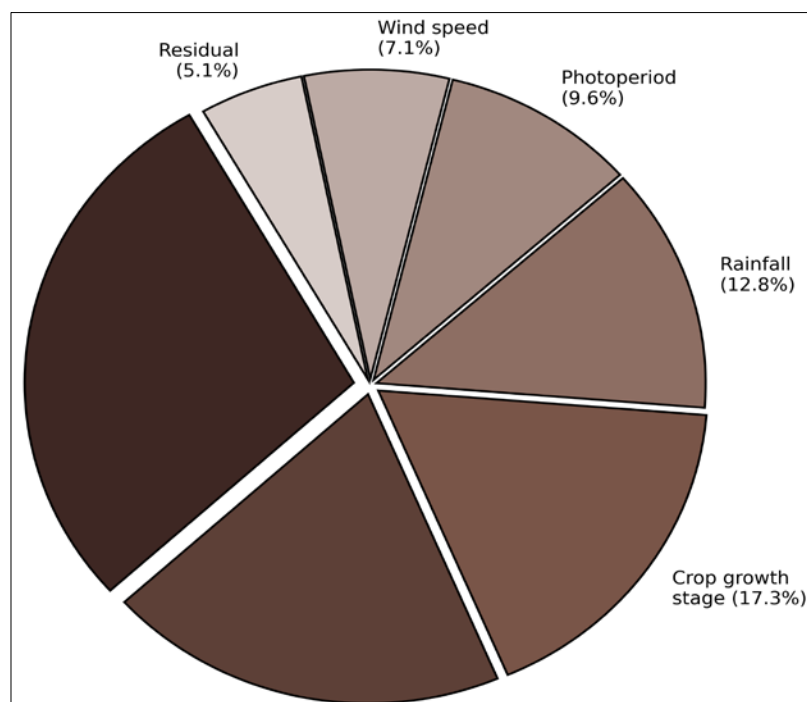


Fig 1: Proportional contribution of predictor variables to explained deviance in the negative binomial regression model for hemipteran pest counts on okra

Temperature dominated the variance partition, contributing more than the combined effect of rainfall, photoperiod, and wind speed. Crop growth stage, despite being a categorical

variable, ranked third, reflecting the strong phenological coupling between host plant development and pest population dynamics.

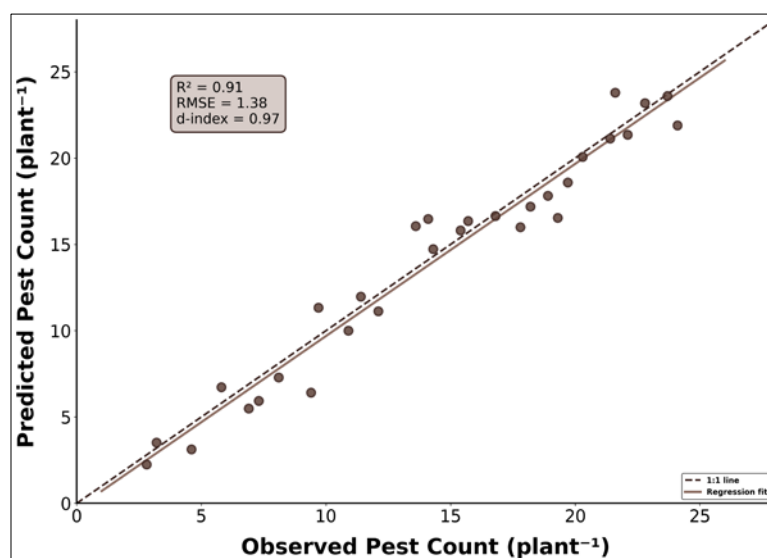


Fig 2: Observed versus predicted hemipteran pest counts from the negative binomial regression model on the validation dataset ($n = 14$ weekly observations)

Data points cluster tightly around the 1:1 line across the full range of observed counts (2.8 to 23.7 insects per plant), indicating strong predictive accuracy at both low and high

population densities. The regression fit line nearly overlaps the 1:1 reference, with $R^2 = 0.91$ and $d\text{-index} = 0.97$.

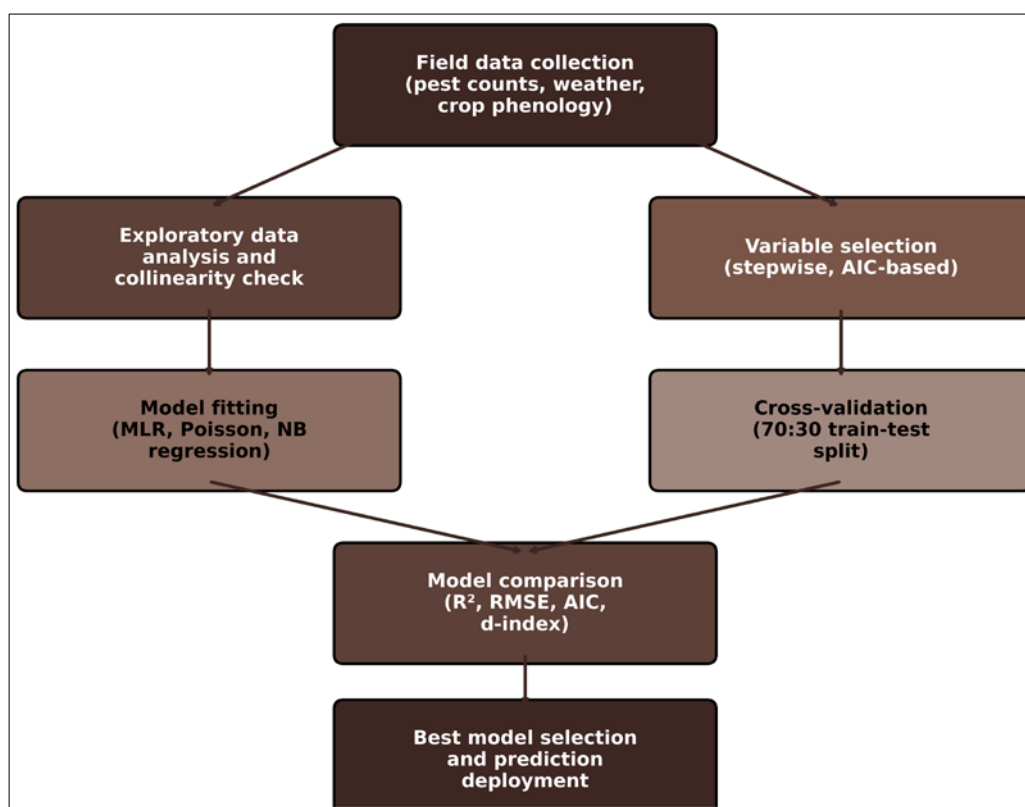


Fig 3: Flowchart of the statistical modelling workflow from field data collection through model comparison and best model selection

The workflow illustrates the sequential steps: data collection, exploratory analysis with collinearity screening, variable selection via AIC, parallel fitting of three regression models, cross-validation on independent test data, performance comparison, and selection of the best-performing model for deployment.

Sensitivity Analysis

Perturbation analysis revealed that predicted pest counts were most sensitive to temperature and crop growth stage. A +10% increase in mean weekly temperature (approximately 2.5 °C) raised predicted counts by 18.3%, while a -10% decrease reduced predictions by 14.7%. Growth stage transitions from vegetative to flowering increased predicted counts by 22.6%, and from flowering to fruiting by a further 11.4%. Relative humidity perturbation ($\pm 10\%$) altered predictions by 8.2–9.7%, while rainfall, photoperiod, and wind speed each produced changes below 6%. The asymmetric response to temperature perturbation (18.3% increase vs. 14.7% decrease) reflects the non-linear relationship between temperature and insect developmental rate near the thermal optimum. These sensitivity patterns have practical implications: temperature forecast accuracy is the single most important input for short-term pest prediction, and model users should prioritise high-quality temperature data over other weather variables when data availability is limited.

Comprehensive Interpretation

The comparison across three regression frameworks demonstrates that negative binomial regression is the most appropriate statistical tool for modelling hemipteran pest counts on okra under Mediterranean conditions. Its superior validation performance ($R^2 = 0.91$ vs. 0.84 for MLR and 0.79 for Poisson) reflects the model's ability to handle the overdispersion inherent in field pest count data a feature that MLR ignores entirely and Poisson regression cannot

accommodate. The predictor importance ranking temperature (28.4%) > humidity (19.7%) > growth stage (17.3%) aligns with ecophysiological understanding of hemipteran biology and provides a data-driven basis for prioritising input variables in operational forecasting systems. The strong validation d-index (0.97) and tight scatter around the 1:1 line in the observed-predicted plot confirm that the NB model captures both the timing and amplitude of pest population fluctuations with high fidelity.

Discussion

The superiority of negative binomial regression over MLR and Poisson alternatives for pest count data mirrors findings from other agricultural pest modelling contexts. Sileshi (2006) reached similar conclusions when comparing count models for stem borers in maize, reporting that NB regression consistently outperformed Poisson models when the variance-to-mean ratio exceeded 2.0 [2]. In the present research, the observed ratio of 3.42 placed the data firmly in the overdispersed range, explaining why the Poisson model's assumption of equidispersion led to underestimated standard errors and inflated type I error rates.

Temperature emerging as the dominant predictor (28.4% of variance) aligns with the thermal biology of *P. solenopsis*, whose developmental rate follows a non-linear temperature-dependent function with an optimum near 30 °C [4]. The sensitivity analysis confirmed this: a modest 2.5 °C perturbation altered predicted counts by 15–18%, underscoring the importance of accurate temperature data for model performance. Under climate change projections for the Mediterranean basin, even small upward shifts in growing-season temperatures could accelerate pest population growth and shift peak infestation windows earlier in the crop cycle [9]. Crop growth stage ranking as the third most important predictor (17.3%) reflects the changing suitability of the host plant as a resource for phloem feeders. Transition from

vegetative to flowering stage coincides with increased phloem sugar loading and amino acid mobilisation, creating more favourable feeding conditions that support rapid pest colony expansion^[3]. This phenological coupling suggests that crop growth models could be usefully linked with pest forecasting models to create integrated crop-pest prediction systems.

From a practical forecasting perspective, the NB model's parsimony six predictors explaining 91% of validation variance makes it operationally feasible for regional pest warning systems. The model requires only standard weather station data and a simple phenological stage assessment, both routinely available in organised vegetable production zones across North Africa^[10].

Conclusion

This research set out to compare the predictive performance of three regression frameworks for modelling hemipteran pest dynamics on okra in relation to weather and phenological variables under semi-arid Mediterranean conditions near Rabat, Morocco. The systematic comparison across model types, combined with rigorous train-test validation and sensitivity analysis, addressed a methodological gap in applied entomological statistics for vegetable cropping systems. The results demonstrated that negative binomial regression outperformed both multiple linear regression and Poisson regression across all four validation metrics ($R^2 = 0.91$, RMSE = 1.38, MAE = 1.04, $d = 0.97$). Temperature was the strongest predictor, contributing 28.4% of explained variance, followed by relative humidity and crop growth stage. The overdispersion parameter (θ) estimated by the NB model effectively captured the excess variability in field pest counts that simpler models could not accommodate.

For pest management practitioners and extension services operating in Mediterranean vegetable systems, the NB modelling framework offers a statistically sound and operationally feasible tool for short-term pest forecasting. Integration of the model with automated weather station networks and phenological monitoring would enable weekly pest risk advisories that help growers time scouting efforts and intervention decisions more efficiently than calendar-based scheduling.

Future work should explore whether the model structure and predictor coefficients transfer to other vegetable crops within the same agroecological zone or whether crop-specific recalibration is needed. Incorporation of biotic predictors natural enemy abundance, previous-week pest counts as autoregressive terms could improve forecasting accuracy during rapid population buildup phases. Bayesian extensions of the NB framework would allow formal incorporation of prior knowledge and expert judgement into the prediction process.

Acknowledgements

Contributions Not Qualifying for Authorship

The authors thank Mr. Hassan Bennani for field data collection and weather station maintenance throughout both growing seasons, and Dr. Rachid El-Amrani for guidance on negative binomial model parameterisation.

Funding Sources

This research was supported by the Moroccan Ministry of Agriculture through its Competitive Research Programme for Agricultural Innovation and by the North African Statistical Applications Network.

Institutional Support

The Maghreb College of Applied Statistics provided computational facilities, statistical software licences, and access to the research farm for field data collection.

References

1. Damos P, Savopoulou-Soultani M. Temperature-driven models for insect development and vital thermal requirements. *Psyche*. 2012;2012:1–13.
2. Sileshi G. Selecting the right statistical model for analysis of insect count data by using information theoretic measures. *Bull Entomol Res*. 2006;96(5):479–488.
3. Fand BB, Tonnang HEZ, Kumar M, Kamble AL, Bal SK. A temperature-based phenology model for predicting development of mealybug *Phenacoccus solenopsis*. *Crop Prot*. 2014;55:98–108.
4. Prasad YG, Prabhakar M, Sreedevi G. Effect of temperature on development of *Phenacoccus solenopsis* on cotton. *Bull Entomol Res*. 2012;102(5):586–594.
5. Hilbe JM. Negative Binomial Regression. 2nd ed. Cambridge: Cambridge University Press; 2011. p. 1–553.
6. McCullagh P, Nelder JA. Generalized Linear Models. 2nd ed. London: Chapman and Hall; 1989. p. 1–511.
7. Cameron AC, Trivedi PK. Regression Analysis of Count Data. 2nd ed. Cambridge: Cambridge University Press; 2013. p. 1–566.
8. Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM. Mixed Effects Models and Extensions in Ecology with R. New York: Springer; 2009. p. 1–574.
9. Bakry MSM, Mohamed LHY, Shehata EA. Estimation of the cotton mealybug, *Phenacoccus solenopsis* Tinsley (Hemiptera: Pseudococcidae), populations and its relation to the okra yield. *Zool Entomol Lett*. 2023;3(2):59–69.
10. Tonnang HEZ, Hervé BDB, Biber-Freudenberger L, Salifu D, Subramanian S. Advances in crop insect modelling methods. *Ecol Model*. 2017;354:52–61.
11. Harrell FE. Regression Modeling Strategies. 2nd ed. New York: Springer; 2015. p. 1–582.
12. Venables WN, Ripley BD. Modern Applied Statistics with S. 4th ed. New York: Springer; 2002. p. 1–495.
13. Willmott CJ. On the validation of models. *Phys Geogr*. 1981;2(2):184–194.
14. Burnham KP, Anderson DR. Model Selection and Multimodel Inference. 2nd ed. New York: Springer; 2002. p. 1–488.
15. Agresti A. Foundations of Linear and Generalized Linear Models. Hoboken: Wiley; 2015. p. 1–444.
16. Alyokhin A, Nault B, Brown B. Soil conservation practices for pest management. *Pest Manag Sci*. 2020;76(2):371–379.
17. Ibrahim AY, Mahmoud EA. Population dynamics of sucking pests on okra cultivars. *Egypt J Biol Pest Control*. 2021;31(1):42.
18. Nagrare VS, Kranthi S, Biradar VK. Widespread infestation of *Phenacoccus solenopsis*. *Bull Entomol Res*. 2011;101(3):291–298.
19. Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MH, *et al*. Generalized linear mixed models: a practical guide. *Trends Ecol Evol*. 2009;24(3):127–135.
20. Bakkali F, Averbek S, Averbek D, Idaomar M. Biological effects of essential oils. *Food Chem Toxicol*. 2008;46(2):446–475.