Logistic Regression is a better Method of Analysis Than Linear Regression of Arcsine Square Root Transformed Proportional Diapause Data of Pieris melete (Lepidoptera: Pieridae)

Authors: Shi, P. J., Sand Hu, H. S., and Xiao, H. J.

Source: Florida Entomologist, 96(3) : 1183-1185

Published By: Florida Entomological Society

URL: https://doi.org/10.1653/024.096.0361

LOGISTIC REGRESSION IS A BETTER METHOD OF ANALYSIS THAN LINEAR REGRESSION OF ARCSINE SQUARE ROOT TRANSFORMED PROPORTIONAL DIAPAUSE DATA OF *PIERIS MELETE* (LEPIDOPTERA: PIERIDAE)

P. J. SHI[1], H. S. SANDHU[2] AND H. J. XIAO[3,*]

[1]Key Laboratory of Sustainable Development of Marine Fisheries, Ministry of Agriculture, Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Qingdao 266071, China

[2]University of Florida, Institute of Food and Agricultural Sciences, Everglades Research and Education Center, Belle Glade, FL, USA

[3]Institute of Entomology, Jiangxi Agricultural University, Nanchang 330045, China

*Corresponding author; E-mail: haijunxiao@hotmail.com

Recently, Xiao et al. (2012) published a study on the effects of daily average temperature and natural day-length on the incidence of summer and winter diapause in the Cabbage Butterfly, *Pieris melete* Ménétriés (Lepidoptera: Pieridae). Under field conditions, the cabbage butterfly, *Pieris melete*, displays a pupal summer diapause in response to relatively low daily temperatures and gradually increasing day-length during spring, and a pupal winter diapause in response to the progressively shorter day-length. To determine whether photoperiod has a stronger role than temperature in the determination of the summer and winter diapause, or vice versa, the effects of the naturally changing day-length and temperature on the initiation of summer and winter diapause were systematically investigated under field conditions for 5 successive years. Field results showed that the incidence of summer diapause significantly declined with the naturally increasing temperature in spring and summer generations. Path coefficient analysis showed that the effect of temperature was much greater than that of photoperiod in the determination of summer diapause. In autumn, the incidence of diapause was extremely low when larvae developed under gradually shortening day-length and high temperatures. However the incidence of winter diapause increased to 60-90% or higher with gradually shortening day-length combined with lower temperatures, i.e., between 20.0 °C and 22.0 °C. Decreasing day-length played a more important role in the determination of winter diapause induction than temperature.

Thus Xiao et al. (2012) provided 5-year proportional data of cabbage butterfly diapause and corresponding temperature and day-length data (Tables 1 and 2 in Xiao et al. 2012). However, only linear regression was used to describe the effects of these 2 predictors on diapause (Table 3 in Xiao et al., 2012). Linear regression simply hypothesizes that the effects of these predictors on the response variable are linear. However, nonlinear effects are ubiquitous in nature. Therefore, linear regression actually neglected the nonlinear effects, which led to low goodness-of-fit to the objective data. Additionally, the proportional data were arcsine square root transformed before performing the linear regression. Although this transformation has long been standard procedure in analyzing proportional data in ecology, logistic regression has greater interpretability and higher power than transformation in data containing binomial and non-binomial response variables (Warton & Hui 2011). The detailed problems with arcsine transformation can be found in Wilson et al. (online), and logistic regression was strongly recommended as an alternative to the arcsine transformation in biological analysis. If the distribution of proportional data is not normal, then the use of arcsine transformation is problematic. In fact, the distribution of the transformed diapause data of Xiao et al. (2012) was still not normalized ($W = 0.8491$, P-value < 0.05) as revealed by the Shapiro-Wilk normality test (Faraway 2005). Thus, we suggest using logistic regression to fit the proportional data of diapause. The following analysis can be considered as an alternative to the analysis performed by Xiao et al. (2012), and also for similar data in future studies.

Diapause and non-diapause response to a combination of temperature and day-length can be exactly described either by the generalized linear model or the generalized additive model (Hastie & Tibshirani 1990). The latter is more flexible in fitting the data. We used the following generalized additive model to describe the effects of temperature and day-length on diapause:

$$\text{logit}(\text{Diapause}) = \alpha + f_1(\text{Temperature}, + f_2(\text{Day-length}), \quad (1)$$

where f_i ($i = 1, 2$) are smooth functions.

We pooled and fitted the data of summer and winter diapause, and found that temperature and day-length both could significantly affect

diapause (P < 2e-16) (Fig. 1). Fig. 2 exhibits the fitted surface of diapause. We found that the temperature > 23 °C led to a very low diapause. Relative to day-length, temperature appeared to be more important in determining diapause of the cabbage butterfly. The goodness-of-fit obtained by using the generalized additive model is satisfactory with r^2 = 0.95. It is higher than the r^2 value calculated in Xiao et al. (2012) by using 2 predictors. Considering the important effect of temperature on diapause, we also explored the effect of a single predictor, i.e., temperature on diapause. The prediction by using temperature only is also satisfactory with r^2 = 0.90, which is greater than the r^2 calculated in Xiao et al. (2012) by using temperature only.

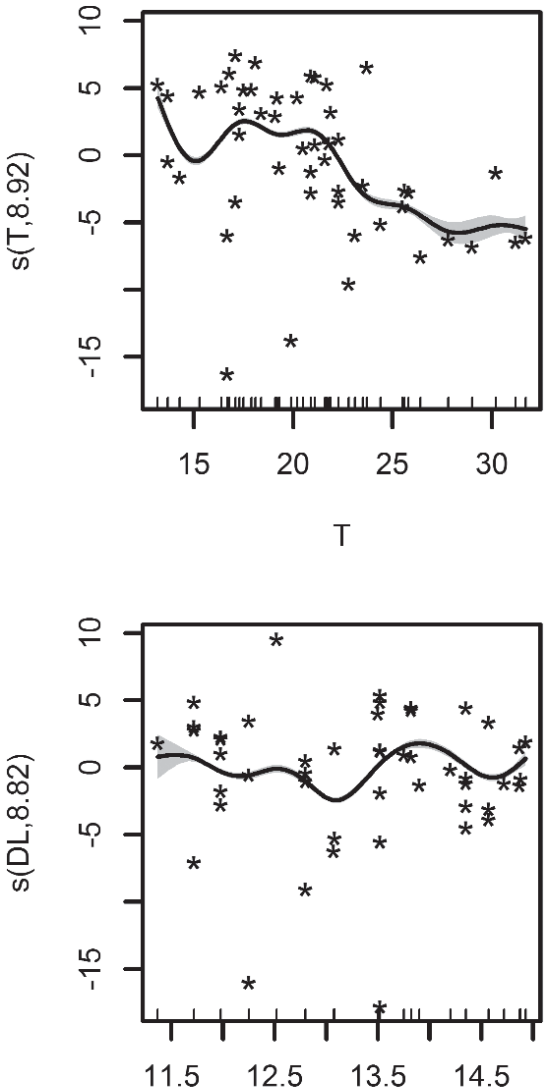

Fig. 1. The solid curves represent the generalized additive model fit to *Pieris melete* diapause data using 2 predictors: temperature and day-length. The asterisks in the figures represent the partial residuals. The gray bands represent 95% confidence intervals. T denotes temperature, and DL denotes day-length.

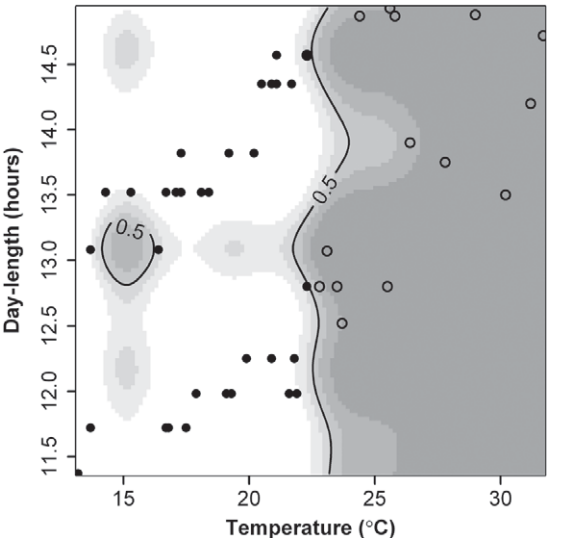

Fig. 2. Fitted surface of *Pieris melete* diapause data by using the generalized additive model. The curves marked with 0.5 represent the combinations of generalized temperature and day-length that can result in 50% diapause. Diapause ≧ 50% are in the white area and diapause < 50% are in the gray area. Points represent the observed data of diapause ≧ 50%, and open circles represent the observed data of diapause < 50%.

Summary

Temperature and day-length are considered to be the 2 important factors that can significantly affect insect diapause, which is a typical proportional dataset. In the previous studies, the method of arcsine square root transformation is widely used to analyze the effect of temperature or day-length or their joint effects on diapause in insects. However, this method has many limitations, for example, the proportional data should be normally distributed. The logistic regression in generalized additive models is a promising method for analyzing the effects of temperature and day-length on diapause. Compared to the arcsine square root transformation method, this method does not require normal distribution of proportional diapause data. The logistic regression also provides better goodness-of-fit by using the non-parametric fitting technique. In this report, we used the diapause data of *Pieris melete* (Xiao et al. 2012) to compare the fitted results of the logistic regression in generalized additive models with arcsine square root transformation. We found that the logistic regression in generalized

additive models is better than linear regression of arcsine square root transformed data in following ways: (1) reasonable predictions about diapause ranging from 0 to 1 can be made without transforming the proportional data; (2) non-linear effects of temperature and day-length on diapause can be determined; (3) the goodness-of-fit can be substantially improved.

Key Words: additive model, binomial response variables, goodness-of-fit, non-linear effects, proportional data set

Resumen

Se considera que la temperatura y la duración del día son los 2 factores importantes que puedan afectar significativamente la diapausa de los insectos, el cual es un conjunto de datos proporcionales típico. En los estudios anteriores, el método de transformación del arcoseno de la raíz cuadrada es ampliamente utilizado para analizar el efecto de la temperatura o la duración del día o su efecto conjunto sobre la diapausa de insectos. Sin embargo, este método tiene muchas limitaciones, por ejemplo, los datos proporcionales deben ser distribuidos normalmente. La regresión logística en los modelos aditivos generalizados es un método prometedor para el análisis de los efectos de la temperatura y la duración del día en diapausa. En comparación con el método de transformación del arcoseno de la raíz cuadrada, este método no requiere que la distribución de los datos de diapausa proporcional sea normal. La regresión logística también proporciona un mayor grado de ajuste mediante el uso de la técnica de ajuste no paramétrico. En este informe, usamos los datos de diapausa de *Pieris melete* (Xiao et al. 2012) para comparar los resultados ajustados de la regresión logística en modelos aditivos generalizados con transformación arcoseno de raíz cuadrada. Se encontró que la regresión logística en los modelos aditivos generalizados es mejor que la regresión lineal de los datos transformados del arcoseno de raíz cuadrada en las siguientes formas: (1) se pueden hacer predicciones razonables sobre diapausa de entre 0 a 1 sin transformar los datos proporcionales; (2) se pueden determinar los efectos no-lineales de temperatura y duración del día en diapausa y (3) puede mejorar el grado de ajuste sustancialmente.

Palabras Clave: modelo aditivo, variables de respuesta binomial, grado de ajuste, efectos no lineales, conjunto de datos proporcionales

Endnotes

We are deeply thankful to the editor, Dr. Waldemar Klassen, and anonymous reviewers for their invaluable comments to improve this manuscript. All three of the authors contributed equally to this work.

References Cited

Faraway, J. J. 2005. Linear Model. Chapman and Hall, CRC, London.

Hastie, T. J., and Tibshirani, R. J. 1990. Generalized Additive Models. Chapman and Hall, London.

Warton, D. I., and Hui, F. K. C. 2011. The arcsine is asinine: the analysis of proportions in ecology. Ecology 92: 3-10.

Wilson, E., Underwood, M., Puckrin, O., Letto, K., Doyle, R., Caravan, H., Camus, S., and Bassett, K. 2013. The arcsine transformation: has the time come for retirement? http://www.mun.ca/biology/dschneider/b7932/B7932Final10Dec2010.pdf

Xiao, H. J., Wu, S. H., He, H. M. Chen, C., and Xue, F. S. 2012. Role of natural day-length and temperature in determination of summer and winter diapause in *Pieris melete* (Lepidoptera: Pieridae). Bull. Entomol. Res. 102: 267-273.