Landscape and plant-pollinator community characteristics differentially shape bumble bee parasite prevalence

Wakeel Adekunle Kasali
wakeel.kasali@stat.ubc.ca

October 19, 2024

Abstract

The prevalence of parasites in bee populations, influenced by factors such as floral abundance, floral diversity, landscape characteristics, and the abundance of native bees and Bombus impatiens, with temporal control and variation across transects and individual bee species, can be approached using multiple open-ended statistical methods. Generalized Linear Mixed Models (GLMM) and Multilevel Structural Equation Models (SEM) are considered for application to understand the relationships within the model's pathways.

Introduction

The co-existence of bee species impacts ecosystem pollinator health. Food resource availability, driven by plant diversity and abundance, and landscape simplification are essential predictors of pollinator populations. Recent studies highlight the need to monitor non-native bee species and their impact on other species' health [4]. Landscape characteristics, floral diversity, and bumble bee community structure may directly or indirectly explain parasite prevalence within pollinator populations. The project aims to achieve three main objectives within this framework. (i) assess how bumble bee abundance and diversity are influenced by landscape composition, diversity, and floral community characteristics; (ii) determine how bumble bee and floral community traits affect the occurrence of common bumble bee parasites, with a focus on the invasive *Bombus impatiens*; (iii) evaluate how landscape characteristics impact the occurrence of bumble bee parasites.

Data description and collection

A field study surveyed *Bombus spp.* across 267 transects in six landscape types in the Lower Fraser Valley, near Vancouver, from May to August 2022. Bumble bee abundance and flowering plant diversity were recorded in 5-minute surveys along 50-meter transects. A total of 3,145 bumble bees were collected, with 749 (6 species) screened for four common parasites (*Crithidia spp., Nosema spp., Apicystis spp., Ascosphaera spp.*). Parasite prevalence was recorded as a binary variable (0 = none, 1 = detected). Landscapes were

classified into 16 categories and analyzed for Shannon diversity, edge area, and blueberry cultivation within a 500-meter buffer. There is no missing data.

Bombus Shannon and floral diversity were calculated using the Shannon Index and subsequently standardized. Native bee and Impatiens abundances were recorded as counts, while floral abundance was log-transformed and then standardized.

Statistical questions

The main statistical questions are as follows:

1. Could lower parasite prevalence with abundant *B. impatiens* be a biological effect or sampling bias? How can we distinguish and ensure model accuracy?

2. How can a non-overfitting model be achieved with Bayesian SEM?

3. What additional validation or residual diagnostics can ensure robustness after fitting models in `lme4` and `brms`, and confirm model validity and interpretability in ecological data?

Proposed statistical methods

The histograms of native bee abundance and impatiens abundance should be presented. In the abundance data, zeros can indicate important ecological patterns (e.g., areas where certain species are absent). If present, the high skewness would suggest the possible need to either log-transform or apply other appropriate transformations that will contribute to a good model structure.

EDA will reveal patterns and relationships between infection status and bee species, guiding hypothesis testing. One practical approach is visually representing the data and applying a statistical test to determine whether the lower parasite prevalence observed with abundant B. impatiens is a true biological effect or simply a result of sampling bias. Creating a bar plot with proportion infected on the y-axis and species status (*B. impatiens* and native bee) on the x-axis will give a clear view of the proportion of infected bees in each species group. A potential biological effect rather than a sampling bias can be verified with a test of the difference in proportions of parasite prevalence in native vs non-native bees. Calculating the difference in proportions and testing it with permutation-based null distribution would remove the influence of confounding factors such as sample size, sampling bias, or variability in how bees were collected.

If in each visit i, I screen N_i bees and observe x_i infected bees, then $x_i \sim \text{Binomial}(N_i, p_i)$ So I can fit binomial generalized linear mixed-effects models (GLMMS) for estimating p_i at the visit level. I recommend using a binomial generalized linear mixed model (GLMM) with a logit link function to test factors such as bee diversity, floral abundance, and landscape features and predict the presence or absence of parasites in bumble bees. The response variable in this model would be binary (1 = parasite present, 0 = parasite absent).

For the fixed effects, consider including bee diversity (Bombus Shannon diversity), floral abundance, and landscape metrics such as the proportion of blueberry cultivation. To account for variability across sites and species, include random effects for sampling locations ($sample\_pt$) and bee species ($final\_id$).

It's essential to check the model's assumptions: (a) Test of the normality of residuals using the Kolmogorov–Smirnov test; and (b) Test for equal variance using Levene's test [10]. Also, before fitting the model, the Variance Inflation Factor (VIF) test for multicollinearity should be carried out; VIF values greater than 5 or 10 indicate the presence of multicollinearity. Dropping one of the correlated variables or creating a new variable representing both predictors is a useful approach to reduce multicollinearity [7]. The results become more tenable and reliable with the non-violation of those binomial GLMM assumptions. To verify model assumptions for binomial GLMMs were met, the 'testResiduals' function in the 'DHARMa' package can be used [5]

Two approaches can employ the prediction of parasite prevalence with a (GLMM): (1) adding a constant to independent variables with zero counts (e.g., native bee and impatiens abundance) before log-transforming the data and fitting GLMM [3], and (2) fitting GLMMs directly using the binomial likelihood, avoiding transformations for zero counts [11]. While the former can have a loss of straightforward interpretation of the model parameters as the effect sizes will be on the log scale, the latter will retain the original scale of the count data, allowing for a more straightforward interpretation of the model coefficients. The constant for zero values in log transforms should be data-driven, not arbitrary, to improve model fit.

GlmmPQL from the MASS package, lmer from the lme4 package, and glmmML from the glmmML package are the primary functions in R that can be used for GLMM. GLMMs handle temporal, spatial effects, non-normal data, clustering, and repeated measures correlations.

Alternative to GLMM is Structural Equation Modeling (SEM), which estimates direct and indirect effects (such as how bee diversity mediates the effect of landscape on parasite prevalence), requiring proper definition of causal pathways regarding exogenous, endogenous, and latent or observed variables. This study modelled predictors (exogenous) such as floral abundance, floral diversity, the proportion of blueberry, and landscape metrics as fixed effects. In contrast, the random effects section summarizes variability due to grouping factors like *sample_pt*, *subsite*, and $final\_id$. The piecewiseSEM R package incorporates SEM with random effects.

Multilevel SEM[1] include measurement invariance across levels and structural invariance in random coefficient models. At lower level[2] (within-transect level), non-normal residuals can lead to biased parameter estimates and incorrect standard errors. This means the relationships between predictors (like floral abundance and diversity) and bee abundance could be inaccurately represented. It's important to check for this in residual diagnostics (e.g. Q-Q plot). Applying transformation or the Bayesian method would be great options to consider. Although if there is non-normality at the higher level[3] of transects (*sample_pt*), it could mean that the variability in bee abundance between transects doesn't follow a normal distribution. While this won't heavily bias the estimates of how floral diversity or other predictors affect bee abundance, it could make the confidence intervals around those effects less reliable[6].

Also, to check if the assumption of constant residuals (Homoscedasticity) is not violated, residuals would be plotted against the fitted values; any observed pattern in the

[1]Multilevel SEM uses both the fixed effects of predictors and finds how the relationships vary across different sampling points (i.e., the random effect).

[2]Lower-level refers to within-group residuals (e.g., variability in bee counts within each transect)

[3]Higher-level in the model refers to between-group residuals (e.g., variability between different transects)

points indicates non-constant residuals. The spread of residuals should be roughly equal across different transects. Hence, group-level residual plots (e.g., residuals plot for each transect) are reasonable. If heteroscedasticity is confirmed, log-transforming the native bee abundance or *impatiens* abundance variables is the most common way to treat the issue. Also, I will allow the residual variance to vary with one or more predictors in the model, i.e. I will assume that the variance in bee abundance is related to floral diversity or landscape characteristics.

Similarly, I can apply the CR2 Estimator (Cluster-Robust Estimator). I would focus on situations where different transects may show different variability in the parasite prevalence, which can lead to non-constant variance. The CR2 estimator adjusts standard errors for these group-level differences.

Nonlinear relationships in the model can lead to biased estimates, necessitating appropriate diagnostic and estimation procedures. In this case, using mixture models of linear structural equation models to approximate the underlying, potentially nonlinear function is suggested. Firstly, the baseline SEM is fitted assuming linearity; if nonlinearity is confirmed in the plot of factor scores for each latent variable (e.g., bee abundance) against its predictors, then fitting mixture models to approximate nonlinearity should be considered next[1]. This approach helps decipher how different factors (like landscape diversity and floral composition) interact in potentially nonlinear ways to influence bee abundance and parasite dynamics.

Linear relationships between latent variables is the default assumption of multilevel SEM. However, suppose the genuine relationship between variables is nonlinear, but it's modeled is linearly. In that case, certain data points might appear as influential points as the model tries to fit a linear line to a nonlinear trend. To this effect, applying mixture models to account for nonlinear relationships between predictors and latent variables is recommendable.

While Bayesian SEM might be an excellent recommendation for fitting the data, there are a few possible approaches to handling them. Fitting Bayesian SEM using brms, I will recommend centering[4] certain variables (e.g., floral abundance, floral diversity, bee abundance) across different transects to account for both within-transect and between-transect variation. By implication, the intercept in the model would be the predicted value of the outcome when other predictors are at their mean. When interaction terms are included, centering makes it easier to interpret the effects of each variable with multicollinearity separately. In this study, centering could be used to separate within a transect from differences between transects.

Conclusion

The study involves a statistical model which can be analyzed and understood based on any function of interest considered for evaluation. A GLMM provides flexibility in incorporating fixed effects, random effects, and group-level variations. However, this might underperform when the mediation effect to be evaluated is complex. Multilevel SEM, on the other hand, can offer room to decompose the components in the causal pathways of the model, allowing it to accommodate the complexity surrounding the direct and indirect effects. When basic assumptions are violated and cannot be addressed in Multilevel SEM,

[4]Grand mean centering: Subtracting the overall mean of the variable from each observation.
Group-mean centering: Subtracting the group-specific mean from each observation within that group

Bayesian SEM might be a better option to consider. Instead of standardizing variables like the Shannon index, centering them might be a better transformation to consider, which might prevent all the models from having high Pareto k values and irregularities within the Bayesian framework.

Further reading

1. Mixed effects models and extensions in ecology with R [12].
2. Applied regression analysis - Chapter 18 [2].
3. Linear causal modeling with structural equations [8].
4. Bayesian SEM with brms in R: Using centering [9].

References

[1] Daniel J. Bauer, Ruth E. Baldasaro, and Nisha C. Gottfredson. Diagnostic procedures for detecting nonlinear relationships between latent variables. *Structural Equation Modeling: A Multidisciplinary Journal*, 19:157 – 177, 2012.

[2] NR Draper. *Applied regression analysis*. McGraw-Hill. Inc, 1998.

[3] John Paul Ekwaru and Paul J Veugelers. The overlooked importance of constants added in log transformation of independent variables with zero values: A proposed approach for determining an optimal constant. *Statistics in Biopharmaceutical Research*, 10(1):26–29, 2018.

[4] Letícia Vanessa Graf, Rafael Dudeque Zenni, and Rodrigo Barbosa Gonçalves. Ecological impact and population status of non-native bees in a Brazilian urban environment. *Revista Brasileira de Entomologia*, 64:e20200006, June 2020. Publisher: Sociedade Brasileira De Entomologia.

[5] Florian Hartig. Dharma: residual diagnostics for hierarchical (multi-level/mixed) regression models. *R Packag version 020*, 2018.

[6] Cora J. M. Maas and Joop J. Hox. The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Comput. Stat. Data Anal.*, 46:427–440, 2004.

[7] Habshah Bt. Midi, Shakuntala Sarkar, and Sohel Rana. Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics*, 13:253 – 267, 2010.

[8] Stanley A Mulaik. *Linear causal modeling with structural equations*. Chapman and Hall/CRC, 2009.

[9] Meghan E Quinn, Qimin Liu, David A Cole, Elizabeth McCauley, Guy Diamond, and Judy Garber. Relations among symptoms of depression over time in at-risk youth. *Journal of psychopathology and clinical science*, 2023.

[10] Holger Schielzeth, Niels J. Dingemanse, Shinichi Nakagawa, David F. Westneat, Hassen Allegue, CÃ©line Teplitsky, Denis RÃ©ale, Ned A. Dochtermann, LÃ¡szlÃ³ Zsolt Garamszegi, and Yimen G. Araya-Ajoy. Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, 11(9):1141–1152, 2020.

[11] Lianne K Siegel, Milena Silva, Lifeng Lin, Yong Chen, Yu-Lun Liu, and Haitao Chu. Choice of link functions for generalized linear mixed models in meta-analyses of proportions. *Research Methods in Medicine & Health Sciences*, page 26320843231224808, 2023.

[12] Alain F. Zuur, Elena N. Ieno, Neil J. Walker, Anatoly A. Saveliev, and Graham M. Smith. *GLMM and GAMM*, pages 323–341. Springer New York, New York, NY, 2009.

Appendix - Test of difference in proportions

```r
library(dplyr)
library(ggplot2)
library(infer)

bee_data <- bee_data
bee_data <- bee_data %>%
  mutate(
    infected = recode_factor(hascrithidia,
                             `1` = "Yes", `0` = "No"),
    species_status = recode_factor(status,
                                   "nonnative" = "Nonnative",
                                   "native" = "Native")
  )

contingency_table <- table(bee_data$species_status,
                           bee_data$infected)

# Calculate the observed difference in proportions
obs_diff_prop <- bee_data %>%
  specify(formula = infected ~ species_status,
          success = "Yes") %>%
  calculate(stat = "diff in props",
            order = c("Native", "Nonnative"))

obs_diff_prop <- round(obs_diff_prop, 3)
set.seed(1234)

# Generate the null distribution using permutation
null_distribution <- bee_data %>%
  specify(formula = infected ~ species_status,
          success = "Yes") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in props",
            order = c("Native", "Nonnative"))

# Calculate the p-value
p_value <- null_distribution %>%
  get_p_value(obs_stat = obs_diff_prop, direction = "both")

# Calculate the proportions of infected bees for each species status
prop_data <- bee_data %>%
  group_by(species_status) %>%
  summarize(prop_infected = mean(any_parasite == 1))
```