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

(ID: 70014261)

Client: Jenna Melanson
Consultant: Wakeel Kasali

Abstract

Pollinators, especially bumble bees (genus Bombus), are essential providers of ecosystem services in agroecosystems, yet they face numerous threats to their population health and persistence. The prevalence of parasites in bee populations is shaped by multiple factors, including floral abundance, floral diversity, landscape characteristics, and the abundance of native bees and *Bombus impatiens*, while accounting for temporal variation. To explore these complex relationships, various open-ended statistical methods may be employed. A proportion-based approach is recommended for modeling the parasite prevalence data, as it allows for the effective application of Generalized Linear Models (GLM). Careful dataset reorganization is important to ensure modeling of bee parasite prevalence. Additionally, Poisson regression is recommended for modeling bee abundance using a Generalized Linear Model (GLM) with the main dataset.

1. 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 the native bee's health [1]. 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, namely: (i) to assess how bumble bee abundance and diversity are influenced by landscape composition, diversity, and floral community characteristics; (ii) to 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) to evaluate how landscape characteristics impact the occurrence of bumble bee parasites.

The remainder of this report is as follows: Following a brief description about the dataset in Section 2, statistical hypotheses and questions are presented in Section 3. Suggestions on exploratory data analysis (EDA) and recommended statistical methods are summarized in Section 4 and 5. Section 6 applies conclusion. Some recommendations are shown in Appendix.

2. Data description and collection

A field study surveyed bumble bees, both native and non-native, across 267 unique transects (sampling units for bee surveys). These transects spanned six landscape types in the Lower Fraser Valley, near Vancouver. Sampling took place from May to August 2022. The dataset records each visit to a transect during 10 sampling rounds. Each sampling unit is a transect visit within a round. Bumble bee abundance and flowering plant diversity were recorded in 5-minute surveys along 50-meter transects. This resulted in a dataset comprising 1,690 rows from the combinations of transect and visit pairs.

Alongside the Bombus surveys, the landscapes surrounding each sampling location was classified into 16 categories (e.g., annual crops, blueberry fields, forest, hedgerows, weedy edges). Rasterized land cover data were used to calculate three landscape metrics within a 500-meter buffer around each transect: Shannon diversity, proportion of edge area (e.g., hedgerows, grassy margins), and proportion of blueberry cultivation.

Bombus and floral diversity were calculated using the Shannon Index and then standardized. Native and non-native bees were recorded as counts. Floral abundance, based on flower species counts, was log-transformed and standardized.

Of the 3,145 bumble bees collected, 749 individuals from six species were screened for four common parasites (*Crithidia spp.*, *Nosema spp.*, *Apicystis spp.*, and *Ascosphaera spp.*). The 749 screened bees came from combinations of (transect, visit), ensuring that all landscape types were represented in the dataset. However, the second dataset for the parasite analysis has significantly fewer rows, n, where $n \ll 749$. Parasite occurrence for each bee was recorded as a binary variable (0 = none, 1 = detected) for any of the four parasites.

3. Statistical Hypotheses and Questions

1. H_0: Simplified landscapes and landscapes with higher proportions of blueberry cultivation do not significantly affect the abundance and diversity of bumble bees.

2. H_0: Floral and bee community diversity, as well as floral and bee abundance, do not significantly affect parasite hosting rates.

3. H_0: Landscape simplification and the proportion of blueberry cultivation do not significantly affect parasite prevalence, either directly or indirectly through impacts on bee diversity.

The primary statistical questions arising from the initial analyses pertain to the second hypothesis and can be summarized as follows:

1. Could lower parasite prevalence with abundant *Bombus impatiens* be a biological effect or sampling variability?

2. Can parasite presence in bee population be predicted by bee diversity, bee abundance, flower abundance, landscape characteristics (e.g., proportion of blueberry cultivation), and seasonal timing (Julian date)?

3. What could be the cause of high pareto k values in the parasite prevalence model ?

4. Exploratory Data Analysis

Before conducting formal analysis, it is important to assess relationship between the variables. In the first hypothesis, since the samples for landscape Shannon diversity and the proportion of blueberry cultivation are transect-based, a scatter plot can be used to visualize the relationship between the two variables. Similar plots should also be created to examine the relationships between other pairs, such as landscape Shannon diversity and total bee abundance or diversity per transect, as well as the proportion of blueberry cultivation and total bee abundance or diversity per transect.

For the second hypothesis, EDA can help explain the possible artificial association between *Bombus impatiens* abundance and infection status by revealing patterns and relationships. One practical approach is to determine whether the lower parasite prevalence observed with abundant *Bombus impatiens* abundance is a true biological effect or simply a result of sampling variability.

For each combination of transect and sampling round, the proportion of native bees infected and the proportion of *Bombus impatiens* infected by each parasite type (e.g., *Crithidia*) can be calculated. These proportions, represented as paired points on a scatter plot, can then be plotted, with each point corresponding to a specific (transect, sampling round) combination. The next step is to assess whether the points predominantly fall on one side of the diagonal line, which would suggest a biological effect where one bee type is more susceptible to infection than the other. Examples of such plots are provided in Figure 1.

5. Recommended statistical methods

The proposed statistical methods corresponding to the three statistical questions and the first hypothesis one are outlined below:

5.1. Statistical Question One

At transects with high *Bombus impatiens* abundance; there seemed to be greater likelihood of collecting and screening *Bombus impatiens* compared to other bee species than at transects where *Bombus impatiens* abundance was low. This resulted in an unexpected effect on parasite prevalence: transects with more *Bombus impatiens* had fewer parasites, contrary to what is expected in the study.

Assuming the second dataset (for the parasite model) has been augmented with eight (8) pairs of data (x, n), where x represents the number of bees infected with a particular parasite, and n is the total number of bees screened, regardless of their infection status. Since there are four (4)types of parasites (e.g., *Crithidia*, *Nosema*), and bees are classified into two categories, native and non-native, this results in 8 pairs of data (4 parasites $\times$ 2 bee classifications). For each row, which represents a visit to a transect, these pairs will capture the number of native and non-native bees infected with each of the four (4) parasites. After obtaining the eight (8) (x, n) pairs, proportions (x/n) can be calculated for each parasite type in both native and non-native bees.

A potential biological effect, rather than sampling variability, can be further verified through a test of the difference in proportions of parasite prevalence between native and non-native bees. Since differences in proportions are not normally distributed, calculating the difference in proportions and testing it using a permutation-based null distribution

would be appropriate, as this approach reduces the influence of confounding factors such as sample size and sampling variability in bee collection methods. A similar code to perform this analysis is provided in the appendix (Test of difference in proportions).

5.2. Statistical Question Two

After augmenting the dataset with the (x, n) pairs and corresponding proportions, a binomial generalized linear model (GLM) can be applied, using bee diversity, bee abundance, flower abundance, and Julian date as covariates. See Appendix for sample R code for the binomial generalized linear model. If the analysis of parasite prevalence did not use a dataset of this form, it might explain the high Pareto k values. Moreso, If sufficient data are available in the second dataset, the analysis can be stratified by landscape type ; although, the likely small sample size in the second dataset may not be sufficient to account for random effects.

For the first hypothesis where the main dataset is used, a poisson regression is recommended for modeling native bee abundance and *Bombus impatiens* abundance as both are count data. An example of this model is shown in equation 1. Each transect visit serves as a sampling unit. For *Bombus* Shannon diversity, a linear model (lm) as in 2 or a linear mixed-effects model (lme) is appropriate. Samples of codes that can be used in GLM are provided in the appendix. **lmer** from the **lme4** package, and **glmmML** from the **glmmML** package are the primary functions in R that can be used for GLMM.

6. Conclusion

For the first hypothesis, poisson regression is recommended for modeling native bee abundance and *Bombus impatiens* abundance, as both are count data, with each transect visit serving as a sampling unit.

The second dataset, augmented with eight (x, n) pairs representing parasite types in native and non-native bees, allows for calculating proportions of infected bees across four parasite types. The observed lower parasite prevalence in transects with high *Bombus impatiens* abundance may be influenced by sampling variability rather than a true biological effect, which can be clarified through a test of difference in proportions between native and non-native bees using a permutation-based null distribution. Similarly, parasite presence in bee populations can be effectively predicted by bee diversity, bee abundance, floral abundance, landscape characteristics (e.g., proportion of blueberry cultivation), and seasonal timing (Julian date), with Generalized Linear Models (GLMs).

If the format of second dataset has been misunderstood, please make another submission to ASDa with more details, and the Stat 551 class might be able to provide better advice.

Further reading

1. Mixed effects models and extensions in ecology with R [2].
2. Applied regression analysis - Chapter 18 [3].

References

[1] L. V. Graf, R. D. Zenni, and R. B. Gonçalves, "Ecological impact and population status of non-native bees in a Brazilian urban environment," *Revista Brasileira de Entomologia*, vol. 64, p. e20200006, Jun. 2020, publisher: Sociedade Brasileira De Entomologia. [Online]. Available: https://www.scielo.br/j/rbent/a/gRxQqWp4Bsyg3QTXzVhgXht/?lang=en&format=html

[2] A. F. Zuur, E. N. Ieno, N. J. Walker, A. A. Saveliev, and G. M. Smith, *GLMM and GAMM*. New York, NY: Springer New York, 2009, pp. 323–341. [Online]. Available: https://doi.org/10.1007/978-0-387-87458-6_13

[3] N. Draper, *Applied regression analysis*. McGraw-Hill. Inc, 1998.

Appendix

Proportion of Infected native bees vs Impatiens bee by parasite type

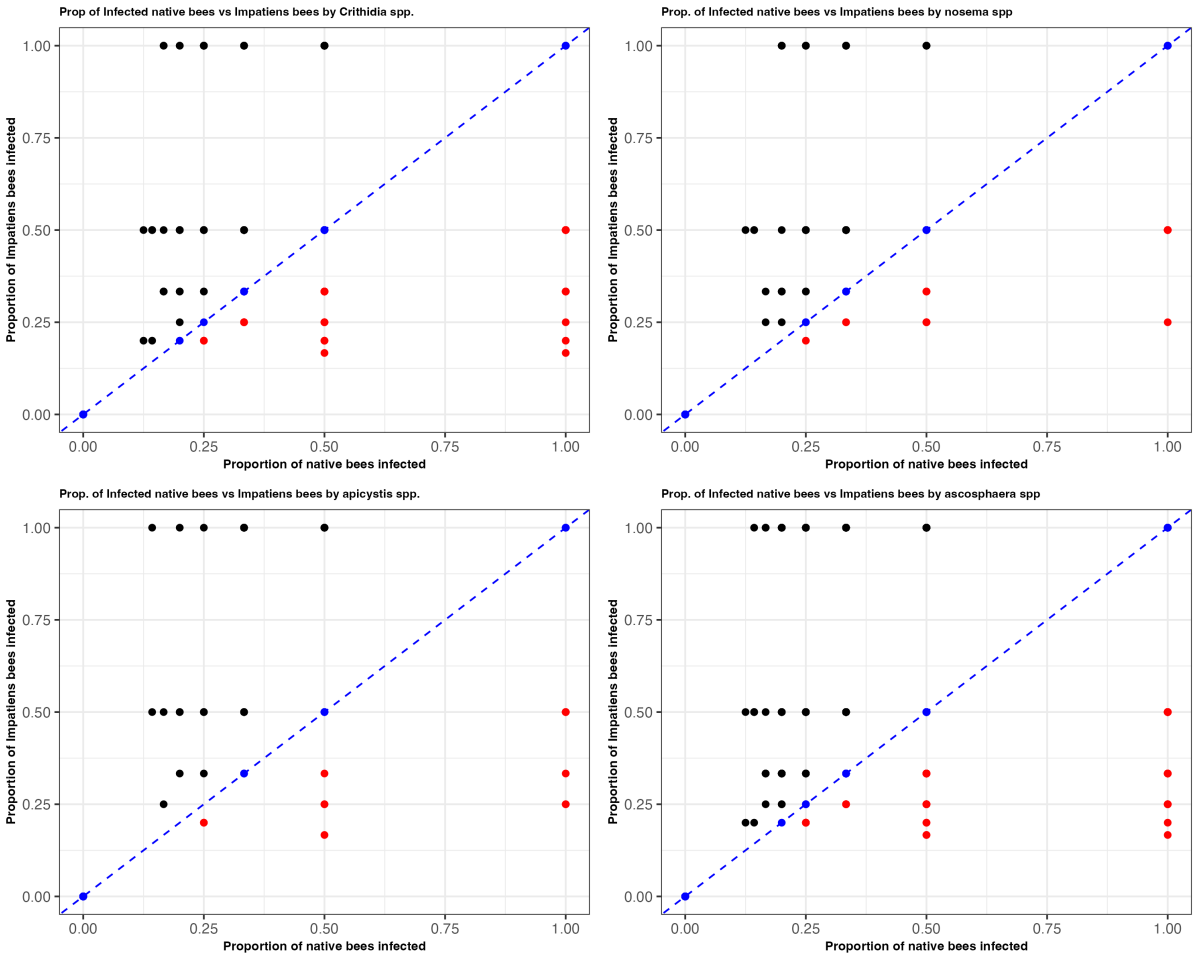

Figure 1: Plot of parasite types by infected proportion between native and impatiens bee for each (transect and sampling round combination)

Model equations

$$\text{native\_bee\_abundance}_{i,j} \sim \text{Poisson}(\lambda_{i,j})$$

$$\begin{aligned} \log(\lambda_{i,j}) = \beta_0 + \beta_1(\text{floral\_abundance}_{i,j}) + \beta_2(\text{floral\_diversity}_{i,j}) \\ + \beta_3(\text{landscape\_shdi}_{i,j}) + \ldots + \epsilon_{i,j} \end{aligned} \tag{1}$$

$$\begin{aligned} \text{bombus\_shannon\_diversity}_{i,j} = \gamma_0 + \gamma_1(\text{floral\_abundance}_{i,j}) + \gamma_2(\text{floral\_diversity}_{i,j}) \\ + \gamma_3(\text{landscape\_shdi}_{i,j}) + \ldots + \epsilon_{i,j} \end{aligned} \tag{2}$$

Test of difference in proportions

```
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))
```

Binomial for how to use glm with success, failure data

```
set.seed(551)
x = rnorm(20)
b0=0.1; b1=0.4
px = 1/(1+exp(-b0-b1*x))
nbinom = floor(runif(20,4,11))
success = rbinom(20,nbinom,px)
failure = nbinom - success
response = cbind(success,failure)
print(response)

fit = glm(response~x, family=binomial)
print(summary(fit))
#Coefficients:
#            Estimate Std. Error z value Pr(>|z|)
#(Intercept)   0.0567     0.1827   0.310    0.756
#x             0.2826     0.1908   1.481    0.139
```