**SUPPLEMENTARY INFORMATION**

**Unhealthy Brood Odor Assay: On Detecting Change Points in Pathogen Load Data**

Samantha A. Alger, M. Sydney Miller, P. Alexander Burnham, Esmaeil Amiri, Corinne Jordan, Kaira Wagoner

*September 25, 2024*

**SUMMARY**

While our main analyses and the intention in this work was to determine if pathogen loads significantly decrease in colonies as UBeeO score increases, the question of threshold values remains an important secondary question. These thresholds have biological relevance as they could serve as reference points for colony identification when selecting for resistance to a given pathogen or pathogens. In her previous findings, Kaira Wagoner, through a combination of statistical inference and biological reasoning, identified 60% at the significance threshold for the *Varroa* mite (Wagoner et al., 2021). Here, we aim to identify thresholds for the remaining significant pathogens identified in this work and using novel methods described below, conduct a reanalysis the *Varroa* count data from Wagoner et al., 2021. We ask two main questions: **(1)** What are the UBeeO values for each pathogen above which pathogen loads are lower? **(2)** If these cutoffs do exist, do they vary by pathogen? In this supplementary document we aim to answer these questions while leaving our explanation of our findings and integration with the main results of this paper to the discussion section of the primary document.

**METHODS**

In order to determine UBeeO thresholds or cutoff points for each significant pathogen, we used change point analysis implemented in the MCP (Multiple Change Point) package in R (Lindeløv, 2020; R Core Team, 2024). This methodology, traditionally used for identifying changes of regime in time-series data is ideally suited for determining thresholds where the processes underlying a dataset have changed resulting in a differential response of the dependent variable.

***The MCP Package:*** The MCP package is a Bayesian tool designed to detect and model change points in time series or other ordered data sequences. Change points refer to locations in the data where statistical properties, such as mean or trend, have shifted significantly. MCP allows users to specify models for different data segments, where each segment is characterized by its own set of statistical parameters (e.g., mean, slope). These segments can fit together in any number of user-specified ways. Joined slope indicates a hinge relationship where the slope has changed at some point along the x-axis. The underlying model might also describe a series of disjointed slopes or plateaus indicative of a shelf-like structure. MCP uses Bayesian inference, which enables the estimation of change point locations by considering prior knowledge about the data. Through Markov Chain Monte Carlo (MCMC) sampling, the package generates posterior distributions for both the change points and the parameters of the segments. This probabilistic framework accounts for uncertainty and variability, making the identification of change points more robust (Lindeløv, 2020).

***Our Implementation:***

We identified change points for seven pathogens. *Varroa*, *Vairimorpha*, Chalkbrood, DWV.A (Deformed Wing Virus-A), DWV.B (Deformed Wing Virus-B), IAPV (Israeli Acute Paralyses Virus), and LSV (Lake Sinai Virus). For all data, missing data were removed from the analysis. For *Vairimorpha*, and Chalkbrood, and the viruses, which were all integer count data, we fit a Poisson model. *Varroa* count was an average of multiple counts and as such, was not of data type integer. However, since Varroa load is naturaly count data, these data were rounded and also fit with a Poisson distribution. For all seven pathogens, raw scatter plots of pathogen load by UBeeO score indicated an underlying shelf or two plateau structure. As such, the MCP algorithm was passed a two-segment model:

Shelf\_model = list(

load ~ 1, # plateau (int\_1)

~ 1 # plateau (int\_2)

)

All models were run for 100,000 iterations. The change point estimate, upper and lower error boundaries, and scatter plot with posterior distributions and test segments were recorded. Additionally, for the virus loads, a second set of change points were determined. It is often the case that, due to the large range in orders of magnitude encountered in virus loads data, a log transformation might be conducted for analysis or visualization purposes. All virus data were log10 + 1 transformed. As these transformed data were relatively normally distributed, a Gaussian distribution was fit. The data, when plot against UBeeO score seemed to exhibit less of a shelf-cutoff structure and seemed more linear in nature. A joined slope was used in this case:

slope\_model = list(

load ~ 1 + UBO\_Score, # intercept + slope

1 ~ 0 + UBO\_Score # joined slope

)

As previously, all models were run for 100,000 iterations. The change point estimate, upper and lower error boundaries, and scatter plot with posterior distributions and test segments were recorded.

**RESULTS**

The change points indicating the threshold where UBeeO score begins to significantly decrease disease loads are shown below (Table S1). Models using a Poisson distribution as well as leveraging a two-plateau model structure most closely detected the apparent change in regime (Figures S1-S4). The log10 transformed virus loads fit with a joined slope model and a Gaussian distribution had a much larger range of error (Figure S5). Change points varied by pathogen ranging from 4.9 to 69% UBeeO scores (Table S1).

**Table S1:** The change point estimates for the seven pathogens that significantly decreased with increasing UBeeO scores in our study and others (Wagoner et al., 2021). Lower and upper describe the uncertainty around each estimate. The model structure indicates the assumed underlying data structure, either a change in slope (joined slope) or a shelf cutoff (two plateau). Count data were modeled on a Poisson distribution and log transformed virus load data were modeled using a Gaussian distribution.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Pathogen** | **change point** | **Lower** | **Upper** | **Model Structure** | **Distribution** |
| Vairimorpha | 69 | 63 | 74 | Two Plateau | Poisson |
| Chalk Brood | 13 | 13 | 14 | Two Plateau | Poisson |
| Varroa\* | 52.84 | 51.60 | 57 | Two Plateau | Poisson |
| DWV.A | 36.2 | 32.7 | 39.9 | Two Plateau | Poisson |
| DWV.B | 4.9 | 0.19 | 9.5 | Two Plateau | Poisson |
| IAPV | 19 | 11 | 26 | Two Plateau | Poisson |
| LSV | 40 | 40 | 40 | Two Plateau | Poisson |
| DWV.A | 56.443 | 8.106 | 99.908 | Joined Slope | Gaussian log10(x) |
| DWV.B | 55.012 | 3.501 | 100 | Joined Slope | Gaussian log10(x) |
| IAPV | 53.976 | 2.303 | 97.015 | Joined Slope | Gaussian log10(x) |
| LSV | 38.697 | 0.017 | 93.938 | Joined Slope | Gaussian log10(x) |

*\* The varroa dataset used to derive these values was taken from Wagoner et. al., 2021.*



**Figure S1:** Change point estimates for Chalkbrood load (cells per colony) plotted by UBeeO score utilizing a two plateau model and an underlying Poisson distribution. Grey segments represent test segments drawn from the model’s posterior distribution. The blue curve represents the posterior distribution itself.



**Figure S2:** Change point estimates for *Vairimorpha* load (spores per bee) plotted by UBeeO score utilizing a two plateau model and an underlying Poisson distribution. Grey segments represent test segments drawn from the model’s posterior distribution. The blue curve represents the posterior distribution itself.



**Figure S3:** Change point estimates for Varroa load (mites per 100 bees) plotted by UBeeO score utilizing a two plateau model and an underlying Poisson distribution. Grey segments represent test segments drawn from the model’s posterior distribution. The blue curve represents the posterior distribution itself.



**Figure S4:** Change point estimates for virus load plotted by UBeeO score. Results represent a two plateau model using an underlying Poisson distribution for **(A)** DWV.A, **(B)** DWV.B, **(C)** IAPV, and **(D)** LSV. Grey segments represent test segments drawn from the model’s posterior distribution. The blue curve represents the posterior distribution itself.



**Figure S5:** Change point estimates for virus load plotted by UBeeO score. Results represent a joined slope model using an underlying Guassian distribution on log-transformed data for **(A)** DWV.A, **(B)** DWV.B, **(C)** IAPV, and **(D)** LSV. Grey segments represent test segments drawn from the model’s posterior distribution. The blue curve represents the posterior distribution itself.

**REFERENCES**

K. Wagoner, J. G. Millar, J. Keller, J. Bello, P. Waiker, C. Schal, M. Spivak, O. Rueppell, Hygiene-Eliciting Brood Semiochemicals as a Tool for Assaying Honey Bee (Hymenoptera: Apidae) Colony Resistance to *Varroa* (Mesostigmata: Varroidae), *Journal of Insect Science*, Volume 21, Issue 6, November 2021, 4, <https://doi.org/10.1093/jisesa/ieab064>

Lindeløv, J. K. (2020). mcp: An R Package for Fitting Bayesian Regression Models with Multiple Change Points. Journal of Open Source Software, 5(51), 2913. <https://doi.org/10.21105/joss.02913>

R Core Team (2024). R: A Language and Environment for Statistical Computing. R

Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.