

Supplementary Material on: Detection and characterization of incidental vibrations from *Drosophila suzukii* in infested fruits – Statistical analysis

Lorenzo Fellin,^{1,2} Giulia Bertagnolli,³ Valerio Mazzoni,^{2,*} Gianfranco Anfora,¹

Claudio Agostinelli,³ Marco Valerio Rossi Stacconi,² and Rachele Neri^{1,3}

¹Center for Agriculture, Food and Environment, University of Trento, Italy

²Research and Innovation Centre, Fondazione Edmund Mach, Italy

³Department of Mathematics, University of Trento, Italy

This Supplementary Material collects detailed information about the statistical analysis. As already described in the main article, the dataset consists of the three-minutes vibrational recordings at five different pest ages: 48, 96, 168, 216 and 264 hours after *D. suzukii* infestation, with a frequency spectrum in the range [100, 2000]Hz. The three-minutes recordings have been summarized in two ways: taking the average over time at each frequency (as described in the main article), and also taking a functional median over time [1]. For the latter, we used a depth measure for functional data, in particular, the integration of the univariate Tukey's depth [2, 3]. The final dataset consists of 36 functional observations, each corresponding to a blueberry, representing a amplitude ($\mu\text{m/s}$) observed over the spectrum of frequencies [100, 2000]Hz. Of the $n = 36$ blueberry, 6 are controls, i.e., not infected by the *Drosophila*, see the main article for more details. In the following, we denote the amplitude by the random variable/function $Y(h)$, with $h \in [100, 2000]\text{Hz}$ and dependent variables, also called *factors* or *control variables* by $\{X_1(h), X_2(h), \dots\}$.

FUNCTIONAL ANALYSIS OF VARIANCE

We firstly test for the effect of infestation and of pest age, on the spectrum of the vibrational recordings. The first factor has two levels, representing control and infested blueberries, while, as already mentioned, the pest age has five levels, corresponding to time intervals (hours) after infestation. The linear model reads

$$Y_{ijk}(h) = \mu(h) + \alpha_i(h) + \beta_j(h) + \varepsilon_{ijk}(h), \quad h \in [100, 2000]\text{Hz} \quad (1)$$

with $k = 1, \dots, n$ indicating the observational unit, $i = 1, 2$ the control and infested groups respectively, and $j = 1, \dots, 5$ the pest ages; $\mu(h)$ is the grand mean of the observations, i.e. is constant for the population, while $\alpha_i(h)$ and $\beta_j(h)$ are the i -th and j -th main-effect functions of the two factors, and finally, $\varepsilon(h)$ represents the error. In the usual context of linear models, it is assumed that errors be normally distributed with mean zero and constant variance. In the permutation framework, we assume ε_{ijk} be exchangeable with zero mean [4, 5]. Observe that we do not consider interaction effects. The usual sum to zero contrasts is used for identifiability of model parameters.

We performed a two-way functional analysis of variance ANOVA [6, 7]

$$\begin{aligned} H_{0,1} &: \alpha_i(h) \equiv 0, i = 1, 2 \text{ and } h \in [100, 2000]\text{Hz}, \\ H_{0,12} &: \alpha_i(h) \equiv 0, i = 1, 2 \text{ and } h \in [100, 2000]\text{Hz}, \\ &\quad \beta_j(t) \equiv 0, j = 1, 2, \dots, 5 \text{ and } h \in [100, 2000]\text{Hz}. \end{aligned}$$

$H_{0,1}$ states that the main effect of infestation is not significant, i.e., on average there are no differences between infested and control berries. Under $H_{0,12}$, the main effects of the two factors are **simultaneously** not significant. The alternative hypothesis is the negation of the null hypothesis.

Let us momentarily drop the dependence on h . ANOVA significance test is based on the decomposition of the total sum of squares

$$\sum_k (y_k - \bar{y})^2 = \sum_k (\hat{y}_k - \bar{y})^2 + \sum_k (y_k - \hat{y}_k)^2 \quad (2)$$

where the first term on the right-hand-side is the regression sum of squares (SSR), i.e. the amount of total variation explained by the regression model. $\hat{y}_k$ represents the k -th fitted value and $\bar{y}$ the overall sample mean. The second

* Corresponding author: marcovalerio.rossistacconi@fmach.it

term involves the residuals $y_k - \hat{y}_k$, and is called the error (or residual) sum of squares (SSE), which does not increase as more explanatory variables are added to the model. The total sum of squares (SST), left-hand-side in (2), does not depend on the model, so we focus on the marginal increase on the SSR = SST - SSE, as more explanatory variables are taken into consideration. Let $SSR(X_1, X_2)$ denote the SSR for the linear model with X_1 and X_2 , then

$$SSR(X_1, X_2) = SSR(X_2) + SSR(X_1|X_2) = SSR(X_1) + SSR(X_2|X_1).$$

In the following, we always consider **partial sum of squares** in ANOVA tables (also called II and III type), $SSR(X_1|X_2)$ and $SSR(X_2|X_1)$ and, in general, $SSR(X_i|\{X_1, \dots, X_p\} \setminus \{X_i\})$, which quantify the additional variability explained adding one variable, when all other variables are already in the model.

Let us go back to the functional case: we choose, as test statistic, the point-wise F -test statistics [6, 7], i.e. the usual F -statistic evaluated at each frequency, and its global version, the globalizing point-wise F -test (GPF test) [8], as these can be easily extended to the more complicated models that follow. The point-wise F -type test statistic reads

$$F_n(h) = \frac{SSR_n(h)/(l-1)}{SSE_n(h)/(n-l)}, \quad (3)$$

where l is the number of groups or, more precisely, the number of levels of the factor variable considered and $l-1$ (resp. $n-l$) are the degrees of freedom of the model (resp. residual). Then, $GPF_n = \int F_n(h)dh$, which, in the empirical case, is a simple finite summation over all frequencies.

The p -values for both point-wise and globalizing statistics are computed through permutation tests with 100,000 permutations, following Manly's approach [4, 5, 9, 10]. Results are shown in Fig. 2 of the main manuscript.

The results of a similar two-way functional ANOVA model with pest age and, this time, *infestation level*, are shown in Fig. 5 of the main article. The *infestation level* factor has three levels: control (0 pupae), low (number of pupae in the range [1, 6]) and high (7 or more pupae).

A. Post-hoc functional analysis

Once we have established that not all group means are equal, we investigate which group, or groups, led to the rejection of the null hypothesis. This is exactly the aim of the post-hoc analysis, for which we chose Quade's multi-comparison test [11]. Since these tests are more powerful for balanced data, we only use a subset of the dataset, dropping the controls, and focus on the effect of pest age.

The Quade test assumes that there are $l \geq 2$ experimental treatments and that the observations are arranged in b blocks. Each treatment is present once and only once within each block (randomized complete block design). In our example we have one block for each berry ($b = 30$) and $l = 5$ treatments, the five pest ages, so we have one observation per cell, as shown in the following table

Berry (block)	pest age (treatment)			
	1 (48h)	2 (96h)	...	5 (264h)
1	Y_{11}	Y_{12}	...	Y_{1l}
2	Y_{21}	Y_{22}	...	Y_{2l}
3	Y_{31}	Y_{32}	...	Y_{3l}
...	...	...	...	...
b	Y_{b1}	Y_{b2}	...	Y_{bl}

The assumptions for the Quade test are:

- The b k -variate random variables are mutually independent.
- There exists an order relation between the observations in each block (i.e. it is possible to compute ranks).
- It is possible to compute the sample ranges of the blocks.

Let $R(Y_{ij})$ be the rank (from 1 to $l = 5$) assigned to Y_{ij} within block i (increasing order, i.e. rank 1 to smallest frequency). Then, we rank the blocks according to their sample range, where the range of block i is $\max_j Y_{ij} - \min_j Y_{ij}$ and call $Q_1, \dots, Q_b$ the ranks of the respective blocks. Consider the statistic

$$S_{ij} = Q_i \left[R(Y_{ij}) - \frac{l+1}{2} \right]$$

where $\frac{l+1}{2}$ is the average rank within blocks: S_{ij} represents the relative size of each observation in each block, weighted on the relative significance of the block. Let us also take the sum over blocks, for each treatment $S_j = \sum_{i=1}^b S_{ij}$.

We are interested in the multiple comparisons between pest ages, hence Quade’s test states that two groups i and j are significantly different at level $1 - \alpha$ if

$$|S_i - S_j| > t_{1-\alpha/2} \left[\frac{2b(A-B)}{(b-1)(k-1)} \right]^{1/2},$$

where $t_{1-\alpha/2}$ is the t -Student’s quantile. p -values are computed according to definition, i.e. for the two-sided test, being $t = \frac{|S_i - S_j|}{\sqrt{\frac{2b(A-B)}{(b-1)(k-1)}}}$ non-negative, the p -value is computed as twice the p -value of the upper-tail: $p = 2\mathbb{P}(T \geq t)$, where T is distributed according to Student’s t -distribution with $(b-1)(k-1)$ degrees of freedom. Again, instead of using the tabulated distribution for our statistics, we implement a permutation procedure for the pointwise and globalizing statistics. Results are shown in Fig. 3 in the main article.

FUNCTIONAL ANALYSIS OF COVARIANCE

Finally, since we also expect that infestation level (zero, low, and high number of pupae and/or eggs) have an effect on amplitude, we perform a functional analysis of covariance with the usual categorical predictor, i.e. pest age *hour*, and additional variates: number of *eggs*, number of *pupae*, and, to account also for possible interactions the ratio between the number of pupae and eggs and the interaction between *pupae* and *eggs*.

Let us look at a sample table from this model, for $h = 1000\text{Hz}$:

```
aovperm(formula = value ~ (hr + eggs * pupae + ratio), method = "manly")$table
## Anova Table
## Resampling test using manly to handle nuisance variables and 5000 permutations.
##
```

	SS	df	F	parametric P(>F)	resampled P(>F)
## hr	0.0638994	4	5.8842	0.000208	0.0004
## eggs	0.0006777	1	0.2496	0.618109	0.6074
## pupae	0.0273900	1	10.0889	0.001833	0.0010
## ratio	0.0040358	1	1.4866	0.224787	0.2284
## eggs:pupae	0.0126089	1	4.6444	0.032853	0.0312
## Residuals	0.3827952	141			

The p -values as functions of frequencies are shown in Fig. 4 of the main manuscript. In this case, since the interaction term is significant (for $\alpha = 0.05$) we should not look at the “pure” main effects of *eggs* and *pupae*. However, we know from previous ANOVA tests that the infestation level—(control, low, high) obtained for different levels of number of pupae in the berries—is significant, so we feel safe dropping the variable *eggs* and keeping the main effect of *pupae*. Also, because the interaction term is not always significant, whereas the main effect of the number of pupae does.

REFERENCES

- [1] S. López-Pintado and J. Romo, On the concept of depth for functional data, *Journal of the American statistical Association* **104**, 718 (2009).
- [2] R. Fraiman and G. Muniz, Trimmed means for functional data, *Test* **10**, 419 (2001).
- [3] M. Febrero-Bande and M. Oviedo de la Fuente, Statistical computing in functional data analysis: The R package *fda.usc*, *Journal of Statistical Software* **51**, 1 (2012).
- [4] P. Good, *Permutation Tests* (Springer New York, 1994).
- [5] D. Basso, F. Pesarin, L. Salmaso, and A. Solari, Permutation tests, in *Permutation Tests for Stochastic Ordering and ANOVA* (Springer New York, 2009) pp. 1–35.
- [6] T. Gorecki and L. Smaga, *fdANOVA: Analysis of Variance for Univariate and Multivariate Functional Data* (2018), r package version 0.1.2.
- [7] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis* (Springer New York, 2005).
- [8] J.-T. Zhang and X. Liang, One-way ANOVA for functional data via globalizing the pointwise F-test, *Scandinavian Journal of Statistics* **41**, 51 (2014).
- [9] B. F. J. Manly, *Randomization, bootstrap and Monte Carlo methods in biology: Texts in statistical science* (Chapman and Hall/CRC, 2018).
- [10] J. Frossard and O. Renaud, Permutation tests for regression, ANOVA, and comparison of signals: The *permuco* package, *Journal of Statistical Software* **99**, 1 (2021).
- [11] W. J. Conover, *Practical nonparametric statistics*, Vol. 350 (John Wiley & Sons, 1999).