1 **Resolving the identification of weak-flying insects during flight: a coupling between rigorous data**

2 **processing and biology.**

3 Kirsty L. Hassall[1*], Alex Dye[2], Ilyas Potamitis[3] & James R. Bell[2]

5 [1] *Computational and Analytical Sciences, Rothamsted Research, West Common, Harpenden, AL5 2JQ,*

6 *UK.*

7 [2] *Rothamsted Insect Survey, Rothamsted Research, West Common, Harpenden, AL5 2JQ, UK.*

8 [3] *Department of Music Technology and Acoustics Engineering, Hellenic Mediterranean University,*

9 *Crete*

10 *corresponding author: kirsty.hassall@rothamsted.ac.uk

12 Running Title: **Identifying weak-flying insects during flight**

14 **Abstract**

15 1. Bioacoustic methods play an increasingly important role for the detection of insects in a range of

16   surveillance and monitoring programs.

17 2. Weak-flying insects evade detection because they do not yield sufficient audio information to

18   capture wingbeat and harmonic frequencies. These inaudible insects often pose a significant

19   threat to food security as pests of key agricultural crops worldwide.

20 3. Automatic detection of such insects is crucial to the future of crop protection by providing critical

21   information to assess the risk to a crop and the need for preventative measures.

22 4.   We describe an experimental setup designed to derive audio recordings from a range of weak-

23   flying aphids and beetles using an LED array.

24   5. A rigorous data processing pipeline was developed to extract meaningful features, linked to

25   morphological characteristics, from the audio and harmonic series for six aphid and two beetle

26   species.

27   6. An ensemble of over 50 bioacoustic parameters was used to achieve species discrimination with

28   a success rate of 80%. The inclusion of the dominant and fundamental frequencies improved

29   prediction between beetles and aphids due to large differences in wingbeat frequencies.

30   7. At the species level, error rates were minimised when harmonic features were supplemented by

31   features indicative of differences in species' flight energies.

32   **Keywords**

33   Insect classification; Random forest classification; Wingbeat harmonics; Aphid; Beetle;

34   **Introduction**

35   Over the last 30 years, agriculture has been in the midst of a digital revolution with the increasing

36   availability of sensor technology and associated collection of 'big data' aiming at improving the

37   sustainability of food production systems. These technologies exist at all levels of the agricultural

38   system, from defining management zones from yield monitor data (Milne et al., 2012), drones and

39   augmented reality (Huuskonen and Oksanen, 2018) to pest detection using UAVs (Tetila et al., 2020).

40   Many of these technologies feed into decision support systems (DSS) further enabling the

41   implementation of precision agriculture (Cancela et al., 2019, Zhai et al., 2020). Despite this, insect

42   pests remain a key challenge. Automatic insect pest detection is a long sought-after goal that began

43   in the 1950s and has still yet to reach maturity (Lowe and Dromgoole, 1958, Cardim Ferreira Lima et

44   al., 2020). For many years, image recognition systems have been at the forefront but increasingly,

45   bioacoustic methods are playing an important role for the detection of insects in a diverse range of

46   surveillance and monitoring programs. Almost exclusively, these programs use sound recordings to

47   detect species and groups across the audible range of human hearing (Chen et al., 2014). Typically,

48   model organisms have included mosquitoes, fruit flies, hawkmoths and crickets using a simple

49   microphone set up (Montealegre-Z et al., 2011, Potamitis and Rigakis, 2016b, Mukundarajan et al.,

50   2017). Whilst the songs of some cicadas can generate in excess of 100 dB of sound pressure (Sanborn

51   and Phillips, 1995), there is an important group of small insects that are effectively silent in flight which

52   pose a much more significant threat. For example, the peach potato aphid, *Myzus persicae* (Sulzer,

53   1776) (Hemiptera: Aphididae), compromises worldwide food security through the transmission of 100

54   different plant viruses, and is consequently one of the world's top 10 pests (Willis and *ed.*, 2017, CABI,

55   2020). Yet, these small insects, no more than a couple of millimetres long and of the upmost

56   agricultural importance, are not the focal interest of bioacoustics monitoring, despite their profound

57   and lasting impact on food quality and quantity.

58   Aphid flights are so weak that they fall well below the lowest human hearing range of 40 dB at 100

59   hertz and are effectively silent, producing a wingbeat frequency that is nearly an order of magnitude

60   weaker than some mosquitoes (Smith, 1999, Byrne et al., 1988, Moore and Miller, 2002, Potamitis

61   and Rigakis, 2016b, Tercel et al., 2018). Further, the rate of progress in flight is very weak, and no more

62   than 0.70 m/sec^{-1} under laboratory conditions for some well-studied aphids (Thomas et al., 1977).

63   Recently, opto-acoustic methods have provided a novel way of capturing the flight of these small

64   insects using phototransistors and infrared light (Ouyang et al., 2015, Potamitis and Rigakis, 2016a,

65   Potamitis and Rigakis, 2016b). Here, the use of both the extinction of light and backscattered light

66   principles have been shown to perform better than audio (Potamitis et al., 2015, Potamitis and Rigakis,

67   2016b).

68   In parallel to the evolution of sensor technologies has been a rapid development in computational

69   techniques. Classification problems once thought to be impossible, can be tackled by any number of

70   different data science techniques. Such techniques have been widely applied to the classification of

71   species from audio measurements and, in general, the literature approaches this problem in two ways;

72   the first approach generates a "dictionary" of features for each species through unsupervised learning

73   methods such as clustering and nearest neighbour classification. A new unknown species is then

74 allocated to the closest "word". Such methods have been shown to perform relatively poorly

75 (Potamitis et al., 2015, Moore and Miller, 2002, Potamitis, 2014). The second, more successful,

76 approach focusses on species classification through supervised learning algorithms that learn from

77 labelled data. These include artificial neural networks (Moore, 1991, Moore and Miller, 2002),

78 Gaussian mixture models (Potamitis, 2014, Ouyang et al., 2015), random forests, support vector

79 machines and gradient boosting classifiers (Potamitis et al., 2015). Deep learning approaches are

80 seeing an exponential increase in their usage but as highlighted in (Chen et al., 2014) and (Kiskin et al.,

81 2020), such methods also require formidable sample sizes. Conversely, convolutional neural networks

82 have been used to good effect in data scarce scenarios (Kiskin et al., 2020).

83 Not only do classification methods vary in the chosen algorithm but also in the choice of input data.

84 Chen *et al.* (2014) attribute the stagnation of insect classification in part to the overreliance on a single

85 feature of wingbeat frequency; often the fundamental frequency or the rate of wing flap. This was

86 observed more than 20 years previously (Moore, 1991) where error rates in classification of

87 mosquitoes increased by 33% when only using the wingbeat frequency, compared to using the entire

88 frequency spectra. Later, however, it was found that using the first 17 harmonics was as effective as

89 using the entire frequency spectra to classify five different aphid species (Moore and Miller, 2002).

90 Similarly, 12 features from the cepstrum have been used as input to the Gaussian mixture models for

91 classifying mosquitoes (Ouyang et al., 2015). This contrasts with (Potamitis et al., 2015) and references

92 therein who argue that the unprocessed spectra are a better choice than more sophisticated features

93 coming from individual harmonics. This is taken further in other studies (Chen et al., 2014) where the

94 frequency spectra are supplemented with additional covariates, such as time of flight and where prior

95 information on insect behaviour is available for inclusion in the Bayes classifier. A less common

96 approach is to move away from the frequency domain and to instead use wavelet transformations of

97 the audio (Kiskin et al., 2020), which arguably loses biological interpretability in relation to insect flight.

98   Literature shows little consensus on the convergence of a single approach but individual studies

99   highlight nuances in specific application areas. It is not known whether the performance of algorithms

100   and associated processing of data differs due to targeted species, experimental conditions, tuning

101   parameters, or most likely a combination of all three. However, the choice of both algorithm and data

102   processing should be made in the context of why discrimination is needed. In this paper, we aim to

103   couple together successful classification along with biological insight and as such, consider both

104   rigorous data processing to extract morphologically meaningful parameters and machine learning

105   algorithms to develop classification models. Furthermore, we aim to show that an ensemble of

106   bioacoustic parameters and indices can be used to distinguish between groups and species of the

107   agriculturally and economically important, but often overlooked, weak-flying insects.

108

109   **Materials and Methods**

110   *Flight experiments*

111   Opto-acoustic recorders capture the variation of light when an insect passes through a light beam.

112   Both the main body and the wings cast a shadow in the emitter's light beam, known as the extinction

113   of light principle, and this shadow is subsequently detected by a receiver photodiode array (Potamitis

114   and Rigakis, 2016a). The Wingbeat Recorder® (Insectronics, Chania, Crete, Greece) was set-up on a

115   work surface in a laboratory, with white dividers either side of the set-up preventing other equipment,

116   lasers and light to interfere with the recording. During experimental conditions, the sensor was placed

117   underneath a 15,000 ml heavy-walled glass beaker (Duran™) which provided sufficient space for

118   insects to behave as normal.

119   For each species studied, insects were either collected from the field or from insectary reared cultures

120   and placed within the up-turned beaker in which the sensor was enclosed (Figure S1). Over a period

121   of two days, insects were free to disperse in and around the sensor. Flights were automatically

122   triggered when an insect entered the field of view of the LED array, generating a recording lasting 0.6

123   seconds. All flights were saved as an audio file on an SD card within the sensor along with average

124   temperature and humidity covariates. Approximately 30 insects were used per experiment run,

125   generating on average 50 recordings (ranging from 0 to more than 500) recordings per run.

126

127   *Audio pre-processing and feature extraction*

128

129   A depiction of the audio processing steps is shown in Figure 1. Amplitude of the audio (volume) was

130   scaled according to the bit rate, *b*, (divided by 0.5×2^b) to be expressed in arbitrary amplitude units

131   (AAU). This allows a direct comparison between the amplitude across recordings sampled with

132   different bit rates. Audio recordings were trimmed to remove the "silence" at either end of the

133   recording. A threshold of 0.0061 AAU was identified at which sound can be considered background

134   noise. However, background noise exhibits stochasticity, and the first sound above this threshold does

135   not necessarily indicate it is an insect flight. Thus, to determine the first and last true sound and

136   therefore flight, any index for which the sound was above the threshold and identified as an outlier

137   (on the temporal scale), defined as more than 3 times the interquartile range (IQR) away from the

138   upper or lower quartiles, was considered stochastic variation above the threshold and not true sound

139   (Figure S2). Trimmed audio recordings consisting of fewer than 128 time points (a total of 0.01

140   seconds) were removed.

141   A filtering to remove background variation through a short-time Fourier transform with Hanning

142   window was applied to the trimmed audio recordings. Summary statistics (maximum amplitude,

143   amplitude range, amplitude interquartile range, see Table 1) and measures of energy (crest factor,

144   energy, power and root mean square (RMS), see Table 1) were obtained from the filtered trimmed

145   audio [Feature extraction box 1 in Figure 1]. Additional summary statistics (amplitude index and

146   temporal entropy, see Table 1) were obtained from zero padded trimmed audio [Feature extraction

147   box 2 in Figure 1].

148 The frequency spectrum was calculated with a window length of 128 time points using a short time

149 Fourier transform and applied to the trimmed audio, padded with zeroes to make a total recording of

150 length 8192 (2^{13}) time points. The dominant frequency was identified as the largest harmonic above

151 0.05 kHz. Harmonics were extracted in order of frequency peak height and the top 10, with no lower

152 frequency limit, were recorded. Spectral summaries (bioacoustic indices at four different frequency

153 ranges; 0-1000 Hz, 50-1000 Hz, 50-300 Hz, 200-3000 Hz, spectral entropy and the acoustic entropy,

154 see Table 1) were calculated [Feature extraction box 3 in Figure 1].

155 A long term trend, often consisting of a single peak and trough, was evident in the majority of

156 recordings. It is thought that this long term trend relates to insect flight movement (e.g. a banking

157 behaviour) rather than to wingbeat frequencies alone. As such, the frequency spectra extracted as

158 above, may have limited interpretability of the resulting harmonics which, in particular, prohibits the

159 estimation of the fundamental frequency. Thus, a second set of features are calculated after removal

160 of this long term trend through smoothing, estimated via a generalized additive model (GAM). The

161 GAM was fitting using thin plate regression splines and a maximum basis dimension of 1/50th of the

162 length of the trimmed audio or of dimension 10, whichever was bigger. Summary statistics (maximum

163 amplitude, amplitude range and amplitude interquartile range, see Table 1) and measures of energy

164 (crest factor, energy, power and root mean square (RMS), see Table 1) and also the maximum

165 amplitude of the estimated GAM were obtained from the filtered trimmed audio [Feature extraction

166 box 4 in Figure 1]. Additional summary statistics (amplitude index and temporal entropy) were

167 obtained from zero padded trimmed audio [Feature extraction box 5 in Figure 1]. Harmonic features

168 (dominant frequency, top 10 harmonic peaks, bioacoustic indices at four different frequency ranges;

169 0-1000 Hz, 50-1000 Hz, 50-300 Hz, 200-3000 Hz, spectral entropy and the acoustic entropy) were

170 extracted from the frequency spectrum calculated on the zero-padded, trend-removed, trimmed

171 audio [Feature extraction box 6 in Figure 1].

To calculate the fundamental frequency, peaks were identified in the modulus of the autocorrelation function of the detrended audio, ensuring peaks were no closer than 10 time points (0.001 seconds) apart. The fundamental frequency was then calculated as the inverse of the time of the first peak [Feature extraction box 7 in Figure 1].

Audio processing was done in the statistical software package, R, using packages seewave (Sueur et al., 2008a) for the Fourier transform, calculation of the harmonics and the calculation of temporal, spectral and acoustic entropies. The package soundecology (Villanueva-Rivera and Pijanowski, 2018) was used to calculate the bioacoustics index and the mgcv package (Wood, 2011) was used for the GAM estimation.

Data

A total of 5026 audio recordings were available. 98 of these did not exceed the required minimum audio length of 0.016 seconds, potentially because individuals did not fly through the whole sensor or were flying vertically through the sensor, and no features were extracted. Of the remaining 4928 observations, for which up to 52 features (as listed in Table 1) were calculated, 70% were randomly allocated to the training set and 30% to the validation set. Four aphid species (*Aphis fabae*, *Sitobion avenae*, *Myzus persicae* and *Rhopalosiphum padi*) were studied because they are global pests and reported on weekly on by the Rothamsted Insect Survey (RIS) to growers (https://insectsurvey.com/aphid-bulletin). *Drepanosiphum platanoidis* and *Periphyllus testudinaceus* are two additional aphid species that are included in our analyses and whilst neither is a crop pest, they are likely to be sampled by a sensor deployed in the field, particularly near sycamores and maples close to field margins. *Psylliodes chrysocephala* and *Brassicogethes aeneus* pose a serious threat to oilseed rape (*Brassica napus*) and other brassicas and are featured weekly in RIS' non-aphid commentary (https://insectsurvey.com/ris-remarks). Due to the small number of recordings for *M. persicae* and *R. padi*, these species were excluded from the random forest analysis, but are included

197   in the basic analyses of flight. The number of observations for each species within each dataset is given

198   in Table 2.

199   A total of 52 features were calculated for each audio recording. Seven different feature sets were

200   considered and are shown in Table S1. The first set considers all 52 features. The second considers the

201   25 features calculated without detrending, whilst the third feature set considers the 27 features

202   calculated on the detrended audio. Feature set 4 considers only the frequencies of the harmonic

203   peaks, calculated both before and after signal detrending. Feature set 5 extends set 4 to also include

204   the frequency indices such as the bioacoustics index, spectral entropy and dominant and fundamental

205   frequencies. Feature sets 6, 7 and 8 consist of the representative features from a hierarchical cluster

206   analysis with complete linkage on the correlation matrix of standardised features with 3, 5 and 14

207   clusters (Figure S3). Representative features were defined to be the feature closest to the cluster

208   centroid.

209

210   *Statistical analysis*

211   A linear model was fitted to each feature including covariates; humidity and temperature and an

212   explanatory variable indicative of species. A Type II ANOVA table was produced showing the effect of

213   dropping each term whilst retaining all others in the model. Where necessary, variables were

214   transformed to ensure homogeneity of variance as listed in Table 1.

215   Random forests (Breiman, 2001) were used to classify observations. Given the high levels of data

216   imbalance across species, balanced random forests were implemented. Balanced random forests

217   resample the data according to a set of defined class-specific sample sizes. Considerable tuning of

218   these class sample sizes is required and our criteria for tuning was to balance the class specific error

219   rates. The chosen set of sample sizes for *A. fabae, P. chrysocephala, S. avenae, P. testudinaceus, B.*

220   *aeneus, D. platanoidis* were 75, 75, 120, 50, 120, 300 for datasets including observations with missing

221   values and 60, 60, 96, 40, 96, 240 for datasets excluding observations with missing values, respectively.

222   Hyperparameters were tuned through an assessment of both the out-of-bag error rate and predictive

223   accuracy. Selected hyperparameters were to grow 1000 trees trying 10 randomly selected variables

224   at each split. Missing values were handled through the inbuilt option na.roughfix which imputes

225   missing values by the variable median. To compute the accuracy on validation data, missing values

226   were replaced by the median of each feature as computed from the training data.

227   Performance measures for classification include the true positive rate (TPR), true negative rate (TNR),

228   weighted accuracy (wAcc) and the class error (clErr) as defined in the supplementary information.

229   The importance of each feature was estimated as the mean decrease in accuracy associated with

230   dropping that variable from the model. This can be calculated for each class separately, the average

231   of which forms the mean decrease accuracy overall. The Gini index is the mean decrease in Gini score

232   associated with dropping the variable from the model, thus the Gini score provides a measure of how

233   well classes are separated.

234   Feature importance was investigated both for the full classification defined in Table 2, but also for

235   separate sub-classifications: Hemiptera (aphids) vs Coleoptera (beetles); within Hemiptera species;

236   within Coleoptera species separately. For the latter two classifications, new training and validation

237   datasets, satisfying the 70:30 split in each case, were defined.

238   Random forest models derived from different feature sets (Table S1) were compared by the out-of-

239   bag error estimates and the predictive accuracy, calculated as the average proportion of correctly

240   classified observations.

241   To simulate the process of identifying previously unidentified species, the random forest model was

242   calibrated on all data excluding all observations of a single nominate species. This excluded species

243   was then used as the validation dataset to form predictions. To investigate this process, the proportion

244 of allocations to each species classification was extracted along with the maximal class probability.

245 This process was repeated for each species in turn.

246 Random forests were fitted using the R package randomForest (Liaw and Wiener, 2002)

247

248 **Results**

249 The average flight duration was 0.17 seconds across all taxa, translating to a speed of 0.41 m/sec^{-1}.

250 Substantial variation in the flight duration was observed with interquartile range of 0.065 – 0.237

251 seconds and could be due to both the speed and direction of flight. Longer flights may involve spiralling

252 as well as turning behaviour. Flight duration differed between species ($F_{5, 4773}$ =19.52, p < 0.001), with

253 *M. persicae* (0.04 seconds; 1.75 m/sec^{-1}) and *P. chrysocephala* (0.076 seconds; 1.0 m/sec^{-1}) exhibiting

254 shorter than average flight durations and thus higher speeds.

255 For each recording, up to 52 features were extracted (Table 1) and all show significant differences

256 between species, on average (Table 3). Furthermore, with the exception of the amplitude index (of

257 the raw audio), all features showed a greater variability between species than with either of the

258 environmental covariates (largest F-statistic is associated with species differences, Table 2). However,

259 there is considerable variability within each feature reducing the chance that any one feature could in

260 isolation discriminate between species without inclusion of additional features (Table S2).

261

262 *Species classification*

263 Classification to species level has varying levels of success with random forest models. An overall out-

264 of-bag error rate of 20.62% (17.88%) on the training set and an error rate of 21.19% (17.75%) on the

265 validation set including (excluding) observations with missing values suggest reasonable success in

266 identifying individual species. A summary of the class specific error rates is given in the supplementary

267 information. Random forest classification is better viewed in the ensemble framework within which

268 it's derived. Figure 2 shows the distribution of the maximal class probability with an indication of

269 whether the maximal probability coincided with the true underlying species. For those species with

270 low misclassification rates (*D. platanoidis, B. aeneus, P. chrysocephala*), a direct correspondence is

271 seen with high maximal class probability (a median of 0.69, 0.62, 0.80, respectively). Furthermore, for

272 those observations of these species that are misclassified, the maximal class probability is lower (a

273 median of 0.42, 0.46, 0.40) indicating greater uncertainty in the final classification. Although the

274 certainty in the correct classification of *A. fabae* is lower (median of 0.51), there is still a pronounced

275 increase in uncertainty when the classification is wrong (median of 0.39). In contrast, the certainty of

276 classification for *S. avenae* and *P. testudinaceus* does not change depending upon whether the

277 classification is correct or not (a median of 0.50, 0.51 for correct classifications and a median of 0.48,

278 0.43 for incorrect classifications).

279

280 *Features for classification*

281 Figure 3 and Figure S5 illustrates the relative importance of the different features in the classification

282 model. The dominant frequency comes out top in both the accuracy (a measure of how well the

283 prediction improves) and the Gini index (a measure of how well class separation improves) when the

284 variable is included in the models. The fundamental frequency is a close second in terms of accuracy.

285 It is clear that the higher order harmonics contribute little in terms of feature importance. In contrast,

286 a number of summary indices of both the frequency and time domain are highlighted as important.

287 These include the spectral and acoustic entropy, Bioacoustic Index (3) over 50-300 Hz, the RMS and

288 power. Figure 3C) shows that the importance of these features differs by species with the acoustic

289 entropy important for identifying *P. chrysocephala* and the fundamental frequency important for *B.*

290 *aeneus* and the dominant frequency for *A. fabae*.

291

292   *Harmonics alone are not enough.*

293   A comparison of out-of-bag error rates shows that classification improves when using all features of

294   both the frequency and time domain compared to using specific subsets of feature variables (Figure

295   S6A). Specifically, the best out-of-bag accuracy rates on the validation set, where missing values are

296   imputed, are seen when using all 52 features (78.9%) and when using the 27 features extracted after

297   detrending (79.4%). A lower accuracy is seen when using the 25 features extracted before detrending

298   (76.8%) and when using only the 20 harmonic features (76.8%). Marginal improvements are seen

299   when supplementing the harmonic features with the additional frequency spectra indices (77.6%).

300   Although the minimal feature sets of 3 and 5 chosen features result in lower accuracy (60.0% and

301   68.5%), the minimal set of chosen 14 features performs relatively well (75.6%). It can be seen that

302   when imputation methods are used, the error rates tend to increase by about 2.5-4%.

303   Further investigation of the class specific error rates (Figure S6B), shows that the high error rates of

304   the minimal feature sets of 3 and 5 chosen features corresponds with poor prediction of pollen beetles

305   in particular. The predictive performance of black bean aphids increases in the feature sets restricted

306   to the harmonics only, 75.0% class specific error rate compared to 44.2% in the full 52 feature set.

307

308   *Features for within order classification differ to between order classification*

309   Features important in classifying between Hemiptera and Coleoptera align very closely with those

310   identified in the full model (Figure 4). However, when data are restricted to a single order, differing

311   patterns of feature importance are revealed. When focussed on aphid species only, the prominent

312   features are the maximum amplitude, the amplitude range and the power or RMS (accuracy decrease)

313   indicating a preference for features of the audio rather than the harmonics. The highly influential

314   features of the full classification reduce to a mid or low level of importance in the within order

315   classification. Similarly, focussing only on classifying between beetle species (albeit on a much smaller

316   dataset), the key features of importance identified are the spectral and acoustic entropy and to a

317   lesser extent, the bioacoustic index (at 50-300 Hz), the frequency of the most prominent harmonic

318   and the temporal entropy, thus indicating a preference for features of the frequency spectra. Thus,

319   harmonics such as the dominant frequency and fundamental frequency appear to be key in identifying

320   between orders, but alternative features of the audio and spectrum are required to identify to a

321   species level.

322

323   *Classifying unknown species results in less certain predictions*

324   In general, the class probability for a misclassified observation is lower than that for a correctly

325   classified observation (Figure 5A). Unobserved species are most commonly classified as sycamore

326   aphids or pollen beetles, likely due to the larger number of observations in these two classes (Figure

327   5B). Lower predictive certainty generally persists when investigating the model performance on

328   predictions of a previously unobserved class (Figure 5B). It is noticeable that when either cabbage

329   stem flea beetles or English grain aphids are excluded from the training set, the class predictions

330   remain relatively high resulting in reasonable certainty that these species are in fact sycamore aphids

331   or pollen beetles respectively. When sycamore aphids are excluded from the training set, they are

332   mostly allocated to the English grain aphid (62.3%) with some to the pollen beetle class (20.9%).

333

334   **Discussion**

335   There is now a wealth of studies having developed classification models of insect flight (Moore, 1991,

336   Chen et al., 2014, Potamitis, 2014, Potamitis et al., 2015, Ouyang et al., 2015, Kiskin et al., 2020), but

337   relatively few have focussed on weak-flying aphids and beetles (Moore and Miller, 2002). None that

338   we know of have attempted to link morphological characteristics to acoustic properties with the

339   exception of (Rajabi et al., 2016) who showed that the corrugated pattern of dragonfly wings explained

340 differences between damsel and dragonfly wingbeat frequencies. The aim of this study has been to

341 provide proof-of-concept for automatic detection methods of aphid and beetle pests via opto-acoustic

342 methods whilst also providing key insight into the drivers that will further this area of science. At first

343 look, the error rates in misclassification in this study appear high at 18-20% and yet, these insect pests

344 are not only inaudible but their wingbeat rate is eight times smaller than the typical model species

345 such as *Anopheles* mosquitoes (≈100 Hz v ≈800 Hz) and have a much weaker flight speed (0.41 m/sec[-]

346 [1] vs ≈ 1 m/sec[-1]) too (Potamitis et al., 2015). As sensors improve and more species observed, it seems

347 inevitable that the overall error rates will improve and yet, we also anticipate specific species

348 comparisons to remain a challenge. This is because of the close species similarity between aphid body

349 plans, their small size (2–5 mm; body mass 1-13 mg) and their simple wings that do not affect the

350 biomechanics of flight profoundly (*cf* dragonflies (Rajabi et al., 2016)). Successful classification to

351 species level should not be the final endpoint though. Rather the highly polyphagous nature of aphids

352 and the differential risk such forms pose make it desirable to classify beyond species. Indeed, (Hardie

353 and Powell, 2002) show substantial variation in flight behaviour though video tracking technology

354 between different forms of *A. fabae.* Although sensor technologies will undoubtedly improve, it is the

355 view of the authors that black-box classification of empirical data will always be limited in its scalability

356 if not coupled with knowledge of morphology and if deployed in-field, phenology.

357 At the highest taxonomic resolution, an ensemble of bioacoustics parameters and indices were used

358 to distinguish between beetle and aphid species. Our models indicate that within a small selection of

359 the Aphididae, measures of flight energy, particularly the maximum amplitude and the amplitude

360 range, are more important than higher order harmonics even though stroke amplitude varies during

361 flight (Tercel et al., 2018). Harmonics alone were shown to perform less well than conjectured by

362 (Moore and Miller, 2002) wherein they proved useful with neural networks. A lack of utility for

363 harmonics in our study is perhaps surprising given that wingbeat frequency and the harmonics are

364 functions of the physical size, shape, stiffness and mass of the wing as well as the wing muscles and

365 stroke amplitude (Tercel et al., 2018, Byrne et al., 1988). However, because of their small size, aphids

366 will likely incur greater relative drag and as a result of their small wings relative to body size will beat

367 their wings comparatively faster than other insects and this appears to be an important discriminator

368 (Tercel et al., 2018, Byrne et al., 1988). Such strengthening phenomena, discussed at length by

369 (Wootton, 1981) and shown to be the cause of variation in insect wing deformation between species,

370 are also observed in birds, where higher wing loadings demand a more substantial humerus (Sullivan

371 et al., 2019).

372 To further improve misclassification rates, a mechanistic understanding of wing acoustics is needed,

373 and whilst a detailed investigation is beyond the scope of this study, some general observations are

374 already profoundly clear; even with major differences in wing and flight apparatus, species are still

375 misclassified between major groups, reducing overall model precision and accuracy; higher order wing

376 beat harmonics do not play a major role in species discrimination, instead fundamental and dominant

377 frequencies as well as audio are more high ranking.

378 Specifically concerning the first point, the forewings of beetles are hardened to form the elytra, such

379 that the hindwing provides the energy and propulsion for flight and are not coupled to the elytra.

380 Instead, both the elytra and hind wing beat in phase during flight, although the former have a smaller

381 stroke angle (Brackenbury and Wang, 1995). Beetle wing venation is also modified to allow folding

382 under the elytra when not in flight. Indeed, for both species of beetle studied here, individuals have

383 poorly developed venation and therefore less stiffness (Kukalová-Peck and Lawrence, 1993, Kukalová-

384 Peck and Lawrence, 2004, Suzuki, 1994). It can be clearly seen how flexible the chyrosomelid beetle

385 *Crepidodera aurata's* wing is without a rigid structure along the complete length of the hind wing

386 (Nadein and Betz, 2016). Compare this flight apparatus with aphids that have coupled fore- and hind-

387 wings, no hardened wing casing and do not fold their wings on landing (Franielczyk-Pietyra and

388 Wegierek, 2017). Aphids, in contrast, have a thickened membrane on the forewing beyond the

389 anterior costal margin, the pterostigma, that increases wing flap performance due to a stiffer leading

390 edge that drives speed (Franielczyk-Pietyra and Wegierek, 2017). Yet, aphid wings still retain a flexible

391 wing membrane due to sparse venation and this flexibility provides greater lift than a completely stiff

392 wing (Mountcastle and Daniel, 2010). We therefore conjecture, ahead of any detailed study of wing

393 bioacoustics, that differences in wing venation and morphology must play a minor role in generating

394 unique wing harmonics for this group (or instead that if present, such differences cannot be detected

395 in the current sensor through the exclusion of light principle), and, our models support this, stressing

396 both the dominant and fundamental frequencies for splitting beetles and aphids, rather than higher

397 harmonics or more complex indices relating to energy or mechanics. It is also possible one source of

398 increased variability in higher order harmonics is the fact that wing movements may not be "clean",

399 for example, wings can touch each other or other parts of the body producing stridulations with

400 ultrasonics or other harmonics, which would modulate the production of wing harmonics.

401 Our study shows that predicted mean (median) wingbeat frequencies for aphids at average

402 temperature and humidity vary between species (Table S2; the exponentiated fundamental

403 frequencies for 6 aphid species: *A. fabae* = 134 Hz (119 Hz); *D. platanoidis* = 104 Hz (95 Hz); *M. persicae*

404 = 130 Hz (130 Hz); *P. testudinaceus* = 113 Hz (101 Hz); *R. padi* = 119 Hz (118 Hz); *S. avenae* = 106 Hz

405 (99 Hz)) but fall within an expected range for hemipterans (90-152 Hz) (Tercel et al., 2018). The

406 wingbeat frequencies of the cabbage stem flea beetle and pollen beetle are not known in the

407 literature, but our values are not remarkably different to other confamilial species recorded by (Tercel

408 et al., 2018) and (Brackenbury and Wang, 1995) (Table S2; the exponentiated fundamental

409 frequencies for *P. chrysocephala* = 121 Hz (119 Hz) *cf Oulema melanopus* 123 Hz Tercel et al. (2018)

410 and *Chalcoides aurata* 118 Hz (Brackenbury and Wang, 1995). *B. aeneus* = 139 Hz (136 Hz)). These

411 means are skewed somewhat when compared to the median, suggesting an increase in variation

412 according to temperature. Previous studies have shown temperature to be positively correlated with

413 the fundamental frequency, for example, female *Aedes aegypti* (L.) mosquitoes with a fundamental

414 wingbeat frequency of ≈ 450-550 Hz, increase their flap rate ≈ 8–13 Hz per unit change in degree

415 centigrade as the air becomes less dense, representing a rate of increase of 1.5-2.9% (Villarreal et al.,

416 2017). 'Frozen flight' is another source of variation that can impact fundamental frequency estimates

417   when the wingbeat is effectively zero (Thomas et al., 1977). Wing muscle autolysis in some of the

418   aphids studied (i.e. *A. fabae*, *M. persicae and R. padi*), during which flight muscle breakdown removes

419   the ability to fly, is yet another source of variation (Johnson, 1953, Leather et al., 1983). Wing muscle

420   autolysis was particularly notable with *R. padi* that were largely grounded once in the flight arena.

421   Despite these covariates, fundamental frequency remains key to discrimination, having a predictable

422   relationship with wing area and to a lesser extent, body mass (Byrne et al., 1988, Tercel et al., 2018).

423   In putting the emphasis of the data processing pipeline on feature extraction, we have been able to

424   link indices of both the temporal and frequency domain to morphological characteristics. Thus,

425   allowing us to gain understanding into the mechanisms contributing to differences in species flight

426   behaviour. Such insight is unavailable in the convolutional neural net approach of inputting frequency

427   spectra only. Furthermore, through our processing, we identified the importance of the flight

428   movement (estimated via smoothing splines). Although this has previously been identified (Potamitis

429   et al., 2015), to our knowledge it has not directly been incorporated into any classification algorithm.

430   We have shown that relatively simple summaries of the temporal domain, such as the power indicative

431   of the energy of a flight, contain important information for classification purposes, and perhaps

432   explains why convolutional neural networks on wavelet transforms perform well (Kiskin et al., 2020)

433   as wavelet transforms will account for both the time and frequency domain. This conceivably indicates

434   that it is not only the short-term flight behaviour such as wing-flaps that are important for species

435   identification but also the longer-term trends in an insect flight. Future studies will investigate this

436   further in the context of these weak-flying agricultural pests.

437   Study design is one of the most important factors in any data collection activity. As with any study,

438   there have been a number of limiting factors not least the high imbalance in observation numbers for

439   each individual species. This does not appear to be uncommon in the literature as (Moore and Miller,

440   2002) also had similarly imbalanced sample sizes ranging from 340 to 3325. Using balanced random

441   forest approaches to account for the data imbalance, class specific error rates can be reduced albeit

442 at the cost of higher out-of-bag error rates. Tuning these algorithms requires a trade-off between false

443 positive and false negative detections balanced across species. The optimal balance will depend on

444 individual study aims, for instance in-field monitoring of pests would require a minimisation of false

445 negative detections of key agricultural pests whilst for population monitoring it is preferable not to

446 bias false detections to any one species. In this study, we have opted for the latter approach and to

447 tune the algorithms aiming to balance class specific error rates. It remains the long-term aim to deploy

448 this technology in-field enabling automatic insect pest-detection at local spatial scales, however,

449 further work is needed in the collation of robust labelled data. Furthermore, we envisage algorithmic

450 development through the incorporation of prior knowledge, such as aphid migration patterns, as an

451 essential component to obtain good accuracy in-field.

452

References

454 Brackenbury, J. & Wang, R. (1995). Ballistics and visual targeting in flea-beetles (Alticinae). *Journal of Experimental Biology,* **198**, 1931-1942.

456 Breiman, L. (2001). Random forests. *Machine Learning,* **45**, 5-32.

457 Byrne, D. N., Buchmann, S. L. & Spangler, H. G. (1988). Relationship between wing loading, wingbeat frequency and body mass in homopterous insects. *Journal of Experimental Biology,* **135**, 9-23.

459 CABI (2020). *Myzus persicae*. Invasive Species Compendium. *Wallingford, UK: CAB International*.

460 Cancela, H., Higgins, A., Pagès-Bernaus, A. & Plà-Aragonès, L. M. (2019). Prologue–BigData and DSS in agriculture. *Computers and Electronics in Agriculture,* **161**, 1-322.

462 Cardim Ferreira Lima, M., Damascena de Almeida Leandro, M. E., Valero, C., Pereira Coronel, L. C. & Gonçalves Bazzo, C. O. (2020). Automatic Detection and Monitoring of Insect Pests—A Review. *Agriculture,* **10**, 161.

465 Chen, Y., Why, A., Batista, G., Mafra-Neto, A. & Keogh, E. (2014). Flying insect classification with inexpensive sensors. *Journal of Insect Behavior,* **27**, 657-677.

467 Franielczyk-Pietyra, B. & Wegierek, P. (2017). The forewing of the Aphis fabae (Scopoli

468   1763)(Hemiptera, Sternorrhyncha): a morphological and histological study. *Zoomorphology,*

469   **136**, 349-358.

470 Hardie, J. & Powell, G. (2002). Video analysis of aphid flight behaviour. *Computers and Electronics in*

471   *Agriculture,* **35**, 229-242.

472 Huuskonen, J. & Oksanen, T. (2018). Soil sampling with drones and augmented reality in precision

473   agriculture. *Computers and Electronics in Agriculture,* **154**, 25-35.

474 Johnson, B. (1953). Flight muscle autolysis and reproduction in aphids. *Nature,* **172**, 813-813.

475 Kiskin, I., Zilli, D., Li, Y., Sinka, M., Willis, K. & Roberts, S. (2020). Bioacoustic detection with wavelet-

476   conditioned convolutional neural networks. *Neural Computing and Applications,* **32**, 915-927.

477 Kukalová-Peck, J. & Lawrence, J. F. (1993). Evolution of the hind wing in Coleoptera. *The Canadian*

478   *Entomologist,* **125**, 181-258.

479 Kukalová-Peck, J. & Lawrence, J. F. (2004). Relationships among coleopteran suborders and major

480   endoneopteran lineages: evidence from hind wing characters. *European Journal of*

481   *Entomology,* **101**, 95-144.

482 Leather, S., Wellings, P. W. & Dixon, A. (1983). Habitat quality and the reproductive strategies of the

483   migratory morphs of the bird cherry-oat aphid, Rhopalosiphum padi (L.), colonizing secondary

484   host plants. *Oecologia*, 302-306.

485 Liaw, A. & Wiener, M. (2002). Classification and regression by randomForest. *R News,* **2**, 18-22.

486 Lowe, A. & Dromgoole, W. (1958). The development of an electronic aphid-counter: I. The problem of

487   counting aphids. *New Zealand Journal of Agricultural Research,* **1**, 903-912.

488 Milne, A., Webster, R., Ginsburg, D. & Kindred, D. (2012). Spatial multivariate classification of an arable

489   field into compact management zones based on past crop yields. *Computers and Electronics*

490   *in Agriculture,* **80**, 17-30.

491 Montealegre-Z, F., Jonsson, T. & Robert, D. (2011). Sound radiation and wing mechanics in stridulating

492   field crickets (Orthoptera: Gryllidae). *Journal of Experimental Biology,* **214**, 2105-2117.

493   Moore, A. (1991). Artificial neural network trained to identify mosquitoes in flight. *Journal of insect*

494   *behavior,* **4**, 391-396.

495   Moore, A. & Miller, R. H. (2002). Automated identification of optically sensed aphid (Homoptera:

496   Aphidae) wingbeat waveforms. *Annals of the Entomological Society of America,* **95**, 1-8.

497   Mountcastle, A. M. & Daniel, T. L. (2010). Aerodynamic and functional consequences of wing

498   compliance. Animal Locomotion. Springer.

499   Mukundarajan, H., Hol, F. J. H., Castillo, E. A., Newby, C. & Prakash, M. (2017). Using mobile phones as

500   acoustic sensors for high-throughput mosquito surveillance. *Elife,* **6**, e27854.

501   Nadein, K. & Betz, O. (2016). Jumping mechanisms and performance in beetles. I. Flea beetles

502   (Coleoptera: Chrysomelidae: Alticini). *Journal of Experimental Biology,* **219**, 2015-2027.

503   Ouyang, T.-H., Yang, E.-C., Jiang, J.-A. & Lin, T.-T. (2015). Mosquito vector monitoring system based on

504   optical wingbeat classification. *Computers and Electronics in Agriculture,* **118**, 47-55.

505   Potamitis, I. (2014). Classifying insects on the fly. *Ecological Informatics,* **21**, 40-49.

506   Potamitis, I. & Rigakis, I. (2016a). Large aperture optoelectronic devices to record and time-stamp

507   insects' wingbeats. *IEEE Sensors Journal,* **16**, 6053-6061.

508   Potamitis, I. & Rigakis, I. (2016b). Measuring the fundamental frequency and the harmonic properties

509   of the wingbeat of a large number of mosquitoes in flight using 2D optoacoustic sensors.

510   *Applied Acoustics,* **109**, 54-60.

511   Potamitis, I., Rigakis, I. & Fysarakis, K. (2015). Insect biometrics: Optoacoustic signal processing and its

512   applications to remote monitoring of McPhail type traps. *PloS One,* **10**, e0140474.

513   Rajabi, H., Rezasefat, M., Darvizeh, A., Dirks, J.-H., Eshghi, S., Shafiei, A., Mostofi, T. M. & Gorb, S. N.

514   (2016). A comparative study of the effects of constructional elements on the mechanical

515   behaviour of dragonfly wings. *Applied Physics A,* **122**, 19.

516   Sanborn, A. F. & Phillips, P. K. (1995). Scaling of sound pressure level and body size in cicadas

517   (Homoptera: Cicadidae; Tibicinidae). *Annals of the Entomological Society of America,* **88**, 479-

518   484.

519   Smith, S. (1999). The Scientist & Engineer's Guide to Digital Signal Processing, *San Diego, CA*, California

520   Technical Publishing.

521   Sueur, J., Aubin, T. & Simonis, C. (2008a). Seewave, a free modular tool for sound analysis and

522   synthesis. *Bioacoustics,* **18**, 213-226.

523   Sueur, J., Pavoine, S., Hamerlynck, O. & Duvail, S. (2008b). Rapid acoustic survey for biodiversity

524   appraisal. *PloS One,* **3**, e4065.

525   Sullivan, T., Meyers, M. A. & Arzt, E. (2019). Scaling of bird wings and feathers for efficient flight.

526   *Science Advances,* **5**, eaat4269.

527   Suzuki, K. (1994). Comparative morphology of the hindwing venation of the Chrysomelidae

528   (Coleoptera). *In:* JOLIVET, P. H., COX, M. L. & PETITPIERRE, E. (eds.) Novel aspects of the biology

529   of Chrysomelidae*. Dordrecht*: Springer Netherlands.

530   Tercel, M. P., Veronesi, F. & Pope, T. W. (2018). Phylogenetic clustering of wingbeat frequency and

531   flight-associated morphometrics across insect orders. *Physiological Entomology,* **43**, 149-157.

532   Tetila, E. C., Machado, B. B., Astolfi, G., Belete, N. A. d. S., Amorim, W. P., Roel, A. R. & Pistori, H. (2020).

533   Detection and classification of soybean pests using deep learning with UAV images. *Computers*

534   *and Electronics in Agriculture,* **179**.

535   Thomas, A., Ludlow, A. & Kennedy, J. (1977). Sinking speeds of falling and flying Aphis fabae Scopoli.

536   *Ecological Entomology,* **2**, 315-326.

537   Villanueva-Rivera, L. J. & Pijanowski, B. C. (2018). Package 'soundecology'.

538   Villarreal, S. M., Winokur, O. & Harrington, L. (2017). The impact of temperature and body size on

539   fundamental flight tone variation in the mosquito vector Aedes aegypti (Diptera: Culicidae):

540   implications for acoustic lures. *Journal of Medical Entomology,* **54**, 1116-1121.

541   Willis, K. J. & *ed.* (2017). State of the World's Plants 2017*, Royal Botanic Gardens, Kew*.

542

543   Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of

544   semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B*

545   *(Statistical Methodology),* **73**, 3-36.

546   Wootton, R. J. (1981). Support and deformability in insect wings. *Journal of Zoology,* **193**, 447-468.

547   Zhai, Z., Martínez, J. F., Beltran, V. & Martínez, N. L. (2020). Decision support systems for agriculture

548   4.0: Survey and challenges. *Computers and Electronics in Agriculture,* **170**, 105256.

549

550

551

Abbreviations

553 Decision Support System (DSS)

554 Light-emitting diode (LED)

555 secure digital (SD)

556 arbitrary amplitude units (AAU)

557 analysis of variance (ANOVA)

558 generalized additive model (GAM)

559 interquartile range (IQR)

560 root mean square (RMS)

561 true positive rate (TPR)

562 true negative rate (TNR)

563 weighted accuracy (wAcc)

564 class error (clErr)

565

Declarations

567 Availability of data and materials: The datasets generated and/or analysed during the current study

568 are publicly available with a DOI on the Rothamsted Research Repository

569 (https://repository.rothamsted.ac.uk/) under a Creative Commons Attribution 4.0 Licence. Raw audio

570 recordings are available at https://doi.org/10.23637/rothamsted.981w7 and extracted features at

571 https://doi.org/10.23637/rothamsted.981w8

572   All code used to analyse the data are available as supplementary materials (see below).

573   Competing interests: The authors declare that they have no competing interests

579   Authors' contributions: KLH developed the pre-processing pipeline, analysed the data, devised the

580   statistical approach and wrote the manuscript. AD collected the insect samples and ran the

581   experiments. IP supplied the entomatic sensor and provided technical advice on the resulting data.

582   JRB managed the project, secured the funding, and wrote the biological components of the

583   manuscript. All authors read and approved the final manuscript.

585

586   **Figure Legends**

587   **Figure 1** Data processing pipeline, indicating where features are extracted. Aphid illustration released

588   under the Creative Commons Licence https://commons.wikimedia.org/wiki/File:Aphid_(PSF).png

589   **Figure 2** Species classification error rates. Box and whisker plots of the maximum class probability

590   predicted for each observation in the validation dataset and split according to whether the

591   classification was correct or not

592   **Figure 3** Species classification. a) and b) give the mean decrease in accuracy and Gini index respectively

593   for the top 10 feature variables considered in the model (a complete set is shown in Figure S5) and c)

594   presents the within species importance of each feature variable. Feature variables denoted by [g] are

595   derived after a detrending step (see Figure 1)

596   **Figure 4** Comparison of class sets. The mean decrease in accuracy due to dropping each feature in turn

597   from a model (using the complete feature set) classifying between all species, between aphids vs

598   beetles, between aphid species and between beetle species

599   **Figure 5** Identifying an unknown species. a) Boxplots of the maximum class probability of the

600   validation set for the full species model, separating between observations correctly and incorrectly

601   classified. b) Confusion matrix of each single species exclusion model showing the proportion of

602   allocations to each of the known species. Boxplots of the maximum class probability are shown for

603   each associated model

604

605   **Additional Files**

606   Additional file 1: Supplementary Information.pdf

607   Supplementary information provides further details on the selection of features and the performance

608   of the classification algorithms. Also included in this file are:

609   Table S1, detailing the chosen features for each random forest model

610   Table S2, containing the summary statistics of each extracted feature calculated per species.

611

612   Additional file 2: Figure S1.pdf

613   Figure S1: Photograph of the experimental setup where the opto-acoustic sensor is contained within

614   a large jar allowing aphids to fly freely through the sensor. Photographs of illustrative wings from *R.*

615   *padi*, *D. platanoidis*, *B. aenus* and *P. chrysocephala*. Each tick mark on the scale bar is 0.1mm.

616

617   Additional file 3: Figure S2.pdf

618   Figure S2: Data processing. Figure illustrates how each recording is trimmed to remove periods of

619   silence at the start or end of a recording. Lower panel is the audio recording, red lines are the threshold

620   of ± 0.0061 arbitrary amplitude units, above and below which sound is considered silence. The upper

621   panel shows a box plot of the temporal indices exceeding this threshold. Audio is then trimmed to the

622   whiskers of the boxplot defined as the largest (smallest) temporal index not exceeding 3 times the

623   interquartile range away from the upper (lower) quartile, shown by the blue lines.

624

625   Additional file 4: Figure S3.pdf

626   Figure S3: Identification of a minimal feature set. A) The correlation matrix between features

627   calculated after standardisation. B) Dendrogram of a hierarchical cluster analysis using complete

628   linkage on $1 - r$, where r is the correlation matrix of the standardised feature set. Features are coloured

629   according to cutting the dendrogram into i) 3 groups, ii) 5 groups iii) 14 groups. Features deemed most

630   representative of each group are indicated by the box.

631

632   Additional file 5: Figure S4.pdf

633   Figure S4. Species classification error rates. A) shows the class specific error rates where clErr is the

634   class error rate (or 1 – true positive rate per class), TNR is the class specific true negative rate, TPR is

635   the class specific true positive rate and wAcc is the weighted accuracy (wAcc=0.5 x TNR + 0.5 x TPR).

636   B) shows the confusion matrices of classification predictions on the validation dataset, presented as a

637   proportion per species.

638

639   Additional file 6: Figure S5.pdf

640   Figure S5. Species classification. A) and B) give the mean decrease in accuracy and Gini index

641   respectively for each feature variable considered in the model and C) presents the within species

642   importance of each feature variable. Feature variables denoted by [g] are derived after a detrending

643   step (see Figure 1)

644

645   Additional file 7: Figure S6.pdf

646   Figure S6. Comparison of feature sets. A) The mean accuracy rate on the out-of-bag predictions from

647   the training set and on the validation set for both omitting and imputing observations with missing

648   values. B) the class specific error rates for each feature set for both omitting and imputing

649   observations with missing values. Each model corresponds to a different subset of feature variables

650   as detailed in Table 3.

651

652   Additional file 8: code.zip

653   Folder containing R scripts for running the random forest models and producing the presented

654   analyses. Files included are:

655   • randomForest_tuning.R: script containing the tuning process of a) the balanced

656   random forest and b) the tuning parameters of the random forest algorithm.

657   • randomForest_speciesModel.R: script containing the random forest model for all

658   species and associated analytics.

659   • randomForest_dataSlice.R: script containing the random forest models for different

660   data subsets (features and species).

661

662 Table 1. A summary of the features extracted from the audio recordings along with their definition. Let a(t) denote the amplitude of the trimmed audio at time t and a*(t)

663 denote the Hilbert amplitude envelope at time t (22). Let f(ξ) denote the spectrum of a(t) at frequency ξ Hz. Let $\bar{f}(\xi)$ denote the mean spectrum. Let g(t) denote the GAM

664 estimate of the long term trend and $\tilde{a}$ and $\tilde{f}$ be the amplitude and frequency of the GAM adjusted audio. Transformations are applied to the formulae listed. All log

665 transformations are the natural logarithm. [g] indicates feature has been calculated after removal of the long-term trend. Indicated in the table is the stage at which the

666 feature was extracted, corresponding to the labelled boxes in Figure 1. The final column provides a description of the biological interpretation of each feature.
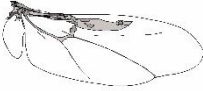
Extraction Stage	Feature	Transformation	Feature Type	Mathematical Description	Description	Biological category
1	Maximum amplitude	logarithm	audio	$\max_t \lvert a(t) \rvert$	Maximum amplitude of a half cycle wingbeat centred at 0.	Flight energy
1	Amplitude range	logarithm	audio	$\max_t a(t) - \min_t a(t)$	The difference between the maximum peak and the minimum trough of a wingbeat recording.	Flight energy
1	Amplitude IQR	logarithm	audio	$q_3\big(a(t)\big) - q_1\big(a(t)\big)$, where q_1 and q_3 denote the first and third quartile.	The difference between recorded amplitude at the 3rd and 1st quartiles. A dampened measure of flight energy ignoring extreme values.	Flight energy
1	Power	logarithm	audio	$\frac{1}{n}\sum_t [a(t)]^2$, where n is the number of time points in the recording.	The average squared amplitude.	Flight energy
1	Root Mean Square (RMS)	logarithm	audio	$\sqrt{\frac{1}{n}\sum_t [a(t)]^2}$	The average squared amplitude, downweighted by a square root transformation, dampening influence of loud sounds.	Flight energy
1	Crest factor	logarithm	audio	$\max_t \lvert a(t) \rvert / \sqrt{\frac{1}{n}\sum_t [a(t)]^2}$	The ratio between the peak value relative to the RMS of the wingbeat cycle in the series.	Speed of wing transition relative to overall flight energy
2	Amplitude index	logarithm	audio	$q_2\big(a^*(t)\big)$, where q_2 denotes the median.	The median value of the amplitude over time.	Flight energy
2	Temporal entropy		audio	$\frac{-1}{\log n}\sum_t a^*(t)\log a^*(t)$	A function of Shannon evenness, the index estimates the variability in amplitude (loudness) over time.	Consistency of flight energy
3	Bioacoustics index (1)		frequency	$\sum_{\xi=0}^{1000} \bar{f}(\xi)$	The area under the curve of the frequency spectra between 0-1000 Hz.	Harmonic information including body oscillations

3					This may include lower frequency body oscillations of the insect between 0-50 Hz.	
3	Bioacoustics index (2)		frequency	$\sum_{\xi=50}^{1000} \bar{f}(\xi)$	The area under the curve of the frequency spectra between 0-1000 Hz. This will exclude lower frequency body oscillations of the insect between 0-50 Hz.	Harmonic information excluding body oscillations
3	Bioacoustics index (3)		frequency	$\sum_{\xi=50}^{300} \bar{f}(\xi)$	The area under the curve of the frequency spectra between 50-300 Hz. This will only include low order frequencies including the fundamental frequency.	Harmonic information at low frequencies
3	Bioacoustics index (4)		frequency	$\sum_{\xi=200}^{3000} \bar{f}(\xi)$	The area under the curve of the frequency spectra between 200-3000 Hz. This will only include higher order frequencies.	Harmonic information at high frequencies
3	Spectral entropy		frequency	$\frac{-1}{\log n}\sum_{\xi} \bar{f}(\xi) \log \bar{f}(\xi)$	A function of Shannon evenness, the index estimates the variability in the frequency spectrum.	Smoothness of the harmonic series
3	Acoustic entropy		audio-frequency	$\frac{-1}{\log n}\sum_{t} a^*(t) \log a^*(t) \times \frac{-1}{\log n}\sum_{\xi} \bar{f}(\xi) \log \bar{f}(\xi)$	An index lying between 0 and 1, with 0 indicating a pure tone and 1 indicating random noise. (Sueur et al., 2008b).	Clarity of the recorded sound
3	Dominant frequency	square root	frequency	$f^{-1}\left(\max_{\xi>50} f(\xi)\right)$	The "loudest" frequency, i.e. the frequency corresponding to the largest changes in amplitude such that the frequency exceeds body oscillations of at least 50 Hz.	(Inverse) Length of the largest oscillations
3	1st harmonic		harmonics	$h_1 = f^{-1}\left(\max_{\xi} f(\xi)\right)$	The "loudest" frequency, i.e. the frequency corresponding to the largest changes in amplitude. This will coincide with the dominant frequency where the largest frequency is over 50 Hz.	(Inverse) Length of the largest oscillations
3	2nd harmonic		harmonics	$h_2 = f^{-1}\left(\max_{\xi,\xi\neq h_1} f(\xi)\right)$	The second to tenth loudest frequency. Higher harmonics may be multiples of lower harmonics, where a repeating oscillation is detected.	(Inverse) Length of the higher order oscillations.
3	3rd harmonic		harmonics	$h_3 = f^{-1}\left(\max_{\xi,\xi\neq h_1,h_2} f(\xi)\right)$		
3	4th harmonic		harmonics	$h_4 = f^{-1}\left(\max_{\xi,\xi\neq h_1,h_2,h_3} f(\xi)\right)$		
3	5th harmonic		harmonics	$h_5 = f^{-1}\left(\max_{\xi,\xi\neq h_1\dots h_4} f(\xi)\right)$		

3	6th harmonic		harmonics	$h_6 = f^{-1}\left(\max_{\xi,\xi \neq h_1...h_5} f(\xi)\right)$		
3	7th harmonic		harmonics	$h_7 = f^{-1}\left(\max_{\xi,\xi \neq h_1...h_6} f(\xi)\right)$		
3	8th harmonic		harmonics	$h_8 = f^{-1}\left(\max_{\xi,\xi \neq h_1...h_7} f(\xi)\right)$		
3	9th harmonic		harmonics	$h_9 = f^{-1}\left(\max_{\xi,\xi \neq h_1...h_8} f(\xi)\right)$		
3	10th harmonic		harmonics	$h_{10} = f^{-1}\left(\max_{\xi,\xi \neq h_1...h_9} f(\xi)\right)$		
4	GAM amplitude range	logarithm	audio	$\max_t g(t) - \min_t g(t)$	The difference between the peak and the trough of a GAM smoothed signal.	Flight behaviour
4	Maximum amplitude [g]	logarithm	audio	$\max_t |a(t)|$	As above but applied to the detrended audio signal	As above but applied to the detrended audio signal
4	Amplitude range [g]	logarithm	audio	$\max_t a(t) - \min_t a(t)$		
4	Amplitude IQR [g]	logarithm	audio	$q_3\big(a(t)\big) - q_1\big(a(t)\big)$, where q_1 and q_3 denote the first and third quartile.		
4	Power [g]	logarithm	audio	$\frac{1}{n}\sum_t [a(t)]^2$, where n is the number of time points in the recording.		
4	RMS [g]	logarithm	audio	$\sqrt{\frac{1}{n}\sum_t [a(t)]^2}$		
4	Crest factor [g]	logarithm	audio	$\max_t |a(t)| / \sqrt{\frac{1}{n}\sum_t [a(t)]^2}$		
5	Amplitude index [g]	logarithm	audio	$q_2\big(\tilde{a}^*(t)\big)$		
5	Temporal entropy [g]		audio	$\frac{-1}{\log n}\sum_t \tilde{a}^*(t) \log \tilde{a}^*(t)$		
6	Bioacoustics index (1) [g]		frequency	$\sum_{\xi=0}^{1000} \bar{\tilde{f}}(\xi)$		
6	Bioacoustics index (2) [g]		frequency	$\sum_{\xi=50}^{1000} \bar{\tilde{f}}(\xi)$		
6	Bioacoustics index (3) [g]		frequency	$\sum_{\xi=50}^{300} \bar{\tilde{f}}(\xi)$		
6	Bioacoustics index (4) [g]		frequency	$\sum_{\xi=200}^{3000} \bar{\tilde{f}}(\xi)$		
6	Spectral entropy [g]		frequency	$\frac{-1}{\log n}\sum_\xi \bar{\tilde{f}}(\xi) \log \bar{\tilde{f}}(\xi)$		
6	Acoustic entropy [g]		audio-frequency	$\frac{-1}{\log n}\sum_t \tilde{a}^*(t) \log \tilde{a}^*(t) \times \frac{-1}{\log n}\sum_\xi \bar{\tilde{f}}(\xi) \log \bar{\tilde{f}}(\xi)$		
6	Dominant frequency [g]	square root	frequency	$\tilde{f}^{-1}\left(\max_{\xi>50} \tilde{f}(\xi)\right)$		

6	1st harmonic [g]		harmonics	$\tilde{h}_1 = \tilde{f}^{-1}\left(\max_{\xi} \tilde{f}(\xi)\right)$		
6	2nd harmonic [g]		harmonics	$\tilde{h}_2 = \tilde{f}^{-1}\left(\max_{\xi,\xi\neq\tilde{h}_1} \tilde{f}(\xi)\right)$		
6	3rd harmonic [g]		harmonics	$\tilde{h}_3 = \tilde{f}^{-1}\left(\max_{\xi,\xi\neq\tilde{h}_1,\tilde{h}_2} \tilde{f}(\xi)\right)$		
6	4th harmonic [g]		harmonics	$\tilde{h}_4 = \tilde{f}^{-1}\left(\max_{\xi,\xi\neq\tilde{h}_1,\tilde{h}_2,\tilde{h}_3} \tilde{f}(\xi)\right)$		
6	5th harmonic [g]		harmonics	$\tilde{h}_5 = \tilde{f}^{-1}\left(\max_{\xi,\xi\neq\tilde{h}_1\ldots\tilde{h}_4} \tilde{f}(\xi)\right)$		
6	6th harmonic [g]		harmonics	$\tilde{h}_6 = \tilde{f}^{-1}\left(\max_{\xi,\xi\neq\tilde{h}_1\ldots\tilde{h}_5} \tilde{f}(\xi)\right)$		
6	7th harmonic [g]		harmonics	$\tilde{h}_7 = \tilde{f}^{-1}\left(\max_{\xi,\xi\neq\tilde{h}_1\ldots\tilde{h}_6} \tilde{f}(\xi)\right)$		
6	8th harmonic [g]		harmonics	$\tilde{h}_8 = \tilde{f}^{-1}\left(\max_{\xi,\xi\neq\tilde{h}_1\ldots\tilde{h}_7} \tilde{f}(\xi)\right)$		
6	9th harmonic [g]		harmonics	$\tilde{h}_9 = \tilde{f}^{-1}\left(\max_{\xi,\xi\neq\tilde{h}_1\ldots\tilde{h}_8} \tilde{f}(\xi)\right)$		
6	10th harmonic [g]		harmonics	$\tilde{h}_{10} = \tilde{f}^{-1}\left(\max_{\xi,\xi\neq\tilde{h}_1\ldots\tilde{h}_9} \tilde{f}(\xi)\right)$		
7	Fundamental frequency [g]	logarithm	harmonics	$1/T$, where T is the period of the autocorrelation function.	A repeating and consistent frequency that relates to the wingbeat frequency (i.e. number of wingflaps per sec)	Wingbeat frequency

667

668   Table 2 Number of audio recordings with feature information listed by species in the complete,

669   training and validation datasets. Illustration of wing venation for each species.

Insect order	Species		Number of observations	Training	Validation	Wing
	Common	Latin				
Hemiptera	Sycamore aphid	*Drepanosiphum platanoidis*	3323	2351	972	
Hemiptera	English grain aphid	*Sitobion avenae*	274	193	81	
Hemiptera	Maple aphid	*Periphyllus testudinaceus*	113	76	37	
Hemiptera	Black bean aphid	*Aphis fabae*	161	120	41	
Hemiptera	Peach-potato aphid	*Myzus persicae*	15			
Hemiptera	Bird cherry-oat aphid	*Rhopalosiphum padi*	8			
Coleoptera	Pollen beetle	*Brassicogethes aeneus*	848	566	282	
Coleoptera	Cabbage stem flea beetle	*Psylliodes chrysocephala*	186	127	59	
	Total		4928	3433	1472	

670

671

Table 3 ANOVA results testing for differences between species for each individual feature. To adjust

for confounding, the covariates humidity and temperature were included in the model. Significant

terms are highlighted in green. F statistics reported are of type II. Features are ordered by the size of

the F statistic associated with dropping species from the model (from largest to smallest). Where

necessary, variables were transformed as listed in Table 1.

Feature	F statistic			P- value		
	Humidity	Temperature	Species	Humidity	Temperature	Species
Maximum amplitude [g]	7.85	4.69	187.17	5.11E-03	3.04E-02	4.55E-242
Bioacoustics index (3) [g]	1.17	1.08	177.00	2.80E-01	2.99E-01	2.95E-230
Spectral entropy [g]	76.92	8.86	161.96	2.55E-18	2.93E-03	1.62E-212
Power [g]	16.16	8.34	138.45	5.93E-05	3.90E-03	4.49E-184
RMS [g]	16.16	8.34	138.45	5.93E-05	3.90E-03	4.49E-184
Amplitude range [g]	10.57	12.60	135.16	1.16E-03	3.89E-04	4.98E-180
Bioacoustics index (1)	12.06	11.07	108.73	5.20E-04	8.84E-04	2.62E-148
Bioacoustics index (2)	8.31	13.05	107.72	3.95E-03	3.07E-04	5.48E-147
GAM amplitude range [g]	2.60	3.32	99.90	1.07E-01	6.87E-02	1.72E-135
Fundamental frequency [g]	0.29	0.07	85.55	5.93E-01	7.89E-01	1.97E-116
Power	4.27	3.38	80.75	3.89E-02	6.61E-02	3.11E-111
RMS	4.27	3.38	80.75	3.89E-02	6.61E-02	3.11E-111
Bioacoustics index (4)	17.53	1.48	77.48	2.88E-05	2.23E-01	7.97E-107
Maximum amplitude	2.42	5.44	76.99	1.20E-01	1.98E-02	3.64E-106
Amplitude range	2.23	5.83	72.70	1.35E-01	1.58E-02	2.44E-100
Dominant frequency [g]	0.05	0.22	71.86	8.15E-01	6.39E-01	1.61E-98
Bioacoustics index (4) [g]	21.64	0.85	69.51	3.40E-06	3.57E-01	2.36E-95
Spectral entropy	18.51	1.30	68.59	1.72E-05	2.54E-01	9.72E-95
Acoustic entropy [g]	0.68	0.05	66.45	4.09E-01	8.17E-01	3.19E-91
Dominant frequency	0.01	0.16	58.95	9.43E-01	6.93E-01	1.85E-81
Temporal entropy [g]	0.01	0.00	57.53	9.16E-01	9.68E-01	4.43E-79
Bioacoustics index (1) [g]	17.79	11.94	48.82	2.53E-05	5.55E-04	4.36E-67
1st harmonic [g]	0.15	0.01	46.58	6.96E-01	9.36E-01	5.60E-64
Bioacoustics index (2) [g]	13.96	13.43	44.70	1.89E-04	2.51E-04	2.34E-61
9th harmonic	2.19	5.68	43.14	1.39E-01	1.72E-02	1.93E-59
6th harmonic	3.39	2.22	42.04	6.55E-02	1.36E-01	6.77E-58
5th harmonic	2.65	3.75	40.60	1.03E-01	5.29E-02	7.25E-56
Acoustic entropy	3.07	0.49	37.73	7.98E-02	4.85E-01	8.21E-52
7th harmonic	2.53	3.28	36.92	1.11E-01	7.01E-02	1.14E-50
8th harmonic	9.80	2.18	35.30	1.76E-03	1.40E-01	2.23E-48
Amplitude IQR [g]	21.25	4.38	32.40	4.15E-06	3.65E-02	3.99E-44
Temporal entropy	5.46	0.36	32.30	1.95E-02	5.48E-01	4.03E-44
10th harmonic	2.34	0.41	31.77	1.26E-01	5.21E-01	2.26E-43
4th harmonic	2.51	2.50	31.19	1.13E-01	1.14E-01	1.48E-42

Bioacoustics index (3)	0.35	4.15	31.08	5.55E-01	4.16E-02	2.14E-42
3rd harmonic	1.05	0.91	30.59	3.07E-01	3.41E-01	1.06E-41
Amplitude IQR	7.01	0.47	29.96	8.13E-03	4.95E-01	8.35E-41
2nd harmonic	0.00	0.30	28.42	9.68E-01	5.87E-01	1.30E-38
10th harmonic [g]	0.70	2.44	19.82	4.03E-01	1.18E-01	2.64E-26
8th harmonic [g]	1.56	1.31	18.88	2.12E-01	2.52E-01	5.63E-25
Crest factor [g]	0.53	4.51	18.35	4.65E-01	3.38E-02	3.20E-24
4th harmonic [g]	0.05	0.15	18.15	8.26E-01	6.99E-01	6.29E-24
3rd harmonic [g]	0.94	0.04	15.59	3.33E-01	8.38E-01	2.65E-20
9th harmonic [g]	6.11	4.39	15.22	1.35E-02	3.62E-02	8.76E-20
2nd harmonic [g]	0.80	0.23	14.70	3.72E-01	6.31E-01	4.80E-19
7th harmonic [g]	2.24	4.66	14.01	1.34E-01	3.10E-02	4.54E-18
Crest factor	1.54	1.92	13.06	2.14E-01	1.66E-01	9.20E-17
5th harmonic [g]	0.82	1.90	12.71	3.66E-01	1.68E-01	3.08E-16
6th harmonic [g]	2.23	0.80	12.44	1.36E-01	3.72E-01	7.40E-16
Amplitude index	11.28	0.04	10.86	7.89E-04	8.50E-01	1.17E-13
1st harmonic	0.10	5.97	8.58	7.54E-01	1.46E-02	1.70E-10
Amplitude index [g]	0.12	0.02	4.96	7.25E-01	8.98E-01	1.34E-05

677

678