Ability of near-infrared spectroscopy and chemometrics to predict the age of mosquitoes reared under different conditions

Oselyne Ong (oselyne.ong@qimrberghofer.edu.au)
QIMR Berghofer Medical Research Institute
Elise Kho
University of Queensland
Pedro Esperança
Imperial College London
Chris Freebairn
Private Contracting Entomologist
Floyd Dowell
US Department of Agriculture
Gregor Devine
QIMR Berghofer Medical Research Institute
Thomas Churcher
Imperial College London

Research

Keywords: Asian Tiger Mosquito, Age, Spectroscopy, Chemometrics, Near-Infrared

Posted Date: December 3rd, 2019

DOI: https://doi.org/10.21203/rs.2.17981/v1

Version of Record: A version of this preprint was published at Parasites & Vectors on March 30th, 2020.
See the published version at https://doi.org/10.1186/s13071-020-04031-3.

Abstract

Background: Practical, field-ready age-grading tools for mosquito vectors of disease are urgently needed because of the impact that daily survival has on vectorial capacity. Previous studies have shown that near-infrared spectroscopy (NIRS), in combination with chemometrics and predictive modeling, can forecast the age of laboratory-reared mosquitoes with moderate to high accuracy. However, it remains it is unclear whether the technique has utility for identifying shifts in the age structure of wild-caught mosquitoes or whether models derived from the laboratory or semi-field mosquitoes can be applied to mosquitoes reared under different environments.

Methods: NIRS spectral data from adult female Aedes albopictus mosquitoes reared in the laboratory (2, 5, 8, 12 and 15 days old) and in semi-field cages populated by wild-caught pupae (resulting in adults of 1, 7 and 14 days old). Spectral data collected from mosquitoes were used to determine if models derived from laboratory material using partial least squares (PLS) regression for the development of predictive models could be effectively applied to mosquitoes from more natural semi-field environments.

Results: Models trained on spectra from laboratory-reared material were able to predict the age of other laboratory-reared mosquitoes with moderate accuracy and successfully differentiated all day 2 and 15 mosquitoes. Models derived with laboratory mosquitoes could not differentiate between semi-field age groups, with age predictions relatively indistinguishable for day 1-14. Pre-processing of spectral data and improving the PLS regression framework to avoid overfitting can increase accuracy, but predictions of mosquitoes reared in different environments remained poor. Principle component analysis confirms substantial spectral variations between laboratory and semi-field mosquitoes despite both being derived from the same island population.

Conclusions: Model trained on laboratory mosquitoes were able to predict ages of laboratory mosquitoes with good sensitivity and specificity, however it was unable to predict age class of semi-field mosquitoes. This study suggests that laboratory-reared mosquitoes do not capture enough environmental variation to accurately predict the age of the same species reared under different conditions. Further research is needed to explore alternative pre-processing methods and machine learning techniques, and to understand factors that affect absorbance in mosquitoes before field application using NIRS.

1. Introduction

Quantifying the average age of a mosquito population would provide cost-effective and compelling entomological evidence for the potential epidemiological impacts of vector control. The mosquito death rate is the most important determinant of vectorial capacity but measuring the age of wild-caught mosquitoes remains impractical and unreliable despite its epidemiological importance. The use of near infrared spectroscopy and chemometrics may offer a solution [1, 2, 3, 4, 5], but its ability to be used in the field remains untested.

Only older mosquitoes are able to transmit disease. This is because pathogens require time to replicate and disseminate in the mosquito after ingestion of an infected blood meal. This extrinsic incubation period (EIP) commonly takes 9–14 days for dengue and Zika [6, 7]. Age-grading methodologies are also required for determining the impact of any vector control intervention that might skew the age structure of a population (i.e. larviciding, insecticide treated materials, indoor residual sprays). Previous methods used to age *Aedes* sp. include transcriptional profiles, morphological differences and cuticular hydrocarbon (CHC) analysis (reviewed in [8]). The work on CHCs in particular [9] gives credence to the idea that the composition of mosquito exoskeletons changes with age and that NIRS might be used to measure the differing absorbances of those surfaces in relation to the organic compounds that they contain [9, 10], however these methods are often laborious, destructive, expensive and inaccurate. The use of near-infrared spectroscopy (NIRS) to age-grade mosquitoes requires no sample preparation and is fast and accurate in distinguishing young and old mosquitoes in laboratory derived samples. In that context, the NIRS method has been used to age grade *Anopheles* sp.1, 4], *Ae. aegypti [*11, 12] and *Ae. albopictus [*5]. Most models to date have been laboratory-derived and typically test their accuracy against mosquitoes from the same origin. The models they yield are therefore likely to be overly optimistic. MP Milali, MT Sikulu-Lord, SS Kiware, FE Dowell, RJ Povinelli and GF Corliss [13] examined spectra collected from laboratory and wild-caught Anopheles mosquitoes and found no significant difference between them. However, the age of the field-collected material appears to be unknown and so the capacity of those similar spectra to reflect age-related differences was not tested. At least for Anopheles, other studies suggest that calibrations generated with one population of mosquitoes are not applicable to combined data sets derived from NIRS studies conducted in different laboratories, on different populations or using different machines [2]. Similarly, models built using laboratory reared mosquitoes had low predictive power in relation to the age of Anopheline adults derived from wild-caught larvae [14].

Machine learning methods are required to convert spectral data into predictive models for mosquito age. This has historically been performed using Partial Least Squares (PLS) regression and the software GRAMS IQ (Thermo Scientific, MA, USA). It has been postulated that prediction accuracies might be improved using more complex analytical and model-building techniques [14, 15] and that pre-processing data cleaning might also improve performance [2].

In the present study, we used a laboratory reared colony of Aedine mosquitoes to attempt to predict the age of mosquitoes collected as pupae in the field and reared to known age in cages held at ambient field conditions. To our knowledge this is the first attempt to use laboratory reared Aedes mosquitoes to develop predictive models of age for mosquitoes derived from field-collected material under ambient environmental conditions. We also examined whether the accuracy of our calibration and prediction models could be improved using different pre-processing and analytical techniques.

The Asian tiger mosquito (*Aedes albopictus)* which is the subject of this study originates from South East Asia, but now has a global distribution facilitated by the international movement of passengers and cargo and its ability to adapt quickly to new environments [16]. *Aedes albopictus* is an important vector

of dengue [17, 18] and chikungunya [19, 20]. It was first noted on the Torres Strait islands of Australia in 2005 and has since facilitated some minor dengue outbreaks in that region [21].

2. Materials And Methods

2.1 Mosquito collection and rearing

2.1.1 Laboratory-reared mosquitoes

Aedes albopictus eggs were collected from Hammond Island, Torres Strait, Australia in June 2016 and used to derive a stable laboratory maintained colony at the quarantine facility in QIMR Berghofer. Larvae hatched from that colony were reared in trays (35 × 15 cm) of de-chlorinated water kept at 27°C and 70% humidity. Larvae were fed with ground fish food *ad libitum* (Tetramin fish food flakes; Blacksburg, VA) and pupae were removed to round containers (9 cm diameter, 130 mls water). Emerging females were transferred daily to cages and provided with 10% sucrose ad libitum.

Those adults were killed 2, 5, 8, 12 and 15 days post-emergence. Individuals of the same age from different generations were pooled to include possible variations in laboratory rearing conditions. Mosquitoes were anaesthetised with CO_2 and placed in individual 1 mL tubes containing RNA*later*® (Ambion, TX, USA), a standard protocol for NIRS characterization [22]. Tween–20 (0.1% v/v) was added to the RNA*later*® to reduce surface tension and allow RNA*later*® to fully penetrate the mosquito. Sample tubes were maintained at room temperature for 24h. The mosquitoes were then preserved at –20°C until spectral collection (< 14 days later). The total number of mosquitoes collected is listed in Table 1.

2.1.2 “Semi-field” mosquitoes

The use of the term “semi-field” simply reflects the fact that these mosquitoes have an origin that is more representative of the field than of the laboratory. They were collected as pupae from a natural habitat (a productive rainwater tank) on Hammond Island, Torres Strait during March 2018. This site is also the origin of the material used to derive the 2016 laboratory colony (see above). Pupae were allowed to emerge in standard rearing cages (60 × 60 × 60 cm, Bugdorm, Megaview, Taiwan) maintained outdoors under ambient conditions. Adults were aspirated from the cages when they were 1, 7 or 14 days old, immobilized by cold (4^{o}C) and placed in RNA*later*® with 0.1% (v/v) of Tween–20 at –20°C until ready for shipping to QIMR Berghofer.

2.2 Mosquito scanning using near-infrared spectroscopy

Preserved, frozen mosquitoes were defrosted at room temperature and excess RNA*later*® removed by placing specimens on paper towelling. A Spectralon plate was used for spectral background collection. Individual mosquitoes were placed on the Spectralon plate laterally, and the head and thorax were

scanned using the LabSpec 5000 NIR spectrometer (Malvern Panalytical, Longmont, CO, USA). NIR spectra were obtained with an attached bifurcated fiber-optic probe that is approximately 2.4 mm above the Spectralon plate; scanning an area of approximately 2 mm. Spectral data was recorded in the 350–2500 nm region. Each spectrum was built using an average of 30 scans at a sampling resolution of 3 nm. Spectral data were collected using RS3 v6.4.3 (Malvern Panalytical, Longmont, CO, USA). Reflectance (R) is converted to absorbance (log 1/R) through RS3 prior to analyses.

2.3 Data analysis

2.3.1 Estimating mosquito age in days

Analyses were performed within the wavelengths of 700 to 2350 nm to disregard background noise at the start and end of the spectra, and any colour differences in mosquitoes (detected at < 700 nm). PLS linear regression was used to convert spectral data into predictive models of mosquito age (in days). Previous mosquito NIRS studies have used GRAMS IQ software (Thermo Scientific, MA, USA) to conduct the PLS analysis GRAMS IQ uses a “leave-one-out” method for internal cross-validation where one sample is taken from the calibration set and the remaining samples are used to develop an equation that would predict that removed sample. This process is repeated for all samples to create a predictive regression model (calibration model). This process is repeated varying the number of PLS components (factors) and the best model has historically been selected by eye [1, 11, 23], choosing the number of components that maximises accuracy whilst trying to minimise over-fitting (inclusion of too many components results in models that fit the sampled data perfectly but that fail to predict new data). This selection process is rather a subjective process making it hard to do reproducible science. Here we repeat the methods of the past (leave-one-out internal cross validation and selecting the number of components by eye) and refer to this method as “Standard PLS”.

An alternative approach for the development of predictive models is to split the dataset into three for training, validation and testing [24, 25]. Here we use 50% of the sample for training (fitting the model to samples of known age using different numbers of PLS components), 25% for validation (selecting an optimum number of that components that effectively predict another subset of known samples) and 25% to the test dataset (evaluating that final model against a blinded sub set of data). This process is repeated 100 times, each time randomly resampling the original dataset to generate different training, validation and testing datasets. The mean model is then selected from the 100 randomisations in order to average out sampling error. Here the number of components selected during the validation exercise is the lowest that permits an average error of 0.5 days of the best fitting model. This value was arbitrarily selected to be a compromise between accuracy and generalizability but has the advantage over previous methods in that this value can be defined and therefore the research is reproducible. This resampling procedure and selection of the number of components is referred to as “resampling PLS” and has been used to optimise models for predicting the presence of malaria parasites in mosquitoes [25]. Results are shown comparing the standard error of the predictions with the true age of the mosquito (root-mean-

square deviation, RMSD). To allow a direct comparison with Standard PLS, RMSD estimates for Resampling PLS were calculated on estimates of individual mosquito age calculated from the mean of the 100 randomisations using the entire dataset.

Mathematical pre-treatment of spectra may reduce noise and increase differentiation between sample properties. To investigate whether the accuracy of the standard PLS models could be improved by pre-processing techniques we examined standard normal variate (SNV), mean normalizing, and detrend-SNV methods to minimize spectral distortion due to scattering. We used second derivative Savitzky-Golay (SG) filtering to remove baseline noise [26, 27].

2.3.2 Classifying mosquitoes as young and old

Previous NIRS studies have estimated mosquito age as above or below a threshold in days (i.e. > or < X days old; [3, 4, 5]. A more robust method would be to directly train the calibration model to classify young and old mosquitoes. Here we use a binomial logistic regression framework to classify mosquitoes as young or old using the same resampling PLS framework outlined above [24]. An 8-day threshold is used to differentiate between young and old mosquitoes as it was the median age of mosquitoes collected thus allowing the calibration dataset to be evenly balanced between outcomes. Misclassification rates (the proportion of test observations incorrectly classified) were used to estimate the optimal boundary threshold (the value of the linear predictor differentiating between age classes). Overall accuracy is assessed by comparing the area under the receiver operating characteristic (ROC) curve (AUC). This is a graphical tool commonly used to illustrate the diagnostic accuracy of binary classification systems, with an AUC of 0.5 signifying the ability of NIRS to classify old and young mosquitoes is no better than chance whilst a value of 1 indicates perfect accuracy. The model with the minimum number of components that is within 0.01 of the model with the highest AUC is selected. Estimates of whether a mosquito is young or old are made by averaging prediction from 100 randomisations and comparing that to the average cut-off for all mosquitoes [24]

2.3.2 Analysis of spectra

Potential outliers in the data were identified and removed using Hotelling T^2 statistics, where samples positioned outside of a 95% confidence interval ellipse are considered outliers. Ten laboratory samples and nine semi-field samples were removed as outliers. Principal component analysis (PCA) was then used to identify spectral differences and clustering within the datasets. Loading plots generated from PCA were analysed to identify key absorbance peaks that may correspond to the age grading of mosquitoes. PCA analysis was conducted in Unscrambler X (v. 10.5.1).

3. Results

3.1 Determining mosquito age in days

NIRS can determine the calendar age of laboratory reared *Ae. albopictus* mosquitoes with moderate accuracy but our laboratory model fails to predict the age of the same mosquito species with the same geographic origin reared *in situ*. The exact predictive accuracy depends on the method of analysis. The best fit calibration model using laboratory data are shown in Figure 1 generated with PLS framework (Figure 1A-C) and with the standard PLS (Additional File 1: Figure S1A-C). Both methods generate regression coefficients with peaks at similar wavelengths (Figure 1A, Additional File 1: Figure S1A) which are broadly the same as those observed previously [12] although they differ in amplitude. Accuracy varies between methods of analyses: the resampling PLS framework gives an average difference between the true age and the predicted age of individual mosquitoes being 2.89 days compared to 2.38 days for the standard method (Table 1). Average estimates for the different age classes were also more accurate in every age group using the resampling method (Table 1). Importantly, the resampling PLS method accurately predicted the age of 2 and 15 day old mosquitoes with minimal bias. Neither technique accurately predicted the age of day 5 or 12 mosquitoes.

Neither PLS model derived from laboratory reared mosquitoes is able to predict the age of semi-field mosquitoes (Figure 1C, Additional File 1: Figure S1C). The standard PLS method has an average error of 5.84 days whilst the resampling method gave an error of 5.42 days. Mean predicted age across the three different semi-field age groups was broadly the same, and age groups were indistinguishable from one another (Table 1). The inability of to predict the age of semi-field mosquitoes is not driven by a lack of signal from the semi-field mosquitoes as training the model on the semi-field mosquitoes alone and then using that to predict the age of a subset of these mosquitoes (internal cross-validation) generates moderately accurate results (average error of 3.41 days and 2.82 days for standard and resampling methods, respectively Table 1).

Preprocessing spectra before standard PLS substantially improved the accuracy of the calibration model for the laboratory derived mosquitoes. The most successful method was Detrend-SNV, which reduced the average error in the calibration (laboratory-derived) dataset to 2.09 days (broadly in line with the accuracy of the resampling PLS framework), however, the accuracy of that model in predicting the age of semi-field mosquitoes remained poor (average error of 4.9 days, see Additional File 3: Table S1).

Table 1. Predictive power of models derived from different *Ae. albopictus* populations. The true age of mosquito groups is shown on the left while the mean predicted age (and variability, given as standard error of the mean, SEM) is shown on the right using standard Partial Least Squares (PLS) regression or a resampling PLS framework. Mosquitoes are classified as young (<8 days) or old (≥ 8 days). First line of each section of the table shows the number of components used in the different models. Accuracy of age estimates is shown by the average difference between the true and predicted age measured in days (root-

mean-square deviation, RMSD), with lowest values indicating a more accurate model. The ability to classifying mosquitoes as young or old is given by the area under the curve (AUC), with higher values indicating greater accuracy.

Actual age (days)	No. scanned	Standard PLS		Resampling PLS	
		Age in days	Classed as old (%)	Age in days	Classed as old (%)
1. Using laboratory-derived models to predict the age of laboratory-reared mosquitoes					
Number of components		8		10	4
2	41	3.58 (0.28)	0	2.13 (0.22)	4
5	42	7.62 (0.30)	36	7.27 (0.23)	9
8	42	8.35 (0.27)	60	8.07 (0.21)	57
12	42	8.47 (0.33)	57	9.84 (0.22)	58
15	44	14.0 (0.33)	100	14.7 (0.26)	68
Overall accuracy		RMSD=2.89	-	RMSD=2.38	AUC=0.88
2. Using semi-field derived models to predict the age of semi-field reared mosquitoes					
Number of components		8		10	15
1	50	4.58 (0.32)	6	3.31 (0.28)	0
7	50	7.71 (0.44)	41	7.28 (0.36)	0
14	100	11.7 (0.34)	90	12.7 (0.26)	100
Overall accuracy		RMSD=3.41	-	RMSD = 2.82	AUC=0.97
3. Using laboratory-derived models to predict the age of semi-field reared mosquitoes					
Number of components		8		10	4
1	50	6.50 (0.20)	12	8.23 (0.18)	36
7	50	6.92 (0.20)	18	9.40 (0.27)	74
14	100	7.00 (0.15)	24	9.18 (0.22)	75

Overall accuracy	RMSD=5.84	-	RMSD = 5.42	AUC=0.60

3.2 Binary classification (young or old)

The ability of NIRS to differentiate between young and old mosquitoes varied substantially according to the method of analysis (Table 1). Most previous work has classified mosquitoes as young or old by estimating the age in days and then using this to classify mosquitoes as young or old. This method performs relatively poorly, overall misclassifying 23.6% mosquitoes reared in the laboratory though it does correctly classify very young and very old laboratory-reared mosquitoes with high accuracy (100% of 2 and 15 day old mosquitoes; Table 1). Standard PLS models derived from laboratory reared mosquitoes also failed to predict the age of semi-field mosquitoes, misclassifying 16.6% of mosquitoes (Table 1).

Training the model to directly classify young or old mosquitoes substantially improves the accuracy of results. The resampling PLS classification model selects different regions of the spectrum (Figure 2A) compared to the continuous age model (Figure 1A), though some regions were informative to both models. Overall, the resampling PLS models ability to predict the age of laboratory mosquitoes was high (Figure 2B, 2C, 2D, AUC=0.88) with good sensitivity (0.75) and specificity (0.86). However, the model trained on laboratory mosquitoes was still unable to predict the age class of mosquitoes reared in the semi-field environment with a sensitivity of 0.55, specificity of 0.55 and a low overall accuracy (AUC=0.60).

There remains a strong age-related signal from semi-field mosquitoes even if they cannot be predicted by laboratory samples. Resampling PLS models derived from semi-field samples accurately differentiated 100% of day 1, 7 and 14 mosquitoes (AUC=0.97; Table 1). The standard PLS framework could only classify only day 1 and day 14 mosquitoes with any accuracy.

3.2 Spectra investigation

Results from an analysis of spectral data identified four principal components that explained 88%, 8%, 2% and 1% of the variance observed. A scatter plot illustrates spectral differences between semi-field and laboratory mosquitoes (Figure 3A, Supporting Figure 3). The clustering of semi-field mosquitoes towards PC-1 could reflect higher water content in those samples, as variances appears to result from absorbance peaks associated with H_2O (1450 nm and 1930 nm) [28] as can be seen in Figure 3B. Younger weevils are

found to have higher H_2O content compared to older individuals [29], indicating that H_2O may influence age grading in insects. The water signals can be detected at these peaks when comparing signals of dried mosquitoes (storage in silica for two days) to mosquitoes treated similarly to this study (Figure 3C). The remaining 12% of the variances observed consisted of some overtones of water absorbance peaks and many weak signals that are difficult to interpret. Scatter plots of PC-2 and PC-3 showed less dramatic spectral differences of unknown cause (Additional File 2: Figure S2). Overall, there are clear differences between spectral outputs that reflect differences in water, protein and other chemical content suggesting predicting age in field mosquitoes of different provenance will be challenging.

4. Discussion

Most outbreak control programs for Aedes-borne diseases rely on a combination of habitat management, larvicides and adulticides. The latter are expected to have dramatic impacts on survival, age structure and consequently the vectorial capacity of the target population [30]. Reductions in the proportion of older mosquitoes in a vector population will dramatically limit disease transmission risks by reducing the proportion of mosquitoes living long enough to feed on an infected host, incubate a pathogen and transmit that pathogen to humans. Fast, cost-effective tools for determining age structure could provide epidemiologically-relevant, entomological measures for program evaluation. The more accurate the age prediction, the more might be inferred about transmission risks in the context of the extrinsic incubation periods (EIPs) of diseases like dengue, Zika and chikungunya. Unfortunately, there are no operationally practical age-grading methods currently available for Aedes [31, 32].

NIRS measures the absorption of organic compounds within a sample using an electromagnetic spectrum in the near-infrared region. The derived spectra are complex and multivariate analytical techniques are required for their interpretation. If these outputs are to be of utility for programmatic field evaluations, the predictive power of the derived models must effectively classify the age of field-collected material of unknown provenance. This validation is clearly challenging to design and test. It requires a comparison of NIRS data derived from mosquitoes of known calendar age, with field collected mosquitoes graded using an independent proxy such as parity, or an alternative age-grading technique such as hydrocarbon analyses or transcriptional profiles [8, 33].

An initial, simpler step in that process is to show that laboratory-derived models are applicable to field-derived material. In this instance, we attempted to correlate the calendar age of mosquitoes from a laboratory colony, with the calendar age of mosquitoes derived from field-collected pupae reared to the adult stage in cages held at ambient field conditions.

We determined that PLS regression models can predict the age of other mosquitoes reared under near-identical conditions. This is consistent with other studies that have used models derived from laboratory-reared *Ae. albopictus*, *Ae aegypti* and *Anopheles gambiae* s.l. colonies to predict the age of other mosquitoes from the same colonies [5, 11, 12, 34]. In an extension of that work, our study demonstrates that models derived from either laboratory or semi-field mosquitoes were unable to predict the age of the other. Similar results were seen when mosquitoes were ascribed a binary age classification (< or > 8 days). This has important implications as the utility of the technique in field programs will rely upon an ability to use data sets from one origin, geographic area or sampling period to accurately predict the age of other samples of a different provenance.

Although our laboratory and semi-field mosquitoes originated from the same site, the former was established in 2016 and is likely to have diverged significantly in profile from those mosquitoes collected in 2018. Their respective histories of nutrition, competition and development times will have been very different, as will a host of other environmental and abiotic factors. We therefore cannot determine whether the failure of models trained on field data to predict the age of semi-field mosquitoes is due to differences in the mosquito population or the rearing conditions. Further work is needed to confirm whether models derived from laboratory-reared mosquitoes can be used to predict the age of semi-field and field mosquitoes. The utility of NIRS will depend on whether models trained on one group of mosquitoes would be able to predict the age of mosquitoes from different times and places. PCA analysis suggest substantial variation in spectra between lab and field-derived material. That seems at odds with the conclusions of a previous study, which suggests no difference between near infrared spectra collected from lab and field mosquitoes [13]. Spectral analyses suggests that water absorbance peaks may contribute to the variation that we have seen and reflect the physiological state of the mosquito or the immediate environment. Water creates strong NIR signals that may dominate other signatures in the cuticle [15] and may be masking important, age-related spectra, however this cannot be confirmed unless there are additional studies performed on dried mosquitoes. All adult mosquitoes used in this study were cage-reared with *ad libitum* access to 10% sugar solution, therefore it is unknown if water signals directly influences spectral data collected for age grading, and whether moisture content in a mosquito is a limitation for NIR mosquito studies.

Exploring alternatives to the standard PLS regression substantially improved internal cross-validation. The resampling method which uses 100 randomisations of the original dataset substantially reduces overfitting in comparison to the leave-one-out methodology typically employed [24]. This is especially the case for relatively small datasets. This study used spectra from ~40 mosquitoes of highly homogenous origin to represent each age category. This is in line with previous studies [1, 5, 11, 23] but the accuracy and robustness of machine learning techniques will improve substantially as more samples are included and as more variability is captured [2]. The resampling PLS framework also enables the number of components to be automatically selected, increasing reproducibility and probably contributing to improvements in the out-of-sample accuracy [25]. Pre-processing of spectral data also appeared to substantially improve accuracy of the standard PLS and is routinely applied to spectral data, especially on solid materials where light scattering often occurs [35]. This reduces background noise, which

consists of random deviations of the spectral measurements and systematic variations within samples which are unimportant in the analysis [36]. In our study, pre-processing produced calibration models with higher accuracy and fewer principle components. There were no significant differences in accuracy of semi-field age predictions using laboratory calibrations but there seems further utility in exploring alternative pre-processing methods and machine learning techniques.

5. Conclusions

There remain many challenges to the development and adoption of NIRS as an age-grading tool with a field application. The application of NIRS and chemometrics to the age classification of insects would benefit from a better understanding of the factors that affect absorbance and the challenges they pose to accurate prediction. A spectral database defined in terms of its causative physiological or biochemical drivers might allow for data analyses to be performed using only the most relevant regions, as can be seen in a mosquito mid-infrared study [15]. This method has also been used in various other NIR studies, where there is emphasis on an individual wavelength related to the detection of a specific chemical bond [37, 38]. Efforts to define its potential and limitations are essential as we consider our priorities for research and development.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

Not applicable.

Competing interests

The authors declare that they have no competing interests. Mention of trade names or commercial products does not imply recommendation or endorsement by the USDA.

Funding

OTWO, GJD and TSC were supported by USAID grant AID-OAA-F-16-00094 and the UK Medical Research Council (MRC) grant MR/P01111X/1. TSC received Centre funding from the UK MRC/UK Department for International Development (DFID) under the MRC/DFID Concordat agreement.

Acknowledgements

Not applicable.

Authors' contributions

OTWO and GJD discussed the contents of the research. OTWO and CF collected and reared laboratory and semi-field mosquitoes, respectively. OTWO collected spectral data from mosquitoes. OTWO, EAK, PME and TSC performed spectral data analyses. FED provided the LabSpec 5000 NIR spectrometer and valuable insight into age-grading insects using NIRS. OTWO, GJD and TSC wrote and revised the manuscript.

References

1. Mayagaya VS, Michel K, Benedict MQ, Killeen GF, Wirtz RA, Ferguson HM, et al. Non-destructive determination of age and species of *Anopheles gambiae s. l.* using near-infrared spectroscopy. The American Journal of Tropical Medicine and Hygiene. 2009;81 4:622-30.
2. Lambert B, Sikulu-Lord MT, Mayagaya VS, Devine G, Dowell F, Churcher TS. Monitoring the age of mosquito populations using near-infrared spectroscopy. Scientific Reports. 2018;8 1:5274.
3. Sikulu M, Killeen GF, Hugo LE, Ryan PA, Dowell KM, Wirtz RA, et al. Near-infrared spectroscopy as a complementary age grading and species identification tool for African malaria vectors. Parasites & Vectors. 2010;3 1:49.
4. Sikulu MT, Majambere S, Khatib BO, Ali AS, Hugo LE, Dowell FE. Using a near-infrared spectrometer to estimate the age of Anopheles mosquitoes exposed to pyrethroids. PLoS One. 2014;9 3:e90657.
5. Sikulu-Lord MT, Devine GJ, Hugo LE, Dowell FE. First report on the application of near-infrared spectroscopy to predict the age of *Aedes albopictus* Skuse. Scientific Reports. 2018;8 1:9590.
6. Ritchie SA, Pyke AT, Hall-Mendelin S, Day A, Mores CN, Christofferson RC, et al. An explosive epidemic of DENV-3 in Cairns, Australia. PLoS One. 2013;8 7:e68137.
7. Hugo RLE, Stassen L, La J, Gosden E, Ekwudu Om, Winterford C, et al. Vector competence of Australian *Aedes aegypti* and *Aedes albopictus* for an epidemic strain of Zika virus. PLoS Neglected Tropical Diseases. 2019;13 4:e0007281.
8. Hugo LE, Quick-Miles S, Kay B, Ryan P. Evaluations of mosquito age grading techniques based on morphological changes. Journal of Medical Entomology. 2014;45 3:353-69.
9. Caputo B, Dani FR, Horne GL, Petrarca V, Turillazzi S, Coluzzi M, et al. Identification and composition of cuticular hydrocarbons of the major afrotropical malaria vector *Anopheles gambiae s. s.* (Diptera: Culicidae): Analysis of sexual dimorphism and age-related changes. Journal of Mass Spectrometry. 2005;40 12:1595-604.
10. Miller C. Chemical Principles of Near-Infrared Technology. St. Paul, USA: American Association of Cereal Chemists Inc.; 2001.

11. Sikulu-Lord MT, Milali MP, Henry M, Wirtz RA, Hugo LE, Dowell FE, et al. Near-infrared spectroscopy, a rapid method for predicting the age of male and female wild-type and Wolbachia infected *Aedes aegypti*. PLoS Neglected Tropical Diseases. 2016;10 10:e0005040.
12. Liebman K, Swamidoss I, Vizcaino L, Lenhart A, Dowell F, Wirtz R. The influence of diet on the use of near-infrared spectroscopy to determine the age of female *Aedes aegypti* mosquitoes. The American Journal of Tropical Medicine and Hygiene. 2015;92 5:1070-5.
13. Milali MP, Sikulu-Lord MT, Kiware SS, Dowell FE, Povinelli RJ, Corliss GF. Do NIR spectra collected from laboratory-reared mosquitoes differ from those collected from wild mosquitoes? PLoS One. 2018;13 5:e0198245.
14. Krajacich BJ, Meyers JI, Alout H, Dabiré RK, Dowell FE, Foy BD. Analysis of near infrared spectra for age-grading of wild populations of *Anopheles gambiae*. Parasites & Vectors. 2017;10 1:552.
15. Gonzalez-Jimenez M, Babayan SA, Khazaeli P, Doyle M, Walton F, Reedy E, et al. Prediction of malaria mosquito species and population age structure using mid-infrared spectroscopy and supervised machine learning. Wellcome Open Research. 2019;4.
16. Benedict MQ, Levine RS, Hawley WA, Lounibos LP. Spread of the Tiger: Global Risk of Invasion by the Mosquito *Aedes albopictus*. Vector-borne and Zoonotic Diseases. 2007;7 1:76-85.
17. Luo L, Jiang L-Y, Xiao X-C, Di B, Jing Q-L, Wang S-Y, et al. The dengue preface to endemic in mainland China: the historical largest outbreak by *Aedes albopictus* in Guangzhou, 2014. Infectious Diseases of Poverty. 2017;6 1:148.
18. Kobayashi D, Murota K, Fujita R, Itokawa K, Kotaki A, Moi ML, et al. Dengue Virus Infection in Aedes albopictus during the 2014 Autochthonous Dengue Outbreak in Tokyo Metropolis, Japan. The American Journal of Tropical Medicine and Hygiene. 2018;98 5:1460-8.
19. Lounibos LP, Kramer LD. Invasiveness of *Aedes aegypti* and *Aedes albopictus* and vectorial capacity for chikungunya virus. The Journal of Infectious Diseases. 2016;214:S453-S8.
20. Lindh E, Argentini C, Remoli ME, Fortuna C, Faggioni G, Benedetti E, et al: **The Italian 2017 outbreak chikungunya virus belongs to an emerging *Aedes albopictus*–adapted virus cluster introduced from the Indian subcontinent**. In: *Open Forum Infectious Diseases 2018*: Oxford University Press US: ofy321.
21. Muzari M, Davis J, Bellwood R, Crunkhorn B, Gunn E, Sabatino U, et al. Dominance of the tiger: The displacement of *Aedes aegypti* by *Aedes albopictus* in parts of the Torres Strait, Australia. Communicable Diseases Intelligence. 2019;43.
22. Dowell FE, Noutcha AE, Michel K. The effect of preservation methods on predicting mosquito age by near-infrared spectroscopy. The American Journal of Tropical Medicine and Hygiene. 2011;85 6:1093-6.
23. Sikulu M, Dowell KM, Hugo LE, Wirtz RA, Michel K, Peiris KH, et al. Evaluating RNAlater® as a preservative for using near-infrared spectroscopy to predict Anopheles gambiae age and species. Malaria Journal. 2011;10 1:186. https://doi.org/10.1186/1475-2875-10-186.

24. Esperança PM, Blagborough AM, Da DF, Dowell FE, Churcher TS. Detection of *Plasmodium berghei* infected *Anopheles stephensi* using near-infrared spectroscopy. Parasites & Vectors. 2018;11 1:377.
25. Esperança PM, Da DF, Lambert B, Dabire RK, Churcher TS. Functional data analysis techniques to improve the generalizability of near-infrared spectral data for monitoring mosquito populations. In press.
26. Agelet LE, Hurburgh Jr CR. A tutorial on near infrared spectroscopy and its calibration. Critical Reviews in Analytical Chemistry. 2010;40 4:246-60.
27. Martens H, Martens M. Multivariate analysis of quality. An introduction. vol. 12. London: John Wiley & Sons; 2001.
28. Ozaki Y, McClure WF, Christy AA. Near-infrared spectroscopy in food science and technology. John Wiley & Sons; 2006.
29. Perez-Mendoza J, Throne JE, Dowell FE, Baker JE. Chronological age-grading of three species of stored-product beetles by using near-infrared spectroscopy. Journal of Economic Entomology. 2004;97 3:1159-67.
30. Cook PE, McMeniman CJ, O'Neill SL. Modifying insect population age structure to control vector-borne disease. Transgenesis and the management of vector-borne disease: Springer; 2008. p. 126-40.
31. Lardeux F, Ung A, Chebret M. Spectrofluorometers are not adequate for aging Aedes and Culex (Diptera: Culicidae) using pteridine fluorescence. Journal of Medical Entomology. 2000;37 5:769-73.
32. Hugo LE, Jeffery JA, Trewin BJ, Wockner LF, Yen NT, Le NH, et al. Adult survivorship of the dengue mosquito *Aedes aegypti* varies seasonally in central Vietnam. PLoS Neglected Tropical Diseases. 2014;8 2:e2669.
33. Cook PE, Hugo LE, Iturbe-Ormaetxe I, Williams CR, Chenoweth SF, Ritchie SA, et al. Predicting the age of mosquitoes using transcriptional profiles. Nature Protocols. 2007;2 11:2796.
34. Sikulu-Lord MT, Maia MF, Milali MP, Henry M, Mkandawile G, Kho EA, et al. Rapid and non-destructive detection and identification of two strains of Wolbachia in *Aedes aegypti* by near-infrared spectroscopy. PLoS Neglected Tropical Diseases. 2016;10 6:e0004759.
35. Rinnan Å, Van Den Berg F, Engelsen SB. Review of the most common pre-processing techniques for near-infrared spectra. Trends in Analytical Chemistry. 2009;28 10:1201-22.
36. Rinnan Å. Pre-processing in vibrational spectroscopy–when, why and how. Analytical Methods. 2014;6 18:7124-9.
37. Fox GP, Onley-Watson K, Osman A. Multiple linear regression calibrations for barley and malt protein based on the spectra of hordein. Journal of the Institute of Brewing. 2002;108:155-9.
38. Bittante G, Cecchinato A. Genetic analysis of the Fourier-transform infrared spectra of bovine milk with emphasis on individual wavelengths related to specific chemical bonds. Journal of Dairy Science. 2013;96:5991-6006.

Figures

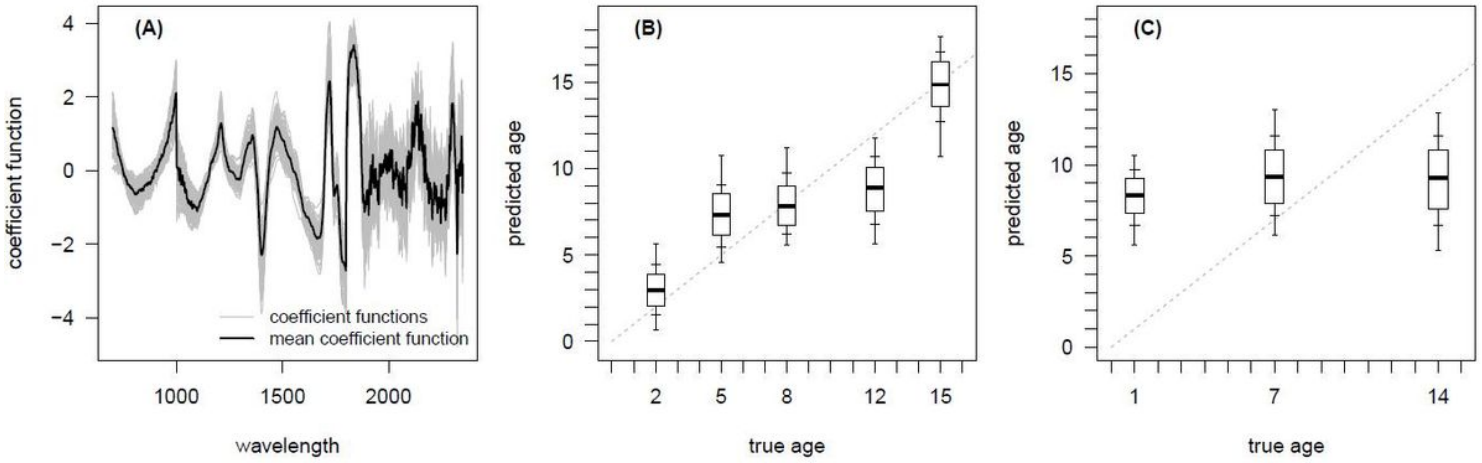

Figure 1

The ability of NIRS to predict the age of Ae. albopictus mosquitoes in days. (A) The best fit regression coefficient function for the resampling PLS model trained on laboratory reared mosquitoes showing the most informative regions of the spectrum. Grey lines show best fit model for each of the 100 dataset randomisations whilst black line indicates the average. (B) Ability of the model to predict age of laboratory reared mosquitoes. Boxplot thick horizontal black line shows the median/50th-percentile whilst the box edges, inner and outer whiskers show 25th/75th, 15th/85th and 5th/95th percentiles, respectively. Grey dashed line shows model with 100% accuracy. (C) Ability of the model trained on laboratory mosquitoes to predict the age of semi-field mosquitoes. Results can be compared to the simple PLS method presented in Additional File 1: Figure S1.

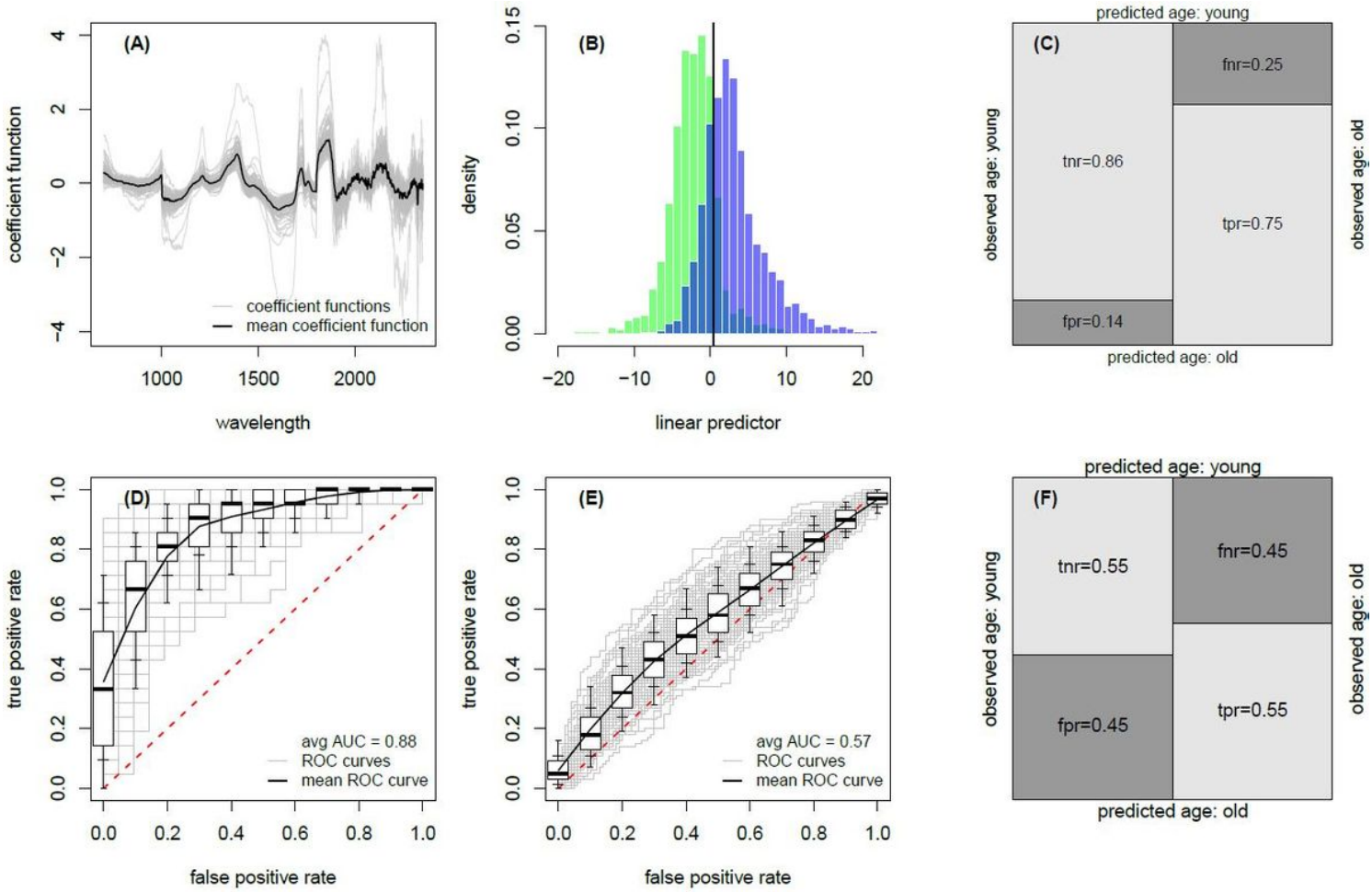

Figure 2

The ability of NIRS to classify Ae. albopictus mosquitoes as being young or old. (A) The best fit regression coefficient function for a model trained on laboratory reared mosquitoes showing the most informative regions of the spectrum. Grey lines show best fit model for each of the 100 dataset randomisations whilst black line indicates average. (B) Ability of the model to predict age classification of laboratory reared mosquitoes. Histogram of the estimated linear predictor for the test observations colour-coded by the true class (green = true young mosquitoes; blue = true old mosquitoes). Vertical black line indicates optimum threshold for classifying mosquitoes as old or young ("left" predicted to be young, "right" predicted to be old). The shaded area where two distributions overlap corresponds to misclassified test observations - false negatives to the left and false positives to the right of the optimal classification threshold. (C) The corresponding confusion matrix for the best model trained and predicting laboratory reared mosquitoes showing the different error rates: tnr, true negative rate; fnr, false negative rate (specificity); fpr, false positive rate; and tpr, true positive rate (sensitivity). (D) The receiver operating characteristic (ROC) curve for the best-fit model predicting laboratory reared mosquitoes showing the false positive and true positive rates achievable for different classification probability thresholds (shifting the black vertical line (B) left or right) whilst the overall performance is given by the area under the ROC curve (AUC). The pink dashed line denotes a model with no predictive ability (a random chance of correct

prediction) whilst a perfect model with 100% sensitivity and specificity would be in the top left corner (coordinates 0, 1). The solid line shows the average ROC curve; boxplots show the variability for 100 randomisations of the training, validation and testing datasets (box edges, inner and outer whiskers show 25th/75th, 15th/85th and 5th/95th percentiles, respectively; black line inside the box showing the median/50th-percentile). (E) The ROC curve showing the ability of the model trained on laboratory rerared mosquitoes to predict the age classification of mosquitoes reared in the semi-field environment. (F) shows the corresponding confusion matrix of the best model.

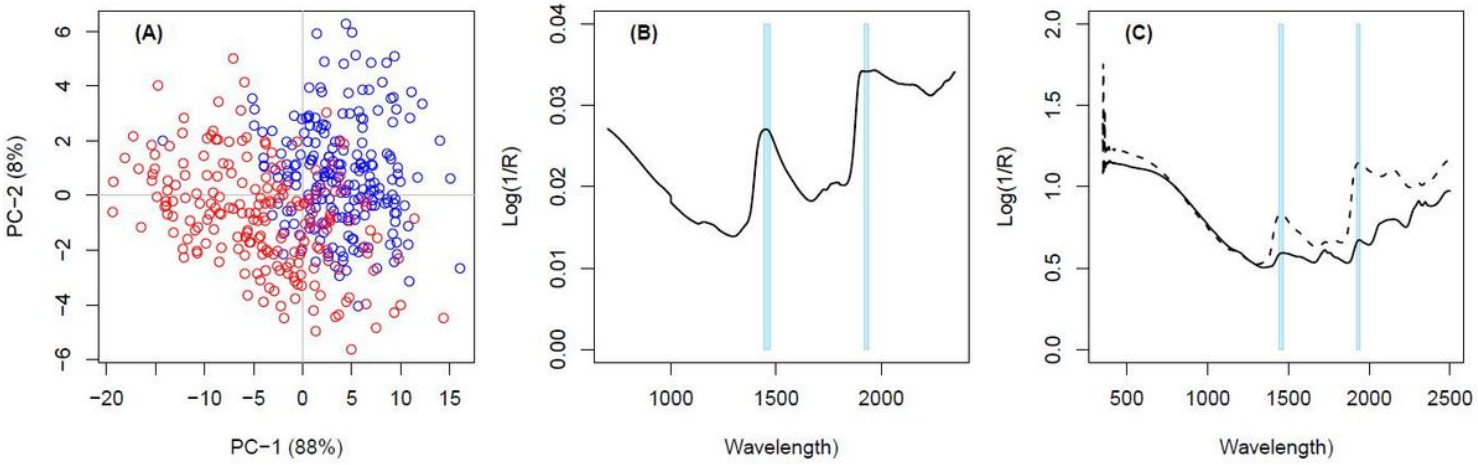

Figure 3

. Differences between laboratory and semi-field spectra. (A) Principle component analysis showing the difference between scores calculated for PC-1 (which explains 88% of the variation) and PC-2 (8% of the variation) for laboratory (blue; square) and semi-field (red; circle) mosquitoes. (B) Loading plot from PCA showing that water bands at 1450 nm and 1930 nm accounted for 88% of the total variance observed. R denotes reflectance (C) difference between undried (dashed line) and dried (solid line) mosquito spectra. In (B) and (C) blue horizontal lines indicate peaks associated with water.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile1.jpg
- ADDITIONALFILES.docx
- AdditionalFile2.jpg
- Graphicalabstract.png