Functional data analysis techniques to improve the generalizability of near-infrared spectral data for monitoring mosquito populations

Pedro M. Esperança[1], Dari F. Da[2], Ben Lambert[1],
Roch K. Dabiré[2] and Thomas S. Churcher[1]

[1] MRC Centre for Global Infectious Disease Analysis,
Department of Infectious Disease Epidemiology,
Imperial College London, Norfolk Place, London W21PG, UK.

[2] Institut de Recherche en Sciences de la Santé,
Direction Régionale, 399 Avenue de la liberté,
Bobo Dioulasso, 01 01 BP 545, Burkina Faso.

April 28, 2020

Abstract

Near infrared spectroscopy is increasingly being used as an economical method to monitor mosquito vector populations in support of disease control. Despite this rise in popularity, strong geographical variation in spectra has proven an issue for generalising predictions from one location to another. Here, we use a functional data analysis approach—which models spectra as smooth curves rather than as a discrete set of points—to develop a method that is robust to geographic heterogeneity. Specifically, we use a penalised generalised linear modelling framework which includes efficient functional representation of spectra, spectral smoothing and regularisation. To ensure better generalisation of model predictions from one training set to another, we use cross-validation procedures favouring smoother representation of spectra. To illustrate the performance of our approach, we collected spectra for field-caught specimens of *Anopheles gambiae* complex mosquitoes – the most epidemiologically important vector species on the planet – in two sites in Burkina Faso. Using these spectra, we show how models trained on data from one site can successfully classify morphologically identical sibling species in another site, over 250km away. Whilst we apply our framework to species prediction, our unified statistical framework can, alternatively, handle regression analysis (for example, to determine mosquito age) and other types of multinomial classification (for example, to determine infection status). To make our methods readily available for field entomologists, we have created an open-source R package `mlevcm`. All data used is publicly also available.

Keywords: entomological monitoring, functional data analysis, malaria, mosquito, near-infrared spectroscopy.

1 Introduction

Mosquito-borne diseases such as malaria, dengue and yellow fever are responsible for huge suffering, death and impose a considerable economic burden in Sub-Saharan Africa, Asia, and Latin America (Sachs and Malaney, 2002; WHO, 2019). The World Health Organization estimated 228 million cases of malaria alone in 2018 resulting in approximately 405,000 deaths. Malaria is transmitted from person to person by female mosquitoes of the *Anopheles* genus. Insecticides which kill mosquitoes, either incorporated into bednets or sprayed on walls, are the most effective method of controlling the disease and prevent millions of cases each year (Bhatt et al., 2015). However, differences in behaviour between mosquito species and the rise of insecticide resistance mean that control interventions increasingly need to be tailored to the local mosquito population. Factors such as species composition, the level of mosquito infection and age distribution in the mosquito population constitute an important direct measure of the efficacy of disease control interventions.

Unfortunately, there is no easy way to cheaply monitor mosquito populations in the field. Molecular techniques, like polymerase chain reaction, are required to determine mosquito species and infection status, which are laborious and require highly trained staff, well equipped laboratories and expensive reagents. By killing mosquitoes, the main effect of insecticides is to reduce mosquito lifespan, shifting the age distribution towards younger mosquitoes. Insecticide resistance may reduce the killing effect of insecticides, increasing the average age in mosquito populations supposedly controlled by insecticides. Therefore, methods to monitor the local age distribution in mosquito populations are critical for knowing whether insecticides remain effective. Yet, there are currently no fast, inexpensive methods for accurately surveying the age distribution of mosquito populations.

Near-infrared spectroscopy (NIRS) is a new, rapid, reagent-free and non-destructive scanning technique, which can determine the species of morphologically indistinguishable mosquitoes, approximate mosquito age and the presence of malaria and dengue infections (Esperança et al., 2018; Lambert et al., 2018; Mayagaya et al., 2009; Ong et al., 2020; Sikulu et al., 2010; Sikulu-Lord et al., 2016). The instrument is portable and battery powered, which means scanning can take place in remote locations. The scanning procedure itself does not require expensive reagents, specialised lab-trained staff or non-portable laboratory equipment and is extremely simple: mosquitoes are killed, placed under a light probe, and scanned to produce a spectrum within seconds. Previous work has successfully predicted characteristics of interest (age, species, infectiousness) from near-infrared (NIR) spectra, using only relatively basic machine learning models based on Partial Least Squares (PLS) regression (Gerlach et al., 1979).

Despite their success in predicting characteristics of interest within a given population of mosquitoes, it has been documented that these methods cannot predict these between populations: that is, models trained to predict (say) age in population A cannot predict age in population B (Lambert et al., 2018). This site specificity is not typically reported in NIRS studies of mosquitoes (e.g. Esperança et al., 2018) and means the reported performance of the method may exceed that in the field. Whilst the exact origins of this between-site performance are unknown, many possible factors may contribute. In a ‘typical’ NIRS study, performance of the method is evaluated by predicting mosquito characteristics in independent test sets. Whilst different mosquitoes may be used in the training and testing sets, they come from the same set of mosquitoes, potentially from the same mother, and are kept in identical rearing conditions. These shared factors mean mosquitoes comprising the test set are much more alike those in the training set than any wild-caught specimens. Part of these issues could be addressed by training models using F0 or F1 mosquitoes derived from collected individuals – although, admittedly, such restrictions on experimental practices would limit the usefulness of NIRS. There is, hence, a demand for machine learning methods that are robust to differences between laboratory training and wild testing sets.

In this article, we use a functional data analysis (FDA) framework to build NIR spectra-based machine learning models which maintain predictive accuracy between populations of mosquitoes. In FDA, individual data points are modelled as originating from (noisy) sampling of unobserved smooth, continuous functions at discrete intervals along them (Ramsay and Silverman, 2005). Since the observed NIR spectra already appear quite smooth (see Fig. 1), this suggests that an FDA approach should be applicable. The statistical framework for functional data has been developed in the past two decades and has been used for both regression (i.e. continuous response) and classification (i.e. categorical response) problems. The flexibility of FDA means that modern techniques such as efficient function representations, smoothing, penalised estimation and dimension reduction can be accommodated seamlessly—all of which we explore here (Morris, 2015; Ramsay and Silverman, 2002, 2005; Reiss et al., 2017; Wang et al., 2016).

To demonstrate the utility of our approach, we generated training and testing samples that mimicked how NIRS could be used in the field: we collected wild *Anopheles gambiae s.l.* mosquito larvae from two locations in Burkina Faso, separated by 283km, which were reared then scanned using near-infrared spectrometers in the laboratory. After scanning, the mosquitoes were killed and their species was determined by PCR. (Specimens were either *An. gambiae s.s.* or *An. arabiensis*, which are morphologically identical species that have epidemiologically important differences in ecology.) We then show that our FDA-based approach trained using paired species-spectra data from each location in isolation can predict the species of individual specimens in the other. To encourage others to replicate and build on our analysis, we make all data (Esperança, 2019a) and code (Esperança, 2019b) publicly available.

2 Methods

2.1 Mosquito data

To train and test our machine learning algorithm, we collected mosquito larvae from two locations in Burkina Faso: Klesso, in the southwest of the country, near Bobo-Dioulasso; and Longo, in the Hauts-Basins region approximately 283km away. Adult mosquitoes were reared from field collected larvae (F0 generation) or from wild, naturally fed mosquitoes caught resting in the eaves of houses which were allowed to lay eggs which were then reared to adult (F1 generation). All mosquitoes were kept in similar 'laboratory colony' conditions and killed using chloroform four days after emergence. Mosquitoes were scanned using a LabSpec4 Standard-Res i (stan-

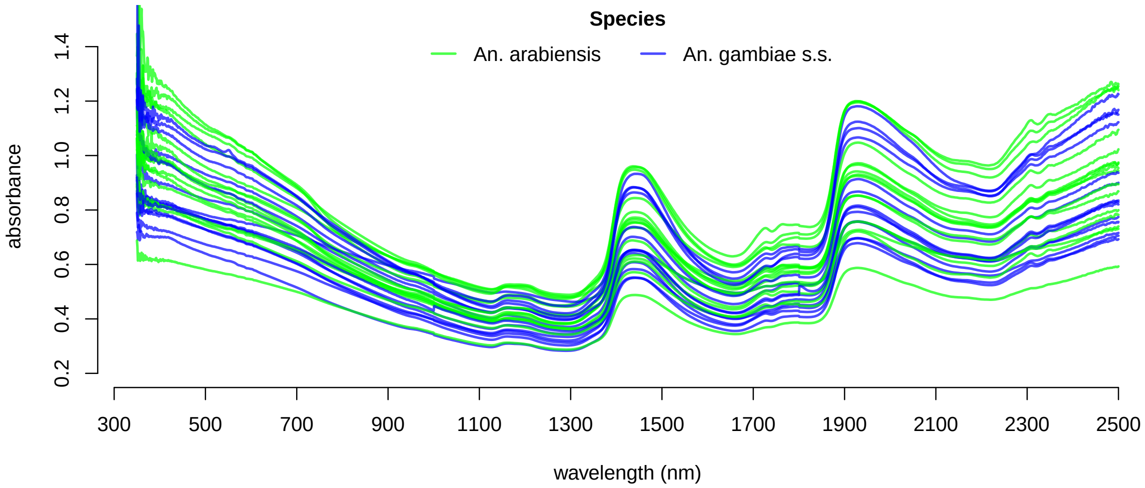

Figure 1: The spectra of 30 mosquitoes within the *Anopheles gambiae s.l.* (sensu lato) species complex, each sampled at discrete wavelengths in the interval $[350, 2500]$. Green lines show *Anopheles arabiensis* whilst blue lines show *Anopheles gambiae s.s.* (sensu stricto).

dard resolution, integrated light source) near-infrared spectrometer and a bifurcated reflectance probe mounted 2mm from a spectralon white reference panel (ASD Inc. (company), 2020). Absorbance was recorded across 350nm–2500nm of the electromagnetic spectrum. Specimens were laid on their side under the focus of the light probe and spectra were recorded with RS3 spectral acquisition software (ASD Inc. (company), 2020), which recorded the average spectra from 20 scans. The light probe was centred on the head and thorax region of the mosquito and each mosquito was scanned up to 4 times, picking the mosquito and replacing on their opposite side after each scan. The average number of scans per mosquito is 2.5 (75% with 2 scans, 24% with 4 scans, 1% with 1 or 3 scans). The mean absorbance across the multiple scans was then used in the analyses (Figure 1). After mosquitoes have been scanned, species was determined by Polymerase Chain Reaction (Fanello et al., 2002). This resulted in 224 spectra samples in Klesso (50 *An. arabiensis* and 174 *An. gambiae s.s.*) and 126 Longo (61 *An. arabiensis* and 65 *An. gambiae s.s.*).

2.2 Statistical Methods

Our approach follows a unified framework for functional data analysis (FDA; Ramsay and Silverman, 2005). As pre-processing, we represent spectra efficiently using basis functions (§2.2.1) and perform smoothing to eliminate measurement noise (§2.2.2). To classifying mosquito species we use a regularised, generalised linear model framework (§2.2.3) with dimension reduction (§2.2.4), and use a cross-validation procedure to optimise hyperparameters (§2.2.5). All variables that we use to describe our method are summarised in Table 1.

2.2.1 Spectra as functional data

Mosquito spectra can be viewed as smooth curves or functions sampled at discrete wavelengths in the near-infrared (NIR) region of the electromagnetic spectrum (Figure 1). Therefore, Func-

Table 1: Notation. Description of variables and their ranges.

Notation	Description	Space
$\boldsymbol{t} = [t_1, \ldots, t_p]$	Vector of discrete integer wavelengths from 350nm-2500nm	$\mathbb{N}^P$
$\mathcal{T}$	Wavelength range surveyed by spectrometer	$\mathbb{R}^\infty$
$P = \lvert\boldsymbol{t}\rvert$	Number of discrete absorbance values measured	$\mathbb{N}$
$X_i(t)$	Absorbance spectrum for mosquito i at wavelength t	$\mathbb{R}^\infty$
$\boldsymbol{X}_i(\boldsymbol{t})$	Vector of absorbances at discrete integer wavelengths	$\mathbb{R}^P$
$\boldsymbol{O}_i(\boldsymbol{t})$	Observed absorbance at discrete integer wavelengths	$\mathbb{R}^P$
$\boldsymbol{Z}_i(\boldsymbol{t})$	Observed non-functional predictors	$\mathbb{R}^S$
$b_k(t)$	kth basis function at wavelength t	$\mathbb{R}^\infty$
$\boldsymbol{b}(t)$	Basis function vector	$K \times \mathbb{R}^\infty$
ν_{ik}	Coefficient for basis function k for mosquito i	$\mathbb{R}$
$\boldsymbol{W}$	Weighting matrix used in estimation	$\mathbb{R}^{P\times P}$
$\beta(t)$	Function giving influence of wavelength t on predictions	$\mathbb{R}^\infty$
$\boldsymbol{\Omega}$	Matrix used to penalise roughness of coefficient function	$\mathbb{R}^{K\times K}$
λ	Parameter used to penalise roughness of coefficient function	$\mathbb{R}_0^+$

tional Data Analysis (FDA) can be used to represent these data (Ramsay and Silverman, 2005). Let $X(t)$ represent the true, underlying absorbance spectrum of a mosquito as a function of wavelength $t \in \mathcal{T}$. For each spectra sample, absorbances are reported at all integer wavelengths between 350nm and 2500nm. We denote the vector of absorbances for mosquito i by $\boldsymbol{X}_i(\boldsymbol{t}) = [X_i(350), \ldots, X_i(2500)], i \in [1{:}N]$.

We follow the standard theoretical framework in FDA, which assumes that functions are real-valued and belong to a Hilbert space containing square-integrable functions over the observed range of wavelengths (Febrero-Bande et al., 2017; Reiss et al., 2017).

2.2.2 Basis function representation and spectra smoothing

Basis functions provide an accurate and efficient way of representing complex functions as combinations of simpler functions, and constitute also to a natural framework for smoothing.

Representation. We express spectra as a linear combination of a set of K basis functions,

$$X_i(t) = \sum_{k=1}^{K} \nu_{ik} b_k(t) = \boldsymbol{\nu}_i^T \boldsymbol{b}(t), \tag{1}$$

where $\boldsymbol{b}(t) = [b_1(t), \ldots, b_K(t)]^T$ is a basis function vector, with $b_k(t)$ denoting the kth basis function evaluated at wavelength t; and $\boldsymbol{\nu}_i = [\nu_{i1}, \ldots, \nu_{iK}]^T$ is a basis coefficient vector, with ν_{ik} denoting the kth basis function coefficient for the ith mosquito spectra. The basis coefficients are estimated from the data, as detailed below, while the basis functions can be either data-dependent (e.g. principal components) or data-independent (e.g. B-splines and wavelets).

B-splines are a natural choice of basis system for NIR spectra. These are constructed from piecewise polynomial functions, typically of low degree n, with continuous derivatives up to derivative degree $n-1$, which makes them appealing both theoretically and computationally (de Boor, 2001; Eubank, 1999; Green and Silverman, 1994; Silverman, 1985). We use cubic B-

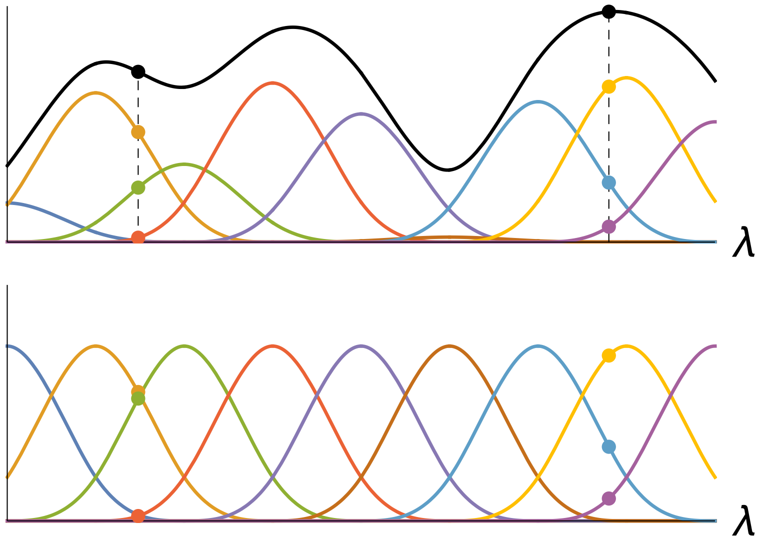

Figure 2: Representing near-infrared spectra using basis functions. BOTTOM: a number of unweighted cubic B-spline basis functions (coloured lines); the sets of points show the largest components of the basis function design matrix in two of its rows (corresponding to two measured wavelengths). TOP: The absorbance spectra (black line) at two wavelengths (black points) is obtained by summing the contributions from weighted individual splines (coloured lines) as in (1).

splines ($n = 3$), which are sufficiently flexible to represent spectra accurately. Figure 2 illustrates how spectral data can be represented using cubic B-splines.

Smoothing. Spectral observations are subject to measurement error due to imperfections in the spectrometer's detection sensors (see Figure 1). Poor signal-to-noise ratios make it harder to avoid overfitting—especially, if noise varies between training and test sets. The measurement error is assumed to be Gaussian and independent for each sample i and wavelength t,

$$O_i(t) = X_i(t) + \varepsilon_i(t) \quad \text{with} \quad \varepsilon_i(t) \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_t), \tag{2}$$

where $O_i(t)$ and $X_i(t)$ represent, respectively, the observed noisy measurements and the underlying unobserved functional process, and $\boldsymbol{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_P)$ represents a diagonal covariance matrix to allow for heteroscedastic measurement noise (Ramsay and Silverman, 2005).

We estimate the unobserved absorbance $\boldsymbol{X}_i(\boldsymbol{t})$ from the vector of observations $\boldsymbol{O}_i(\boldsymbol{t})$. Using the basis representation of $X_i(t)$ in (1), the measurement error model in (2) reduces to a linear regression model, $\mathbb{E}[\boldsymbol{O}_i(\boldsymbol{t})] = \boldsymbol{B}\boldsymbol{\nu}$, where $\boldsymbol{B} = [\boldsymbol{b}(t_1), \ldots, \boldsymbol{b}(t_P)]^T$ is the $P \times K$ basis function design matrix and $\boldsymbol{\nu}$ is the basis coefficient vector which can be estimated via least squares.

To correct for heteroscedastic measurement noise we introduce a $P \times P$ weight matrix $\boldsymbol{W} = \boldsymbol{\Sigma}^{-1}$. Additionally, we penalise discontinuous jumps in consecutive basis coefficient values through a $K \times K$ regularisation matrix $\boldsymbol{\Omega}$, and estimate ν by solving

$$\underset{\boldsymbol{\nu}_i \in \mathbb{R}^K}{\arg\min} \left\{ \left\| \boldsymbol{W}^{1/2}(\boldsymbol{O}_i - \boldsymbol{B}\boldsymbol{\nu}_i) \right\|^2 + \lambda \boldsymbol{\nu}_i^T \boldsymbol{\Omega} \boldsymbol{\nu}_i \right\}, \tag{3}$$

where $|| \cdot ||$ denotes the L_2 norm. The $K \times K$ penalty matrix $\boldsymbol{\Omega}$ has elements (k, l) equal to $\int_{\mathcal{T}} b_k''(t) b_l''(t) dt$, such that $\boldsymbol{\nu}_i^T \boldsymbol{\Omega} \boldsymbol{\nu}_i$ approximates the curvature of $X_i(t)$ as measured by the integrated squared second derivative, $\int_{\mathcal{T}} [X_i''(t)]^2 dt$ (Cardot et al., 2003; Eubank, 1999; Green and Silverman, 1994; Marx and Eilers, 1999; Ramsay and Silverman, 2005). The criterion (3) therefore enforces smoothness by penalising roughness in the least-squares estimate of $X_i(t)$. The penalty parameter $\lambda \geq 0$ regulates the degree of smoothness and is optimally chosen using cross-validation (Wahba, 1990).

2.2.3 Statistical Model

Model specification. The exponential family of statistical models can be extended to the case of functional data, providing a comprehensible and unified framework to tackle regression and classification tasks (Cardot and Sarda, 2005; Goldsmith et al., 2011; James, 2002; Müller and Stadtmüller, 2005). These models can be written as follows:

$$y_i = g^{-1}(\eta_i) \quad \text{where} \quad \eta_i = \alpha + \int_{\mathcal{T}} X_i(t)\beta(t)dt + \sum_{s=1}^{S} \gamma_s U_{is}, \tag{4}$$

where α is a constant intercept; $X_i(t)$ is the spectrum of mosquito i (represented as a function); $\beta(t)$ is a functional slope coefficient giving the influence of different wavelength regions on the response; y_i is a scalar response with distribution belonging to the exponential family; U_{is} is a real-valued non-functional predictor; and γ_s is the corresponding slope coefficient.

The invertible link function g relates the subject-specific mean response μ_i to the linear predictor η_i as follows: $g(\mathrm{E}[y_i|X_i(\boldsymbol{t})]) = \eta_i$, or, equivalently, $\mu_i = \mathrm{E}[y_i|X_i(\boldsymbol{t})] = g^{-1}(\eta_i)$. The functional form of g depends on the distribution of the response, and determines the type of statistical model, as follows:

I. REGRESSION: when the response is real-valued and assumed to follow a Gaussian distribution, the link function is just the identity, that is $g(\mu_i) = \mu_i$, and so $\mu_i = \eta_i$, leading to the functional linear model ("*functional LM*"). This is the case when determining mosquito age.

II. CLASSIFICATION: when the response is binary and assumed to follow a Bernoulli distribution, the link function is equal to $g(\mu_i) = \log(\mu_i/(1-\mu_i))$, and so $\mu_i = 1/(1+e^{-\eta_i})$, leading to the logistic-link functional generalised linear model ("*functional GLM*"). This is the case when determining mosquito species or infection.

In some applications, we may also be interested in the multi-class classification problem, for instance when information on the severity of infection is available (e.g. low, medium and high levels) or differentiating between more than two mosquito species. This corresponds to a multinomial distribution for the response, which can be reduced to a set of binary functional logistic models and therefore tackled within this same framework (McCullagh and Nelder, 1989).

Basis functions for $\beta(t)$. The coefficient function $\beta(t)$ is modelled as a smooth function (like the spectra themselves) using cubic B-splines,

$$\beta(t) = \sum_{k=1}^{K} \zeta_k b_k(t) = \boldsymbol{\zeta}^T \boldsymbol{b}(t), \tag{5}$$

where $\boldsymbol{b}(t)$ is the basis function vector defined as in (1); and $\boldsymbol{\zeta} = [\zeta_1, \dots, \zeta_K]^T$ is a basis coefficient vector. The functional term in (4) then becomes:

$$\int_{\mathcal{T}} X_i(t)\beta(t)dt = \sum_{k=1}^{K} \zeta_k \int_{\mathcal{T}} X_i(t) b_k(t)dt \approx \sum_{k=1}^{K} \zeta_k \bar{x}_{ik}, \tag{6}$$

where $\bar{x}_{ik} = \sum_{j=1}^{P} x_{ij} b_k(t_j)$ is a discretised approximation to $\int_{\mathcal{T}} X_i(t) b_k(t)dt$. In this way, the functional model can be reduced to a multivariate model, for which estimation and inference procedures are well known. Notice that despite this discretisation, the functional representation of $\beta(t)$ is easily recovered from (5), given estimates $\hat{\boldsymbol{\zeta}}$ of $\boldsymbol{\zeta}$.

Model estimation. We assume independent and identically distributed pairs of observations $\{(X_i(\boldsymbol{t}), y_i)\}_{i \in \{1:N\}}$. Let $\boldsymbol{y} = [y_1, \dots, y_N]^T$ denote the responses vector of length N; and let $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_N]^T$ denote the functional predictor design matrix of size $N \times P$, where $\boldsymbol{x}_i = [x_{i1}, \dots, x_{iP}]$ and $x_{ij} = X_i(t_j)$ for all $i \in \{1:N\}$ and $j \in \{1:P\}$. The design matrix $\boldsymbol{X}$ can denote the raw observations of the functional predictor or the corresponding smoothed version as in §2.2.2. Also, let $\boldsymbol{B} = [\boldsymbol{b}(t_1), \dots, \boldsymbol{b}(t_P)]^T$ denote a $P \times K$ basis function design matrix, with $\boldsymbol{b}(t_j) = [b_1(t_j), \dots, b_K(t_j)]$ for all $j \in \{1:P\}$. Finally, let $\boldsymbol{Z} = [\boldsymbol{z}_1, \dots, \boldsymbol{z}_N]^T$ denote the non-functional predictor design matrix of size $N \times S$, where $\boldsymbol{z}_i = [z_{i1}, \dots, z_{iS}]$ for all $i \in \{1:N\}$.

Estimation. Projecting the coefficient function onto the space spanned by the K B-splines, as defined in (5), leads to a model with design matrix $\boldsymbol{XB}$ instead of $\boldsymbol{X}$. Here we consider penalised likelihood estimation. In the linear case, the least squares solution gives the maximum likelihood estimator:

$$\underset{\alpha \in \mathbb{R}, \boldsymbol{\zeta} \in \mathbb{R}^K, \gamma \in \mathbb{R}^S}{\arg\min} \left\{ ||\boldsymbol{y} - \alpha \mathbf{1} - \boldsymbol{XB\zeta} - \boldsymbol{Z}\gamma||^2 + \lambda \boldsymbol{\zeta}^T \boldsymbol{\Omega} \boldsymbol{\zeta} \right\} \tag{7}$$

where λ and $\boldsymbol{\Omega}$ are as in (3) and the term $\boldsymbol{\zeta}^T \boldsymbol{\Omega} \boldsymbol{\zeta}$ gives the curvature of the projected coefficient function $\dot{\boldsymbol{\zeta}} = \boldsymbol{B\zeta}$, which approximates the curvature of the original coefficient function $\beta(t)$ as measured by the integrated squared second derivative, $\int_{\mathcal{T}} [\beta''(t)]^2 dt$ (Cardot et al., 2003; Marx and Eilers, 1999; Reiss and Ogden, 2007). In the generalised linear case, the squared norm term in (7) is replaced with the negative of the model likelihood and the resulting penalised likelihood criterion is optimised (Gertheiss et al., 2013).

The problem (7) will typically be ill-posed as a result of the high dimensional nature of spectral data and the small sample sizes usually available ($N \ll P$). Variable selection and/or dimension reduction techniques provide a solution which we explore below.

2.2.4 Dimension reduction and feature selection

The projection of $\beta(t)$ onto a B-spline system defined in (5) can provide some dimension reduction when $K < P$. However, this comes at a cost of loss of information which can lead to poor predictive performance. Here we assume a rich basis system, capable of representing spectra without any considerable loss of information. In practice, this means that the design matrix $\boldsymbol{XB}$ may still present an ill-posed problem $(K > N)$ and further dimension reduction is then required.

We consider projecting the coefficient function onto $\boldsymbol{D}$ after the projection onto $\boldsymbol{B}$. The dimension reduction projection matrix $\boldsymbol{D}$, with dimension $Q < K$, is derived from the data such that $\boldsymbol{XBD}$ captures the essential features of $\boldsymbol{XB}$ (and therefore of $\boldsymbol{X}$), but in a lower dimensional space. Below we give details on two methods to compute $\boldsymbol{D}$, namely functional principal component analysis and functional partial least squares. For further comparisons of the two methods see for instance Febrero-Bande et al. (2017); Frank and Friedman (1993); Reiss and Ogden (2007).

These two projection-based dimension reduction approaches can be viewed as plug-in methods since $\boldsymbol{D}$ is independent of the parameters in the statistical model (details are given in Appendix A). Parameter estimation follows essentially the same approach as in (7), except that we now use the reduced spectral data $\boldsymbol{XBD}$ instead of $\boldsymbol{XB}$:

$$\underset{\alpha\in\mathbb{R},\boldsymbol{\omega}\in\mathbb{R}^Q,\boldsymbol{\gamma}\in\mathbb{R}^S}{\arg\min} \left\{ ||\boldsymbol{y} - \alpha\boldsymbol{1} - \boldsymbol{XBD\omega} - \boldsymbol{Z\gamma}||^2 + \lambda\boldsymbol{\omega}^T\boldsymbol{D}^T\boldsymbol{\Omega}\boldsymbol{D\omega} \right\} \tag{8}$$

where λ and $\boldsymbol{\Omega}$ are as in (3) and the term $\boldsymbol{\omega}^T\boldsymbol{D}^T\boldsymbol{\Omega}\boldsymbol{D\omega}$ gives the curvature of projected coefficient function $\dot{\boldsymbol{\omega}} = \boldsymbol{BD\omega}$. Provided that $Q + S < N$, the problem is well-posed and a solution can be found via either penalised least squares or penalised likelihood estimation as before, depending on the distribution of the response.

Importantly, note that from the estimation procedure in (8) it is possible to recover the coefficient function $\beta(t)$ by first computing $\hat{\boldsymbol{\zeta}} = \boldsymbol{D}\hat{\boldsymbol{\omega}}$ and then plugging this into (5). This tells us which regions of the spectra—in the original P-dimensional space—are more important.

2.2.5 Cross-validation

We use a two-stage cross-validation procedure which explicitly enforces a smooth coefficient function $\beta(t)$. Performance is measured by RMSD for the functional LM and by AUC for the functional GLM.

Datasets. To show the full potential of the techniques proposed—functional representation, smoothing, penalisation—we use two datasets. The first, called the *cross-validation dataset*, is split into training, validating and testing subsets, respectively used to train the models, cross-validate model parameters as detailed below, and estimate the generalisation error. The second, called the *alternative dataset*, is composed of a testing set which is used to evaluate the quality of predictions on slightly different samples using the model trained with the cross-validation dataset. For the application presented in this paper, this means samples collected from different regions (see §3 for more details).

Choosing K. In the first stage of cross-validation, the number of basis functions K used to represent the spectra is chosen to maximise the performance in a model without penalisation. To do this, K is decreased from P until the loss in accuracy exceeds the threshold τ_K. This guarantees that there is virtually no information loss while at the same time giving the most efficient representation of the spectra. The technique provides non only an efficient representation but also a small degree of smoothing. We use $\tau_K = 0.01$ in the following.

Choosing Q and λ. In the second stage of cross-validation, the number of PCA/PLS components Q and the penalty parameter λ are chosen jointly to give the smoothest coefficient function whose predictive performance is within a margin $\tau_{\lambda,Q}$ of the predictive performance of the optimal non-penalised model, which we denote by $a_\star$. That is, we compute models with different combinations of parameters (λ, Q) and from those with acceptable performance we select the one having the smoothest coefficient function $\beta(t)$, as measured by the integrated squared second derivative, $R_\beta = \int_\mathcal{T}[\beta''(t)]^2 dt$, where larger values correspond to rougher coefficient functions. Acceptable performance is defined here as an RMSD between $[a_\star, a_\star + \tau_{\lambda,Q}]$ in the case of the functional LM, or an AUC between $[a_\star - \tau_{\lambda,Q}, a_\star]$ in the case of the functional GLM. We use $\tau_{\lambda,Q} = 0.01$ in the following.

Ensemble models. We also test the performance of ensemble models where the error rate is averaged over a set of models, chosen as follows: first, we select the top n_a models that perform within a margin $\tau_{\lambda,Q}$ of the optimal non-penalised model (similarly to the procedure used to choose λ and Q); and from this set of acceptable models we select the smoothest n_e models. We use $n_\mathrm{a} = 25$ and $n_\mathrm{e} = 5$ in the following.

Cross-validation details. We average the cross-validation results over 100 randomisation of the data subsets to reduce the effect of sampling error. The proportions of observations used in each subset of the cross-validation dataset were: 50% for training, 25% for validating, and 25% for testing. We use as the benchmark the Generalised Linear Model (`GLM` in Table 2).

3 Results

We compare the performance of 16 different models, arising from the use of the two different dimension reduction methods (PLS and PCA) and the use, or not, of the FDA techniques presented. We will show results for a classification task, thus all models will be generalised linear models (`GLM`), prefixed as follows: `f` (e.g., `fGLM`) when making use of the functional representation in (1); `s` (e.g., `sGLM`) when making use of spectra smoothing as in (3); `p` (e.g., `pGLM`) when making use of penalisation for the coefficient function estimation as in (8). Additionally we evaluate ensembles of the smoothest models that use penalisation.

3.1 Improving generalisation

The techniques used—spectra smoothing, functional representation and penalised estimation of the coefficient function—all improve the AUC and test error on the testing subset of the cross-validation dataset, if only slightly, with the exception of penalisation-only (`pGLM`) with PLS

Table 2: Performance of different GLM models and different feature selection methods (details in §2.2) for determining mosquito species (*An. arabiensis* vs. *An. gambiae s.s.*). Measures given are: number of basis functions as a fraction of the number of predictors/wavelengths (K/P), number of features (Q), roughness of the coefficient function (R_β), penalty parameter (λ), area under the ROC curve (AUC; 0–100), and cross-validation/alternative testing set errors ($ERR_{\text{cv}}/ERR_{\text{alt}}$, % misclassification rate with standard deviation in parenthesis). We show two sets of models: best-performing (highest AUC) and smoothest (lowest R_β) as determined by their performance on the cross-validation dataset.

	PLS												
		best-performing						**smoothest**					
	K/P	Q	R_β	λ	AUC	ERR_{cv}	ERR_{alt}	Q	R_β	λ	AUC	ERR_{cv}	ERR_{alt}
GLM	1	2	11.62	–	72	34 (.08)	32 (.04)	–	–	–	–	–	–
sGLM	1	8	0.04	–	75	30 (.07)	30 (.05)	–	–	–	–	–	–
pGLM	1	37	0.20	8.9	65	53 (.07)	62 (.25)	32	0.07	6.8	75	44 (.11)	38 (.24)
spGLM	1	19	0.001	61.1	76	30 (.08)	24 (.05)	20	0.001	94.4	76	30 (.08)	25 (.05)
fGLM	0.1	8	2.00	–	75	29 (.07)	30 (.06)	–	–	–	–	–	–
fsGLM	0.1	8	2.11	–	75	30 (.07)	30 (.06)	–	–	–	–	–	–
fpGLM	0.1	22	0.09	58.3	75	32 (.08)	24 (.06)	23	0.09	58.3	75	32 (.09)	24 (.06)
fspGLM	0.1	23	0.08	102	75	32 (.09)	24 (.05)	23	0.07	134	75	32 (.08)	24 (.07)
	PCA												
		best-performing						**smoothest**					
	K/P	Q	R_β	λ	AUC	ERR_{cv}	ERR_{alt}	Q	R_β	λ	AUC	ERR_{cv}	ERR_{alt}
GLM	1	2	35.07	–	71	36 (.08)	30 (.05)	–	–	–	–	–	–
sGLM	1	10	0.18	–	75	30 (.08)	29 (.06)	–	–	–	–	–	–
pGLM	1	48	0.02	3.9	73	34 (.08)	29 (.04)	63	0.02	3.2	74	33 (.08)	27 (.04)
spGLM	1	31	0.001	53.4	76	29 (.08)	26 (.04)	27	0.001	53.4	77	29 (.08)	26 (.04)
fGLM	0.1	10	3.30	–	75	30 (.08)	30 (.05)	–	–	–	–	–	–
fsGLM	0.1	10	1.38	–	75	30 (.08)	30 (.06)	–	–	–	–	–	–
fpGLM	0.1	30	0.09	7.8	75	32 (.09)	25 (.04)	22	0.10	9.3	75	32 (.08)	26 (.04)
fspGLM	0.1	30	0.09	7.8	75	32 (.08)	24 (.04)	25	0.10	6.2	75	31 (.08)	25 (.04)

reduction which introduces considerable variation in the estimates as shown by the standard error of the error rate (Table 2). More importantly, however, substantial improvement of the error rate can be observed on alternative testing set, from 32/30% for the benchmark model (GLM) to 24/25% for the model using all three techniques (fspGLM) with PLS/PCA reduction.

The fpGLM and fspGLM are the best performing models with very similar error rates on the alternative test set, showing that smoothing becomes only marginally important when a functional representation is used. This is not surprising seeing that a functional representation provides smoothing alongside dimension reduction. It is worth noting that a functional representation provides marginally better results then smoothing when only one of the techniques is used in conjunction with penalisation, which can be seen by comparing spGLM and fpGLM.

The relationship between smoothness and performance is as expected. Specifically, the smoothest models—here defined by a low value of R_β, the roughness of the coefficient function—tend to perform better on the alternative testing set than rougher models.

The ensemble approach does not improve results w.r.t. the corresponding smoothest models with the exception of the pGLM with PLS reduction, where both error rate and standard deviation are improved substantially (Table 3). However, this still does not constitute an improvement w.r.t.

Table 3: Performance of different penalised ensemble models. See Table 2 for a description of the measures reported.

		PLS						PCA					
	K/P	Q	R_β	λ	AUC	ERR_{cv}	ERR_{alt}	Q	R_β	λ	AUC	ERR_{cv}	ERR_{alt}
`pGLM`	1	31	0.09	4.5	75	41 (.11)	33 (.18)	60	0.02	3.3	74	33 (.08)	27 (.04)
`spGLM`	1	20	0.001	87.8	76	30 (.08)	25 (.05)	28	0.001	53.4	77	29 (.08)	26 (.04)
`fpGLM`	0.1	22	0.09	51.7	75	32 (.08)	24 (.06)	24	0.11	6.5	75	31 (.08)	26 (.04)
`fspGLM`	0.1	23	0.08	105	75	32 (.09)	24 (.05)	23	0.10	8.1	75	31 (.08)	25 (.04)

the smoothest `GLM` (the benchmark). Additionally, the ensemble `pGLM` does not outperform `sGLM`, `fGLM` or `fsGLM`, suggesting that smoothing spectra is essential to avoid overfitting.

The optimal (i.e. lossless) functional representation affords a 90% compression level, which reduces the computational costs of computing the dimension reduction matrix $\boldsymbol{D}$. The resulting optimal number of PLS components Q increases although only marginally, for instance from 20 in the `spGLM` to 23 in the `fspGLM`. For the model using PCA components this number decreases from 27 to 25.

Importantly, the standard deviation of error rate on the alternative test set is left virtually unchanged or only minimally increased as a result of the three techniques used, again with the exception of the `pGLM` with PLS reduction.

In general, PLS gives slightly smaller error rates than PCA on the alternative testing set and requires a smaller number of components, thus performing better in both accuracy and efficiency.

The AUC is a less important performance measure since it is computed with the testing subset of the cross-validation dataset, while we are primarily interested in the performance of the model on the alternative testing set. Nonetheless, we also see an improvement between 3–4 p.p. in the AUC with the use of functional techniques.

3.2 Visualisation and Diagnostics

Cross-validation illustrated. The results of (λ, Q) cross-validation evaluated on the testing subset of the cross-validation dataset are illustrated in Figure 3. Accuracy is measured by the AUC and the models with acceptable performance, as defined in §2.2.5, are shown within the boxed area (Figure 3a). From the set of acceptable models, the smoothest or an ensemble are picked to produce the predictions on the alternative dataset, reported in Table 2.

Diagnostics and results illustrated. We provide extensively annotated plots for quick assessment of model fit and prediction accuracy. These are designed to be as informative as possible so that potential issues with the training process can be easily identified. An example is shown in Figure 4 for a binary classification problem. Multinomial classification and regression problems have similar outputs.

The Q cross-validation curves evaluated on the validating subset of the cross-validation dataset shows the performance achieved for a given number of PLS/PCA components (Figure 4a). When

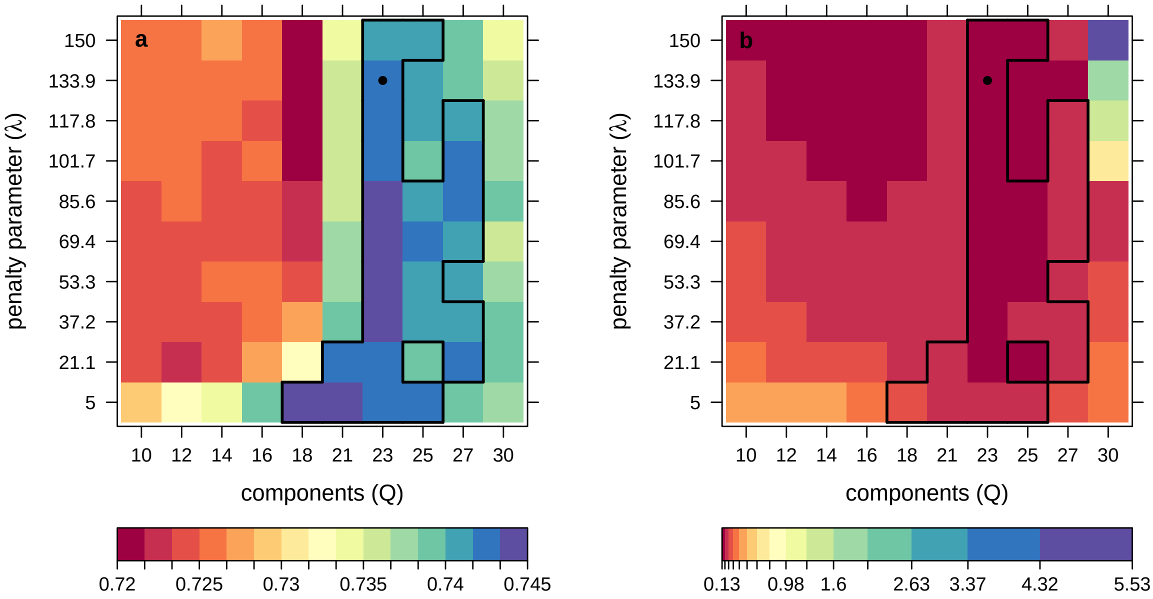

Figure 3: Maps showing (**a**) Area under the ROC curve (AUC) and (**b**) roughness of the coefficient function (R_β), for `fspGLM` with different numbers of PLS components and penalty parameter, on the validating subset of the cross-validation dataset. The boxed area denotes the n_a models with acceptable performance, and the black circle denotes the smoothest model among those.

these curves are close to flat, it is useful to set a margin parameter $\tau_{Q_{\text{opt}}}$ so that we choose not the Q corresponding to the largest AUC but one that gives an AUC within $\tau_{Q_{\text{opt}}}$ of the optimal. A small value such as $\tau_{Q_{\text{opt}}} = 0.01$ can often reduce the number of components substantially without considerable loss in performance, and furthermore can help prevent overfitting.

The classification cutoff–error curves depict the optimal probability cutoff that minimises the misclassification rate, giving equal weight to false positives and false negatives (Figure 4b). Models with optimal probability cutoff equal to 0 or 1 will always predict the same class, which can mean that a particular split of the dataset has come out highly imbalanced in terms of the responses classes and the model minimises the error by always predicting the majority frequency. Such models may deserve further investigation. The package also includes an option to enforce balanced dataset splits. An important quality requirement for any classifier using predictors (C_p) is that it outperforms a naive classifier (C_N) trained only on the frequency of the response variable (i.e., without predictors). The later will always predict the majority class, and so if our classifier outperforms this naive strategy we can be assured that the predictors contain valuable information. In other words, if the spectra is predictive of mosquito species, C_p must outperform C_n. The color-coded dot in Figure 4b provides this information: green if C_p outperforms the C_n; red otherwise. To fairly access performance we compute the misclassification rate using the average probability cutoff, although we also show the curve-specific cutoff for reference.

The ROC curves evaluated on the testing subset of the cross-validation dataset, together with dispersion measures and the AUC corresponding to the optimal classification cutoff as given in panel 4b give an overview of the performance of the model (Figure 4c).

The coefficient function $\beta(t)$ can be used to identify the spectral regions of more importance for prediction, and is the key output of the model (Figure 4d).

A histogram of the estimated linear predictor for the test observations illustrates the models' ability to separate the two classes (Figure 4e). The shaded area corresponds to misclassified observations, with false negatives to the left of the optimal cutoff line (*An. gambiae* incorrectly predicted to be *An. arabiensis*) and false positives to the right (*An. arabiensis* incorrectly predicted to be *An. gambiae*). The inset plot shows the confusion matrix with the breakdown of the classification results: true negative rate (tnr), false negative rate (fnr), true positive rate (tpr), false positive rate (fpr).

3.3 Identifying important predictors

Experiments to generate spectra from laboratory-reared mosquitoes remain relatively time-consuming, although necessary to train predictive models. With the view of simplifying and accelerating the process of gathering data, it is of interest to determine which variables under the experimenter's control have an effect on spectra. The statistical framework presented, encapsulated in (8), can handle this type of hypothesis testing straightforwardly by testing the significance of the parameters $\boldsymbol{\gamma}$ associated with the non-functional predictors $\boldsymbol{Z}$.

To to illustrate this using the same dataset, we can determine whether the location of collection is a statistically significant. We use a binary variable `Location` encoding the location where samples were collected (Longo or Klesso). The average p-value for this variable in a `fsGLM` with balanced classes ($N = 222$) is $p = 0.04$, which provides some evidence that mosquitoes from the two collection locations have some differences which are not captured by the spectra. This result supports the use of penalised estimation and smoothing methods to prevent overfitting when doing cross-location prediction.

4 Discussion

NIRS has the potential to revolutionise entomological monitoring of mosquito-borne diseases though there is a need to refine the statistical methods used to translate spectral information into quantities of epidemiological interest. Spectra from mosquitoes with the same characteristics are also likely to vary from site to site reflecting the genetic heterogeneity in the mosquito population, local environmental factors and procedural differences between teams collecting and processing samples. If the NIRS is to become a widely used there is therefore a need to prevent statistical models converting spectra into mosquito characteristics to be generalizable and not overfitting to the local training dataset. Here we have identified a number of statistical techniques to support this process which should be adopted to increase the rigour of NIRS entomological monitoring. Spectra functional representation, spectra smoothing and penalisation for the coefficient function all improve the accuracy of NIRS models predicting mosquito species in the test dataset (independent mosquitoes collected from the same location) and more importantly on the alternative test dataset (mosquitoes collected 283km away). All the techniques provide a level of spectra smoothing, though the optimum use of these different methods (in combination or individually) will vary depending on the characteristics of the training and unknown dataset.

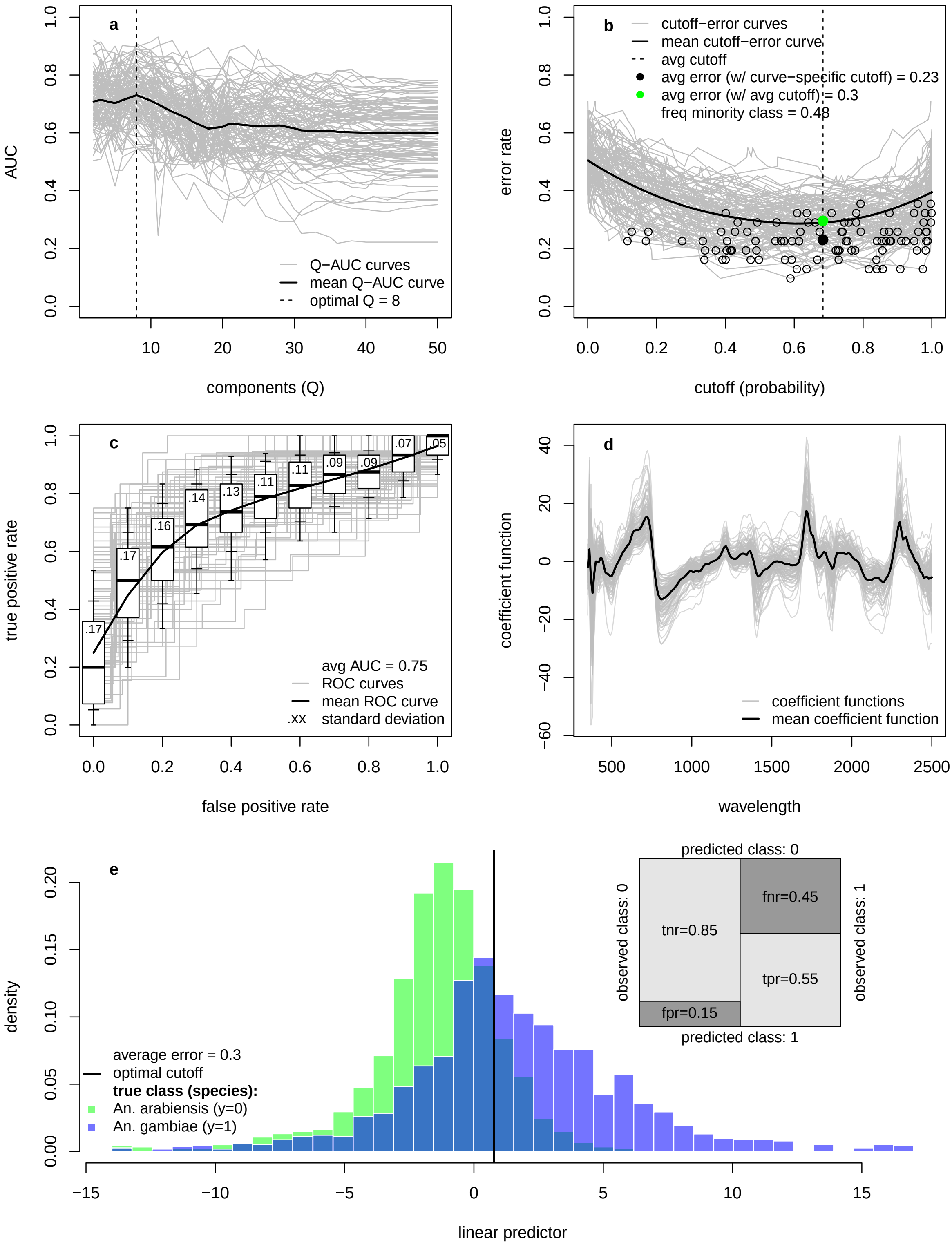

Figure 4: Diagnostic plots for a `fsGLM` with PLS components, showing (**a**) cross-validation for the number of components (Q); (**b**) cutoff-error plot displaying the choice of probability cutoff for classification; (**c**) ROC curves with variability shown in the boxplots (black line, box edges, inner and outer whiskers show 50th (median), 25th/75th, 15th/85th and 5th/95th percentiles, respectively); (**d**) coefficient function $\beta(t)$; (**e**) histogram of the estimated linear predictor for the test observations, colour-coded by the true class. Results are averaged over 100 randomisations of the training/validating/testing subsets, shown individually in grey.

Smoother coefficient functions tend to give conservative estimates and therefore prevent overfitting to individual peaks which may result from measurement error generated by the machinery or represent a minor deviation specific to the local mosquito population. We would recommend all these methods are trialled with new and expanded datasets and the optimum model chosen using the methods outlined here.

The regularisation framework proposed here has several advantages over the standard methods used in the literature. The functional representation of spectra is more computationally efficient allowing models to be trained and fit quicker. Though this isn't necessarily an issue with the dataset presented here (with single models being fit within a few minutes) this is likely to get more important as datasets grow and samples from multiple sites are used within the same model. Regularisation also provides a smoother coefficient functions which generalises better preventing overfitting to noisy spectra. This is particularly important when sample sizes are small and where instruments have high noise-to-signal ratio in some regions of the spectrum (for example at the ends their spectral ranges).

In these data both PCA and PLS as a method of dimension reduction tend to give similar results, but in some cases PLS requires fewer components than PCA to achieve a given accuracy as has been seen previously (de Jong, 1993a). In addition to the standard penalisation approach, the methods presented here also enable predictions to be made using either the smoothest or top 5 smoothest models selected from the best performing models. Here they were selected for by choosing the models with either the smoothest or the 5 smoothest coefficient functions which were drawn from the top 25 most accurate models as evaluated on the validating subset of the cross-validation dataset (from mosquitoes within the same village). In these data this did not substantially improve the accuracy when predicted the species of the second village. This confirms the robustness of the strategy employed here of selecting the most appropriate smoothing method in order to obtain good generalizability. Further work with larger more diverse datasets are needed to understand the benefit of selecting the smoothest over the best fitting models. Here we defined an acceptably accurate model as being within 1 percentage point of the most accurate model, although this parameter will need to be refined according to the question under investigation which will determine the trade-off between accuracy and generalisability. Similarly, the ensemble method proposed here which selected the 5 smoothest models from the top 25 most accurate did not perform better than the single best fit model. The added benefit of this ensemble approach needs to be investigated further using larger datasets collected from more diverse geographical locations as it may be expected to perform better in these scenarios.

Our results indicate that the accuracy of NIRS ability to determine the sibling species of mosquito within the An. gambiae complex is lower than previous estimates. This work was intended to showcase the different statistical methods and not evaluate the technique and there are a number of reasons why the moderate accuracy should not be overly interpreted. Firstly, the sample size used in this study is very small with only 126 samples available. This means that only 63 samples were used to train each model (a different set of 63 samples for each of the 100 models), which is a very low number given the diversity of spectra. Future work may have sample sizes an order or two of magnitude larger if the technique is adopted further. Secondly, there were a mixture of F0 and F1 mosquitoes used in this analysis. Spectra collected from mosquitoes which were caught in different ways may vary, though the number of mosquitoes available for this analysis was insufficient to test this here. Lastly, other characteristics of interest which might influence spectra were not collected or included in the model. For example the level of insecticide resistance may vary substantially within the same species within the same population and has been shown to influence specta ( ). It is worth noting however that high accuracy isn't necessarily a prerequisite for NIRS to be a useful tool (Lambert et al., 2018). NIRS could also be used as a pre-scanning tool, for instance to determine if mosquitoes are infected before parasite genetic sequencing. In that case, we are interested in maximising the ratio of truly infected to truly uninfected TP/(TP +FP) in order to minimise the cost per mosquito sequenced. This can be done with ROCR package by choosing the ppv (positive predictive value) criterion in the function performance(). Additionally, there can be class imbalance which leads to imbalanced misclassification rates. Tuning the importance of false positive rates to false negative rates can help giving balanced misclassification rates for the different classes according to the question under investigation.

Acknowledgements

The work was supported by UK Medical Research Council (MRC) Project Grant (MR/P01111X/1). We acknowledge joint Centre funding from the UK Medical Research Council and Department for International Development (MR/R015600/1).

References

Aguilera, A. M., Escabias, M., Preda, C., and Saporta, G. (2010). Using basis expansions for estimating functional PLS regression: applications with chemometric data. *Chemometrics and Intelligent Laboratory Systems*, 104(2):289–305.

ASD Inc. (company) (2020). ASD Inc., Boulder, Colorado, USA. https://www.asdi.com.

Bhatt, S., Weiss, D. J., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., Battle, K. E., Moyes, C. L., Henry, A., Eckhoff, P. A., Wenger, E. A., Briët, O., Penny, M. A., Smith, T. A., Bennett, A., Yukich, J., Eisele, T. P., Griffin, J. T., Fergus, C. A., Lynch, M., Lindgren, F., Cohen, J. M., Murray, C. L. J., Smith, D. L., Hay, S. I., Cibulskis, R. E., and Gething, P. W. (2015). The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. *Nature*, 526(7572):207–211.

Cai, T. T. and Hall, P. (2006). Prediction in functional linear regression. *The Annals of Statistics*, 34(5):2159–2179.

Cardot, H., Ferraty, F., and Sarda, P. (1999). Functional linear model. *Statistics and Probability Letters*, 45:11–22.

Cardot, H., Ferraty, F., and Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica*, 13:571–591.

Cardot, H. and Sarda, P. (2005). Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis*, 92(1):24–41.

de Boor, C. (2001). *A practical guide to splines*. Springer.

de Jong, S. (1993a). PLS fits closer than PCR. *Journal of Chemometrics*, 7(6):551–557.

de Jong, S. (1993b). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3):251–263.

Delaigle, A. and Hall, P. (2012). Methodology and theory for partial least squares applied to functional data. *Annals of Statistics*, 40(1):322–352.

Esperança, P. M. (2019a). Experiments for determination of species of field-collected lab-rared mosquitoes using near-infrared spectroscopy (A1T1) [Data set]. *Zenodo*. doi.org/10.5281/zenodo.2557559.

Esperança, P. M. (2019b). ML-based Epidemiological Vector Control Monitoring using FDA techniques for NIRS data [Software]. *GitHub*. github.com/pmesperanca/mlevcm.

Esperança, P. M., Blagborough, A. M., Da, D. F., Dowell, F. E., and Churcher, T. S. (2018). Detection of *Plasmodium berghei* infected *Anopheles stephensi* using near-infrared spectroscopy. *Parasites & Vectors*, 11:377.

Eubank, R. L. (1999). *Nonparametric regression and spline smoothing*. Marcel Dekker, 2nd edition.

Fanello, C., Santolamazza, F., and Della Torre, A. (2002). Simultaneous identification of species and molecular forms of the anopheles gambiae complex by PCR-RFLP. *Medical and veterinary entomology*, 16(4):461–464.

Febrero-Bande, M., Galeano, P., and González-Manteiga, W. (2017). Functional principal component regression and functional partial least-squares regression: an overview and a comparative study. *International Statistical Review*, 85(1):61–83.

Frank, l. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Geladi, P. and Kowlaski, B. (1986). Partial least-squares regression: a tutorial. *Analytica Chemica Acta*, 185:1–17.

Gerlach, R. W., Kowalski, B. R., and Wold, H. O. (1979). Partial least-squares path modelling with latent variables. *Analytica Chimica Acta*, 112(4):417–421.

Gertheiss, J., Maity, A., and Staicu, A.-M. (2013). Variable selection in generalized functional linear models. *Stat*, 2(1):86–101.

Goldsmith, J., Bobb, J., Crainiceanu, C. M., Caffo, B., and Reich, D. (2011). Penalized functional regression. *Journal of Computational and Graphical Statistics*, 20(4):830–851.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach*. Chapman & Hall/CRC.

Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):109–126.

James, G. M. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):411–432.

Jolliffe, I. T. (2002). *Principal component analysis*. Springer, 2nd edition.

Lambert, B., Sikulu-Lord, M. T., Mayagaya, V. S., Devine, G., Dowell, F., and Churcher, T. S. (2018). Monitoring the age of mosquito populations using near-infrared spectroscopy. *Scientific Reports*, 8:5274.

Marx, B. D. and Eilers, P. H. C. (1999). Generalized linear regression on sampled signals and curves: A P-spline approach. *Technometrics*, 41(1):1–13.

Massy, W. F. (1965). Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 60(309):234–256.

Mayagaya, V. S., Michel, K., Benedict, M. Q., Killeen, G. F., Wirtz, R. A., Ferguson, H. M., and Dowell, F. E. (2009). Non-destructive determination of age and species of *Anopheles gambiae* s.l. using near-infrared spectroscopy. *American Journal of Tropical Medicine and Hygiene*, 81(4):622–630.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall/CRC, 2nd edition.

Mevik, B. and Wehrens, R. (2015). The pls package: principal component and partial least squares regression in R. *Journal of Statistical Software*, 18(2):1–23.

Morris, J. S. (2015). Functional regression. *Annual Review of Statistics and Its Application*, 2:321–359.

Müller, H.-G. and Stadtmüller, U. (2005). Generalized functional linear models. *Annals of Statistics*, 33(2):774–805.

Ong, O. T., Kho, E. A., Esperança, P. M., Freebairn, C., Dowell, F. E., Devine, G. J., and Churcher, T. S. (2020). Ability of near-infrared spectroscopy and chemometrics to predict the age of mosquitoes reared under different conditions. *Parasites & vectors*, 13:160.

Preda, C. and Saporta, G. (2005). PLS regression on a stochastic process. *Computational Statistics and Data Analysis*, 48(1):149–158.

Ramsay, J. O. and Silverman, B. W. (2002). *Applied functional data analysis. methods and case studies*. Springer-Verlag, 2nd edition.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis*. Springer, 2nd edition.

Reiss, P. T., Goldsmith, J., Shang, H. L., and Ogden, R. T. (2017). Methods for scalar-on-function regression. *International Statistical Review*, 85(2):228–249.

Reiss, P. T. and Ogden, R. T. (2007). Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*, 102(479):984–996.

Sachs, J. and Malaney, P. (2002). The economic and social burden of malaria. *Nature*, 415(6872):680–685.

Shang, H. L. (2014). A survey of functional principal component analysis. *AStA Advances in Statistical Analysis*, 98(2):121–142.

Sikulu, M. T., Killeen, G. F., Hugo, L. E., Ryan, P. A., Dowell, K. M., Wirtz, R. A., Moore, S. J., and Dowell, F. E. (2010). Near-infrared spectroscopy as a complementary age grading and species identification tool for African malaria vectors. *Parasites & Vectors*, 3:49.

Sikulu-Lord, M. T., Milali, M. P., Henry, M., Wirtz, R. A., Hugo, L. E., Dowell, F. E., and Devine, G. J. (2016). 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*, 10(10):e0005040.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 47(1):1–52.

Wahba, G. (1990). *Spline models for observational data*. Society for Industrial and Applied Mathematics.

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295.

WHO (2019). *World malaria report 2019*. World Health Organization.

Wold, S., Ruhe, A., Wold, H., and Dunn III, W. J. (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3):735–743.

Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130.

Appendix

A Dimension reduction

Functional principal component analysis (fPCA). Principal component analysis was originally used to overcome multicollinearity in the linear model (Jolliffe, 2002; Massy, 1965) and subsequently extended to functional data (Cai and Hall, 2006; Cardot et al., 1999; Hall and Hosseini-Nasab, 2006; Müller and Stadtmüller, 2005; Shang, 2014; Wang et al., 2016). The objective of fPCA is to compute directions that maximise the variance of the functional data $X(t)$ when projected onto these directions. Assuming $\mathrm{E}[X(t)] = 0, \forall t \in \mathcal{T}$ for notational simplicity, we can formalise fPCA as solving the following problem:

$$\phi_k = \underset{\phi \in L^2(\mathcal{T})}{\arg\max} \ \mathrm{Var}\left[\int_{\mathcal{T}} X(t)\phi(t)dt\right] \tag{9}$$

subject to $||\phi|| = 1$ (normalisation) and $\int_{\mathcal{T}} \phi_l(t)\phi(t) = 0, \forall l < k$ (orthogonality). Here, ϕ_k is the kth orthogonal fPCA direction (*loading*) associated with the covariance function $K(s,t) = \mathrm{Cov}[X(s), X(t)]$; and $v_{ik} = \int_{\mathcal{T}} X_i(t)\phi_k(t)dt$ is the kth fPCA component (*score*), that is, the projection of X_i onto ϕ_k. By construction, ϕ_1 gives the direction of highest variation; ϕ_2 the direction of next highest variation that is orthogonal (uncorrelated) to ϕ_1; and so on.

Even in high dimensional data, often a small number of these components is sufficient to capture most of the variation in X. This feature selection procedure can therefore accommodate dimension reduction with minimal information loss, the trade-off being regulated by the tuning parameter Q, which can be chosen by cross-validation.

Computation. Following (5), we derive the fPCA components not from $\boldsymbol{X}$ but from $\boldsymbol{XB}$. The fPCA components are estimated by singular value decomposition, $\boldsymbol{XB} = \boldsymbol{U\Sigma V}^T$. This produces a matrix $\boldsymbol{V}$ whose columns $[\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots,]$ are the fPCA components or eigenvectors of the covariance matrix of $\boldsymbol{XB}$, approximating the eigenfunctions $[\phi_1, \phi_2, \ldots]$. The dimension reduction projection matrix in (8) is then $\boldsymbol{D} = \boldsymbol{V}_Q$, where $\boldsymbol{V}_Q$ denotes the matrix whose columns are the first Q columns of $\boldsymbol{V}$.

Functional partial least squares (fPLS). Partial least squares was also originally used to solve multicollinearity among predictors in the context of linear models, and has been widely used in chemometrics (Geladi and Kowlaski, 1986; Wold et al., 1984, 2001) and extended to functional data (Aguilera et al., 2010; Delaigle and Hall, 2012; Preda and Saporta, 2005). The objective of fPLS is to identify directions which maximise the covariance between the response y and the functional data $X(t)$ when projected onto those directions:

$$\psi_k = \underset{\psi \in L^2(\mathcal{T})}{\arg\max} \ \mathrm{Cov}^2\left[y, \int_{\mathcal{T}} X(t)\psi(t)dt\right] \tag{10}$$

subject to $||\psi|| = 1$ (normalisation) and $\int_{\mathcal{T}}\int_{\mathcal{T}} \psi_l(s)\Sigma(s,t)\psi(t)ds\,dt = 0, \forall l < k$ (covariance-orthogonality), where $\Sigma(s,t)$ denotes the covariance function of X. Here, ψ_k is the kth covariance-orthogonal fPLS direction; and $r_{ik} = \int_{\mathcal{T}} X_i(t)\psi_k(t)dt$ is the kth fPLS component, that is, the projection of X_i onto ψ_k.

The interpretation of the sequential optimisation problem is similar to the case of fPCA, except that fPLS maximises the covariance between response and predictor instead of the predictor variance. This addresses an important concern, namely that in fPCA the response is not considered and, therefore, there are no guarantees that the components explaining the most variation in the functional predictor are also the best at explaining the relation between the predictor and the response—which is the ultimate goal of the analysis—although the two tend to be related (de Jong, 1993a; Mevik and Wehrens, 2015).

Computation. As before, we derive the fPLS components from $\boldsymbol{XB}$. Several algorithms have been proposed, for example NIPALS (Wold et al., 1984) or SIMPLS (de Jong, 1993b). These produce a matrix $\boldsymbol{R}$ whose columns $[\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots,]$ are the fPLS components approximating $[\psi_1, \psi_2, \ldots]$. The dimension reduction projection matrix in (8) is then $\boldsymbol{D} = \boldsymbol{R}_Q$, where $\boldsymbol{R}_Q$ denotes the matrix whose columns are the first Q columns of $\boldsymbol{R}$.