Research Article

JOURNAL OF CHEMOMETRICS

Received: 5 September 2007, Revised: 22 December 2007, Accepted: 28 December 2007, Published online in Wiley InterScience: 20 March 2008

(www.interscience.wiley.com) DOI: 10.1002/cem.1129

Potential applications of functional data analysis in chemometrics

Wouter Saeys[a]*, Bart De Ketelaere[a,b] and Paul Darius[a,b]

In spectroscopy the measured spectra are typically plotted as a function of the wavelength (or wavenumber), but analysed with multivariate data analysis techniques (multiple linear regression (MLR), principal components regression (PCR), partial least squares (PLS)) which consider the spectrum as a set of *m* different variables. From a physical point of view it could be more informative to describe the spectrum as a function rather than as a set of points, hereby taking into account the physical background of the spectrum, being a sum of absorption peaks for the different chemical components, where the absorbance at two wavelengths close to each other is highly correlated. In a first part of this contribution, a motivating example for this functional approach is given. In a second part, the potential of functional data analysis is discussed in the field of chemometrics and compared to the ubiquitous PLS regression technique using two practical data sets. It is shown that for spectral data, the use of B-splines proves to be an appealing basis to accurately describe the data. By applying both functional data analysis and PLS on the data sets the predictive ability of functional data analysis is found to be comparable to that of PLS. Moreover, many chemometric datasets have some specific structure (e.g. replicate measurements, on the same object or objects that are grouped), but the structure is often removed before analysis (e.g. by averaging the replicates). In the second part of this contribution, we suggest a method to adapt traditional analysis of variance (ANOVA) methods to datasets with spectroscopic data. In particular, the possibilities to explore and interpret sources of variation, such as variations in sample and ambient temperature, are examined. Copyright © 2008 John Wiley & Sons, Ltd.

Keywords: functional data analysis; B-splines; PLS; functional regression; FANOVA

1. INTRODUCTION

In the past decades spectroscopy has steadily gained importance as a rapid and non-destructive analytical technique in the domains of medicine, chemistry and pharmaceutical, environmental, agricultural and food sciences [1–2]. Nowadays, it is even considered the standard technique for the measurement of various parameters and it is introduced in on-line process monitoring [3].

Spectroscopic techniques are based on the fact that the absorption of electromagnetic energy by chemical substances (atoms, molecules, ...) is both specific and proportional. The proportionality is known as Beer's law [4], which states that the absorption of light in a medium is proportional to the pathlength and the concentration of the absorbing agent. The specificity originates from the fact that each atom or molecule will only absorb or emit radiation at the wavelengths corresponding to the energy necessary to move it from one energy state into another. Each molecule has several possible energy transitions (electronic, vibrational and rotational) which can occur simultaneously, such that the resulting energy levels are the sum of the separate transition energies. Moreover, the energy levels are slightly changed by both external influences and energy losses (damping of the oscillator vibrations; collisions between the molecules during absorption and Doppler effects) [5]. In liquids and solids the increased structure of the molecules (e.g. by hydrogen bonding) also influences the energy levels, which results in peak shifts to longer, less energetic wavelengths. The combination of all these phenomena results in the observation of broad absorption peaks, which will often overlap for complex mixtures.

Traditionally, spectral data are analysed by means of multivariate statistical techniques such as multiple linear regression (MLR), principal components regression (PCR) and partial least squares regression (PLS) [6,7]. Historically, Principal Component Analysis has been derived as a data reduction technique, and became rather popular as a data visualization tool due to the presentation of the biplot [8]. As such, PCA and related techniques were not specifically intended for the analysis of spectral data: they consider the multidimensional data block as a set of *m* different variables and the way they have been ordered has no influence on the results. We believe that this is a major shortcoming of these techniques, and advocate the use of analysis tools that take advantage of the peculiar characteristics of spectral data, namely being the result of a sum of absorption peaks caused by the different chemical constituents present in the sample under study.

Ramsay and Silverman [9,10] proposed such a framework, called functional data analysis, to deal with data of a functional nature. So far, functional data analysis was mainly used to handle

* *BIOSYST-MeBio S, Katholieke Universiteit Leuven, K.U. Leuven, Kasteelpark Arenberg 30, B-3001 Heverlee, Belgium.*
E-mail: wouter.saeys@biw.kuleuven.be

a *W. Saeys, B. De Ketelaere, P. Darius*
Division of Mechatronics Biostatistics and Sensors, Department of Biosystems, Katholieke Universiteit Leuven, Leuven, Belgium

b *B. De Ketelaere, P. Darius*
Leuven Statistics Research Centre (LStat), W. de Croylaan 54, B-3001 Heverlee, Belgium

longitudinal data (e.g. repeated measures across time) for applications in various fields, such as medicine, economics, agricultural and behavioural sciences [Wang S, Jank W, Shmueli G. Forecasting eBay's online auction prices using functional data analysis. *J Bus Econ Stat* [14]] [11–16]. Such longitudinal data have in common with spectrometric data the fact that a high correlation exists among neighbouring data points. However, the covariance structure for spectrometric data is more complicated, since chemical constituents have absorption peaks at different wavelengths (e.g. overtones of the same molecular vibration), whereas with longitudinal data it is usually assumed that the correlation decreases with increasing time steps [17].

Although Alsberg [18] already introduced the idea to represent spectra by continuous functions in 1993, the potential of functional data analysis is still not well known to most chemometricians. Therefore, the intent of this article is to show the potential of functional data analysis for the chemometric analysis of spectroscopic data. It should be noted that our main aim is to obtain models that make better use of our knowledge of the physical background of the data and are therefore expected to be more robust and better interpretable. Although it is not our goal to obtain lower RMSEPs, we aim at models that are non-inferior to the purely data based (black box) PLS models.

In a first part, a motivating example will be given. Second, the functional representation of spectral data will be elucidated. In a third part, the functional data will be used in a regression context and will be compared to PLS. Finally, an example for datasets having a specific structure will be handled as an alternative for classical analysis of variance (ANOVA). Throughout all sections, examples will be used.

2. MOTIVATING EXAMPLE

As a motivating example for the use of functional tools for the analysis of spectral data, consider NIR spectra from hog manure samples, detailed in [19]. The data set consists of 420 samples (Flemish hog manures collected and analysed in 2004). All samples were scanned in reflectance mode on a diode array Vis/NIR spectrophotometer. After conversion into absorbance units and removal of the noisy parts at the lower and higher end, the absorbance spectra ranging from 426 to 1686 nm serve as predictor variables. Figure 1 (top) presents the average absorbance spectrum together with pointwise 95% confidence limits.

After removal of six outliers (four spectral, one compositional and one for both) the remaining 414 hog manure samples are divided into a training set of 276 samples and a test set of 138 samples. The dry matter content is considered as the variable to be predicted in this case study. PLS models are built between the absorbance spectra and the mean centred reference data. The optimal number of latent variables is chosen for the training set at the first local minimum in the root mean square error of cross-validation (RMSECV) obtained in a five-fold cross-validation with contiguous blocks. The predictability of the resulting model (trained on the 276 samples of the training set) is then evaluated by calculating the root mean square error of prediction (RMSEP) for the test set of 138 samples, which have not been used during the model building. An RMSEP of 12.316 g L^{-1} is obtained for the selected PLS model with 10 latent variables.

Let us now alter the dataset by random permutation of the covariates, i.e. datavector $x_i = x_i(\lambda)$ is altered to $x_i = x_i(\kappa)$ with κ being a random permutation of λ. By doing so, the natural ordering of the covariates along the corresponding wavelengths, and hence, spectral information, is destroyed. The permuted average spectrum and 95% confidence limits are given in Figure 1 (bottom). However, using $x_i(\kappa)$ instead of $x_i(\lambda)$ yields exactly the same results if PLS is used with an RMSEP of 12.316 g L^{-1} for a 10 component model. One could conclude that PLS does not take advantage of the natural ordering of the covariates and that PLS is thus in a way suboptimal. As a result, the obtained model could yield predictions that are not optimal in terms of RMSE. More often, not including the natural ordering of the covariates gives the model too much flexibility, such that over-fitting and/or lack of generality are potential threats.

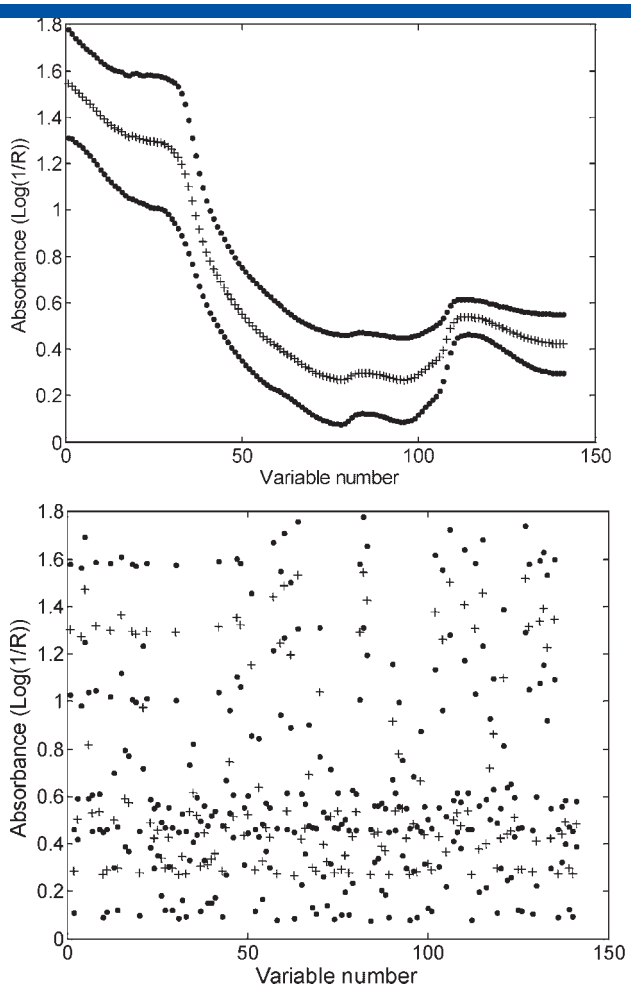

Figure 1. Visualization of the mean reflectance values (+) and their 95% confidence limits (.) for the 141 wavelength variables measured for the 420 hog manure samples in order of increasing wavelength (top) and in random order (bottom).

The proposed methodology takes into account the natural ordering of the covariates, and will be compared to PLS in the next section.

3. FUNCTIONAL REPRESENTATION OF OPTICAL SPECTRA

As mentioned before, the basic idea of functional data analysis is to think of the observed data functions as single entities rather than merely a sequence of individual observations. As the

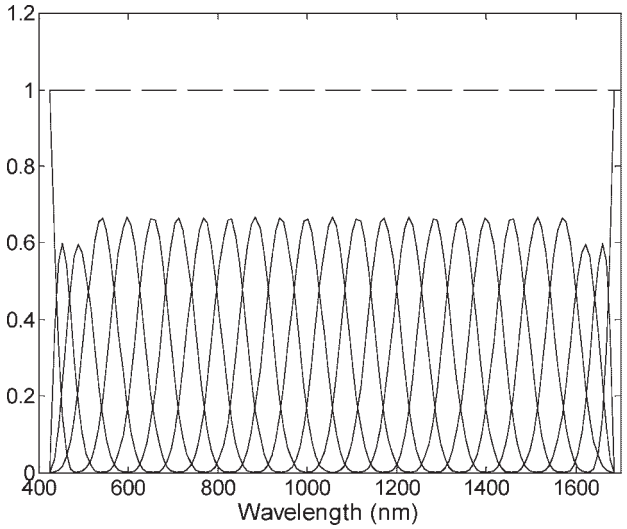

Figure 2. A functional basis of 25 B-splines (solid) in the wavelength range from 426 to 1686 nm with equally spaced knots and their summed effect (dashed).

functional data usually are observed and recorded discretely, as is the case with spectroscopic data, we need techniques to retrieve the underlying function from the raw (discrete) functional data. Therefore, the first step in any functional data analysis consists of the functional representation of the observed data.

There exists a variety of methods (*bases*) for the functional representation of the discrete data set, all of which are typically based on some kind of smoothing. Examples of widespread functional bases are Fourier series, polynomial bases, wavelet bases and spline bases. Hence, the first step in recovering the functional object is to choose an appropriate family of basis functions matching the underlying function(s) to be estimated.

For spectroscopic data, especially in the NIR range, a basis of B-splines seems to be most appropriate because of their natural resemblance with absorption peaks and their compact support, implying interesting computational properties. Except near the boundaries, the B-splines are all identical bell-shaped curves. The non-zero part of each B-spline consists of a piecewise cubic polynomial, with four cubic pieces that fit together smoothly [20]. In Figure 2 a basis of 25 of such cubic B-splines in the wavelength range (426–1686 nm) with equally spaced knots is illustrated. To make this basis of 25 B-splines 21 knots at equal distances of 57.27 nm are defined between the start point at 426 nm and the end point at 1686 nm. As can be seen from Figure 2, each B-spline runs from one knot to another and the different splines overlap. The splines in the middle part are of equal magnitude and symmetrical, and run over a distance of 229.10 nm or 4 knots. At the sides three asymmetric splines are used which all start (end) in the start (end) point, but run to different knots. This is done to avoid boundary effects, such that adding all 25 splines results in constant amplitude of 1 over the whole considered spectral range.

Choosing the appropriate magnitudes for all splines makes it possible to approximate any continuous functional form and is discussed in the next paragraph. The functional basis taken as an example in Figure 2, with equally spaced knots, is only a subclass of all possible B-spline bases; indeed, one could also define a basis with knots that are not equally spaced. It could for instance be worthwhile considering knot placements based on spectral curvature, or on the known properties of the absorption peaks of the major constituents in the sample under study. This flexibility is among others the strength of this method.

Denote the basis of B-splines $\{\phi_k\}_{k=1}^{K}$ with K the number of basis functions. The degree to which the data is smoothed is determined by the number K. A basis consisting of a large number of splines will result in a low degree of smoothing and thus a high resolution of the spectral details. On the other hand, this could cause spectral noise being included in the analysis. There are several possible ways to determine the degree of smoothing and thus the number of data functions (splines). One can use a leave one out procedure that considers the approximation error for a wavelength variable left out of the functional fitting procedure [21]. Other possible approaches consist of applying a scree plot procedure similar to the way the number of latent variables is selected for PLS or the construction of a B-spline basis from expert knowledge of the underlying absorption peaks.

Consider the $n \times m$ matrix $\mathbf{X}$ where x_{ij} is the intensity, or equivalently the reflectance or absorbance at wavelength λ_j, measured for sample i. The spectral function $x_i = x_i(\lambda)$ for sample i can then be described as a linear combination of the basis functions

$$x_i = \sum_{k=1}^{K} c_{ik}\phi_k. \tag{1}$$

The regression coefficients c_{ik}, also called the B-spline weights, are obtained by minimizing the following least squares criterion:

$$\min \sum_{j=1}^{m} \left(x_{ij} - \sum_{k=1}^{K} c_{ik}\phi_k(\lambda_j) \right)^2 \quad \forall i \in [1, n] \tag{2}$$

which corresponds to minimizing the vertical distance between the observed spectral information and the fitted curve. This criterion can be augmented with penalty terms to tune the smoothness [20].

Denote the $m \times K$ matrix $\{\phi_k(\lambda_j)\} = \Phi$. Then the criterion is minimized by the solution

$$c_i = (\Phi'\Phi)^{-1}\Phi'\mathbf{X'_i} \quad \forall i \in [1, n] \tag{3}$$

where the K dimensional vector c_i contains the coefficients c_{ik}. In the remainder of the text, all computations will be based on the B-spline weights, which corresponds to a powerful dimension reduction.

In Figure 3 the absorbance values measured at 141 spectral variables x_i for a hog manure sample with 48.22 g L^{-1} dry matter content are presented together with the functional representation of the spectrum and the corresponding B-spline weights. These B-spline weights c_{ik} are multiplied by the B-spline basis $\{\phi_k\}_{k=1}^{K}$ from Figure 2 ($K = 25$) using Eqn. 3 to obtain the functional approximation of the spectrum by summation of those contributions (Eqn. 1). The functional representation of the spectra combines dimensionality reduction and smoothing in one step while preserving the information in the derivatives. This combination is well adapted to the situation we have in spectroscopy where we expect fairly smooth spectra and need dimensionality reduction prior to regression.

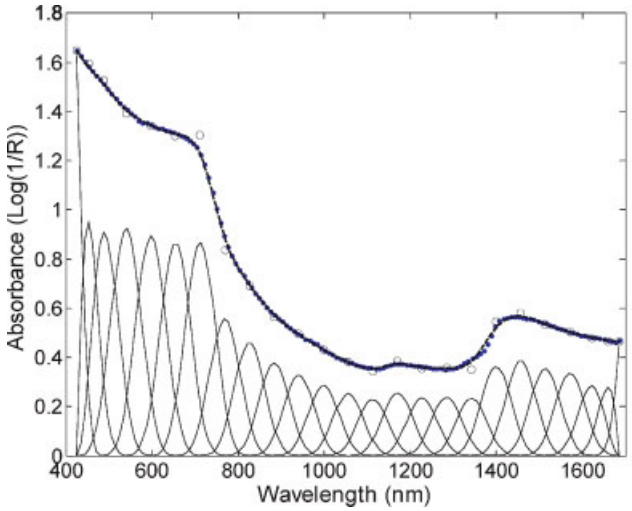

Figure 3. Discrete absorbance values (diamonds) at 141 wavelengths for a hog manure sample with 48.22 g L^{-1} dry matter and 5.29 g L^{-1} total nitrogen, together with its functional representation (dashed line). This functional representation is the sum of the B-spline contributions (solid lines) obtained by multiplying the B-spline basis illustrated in Fig. 2 by the appropriate B-spline coefficients (circles). This figure is available in colour online at www.interscience.wiley.com/journal/cem

4. FUNCTIONAL REGRESSION MODELS

We consider a linear model defined by a set of functions where the response variable y is a scalar. Let us recall some aspects of ordinary linear regression. Suppose $y_1, \ldots, y_n$ are observations of a response variable at values of $x_1, \ldots, x_n$ of a multivariate covariate x of dimension m. Multiple linear regression then fits a model of the form

$$y_i = a + \sum_j \beta_j x_{ij} + \varepsilon_i = a + \langle \beta | x_i \rangle + \varepsilon_i \tag{4}$$

where a is an intercept, β_j is the j-th regression coefficient, ε_i is a residual or disturbance term and $\langle \beta | x_i \rangle$ represents the vector dot product between the vectors β and $\mathbf{x}_i$ consisting of the elements β_j and x_{ij} [17]. We can now consider a functional extension of linear regression where the prediction of the scalar values y_i is based on functions $x_i(s)$ with s the function argument. To extend the idea of linear regression, we must replace the regression coefficient vector β in Eqn. (4) by a regression function $\beta(s)$. Since the functions $x_i(s)$ and $\beta(s)$ are continuous, the sum of the inner product becomes an integral. This results in the following expression for the functional regression model:

$$y_i = a + \int_{\lambda_1}^{\lambda_m} x_i(s)\beta(s)\mathrm{d}s + \varepsilon_i = a + \langle x_i | \beta \rangle + \varepsilon_i \tag{5}$$

where λ_1 and λ_m are the start and end points (i.e. wavelengths) of the functions.

Analogous to the spectral functions $x_i(\lambda)$ the regression function $\beta(s)$ can be described as a linear combination of the K basis functions

$$\beta(s) = \sum_{\nu=1}^{K} b_\nu \phi_\nu(s) \tag{6}$$

where b_ν is the B-spline coefficient corresponding to the νth B-spline basis function ϕ_ν.

Let us define the $K \times K$ matrix $\mathbf{J}$ with entries

$$\mathbf{J}_{jk} = \int \phi_j(s)\phi_k(s)\mathrm{d}s = \langle \phi_j | \phi_k \rangle \tag{7}$$

Since the B-splines in the basis are not orthonormal, $\mathbf{J}$ will also have off-diagonal elements. However, the matrix $\mathbf{J}$ will be diagonal dominant such that good estimates can be obtained for the regression coefficients [18].

Thanks to this projection of the spectral and regression functions onto the functional basis the integral of the functional scalar product in Eqn. (5) can be approximated as

$$\langle x_i | \beta \rangle = \sum_{k=1}^{K} \sum_{\nu=1}^{K} c_{ik} \mathbf{J}_{k\nu} b_\nu \tag{8}$$

The notation can be further simplified by defining the $(K+1)$ regression coefficient vector $\varsigma = [a, b_1, \ldots, b_K]^{\mathrm{T}}$, which incorporates the intercept a, and defining the coefficient matrix $\mathbf{Z}$ to be the $n \times (K+1)$ matrix $\mathbf{Z} = [\,1 \quad \mathbf{CJ}\,]$ where $\mathbf{C}$ contains the B-spline coefficients c_{ik} of Eqn. (1). In this way Eqn. (5) can be rewritten in the following form

$$y = \mathbf{Z}\varsigma + \varepsilon \tag{9}$$

And the least squares estimate $\hat{\varsigma}$ of the augmented parameter vector ς is calculated as

$$\hat{\varsigma} = (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1} Z^{\mathrm{T}} y. \tag{10}$$

The prediction of a new observation vector y_{new} is given by the following formula

$$\hat{y}_{\mathrm{new}} = \mathbf{Z}_{\mathrm{new}}\hat{\varsigma} \tag{11}$$

where $\mathbf{Z}_{\mathrm{new}} = [\,1 \quad \mathbf{C}_{\mathrm{new}}\mathbf{J}\,]$ and $\mathbf{C}_{\mathrm{new}}$ is the matrix of the B-spline weights for the spectra of the new observations projected on the same functional basis. Since MLR is performed between the B-spline coefficients and the predictor variable, functional regression has some mathematical analogy to other projection methods like PCR and PLS. It should, however, be noted that the B-spline basis functions take advantage of the correlation between neighbour variables which is totally ignored in the other projection methods (PCR and PLS). The fact that these basis functions only run over a limited wavelength interval also creates interesting possibilities with respect to variable selection and interpretability [21].

This functional regression procedure will now be illustrated on two examples: the prediction of dry matter content from NIR spectra of hog manure (motivating example) [19] and the prediction of the cetane number of Diesel [22].

For the motivating example of hog manure spectra functional regression models are built between the B-spline weights and the standardized reference data. This is done for different numbers of B-splines in the basis and the RMSECV values obtained in a five-fold cross-validation with contiguous blocks are plotted against the number of B-splines (Figure 4). Based on the RMSECV curve a functional regression model with 11 B-splines is selected. The predictive power of the resulting model (trained on the 276 samples of the training set) is then evaluated by calculating the RMSEP for the test set of 138 samples, which have not been used

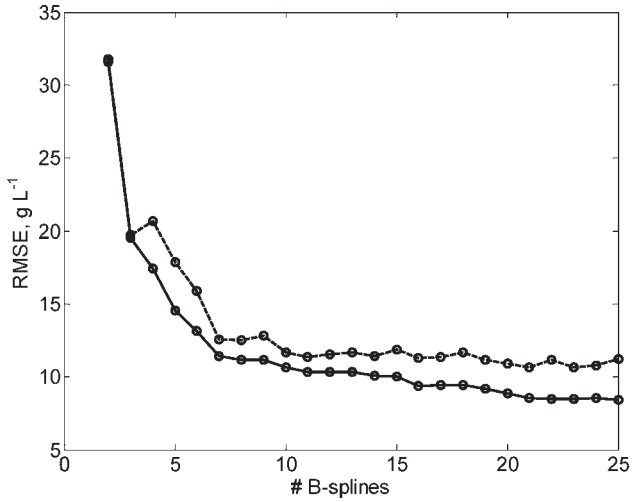

Figure 4. RMSEC (solid line) and RMSECV (dashed line) curves for the selection of the number of B-splines in the functional regression model for the hog manure example.

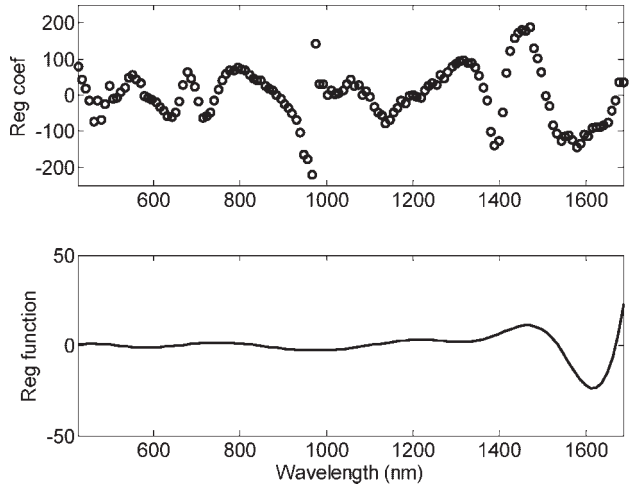

Figure 5. Regression coefficients for the partial least squares models (top) and regression function for the functional regression models (bottom) built on the training set.

during the model building. An RMSEP of 11.928 g L^{-1} is obtained a for the functional regression model with 11 B-splines. The obtained RMSEs are listed in Table I together with the corresponding values obtained for the PLS regression model. Comparison of the results in Table I shows that functional regression and PLS perform equally well. The regression function for the functional regression model and the regression coefficient for the PLS regression model are displayed in Figure 5. As could be expected, the regression function and the PLS regression coefficient plot have some similarity in their curvature, but the regression function is far smoother than the PLS regression coefficient plot.

The second example consists of the Diesel data set which can be downloaded from the eigenvector website [22]. This data was measured by the Southwest Research Institute (SWRI) and consists of a training set of 133 samples (20 high leverage samples and 113 low leverage samples) and a validation set of 112 samples (low leverage). In this study, both PLS and functional regression models are built to predict the cetane number from the NIR transmission spectra. Models are built for different numbers of PLS components, respectively B-splines and the optimal number of PLS component/B-splines is chosen based on the curves for the RMSECV obtained in a random five-fold cross-validation. Figure 6 displays RMSEC and RMSECV curves for the functional regression, where a clear minimum in the RMSECV can be seen for a functional regression model with 15 B-splines in the basis. A PLS model with six components and a functional regression model with 15 B-splines are selected, because these

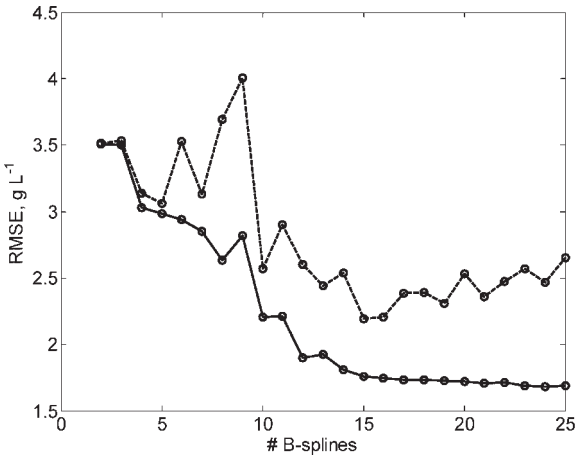

Figure 6. RMSEC (solid lines) and RMSECV (dashed lines) curves for the selection of the number of B-splines in the functional regression model for the Diesel example.

give minimal RMSECV values of 2.016 CN, respectively 2.195 CN. The predictive power of the resulting models (trained on the 133 training samples) is then evaluated by calculating the RMSEP for the test set of 112 samples, which have not been used during the model building. An RMSEP of 2.139 CN is obtained for the PLS model with six latent variables and an RMSEP of 2.111 CN for the

Table I. Predictability of the partial least squares and functional regression models obtained on the training and test set for the hog manure example

Partial least squares regression				Functional regression			
# LV's	RMSEC g L^{-1}	RMSECV g L^{-1}	RMSEP g L^{-1}	# B-splines	RMSEC g L^{-1}	RMSECV g L^{-1}	RMSEP g L^{-1}
10	9.353	10.957	12.316	11	10.325	11.380	11.928

Table II. Predictability of the partial least squares and functional regression models obtained on the training and test set for the Diesel example

Partial least squares regression				Functional regression			
# LV's	RMSEC CN	RMSECV CN	RMSEP CN	# B-splines	RMSEC CN	RMSECV CN	RMSEP CN
6	1.788	2.016	2.139	15	1.761	2.195	2.111

functional regression model with 15 B-splines. An overview of the obtained RMSEs for both models is given in Table II. Comparison of the results in Table II shows that functional regression and PLS perform equally well. The regression function for the functional regression model and the regression coefficients for the PLS regression model are displayed in Figure 7. Different to what one would expect, there is very little similarity between the PLS regression coefficient plot and the regression function. The regression function is again far smoother than the PLS regression coefficient plot. To get an idea of the degree of smoothness that has been introduced by the functional representation the measured spectrum for a sample with a cetane number of 59.100 CN is displayed together with the corresponding function (Figure 8). As could have been expected, the functional description using only 15 B-splines is not able to fit these Diesel spectra with high curvature accurately. One might say that the measured spectrum has been smoothed too hard. However, as can be seen from Figure 6, using more B-splines in the functional description would result in over-fitting by the functional regression model. It should also be noted that the number of B-splines used in the functional regression is considerably higher than the selected number of [16] PLS components (6). Clearly a trade off has been made between having sufficient B-splines to accurately fit the curvature and not having too many dimensions to avoid over-fitting and loss of generalizability due to the higher number of regression coefficients to be estimated. From this observation it can be questioned whether selection of the number of B-splines based on the cross-validation error, which is commonly used for selecting the number of latent variables in PLS, is the best way to follow. This approach makes a decision about the degree of smoothing and the dimensionality reduction based on one criterion, while an independent optimization of both aspects might be better.

Thanks to the resemblance of the B-splines with absorption peaks and its use of the spatial correlation the functional description also takes into account the correlation between neighbour variables which is totally ignored by PLS. Since functional regression performs equally well as PLS for both examples, while incorporating the spatial correlation between the wavelength variables, functional regression could be a valuable alternative for PLS to handle spectroscopic data in a regression context. However, the high number of B-splines needed to obtain a good functional fit of a spectrum with strong curvature poses a risk of over-fitting which forces us to make a trade off between a good functional fit of the spectra and high prediction power. The choice of the number of B-splines based on the cross-validation error might therefore be suboptimal for spectra with high curvature such as the Diesel spectra. In this case variable selection techniques such as a forward-backward procedure [21] or a further dimensionality reduction by performing PCR or PLS on the B-spline regression coefficients [23] might be good options which have to be further investigated. The selection of the degree of smoothness and the amount of dimensionality reduction could then be optimized independently by using a cross-model validation scheme. It would also be interesting to compare this approach to other methodologies for the creation of parsimonious models such as O-PLS [24] and wavelet and Fourier transforms [25,26].

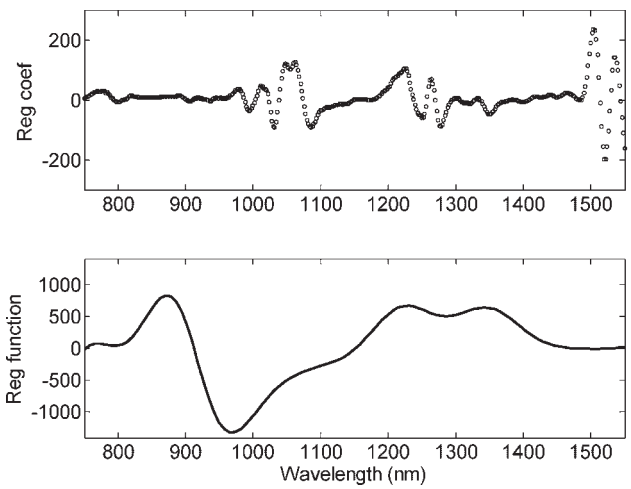

Figure 7. Regression coefficients for the partial least squares models (top) and regression function for the functional regression model (bottom) built for the Diesel training set.

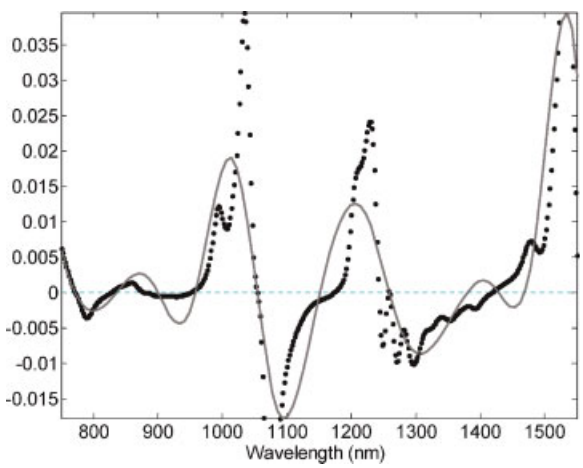

Figure 8. Overlay of the measured spectrum for a Diesel sample with a cetane number of 59.100 CN and its functional representation on the basis of 15 B-splines.

5. FUNCTIONAL ANOVA

ANOVA models are versatile statistical tools for studying the relation between a response variable and one or more explanatory (typically categorical) variables. In this study, we consider the case where the response variable, i.e. the spectral information, is decomposed into contributions of the overall mean and the main effects, which are functional. We illustrate the method for the case of one-way ANOVA, but the generalization to other cases is straightforward. The corresponding functional ANOVA model can be written as

$$\begin{aligned} &x_{ig}(\lambda) = \mu(\lambda) + \alpha_g(\lambda) + \varepsilon_{ig}(\lambda) \\ &\text{with } i \in [1, n_g]; \quad g \in [1, g]; \quad n_T = \sum_{g=1}^{G} n_g \end{aligned} \tag{12}$$

where $x_{ig}(\lambda)$ is the functional representation of the spectrum for sample i of group g, n_g is the number of samples in group g, G is the number of groups and n_T is the total number of samples. The function $\mu(\lambda)$ is the grand mean function, which corresponds to the average spectrum. The terms $\alpha_g(\lambda)$ are the specific functional effects of belonging to a certain group g. To be able to identify them uniquely, we require that they satisfy the constraint

$$\sum_g \alpha_g(\lambda) = 0 \quad \forall \lambda \in [\lambda_1, \lambda_m] \tag{13}$$

The residual function $\varepsilon_{ig}(\lambda)$ is the unexplained variation specific to the ith sample within group g. We can now define a corresponding set of $(g+1)$ functional effects $\beta_q(\lambda)$ by setting $\beta_1(\lambda) = \mu(\lambda)$, $\beta_2(\lambda) = \alpha_2(\lambda)$ and so on to $\beta_{g+1}(\lambda) = \alpha_g(\lambda)$ such that the functional vector $\beta = [\mu, \alpha_1, \ldots, \alpha_g]^{\mathsf{T}}$. By defining an $n_T \times (g+1)$ design matrix $\mathbf{Z}$ where each row corresponds to a sample with a one in the first column for the average effect and a one in the column of the respective group effect, Eqn. (12) can be rewritten as

$$x(\lambda) = \mathbf{Z}\beta(\lambda) + \varepsilon(\lambda) \tag{14}$$

The constraint (Eqn. (13)) is implemented by adding a row to the matrix $\mathbf{Z}$ and a corresponding zero entry in the x vector to obtain $\mathbf{Z}^*$ and x^*. By describing both the spectral functions $x(\lambda)$ and the regression functions $\beta(\lambda)$ as a linear combination of the same K B-spline basis functions, $x(\lambda) = \mathbf{C}\phi(\lambda)$, respectively $\beta(\lambda) = \mathbf{B}\phi(\lambda)$, the estimation of the functional effects from the spectral functions comes down to estimating the matrix $\mathbf{B}$ from the matrix $\mathbf{C}$. The least squares estimate $\hat{\mathbf{B}}$ for the matrix $\mathbf{B}$ is then obtained by the following matrix calculation:

$$\hat{\mathbf{B}} = (\mathbf{Z}*^{\mathsf{T}}\mathbf{Z}*)^{-1}\mathbf{Z}*^{\mathsf{T}}\mathbf{C}* \tag{15}$$

where a row of zeroes has been added to the matrix $\mathbf{C}$, denoted by $\mathbf{C}^*$, to satisfy the constraint on the functional effects (Eqn. (13)). In this way, the projection onto the functional basis has again been used efficiently.

After estimating and plotting the main effects, we can also investigate whether this effect is substantial, both locally and globally. To determine whether the main effect is significant locally, one can compute the estimated pointwise standard error for that effect and construct a confidence interval by adding and subtracting two estimated standard errors. The estimated pointwise standard error is given by the square root of the following mean square of error function MSE (λ)

$$\mathrm{MSE}(\lambda) = \frac{1}{df(\text{error})} \sum_{i,g} \left[x_{ig}(\lambda) - (\mathbf{Z}\hat{\beta})_{ig}(\lambda)\right]^2 \tag{16}$$

where df(error) is the number of degrees of freedom for error, or the sample size minus the number of mathematically independent functions $\beta_q(\lambda)$ in the model.

To judge whether a main effect is significant globally, one defines contrasts between the different levels of the explanatory variables and can consider the following test statistic

$$M = \sup_\lambda \left| \mathrm{Contrast}(\lambda) \Big/ \sqrt{\mathrm{MSE}_C(\lambda)} \right| \tag{17}$$

where Contrast(λ) is the considered contrast function and $\mathrm{MSE}_C(\lambda)$ is the mean square error of the contrast. Let $\mathbf{u}$ be the vector which relates the estimated contrast to the vector of functional effects

$$\mathrm{Contrast}(\lambda) = \mathbf{u}^{\mathsf{T}}\hat{\beta}(\lambda) \tag{18}$$

and define the scalar a by

$$a^2 = \mathbf{u}^{\mathsf{T}}(\mathbf{Z}^{\mathsf{T}}\mathbf{Z})\mathbf{u} \tag{19}$$

such that the squared pointwise standard errors of the estimated contrasts equal

$$\mathrm{MSE}_C(\lambda) = a^2\mathrm{MSE}(\lambda) \tag{20}$$

A permutation-based significance value for the test statistic M can be obtained by randomly assigning the group labels to the data, keeping the number of elements n_g in each group g the same. The statistic M is then calculated for each random permutation of the data and the p-value is computed.

An alternative test statistic was developed by Cuevas *et al.* [26] for testing the hypothesis:

$$H_0 : \alpha_1 = \cdots = \alpha_G = 0 \tag{21}$$

This test is considered less informative and was therefore not implemented here.

This functional analysis of variance (FANOVA) approach will be illustrated on the data of a study to test the effect of ambient temperature and sample temperature on the spectra measured with a mobile diode array instrument (Zeiss Corona VISNIR 1.7 fibre). The experiment followed a full factorial design with spectra from manure samples of four types of animals (dairy, beef, calf and hog), preserved at three temperatures (4, 12, +20°C) and measured at three ambient temperatures (4, 12, +20°C), having nine replicates, yielding 324 samples in total. We work with the intensity spectra (pure spectra) measured in reflectance mode instead of the relative reflectance spectra, because it is expected that the reference intensity might also depend on the ambient temperature.

The following functional FANOVA model is fit to the data:

$$\begin{aligned} &I_{ijkl}(\lambda) = \mu(\lambda) + T_i(\lambda) + A_j(\lambda) + S_k(\lambda) + \varepsilon_{ijkl}(\lambda) \\ &\forall i \in [1, 4]; \quad j \in [1, 3]; \quad k \in [1, 3]; \quad l \in [1, 9] \end{aligned} \tag{22}$$

where $I_{ijkl}(\lambda)$ is the measured intensity at wavelength λ, $\mu(\lambda)$ is the overall mean, $T_i(\lambda)$ is the main effect of the ith type of animal, $A_j(\lambda)$

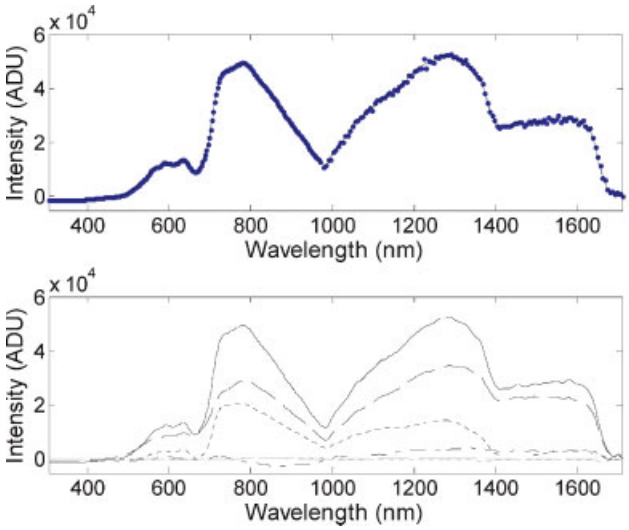

Figure 9. Upper plot: discrete intensity values (dots) of a dairy manure sample preserved at 4°C and measured at room temperature (+20°C) together with its corresponding functional object (solid line). Lower plot: the contribution of the overall mean function (black dashed line), the functional effects of diary type (black dotted line), 4°C sample temperature (grey dashed line), and 20°C ambient temperature (grey dotted line), and the residual (black dash-dotted line) to the spectral function (black solid line). This figure is available in colour online at www.interscience.wiley.com/journal/cem

is the main effect of the *j*th ambient temperature, $S_k(\lambda)$ is the main effect of the *k*th sample temperature and $\varepsilon_{ijkl}(\lambda)$ the residual.

In order to obtain a unique solution, the following constraints are added

$$\sum_i T_i(\lambda) = 0 \quad \sum_j A_j(\lambda) = 0 \quad \sum_k S_k(\lambda) = 0 \qquad \forall \lambda \in [\lambda_1, \lambda_m] \tag{23}$$

The effect of type is expected to be dominant since spectra for the different kinds of animals are very different from one another. In Figure 9 the intensity spectrum of a sample of dairy manure preserved at 4°C and measured at room temperature is shown. In the upper pane the functional object is fit to the intensity values measured at discrete wavelengths. The two distinct peaks originate from the wavelength dependent sensitivity of the detector diodes, which is maximal around 750 nm for the Si diodes and around 1350 nm for the InGaAs diodes. The clear decrease towards 1450 nm indicates the presence of water in the manure, while the local minimum around 680 nm indicates the presence of chlorophyll. The low intensity in the visible range can be explained by the dark brown colour of manure.

In the lower pane the contribution of the different effects is shown. The most important effect is the overall mean, which represents the features that are common for all manure samples: absorption peaks at 680 and 1450 nm due to the presence of some chlorophyll and water, and low reflected intensities in the visible due to the dark brown colour. Second most important is the effect of manure type. The clear local minima at 680 and 1450 nm correspond to the higher chlorophyll and water content in diary manure compared to the average manure sample. The

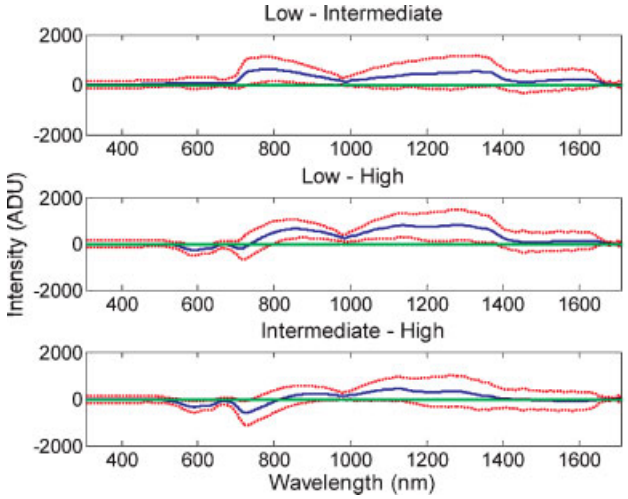

Figure 10. Contrasts for sample temperature: pointwise estimates (solid line) and 95% confidence intervals (dotted line) around the estimates. This figure is available in colour online at www.interscience.wiley.com/journal/cem

effects of ambient and sample temperature are small compared to the other effects, while the residual function is considerable.

We wish to investigate whether these effects are significant. Considering the practical applications of Vis/NIR spectroscopy, we are only interested in the effects of ambient and sample temperature. We define three contrasts: low–intermediate temperature (4–12°C), low–high temperature (4–20°C) and intermediate–high temperature (12–20°C). The pointwise estimated values of the contrasts and the 95% confidence intervals around these are plotted in Figures 10 and 11 for sample temperature and ambient temperature, respectively. When zero (no contrast) lies outside the confidence interval around the measured contrasts, the H_0 hypothesis is rejected and the contrast is considered significant.

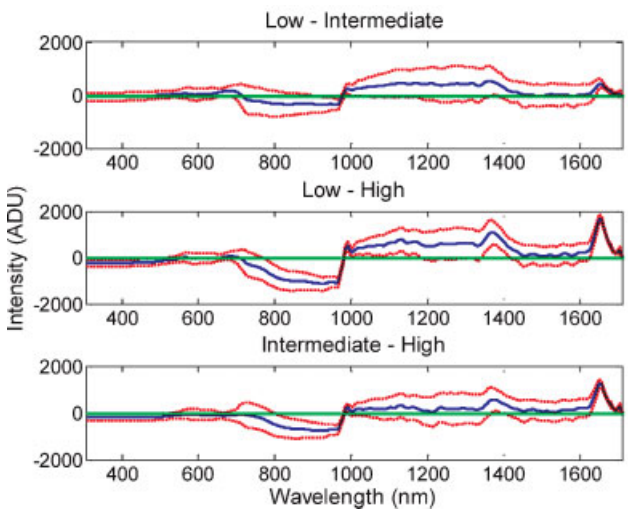

Figure 11. Contrasts for ambient temperature: pointwise estimates (solid line) and 95% confidence intervals (dotted line) around the estimates. This figure is available in colour online at www.interscience.wiley.com/journal/cem

Table III. Test statistics and corresponding *p*-values for contrasts for ambient and sample temperature

Effect of sample temperature	Effect of ambient temperature
$M_{L\text{-}I} = 2.5923$ ($p = 0.100$)	$M_{L\text{-}I} = 7.1555$ ($p < 0.001$)
$M_{L\text{-}H} = 3.9837$ ($p = 0.003$)	$M_{L\text{-}H} = 25.8461$ ($p < 0.001$)
$M_{I\text{-}H} = 2.9288$ ($p = 0.051$)	$M_{I\text{-}H} = 18.6907$ ($p < 0.001$)

From Figure 10 it can be seen that the contrast between low and intermediate sample temperatures is significant in the 700–1000 nm region, and the contrast between low and high sample temperature is significant in the region around 600 nm and the region from 800 to 1350 nm, while the one between intermediate and high temperature is only significant around 600 and 700 nm. It should be noted that although both the contrasts low-intermediate and intermediate-high are significant in the region around 700 nm, the contrast between low and high sample temperature is not. The contrasts in Figure 11 show that the contrast between low and intermediate ambient temperature is only significant in the 920–975 and 1640–1700 nm regions, and the contrast between intermediate and high temperature is significant in the range from 800 to 1000 nm and above 1630 nm. The contrast between low and high ambient temperatures is found to be significant in the regions below 500 nm, between 770 and 1180 nm, between 1340 and 1400 nm and above 1630 nm.

To investigate whether the effects of sample and ambient temperature are overall significant, the permutation based overall significance test is performed. The test statistics for the defined contrasts are summarized in Table III. A permutation-based significance value for each of these statistics is obtained by randomly permuting the temperatures, keeping the totals the same within each type. The test statistics M are calculated for each random permutation of the data and the models are ordered by decreasing values of this statistic. The p-value for the contrast under study is then obtained by dividing its position by the number of randomizations (e.g. 5th place out of 1000 results in a p-value of 0.005; and a higher M value than all 1000 randomizations gives a p-value smaller than 0.001).

From the overall test we can conclude the following: since the probability of the observed contrasts for sample temperature under the H_0 hypothesis of no contrast is more than 5% the effect of switching from low to intermediate, and from intermediate to high sample temperature is not considered significant, while the contrast between low and high sample temperature with its p-value of 0.3% is. For ambient temperature all contrasts are, however, found to be highly significant, with probabilities under the H_0 hypothesis of less than 0.1%.

These findings have consequences for the practical implementation of an on-line Vis/NIR spectroscopic manure composition sensor. While limited variations in the sample temperature can be tolerated, it is obvious that the ambient temperature around this diode array spectroscopic sensor should be kept within narrow limits to keep the sensor robust. Omitting the spectral regions which are significantly sensitive for ambient temperature variations could make the sensor more robust against varying ambient (and to a certain extent also sample) temperatures. However, what the effect of this would be on the predictability of the resulting regression models should be further investigated. This case study has shown that functional analysis of variance makes it possible to quantify the contribution of different main effects and to judge the significance of these contributions. The main benefit of using functional analysis of variance in this context rather than pointwise ANOVA on the measured variables, is the fact that the spectra are described as smooth functions, which makes the detection of significantly influenced spectral regions less sensitive to the noise level on each wavelength. A pointwise ANOVA on the original variables often gives jumpy effect-spectra and significance tests which are uninteresting from a spectroscopic point of view. On the other hand, alternative MANOVA techniques for spectra, such as 50-50 MANOVA [28], try to avoid this problem by restricting the analysis to the variation described by the first few principal components (PCA, PLS) and thus totally ignore the correlation between the variables. A thorough comparison of these different approaches for analysis of variance on spectra would be very interesting and will form the topic of a future study.

6. SOFTWARE

- All calculations in this research were performed in Matlab making use of the functional data analysis toolbox which can be downloaded from Professor Ramsay's ftp-site (ftp://ego.psych.mcgill.ca/pub/ramsay/fdafuns/). A toolbox in R and SPLUS are also available from this site.
- The data and specific code for the analysis of the examples presented in this study are made available to the reader (http://www.biw.kuleuven.be/aee/amc/staff/wouters/wouter.htm).
- All PLS calculations were performed using the PLS toolbox (www.eigenvector.com).

7. CONCLUSION

In this article an alternative approach to deal with spectrometric data has been suggested. This approach considers a spectrum as a function of the wavelength or wavenumber rather than as a set of separate points. Moreover, the use of a B-spline basis was shown to have some similarity with the physical origin of a spectrum as the result of light (energy) absorbance by molecular bonds. The main advantage of this functional description of spectra lies in its combination of dimensionality reduction and smoothing.

By applying both functional data analysis and PLS regression to two example data sets it was shown that this technique obtains predictive power comparable to that of PLS while taking into account the spatial correlation between the variables, which is related to the physics underlying the spectra. However, it was noted that simultaneous selection of the degree of smoothness and the dimensionality reduction based on the cross-validation error may be suboptimal.

A second advantage of functional data analysis over PLS is that it provides a way to deal with designed experiments and to perform an ANOVA for spectroscopic data. This FANOVA approach has been applied to a designed manure data set and revealed a significant effect of the ambient temperature on the intensity spectrum of manure measured using a diode array instrument. This technique could be a valuable tool for the investigation and improvement of the robustness of spectroscopic techniques.

Acknowledgements

Wouter Saeys is funded as a Postdoctoral Researcher of the Research Foundation - Flanders (FWO). Bart De Ketelaere is an Industrial Research Fellow of the K. U. Leuven and Paul Darius is an associate professor at K. U. Leuven.

REFERENCES

1. Burns DA, Ciurczak EW. Handbook of Near-Infrared Analysis. Marcel Dekker Inc.: New York, 1992; 681.
2. Williams PC, Norris K. Near-Infrared Technology in the Agricultural and Food Industries (2nd edn). American Association of Cereal Chemists: St Paul, MN, 2001.
3. Jorgensen P, Pedersen JG, Jensen EP, Esbensen KH. On-line batch fermentation process monitoring (NIR)-introducing 'biological process time'. *J Chemometrics* 2004; **18**: 1–11.
4. Beer A. Bestimmung der absorption des rothen Lichts in farbigen Flüssigkeiten. *Annalen der Physik Chimie* 1852; **86**: 78–88.
5. Liou KN. Introduction to Atmospheric Radiation (2nd edn). Academic Press: London, UK, 2002.
6. Martens H, Naes T. Multivariate Calibration (2nd edn). Wiley: Chichester, UK, 1989.
7. Frank I, Friedman J. A statistical view of some chemometrics regression tools (with discussion). *Technometrics* 1993; **35**: 109–147.
8. Gabriel KR. The biplot graphical display of matrices with applications to principal component analysis. *Biometrika* 1971; **58**: 453–467.
9. Ramsay J, Silverman B. Functional Data Analysis. Springer: New York, USA, 1997.
10. Ramsay J, Silverman B. Applied Functional Data Analysis. Springer: New York, USA, 2002.
11. Ratcliffe S, Leader L, Heller G. Functional data analysis with application to periodically stimulated foetal heart rate data. I: Functional regression. *Stat Med* 2002; **21**: 1103–1114.
12. Pfeiffer R, Bura E, Smith A, Rutter J. Two approaches to mutation detection based on functional data. *Stat Med* 2002; **21**: 3447–3464.
13. Ramsay J, Ramsey J. Functional data analysis of the dynamics of the monthly index of nondurable goods production. *J Econom* 2002; **107**: 327–344.
14. Wang S, Jank W, Shmueli G. (2007) "Explaining and Forecasting Online Auction Prices and their Dynamics using Functional Data Analysis." Forthcoming at the Journal of Business and Economic Statistics.
15. Rossi N, Wang X, Ramsay J. Nonparametric item response function estimates with the EM algorithm. *J Educ Behav Stat* 2002; **27**: 291–317.
16. Ogden R, Miller C, Takezawa K, Ninomiya S. Functional regression in crop lodging assessment with digital images. *J Agric Biol Environ Stat* 2002; **7**: 389–402.
17. Verbeke G, Molenberghs G. Linear Mixed Models for Longitudinal Data. Springer: New York, USA, 2000.
18. Alsberg BK. Representation of spectra by continuous functions. *J Chemometrics* 1993; **7**: 177–193.
19. Saeys W, Mouazen AM, Ramon H. Potential for onsite and online analysis of pig manure using visible and near infrared spectroscopy. *Biosyst Eng* 2005; **91**(4): 393–402.
20. Eilers P, Marx B. Flexible smoothing with B-splines and penalties. *Stat Sci* 1996; **11**: 89–121.
21. Rossi F, Francois D, Wertz V, Meurens M, Verleysen M. Fast selection of spectral variables with B-spline compression. *Chemometrics Intelligent Lab Syst* 2007; **86**: 208–218.
22. http://software.eigenvector.com/Data/SWRI/index.html
23. Reiss PT, Ogden RT. Functional principal component regression and functional partial least squares. *J Am Stat Assoc* 2007; **102**(479): 984–996.
24. Trygg J, Wold S. Orthogonal projections to latent structures, O-PLS. *J Chemometrics* 2002; **16**(3): 119–128.
25. Alsberg BK, Woodward AM, Kell DB. Tutorial an introduction to wavelet transforms for chemometricians: A time-frequency approach. *Chemometrics Intelligent Lab Syst* 1997; **37**: 215–239.
26. Harrington PD, Rauch PJ, Cai CS. Multivariate curve resolution of wavelet and Fourier compressed spectra. *Anal Chem* 2001; **73**: 3247–3256.
27. Cuevas A, Febrero M, Fraiman R. An anova test for functional data. *Comput Stat Data Anal* 2004; **47**: 111–122.
28. Langsrud Ø. 50–50 Multivariate analysis of variance for collinear responses. *Statistician* 2002; **51**: 305–317.