Talanta 146 (2016) 155–165

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Selection of the most informative near infrared spectroscopy wavebands for continuous glucose monitoring in human serum

Mohammad Goodarzi*, Wouter Saeys

KU Leuven, Department of Biosystems, MeBioS Division, Kasteelpark Arenberg 30 Box 2456, B-3001 Leuven, Belgium

ARTICLE INFO

Article history:
Received 9 June 2015
Received in revised form
13 August 2015
Accepted 15 August 2015
Available online 18 August 2015

Keywords:
Diabetes
Continuous glucose monitoring
Human serum
Non-invasive measurements
Variable selection
Near infrared spectroscopy

ABSTRACT

By controlling the blood glucose levels of diabetics permanent diabetes-related problems such as blindness and loss of limbs can be delayed or even avoided. Therefore, many researchers have aimed at the development of a non-invasive sensor to monitor the blood glucose level continuously. As non-invasive measurements through the skin, the ear lobe or the gums have proven to be either unreliable or impractical, attention has recently shifted to minimally invasive sensors which measure the glucose content in serum or interstitial fluid. Thanks to the development of on-chip spectrometers minimally invasive, implantable devices are coming within reach. However, this technology does not allow to acquire a large number of wavelengths over a broad range. Therefore, the most informative combination of a limited number of variables should be selected. In this study, Interval PLS (iPLS), Variable Importance in Projection (VIP), Uninformative Variable Elimination (UVE), Bootstrap-PLS coefficients, Moving window, CorXyPLS, Interval Random Frog-PLS and combinations of these methods were used in order to address the question whether the short wave band (800–1500 nm), first overtone band (1500–1800 nm), the combination band (2050–2300 nm) or a combination of them is the most informative region for glucose measurements and which wavebands should be measured within these wavelength ranges. The three different data sets employed focus on the determination of (1) glucose in aqueous solutions over the 1–30 mM range in presence of urea and sodium D-lactate, (2) glucose in aqueous solutions over the 2–16 mM range in presence of icodextrin and urea and (3) glucose in human serum samples. The best results for the first, second and third data sets were obtained by selecting 40, 130 and 20 variables resulting in a PLS model with an RMSEP of 0.56, 0.59 and 1.5 mM, respectively. It was found that the first overtone band is most informative for aqueous solutions, while for glucose measurement of serum samples the combination band was found to be the better choice.

1. Introduction

Diabetes, a disorder in the control of the blood glucose level is considered to be one of the most important metabolic diseases worldwide [1]. As the natural control system regulating the blood glucose level is deficient in diabetes patients, diet adjustment and insulin therapy are needed to avoid toxic blood glucose levels leading to diabetes-related complications such as blindness and loss of limbs [2]. As the blood glucose level should be kept within the physiological range from 4 mM to 7 mM the diet and application of insulin should be based on frequent measurements of the blood glucose level. Nowadays, self-monitoring of blood glucose based on painful finger pricking is typically used. This method, however, is not suitable for tight insulin control, because isolated glucose values do not reflect the variations occurring throughout the day and night [3]. Ideally, the blood glucose level should be monitored continuously in an automatic, painless and non-invasive way. Therefore, many research initiatives have focused on the development of sensor systems for Continuous Glucose Monitoring (CGM) [4]. Among the investigated technologies, Near-Infrared Spectroscopy (NIR) coupled with chemometrics has received the most attention.

NIR spectroscopy is already widely used for composition measurement in different industries (e.g. food, chemical and pharmaceutical), because it is rapid, nondestructive and requires little or no sample preparation [5]. As the absorption in the NIR corresponds to overtones and combinations of the fundamental molecular vibrations, the absorption peaks are relatively weak and highly overlapping. This makes that there is typically no single wavelength variable which is only influenced by the analyte of interest, but there are many variables which are weakly correlated to it. Therefore, multivariate data analysis techniques are needed

* Correspondence to: Department of Biosystems, Faculty of Bioscience Engineering, Katholieke Universiteit Leuven – KU Leuven, Kasteelpark Arenberg 30, B-3001 Heverlee, Belgium. Fax: +32 16328590.

E-mail addresses: mohammad.godarzi@gmail.com, mohammad.goodarzi@biw.kuleuven.be (M. Goodarzi).

http://dx.doi.org/10.1016/j.talanta.2015.08.033
0039-9140/© 2015 Elsevier B.V. All rights reserved.

to extract the information on the analyte of interest from the acquired NIR spectra. The main goal of Multivariate Calibration is to establish a regression model linking the measured signals to certain properties of samples. Afterwards, the prediction ability of such model is validated by applying it to a set of samples, which are not involved in the calibration. Many Multivariate Calibration methods such as Multiple Linear Regression (MLR), Principal Component Regression (PCR), Partial Least Squares (PLS), Ridge Regression (RR), Artificial Neural Networks (ANN), Support Vector Machines (SVM), etc. have been used in NIR spectral analysis [6,7].

Although biased regression methods such as RR, PCR and PLS can be used with all measured wavelengths, many of these wavelength variables are not essential for predicting the analyte of interest. Therefore, a simplified model could be built by including only the most informative wavelengths [8–11]. By removing those irrelevant features the prediction performance and robustness of the calibration models can be improved [12–15].

As a consequence, variable selection/reduction techniques are used to: simplify the interpretation of models with few variables, to improve the prediction performance, and to decrease the risk of overfitting/overtraining [16,17].

Different strategies have been used to find the most informative part of NIR spectra for glucose measurement. These strategies can be divided in two categories: manual splitting of the spectra into several wavelength ranges followed by model building on each of these splits, and use of a variable selection technique (e.g. genetic algorithm). However, all the studies reported so far only considered one specific data set, which may have led to the selection of a variable set, which is only optimal for that data set. For instance, one group examined seven different spectral regions based on three different glucose bands around 2128, 2273 and 2325 nm for glucose measurement [18]. They reported that the most informative wavelength range is from 2174 to 2326 nm, while the prediction ability of the PLS model decreased by using the full spectrum. The same conclusions were drawn from glucose measurements under changing temperatures [19]. They concluded that the C–H combination band at 2273 nm is the most useful band for glucose measurement [18,20–23]. On the other hand, in [24] the combination of four informative regions of 1373–1429, 1495–1545, 1565–1696, and 1790–1805 nm was found to be most informative. In [25], a comparison was made between PLS models built using the first overtone (1500–1800 nm) and combination band (2050 to 2300 nm) to measure glucose, lactate, urea, ascorbate, triacetin, and alanine in aqueous solutions [25]. The PLS models for glucose resulted in Standard Errors of Prediction (SEPs) of 1.12 and 0.45 mM for the first overtone and combination band spectra, respectively. In yet another study, an aqueous solution consisting of glucose, lactate, urea, ascorbate triacetin and alanine was used [26–28]. In [26], the best region was reported to be from 2036 to 2324 nm, while in [27], a combination of the regions from 1473 to 1831 nm and from 2111 to 2374 nm provided the best glucose prediction accuracy. More details on the different studies done for selection of the most informative wavelength region can be found in [3], where the multivariate calibration, variable selection and preprocessing strategies for blood glucose measurement by NIR spectroscopy are critically reviewed.

Although many researches have aimed to find the most informative wavebands in the NIR range for CGM, they came to different conclusions. This suggests that the selected set of wavebands might depend on the used variable selection technique and the concentrations of the spectral interferents in the considered samples. Therefore, in this study we applied different variable selection techniques on 3 different data sets with different composition structure to clarify which wavebands in the NIR region are most informative for glucose measurement in the presence of the major interferents which can be expected in the serum of diabetics.

2. Theory

2.1. Interval PLS (iPLS)

Interval PLS is a technique, which was proposed by Nérgaard et al. [29], and develops local PLS models on equidistant subintervals of the full-spectrum region. Interval PLS (iPLS) selects a subset of variables which give superior prediction performance compared to when all the variables are used. This approach can be applied either based on a forward or reverse strategy. In any case, the algorithm performs a sequential, exhaustive search for the best variable or combination of variables by splitting the spectra into small intervals with equal distance and fitting a PLSR model to each combination of intervals. The combination of intervals, which gives the lowest prediction error, is then selected. In forward iPLS the single interval which gives the lowest prediction error in cross validation (RMSECV) is selected in the first step and in each subsequent step the interval which gives the largest improvement in the prediction error is added to the set of intervals which have been selected in the previous steps. In reverse iPLS all intervals are included in the first step and in every subsequent step the interval is removed for which removal leads to the largest reduction of the prediction error in cross validation [29].

2.2. Variable importance in projection (VIP)

The VIP method estimates the importance of each variable based on the weight of the loading factors from each component. The VIP scores can be calculated as follows:

$$\mathrm{VIP}_j = \sqrt{\frac{n\sum_{a=1}^{k}[(q_a^2 \grave{t}_a t_a)(\frac{w_{aj}}{\|{w_a}^2\|})]}{\sum_{a=1}^{k} q_a^2 \grave{t}_a t_a}} \quad a = 1, 2, \ldots k \tag{1}$$

where n is the number of predictor variables, $\boldsymbol{w}_{aj}$ is the loading weight for variable j using a components and $\boldsymbol{w}_a$, $\boldsymbol{q}_a$ and $\boldsymbol{t}_a$ are PLS loading weight, $\boldsymbol{y}$-loadings and scores, respectively, corresponding to the a^{th} component. A variable with a squared VIP score close to or above 1 is then considered as important for the model and thus selected for further use [30].

2.3. Uninformative variable elimination (UVE)

UVE aims at the reduction of the dimension of a data set by discarding uninformative variables (i.e. those that have high variance, but small covariance) [31]. By building a PLS model with the full set of variables, the optimum number of latent variables is determined. Then, the calibration data matrix $\mathbf{X}$ is augmented with a matrix of the same size where the variables contain Gaussian random noise multiplied with a small constant. The purpose of this addition is to eliminate any possible interaction with the original variables of the $\mathbf{X}$ matrix. Centner et al. [31] suggested that the constant value should be lower than the level of inaccuracy of the instrument. For instance, when the magnitude of the real variables is in the range of 0.0–1.0, the constant value is selected as an order of magnitude smaller than the imprecision of the instrument, i.e., 1×10^{-4}. A collection of PLS regression models is then built on this augmented data matrix by performing leave-one-out crossvalidation. A normalized regression coefficient value is then calculated for every variable by calculating the mean value and dividing it by the standard deviation. Although there are many options to tune the threshold, the most commonly used method is to determine the cut-off value as the maximum normalized regression coefficient value for the Gaussian random noise variables. This method was also used in this study. Finally, all spectral variables with a normalized regression coefficient equal to or smaller than the cut-off value are considered to be uninformative and eliminated.

In Monte Carlo UVE, the leave-one-out cross validation is replaced by iterations of random sampling of *M* calibration samples [32].

2.4. Bootstrap-PLS coefficients

Bootstrapping involves sampling multiple datasets, each consisting of *n* observations randomly selected with replacement from the original data set and building a PLS regression model for each bootstrap sample. Based on all bootstrap samples, confidence intervals for the PLS regression coefficients can be estimated. Wavelength variables whose 95% confidence intervals for the PLS regression coefficients include the zero value are considered not significantly different from zero and thus eliminated [33]. This corresponds to a hypothesis test where one investigates whether the regression coefficient is significantly different from zero. If there is a significant linear relationship between the dependent ($\boldsymbol{y}$) and independent ($\boldsymbol{X}$) variables, the slope will not be zero. The null hypothesis states that the slope is equal zero (H_0: $\boldsymbol{b}=0$) or in alternative hypothesis states that H_a: $\boldsymbol{b}$ different from zero. In order to accept or reject the defined hypothesis,'a significance levels of 0.05' was used.

2.5. Moving window-PLS

This technique starts with fixing a window, which moves throughout the entire spectrum. At each position, a PLS model is built for which the RMSECV based on Cross Validation is calculated. In the next step, the prediction error of each model is plotted against the window position to select the region for which the corresponding model gave an acceptable RMSECV level with the lowest number of latent variables [34].

2.6. CorXyPLS

This technique calculates the mutual correlation (correlation coefficients) between each of the independent variables (wavelength variables in this case) and the dependent variable (response variable). Then, those with the highest correlation coefficients are selected as best variables. Although this method may be intuitively appealing, it may not always lead to the selection of the best combination of variables. This one variable at the time selection strategy might be suboptimal when the variables are highly correlated, as is the case for NIR spectra, because it might select highly correlated variables which have very limited added value with respect to the others.

2.7. Interval random frog-PLS

The Interval Random Frog (IRF) method investigates all possible combinations of spectral intervals to find the best set of informative variables. At first, the spectra are split based on interval value and then these intervals are ranked applying random frog-PLS and the optimal ones are chosen. This algorithm works in an iterative manner in three steps: First, a variable subset $\boldsymbol{X}_0$ containing *n* variables is randomly initialized from a given data set. Then, based on $\boldsymbol{X}_0$, a candidate variable subset $\boldsymbol{X}^*$ consisting of *m* variables is proposed. Afterwards, accept $\boldsymbol{X}^*$ with a certain probability as $\boldsymbol{X}_1$, and replace $\boldsymbol{X}_0$ using $\boldsymbol{X}_1$. This step continues until the predefined number of iterations has finished. The last step is to calculate a selection probability for each variable, which can be used as a measure of variable importance. For details on this technique the reader is referred to [35].

3. Experimental

3.1. Data set-1

People with diabetes who are also peritoneal dialysis patients are treated with icodextrin. Icodextrin is a glucose polymer and is used as an osmotic agent in dialysis solutions since glucose solutions are not suitable for these patients and may furthermore lead to metabolic complications [36]. As the absorption spectrum of icodextrin is highly similar to that of glucose, it is a spectral interferent for NIR spectroscopic measurement of glucose, which might lead to an over-estimation of the glucose level if the calibration models are not able to distinguish its spectral signature from that of glucose [37]. Therefore, a data set was designed to investigate the possibility to extract glucose information independent of the icodextrin concentration. Fourty-five aqueous solutions of glucose, icodextrin and urea were prepared based on a full factorial design covering the physiological ranges of glucose (5 levels: 2 mM, 4 mM, 8 mM, 12 mM and 16 mM), urea (3 levels: 3 mM, 6 mM and 9 mM) and icodextrin (2 levels: 50 and 200 mg/dL). NIR spectra in the 800–2500 nm range were acquired for these samples with a Bruker MPA FT-NIR spectrometer (Bruker, Ettlingen, Germany) in transmission mode using a 1 mm cuvette. The samples were preheated to 37 °C by placing them in a water bath and the temperature of the cuvette was controlled at 37 ± 1.0 °C. The spectra for each sample were recorded in triplicate resulting in 252 spectra for 84 aqueous solution samples. This data set consists of 45 samples, which were each measured three times resulting in 135 spectra. The data set was split into a calibration set of 31 samples and a test set of 14 samples.

3.2. Data set-2

This data set consists of NIR spectra of three parallel sets of aqueous solutions containing similar concentrations for glucose (1, 3, 7, 12, 15, 22 and 30 mM), urea (5 and 6 mM) and D-lactate (1 and 5 mM) [38]. A full factorial design of these concentrations was prepared resulting in 28 samples for every set of aqueous solutions. In total, 84 samples were produced for three sets. NIR spectra in the range 800–2500 nm were acquired for these samples with a Bruker MPA FT-NIR spectrometer (Bruker, Ettlingen, Germany) with a 1 mm transflectance probe. All the measurements were carried out in a temperature controlled facility at 37 ± 1.0 °C. The spectra for each sample were recorded in triplicate resulting in 252 spectra for 84 aqueous solution samples.

3.3. Data set-3

The NIR spectra of human blood serum are the result of overlapping strong absorption bands of not only water, but also proteins, which can affect the prediction ability of a multivariate calibration model built for glucose measurement. In a previous study, we investigated the effect of the total protein concentration and the glycated protein concentration in blood serum on the prediction ability of PLSR calibration models [39]. Overall, it was concluded that the glycated serum proteins do not affect the accuracy of the glucose prediction from the NIR spectra of human serum. The main aim of this study was to find the most informative wavebands in the NIR region for glucose monitoring in human serum. So, while the two previous data sets were designed to investigate the impact of specific interferents, this human serum data set was used to evaluate if the same wavebands would be selected for real human serum samples.

Fifty human serum samples were selected from a large database of human serum samples with special attention to cover the glycated hemoglobin (HbA1c) range as widely as possible (in the present case, 24–105 mM) [39]. The HbA1c is developed when the hemoglobin in the blood becomes "glycated" by joining glucose in the blood. This glycation process is promoted by the high blood glucose levels, which can occur in diabetics, such that the HbA1c concentration can be seen as a memory of the blood glucose level history of a patient. The glucose molecule attached to the hemoglobin can no longer be used as energy source by the cells and thus should not be included in the blood glucose level. So, the multivariate calibration models should be able to distinguish between the spectral signature of the HbA1c and that of glucose. Therefore, special attention was paid to cover a wide range of HbA1c in this experiment. The samples were frozen at −18 °C to preserve their chemical composition and integrity until the time of spectral measurement. The serum samples and their reference concentration measurements for glucose, HbA1c, and total proteins were obtained from the Clinical Chemistry Laboratory, University Hospital (UZ), Ghent, Belgium. The NIR spectra of these samples were measured with a Bruker MPA FT-NIR spectrometer (Bruker, Ettlingen, Germany) in transmittance mode in a 1 mm flow through cuvette. All the measurements were carried out in a temperature controlled cell at 37 ± 1.0 °C. Among all 150 measured spectra, only one sample was found to have a measurement artifact due to the instrumental error and was removed from the data set.

4. Data analysis

The first two data sets contain NIR spectra measured from designed aqueous solutions with specific interfering structures, while the last one contains NIR spectra of real human serum samples. In all three data sets glucose is the analyte of interest. For the first data set, urea and icodextrin and in the second data set, urea and D-lactate are interferents.

The NIR spectra were split into three major areas: Short wave, First overtone and combination band ranging from 780 to 1500 nm, 1500 to 1800 nm and 2050 to 2300 nm, respectively. As shown in Fig. 1, in order to investigate the effect of different regions of NIR spectra on glucose measurement, seven subsets were created from each data set. The first one consists of the full range of NIR spectra ranging from 800 to 2500 nm. Note that for all sub sets the range from 1850 to 2050 nm, which corresponds to a characteristic absorption peak of water related to the combination band of O–H, was removed from the data set. The second set consists of the wavelength variables below 1500 nm belonging to the short wave range. First overtone and combination band were from 1500 to 1800 nm and 2050 to 2300 nm, respectively. The fifth, sixth and seventh subsets are combinations of short wave, first overtone and combination band.

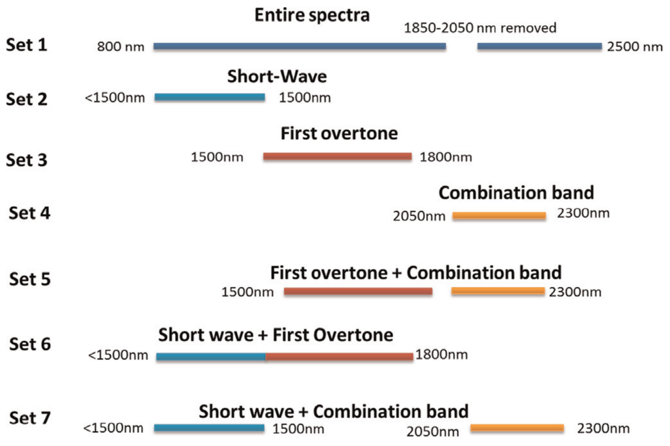

Fig. 1. This figure Illustrates the strategy to set different subsets which applied on the three data sets.

In spite of the fact that preprocessing techniques could help to improve the robustness of calibration models, no other preprocessing except mean-centering was applied for modeling as each of the preprocessing techniques would need to be optimized and may have an impact on the variable selection by redistributing information from one variable to another one.

The duplex method was used to split the three data sets for 70% and 30% into a calibration and test set, respectively. The duplex technique aims at setting up a calibration set which is as representative as possible and covers the calibration (***X*** and ***y***) space as uniformly as possible. This technique was applied on the **X** matrix and after the data had been split, the range of ***y*** values for both calibration and test sets were checked in order to assure their uniformity. It starts by assigning the two samples, which are farthest away from each other to the calibration set. Then, among the remaining samples, the two samples, which are farthest away from each other are placed into the test set. This procedure continues until all samples have been assigned to the calibration and test sets.

Each model built on the calibration set was internally validated using a 10-fold Contiguous Block Cross Validation (CBCV) method. This cross validation method was selected to keep all three spectral replicates for the same sample together. The 10-fold CBCV was used to ensure that the model was trained for the available variation in the calibration set during the training phase. After tuning the model complexity in cross validation the performance of the resulting model was tested on a separate test set consisting of 30% of the entire data set, which had not been used for building or tuning the model. For all data sets, the prediction performance of PLS regression models built on different subsets of variables selected with the variable selection techniques described in Section 2 was compared to the performance of a PLS regression model using the full spectral range. The Root Mean Squared Error of Calibration (RMSEC), Cross Validation (RMSECV) and Prediction (RMSEP) were used as the performance criteria to assess and compare the predictive ability of the different models.

All calibrations were performed in MATLAB®, 7.10.0 (R2010a) (The Mathworks, Natick, MA, USA). For PLS regression, the PLS toolbox version 7.8.2 was used (Eigenvector Research, Wenatchee, WA, USA).

Matrices are shown in bold capital letters (e.g. **X**), vectors in bold lowercase (e.g. **x**) and scalars in italic characters (e.g. *x*). Note that in order to illustrate the distribution of samples, at least described by spectra recorded in the entire spectral range, score plots of the first two PCs (including replicated samples) are shown in Fig. 1s in supplementary materials.

The application of different variable selection strategies to the different data sets generated a large number of graphs. Due to space limitations, only the final results are presented here. For more details the reader is referred to the supplementary material. All figures with a subscript "s" can be found in the supplementary material.

5. Results and discussion

5.1. Data set-1

The best PLS model for the full spectral range, excluding the 1850–2050 nm region gave RMSECV and RMSEP values of respectively 3.98 mM and 2.63 mM. By applying forward iPLS only 40 from the 2068 wavelength variables were selected to obtain RMSECV and RMSEP values of 0.56 mM and 0.56 mM, respectively with a PLS model using 8 LVs. Although the selected wavelength variables are distributed over the SW, FOT and CB, they are mainly situated in the combination band over the range from 2050 to 2200 nm. In Fig. 2a the RMSEC, RMSECV and RMSEP values obtained by different variable selection techniques coupled with PLS are shown. In Fig. 2b the spectral regions selected by the different variable selection techniques are illustrated. As can be seen in Fig. 2, all variable selection techniques except VIP resulted in a better prediction performance than that obtained by the PLS regression model built on the entire spectra. This indicates that some wavelength variables are detrimental for the prediction performance. However, as all other techniques did not give a better prediction performance, while selecting a larger number of variables, the set of wavelength variables selected by forward iPLS was chosen for this.

The second subset, which was tested, is the shortwave range from 800 nm to 1500 nm consisting of 1511 spectral points for this data set. The PLS regression model using the full spectrum, gave RMSECV and RMSEP values of respectively 1.07 mM and 1.11 mM. Forward iPLS with an interval width of 5 wavelength variables resulted in the selection of 80 wavelength variables. A PLS regression model with 5 LVs built on this selection resulted in a large improvement in the prediction performance with RMSECV and RMSEP values of respectively 0.62 and 0.76 mM. This prediction performance is only slightly worse than that obtained with the variables selected from the full spectral range by forward iPLS. Note that the variables selected by all techniques mainly range from 1300 nm to 1500 nm (Fig. 2s supplementary material).

A PLS model was built based on the third subset consisting of only the FOT region. For conventional PLS calibration, the RMSECV was 1.43 mM for 6 LVs and the corresponding RMSEP value was 1.21 mM when only 289 variables (first overtone) were used. For this third subset, bootstrapPLS showed to select fewer variables (112), resulting in RMSECV and RMSEP values of 1.07 mM and 0.9 mM, respectively. The spectral range from 1700 to 1800 nm was selected by all techniques (Fig. 3s, supplementary materials). This underlines the importance of this range for glucose measurement. In the NIR region, the first harmonic of the C–H stretching vibrations is observed at about 1700 nm. In order to get an idea about where the glucose information is higher than that of water, their molar absorptivities have been calculated. The molar absorptivities of water and glucose in the first overtone band and combination bands are presented in Fig. 4s. These pure component spectra of glucose and water were measured experimentally by following the method described by Amerov et al. [40]. In Fig. 4s, it can be seen that the region between 1500 and 1850 nm is an important region for glucose measurement since glucose has a strong absorbance in that region while water absorbance is rather weak. Therefore, it could be a suitable range to extract glucose information.

A fourth subset containing only combination band information was used to investigate the value of this region for glucose measurement in the presence of icodextrin and urea. A PLS model was built on this subset of 139 variables resulting in an RMSECV and RMSEP of 1.51 mM and 1.55 mM, respectively for 6 LVs. Forward iPLS was found to select only 50 variables and to obtain similar prediction performance with 5 LVs (RMSECV = 1.47 mM and RMSEP = 1.61 mM). The wavelength variables selected by the different techniques were not limited to a specific range in the combination band, but spread over the range from 2050 to 2300 nm. Combination bands involving stretching and bending of the C–H bond may be identified between 2000 and 2500 nm and, with a lower intensity in the short wave range, between 1300 and 1440 nm. The molar absorptivities of water and glucose in the combination band are also presented in Fig. 4s [40].

By combining both FOT and CB, 428 variables were selected on which a PLS model was built which resulted in an RMSECV of 1.32 mM and an 'RMSEP 1.19 mM'. In this case the IRF-PLS selected 125 variables to obtain an RMSECV of 0.83 mM and an RMSEP of 0.65 mM. As can be seen in Fig. 5s (supplementary material) all different variable selections resulted in a rather similar model performance. Although some techniques selected slightly less variables (e.g. Forward iPLS selected 120 variables), the prediction performance of the model built on the variable set selected with IRF-PLS was better.

A PLS model with 6 LVs using all variables in the short wave and first overtone regions gave RMSECV and RMSEP values of 0.54 mM and 0.48 mM, respectively. Although the prediction performance is very good, this model uses 1800 variables. By

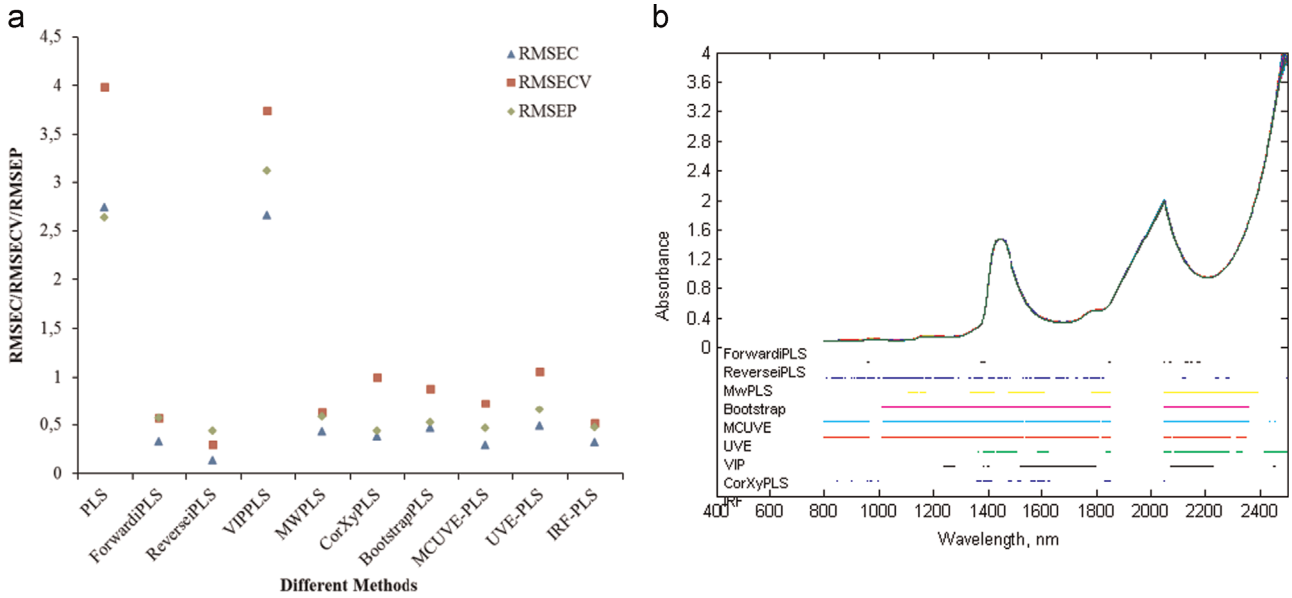

Fig. 2. (a) The figure represents the RMSEC, RMSECV and RMSEP for the first variable subset of the first data set. (b) This figure shows graphically which part of spectra is selected the most by different techniques for the first set of the first data set.

performing forward iPLS on this subset, 85 variables were selected to obtain an RMSECV of 0.44 mM and an RMSEP of 0.41 mM with a PLS model with 5 LVs. It can be seen in Fig. 6s in supplementary material that forward iPLS selected the smallest number of variables, while the prediction performance significantly improved in comparison with that of PLS using all variables from this subset. It can be seen that the selected variables are mainly situated in the FOT band.

The last subset that was defined for this data set was built by combining all 1650 variables from the SW and CB range. The RMSECV and RMSEP values obtained with a PLS model built on this subset were respectively 0.88 and 0.61 mM. All variable selection techniques resulted in an improvement in the prediction performance. Also in this case forward iPLS selected the smallest number of variables i.e. 200. The PLS model built using these selected variables resulted in an RMSECV of 0.57 mM and an RMSEP of 0.43 mM, which is considerably lower, but not significantly different from that which is obtained by a PLS model using all 1650 variables of this subset. In Fig. 7s (supplementary materials) it can be seen that the prediction performance of all built models using different variable selection methods is rather similar here as well. However, all other techniques selected a larger number of variables to obtain a prediction performance similar to that obtained by forward iPLS.

Based on the Tukey Honestly Significant Difference (HSD) multiple comparison test, it was shown that the results obtained by a PLS model built on the entire spectra is significantly different from all other models (Fig. 9s). In Table 1 is illustrated where the result of PLS was statistically significantly different from other techniques using various subsets of the first data set.

If we compare the results obtained for the different subsets we can conclude that ForwardiPLS typically selected the smallest number of variables. In most cases, the prediction performance either significantly improved or stayed similar with that of PLS using the entire spectral range or with those which were obtained with other variable selection techniques.

Prediction performance and number of variables selected by the best performing methods on the different subsets are summarized in Fig. 8s and Table 2.

This table shows that the best prediction performance was obtained by a PLS model with 5 LVs using 85 variables selected from the short wave and first overtone band. However, it should be noted that a smaller number of variables was selected for the first subset (SW+FOT+CB) and the results are not significantly different (Fig. 9s). Using only variables from the combination band (CB) led to worse results. This suggests that the interference by icodextrin is the highest in this region. The wavelength variables, which have been selected from the different subsets, are illustrated in Fig. 3. As can be seen in this figure, most methods selected variables from the FOT and CB region. Here, the HSD multiple comparison test was applied to all “best” models from different subsets and compared to a PLS model built using the entire spectral range. It was found that the average prediction error for the PLS model built on the full range was significantly higher from that of the models built on the subsets. On the other hand, application of Forward iPLS on the fourth subset also gave an average prediction performance which was significantly lower than that obtained with the other techniques. This is in agreement with Fig. 8s where the RMSEC, RMSECV and RMSEP obtained from the fourth subset are shown to be much higher than those obtained for other subsets. Therefore, a combination band alone cannot be used for this data set while a quite good prediction performance based on FOT or SW alone was obtained. However, it helped to combine the information present in different regions together.

Table 1
Tukey honestly significant difference (HSD) multiple comparison test based on two way ANOVA.

	Data set 1	Data set 2	Data set 3
Set 1	VIP	All techniques	CorXy
Set 2	ForwardiPLS, ReverseiPLS and MwPLS	–	Bootstrap, MwPLS, CorXy
Set 3	–	Bootstrap	CorXy
Set 4	–	VIP and CorXy	–
Set 5	ForwardiPLS, ReverseiPLS, MwPLS and IRF	MCUVE, UVE, VIP, CorXy and IRF	–
Set 6	–	MCUVE, UVE, CorXy	CorXy, Bootstrap
Set 7	MwPLS	ReverseiPLS	forwardiPLS, CorXy

Significance level was 5%.

Table 2
This table illustrates the best-selected subsets from each data set and the model built. This table indicates which technique resulted in the optimal subset. Number of variables, RMSEC, RMSECV, RMSEP and latent variables (LVs) are shown in this table. The RMSEC, RMSECV and RMSEP are in mM.

Used techniques	No. variables	RMSEC	RMSECV	RMSEP	LV
First data set					
Forward iPLS (set 1)	40	0.32	0.56	0.56	8
Forward iPLS (set 2)	80	0.54	0.62	0.76	5
BootstrapPLS (set 3)	112	0.90	1.07	0.90	5
Forward iPLS (set 4)	50	1.27	1.47	1.61	5
IRF-PLS (set 5)	125	0.56	0.83	0.65	7
Forward iPLS (set 6)	85	0.33	0.44	0.41	5
Forward iPLS (set 7)	200	0.33	0.57	0.43	6
Second data set					
Forward iPLS (set 1)	190	0.47	0.59	0.82	9
Forward iPLS (set 2)	190	1.13	1.27	1.50	7
Forward iPLS (set 3)	120	1.39	1.81	1.95	6
MCUVE-PLS (set 4)	82	1.42	1.54	1.59	4
IRF-PLS (set 5)	130	0.49	0.55	0.59	6
Forward iPLS (set 6)	190	0.56	0.67	0.78	8
IRF-PLS (set 7)	197	0.87	0.99	1.16	7
Third data set					
Forward iPLS (set 1)	50	0.89	1.33	1.91	10
Forward iPLS (set 2)	25	2.02	2.53	2.81	7
Forward iPLS (set 3)	25	1.79	2.32	3.27	10
CorXyPLS (set 4)	35	1.16	1.42	1.83	7
Forward iPLS (set 5)	20	1.12	1.77	1.51	10
IRF-PLS (set 6)	92	1.13	1.61	1.78	8
Forward iPLS (set 7)	30	1.00	1.33	1.57	8

5.2. Data set-2

The first subset consists of 2068 variables and a PLS model using all these variables resulted in an RMSECV of 4.87 mM and an RMSEP of 4.34 mM were obtained. The best result was obtained using forwardiPLS as variable selection technique. This technique selected 190 wavelengths resulting in a considerable decrease in-the prediction error with RMSECV and RMSEP values of 0.59 mM and 0.82 mM, respectively. This indicates that some of the variables in the original data set are detrimental for the prediction performance. The performance of all variable selection methods on this variable subset of the second data set was rather similar except for VIP, which gave the worst results (Fig. 10s supplementary material). All techniques led to a higher number of selected variables and thus their results were not chosen for further study. It is seen that all techniques mainly selected variables from the short wave (SW), the first overtone (FOT) and the combination band (CB) regions. Note that Forward iPLS selected mainly

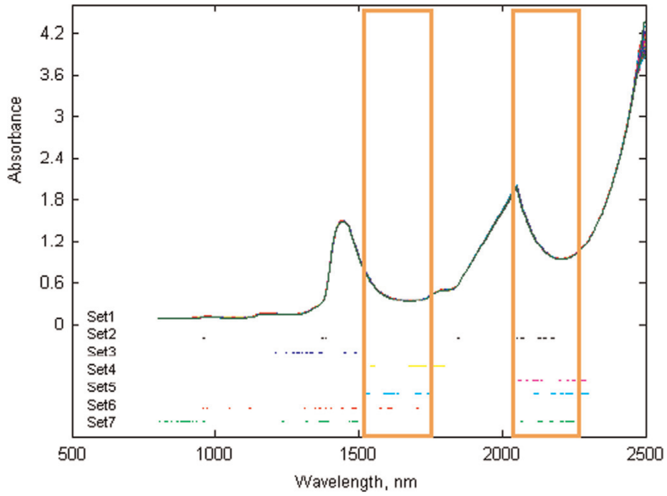

Fig. 3. This figure indicates that the selected variables for all seven subsets from the first data set.

wavelength variables from the FOT region. This suggests that accurate glucose prediction might also be possible based on the FOT alone.

The second subset consists of all variables below 1500 nm, known as the Short Wave region. This set consists of 1511 variables. A PLS model was built using all variables resulting in an RMSECV and RMSEP of 1.33 mM and 1.48 mM, respectively when 9 latent variables were used. The same as above, forwardiPLS selected the smallest number of variables (190) to obtain comparable RMSECV and RMSEP values of 1.27 mM and 1.50 mM, respectively. It can be seen from Fig. 11s in the supplementary material that the prediction performance for the other variable selection techniques was similar, but they all selected a larger number of variables than for forward iPLS.

The third subset consists of only the variables in the first overtone region ranging from 1500 to 1800 nm (289 variables). The RMSECV and RMSEP obtained for the PLS model using all variables from this subset were respectively 2.33 mM and 2.12 mM. Forward iPLS selected 120 variables resulting in RMSECV and RMSEP values of 1.81 mM and 1.95 mM, respectively. Bootstrap PLS selected only 36 variables, but this led to worse RMSECV and RMSEP values. The results obtained using different techniques are summarized in Fig. 12s (supplementary material). It can be seen that less latent variables are needed when only the first overtone region is used in comparison to that of short wave and entire spectra range for this subset. However, by using only this range the prediction error increased by 33% compared to using variables selected from the full spectral range.

The combination band spectra consisting of 139 wavelengths variables were used as the fourth subset for analysis. In Fig. 13s in the supplementary material the prediction performances of the different variable selection techniques on this data set are summarized.

A PLS model using all 139 variables required 4 latent variables resulted in an RMSECV and RMSEP of 1.77 mM and 1.77 mM, respectively. As it can be seen, the built model using the entire CB is less predictive than that of Short Wave, but better than that of FOT and entire spectra range. Note that in this region only 139 variables are available while in the Short Wave 1511 variables are available. On the other hand, only 4 latent variables are used by the PLS model, while a PLS model using the entire short wave region required 9 latent variables. Monte Carlo UVE showed to perform the best for this subset by selecting only 82 variables, resulting in RMSECV and RMSEP values of 1.54 mM and 1.59 mM, respectively.

In order to investigate whether a combination of variables selected from different spectral regions improves he prediction performance, different combinations of SW, FOT and CB were evaluated. A combination of FOT and CB includes 428 variables. A PLS model with 5 Latent variables using all these wavelength variables resulted in an RMSECV of 1.13 mM and an RMSEP of 1.20 mM. For this subset IRF-PLS selected the smallest number of variables (130 variables) resulting in an RMSECV and RMSEP of 0.55 mM and 0.59 mM, respectively. This clearly shows that the removal of irrelevant information improved the prediction performance significantly. It can be seen from Fig. 4 that the IRF-PLS method selected mainly variables from the first overtone region. Although the other methods resulted in a similar performance for this data set, they selected a higher number of variables. On the other hand, the HSD test indicated that the results obtained on subsets 1, 5, 6 and 7 are not significantly different from each other, while their prediction abilities are statistically significantly different from a PLS model built using the entire spectra and those

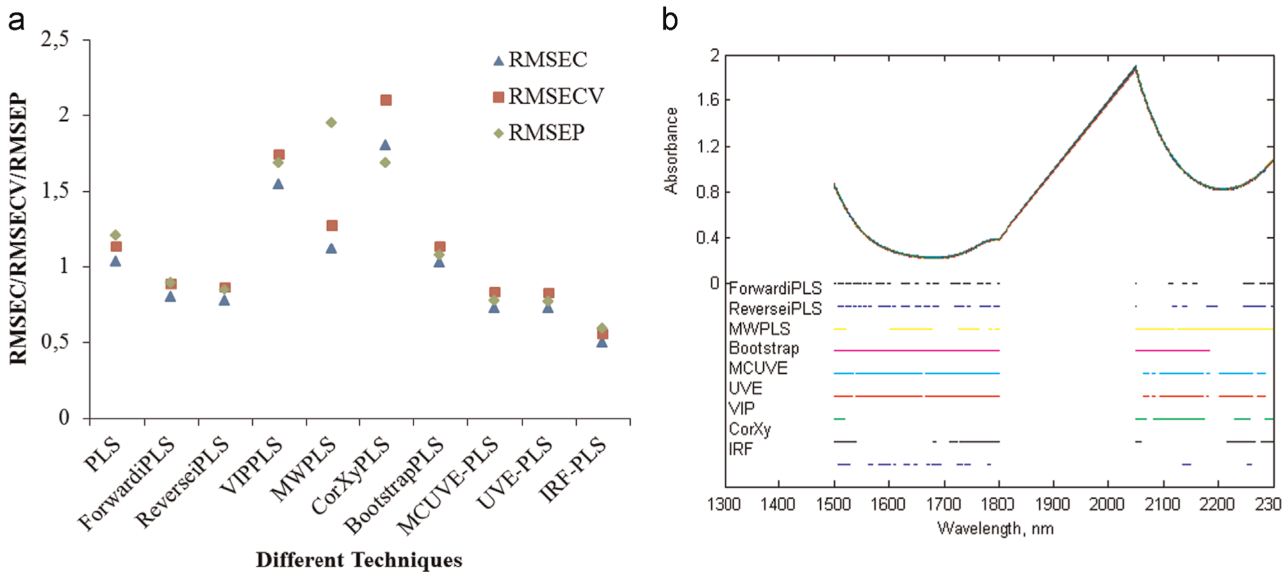

Fig. 4. (a) This figure displays the RMSEC, RMSECV and RMSEP for the fifth subset of the second data set. (b) This figure displays graphically which part of spectra is selected the most by different techniques for the fifth subset of the second data set.

models built using subsets 2, 3 and 4.

The sixth subset corresponds to the combination of Short Wave and FOT information of the second data set. This subset consists of 1800 variables resulting in an RMSECV and RMSEP of 1.02 mM and 1.03 mM, respectively. By selecting 190 variables using forwardiPLS, the RMSECV and RMSEP reduced respectively to 0.67 mM and 0.78 mM. As can be seen from Fig. 14s in the supplementary material, the forward iPLS method selected mainly the variables from the FOT region rather than from the SW region. It can be concluded that significant information corresponding to glucose is present in the FOT region.

The last subset of this data set is a combination of SW and CB consisting of 1650 variables. A PLS model using 7 latent variables resulted in an RMSECV and RMSEP of 1.20 mM and 1.36 mM, respectively. IRF-PLS selected only 197 variables resulting in an RMSECV and RMSEP of 0.99 mM and 1.16 mM, respectively. This method selected mainly the variables from the CB region rather than those from the SW region as can be seen in Fig. 15s in the supplementary material.

The best results obtained for all subsets are summarized in Fig. 16s and Table 2. The combination of FOT and CB regions (consisting of 197 selected variables) led to the lowest RMSECV and RMSEP values, while the number of latent variables used was also lower than for other combinations.

In Table 1 the results of the HSD multiple comparisons test for all different subsets are shown. This test indicated that the average prediction error of the PLS model was significantly different from that obtained with MCUVE, UVE, VIP, CorXy and IRF. On the other hand, all best models from each subset were compared to a PLS model built using the entire spectral range (Fig. 17s in supplementary material). It was found that the PLS model using all variables was significantly different in average prediction error from all other built models. Moreover, the performance of the models built based on the second, third and fourth subsets are significantly different from those built on the first and fifth subsets. This indicates that a combination of wavebands from the first overtone and combination band leads to the best performance.

5.3. Data set-3

The first variable subset covers the full spectra and consists of 711 variables. All PLS models coupled with variable selection for this set, require 7–10 LVs. The PLS calibration built using all wavelength variables resulted in an RMSECV of 1.85 mM and an RMSEP for the independent test set of 2.45 mM. After reducing the number of variables to 50 by forward iPLS, the RMSECV and RMSEP values improved to 1.33 mM and 1.91 mM, respectively. It can be seen in Fig. 18s (supplementary material) that the different techniques selected largely the same variable ranges, but that Forward iPLS selected considerably less variables than the others. These variables are mainly selected variables from FOT between 1600 and 1800 nm and CB from 2050 to 2200 nm. It is obvious that any model should lead to the smallest possible error of prediction. However, a prediction error of 1 mM would be acceptable for a new NIR spectroscopy based device used for CGM.

The second subset, which only covers the short wave spectral range, consists of 432 wavelength variables. A PLS model with 11 LVs using all wavelength variables results in RMSECV and RMSEP values of 2.41 mM and 1.92 mM, respectively. The prediction performance using this subset is only marginally worse than that obtained by a PLS model built on the entire spectral range. This indicates that a PLS model built using only the FOT region roughly has the same prediction performance as when it is built using the entire spectral range. The prediction performance is decreased in comparison to that of a PLS model built using entire spectra. Forward iPLS selected only 22 variables to obtain a worse prediction performance with RMSECV and RMSEP values of respectively 2.53 mM and 2.81 mM. This can show that less or insufficient glucose information can be found in the SW for building a robust PLS model. Although all techniques reduced the number of variables for this subset to less than 200 variables, the prediction performance was also reduced in comparison with when all variables in this range were used (Fig. 19s).

The FOT consisting of 144 variables was used as the third subset. The RMSECV and RMSEP for the full model built on this subset were respectively 2.85 mM and 3.74 mM. Forward iPLS selected only 25 relevant wavelengths with which an RSMECV and RSMEP of 2.32 mM and 3.26 mM were obtained, respectively. Some of the techniques led to a slightly better Cross Validation error than the other methods. However, all of them selected not only a larger number of variables, but they either resulted in worse RMSEC, RMSECV or RMSEP values than those obtained with Forward iPLS (Fig. 20s).

Next, the importance of the CB region was examined for the prediction of the glucose content in human serum samples. This region consisted of 70 variables and a PLS model with 6 LVs resulted in RMSECV and RMSEP values of respectively 1.60 mM and 2.05 mM. These prediction errors comparable to those obtained for the full spectrum show the importance of the CB region for glucose measurement in human serum samples. By reducing the number of variables to 35 using the CorXy method, the best PLS model was obtained, resulting in RMSECV and RMSEP values of 1.42 mM and 1.83 mM, respectively. As can be seen in Fig. 21s, the region between 2050 and 2200 nm is the most informative for glucose measurement in human serum. This motivated us to investigate the combination of different ranges with the question whether the prediction performance can be improved by including variables from an additional range. Therefore, the same strategy as for two other data sets was also followed here. From the results of the HSD multiple comparisons test (Table 1), it can be concluded that the prediction performance of a PLS model built using the entire spectral range is not significantly different from all techniques except from CorXy, Bootstrap and MwPLS for different subsets.

The combination of the FOT and CB ranges consists of 214 variables. A PLS model built on these variables resulted in RMSECV and RMSEP values of 1.75 mM and 1.73 mM, respectively. This shows that the prediction performance can be improved by adding some glucose information from the FOT region to that of the CB. More importantly, by reducing the number of variables to only 20 forward iPLS was able to obtain RMSECV and RMSEP values of respectively 1.77 mM and 1.50 mM. As can be seen in Fig. 5, two spectral ranges around 1600 and 1700 nm which belong to the FOT and two spectral ranges around 2100 and 2200 nm belonging to the CB were selected by the forward iPLS method. These results indicate that a combination of FOT and CB certainly leads to a model with higher prediction performance. However, a combination of SW and FOT range consisting of 576 variables led to a PLS model with RMSECV and RMSEP of 1.85 mM and 2.24 mM, respectively. It is seen that the prediction performance is deteriorated in comparison to that of using the FOT and CB ranges. In this case, IRF-PLS showed to select the smallest number (92) of informative wavelengths for this subset. The built PLS model using those 92 variables resulted in an RMSECV of 1.61 mM, while the RMSEP was 1.78 mM. It can be seen from Fig. 22s that most selected variables correspond to the FOT region. However, based on the results obtained by using only wavelength variables from the FOT region, it can be concluded that neither this region alone is sufficient for good prediction performance.

The last subset was the combination of SW and CB of NIR spectra for glucose measurement of human serum samples. This subset consists of 502 variables. The PLS model built using this

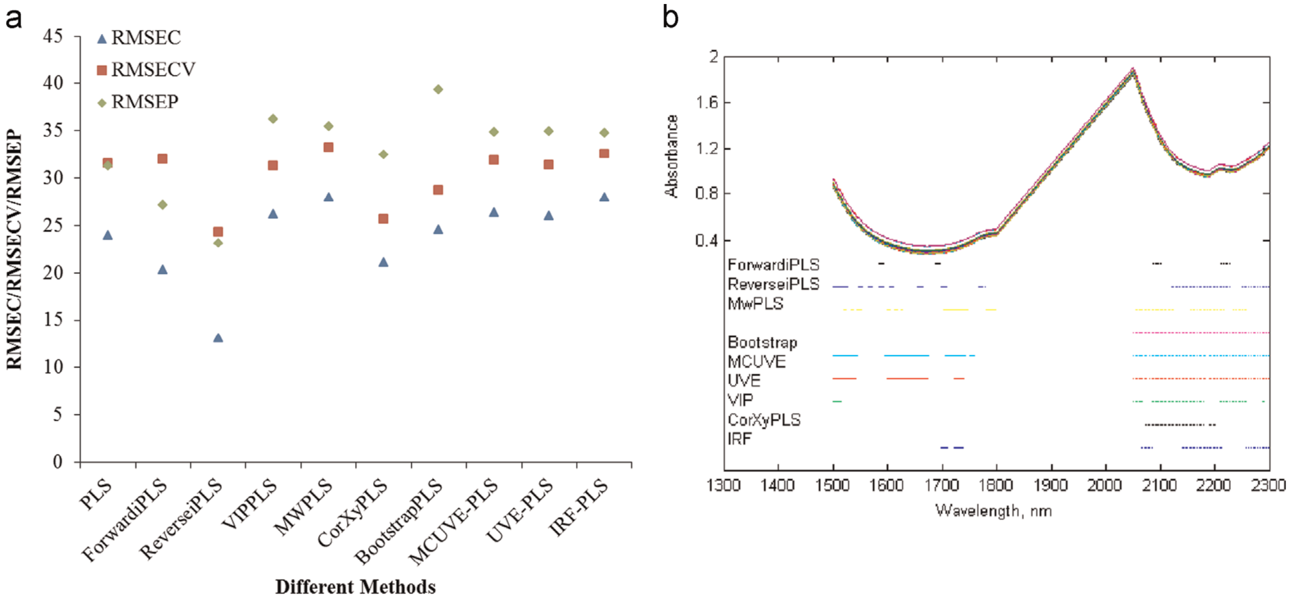

Fig. 5. (a). This figure represents the RMSEC, RMSECV and RMSEP for the fifth subset of third data set. (b) this figure shows graphically which part of spectra is selected the most by different techniques for the fifth subset of the third data set.

subset resulted in RMSECV and RMSEP values of respectively 1.81 mM and 2.61 mM. By selecting the most informative 30 variables using forwardiPLS, a better prediction performance was obtained, resulting in RMSECV and RMSEP values of respectively 1.33 mM and 1.56 mM. Fig. 22s shows the selected variables are mainly from the CB region. This indicates that the CB has more information for glucose measurement. However, the better prediction performance than for the CB region alone suggests that adding additional variables from other glucose related regions to the CB region has an added value. The HSD results shown in Fig. 23s, indicate that the average prediction performance for a PLS model built on the entire spectra is significantly different from that obtained on subset 3 and 5. The prediction performance obtained on subset 3 (FOT spectra range) was worse than that obtained by PLS using the entire spectral range as well as a combination of both FOT and CB (subset 5). Again, it is shown that the best combination for glucose measurement contains wavebands from both the FOT and the CB range.

In Fig. 24s and Table 2 the prediction performances are summarized for the bestmodels built using different variables selection techniques applied on all seven subsets of the human serum data set. Except for subset 6, the best models used less than 50 variable wavelengths for glucose measurement. When comparing the prediction performances it becomes clear that using only variables from the SW (subset 2) or the FOT (subset 3) region results in a lower prediction performance, while combination of 20 variables from the FOT and CB regions gave the best prediction results.

An overview of the wavelength variables selected by the best models for data set-3 is presented in Fig. 6. This figure shows that when CB information was in a subset, the variable selection selected mainly information from that region. Based on these results, it is confirmed that a combination of FOT and CB can lead to an accurate model for glucose measurement.

A HSD multiple comparisons test was applied on the best-selected models for each subset and compared with a PLS model built using the entire spectral range. The results presented in Fig. 25s indicate that the prediction ability of a PLS model built on the entire spectra is significantly different from that obtained on the third and fifths subsets. The third subset (only first overtone) gave a worse prediction ability, while a combination of both FOT and CB gave the best prediction performance.

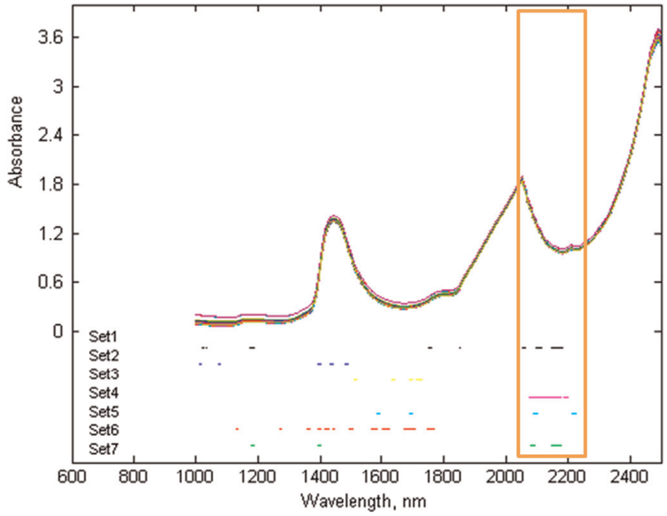

Fig. 6. This figure represents the selected variables for all seven subsets from the third data set.

Several previous studies identified the wavelength region from 2000 to 2500 nm (CB) as the most informative region for glucose measurement [41,42]. In agreement with these studies, the analyses in this study indicated that the region between 2100 and 2300 nm is very informative for glucose measurement. Furthermore, several research groups investigated whether the first overtone band or the combination band is the most informative for glucose measurement [25–28]. For example, in [26] the best region was mentioned to be from 2036 to 2324 nm [26], while in [27] a combination of the regions from 1473–1831 nm and from 2111 to 2374 nm provided the best glucose prediction accuracy. Based on the results obtained in this study, it can be concluded that a combination of wavebands in the first overtone between 1600 and 1700 nm, and the combination band from 2100 to 2200 nm would be the best choice for glucose measurement in human serum when the different spectral interferents are taken into account.

6. Conclusion

This research has focused on selecting the most informative wavelengths for Continuous Glucose Monitoring using Near Infared Spectroscopy. Three NIR data sets with different interference structures were used for this study and different spectral regions were examined to find the most relevant ranges and combinations of variables. It was found that a combination of variables from the first overtone band between 1600 and 1700 nm and the combination band from 2100 to 2300 nm gives the best prediction performance. Moreover, a comparison was made between the prediction performances based on both calibration and prediction for full spectra in different regions (e.g. short wave, first overtone and combination bands) and reduced sets. Different variable selection techniques were considered in this research. It was observed that in most cases Forward iPLS selected the smallest number of variables while providing equally good prediction accuracy as the other methods. It was also found that the combination band from 2100 to 2300 nm was mainly selected by the different variable selection techniques. However, including variables from other regions together with those from the combination band can improve the glucose prediction accuracy. In order to obtain accurate glucose measurement, which is robust against the different interferences, it is recommended to use wavebands from both the first overtone and the combination band.

Acknowledgments

The authors gratefully acknowledge Dr. S. Sharma, former member of our research group, and Dr. P. Dardenne, Dr. V. Baeten and Dr. J.-A. Fernandez-Piérna from CRA-W for their cooperation in acquiring the different data sets and finally to Prof. Joris Delanghe for providing us the icodextrin solution sample. The authors gratefully acknowledge I.W.T.-Flanders for the financial support through the Chameleon project (IWT 100021) and the GlucoSens project (SB-090053).

Appendix A. Supplementary information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.talanta.2015.08.033.

References

[1] J.P. Auses, S.L. Cook, J.T. Maloy, Chemiluminescent enzyme method for glucose, Anal. Chem. 47 (1975) 244–249.

[2] S. Auxter, Disease management models of diabetes take root, Clin. Chem. News (1996).

[3] M. Goodarzi, S. Sharma, H. Ramon, W. Saeys, Multivariate Calibration of NIR Spectroscopic Sensors for Continuous Glucose Monitoring, TrAC Trends in Analytical Chemistry 67 (2015) 147–158.

[4] T. Koschinsky, L. Heinemann, Sensors for glucose monitoring: technical and clinical aspects, Diabetes Metab. Res. Rev. 17 (2001) 113–123.

[5] X. Shao, X. Bian, J. Liu, M. Zhang, W. Cai, Multivariate calibration methods in near infrared spectroscopic analysis, Anal. Methods 2 (2010) 1662–1666.

[6] T.M. Venas, A. Rinnan, Determination of weight percent gain in solid wood modified with in situ cured furfuryl alcohol by near-infrared reflectance spectroscopy, Chemom. Intell. Lab. Syst. 92 (2008) 125–130.

[7] R.G. Brereton, G.R. Lloyd, Support vector machines for classification and regression, Analyst 135 (2010) 230–267.

[8] R.K.H. Galvao, M.C.U. Araujo, Variable selection, in: S.D. Brown, R. Tauler, B. Walczak, J.H. Kalivas (Eds.), Comprehensive Chemometrics: Chemical and Biochemical Data Analysis, section ed., vol. 3, Elsevier, Amsterdam, 2009, pp. 233–283 (Chapter 5).

[9] H. Li, Y. Liang, Q. Xu, D. Cao, Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration, Anal. Chim. Acta 648 (2009) 77–84.

[10] U. Norinder, Single and domain mode variable selection in 3D QSAR applications, J. Chemom. 10 (1996) 95–105.

[11] H. Ojelund, P.J. Brown, H. Madsen, P. Thyregod, Prediction based on mean subset, Technometrics 44 (2002) 369.

[12] I. Chong, C. Jun, Performance of some variable selection methods when multicollinearity is present, Chemom. Intell. Lab. Syst. 78 (2005) 103–112.

[13] F. Stout, J.H. Kalivas, K. Heberger, Wavelength selection for multivariate calibration using tikhonov regularization, Appl. Spectrosc. 61 (2007) 85–95.

[14] P.J. de Groot, H. Swierenga, G.J. Postma, W.J. Melssen, L.M.C. Buydens, Effect on the partial least-squares prediction of yarn properties combining raman and infrared measurements and applying wavelength selection, Appl. Spectrosc. 57 (2003) 642.

[15] J. Jiang, R.J. Berry, H.W. Siesler, Y. Ozaki, Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with applications to mid-infrared and near-infrared spectroscopic data, Anal. Chem. 74 (2003) 3555.

[16] M. Goodarzi, S. Funar-Timofei, Y. Vander Heyden, Towards better understanding of feature-selection or reduction techniques for quantitative structure–activity relationship models, TrAC Trends Anal. Chem. 42 (2013) 49–63.

[17] M. Goodarzi, B. Dejaegher, Y. Vander Heyden, Feature selection methods in QSAR studies, J. AOAC Int. 95 (2012) 636–651.

[18] L.A. Marquardt, M.A. Arnold, G.W. Small, Near-infrared spectroscopic measurement of glucose in a protein matrix, Anal. Chem. 65 (1993) 3271–3278.

[19] S. Pan, H. Chung, M.A. Arnold, G.W. Small, Near-infrared spectroscopic measurement of physiological glucose levels in variable matrices of protein and triglycerides, Anal. Chem. 68 (1996) 1124–1135.

[20] S.F. Malin, T.L. Ruchti, T.B. Blank, S.N. Thennadil, S.L. Monfre, Noninvasive prediction of glucose by near-infrared diffuse reflectance spectroscopy, Clin. Chem. 45 (1999) 1651–1658.

[21] K.H. Hazen, M.A. Arnold, G.W. Small, Temperature-insensitive near-infrared spectroscopic measurement of glucose in aqueous solutions, Appl. Spectrosc. 48 (1994) 477–483.

[22] G.W. Small, M.A. Arnold, L.A. Marquardt, Strategies for coupling digital filtering with partial least-squares regression: application to the determination of glucose in plasma by Fourier transform near-infrared spectroscopy, Anal. Chem. 65 (1993) 3279–3289.

[23] D.M. Haaland, M.R. Robinson, G.W. Koepp, E.V. Thomas, R.P. Eaton, Reagentless near-infrared determination of glucose in whole blood using multivariate calibration, Appl. Spectrosc. 46 (1992) 1575–1578.

[24] S. Kasemsumran, Y.P. Du, K. Maruo, Y. Ozaki, Improvement of partial least squares models for in vitro and in vivo glucose quantifications by using near-infrared spectroscopy and searching combination moving window partial least squares, Chemom. Intell. Lab. Syst. 82 (2006) 97–103.

[25] J. Chen, M.A. Arnold, G.W. Small, Comparison of combination and first overtone spectral regions for near-infrared calibration models for glucose and other biomolecules in aqueous solutions, Anal. Chem. 76 (2004) 5405–5413.

[26] H.M. Heise, R. Marbach, A. Bittner, T. Koschinsky, Clinical chemistry and near infrared spectroscopy: multicomponent assay for human plasma and its evaluation for the determination of blood substrates, J. Near Infrared Spectrosc. 6 (1998) 361–374.

[27] S. Kasemsumran, Y. Du, K. Murayama, M. Huehne, Y. Ozaki, Simultaneous determination of human serum albumin, gamma-globulin, and glucose in a phosphate buffer solution by near-infrared spectroscopy with moving window partial least-squares regression, Analyst 128 (2003) 1471–1477.

[28] A.K. Amerov, J. Chen, G.W. Small, M.A. Arnold, Scattering and absorption effects in the determination of glucose in whole blood by near-infrared spectroscopy, Anal. Chem. 77 (2005) 4587–4594.

[29] L. Nérgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy, Appl. Spectrosc. 54 (2000) 413–419.

[30] R. Gosselin, D. Rodrigue, C. Duchesne, A bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications, Chemom. Intell. Lab. Syst. 100 (2010) 12–21.

[31] V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B.G.M. Vandeginste, Elimination of uninformative variables for multivariate calibration, Anal. Chem. 68 (1996) 3851.

[32] Q.J. Han, H.L. Wu, C.B. Cai, L. Xu, R.Q. Yu, An ensemble of Monte Carlo uninformative variable elimination for wavelength selection, Anal. Chim. Acta 612 (2008) 121–125.

[33] L.P. Bras, M. Lopes, A.P. Ferreira, J.C. Menezes, A bootstrap-based strategy for spectral interval selection in PLS regression, J. Chemom. 22 (2008) 695–700.

[34] Z. Xiaobo, Z. Jiewen, M.J.W. Povey, M. Holmes, M. Hanpin, Variables selection methods in near-infrared spectroscopy, Analytica Chimica Acta 667 (2010) 14–32.

[35] Y.-H. Yun, H.D. Li, L.R.E. Wood, W. Fan, J.J. Wang, D.S. Cao, Q.S. Xu, Y.Z. Liang, An efficient method of wavelength interval selection based on random frog for multivariate spectral calibration, Spectrochim. Acta Part A: Mol. Biomol. Spectrosc. 111 (2013) 31–36.

[36] H. Qi, C. Xu, H. Yan, J. Ma, Comparison of icodextrin and glucose solutions for long dwell exchange in peritoneal dialysis: a meta-analysis of randomized controlled trials, Perit. Dial. Int. 31 (2011) 179–188.

[37] T.G. Schleis, Interference of maltose, icodextrin, galactose, or xylose with some blood glucose monitoring systems, Pharmacotherapy 27 (2007) 1313–1321.

[38] S. Sharma, M. Goodarzi, H. Ramon, W. Saeys, Performance evaluation of preprocessing techniques utilizing expert information in multivariate calibration, Talanta 121 (2014) 105–112.

[39] S. Sharma, M. Goodarzi, J. Delanghe, H. Ramon, W. Saeys, Using experimental data designs and multivariate modeling to assess the effect of glycated serum protein concentration on glucose prediction from near-infrared spectra of human serum, Appl. Spectrosc. 68 (2014) 398–405.

[40] A.K. Amerov, J. Chen, M.A. Arnold, Molar absorptivities of glucose and other biological molecules in aqueous solutions over the first overtone and combination regions of the near-infrared spectrum, Appl. Spectrosc. 58 (2004) 1195–1204.

[41] A.S. Bangalore, R.E. Shaffer, G.W. Small, M.A. Arnold, Genetic algorithm-based method for selecting wavelengths and model size for use with partial least-squares regression: application to near-infrared spectroscopy, Anal. Chem., 68, (1996) 4200–4212.

[42] Q. Ding, G.W. Small, M.A. Arnold, Genetic algorithm-based wavelength selection for the near-infrared determination of glucose in biological matrixes: initialization strategies and effects of spectral resolution, Anal. Chem., 70, (1998) 4472–4479.