

GLOBAL REGRESSION MODEL FOR MOISTURE CONTENT DETERMINATION USING NEAR-INFRARED SPECTROSCOPY

Matthieu Clavaud ^{a,b,*}, Yves Roggo ^a, Klara Dégardin ^a, Pierre-Yves Sacré ^b, Philippe Hubert ^b, Eric Ziemons ^b

^a *F. Hoffmann-La Roche Ltd., Wurmisweg, CH-4303 Kaiseraugst, Switzerland*

^b *University of Liege (ULg), CIRM, Department of Pharmacy, Laboratory of Analytical Chemistry, Quartier Hôpital, Avenue Hippocrate 15, B36, B-4000 Liege, Belgium*

**Corresponding author at: F.Hffmann-La Roche Ltd., Wurmisweg, CH-4303 Kaiseraugst, Switzerland.*

Keywords:

Freeze-dried medicine, Moisture content, Near-infrared spectroscopy, Support vector regression, Global model

Abstract

Near-infrared (NIR) global quantitative models were evaluated for the moisture content (MC) determination of three different freeze-dried drug products. The quantitative models were based on 3822 spectra measured on two identical spectrometers to include variability. The MC, measured with the reference Karl Fischer (KF) method, were ranged from 0.05% to 4.96%. Linear and non-linear regression models using Partial Least Square (PLS), Decision Tree (DT), Bayesian Ridge Regression (Bayes-RR), K-Nearest Neighbors (KNN), and Support Vector Regression (SVR) algorithms were created and evaluated. Among them, the SVR model was retained for a global application. The Standard Error of Calibration (SEC) and the Standard Error of Prediction (SEP) were respectively 0.12% and 0.15%. This model was then evaluated in terms of total error and risk-based assessment, linearity, and accuracy. It was observed that MC can be fastly and simultaneously determined in freeze-dried pharmaceutical products thanks to a global NIR model created with different medicines. This innovative approach allows to speed up the validation time and the in-lab release analyses.

1. Introduction

The freeze-drying process is widely used in the manufacturing of medicines. This operation consists in reducing the water content by applying sublimation and desorption [1]. Lab analyses are required to quantify the remaining water in the freeze-dried products. While Karl Fischer (KF) titration is the most employed method [2], studies demonstrated the efficiency of Near-infrared (NIR) spectroscopy to determine the remaining part of water in the freeze-dried products, as well as in powders, tablets or capsules [3–6]. The choice of NIR for water determination is first justified because it is a safe, fast, and non-destructive method. Secondly, water has an important signal in the NIR spectral range at around 6900 cm^{-1} and 5150 cm^{-1} , which is suitable for quantitative applications [7]. The NIR studies published so far are mainly product specific and creating a NIR model for one product is time-consuming. Thus, a solution would be to create a dataset composed of spectra of different products for one and unique quantitative application.

The majority of the published applications for MC determination are focused on single models [5,8] and few of them on universal models [9,10]. Nevertheless, some multivariate global models have been investigated for fast quantitative determination of compounds.

The first type of global models published deals with the quantification of compounds in the pharmaceutical or agricultural fields manufactured at different sites. In the agricultural field, universal NIR models were investigated for the quantification of compounds in pesticides, in milk (e.g. fat and protein content), or in feedstock (e.g. banana, coffee) [11–13]. In the pharmaceutical field, models were presented for active pharmaceutical ingredient (API) quantification (e.g. in capsules, tablets), and for vitamins quantification [14–18]. In all the studies, the possibility of replacing the case-by-case strategy conventionally used by a NIR universal model was demonstrated. A strategy has also proposed to update a universal quantitative model when products with different excipients and manufacturing processes are involved [19], as well as strategy to select a validation set [20].

The second type of global models that were published is the instrument-to-instrument transfer. Applying a bias adjustments [21] or adding noise in order to deteriorate data [22] are two examples of computations to show that model developed with multiple instruments can be transferred to other ones.

In this paper, three different freeze-dried medicines were chosen to investigate a global regression model for MC determination. The main novelty of this work lies in the investigation of linear and non-linear chemometrics regression methods for a global application, namely the Partial least squares (PLS), the decision tree (DT), the Bayesian ridge regression (Bayes-RR), the K-Nearest Neighbors (KNN), and the Support vector regression (SVR). A dataset composed of 3822 spectra acquired on two identical NIR spectrometers was used to create and validate a model common to the three drug products. Each regression method was evaluated, and the best one based on SVR algorithm was retained.

2. Material and methods

2.1. SAMPLES AND VALIDATION CONCEPT

2.1.1. SELECTION OF SAMPLES

2.1.1.1. Calibration and validation sets. Three different freeze-dried medicines, filled into transparent glass vials, have been measured. The products were annotated A, B, and C for confidentiality reasons. For the product B, two dosages B₁ and B₂ were analyzed that were composed of the same ingredients but presented a different active pharmaceutical ingredient (API) concentration. All the products differed in terms of type and content of API, diameter size of the glass vial, and excipients amount (**Table 1**).

At least 5 batches and 100 samples per product were selected to include batch-to-batch, product-to-product, and sample-to-sample variability in the dataset (**Table 1**).

In addition, a special attention was paid to the quality of the cake. Only sample passing the visual appearance test were used. The vials were visually inspected to retain the cakes presenting the best appearance in order to avoid meltbacks [1] (collapse resulting from the presence of liquid during the primary drying), tilted cake (non-correct position of the cake in the glass vial), and cake breaks or cracks (due to transport malfunction). All of them could have an impact on the NIR measurement. The NIR spectra will be impacted and consequently the prediction results will be incorrect.

Table 1

Main characteristics of the products used to calibrate and validate the multivariate models.

Products	Model calibration and validation				External validation
	A	B ₁	B ₂	C	D
Galenic form	Freeze-dried			Chemically manufactured	
Type of product	Biopharmaceutical (protein based)			Antiviral	
API type	Monoclonal antibody			Antiviral	
API content (mg/vial)	150	100	160	100	500
Vial format (mL)	15	15	20	3	15
Vial diameter size (cm)	2.5	2.5	3	1.5	2.5
Total amount of product in a vial (mg)	50	433	700	50	60
Number of batches	5	10	7	15	5
Number of vials	100	194	238	105	28
(calibration/validation)	(67/33)	(124/70)	(158/80)	(70/35)	(0/28)
Humid	58	126	166	95	21
Dry	30	14	15	n/a	n/a
Unchanged	12	54	57	10	7

2.1.1.2. External validation set. One freeze-dried medicine, named D, also filled into transparent glass vials, has been tested during the external validation. The product was different to the ones used to create the models (**Table 1**).

2.1.2. SAMPLE PREPARATION

In normal circumstances, it is not required to prepare the samples prior to NIR measurements. Nevertheless, since each of the measured product contains on average 1.00% of residual moisture on the routine, a preparation step was necessary in order to create a wider MC range.

On the whole, 140 vials were kept intact (unchanged) and stored at 5 °C, storage temperature of the studied products.

Then, 466 vials were humidified using the strategy protocol described in a previous work [23]. The vials were humidified from 5 to 105 min for the product A, from 5 to 240 min for the product B, from 5 to 180 min for product C, and from 15 to 150 min for product D, with 5 min interval. After this period of time, the samples were closed and stored at 5 °C. The humidification times were selected after feasibility studies which showed those were the necessary times to reach a wide MC range.

In parallel, 59 vials were dried using also the strategy protocol described in a previous work [23]. The samples of product A were kept inside desiccators from 2 to 22 days, with 2 days interval. In parallel, the samples of product B were kept inside desiccators during: 2, 6, 10, 14, 18 and 22 days. Then, when the indicated time was reached, the samples were closed and stored in a second desiccator containing silica gel. The drying times were selected after feasibility studies which showed those were the necessary times to reach a wide MC range. The samples of product C and D were not dried due to their to initial low MC.

2.1.3. CALIBRATION AND VALIDATION CONCEPT

For the global model evaluation, the whole dataset composed of products A, B, and C, was split into two independent calibration and validation sets. Each set contained samples from all the products. The calibration set was composed of two thirds of the dataset and the validation set of the remaining third. Therefore, 419 vials were used for the calibration and 218 vials for the validation (**Table 1**).

For the individual model evaluation, each product dataset (products A, B, and C) was also split into independent calibration and validation sets, with two thirds of the samples in the calibration set and one third in the validation set (**Table 1**).

The global and individual models were created using the calibration set and validated with the validation set. An external validation set composed of 28 vials from the product D was also used to evaluate the prediction effects of the best global model on an independent product.

2.2. ANALYTICAL METHODS

2.2.1. NEAR-INFRARED EQUIPMENT AND SPECTRA RECORDING

Two identical Antaris Fourier Transform spectrometers (Thermo Fisher Scientific®, Madison, WI, USA) were used to perform the measurements. The instruments were equipped with an InGaAs detector, a Michelson interferometer, an halogen NIR source, and an integrating sphere module.

The spectra were acquired in diffuse reflection mode with 32 scans, a resolution of 8 cm⁻¹, a gain of 1 and no attenuator. A spectral range of 4000–10,000 cm⁻¹ was used. A background spectrum was acquired every hour. Each medicine vial was measured directly on the measurement window. The measurements were performed on the bottom of the glass vial. Three repetitions per vial were performed on each instrument, meaning that each vial was measured six times. Therefore, a total of 3822 spectra for the 637 vials (products A, B, and C) were obtained for the calibration and validation sets. In addition, a total of 168 spectra was acquired for the product D.

2.2.2. KARL FISCHER TITRATION METHOD

After the NIR measurements, the MC of each vial was determined using the KF titration reference method. A Metrohm 901 Volumetric Karl Fischer Titrand (Metrohm, Herisau, Switzerland) instrument was used. The apparatus was composed of a KFTitrand 901, equipped with a 30 mL/min dosing system and coupled with a 803 Ti Stand stirring system. Each vial was titrated twice by KF and an averaged value (expressed as percentage) was obtained.

2.3. CHEMOMETRIC REGRESSION METHODS

2.3.1. DATA PREPROCESSING

Data preprocessing were applied on NIR spectra prior the computation of the regression methods, in order to correct the effects not related to the moisture. Mathematical treatments such as Standard normal variate (SNV) [24] and derivatives [25] were applied to correct the baseline drift and the instrumental changes.

Partial least squares (PLS), Bayesian Ridge regression (BayesRR), Decision tree (DT), and K-Nearest Neighbors (KNN) models were computed with a SNV as first preprocessing and followed by a Savitzky-Golay first derivative (polynomial order of 5, windows size of 15). The Support Vector Regression (SVR) model was computed using only SNV.

2.3.2. PARTIAL LEAST SQUARES (PLS)

The PLS approach is the most popular algorithm for the creation of models [26], especially for MC determination in pharmaceutical products [5]. Least squares algorithms are used to develop a linear link between two matrices, in the present case between the NIR spectral data and the KF reference values.

The PLS algorithm was used to create individual models for each product and then for the computation of the global model. A mean-centered PLS was applied. The number of PLS components were chosen by looking at the regression coefficients, and the variation explained by the calibration and cross validation.

2.3.3. BAYESIAN RIDGE REGRESSION (BAYES-RR)

Another linear regression method used was the Bayes-RR. Bayesian regression is widely used in machine learning area [27] and its performances for spectroscopic calibration have been proven [28]. The Bayes-RR is a variant of the Bayesian regression based on the ridge regression. The method presents similarities with least squares, however the estimated coefficients tend towards zero to avoid multicollinearity [29].

The Bayes-RR approach was applied to the present data for the computation of a global model

2.3.4. DECISION TREE (DT)

The DT method is a non-linear regression technique. The principle is to compute a regression in a tree structure form breaking down a dataset into smaller and smaller subsets [30].

The DT approach was applied to evaluate a global model.

2.3.5. K-NEAREST NEIGHBORS (KNN)

An additional non-linear regression method named KNN was also tested [31]. Applied to a set of NIR spectra, the KNN algorithm predicts an unknown spectrum in function of the K-closest spectra available (the K-Nearest Neighbors).

The KNN was also applied for the evaluation of a global model.

2.3.6. SUPPORT VECTOR REGRESSION (SVR)

The Support Vector Machine (SVM) is a non-linear algorithm mostly known as a classification method. It can also be used as a regression method, called the Support Vector Regression (SVR). The SVR uses the same principles as the SVM [32], based on finding the hyperplane maximizing the margin between classes. The regression is made possible by the use of kernel functions. The functions transform the data into a higher dimensional feature space to make a linear separation possible. Different types of Kernel functions can be applied. In the present study, the Radial Basis Function has been used. The performance of the SVR algorithm depends on the regularization and gamma parameters [31] which are optimized to achieve the lowest error.

The SVR was used for the computation of a global model, and then for the creation of the individual models.

2.4. EVALUATION OF MODEL PERFORMANCE

Three individual PLS models corresponding to the three products A, B, and C and a PLS global model were created. Individual and global PLS models were then compared.

The DT, Bayes-RR, KNN, and SVR algorithms were used to compute a global model and compared to the PLS global model in order to select the best predictive model.

The performance of the models was evaluated using the slope, the bias, the intercept, the coefficient of determination (R^2), the standard error of calibration (SEC), the standard error of prediction (SEP), the Ratio of performance deviation (RPD), and the Range error ratio (RER) [23]. The slope, the bias, the intercept, and the R^2 were used to detect systematic errors in the models. The SEC and the SEP values were used to evaluate the accuracy of the models. In addition, RPD and RER were also calculated to assess the performance of the models. A reliable model should yield high R^2 , RPD and RER values and low SEC, SEP, and bias values. In addition, the difference between SEC and SEP values should be low in order to avoid overfitting.

Moreover, a total error and risk-based approach using the accuracy profile was also used to demonstrate the capacity of the NIR method to fulfill its MC analyses objective [33]. A part of the validation set was selected among dried, unchanged and humidified vials and 6 levels of moisture were defined. For each level of moisture, 4 series combining equipment, product, formulation, and lots effects were used. For each series, 18 samples were analyzed in triplicate. The acceptance limit for the accuracy profile was set at $\pm 25\%$. The method is considered valid within the range for which the β -expectation tolerance limits are within the acceptance limits. The objective of this approach is to

guarantee that a defined proportion of future measurement of unknown samples will be included within the tolerance limits.

2.5. SOFTWARE

The “Result NIR Suite 3” package (Thermo Fisher Scientific®, Madison, WI, USA) with the Result Integration software was used for spectral acquisition.

The regression models were computed with the programming language Python 3.4 (Anaconda Package, Continuum Analytics) [34,35] and the Spyder 2.3.5.2 interface (The ScientificPython Development Environment) [36]. Pandas 0.16.2 and Numpy 1.9.2. were used to structure and analyze the data. The derivative pretreatment were calculated with Scipy 0.15.1.

The graphics were generated with R version 3.2.3. E.noval V3.0b® (Arlenda, Liege, Belgium) was used for accuracy profile computation.

3. Results and discussion

3.1. NIR SPECTRA FEATURES

The dataset was composed of 3822 spectra of 3 different freeze-dried products. The amount of spectra and the differences between the products (**Table 1**) might constitute a challenge for the creation of a common quantitative model. That is why the NIR spectra had first to be interpreted.

The preprocessed NIR spectra have been displayed in **Fig. 1**. The combination band of OH stretching and HOH bending at around 5150 cm^{-1} and first overtone of OH stretch at around 6900 cm^{-1} are specific to water. They were observed in the spectra of all the products and increased with the humidity preparation time. Both moisture peaks showed that the sample preparation was successfully achieved.

In the region comprised between 8500 and $10,000\text{ cm}^{-1}$, no particular peaks were observed and most of the spectral information is observed in the region between 4000 and 8500 cm^{-1} . In this last region, spectral differences were detected between each product. For instance in the 4000 – 5000 cm^{-1} region, the product C presents three sharp peaks at 4418 cm^{-1} , 4340 cm^{-1} , and 4260 cm^{-1} , corresponding to the CAH stretching combination. Similarly, two sharp peaks were observed for the product B at 4760 cm^{-1} and 4370 cm^{-1} . In addition, in the region between 6000 cm^{-1} and 5600 cm^{-1} , corresponding to the first overtone C–H stretching, spectral differences were observed between each product.

The sample post-preparation had an impact on the composition of the product C. A change in the chemical structure of the freeze-dried cake was observed. Additional peaks were detected at 4600 cm^{-1} and 4900 cm^{-1} in some of the spectra (**Fig. 1c**) characteristic of amine functions. These spectra correspond to the vials which were humidified and provided MC values superior than 3%.

Although spectra differences were observed for all of the products, a full spectrum approach over the range 4000 and $10,000\text{ cm}^{-1}$ was chosen to create the regression models.

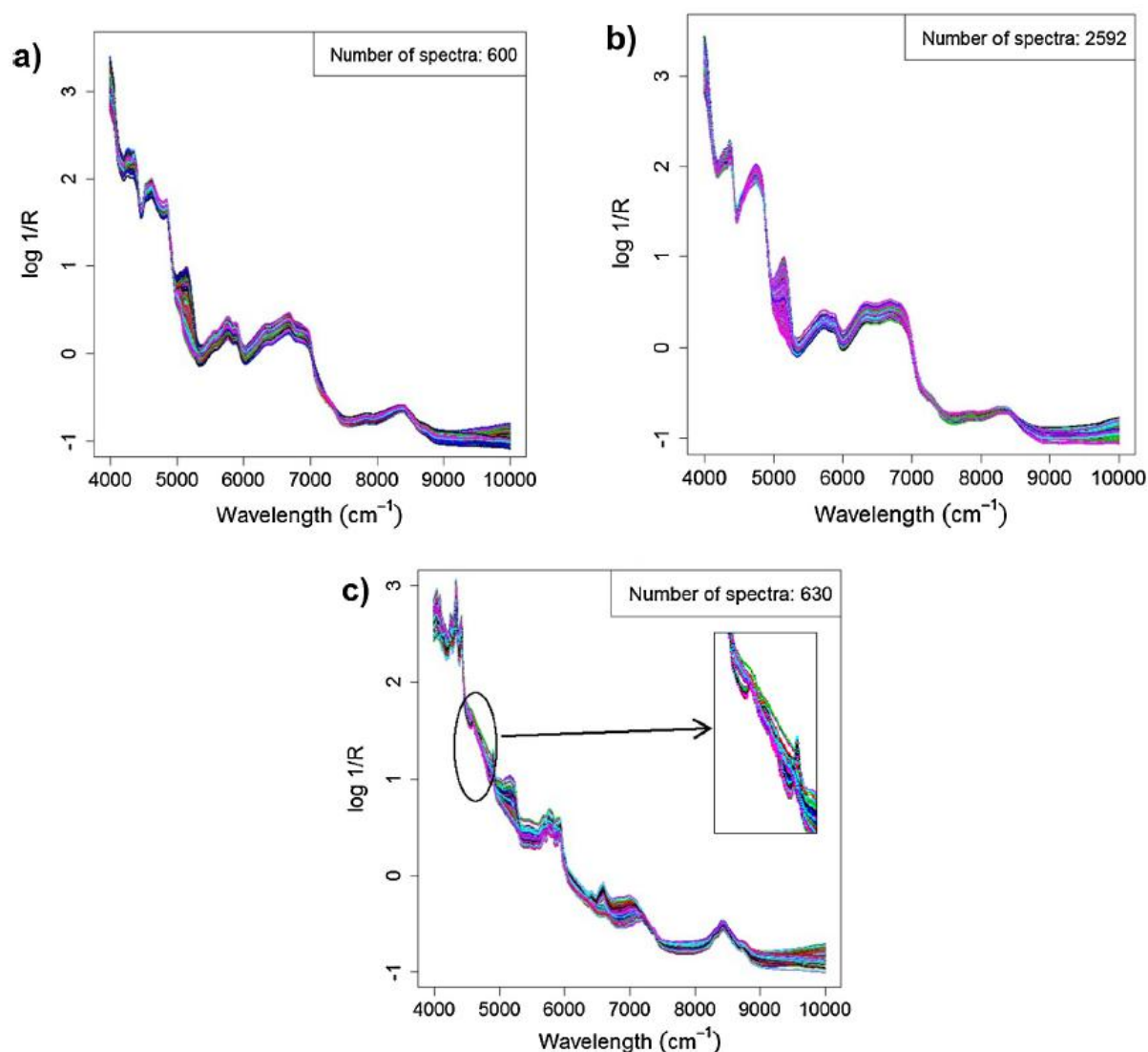


Fig. 1. NIR spectra of the freeze-dried products after SNV preprocessing with the moisture bands at 6900 and 5150 cm^{-1} . (a) Spectra of the product A, (b) Spectra of the products B – The differences in the formulation B₁ and B₂ are not observed, (c) Spectra of the product C.

3.2. INDIVIDUAL MODELS USING THE PLS

Prior to global modeling, calibrations for each of the 3 types of products have been created. The objective was to highlight the performance of each calibration in order to be able to compare them with the global model.

Five latent variables were selected for each individual model. The **Fig. 2a, b, and c** display the regression coefficients for the five selected latent variables. The regions of large coefficients are coincident with both moisture wavelength bands at around 5150 cm^{-1} and 6900 cm^{-1} . This demonstrated the specificity of each model.

The three individual calibration curves - $\text{PLS}_{A/\text{Cal}}$, $\text{PLS}_{B/\text{Cal}}$, $\text{PLS}_{C/\text{Cal}}$ - are presented in **Fig. 2d, e and f**. The SEC and R^2_{Cal} values obtained for the individual models were ranged between 0.15% and 0.22%, and

0.992 and 0.961, respectively (**Table 2**). These values demonstrated the correlation between NIR and KF value. They were confirmed with intercept and slope values close to 0 and 1, respectively. No bias was observed for the three calibration models.

The three individual validation plots - $PLS_{A/Val}$, $PLS_{B/Val}$, $PLS_{C/Val}$ are presented in **Fig. 2g, h, and i**. The SEP and R^2_{Val} values were ranged between 0.17 and 0.25, and 0.972 and 0.986, respectively (**Table 2**). Concerning the product C, a high prediction error was observed around 4.0–5.0%. This phenomenon can be explained by the differences observed in the spectra of the samples C in this region. Indeed, as showed during the spectra analysis, a change in the chemical structure of the freeze-dried cake was noticed for samples C presenting a MC superior than 3% between 5000 cm^{-1} and 4400 cm^{-1} . These spectra were not considered as outliers and were kept in the database.

The individual PLS models created for the products A, B, and C provided satisfactory results in terms of accuracy and precision.

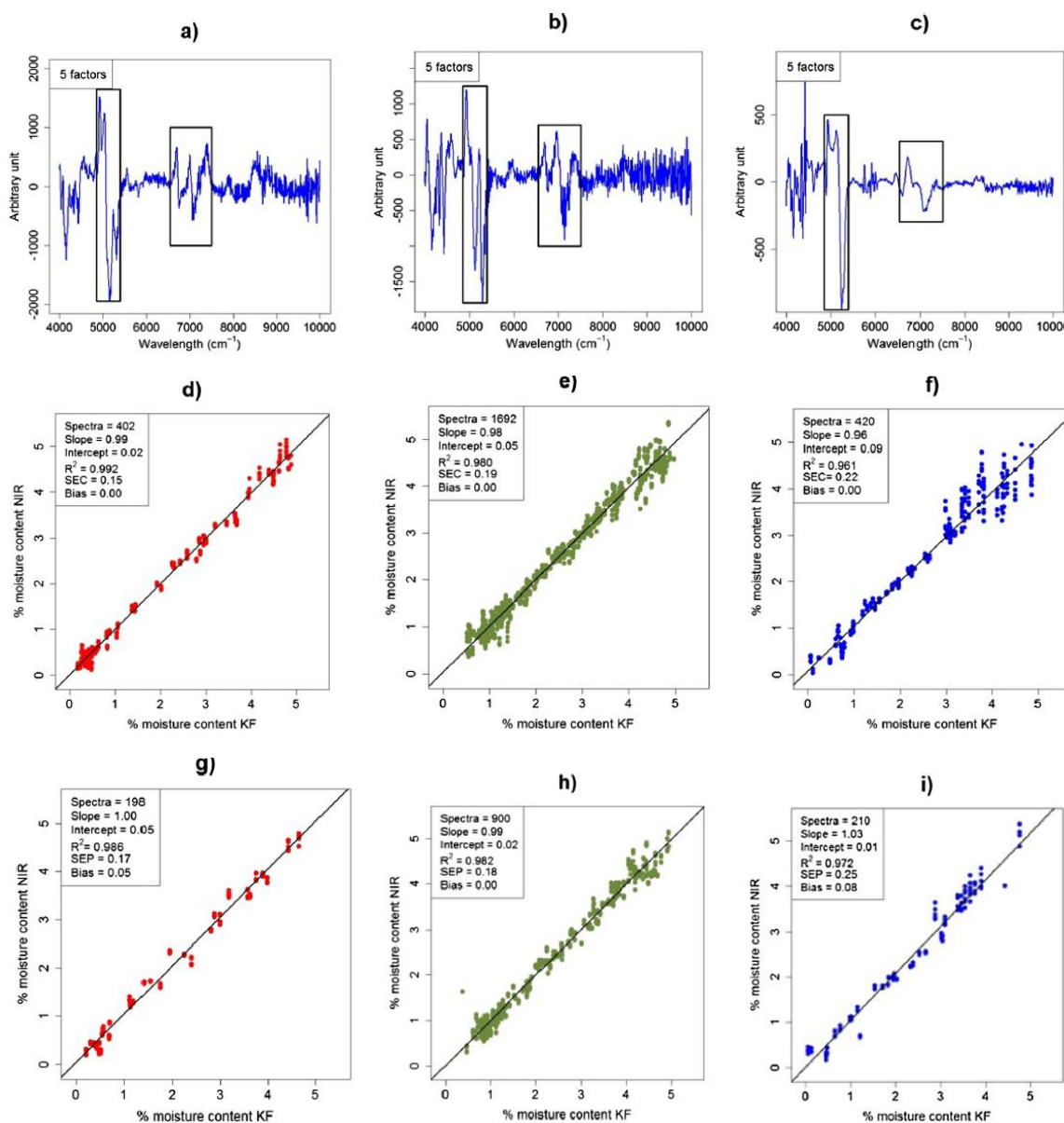


Fig. 2. Individual PLS plots - Representation of the regression coefficients plots of the PLS models obtained (a) for the product A, (b) for the product B, (c) for the product C, with the moisture bands in the black squares. The calibration curves are represented (d) for product A, (e) product B, (f) product C. The validation samples predicted against their respective calibration curves are represented (g) for product A, (h) product B, (i) product C.

Table 2

Statistics for the calibration and validation sets obtained with PLS algorithm.

Product Set	Individual models						Global model	
	A		B		C		A, B, and C	
	Calibration	Validation	Calibration	Validation	Calibration	Validation	Calibration	Validation
Model name	PLS _{A/Cal}	PLS _{A/Val}	PLS _{B/Cal}	PLS _{B/Val}	PLS _{C/Cal}	PLS _{C/Val}	PLS _{Uni/Cal}	PLS _{Uni/Val}
Number of spectra	402	198	1692	900	420	210	2514	1308
KF Concentration range (%)	[0.17; 4.88]	[0.21; 4.65]	[0.51; 4.96]	[0.46; 4.92]	[0.05; 4.87]	[0.05; 4.75]	[0.05; 4.96]	[0.05; 4.92]
Mean	1.87	1.84	2.34	2.35	2.27	2.06	2.25	2.22
Standard deviation	1.62	1.40	1.32	1.34	1.27	1.21	1.38	1.34
Wavelength range (cm ⁻¹)	[4000; 10,000]		[4000; 10,000]		[4000; 10,000]		[4000; 10,000]	
Preprocessing	SNV + 1st derivative		SNV + 1st derivative		SNV + 1st derivative		SNV + 1st derivative	
Number of factors	5		5		5		8	
Slope	0.99	1.00	0.98	0.99	0.96	1.03	0.97	1.00
Intercept	0.02	0.05	0.05	0.02	0.09	0.01	0.07	0.04
R^2_{Cal} (R^2_{Val})	0.992	0.986	0.980	0.982	0.961	0.972	0.970	0.987
SEC(P)	0.15	0.17	0.19	0.18	0.22	0.25	0.24	0.23
Ratio SEP/SEC	1.13		0.95		1.14		0.96	
Bias	0.00	0.05	0.00	0.00	0.00	0.08	0.00	0.03
Residues [min; max]	[-0.39; 0.41]	[-0.33; 0.43]	[-0.72; 0.60]	[-0.52; 1.26]	[-1.00; 1.02]	[-0.55; 0.77]	[-0.99; 1.47]	[-0.67; 1.15]
RPD	9.53		7.33		5.08		6	
RER	27.71		24.78		19.28		21.35	

Table 3

Statistics for the calibration and validation sets obtained with SVR algorithm for the individuals products.

Product Set	A		B		C	
	Calibration	Validation	Calibration	Validation	Calibration	Validation
Model name	SVR _{A/Cal}	SVR _{A/Val}	SVR _{B/Cal}	SVR _{B/Val}	SVR _{C/Cal}	SVR _{C/Val}
Number of spectra	402	198	1692	900	420	210
KF Concentration range (%)	[0.17; 4.88]	[0.21; 4.65]	[0.51; 4.96]	[0.46; 4.92]	[0.05; 4.87]	[0.05; 4.75]
Mean	1.87	1.84	2.34	2.35	2.27	2.06
Standard deviation	1.62	1.40	1.32	1.34	1.27	1.21
Wavelength range (cm ⁻¹)	[4000; 10,000]		[4000; 10,000]		[4000; 10,000]	
Preprocessing	SNV		SNV		SNV	
Slope	1.00	1.03	0.99	0.99	0.97	1.02
Intercept	0.00	0.02	0.02	0.03	0.07	0.03
R^2_{Cal} (R^2_{Val})	0.996	0.988	0.995	0.992	0.979	0.965
SEC(P)	0.10	0.16	0.10	0.12	0.19	0.24
Ratio SEP/SEC	1.6		1.2		1.26	
Bias	-0.01	0.03	0.00	0.02	0.00	0.06
Residues [min; max]	[-0.25; 0.97]	[-0.33; 0.48]	[-0.35; 0.29]	[-0.33; 0.32]	[-0.94; 0.89]	[-0.49; 0.90]
RPD	10.13		11.00		5.29	
RER	29.44		37.08		20.08	

3.3. GLOBAL MODEL USING THE PLS

Eight regression coefficient were selected as displayed in **Fig. 3a**. The regions of large coefficients are coincident with both moisture wavelength bands, as already observed for the individual PLS models. The huge number of factors regression coefficients is a sign that the spectral data are non-linear in their nature.

The calibration curve - PLS_{Uni/Cal} - is presented in **Fig. 3b**. A SEC value of 0.24 and a R^2_{Cal} value of 0.97 were obtained (**Table 2**) confirming the correlation between NIR and KF value. This observation was confirmed with an intercept value of 0.07 and a slope value of 0.97.

In addition, no bias was observed.

The validation plot - $PLS_{Uni/val}$ - is presented in **Fig. 3c**. A SEP value of 0.23 and a R^2_{Val} value of 0.987 were observed (**Table 2**). As for the calibration curve, the correlation and slope values were close to 1 and the intercept value was close to 0.

The SEC and SEP values of the global model were higher than the SEC and SEP values of the individual models. The prediction of the results were not improved compared to individual models. Moreover, a high prediction error was observed for low and high MC values at around 1% and 4%. The degradation of the accuracy is explained by the spectral influence of the product C, the presence of several products and thus the variability coming from each product.

Nevertheless, these encouraging results showed the possibility to use a global model for MC determination. The challenge would be then to propose a model with prediction results presenting a better accuracy.

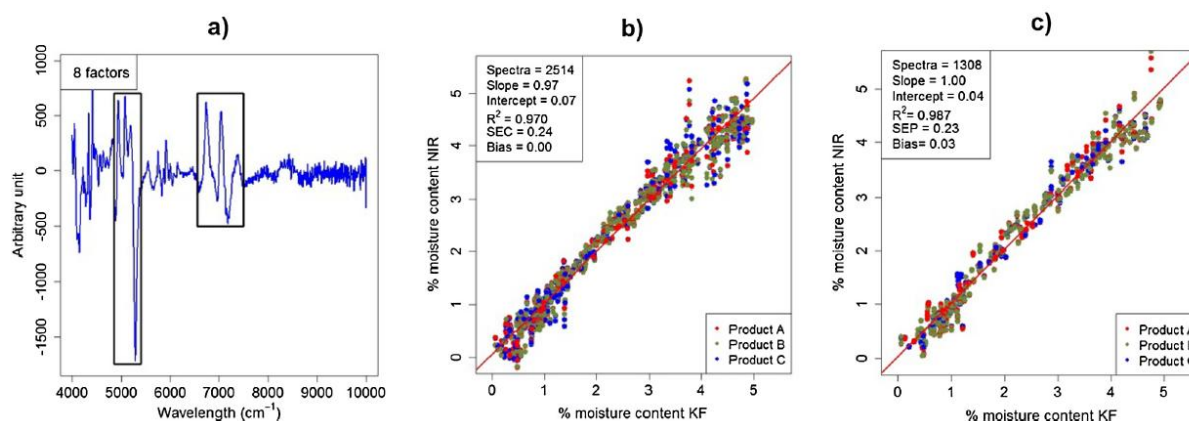


Fig. 3. Global PLS plots – (a) Representation of the regression coefficients plot for 8 latent variables with the two water bands in the black squares. (b) Representation of the calibration curve. (c) Representation of the validation curve.

3.4. SCREENING OF OTHER LINEAR AND NON-LINEAR METHODS FOR THE GLOBAL MODEL

The DT, Bayes-RR, KNN, and SVR chemometrics methods were then evaluated for global modeling. The **table 4** presents a summary of the calibration and validation results obtained for each of the four models.

The performance overview for each model was first evaluated based on the RPD and RER criteria. Out of the four the models, the RPD of the SVR model was superior than 8 and the RER superior than 30 meaning that this model was the most accurate one. Indeed, a RPD and a RER greater than 8 and 10 respectively indicate that the model can be used for any quantitative application [37,38]. The DT, Bayes-RR, and the KNN, all presented lower RPD and RER values than the SVR model, indicating a higher error of prediction. The chemometric models could thus be ranked (from the best to the worst) in the following order: SVR, KNN, Bayes-RR, DT. According to the R^2_{Val} and SEP values, the same ranking was obtained and R^2_{Val} and SEP values of 0.987 and 0.15 were obtained for the best model (SVR).

In addition, the model DT_{Uni/Cal} presented a very good equation (R_{Cal}^2 of 0.999, SEC of 0.04) which was not confirmed with the validation set. Indeed, a SEP 5 times higher than the SEC was observed, suggesting an overfitting of the model.

Consequently, SVR_{Uni/Cal} was retained as the best chemometric model and its performances will be evaluated in the next paragraph.

Table 4

Prediction and statistics of the global calibration and validation results obtained with the DT, Bayes-RR, KNN, and SVR algorithms. The best predictions results were achieved with the SVR algorithm.

Type of model	DT		Bayes-RR		KNN		SVR	
Set	Non linear		Linear		Non linear		Non linear	
Model ID	Calibration	Validation	Calibration	Validation	Calibration	Validation	Calibration	Validation
Preprocessing	DT _{Uni/Cal}	DT _{Uni/Val}	Bayes-RR _{Uni/Cal}	Bayes-RR _{Uni/Val}	KNN _{Uni/Cal}	KNN _{Uni/Val}	SVR _{Uni/Cal}	SVR _{Uni/Val}
c/gamma	SNV + 1st derivative		SNV + 1st derivative		SNV + 1st derivative		SNV	
Slope	n/a		n/a		n/a		100/100	
Confidence interval	1.00	1.00	0.99	1.01	0.98	1.00	0.99	1.00
Intercept	[0.998; 1.000]	[0.996; 1.013]	[0.981; 0.990]	[1.001; 1.017]	[0.974; 0.983]	[0.991; 1.007]	[0.988; 0.995]	[0.997; 1.009]
Confidence interval	0.00	0.04	0.03	0.00	0.05	0.04	0.02	0.02
R_{Cal}^2 (R_{Val}^2)	[−0.001; 0.006]	[0.016; 0.060]	[0.021; 0.044]	[−0.024; 0.019]	[0.036; 0.061]	[0.021; 0.061]	[0.013; 0.023]	[0.008; 0.041]
SEC(P)	0.999	0.978	0.988	0.978	0.984	0.980	0.993	0.987
Ratio SEP/SEC	0.04	0.20	0.15	0.20	0.17	0.19	0.12	0.15
Bias	5		1.33		1.12		1.25	
Residues [min; max]	0.00	0.04	0.00	0.02	0.00	0.04	0.00	0.03
RPD	[−0.44; 0.32]	[−0.56; 1.07]	[−0.80; 0.56]	[−1.19; 0.80]	[−0.90; 0.99]	[−0.53; 1.20]	[−0.96; 0.89]	[−0.48; 0.90]
RER	6.9		6.9		7.26		9.2	
	24.55		24.55		25.84		32.73	

3.5. SVR PERFORMANCES

The SVR algorithm provided a global model with better results in terms of accuracy compared to the other approaches. The aim of this paragraph part is to demonstrate that this model can be validated and brings an equivalent or a better precision than the conventional PLS approach and the individual SVR models.

3.5.1. GLOBAL MODEL

The calibration and validation curves obtained with the SVR global model have been presented in **Fig. 4**. A linear relationship is observed between the NIR and KF data. The statistics presenting the efficiency of the model have already been displayed in **Table 4**. The advantage of SVR is its ability to deal with non-linear data which was supposed with the global PLS model. Moreover, the SVR is a quite new method in the pharmaceutical field and this study should be bring information about its use on freeze-dried products.

Additionally, since the calibration range was of 4.91% and the SEP value of 0.15, the requirement that the calibration range should be higher or equal than 10 times the SEP was fulfilled. Moreover, the concentration limit for each of the 3 types of products was 3%. With a calibration limit at 4.96%, the requirement that the calibration limit should be higher or equal than 120% of the concentration limit was also fulfilled. The slope value was included in the range [0.95; 1.05], and the intercept and the bias values were inferior than 0.03, which demonstrated the linearity for the calibration and validation samples.

The statistics of the SVR model fulfilled the acceptance criteria defined for the validation of a NIR method intended to the measurement of the MC in freeze-dried drug products [23].

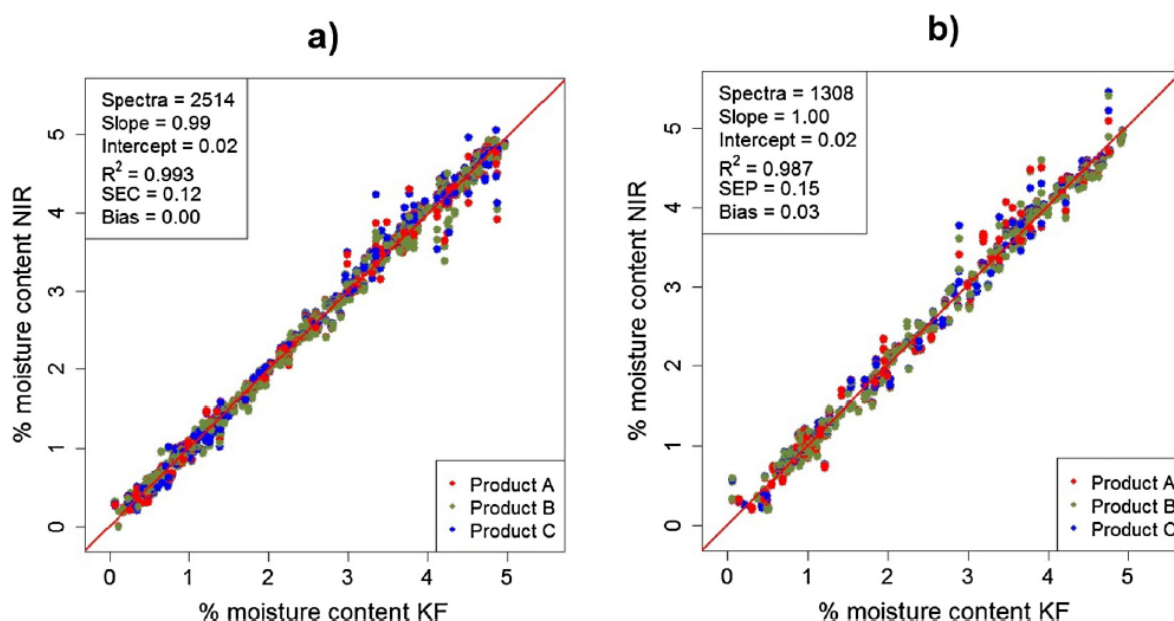


Fig. 4. Global SVR plots – (a) Representation of the calibration curve. (b) Representation of the validation curve.

3.5.2. COMPARISON OF GLOBAL PLS AND SVR PERFORMANCES, AND GLOBAL AND INDIVIDUAL SVR EFFECTS

The PLS and the SVR also proved their efficiency for a global quantification prediction model with NIR spectroscopy. Both models were accurate and could be validated. However, the performance of the SVR approach was better than the one obtained with the PLS. Indeed, the RPD and RER values were higher for the SVR model than the PLS, meaning that SVR model was the most accurate. In addition, the R^2_{Cal} value of the SVR model is higher than the one of PLS model and the R^2_{Val} were equivalent. The same observation could be done for the SEC and the SEP values with lower values for the SVR than for the PLS approach while the bias values were equivalent. In addition, the range of the residues is wider for the PLS model (2.46 and 1.82 for the calibration and validation samples, respectively) than for the SVR model (1.85 and 1.38 for the calibration and validation samples). The lower the range of residues, the better the model can be accurate.

The accuracy profile obtained with the validation data using the PLS is illustrated in **Fig. 5** while **Fig. 6** showed the accuracy profile obtained with the validation data using the SVR. The tolerance interval for the PLS was included into the acceptance limits in the range 2–4%. For the SVR, the tolerance interval was wider and included into the acceptance limits in the range 1–4%. Furthermore, the risk to obtain a result outside the acceptance limits between 0 and 2% is very high with the PLS while it is half reduced with the SVR. With small moisture content (<1%) the model is not reliable enough. 1% is the limit of quantification for the SVR model and 2% for the PLS model. According to KF references values the moisture content of the real-life freeze-drying products is mainly below or around 1%. Under the limit of quantification values (i.e. 1% or 2%) the recommendation for real-life freeze-drying products should be to provide the results with the mention “results less than 1%” or “result less than 2%”. ICH Q2(R1) validation criteria are presented in **Table 5**. These criteria are in terms of trueness, precision

and accuracy in accordance with the validation results obtained in the previous sections for the global SVR model.

The SVR individual analytical performances were evaluated for each product (**Table 3**). SEC values of 0.10, 0.10, and 0.19 were observed for the product A, B, and C, respectively. SEP values of 0.16, 0.12, and 0.24 were then observed. The prediction errors were not better than the ones observed with SVR_{Uni/Cal} and SVR_{Uni/Val} confirming that using a global model did not degrade the prediction accuracy.

Moreover, the minimum amount of NIR spectra required for establishing an almost same global SVR model was not evaluated. For a PLS application, it was observed that a reduced number of spectra can provide similar performances than a bigger amount of spectra [23]. Consequently, it could be expected that statistics obtained with SVR model created with reduced spectra can be similar to the ones obtained in this study. This criteria can be evaluated in another study.

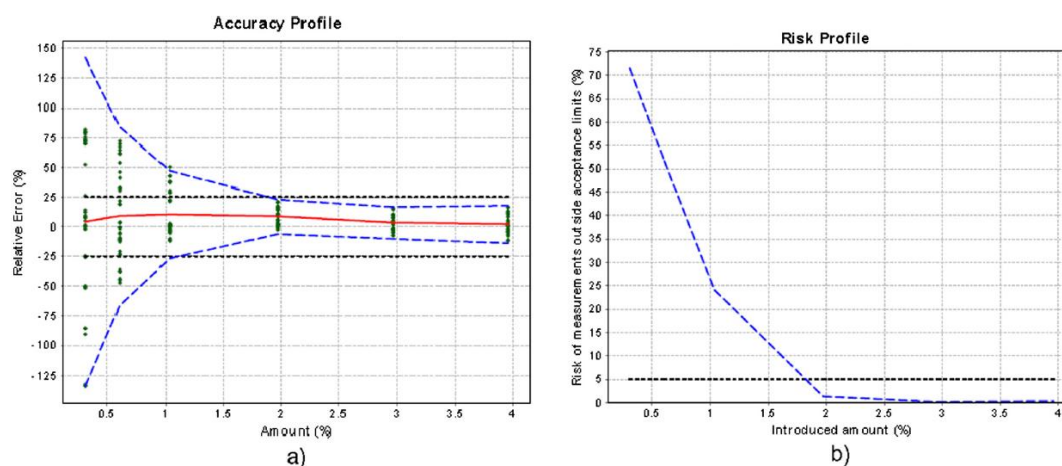


Fig. 5. (a) Accuracy profile of the validation set obtained with PLS model. The plain red line is the relative bias, the dashed lines are the β -expectations tolerance limits ($\beta = 95\%$) and the dotted lines represent the acceptance limits ($\pm 25\%$). The green dots represent the relative error of the calculated concentrations and are plotted against their KF reference values. (b) Risk profile – the dotted line represents the maximum risk level chosen at 5%. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

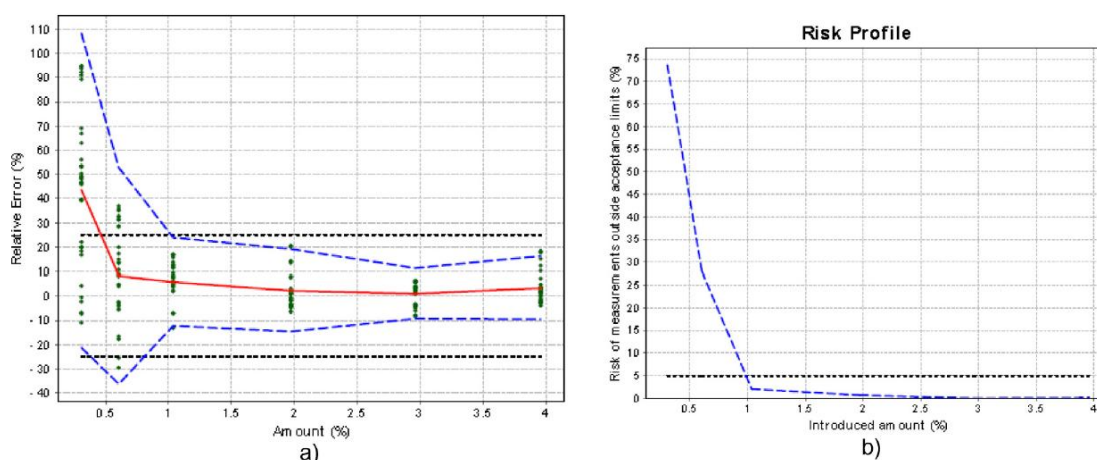


Fig. 6. (a) Accuracy profile of the validation set obtained with the SVR regression model. The plain red line is the relative bias, the dashed lines are the β -expectations tolerance limits ($\beta = 95\%$) and the dotted line represents the acceptance limits ($\pm 25\%$). The green dots represent the relative error of the calculated concentrations and are plotted against their KF

reference value. (b) Risk profile – the dotted line represents the maximum risk level chosen at 5%. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5

ICH Q2(R1) validation criteria of global NIR method for MC determination using SVR based on the total and risk-based approach.

Mean content level (%)	Relative bias (%)	Repeatability (RSD %)	Intermediate precision (RSD %)	β -expectation relative tolerance limits (%)	Relative uncertainty (%)
0.31	43.51	30.58	31.26	[−21.23; 108.3]	63.67
0.60	8.27	15.53	19.59	[−36.21; 52.75]	41.29
1.04	5.82	8.07	8.59	[−12.22; 23.87]	17.64
1.97	2.29	7.21	7.92	[−14.54; 19.11]	16.34
2.96	1.03	4.44	4.85	[−9.28; 11.33]	10.01
3.96	3.34	5.86	6.16	[−9.55; 16.23]	12.63

3.5.3. PERFORMANCE OF THE GLOBAL SVR MODEL WITH THE EXTERNAL VALIDATION SET

The 168 spectra of the product D were then predicted with the SVR_{Uni/Cal} model. The idea was to evaluate the prediction effects on an independent product when using the SVR global model. A large residues range [0.4; 2.38] was observed. The accuracy was not demonstrated since the following values were observed: R^2 of 0.45, RER of 4.03, RPD of 1.31, SEP of 0.64, and bias of 1.17. Therefore, the product D could not be correctly predicted with the developed global model.

Furthermore, a SVR_{Uni/Cal} model was created using only the two moisture bands by applying the same preprocessing. The accuracy was also not demonstrated with SEC and SEP values of 0.07 and 1.4, respectively.

Since the accuracy is not demonstrated in such case, this is highly recommended that a qualitative model should be created to insure that the measured product is included in the database, or an update of the database should be performed. A method to exclude NIR spectra to be predicted by the model should be in place. For example, a PCA or any classification methods can be investigated. In this study, a classification method was not established since the scope was focused on prediction possibilities. Consequently, the developed SVR model has some limitations. The model can be applied only for products A, B, and C. A product D is badly predicted since it is not part of the calibration set. In addition, only spectra on the same nature of the ones trained in the calibration (e.g. same products, same composition, same sampling) should be predicted. If a new product has to be predicted by the model, it must be first included in the calibration set and the regression model must be reconstructed with these samples. Finally, the model can be applied only for product A, B, and C produced with freeze-dried cycles trained in the calibration.

4. Conclusion

The use of NIR spectroscopy was presented for the determination of the MC in three different freeze-dried drugs products. The aim was to develop a global regression model to speed up the validation time and the in-lab release analyses. A huge spectral database was created to evaluate several global regression models. For that purpose, the choice of a SVR model was demonstrated with a very good reliability. A SEC value of 0.12% and an SEP value of 0.15% were obtained.

The possibility of using a unique model applied to several products for the rapid and non-destructive analysis of MC represents a step forward on the analysis of in-lab products. The lead time reduction concept is indeed a challenge in the pharmaceutical field. This is why innovative tools are implemented to speed up the quality control of medicines. The faster the analysis, the faster the manufacturer can release the product on the market. The possibility of using a unique method for MC determination instead of several ones is seen as positive not only for the quickness of the analyses, but also for the quality of the results. Indeed, the concrete use of the global approach is similar than the one of individual models and the results remain fully comparable to the conventional KF technique. Moreover, the overall in-lab organization can thus be optimized and the in-lab costs reduced.

References

- [1] Food and Drug Administration (FDA), Guide to inspections of lyophilization of parenterals, 2014 (accessed 01.06.2016). <<http://www.fda.gov/ICECI/Inspections/InspectionGuides/ucm074909.htm>>.
- [2] European Pharmacopoeia 8th Edition 2016 (8.7), Chapter 2.5.12. Water: semimicro determination, 2016.
- [3] A. Kauppinen, M. Toivainen, M. Lehtonen, K. Järvinen, J. Paaso, M. Juuti, J. Ketolainen, Validation of a multipoint near-infrared spectroscopy method for in-line moisture content analysis during freeze-drying, *J. Pharm. Biomed. Anal.* 95 (2014) 229–237.
- [4] K.P. Malik, C. Duru, M. Ahmed, P. Matejtschuk, Analytical Options for the Measurement of Residual Moisture Content in Lyophilized Biological Materials, *American Pharmaceutical review*, 2010.
- [5] Y. Roggo, P. Chalus, L. Maurer, C. Lema-Martinez, A. Edmond, N. Jent, A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies, *J. Pharm. Biomed. Anal.* 44 (2007) 683–700.
- [6] T.P. Lin, C. Hsu, Determination of residual moisture in lyophilized protein using a rapid and non invasive method: near infrared spectroscopy, *PDA J. Pharm. Sci. Technol.* 56 (2002) 196–205.
- [7] J. Jerome, J.R. Workman, Interpretive spectroscopy for near infrared applied spectroscopy reviews, *Appl. Spectrosc. Rev.* 31 (1996) 251–320.
- [8] C. De Bleye, P.-F. Chavez, J. Mantanus, R. Marini, P. Hubert, E. Rozet, E. Ziemons, Critical review of near-infrared spectroscopy methods validations in pharmaceutical applications, *J. Pharm. Biomed. Anal.* 69 (2012) 125–132.
- [9] Y. Jiang, M. Yang, X.-B. Zhang, L.-H. Yin, Research on quantitative models of water content for Chinese herbal pieces by NIR Chinese, *J. Pharm. Anal.* 33 (2013) 1572–1577.
- [10] Z. Xb, F. Yc, H. Cq, Feasibility and extension of universal quantitative models for moisture content determination in beta-lactam powder injections by nearinfrared spectroscopy, *Anal. Chim. Acta* 630 (2008) 131–140.
- [11] X. Yan Mei, S. Xiang Zhong, C. Chang Zhou, L. Hong Ping, T. Guo, H. Yue, D. Jia, M. Shun Geng, The establishment and evaluation of near infrared universal model to determinate the effective ingredient content in pesticide rapidly, *Chin. Chem. Lett.* 23 (2012).
- [12] M. Anastasia, G. Vladislav, S. Elena, B. Andrey, Building global models for fat and total protein content in raw milk based on historical spectroscopic data in the visible and short-wave near infrared range, *Food Chem.* 203 (2016) 190–198.
- [13] M.K.D. Rambo, M.M.C. Ferreira, E.P. Amorim, Multi-product calibration models using NIR spectroscopy, *Chemometr. Intell. Lab. Syst.* 151 (2016) 108–114.
- [14] X.-P. Liu, Y.-C. Feng, C.-Q. Hu, L. Ding, Construction of universal quantitative models for determination of cefradine capsules, *Chin. J. Pharm. Anal.* 28 (2008).
- [15] H.h. Pang, Y.C. Feng, C.Q. Hu, X.B. Ren, Construction of universal quantitative models for determination of cefoperazone sodium for injection from different manufacturers using near infrared reflectance spectroscopy, spectroscopy and spectral analysis, 26 (2006) 2214–2218.
- [16] W. Xiaoliang, F. Qiang, S. Jinfang, Y. Xin, J. Jianzhong, D. Wei, Construction of a universal quantitative model for ibuprofen sustained-release capsules from different manufacturers using near-infrared diffuse reflection spectroscopy, *Vib. Spectroscopy* 53 (2010) 214–217.

- [17] Y. Lei, Z.-Y. Luo, C.Q. Hu, Rapidly screening counterfeit drugs using near infrared spectroscopy: combining qualitative analysis with quantitative analysis to increase effectiveness, *J. Near Infrared Spectroscopy* 16 (2008) 349–355.
- [18] F. Yan-Chun, H. Chang-Qin, Construction of universal quantitative models for determination of roxithromycin and erythromycin ethylsuccinate in tablets from different manufacturers using near infrared reflectance spectroscopy, *J. Pharm. Biomed. Anal.* 41 (2006) 373–384.
- [19] Z. WenBo, F. YanChun, D. JiXiong, S. DanQing, H. ChangQin, A new strategy to iteratively update scalable universal quantitative models for the testing of azithromycin by near infrared spectroscopy, *Sci. China Chem.* 56 (2013) 533–540.
- [20] J. Yh, L. Xp, F. Yc, H. Cq, A training set selection strategy for a universal near-infrared quantitative model, *AAPS PharmSciTech.* 12 (2011) 738–745.
- [21] B.G. Osborne, T. Fearn, Collaborative evaluation of universal calibrations for the measurement of protein and moisture in flour by near infrared reflectance, *J. Fd Technol.* 18 (1983) 453–460.
- [22] E. Carl Emil, H. Per Waaben, S. Thomas, M. Federico, N. Lars, Evaluation of multivariate calibration models transferred between spectroscopic instruments: Applied to near infrared measurements of flour samples, *J. Near Infrared Spectroscopy* 24 (2016).
- [23] M. Clavaud, Y. Roggo, K. Dégardin, P.-Y. Sacré, P. Hubert, E. Ziemons, Moisture content determination in an antibody-drug conjugate freeze-dried medicine by near-infrared spectroscopy: a case study for release testing, *J. Pharm. Biomed. Anal.* 131 (2016) 380–390.
- [24] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, *Appl. Spectroscopy* 43 (1989) 772–777.
- [25] M. Blanco, J. Coello, H. Iturriaga, S. Maspocho, C. De la Pezuela, Effect of data preprocessing methods in near-infrared diffuse reflectance spectroscopy for the determination of the active compound in a pharmaceutical preparation, *Appl. Spectroscopy* 51 (1997) 240–246.
- [26] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, *Chemometr. Intell. Lab.* 58 (2001) 109–130.
- [27] Springer, Pattern recognition and machine learning, in: M. Jordan, J. Kleinberg, B. Schölkopf (Eds.), Singapore, 2006.
- [28] T. Chen, E. Martin, Bayesian linear regression and variable selection for spectroscopic calibration, *Anal. Chim. Acta* 631 (2009) 13–21.
- [29] A.E. Hoerl, R.W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, *Technometrics* 12 (1970) 55–67.
- [30] M. Czajkowski, M. Kretowski, The role of decision tree representation in regression problems – an evolutionary perspective, *Appl. Soft Comput.* 48 (2016) 458–475.
- [31] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning - Data Mining, Inference, and Prediction - Second Edition, in: S.S.i. Statistics (Ed.), 2008.
- [32] N. Christianini, J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge University Press, New York, 2000.
- [33] P. Hubert, J.-J. Nguyen-Huu, B. Boulanger, E. Chapuzet, P. Chiap, N. Cohen, P.-A. Compagnon, W. Dewe, M. Feinberg, M. Lallier, M. Laurentie, N. Mercier, G. Muzard, C. Nivet, L. Valat, Harmonization of strategies for the validation of quantitative analytical procedures. A SFSTP proposal—part I, *J. Pharm. Biomed. Anal.* 36 (2004) 579–586.
- [34] D. Ascher, P.F. Dubois, K. Hinsén, J. Hugunin, T. Oliphant, Numerical Python, 1999 (accessed 24.01.2017) <<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.476.2204&rep=rep1&type=pdf>>.
- [35] T.E. Oliphant, Python for scientific computing, *Comput. Sci. Eng.* 1521–9615 (2007) 10–20.
- [36] S. Van der Walt, S.C. Colbert, G. Varoquaux, The NumPy array: a structure for efficient numerical computation, *Comput. Sci. Eng.* 1521–9615 (2011) 22–30.
- [37] G.D. Batten, Plant analysis using near infrared reflectance spectroscopy: the potential and the limitations, *Aust. J. Exp. Agr.* 38 (1998) 697–706.
- [38] P. Williams, K. Norris (Eds), Variables affecting near-infrared reflectance spectroscopic analysis, in: P.C. Williams (Ed.), *Am. Cereal Assoc. Cereal Chem.*, St. Paul, 1987, pp. 143–167.