Feasibility study of smartphone-based Near Infrared Spectroscopy (NIRS) for salted minced meat composition diagnostics at different temperatures

Andreas KARTAKOULLIS[1], Josep COMAPOSADA[1], Alvaro CRUZ-CARRIÓN[1,2], Xavier SERRA[1], Pere GOU[1,3]

andreas.kartakoullis@irta.cat, josep.comaposada@irta.cat, alvarojavier.cruz@urv.cat, xavier.serra@irta.cat, pere.gou@irta.cat

Abstract

This research work evaluates the feasibility of a smartphone-based spectrometer (740-1070 nm) for salted minced meat composition diagnostics at industrial scale. A commercially available smartphone-based spectrometer and a benchtop NIR spectrometer (940-1700 nm) were used for acquiring 1312 spectra from minced meat samples stored at four different temperatures ranging from -14 ºC to 25 ºC. Thereafter, for each spectrometer, PLS and Random Forest regression models specific for each temperature and global models were created to predict the fat, moisture and protein contents. Fat and moisture can be estimated with the global model in a wide range of temperatures by using the smartphone-based spectrometer, which has an acceptable accuracy for quality control purposes (RPD>7) and comparable to the accuracy of a benchtop spectrometer.

Highlights:

- Smartphone-based NIR was used to estimate salted meat composition
- Accuracy of fat and moisture estimates are acceptable for quality control purposes
- Predictive ability of smartphone-based NIR is similar to a benchtop spectrometer

[1] IRTA, Food Technology Program, Finca Camps i Armet 17121 Monells, Girona, Spain
[2] Universitat Rovira i Virgili, C/ Marcel·lí Domingo, 2-4-6. 43007 Tarragona, Spain
[3] Corresponding author: +34-902789449 Ext. 1407

- Random Forest predictive models are robust at temperatures from -14 ºC to 25 ºC

1. Introduction

Smartphone-based food diagnostic technologies have the potential to revolutionize the food sector by allowing rapid, on-site and inexpensive analysis of food and food products (Rateni, Dario, & Cavallo, 2017). This potential has been recognized by the European Union (EU) and a Horizon Prize competition called "Food Scanner" took place in 2016 for developing affordable and non-invasive mobile solutions that will enable users to measure and analyze their food intake. The winners were three European startups: Spectral Engines (winner), SCIO (runner-up) and TellSpec (runner-up) (Horizon, 2016). These three companies have developed novel smartphone-based food diagnostic applications by utilizing the analytical power of Near Infrared Spectroscopy (NIRS). NIRS utilize the spectral range from 800nm to 2500nm in order to provide chemical information through the different response of molecular bonds and structural information of the scanned sample (Burns & Ciurczak, 2007).

Several research studies have been published in the last years in the agrifood sector that are taking advantage of the combined use of the rear smartphone camera with augmented sensors. For example, bio-receptors (e.g. fluorescent dyes) and colorimetric assays have been used for the on-site detection of contaminants (Coskun, et al., 2017); (Chen, et al., 2014), pesticide residues (Levin, Krishnan, Rajkumar, Halery, & Balkunde, 2016), (Wang, et al., 2016) and foodborne pathogens (Zhu, Sikora, & Ozcan, 2012), (Seo, et al., 2016). As for commercial applications, a great number of startups worldwide are offering smartphone-based NIRS by exploding the technological advancements in the high-volume fabrication of microelectromechanical systems (MEMS) and micro-mirror arrays. These advancements have made possible the construction of miniaturized and inexpensive NIR spectrometers while keeping the performance at similar levels with the benchtop equipment (Ozcan, 2016). Device schematics and details

on how to design a low-cost smartphone-based spectrometers from off-the-shelf components can be found in the literature (Das, Wahi, Kothari, & Raskar, 2016).

Spectral Engines solution uses a variety of Bluetooth-connected NIR sensors together with advanced analytics, cloud-connectivity and spectral libraries to reveal the fat, protein, sugar, and calories of food items (Antila, Kantojärvi, & Mäkynen, 2016). TellSpec is using a pocket-sized NIR spectrometer based on the Texas Instruments DLP module for scanning food through a mobile application in order to identify calories, allergens, contaminants and food frauds (Beam Your Health Up–TellSpec, 2018). Consumer Physics has the in-house-developed SCIO, a pocket-sized NIR spectrometer for molecular analysis of different samples, including food (Goldring & Sharon, 2011). In addition to the advancements in miniaturizing spectrometers, a mobile phone-compatible hyperspectral imager based on a tunable MEMS Fabry-Perot interferometer has been demonstrated for food sensing applications among other applications (Rissanen, et al., 2016).

The abovementioned innovations have developed reliable food scanning mobile applications but are mostly targeting the customer. There is a need to develop more robust smartphone-based food diagnostic applications especially for the meat industry in order to ensure food quality and safety. NIRS is a widely used analytical technique in the food industry for the determination of the chemical composition because of its non-destructive, fast and sensitive abilities, and requires little or no sample preparation (Prieto, Roehe, Lavin, Batten, & Andres, 2009) (Van den Berg, Lyndgaard, Sørensen, & Engelsen, 2013). Furthermore, NIRS coupled with multivariate analysis can determine the water and salt content at the surface of fermented sausages during the drying process (Collell C. , Gou, Arnau, Muñoz, & Comaposada, 2012), water and salt content at the surface of dry-cured ham during the resting and drying processes (Collell C. , Gou, Arnau, & Comaposada, 2011) and the authentication of meat products (Chiesa, et al., 2016).

A limitation for the adoption of NIRS-based innovations is the sample temperature because it can cause significant variation to the acquired spectra. When the light of wavelengths longer than 450 nm is absorbed by liquid water, the energy is transferred to one or more of the vibrational modes of the O–H bond.

Spectroscopic studies of liquid water have shown that the shift in frequency is attributed to the weakness of intermolecular H-bonds due to the rise of temperature, which strengthens the covalent O-H bonds and consequently causes the water molecule to vibrate at higher frequencies (Segtnan, Sasic, Isaksson, & Ozaki, 2001). Furthermore, the absorption of light from liquid water at the visible and NIR region is getting stronger due to the rise of temperature above the freezing point of water (Pegau, Deric, & Zaneveld, 1997). Unfortunately, the temperature influence is usually ignored in NIRS predictive modeling because the temperature control, especially in industrial conditions, is quite difficult. However, recent studies have used compensation methods to mitigate the temperature influence and to improve the accuracy of the prediction by employing temperature-compensated PLSR global models (Yao, Chen, Xie, & Rao, 2013) (Campos, Antolin, Deban, & Pardo, 2018).

The aim of this paper is to study the feasibility of smartphone-controlled near-infrared spectrometers for salted minced meat composition prediction at different temperatures in comparison to a benchtop spectrometer. Additionally, linear and nonlinear regression methods have been used for constructing global models with the smartphone-based spectrometer that are robust for a great range of temperatures, from -14^{o}C to 25^{o}C. Furthermore, the effect of different data partitioning methods on how they affect the construction of predictive models is also investigated.

2. Material and Methods

2.1 Experimental Design

Green hams (n= 328) from different origins were purchased: 171 Serrano-type hams from Jamonificio Subirats, S.L. (Els Hostalets de Balenyà, Barcelona, Spain), 72 Serrano-type hams from Càrniques de Juià, S.A. (Juià, Girona, Spain), 30 Duroc Serrano-type hams from Cárnica Batallé, S.A. (Riudarenes, Girona, Spain) and 55 Iberian hams from Corporación Alimentaria Guissona, S.A., (Guissona, Lleida, Spain).

Green hams were salted by covering them with a layer of salt of at least 10 cm for different periods of time (7 days to 16 days) in order to obtain a high variability in

salt uptake, and therefore, in composition. Each green ham was carefully dissected and the subcutaneous fat, lean and intermuscular fat were minced together in a bowl chopper until a homogeneous meat paste was obtained. Four samples from each ham were vacuum packed and stored at different temperatures (-14 ºC, +5 ºC, +12 ºC and +25 ºC) until data acquisition. The whole procedure resulted in samples with fat content varying from 5% to 43%, protein content varying from 12% to 23% and moisture content varying from 35% to 69%.**2.2 Instrumentation**

Two NIR instruments were used. The first one is the smartphone-based SCiO (Consumer Physics, Israel) that can acquire 331-points reflection spectra that are ranging from 740 to 1070 nm. For controlling the SCiO spectrometer a Samsung Galaxy Core Prime with an Android 5.1.1 operating system was used. The second NIR instrument is a diode array Polychromix Spectral Probe (Polychromix Inc., Wilmington, USA) with InGaAs detector type that is covering the spectral region between 940 and 1700 nm with a resolution of 8 nm. Polychromix was used to compare the results obtained with SCiO. Polychromix reflectance data were stored as the logarithm of reciprocal of reflectance (1/R). For both instruments, three spectra per sample were collected and the mean spectrum per sample was used for further analysis. All the equipment was set in reflectance mode and operated at room temperature (20 ± 3 °C).

The moisture, protein and fat contents of the meat samples were determined by using a FoodScan™ Lab (Foss Analytical, Hillerød, Denmark), which is an approved method for the analysis of moisture, fat and protein in meat and meat products by the Association of Analytical Communities, labeled as 2007.04 method (AOAC, 2007).

2.3 NIRS Analysis

2.3.1 Data Acquisition

Data acquisition was performed using the two different NIRS systems. Each sample was removed from the storage rooms and immediately three SCiO scans and three Polychromix scans were acquired at three different random spots and subsequently averaged. Great attention was given to keep the total acquisition

time for both the instruments less than 2 minutes in order to ensure that the temperature did not change significantly the sample composition during the spectra acquisition.

2.3.2 Data partitioning

Dataset from each acquisition temperature was divided in two: a training dataset with 80% of the initial data and a testing dataset with 20% of the initial data. The general aim of performing the data partitioning is to prevent knowledge from the training dataset to "leak" to the testing dataset, which will affect the predictive power of the PLSR model (Wold, Sjöström, & Eriksson, 2001). Two different methods for data partitioning have been selected, the first one is using only the responses **Y** to construct the two datasets by separating the values of **Y** with random stratification into groups. Each of the stratified groups contains approximately the same percentage of each response as in the original **Y** dataset (Molinaro, Simon, & Pfeiffer, 2005). For constructing the global model, the datasets of each temperature were merged and the training and testing datasets were obtained with the random stratification on the Y-dataset.

The second method is to use the properties of the NIR spectra from the predictors **X** dataset. The Kennard-Stone (KS) algorithm was used for this partitioning, which is a sequential method that covers the spectral space uniformly in order to build a training dataset with flat distribution over the spectral space. It starts by computing the geometric distances (e.g. Euclidean) between all the pairs of spectra and the two spectra that are the farthest apart are assigned to the training dataset. The selection procedure is repeated until to build the training dataset according to the following maximin criterion, the next selected spectra must have the least distance with respect to any other spectra already selected (Puzyn, Mostrag-Szlichtyng, Gajewicz, Skrzyński, & Worth, 2011). Mahalanobis distance can be used also for satisfying the maximin criterion by performing Principal Component Analysis (PCA) on the spectra and computing the Euclidean distance of the PCA scores according to the following definition of the Mahalanobis distance:

$$H_{ij}^2 = \sum_{n=1}^{N} \left(\hat{t}_{i,n} - \hat{t}_{j,n}\right)^2 / \hat{\lambda}_n \quad (1)$$

where $\hat{t}_{i,n}$ is the n-th principal component score of point i, $\hat{t}_{j,n}$ the corresponding principal component score for point j, $\hat{\lambda}_n$ is the eigenvalue of principal component n and N is the number of the selected principal components.

2.3.3 Data pretreatment

The scatter correction methods include Multiplicative Scatter Correction (MSC), Inverse MSC, Extended MSC, Extended Inverse MSC, detrending, Standard Normal Variate (SNV) and different kinds of normalization to unit-vector length (e.g. scaling). The most common spectral derivatives methods are the gap-segment (also known as Norris-Williams) derivatives and the Savitzky-Golay (SG) polynomial derivative filters (Rinnan, van den Berg and Engelsen 2009).

The SG method is fitting polynomials to windows around each point in the spectrum and these polynomials are then used to smooth the obtained data and subsequently differentiate them. Finite difference method can be used instead of the polynomial fitting for the computation of the first and second derivative by considering two different points of the spectrum:

$$x_i' = x_i - x_{i-1} \quad (2)$$

$$x_i'' = x_{i-1} - 2x_i + x_{i+1} \quad (3)$$

This numerical subtraction deemphasizes lower frequencies and emphasizes higher frequencies (e.g. noise). Data smoothing methods are needed in order to improve the signal-to-noise ratio of the obtained data. The gap-segment method first performs a smoothing under a given segment size, followed by a gap derivative:

$$x_i' = x_{i+k} - x_{i-k} \quad (4)$$

$$x_i'' = x_{i-k} - 2x_i + x_{i+k} \quad (5)$$

where k is the gap size between the points.

2.3.4 Predictive modelling

Partial Least Squares (PLS) Regression (also known as Projection of Latent Structures) is one of the most popular tools for multivariate analysis (Wold H. ,

1966). PLSR derives its popularity from the ability to analyze data with multiple, noisy, collinear (e.g. NIR spectra), and even incomplete variables. Its goal is to predict a set of dependent variables (denoted as **Y**) from a set of predictors (denoted as **X**). The prediction is achieved by extracting from the predictors a new set of orthogonal factors called Latent Variables (LVs) that have the best predictive power on unseen data. More specifically, PLSR tries to find two sets of weights **w** and **c** in order to extract two vectors from **X** and **Y, t**=**Xw** and **u=Yc** such that maximize the covariation between **X** and **Y**. It can be achieved with the following cost function:

$$\text{maximize } \mathrm{cov}(\mathbf{Xw}, \mathbf{Y}) \text{ subject to } \mathbf{w^T w}=1 \quad (6)$$

The number of LVs defines the complexity of the predictive model and selecting the optimal number is one the most important steps because the amount of variance explained by a LV indicates its importance in the prediction of **Y**. For example, selecting too many LVs will result in an over-fitted model that takes into consideration not only the variance of the data but also noise. While selecting too few LVs implies an under-fitted model, which incorporates insufficient information of the data.

In the current study, the target **Y** variables describe the concentration of certain components and **X** are the spectral data. The PLSR analysis was applied to the training dataset for each acquisition temperature. We also constructed a global model for investigating further the temperature dependency on the predictive ability of the PLSR model. For each instrument, the four training datasets acquired at the four different temperatures were merged to create a global model for each instrument. The Root Mean Squared Error of Prediction (RMSEP), the coefficient of the determination (R^2) and the residual predictive deviation (RPD) were used to decide the number of LVs.

The RMSE of calibration was calculated:

$$RMSEC = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{N - LVs - 1}} \quad (7)$$

2.3.5 Validation of predictive models

Predictive models obtained after PLSR analysis of training dataset were applied to the testing data set. The RMSEP, R^2 and RPD were calculated:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{N}} \quad (8)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2} \quad (9)$$

where N is the size of the dataset, Y_i is the experimentally measured reference value for sample i, $\hat{Y}_i$ is the predicted reference value for sample i and $\bar{Y}$ is the arithmetic mean of the corresponding dataset.

RPD is the ratio between the standard deviation of the reference values and the error of prediction of the testing. It has been suggested that RPD values >3 are considered good for screening purpose; RPD values >5 are good for quality control, whereas RPD values >8 are considered excellent for all analytical tasks (Conzen, 2006).

2.4 Software and Algorithms

The open-source programming language R (version 3.4.0, codename: "You Stupid Darkness") (The R Core Team, 2018) was used to develop in-house codes based on the PLS and randomForest packages (Mevik & Wehrens, 2007) (Liaw & Wiener, 2002).

3. Results and Discussion

3.1 Spectral description of the samples

Figure 1 shows the raw mean spectrum of all the collected spectra at each one of the acquisition temperatures. As expected, there are differences between the different temperatures even with a shift in frequency (frozen sample at -14 °C) due to perturbations caused by the increase in temperature (Fig. 1B). The primary vibrational modes of liquid water are in the mid-infrared region at 3049 nm (ν_1, symmetric stretch), 6079 nm (ν_2, bend), and 2865 nm (ν_3, asymmetric stretch). Several higher overtones of the vibrational modes can be found at the detection area of our equipment between 750-1800nm and therefore the whole spectrum

from each sample will be used for the construction of the predictive model in order not to exclude any informative part of the spectrum.

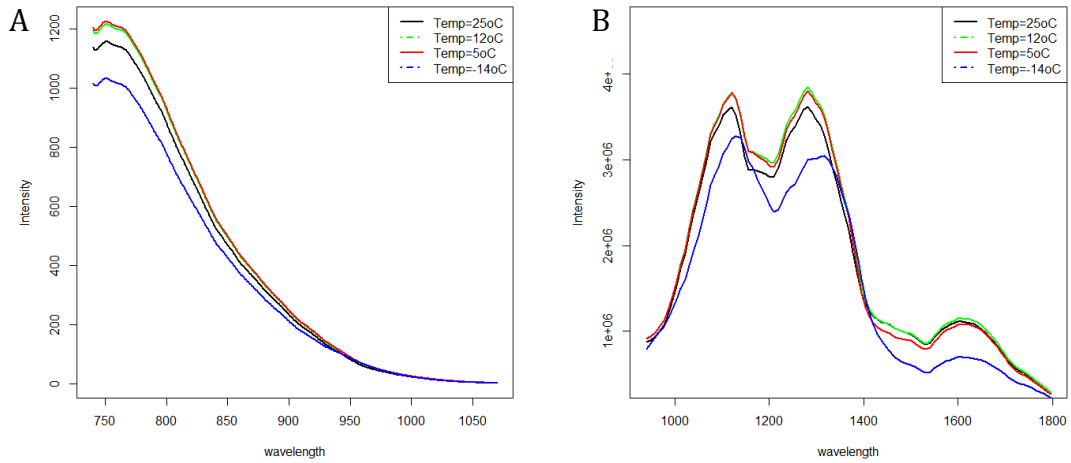
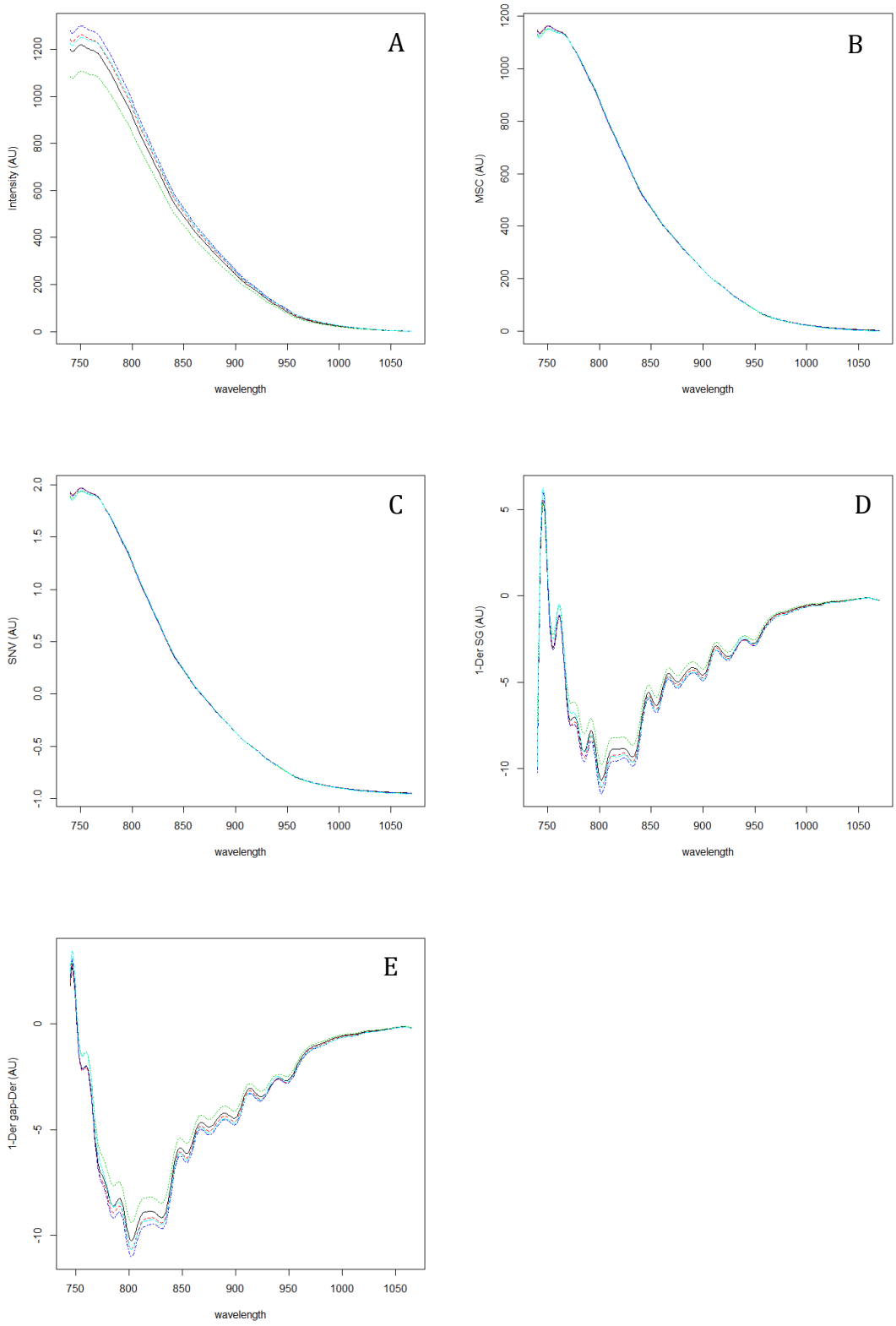


Figure 1: The mean spectrum for each temperature for SCiO (A) and Polychromix (B).

3.2 Pretreatment of raw spectra

Figure 2 shows the effect of the pretreating methods upon 5 different SCiO spectra collected from samples stored at 25 °C. As expected the SNV and MSC are returning almost identical spectra with the only difference to be the offset of the intensity. The spectra obtained from the two spectral derivatives methods shows distinct differences between them, because of the different methods used for the computation of the derivatives. The spectral derivatives need thoroughly tuning because they have a number of parameters that are critical for building a predictive model, such as the order and length of the polynomial for the SG, and the size of the gap and the segment size of the smoothing filter for the gap-segment.



277 *Figure 2: SCiO spectra without any pretreatment (A), with MSC pretreatment (B), with SNV pretreatment (C),*
278 *with SG-1Der pretreatment (D), and with gap-segment derivative pretreatment (E).*

279 Several pretreatments have been used in order to build a PLSR model with the
280 best overall predictive ability for the two different spectrometers. For scatter

correction, we have selected to use only SNV for practical reasons. The signal-correction concept behind the SNV and MSC is similar except that a reference spectrum is needed for the MSC correction. Usually, the reference spectrum is the mean spectrum of all the measurements and that can lead to the so-called *data leakage.* Data leakage is when unexpected information is being added to the training dataset that can allow the model to learn something that it otherwise would not know. This will hinder the performance of the predictive model under construction (Nisbert, Elder, & Miner, 2009). Therefore, the selection of the MSC reference spectrum is a great challenge and that gives to SNV a practical advantage (Rinnan, van den Berg, & Engelsen, 2009). As for spectral derivatives, the gap-segment method performs better with the SCiO data compared to the SG method that performs better with the Polychromix data.

3.3 Predictive ability of NIR devices with two different data partitioning methods

Table 1 and Table 2 show, for both NIR devices the predictive ability of the different PLSR models created by using the random stratification method and the KS method respectively for partitioning the data into training and testing datasets.

The predictive models were constructed on the SNV+gap-seg and SNV+SG pretreated datasets. Fat, protein and moisture content of the samples were accurately predicted with the lowest value of the coefficient of the determination to be R^2=0.92 and with 4 LVs for protein estimation at 25 °C. The RPD values for all the predictions are above 3 and in some cases above 8, which means that our PSLR modelling could be accurate enough for quality control of fat and moisture contents and for screening purposes of protein content. It should be noted that RMSEP, R^2 and RPD for fat and moisture prediction are worst in frozen samples, especially with SCiO device. RMSEP values for fat and moisture prediction are similar to the results reported in previous studies where benchtop spectrometers have been used in minced mass of pork sausages, for example SEP: 0.94-1.41% for fat and 0.76–1.01% for moisture (Ortiz-Somovilla, Espana-Espana, Gaitan-Jurado, Perez-Aparicio, & De Pedro-Sanz, 2007), RMSEP= 0.622-0-675% for moisture (Collell C. , Gou, Picouet, Arnau, & Comaposada, 2010), and in minced *Biceps*

femoris muscle of dry-cured ham RMSEP: 0.31-0.43% for fat and 0.44-0.51% for moisture (Prevolnik, et al., 2011).

The results obtained with datasets built with the KS method for partitioning the data (Table 2) are comparable to those obtained with datasets built with the stratified random partitioning based on Y values (Table 1). KS method is concentrating most of the diversity of the data at the training set that in some cases can lead to overoptimistic results (Wested & Marini, 2015), which is not the case for our study since the two data partitioning methods are returning similar results. Similar methods to KS, such as the Duplex algorithm (Daszykowski, Walczak, & Massart, 2002), have great potential for building new predictive models with smartphones because they require less computational effort compared to the random stratification.

Table 1: Results of PLSR analysis after dataset partitioning based on random stratification sampling applied to response values (Y) for each temperature of spectra acquisition.

Parameter	Pretreatment	LVs	RMSEC	RMSEP	R^2	RPD
SCiO, Temp=-14°C						
Fat	SNV+gap-seg	6	1.26%	1.47%	0.97	6.89
Moisture		6	1.23%	1.36%	0.96	5.48
Protein		6	0.43%	0.54%	0.93	3.89
Polychromix, Temp=-14°C						
Fat	SNV+SG (1D)	4	0.99%	0.92%	0.99	10.34
Moisture		5	0.78%	0.76%	0.99	7.46
Protein		5	0.49%	0.48%	0.94	3.91
SCiO, Temp=5°C						
Fat	SNV+gap-seg	5	0.97%	1.00%	0.99	6.99
Moisture		5	0.89%	1.06%	0.98	6.86
Protein		5	0.49%	0.43%	0.95	3.87
Polychromix, Temp=5°C						
Fat	SNV+SG (1D)	6	0.85%	0.74%	0.99	14.11
Moisture		6	0.82%	0.88%	0.98	7.99
Protein		5	0.42%	0.49%	0.93	4.76

SCiO, Temp=12oC						
Fat	SNV+gap-seg	4	1.05%	0.94%	0.99	12.05
Moisture		5	0.96%	0.83%	0.98	8.55
Protein		5	0.41%	0.45%	0.96	4.18
Polychromix, Temp=12°C						
Fat	SNV+SG (1D)	6	0.86%	0.92%	0.99	10.79
Moisture		4	0.79%	0.88%	0.98	9.76
Protein		4	0.46%	0.45%	0.95	5.04
SCiO, Temp=25°C						
Fat	SNV+gap-seg	6	0.97%	0.74%	0.99	12.65
Moisture		4	0.85%	0.71%	0.98	9.05
Protein		5	0.45%	0.47%	0.95	3.99
Polychromix, Temp=25°C						
Fat	SNV+SG (1D)	5	0.92%	0.86%	0.99	10.65
Moisture		5	0.76%	1.01%	0.98	9.56
Protein		6	0.38%	0.36%	0.96	4.09

Table 2: Results of PLSR after dataset partitioning based on the KS method applied to X descriptors (NIR spectra).

Parameter	Pretreatment	LVs	RMSEC	RMSEP	R^2_p	RPD
SCiO, Temp=-14°C						
Fat	SNV+gap-seg	6	1.23%	1.43%	0.98	7.12
Moisture		6	1.29%	1.02%	0.97	6.34
Protein		5	0.48%	0.49%	0.92	3.29
Polychromix, Temp=-14°C						
Fat	SNV+SG (1D)	4	1.13%	0.76%	0.99	7.51
Moisture		5	0.82%	0.72%	0.98	8.79
Protein		4	0.50%	0.45%	0.94	2.64
SCiO, Temp=5°C						
Fat	SNV+gap-seg	4	1.11%	0.83%	0.99	11.64
Moisture		4	0.99%	0.78%	0.99	8.57
Protein		6	0.46%	0.43%	0.94	3.87

Polychromix, Temp=5°C						
Fat	SNV+SG (1D)	4	0.95%	0.90%	0.99	9.38
Moisture		5	0.83%	0.90%	0.98	7.96
Protein		6	0.42%	0.47%	0.93	4.11
SCiO, Temp=12°C						
Fat	SNV+gap-seg	5	1.00%	1.19%	0.98	10.15
Moisture		5	0.93%	1.12%	0.97	7.95
Protein		6	0.46%	0.52%	0.93	3.21
Polychromix, Temp=12°C						
Fat	SNV+SG (1D)	4	1.02%	0.76%	0.99	8.54
Moisture		5	0.83%	0.79%	0.98	6.83
Protein		4	0.48%	0.43%	0.93	3.35
SCiO, Temp=25°C						
Fat	SNV+gap-seg	6	1.01%	0.80%	0.99	10.50
Moisture		6	0.87%	0.78%	0.98	7.54
Protein		5	0.49%	0.37%	0.95	3.37
Polychromix, Temp=25°C						
Fat	SNV+SG (1D)	4	0.98%	0.68%	0.99	9.46
Moisture		4	0.76%	0.84%	0.99	7.82
Protein		4	0.44%	0.47%	0.92	3.41

3.4 Global Model

For constructing the global model, the datasets of different temperatures were merged and the training and testing datasets were obtained with the random stratification on the Y-dataset. Table 3 shows that when a global temperature model is applied, in most of the cases the SCiO has results similar to those obtained with local temperature models. The predictive ability for the fat content of the SCiO global model has slightly improved for the -14 °C samples from RMSEP=1.47% to RMSEP=1.18% and for the 5 °C samples from RMSEP=1.00% to RMSEP=0.88%. Predictive errors do not show any specific pattern related to the temperature (Figure 3a-c).

However, Figure 3b shows that the predicted values of moisture are always lower than the measured values at the midrange (53-59%). Figure 3a for fat predictions has a similar trend at the midrange but to a lesser extent. There is a non-linear relationship between moisture or fat content and spectral response over large ranges of fat or moisture. PLS, as a linear regression method, cannot handle this non-linear relationship and some part of the explanatory information is going to the residuals of the model. Therefore, for estimating correctly the predicted values of moisture and fat, non-linear regression methods are needed such as Random Forest (RF) regression that has the ability of fully utilizing the explanatory information from various types of data (Hastie, Tibshirani, & Friedman, 2009).

RF is an ensemble learning algorithm that can be used for both classification and regression. It is using a large number of decision trees (ntree) to split the data into an equal number of bootstrap samples that will be randomly sampled by a number of predictors (mtry). This adds randomness to the model while growing the number of the decision trees, and the algorithm chooses the best split between the sampled variables leading to a wide diversity that generally results in a better model (Breiman, 2001).

Figure 3e-f depicts the new prediction plots and it is obvious that the RF regression has corrected the non-linear relationship between moisture and fat content. The RMSEC values at Table 2 have significantly improved because the model captures all the explanatory information from the data, but the RMSEP values are at the same levels as PLS results.

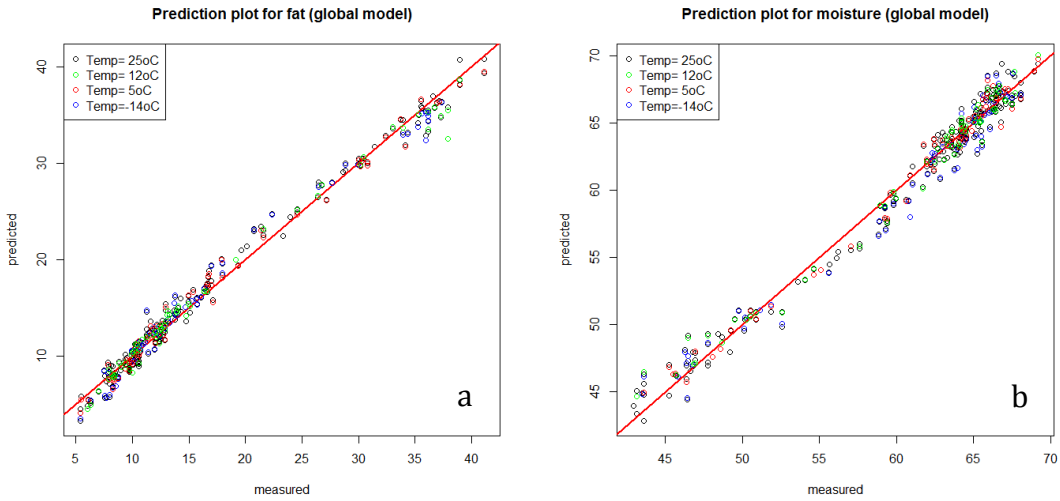
Therefore, the RF regression global temperature model build with smartphone-based NIR spectrometers has the potential to predict fat and moisture contents at different temperatures, even in frozen samples, with high accuracy and protein content with good accuracy.

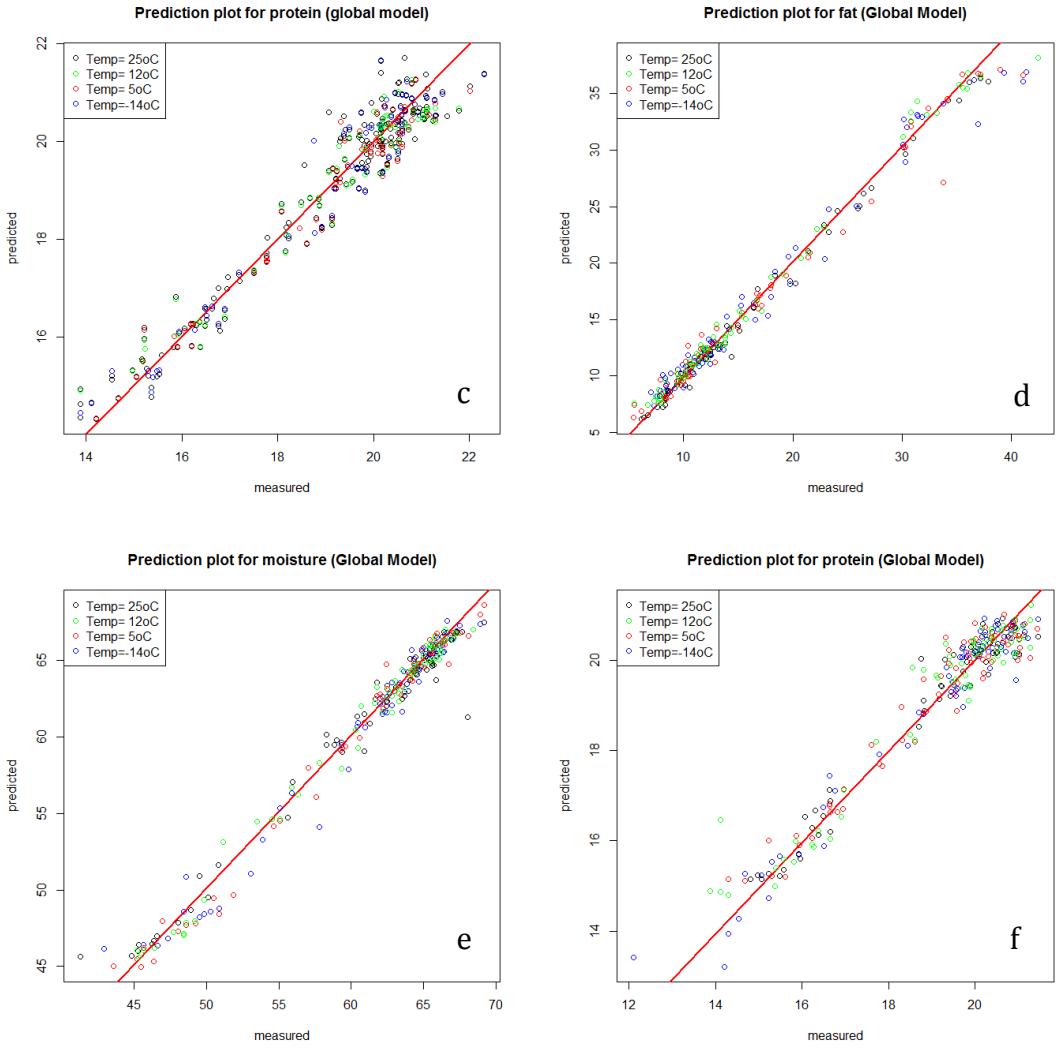
Table 3: Results of the global PLS and RF regression after datasets selection based on random stratification sampling applied to response values (Y) for all temperatures of spectra acquisition.

Parameter	Pretreatment	Parameter	RMSEC	RMSEP	R^2_p	RPD
PLS SCiO Global						
Fat	SNV+gap-seg	8 LVs	1.21%	1.07%	0.99	8.89

Moisture		7 LVs	1.01%	1.03%	0.98	7.16
Protein		7 LVs	0.46%	0.49%	0.94	3.94
RF SCiO Global						
Fat	SNV+gap-seg	ntree=500 mtry=104	0.58%	1.18%	0.98	7.91
Moisture		ntree=500 mtry=106	0.47%	0.98%	0.98	6.88
Protein		ntree=500 mtry=105	0.21%	0.48%	0.94	4.21

366





367 *Figure 3: Prediction plots for the SCiO global model using PLS regression (a-c) and RF regression (d-f).*

368 **3.5 Prediction error**

369 The whole procedure of this study has resulted in a predictive model with high
370 RPD values. However, NIRS-based predictions were used for obtaining the
371 reference values of **Y** instead of using wet chemistry methods and this introduces
372 an additional error to our final predictive model. The Foss FoodScan has the
373 following Standard Errors of Prediction (SEP): for fat 1.01%, for moisture 0.72%
374 and for protein 0.62%. By following simple error propagation, our global RF
375 predictive model has the following SEPs: for fat 1.56%, for moisture 1.21% and for
376 protein 0.78%.

377 **4. Conclusion**

Robust, efficient and reliable methods for predicting minced meat composition at four temperatures (-14^{o}C, 5^{o}C, 12^{o}C and 25^{o}C) have been demonstrated by using a smartphone-based spectrometer. The predictive ability of the smartphone-based spectrometer is comparable with the one obtained from a benchtop spectrometer with in-house developed predictive analytics. Two different data partition techniques were applied to construct a PLSR predictive model, which perform well for moisture and fat contents even for frozen meat samples (RPD>7). Furthermore, a global model with all the temperatures was constructed for further examining the abilities of the smartphone-based spectrometer. In this case, the PLS modelling has shown problems to capture all the explanatory information from the global model, which was leaked to the residuals of the model. Random Forest regression was used instead of PLS, which improved the predictive modelling.

It is expected that as the technological improvements will continue, the smartphones will include greater computational power, high-end sensing technologies and higher connectivity. This together with the massive volume of smartphone users, which is over than 7 billion, will provide a paradigm shift in how measurement science for food quality and safety will be practiced globally.

Acknowledgement

The research leading to these results has received funding from the project "D1-17 NIR low cost to control quality parameters of dry fermented meat products" of the Innovative Business Groups of the Spanish Ministry of Economy, Industry and from the CERCA programme from Generalitat de Catalunya. Andreas Kartakoullis acknowledges financial support from the Marie Sklodowska-Curie COFUND P-SPHERE project, which has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 665919.

Conflicts of Interest Statement

The authors whose names are listed on the first page of this manuscript certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in

speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

References

Antila, J., Kantojärvi, U., & Mäkynen, J. (2016). *USA Patent No. WO2016071572A1.*

AOAC. (2007). Determination of fat, moisture and protein in meat and meat products by using the Foss Foodscan Near-Infrared Spectrophotometer with Foss Artificial Neural Network calibration model and associated database. *Journal of AOAC International*, 1073-1083.

Beam Your Health Up–TellSpec. (2018). Retrieved from Tellspec: http://tellspec.com/en/

Breiman, L. (2001). Random Forests. *Machine Learning, 45*, 5-32.

Burns, D., & Ciurczak, E. (2007). *Handbook of Near-Infrared Analysis, Third Edition.* CRC Press.

Campos, M., Antolin, G., Deban, L., & Pardo, R. (2018). Assessing the influence of temperature on NIRS prediction models for the determination of sodium content in dry-cured ham slices. *Food Chemistry, 257*, 237-242.

Chen, A., Wang, R., Bever, C., Xing, S., Hammock, B., & Pan, T. (2014). Smartphone-interfaced lab-on-a-chip devices for field-deployable enzyme-linked immunosorbent assay. *Biomicrofluidics, 8*, 064101.

Chiesa, L., Panseri, S., Bonacci, S., Procopio, A., Zecconi, A., Arioli, F., . . . Moreno-Rojas, J. (2016). Authentication of Italian PDO lard using NIR spectroscopy, volatile profile and fatty acid composition combined with chemometrics. *Food Chemistry, 212*, 296-304.

Collel, C., Gou, P., Arnau, J., Muñoz, I., & Comaposada, J. (2012). NIR technology for on-line determination of superficial aw and moisture content during the drying process of fermented sausages. *Food Chemistry, 135*, 1750–1755.

Collell, C., Gou, P., Arnau, J., & Comaposada, J. (2011). Non-destructive estimation of moisture, water activity and NaCl at ham surface during resting and drying using NIR spectroscopy. *Food Chemistry, 129*, 601-607.

Collell, C., Gou, P., Picouet, P., Arnau, J., & Comaposada, J. (2010). Feasibility of near-infrared spectroscopy to predict aw and moisture and NaCl contents of fermented pork sausages. *Meat Science, 85*, 325-330.

444 Conzen, J. P. (2006). *Multivariate calibration. A practical guide for developing*
445 *methods in the quantitative analytical chemistry.* Ettlingen, Germany:
446 Bruker Optik GmbH.

447 Coskun, A., Wong, J., Khodadadi, D., Nagi, R., Tey, A., & Ozcan, A. (2017). A
448 personalized food allergen testing platform on a cellphone,. *Lab Chip, 13*,
449 636–640.

450 Das, A. J., Wahi, A., Kothari, I., & Raskar, R. (2016). Ultra-portable, wireless
451 smartphone spectrometer for rapid, non-destructive testing of fruit
452 ripeness. *Scientific Reports, 6*, 32504 .

453 Daszykowski, M., Walczak, B., & Massart, D. L. (2002). Representative subset
454 selection. *Analytica Chimica Acta, 468*, 91-103.

455 Goldring, D., & Sharon, D. (2011). *USA Patent No. WO2013065035A1.*

456 Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical*
457 *Learning.* New York: Springer-Verlag.

458 Horizon. (2016). Retrieved from The Horizon Prize for a food scanner:
459 http://ec.europa.eu/research/horizonprize/index.cfm?prize=food-
460 scanner

461 Levin, S., Krishnan, S., Rajkumar, S., Halery, N., & Balkunde, P. (2016). Monitoring
462 of fluoride in water samples using a smartphone. *Sci. Total Environ., 551–*
463 *552*, 101–107.

464 Liaw, A., & Wiener, M. (2002). Classification and Regression by RandomForest. *R*
465 *News, 2/3*, 18-22.

466 Mevik, B. H., & Wehrens, R. (2007). The pls Package: Principal Component and
467 Partial Least Squares Regression in R. *Journal of Statistical Software, 18*, 1-
468 24.

469 Molinaro, A., Simon, R., & Pfeiffer, R. (2005). Prediction error estimation: a
470 comparison of resampling methods. *Bioinformatics, 21*(15), 3301-3307.

471 Nisbert, R., Elder, J., & Miner, G. (2009). *Handbook of Statistical Analysis and Data*
472 *Mining Applications.* Academic Press.

473 Ortiz-Somovilla, V., Espana-Espana, F., Gaitan-Jurado, A., Perez-Aparicio, J., & De
474 Pedro-Sanz, E. (2007). Proximate analysis of homogenized and minced
475 mass of pork sausages by NIRS. *Food Chemistry, 101*, 1031-1040.

476 Ozcan, A. (2016). Mobile phones democratize and cultivate next-generation
477 imaging, diagnostics and measurement tools. *Lab Chip, 14*, 3187.

478 Pegau, S., Deric, G., & Zaneveld, V. (1997). Absorption and attenuation of visible
479 and near-infrared light in water: dependence on temperature and salinity.
480 *Applied Optics, 36*, 6035-6046.

Prevolnik, M., Skrlep, M., Janes, L., Velikonja-Bolta, S., Skorjanc, D., & Candek-Potokar, M. (2011). Accuracy of near infrared spectroscopy for prediction of chemical composition, salt content and free amino acids in dry-cured ham. *Meat Science, 88*, 299-304.

Prieto, N., Roehe, R., Lavin, P., Batten, G., & Andres, S. (2009). Application of near infrared reflectance spectroscopy to predict meat and meat products quality: A review. *Meat Science, 83*, 175–186.

Puzyn, T., Mostrag-Szlichtyng, A., Gajewicz, A., Skrzyński, M., & Worth, A. (2011). Investigating the influence of data splitting on the predictive ability of QSAR/QSPR models. *Struct Chem, 22*(4), 795-804.

Rateni, G., Dario, P., & Cavallo, F. (2017). Smartphone-Based Food Diagnostic Technologies: A Review. *Sensors, 17*, 1453.

Rinnan, A., van den Berg, F., & Engelsen, S. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *Trends in Analytical Chemistry, 28*, 1201-1222.

Rissanen, A., Saari, H., Rainio, K., Stuns, I., Viherkanto, K., Holmlund, C., . . . Ojanen, H. (2016). MEMS FPI-based smartphone hyperspectral imager. *In Proceedings of the SPIE Next-Generation Spectroscopic Technologies IX* (p. 985507). Baltimore, MD, USA: SPIE.

Segtnan, V., Sasic, S., Isaksson, T., & Ozaki, Y. (2001). Studies on the Structure of Water Using Two-Dimensional Near-Infrared Correlation Spectroscopy and Principal Component Analysis. *Analytical Chemistry, 73*, 3153–3161.

Seo, S., Kim, S., Jeon, J., Kim, J., Kim, H., Cho, J., . . . Paek, S. (2016). Food contamination monitoring via internet of things, exemplified by using pocket-sized immunosensor as terminal unit. *Sens. Actuators B Chem., 233*, 148–156.

The R Core Team. (2018). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Retrieved from R-project: https://www.r-project.org/

Van den Berg, F., Lyndgaard, C., Sørensen, K., & Engelsen, S. (2013). Process analytical technology in the food industry. *Trends Food Sci. Technol., 31*, 27–35.

Wang, Y., Li, Y., Bao, X., Han, J., Xia, J., Tian, X., & Ni, L. (2016). A smartphone-based colorimetric reader coupled with a remote server for rapid on-site catechols analysis. *Talanta, 160*, 194–204.

Wested, F., & Marini, F. (2015). Validation of chemometric models e A tutorial. *Analytica Chimica Acta, 893*, 14-24.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In H. Wold, *In Proceedings s of international symposium on multivariate analysis* (Vol. 2, pp. 391-420). New York: Academic Press.

522 Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of
523 chemometrics. *Chemom. Intell. Lab. Syst., 58*, 109 – 130.

524 Yao, Y., Chen, H., Xie, L., & Rao, X. (2013). Assessing the temperature influence on
525 the soluble solids content of watermelon juice as measured by visible and
526 near-infrared spectroscopy and chemometrics. *Journal of Food*
527 *Engineering, 119*, 22-27.

528 Zhu, H., Sikora, U., & Ozcan, A. (2012). Quantum dot enabled detection of
529 Escherichia coli using a cell-phone. *Analyst, 137*, 2541–2544.

530