Domain invariant covariate selection (Di-CovSel) for selecting generalized features across domains

Valeria Fonseca Diaz, Puneet Mishra, Jean-Michel Roger, Wouter Saeys

Chemometrics and Intelligent Laboratory Systems 222 (2022) 104499

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Domain invariant covariate selection (Di-CovSel) for selecting generalized features across domains

Valeria Fonseca Diaz [a], Puneet Mishra [b,*], Jean-Michel Roger [c,d], Wouter Saeys [a]

[a] KU Leuven Department of Biosystems, Division of Mechatronics, Biostatistics and Sensors, Kasteelpark Arenberg 30, 3001, Leuven, Belgium
[b] Wageningen Food and Biobased Research, Bornse Weilanden 9, P.O. Box 17, 6700AA, Wageningen, the Netherlands
[c] ITAP, Univ Montpellier, INRAE, Institut Agro, Montpellier, France
[d] ChemHouse Research Group, Montpellier, France

ARTICLE INFO

Keywords:
Feature selection
Multivariate
Spectroscopy
Domain adaptation

1. ABSTRACT

Multivariate spectral signals are highly correlated. Often, variable selection techniques are deployed, aiming at model optimization, identification of key variables to explore the underlying physicochemical system or development of a cheap multi-spectral system based on key variables. However, many times the selected variables do not supply a good estimate of properties when tested on a new setting such as new measurements performed on a different spectrometer, different physical or chemical state of the samples and difference in the environmental factors around the experiment. Often the model based on variables selected in the first domain (specific conditions/instrument) does not generalize on the new domain (specific conditions/instrument). To deal with it, in the present work a new method to variable selection called domain invariant covariate selection (di-CovSel) is proposed. The method selects the most informative variables which are invariant to the differences in the instruments, physical or chemical state of the samples and the differences in the environmental factors around the experiment. The method is inspired by domain invariant partial least-square (di-PLS) and the covariate selection (CovSel). The potential of the method is demonstrated on four real cases related to the calibration of near-infrared (NIR) spectroscopy on agri-food materials. The results show that in all the cases, the domain invariant features selected by the di-CovSel have low prediction error compared to the standard variable selection with the CovSel approach when the models are tested on a new data domain. In summary, domain invariant features selected across domains support the development of calibration models with good generalization and supply a better understanding of the system by bypassing the external factors originating from differences in the instruments, physical or chemical states of the samples and the differences in the environmental factors around the experiment. Note that one key feature of the proposed method is that the most important variables which generalize well across domains can be identified without requiring reference measurements in the target domain.

1. Introduction

Optical spectroscopy techniques are widely used in diverse scientific domains for non-destructive and non-contact analysis of material properties [1–3]. Optical spectroscopy can be performed in different parts of the electromagnetic radiation (EMR) spectrum ranging from X-rays [4] to terahertz [5]. Furthermore, it allows to capture information on different EMR interactions related to the chemical components and those related to the microstructure [2]. Application of optical spectroscopy can be found in several areas of research such as fruit and vegetables [6], pharmaceutical manufacturing [7,8], medicine [9], forensics [10], agricultural plants [11,12], veterinary [13] etc.

A key feature of the optical spectroscopy sensing techniques is that all the techniques capture the material responses as multivariate signals which are acquired as responses of the interactions between EMR and the material. For example, in near infrared spectroscopy (NIRS) the attenuation of EMR in the wavelength range from 700 to 2500 nm upon interaction with the sample is quantified at different wavelengths [1]. Multivariate signals acquired with spectral sensing technique are highly collinear [14] and especially the responses of neighboring spectral bands are highly correlated in low energy EMR ranges such as UV, visible (Vis), NIR, MIR and terahertz. As these spectra typically do not contain wavelength variables which are specific for the component of interest, the data must be processed with chemometric or machine learning tools to

* Corresponding author.
E-mail address: puneet.mishra@wur.nl (P. Mishra).

https://doi.org/10.1016/j.chemolab.2022.104499
Received 7 November 2021; Received in revised form 13 January 2022; Accepted 16 January 2022
Available online 19 January 2022

identify the most informative wavelengths and for predictive modelling [14]. While a wide range of methods are used for exploring the spectral signals, the latent space modelling approaches such as principal component analysis (PCA) [15] and partial least-squares (PLS) [16,17] based approaches are widely used when the aim is not only to model the data but also to gain insights in the underlying chemistry. The main asset of the latent space modelling approaches for modelling highly correlated spectral data is their ability to extract the key orthogonal subspaces which are most explanative of the data (for example variation for PCA) or predictive (for example covariation for PLS) for the property of interest [16–18].

While many researchers have reported on the use of NIR spectral sensing and chemometrics for both qualitative and quantitative analysis of samples, most studies were limited to primary calibration of models on a limited number of samples where the NIR spectroscopy works very well [1]. However, due to its sensitivity to the unmodelled physical and chemical disturbances in the systems, obtaining long-term performance with NIR models turns out to be more challenging [19–21]. These disturbances can be due to a range of factors such as a change in temperature of the sensor or the samples compared to the temperature of sensor or samples of the calibration data set [19], changes in the physical form of samples such as from solid to powder form [22], seasonal differences for fresh fruit analysis related applications [21,23,24] and due to changes in the instruments where the model is made on one instrument and required to be used on a new instrument [25]. All these disturbances may lead to failure of the predictive models. Therefore, different methods have been proposed for correcting models for the disturbances, such as external parameter orthogonalization (EPO) [19], dynamic orthogonal projection (DOP) [20,26], domain invariant partial least-squares (di-PLS) [27–29] and transfer component analysis (TCA) [30]. EPO and DOP aim to eliminate the differences which can cause model failure, such as spectral differences originating from temperature variation [31]. Di-PLS, a method inspired by domain adaptation, was recently proposed to achieve domain invariant NIR calibrations that generalize well when used on data acquired in the new domain [27–29]. A key point to note is that the domain adaptation approaches such as di-PLS and transfer component analysis (TCA) [30] require some data from the new domain to achieve the domain adaptation.

Apart from the development of predictive models, it is also of interest to gain more insight into the key region of interest of the multivariate signals which are explanative of the data or the property of interest. The task is usually termed variable selection and particularly for spectral data aims to select key wavelengths of interest [32,33]. In the chemometric scientific literature, a wide range of variable selection methods for spectral data can be found and can be broadly classified as wrapper, embedded, filter and a combination of them called hybrid approaches [33]. Variable selection can be aimed at enhanced understanding of the background chemistry or at developing low cost and fast multi-spectral sensors. However, like full range calibration models, the models and conclusions based on selected variables may not hold true when met with a new batch of data carrying unseen variation. Hence, there is also a need to tune for the variable selection task specific for a batch carrying unseen variation. This might be achieved by adding domain invariance to the variable selection approaches. Covariate selection (CovSel) [34] is a widely used variable selection technique with strong similarity to PLS thanks to the involvement of simple orthogonalization steps. Thanks to this similarity, it could be possible to make it domain invariant based on the principles of di-PLS [27–29]. Therefore, the aim of this study was to develop and test a domain invariant version of the CovSel variable selection method. The development of di-CovSel is inspired by the existing di-PLS and CovSel variable selection approaches. The main hypotheses of the study are as follow:

- Just like the standard PLS model that suffers from unseen variation present in the new domains, the models based on CovSel selected variables will not generalize well to the new domain.
- Adding domain invariance to CovSel will allow to select variables that provide models which generalize better to the new domain of the data.

To test the above two hypotheses, CovSel and di-CovSel were compared on a wide range of NIR data sets. Furthermore, the advantages and disadvantages of the newly developed di-CovSel method are also discussed and recommendations for future use are provided.

2. Theory

In this section, a theoretical background to both di-PLS and CovSel is provided. Next, the di-CovSel method is introduced starting from the concepts of di-PLS and CovSel. All matrices are presented in bold upper case. All vectors are in bold lower case and constants are in lower case. We symbolize the elements to construct di-CovSel as follows: $\boldsymbol{X}_s \in R^{n_s \times k}$ and $\boldsymbol{y} \in R^{n_s}$ rrepresent the calibration data belonging to the source domain. $\boldsymbol{X}_t \in R^{n_t \times k}$ represents the data for adaptation to the target domain. Double sub indexed vectors $\boldsymbol{x}_{sj} \in R^{n_s}$, $\boldsymbol{x}_{tj} \in R^{n_t}$, represent the j-th column of the corresponding matrices which relate to the j-th spectral variable. Super indexed vectors $\boldsymbol{x}_s^i$ and $\boldsymbol{x}_t^i \in R^k$ represent an observation in the source and target domain, respectively. $\boldsymbol{X}_s^T$, $\boldsymbol{y}^T$ and $\boldsymbol{X}_t^T$ are used for the respective transpose elements. One-dimensional arrays represented in bold lower case are assumed to be arranged as columns and therefore their transpose represents a one-dimensional row array. Regular italic low case notation represents scalar values. The matrices are assumed to be centered by columns, with $\boldsymbol{\mu}_{x_s}$, μ_y and $\boldsymbol{\mu}_{x_t}$, representing the means.

2.1. Domain invariant PLS analysis

The underlying idea to construct a domain invariant model based on PLS relies on the association of $\boldsymbol{X}$ and $\boldsymbol{y}$ by a bilinear model constrained to an invariability of the projected domains [35]. The objective function considering the domain invariability for PLS [35] is given by Eq. (1)

$$\min_{w} \left\| \boldsymbol{X} - \boldsymbol{y}\boldsymbol{w}^T \right\|^2, \; subject\ to\ |var(\boldsymbol{t}_s) - vat(\boldsymbol{t}_t)| = 0 \tag{1}$$

where $\boldsymbol{w} \in R^k$ corresponds to the weight vector of the bilinear model and $\boldsymbol{t}_s$ and $\boldsymbol{t}_t$ represent the scores, that is, the projection of matrices $\boldsymbol{X}_s$ and $\boldsymbol{X}_t$ onto $\mathbf{w}$. This optimization criterion is redefined in di-PLS as Eq. (2)

$$\min_{w} \left\| \boldsymbol{X} - \boldsymbol{y}\boldsymbol{w}^T \right\|^2 + \lambda(\boldsymbol{w}^T \boldsymbol{\Delta} \boldsymbol{w}) \tag{2}$$

where $\boldsymbol{w}^T\boldsymbol{\Delta}\boldsymbol{w}$ is an upper bound of $|var(\boldsymbol{t}_s) - var(\boldsymbol{t}_t)|$, λ becomes a domain regularization parameter and matrix $\boldsymbol{\Delta}$ is defined by the eigenvectors and the absolute value of the eigenvalues of $\boldsymbol{D} = cov(\boldsymbol{X}_s) - cov(\boldsymbol{X}_t)$. Following an iterative procedure as in classical PLS, the complete di-PLS model is defined by a latent variables. At each step, $\boldsymbol{w}$ is calculated based on Eq. (2). Latent variable scores $\boldsymbol{t}_s = \boldsymbol{X}_s\boldsymbol{w}$ and $\boldsymbol{t}_t = \boldsymbol{X}_t\boldsymbol{w}$ are calculated followed by loading vectors $\boldsymbol{p}_s^T = \boldsymbol{t}_s^{-1}\boldsymbol{t}_s^T\boldsymbol{X}_s$, $\boldsymbol{p}_t^T = \boldsymbol{t}_t^{-1}\boldsymbol{t}_t^T\boldsymbol{X}_t$ and regression coefficient $c = (\boldsymbol{t}_s^T\boldsymbol{t}_s)^{-1}\boldsymbol{t}_s^T\boldsymbol{y}$. Such scores, loadings and regression coefficient are finally used to deflate the elements as $\boldsymbol{X}_s = \boldsymbol{X}_s - \boldsymbol{t}_s\boldsymbol{p}_s^T$, $\boldsymbol{y} = \boldsymbol{y} - c\boldsymbol{t}_s$ and $\boldsymbol{X}_t = \boldsymbol{X}_t - \boldsymbol{t}_t\boldsymbol{p}_t^T$.

With the deflated elements, a new iteration starts until the target number of a latent variables have been calculated. Using the corresponding obtained matrices $\boldsymbol{W}$ and $\boldsymbol{P}$ and regression coefficients $\boldsymbol{c}$, the net analyte signal vector for the linear regression and intercept are calculated as Eqs. (3) and (4)

$$\boldsymbol{b} = \boldsymbol{W}(\boldsymbol{P}^T\boldsymbol{W})^{-1}\boldsymbol{c} \tag{3}$$

$$b_o = \mu_y - \mu_{x_t}^T\boldsymbol{b} \tag{4}$$

In this way, di-PLS delivers the estimation of a linear regression model which can be subsequently used for prediction in the target

Algorithm for di-CovSel

Input: source domain matrices $\boldsymbol{X}_s \in R^{n_s \times k}$, $\boldsymbol{y} \in R^{n_s}$ and target domain $\boldsymbol{X}_t \in R^{n_t \times k}$ centered, number of variables to select m and the value of λ for the domain invariant regularization parameter.

Output: Selected variables and order of selection $selvars := [\,]$

$for\ l = 1\ to\ m$ do:

$$\mathbf{D} = \mathrm{cov}(\boldsymbol{X}_s) - \mathrm{cov}(\boldsymbol{X}_t) = \boldsymbol{V} diag(\sigma_1, \sigma_2, \ldots, \sigma_k)\boldsymbol{V}^T$$
$$\Delta = \boldsymbol{V} diag(|\sigma_1|, |\sigma_2|, \ldots |\sigma_k|)\boldsymbol{V}^T$$
$$\boldsymbol{A}_\lambda = (\boldsymbol{y}^T\boldsymbol{y})\boldsymbol{I} + \lambda\Delta)^{-1}$$
$$j = ArgMax\big(diag(\boldsymbol{A}_\lambda \boldsymbol{X}_s^T \boldsymbol{y}\boldsymbol{y}^T \boldsymbol{X}_s \boldsymbol{A}_\lambda)\big); \qquad \boldsymbol{x}_{s(1)} = \boldsymbol{x}_{sj}$$
$$selvars := [selvars, j]$$
$$\boldsymbol{P}_s = \boldsymbol{x}_{s(l)}(\boldsymbol{x}_{s(l)}^T \boldsymbol{x}_{s(l)})^{-1}\boldsymbol{x}_{s(l)}^T$$
$$\boldsymbol{P}_t = \boldsymbol{x}_{t(l)}(\boldsymbol{x}_{t(l)}^T \boldsymbol{x}_{t(l)})^{-1}\boldsymbol{x}_{t(l)}^T$$
$$\boldsymbol{X}_s = \boldsymbol{X}_s - \boldsymbol{P}_s\boldsymbol{X}_s$$
$$\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{P}_s\boldsymbol{y}$$
$$\boldsymbol{X}_t = \boldsymbol{X}_t - \boldsymbol{P}_t\boldsymbol{X}_t$$

end

domain using Eq. (5)

$$\hat{y} = \boldsymbol{x}_t^{newT}\boldsymbol{b} + b_o \tag{5}$$

2.2. Covariate selection approach

Variable selection by covariate selection (CovSel) relies on the maximization of the covariance between $\boldsymbol{X}$ and $\boldsymbol{y}$, analogous to the definition of the latent variables in the case of PLS. We use the source domain to refer to CovSel. The first variable selected by CovSel i.e., $\boldsymbol{x}_{s(1)}$ corresponds to the j-th variable i.e., $\boldsymbol{x}_{sj}$ which maximizes the criterion given by Eq. (6)

$$cov(\boldsymbol{x}_{sj}, \boldsymbol{y})^2 = \boldsymbol{x}_{sj}^T \boldsymbol{y}\boldsymbol{y}^T \boldsymbol{x}_{sj} \tag{6}$$

This criterion is equivalent to finding $\boldsymbol{x}_{sj}$ such that,

$$j = ArgMax\big(diag\big(\boldsymbol{X}_s^T \boldsymbol{y}\boldsymbol{y}^T \boldsymbol{X}_s\big)\big); \quad \boldsymbol{x}_{s(1)} = \boldsymbol{x}_{sj}$$

Once $\boldsymbol{x}_{s(1)}$ has been selected, a deflation step takes place to orthogonalize the matrices by the selected variable. To do so, the projector $\boldsymbol{P}$ is calculated as Eq. (7)

$$\boldsymbol{P}_1 = \boldsymbol{x}_{s(1)}\left(\boldsymbol{x}_{s(1)}^T \boldsymbol{x}_{s(1)}\right)^{-1}\boldsymbol{x}_{s(1)}^T \tag{7}$$

and matrices are deflated as $\boldsymbol{X}_s = \boldsymbol{X}_s - \boldsymbol{P}_1\boldsymbol{X}_s$ and $\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{P}_1\boldsymbol{y}$. The process is repeated to select $\boldsymbol{x}_{s(1)}, \boldsymbol{x}_{s(2)}, \ldots, \boldsymbol{x}_{s(k)}$.

2.3. Domain invariant covariate selection

To extend CovSel to select variables that are invariant across domains, the principle used by di-PLS is inherited to develop domain invariant covariate selection (di-CovSel). This method is based on calibration data from the source domain $\boldsymbol{X}_s$ and $\boldsymbol{y}$, and a separate set for domain adaptation $\boldsymbol{X}_t$. Note that the use of unsupervised samples in the target domain builds on the assumption that the marginal distributions of the domains are different ($P(\boldsymbol{X}_s) \neq P(\boldsymbol{X}_t)$), while the conditional distribution that defines the relationship between $\boldsymbol{X}$ and $\boldsymbol{y}$ remains unchanged [27,35]. The procedure in di-CovSel consists of three steps: (*i*) calculate domain invariant matrix Δ, (*ii*) select the most informative variable according to CovSel and di-PLS criteria combined, (*iii*) deflate source and target domain matrices by their indicative selected variable. The initial step is executed by calculating matrix Δ as in di-PLS making use of $\boldsymbol{X}_s$ and $\boldsymbol{X}_t$. For the second step, the criterion for variable selection takes the same criterion from CovSel [34] inheriting the solution of the optimization criterion of di-PLS as in Eq. (2). The solution to the latter corresponds to the vector calculated with Eq. (8)

$$\boldsymbol{w}^T = \boldsymbol{y}^T \boldsymbol{X}_s((\boldsymbol{y}^T\boldsymbol{y})\boldsymbol{I} + \lambda\Delta)^{-1} \tag{8}$$

where $\boldsymbol{I}$ is the identity matrix of the adequate order. This is equivalent to defining the solution as $\boldsymbol{w} = \boldsymbol{A}_\lambda \boldsymbol{X}_s^T \boldsymbol{y}$, where $\boldsymbol{A}_\lambda$ is a symmetric matrix defining the rightest term becoming a weighting matrix for the columns of $\boldsymbol{X}_s$, that is, a weighting for the variables. Therefore, to select the most informative domain invariant variable j, $\boldsymbol{X}_s$ is substituted by the weighted matrix $\boldsymbol{X}_s\boldsymbol{A}_\lambda$ in the criterion of CovSel, rendering the criterion for di-CovSel as Eq. (9)

$$j = ArgMax\big(diag\big(\boldsymbol{A}_\lambda \boldsymbol{X}_s^T \boldsymbol{y}\boldsymbol{y}^T \boldsymbol{X}_s \boldsymbol{A}_\lambda\big)\big); \quad \boldsymbol{x}_{s(1)} = \boldsymbol{x}_{sj} \tag{9}$$

Once j has been found, projectors for the source and target domain are calculated as Eqs. (10) and (11)

$$\boldsymbol{P}_{s1} = \boldsymbol{x}_{s(1)}\left(\boldsymbol{x}_{s(1)}^T \boldsymbol{x}_{s(1)}\right)^{-1}\boldsymbol{x}_{s(1)}^T \tag{10}$$

$$\boldsymbol{P}_{t1} = \boldsymbol{x}_{t(1)}\left(\boldsymbol{x}_{t(1)}^T \boldsymbol{x}_{t(1)}\right)^{-1}\boldsymbol{x}_{t(1)}^T \tag{11}$$

These projectors are finally used to carry over the deflation step as $\boldsymbol{X}_s = \boldsymbol{X}_s - \boldsymbol{P}_{s1}\boldsymbol{X}_s$, $\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{P}_{s1}\boldsymbol{y}$, and $\boldsymbol{X}_t = \boldsymbol{X}_t - \boldsymbol{P}_{t1}\boldsymbol{X}_t$. Thereafter, the process is repeated for $[\boldsymbol{x}_{s(2)}, \boldsymbol{x}_{t(2)}]$, $[\boldsymbol{x}_{s(3)}, \boldsymbol{x}_{t(3)}]$, …, $[\boldsymbol{x}_{s(k)}, \boldsymbol{x}_{t(k)}]$. Note that the

Table 1
A summary of samples and reference properties in calibration, domain adaption and test set for the 4 cases.

Datasets (reference property)	Wavelength range (nm)	Samples in source calibration set/ reference property range (mean ± std)	Samples in source test set/ reference property range (mean ± std)	Samples for adaption in target set/reference property range (mean ± std)	Samples for test in target set/ reference property range (mean ± std)
Rice (Protein %)	950–1200	140/(9.07 ± 1.46)	60/(9.02 ± 1.57)	60/(9.19 ± 1.48)	140/(8.99 ± 1.50)
Wheat (Nitrogen %)	397–2497	380/(4.15 ± 1.29)	163/(4.37 ± 1.32)	162/(4.31 ± 1.36)	381/(4.18 ± 1.27)
Pear (Moisture %)	720–997	167/(85.79 ± 1.42)	72/(85.79 ± 1.22)	69/(85.59 ± 1.02)	161/(85.53 ± 1.02)
Mango (Moisture %)	684–990	2584/(83.63 ± 2.52)	1108/(83.51 ± 2.58)	150/(85.33 ± 1.90)	351/(85.47 ± 1.97)

term in Eq. (9) corresponds to $j = ArgMax(\boldsymbol{w}^2)$, where $\boldsymbol{w}$ is the domain invariant version of the covariance direction $\boldsymbol{X}_s^T\boldsymbol{y}$ which is also squared in CovSel as shown in Eq. (6). The di-CovSel depends on the domain invariant regularization parameter λ as it is the case in di-PLS [35]. The role of this parameter in di-CovSel becomes the trade-off between selecting variables that contain purely the highest covariance with $\boldsymbol{y}$ in the source domain and variables that are the most invariant between the domains. In this regard, for $\lambda \to 0$, $\boldsymbol{A}_\lambda \to (\boldsymbol{y}^T\boldsymbol{y})\boldsymbol{I}^{-1}$ and the di-CovSel solution converges to the CovSel [34] solution.

Algorithm. for di-CovSel

3. Datasets

3.1. Different physical forms of rice samples

The rice data set was used to show the capability of the di-CovSel model to select the variables that allow a model to be used in a different physical form of samples. The source was the solid rice kernels, while the target was rice powder. This data set was first published in Ref. [36] and later used in Ref. [22] to show the application of di-PLS for domain adaption of the model made on rice kernels to use it on rice powder. The data set consists of NIR spectra measured on 200 individual rice kernels followed by spectral measurement on the ground rice kernels in powder form. The spectral measurements were performed with a FT-NIR spectrometer (MPA, Bruker, Germany). The reference property was the protein content which according to the primary study was measured using the Dumas combustion method [36].

3.2. Wheat data set for model adaption from point to spectral camera for digital phenotyping

The wheat data set was used to show the capability of di-CovSel to select variables that lead to generalized models between different modes of Vis-NIR spectroscopy. The source was the point spectrometer, while the target was the spectral camera. The data set is the same as published in Ref. [37] and later used in Ref. [38] to show the capability of calibration transfer to transfer models from point spectrometers to spectral cameras. It consists of spectral and reference nitrogen content measurements performed on 200 plants at 3 different time points. The spectral measurements were performed in the Plant Accelerator facility at the University of Adelaide, Australia. To obtain wide nitrogen variation in plants, the plants were treated with four treatments of 25, 50, 100 and 200 mg N/kg. The point spectral measurements were performed with a combination of a diode array with a monochromator in the 400–2500 nm range (FieldSpec 3, Analytic Spectral Devices, Boulder, USA) with a leaf clip, and the spectral imaging was performed using a high-throughput spectral imaging setup (WIWAM, Ghent, Belgium) with Vis-NIR and SWIR spectral cameras from Specim, Finland. The data set contains 600 spectral measurements for each spectral sensor and 600 reference values for the N content (%). Based on the earlier study, 57 measurements were outliers and removed as suggested in Ref. [37]. Finally, the data set had 543 spectral and 543 reference nitrogen measurements.

3.3. Moisture prediction in pear fruit of different seasons

The pear data set was used to show the capability of di-CovSel to adapt models based on different seasons for fruit moisture content (MC) prediction. Currently, the season variability is one of the main challenges in NIR spectroscopy of fruit which leads to model failure. The pear data set consists of pear 'Conference' fruit harvest of two seasons 2019 (season 1) and 2020 (season 2), measured with a portable fruit spectrometer (Felix F-750, Camas, WA, USA). This data set is the same as used in Ref. [39]. Season 1 data consists of spectra and reference MC measurements on 239 pear fruits, while the season 2 data set consists of 230 fruits. Spectral measurements were performed at the center belly part of the fruit. For reference MC measurements, a 1 cm thick slice was cut from the fruit equator and divided into four equal parts. One of these parts without peel was used to determine MC by recording the weight of the part before and after drying in a hot-air oven (FP 720, Binder GmbH, Tuttlingen, Germany) at 80 °C for 96 h. Details of the reference MC range can be found in Table 1.

3.4. Model transfer between point spectrometers to predict moisture content in mango fruit

The mango fruit data set was used to demonstrate di-CovSel for using free access spectral datasets for local use in new experiments. Basically, it is a case of standard free calibration transfer where the free access dataset measured with a similar instrument was generalized to a new local instrument. The free access data set used in this study is the mango data set [40–42] for which both spectra and reference moisture content values are available. Please note that the mango data set is based on mango samples measured in Australia. The local mango data set measured with a similar but different instrument was the same as used in Ref. [43] and consisted of spectral and moisture measurements performed on 501 mango fruit of 'Keitt' and 'Kent' cultivar performed at Wageningen University & Research, The Netherlands. The instrument model used for the acquisition of the open-access and local data set was a portable fruit spectrometer (Felix F-750, Camas, WA, USA). For MC measurements, the peel of the part was removed and later the fresh weight of the fruit's flesh was measured using an electronic balance and dried in a hot-air oven (FP 720, Binder GmbH, Tuttlingen, Germany) at 80 °C for 96 h. After drying, the dried fruit weight was measured, and the MC was estimated and expressed in %. The spectral range used for modelling was the same (684–990 nm) as recommended in earlier publications [40,42,43] on the open-access mango data set. Please note that the original free access mango data set has more than 10k measurements. However, in this study we only used measurements from the year 2016. Since we used data from season 2016 and the new data set measured in a local setting based on fruit from season 2020, the mango case also forms a basis for model adaptation related to seasonal differences. More details on the sample number and reference property range can be found in Table 1.

4. Data analysis

The datasets of the source domain were randomly divided into calibration and test sets with a 70-30 rule. A detail on total samples and

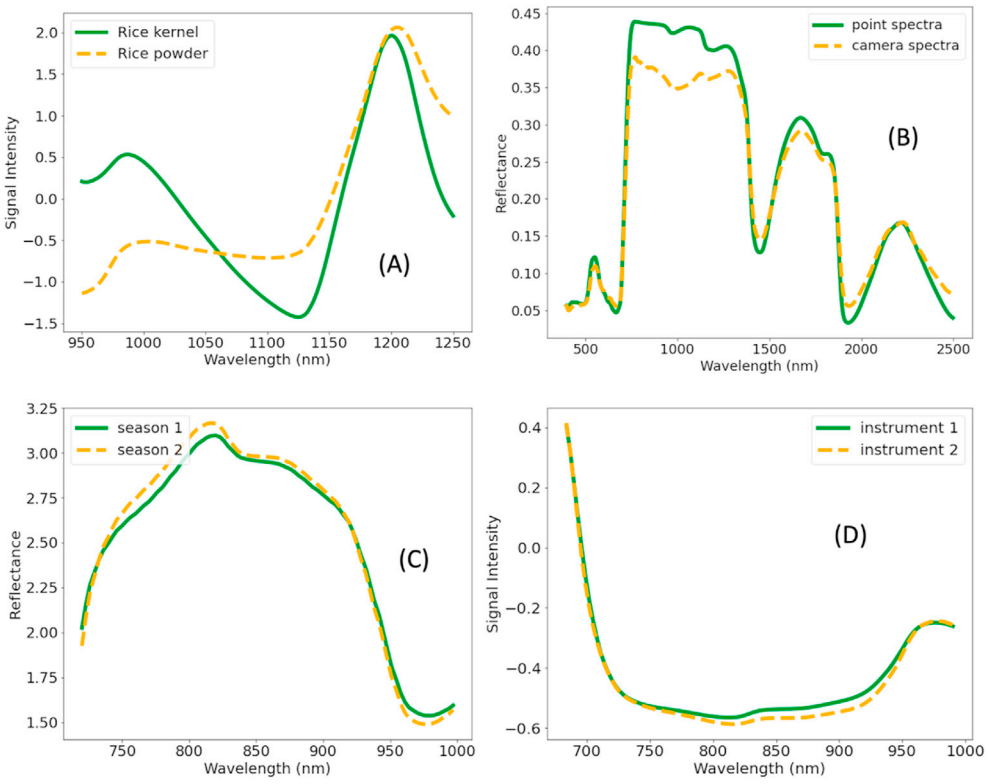

Fig. 1. Mean spectral differences between the different domains considered in the 4 cases: (A) rice, (B) wheat plants, (C) pear fruit, and (D) mango fruit.

before and after the partition is provided in Table 1. The datasets from the target domain were randomly divided into target and test targets with a 30–70 rule. The minority was chosen for the sake of using as few samples from the target domain as possible during domain invariant variable selection. Calibration models were trained using ordinary least squares (OLS) with the selected variables based on the calibration data of the source domain and tested on the source and target domains for CovSel and di-CovSel.

To account for a suitable bias term in the OLS models to predict the target domain, recentering of the source domain was performed. For this, 10 samples were selected from the adaptation target data with a Kennard Stone [44] algorithm. These samples were used to recenter only the source matrices $\boldsymbol{X}_s$ in the case of mango, rice, and wheat. On the other hand, as indicated in section 3.3, the final di-CovSel using unsupervised samples and subsequent OLS models remains valid under the assumption that only the marginal distributions of $\boldsymbol{X}_s$ and $\boldsymbol{X}_t$ need to be adjusted. For the current application of moisture prediction in pear, it was detected from the predictions that a bias correction was necessary also for $\boldsymbol{y}$ suggesting that a shift occurred in the relationship between $\boldsymbol{X}$ and $\boldsymbol{y}$. Therefore, in this case, the $\boldsymbol{y}$ values were also recentered using the 10 selected samples.

The number of variables was tuned from 1 up to 30 and the domain invariant parameter in di-CovSel was tuned for the values $\lambda = [1, 10^1, 10^2, 10^3, 10^4, 10^5, 10^6, 10^7, 10^8]$ taking reference from the tuning reported in applications with di-PLS [27,35]. For the case of wheat, intermediate values $\lambda = [5.10^4, 5.10^5, 5.10^6, 5.10^7]$ were also tuned as large changes were detected in the cross-validation performance. The optimal number of variables and optimal value of λ were selected based on a 10-fold cross-validation. The error was quantified in terms of the root mean squared error in cross-validation (RMSECV) and prediction on the test set (RMSEP). The squared coefficient of correlation r^2 was also calculated in cross-validation and on the test set. Furthermore, bias, and standard errors were also estimated to access the capability of di-CovSel to achieve lower bias predictive models. All analyses were performed in Python language 3.8 using in-house codes with functions from the SciKit Learn Library version 0.24 (https://scikit-learn.org/stable/about.html).

5. Results

The spectra for the different cases used to demonstrate the performance of di-CovSel and CovSel are illustrated in Fig. 1. The mean rice kernel spectra and rice powder spectra are shown in Fig. 1A, where the differences in the spectra for different rice physical forms can be noted. At first, three main peaks and valleys can be noted at 960 nm, 1140 nm, and 1200 nm. These peaks can be related to OH, RNH_2 and CH bond overtones, which are abundant in molecules such as starch, cellulose, and protein present in rice [45]. However, apart from the peaks, the main thing to note was that the key difference between the rice physical forms was mainly related to the absorption bands at 960 nm and 1140 nm, which could show that the rice grinding brought changes in the spectral zones related to OH and NH bonds compared to the CH bonds. This might be attributed to differences in the light scattering as the powder particles scatter more light compared to the solid kernels [22]. The wheat data set constitutes a pure case of instrument differences, where the spectra measured with the point spectrometer have higher spectral intensity in the NIR range compared to those acquired with the spectral camera (Fig. 1B). The smaller absorption bands also seem to have been smoothed out due to the lower spectral resolution in the spectral camera. For the pear data set, differences in the mean spectra of fruit from the different seasons can be noted around 800 nm and 960 nm (Fig. 1C). These can be attributed to overtones of OH molecules present in abundance in fresh fruit. The mango data set (Fig. 1D) is a mixed case involving both instrument and seasonal differences, where the main differences can be noted in the spectral range from 800 to 950 nm, again a domain corresponding to the overtones of OH and CH bonds present in abundance in macromolecules present in fresh fruit. Please note that in real-life situations there is usually very little information available about the difference between scenarios, for example, in the case of seasonal differences, the underlying cause is hardly known due to the high biological

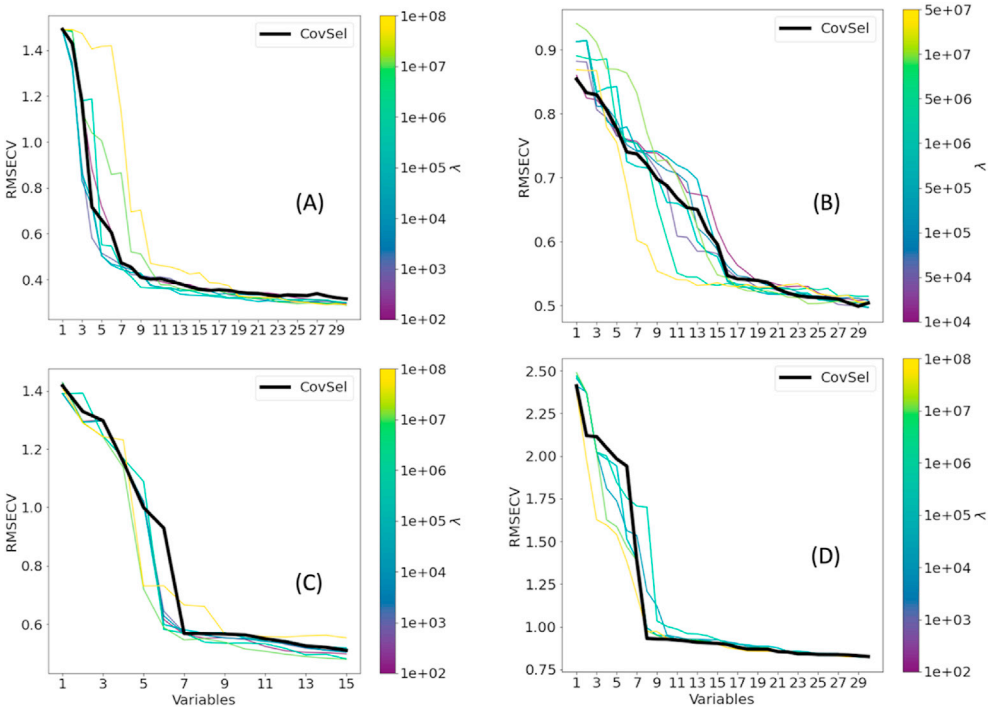

Fig. 2. 10-fold cross-validation plots to select optimal number of variables for CovSel and di-CovSel for the 4 cases: (A) rice, (B) plants, (C) pear fruit, and (D) mango fruit.

variability. Hence, a major focus of the scientific community working on fruit spectroscopy is to bypass the difference [21–23], which is also explored with the di-CovSel approach proposed in this study. The properties of the different data sets are summarized in Table 1. The reference property ranges for the target domain for all data sets were in the range of the target property ranges of the source samples. Also, note that the di-CovSel method was applied in this study using as low as 60 samples for the rice data set to up to a maximum of 150 samples for the mango data set. Sample sizes of 60–150 for model adaptation without the need for new reference property analyses can be considered as a profitable situation in practice, as typically the reference analysis is the main limiting factor for model adaptation with traditional chemometric approaches [25].

The optimizations of the CovSel and the di-CovSel for all four data sets are illustrated in Fig. 2. It should be noted that for di-CovSel there is an extra λ parameter which needs to be tuned and in Fig. 2 corresponds to the distinct color error curves. For the rice data set (Figs. 2A), 10 variables were selected for CovSel and di-CovSel based on the evolution of the cross-validation error as a function of the number of selected variables. The domain invariant regularization parameter was set to 10^6 as the RMSECV values started to increase for larger values. With the same number of selected variables for CovSel and di-CovSel, the OLS model made on selected variables to predict protein content in rice performed differently. Particularly, the model made on variables selected with CovSel on the rice kernel data set (Fig. 3A) performed poor (Fig. 3B) when tested on the rice flour data set. On the other hand, the model made on di-CovSel selected variables achieved lower prediction error (Fig. 3C) compared to CovSel tested on rice flour samples. Achieving up to 65% lower RMSEP than the CovSel approach, the di-CovSel demonstrated the importance of selecting domain invariant variables for generalized variable selection across the different physical forms of samples.

For the wheat data set, the cross-validation results for CovSel and di-CovSel (Fig. 2B) suggested that the optimal number of variables by CovSel was 16, which was similar for values of $\lambda < 5.10^6$ in di-CovSel. For larger values of the domain invariant regularizer, the optimal number of variables became more clearly ~11. As the aim was to compare di-CovSel and CovSel for selecting variables that generalize well across different modalities of NIR spectroscopy i.e., point spectroscopy and spectral imaging, the models were created with the same number of selected variables for CovSel and di-CovSel. The OLS model made with CovSel selected variables on point spectroscopy data (Fig. 3D) performed poor when used on the spectral imaging data (Fig. 3E). The prediction performance showed a large dispersion together with a clear problem of bias. However, with di-CovSel the RMSEP was decreased from 1.29% to 0.90%, showing the bias problem (Fig. 3E and F) was fully corrected and a lower dispersion in the predictions was obtained. This result suggests the ability of di-CovSel to select better generalizing variables when the calibration model is to perform across different modalities of NIR spectroscopy, especially for the correction of bias in the predictions.

For the pear data set, the optimal number of variables appears to be 7 for CovSel and for di-CovSel with values of $\lambda < 10^8$. The effect of the λ value on the optimal number of variables can be noted as some smaller values of this parameter suggested to use 6 variables. The resulting performance of CovSel is illustrated in Fig. 3H and in Fig. 3I for di-CovSel with the same number of variables i.e., 7 and setting $\lambda = 10^7$. The OLS model made with equal number of variables showed that di-CovSel achieved a lower RMSEP of 0.51% compared to the RMSEP of 0.55% for CovSel.

For the mango data set, CovSel suggested 8 variables while for di-CovSel some values of λ suggested up to 11 variables. Please note that the mango data set was a complex case involving both instrument and seasonal differences, where the two batches correspond to mango samples of different year harvests measured with two different instruments. The model made with CovSel selected variables resulted in higher RMSEP value (Fig. 3K) compared to di-CovSel selected variables (Fig. 3L) when the model was tested on a new season/instrument data.

In summary, the better predictive performances of the OLS models made with di-CovSel selected variables compared to CovSel selected variables shows the potential of selecting domain invariant variables to handle common model adaptation tasks such as model adaptation for physical forms of samples, instruments, and seasonal effects.

The performance of the final OLS models built with the variables selected by CovSel and di-CovSel is summarized in Table 2 and Table 3, respectively. For all four cases, the performance obtained in the target

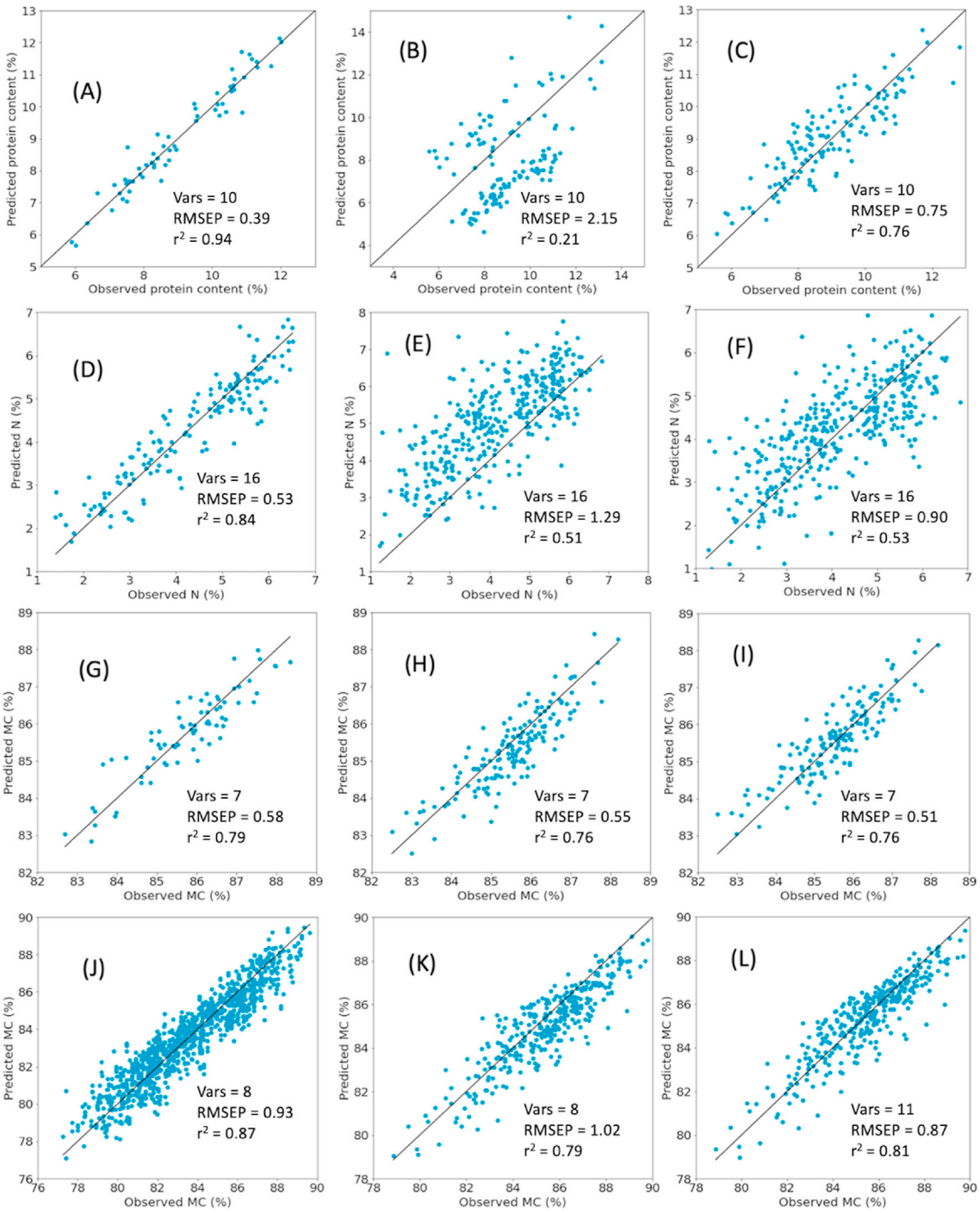

Fig. 3. Summary of OLS models built on variables selected with CovSel and di-CovSel for the rice case (A, B, C), the wheat plants (D, E, F), the pear fruit (G, H, I) and the mango fruit (J, K, L); with results for CovSel built and tested on the source test set (left), CovSel tested on the target test set (middle) and di-CovSel tested on the target test set (right).

Table 2
A summary of performance of CovSel on the source and target domain test sets.

Data sets	Variables	$Bias_{cv}$	SECV	$Bias_p$ (Source)	SEP (Source)	$Bias_p$ (Target)	SEP (Target)
Rice	10	0.01	0.40	−0.03	0.39	1.12	1.84
Plant	16	0.00	0.55	0.03	0.53	−0.89	0.94
Pear	7	0.00	0.57	−0.03	0.58	0.19	0.52
Mango	8	0.00	0.93	−0.01	0.93	0.44	0.92

domain with di-CovSel was better than with CovSel in terms of lower SEP (Tables 2 and 3). One of the key benefits of the di-CovSel was the reduction of the bias for all the cases. The bias reduction was more dominant for the Rice and Plant cases, where the model obtained based

Table 3
A summary of performance of di-CovSel on the source and target domain test sets.

Data sets	Variables	$Bias_{cv}$	SECV	$Bias_p$ (Source)	SEP (Source)	$Bias_p$ (Target)	SEP (Target)
Rice	10	0.00	0.36	0.00	0.35	−0.09	0.74
Plant	16	0.00	0.53	0.04	0.50	−0.04	0.89
Pear	7	0.00	0.55	0.02	0.56	−0.06	0.51
Mango	8	0.00	0.93	−0.01	0.96	0.11	0.87

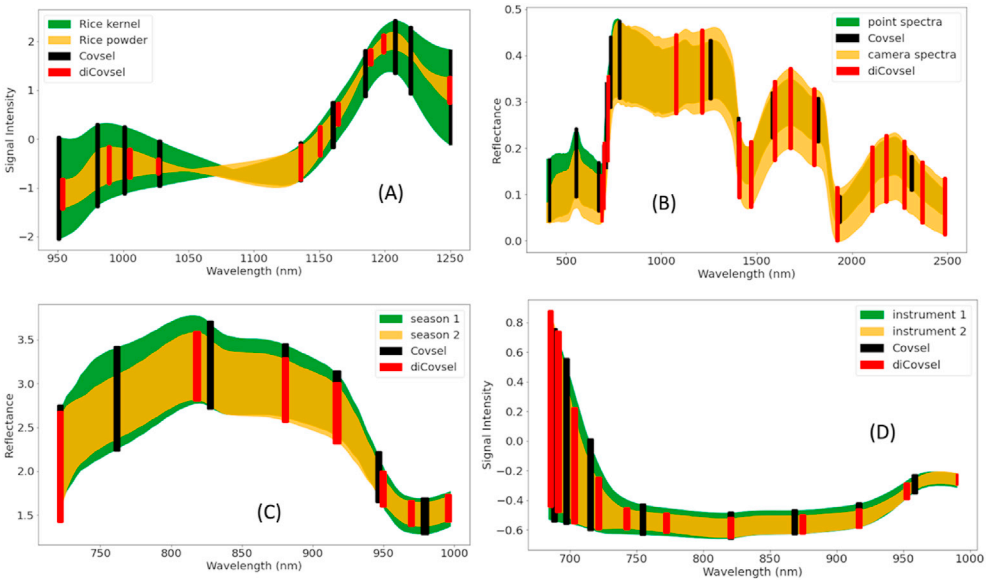

Fig. 4. A summary of variables selected with CovSel and di-Covsel. (A) Rice, (B) wheat plants, (C) pear fruit, and (D) mango fruit. The variables selected by CovSel and di-CovSel are respectively highlighted as black and red vertical lines. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

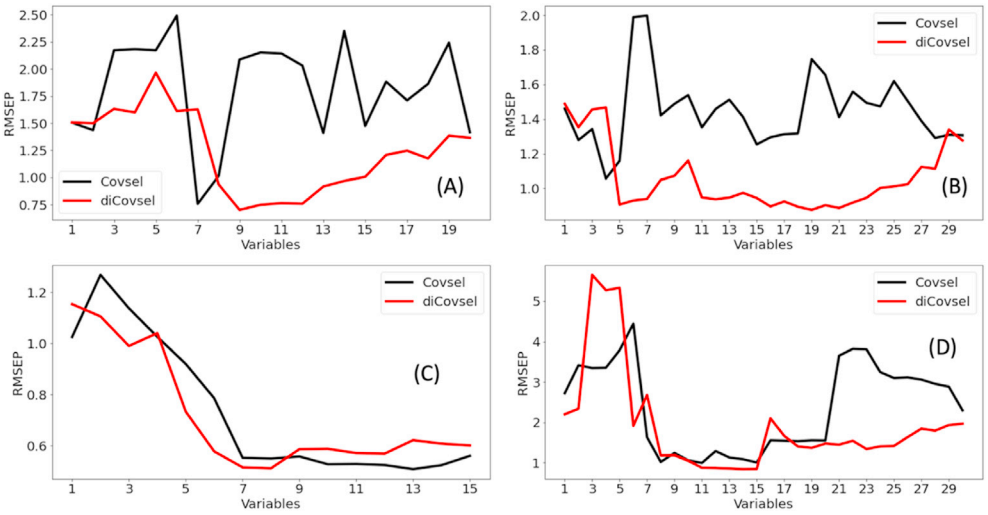

Fig. 5. Posterior analysis of the evolution of prediction performance on the test set (RMSEP) as a function of the number of selected variables for CovSel and di-CovSel for the 4 cases: (A) rice, (B) wheat plants, (C) pear fruit, and (D) mango fruit.

on the CovSel variables achieved high bias (Tables 2 and 3). However, it should be noted that for all cases the performance in the target domain was poorer than the performance on the source domain. This finding related to di-CovSel aligns with the conclusions of di-PLS, which states that the domain invariance does not guarantee similar performances across domains, but better performance across domains compared to models that rely on monodomain (source domain only) information. Complementary, the performances on the source domain in cross-validation and on the test set indicate that the models based on di-CovSel selected variables maintained their performance on the source domain. This indicates that the variables selected by di-CovSel were considerably less sensitive to the domain variance, and the models based on them can be used both in the source and target domain. We also compared the performance of standard PLS calibrations based on full spectral data with the OLS models based on CovSel and di-CovSel selected variables. We found that OLS models based on CovSel and di-CovSel selected variables worked either similar or slightly better than the PLS modelling on full spectra of source calibration (Supplementary Table 1) and source test set (Supplementary Table 2), rendering a lower bias and prediciton error. However, we found out that testing the OLS

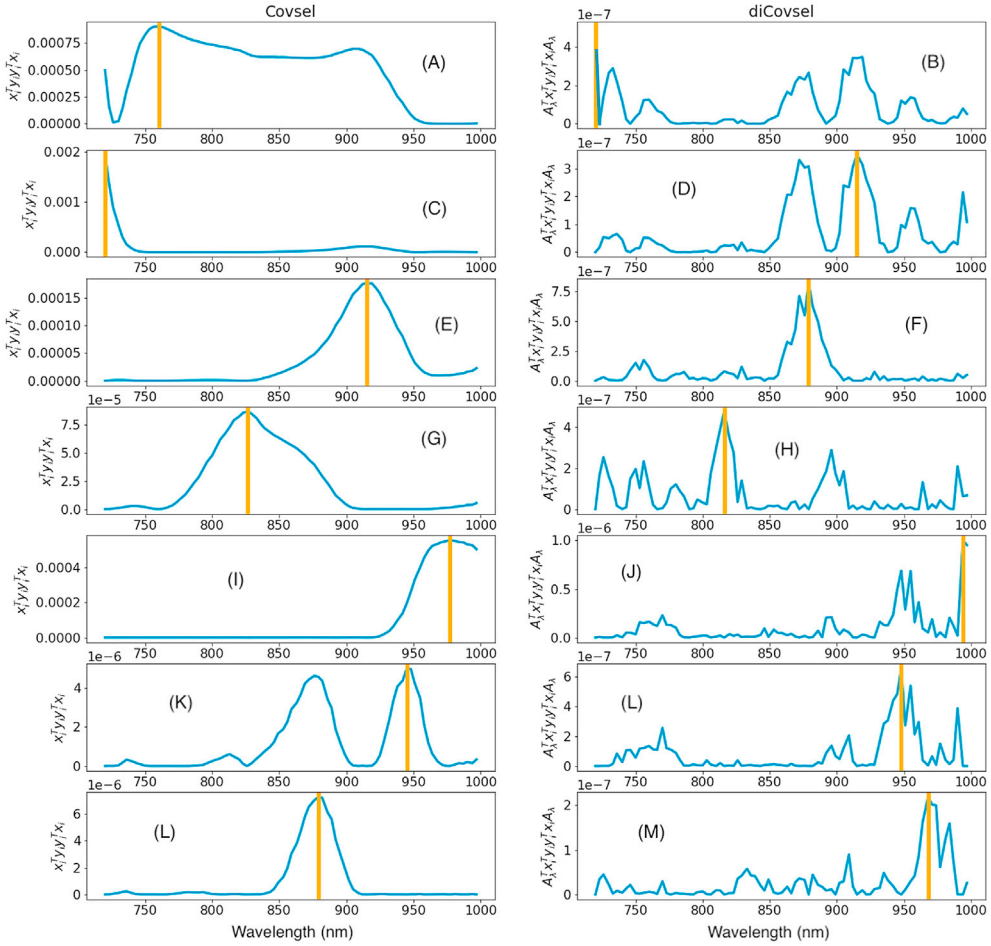

Fig. 6. Illustration of the variable selection steps (top to bottom) for CovSel (left column) and di-CovSel (right column) for the pear fruit data set. In total, 7 variables were selected by both methods.

model based on CovSel selected variables on the target test set showed either similar perfomance to PLS based on full spectra or poorer performance (Supplementary Table 3). The OLS model based on di-CovSel selected variables on the target test set always led to lower prediction bias and for three out of four cases lower prediction error than the PLS analysis based on full spectra (Supplementary Table 3).

The selected variables for CovSel and di-CovSel are illustrated in Fig. 4 for the different cases. For the rice data set, the selected variables by both methods are shown in Fig. 4A where the spectra are plotted as recentered to the mean of the target domain. In Fig. 4A, it can be noted that in the range 950–1050 nm, di-CovSel led to the selection of variables slightly shifted from variables selected by CovSel. Such a shift could show that the rice grinding causing light scattering may have deformed the spectra. In the range from 1150 to 1250 nm, the di-CovSel selected several new variables which could be related to overtones of CH bonds in several macromolecules present in rice kernels [36]. For the wheat data set (case of model adaptation from point spectrometer to spectral camera), the recentered spectra with the selected variables for CovSel and di-CovSel are shown in Fig. 4B. While the mean shift (Fig. 4B) around 1000 nm was corrected with recentering, a clear difference in the domains can be seen around 500 nm. This is in line with the variable selection by di-CovSel where the spectral range around 500 nm was not included in the final selection by di-CovSel but was included by CovSel leading to inferior performance of the OLS model made on CovSel selected variables. For the pear and mango data sets, the variables selected by di-CovSel were slightly shifted compared to the CovSel selected variables.

From the above, it can be concluded that the OLS models based on di-CovSel selected variables led to better predictive performance compared to models based on CovSel selected variables for all the four data sets. To explore the performance of di-CovSel and CovSel as a function of selected variables, a posterior analysis was performed where the RMSEP was estimated on the test set from the target domain as a function of the increasing number of selected variables. The test set was the same as used to evaluate the performance of the CovSel and di-CovSel. The posterior analysis for all four data sets is shown in Fig. 5. For the rice data set (Fig. 5A), the performance of the model based on di-CovSel selected variables was better from the start compared to the model based on CovSel selected variables, although the CovSel variables-based model could have performed better than the di-CovSel with 7 selected variables. However, it should be noted that one might not have found the optimal number of variables as 7 from the CV plot (Fig. 2A). On the other hand, the di-CovSel selected variables (>7) always performed better than the CovSel selected variables. Similarly, for the wheat data set (Fig. 5B), the model based on di-CovSel selected variables performed better than the model based on CovSel selected variables (for >5 variables). For the pear data set (Fig. 5C), the model based on di-CovSel selected variables performed better than the model based on CovSel selected variables <8. For the mango data set (Fig. 5D), the di-CovSel and CovSel selected variable models performed similarly in the initial range of variables <9. However, the di-CovSel selected variables model outperformed the model based on CovSel selected variables for higher numbers of variables.

To illustrate the difference in the functionality of the CovSel and di-CovSel, each step in the variable selection related to the pear data set is shown in Fig. 6. As a total of seven variables were selected for the pear fruit case, there are seven rows in Fig. 6. Furthermore, the two columns are related to CovSel (left column) and di-CovSel (right column). It can be

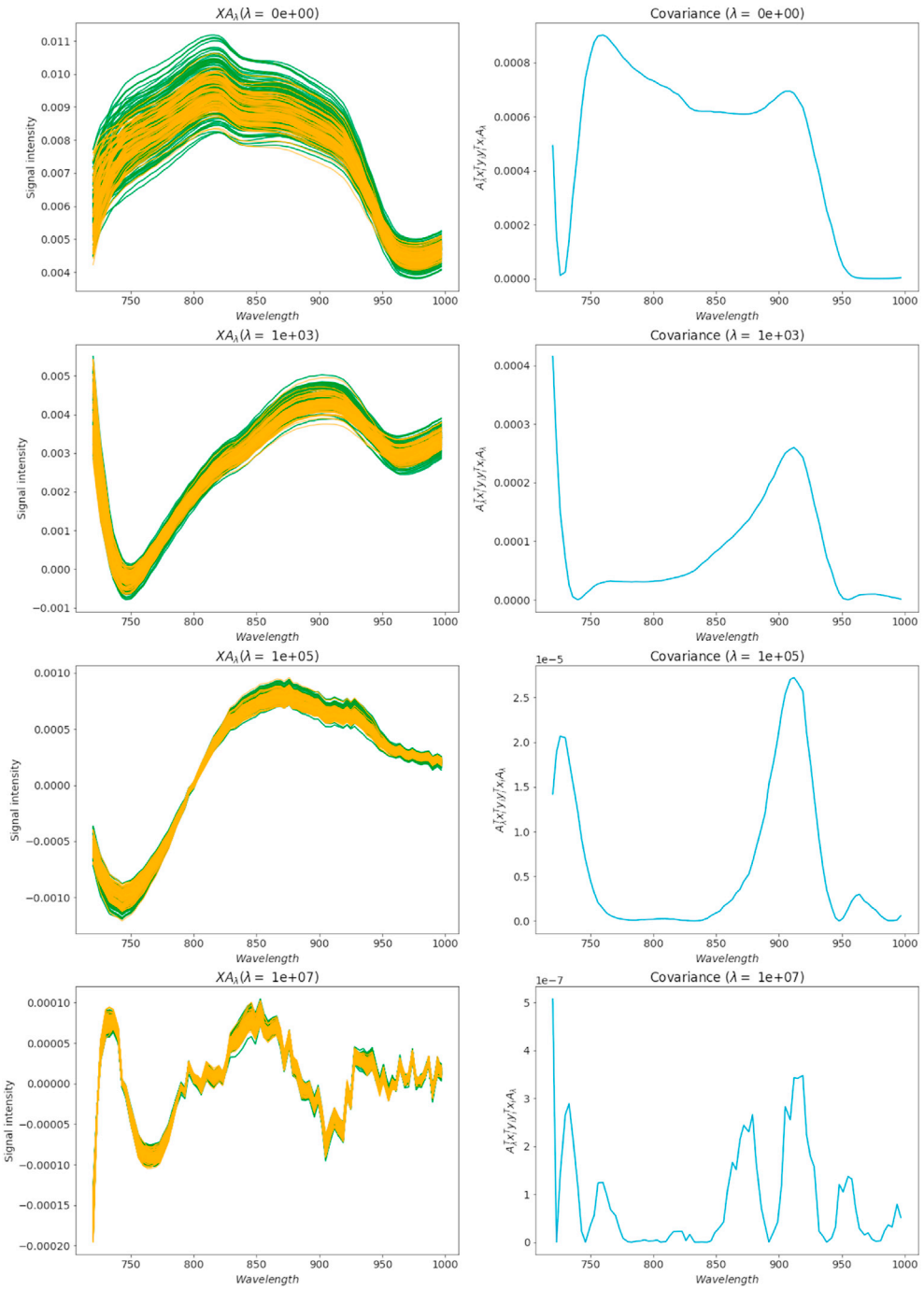

Fig. 7. Illustration of the variable selection criterion as a function of the regularization parameter in di-CovSel for the pear data set. Weighted source (yellow) and target spectra (green) by matrix **A** (left column) and squared covariance criterion (right column) for four values of regularization (top to bottom). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

noted that the first variable selected by the CovSel method was related to the overtone bands (Fig. 6A). In the case of di-CovSel, it was the shortest wavelength in the dataset (Fig. 6B) which may be attributed to an offset correction [34]. The CovSel selected such an offset as a second variable (Fig. 6C), while the second variable selected by di-CovSel was related to the chemical overtones (Fig. 6D). It should be noted that the spectral variable around 760 nm selected by CovSel was never selected by the di-CovSel method. This is remarkable as the region around 760 nm was identified as the one with the largest differences between the two data batches (Fig. 1C). Hence, avoidance of that region by the di-CovSel method is in line with our expectations. The variable 760 nm included in the CovSel selection could be the reason for the inferior performance of the CovSel method on the target domain compared to the di-CovSel, as all other variables selected by CovSel and di-Covsel were neighboring bands. It should also be noted that the di-CovSel method selected better-resolved peaks which can be related to chemical overtones compared to the broader global peaks selected by the CovSel method (Fig. 6).

The behavior of the covariance criterion for variable selection as a function of the regularization parameter is demonstrated for the first iteration of the di-CovSel algorithm related to the pear data set in Fig. 7. The first row of the figure shows the original spectra for $\lambda = 0$ which is equivalent to CovSel. As the value of λ increases, the weighted spectra show less variability between the domains, while showing different

peaks according to the covariance criterion. In the third row of Fig. 7, with a larger regularization value there are specifically 2 regions, around 750 nm and between 850 nm and 950 nm, which are still not invariant. With the last increase of λ, all the regions became invariant. However, the degree of noise is of important notice in the last row of Fig. 7. It can be noted that from 700 nm to 800 nm the weighted spectra remained smoother than the resulting signal across the rest of the spectrum. This was in line with the highest peak in the covariance criterion obtained around 700 nm making it the first selected variable, in contrast to the less regularized covariances. Therefore, we highlight that the method can select wavelengths that remain more stable across domains and the regularization parameter has the potential to find such stable bands. It is of important notice here that the domain invariant criterion supports the identification of such bands, which are further used to build a model using the original source domain data with reference values.

6. Conclusions

A new method called domain invariant covariate selection (di-CovSel) for generalized feature selection across different data domains was presented, which combines the principles of domain invariant PLS (di-PLS) and covariate selection (CovSel). The method was tested on diverse cases of domain adaptation such as adapting models for different physical forms of samples, adapting models for seasonal differences and performing tasks such as calibration transfer for similar modality (point to point spectrometer), as well as different modalities of NIR spectroscopy (point spectrometer to spectral camera). In all the presented cases, the multivariate predictive models based on variables selected by adding the domain invariance to the CovSel approach led to lower standard error of prediction and prediction bias compared to the models based on variables selected with the original CovSel method. However, it should be noted that for all cases the performance in the target domain was poorer than the performance on the source domain. This finding related to di-CovSel aligns with the conclusions of di-PLS, which states that the domain invariance does not guarantee similar performances across domains, but better performance across domains compared to models that rely on monodomain (source domain only) information. Furthermore, in three out of four cases better prediction performance was obtained with the same number of variables as selected by the CovSel method. Moreover, di-CovSel suggested that even with a lower number of variables a better performance compared to CovSel can be achieved for several cases. Interpreting the selected variables allowed us to understand that the di-CovSel method utilizing the knowledge about the spectra from the target domain was able to avoid the spectral regions which were most influenced by the domain differences and to select slightly shifted peaks compared to the CovSel based approach. Avoidance of the spectral region that has domain differences led to bias correction, hence, reduction in the prediction error. A key point to note is that the di-CovSel variable selection approach does not require any new reference analytical measurements and solely relies on the spectra from the target domain. Hence, solely relying on the spectra makes it a very practical tool to be used in practice as often the reference analysis is the main limiting factor to update models with traditional chemometric approaches. In the current application of di-CovSel, it was possible to perform a successful tuning of the number of variables and the value of the domain invariant regularization parameter λ based on cross-validation with the source domain calibration data. Nonetheless, we suggest the practitioner to validate the selected values of these tuning parameters using reference measurements in the target domain. In addition, with the availability of a few reference measurements, the performance of the adapted models was improved by correcting the degradation in the bias term of the calibration models. In this work, it was shown that as low as 60 samples were sufficient to allow di-CovSel to select generalized variables. Although in this study all the presented cases were related to NIR spectroscopy, the method is applicable to any area of multivariate data modelling where often the models fail when they need to be used in a new domain. The new domain can be anything related to the different physical forms of samples, different instruments and modalities and seasonal effects. Note that, in general, the domain invariant techniques such as di-PLS and di-CovSel, are bound to succeed if the required adaptation concerns the spectral variation and no changes in the relationship between spectral values and reference analysis occur.

Author statement

Valeria Fonseca Diaz: Conceptualization, Methodology, Software, Formal analysis, Writing - Original Draft. Puneet Mishra: Conceptualization, Methodology, Writing - Original Draft. Jean Michel Roger: Conceptualization, Methodology, Writing - Original Draft. Wouter Saeys: Conceptualization, Methodology, Writing - Original Draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Valeria Fonseca Diaz was funded as a doctoral fellow of the Research Foundation-Flanders (FWO, Brussels, Belgium). The codes will be made available at: https://github.com/vfonsecad?tab=repositories.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2022.104499.

References

[1] C. Pasquini, Near infrared spectroscopy: a mature analytical technique with new perspectives – a review, Anal. Chim. Acta 1026 (2018) 8–36.
[2] S. Lohumi, M.S. Kim, J. Qin, B.-K. Cho, Raman imaging from microscopy to macroscopy: quality and safety control of biological materials, Trac. Trends Anal. Chem. 93 (2017) 183–198.
[3] R.F. Lu, R. Van Beers, W. Saeys, C.Y. Li, H.Y. Cen, Measurement of optical properties of fruits and vegetables: a review, Postharvest Biol. Technol. 159 (2020).
[4] Ó. López-Campos, J.C. Roberts, I.L. Larsen, N. Prieto, M. Juárez, M.E.R. Dugan, J.L. Aalhus, Rapid and non-destructive determination of lean fat and bone content in beef using dual energy X-ray absorptiometry, Meat Sci. 146 (2018) 140–146.
[5] A.A. Gowen, C. O'Sullivan, C.P. O'Donnell, Terahertz time domain spectroscopy and imaging: emerging techniques for food process monitoring and quality control, Trends Food Sci. Technol. 25 (2012) 40–46.
[6] B.M. Nicolai, K. Beullens, E. Bobelyn, A. Peirs, W. Saeys, K.I. Theron, J. Lammertyn, Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: a review, Postharvest Biol. Technol. 46 (2007) 99–118.
[7] R. Rocha de Oliveira, A. de Juan, SWiVIA – Sliding window variographic image analysis for real-time assessment of heterogeneity indices in blending processes monitored with hyperspectral imaging, Anal. Chim. Acta 1180 (2021) 338852.
[8] L. Coic, P.-Y. Sacré, A. Dispas, C. De Bleye, M. Fillet, C. Ruckebusch, P. Hubert, E. Ziemons, Pixel-based Raman hyperspectral identification of complex pharmaceutical formulations, Anal. Chim. Acta 1155 (2021) 338361.
[9] B. Fei, J.M. Amigo, Chapter 3.6 - Hyperspectral Imaging in Medical Applications, Data Handling in Science and Technology, 2020, pp. 523–565. Elsevier.
[10] A. Martyna, A. Menżyk, A. Damin, A. Michalska, G. Martra, E. Alladio, G. Zadora, Improving discrimination of Raman spectra by optimising preprocessing strategies on the basis of the ability to refine the relationship between variance components, Chemometr. Intell. Lab. Syst. 202 (2020) 104029.
[11] P. Mishra, S. Lohumi, H. Ahmad Khan, A. Nordon, Close-range hyperspectral imaging of whole plants for digital phenotyping: recent applications and illumination correction approaches, Comput. Electron. Agric. 178 (2020) 105780.
[12] P. Mishra, M.S.M. Asaari, A. Herrero-Langreo, S. Lohumi, B. Diezma, P. Scheunders, Close range hyperspectral imaging of plants: a review, Biosyst. Eng. 164 (2017) 49–67.
[13] D. Pérez-Marín, E. De Pedro Sanz, J.E. Guerrero-Ginel, A. Garrido-Varo, A feasibility study on the use of near-infrared spectroscopy for prediction of the fatty acid profile in live Iberian pigs and carcasses, Meat Sci. 83 (2009) 627–633.
[14] W. Saeys, N.N. Do Trong, R. Van Beers, B.M. Nicolai, Multivariate calibration of spectroscopic sensors for postharvest quality evaluation: a review, Postharvest Biol. Technol. (2019) 158.
[15] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods 6 (2014) 2812–2831.

[16] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta 185 (1986) 1–17.
[17] S. Wold, PLS Modeling with Latent Variables in Two or More Dimensions, 1987.
[18] R.D. Cramer, Partial least squares (PLS): its strengths and limitations, Perspect. Drug Discov. Des. 1 (1993) 269–278.
[19] J.-M. Roger, F. Chauchard, V. Bellon-Maurel, EPO–PLS external parameter orthogonalisation of PLS application to temperature-independent measurement of sugar content of intact fruits, Chemometr. Intell. Lab. Syst. 66 (2003) 191–204.
[20] J.-M. Roger, F. Chauchard, P. Williams, Removing the block effects in calibration by means of dynamic orthogonal projection. Application to the year effect correction for wheat protein prediction, J. Near Infrared Spectrosc. 16 (2008) 311–315.
[21] P. Mishra, J.M. Roger, D.N. Rutledge, E. Woltering, Two standard-free approaches to correct for external influences on near-infrared spectra to make models widely applicable, Postharvest Biol. Technol. 170 (2020) 111326.
[22] P. Mishra, R. Nikzad-Langerodi, A brief note on application of domain-invariant PLS for adapting near-infrared spectroscopy calibrations between different physical forms of samples, Talanta 232 (2021) 122461.
[23] P. Mishra, R. Nikzad-Langerodi, Partial Least Square Regression versus Domain Invariant Partial Least Square Regression with Application to Near-Infrared Spectroscopy of Fresh Fruit, Infrared Physics & Technology, 2020, p. 103547.
[24] P. Mishra, J.M. Roger, F. Marini, A. Biancolillo, D.N. Rutledge, FRUITNIR-GUI: a graphical user interface for correcting external influences in multi-batch near infrared experiments related to fruit quality prediction, Postharvest Biol. Technol. (2020) 111414.
[25] P. Mishra, R. Nikzad-Langerodi, F. Marini, J.M. Roger, A. Biancolillo, D.N. Rutledge, S. Lohumi, Are standard sample measurements still needed to transfer multivariate calibration models between near-infrared spectrometers? The answer is not always, Trac. Trends Anal. Chem. (2021) 116331.
[26] M. Zeaiter, J.M. Roger, V. Bellon-Maurel, Dynamic orthogonal projection. A new method to maintain the on-line robustness of multivariate calibrations. Application to NIR-based monitoring of wine fermentations, Chemometr. Intell. Lab. Syst. 80 (2006) 227–235.
[27] R. Nikzad-Langerodi, W. Zellinger, E. Lughofer, S. Saminger-Platz, Domain-invariant partial-least-squares regression, Anal. Chem. 90 (2018) 6693–6701.
[28] R. Nikzad-Langerodi, W. Zellinger, S. Saminger-Platz, B. Moser, Domain-Invariant Regression under Beer-Lambert's Law, 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA), 2019, pp. 581–586.
[29] G. Huang, X. Chen, L. Li, X. Chen, L. Yuan, W. Shi, Domain adaptive partial least squares regression, Chemometr. Intell. Lab. Syst. 201 (2020) 103986.
[30] S.J. Pan, I.W. Tsang, J.T. Kwok, Q. Yang, Domain adaptation via transfer component analysis, IEEE Trans. Neural Network. 22 (2011) 199–210.
[31] J.M. Roger, Orthogonal projections in the row and the column spaces, NIR News 27 (2016) 15–20.
[32] T. Mehmood, K.H. Liland, L. Snipen, S. Sæbø, A review of variable selection methods in Partial Least Squares Regression, Chemometr. Intell. Lab. Syst. 118 (2012) 62–69.
[33] T. Mehmood, S. Sæbø, K.H. Liland, Comparison of variable selection methods in partial least squares regression, J. Chemometr. (2020) e3226, n/a.
[34] J.M. Roger, B. Palagos, D. Bertrand, E. Fernandez-Ahumada, CovSel: variable selection for highly multivariate and multi-response calibration Application to IR spectroscopy, Chemometr. Intell. Lab. Syst. 106 (2011) 216–223.
[35] R. Nikzad-Langerodi, W. Zellinger, S. Saminger-Platz, B.A. Moser, Domain adaptation for regression under Beer–Lambert's law, Knowl. Base Syst. 210 (2020) 106447.
[36] Z. Xu, S. Fan, J. Liu, B. Liu, L. Tao, J. Wu, S. Hu, L. Zhao, Q. Wang, Y. Wu, A calibration transfer optimized single kernel near-infrared spectroscopic method, Spectrochim. Acta Mol. Biomol. Spectrosc. 220 (2019) 117098.
[37] H. Liu, B. Bruning, T. Garnett, B. Berger, The Performances of Hyperspectral Sensors for Proximal Sensing of Nitrogen Levels in Wheat, Sensors, 2020.
[38] P. Mishra, Chemometric approaches for calibrating high-throughput spectral imaging setups to support digital plant phenotyping by calibrating and transferring spectral models from a point spectrometer, Anal. Chim. Acta 1187 (2021) 339154.
[39] P. Mishra, E. Woltering, Handling batch-to-batch variability in portable spectroscopy of fresh fruit with minimal parameter adjustment, Anal. Chim. Acta 1177 (2021) 338771.
[40] N.T. Anderson, K.B. Walsh, P.P. Subedi, C.H. Hayes, Achieving robustness across season, location and cultivar for a NIRS model for intact mango fruit dry matter content, Postharvest Biol. Technol. 168 (2020) 111202.
[41] N. Anderson, K. Walsh, P. Subedi, Mango DMC and spectra Anderson et al. 2020, Mendley, Mendley data, 2020.
[42] N.T. Anderson, K.B. Walsh, J.R. Flynn, J.P. Walsh, Achieving robustness across season, location and cultivar for a NIRS model for intact mango fruit dry matter content. II. Local PLS and nonlinear models, Postharvest Biol. Technol. 171 (2021) 111358.
[43] P. Mishra, D. Passos, Deep chemometrics: validation and transfer of a global deep near-infrared fruit model to use it on a new portable instrument, J. Chemometr. (2021) e3367, n/a.
[44] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137–148.
[45] B.G. Osborne, Near-Infrared Spectroscopy in Food Analysis, Encyclopedia of Analytical Chemistry, 2006.