Analytica Chimica Acta xxx (2016) 1–9

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Regression models based on new local strategies for near infrared spectroscopic data

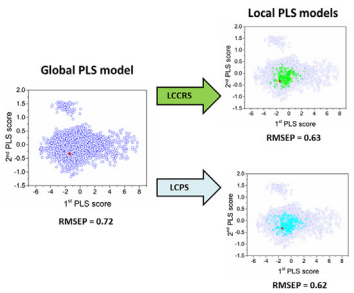
F. Allegrini [a, *], J.A. Fernández Pierna [b], W.D. Fragoso [c], A.C. Olivieri [a], V. Baeten [b], P. Dardenne [b]

[a] Univ. Nacional de Rosario, Facultad de Ciencias Bioquímicas y Farmacéuticas, IQUIR, CONICET, Argentina
[b] Valorisation of Agricultural Products Dpt, Walloon Agricultural Research Centre, Gembloux, Belgium
[c] Departamento de Química, Universidade Federal da Paraíba, Campus I, João Pessoa, Brazil

HIGHLIGHTS

- New local regression models based on PLS scores.
- Rapid quantification of five major constituents in corn seeds using near infrared spectroscopy.
- Statistically significant predictions improvement with respect to global PLS.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 7 April 2016
Received in revised form
23 June 2016
Accepted 5 July 2016
Available online xxx

Keywords:
Local regression models
Near infrared spectroscopy
Partial least squares regression

ABSTRACT

In this work, a comparative study of two novel algorithms to perform sample selection in local regression based on Partial Least Squares Regression (PLS) is presented. These methodologies were applied for Near Infrared Spectroscopy (NIRS) quantification of five major constituents in corn seeds and are compared and contrasted with global PLS calibrations. Validation results show a significant improvement in the prediction quality when local models implemented by the proposed algorithms are applied to large data bases.

1. Introduction

Since years, near infrared spectroscopy has been used to provide calibration models based on collected databases of many kinds of samples such as soils [1–3], fuels [4,5], grains [6–8] and others. Recent studies indicate the necessity to define the optimal size of a sample set in order to build calibration models with the highest possible accuracy. Although a large number of samples might be better in terms of product variability characterization, the cost of analyzing such large number of samples would be significantly higher [9]. This is often also justified by the accuracy required on the calibration models, which requires considerable research to optimize the calibration procedure and then the development of robust models. Anyhow, quite often, large databases have been collected by companies during years, avoiding the development of robust protocols due to the huge variability covered for the studied

* Corresponding author.
E-mail address: allegrini@iquir-conicet.gov.ar (F. Allegrini).

http://dx.doi.org/10.1016/j.aca.2016.07.006
0003-2670/

products. This kind of data bases leads to what is normally known as global calibrations.

Until now, large datasets lead to global calibrations, which in principle are expected to be very robust to sample composition variation. However often nature presents dependencies that are much more complex than those that can be captured with a simple linear and global parametric model. As databases get larger, they may increase the complexity in terms of variability and thus, what is normally seen as an advantage of global calibrations turns into a problem. This could be the case when samples to be predicted have not been observed because of the inclusion of new components or changes/drifts on the instruments, etc. A solution could be to accommodate this by specifying a complex parametric model with many parameters. Sometimes this can be successfully done, but finding the appropriate model can be very challenging, and can result in difficult-to-interpret coefficients. As a way to solve the situation previously stated, one of the main alternatives is a group of methods based on local regression principles [10–12]. These methods attempt to improve the prediction of unknown samples by means of calibration models, which are built according to the similarity between the sample to be predicted and the calibration samples. Instead of constraining to have a model with a parametric form, assume that the data locally, around some neighborhood of x, can be well approximated by a parametric form –namely low order polynomials.

Although there is a large heterogeneity of local regression methods, their basic principles come from a methodology known as Locally Weighted Regression (LWR) [13]. In LWR, the influence of each calibration sample on the quantification of an unknown sample is weighted according to their proximity. This leads to a customized PCR model for each test sample. A simplification of LWR consists on LOCAL algorithms [14,15], where instead of a continuous weighing, samples are selected or discarded for the prediction in a discrete way according to a distance limit. This means that if the distance to a test sample calculated for a calibration sample is below the established limit, it will be included in the model and if it is above that limit it will be taken out. The main difference among different LWR and LOCAL methods is the way the distance between samples and the weights are determined.

Another alternative to perform local regression is the Comparison Analysis using Reconstructed Near Infrared and Constituent Data (CARNAC), which instead of using metric distances between spectral samples uses a method combining Fast Fourier Transformed spectra and reference values [16].

Some recent applications of local regression methods showed their utility when applied to soil samples [3]. In these cases, the metrics were built as complex indexes which combine different ways of measuring distances between samples. However, the way to select the optimal number of samples is left to the user criteria, despite the fact this latter aspect is essential in local methods. Moreover, there is no strict comparison of the performance of the local regression methods proposed with respect to global calibration. Together with the latter, many other examples of applications of local calibrations to different types of data sets can be found in the literature [17–19].

In this work, two novel approaches to perform sample selection in local regression methods are presented and compared: the Local Calibration by Percentile Selection (LCPS) and the Local Calibration by Customized Radii Selection (LCCRS). These local selection methodologies share two important features: (1) they are based on an attempt to rationalize and automatize the decision about the number of samples used to build each local model and, (2) they operate on PLS scores space, meaning that the distance between samples is measured considering spectral similarities but also reference values coincidences. To test both approaches the results in the context of the quantification of five constituents in corn seeds were analyzed.

2. Material and methods

2.1. Sample set

As presented in Table 1 datasets of more than 3000 corn samples were available for each parameter (moisture, protein, fiber, fat and ash). These parameters, which represent the major constituents of corn seeds and were measured by the corresponding reference methods. NIR spectra for all samples were collected using a FOSS NIRSystem 5000 working in the 1100–2498 nm range every 2 nm. The set named as “protein extra” refers to an extra set of samples that was only available for protein determination and that was used to test the robustness of the models presented in Section 2. For all the data sets, outliers were detected using a Mahalanobis distance based criterion:

$$MD_{ik} = \sqrt{(t_k - t_i)\Sigma_{T_{cal}}(t_k - t_i)^T}$$

$$MD_{i0} > \overline{MD}_{cal0} + 3\sigma_{MD_{cal}}$$

where MD_{ik} is the Mahalanobis distance between samples i and k, calculated in the PLS latent variable space, through the corresponding score vectors ($\mathbf{t}_k$ and $\mathbf{t}_i$). $\Sigma_{T_{cal}}$ is the covariance matrix calculated from the calibration scores matrix ($\mathbf{T}_{cal}$). Finally, MD_{i0} and $\overline{MD}_{cal0}$ are the Mahalanobis distances for a given sample and for the remaining samples in the set to the calibration center respectively, ‘¯’ indicates mean value and ‘σ’ standard deviation. For each sample, MD_{cal0} is computed by excluding the sample under study. This outlier filter is similar to the well-known GH criterion [20,21] with the difference that it does not divide the Mahalanobis distance by the number of dimensions, in order to be consistent with the way the algorithms used to perform local calibration work. The practical significance of the GH is that it indicates when a predicted value for a given sample is outside the limits defined by the population that make up the calibration set. Since the MD distances were calculated from the projections of the sample spectra onto the PLS latent space, the outlier detection criterion is taking into account both spectral and concentration information.

To finish this section, it is important to remark that the number of samples presented in Table 1 correspond the result of applying the previous outlier detection methodology both, over the calibration and test sets, i.e. they refer to the filtered datasets and not to the original ones.

2.2. Local Calibration by Percentile Selection (LCPS)

The LCPS strategy consists on first ranking, for each test sample to be predicted, the calibration samples according to their proximity to that unknown sample, using the Mahalanobis distance as farness measure. A number of samples is then selected to build a local calibration model adapted to the test sample under study. This local calibration set includes all samples within a certain percentile of all the previously computed Mahalanobis distances. Several alternatives have been studied and concluded that the best choice for the percentile was a value of 15% for the five calibrated parameters (see Section 3 for additional details). Fig. 1 shows a flow chart for the LCPS algorithm.

Please cite this article in press as: F. Allegrini, et al., Regression models based on new local strategies for near infrared spectroscopic data, Analytica Chimica Acta (2016), http://dx.doi.org/10.1016/j.aca.2016.07.006

Table 1
Description of the data sets for each parameter.

	Moisture	Protein	Protein extra	Fiber	Fat	Ash
Number of calibration samples	3341	3394	1469	1669	2955	2939
Number of test samples	2302	2430	0	1333	1957	1791
Calibration mean value[a]	11.81	46.49	46.87	4.31	1.79	6.72
Calibration standard deviation[a]	1.39	2.02	1.62	1.20	1.28	0.62
Calibration minimum[a]	4.55	38.09	40.77	2.48	0.1	4.99
Calibration maximum[a]	15.5	52.86	52.52	8.86	13.05	9.04

[a] Expressed as % w/w.

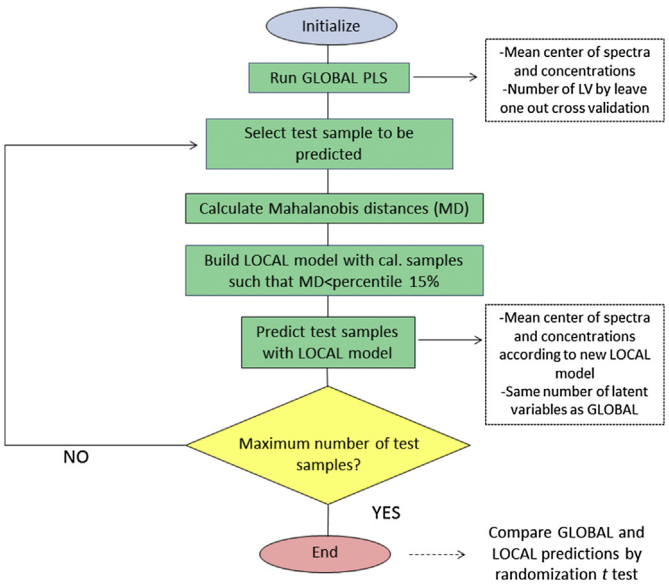


Fig. 1. Flow chart for LCPS algorithm.

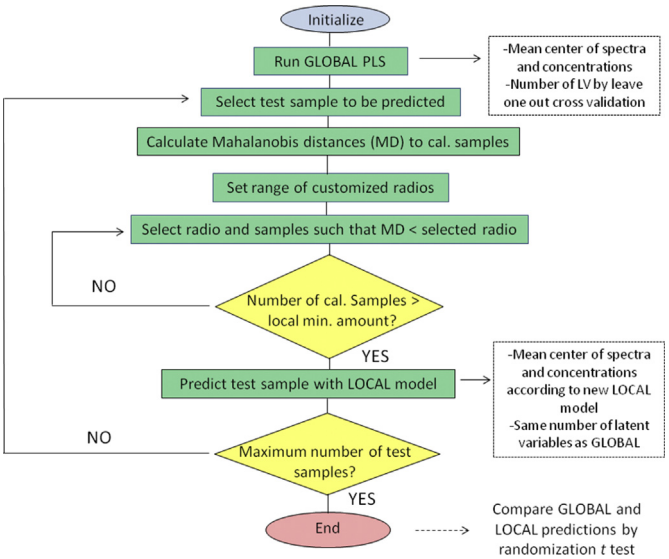


Fig. 2. Flow chart for LCCRS algorithm.

2.3. Local Calibration by Customized Radii Selection (LCCRS)

The LCCRS algorithm does also use the Mahalanobis distance to build local calibration models. However, the main difference with LCPS, is that a variable number of calibration samples is selected for each test sample. Specifically, a customized range of radii around a given test sample is set through a two-step procedure: (1) the mean and the standard deviation of the Mahalanobis distances ($\overline{MD}$ and σ_{MD}) are computed for the test sample with respect to all calibration samples, (2) a set of radii is defined by dividing the interval $MD \pm 3\sigma_{MD}$ in six equally spaced values. These radii are then tested by gradually increasing the distance from the test sample, until a user-defined number of calibration samples is reached. Fig. 2 shows a flow chart for the LCCRS algorithm.

2.4. LCPS, LCCRS and other clustering methods

A local calibration sample subset can be understood as a cluster representing a particular test sample. Just as the classical aggregation k-NN algorithm [22,23], here the similarity between samples is understood as the proximity between them in the multivariate space. LCPS takes a closer given percentile, determining the size of k (number of neighbors included in the model) as a fraction of the size of the global set of samples. LCCRS does not fix the size of k as it only sets a minimum amount that must be achieved. However, LCCRS operates as close as possible to the test sample, increasing the radius only if the minimum value for k is not reached in the shortest distance. In both methodologies presented the objective is always to find a calibration set of samples spatially nearest the test sample.

These methods follow the ideas of unsupervised density-based strategies such as DBSCAN (Density Based Spatial Clustering of Applications with Noise) and OPTICS (Ordering Points to Investigate the Clustering Structure) [24–27]. In these methods, a cluster is defined as a region in the multivariate space with high density of samples that is separated from other clusters by regions with low sample density. Both approaches, i.e., DBSCAN and OPTICS, can be related to the single linkage clustering techniques and the common idea with LCCRS is that for each object of a cluster the neighborhood of a given radius has to contain at least a minimal number of other objects. However, the final aim of LCPS and LCCRS algorithms is not the same. While unsupervised density based methods try to reveal clusters of different shapes and densities, the local methods proposed here work in a supervised way and what they try to do, is to identify, in the PLS score space, a cluster in the global calibration set containing the test sample, and then build a calibration using the samples of that cluster, avoiding the samples with different characteristics which may lead to instabilities in the model. The LCPS and LCCRS algorithms accept less dense clusters, meaning that if a test sample is projected into the latent variable space and the density of samples around it is not high enough, this will not limit the selection of a particular subset to perform local calibration.

3. Results and discussion

Fig. 3 shows the plots of the first vs. second PLS scores corresponding to the global calibrations sets for each studied parameter, together with the projected scores for the validation sets. The distributional match between the two sets is due to the usage of

Please cite this article in press as: F. Allegrini, et al., Regression models based on new local strategies for near infrared spectroscopic data, Analytica Chimica Acta (2016), http://dx.doi.org/10.1016/j.aca.2016.07.006

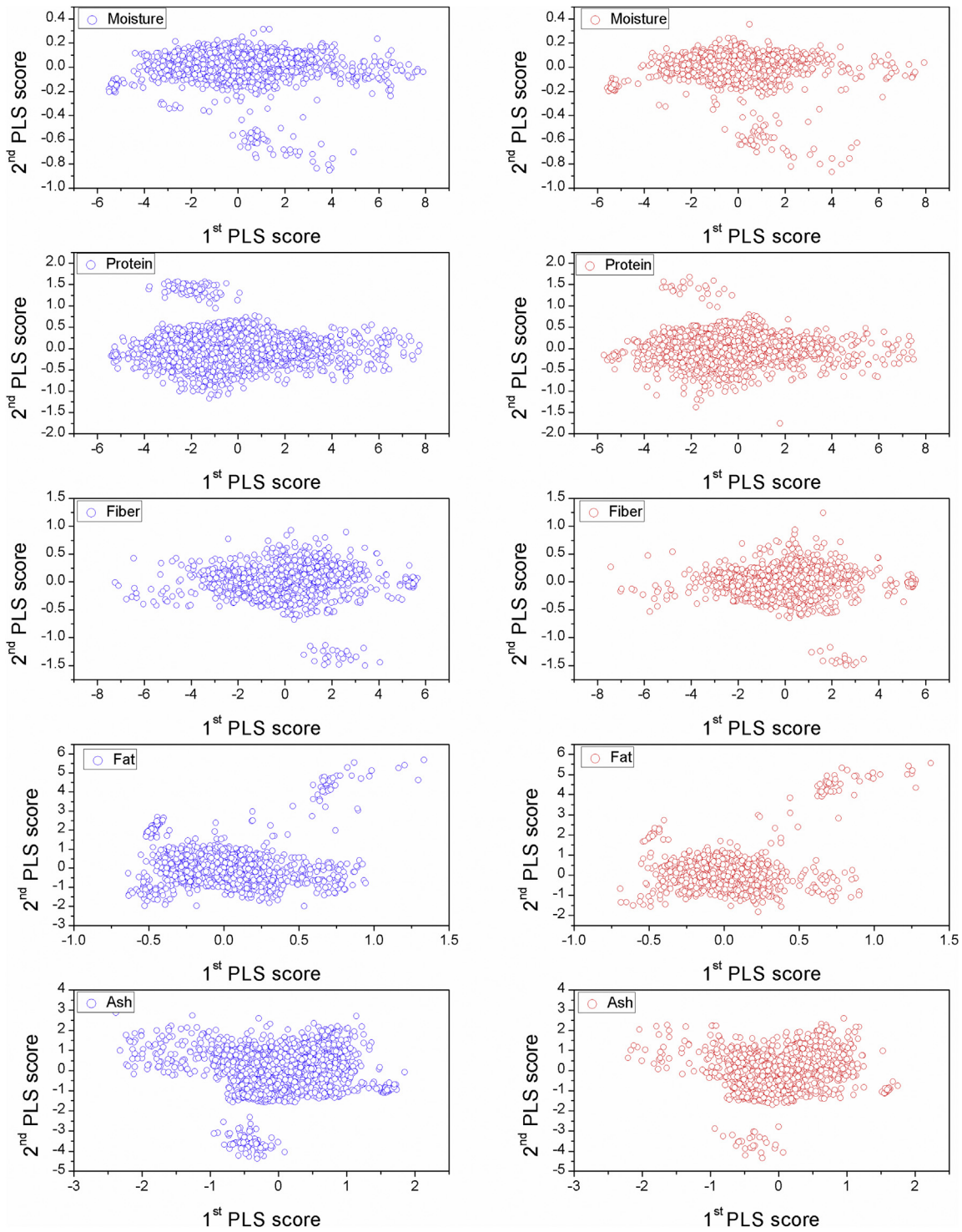


Fig. 3. Plots of first vs. second global PLS scores for the five parameters. Left panels, calibration sets, right panels, projected score plots for the validation sets.

Kennard Stone algorithm [28], to separate samples in a homogeneous way and to ensure a basic condition of first order calibration which is that the test samples should be well represented by the universe of calibration samples.

In large datasets as the presently studied, samples may often present a rather wide dispersion in the multivariate space. In addition, it is apparent from Fig. 3 that the inclusion of a large variety of samples belonging to different sub-populations, leads to the formation of clusters which appear isolated from the main bulk. In consequence, a global PLS model including all calibration samples, does not take into account that a specific test sample could be significantly different from most of the calibration samples, even if it the former is not an outlier. This sample heterogeneity introduces non-linearities in the system [29,30], which are difficult to model by a conventional PLS algorithm, even if the number of latent variables is increased. Thus global models should be less accurate

Please cite this article in press as: F. Allegrini, et al., Regression models based on new local strategies for near infrared spectroscopic data, Analytica Chimica Acta (2016), http://dx.doi.org/10.1016/j.aca.2016.07.006

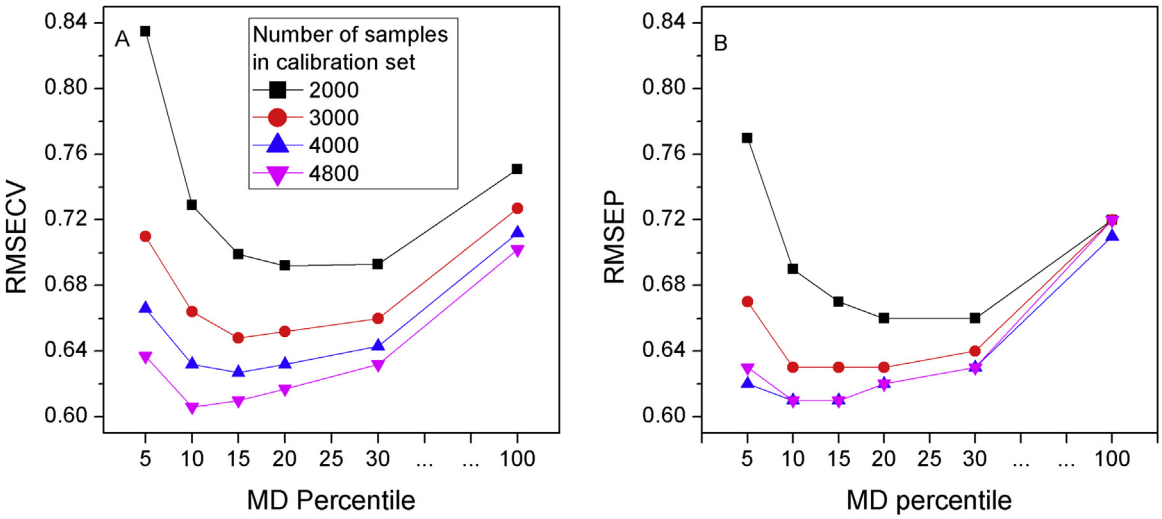


Fig. 4. RMSECV (A) and RMSEP (B) vs. MD percentile for different calibration set sizes for the determination of protein content.

compared with local models built with samples located in the neighborhood of the test sample under scrutiny.

As explained above, for the LCPS algorithm, the main parameter to be tuned is the percentile, i.e. the relative number of calibration samples to be selected for local model building. To rationalize this selection, a leave one out cross validation strategy changing the percentile value was applied. After that, in order to verify this selection, the results were contrasted with the RMSEP values calculated over an independent test data set. The study was first focused on the determination of the protein content, as for this parameter, an extra set of samples was available (see description as "protein extra" in Table 1) to be added to the total number of samples presented in Table 1. This larger number of samples provided, then, the opportunity to test the robustness of the chosen percentile with respect to the number of selected samples. Fig. 4A and B respectively show the changes in RMSECV and RMSEP for increasing number of samples in the calibration set, as the percentile of Mahalanobis distance is gradually increased. It can be seen that the variation of these values for the global model (percentile = 100%) is almost independent on the size of the calibration set.

On the other hand, when the LCPS strategy is applied, the corresponding RMSE values strongly varies as a function of the calibration size. Smaller RMSECV and RMSEP values are observed for percentiles in the range 10–20%. In Fig. 5A and B, the results when the same analysis is extended to the remaining parameters (using the total number of samples specified in Table 1) are displayed. This figure clearly shows that prediction error values for validation and test set, changes with the percentile for all studied parameters (see Table 1 for the number of calibration samples for these models).

From Figs. 4 and 5, we suggest the 15th percentile as a general good choice for the five parameters evaluated, although one may consider using the 10th percentile if the calibration set is large enough (as is the case for ash, moisture and protein). Only fiber (with less than 2000 samples in the calibration set) and oil showed significant increase in the RMSECV and RMSEP when the 10th percentile was used instead of the 15th percentile for the data sets detailed in Table 1 and represented in Fig. 4.

As previously discussed, in the LCCRS strategy there are two parameters affecting the sample selection: (1) the radii, and (2) the minimal number of local selected samples. The former was fixed to 6, which is considered an adequate value to reach a reasonable

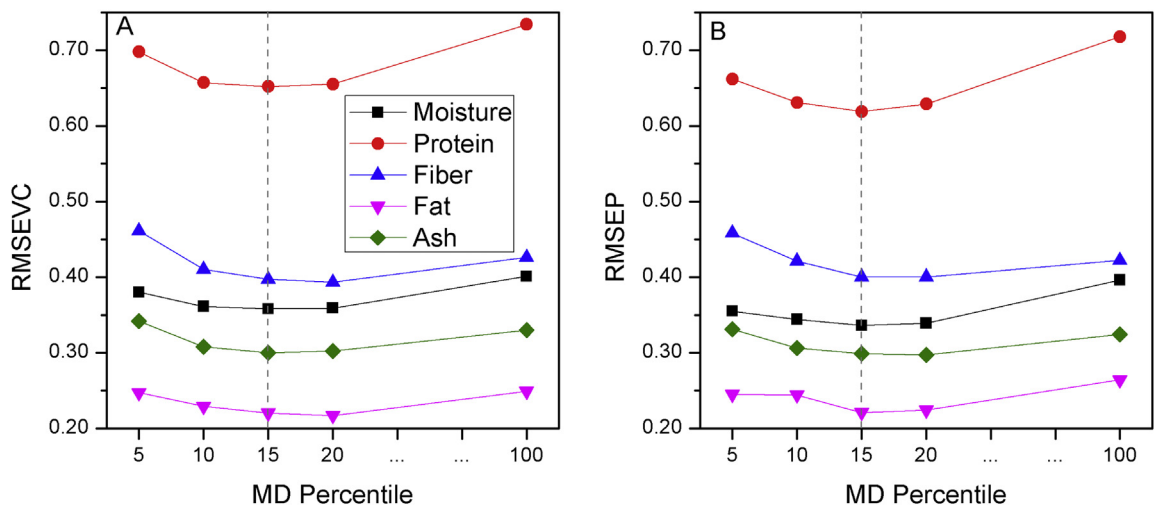


Fig. 5. RMSECV (A) and RMSEP (B) vs. MD percentile for the five parameters.

Please cite this article in press as: F. Allegrini, et al., Regression models based on new local strategies for near infrared spectroscopic data, Analytica Chimica Acta (2016), http://dx.doi.org/10.1016/j.aca.2016.07.006

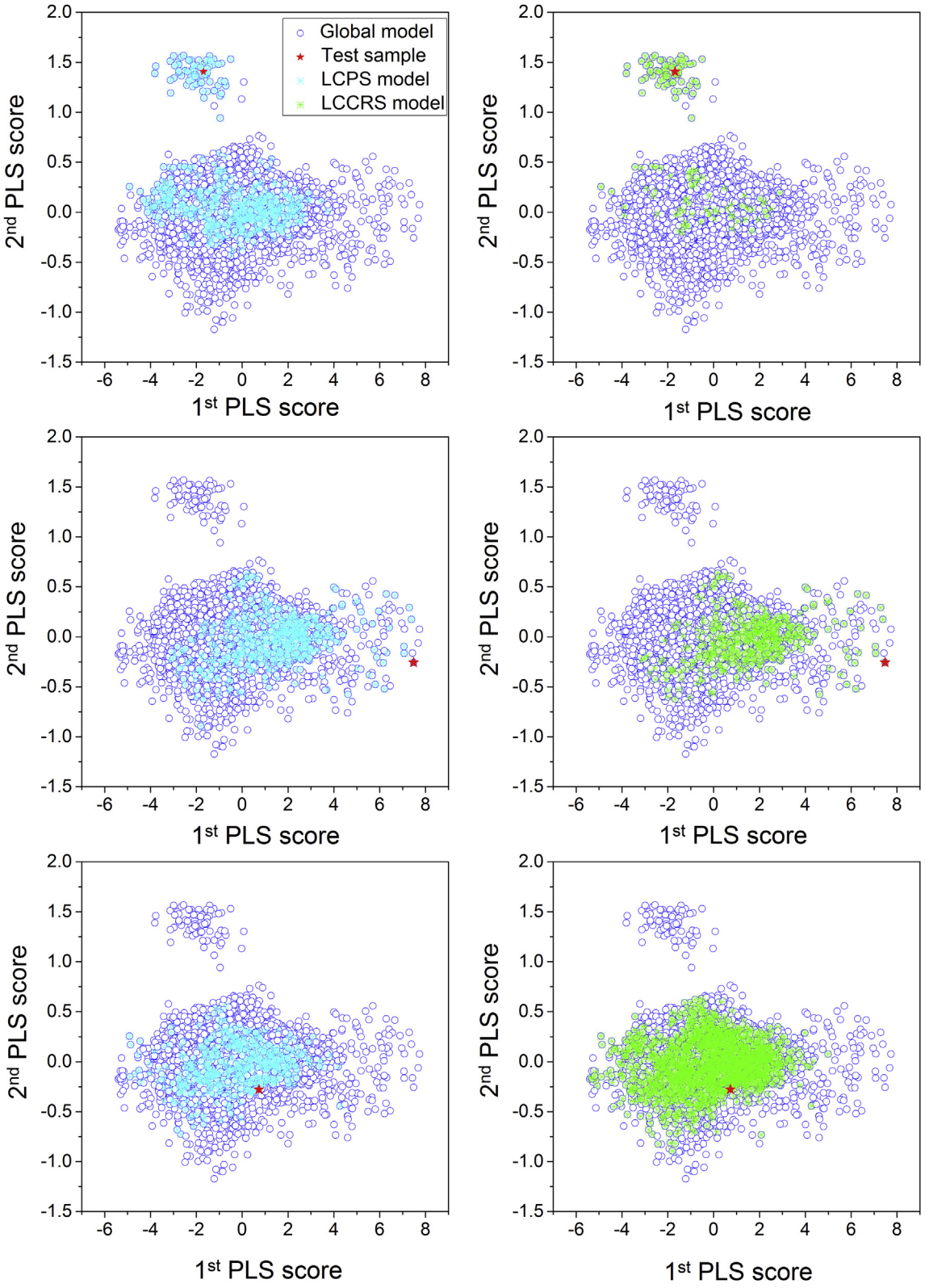


Fig. 6. Samples selected for LCPS and LCCRS for three different test samples.

trade-off between computational cost and exhaustive-enough interval scanning. This latter aspect has to do with the fact that, as long as the number of radii is larger, it will allow for a more detailed scan of the interval established for each test sample. In consequence the number of calibration samples will remain as close as possible to the minimum amount set by the user to build each local model. To tune this second parameter, the previous approach employed to select the optimal percentile for LCPS was

Please cite this article in press as: F. Allegrini, et al., Regression models based on new local strategies for near infrared spectroscopic data, Analytica Chimica Acta (2016), http://dx.doi.org/10.1016/j.aca.2016.07.006

Table 2
Statistical predictive results obtained by each model for different parameters.

Parameter	Model			
	Global	LCPS	LCCRS	15% random
Moisture				
A			24	
RMSEP	0.40	0.34	0.33	0.42
REP %	3.4	2.8	2.8	3.6
R^2	0.92	0.95	0.95	0.91
Q^2	0.92	0.93	0.94	0.90
Protein				
A			24	
RMSEP	0.72	0.62	0.63	0.78
REP %	1.5	1.3	1.4	1.7
R^2	0.87	0.90	0.90	0.85
Q^2	0.87	0.90	0.89	0.85
Fiber				
A			22	
RMSEP	0.42	0.40	0.41	0.47
REP %	9.9	9.3	9.4	10.8
R^2	0.88	0.89	0.89	0.85
Q^2	0.87	0.89	0.89	0.84
Fat				
A			22	
RMSEP	0.27	0.22	0.23	0.29
REP %	14.8	12.4	12.7	16
R^2	0.97	0.98	0.98	0.97
Q^2	0.96	0.97	0.97	0.95
Ash				
A			24	
RMSEP	0.32	0.30	0.29	0.34
REP %	4.8	4.5	4.4	5.2
R^2	0.74	0.78	0.79	0.71
Q^2	0.72	0.77	0.77	0.70

p(t): probability associated to the t value obtained by a one tail randomization t-test for the comparison between each local strategy and the GLOBAL predictions; REP%: Relative Error of Prediction, calculated as [RMSEP/(mean calibration reference values)]×100; R^2: correlation coefficient for predicted vs. actual values plot; Q^2: cross-validation correlation coefficient, i.e., correlation coefficient for predicted vs. actual values plot obtained after a leave one out cross validation performed using a particular model.

applied. Using again the protein data set, this analysis led to a minimal number of 100 local samples, which showed to be the smaller number of samples for which the RMSECV and the RMSEP is not significantly different than the minimum in all five parameters.

As it was anticipated during the methods description section, these two strategies led different number of samples chosen to build the local model. LCPS is based in percentiles, and always selects the same number of samples, independently of the test sample, because this number is a fixed fraction of the global calibration set. In contrast, in LCCRS the number of samples included in the model is strongly dependent on the position of the test sample in the global PLS score space. If the first region around a sample, defined by the first radius, contains lesser samples than the minimal number of samples required to build the local model, LCCRS will expand the search to a second radius value, and so on. As a consequence, it is likely that more samples than the minimal limit set by the user will be included. Fig. 6 shows this behavior for three different test samples, representing three possible situations: the numbers of selected samples by LCCRS is smaller, similar or larger than for LCPS. As in Fig. 3, to represent the samples in the latent variable space, first and second scores values were chosen. As expected, this selection can be explained on the basis that the first two latent variables gather the largest possible amount of explained variance. However, is important to make clear that to decide this latter number that this was not the final number of latent variables used to run the models tested. Then, firstly a leave-one-out cross validation for each parameter was performed, using the global data set (the optimal number of factors obtained by this procedure is shown in Table 2). After that, as doing a cross validation for each new test sample would demand too much considering that the proposed models are intended to be applied at online prediction systems, we employed two different strategies. For LCPS the same number of latent variables as the global model was used together with classical PLS calibration. On the other hand, for LCCR a weighted version of the PLS algorithm was applied. This version uses the number of factors determined over the global model, but it weights the prediction for each latent variable by taking into account the stability of the regression coefficients and the size of the residuals. This means that the lower the standard deviation of the regression coefficients and the residuals obtained for a particular number of latent variables, the highest the weighting value that the PLS prediction using that amount of latent variables will receive. The other combinations (LCPS with weighting PLS and LCCRS with classical PLS) were also tested but the best results were obtained by the versions presented here.

Fig. 7 shows plots of predicted vs. reference values for the five studied parameters, considering the two local strategies applied, and comparing with the results obtained using the global data set. From the inspection of these plots it is evident that predicted parameters using the presently described local strategies, show a smaller dispersion with respect to nominal values. This latter observation is a graphical representation of what is then confirmed by the numerical results shown in Table 2: LCPS and LCCRS lead to a smaller RMSEP than the global calibration.

To verify that the results shown in Fig. 7 for the comparison of LCPS and LCCRS to global models, were not a consequence of a particular combination of samples in the calibration set, from the maximum number of samples available in the protein data set (4863 samples), 30 different calibrations sets were built by randomly taking the 70% of the total number of samples (3394 samples). For the global models, the average error ± standard deviation for these 30 models was 0.720 ± 0.002. On the other hand, for LCPS and LCCRS they were 0.627 ± 0.005 and 0.657 ± 0.010, respectively. This is a way to confirm that the judicious selection of local calibration sets by the strategies proposed in this work, leads to better prediction ability than the global model, with that improvement being independent from random variations of the samples used to generate the global calibration.

Table 2 collects the statistical predictive RMSEP results for LCPS, LCCRS and global strategies. As expected for any local model, the distribution of the calibration set as close as possible around the test sample, and no the reduction of the number of samples in calibration set itself, is the critical factor leading to RMSEP improvement when comparing global to local models. To verify this assumption for the proposed algorithms, additional models were built with 15% of samples selected randomly (results presented in the last column of Table 2) from the whole calibration set. This means, that for these “random” models, calibration samples were not necessarily locally distributed around the test sample. It could be argued that comparison with a random selection of 15% is unfair because to give random selection the same chance, it would be necessary to try a range of different random percentages and pick the one with the best performance on the test set. However it should be noted that the higher the percentile when doing a random selection the closer the system is to global calibration and consequently the RMSEP over the test set tends to decrease as the percentile gets larger. This is the reason why fixing this value at 15% is, in principle, a reasonable choice to differentiate this random comparison trial, with the one performed with the complete global data set.

As expected, for the five studied parameters, LCPS and LCCRS

Please cite this article in press as: F. Allegrini, et al., Regression models based on new local strategies for near infrared spectroscopic data, Analytica Chimica Acta (2016), http://dx.doi.org/10.1016/j.aca.2016.07.006

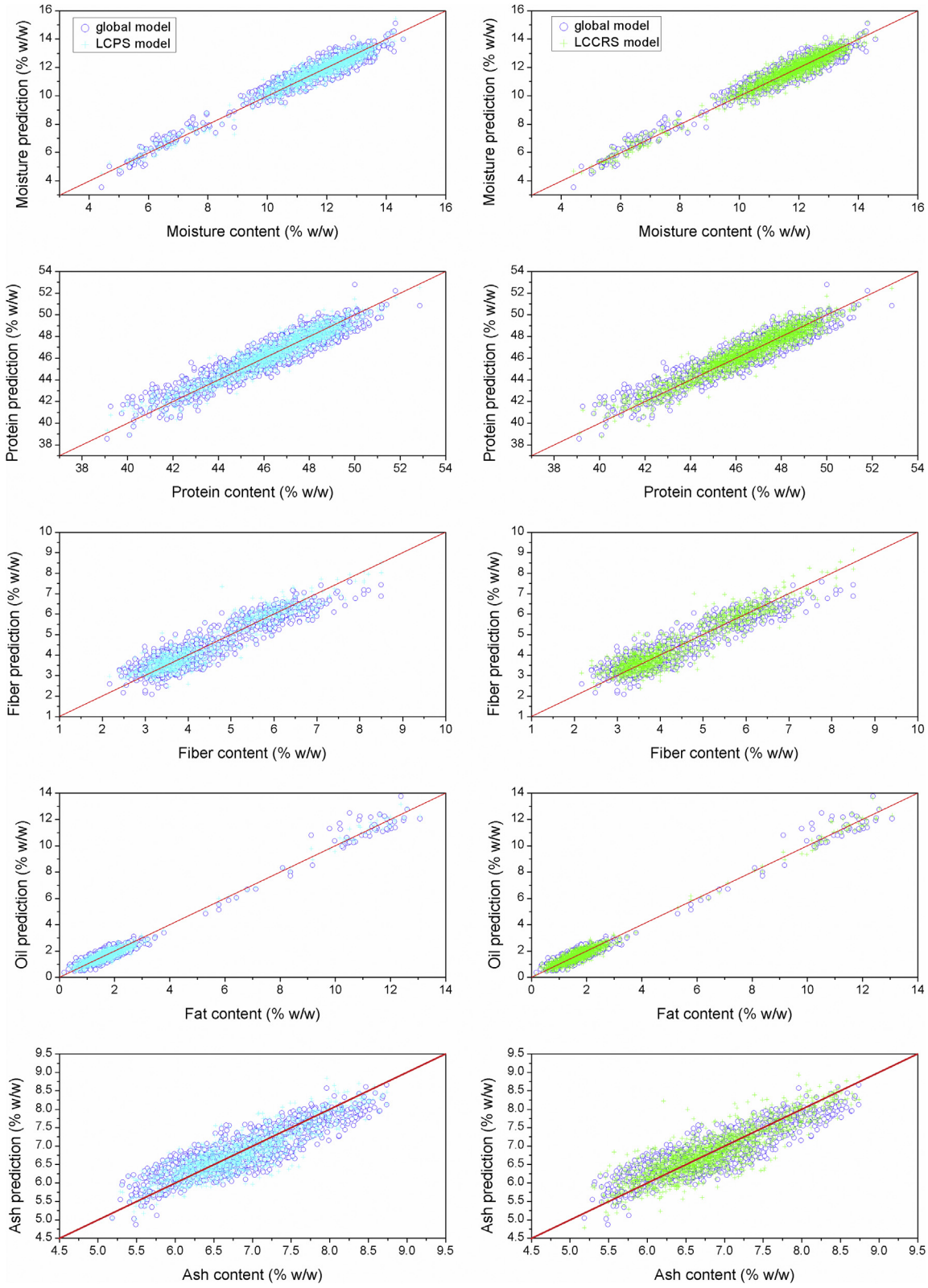


Fig. 7. Predicted vs. reference values for the five parameters and two models.

furnished smaller RMSEP values than both the global and random models. Moreover, although random models were built with the same number of calibration samples as the LCPS model, they presented the worst results.

In order to check the statistical significance of the improvement in the RMSEP values for each local strategy when compared to the global model, a randomization *t*-test [[31], [32]] was applied. When the one-tail version of this test is used, an associated probability *p*

Please cite this article in press as: F. Allegrini, et al., Regression models based on new local strategies for near infrared spectroscopic data, Analytica Chimica Acta (2016), http://dx.doi.org/10.1016/j.aca.2016.07.006

smaller than the critical value of 0.05 means that the RMSEP of one of the methods is significantly lower than the RMSEP obtained from the other method being compared. This was the case for the five parameters when either LCPS or LCCRS were compared to the global calibration. On the other side, if p exceeds the critical value means that no significant differences were found. This is the case when LCPS and LCCRS are compared between each other. Finally the same test was used to verify that the random local models lead to worse predictions than the global model.

4. Conclusion

In this study, two different local strategies based on the PLS algorithm, were carefully investigated in the context of the quantification of five parameters in corn seeds. These strategies showed to be an efficient alternative to optimize predictions, when compared to global models, allowing a statistically significant reduction of the RMSEP without need of preprocessing methods. In addition to this feature, the simplicity and speed of the algorithms developed, based on Mahalanobis distance measures in PLS scores space, allows their application to on-line predictions.

Finally, the following perspectives of the present work can be outlined: (1) extension of the proposed methodologies to the prediction of more than one product from a unique and large data set, i.e., as there is a local selection of the most spectroscopically similar samples, the spectral library can be multi-products, which can also drive to the development of unique predictions with consequent savings in time and effort required to develop and maintain individual calibration models and (2) design an algorithm-fast method to find from an interval of percentiles, the optimal value according to each new test sample to be predicted.

Acknowledgments

F. Allegrini and A. C. Olivieri thank Universidad Nacional de Rosario, CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas) and ANPCyT (Agencia Nacional de Promoción Científica y Tecnológica, Project No. PICT-2013-0136). F.A. thanks CONICET for a postdoctoral fellowship.

W. D. Fragoso thanks Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), process number 305760/2014-9 and special thanks to brazilian program “Ciência sem Fronteiras”, process number 203474/2014-7.

J. A. Fernández Pierna, V. Baeten and P. Dardenne thank B. Lecler, O. Minet and the technical staff from the Valorisation of Agricultural Products Department of the CRA-W in Belgium for suplying the data.

References

[1] C.W. Chang, D.A. Laird, M.J. Mausbach, C.R. Hurburgh Jr., Near-infrared reflectance spectroscopy – principal components regression analyses of soil properties, Soil Sci. Soc. Am. J. 65 (2001) 480–490.
[2] S.R. Araújo, J. Wetterlind, J.A.M. Demattê, B. Stenberg, Improving the prediction performance of a large tropical vis-NIR spectroscopic soil library from Brazil by clustering into smaller subsets or use of data mining calibration techniques, Eur. J. Soil Sci. 65 (2014) 718–729.
[3] F. Gogé, R. Joffre, C. Jolivet, I. Ross, L. Ranjard, Optimization criteria in sample selection step of local regression for quantitative analysis of large soil NIRS database, Chemom. Intell. Lab. Syst. 110 (2012) 168–176.
[4] J.C.L. Alves, R.J. Poppi, Simultaneous determination of hydrocarbon renewable diesel, biodiesel and petroleum diesel contents in diesel fuel blends using near infrared (NIR) spectroscopy and chemometrics, Analyst 138 (2013) 6477–6487.
[5] F.V.C. de Vasconcelos, P.F.B. de Souza, M.F. Pimentel, M.J.C. Pontes, C.F. Pereira, Using near-infrared overtone regions to determine biodiesel content and adulteration of diesel/biodiesel blends with vegetable oils, Anal. Chim. Acta 716 (2012) 101–107.
[6] T.B. Bagchi, S. Sharma, K. Chattopadhyay, Development of NIRS models to predict protein and amylose content of brown rice and proximate compositions of rice bran, Food Chem. 191 (2016) 21–27.
[7] S. Wright, L. Hagen, Oleic acid content in ground corn by NIR spectroscopy with an indirect calibration method, J. Am. Oil Chemists' Soc. 80 (2003) 1163–1167.
[8] I.V. Kovalenko, G.R. Rippke, C.R. Hurburgh, Measurement of soybean fatty acids by near-infrared spectroscopy; linear and nonlinear calibration methods, J. Am. Oil Chemists' Soc. 83 (2006) 421–427.
[9] B. Kuang, A.M. Mouazen, Influence of the number of samples on prediction error of visible and near infrared spectroscopy of selected soil properties at the farm scale, Eur. J. Soil Sci. 63 (2012) 421–429.
[10] W.S. Cleveland, S.J. Devlin, E. Grosse, Regression by local fitting, J. Econ. 37 (1988) 87–114.
[11] T. Fearn, A.M.C. Davies, Locally-biased regression, J. Near Infrared Spectrosc. 11 (2003) 467–478.
[12] D.A. Burns, E.W. Ciurczak, Handbook of near-infrared analysis, in: Practical Spectroscopy Series, third ed., vol. 40, CRC Press, Boca Raton, USA, 2008.
[13] F.E. Barton, J.S. Shenk, M.O. Westerhaus, D.B. Funk, The development of near infrared wheat quality models by locally weighted regressions, J. Near Infrared Spectrosc. 8 (2000) 201–208.
[14] J.S. Shenk, P. Berzaghi, M.O. Westerhaus, Investigation of a LOCAL calibration procedure for near infrared instruments, J. Near Infrared Spectrosc. 5 (1997) 223–232.
[15] P. Berzaghi, J.S. Shenk, M.O. Westerhaus, LOCAL prediction with near infrared multi-product databases, J. Near Infrared Spectrosc. 8 (2000) 1–9.
[16] A.M.C. Davies, H.V. Britcher, J.G. Franklin, S.M. Ring, A. Grant, W.F. McClure, The application of Fourier-transformed near-infrared spectra to quantitative analysis by comparison of similarity indexes (CARNAC), Mikrochim. Acta 1 (1988) 61–64.
[17] S. Navea, R. Tauler, A. de Juan, Application of the local regression method interval partial least-squares to the elucidation of protein secondary structure, Anal. Biochem. 15 336 (2) (2005) 231–242.
[18] T. Öberg, T. Liu, Global and local PLS regression models to predict vapor pressure, QSAR Comb. Sci. 27 (3) (2008) 273–279.
[19] X. Tian, Z. Wu, E.S. Chang, Local partial least square regression for spectral mapping in voice conversion, in: Signal and Information Processing Association Annual Summit and Conference (APSIPA), Asia-Pacific, Kaohsiung, 2013.
[20] N. Ruiz, Near Infrared Spectroscopy: present and future applications, Tech. Bull. Am. Soybean Assoc. FT52 (2001).
[21] R.G. Whitfield, M.E. Gerger, R.L. Sharp, Near-infrared spectrum qualification via Mahalanobis distance determination, Appl. Spectrosc. 41 (7) (1987) 1204–1213.
[22] T.M. Cover, P.E. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. theory, IT 13 (1) (1967) 21–27.
[23] D. Coomans, D.L. Massart, Alternative k-nearest neighbor rules in supervised pattern recognition. Part 2. Probabilistic classification on the basis of the kNN method modified for direct density estimation, Anal. Chim. Acta 138 (1982) 153–165.
[24] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: E. Simoudis, J. Han, U.M. Fayyad (Eds.), Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, AAAI, Menlo Park, CA, 1996, pp. 226–231.
[25] M. Daszykowski, B. Walczak, D.L. Massart, Looking for natural patterns in data Part 1. Density based approach, Chemom. Intell. Lab. Syst. 56 (56) (2001) 83–92.
[26] M. Ankerst, M.M. Breunig, H.-P. Kriegel, J. Sander, OPTICS: ordering points to identify the clustering structure, in: A. Delis, C. Faloutsos, S. Ghandeharizadeh (Eds.), Proceedings of ACM SIGMOD International Conference on Management of Data Philadelphia, PA, ACM, New York, 1999, pp. 49–60.
[27] M. Daszykowski, B. Walczak, D.L. Massart, Looking for natural patterns in analytical data. 2. Tracing local density with OPTICS, J. Chem. Inf. Comput. Sci. 42 (2002) 500–507.
[28] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137–148.
[29] H. Martens, T. Naes, Multivariate calibration II. Chemometrics methods, Trends Anal. Chem. 3 (8) (1985) 266–271.
[30] Z. Wang, T. Isaksson, B. Kowalski, New approach for distance measurement in locally weighted regression, Anal. Chem. 66 (1994) 249–260.
[31] H. van der Voet, Comparing the predictive accuracy of models using a simple randomization test, Chemom. Intell. Lab. Syst. 25 (1994) 313–323.
[32] A.C. Olivieri, Practical Guidelines for Reporting Results in a Single- and Multi-Component Analytical Calibration: a Tutorial,, 868, 2015, pp. 10–22.

Please cite this article in press as: F. Allegrini, et al., Regression models based on new local strategies for near infrared spectroscopic data, Analytica Chimica Acta (2016), http://dx.doi.org/10.1016/j.aca.2016.07.006