Received: 18 January 2015 ■ *Revised: 2 May 2016* ■ *Accepted: 3 May 2016* ■ *Publication: 9 May 2016*

JOURNAL OF NEAR INFRARED SPECTROSCOPY

Virtual Issue: Papers Presented at NIR-2015, October 2015, Foz do Iguassu, Brazil

LOCAL regression algorithm improves near infrared spectroscopy predictions when the target constituent evolves in breeding populations

F. Davrieux,[a,*] D. Dufour,[b,e] P. Dardenne,[c] J. Belalcazar,[d] M. Pizarro,[d] J. Luna,[d] L. Londoño,[e] A. Jaramillo,[e] T. Sanchez,[d] N. Morante,[d] F. Calle,[d] L.A. Becerra Lopez-Lavalle [d] and H. Ceballos[d]

[a]Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), UMR Qualisud, St Pierre, 97455, Reunion Island, France. E-mail: davrieux@cirad.fr

[b]Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), UMR Qualisud, Cali, Colombia

[c]Walloon Agricultural Research Centre (CRAW), Gembloux, Belgium

[d]Centro Internacional de Agricultura Tropical (CIAT), Cassava program, Cali, Colombia

[e]Centro Internacional de Agricultura Tropical (CIAT), Harvestplus LAC, Cali, Colombia

The CGIAR Harvest Plus Challenge Program began in the mid-2000s to support the genetic improvement of nutritional quality in various crops, including the carotenoids content of cassava roots. Successful conventional breeding requires a large number of segregating progenies. However, only a few samples can be quantified by high performance liquid chromatography each day for total carotenoids (TCC) and β-carotene (TBC) contents, limiting the gains from breeding. This study describes the usefulness of near infrared (NIR) spectroscopy and the efficiency of a large database coupled to a LOCAL regression algorithm to reach accurate TCC/TBC predictions on fresh cassava roots. The cassava database (6026 samples) was built over six years. TCC values ranged from 0.11 µg g^{-1} to 29.0 µg g^{-1}, whereas TBC ranged from negligible values up to 20.1 µg g^{-1}. All values were measured and expressed on a fresh weight basis. Between 2009 and 2014 increases in TCC and TBC were 86% and 122%, respectively. A comparison of calibrations using partial least squares (PLS) regression and LOCAL regression was done. The standard error of prediction were 1.82 µg g^{-1} for TCC and 1.28 µg g^{-1} for TBC using PLS model and 1.38 µg g^{-1} and 1.02 µg g^{-1}, respectively, using LOCAL regression. The specificity of the data, with increasing content of the constituent of interest year after year, clearly showed the limitation of the classical partial least squares regression approach. The LOCAL regression algorithm takes advantage of large databases; this study highlighted the efficiency of this concept. NIR spectroscopy coupled to LOCAL regression led to efficient models for breeding programmes aiming at increasing carotenoids content in fresh cassava roots. NIR spectroscopy can also be used to predict other important constituents such as dry matter content and cyanogenic glucosides.

Keywords: LOCAL regression, PLS regression, fresh cassava, nutritional quality, breeding, near infrared spectroscopy, carotenoids, crop biofortification

Introduction

Cassava (*Manihot esculenta* Crantz) contributes importantly as a food source for millions of people in developing countries. Cassava is the second most important food staple (in terms of calories consumed) in Sub-Saharan Africa.[1] In addition to being consumed in a fresh form as food, cassava can be used as a source of raw material for the starch, animal feed

ISSN: 0967-0335
doi: 10.1255/jnirs.1213

and ethanol industries.[2] However, cassava presents various drawbacks, such as the relatively low nutritional quality (low proteins, fat, minerals and bioactive compounds[3]) of its roots, which limits the crop to providing only dietary energy.[4] Bioactive compounds may be grouped into different categories[5] among which the carotenoids represent a major interest. Carotenoids are plant pigments[6] with important roles such as photosynthesis, and their antioxidant properties have a positive impact in human health preventing heart disease and cancer.[7] Carotenoids are isoprenoid compounds biosynthesised from geranyl pyrophosphate and are precursors to vitamin A. β-carotene (Figure 1) is the carotenoid most efficiently converted into vitamin A. The long chain of alternating double bonds (conjugated) is responsible for the yellow orange colour of β-carotene. Carotenoids are extremely hydrophobic molecules with little or no solubility in water, so they are extracted and/or separated with non-polar solvents such as hexane.

The Harvest Plus Challenge Program from the Consultative Group of International Agriculture Research (CGIAR) began in the mid-2000s to support genetic improvement of nutritional quality in various crops, including carotenoids content of cassava roots.[8]

Over this period, cassava biofortification has led to a threefold increase of the original concentration of carotenoids in cassava roots.[9] Successful conventional breeding requires large numbers of segregating progenies. However, only a few samples can be quantified by HPLC each day for total carotenoids (TCC) and β-carotene (TBC) contents, limiting the gains from breeding. Moreover, the sensitivity of carotenoids to heat, light and oxygen[10,11] and postharvest physiological deterioration makes the storage of fresh roots difficult or impossible.

Near infrared (NIR) spectroscopy, which has found many applications in research involving plant tissues, offers a solution to this problem. The review of NIR spectroscopy applications related to bioactive compounds, including carotenoids quantification, done by M. McGoverin[5] *et al.* inventoried studies on carotenoids in various crops[12–15] and one related to carotenoids in potato.[16] Other works aimed at predicting functional properties in tropical root and tuber crops.[17–19] Few studies have been conducted on nutritional properties of fresh tubers or roots using NIR spectroscopy. A previous study[20] demonstrated the efficiency of NIR spectroscopy for predicting TCC, TBC and dry matter content (DMC) in fresh cassava roots. The models were based on partial least squares (PLS) regression.[21] In fact, to be effective a global calibration (based on all available samples)[22] must take into consideration the full range of expected values. This includes the variability of the independent variables (spectra) and of dependent variables (range of the constituent to be predicted). PLS regression ranged within the linear methods which assume that the relationship between the independent and dependent variables are linear in nature.[23] However, predictions of a new harvest based on the PLS models were actually "extrapolations" because, year after year, cassava genotypes with higher carotenoids content were obtained by the breeding project. This resulted in a non-linear response and a sub-determination for the highest contents. There are several potential sources of non-linearity:[24] the instrument used, the chemical nature of the target constituent,[25] high moisture content of the products[23] or when new samples to be predicted fall outside the calibration domain.[26]

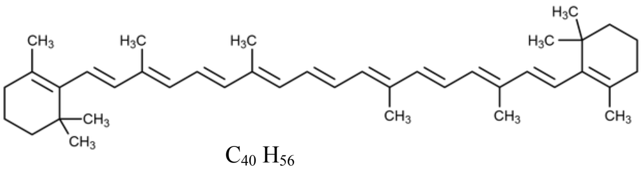

Figure 1. Developed formula of β-carotene

To overcome non-linearity problems different strategies have been considered, and in particular the application of non-linear methods.[24] Local approaches,[27] based on a selection of a specific training set for each new single sample to fit a "virtual" model (PLS regression) and to perform a prediction, have been successfully applied for different products[28–31] and heterogeneous databases.[27] The present study assesses the efficiency of the LOCAL regression proposed by Schenk *et al.*,[32,33] compared to global PLS regression, to cope with the non-linearity due to extrapolations in predictions of TCC and TBC in fresh cassava roots. This approach has allowed selecting the best fresh cassava root samples based on more precise NIR spectroscopy predictions of carotenoids contents.

Materials and methods

Samples and reference analysis

Cassava plants were produced and grown by the International Center for Tropical Agriculture (CIAT) cassava breeding programme in Palmira, Colombia. Samples came from a rapid cycling recurrent selection that has been implemented to develop biofortified cassava.[9] Data for the present study was generated over a six-year period. During this time, 6026 samples were collected and analysed for their NIR spectra (the number of samples are given in parenthesis): 2009 (651), 2010 (664), 2011 (707), 2012 (1372), 2013 (1762) and 2014 (870). Samples were analysed for their DMC, TCC and TBC contents.[9,20] Quantifications were made on fresh root tissue, not lyophilised or stored. The numbers of reference values per year and per constituent are reported in Table 1.

Harvest took place at sunrise and NIR analysis, dry matter quantification and carotenoid extraction were done in the morning. Root samples and extracts were protected from the light.[34] High-performance liquid chromatography (HPLC) quantifications were carried out in the afternoon.

Root harvest and processing

Harvesting was done by hand, pulling the plant out of the ground. For each genotype, two to three commercial-size roots were taken to the lab where they were washed, peeled and homogenised with a food processor into a homogenous

Table 1. Number of reference analyses per year.

	Constituent		
Year	DM	TCC	TBC
2009	650	572	572
2010	645	641	640
2011	693	605	611
2012	1369	1264	1270
2013	1758	732	732
2014	463	463	463
Total	5578	4277	4288

paste. Further analyses were made using aliquots from this homogeneous paste.

Dry matter content

A subsample from the root material was taken for the quantification of DMC. Two samples of the ground root tissue were weighed, dried in an oven at 105°C for 24 h and then weighed again. Dry matter was expressed as the percentage of dry weight relative to fresh weight.

Carotenoid extraction and quantification

Carotenoids were extracted following the protocol described elsewhere.[20] Separation and quantification of (TCC) and total β-carotene (TBC) were achieved using a YMC Carotenoid S-5 C30 reversed-phase column (4.6 mm × 150 mm: particle size, 5 µm), with a YMC Carotenoid S-5 guard column (4.0 mm × 23 mm) in a HPLC system (Agilent Technologies 1200 series, Waldbronn, Germany), using a diode array detector with the wavelength set at 450 nm.

NIR spectroscopy

Fresh root samples were ground with a food processor, as described above, prior to NIR spectroscopy analysis. Each sample was duplicated. Therefore, spectra from two root subsamples were obtained per genotype. Each of these two samples was measured once. Approximately 8 g of ground root tissue were placed in NIR spectroscopy capsules for analysis using a FOSS 6500 monochromator with autocup sampling module (FOSS, Hillerod, Denmark). All spectra were recorded from 400 nm to 2498 nm at 2 nm intervals and saved as the average of 32 scans. Further analyses were made on the average of the two spectra available per genotype.

Data analysis

Data and statistical analyses were performed using Win-ISI 4.6 software (Infrasoft International and FOSS, Hillerod, Denmark) and XLstat software (Addinsoft, Paris, France). Different pretreatments were initially tested and the one selected was a mathematical correction for light scattering using the standard normal variate and de-trend (SNVD) correction. Then, depending on the constituent studied, log (1/*R*) or the second derivative calculated on five data points and smoothed using Savitzky–Golay polynomial smoothing on five data points based on full (visible and NIR) or reduced wavelength range (NIR only) were used for calibrations. The PLS [WinISI Modified-PLS (MPLS) algorithm] and LOCAL regressions were used to develop prediction models. The specific factors for each PLS or LOCAL model were optimised according to WinISI 4.6 software. Calibration statistics included the following parameters: standard deviation (*SD*), coefficient of determination (R^2 for calibration models and r^2 for validation), standard error of calibration (*SEC*), standard error of cross-validation (*SECV*) and standard error of prediction (*SEP*). For MPLS models, cross-validation was used during calibration development in order to select the optimum number of latent variables and to minimise overfitting the equations. The LOCAL procedure[32] is designed to search and select *n* samples similar to the sample to predict. The *n* samples are then used to develop a model (based on PLS regression) specific to the sample being analysed. The similarity index, used to select samples, is the correlation coefficient between the spectrum of the unknown sample and the spectra from the database. Two calibration parameters were optimised for LOCAL regression models: the maximum number of samples to select and the number of PLS terms. This was achieved by testing all the combinations with 50, 100, 150, 200, 250, 300 and 350 samples and with a minimum of 4 and a maximum of 13 PLS terms. The best treatments in the MPLS calibration step were assumed to be the best for LOCAL. Therefore the same segments, same treatments and same number of elimination passes were applied to both PLS and LOCAL models. The Student (t) test was used to identify t-outlier samples during calibration development. Outlier detection was based on the standardised residuals (= error/*SECV*) with a cutoff of 2.5. Three passes of outlier elimination were used.

The standard errors of laboratory (*SEL*) were estimated for DMC, TCC and TBC using seven different genotypes, with three to five replicates per genotype per parameter. The *SEL* were calculated as the pooled standard deviations.

Results and discussion

TCC values (*n* = 4277) ranged from 0.11 µg g^{-1} to 29.0 µg g^{-1} with an average value of 11.6 µg g^{-1}. TBC values (*n* = 4288), on the other hand, ranged from negligible to 20.1 µg g^{-1} with an average value of 6.9 µg g^{-1}. The standard deviation (*SD*) was 5.1 µg g^{-1} for TCC and 3.6 µg g^{-1} for TBC. The DMC values (*n* = 5578) ranged between 12.3% and 52.4% with a *SD* of 5.9%. The descriptive statistics per constituent and year are reported in Table 2. These figures highlight the efficiency of the breeding process between 2009 and 2014. The increases in TCC and TBC were 86% and 122%, respectively. The average DMC content was constant for the same period, with an overall average of 33.8%. This observation is relevant because it would demonstrate that TCC and TBC contents can be increased without any reduction in DMC. Many African programmes currently developing biofortified cassava face a problem of

Table 2. Descriptive statistics per constituent and per year.

	Year	*N*	Minimum	Maximum	Mean	*SD*
DM (%)	2009	650	13.96	52.44	30.39	6.23
	2010	645	12.29	46.70	28.94	6.18
	2011	693	13.88	45.99	35.52	5.03
	2012	1369	20.01	49.42	37.27	4.36
	2013	1758	15.77	45.74	32.79	5.15
	2014	463	21.1	52.3	37.0	3.8
TCC (μg g^{-1}) of fresh weight	2009	572	2.0	19.4	9.1	2.8
	2010	641	0.1	24.7	10.0	4.3
	2011	605	0.1	25.8	10.7	6.1
	2012	1264	0.4	24.3	11.2	4.4
	2013	732	0.6	29.0	13.0	5.8
	2014	463	6.0	25.8	17.0	2.9
TBC (μg g^{-1}) of fresh weight	2009	572	0.7	13.4	5.1	2.3
	2010	640	0.0	14.2	6.3	2.8
	2011	611	0.0	15.0	5.7	3.8
	2012	1270	0.2	16.6	6.5	2.9
	2013	732	0.2	20.1	8.1	4.1
	2014	463	2.8	19.0	11.3	2.5

lower than desirable levels of DMC. The impact of the selection programme has been significant, particularly after 2012, with an increase of 2 points in 2013 and 4 points in 2014 for TCC and TBC (Figure 2), respectively. The trends for increased TCC and TBC average values over years were linear with R^2 (average contents versus years) of 0.85 for TCC and 0.77 for TBC. These results highlight the efficiency of the conventional breeding programme.

In a previous study[20] we demonstrated that PLS regressions based on 2009–2011 samples were judged efficient for DMC, TCC and TBC contents. The *SEC* and *SECV* were close to *SEP* when predicting 2012 samples. For example, the *SECV* for TCC was 1.48 μg g^{-1} (calibration 2009–2011), and the *SEP* estimated on 2012 samples was 1.81 μg g^{-1}, with no significant bias. However, this apparent robustness of the models did not take into account the remaining source of variation due to the ongoing breeding progress (development of new genotypes with higher carotenoid content). In fact, the evolution of average content (TCC and TBC) continued increasing after 2012 (Figure 2). The gains in TCC and TBC were accentuated after 2012. Samples from 2014 had a lower dispersion than for 2013 (*SD* for TTC values of samples from 2014 was equal

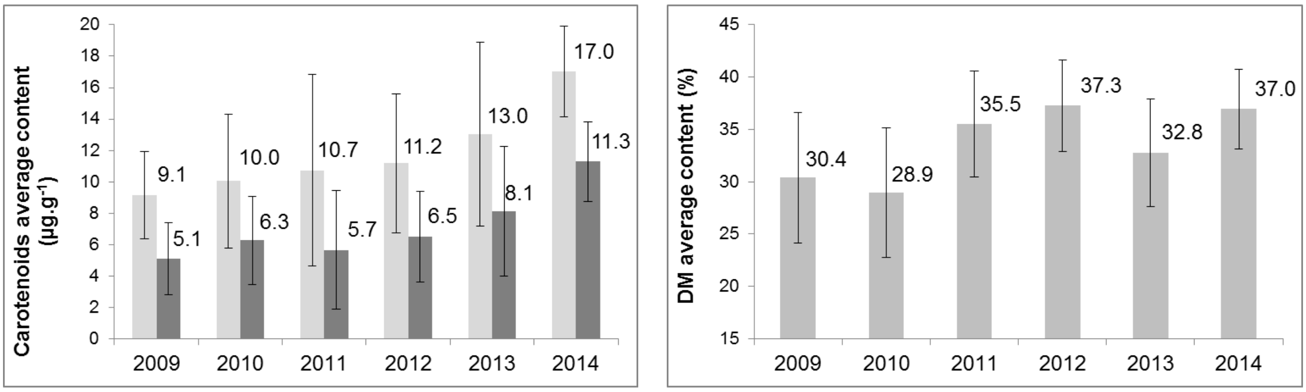

Figure 2. Evolution of average contents over years for TCC (grey bars) and TBC (black bars) left graphic and DM (grey bars) right graphic.

Table 3. Performances of PLS regression models based on 2009–2012 data.

	Constituent	*N*	Mean	*SD*	*SEC*	R^2	*SECV* (μg g-1)	PLS terms
Calibration (2009–2012)	TCC (μg g−1)	2980	10.39	4.45	1.36	0.91	1.43	12
	TBC (μg g−1)	2955	5.94	2.91	0.81	0.92	0.86	9
	DM (%)	3086	34.32	6.14	1.28	0.96	1.49	8
Validation (2013)	Constituent	*N*	*SEP*	Bias	Slope	SEP_c	r^2	
	TCC (μg g−1)	732	2.63	1.00	1.26	2.44	0.86	
	TBC (μg g−1)	732	1.75	0.60	1.64	1.33	0.90	
	DM (%)	1758	1.20	-0.38	1.11	1.14	0.96	

Table 4. Performances of PLS versus LOCAL. Models using samples from 2009–2012 and predictions of samples from 2013.

	PLS		LOCAL PLS R	
Constituent	r^2	SEP	r^2	SEP
TCC (μg g^{-1})	0.863	2.63	0.941	1.59
TBC (μg g^{-1})	0.898	1.75	0.947	1.11

to 2.9 μg g^{-1}), which meant higher content for almost all of the genotypes.

In order to compare the two approaches of global PLS and LOCAL regressions, calibrations based on these two algorithms were developed for DMC, TCC and TBC using samples from the 2009–2012 collection. The different models were compared using the *SEP* estimated when predicting the samples from 2013.

The PLS regression performances (Table 3) were similar to those from the previous study[20] in terms of *SEC* and R^2 (e.g. $SEC_{2009–2011}$ was 1.38 μg g^{-1} for TCC and $SEC_{2009–2012}$ was 1.36 μg g^{-1}). But the models clearly failed to properly predict samples from 2013 for TCC and TBC. The result of this was a non-linear fitting for the higher contents as illustrated for 2013 TCC values in Figure 3A. In this case, the linear fitting led to r^2 of 0.86 while a quadratic fitting resulted in a r^2 of 0.90. The same pattern was observed for TBC values with r^2 = 0.89 for linear fitting and r^2 = 0.93 for the quadratic. The PLS model based on 2009–2012 DMC values was efficient to predict the 2013 and 2014 samples with an *SEP* = 1.20% and r^2 = 0.87. This made sense as DMC remained more or less unchanged over the years.

The LOCAL regressions (LR), based on samples from 2009 to 2012, were first optimised for the minimum and maximum numbers of samples and PLS terms to be used: for both TCC and TBC the minimum number of samples was 50 and the maximum, 300.

The LR was not investigated for DMC, as the PLS model was efficient with no bias and no non-linearity. The performances of LR, expressed as *SEP* and r^2 when predicting 2013 samples, were slightly better than PLS regression (Table 4). The non-linear effect was largely corrected by LR (Figure 3B), there was no significant difference in terms of r^2 between linear and second order fitting. However, the remaining errors are still high and similar. In both cases extreme values (especially high values) were underestimated. This lack of gain using LR could be due to the gap observed in TCC and TBC content after 2012, which resulted in too few samples being available with high TCC and TBC for the 2009–2012 period. Actually, for TCC > 18 μg g^{-1}, 56 samples were analysed during the 2009–2012 period, while 177 samples were analysed in 2013.

A new comparison was done using the 2009–2013 samples for calibration (PLS and LOCAL regressions) and the samples harvested in 2014 for validation (*n* = 463). For TCC, the *SEP* was 1.82 μg g^{-1} with PLS regression (R^2 = 0.64) while LOCAL regres-

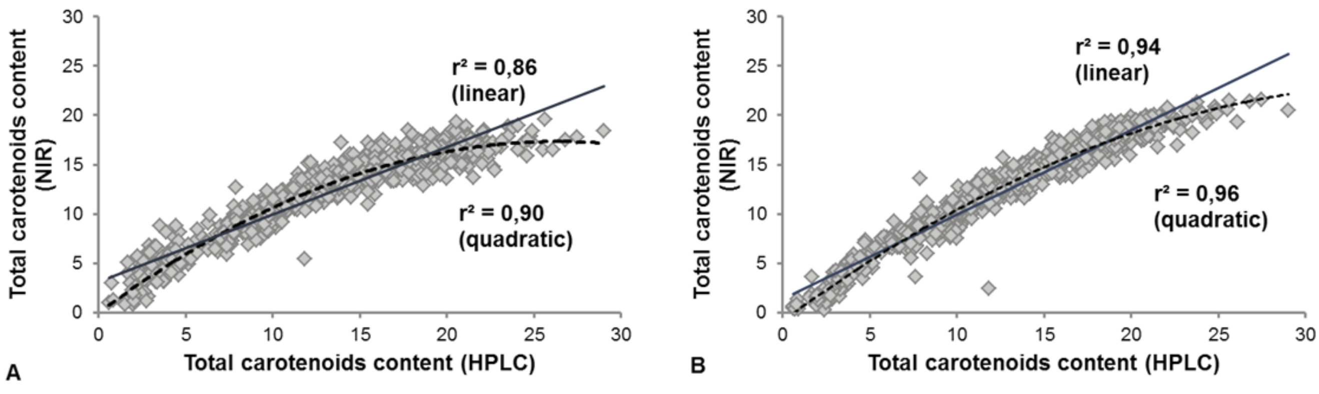

Figure 3. Scatter plots of TCC NIR spectroscopy values versus TCC HPLC values of year 2013 A) using PLS model and B) LOCAL regression, based on 2009–2012 samples.

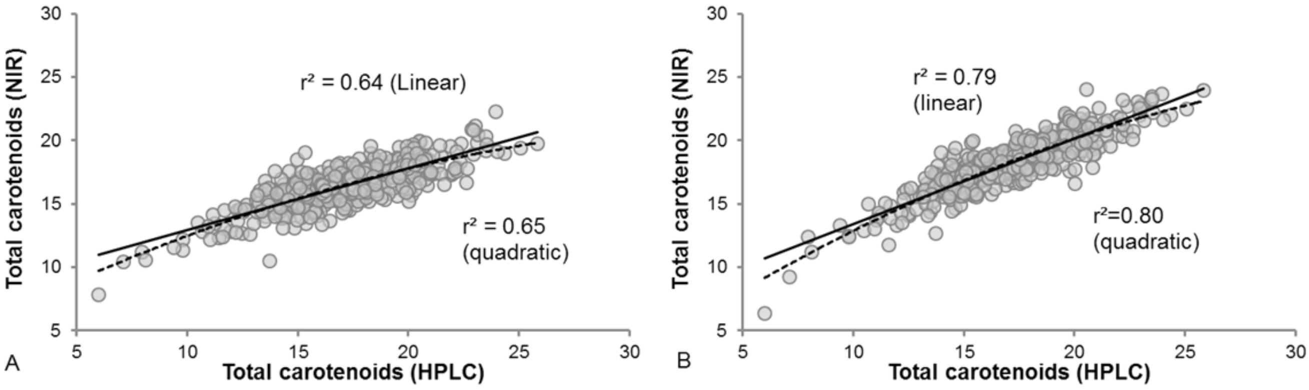

Figure 4. Scatter plots of TCC NIR spectroscopy values versus TCC HPLC values of year 2014 A) using PLS model and B) LOCAL regression, based on 2009–2013 samples.

sion led to $SEP = 1.38\,\mu g\,g^{-1}$ with $R^2 = 0.79$. For TBC, the *SEP* was $1.28\,\mu g\,g^{-1}$ with PLS regression ($r^2 = 0.78$), while LOCAL regression resulted in a $SEP = 1.02\,\mu g\,g^{-1}$ with $r^2 = 0.84$. The LOCAL regression, therefore, improved the *SEP* for both constituents. Figures 4 A and B illustrate the performance of linear and quadratic models based on the PLS and LOCAL predictions.

The comparison of the three data sets, HPLC data, and PLS and LOCAL predicted values, was done through a one-way ANOVA. Type of value was used as a factor. The ANOVA was completed by a Dunnett's test[35] (unilateral on the right with 95% confidence and HPLC as control). The null hypothesis (H_0) was that the average of the control (HPLC values) was not different from the averages of the two NIR prediction approaches, the alternative hypothesis (Ha) was that the mean control was higher than the tested modality. The ANOVA resulted in a significant effect of the factor (type of values). The Dunnett's test, however, allowed a conclusion, with 95% confidence, to accept H_0 when comparing HPLC and LOCAL regression values (no significant difference). On the other hand, H_0 was rejected when comparing HPLC and PLS values through the same test. In other words, the average PLS predicted value was significantly lower than the respective HPLC data.

The results showed an improvement of the *SEP* when the LOCAL regression was applied, but the errors ($1.38\,\mu g\,g^{-1}$ for TCC and $1.02\,\mu g\,g^{-1}$ for TBC) are still high compared to the *SEL*: respectively, $0.59\,\mu g\,g^{-1}$ and $0.44\,\mu g\,g^{-1}$. A closer look at the data highlighted that a large part of the error for the LOCAL model was due to low content values, this was illustrated by the adjusted normal laws (Figure 5). For TCC, LOCAL predictions and HPLC values showed excellent agreement on the right tails, contrary to what can be observed for the left tails. On the other hand, PLS failed to properly predict values on both ends of the distribution curve. The HPLC and LOCAL curves were similar with means, respectively, equal to $17.0\,\mu g\,g^{-1}$ and $18.1\,\mu g\,g^{-1}$ and standard deviations, respectively, equal to $2.89\,\mu g\,g^{-1}$ and $2.21\,\mu g\,g^{-1}$. The higher *SD* for HPLC values was due to a higher spread for low TCC. The PLS curve was characterised by a lower standard deviation ($SD = 1.76\,\mu g\,g^{-1}$) and an average value equal to $16.4\,\mu g\,g^{-1}$ close to the HPLC average value. This means that PLS models predicted the average value well.

The efficiency of the LOCAL regression model was tested for TCC by counting the number of missed samples, that is to say when the prediction was lower than the HPLC values (Table 5). Within the 20 samples missed by LOCAL when HPLC was $\geqslant 20.0\,\mu g\,g^{-1}$ only one sample was predicted considerably lower ($16.6\,\mu g\,g^{-1}$). Similarly, for the 11 samples missed when HPLC was $\geqslant 22.0\,\mu g\,g^{-1}$, seven samples were higher than $21.2\,\mu g\,g^{-1}$ and only four had predicted values ranging between $20.1\,\mu g\,g^{-1}$ and $22.1\,\mu g\,g^{-1}$. On the other hand, LOCAL regression led to 26 false positives, i.e. predicted $\geqslant 20.0\,\mu g\,g^{-1}$ when the HPLC value was $<20.0\,\mu g\,g^{-1}$. The HPLC values for these 26 samples ranged between $17.0\,\mu g\,g^{-1}$ and $20.0\,\mu g\,g^{-1}$, with 11 samples between $17.0\,\mu g\,g^{-1}$ and $19.0\,\mu g\,g^{-1}$ and the remaining 15 samples between $19.0\,\mu g\,g^{-1}$ and $20.0\,\mu g\,g^{-1}$. This result is of interest in the frame of a breeding programme, the LOCAL procedure

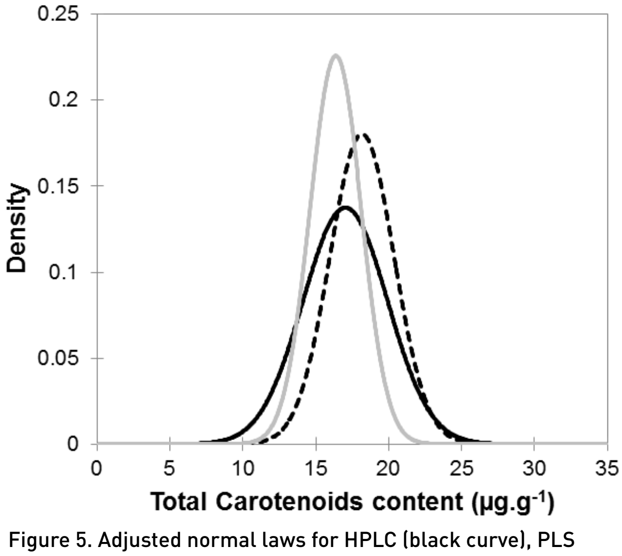

Figure 5. Adjusted normal laws for HPLC (black curve), PLS (grey curve) and LOCAL (black dashed curve) values. Parameters estimated (Xlstat software).

Table 5. Number of samples missed by PLS and LOCAL models according to threshold of TTC values (HPLC).

Threshold of TTC values	Number of samples HPLC values	Number of missed samples	
		PLS regression(<)	LOCAL regression (<)
⩾ 15 μg.g^{-1}	357	32	0
⩾ 18 μg.g^{-1}	163	102	11
⩾ 20 μg.g^{-1}	77	71	20
⩾ 22 μg.g^{-1}	23	22	11

allowed a successful selection of genotypes according to TCC or TBC. The comparison of performances between PLS and LOCAL is illustrated by the graphic presented in Figure 6. In this plot, data are ordered for TCC values ⩾18 μg g^{-1} (HPLC, PLS and LOCAL). This graphic clearly indicated that PLS regression based on the whole data set failed to predict new samples with high TCC values; the predictions were systematically lower than the actual values. Moreover, 67.5% of the TCC values predicted using LOCAL regression fell inside the 95% confidence interval, calculated as ±2 × *SEL*, for each individual HPLC TCC value (Figure 7).

The PLS and LOCAL models presented higher performances for TCC and TBC when the whole segment (visible and IR) was used, while for DMC the best models were with IR only. The b coefficients of the models showed that high coefficients were located in the visible part of the spectrum (data not shown). This is not surprising as carotenoids are pigments coloured yellow, orange or red with a conjugated double-bond system (Figure 1). In this kind of system the π-electrons are highly delocalised and their excited state is of comparatively low energy.[6] Therefore the energy required for excitation is relatively low and corresponds to the visible part of the electromagnetic spectrum (400–500 nm). Models for carotenoids were in accordance with the literature in terms of performance[5] and the use of the whole spectral region (visible and NIR). However, models developed without the NIR range were slightly less efficient than models based on the whole spectrum. As an example, the *SEP* was 1.38 μg g^{-1} when predicting TCC for 2014 samples using a full spectrum LOCAL model (based on 2009–2013 samples), whereas it was 1.43 μg g^{-1} using only the "visible" spectrum, other things being equal. This result confirmed that a part of the information remains in the NIR range of the spectra and, as suggested M. McGoverin *et al.*,[5] the investigation of indirect correlations, with proteins, starch or carbohydrates will help to better understand the models and improve them. A model based solely on the NIR range of the spectrum was obviously less accurate (*SEP* = 2.58 μg g^{-1}, when predicting TCC of 2014 samples using LOCAL with 2009–2013 samples). However, the PLS loadings, as underlined in the previous study,[20] presented for the first and second terms high coefficients at 2284 nm and 1148 nm (assigned, respectively, to starch and conjugated double-bond systems).

The results of this study were in accordance with the results obtained by G. Sinnaeve *et al.*[28] The LR can exploit successfully non-linearity by using a restricted range of the parameter (*Y* variation). Moreover, the LR method, based on PLS regression, led to an efficient localisation in both domains (spectrum and property) as was demonstrated by R.S. Andersen *et al.*[36]

Further investigations can be done by testing the gain in accuracy and robustness when using non-linear methods, such as artificial neural networks, least squares support vector machines or comparison analysis using restructured near infrared and constituent data (CARNAC).[24]

The next step is to test the accuracy and stability of the models with the genotypes grown in 2015, in particular the ability to select the best genotypes according to TCC and TBC.

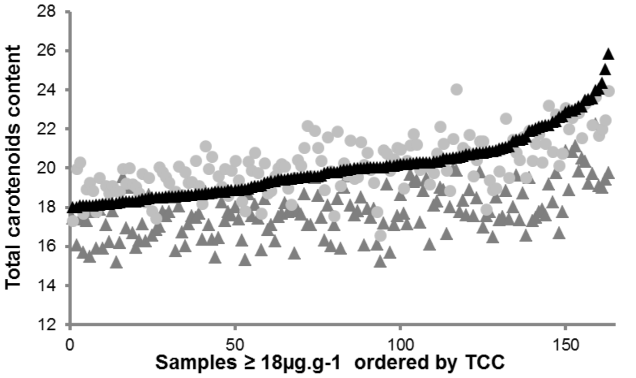

Figure 6. Ordered TCC values higher than 18 μg g^{-1}, for HPLC (black triangles), PLS (grey triangles) and LOCAL (grey circles).

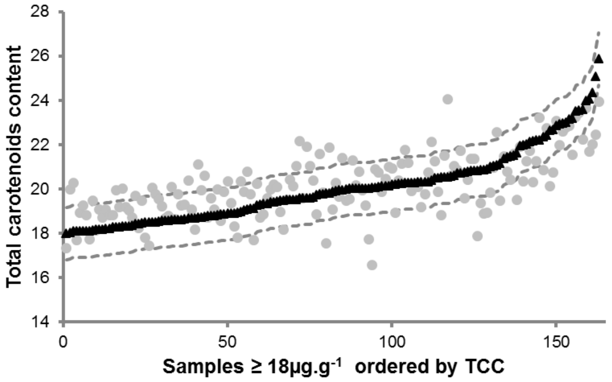

Figure 7. Ordered TCC HPLC values (black triangles) higher than 18.0 μg g^{-1} with 95% confidence interval and corresponding TCC LOCAL predicted values (grey circles).

Conclusion

The specificity of the data, with increasing content of the constituent of interest year after year, clearly illustrated the limitation of a classical PLS regression approach, which resulted in a non-linear fitting for the highest contents. The LOCAL regression algorithm takes advantage of large databases (thousands of samples) of spectra and reference values. This study highlighted the efficiency of this concept which led to models that were able to manage the non-linearity observed with PLS regression for samples with a high constituent content. The single sample prediction concept provided the highest level of accuracy.

After five years of harvest and database building, NIR spectroscopy coupled to LOCAL regression led to more efficient models for breeding programmes aiming at increasing carotenoid content in fresh cassava roots.

Acknowledgements

CIRAD and CIAT authors gratefully acknowledge the financial support of HarvestPlus (www.harvestplus.org), part of the Consultative Group on International Agriculture Research (CGIAR), Program on Agriculture for Nutrition and Health (A4NH), project: "Enhancing the nutritional quality of cassava roots to improve the livelihoods of farmers in marginal agriculture land in Africa, Haiti, and North-Colombia and the CGIAR Research Program on Roots, Tubers and Bananas (RTB)", www.rtb.cgiar.org, project: "Driving livelihood improvements through demand-oriented interventions for competitive production and processing of RTBs".

References

1. G. Tarawali, C. Iyangbe, U.E. Udensi, P. Ilona, T. Osun, C. Okater and G.N. Asumugha, "Commercial-scale adoption of improved cassava varieties: a baseline study to highlight constraints of large-scale cassava based agroprocessing industries in Southern Nigeria", *J. Food Agric. Environ.* **10,** 689 (2012).
2. H. Ceballos, C. Hershey and L.A. Becerra-López-Lavalle, "New approaches to cassava breeding", *Plant Breed. Rev.* **36,** 427 (2012). doi: http://dx.doi.org/10.1002/9781118358566.ch6
3. R.E. Wildman, R. Wildman and T.C. Wallace, *Handbook of Nutraceuticals and Functional Foods*. CRC Press (2006). doi: http://dx.doi.org/10.1201/9781420006186
4. J.A. Montagnac, C.R. Davis and S.A. Tanumihardjo, "Nutritional value of cassava for use as a staple food and recent advances for improvement", *Comp. Rev. Food Sci. Food Safe.* **8,** 181 (2009). doi: http://dx.doi.org/10.1111/j.1541-4337.2009.00077.x
5. C. McGoverin, J. Weeranantanaphan, G. Downey and M. Manley, "Review: The application of near infrared spectroscopy to the measurement of bioactive compounds in food commodities", *J. Near Infrared Spectrosc.* **18,** 87 (2010). doi: http://dx.doi.org/10.1255/jnirs.874
6. G. Britton, "Structure and properties of carotenoids in relation to function", *FASEB J.* **9,** 1551 (1995). Pubmed: http://www.ncbi.nlm.nih.gov/pubmed/8529834
7. C.O. Perera and G.M. Yen, "Functional properties of carotenoids in human health", *Int. J. Food Prop.* **10,** 201 (2007). doi: http://dx.doi.org/10.1080/10942910601045271
8. A. Saltzman, E. Birol, H.E. Bouis, E. Boy, F.F. De Moura, Y. Islam and W.H. Pfeiffer, "Biofortification: progress toward a more nourishing future", *Global Food Secur.* **2,** 9 (2013). doi: http://dx.doi.org/10.1016/j.gfs.2012.12.003
9. H. Ceballos, N. Morante, T. Sánchez, D. Ortiz, I. Aragón, A. Chávez, M. Pizarro, F. Calle and D. Dufour, "Rapid cycling recurrent selection for increased carotenoids content in cassava roots", *Crop Sci.* **53,** 2342 (2013). doi: http://dx.doi.org/10.2135/cropsci2013.02.0123
10. R. Alcides Oliveira, J. De Carvalho Lv and W.G. Fukuda, "Assessment and degradation study of total carotenoid and beta carotene in bitter yellow cassava varieties", *African J. Food Sci.* **4,** 148 (2010).
11. A. Chavez, T. Sanchez, H. Ceballos, D. Rodriguez-Amaya, P. Nestel, J. Tohme and M. Ishitani, "Retention of carotenoids in cassava roots submitted to different processing methods", *J. Sci. Food Agric.* **87,** 388 (2007). doi: http://dx.doi.org/10.1002/jsfa.2704
12. M.W. Davey, W. Saeys, E. Hof, H. Ramon, R.L. Swennen and J. Keulemans, "Application of visible and near-infrared reflectance spectroscopy (Vis/NIRS) to determine carotenoid contents in banana (*Musa* spp.) fruit pulp", *J. Agric. Food Chem.* **57,** 1742 (2009). doi: http://dx.doi.org/10.1021/jf803137d
13. A. Clement, M. Dorais and M. Vernon, "Nondestructive measurement of fresh tomato lycopene content and other physicochemical characteristics using Visible-NIR spectroscopy", *J. Agric. Food Chem.* **56,** 9813 (2008). doi: http://dx.doi.org/10.1021/jf801299r
14. M. Moh, Y.C. Man, B. Badlishah, S. Jinap, M. Saad and W. Abdullah, "Quantitative analysis of palm carotene using Fourier transform infrared and near infrared spectroscopy", *J. Amer. Oil Chem. Soc.* **76,** 249 (1999). doi: http://dx.doi.org/10.1007/s11746-999-0226-9
15. H. Schulz, H. Drews, R. Quilitzsch and H. Krüger, "Application of near infrared spectroscopy for the quantification of quality parameters in selected vegetables and essential oil plants", *J. Near Infrared Spectrosc.* **6A,** 125 (1998). doi: http://dx.doi.org/10.1255/jnirs.179
16. M. Bonierbale, W. Grüneberg, W. Amoros, G. Burgos, E. Salas, E. Porras and T. zum Felde, "Total and individual carotenoid profiles in solanum phureja cultivated potatoes: II. Development and application of near-infrared reflectance spectroscopy (NIRS) calibrations for germplasm characterization", *J. Food Compos. Anal.* **22,** 509 (2009). doi: http://dx.doi.org/10.1016/j.jfca.2008.08.009
17. V. Lebot, A. Champagne, R. Malapa and D. Shiley, "NIR determination of major constituents in tropical root and tuber crop flours", *J. Agric. Food Chem.* **57,** 10539 (2009). doi: http://dx.doi.org/10.1021/jf902675n

18. S. Tumwegamire, R. Kapinga, P.R. Rubaihayo, D.R. LaBonte, W.J. Grüneberg, G. Burgos, T. zum Felde, R. Carpio, E. Pawelzik and R.O. Mwanga, "Evaluation of dry matter, protein, starch, sucrose, β-carotene, iron, zinc, calcium, and magnesium in East African sweetpotato [*Ipomoea batatas* (L.) Lam] germplasm", *HortScience* **46,** 348 (2011).

19. A. López, S. Arazuri, I. García, J. Mangado and C. Jarén, "A review of the application of near-infrared spectroscopy for the analysis of potatoes", *J. Agric. Food Chem.* **61,** 5413 (2013). doi: http://dx.doi.org/10.1021/jf401292j

20. T. Sánchez, H. Ceballos, D. Dufour, D. Ortiz, N. Morante, F. Calle, T. Zum Felde, M. Domínguez and F. Davrieux, "Prediction of carotenoids, cyanide and dry matter contents in fresh cassava root using NIRS and Hunter color techniques", *Food Chem.* **151,** 444 (2014). doi: http://dx.doi.org/10.1016/j.foodchem.2013.11.081

21. S. Wold, M. Sjöström and L. Eriksson, "PLS-regression: a basic tool of chemometrics", *Chemometr. Intell. Lab. Syst.* **58,** 109 (2001). doi: http://dx.doi.org/10.1016/S0169-7439(01)00155-1

22. J. Shenk and M. Westerhaus, "Population definition, sample selection, and calibration procedures for near infrared reflectance spectroscopy", *Crop Sci.* **31,** 469 (1991). doi: http://dx.doi.org/10.2135/cropsci1991.0011183X003100020049x

23. P. Dardenne, G. Sinnaeve and V. Baeten, "Multivariate calibration and chemometrics for near infrared spectroscopy: which method?", *J. Near Infrared Spectrosc.* **8,** 229 (2000). doi: http://dx.doi.org/10.1255/jnirs.283

24. D. Perez-Marin, A. Garrido-Varo and J. Guerrero, "Nonlinear regression methods in NIRS quantitative analysis", *Talanta* **72,** 28 (2007). doi: http://dx.doi.org/10.1016/j.talanta.2006.10.036

25. E. Bertran, M. Blanco, S. Maspoch, M. Ortiz, M. Sánchez and L. Sarabia, "Handling intrinsic non-linearity in near-infrared reflectance spectroscopy", *Chemometr. Intell. Lab. Syst.* **49,** 215 (1999). doi: http://dx.doi.org/10.1016/S0169-7439(99)00043-X

26. F. Estienne, L. Pasti, V. Centner, B. Walczak, F. Despagne, D.J. Rimbaud, O. De Noord and D. Massart, "A comparison of multivariate calibration techniques applied to experimental NIR data sets: Part II. Predictive ability under extrapolation conditions", *Chemometr. Intell. Lab. Syst.* **58,** 195 (2001). doi: http://dx.doi.org/10.1016/S0169-7439(01)00159-9

27. E. Zamora-Rojas, A. Garrido-Varo, F. Van den Berg, J. Guerrero-Ginel and D. Pérez-Marín, "Evaluation of a new local modelling approach for large and heterogeneous NIRS data sets", *Chemometr. Intell. Lab. Syst.* **101,** 87 (2010). doi: http://dx.doi.org/10.1016/j.chemolab.2010.01.004

28. G. Sinnaeve, P. Dardenne and R. Agneessens, "Global or Local? A choice for NIR calibrations in analyses of forage quality", *J. Near Infrared Spectrosc.* **2,** 163 (1994). doi: http://dx.doi.org/10.1255/jnirs.43

29. B. Godin, R. Agneessens, J. Delcarte and P. Dardenne, "Prediction of chemical characteristics of fibrous plant biomasses from their near infrared spectrum: comparing Local versus partial least square models and cross-validation versus independent validations", *J. Near Infrared Spectrosc.* **23,** 1 (2015). doi: http://dx.doi.org/10.1255/jnirs.1138

30. R. Dambergs, D. Cozzolino, W. Cynkar, L. Janik and M. Gishen, "The determination of red grape quality parameters using the LOCAL algorithm", *J. Near Infrared Spectrosc.* **14,** 71 (2006). doi: http://dx.doi.org/10.1255/jnirs.593

31. M.-T. Sánchez, M.-J. De la Haba, J.-E. Guerrero, A. Garrido-Varo and D. Pérez-Marín, "Testing of a Local approach for the prediction of quality parameters in intact nectarines using a portable NIRS instrument", *Postharvest Biol. Technol.* **60,** 130 (2011). doi: http://dx.doi.org/10.1016/j.postharvbio.2010.12.006

32. J.S. Shenk, M.O. Westerhaus and P. Berzaghi, "Investigation of a LOCAL calibration procedure for near infrared instruments", *J. Near Infrared Spectrosc.* **5,** 223 (1997). doi: http://dx.doi.org/10.1255/jnirs.115

33. P. Berzaghi, J.S. Shenk and M.O. Westerhaus, "LOCAL prediction with near infrared multi-product databases", *J. Near Infrared Spectrosc.* **8,** 1 (2000). doi: http://dx.doi.org/10.1255/jnirs.258

34. D. Ortiz, T. Sánchez, N. Morante, H. Ceballos, H. Pachón, M.C. Duque, A.L. Chávez and A.F. Escobar, "Sampling strategies for proper quantification of carotenoid content in cassava breeding", *Plant Breed. Crop Sci.* **3,** 14 (2011).

35. C.W. Dunnett, "A multiple comparison procedure for comparing several treatments with a control", *J. Amer. Stat. Assoc.* **50,** 1096 (1955). doi: http://dx.doi.org/10.1080/01621459.1955.10501294

36. R. Anderssen, B. Osborne and I. Wesley, "The application of localisation to near infrared calibration and prediction through partial least squares regression", *J. Near Infrared Spectrosc.* **11,** 39 (2003). doi: http://dx.doi.org/10.1255/jnirs.352