Analysis of uncertainties affecting clay contents predictions obtained from airborne VNIR/SWIR hyperspectral data

Cecile Gomez, Arthur Drost, J.M. Roger

Author-produced version of the article published in Remote Sensing of Environment, 2015, N°156, p. 58-70.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0034425714003861
Doi: 10.1016/j.rse.2014.09.032

Analysis of uncertainties affecting clay contents predictions obtained from airborne VNIR/SWIR Hyperspectral data

Gomez C. [1], Drost A.P.A. [1], Roger J-M[2]

[1] IRD, UMR LISAH (INRA-IRD-SupAgro), F-34060 Montpellier, France (cecile.gomez@ird.fr)

[2] IRSTEA, UMR ITAP, Montpellier, France

Abstract:

Visible, Near-Infrared and Short Wave Infrared (VNIR/SWIR, 350-2500 nm) hyperspectral imaging spectroscopy may provide estimated soil properties maps. The performance of the estimations is usually assessed with figures of merit such as the Standard Error of Calibration, the Standard Error of Prediction or the Ratio of Performance Deviation. All of these parameters are estimated during model building and validation stages to evaluate global model performance. Beyond these global indicators, evaluation of uncertainty affecting each prediction is a major trend in analytical chemistry and chemometrics, but not yet in hyperspectral imagery. Several approximate expressions and resampling methods have been proposed for the estimation of prediction uncertainty when using multivariate calibration from laboratory spectra. Based on these works, this paper proposes a mapping and analysis of the uncertainties affecting predictions obtained from VNIR/SWIR airborne data, using several methods. An application to real VNIR/SWIR airborne data related to clay content allowed us to compare these different methods. A focus on different specific cases yielded some insights on the uncertainty sources and showed that uncertainty analysis could guide the user to better sampling, better calibration and finally better mapping.

Keyword:

Airborne hyperspectral imagery ; VNIR/SWIR spectroscopy ; uncertainty ; Multivariate calibration; Clay content mapping;

Author-produced version of the article published in Remote Sensing of Environment, 2015, N°156, p. 58-70.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0034425714003861
Doi: 10.1016/j.rse.2014.09.032

1- Introduction

Since two decades, laboratory Visible, Near-Infrared and Short Wave Infrared (VNIR/SWIR, 350-2500 nm) spectroscopy has been proven as a good alternative to costly physical and chemical laboratory soil analysis for the estimation of a large range of soil properties (e.g. Ben-Dor & Banin, 1995, Viscarra Rossel, 2006, Cécillon et al., 2009). These works opened the way to VNIR/SWIR hyperspectral imaging spectroscopy for estimating soil properties. VNIR/SWIR hyperspectral imaging spectroscopy may provide a global view of the area under study at high spatial resolutions, and may provide characterization of several soil properties simultaneously. Ben Dor et al. (2002) started this research way, with multivariate calibration statistics applied to remotely-sensed data. Since then, the VNIR/SWIR hyperspectral imaging spectroscopy benefits from an increasing number of methodologies developed in lab soil property prediction studies, including Partial Least Square Regression (e.g. Gomez et al., 2008a), Support Vector Machine Regression (e.g. Stevens et al., 2010), Multiple Linear Regression (e.g. Bayer et al., 2012), Stepwise Multiple Linear Regression (e.g. Lu et al., 2013) and Regression Rules (Budiman & McBratney, 2008). And several studies have been successfully conducted to map soil properties such as Clay, Calcium Carbonate, Iron, Soil Organic Carbon… (e.g., Ben-Dor et al., 2007, Selige et al., 2006, Gomez et al., 2008b, Gomez et al., 2012a, Stevens et al., 2010).

Nevertheless, whatever the regression methods, the number of samples in the calibration database, the scales of study or the studied soil properties, the quality of the mapping results are expressed by using the performance of the models with global figures of merit such as standard error of calibration (SEC) and standard error of prediction (SEP), respectively referred to as Root Mean Squared Error of Calibration (RMSEC) and Root Mean Squared Error of Prediction (RMSEP). The coefficient of Determination (R^2) and the Ratio of Performance Deviation (RPD) are also used as a measure of multivariate model prediction accuracy. The ratio of performance to interquartile (RPIQ), which is the ratio of the interquartile (IQ = Q3–Q1) to the RMSEP has been recently proposed as an alternative to the RPD to better take into account the shape of the distribution (Bellon-Maurel et al., 2010). All of these parameters are estimated during model building and validation stages to evaluate global model performance. Selige et al. (2006) used the Partial Least Square Regression (PLSR) method with 72 soil samples in their calibration database to map Sand content over 7 km^2, from HYMAP VNIR/SWIR hyperspectral data. Gomez et al. (2012b) used the PLSR method with 95 soil samples in their calibration database to map Clay content over 300 km^2, from AISA-DUAL VNIR/SWIR hyperspectral data. And both researches expressed the

Author-produced version of the article published in Remote Sensing of Environment, 2015, N°156, p. 58-70.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0034425714003861
Doi: 10.1016/j.rse.2014.09.032

quality of their mapping results by using the performance of their PLSR models (calculation of the figures of merit) associated to a visual pedological expertise of the soil maps. Finally, some researches added a geostatistical analysis to study the spatial structure of the predicted soil property (e.g. Schwanghart & Jarmer 2011, Gomez et al., 2012a).

Added to the global performance indicators, visual pedological expertise and geostatistic analysis, the accuracy, precision, trueness and uncertainty of each new prediction should have to be provided, in association to the predicted values, to better characterize the quality of the maps of predictions. In this paper, we chose the following definitions, also described by Zeaiter et al., 2004. The "accuracy" of new predictions is the distance between the predicted value and the "true" value (ISO 5725-1:1994). And the "uncertainty" of new predictions is defined as "a parameter associated with the result of a measurement that characterizes the dispersion of the values that could reasonably be attributed to the measurand" (AFNOR NFX07-001:1994), and is related to the variance of the prediction. The accuracy and the uncertainty are two uncorrelated indicators of quality of the predictions. A predicted value can be close to the "true" value (high accuracy), but associated to a high uncertainty (high variance). And inversely, a predicted value can be far to the "true" value (low accuracy), but associated to a low uncertainty (low variance). The accuracy would be the more interesting criteria to obtain, but the estimation of the accuracy for each new prediction is still an inaccessible scientific challenge.

The uncertainty of new predictions has been firstly studied by Phatak et al. (1993), Denham (1997), and Hoskuldsson (1988) who assumed the hypothesis of negligible errors in predictors. Those works were expanded by Faber and Kowalski (1997) who included errors in the predictors under the general errors-in-variables model. A drawback of their approach is that the original expression is derived under the assumption that the errors in the predictors have constant variance (the homoscedastic case). Later Faber and Bro (2002) proposed a new expression which accommodated for heteroscedastic and correlated errors. But in fact, the expression was derived under the assumption that the errors in predictors are identically and independently distributed (i.i.d.) and the authors conjectured that it applied to most types of heteroscedasticity. More recently, Fernandez-Ahumada et al. (2012) proposed a new expression for the variance of the prediction adapted to any linear calibration models, like, e.g. PLSR. This formulation respects the specificities of spectrometry and particularly the spectral error structure which is induced by the high colinearity of the variables.

Based on these works, this paper proposes an analysis process of the uncertainty affecting predictions obtained from VNIR/SWIR airborne data. A bootstrap procedure allows

Author-produced version of the article published in Remote Sensing of Environment, 2015, N°156, p. 58-70.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0034425714003861
Doi: 10.1016/j.rse.2014.09.032

the calculation of a variance of predictions of each VNIR/SWIR airborne spectrum, which would represent the sum of all the sources on uncertainty and would be considered as the "true" variance. In addition, the distance between the spectral predictors and the spectral calibration samples are calculated as two uncertainties expressions: i) in a Principal Component Analysis (PCA) space (Mahalanobis Distance) and ii) in the multivariate model space (Leverage). Finally, the expression for the variance of prediction proposed by Fernandez-Ahumada et al. (2012) is adapted to the VNIR/SWIR airborne data. The adaptation of this formula allows taking into account the spatial dimension of these remote sensing data and allows the spatialisation of the total variance of predictions, and each term of the formula: *i)* variance of predictions due to the multivariate model, *ii)* variance of predictions due to the spectra and *iii)* interaction between these two effects. The proposed analysis process uses the spatial dimension of the VNIR/SWIR airborne data to express the uncertainties in the form of maps.

This process of uncertainty analysis has been performed on recent study of a clay content mapping by VNIR/SWIR AISA-Dual airborne data over a Tunisia area (Gomez et al., 2012b). First, all the uncertainty expressions are calculated and analyzed on a validation database. Secondly, all the uncertainty expressions are calculated, mapped and analyzed on a test area and different specific cases highlighted some specific interest of these uncertainty expressions.

2. Materials

2.1 Notations

In the following paper, capital bold characters will be used for matrices, e.g. **A**; small bold characters for column vectors, e.g. $\mathbf{a}_i$ will denote the i^{th} column of **A**; row vectors will be denoted by the transpose notation, e.g. **a**'. Lowercase non bold italic characters will be used for scalar variables, e.g. indices *i*. Uppercase non bold italic characters will be used for scalar constants, e.g the number of samples *N*. If needed, matrix dimensions will be indicated, e.g **A** (*N*×*P*). The trace of a square matrix **A** will be noted tr(**A**).

2.2 Site description

The study area is located in the Cap Bon region in northern Tunisia (36°24'N to 36°53'N; 10°20'E to 10°58'E), 60 km east of Tunis, Tunisia (Figure 1). This 300 km² area includes the Lebna catchment, which is mainly rural (>90%) and devoted to cereals in addition to legumes,

Author-produced version of the article published in Remote Sensing of Environment, 2015, N°156, p. 58-70.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0034425714003861
Doi: 10.1016/j.rse.2014.09.032

olive trees, natural vegetation for breeding and vineyards. It is characterized by rolling areas, with an altitude between 0 and 226 m. The climate varies from humid to semi-arid, with an inter-annual precipitation of 600 mm and an inter-annual potential evapotranspiration of 1500 mm. The soil pattern of the Lebna catchment arises mainly from variations in lithology. The changes in the landscape between Miocene sandstone and marl outcrops induce large variations in the physical and chemical soil properties (map of Zante et al., IRD production). Furthermore, the distance between successive sandstone outcrops decreases steeply along a sea-mountain direction, which results in variations in the soil property patterns as well (Gomez et al., 2012b). The soil materials were redistributed laterally along the slopes during the Holocene, which add to the complexity of the soil patterns. The main soil types are Regosols, Eutric Regosols (9.6%) preferentially associated with sandstone outcrops, Calcic Cambisol, and Vertisol preferentially formed on marl outcrops and lowlands. The southeastern region of the study area has a flatter landscape with sandy Pliocene deposits yielding Calcosol and Rendzina.

[Figure 1]

2.3 VNIR/SWIR hyperspectral data

On November 2, 2010, an AISA-Dual hyperspectral airborne image was acquired for the study area with a spatial resolution of 5 m (Figure 1b). The area of the image is approximately 12 km x 24 km. The AISA-Dual spectrometer measured the reflected radiance in 359 non-contiguous bands covering the 400- to 2450-nm spectral domain, with 4.6 nm bandwidths between 400 and 970 nm and 6.5 nm bandwidths between 970 and 2450 nm. The instantaneous field of view (IFOV) was 24 degrees. The radiance units were converted to reflectance units using ASD spectrometer measurements of uniform surfaces (parking lots, asphalt, concrete) that were collected at the same time during the over flight. An empirical line correction method was used to calibrate each flight line to the reflectance. Topographic corrections were performed using a digital elevation model built from the ASTER data and ground control points. In this study, we removed 1) the spectral bands in the blue part of the spectral domain (between 400 and 484 nm) due to noise in these bands and 2) the spectral bands between 1339 and 1464 nm as well as between 1772 and 2004 nm due to vibrational-rotational H2O absorption bands. Consequently, 280 AISA-Dual spectral bands were retained.

When the image was acquired (November 2010) the bare soils represented 46.3 % of the study area. The rest of the area was covered by green vegetation, consisting mainly of olive

Author-produced version of the article published in Remote Sensing of Environment, 2015, N°156, p. 58-70.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0034425714003861
Doi: 10.1016/j.rse.2014.09.032

trees, native forests, green plants and vineyards. To isolate the bare soil areas, pixels with normalized difference vegetation index (NDVI) values over an expert-calibrated threshold were masked: a value of 0.20 was determined after considering twenty parcels that had been visually inspected on the field. Water areas were also masked using an expert-calibrated threshold: pixels with a reflectance of less than 8% at 1665 nm were removed. Finally, urban areas were masked using a map of urban areas. A first clay content predicted map over the Lebna Catchment was obtained from a multivariate PLSR model using these AISA-Dual airborne data (Gomez et al., 2012b).

In this paper, the attention will be focused on a test area for the analysis of the uncertainty maps. This test area is a 6.67 km² area centered on the Kamech catchment, which had high percentage of bare soils during the image acquisition and exhibited contrasting soil patterns (Figure 1b, Figure 2). The bare soils represent 49.2% of the test area and 10705 AISA-Dual pixels. Over this area, six pixels have been selected to study in detail the expressions of uncertainty affecting predictions proposed in this paper. Three of these test-pixels are located on the center of bare soil fields and three other test-pixels are located on the boundary of fields (Figure 2).

[Figure 2]

2.4 Soil samples data base

129 soil samples were collected on the Lebna catchment. Of these samples, 58 were collected in October 2008, 30 in October 2009, and 41 in November 2010. All of the samples were composed of five sub-samples collected to a depth of 5 cm at random locations within a 10×10 m square centered on the geographical position of the sampling plot, as recorded by a Garmin GPS instrument. All of these soil samples were collected in fields that were bare during the hyperspectral data acquisition in November 2010, between 20 and 180 m of altitude. After homogenizing the sample, approximately 20 g was devoted to soil property analysis. The initial samples were air-dried and sieved with a 2 mm sieve prior to being transported to the laboratory for analysis. The determination of clay content (granulometric fraction < 2 µm) was determined according to the method NF X 31-107 (Baize and Jabiol, 1995). The clay content of the 129 soil samples varies between 46 and 777g/kg and follows a normal distribution.

3. Prediction Model: PLSR

Author-produced version of the article published in Remote Sensing of Environment, 2015, N°156, p. 58-70.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0034425714003861
Doi: 10.1016/j.rse.2014.09.032

The Partial Least Square Regression (PLSR) method (Wold et al.; 2001) was used to establish relationships between the soil clay content and the VNIR/SWIR hyperspectral imaging data. The spectroscopic and chemometric analyses were implemented in R (Version 1.17).

Prior to quantitative statistical analysis, the reflectance was converted into "pseudo absorbance" (log [1/reflectance]). Noise reduction was achieved through standard pre-treatments: a Savitzky–Golay filter with second-order polynomial smoothing and window widths of 30 nm (Savitzky and Golay, 1964) for noise removal, plus a Standard Normal Variate correction (Barnes et al., 1993) for additive and multiplicative effect removal.

The dataset was split into a calibration set (97 = 3 /4 of the total database, denoted as BD_Calib) and a validation set (32 = 1 /4 of the total database, denoted as BD_Valid). For each PLSR model, the reference values were sorted in an ascending order. The method *i)* starts by selecting the sample with the lowest reference value and put it in a calibration set, *ii)* then the next three samples are put in the validation set, *iii)* then the procedure is continued by alternately placing the following in the validation set and the next three samples in the calibration set. Because a limited number of samples were available, a leave-one-out cross-validation procedure was adopted to verify the prediction capability of the PLSR model for the calibration set (Wold, 1978). Each time, $N-1$ samples were used to build the regression model from all N samples within the dataset. Based on this model, the value for the soil property of the sample not used in developing the model was predicted. This procedure was repeated for all N samples, resulting in predictions for all of the calibration samples.

Outliers are commonly defined as observations that are not consistent with the majority of the data (Chiang et al., 2003; Pearson, 2002), such as observations that deviate significantly from normal values. An outlier can be defined as (i) a spectral outlier, when the sample is spectrally different from the rest of the samples or (ii) a concentration outlier, when the predicted value has a residual difference significantly greater than the mean of the predicted values. One method for identifying spectral outliers uses the principle of the Mahalanobis distance (Mark and Tunnell, 1985) applied to PCA-reduced data. In the present study, a value of 3 based on the Mahalanobis distance was selected for the identification of outliers. An analysis was carried out to detect outliers for all of the samples in the calibration set, and the detected spectral and concentration outliers were deleted from the calibration set.

In this paper, we only provide a brief description of the PLSR model that is fully detailed in Wold et al. (2001). The PLSR model is developed from a training set of N observations (number of spectra in the calibration dataset) with K **X**-variables (number of

Author-produced version of the article published in Remote Sensing of Environment, 2015, N°156, p. 58-70.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0034425714003861
Doi: 10.1016/j.rse.2014.09.032

wavelengths in the spectra) denoted $\mathbf{x}_k$ ($k=1,\ldots,K$), and M **Y**-variables (number of soil property) denoted $\mathbf{y}_m$ ($m = 1,\ldots M$). These training data forms the two matrices **X** and **Y** of dimensions ($N \times K$) and ($N \times M$) respectively. As all the factorial methods, the main principle of PLSR is: i) to find a subspace of the spectral space $\mathbb{R}^K$ on which the spectra are projected, yielding a matrix of N scores **T** ($N \times k$) ; ii) to perform a linear regression between **T** and **Y**. The first step is carried out iteratively, by searching loadings **u** and **v**, respectively from $\mathbb{R}^K$ and $\mathbb{R}^M$ such as $cov^2(\mathbf{Xu}, \mathbf{Yv})$ is maximal. The scores **T** are thus given by **T**=**XU**. The linear regression between **T** and **Y** produces an estimate $\hat{\mathrm{Y}} = \mathbf{T}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{Y}$, so that replacing **T** by **XU** finally yields: $\hat{\mathrm{Y}} = \mathbf{XU}(\mathbf{U}'\mathbf{X}'\mathbf{XU})^{-1}\mathbf{U}'\mathbf{X}'\mathbf{Y}$. The final regression coefficients are given by $\mathbf{b} = \mathbf{U}(\mathbf{U}'\mathbf{X}'\mathbf{XU})^{-1}\mathbf{U}'\mathbf{X}'\mathbf{Y}$. In the case where **Y** contains only one response, **b** is a K-vector, usually called "b-coefficients" which is generally analyzed as a spectrum. The k columns of **U** also are K-vectors and are called latent variables.

The prediction performances of PLSR models were evaluated using the coefficient of determination R^2_{cal} and R^2_{val} of predicted against measured values in the calibration and validation set respectively. The root mean square errors of calibration (RMSEC) and the root mean square errors in the validation set (RMSEP) were also analyzed for all the models. The ratio of performance to deviation (RPD), which is the ratio of the standard deviation in the validation set to the RMSEP, was used as an index of model accuracy.

4. Measurements of uncertainty affecting predictions

The estimation $\hat{y}$ of the y value for a new sample **x** can be written as:

$$\hat{y} = f(\mathbf{X}c, \mathbf{Y}c, Model, \mathbf{x}) \qquad (1)$$

where($\mathbf{X}c, \mathbf{Y}c$) contains the calibration spectra and the calibration responses; *Model* represents the calibration action, including preprocessing, choice of dimensions, etc. Thus, each prediction relies on a chain of operations, each of them adding a source of uncertainty:

- Uncertainty on spectra was assumed to be identical for calibration and test spectra. It is mainly due to the device repeatability (Ud) and to the spatial positioning (Us)
- Uncertainty on reference lab values (Uy)
- Uncertainty on model building (Um) may originate from 2 main causes, with Um = Uc + Ul. The first one (Uc) is related to the choice made for building the calibration set. The second one (Ul) is related to the choice of model dimension.

Author-produced version of the article published in Remote Sensing of Environment, 2015, N°156, p. 58-70.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0034425714003861
Doi: 10.1016/j.rse.2014.09.032

In this study, the following choices were made: Ud was considered negligible in comparison with Us and Uy was neglected. Us was modeled as a variance-covariance matrix $\mathbf{\Sigma}_x$ computed on the 9 point neighborhood of **x**. Uc and Ul were modeled and merged in a unique variance-covariance matrix $\mathbf{\Sigma}_b$ computed by means of bootstrap (Efron, 1982).

Seven expressions related to the uncertainty affecting predictions were computed, as described in the following sections.

4.1. Bootstrap procedure

A bootstrap procedure was performed to obtain a variance value which integrated all the sources of uncertainty previously described. Figure 3a describes the workflows and can be summary as:

1st step: *N* drawing with replacement of *N* samples among the calibration dataset.

2sd step: For each of the *N* selected samples, the associated AISA-Dual spectrum was randomly sampled among a grid of 3x3 pixels, centered on the location of the selected samples. To favor the central pixel, the sampling followed a normal distribution in line and column, with a standard deviation of 0.6.

3rd step: Drawing the number of latent variable, from 3 to 7, following a normal distribution centered on 5 with a standard deviation of 0.97.

4th step: For each new AISA-Dual spectrum **x** for which is looked for the clay content estimated value, the associated AISA-Dual spectrum was randomly sampled among a grid of 3x3 pixels, centered on the pixel **x**. To favor the central pixel, the sampling followed a normal distribution in line and column, with a standard deviation of 0.6. This 4th step is the same for the validation dataset *BD_Valid*, than for the entire image.

These four steps were done *R* times, producing *R* bootstrap (training) data sets and so producing *R* PLSR models, *R* prediction values for each new pixels and (*R* x *K)* b-coefficients. In this study the number of iterations *R* was 999. Predicted outcomes by cross-validation and corresponding performance indicators R^2_{cal}, and RMSEC were calculated for the calibration set at each *R* bootstrap iteration. Predicted outcomes and corresponding performance indicators R^2_{val}, RMSEP, and RPD were also calculated for the validation set, at each of the *R* bootstrap iterations.

4.2. Expressions of the uncertainty using bootstrap

Author-produced version of the article published in Remote Sensing of Environment, 2015, N°156, p. 58-70.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0034425714003861
Doi: 10.1016/j.rse.2014.09.032

Bootstrap was used to obtain a variance value $\boldsymbol{var}(\hat{\boldsymbol{y}})_{\boldsymbol{BS}}$, where no assumptions were considered. At the end of the *R* iterations, the variance $\boldsymbol{var}(\hat{\boldsymbol{y}})_{\boldsymbol{BS}}$ of the *R* predictions for each pixel of the AISA-Dual image was calculated (Figure 3). In case of the bootstrap would integrate all the sources on uncertainty (on spectra, reference lab values, model building), the variance $\boldsymbol{var}(\hat{\boldsymbol{y}})_{\boldsymbol{BS}}$ would represent the sum of all the sources on uncertainty and would be considered as the "true" variance, following:

$$\mathrm{var}(\hat{\boldsymbol{y}})_{\mathrm{BS}} = \mathrm{Ud} + \mathrm{Us} + \mathrm{Uy} + \mathrm{Um} \qquad (2)$$

4.2 Expressions of the uncertainty using spectral distance

a. Mahalanobis Distance

The Mahalanobis distance (*MD*) may be used to detect the outliers that are observations sitting at the periphery of the data cloud, which can have a stronger influence than average on the fitted model. In the framework of linear regression, the farer a sample from the center of the model, the higher its prediction uncertainty. The *MD* of a spectrum is calculated independently of the bootstrap procedure (Figure 3b), as the distance of this spectrum to the center of the calibration set, in a PCA-reduced dimension. So the *MD* of a VNIR/SWIR spectrum **x** is calculated from the calibration database, following:

$$MD(\mathbf{x}) = \sqrt{(\mathbf{x} - \bar{\mu})' \mathbf{S}^{-1} (\mathbf{x} - \bar{\mu})} \qquad (3)$$

Where **x** ($K \times 1$) is a VNIR/SWIR spectrum, $\bar{\mu}$ ($K \times 1$) is the mean of the *N* calibration VNIR/SWIR spectra and **S** ($K \times K$) is the variance-covariance matrix of the *N* calibration spectra. The principle of the Mahalanobis distance is described in detail in Mark and Tunnell, 1985. Generally, a *MD* value superior to 3 is considered to reflect an outlier with regard to the calibration set (Mark and Tunnell, 1985). The *MD* is related to the uncertainty on model building (Um), over the hypothesis that the PCA space is close to the PLSR space.

b. Leverage

The leverage (*H*) may be also used to diagnose how atypical a new vector of predictor is. The leverage value of a spectrum $\boldsymbol{x}$ is the distance of this spectrum to the center of the calibration set, within the multivariate model. The leverage measures the variation within the multivariate model, so the *H* measurement is dependent on the model, contrary to the Mahalanobis distance. Although the *H* measurement is dependent on the model, it is calculated independently of the bootstrap procedure (Figure 3b). The leverage of a VNIR/SWIR spectrum $\boldsymbol{x}$ is calculated following (Martens, 1991):

Author-produced version of the article published in Remote Sensing of Environment, 2015, N°156, p. 58-70.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0034425714003861
Doi: 10.1016/j.rse.2014.09.032

$$H(\mathbf{x}) = \mathbf{t_x}' \, (\mathbf{T}'\mathbf{T})^{-1} \, \mathbf{t_x} \qquad (4)$$

Where $\mathbf{x}$ $(K \times 1)$ is a VNIR/SWIR spectrum, $\mathbf{t_x}$ $(K \times 1)$ is the score vector of $\mathbf{x}$ for the multivariate model, and $\mathbf{T}$ $(N \times A)$ is the $\mathbf{X}$-scores matrix of the multivariate model (see section 3). The leverage is related to the uncertainty on model building (Um).

4.3 Expressions of the uncertainty using the Fernandez-Ahumada et al. (2012) formula

a. Total Variance of predictions

As a measure of uncertainty of the estimations by PLSR model, a variance model, proposed by Fernandez-Ahumada et al. (2012), has been applied in this study. This proposal for variance modeling is based on the EIV-model initialy proposed by Faber and Kowalski (1997), which considers all sources of uncertainty affecting predictors, the dependent variables and model coefficients. The difference between the classical EIV-model and the new variance model proposed by Fernandez-Ahumada et al. (2012) is that the latter model is constructed based on assumptions related to the predictors, the spectral bands. The formulation is described in detail in Fernandez-Ahumada et al. (2012), and can be resumed by:

$$\mathrm{var}(\hat{y})_{\mathbf{Ahumada}} = \left(1 + \frac{1}{N}\right) \mathbf{b}'\mathbf{\Sigma_x}\mathbf{b} + \mathbf{z}'\mathbf{\Sigma_b}\mathbf{z} \; + \left(1 + \frac{1}{N}\right) \mathrm{tr}(\mathbf{\Sigma_x}\mathbf{\Sigma_b}) + \frac{\sigma^2_{\mathrm{lab}}}{N} \qquad (5)$$

Where $\hat{y}$ is the prediction, $\mathbf{z}$ is the centered $\mathbf{x}$ spectrum used for the prediction, $\mathbf{b}$ is the K-vector of the b-coefficients of the model, N is the number of the calibration samples, $\mathbf{\Sigma_b}$ is the variance-covariance matrix of the b-coefficients, $\mathbf{\Sigma_x}$ is the variance-covariance matrix of the spectrum $\mathbf{x}$, and σ^2_{lab} is the laboratory y variance,.

The four terms of equation (5) from left to right are related to: *T1*) the hyper spectral data uncertainty, *T2*) the model coefficients uncertainty, *T3*) the dependency between spectral and model uncertainties and *T4*) the laboratory reference measurement uncertainty. Term *T4* is constant all over the image and is certainly small with regards to the three other terms. So in this study the variance model was confined to the first three terms, formulated by:

$$\mathrm{var}(\hat{y})_{\mathbf{Ahumada}} = \left(1 + \frac{1}{N}\right) \mathbf{b}'\mathbf{\Sigma_x}\mathbf{b} + \mathbf{z}'\mathbf{\Sigma_b}\mathbf{z} \; + \left(1 + \frac{1}{N}\right) \mathrm{tr}(\mathbf{\Sigma_x}\mathbf{\Sigma_b}) \qquad (6)$$

Each term of the equation (6) and their calculation are described in the following sections.

b. Variance of predictions due to the spectrum *x*

Author-produced version of the article published in Remote Sensing of Environment, 2015, N°156, p. 58-70.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0034425714003861
Doi: 10.1016/j.rse.2014.09.032

The first term of the formula of Fernandez-Ahumada et al. (2012) (denoted *T1*, Figure 3d) expresses the variance of predictions caused by the uncertainty of hyperspectral data **x** for each pixel:

$$T1 = \left(1 + \frac{1}{N}\right) \mathbf{b}' \mathbf{\Sigma_x} \mathbf{b} \qquad (7)$$

Where N is the number of samples in the calibration set, **b** is the model coefficients (here the mean of the R b-coefficients vectors obtained in the bootstrap procedure), and $\mathbf{\Sigma_x}$ is a variance-covariance matrix that describes the uncertainty of the spectrum **x**. In this study, the variance-covariance matrix $\mathbf{\Sigma_x}$ of the spectrum ***x*** was calculated following:

$$\Sigma_x = \mathrm{cov}(\mathbf{G(x)}) \qquad (8)$$

Where **G(x)** is a $(9 \times K)$ matrix where each line is a spectrum belongs to the grid of 3x3 pixels centered on the spectrum **x**. So the matrix $\mathbf{\Sigma_x}$ $(K \times K)$ attempts to reflect the influence of surrounding areas on reflection values of the center pixel **x** (Figure 3c).

Since especially reflectance values of vegetated or surrounding bare soil parcels are hypothetically influential for the variance of the clay content, the algorithm used the full image, i.e. the variance-covariance matrix $\mathbf{\Sigma_x}$ of a pixel at the edge of a parcel, and also included spectral information of the pixels that are normally masked because they are considered vegetated area. The term *T1* is related to the uncertainty on spectra due to the spatial positioning (Us). It reflects how the uncertainty of the spectrum **x** is amplified by the model.

c. Variance of predictions due to the PLSR model

The second term of the formula of Fernandez-Ahumada et al. (2012) (denoted *T2*, Figure 3d) expresses the variance of predictions due to the uncertainty on the multivariate model:

$$T2 = \mathbf{z}' \mathbf{\Sigma_b} \mathbf{z} \qquad (9)$$

Where **z** is the centered spectrum of **x** and $\mathbf{\Sigma_b}$ is the variance-covariance matrix $(K \times K)$ of the b-coefficients. In this study, the variance-covariance matrix $\mathbf{\Sigma_b}$ was calculated using the b-coefficients obtained at each of the R iterations of the bootstrap (Figure 3a), following:

$$\mathbf{\Sigma}_\mathrm{b} = \mathrm{cov}(\mathbf{B}) \qquad (10)$$

Where **B** is the matrix $(R \times K)$ where the i^{th} line contains the b-coefficients of the i^{th} iteration of the bootstrap. This second term is considered as a distance between the spectrum and the model center, weighted by the model noise. It depends mainly on three factors: (i) the length of the centred spectra **z**, which is related to the classical concept of leverage, (ii) the norm of $\mathbf{\Sigma_b}$ and (iii) the colinearity between **z** and $\mathbf{\Sigma_b}$. The term *T2* is related to the uncertainty on

Author-produced version of the article published in Remote Sensing of Environment, 2015, N°156, p. 58-70.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0034425714003861
Doi: 10.1016/j.rse.2014.09.032

model building (Um). It reflects how the uncertainty of the model is amplified by the spectrum **x**.

d. Variance of predictions due to the intersection of the multivariate model and the spectrum variances

The third term of the formula of Fernandez-Ahumada et al. (2012) (denoted *T3*, Figure 3d) expresses the intersection of the variance of predictions due to the multivariate model and the spectrum ***x***:

$$T3 = \left(1 + \frac{\mathbf{1}}{N}\right) \mathrm{tr}(\mathbf{\Sigma_x \Sigma_b}) \qquad (11)$$

The quantity $\mathrm{tr}(\mathbf{\Sigma_x \Sigma_b})$ measures the common part of the instances of $\mathbf{\Sigma_x}$ and $\mathbf{\Sigma_b}$ (Fernandez-Ahumada et al., 2012). A low value of the term *T3* would signify that the uncertainty of the model and the spectrum, $\mathbf{\Sigma_b}$ and $\mathbf{\Sigma_x}$ respectively, are different and could cancel each other out. A high value of the term *T3* would signify that the uncertainty of the model and the spectrum, $\mathbf{\Sigma_b}$ and $\mathbf{\Sigma_x}$ respectively, are similar and could be accumulated.

[Figure 3]

5. Results

5.1 Global performances of the models – preliminary results

The VNIR/SWIR spectra extracted from the AISA-DUAL image at the locations of the 97 soil samples of the calibration data set were used to build PLSR-based prediction models. Only one spectral outlier was identified among the 97 calibration data. So all the PLSR models have been built from 96 VNIR/SWIR spectra. Moreover, the VNIR/SWIR spectra extracted from the AISA-DUAL image at the locations of the 32 soil samples of the validation data set were used to validate the PLSR-based prediction models.

Following the classes of RPD defined by Chang and Laird (2002), a correct prediction was obtained by PLSR model from the 96 AISA-DUAL spectra for clay content prediction, with R^2_{val} and RPD values greater than 0.7 and 1.4, respectively, and with RMSEP around 94g/kg (Table 1). A correct prediction was also obtained by bootstrap-PLSR models with R^2_{val} and RPD values greater than 0.5 and 1.4, respectively, and with RMSEP around 109 g/kg (Table 1).

[Table 1]

5.2 Uncertainty analysis on validation data set

Author-produced version of the article published in Remote Sensing of Environment, 2015, N°156, p. 58-70.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0034425714003861
Doi: 10.1016/j.rse.2014.09.032

The seven uncertainty expressions described in section 4 have been firstly calculated for the 32 samples of the BD_Valid. The correlations between these uncertainty expressions (Table 2) and their ranges (Figure 4) have been studied.

The variance model $\mathrm{var}(\hat{y})_{\mathrm{Ahumada}}$ is very highly correlated to *T2* (R ≈ 1, Table 2). So the most part of $\mathrm{var}(\hat{y})_{Ahumada}$ is due to the uncertainty on the multivariate model. Moreover, the variance $\mathrm{var}(\hat{y})_{\mathrm{BS}}$ of the *R* predictions for each AISA-DUAL spectra of the validation data set is highly correlated to the variance model $\mathrm{var}(\hat{y})_{\mathrm{Ahumada}}$ and to the term *T2* (R = 0.91, Table 2). So the "true" variance of predictions, expressed by $\mathrm{var}(\hat{y})_{\mathrm{BS}}$ is mostly due to the uncertainty on the multivariate model, and in minority to the uncertainty on the spectra.

The correlation between the "true" variance of predictions $\mathrm{var}(\hat{y})_{\mathrm{BS}}$ and the Mahalanobis Distance *MD* is modest (R = 0.52, Table 2). As the Mahalanobis Distance is related to the uncertainty on model building (Um), over the hypothesis that the PCA space is close to the PLSR space, this low correlation is coherent. Indeed the PCA space is not the same than the PLSR space, so the Mahalanobis Distance does not represent correctly the uncertainty on model building (Um).

The absence of correlation between the "true" variance of predictions $\mathrm{var}(\hat{y})_{\mathrm{BS}}$ and the Leverage *H* reveals that *H* represents a minor part of Um. The leverage *H* is an adaptation of the formula of $\mathrm{var}(\hat{y})$ in classical linear regressions to factorial regressions. In classical linear regressions, the only source of uncertainty is associated to *y*, so $\mathrm{var}(\hat{y}) = \mathbf{z}'\,(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{z}$, because $\mathrm{var}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}$. The leverage *H* uses this formula in the latent variables space. So *H* is a simplified version of *T2*, which express the uncertainty on model building over the hypothesis that the only source of uncertainty comes from *y*. In this case of spatialisation of uncertainties from VNIR/SWIR hyperspectral data, this expression *H* of the uncertainty seems to be inadequate or, a minima too limited to express reality of source of uncertainty on model building.

Finally, we can observe that uncertainty values of *T1* were lower compared to values of *T2* which is 200 times superior, and the range of the *T1* values was also low (Figure 4). And no correlation appears between the uncertainty expressions and the residues (R<0.4, Table 2).

Thresholds corresponding to the 95th percentile of variance distributions were set, so that abnormally high values can be filtered. Values of 3, 0.2, 5000, 5000 and 5000 were calculated for *MD*, *H*, *T2*, $\mathrm{var}(\hat{y})_{\mathrm{BS}}$ and $\mathrm{var}(\hat{y})_{\mathrm{Ahumada}}$ respectively (Figure 4). These

Author-produced version of the article published in Remote Sensing of Environment, 2015, N°156, p. 58-70.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0034425714003861
Doi: 10.1016/j.rse.2014.09.032

thresholds have been used in next sections to study and identify some area associated to high uncertainties.

[Table 2]

[Figure 4]

5.3 Uncertainty mapping on Kamech area

The higher values of the seven uncertainty expressions were gathered on the East part of the area, which may correspond to some isolated houses (black rectangle on Figure 5a). Urban pixels have been masked in the pre-treatment of the AISA-DUAL data, using an urban vector layer of the region which took into account the bigger towns of the region. But some isolated houses, not referenced in the urban map of the region, could be missed by this mask pre-treatment and seem to be identified by each of these uncertainty maps (Figure 5 and 6). Over the 10750 studied pixels, only 108 pixels had the combination $MD > 3$, $H > 0.2$, $T2 > 6000$, $\mathrm{var}(\hat{y})_{\mathrm{BS}} > 5000$ and $\mathrm{var}(\hat{y})_{\mathrm{Ahumada}} > 5000$. So the thresholds defined from the validation database may be used to mask some missed no-soil area.

The higher values of *T1* (except those on the urban area) were located at the boundary of the fields (Figure 6a) which is coherent. Indeed these high values of *T1* were the consequence of the calculation of $\mathbf{\Sigma_x}$ which used the grid of (3×3) pixels centered of the studied pixel associated to the spectrum **x**. If the studied pixel associated to the spectrum **x** is located on the boundary of the field, its $\mathbf{\Sigma_x}$ will include the heterogeneity of the boundary (field, ditch, road or path).

The new maps of uncertainty expressions, using the thresholds defined from the validation database, offered more details (Figure 7). Two parts of the Kamech area have high values of $\mathrm{var}(\hat{y})_{\mathrm{BS}}$, *H*, *T2* and $\mathrm{var}(\hat{y})_{\mathrm{Ahumada}}$: one field on the south of the area and the band North-East South-West highlighted by black boxes in Figure 7d. These areas are covered by cultivated fields, mostly bare during the hyperspectral flight. A low proportion of dry vegetation may be missed by the masking pre-treatment and may modify the signal that we consider as “soil signal”. As well soil humidity can modify the signal and no mask can be done to classify and remove specific humid area. As no ground truth exists on both areas, no explanation can be affirmed. But these areas of high uncertainty of prediction could be considered as interesting sites for a next and additional sampling.

High values of $\mathrm{var}(\hat{y})_{\mathrm{BS}}$ were also observed at the boundary of the fields (Figure 7a). As the bootstrap includes a random sample of the spectra among the pixels of the grid of (3×3) pixels centered of the studied pixel **x** (Figure 3), the $\mathrm{var}(\hat{y})_{\mathrm{BS}}$ uncertainty expression

Author-produced version of the article published in Remote Sensing of Environment, 2015, N°156, p. 58-70.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0034425714003861
Doi: 10.1016/j.rse.2014.09.032

integrates the spectral heterogeneity of the studied pixels. The map of term *T1*, even restricted to a small range, did not highlight particular areas with high uncertainties (map not shown). So the uncertainties seem to be due to the PLSR models, more than to the spectral heterogeneity of the neighborhood. The $\mathrm{var}(\hat{y})_{\mathrm{BS}}$ and $\mathrm{var}(\hat{y})_{\mathrm{Ahumada}}$ maps which are very similar (Figure 7a and d) confirm the high correlation found between both uncertainty expressions from the validation database (Table 2). So the uncertainty expression developed by Ahumada-Fernandez et al. (2012) expressed here the same trends than the "true" variance.

[Figure 5]

[Figure 6]

[Figure 7]

5.4 Analysis of selected pixels

To understand the sources of uncertainty and their links, the uncertainty expressions of 6 test-pixels were analyzed (Figure 2). Three of them are located on the center of bare soil fields for which homogeneous surface conditions might be supposed. And three of them are located on field boundaries for which heterogeneous surface conditions might be supposed.

The pixel *Pix_Soil1* has low values of uncertainties, whatever the expressions (*MD*, *H*, *T1*, *T2*, *T3*, $\mathrm{var}(\hat{y})_{\mathrm{BS}}$ and $\mathrm{var}(\hat{y})_{\mathrm{Ahumada}}$, Table 3). This pixel may represent the ideal pixel, a "no risky" pixel. It is located in the center of a bare soil field, in an area with stable surface conditions which involves stable spectral conditions. The uncertainties *MD* and *H* are weak, which means that the spectrum **x** of *Pix_Soil1* might be close to spectra of the calibration data base (verified but not shown). The uncertainty *T1* of *Pix_Soil1* is weak with a low value of $\mathrm{tr}(\mathbf{\Sigma_x})$ (0.2), which means that the neighbor spectra are weakly variable (verified but not shown), and the uncertainty $\mathbf{\Sigma_x}$ on the spectrum **x** is not amplified by the model coefficients **b**. Finally, the uncertainty *T2* of the model is weak with a less intense spectrum **z**, and it is not amplified by the model variance covariance matrix $\mathbf{\Sigma_b}$.

The pixel *Pix_Urban* has high values of uncertainties, whatever the expressions (*MD*, *H*, *T1*, *T2*, *T3*, $\mathrm{var}(\hat{y})_{\mathrm{BS}}$ and $\mathrm{var}(\hat{y})_{\mathrm{Ahumada}}$, Table 3). This pixel which is located in an urban area, represents an "incorrectly masked" pixel. The uncertainties *MD* and *H* are high, which means that the spectrum **x** of *Pix_Urban* might be far from spectra of the calibration data base. Indeed, the albedo of *Pix_Urban* spectrum is twice as high as spectra of the five others test-spectra (Figure 8) and the shape of this spectrum does not correspond to a soil spectrum, which can explain the high values of uncertainty expressions *MD*, *H* and *T2*. Located in an

Author-produced version of the article published in Remote Sensing of Environment, 2015, N°156, p. 58-70.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0034425714003861
Doi: 10.1016/j.rse.2014.09.032

urban area with unstable surface conditions on 15x15m, the neighbor spectra of *Pix_Urban* are more variable which explain that $\mathbf{\Sigma_x}$ and *T1* are high.

The pixel *Pix_BoundaryField1* has high *T1*, whereas the six other uncertainty expressions are weak (Table 3). This pixel which is located at the boundary between two fields with different surface soil conditions (due to plowed, rugosity, dry and/or green vegetation …) may represent a "to be monitored" pixel. The high value of *T1* which is related to the neighbor spectra variability and the amplification of $\mathbf{\Sigma_x}$ by the model coefficients $\mathbf{b}$, is a logical result of this location.

The pixel *Pix_Soil2*, even though located in a center of bare soil field as *Pix_Soil1* (Figure 2), has high *T2* and $\mathrm{var}(\hat{y})_{\mathrm{BS}}$, whereas the five others uncertainty expressions are weak (Table 3). So high uncertainties linked to the PLSR models, are revealed for the clay prediction of this pixel *Pix_Soil2*. As seen in section 5.2, *H* is a simplified version of *T2*, which expresses the uncertainty on model building (Um) over the hypothesis that the only source of uncertainty comes from *y*. For this pixel, the term *T2*, which takes into account the colinearity between $\mathbf{z}$ and $\mathbf{\Sigma_b}$, reveals more uncertainty than classical expression *H* (which is under the threshold of 0.2, Table 3). So the spectrum $\boldsymbol{x}$ associated to the pixel *Pix_Soil2*, although associated to low spectral distances *H* and *MD* to the calibration soil spectra, is highly sensitive to the PLSR models.

The pixel *Pix_Soil3* has moderate *T1* uncertainty, whereas the six others uncertainty expressions are weak (Table 3), as obtained for *Pix_BoundaryField1*. This moderate uncertainty *T1* leads to focus the attention to the neighbor spectra. $\mathrm{tr}(\mathbf{\Sigma_x})$ is around 0.5 so superior to the value obtained for *Pix_Soil1*, which means that the neighbor spectra of *Pix_Soil3* are more variable than neighbor spectra of *Pix_Soil1*. Moreover the uncertainty $\mathbf{\Sigma_x}$ on the spectrum $\mathbf{x}$ of *Pix_Soil3* is more amplified by the model coefficients $\mathbf{b}$ than uncertainty $\mathbf{\Sigma_x}$ on the spectrum $\mathbf{x}$ of *Pix_Soil1*. This amplification is due to higher albedo of neighbor spectra of *Pix_Soil3*, compared to the calibration soil spectra**.**

The pixel *Pix_BoundaryField2*, even though located in supposed heterogeneous surface conditions, has low values of uncertainties, whatever the expressions (*MD*, *H*, *T1*, *T2*, *T3*, $\mathrm{var}(\hat{y})_{\mathrm{BS}}$ and $\mathrm{var}(\hat{y})_{\mathrm{Ahumada}}$, Table 3), as obtained for *Pix_Soil1*. Due to the location at the boundary of fields, higher *T1* are expected as *Pix_BoundaryField1*. This low uncertainty *T1* is due to a weak difference of soil surfaces conditions, in particular a weak difference of vegetated conditions, between both fields on both sides of *Pix_BoundaryField2*. Indeed, the NDVI values of the masked pixels of the grid centered on the *Pix_BoundaryField2* are

Author-produced version of the article published in Remote Sensing of Environment, 2015, N°156, p. 58-70.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0034425714003861
Doi: 10.1016/j.rse.2014.09.032

included from 0.21 to 0.28, whereas those of the no-masked pixels of this grid are included from 0.18 to 0.19. For the comparison, the NDVI values of the masked pixels of the grid centered on the *Pix_BoundaryField1* are included from 0.21 to 0.28 (as *Pix_BoundaryField2)*, whereas those of the no-masked pixels of this grid are significantly lower (from 0.11 to 0.15), indicating different vegetated conditions.

[Figure 8]

[Table 3]

6. Discussion

Values of uncertainty expression *T2*, which is related to the uncertainty on model building (Um), are higher than values of uncertainty expression T1, which is related to the uncertainty on spectra due to the spatial positioning (Us) (Figure 4). These high values of uncertainty linked to model building (Um), may be explained by the modest performance of the PLSR model. Indeed as described in section 5.1, the clay prediction model obtained from the 96 AISA-DUAL spectra is characterized by R^2_{val} and RPD values around 0.7 and 1.4, respectively, and RMSEP value around 94g/kg (Table 1). So, these prediction models provide correct clay estimations, but cannot be qualified as accurate models. These model performances are in accordance with literature in which soil property estimated from VNIR/SWIR hyperspectral imagery data are less accurate than those estimated from Lab VNIR/SWIR data (e.g. Stevens et al., 2006, Lagacherie et al., 2008).

The pixel *Pix_Soil1* is representative of the “no risky” pixels. The uncertainty expressions obtained for the *Pix_Soil1* (low values whatever the uncertainty expression, Table 3) may be obtained for all the pixels of the AISA-DUAL image, if the mask process (section 2.3) would be well done. Although this mask process was done carefully in this work, respecting field expertise to characterize the vegetated areas, using urban maps to identify the urban areas, and using spectral knowledge to remove water areas, the mask is inaccurate and some studied pixels do not correspond to soil. The pixel *Pix_Urban* is representative of these “incorrectly masked” pixels. The uncertainty expressions obtained for the *Pix_Urban* (high values whatever the uncertainty expression, Table 3) might be also obtained for all the pixels located in urban area, over roads and vegetated areas... From this result, we could imagine a new mapping process in four steps: 1) application of the process described in Figure 3 for a soil property mapping over all the pixels, whatever the pixel compositions, associated to the uncertainty expressions calculation 2) analysis of the uncertainty expressions, would allow to identify the “no-soil” pixels, which correspond to those associated to the seven uncertainty

Author-produced version of the article published in Remote Sensing of Environment, 2015, N°156, p. 58-70.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0034425714003861
Doi: 10.1016/j.rse.2014.09.032

expressions superior to the thresholds, 3) mask of these "no-soil" pixels and 4) new application of the process described in Figure 3 for a soil property mapping over the "soil" pixels, associated to the uncertainty expressions calculation. At this stage, an analysis of the uncertainty expressions *T1*, *T2* and $var(\hat{y})_{BS}$ allows to identify the "to be monitored" pixels. This might allow new and better monitored soil sampling over the field, to take into account in the calibration database more soil variability.

Another alternative of this new mapping process (or an additional first step) would be to build first, the maps of *H* and *MD* uncertainty expressions. Rejecting all pixels with *H* and *MD* superior to a threshold defined by validation data study would allow masking major parts of urban areas (Figure 5). Both maps could be obtained faster than other uncertainty expressions maps, because no bootstrap is needed (Figure 3). This alternative would improve the mask process.

The VNIR/SWIR hyperspectral airborne imaging has been one of the emerging technologies selected for soil mapping and predicting soil properties by the Global Soil Map project GSM (www.globalsoilmap.net) (Lagacherie & Gomez, 2014). Indeed, the GSM project has proposed the construction of a new digital soil map of the world at a the spatial resolution of 90 m to assist in decision making for a range of global issues such as food production, climate change and environmental degradation (Mc Bratney et al., 2003, Sanchez et al., 2009). Digital Soil Mapping (DSM) approaches were recently developed for using the estimated soil properties maps restricted to bare soils, allowing the exhaustive mapping of some topsoil properties, based on block co-kriging methods (Lagacherie et al, 2012) and subsurface properties, based on statistical functions (Lagacherie et al, 2013). An improvement of the mask process thanks to the uncertainty maps, which is an unavoidable and recurrent step in remote sensing, might avoid misleading results and might improve results of these DSM approaches. Moreover, the uncertainty maps could be used to weight each pixel of the estimated soil properties maps, in the DSM approach, still to avoid misleading results and improve results.

7. Conclusion

The quality of a soil property map obtained by VNIR-SWIR hyperspectral imagery is usually assessed with an analysis of the regression model performances (calculation of the figures of merit whatever the model: PLSR, MLR …) associated to a visual pedological expertise of the soil property map and sometimes to geostatistical analysis to study the spatial structure of the predicted soil property. This paper shows the benefits of prediction uncertainties maps to

Author-produced version of the article published in Remote Sensing of Environment, 2015, N°156, p. 58-70.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0034425714003861
Doi: 10.1016/j.rse.2014.09.032

better mask the no-soil pixels, better define the soil sampling and so the calibration data set and better characterize the quality of the mapping results. In this way, this study uses twice information contained in the airborne hyperspectral data: *i)* the spectral information to create an estimated clay map, and *ii)* the spatial information to produce prediction uncertainties maps. To reinforce these conclusions and the guideline, it would be interesting and important to enlarge this study to additional soil properties (one-to-one and also together in prediction models), additional multivariate models, and additional pedological context.

In order to use the estimated soil property map as input data in environmental models (soil vulnerability to erosion, wheat yield in water limited situations…), the prediction uncertainties maps might provide confidence level on the input data. As well, in order to use the estimated soil property map as input data in Digital Soil Mapping approaches to exhaustive mapping, the prediction uncertainties maps might provide weighting information.

Finally, future hyperspectral satellite sensors (HYPXIM, PRIMSA, ENMAP and HyspIRI) will offer huge coverage of Earth surface. Using uncertainty maps on these data would help to build correct calibrations without too much sampling effort.

Acknowledgments

This research was granted by the TOSCA- CNES project « HUMPER - Mission HYPXIM : Apport de la résolution spatiale de la mission HYPXIM pour l'étude des propriétés pérennes des sols et de leur humidité de surface » (2013-2014). The authors are indebted to UMR LISAH (IRD, France) and to CNCT (Centre National de Cartographie et de Télédétection, Tunisia), for providing the AISA-Dual images for this study. This hyperspectral data acquisition was granted by IRD, INRA and the French National Research Agency (ANR) (ANR-O8-BLAN-C284-01) ». We are also indebted to Yves Blanca (IRD-UMR LISAH Montpellier), Zakia Jenhaoui (IRD-UMR LISAH Tunis) for the soil sampling in 2009 and 2010 over the Lebna catchment and to Hedi Hamrouni (DG/ACTA Sol, Tunis) for his significant support to this study.

References:

Baize, D., & Jabiol, B. (1995). Guide pour la description des sols. *INRA édition*, Paris.

Barnes, R.J., Dhanoa, M.S., and Lister, S.J. (1993). Correction to the description of standard normal variate (snv) and de-trend transformations in practical spectroscopy with applications in food and beverage analysis – 2nd edition. J. Near Infrared Spectrosc., 1:185–186.

Bayer, A., Bachmann, M., Müller, A., & Kaufmann, H. (2012). A Comparison of Feature-Based MLR and PLS Regression Techniques for the Prediction of Three Soil Constituents in a Degraded South African Ecosystem. *Applied and Environmental Soil Science*, vol. 2012, Article ID 971252, 20 pages, 2012. doi:10.1155/2012/971252.

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, & J.M., McBratney, A. (2010). Prediction of soil attributes by NIR spectroscopy. A critical review of chemometric indicators commonly used for assessing the quality of the prediction. *Trac-Trends in Analytical Chemistry*. 29 (9), 1073–1081.

Ben-Dor, E. & Banin, A. (1995). Near infrared analysis (NIRA) as a simultaneously method to evaluate spectral featureless constituents in soils. *Soil Science*, 159:259‑269.

Ben-Dor, E., Patkin, K., Banin, A. & Karnieli, A. (2002). Mapping of several soil properties using DAIS-7915 hyperspectral scanner data. A case study over clayey soils in Israel. *International Journal of Remote Sensing*. 23:1043-1062

Ben-Dor, E., Taylor, R.G., Hill, J., Demattê, J.AM., Whiting, M.L., Chabrillat, S., & Sommer, S. (2007). Imaging spectrometry for soil applications, *Agronomy Journal,* 97:323-381.

Budiman M. & McBratney, A.B. (2008). Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy, *Chemometrics and Intelligent Laboratory Systems*, 94(1), 72-79.

Cécillon, L., Barthès, B.G., Gomez, C., Ertlen, D., Genot, V., Hedde, M., Stevens, A., & Brun, J.J. (2009). Assessment and monitoring of soil conditions using indexes based on near infrared reflectance (NIR) spectroscopy. *European Journal of Soil Science,* 60, 770-784.

Chang, C.-W. & Laird, D.A., (2002). Near-infrared reflectance spectroscopic analysis of soil C and N. *Soil Science*, 167 (2), 110–116.

Chiang, L.H., Pell, R.J., & Seasholtz, M.B. (2003). Exploring process data with the use of robust outlier detection algorithms. *Journal of Process Control*, 13 (5), 437–449.

Denham, M. (1997) Prediction intervals in partial least squares. *Journal of Chemometrics*. 11 (1), 39–52.

Author-produced version of the article published in Remote Sensing of Environment, 2015, N°156, p. 58-70.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0034425714003861
Doi: 10.1016/j.rse.2014.09.032

Efron, B., & Efron, B. (1982). The jackknife, the bootstrap and other resampling plans (Vol. 38). Philadelphia: Society for industrial and applied mathematics.

Faber, N.K.M., & Kowalski B.R. (1997). Propagation of measurement errors for the validation of predictions obtained by principal component regression and partial least squares. *Journal of Chemometrics*, 11 (1997) 181-238

Faber, N.K.M., & Bro, R. (2002) Standard error of prediction for multiway PLS. 1. Background and a simulation study, *Chemometrics and Intelligent Laboratory Systems*, 61, 133—149.

Fernandez-Ahumada, E., Roger, J.M., & Palagos, B. (2012). A new formulation to estimate the variance of model prediction. Application to near infrared spectroscopy calibration. *Analytica Chimica Acta*. 721:28-34.

Gomez, C., Lagacherie, P., & Coulouma, G. (2008a). Continuum removal versus PLSR method for clay and calcium carbonate content estimation from laboratory and airborne hyperspectral measurements. *Geoderma* 148 (2), 141–148.

Gomez, C., Viscarra Rossel, R. A., & McBratney, A.B. (2008b). Soil organic carbon prediction by hyperspectral remote sensing and field VNIR/SWIR spectroscopy: an Australian case study. *Geoderma*, 146 (3-4), 403-411.

Gomez, C., Coulouma G., Lagacherie P. (2012a). Regional predictions of eight common soil properties and their spatial structures from hyperspectral Vis–NIR data, *Geoderma*, Volumes 189–190, November 2012, Pages 176-185**.**

Gomez C., Lagacherie P. & Bacha S. (2012b). Using an VNIR/SWIR hyperspectral image to map topsoil properties over bare soil surfaces in the Cap Bon region (Tunisia). In *"Digital Soil Assessments and Beyond"* Minasny B., Malone B.P., McBratney A.B. (Ed.). Springer. pp 387-392.

Höskuldsson A. (1988). PLS regression methods. *Journal of Chemometrics*. 2:211–228. 54.

Lagacherie, P., Baret, F., Feret, J-B, Madeira Netto, J., & Robbez-Masson, J.-M. (2008). Estimation of soil clay and calcium carbonate using laboratory, field and airborne hyperspectral measurements. *Remote Sensing of Environment*, 112 (3), 825-835.

Lagacherie, P., Bailly J.S., Monestiez P., & Gomez C. (2012). Using scattered hyperspectral imagery data to map the soil properties of a region, , *European Journal of Soil Science*, Volume 63, Number 1, p.110-119, (2012).

Lagacherie, P., Sneep, A.R., Gomez, C., Bacha, S., Coulouma, G., Hamrouni, M.H., & Mekki, I. (2013). Combining Vis-NIR hyperspectral imagery and legacy measured soil

Author-produced version of the article published in Remote Sensing of Environment, 2015, N°156, p. 58-70.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0034425714003861
Doi: 10.1016/j.rse.2014.09.032

profiles to map subsurface soil properties in a Mediterranean area (Cap-Bon, Tunisia). *Geoderma*, 209-210, 168-176.

Lagacherie, P. & Gomez, C. (2014). What can GlobalSoilMap expect from Vis-NIR hyperspectral imagery in the near future? In book: *GlobalSoilMap: Basis of the global spatial soil information system*, Publisher: CRC Press, Editors: Dominique Arrouays, Neil McKenzie, Jon Hempel, Anne Richer de Forges, Alex B. McBratney, pp.387-392

Lu, P., Wang, L., Niu, Z., Li, L., & Zhang, W. (2013). Prediction of soil properties using laboratory VIS–NIR spectroscopy and Hyperion imagery, *Journal of Geochemical Exploration*, Volume 132, September 2013, Pages 26-33

Mark, H.L., & Tunnell, D. (1985). Qualitative near infrared reflectance analysis using Mahalanobis distances. *Analytical Chemistry*, 57 (7), 1449–1456.

Martens, H. (1991). Multivariate calibration. John Wiley & Sons.

McBratney, A.., Mendonça Santos, M.., & Minasny, B. (2003). On digital soil mapping. *Geoderma*, 117, 3–52.

Pearson, R.K., (2002). Outliers in process modeling and identification. *IEEE Transactions on Control Systems Technology*, 10 (1), 55–63.

Phatak, A., Reilly, P. & Penlidis, A. (1993) An approach to interval estimation in partial least squares regression, *Analytica Chimica Acta*, 277(2), 495-501.

Sanchez, P.A., Ahamed, S., Carré, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., Mendonça-Santos, M.D.L., Minasny, B., Montanarella, L., Okoth, P., Palm, C.A., Sachs, J.D., Shepherd, K.D., Vågen, T.-G., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A., & Zhang, G.-L. (2009). Digital Soil Map of the World. *Science,* 325, 680–681.

Savitzky, A., & Golay, M.J.E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36 (8), 1627–1639.

Schwanghart, W., & Jarmer, T. (2011). Linking spatial patterns of soil organic carbon to topography - a case study from south-eastern Spain. *Geomorphology*, 126, 252-263.

Selige, T., Bohner, J., & Schmidhalter, U. (2006). High resolution topsoil mapping using hyperspectral image and field data in multivariate regression modeling procedures. *Geoderma*, 136, n°1-2, pp. 235-244

Stevens, A., Van Wesemael, B.,Vandenschrick, G., Touré, S. & Tychon, B. (2006) Detection of Carbon Stock Change in Agricultural Soils Using Spectroscopic Techniques, *Soil Science Society of America journal*, 70, 844–850.

Author-produced version of the article published in Remote Sensing of Environment, 2015, N°156, p. 58-70.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0034425714003861
Doi: 10.1016/j.rse.2014.09.032

762 Stevens, A., Udelhoven, T., Denis, A., Tychon, B., Lioy, R., Hoffmann, L., & Wesemael, B.
763 (2010). Measuring soil organic carbon in croplands at regional scale using airborne
764 imaging spectroscopy, *Geoderma*, 158, 1-2.
765 Udelhoven, T., Emmerling, C. & Jarmer T. (2003) Quantitative analysis of soil chemical
766 properties with diffuse reflectance spectrometry and partial least-square regression: A
767 feasibility study, Plant and Soil, 251: 319–329.
768 Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., & Skjemstad, J.O.,
769 2006. Visible, near-infrared, mid-infrared or combined diffuse reflectance spectroscopy
770 for simultaneous assessment of various soil properties. *Geoderma*, 131, 59–75.
771 Wold, S. (1978) Cross-validatory estimation of the number of components in factor and
772 principal components models. Technometrics, 20:397–405, 1978.
773 Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of Chemometrics.
774 *Chemometrics and Intelligent Laboratory Systems*, 58, 109-130.
775 Zante, P., Collinet, J. & Pepin, Y., 2005. Caractéristiques pédologiques et
776 hydrométéorologiques du bassin versant de Kamech, Cap Bon, Tunisie. *UMR LISAH*
777 *IRD Tunis, DG ACTA Direction des Sols Tunis, INRGREF Tunis*. (21 p. + 6 annexes).
778 Zeaiter, M., Roger, J.M., & Bellon-Maurel, V. (2004). Robustness of models developed by
779 multivariate calibration. Part I: The assessment of robustness. *Trends in Analytical*
780 *Chemistry*, Vol. 23, No. 2, 157–170.
781