



HAL
open science

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

Cecile Gomez, Arthur Drost, J.M. Roger

► **To cite this version:**

Cecile Gomez, Arthur Drost, J.M. Roger. Analysis of uncertainties affecting clay contents predictions obtained from airborne VNIR/SWIR hyperspectral data. *Remote Sensing of Environment*, 2015, 156, pp.58-70. 10.1016/j.rse.2014.09.032 . hal-01082834

HAL Id: hal-01082834

<https://hal.science/hal-01082834>

Submitted on 14 Nov 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

32 **1- Introduction**

33 Since two decades, laboratory Visible, Near-Infrared and Short Wave Infrared (VNIR/SWIR,
34 350-2500 nm) spectroscopy has been proven as a good alternative to costly physical and
35 chemical laboratory soil analysis for the estimation of a large range of soil properties (e.g.
36 [Ben-Dor & Banin, 1995](#), [Viscarra Rossel, 2006](#), [Cécillon et al., 2009](#)). These works opened
37 the way to VNIR/SWIR hyperspectral imaging spectroscopy for estimating soil properties.
38 VNIR/SWIR hyperspectral imaging spectroscopy may provide a global view of the area under
39 study at high spatial resolutions, and may provide characterization of several soil properties
40 simultaneously. [Ben Dor et al. \(2002\)](#) started this research way, with multivariate calibration
41 statistics applied to remotely-sensed data. Since then, the VNIR/SWIR hyperspectral imaging
42 spectroscopy benefits from an increasing number of methodologies developed in lab soil
43 property prediction studies, including Partial Least Square Regression (e.g. [Gomez et al.,](#)
44 [2008a](#)), Support Vector Machine Regression (e.g. [Stevens et al., 2010](#)), Multiple Linear
45 Regression (e.g. [Bayer et al., 2012](#)), Stepwise Multiple Linear Regression (e.g. [Lu et al.,](#)
46 [2013](#)) and Regression Rules ([Budiman & McBratney, 2008](#)). And several studies have been
47 successfully conducted to map soil properties such as Clay, Calcium Carbonate, Iron, Soil
48 Organic Carbon... (e.g., [Ben-Dor et al., 2007](#), [Selige et al., 2006](#), [Gomez et al., 2008b](#),
49 [Gomez et al., 2012a](#), [Stevens et al., 2010](#)).

50 Nevertheless, whatever the regression methods, the number of samples in the
51 calibration database, the scales of study or the studied soil properties, the quality of the
52 mapping results are expressed by using the performance of the models with global figures of
53 merit such as standard error of calibration (SEC) and standard error of prediction (SEP),
54 respectively referred to as Root Mean Squared Error of Calibration (RMSEC) and Root Mean
55 Squared Error of Prediction (RMSEP). The coefficient of Determination (R^2) and the Ratio of
56 Performance Deviation (RPD) are also used as a measure of multivariate model prediction
57 accuracy. The ratio of performance to interquartile (RPIQ), which is the ratio of the
58 interquartile ($IQ = Q3 - Q1$) to the RMSEP has been recently proposed as an alternative to the
59 RPD to better take into account the shape of the distribution ([Bellon-Maurel et al., 2010](#)). All
60 of these parameters are estimated during model building and validation stages to evaluate
61 global model performance. [Selige et al. \(2006\)](#) used the Partial Least Square Regression
62 (PLSR) method with 72 soil samples in their calibration database to map Sand content over 7
63 km², from HYMAP VNIR/SWIR hyperspectral data. [Gomez et al. \(2012b\)](#) used the PLSR
64 method with 95 soil samples in their calibration database to map Clay content over 300 km²,
65 from AISA-DUAL VNIR/SWIR hyperspectral data. And both researches expressed the

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

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

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

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

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

113 This process of uncertainty analysis has been performed on recent study of a clay
114 content mapping by VNIR/SWIR AISA-Dual airborne data over a Tunisia area ([Gomez et al.,](#)
115 [2012b](#)). First, all the uncertainty expressions are calculated and analyzed on a validation
116 database. Secondly, all the uncertainty expressions are calculated, mapped and analyzed on a
117 test area and different specific cases highlighted some specific interest of these uncertainty
118 expressions.

120 **2. Materials**

122 **2.1 Notations**

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

130 **2.2 Site description**

131 The study area is located in the Cap Bon region in northern Tunisia (36°24'N to 36°53'N;
132 10°20'E to 10°58'E), 60 km east of Tunis, Tunisia ([Figure 1](#)). This 300 km² area includes the
133 Lebna catchment, which is mainly rural (>90%) and devoted to cereals in addition to legumes,

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

148 [Figure 1]

150 2.3 VNIR/SWIR hyperspectral data

151 On November 2, 2010, an AISA-Dual hyperspectral airborne image was acquired for the
152 study area with a spatial resolution of 5 m ([Figure 1b](#)). The area of the image is approximately
153 12 km x 24 km. The AISA-Dual spectrometer measured the reflected radiance in 359 non-
154 contiguous bands covering the 400- to 2450-nm spectral domain, with 4.6 nm bandwidths
155 between 400 and 970 nm and 6.5 nm bandwidths between 970 and 2450 nm. The
156 instantaneous field of view (IFOV) was 24 degrees. The radiance units were converted to
157 reflectance units using ASD spectrometer measurements of uniform surfaces (parking lots,
158 asphalt, concrete) that were collected at the same time during the over flight. An empirical
159 line correction method was used to calibrate each flight line to the reflectance. Topographic
160 corrections were performed using a digital elevation model built from the ASTER data and
161 ground control points. In this study, we removed 1) the spectral bands in the blue part of the
162 spectral domain (between 400 and 484 nm) due to noise in these bands and 2) the spectral
163 bands between 1339 and 1464 nm as well as between 1772 and 2004 nm due to vibrational-
164 rotational H₂O absorption bands. Consequently, 280 AISA-Dual spectral bands were
165 retained.

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

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

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

184 [Figure 2]

186 2.4 Soil samples data base

187 129 soil samples were collected on the Lebna catchment. Of these samples, 58 were collected
188 in October 2008, 30 in October 2009, and 41 in November 2010. All of the samples were
189 composed of five sub-samples collected to a depth of 5 cm at random locations within a
190 10×10 m square centered on the geographical position of the sampling plot, as recorded by a
191 Garmin GPS instrument. All of these soil samples were collected in fields that were bare
192 during the hyperspectral data acquisition in November 2010, between 20 and 180 m of
193 altitude. After homogenizing the sample, approximately 20 g was devoted to soil property
194 analysis. The initial samples were air-dried and sieved with a 2 mm sieve prior to being
195 transported to the laboratory for analysis. The determination of clay content (granulometric
196 fraction < 2 μm) was determined according to the method NF X 31-107 (Baize and Jabiol, 1995).
197 The clay content of the 129 soil samples varies between 46 and 777g/kg and follows a normal
198 distribution.

200 3. Prediction Model: PLSR

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

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

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

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

231
232 In this paper, we only provide a brief description of the PLSR model that is fully
233 detailed in Wold et al. (2001). The PLSR model is developed from a training set of N
234 observations (number of spectra in the calibration dataset) with K \mathbf{X} -variables (number of

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

247
248 The prediction performances of PLSR models were evaluated using the coefficient of
249 determination R^2_{cal} and R^2_{val} of predicted against measured values in the calibration and
250 validation set respectively. The root mean square errors of calibration (RMSEC) and the root
251 mean square errors in the validation set (RMSEP) were also analyzed for all the models. The
252 ratio of performance to deviation (RPD), which is the ratio of the standard deviation in the
253 validation set to the RMSEP, was used as an index of model accuracy.

254 255 **4. Measurements of uncertainty affecting predictions**

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

$$257 \quad \hat{y} = f(\mathbf{Xc}, \mathbf{Yc}, \text{Model}, \mathbf{x}) \quad (1)$$

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

- 261 - Uncertainty on spectra was assumed to be identical for calibration and test spectra. It
262 is mainly due to the device repeatability (U_d) and to the spatial positioning (U_s)
- 263 - Uncertainty on reference lab values (U_y)
- 264 - Uncertainty on model building (U_m) may originate from 2 main causes, with $U_m =$
265 $U_c + U_l$. The first one (U_c) is related to the choice made for building the calibration
266 set. The second one (U_l) is related to the choice of model dimension.

267 In this study, the following choices were made: U_d was considered negligible in comparison
268 with U_s and U_y was neglected. U_s was modeled as a variance-covariance matrix Σ_x computed
269 on the 9 point neighborhood of \mathbf{x} . U_c and U_l were modeled and merged in a unique variance-
270 covariance matrix Σ_b computed by means of bootstrap (Efron, 1982).

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

273 274 **4.1. Bootstrap procedure**

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

278 1st step: N drawing with replacement of N samples among the calibration dataset.

279 2^{sd} step: For each of the N selected samples, the associated AISA-Dual spectrum was
280 randomly sampled among a grid of 3x3 pixels, centered on the location of the
281 selected samples. To favor the central pixel, the sampling followed a normal
282 distribution in line and column, with a standard deviation of 0.6.

283 3rd step: Drawing the number of latent variable, from 3 to 7, following a normal
284 distribution centered on 5 with a standard deviation of 0.97.

285 4th step: For each new AISA-Dual spectrum \mathbf{x} for which is looked for the clay content
286 estimated value, the associated AISA-Dual spectrum was randomly sampled among a
287 grid of 3x3 pixels, centered on the pixel \mathbf{x} . To favor the central pixel, the sampling
288 followed a normal distribution in line and column, with a standard deviation of 0.6.
289 This 4th step is the same for the validation dataset *BD_Valid*, than for the entire
290 image.

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

298 299 **4.2. Expressions of the uncertainty using bootstrap**

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

$$\text{var}(\hat{\mathbf{y}})_{BS} = U_d + U_s + U_y + U_m \quad (2)$$

308 4.2 Expressions of the uncertainty using spectral distance

309 a. Mahalanobis Distance

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

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

317 Where \mathbf{x} ($K \times 1$) is a VNIR/SWIR spectrum, $\bar{\mu}$ ($K \times 1$) is the mean of the N calibration
318 VNIR/SWIR spectra and \mathbf{S} ($K \times K$) is the variance-covariance matrix of the N calibration
319 spectra. The principle of the Mahalanobis distance is described in detail in [Mark and Tunnell, 1985](#).
320 Generally, a MD value superior to 3 is considered to reflect an outlier with regard to the
321 calibration set ([Mark and Tunnell, 1985](#)). The MD is related to the uncertainty on model
322 building (U_m), over the hypothesis that the PCA space is close to the PLSR space.
323

325 b. Leverage

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

333
$$H(\mathbf{x}) = \mathbf{t}_x' (\mathbf{T}'\mathbf{T})^{-1} \mathbf{t}_x \quad (4)$$

334 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
 335 multivariate model, and \mathbf{T} ($N \times A$) is the \mathbf{X} -scores matrix of the multivariate model (see
 336 section 3). The leverage is related to the uncertainty on model building (Um).
 337

338 **4.3 Expressions of the uncertainty using the Fernandez-Ahumada et al. (2012)**
 339 **formula**

340 **a. Total Variance of predictions**

341 As a measure of uncertainty of the estimations by PLSR model, a variance model, proposed
 342 by Fernandez-Ahumada et al. (2012), has been applied in this study. This proposal for
 343 variance modeling is based on the EIV-model initially proposed by Faber and Kowalski
 344 (1997), which considers all sources of uncertainty affecting predictors, the dependent
 345 variables and model coefficients. The difference between the classical EIV-model and the
 346 new variance model proposed by Fernandez-Ahumada et al. (2012) is that the latter model is
 347 constructed based on assumptions related to the predictors, the spectral bands. The
 348 formulation is described in detail in Fernandez-Ahumada et al. (2012), and can be resumed
 349 by:

350
$$\text{var}(\hat{y})_{\text{Ahumada}} = \left(1 + \frac{1}{N}\right) \mathbf{b}'\boldsymbol{\Sigma}_x\mathbf{b} + \mathbf{z}'\boldsymbol{\Sigma}_b\mathbf{z} + \left(1 + \frac{1}{N}\right) \text{tr}(\boldsymbol{\Sigma}_x\boldsymbol{\Sigma}_b) + \frac{\sigma_{lab}^2}{N} \quad (5)$$

351 Where \hat{y} is the prediction, \mathbf{z} is the centered \mathbf{x} spectrum used for the prediction, \mathbf{b} is the K -
 352 vector of the b-coefficients of the model, N is the number of the calibration samples, $\boldsymbol{\Sigma}_b$ is the
 353 variance-covariance matrix of the b-coefficients, $\boldsymbol{\Sigma}_x$ is the variance-covariance matrix of the
 354 spectrum \mathbf{x} , and σ_{lab}^2 is the laboratory y variance,.

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

360
$$\text{var}(\hat{y})_{\text{Ahumada}} = \left(1 + \frac{1}{N}\right) \mathbf{b}'\boldsymbol{\Sigma}_x\mathbf{b} + \mathbf{z}'\boldsymbol{\Sigma}_b\mathbf{z} + \left(1 + \frac{1}{N}\right) \text{tr}(\boldsymbol{\Sigma}_x\boldsymbol{\Sigma}_b) \quad (6)$$

361 Each term of the equation (6) and their calculation are described in the following sections.
 362

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

364 The first term of the formula of [Fernandez-Ahumada et al. \(2012\)](#) (denoted $T1$, [Figure 3d](#))
365 expresses the variance of predictions caused by the uncertainty of hyperspectral data \mathbf{x} for
366 each pixel:

$$367 \quad T1 = \left(1 + \frac{1}{N}\right) \mathbf{b}' \boldsymbol{\Sigma}_{\mathbf{x}} \mathbf{b} \quad (7)$$

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

$$372 \quad \boldsymbol{\Sigma}_{\mathbf{x}} = \text{cov}(\mathbf{G}(\mathbf{x})) \quad (8)$$

373 Where $\mathbf{G}(\mathbf{x})$ is a $(9 \times K)$ matrix where each line is a spectrum belongs to the grid of 3x3
374 pixels centered on the spectrum \mathbf{x} . So the matrix $\boldsymbol{\Sigma}_{\mathbf{x}}$ ($K \times K$) attempts to reflect the influence
375 of surrounding areas on reflection values of the center pixel \mathbf{x} ([Figure 3c](#)).

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

384 c. Variance of predictions due to the PLSR model

385 The second term of the formula of [Fernandez-Ahumada et al. \(2012\)](#) (denoted $T2$, [Figure 3d](#))
386 expresses the variance of predictions due to the uncertainty on the multivariate model:

$$387 \quad T2 = \mathbf{z}' \boldsymbol{\Sigma}_{\mathbf{b}} \mathbf{z} \quad (9)$$

388 Where \mathbf{z} is the centered spectrum of \mathbf{x} and $\boldsymbol{\Sigma}_{\mathbf{b}}$ is the variance-covariance matrix ($K \times K$) of
389 the b-coefficients. In this study, the variance-covariance matrix $\boldsymbol{\Sigma}_{\mathbf{b}}$ was calculated using the b-
390 coefficients obtained at each of the R iterations of the bootstrap ([Figure 3a](#)), following:

$$391 \quad \boldsymbol{\Sigma}_{\mathbf{b}} = \text{cov}(\mathbf{B}) \quad (10)$$

392 Where \mathbf{B} is the matrix ($R \times K$) where the i^{th} line contains the b-coefficients of the i^{th} iteration
393 of the bootstrap. This second term is considered as a distance between the spectrum and the
394 model center, weighted by the model noise. It depends mainly on three factors: (i) the length
395 of the centred spectra \mathbf{z} , which is related to the classical concept of leverage, (ii) the norm of
396 $\boldsymbol{\Sigma}_{\mathbf{b}}$ and (iii) the colinearity between \mathbf{z} and $\boldsymbol{\Sigma}_{\mathbf{b}}$. The term $T2$ is related to the uncertainty on

397 model building (U_m). It reflects how the uncertainty of the model is amplified by the
398 spectrum \mathbf{x} .

400 **d. Variance of predictions due to the intersection of the multivariate model** 401 **and the spectrum variances**

402 The third term of the formula of [Fernandez-Ahumada et al. \(2012\)](#) (denoted $T3$, [Figure 3d](#))
403 expresses the intersection of the variance of predictions due to the multivariate model and the
404 spectrum \mathbf{x} :

$$405 \quad T3 = \left(1 + \frac{1}{N}\right) \text{tr}(\Sigma_{\mathbf{x}}\Sigma_{\mathbf{b}}) \quad (11)$$

406 The quantity $\text{tr}(\Sigma_{\mathbf{x}}\Sigma_{\mathbf{b}})$ measures the common part of the instances of $\Sigma_{\mathbf{x}}$ and $\Sigma_{\mathbf{b}}$ ([Fernandez-](#)
407 [Ahumada et al., 2012](#)). A low value of the term $T3$ would signify that the uncertainty of the
408 model and the spectrum, $\Sigma_{\mathbf{b}}$ and $\Sigma_{\mathbf{x}}$ respectively, are different and could cancel each other out.
409 A high value of the term $T3$ would signify that the uncertainty of the model and the spectrum,
410 $\Sigma_{\mathbf{b}}$ and $\Sigma_{\mathbf{x}}$ respectively, are similar and could be accumulated.

411 [\[Figure 3\]](#)

413 **5. Results**

414 **5.1 Global performances of the models – preliminary results**

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

421 Following the classes of RPD defined by [Chang and Laird \(2002\)](#), a correct prediction
422 was obtained by PLSR model from the 96 AISA-DUAL spectra for clay content prediction,
423 with R^2_{val} and RPD values greater than 0.7 and 1.4, respectively, and with RMSEP around
424 94g/kg ([Table 1](#)). A correct prediction was also obtained by bootstrap-PLSR models with R^2_{val}
425 and RPD values greater than 0.5 and 1.4, respectively, and with RMSEP around 109 g/kg
426 ([Table 1](#)).

427 [\[Table 1\]](#)

429 **5.2 Uncertainty analysis on validation data set**

430 The seven uncertainty expressions described in section 4 have been firstly calculated for the
431 32 samples of the BD_Valid. The correlations between these uncertainty expressions (Table
432 2) and their ranges (Figure 4) have been studied.

433 The variance model $\text{var}(\hat{y})_{\text{Ahumada}}$ is very highly correlated to $T2$ ($R \approx 1$, Table 2).
434 So the most part of $\text{var}(\hat{y})_{\text{Ahumada}}$ is due to the uncertainty on the multivariate model.
435 Moreover, the variance $\text{var}(\hat{y})_{\text{BS}}$ of the R predictions for each AISA-DUAL spectra of the
436 validation data set is highly correlated to the variance model $\text{var}(\hat{y})_{\text{Ahumada}}$ and to the term
437 $T2$ ($R = 0.91$, Table 2). So the “true” variance of predictions, expressed by $\text{var}(\hat{y})_{\text{BS}}$ is mostly
438 due to the uncertainty on the multivariate model, and in minority to the uncertainty on the
439 spectra.

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

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

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

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

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

465 [Table 2]

466 [Figure 4]

467 468 **5.3 Uncertainty mapping on Kamech area**

469 The higher values of the seven uncertainty expressions were gathered on the East part of the
470 area, which may correspond to some isolated houses (black rectangle on Figure 5a). Urban
471 pixels have been masked in the pre-treatment of the AISA-DUAL data, using an urban vector
472 layer of the region which took into account the bigger towns of the region. But some isolated
473 houses, not referenced in the urban map of the region, could be missed by this mask pre-
474 treatment and seem to be identified by each of these uncertainty maps (Figure 5 and 6). Over
475 the 10750 studied pixels, only 108 pixels had the combination $MD > 3$, $H > 0.2$, $T2 > 6000$,
476 $\text{var}(\hat{y})_{BS} > 5000$ and $\text{var}(\hat{y})_{Ahumada} > 5000$. So the thresholds defined from the validation
477 database may be used to mask some missed no-soil area.

478 The higher values of TI (except those on the urban area) were located at the boundary
479 of the fields (Figure 6a) which is coherent. Indeed these high values of TI were the
480 consequence of the calculation of Σ_x which used the grid of (3×3) pixels centered of the
481 studied pixel associated to the spectrum x . If the studied pixel associated to the spectrum x is
482 located on the boundary of the field, its Σ_x will include the heterogeneity of the boundary
483 (field, ditch, road or path).

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

494 High values of $\text{var}(\hat{y})_{BS}$ were also observed at the boundary of the fields (Figure 7a).
495 As the bootstrap includes a random sample of the spectra among the pixels of the grid of
496 (3×3) pixels centered of the studied pixel x (Figure 3), the $\text{var}(\hat{y})_{BS}$ uncertainty expression

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

504 [Figure 5]

505 [Figure 6]

506 [Figure 7]

507 508 **5.4 Analysis of selected pixels**

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

513 The pixel *Pix_Soill* has low values of uncertainties, whatever the expressions (MD , H ,
514 $T1$, $T2$, $T3$, $\text{var}(\hat{y})_{\text{BS}}$ and $\text{var}(\hat{y})_{\text{Ahumada}}$, Table 3). This pixel may represent the ideal pixel, a
515 “no risky” pixel. It is located in the center of a bare soil field, in an area with stable surface
516 conditions which involves stable spectral conditions. The uncertainties MD and H are weak,
517 which means that the spectrum \mathbf{x} of *Pix_Soill* might be close to spectra of the calibration data
518 base (verified but not shown). The uncertainty $T1$ of *Pix_Soill* is weak with a low value of
519 $\text{tr}(\Sigma_{\mathbf{x}})$ (0.2), which means that the neighbor spectra are weakly variable (verified but not
520 shown), and the uncertainty $\Sigma_{\mathbf{x}}$ on the spectrum \mathbf{x} is not amplified by the model coefficients \mathbf{b} .
521 Finally, the uncertainty $T2$ of the model is weak with a less intense spectrum \mathbf{z} , and it is not
522 amplified by the model variance covariance matrix $\Sigma_{\mathbf{b}}$.

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

530 urban area with unstable surface conditions on 15x15m, the neighbor spectra of *Pix_Urban*
531 are more variable which explain that Σ_x and *T1* are high.

532 The pixel *Pix_BoundaryField1* has high *T1*, whereas the six other uncertainty
533 expressions are weak (Table 3). This pixel which is located at the boundary between two
534 fields with different surface soil conditions (due to plowed, rugosity, dry and/or green
535 vegetation ...) may represent a “to be monitored” pixel. The high value of *T1* which is related
536 to the neighbor spectra variability and the amplification of Σ_x by the model coefficients **b**, is a
537 logical result of this location.

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

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

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

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

568 [Figure 8]

569 [Table 3]

570 571 **6. Discussion**

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

583 The pixel *Pix_Soill* is representative of the “no risky” pixels. The uncertainty
584 expressions obtained for the *Pix_Soill* (low values whatever the uncertainty expression, Table
585 3) may be obtained for all the pixels of the AISA-DUAL image, if the mask process (section
586 2.3) would be well done. Although this mask process was done carefully in this work,
587 respecting field expertise to characterize the vegetated areas, using urban maps to identify the
588 urban areas, and using spectral knowledge to remove water areas, the mask is inaccurate and
589 some studied pixels do not correspond to soil. The pixel *Pix_Urban* is representative of these
590 “incorrectly masked” pixels. The uncertainty expressions obtained for the *Pix_Urban* (high
591 values whatever the uncertainty expression, Table 3) might be also obtained for all the pixels
592 located in urban area, over roads and vegetated areas... From this result, we could imagine a
593 new mapping process in four steps: 1) application of the process described in Figure 3 for a
594 soil property mapping over all the pixels, whatever the pixel compositions, associated to the
595 uncertainty expressions calculation 2) analysis of the uncertainty expressions, would allow to
596 identify the “no-soil” pixels, which correspond to those associated to the seven uncertainty

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

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

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

624 625 **7. Conclusion**

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

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

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

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

648 **Acknowledgments**

649 This research was granted by the TOSCA- CNES project « HUMPER - Mission HYPXIM :
650 Apport de la résolution spatiale de la mission HYPXIM pour l'étude des propriétés pérennes
651 des sols et de leur humidité de surface » (2013-2014). The authors are indebted to UMR
652 LISAH (IRD, France) and to CNCT (Centre National de Cartographie et de Télédétection,
653 Tunisia), for providing the AISA-Dual images for this study. This hyperspectral data
654 acquisition was granted by IRD, INRA and the French National Research Agency (ANR)
655 (ANR-O8-BLAN-C284-01) ». We are also indebted to Yves Blanca (IRD-UMR LISAH
656 Montpellier), Zakia Jenhaoui (IRD-UMR LISAH Tunis) for the soil sampling in 2009 and
657 2010 over the Lebna catchment and to Hedi Hamrouni (DG/ACTA Sol, Tunis) for his
658 significant support to this study.

662 **References:**

- 663 Baize, D., & Jabiol, B. (1995). Guide pour la description des sols. *INRA édition*, Paris.
- 664 Barnes, R.J., Dhanoa, M.S., and Lister, S.J. (1993). Correction to the description of standard
665 normal variate (snv) and de-trend transformations in practical spectroscopy with
666 applications in food and beverage analysis – 2nd edition. *J. Near Infrared Spectrosc.*,
667 1:185–186.
- 668 Bayer, A., Bachmann, M., Müller, A., & Kaufmann, H. (2012). A Comparison of Feature-
669 Based MLR and PLS Regression Techniques for the Prediction of Three Soil
670 Constituents in a Degraded South African Ecosystem. *Applied and Environmental Soil
671 Science*, vol. 2012, Article ID 971252, 20 pages, 2012. doi:10.1155/2012/971252.
- 672 Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, & J.M., McBratney, A.
673 (2010). Prediction of soil attributes by NIR spectroscopy. A critical review of
674 chemometric indicators commonly used for assessing the quality of the prediction.
675 *Trac-Trends in Analytical Chemistry*. 29 (9), 1073–1081.
- 676 Ben-Dor, E. & Banin, A. (1995). Near infrared analysis (NIRA) as a simultaneously method
677 to evaluate spectral featureless constituents in soils. *Soil Science*, 159:259-269.
- 678 Ben-Dor, E., Patkin, K., Banin, A. & Karnieli, A. (2002). Mapping of several soil properties
679 using DAIS-7915 hyperspectral scanner data. A case study over clayey soils in Israel.
680 *International Journal of Remote Sensing*. 23:1043-1062
- 681 Ben-Dor, E., Taylor, R.G., Hill, J., Demattê, J.A.M., Whiting, M.L., Chabrilat, S., & Sommer,
682 S. (2007). Imaging spectrometry for soil applications, *Agronomy Journal*, 97:323-381.
- 683 Budiman M. & McBratney, A.B. (2008). Regression rules as a tool for predicting soil
684 properties from infrared reflectance spectroscopy, *Chemometrics and Intelligent
685 Laboratory Systems*, 94(1), 72-79.
- 686 Cécillon, L., Barthès, B.G., Gomez, C., Ertlen, D., Genot, V., Hedde, M., Stevens, A., &
687 Brun, J.J. (2009). Assessment and monitoring of soil conditions using indexes based on
688 near infrared reflectance (NIR) spectroscopy. *European Journal of Soil Science*, 60,
689 770-784.
- 690 Chang, C.-W. & Laird, D.A., (2002). Near-infrared reflectance spectroscopic analysis of soil
691 C and N. *Soil Science*, 167 (2), 110–116.
- 692 Chiang, L.H., Pell, R.J., & Seasholtz, M.B. (2003). Exploring process data with the use of
693 robust outlier detection algorithms. *Journal of Process Control*, 13 (5), 437–449.
- 694 Denham, M. (1997) Prediction intervals in partial least squares. *Journal of Chemometrics*. 11
695 (1), 39–52.

- 696 Efron, B., & Efron, B. (1982). The jackknife, the bootstrap and other resampling plans (Vol.
697 38). Philadelphia: Society for industrial and applied mathematics.
- 698 Faber, N.K.M., & Kowalski B.R. (1997). Propagation of measurement errors for the
699 validation of predictions obtained by principal component regression and partial least
700 squares. *Journal of Chemometrics*, 11 (1997) 181-238
- 701 Faber, N.K.M., & Bro, R. (2002) Standard error of prediction for multiway PLS. 1.
702 Background and a simulation study, *Chemometrics and Intelligent Laboratory Systems*,
703 61, 133—149.
- 704 Fernandez-Ahumada, E., Roger, J.M., & Palagos, B. (2012). A new formulation to estimate
705 the variance of model prediction. Application to near infrared spectroscopy calibration.
706 *Analytica Chimica Acta*. 721:28-34.
- 707 Gomez, C., Lagacherie, P., & Coulouma, G. (2008a). Continuum removal versus PLSR
708 method for clay and calcium carbonate content estimation from laboratory and airborne
709 hyperspectral measurements. *Geoderma* 148 (2), 141–148.
- 710 Gomez, C., Viscarra Rossel, R. A., & McBratney, A.B. (2008b). Soil organic carbon
711 prediction by hyperspectral remote sensing and field VNIR/SWIR spectroscopy: an
712 Australian case study. *Geoderma*, 146 (3-4), 403-411.
- 713 Gomez, C., Coulouma G., Lagacherie P. (2012a). Regional predictions of eight common soil
714 properties and their spatial structures from hyperspectral Vis–NIR data, *Geoderma*,
715 Volumes 189–190, November 2012, Pages 176-185.
- 716 Gomez C., Lagacherie P. & Bacha S. (2012b). Using an VNIR/SWIR hyperspectral image to
717 map topsoil properties over bare soil surfaces in the Cap Bon region (Tunisia). In
718 “*Digital Soil Assessments and Beyond*” Minasny B., Malone B.P., McBratney A.B.
719 (Ed.). Springer. pp 387-392.
- 720 Höskuldsson A. (1988). PLS regression methods. *Journal of Chemometrics*. 2:211–228. 54.
- 721 Lagacherie, P., Baret, F., Feret, J-B, Madeira Netto, J., & Robbez-Masson, J.-M. (2008).
722 Estimation of soil clay and calcium carbonate using laboratory, field and airborne
723 hyperspectral measurements. *Remote Sensing of Environment*, 112 (3), 825-835.
- 724 Lagacherie, P., Bailly J.S., Monestiez P., & Gomez C. (2012). Using scattered hyperspectral
725 imagery data to map the soil properties of a region, , *European Journal of Soil Science*,
726 Volume 63, Number 1, p.110-119, (2012).
- 727 Lagacherie, P., Sneepe, A.R., Gomez, C., Bacha, S., Coulouma, G., Hamrouni, M.H., &
728 Mekki, I. (2013). Combining Vis-NIR hyperspectral imagery and legacy measured soil

- 729 profiles to map subsurface soil properties in a Mediterranean area (Cap-Bon, Tunisia).
730 *Geoderma*, 209-210, 168-176.
- 731 Lagacherie, P. & Gomez, C. (2014). What can GlobalSoilMap expect from Vis-NIR
732 hyperspectral imagery in the near future? In book: *GlobalSoilMap: Basis of the global*
733 *spatial soil information system*, Publisher: CRC Press, Editors: Dominique Arrouays,
734 Neil McKenzie, Jon Hempel, Anne Richer de Forges, Alex B. McBratney, pp.387-392
- 735 Lu, P., Wang, L., Niu, Z., Li, L., & Zhang, W. (2013). Prediction of soil properties using
736 laboratory VIS–NIR spectroscopy and Hyperion imagery, *Journal of Geochemical*
737 *Exploration*, Volume 132, September 2013, Pages 26-33
- 738 Mark, H.L., & Tunnell, D. (1985). Qualitative near infrared reflectance analysis using
739 Mahalanobis distances. *Analytical Chemistry*, 57 (7), 1449–1456.
- 740 Martens, H. (1991). Multivariate calibration. John Wiley & Sons.
- 741 McBratney, A., Mendonça Santos, M., & Minasny, B. (2003). On digital soil mapping.
742 *Geoderma*, 117, 3–52.
- 743 Pearson, R.K., (2002). Outliers in process modeling and identification. *IEEE Transactions on*
744 *Control Systems Technology*, 10 (1), 55–63.
- 745 Phatak, A., Reilly, P. & Penlidis, A. (1993) An approach to interval estimation in partial least
746 squares regression, *Analytica Chimica Acta*, 277(2), 495-501.
- 747 Sanchez, P.A., Ahamed, S., Carré, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie,
748 P., McBratney, A.B., McKenzie, N.J., Mendonça-Santos, M.D.L., Minasny, B.,
749 Montanarella, L., Okoth, P., Palm, C.A., Sachs, J.D., Shepherd, K.D., Vågen, T.-G.,
750 Vanlauwe, B., Walsh, M.G., Winowiecki, L.A., & Zhang, G.-L. (2009). Digital Soil
751 Map of the World. *Science*, 325, 680–681.
- 752 Savitzky, A., & Golay, M.J.E. (1964). Smoothing and differentiation of data by simplified
753 least squares procedures. *Analytical Chemistry*, 36 (8), 1627–1639.
- 754 Schwanghart, W., & Jarmer, T. (2011). Linking spatial patterns of soil organic carbon to
755 topography - a case study from south-eastern Spain. *Geomorphology*, 126, 252-263.
- 756 Selige, T., Bohner, J., & Schmidhalter, U. (2006). High resolution topsoil mapping using
757 hyperspectral image and field data in multivariate regression modeling
758 procedures. *Geoderma*, 136, n°1-2, pp. 235-244
- 759 Stevens, A., Van Wesemael, B., Vandenschrick, G., Touré, S. & Tychon, B. (2006) Detection
760 of Carbon Stock Change in Agricultural Soils Using Spectroscopic Techniques, *Soil*
761 *Science Society of America journal*, 70, 844–850.

- 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