Near infrared spectroscopy to quantify the temperature reached in burned soils: importance of calibration set variability

Andrea Pérez-Bejarano, César Guerrero*

Departamento de Agroquímica y Medio Ambiente, Universidad Miguel Hernández de Elche, Avenida de la Universidad, s/n. E-03202 (Elche, Alicante), Spain

* Corresponding author e-mail: cesar.guerrero@umh.es

https://doi.org/10.1016/j.geoderma.2018.03.038

Abstract

Near infrared (NIR) spectroscopy is a potential technique for the quantification of the temperature reached (TR) in burned soils. Due to spatial variation, inaccurate predictions can result from calibrating a model with heat-sensitive compounds that are not present in the samples of the burned area. Therefore, we investigated how to develop robust models. The progressive augmentation of the model size successively enhanced the precision, while the increase of the calibration set's variability gradually improved the accuracy through decreases in bias. The increase in calibration set variability enhances the probability of calibration using only the most common heat-sensitive compounds, facilitating reliable predictions of TR regardless of the spatial variation. On the other hand, models calibrated with heated aliquots from a unique sample, even from a composite sample, should be totally avoided because, regardless of their apparent utility, they are prone to inaccurate predictions.

Keywords: wildfire effects; heat-sensitive compounds; fire intensity; NIR spectroscopy; robust models; postfire assessment

1. Introduction

Fire is an important process in several terrestrial ecosystems throughout the world. The immediate fire effects on soil properties are mainly related with the fire severity, where the TR in soil is a crucial parameter (Neary et al., 1999; Vieira et al., 2015). Despite its importance, soil heating or TR in soil is a parameter neither measured by post-fire assessment teams, such as the Burned Area Emergency Response (BAER), nor by researchers, mostly due to the lack of a standard, easy, cheap, rapid and accurate method (Fernández and Vega et al., 2016; Parson et al., 2010). Some of the existing methods are subjective (Vega et al., 2013), complex and expensive (Merino et al., 2014, 2015; Neris et al., 2014; Santín et al., 2016; Verdes and Salgado, 2011), have moderate accuracy (Melquiades and Thomaz, 2016; Pérez and Moreno, 1998), or need several indicators to derive wide classes or levels, such as the soil burn severity index (Jain et al. 2008; Morgan et al., 2014; Parsons et al., 2010). To fill this gap, Guerrero et al. (2007) proposed the use of near infrared (NIR) reflectance spectroscopy as a potential technique for the quantification of TR in burned soils. The approach suggested by Guerrero et al. (2007) is based on two basic premises: i) the NIR spectrum of a soil sample contains information about the organic matter (quantity and quality), clay (quantity and quality), minerals (such as carbonates and iron oxides) and water content (Nocita et al., 2015; Stenberg et al., 2010; Viscarra Rossel and Behrens, 2010; Viscarra Rossel et al., 2016); and ii) these components have different sensitivity to thermal shocks (DeBano et al., 1998; Knicker, 2007; Neary et al., 1999; Raison, 1979; Santín et al., 2016). Consequently, each temperature causes a group of changes in soil properties, leaving a typical fingerprint in the NIR spectra (Guerrero et al., 2007; Lugassi et al., 2010, 2014). Therefore, in this approach, the NIR spectrum is used as an integrative measurement of soil properties, which can be modified by the temperature (i.e., heat-sensitive compounds). However, as in other empirical approaches using NIR, a model is needed to relate the TR with

its fingerprint in the NIR spectrum. For the development of such a model, soil samples are heated at known temperatures, which are used as standards (i.e., as calibration set) to calibrate the model. Hence, once the model has been calibrated (using chemometrics), the TR can be efficiently measured in very large numbers of samples because the NIR spectrum of a soil sample is obtained in seconds, easily and without the need of chemical reagents (avoiding the generation of toxic wastes in laboratories). To obtain high resolution maps of TR, which can serve to locate prior intervention areas in burned sites, it is necessary to measure this parameter (TR) in hundreds or even thousands of soil samples (Parsons et al., 2010; Jain et al., 2012). Additionally, the TR can provide relevant information for a better description of fire effects at very fine scale and to enhance the understanding of fire ecology, such as for the *in situ* analysis of fire-mediated germination patterns (Keeley et al., 2008; Lentile et al., 2007; Pausas et al., 2003).

The assessment of TR using NIR in a wildfire-affected area implies the collection of burned samples in those target points where TR is needed (for instance, for mapping) but also the collection of unburned soil samples to calibrate the model (Guerrero et al., 2007). As a consequence of the short-scale natural variation of the soil properties, a sample located at an arbitrary position (regardless if burned or not) may present its particular composition of heat-sensitive compounds, and this composition can differ with respect to other samples located at different positions. Two issues are imposed by the presence of spatial variation, which, through its design, a suitable model should overcome. The first is the impossibility of constructing a model using samples with similar composition to those to be predicted (except for planned events, such as prescribed fires). This limitation may exert a negative effect on prediction accuracy because the model might be fitted with heat-sensitive compounds that might not be the same as those present in samples located in the wildfire-affected area. The second implication is that the wildfire will affect samples with different heat-sensitive

compounds; therefore, the model should be able to properly predict samples with different spectral responses. For these reasons, we found it interesting to investigate how we can develop robust models able to overcome such problems linked to the natural short-scale spatial variation. The calibration set characteristics play an important role in model performance. If the calibration set is composed by several different samples, only common heat-sensitive compounds (i.e., common spectral changes) should be used to fit the model. Consequently, accurate predictions are expected in those samples where the common heat-sensitive compounds are present. Therefore, the higher the variability included, the larger the commonality of the predictors, and the wider the model applicability. Thus, we hypothesised that increasing the calibration set variability should result in a progressive improvement of the predictions accuracy. To test the hypothesis, different model types were constructed and were then used to predict TR. ~~The quality of the predictions was analysed with the coefficient of determination (R^2), root mean square error of prediction (RMSEP), bias, standard error of prediction (SEP) and ratio of performance to deviance (RPD). The study was repeated in three sites located in Alicante province (Spain).~~

2. Materials and methods

2.1. Sites

This study was performed using forest soil samples collected in three different sites located in Alicante province (Spain): Aitana, Maigmó and Pinoso. These sites were approximately 30 km apart from each other. The vegetation of these sites is composed of *Pinus halepensis* Mill., as the dominant species in the tree stratum, the understory vegetation being dominated by species such as *Quercus coccifera* L., *Rosmarinus officinalis* L., *Juniperus oxycedrus* L., *Stipa tenacissima* L. and *Brachypodium retusum* (Pers.) P. Beauv. The main characteristics of the three study sites are given in Table 1.

100

101 **2.2. Sample collection**

102 In each site, five microplots (1 m^2) were randomly selected in a small area of
103 approximately 2500 m^2. The minimum and maximum distance between microplots was 5 and
104 50 m, respectively. In each microplot, a topsoil sample (0–5 cm depth) of approximately three
105 kilograms was collected. These microplot-scale soil samples (hereafter MPS) collected in
106 Aitana site were identified as A1, A2, A3, A4 and A5. Those MPS collected in Maigmó were
107 identified as M1, M2, M3, M4 and M5. Similarly, the MPS collected in the Pinoso site were
108 identified as P1, P2, P3, P4 and P5. In each site, an additional "composite sample" was
109 obtained by bulking several subsamples collected at different points across the whole area;
110 these composite samples were identified as Ac, Mc and Pc for Aitana, Maigmó and Pinoso,
111 respectively.

112

113 **2.3. Obtaining standards (laboratory-heated samples)**

114 Once in the laboratory, the 18 samples (15 MPS + 3 composite samples) were air-
115 dried for two weeks (at 25°C) and sieved to <2 mm. These 18 samples were the "sources of
116 standards." Twenty-four aliquots of approximately 10 g were obtained from each sample.
117 These aliquots were heated in a muffle furnace at 24 different combinations of temperatures
118 (70°C, 100°C, 200°C, 300°C, 400°C, 500°C, 600°C and 700°C) and exposure times (10, 20
119 and 40 minutes). In this way, a set of 24 heated aliquots was obtained from each MPS sample
120 (also from each composite sample). A total of 432 heated aliquots (24 heating combinations
121 per sample × 6 samples [5 MPS + 1 composite] × 3 sites) were obtained, which were used as
122 standards for the models (see section 2.6). Each aliquot was introduced in the pre-heated
123 furnace as a 1-mm layer in order to guarantee homogeneous heating (Guerrero, 2010). During
124 the heating, to register the exact TR, the temperature of the aliquot was monitored and

125 recorded every 10 seconds using a thermocouple (Guerrero et al., 2007).

126

127 **2.4. Obtaining the NIR spectra of the heated aliquots**

128 Once cooled, the NIR spectra (12000–3800 cm^{-1}) of the 432 heated aliquots were
129 obtained using a FT-NIR diffuse reflectance spectrophotometer (MPA Bruker, Germany).
130 Further details about the scanning can be found in Guerrero et al. (2007). The x-scale of the
131 spectra was transformed to nanometres (834–2630 nm) and resampled to 1 nm. All the spectra
132 were transformed to absorbance, and then were pre-processed with the first derivative
133 (Savitzky–Golay, 25 points) and vector normalization. (Standard Normal Variate). The
134 absorbance spectra of the 432 heated aliquots can be found as supplementary content. The
135 OPUS spectroscopic software (OPUS version 6.5 software; BrukerOptik GmbH, Ettlingen,
136 Germany) was used for spectral pre-processing.

137

138 **2.5. Regression method**

139 We have constructed different types of models relating the TR with the NIR spectra
140 (see section 2.6). In all cases, the aliquots heated in laboratory (see section 2.3) were used as
141 the "standards" to calibrate the models. All models were calibrated with partial least squares
142 (PLS) as the multivariate regression method, using the OPUS spectroscopic software (OPUS
143 version 6.5 software; BrukerOptik GmbH, Ettlingen, Germany). The leave-one-out cross-
144 validation was used to decide the number of PLS vectors to be included in models on the
145 basis of the root mean square error of cross-validation (RMSECV) in a scree plot.

146

147 **2.6. Model types**

148 Ten different model types were constructed in this study. The models differed in the
149 size of the calibration set, and in the variability included. In this study, the *size* is the number

of standards (i.e., number of heated aliquots) used to calibrate the model, and the *variability* is the number of different samples used as source(s) of heated aliquots. The "*label*" of the model type provides information on the size and the variability (Table 2). For instance, the label 72s–v3 denotes a model calibrated with 72 heated aliquots (72 standards) which were obtained from three different samples (three MPS have been used as sources of heated aliquots). At each site, ten different model types were constructed as follows (Table 2):

1) Models 24s–v1: these models were calibrated with 24 heated aliquots (24 standards) obtained from one MPS. In each site, five models were constructed (models #1 to #5; Table 2).

2) Models 48s–v2: these models were calibrated with 48 heated aliquots (48 standards) obtained from two MPS. In each site, ten models were constructed (#6 to #15; Table 2).

3) Models 72s–v3: these models were calibrated with 72 heated aliquots (72 standards) obtained from three MPS. In each site, ten models were constructed (#16 to #25; Table 2).

4) Models 96s–v4: models calibrated with 96 heated aliquots (96 standards) obtained from four MPS. In each site, five models were constructed (#26 to #30; Table 2).

5) Models 24s–v2: models were calibrated with 24 standards selected from two MPS. In each site, ten models were constructed (#31 to #40; Table 2).

6) Models 24s–v3: models calibrated with 24 standards selected from three MPS. In each site, ten models were constructed (#41 to #50; Table 2).

7) Models 24s–v4: models calibrated with 24 standards selected from four MPS. In each site, five models were constructed (#51 to #55; Table 2).

8) Models 48s–v4: models calibrated with 48 standards selected from four MPS. In each site, five models were constructed (#56 to #60; Table 2).

9) Models 72s–v4: models calibrated with 72 standards selected from four MPS. In each site, five models were constructed (#61 to #65; Table 2).

10) Models 24s–v1c: models calibrated with the 24 heated aliquots (24 standards) obtained from the composite sample. In each site, one model was constructed (model #66; Table 2).

A total of 66 models were constructed for each site (Table 2). The description of the MPS used for each of model can be found in Tables S1, S2 and S3 for Aitana, Maigmó and Pinoso site, respectively. In some cases (those marked with an asterisk in Table 2), the required model size is smaller than the total number of available spectra. In these cases, a principal component analysis (PCA) was performed with the NIR spectra of the available spectra. Then, the Kennard–Stone algorithm was used to select the desired number of spectra. The selected spectra were those whose scores were evenly distributed across the space defined by the first three principal components. For instance, a model 48s–v4 (as an example) is a model calibrated with 48 heated aliquots (48 standards), which were obtained from four different samples. Since four MPS were used as sources of heated aliquots, a total of 96 spectra were available. Thus, a PCA was conducted with the 96 spectra, but only 48 spectra were used to calibrate the model. These 48 spectra were selected with the Kennard–Stone algorithm. The minimum variability would have been reached when all the standards used to calibrate the model had been derived from a unique MPS sample (such as in models 24s–v1). The maximum variability would have been reached when each standard included in the calibration set has been obtained from a different sample (i.e., only one heated aliquot per MPS sample). Such maximum variability has not been evaluated in this study, as the maximum number of different MPS used as sources of heated aliquots was four (Table 2).

2.7 Predictions and data analysis

The previously described models were used to predict the TR. Several prediction performance parameters were computed to analyse the quality of the predictions:

determination coefficient (R^2), root mean square error of prediction (RMSEP), standard error of prediction (SEP), bias and ratio of performance to deviance (RPD~~).~~), which is the standard deviation of the prediction set divided by the RMSEP (Stenberg et al., 2010; Bellon-Maurel et al., 2010). The RMSEP was used to measure the accuracy, and the SEP was used to measure the precision (Bellon-Maurel et al., 2010; Næs et al., 2002).

These prediction performance parameters were computed using predictions obtained in 24 heated aliquots from a MPS not used to calibrate the model. Therefore, the setup provided independent predictions, similar to a MPS-hold-out cross-validation or leave-one-MPS-out cross-validation. In this way (although with some limitations), the approach tried to mimic a realistic scenario, where the samples used to construct the models, and those to be predicted, are irremediably located on different plots, since the former would be located in the unburned area and the latter in the burned area.

A complete identification of the MPS used as prediction set in each case can be found in Tables S1, S2 and S3 (indicated by a cross in the right side of the tables). At each site, the 66 models were used 145 times, and therefore, 145 different values of R^2, RMSEP, SEP, bias and RPD were obtained for each site (see supplementary Tables S1, S2 and S3). In each site, those values of R^2, RMSEP, SEP, bias and RPD obtained with the same model type were considered as replicates and were therefore averaged by model type. Since the bias can be positive or negative, in order to get a meaningful average, we used its absolute value.

The prediction performance parameters, once averaged by model type, were arranged in four groups, to facilitate the analysis of the results (Table 2). To facilitate the description and meaning of groups and its comparisons, some results included in the Group *i* have also been included in Group *ii* (24s–v1) and Group *iii* (96s–v4), as follows:

– Group *i*: this group contains results from models that differed in both the variability (number of MPS used as source of standards) and model size (number of standards included

in the calibration set; Table 2).

– Group *ii*: this group contains results from models that differed only in variability because the model size was constant (Table 2).

– Group *iii*: this group contains results from models that differed only in size because the variability was constant (Table 2).

– Group *iv*: this group contains results obtained from the model calibrated with the composite sample (Table 2).

3. Results

3.1. Changes in soils and NIR spectra

The colour of the soil samples was modified as a consequence of heating. Due to the carbonization of the soil organic matter, a progressive darkening was observed when the exposure temperature was increased (up 450°C). As a consequence of such darkening, the baseline of the NIR spectra (absorbance) was increased, especially at shorter wavelengths close to the visible spectral range (see supplementary figures Figs. S1-S27). The dark colour abruptly disappeared when the temperature was above ~450°C, mainly due to the combustion of the organic compounds. The transformation of some iron oxides into hematite was an additional change in the NIR spectra (approximately 890–892 nm; Fig. 1a) and in soil colour, causing sample reddening (Ketterings and Bigham, 2000; Torrent and Barrón, 2002; Terefe et al., 2005; Ulery and Graham, 1993), which was especially evident at high temperatures (>500°C).

While changes at naked eye are the basis of visual estimators of fire severity (Pérez and Moreno, 1998; Vega et al., 2013), the NIR spectra can provide further additional information about the changes in soil properties, since NIR spectroscopy has been used to quantify a large number of soil properties, which in turn are affected by temperature (Santín et

al., 2016; Stenberg et al., 2010). Other important changes in the NIR spectra were those observed in features located approximately 1414 nm (Fig. 1b), 1920 nm (Fig. 1c) and 2210 nm (Fig. 1d), related with adsorbed and free water –OH (at 1414 nm and 1920 nm) and clay –OH (at 1414 nm and 2210 nm). These spectral features tended to decrease with the increase in TR, mostly due to dehydration and dehydroxylation processes but, as other authors have observed (Guerrero et al., 2007; Lugassi et al., 2010, 2014), also due to the combustion of the organic matter. A detailed discussion of changes in soil properties can be found in Certini (2005), and a detailed discussion of changes in the NIR spectra can be found in Lugassi et al. (2014).

3.2. Calibrations

All the above mentioned changes, and others, facilitated the calibration of models to quantify TR. Therefore, we could surmise that several heat-sensitive compounds were involved in the calibrations. Despite the large differences between calibration sets, the 198 models showed high R^2 values (ranging from 0.92 to 0.99), and low RMSECV (RMSE of cross-validation) values, ranging from 21°C to 58°C (data not shown). These values of R^2 and RMSECV obtained in the cross-validations were similar to those from previous studies (Guerrero et al., 2007; Guerrero, 2010; Maia et al., 2012). Fig. 2 shows the cross-validation results of four randomly selected models, as a representative illustration of the 198 models calibrated for this study. These results indicate the great capacity of NIR spectroscopy to measure some of the soil properties that change with temperature (i.e., heat-sensitive compounds). Apparently, they all seem to be useful to predict TR.

3.3. Predictions

The 435 values of RMSEP obtained with the different 198 models ranged from 14.3°C to 99.7°C. Three out of these 435 cases are shown in Fig. 3, as representative predictions obtained with three different types of models. Fig. 4 shows the RMSEP obtained with the models included in Group *i*, which were models constructed with all the heated aliquots from one (24s–v1), two (48s–v2), three (72s–v3) and four MPS (96s–v4). The grey bars in Fig. 4 represent the RMSEP values averaged by model type (separately for each study site), whereas the dots and triangles are used to show the minimum and the maximum RMSEP observed for each model type, respectively. For the Aitana site (Fig. 4a), the RMSEP of predictions obtained with the 20 models labelled as 24s–v1 ranged from 31.2°C to 60.2°C, with an average of 40.9°C. In Maigmó (Fig. 4b), these 20 RMSEP values were slightly lower than in Aitana, and ranged from 18.6°C to 44.9°C, with an average of 31.2°C. In Pinoso, the 20 RMSEP values obtained with models 24s–v1 were slightly higher than the other study sites, and ranged from 24.8°C to 99.7°C, with an average of 49.4°C (Fig. 4c). Regardless of the site considered, the progressive increase of the calibration set variability (i.e., number of MPS) and size (i.e., number of standards) coincided with the gradual decline of RMSEP (Fig. 4). Other prediction performance parameters (such as R^2, SEP, etc.) also indicated an improvement of predictions with the increase in variability and size of the calibration sets. Since similar patterns were observed for the three study sites (data not shown), the values of the three sites were pooled together, and averaged by model type (Fig. 5). Once averaged by model type, the RMSEP dropped progressively from 40.5°C in models 24s–v1 to 29.6°C in models 96s–v4 (Fig. 5a). A similar gradual change was also observed for the other prediction performance parameters, such as the R^2, SEP, bias and RPD (Fig. 5). For instance, RPD was 6.2 in models 24s–v1, and it increased to RPD 8.1 in models 96s–v4 (Fig. 5b). The improvement of the predictions might be attributed to the higher variability included in the calibration set, which was four times higher in models 96s–v4 than in models 24s–v1,

299 although such improvement could also be attributed to the calibration set size, which was also
300 four times higher. To discern whether the calibration set variability affects accuracy, we
301 constructed another group of models (Group *ii*) where the number of MPS varied while the
302 model size was constant at 24 in all cases (Table 2). For the average of the three sites, the
303 RMSEP decreased from 40.5°C in 24s–v1 to 35.7°C in 24s–v4, confirming the positive effect
304 of the increase in the calibration set variability on accuracy (Fig. 5a). The decrease in bias
305 from 15.1°C in 24s–v1 to 8.2°C in 24s–v4 was the main contributor to the RMSEP reduction
306 observed in Group *ii*, because SEP only decreased by 2°C (Fig. 5a). The increase of the
307 calibration set size was the additional factor explaining the progressive decrease of RMSEP
308 observed in the models from Group *i* (Fig. 4 and Fig. 5). Nevertheless, the size effect was
309 clearly observed when the RMSEP was compared for the models included in Group *iii* (Table
310 2), where the size was the unique difference between them. For that group of models (Group
311 *iii*), the decrease in RMSEP from 35.7°C in 24s–v4 to 29.6°C in 96s–v4 is mostly attributable
312 to an improvement in precision (lower SEP) because the bias remained fairly stable (only
313 decreased less than 1°C), whereas the SEP decreased by approximately 6.5°C (Fig. 5a).
314 The less accurate predictions were obtained with models constructed with the heated
315 aliquots from the composite sample (model type labelled 24s–v1c). On average across the
316 three sites, the bias was 22°C, explaining a substantial portion of the RMSEP (47.6°C). These
317 errors were clearly higher than those obtained with models of similar size, such as models
318 24s–v4. Furthermore, these errors were also higher than those obtained with models having a
319 similar degree of variability (i.e., same number of MPS) such as the models 24s–v1.
320 The R^2 values shown in Fig. 5b should be interpreted with some caution since they are
321 somehow affected by the approach used to compute results. In this study, the prediction
322 performance parameters (such as R^2) have been computed using the predictions obtained in
323 sets of 24 heated aliquots, which belong to the same MPS (see tables S1, S2 and S3).

Consequently, the prediction performance parameters have been obtained at MPS scale (and were then averaged by model type). At MPS scale, biased predictions do not affect the R^2. This can be seen in Figs. 6a to 6d. Fig. 6a shows TR predictions obtained in 24 heated aliquots from one randomly selected MPS. The cases shown in Figs. 6b, 6c and 6d are the same predictions shown in Fig. 6a but after the manipulation of the predicted values in order to have differently biased predictions. In Fig. 6b, 100°C was added to each predicted value shown in Fig. 6a; in Fig. 6c, 50°C was subtracted from each predicted value; in Fig. 6d, each predicted value was multiplied by 0.5 (divided by two). In all cases (Figs. 6a to 6d), the R^2 is the same (R^2=0.99) regardless of the bias (due to slope or due to offset). If we compute the mean value of these four cases, the R^2 is 0.99 despite the different bias in each case. However, under realistic conditions, the evaluation of the burned area implies making predictions from samples collected at several positions. If the model predictions have a different bias depending on the sample composition (which may vary with position), then we must expect patterns as those shown in Figs. 6e and 6f, where due to differences in bias, such heterogeneity in the prediction set results in much lower R^2 values. Figs. 6e and 6f contain a random selection of 24 cases from those shown in Figs. 6a to 6d, and simulate two prediction sets composed by samples with different bias. In these cases (Figs. 6e and 6f), the different values of bias (by slope or by offset) were negatively affecting the R^2. In these examples, the values were R^2=0.72 and R^2=0.63, resulting in a mean R^2=0.675, which clearly contrasts with the R^2=0.99, obtained as a mean of four values of R^2 obtained at MPS scale. To have a direct measure of such discrepancy, we also computed the R^2 after pooling predictions from the different MPS, mimicking prediction sets composed by samples that can have a different bias. For that, predictions from two illustrative model types were selected: 1) predictions obtained with models 24s–v1c (Fig. 6g), which was the worst option in terms of bias (highest bias in Fig. 5a); 2) predictions obtained with models 96s–v4 (Fig. 6h), which was the best option in

terms of bias (lowest bias in Fig. 5a). A small discrepancy between the approaches was observed when the predictions were obtained with models 96s–v4 (low bias; Fig. 6h), with R^2=0.981 when the R^2 was computed once predictions were pooled (Fig. 6h) and R^2=0.985 when R^2 was computed as an average of values obtained at MPS scale (Fig. 5b). As expected, a larger discrepancy was observed when predictions were obtained with models 24s–v1c (high bias; Fig. 6g), with R^2=0.954 when predictions were pooled (Fig. 6g) while R^2=0.971 when it was computed as an average of the five cases (Fig. 5b). Regardless of the approach, the R^2 values were high because the bias values were not very large—clearly lower than those shown in Figs. 6b or 6d. Moreover, the wide range included in the prediction sets also contributed to the high R^2 values (Davies and Fearn, 2006). Despite their absolute values, the R^2 values shown in Fig. 5b were in concordance with other performance parameters and therefore support the same conclusions about the importance of the calibration set size and variability. It is worth highlighting that other prediction performance parameters, such as RMSEP, in contrast to R^2, are minimally affected by the way results were computed because RMSEP is a parameter related to the residuals, given that it is linked to each predicted value.

4. Discussion

Soil contains heat-sensitive compounds, which can be can be measured with NIR spectroscopy, allowing the development of models to quantify the TR. However, the composition of soils, including its heat-sensitive compounds, varies with space, even at short distances. Therefore, each sample might possess its particular composition of heat-sensitive compounds and that can exert an important effect on prediction reliability. We have observed an effect of the calibration set variability on accuracy, which highlights important issues about how adequate calibration sets should be developed to manage the spatial variation effects. To facilitate the explanation of mechanisms connecting the calibration set variability with the

374 robustness of predictions, we discuss the results with the help of a schematic representation of
375 spectra, models and predictions (Fig. 7), since a direct analysis of the *b*–coefficients of PLS–
376 models is not straightforward. Such simplification enhances the illustration and understanding
377 of mechanisms involved. A schematic representation of the NIR spectra of three MPS and one
378 composite sample is shown in Fig. 7a: a black circle denotes the wavelength (λ) where the
379 spectral feature of a heat-sensitive compound appears, and a grey circle denotes its absence in
380 the sample. Fig. 7b contains the schematic representation of seven models calibrated with
381 aliquots from samples shown in Fig. 7a. The black arrows denote those wavelengths included
382 in the models, which depend on the heat-sensitive compounds present in the calibration set.
383 Thus, a black arrow at λ_n denotes where the *b*–coefficient has a large contribution in
384 predicting TR. As consequence of heating, some compounds trend to disappear, decreasing
385 their spectral features as temperature increases, such those related with –OH loss by
386 dehydroxilation (and organic matter combustion, in general). However, other compounds can
387 be generated by heating, such as certain iron oxides (as hematite), and thus they appear as
388 new spectral features as the temperature increases. Consequently, the sign of *b*–coefficients
389 could be negative or positive, being denoted as down or up arrows, respectively.

390 The first case shown at the top of Fig. 7b (case *i*) represents an example of a model
391 whose calibration set had minimal variability (variability=1) because all the heated aliquots
392 included in the calibration set had been obtained from the same sample (MPS1). Therefore,
393 this is a sample-specific model. This model had been fitted only on the basis on changes
394 produced in the heat-sensitive compounds present in that sample (those located at λ_3 and λ_4;
395 see black dots in Fig. 7a). As consequence, this model will predict TR using two *b*–
396 coefficients located at λ_3 and λ_4 (see black arrows at λ_3 and λ_4 in Fig. 7b, case *i*). The second
397 and third models shown in Fig. 7b (cases *ii* and *iii*) are also examples of models constructed
398 with heated aliquots obtained from a unique sample (MPS2 for case *ii*; MPS3 for case *iii*),

399 therefore, these are also sample-specific models (variability=1). Since each MPS sample
400 would have had a different composition (Fig. 7a), we would therefore expect that the models
401 were based on the different wavelengths from each sample, as shown in Fig. 7b for cases *i* to
402 *iii*. The models shown in cases *iv* and *v* of Figure 7b represent models constructed with heated
403 aliquots from two MPS samples (i.e., variability=2). While these two MPS samples might
404 have had a different composition, only the common changes (common heat-sensitive
405 compounds) were used to fit the model. Similarly, when three MPS were used as the source of
406 the aliquots (Fig. 7b, case *vii*), three different sets of heat-sensitive compounds might have
407 been be present in the calibration set, but the model was fitted only with the common changes
408 (i.e., common heat-sensitive compounds, as in that located at λ_3). Thus, the increase in the
409 number of MPS used as sources of standards (i.e., the increase of the calibration set
410 variability) forces the calibration of models on the basis of the most common heat-sensitive
411 compounds. It is worth mentioning that a model based on the most common heat-sensitive
412 compounds could also have been obtained with only two samples (MPS1 and MPS2), as
413 illustrated in Fig. 7b *iv*. However, two samples might not be enough, as in the case shown in
414 Fig. 7b *v*, explaining why a gradual augmentation in calibration set variability progressively
415 increases the chance to fit a model based only on the most common heat-sensitive
416 compounds. Fig. 7b clearly illustrates that several different models can be obtained with the
417 samples from a given site. Apparently, all of them seem to be useful since they relate NIR and
418 TR; however, most of them would have restricted applicability.
419 Figs. 7c to 7f represent the predictions obtained in MPS1, MPS2 and MPS3 from four
420 different models. The model used in each case is described by its *b*–coefficients (black
421 arrows). When the model is used in each of the three samples, a pass or fail symbol is added
422 to denote unbiased or biased predictions, respectively. We can expect accurate predictions
423 when the sample used for prediction has the same heat-sensitive compounds as the predictors

424 in the model. For instance, Model 1 (shown in Fig. 7c) predicts TR using the changes in the
425 heat-sensitive compounds located at λ_1 and λ_3 (black arrows). Thus, accurate predictions are
426 expected when Model 1 is used for sample MPS2 (Fig. 7c). Conversely, biased predictions
427 are expected when Model 1 (Fig. 7c) is used for samples MPS1 and MPS3 because these
428 samples lack the heat-sensitive compound located at λ_1. Since the contribution of λ_1 to the
429 prediction is missing, then we would expect negatively biased predictions in MPS1 and MPS3
430 (a 50% underestimation assuming a similar contribution from both heat-sensitive
431 compounds). However, Figs. 7d and 7e show how we can obtain accurate predictions from
432 samples that contain a different composition of heat-sensitive compounds. This is the case for
433 Model 2 when used for sample MPS2 (Fig. 7d). The presence of an additional heat-sensitive
434 compound in MPS2 (the one located at λ_3) does not distort the predictions obtained with
435 Model 2, because this model does not consider the changes in wavelength λ_3 when predicting
436 TR. Whereas Model 2 can predict without bias only for one sample (MPS2; Fig. 7d), Model 3
437 can do so for three samples (Fig. 7e). The reason is that Model 3 (Fig. 7e) predicts TR with
438 the heat-sensitive compound located at λ_3, which is the most common. Therefore, a model
439 such as this can predict without bias in a wide number of different samples. The unique
440 requirement is the presence of that heat-sensitive compound, which most of the samples
441 would contain since it is a very common compound. Therefore, the better identification of the
442 most common heat-sensitive compounds might be the link which explains why the
443 progressive augmentation of the calibration set variability successively improved prediction
444 accuracy.

445 In agreement with this supposed mechanism, a high variability in the calibration set
446 would allow obtaining reliable (not biased) predictions in dissimilar samples as well (Fig. 7e),
447 overcoming the restriction of predictions based only on similar samples. Consequently, it
448 increases the spatial applicability of the models, since they would not only be restricted to

449 adjacent sites, where similar samples would be expected. A high variability does not
450 completely exclude the possibility of bias in predictions, although this would be expected
451 only for those samples where the most common heat-sensitive compound is absent. Especially
452 if the collection of unburned samples used as sources of aliquots is sufficiently representative
453 of the unburned area, this should be a minor proportion of the total samples. Additionally, in
454 turn, the unburned area would also have to be representative of the burned area. Under these
455 conditions, a calibration set constituted by a number of different samples could be considered
456 a site-specific model.

457 This theory also explains why the most biased predictions were obtained with the 24s–
458 v1c models. For an arbitrary study site, all the heated aliquots used for the model 24s–v1c
459 were obtained from the composite sample; therefore, it can considered as a sample-specific
460 model. The composite sample, as a bulked mixture of several subsamples, should contain
461 most of the heat-sensitive compounds (Fig. 7a). As a consequence, all these compounds
462 would probably have contributed to the calibration of the model (Fig. 7b, case *vi*). Thus, this
463 model (such as Model 4, Fig. 7f) should produce biased predictions in those samples where at
464 least one of the heat-sensitive compounds is absent. The observed results fully support this
465 theory, since the most biased predictions were obtained with models 24s–v1c. Clearly, any
466 attempt to include site variability through the use of a composite sample is not recommended.
467 A composite sample would describe a representative fingerprint of heat on soil properties, but
468 this would not be useful to obtain robust models because they would be prone to produce
469 biased predictions in many samples. Therefore, the development of models using a composite
470 sample should be totally avoided. This result might explain why, when using predictions from
471 this type of model, some authors did not observe the expected relationships between changes
472 in soil properties and temperature (Maia et al., 2012; Mataix-Solera et al., 2013).

473 The number of heated aliquots (model size) also had an impact on the accuracy.

Whereas the number of MPS decreased the bias, the number of heated aliquots increased the accuracy through the improvement of precision (i.e., the decrease of SEP). A high number of heated aliquots in the calibration set of a model should better describe the heat effect on the heat-sensitive compounds. Consequently, it enhances the quantification of the *b*–coefficient values of the implied heat-sensitive compounds, improving the calibration of models, which is a basic premise for obtaining accurate predictions. Clearly, the best option is the development of models with a large number of heated aliquots from several samples. The importance of the variability and the number of samples used for calibration has been already demonstrated by many authors with other soil properties (Brown, 2007; Shepherd and Walsh, 2002).

Very accurate predictions have been obtained in this study, even when the worst models were used, which were those calibrated with low variability (such as those 24s–v1 and 24s–v1c). This is probably due to the very small size of the area (2500 m^2) where the five MPS were sampled in each site. Thus, these samples were separated by short distances (less than 50 m; see section 2.2), and consequently, they were quite similar to each other. Hence, it is reasonable to expect a substantially lower accuracy when this model type is used under realistic conditions, where larger distances (and larger dissimilarity) may exist between burned and unburned samples. Thus, we should also expect a larger difference in the accuracy of predictions obtained with models calibrated with low and high variability.

This study provides valuable information about factors affecting the reliability of models. Most of these results are very interesting because they have a direct impact on how to design the calibration set, which even affects the collection of the unburned samples to be used as sources of aliquots, and the built of a calibration set for site-specific models. Nevertheless, more studies are needed to understand the limitations and capabilities of NIR spectroscopy to predict TR and how to construct models in a more efficient way. For instance, the transferability of models across sites and scales is an important issue because it may

499 facilitate a faster development of models, which may substantially decrease the time needed
500 to evaluate wildfire-affected sites. Dedicated experiments should be designed for that
501 purpose, especially because some of the results obtained in previous studies (Guerrero et al.,
502 2007) have suggested its feasibility. These authors constructed several hundred different
503 models using soil samples from four sites. These models differed from each other in their
504 basic characteristics such as rank (number of PLS vectors), spectral range included and pre-
505 processing (normalization, derivatives, etc.). Then, a few of these models (but not all of them)
506 were able to predict TR in samples from a new independent site (as left-out site). Similar
507 results were observed when repeated with each of the five sites as a prediction set. Thus, the
508 results undoubtedly demonstrated the presence of common heat-sensitive compounds in soils
509 from five different sites in Alicante province (southeast Spain). Anyway, it should be noted
510 that although independent sites were used in Guerrero et al. (2007), the identification of the
511 *best* model (correct design) was based on the results from the leave-out site and thus not an
512 independent selection. Therefore, these results should not be considered as the expected
513 accuracy for a true independent prediction, since hundreds of models (varying in its design)
514 can be fitted, but not all of them would produce accurate predictions. Nonetheless, the results
515 shown in Guerrero et al. (2007) were useful to illustrate the potential of NIR to predict TR.
516 For that reason, the results of the present study have an additional value because they provide
517 the first measure of the accuracy obtained by independent predictions.

518 The improvement of the ability to obtain more realistic standards (mimicking naturally
519 burned soils) is also a step needed in the way towards the use of NIR as a thermometer. For
520 instance, the convenience to built models including standards generated under different
521 oxygen limitation levels should be also evaluated in future studies. Oxygen limitation level
522 during heating could be an important variable since it controls the type of thermal changes in
523 soil organic matter during heating (from combustion to pyrolysis), and it may represent one of

524 the most important difference between naturally burned soils respect to the artificial
525 standards, which act as their surrogates to calibrate models.
526
527 **5. Conclusions**
528 The models to predict TR through NIR are calibrated with standards generated by
529 heating soil samples at known temperatures under controlled conditions in the laboratory.
530 When all the standards are obtained from a unique sample, the variability of the calibration set
531 is minimal. In this case, the model is calibrated with the heat-sensitive compounds present in
532 the sample. However, due to spatial variation, these compounds can differ in samples located
533 at distant positions. Therefore, each sample-specific model may be based on a particular,
534 unique and singular group of predictors, which are virtually different in each model. Thereby,
535 we can fit a very large number of different models, although the applicability might be
536 restricted to similar samples. On the other hand, when the calibration set includes standards
537 from different samples (i.e., high variability), then the model is fitted only with the common
538 heat-sensitive compounds. If the number of different samples is high enough to cover the site
539 variability, then the model can be considered site-specific. In contrast with a sample-specific
540 model, a site-specific model can predict without bias in a large number of different samples,
541 being able to overcome the negative effects of the spatial variation. Despite their apparent
542 utility, models calibrated with low variability or those based on a composite sample should be
543 avoided since they might be not useful. In addition, the model size (number of standards) is
544 also an important factor to consider during the calibration of models. While the increase in
545 variability improved prediction accuracy (lower RMSEP) by a decrease in bias, the increase
546 of model size did so through the enhancement of precision (i.e., decrease of SEP).
547
548 **References**

549 Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.M. McBratney, A. 2010.
550 Critical review of chemometric indicators commonly used for assessing the quality of the
551 prediction of soil attributes by NIR spectroscopy. TrAC Trends in Analytical Chemistry
552 29, 1073-1081. DOI: 10.1016/j.trac.2010.05.006.
553 Brown, D.J. 2007. Using a global VNIR soil-spectral library for local soil characterization
554 and landscape modelling in a 2nd-order Uganda watershed. Geoderma 140, 444–453.
555 DOI: 10.1016/j.geoderma.2007.04.021
556 Certini, G. 2005. Effects of fire on properties of forest soils: A review. Oecologia 143, 1–10.
557 DOI: 10.1007/s00442-004-1788-8
558 Davies, A., Fearn, T. 2006. Back to basics: calibration statistics. Spectroscopy Europe 18, 31–
559 32
560 DeBano, L.F., Neary, D.G., Ffolliott, P.F. 1998. Fire's effects on ecosystems. John Wiley &
561 Sons, New York.
562 Fernández, C., Vega, J.A. 2016. Modelling the effect of soil burn severity on soil erosion at
563 hillslope scale in the first year following wildfire in NW Spain. Earth Surface Processes
564 and Landforms 41, 928–935. DOI: 10.1002/esp.3876
565 Guerrero, C. 2010. Espectroscopía de infrarrojo cercano (NIR) para la estimación de las
566 temperaturas alcanzadas en suelos quemados. In: Actualización en Métodos y Técnicas
567 para el Estudio de los Suelos Afectados por Incendios Forestales. Pp. 259–287. Ed.
568 Cátedra de divulgación de la Ciencia. Universitat de Valencia. ISBN: ISBN 978-84-370-
569 7887-8.
570 Guerrero, C., Mataix-Solera, J., Arcenegui, V., Mataix-Beneyto, J., Gómez, I. 2007. Near-
571 infrared spectroscopy to estimate the maximum temperatures reached on burned soils.
572 Soil Science Society of America Journal 71, 1029–1037. DOI: 10.2136/sssaj2006.0187

573 Jain, T.B., Gould, W.A., Graham, R.T., Pilliod, D.S., Lentile, L.B., González, G. 2008. A soil
574 burn severity index for understanding soil-fire relations in tropical forests. Ambio 37,
575 563–568.
576 Jain, T.B., Pilliod, D.S., Graham, R.T., Lentile, L.B., Sandquist, J.E. 2012. Index for
577 characterizing post-fire soil environments in temperate coniferous forests. Forest 3, 445–
578 466. DOI: 10.3390/f3030445
579 Keeley, J.E., Brennan, T., Pfaff, A.H. 2008. Fire severity and ecosystem responses following
580 crown fires in California shrublands. Ecological Applications 18, 1530–1546. DOI:
581 10.1890/07-0836.1
582 Ketterings, Q.M., Bigham, J.M. 2000. Soil Color as an Indicator of Slash-and-Burn Fire
583 Severity and Soil Fertility in Sumatra, Indonesia. Soil Science Society of America
584 Journal 64, 1826–1833. DOI:10.2136/sssaj2000.6451826x
585 Knicker H. 2007. How does fire affect the nature and stability of soil organic nitrogen and
586 carbon? A review. Biogeochemistry 85, 91–118. DOI: 10.1007/s10533-007-9104-4
587 Lentile, L.B., Morgan, P., Hudak, A.T., Bobbitt, M.J., Lewis, S.A., Smith, A.M., Robichaud,
588 P.R. 2007. Post-fire burn severity and vegetation response following eight large wildfires
589 across the western United States. Fire Ecology 3, 91–108.
590 Lugassi, R., Ben-Dor, E., Eshel, G. 2010. A spectral-based method for reconstructing spatial
591 distributions of soil surface temperature during simulated fire events. Remote Sensing of
592 Environment 114, 322–331. DOI: 10.1016/j.rse.2009.09.015
593 Lugassi, R., Ben-Dor, E., Eshel, G. 2014. Reflectance spectroscopy of soils post-heating—
594 Assessing thermal alterations in soil minerals. Geoderma 213, 268–279. DOI:
595 10.1016/j.geoderma.2013.08.014
596 Maia, P., Pausas, J.G., Arcenegui, V., Guerrero, C., Pérez-Bejarano, A., Mataix-Solera, J.,
597 Varela, M.E.T., Fernandes, I., Pedrosa, E.T., Keizer, J.J. 2012. Wildfire effects on the

598 soil seed bank of a maritime pine stand - The importance of fire severity. Geoderma 191,
599 80–88. DOI: 10.1016/j.geoderma.2012.02.001
600 Mataix-Solera, J., Arcenegui, V., Tessler, N., Zornoza, R., Wittenberg, L., Martínez, C.,
601 Caselles, P., Pérez-Bejarano, A., Malkinson, D., Jordán, M. 2013. Soil properties as key
602 factors controlling water repellency in fire-affected areas: evidences from burned sites in
603 Spain and Israel. Catena 108, 9–16. DOI: 10.1016/j.catena.2011.12.006
604 Melquiades, F.L., Thomaz, E.L. 2016. X-ray fluorescence to estimate the maximum
605 temperature reached at soil surface during experimental slash-and-burn fires. Journal of
606 Environmental Quality 45, 1104–1109. DOI: 10.2134/jeq2015.06.0305
607 Merino, A., Chávez-Vergara, B., Salgado, J., Fonturbel, M.T., García-Oliva, F., Vega, J.A.
608 2015. Variability in the composition of charred litter generated by wildfire in different
609 ecosystems. Catena 133, 52–63. DOI: 10.1016/j.catena.2015.04.016
610 Merino, A., Ferreiro, A., Salgado, J., Fontúrbel, M.T., Barros, N., Fernández, C., Vega, J.A.
611 2014. Use of thermal analysis and solid-state C CP-MAS NMR spectroscopy to diagnose
612 organic matter quality in relation to burn severity in Atlantic soils. Geoderma 226, 376–
613 386. DOI: 10.1016/j.geoderma.2014.03.009
614 Morgan, P., Keane, R.E., Dillon, G.K., Jain, T.B., Hudak, A.T., Karau, E.C., Sikkink, P.G.,
615 Holden, Z.A., Strand, E.K. 2014. Challenges of assessing fire and burn severity using
616 field measures, remote sensing and modelling. International Journal of Wildland Fire 23,
617 1045–1060. DOI: 10.1071/WF13058
618 Næs, T., Isakson, T., Fearn, T., Davies, T. 2002. A user-friendly guide to multivariate
619 calibration and classification. NIR Publications, Chichester. ISBN 9780952866626
620 Neary, D.G., Klopatek, C.C., DeBano, L.F., Ffolliott, P.F., 1999. Fire effects on belowground
621 sustainability: a review and synthesis. Forest Ecology and Management 122, 51–71. DOI:
622 10.1016/S0378-1127(99)00032-8

623 Neris, J., Doerr, S.H., Tejedor, M., Jiménez, C., Hernández-Moreno, J.M. 2014. Thermal
624 analysis as a predictor for hydrological parameters of fire-affected soils. Geoderma 235,
625 240–249. DOI: 10.1016/j.geoderma.2014.07.018
626 Nocita, M., Stevens, A., vanWesemael, B., Aitkenhead,M., Bachmann, M., Barthès, B., Ben-
627 Dor, E., Brown, D.J., Clairotte, M., Csorba, A., Dardenne, P., Demattê, J.A., Genot, V.,
628 Guerrero, C., Knadel, M., Montanarella, L., Noon, C., Ramirez-Lopez, L., Robertson, J.,
629 Sakai, H., Soriano-Disla, J.M., Shepherd, K.D., Stenberg, B., Towett, E.K., Vargas, R.,
630 Wetterlind, J., 2015. Soil spectroscopy: an alternative to wet chemistry for soil
631 monitoring. Advances in Agronomy 132, 139–159. DOI: 10.1016/bs.agron.2015.02.002
632 Parson, A.; Robichaud, P.R.; Lewis, S.A.; Napper, C.; Clark, J.T. 2010. Field guide for
633 mapping post-fire soil burn severity. Gen. Tech. Rep. RMRS-GTR-243. Fort Collins,
634 CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station.
635 49 p.
636 Pausas, J.G., Ouadah, N., Ferran, A., Gimeno, T., Vallejo, R. 2003. Fire severity and seedling
637 establishment in Pinus halepensis woodlands, eastern Iberian Peninsula. Plant Ecology
638 169, 205–213. DOI: 10.1023/A:1026019528443
639 Perez, B., Moreno, J.M. 1998. Methods for quantifying fire severity in shrubland-fires. Plant
640 Ecology 139, 91–101. DOI: 10.1023/A:1009702520958
641 Raison, R.J. 1979. Modifications of the soil environment by vegetation fires, with particular
642 reference to nitrogen transformations: A review. Plant and Soil 51, 73–108. DOI:
643 10.1007/BF02205929
644 Santín, C., Doerr, S.H., Merino, A., Bryant, R., Loader, N.J. 2016. Forest floor chemical
645 transformations in a boreal forest fire and their correlations with temperature and heating
646 duration. Geoderma 264, 71–80. DOI: 10.1016/j.geoderma.2015.09.021

647 Shepherd, K.D., Walsh, M.G. 2002. Development of reflectance spectral libraries for
648 characterization of soil properties. Soil Science Society of America Journal 66, 988–998.
649 DOI:10.2136/sssaj2002.9880
650 Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J. 2010. Visible and near
651 infrared spectroscopy in soil science. Advances in Agronomy 107, 163–215. DOI:
652 10.1016/S0065-2113(10)07005-7
653 Terefe T., Mariscal I., Gómez M., Espejo R. 2005. Relationship between soil color and
654 temperature in the surface horizon of Mediterranean soils. A laboratory study. Soil
655 Science 170, 495–503. DOI: 10.1097/01.ss.0000175341.22540.93
656 Torrent, J., Barrón, V. 2002. Diffuse reflectance spectroscopy of iron oxides. Encyclopedia of
657 Surface and Colloid Science. Ed. Marcel Dekker. Pp. 1438-1446.
658 Ulery, A.L., Graham, R.C. 1993. Forest-fire effects on soil color and texture. Soil Science
659 Society of America Journal 57, 135–140.
660 Vega, J.A., Fontúrbel, T., Merino, A., Fernández, C., Ferreiro, A., Jiménez, E. 2013. Testing
661 the ability of visual indicators of soil burn severity to reflect changes in soil chemical and
662 microbial properties in pine forest and shrubland. Plant and Soil 369, 73–91. DOI:
663 10.1007/s11104-012-1532-9
664 Verdes, P.V., Salgado, J. 2011. Changes induced in the thermal properties of Galizian soils by
665 the heating in laboratory conditions: Estimation of the soil temperature during a wildfire.
666 Journal of Thermal Analysis and Calorimetry 104 (1), 177-186. DOI: 10.1007/s10973-
667 010-1173-2
668 Vieira, D.C.S., Fernández, C., Vega, J.A., Keizer, J.J. 2015. Does soil burn severity affect the
669 post-fire runoff and interrill erosion response? A review based on meta-analysis of field
670 rainfall simulation data. Journal of Hydrology 523, 452–464. DOI:
671 10.1016/j.jhydrol.2015.01.071

672 Viscarra Rossel, R.A., Behrens, T. 2010. Using data mining to model and interpret soil
673 diffuse reflectance spectra. Geoderma 158, 46–54. DOI: 10.1016/j.geoderma.2009.12.025
674 Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J.A.M., Shepherd,
675 K.D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aïchi, H., Barthès, B.G.,
676 Bartholomeus, H.M., Bayer, A.D., Bernoux, M., Böttcher, K., Brodský, L., Du, C.W.,
677 Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A., Guerrero, C.,
678 Hedley, C.B., Knadel, M., Morrás, H.J.M., Nocita, M., Ramirez-Lopez, L., Roudier, P.,
679 Campos, E.M.R., Sanborn, P., Sellitto, V.M., Sudduth, K.A., Rawlins, B.G., Walter, C.,
680 Winowiecki, L.A., Hong, S.Y., Ji, W. 2016. A global spectral library to characterize the
681 world's soil. Earth-Science Reviews 155, 198–230. DOI: 10.1016/j.earscirev.2016.01.012
682

Figures captions

Figure 1. Four spectral regions (a-d) displaying in detail the NIR spectra (absorbance after first derivative) of a soil sample, before heating (black colour, thin line) and after exposition to 700°C for 20 minutes (red colour, thick line).

Figure 2. Measured *vs* predicted TR values obtained during the cross-validation (leave-one-out) of four randomly selected models: a) model #5 (type 24s–v1) from Maigmó site; b) model #61 (type 72s–v4) from Aitana site; c) model #27 (type 96s–v4) from Pinoso site; d) model #57 (type 48s–v4) from Aitana site. See more details in Table 2, and supplementary content (Tables S1, S2 and S3). The black line denotes the 1:1, and the grey dashed line is the linear regression.

Figure 3. Representative predictions obtained with three models: a) predictions obtained in A5 with model #1, which was constructed with 24 heated aliquots from A1 and is therefore labelled 24s–v1; b) predictions obtained in A5 with model #35, which was constructed with 24 heated aliquots from A2 and A3 and is therefore labelled 24s–v2; c) predictions obtained in A2 with the model #62, which was constructed with 72 heated aliquots from A1, A3, A4 and A5 and is therefore labelled 72–v4. The black line denotes the 1:1, and the grey dashed line is the linear regression.

Figure 4. Values of the root mean square error of prediction (RMSEP) obtained with different model types in Aitana site (a), Maigmo site (b) and Pinoso site (c). The grey bars denote the mean value, whereas dots and triangles are used for minimum and maximum values respectively. Please note that the scale of the y-axis varies for each site.

711

712

713 **Figure 5.** Prediction performance parameters obtained with the different model types (results of the
714 three studies were averaged): a) RMSEP, SEP and bias; b) R^2 and RPD. The first number in the model
715 type label (24, 48, 72 or 96) denotes the number of heated aliquots (=number of spectra) used to
716 calibrate the model (i.e., model size). The number after –v denotes the variability, which is the number
717 of micro-plot scale soil (MPS) sample(s) used as source(s) of aliquots.

718

719

720 **Figure 6.** Measured *vs* predicted TR values: a) original data; b) 100ºC was added to original predicted
721 values; c) 50ºC was subtracted to original predicted values; d) original predicted values divided by two
722 (×0.5); e) random selection of 24 cases shown in panels a to d; f) random selection of 24 cases shown
723 in panels a to d; g) predictions obtained with models 24s–v1c (three sites pooled); h) predictions
724 obtained with models 96s–v4 (three sites pooled).

725

726

727 **Figure 7.** Panel 7a contains the schematic representation of four NIR spectra (stacked), where black
728 and grey circles denote presence and absence, respectively, of the heat-sensitive compounds at four
729 wavelengths. The black arrows in panel 7b represent wavelengths selected in seven stacked models
730 (relevant *b*–coefficients), depending on the composition of heat-sensitive compounds present in the
731 MPS used as source(s) of aliquots. The sign of *b*–coefficients could be negative or positive, being
732 denoted as down or up arrow, respectively. The panels 7c to 7f represents predictions obtained with
733 four models (Models 1 to 4) applied to MPS1, MPS2 and MPS3, and the fail and pass symbols denote
734 biased and unbiased predictions, respectively.

Highlights:

- NIR used to predict the temperature reached with different model types
- The higher the variability included in the calibration set, the better the accuracy
- High variability helps to fit with commonest thermosensible compounds
- Model size improves accuracy by enhancing precision (SEP)
- Calibration set should avoid low variability and composite sample

Near infrared spectroscopy to quantify the temperature reached in burned soils: importance of calibration set variability

Andrea Pérez-Bejarano, César Guerrero*

Departamento de Agroquímica y Medio Ambiente, Universidad Miguel Hernández de Elche, Avenida de la Universidad, s/n. E-03202 (Elche, Alicante), Spain

* Corresponding author e-mail: cesar.guerrero@umh.es

Abstract

Near infrared (NIR) spectroscopy is a potential technique for the quantification of the temperature reached (TR) in burned soils. Due to spatial variation, inaccurate predictions can result from calibrating a model with heat-sensitive compounds that are not present in the samples of the burned area. Therefore, we investigated how to develop robust models. The progressive augmentation of the model size successively enhanced the precision, while the increase of the calibration set's variability gradually improved the accuracy through decreases in bias. The increase in calibration set variability enhances the probability of calibration using only the most common heat-sensitive compounds, facilitating reliable predictions of TR regardless of the spatial variation. On the other hand, models calibrated with heated aliquots from a unique sample, even from a composite sample, should be totally avoided because, regardless of their apparent utility, they are prone to inaccurate predictions.

Keywords: wildfire effects; heat-sensitive compounds; fire intensity; NIR spectroscopy; robust models; postfire assessment

1. Introduction

Fire is an important process in several terrestrial ecosystems throughout the world. The immediate fire effects on soil properties are mainly related with the fire severity, where the TR in soil is a crucial parameter (Neary et al., 1999; Vieira et al., 2015). Despite its importance, soil heating or TR in soil is a parameter neither measured by post-fire assessment teams, such as the Burned Area Emergency Response (BAER), nor by researchers, mostly due to the lack of a standard, easy, cheap, rapid and accurate method (Fernández and Vega et al., 2016; Parson et al., 2010). Some of the existing methods are subjective (Vega et al., 2013), complex and expensive (Merino et al., 2014, 2015; Neris et al., 2014; Santín et al., 2016; Verdes and Salgado, 2011), have moderate accuracy (Melquiades and Thomaz, 2016; Pérez and Moreno, 1998), or need several indicators to derive wide classes or levels, such as the soil burn severity index (Jain et al. 2008; Morgan et al., 2014; Parsons et al., 2010). To fill this gap, Guerrero et al. (2007) proposed the use of near infrared (NIR) reflectance spectroscopy as a potential technique for the quantification of TR in burned soils. The approach suggested by Guerrero et al. (2007) is based on two basic premises: i) the NIR spectrum of a soil sample contains information about the organic matter (quantity and quality), clay (quantity and quality), minerals (such as carbonates and iron oxides) and water content (Nocita et al., 2015; Stenberg et al., 2010; Viscarra Rossel and Behrens, 2010; Viscarra Rossel et al., 2016); and ii) these components have different sensitivity to thermal shocks (DeBano et al., 1998; Knicker, 2007; Neary et al., 1999; Raison, 1979; Santín et al., 2016). Consequently, each temperature causes a group of changes in soil properties, leaving a typical fingerprint in the NIR spectra (Guerrero et al., 2007; Lugassi et al., 2010, 2014). Therefore, in this approach, the NIR spectrum is used as an integrative measurement of soil properties, which can be modified by the temperature (i.e., heat-sensitive compounds). However, as in other empirical approaches using NIR, a model is needed to relate the TR with

its fingerprint in the NIR spectrum. For the development of such a model, soil samples are heated at known temperatures, which are used as standards (i.e., as calibration set) to calibrate the model. Hence, once the model has been calibrated (using chemometrics), the TR can be efficiently measured in very large numbers of samples because the NIR spectrum of a soil sample is obtained in seconds, easily and without the need of chemical reagents (avoiding the generation of toxic wastes in laboratories). To obtain high resolution maps of TR, which can serve to locate prior intervention areas in burned sites, it is necessary to measure this parameter (TR) in hundreds or even thousands of soil samples (Parsons et al., 2010; Jain et al., 2012). Additionally, the TR can provide relevant information for a better description of fire effects at very fine scale and to enhance the understanding of fire ecology, such as for the *in situ* analysis of fire-mediated germination patterns (Keeley et al., 2008; Lentile et al., 2007; Pausas et al., 2003).

The assessment of TR using NIR in a wildfire-affected area implies the collection of burned samples in those target points where TR is needed (for instance, for mapping) but also the collection of unburned soil samples to calibrate the model (Guerrero et al., 2007). As a consequence of the short-scale natural variation of the soil properties, a sample located at an arbitrary position (regardless if burned or not) may present its particular composition of heat-sensitive compounds, and this composition can differ with respect to other samples located at different positions. Two issues are imposed by the presence of spatial variation, which, through its design, a suitable model should overcome. The first is the impossibility of constructing a model using samples with similar composition to those to be predicted (except for planned events, such as prescribed fires). This limitation may exert a negative effect on prediction accuracy because the model might be fitted with heat-sensitive compounds that might not be the same as those present in samples located in the wildfire-affected area. The second implication is that the wildfire will affect samples with different heat-sensitive

compounds; therefore, the model should be able to properly predict samples with different spectral responses. For these reasons, we found it interesting to investigate how we can develop robust models able to overcome such problems linked to the natural short-scale spatial variation. The calibration set characteristics play an important role in model performance. If the calibration set is composed by several different samples, only common heat-sensitive compounds (i.e., common spectral changes) should be used to fit the model. Consequently, accurate predictions are expected in those samples where the common heat-sensitive compounds are present. Therefore, the higher the variability included, the larger the commonality of the predictors, and the wider the model applicability. Thus, we hypothesised that increasing the calibration set variability should result in a progressive improvement of the predictions accuracy. To test the hypothesis, different model types were constructed and were then used to predict TR.

2. Materials and methods

2.1. Sites

This study was performed using forest soil samples collected in three different sites located in Alicante province (Spain): Aitana, Maigmó and Pinoso. These sites were approximately 30 km apart from each other. The vegetation of these sites is composed of *Pinus halepensis* Mill., as the dominant species in the tree stratum, the understory vegetation being dominated by species such as *Quercus coccifera* L., *Rosmarinus officinalis* L., *Juniperus oxycedrus* L., *Stipa tenacissima* L. and *Brachypodium retusum* (Pers.) P. Beauv. The main characteristics of the three study sites are given in Table 1.

2.2. Sample collection

In each site, five microplots (1 m^2) were randomly selected in a small area of

approximately 2500 m^2. The minimum and maximum distance between microplots was 5 and 50 m, respectively. In each microplot, a topsoil sample (0–5 cm depth) of approximately three kilograms was collected. These microplot-scale soil samples (hereafter MPS) collected in Aitana site were identified as A1, A2, A3, A4 and A5. Those MPS collected in Maigmó were identified as M1, M2, M3, M4 and M5. Similarly, the MPS collected in the Pinoso site were identified as P1, P2, P3, P4 and P5. In each site, an additional "composite sample" was obtained by bulking several subsamples collected at different points across the whole area; these composite samples were identified as Ac, Mc and Pc for Aitana, Maigmó and Pinoso, respectively.

2.3. Obtaining standards (laboratory-heated samples)

Once in the laboratory, the 18 samples (15 MPS + 3 composite samples) were air-dried for two weeks (at 25°C) and sieved to <2 mm. These 18 samples were the "sources of standards." Twenty-four aliquots of approximately 10 g were obtained from each sample. These aliquots were heated in a muffle furnace at 24 different combinations of temperatures (70°C, 100°C, 200°C, 300°C, 400°C, 500°C, 600°C and 700°C) and exposure times (10, 20 and 40 minutes). In this way, a set of 24 heated aliquots was obtained from each MPS sample (also from each composite sample). A total of 432 heated aliquots (24 heating combinations per sample × 6 samples [5 MPS + 1 composite] × 3 sites) were obtained, which were used as standards for the models (see section 2.6). Each aliquot was introduced in the pre-heated furnace as a 1-mm layer in order to guarantee homogeneous heating (Guerrero, 2010). During the heating, to register the exact TR, the temperature of the aliquot was monitored and recorded every 10 seconds using a thermocouple (Guerrero et al., 2007).

2.4. Obtaining the NIR spectra of the heated aliquots

Once cooled, the NIR spectra (12000–3800 cm^{-1}) of the 432 heated aliquots were obtained using a FT-NIR diffuse reflectance spectrophotometer (MPA Bruker, Germany). Further details about the scanning can be found in Guerrero et al. (2007). The x-scale of the spectra was transformed to nanometres (834–2630 nm) and resampled to 1 nm. All the spectra were transformed to absorbance, and then were pre-processed with the first derivative (Savitzky–Golay, 25 points) and vector normalization (Standard Normal Variate). The absorbance spectra of the 432 heated aliquots can be found as supplementary content. The OPUS spectroscopic software (OPUS version 6.5 software; BrukerOptik GmbH, Ettlingen, Germany) was used for spectral pre-processing.

2.5. Regression method

We have constructed different types of models relating the TR with the NIR spectra (see section 2.6). In all cases, the aliquots heated in laboratory (see section 2.3) were used as the "standards" to calibrate the models. All models were calibrated with partial least squares (PLS) as the multivariate regression method, using the OPUS spectroscopic software (OPUS version 6.5 software; BrukerOptik GmbH, Ettlingen, Germany). The leave-one-out cross-validation was used to decide the number of PLS vectors to be included in models on the basis of the root mean square error of cross-validation (RMSECV) in a scree plot.

2.6. Model types

Ten different model types were constructed in this study. The models differed in the size of the calibration set, and in the variability included. In this study, the *size* is the number of standards (i.e., number of heated aliquots) used to calibrate the model, and the *variability* is the number of different samples used as source(s) of heated aliquots. The "*label*" of the model type provides information on the size and the variability (Table 2). For instance, the label

72s–v3 denotes a model calibrated with 72 heated aliquots (72 standards) which were obtained from three different samples (three MPS have been used as sources of heated aliquots). At each site, ten different model types were constructed as follows (Table 2):

1) Models 24s–v1: these models were calibrated with 24 heated aliquots (24 standards) obtained from one MPS. In each site, five models were constructed (models #1 to #5; Table 2).

2) Models 48s–v2: these models were calibrated with 48 heated aliquots (48 standards) obtained from two MPS. In each site, ten models were constructed (#6 to #15; Table 2).

3) Models 72s–v3: these models were calibrated with 72 heated aliquots (72 standards) obtained from three MPS. In each site, ten models were constructed (#16 to #25; Table 2).

4) Models 96s–v4: models calibrated with 96 heated aliquots (96 standards) obtained from four MPS. In each site, five models were constructed (#26 to #30; Table 2).

5) Models 24s–v2: models were calibrated with 24 standards selected from two MPS. In each site, ten models were constructed (#31 to #40; Table 2).

6) Models 24s–v3: models calibrated with 24 standards selected from three MPS. In each site, ten models were constructed (#41 to #50; Table 2).

7) Models 24s–v4: models calibrated with 24 standards selected from four MPS. In each site, five models were constructed (#51 to #55; Table 2).

8) Models 48s–v4: models calibrated with 48 standards selected from four MPS. In each site, five models were constructed (#56 to #60; Table 2).

9) Models 72s–v4: models calibrated with 72 standards selected from four MPS. In each site, five models were constructed (#61 to #65; Table 2).

10) Models 24s–v1c: models calibrated with the 24 heated aliquots (24 standards) obtained from the composite sample. In each site, one model was constructed (model #66; Table 2).

A total of 66 models were constructed for each site (Table 2). The description of the MPS used for each of model can be found in Tables S1, S2 and S3 for Aitana, Maigmó and Pinoso site, respectively. In some cases (those marked with an asterisk in Table 2), the required model size is smaller than the total number of available spectra. In these cases, a principal component analysis (PCA) was performed with the NIR spectra of the available spectra. Then, the Kennard–Stone algorithm was used to select the desired number of spectra. The selected spectra were those whose scores were evenly distributed across the space defined by the first three principal components. For instance, a model 48s–v4 (as an example) is a model calibrated with 48 heated aliquots (48 standards), which were obtained from four different samples. Since four MPS were used as sources of heated aliquots, a total of 96 spectra were available. Thus, a PCA was conducted with the 96 spectra, but only 48 spectra were used to calibrate the model. These 48 spectra were selected with the Kennard–Stone algorithm. The minimum variability would have been reached when all the standards used to calibrate the model had been derived from a unique MPS sample (such as in models 24s–v1). The maximum variability would have been reached when each standard included in the calibration set has been obtained from a different sample (i.e., only one heated aliquot per MPS sample). Such maximum variability has not been evaluated in this study, as the maximum number of different MPS used as sources of heated aliquots was four (Table 2).

2.7 Predictions and data analysis

The previously described models were used to predict the TR. Several prediction performance parameters were computed to analyse the quality of the predictions: determination coefficient (R^2), root mean square error of prediction (RMSEP), standard error of prediction (SEP), bias and ratio of performance to deviance (RPD), which is the standard deviation of the prediction set divided by the RMSEP (Stenberg et al., 2010; Bellon-Maurel et

200 al., 2010). The RMSEP was used to measure the accuracy, and the SEP was used to measure
201 the precision (Bellon-Maurel et al., 2010; Næs et al., 2002).

202 These prediction performance parameters were computed using predictions obtained
203 in 24 heated aliquots from a MPS not used to calibrate the model. Therefore, the setup
204 provided independent predictions, similar to a MPS-hold-out cross-validation or leave-one-
205 MPS-out cross-validation. In this way (although with some limitations), the approach tried to
206 mimic a realistic scenario, where the samples used to construct the models, and those to be
207 predicted, are irremediably located on different plots, since the former would be located in the
208 unburned area and the latter in the burned area.

209 A complete identification of the MPS used as prediction set in each case can be found
210 in Tables S1, S2 and S3 (indicated by a cross in the right side of the tables). At each site, the
211 66 models were used 145 times, and therefore, 145 different values of R^2, RMSEP, SEP, bias
212 and RPD were obtained for each site (see supplementary Tables S1, S2 and S3). In each site,
213 those values of R^2, RMSEP, SEP, bias and RPD obtained with the same model type were
214 considered as replicates and were therefore averaged by model type. Since the bias can be
215 positive or negative, in order to get a meaningful average, we used its absolute value.

216 The prediction performance parameters, once averaged by model type, were arranged
217 in four groups, to facilitate the analysis of the results (Table 2). To facilitate the description
218 and meaning of groups and its comparisons, some results included in the Group *i* have also
219 been included in Group *ii* (24s–v1) and Group *iii* (96s–v4), as follows:

220 – Group *i*: this group contains results from models that differed in both the variability
221 (number of MPS used as source of standards) and model size (number of standards included
222 in the calibration set; Table 2).

223 – Group *ii*: this group contains results from models that differed only in variability because
224 the model size was constant (Table 2).

225 – Group *iii*: this group contains results from models that differed only in size because the
226 variability was constant (Table 2).
227 – Group *iv*: this group contains results obtained from the model calibrated with the composite
228 sample (Table 2).
229
230 **3. Results**
231 **3.1. Changes in soils and NIR spectra**
232 The colour of the soil samples was modified as a consequence of heating. Due to the
233 carbonization of the soil organic matter, a progressive darkening was observed when the
234 exposure temperature was increased (up 450°C). As a consequence of such darkening, the
235 baseline of the NIR spectra (absorbance) was increased, especially at shorter wavelengths
236 close to the visible spectral range (see supplementary figures Figs. S1-S27). The dark colour
237 abruptly disappeared when the temperature was above ~450°C, mainly due to the combustion
238 of the organic compounds. The transformation of some iron oxides into hematite was an
239 additional change in the NIR spectra (approximately 890–892 nm; Fig. 1a) and in soil colour,
240 causing sample reddening (Ketterings and Bigham, 2000; Torrent and Barrón, 2002; Terefe et
241 al., 2005; Ulery and Graham, 1993), which was especially evident at high temperatures
242 (>500°C).
243 While changes at naked eye are the basis of visual estimators of fire severity (Pérez
244 and Moreno, 1998; Vega et al., 2013), the NIR spectra can provide further additional
245 information about the changes in soil properties, since NIR spectroscopy has been used to
246 quantify a large number of soil properties, which in turn are affected by temperature (Santín et
247 al., 2016; Stenberg et al., 2010). Other important changes in the NIR spectra were those
248 observed in features located approximately 1414 nm (Fig. 1b), 1920 nm (Fig. 1c) and 2210
249 nm (Fig. 1d), related with adsorbed and free water –OH (at 1414 nm and 1920 nm) and clay –

250 OH (at 1414 nm and 2210 nm). These spectral features tended to decrease with the increase in
251 TR, mostly due to dehydration and dehydroxylation processes but, as other authors have
252 observed (Guerrero et al., 2007; Lugassi et al., 2010, 2014), also due to the combustion of the
253 organic matter. A detailed discussion of changes in soil properties can be found in Certini
254 (2005), and a detailed discussion of changes in the NIR spectra can be found in Lugassi et al.
255 (2014).
256
257 **3.2. Calibrations**
258 All the above mentioned changes, and others, facilitated the calibration of models to
259 quantify TR. Therefore, we could surmise that several heat-sensitive compounds were
260 involved in the calibrations. Despite the large differences between calibration sets, the 198
261 models showed high R^2 values (ranging from 0.92 to 0.99), and low RMSECV (RMSE of
262 cross-validation) values, ranging from 21°C to 58°C (data not shown). These values of R^2 and
263 RMSECV obtained in the cross-validations were similar to those from previous studies
264 (Guerrero et al., 2007; Guerrero, 2010; Maia et al., 2012). Fig. 2 shows the cross-validation
265 results of four randomly selected models, as a representative illustration of the 198 models
266 calibrated for this study. These results indicate the great capacity of NIR spectroscopy to
267 measure some of the soil properties that change with temperature (i.e., heat-sensitive
268 compounds). Apparently, they all seem to be useful to predict TR.
269
270 **3.3. Predictions**
271 The 435 values of RMSEP obtained with the different 198 models ranged from 14.3°C
272 to 99.7°C. Three out of these 435 cases are shown in Fig. 3, as representative predictions
273 obtained with three different types of models. Fig. 4 shows the RMSEP obtained with the
274 models included in Group *i*, which were models constructed with all the heated aliquots from

one (24s–v1), two (48s–v2), three (72s–v3) and four MPS (96s–v4). The grey bars in Fig. 4 represent the RMSEP values averaged by model type (separately for each study site), whereas the dots and triangles are used to show the minimum and the maximum RMSEP observed for each model type, respectively. For the Aitana site (Fig. 4a), the RMSEP of predictions obtained with the 20 models labelled as 24s–v1 ranged from 31.2°C to 60.2°C, with an average of 40.9°C. In Maigmó (Fig. 4b), these 20 RMSEP values were slightly lower than in Aitana, and ranged from 18.6°C to 44.9°C, with an average of 31.2°C. In Pinoso, the 20 RMSEP values obtained with models 24s–v1 were slightly higher than the other study sites, and ranged from 24.8°C to 99.7°C, with an average of 49.4°C (Fig. 4c). Regardless of the site considered, the progressive increase of the calibration set variability (i.e., number of MPS) and size (i.e., number of standards) coincided with the gradual decline of RMSEP (Fig. 4). Other prediction performance parameters (such as R^2, SEP, etc.) also indicated an improvement of predictions with the increase in variability and size of the calibration sets. Since similar patterns were observed for the three study sites (data not shown), the values of the three sites were pooled together, and averaged by model type (Fig. 5). Once averaged by model type, the RMSEP dropped progressively from 40.5°C in models 24s–v1 to 29.6°C in models 96s–v4 (Fig. 5a). A similar gradual change was also observed for the other prediction performance parameters, such as the R^2, SEP, bias and RPD (Fig. 5). For instance, RPD was 6.2 in models 24s–v1, and it increased to RPD 8.1 in models 96s–v4 (Fig. 5b). The improvement of the predictions might be attributed to the higher variability included in the calibration set, which was four times higher in models 96s–v4 than in models 24s–v1, although such improvement could also be attributed to the calibration set size, which was also four times higher. To discern whether the calibration set variability affects accuracy, we constructed another group of models (Group *ii*) where the number of MPS varied while the model size was constant at 24 in all cases (Table 2). For the average of the three sites, the

RMSEP decreased from 40.5°C in 24s–v1 to 35.7°C in 24s–v4, confirming the positive effect of the increase in the calibration set variability on accuracy (Fig. 5a). The decrease in bias from 15.1°C in 24s–v1 to 8.2°C in 24s–v4 was the main contributor to the RMSEP reduction observed in Group *ii*, because SEP only decreased by 2°C (Fig. 5a). The increase of the calibration set size was the additional factor explaining the progressive decrease of RMSEP observed in the models from Group *i* (Fig. 4 and Fig. 5). Nevertheless, the size effect was clearly observed when the RMSEP was compared for the models included in Group *iii* (Table 2), where the size was the unique difference between them. For that group of models (Group *iii*), the decrease in RMSEP from 35.7°C in 24s–v4 to 29.6°C in 96s–v4 is mostly attributable to an improvement in precision (lower SEP) because the bias remained fairly stable (only decreased less than 1°C), whereas the SEP decreased by approximately 6.5°C (Fig. 5a).

The less accurate predictions were obtained with models constructed with the heated aliquots from the composite sample (model type labelled 24s–v1c). On average across the three sites, the bias was 22°C, explaining a substantial portion of the RMSEP (47.6°C). These errors were clearly higher than those obtained with models of similar size, such as models 24s–v4. Furthermore, these errors were also higher than those obtained with models having a similar degree of variability (i.e., same number of MPS) such as the models 24s–v1.

The R^2 values shown in Fig. 5b should be interpreted with some caution since they are somehow affected by the approach used to compute results. In this study, the prediction performance parameters (such as R^2) have been computed using the predictions obtained in sets of 24 heated aliquots, which belong to the same MPS (see tables S1, S2 and S3). Consequently, the prediction performance parameters have been obtained at MPS scale (and were then averaged by model type). At MPS scale, biased predictions do not affect the R^2. This can be seen in Figs. 6a to 6d. Fig. 6a shows TR predictions obtained in 24 heated aliquots from one randomly selected MPS. The cases shown in Figs. 6b, 6c and 6d are the

same predictions shown in Fig. 6a but after the manipulation of the predicted values in order to have differently biased predictions. In Fig. 6b, 100°C was added to each predicted value shown in Fig. 6a; in Fig. 6c, 50°C was subtracted from each predicted value; in Fig. 6d, each predicted value was multiplied by 0.5 (divided by two). In all cases (Figs. 6a to 6d), the R^2 is the same (R^2=0.99) regardless of the bias (due to slope or due to offset). If we compute the mean value of these four cases, the R^2 is 0.99 despite the different bias in each case. However, under realistic conditions, the evaluation of the burned area implies making predictions from samples collected at several positions. If the model predictions have a different bias depending on the sample composition (which may vary with position), then we must expect patterns as those shown in Figs. 6e and 6f, where due to differences in bias, such heterogeneity in the prediction set results in much lower R^2 values. Figs. 6e and 6f contain a random selection of 24 cases from those shown in Figs. 6a to 6d, and simulate two prediction sets composed by samples with different bias. In these cases (Figs. 6e and 6f), the different values of bias (by slope or by offset) were negatively affecting the R^2. In these examples, the values were R^2=0.72 and R^2=0.63, resulting in a mean R^2=0.675, which clearly contrasts with the R^2=0.99, obtained as a mean of four values of R^2 obtained at MPS scale. To have a direct measure of such discrepancy, we also computed the R^2 after pooling predictions from the different MPS, mimicking prediction sets composed by samples that can have a different bias. For that, predictions from two illustrative model types were selected: 1) predictions obtained with models 24s–v1c (Fig. 6g), which was the worst option in terms of bias (highest bias in Fig. 5a); 2) predictions obtained with models 96s–v4 (Fig. 6h), which was the best option in terms of bias (lowest bias in Fig. 5a). A small discrepancy between the approaches was observed when the predictions were obtained with models 96s–v4 (low bias; Fig. 6h), with R^2=0.981 when the R^2 was computed once predictions were pooled (Fig. 6h) and R^2=0.985 when R^2 was computed as an average of values obtained at MPS scale (Fig. 5b). As expected,

a larger discrepancy was observed when predictions were obtained with models 24s–v1c (high bias; Fig. 6g), with R^2=0.954 when predictions were pooled (Fig. 6g) while R^2=0.971 when it was computed as an average of the five cases (Fig. 5b). Regardless of the approach, the R^2 values were high because the bias values were not very large—clearly lower than those shown in Figs. 6b or 6d. Moreover, the wide range included in the prediction sets also contributed to the high R^2 values (Davies and Fearn, 2006). Despite their absolute values, the R^2 values shown in Fig. 5b were in concordance with other performance parameters and therefore support the same conclusions about the importance of the calibration set size and variability. It is worth highlighting that other prediction performance parameters, such as RMSEP, in contrast to R^2, are minimally affected by the way results were computed because RMSEP is a parameter related to the residuals, given that it is linked to each predicted value.

4. Discussion

Soil contains heat-sensitive compounds, which can be can be measured with NIR spectroscopy, allowing the development of models to quantify the TR. However, the composition of soils, including its heat-sensitive compounds, varies with space, even at short distances. Therefore, each sample might possess its particular composition of heat-sensitive compounds and that can exert an important effect on prediction reliability. We have observed an effect of the calibration set variability on accuracy, which highlights important issues about how adequate calibration sets should be developed to manage the spatial variation effects. To facilitate the explanation of mechanisms connecting the calibration set variability with the robustness of predictions, we discuss the results with the help of a schematic representation of spectra, models and predictions (Fig. 7), since a direct analysis of the *b*–coefficients of PLS–models is not straightforward. Such simplification enhances the illustration and understanding of mechanisms involved. A schematic representation of the NIR spectra of three MPS and one

composite sample is shown in Fig. 7a: a black circle denotes the wavelength (λ) where the spectral feature of a heat-sensitive compound appears, and a grey circle denotes its absence in the sample. Fig. 7b contains the schematic representation of seven models calibrated with aliquots from samples shown in Fig. 7a. The black arrows denote those wavelengths included in the models, which depend on the heat-sensitive compounds present in the calibration set. Thus, a black arrow at λ_n denotes where the *b*–coefficient has a large contribution in predicting TR. As consequence of heating, some compounds trend to disappear, decreasing their spectral features as temperature increases, such those related with –OH loss by dehydroxilation (and organic matter combustion, in general). However, other compounds can be generated by heating, such as certain iron oxides (as hematite), and thus they appear as new spectral features as the temperature increases. Consequently, the sign of *b*–coefficients could be negative or positive, being denoted as down or up arrows, respectively.

The first case shown at the top of Fig. 7b (case *i*) represents an example of a model whose calibration set had minimal variability (variability=1) because all the heated aliquots included in the calibration set had been obtained from the same sample (MPS1). Therefore, this is a sample-specific model. This model had been fitted only on the basis on changes produced in the heat-sensitive compounds present in that sample (those located at λ_3 and λ_4; see black dots in Fig. 7a). As consequence, this model will predict TR using two *b*–coefficients located at λ_3 and λ_4 (see black arrows at λ_3 and λ_4 in Fig. 7b, case *i*). The second and third models shown in Fig. 7b (cases *ii* and *iii*) are also examples of models constructed with heated aliquots obtained from a unique sample (MPS2 for case *ii*; MPS3 for case *iii*), therefore, these are also sample-specific models (variability=1). Since each MPS sample would have had a different composition (Fig. 7a), we would therefore expect that the models were based on the different wavelengths from each sample, as shown in Fig. 7b for cases *i* to *iii*. The models shown in cases *iv* and *v* of Figure 7b represent models constructed with heated

aliquots from two MPS samples (i.e., variability=2). While these two MPS samples might have had a different composition, only the common changes (common heat-sensitive compounds) were used to fit the model. Similarly, when three MPS were used as the source of the aliquots (Fig. 7b, case *vii*), three different sets of heat-sensitive compounds might have been be present in the calibration set, but the model was fitted only with the common changes (i.e., common heat-sensitive compounds, as in that located at λ_3). Thus, the increase in the number of MPS used as sources of standards (i.e., the increase of the calibration set variability) forces the calibration of models on the basis of the most common heat-sensitive compounds. It is worth mentioning that a model based on the most common heat-sensitive compounds could also have been obtained with only two samples (MPS1 and MPS2), as illustrated in Fig. 7b *iv*. However, two samples might not be enough, as in the case shown in Fig. 7b *v*, explaining why a gradual augmentation in calibration set variability progressively increases the chance to fit a model based only on the most common heat-sensitive compounds. Fig. 7b clearly illustrates that several different models can be obtained with the samples from a given site. Apparently, all of them seem to be useful since they relate NIR and TR; however, most of them would have restricted applicability.

Figs. 7c to 7f represent the predictions obtained in MPS1, MPS2 and MPS3 from four different models. The model used in each case is described by its *b*–coefficients (black arrows). When the model is used in each of the three samples, a pass or fail symbol is added to denote unbiased or biased predictions, respectively. We can expect accurate predictions when the sample used for prediction has the same heat-sensitive compounds as the predictors in the model. For instance, Model 1 (shown in Fig. 7c) predicts TR using the changes in the heat-sensitive compounds located at λ_1 and λ_3 (black arrows). Thus, accurate predictions are expected when Model 1 is used for sample MPS2 (Fig. 7c). Conversely, biased predictions are expected when Model 1 (Fig. 7c) is used for samples MPS1 and MPS3 because these

425 samples lack the heat-sensitive compound located at λ_1. Since the contribution of λ_1 to the
426 prediction is missing, then we would expect negatively biased predictions in MPS1 and MPS3
427 (a 50% underestimation assuming a similar contribution from both heat-sensitive
428 compounds). However, Figs. 7d and 7e show how we can obtain accurate predictions from
429 samples that contain a different composition of heat-sensitive compounds. This is the case for
430 Model 2 when used for sample MPS2 (Fig. 7d). The presence of an additional heat-sensitive
431 compound in MPS2 (the one located at λ_3) does not distort the predictions obtained with
432 Model 2, because this model does not consider the changes in wavelength λ_3 when predicting
433 TR. Whereas Model 2 can predict without bias only for one sample (MPS2; Fig. 7d), Model 3
434 can do so for three samples (Fig. 7e). The reason is that Model 3 (Fig. 7e) predicts TR with
435 the heat-sensitive compound located at λ_3, which is the most common. Therefore, a model
436 such as this can predict without bias in a wide number of different samples. The unique
437 requirement is the presence of that heat-sensitive compound, which most of the samples
438 would contain since it is a very common compound. Therefore, the better identification of the
439 most common heat-sensitive compounds might be the link which explains why the
440 progressive augmentation of the calibration set variability successively improved prediction
441 accuracy.

442 In agreement with this supposed mechanism, a high variability in the calibration set
443 would allow obtaining reliable (not biased) predictions in dissimilar samples as well (Fig. 7e),
444 overcoming the restriction of predictions based only on similar samples. Consequently, it
445 increases the spatial applicability of the models, since they would not only be restricted to
446 adjacent sites, where similar samples would be expected. A high variability does not
447 completely exclude the possibility of bias in predictions, although this would be expected
448 only for those samples where the most common heat-sensitive compound is absent. Especially
449 if the collection of unburned samples used as sources of aliquots is sufficiently representative

of the unburned area, this should be a minor proportion of the total samples. Additionally, in turn, the unburned area would also have to be representative of the burned area. Under these conditions, a calibration set constituted by a number of different samples could be considered a site-specific model.

This theory also explains why the most biased predictions were obtained with the 24s–v1c models. For an arbitrary study site, all the heated aliquots used for the model 24s–v1c were obtained from the composite sample; therefore, it can considered as a sample-specific model. The composite sample, as a bulked mixture of several subsamples, should contain most of the heat-sensitive compounds (Fig. 7a). As a consequence, all these compounds would probably have contributed to the calibration of the model (Fig. 7b, case *vi*). Thus, this model (such as Model 4, Fig. 7f) should produce biased predictions in those samples where at least one of the heat-sensitive compounds is absent. The observed results fully support this theory, since the most biased predictions were obtained with models 24s–v1c. Clearly, any attempt to include site variability through the use of a composite sample is not recommended. A composite sample would describe a representative fingerprint of heat on soil properties, but this would not be useful to obtain robust models because they would be prone to produce biased predictions in many samples. Therefore, the development of models using a composite sample should be totally avoided. This result might explain why, when using predictions from this type of model, some authors did not observe the expected relationships between changes in soil properties and temperature (Maia et al., 2012; Mataix-Solera et al., 2013).

The number of heated aliquots (model size) also had an impact on the accuracy. Whereas the number of MPS decreased the bias, the number of heated aliquots increased the accuracy through the improvement of precision (i.e., the decrease of SEP). A high number of heated aliquots in the calibration set of a model should better describe the heat effect on the heat-sensitive compounds. Consequently, it enhances the quantification of the *b*–coefficient

values of the implied heat-sensitive compounds, improving the calibration of models, which is a basic premise for obtaining accurate predictions. Clearly, the best option is the development of models with a large number of heated aliquots from several samples. The importance of the variability and the number of samples used for calibration has been already demonstrated by many authors with other soil properties (Brown, 2007; Shepherd and Walsh, 2002).

Very accurate predictions have been obtained in this study, even when the worst models were used, which were those calibrated with low variability (such as those 24s–v1 and 24s–v1c). This is probably due to the very small size of the area (2500 m^2) where the five MPS were sampled in each site. Thus, these samples were separated by short distances (less than 50 m; see section 2.2), and consequently, they were quite similar to each other. Hence, it is reasonable to expect a substantially lower accuracy when this model type is used under realistic conditions, where larger distances (and larger dissimilarity) may exist between burned and unburned samples. Thus, we should also expect a larger difference in the accuracy of predictions obtained with models calibrated with low and high variability.

This study provides valuable information about factors affecting the reliability of models. Most of these results are very interesting because they have a direct impact on how to design the calibration set, which even affects the collection of the unburned samples to be used as sources of aliquots, and the built of a calibration set for site-specific models. Nevertheless, more studies are needed to understand the limitations and capabilities of NIR spectroscopy to predict TR and how to construct models in a more efficient way. For instance, the transferability of models across sites and scales is an important issue because it may facilitate a faster development of models, which may substantially decrease the time needed to evaluate wildfire-affected sites. Dedicated experiments should be designed for that purpose, especially because some of the results obtained in previous studies (Guerrero et al., 2007) have suggested its feasibility. These authors constructed several hundred different

models using soil samples from four sites. These models differed from each other in their basic characteristics such as rank (number of PLS vectors), spectral range included and pre-processing (normalization, derivatives, etc.). Then, a few of these models (but not all of them) were able to predict TR in samples from a new independent site (as left-out site). Similar results were observed when repeated with each of the five sites as a prediction set. Thus, the results undoubtedly demonstrated the presence of common heat-sensitive compounds in soils from five different sites in Alicante province (southeast Spain). Anyway, it should be noted that although independent sites were used in Guerrero et al. (2007), the identification of the *best* model (correct design) was based on the results from the leave-out site and thus not an independent selection. Therefore, these results should not be considered as the expected accuracy for a true independent prediction, since hundreds of models (varying in its design) can be fitted, but not all of them would produce accurate predictions. Nonetheless, the results shown in Guerrero et al. (2007) were useful to illustrate the potential of NIR to predict TR. For that reason, the results of the present study have an additional value because they provide the first measure of the accuracy obtained by independent predictions.

The improvement of the ability to obtain more realistic standards (mimicking naturally burned soils) is also a step needed in the way towards the use of NIR as a thermometer. For instance, the convenience to built models including standards generated under different oxygen limitation levels should be also evaluated in future studies. Oxygen limitation level during heating could be an important variable since it controls the type of thermal changes in soil organic matter during heating (from combustion to pyrolysis), and it may represent one of the most important difference between naturally burned soils respect to the artificial standards, which act as their surrogates to calibrate models.

5. Conclusions

525 The models to predict TR through NIR are calibrated with standards generated by
526 heating soil samples at known temperatures under controlled conditions in the laboratory.
527 When all the standards are obtained from a unique sample, the variability of the calibration set
528 is minimal. In this case, the model is calibrated with the heat-sensitive compounds present in
529 the sample. However, due to spatial variation, these compounds can differ in samples located
530 at distant positions. Therefore, each sample-specific model may be based on a particular,
531 unique and singular group of predictors, which are virtually different in each model. Thereby,
532 we can fit a very large number of different models, although the applicability might be
533 restricted to similar samples. On the other hand, when the calibration set includes standards
534 from different samples (i.e., high variability), then the model is fitted only with the common
535 heat-sensitive compounds. If the number of different samples is high enough to cover the site
536 variability, then the model can be considered site-specific. In contrast with a sample-specific
537 model, a site-specific model can predict without bias in a large number of different samples,
538 being able to overcome the negative effects of the spatial variation. Despite their apparent
539 utility, models calibrated with low variability or those based on a composite sample should be
540 avoided since they might be not useful. In addition, the model size (number of standards) is
541 also an important factor to consider during the calibration of models. While the increase in
542 variability improved prediction accuracy (lower RMSEP) by a decrease in bias, the increase
543 of model size did so through the enhancement of precision (i.e., decrease of SEP).
544
545 **References**
546 Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.M. McBratney, A. 2010.
547 Critical review of chemometric indicators commonly used for assessing the quality of the
548 prediction of soil attributes by NIR spectroscopy. TrAC Trends in Analytical Chemistry
549 29, 1073-1081. DOI: 10.1016/j.trac.2010.05.006.

Brown, D.J. 2007. Using a global VNIR soil-spectral library for local soil characterization and landscape modelling in a 2nd-order Uganda watershed. Geoderma 140, 444–453. DOI: 10.1016/j.geoderma.2007.04.021

Certini, G. 2005. Effects of fire on properties of forest soils: A review. Oecologia 143, 1–10. DOI: 10.1007/s00442-004-1788-8

Davies, A., Fearn, T. 2006. Back to basics: calibration statistics. Spectroscopy Europe 18, 31–32

DeBano, L.F., Neary, D.G., Ffolliott, P.F. 1998. Fire's effects on ecosystems. John Wiley & Sons, New York.

Fernández, C., Vega, J.A. 2016. Modelling the effect of soil burn severity on soil erosion at hillslope scale in the first year following wildfire in NW Spain. Earth Surface Processes and Landforms 41, 928–935. DOI: 10.1002/esp.3876

Guerrero, C. 2010. Espectroscopía de infrarrojo cercano (NIR) para la estimación de las temperaturas alcanzadas en suelos quemados. In: Actualización en Métodos y Técnicas para el Estudio de los Suelos Afectados por Incendios Forestales. Pp. 259–287. Ed. Cátedra de divulgación de la Ciencia. Universitat de Valencia. ISBN: ISBN 978-84-370-7887-8.

Guerrero, C., Mataix-Solera, J., Arcenegui, V., Mataix-Beneyto, J., Gómez, I. 2007. Near-infrared spectroscopy to estimate the maximum temperatures reached on burned soils. Soil Science Society of America Journal 71, 1029–1037. DOI: 10.2136/sssaj2006.0187

Jain, T.B., Gould, W.A., Graham, R.T., Pilliod, D.S., Lentile, L.B., González, G. 2008. A soil burn severity index for understanding soil-fire relations in tropical forests. Ambio 37, 563–568.

573 Jain, T.B., Pilliod, D.S., Graham, R.T., Lentile, L.B., Sandquist, J.E. 2012. Index for
574 characterizing post-fire soil environments in temperate coniferous forests. Forest 3, 445–
575 466. DOI: 10.3390/f3030445
576 Keeley, J.E., Brennan, T., Pfaff, A.H. 2008. Fire severity and ecosystem responses following
577 crown fires in California shrublands. Ecological Applications 18, 1530–1546. DOI:
578 10.1890/07-0836.1
579 Ketterings, Q.M., Bigham, J.M. 2000. Soil Color as an Indicator of Slash-and-Burn Fire
580 Severity and Soil Fertility in Sumatra, Indonesia. Soil Science Society of America
581 Journal 64, 1826–1833. DOI:10.2136/sssaj2000.6451826x
582 Knicker H. 2007. How does fire affect the nature and stability of soil organic nitrogen and
583 carbon? A review. Biogeochemistry 85, 91–118. DOI: 10.1007/s10533-007-9104-4
584 Lentile, L.B., Morgan, P., Hudak, A.T., Bobbitt, M.J., Lewis, S.A., Smith, A.M., Robichaud,
585 P.R. 2007. Post-fire burn severity and vegetation response following eight large wildfires
586 across the western United States. Fire Ecology 3, 91–108.
587 Lugassi, R., Ben-Dor, E., Eshel, G. 2010. A spectral-based method for reconstructing spatial
588 distributions of soil surface temperature during simulated fire events. Remote Sensing of
589 Environment 114, 322–331. DOI: 10.1016/j.rse.2009.09.015
590 Lugassi, R., Ben-Dor, E., Eshel, G. 2014. Reflectance spectroscopy of soils post-heating—
591 Assessing thermal alterations in soil minerals. Geoderma 213, 268–279. DOI:
592 10.1016/j.geoderma.2013.08.014
593 Maia, P., Pausas, J.G., Arcenegui, V., Guerrero, C., Pérez-Bejarano, A., Mataix-Solera, J.,
594 Varela, M.E.T., Fernandes, I., Pedrosa, E.T., Keizer, J.J. 2012. Wildfire effects on the
595 soil seed bank of a maritime pine stand - The importance of fire severity. Geoderma 191,
596 80–88. DOI: 10.1016/j.geoderma.2012.02.001

597 Mataix-Solera, J., Arcenegui, V., Tessler, N., Zornoza, R., Wittenberg, L., Martínez, C.,
598 Caselles, P., Pérez-Bejarano, A., Malkinson, D., Jordán, M. 2013. Soil properties as key
599 factors controlling water repellency in fire-affected areas: evidences from burned sites in
600 Spain and Israel. Catena 108, 9–16. DOI: 10.1016/j.catena.2011.12.006
601 Melquiades, F.L., Thomaz, E.L. 2016. X-ray fluorescence to estimate the maximum
602 temperature reached at soil surface during experimental slash-and-burn fires. Journal of
603 Environmental Quality 45, 1104–1109. DOI: 10.2134/jeq2015.06.0305
604 Merino, A., Chávez-Vergara, B., Salgado, J., Fonturbel, M.T., García-Oliva, F., Vega, J.A.
605 2015. Variability in the composition of charred litter generated by wildfire in different
606 ecosystems. Catena 133, 52–63. DOI: 10.1016/j.catena.2015.04.016
607 Merino, A., Ferreiro, A., Salgado, J., Fontúrbel, M.T., Barros, N., Fernández, C., Vega, J.A.
608 2014. Use of thermal analysis and solid-state C CP-MAS NMR spectroscopy to diagnose
609 organic matter quality in relation to burn severity in Atlantic soils. Geoderma 226, 376–
610 386. DOI: 10.1016/j.geoderma.2014.03.009
611 Morgan, P., Keane, R.E., Dillon, G.K., Jain, T.B., Hudak, A.T., Karau, E.C., Sikkink, P.G.,
612 Holden, Z.A., Strand, E.K. 2014. Challenges of assessing fire and burn severity using
613 field measures, remote sensing and modelling. International Journal of Wildland Fire 23,
614 1045–1060. DOI: 10.1071/WF13058
615 Næs, T., Isakson, T., Fearn, T., Davies, T. 2002. A user-friendly guide to multivariate
616 calibration and classification. NIR Publications, Chichester. ISBN 9780952866626
617 Neary, D.G., Klopatek, C.C., DeBano, L.F., Ffolliott, P.F., 1999. Fire effects on belowground
618 sustainability: a review and synthesis. Forest Ecology and Management 122, 51–71. DOI:
619 10.1016/S0378-1127(99)00032-8

620 Neris, J., Doerr, S.H., Tejedor, M., Jiménez, C., Hernández-Moreno, J.M. 2014. Thermal
621 analysis as a predictor for hydrological parameters of fire-affected soils. Geoderma 235,
622 240–249. DOI: 10.1016/j.geoderma.2014.07.018
623 Nocita, M., Stevens, A., vanWesemael, B., Aitkenhead,M., Bachmann, M., Barthès, B., Ben-
624 Dor, E., Brown, D.J., Clairotte, M., Csorba, A., Dardenne, P., Demattê, J.A., Genot, V.,
625 Guerrero, C., Knadel, M., Montanarella, L., Noon, C., Ramirez-Lopez, L., Robertson, J.,
626 Sakai, H., Soriano-Disla, J.M., Shepherd, K.D., Stenberg, B., Towett, E.K., Vargas, R.,
627 Wetterlind, J., 2015. Soil spectroscopy: an alternative to wet chemistry for soil
628 monitoring. Advances in Agronomy 132, 139–159. DOI: 10.1016/bs.agron.2015.02.002
629 Parson, A.; Robichaud, P.R.; Lewis, S.A.; Napper, C.; Clark, J.T. 2010. Field guide for
630 mapping post-fire soil burn severity. Gen. Tech. Rep. RMRS-GTR-243. Fort Collins,
631 CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station.
632 49 p.
633 Pausas, J.G., Ouadah, N., Ferran, A., Gimeno, T., Vallejo, R. 2003. Fire severity and seedling
634 establishment in Pinus halepensis woodlands, eastern Iberian Peninsula. Plant Ecology
635 169, 205–213. DOI: 10.1023/A:1026019528443
636 Perez, B., Moreno, J.M. 1998. Methods for quantifying fire severity in shrubland-fires. Plant
637 Ecology 139, 91–101. DOI: 10.1023/A:1009702520958
638 Raison, R.J. 1979. Modifications of the soil environment by vegetation fires, with particular
639 reference to nitrogen transformations: A review. Plant and Soil 51, 73–108. DOI:
640 10.1007/BF02205929
641 Santín, C., Doerr, S.H., Merino, A., Bryant, R., Loader, N.J. 2016. Forest floor chemical
642 transformations in a boreal forest fire and their correlations with temperature and heating
643 duration. Geoderma 264, 71–80. DOI: 10.1016/j.geoderma.2015.09.021

644 Shepherd, K.D., Walsh, M.G. 2002. Development of reflectance spectral libraries for
645 characterization of soil properties. Soil Science Society of America Journal 66, 988–998.
646 DOI:10.2136/sssaj2002.9880
647 Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J. 2010. Visible and near
648 infrared spectroscopy in soil science. Advances in Agronomy 107, 163–215. DOI:
649 10.1016/S0065-2113(10)07005-7
650 Terefe T., Mariscal I., Gómez M., Espejo R. 2005. Relationship between soil color and
651 temperature in the surface horizon of Mediterranean soils. A laboratory study. Soil
652 Science 170, 495–503. DOI: 10.1097/01.ss.0000175341.22540.93
653 Torrent, J., Barrón, V. 2002. Diffuse reflectance spectroscopy of iron oxides. Encyclopedia of
654 Surface and Colloid Science. Ed. Marcel Dekker. Pp. 1438-1446.
655 Ulery, A.L., Graham, R.C. 1993. Forest-fire effects on soil color and texture. Soil Science
656 Society of America Journal 57, 135–140.
657 Vega, J.A., Fontúrbel, T., Merino, A., Fernández, C., Ferreiro, A., Jiménez, E. 2013. Testing
658 the ability of visual indicators of soil burn severity to reflect changes in soil chemical and
659 microbial properties in pine forest and shrubland. Plant and Soil 369, 73–91. DOI:
660 10.1007/s11104-012-1532-9
661 Verdes, P.V., Salgado, J. 2011. Changes induced in the thermal properties of Galizian soils by
662 the heating in laboratory conditions: Estimation of the soil temperature during a wildfire.
663 Journal of Thermal Analysis and Calorimetry 104 (1), 177-186. DOI: 10.1007/s10973-
664 010-1173-2
665 Vieira, D.C.S., Fernández, C., Vega, J.A., Keizer, J.J. 2015. Does soil burn severity affect the
666 post-fire runoff and interrill erosion response? A review based on meta-analysis of field
667 rainfall simulation data. Journal of Hydrology 523, 452–464. DOI:
668 10.1016/j.jhydrol.2015.01.071

669 Viscarra Rossel, R.A., Behrens, T. 2010. Using data mining to model and interpret soil
670 diffuse reflectance spectra. Geoderma 158, 46–54. DOI: 10.1016/j.geoderma.2009.12.025
671 Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J.A.M., Shepherd,
672 K.D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aïchi, H., Barthès, B.G.,
673 Bartholomeus, H.M., Bayer, A.D., Bernoux, M., Böttcher, K., Brodský, L., Du, C.W.,
674 Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A., Guerrero, C.,
675 Hedley, C.B., Knadel, M., Morrás, H.J.M., Nocita, M., Ramirez-Lopez, L., Roudier, P.,
676 Campos, E.M.R., Sanborn, P., Sellitto, V.M., Sudduth, K.A., Rawlins, B.G., Walter, C.,
677 Winowiecki, L.A., Hong, S.Y., Ji, W. 2016. A global spectral library to characterize the
678 world's soil. Earth-Science Reviews 155, 198–230. DOI: 10.1016/j.earscirev.2016.01.012
679

Figures captions

Figure 1. Four spectral regions (a-d) displaying in detail the NIR spectra (absorbance after first derivative) of a soil sample, before heating (black colour, thin line) and after exposition to 700°C for 20 minutes (red colour, thick line).

Figure 2. Measured *vs* predicted TR values obtained during the cross-validation (leave-one-out) of four randomly selected models: a) model #5 (type 24s–v1) from Maigmó site; b) model #61 (type 72s–v4) from Aitana site; c) model #27 (type 96s–v4) from Pinoso site; d) model #57 (type 48s–v4) from Aitana site. See more details in Table 2, and supplementary content (Tables S1, S2 and S3). The black line denotes the 1:1, and the grey dashed line is the linear regression.

Figure 3. Representative predictions obtained with three models: a) predictions obtained in A5 with model #1, which was constructed with 24 heated aliquots from A1 and is therefore labelled 24s–v1; b) predictions obtained in A5 with model #35, which was constructed with 24 heated aliquots from A2 and A3 and is therefore labelled 24s–v2; c) predictions obtained in A2 with the model #62, which was constructed with 72 heated aliquots from A1, A3, A4 and A5 and is therefore labelled 72–v4. The black line denotes the 1:1, and the grey dashed line is the linear regression.

Figure 4. Values of the root mean square error of prediction (RMSEP) obtained with different model types in Aitana site (a), Maigmo site (b) and Pinoso site (c). The grey bars denote the mean value, whereas dots and triangles are used for minimum and maximum values respectively. Please note that the scale of the y-axis varies for each site.

Figure 5. Prediction performance parameters obtained with the different model types (results of the three studies were averaged): a) RMSEP, SEP and bias; b) R^2 and RPD. The first number in the model type label (24, 48, 72 or 96) denotes the number of heated aliquots (=number of spectra) used to calibrate the model (i.e., model size). The number after –v denotes the variability, which is the number of micro-plot scale soil (MPS) sample(s) used as source(s) of aliquots.

Figure 6. Measured *vs* predicted TR values: a) original data; b) 100ºC was added to original predicted values; c) 50ºC was subtracted to original predicted values; d) original predicted values divided by two (×0.5); e) random selection of 24 cases shown in panels a to d; f) random selection of 24 cases shown in panels a to d; g) predictions obtained with models 24s–v1c (three sites pooled); h) predictions obtained with models 96s–v4 (three sites pooled).

Figure 7. Panel 7a contains the schematic representation of four NIR spectra (stacked), where black and grey circles denote presence and absence, respectively, of the heat-sensitive compounds at four wavelengths. The black arrows in panel 7b represent wavelengths selected in seven stacked models (relevant *b*–coefficients), depending on the composition of heat-sensitive compounds present in the MPS used as source(s) of aliquots. The sign of *b*–coefficients could be negative or positive, being denoted as down or up arrow, respectively. The panels 7c to 7f represents predictions obtained with four models (Models 1 to 4) applied to MPS1, MPS2 and MPS3, and the fail and pass symbols denote biased and unbiased predictions, respectively.

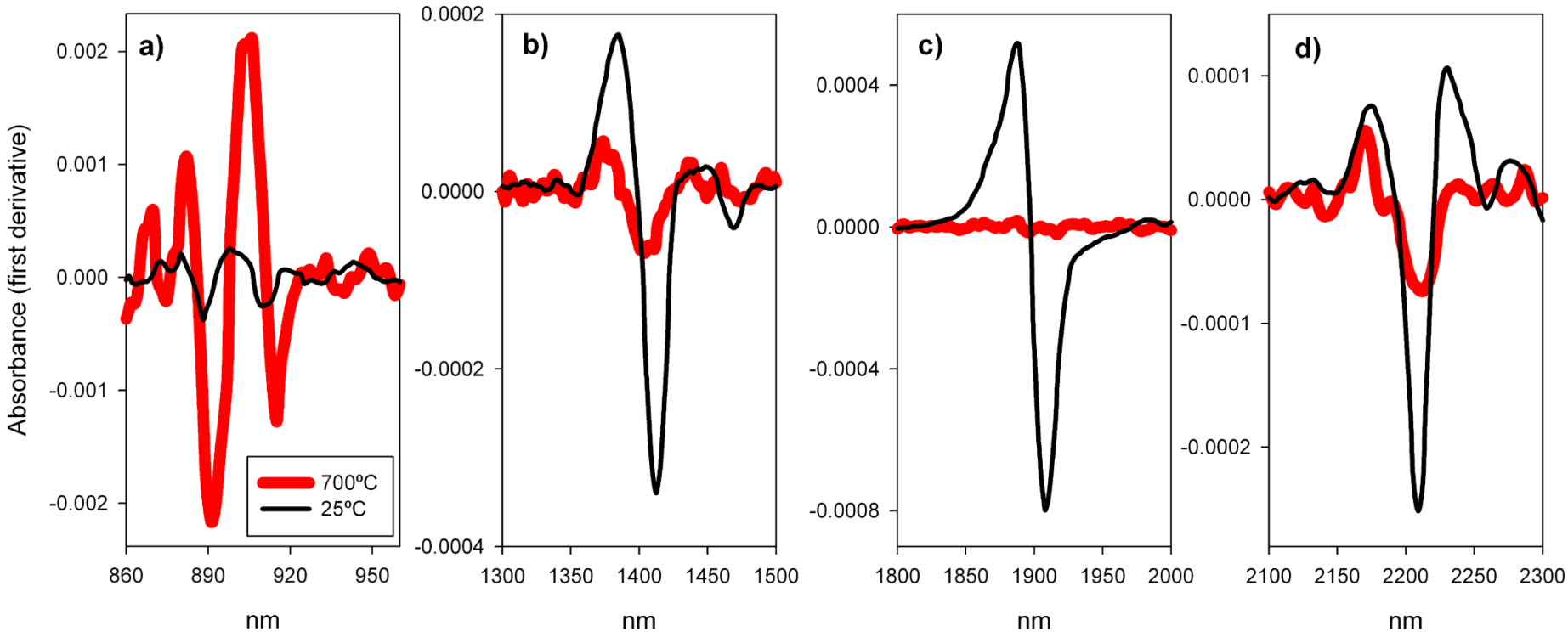
a)
b)
c)
d)
Absorbance (first derivative)
0.002
0.001
0.000
-0.001
-0.002
860
890
920
950
nm
700ºC
25ºC
0.0002
0.0000
-0.0002
-0.0004
1300
1350
1400
1450
1500
nm
0.0004
0.0000
-0.0004
-0.0008
1800
1850
1900
1950
2000
nm
0.0001
0.0000
-0.0001
-0.0002
2100
2150
2200
2250
2300
nm

Figure 1.

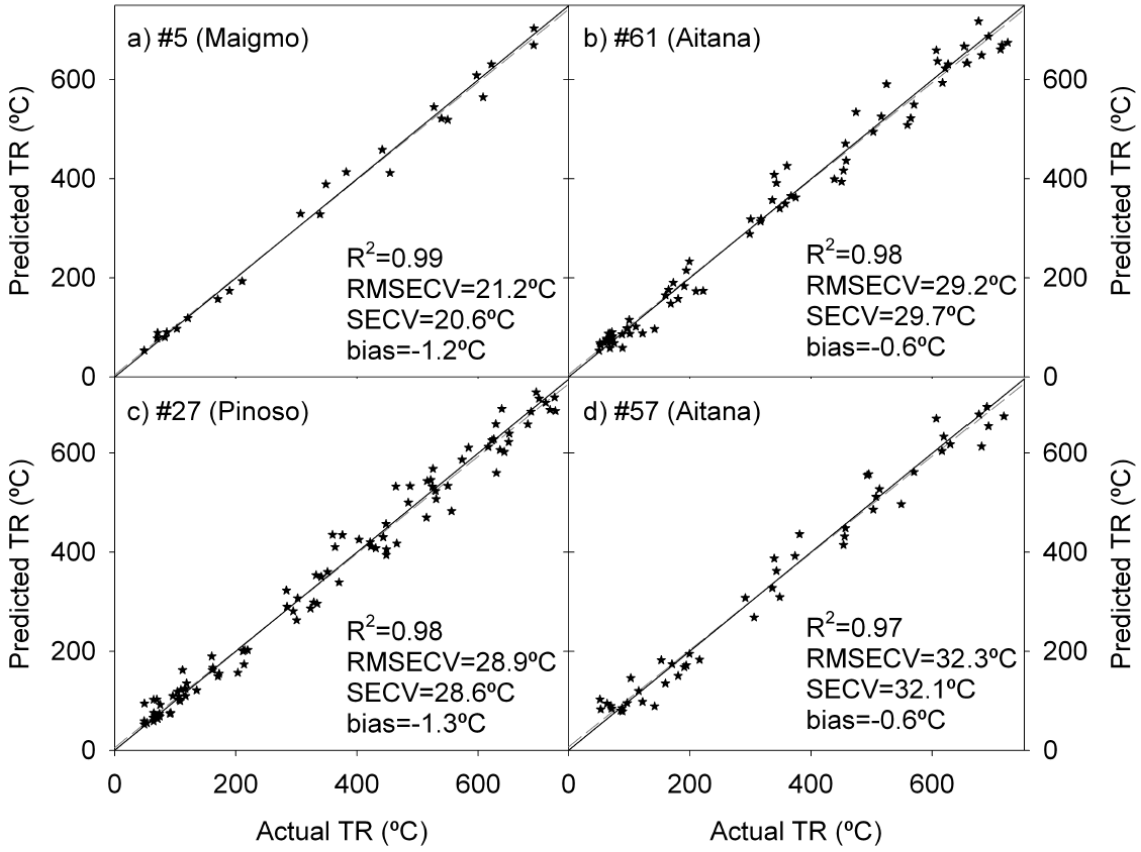
a) #5 (Maigmo)
R2=0.99
RMSECV=21.2ºC
SECV=20.6ºC
bias=-1.2ºC
b) #61 (Aitana)
R2=0.98
RMSECV=29.2ºC
SECV=29.7ºC
bias=-0.6ºC
c) #27 (Pinoso)
R2=0.98
RMSECV=28.9ºC
SECV=28.6ºC
bias=-1.3ºC
d) #57 (Aitana)
R2=0.97
RMSECV=32.3ºC
SECV=32.1ºC
bias=-0.6ºC
Predicted TR (ºC)
Actual TR (ºC)
0
200
400
600

Figure 2.

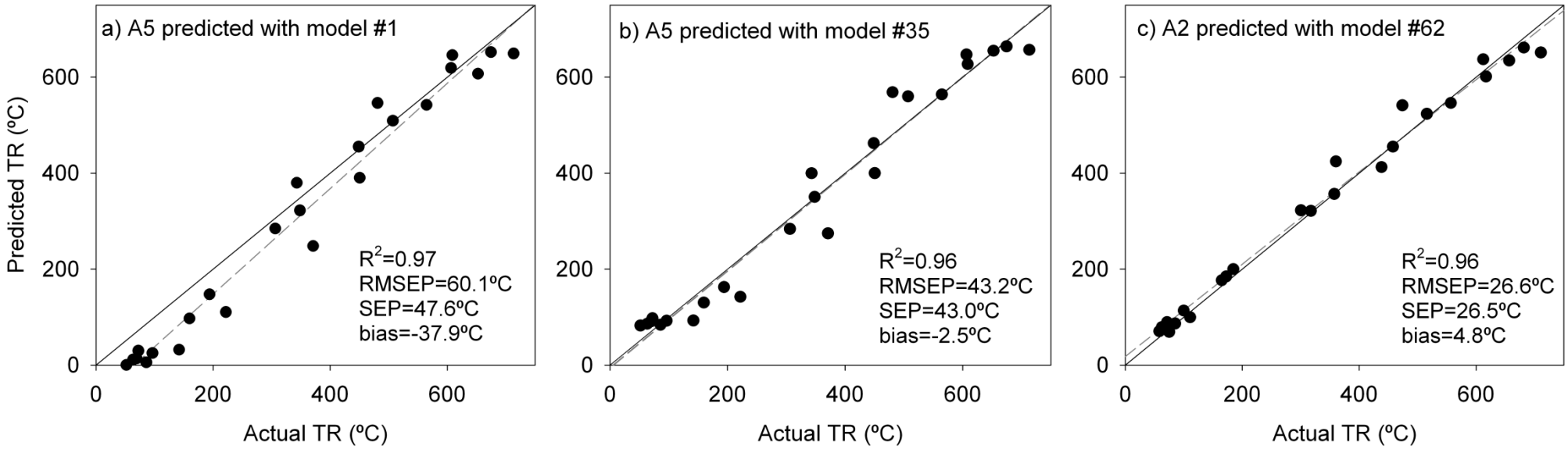
a) A5 predicted with model #1
R^2=0.97
RMSEP=60.1ºC
SEP=47.6ºC
bias=-37.9ºC
b) A5 predicted with model #35
R^2=0.96
RMSEP=43.2ºC
SEP=43.0ºC
bias=-2.5ºC
c) A2 predicted with model #62
R^2=0.96
RMSEP=26.6ºC
SEP=26.5ºC
bias=4.8ºC
Predicted TR (ºC)
Actual TR (ºC)
0
200
400
600

Figure 3

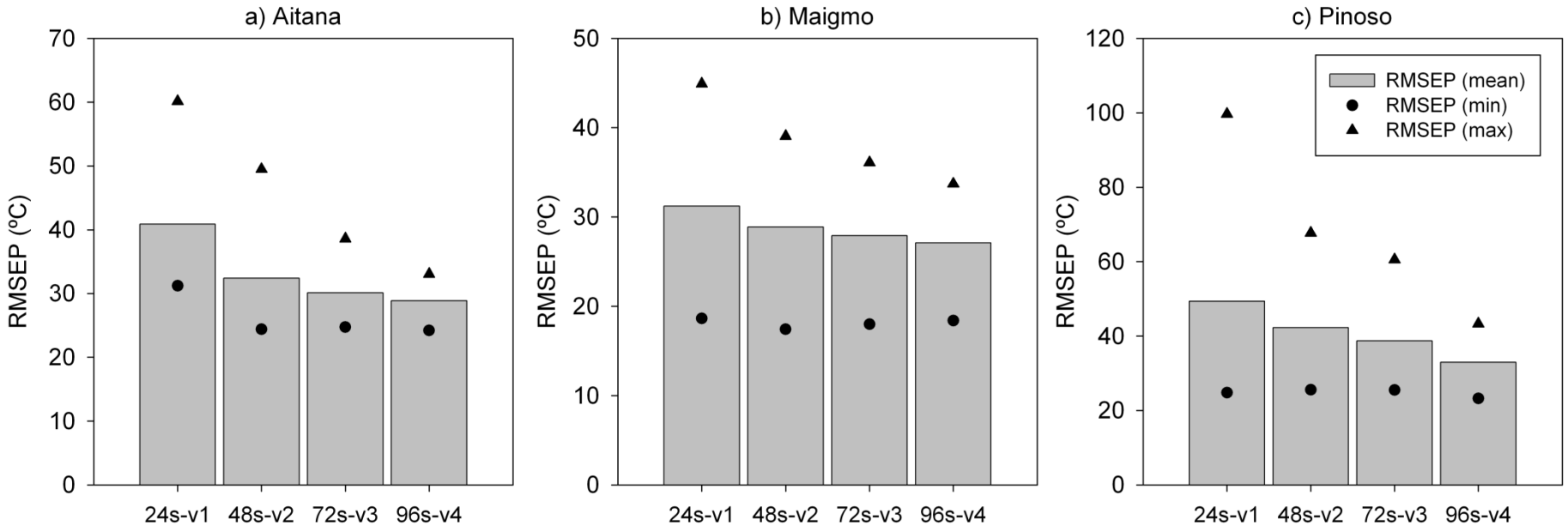
a) Aitana
b) Maigmo
c) Pinoso
RMSEP (ºC)
RMSEP (mean)
RMSEP (min)
RMSEP (max)
24s-v1
48s-v2
72s-v3
96s-v4

Figure 4

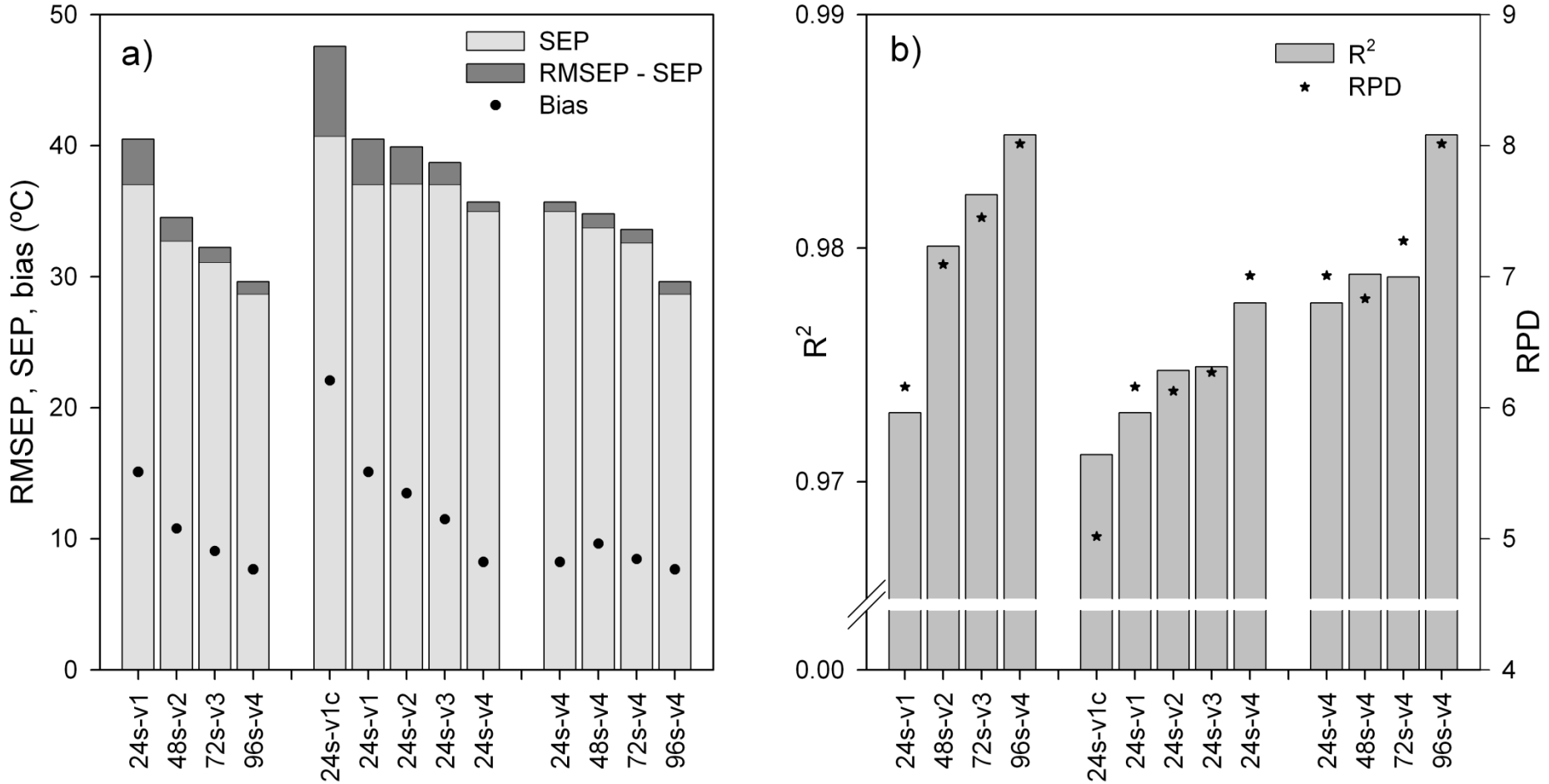
a)
SEP
RMSEP - SEP
Bias
RMSEP, SEP, bias (ºC)
50
40
30
20
10
0
24s-v1
48s-v2
72s-v3
96s-v4
24s-v1c
24s-v1
24s-v2
24s-v3
24s-v4
24s-v4
48s-v4
72s-v4
96s-v4
b)
R2
RPD
0.99
0.98
0.97
0.00
9
8
7
6
5
4
24s-v1
48s-v2
72s-v3
96s-v4
24s-v1c
24s-v1
24s-v2
24s-v3
24s-v4
24s-v4
48s-v4
72s-v4
96s-v4

Figure 5

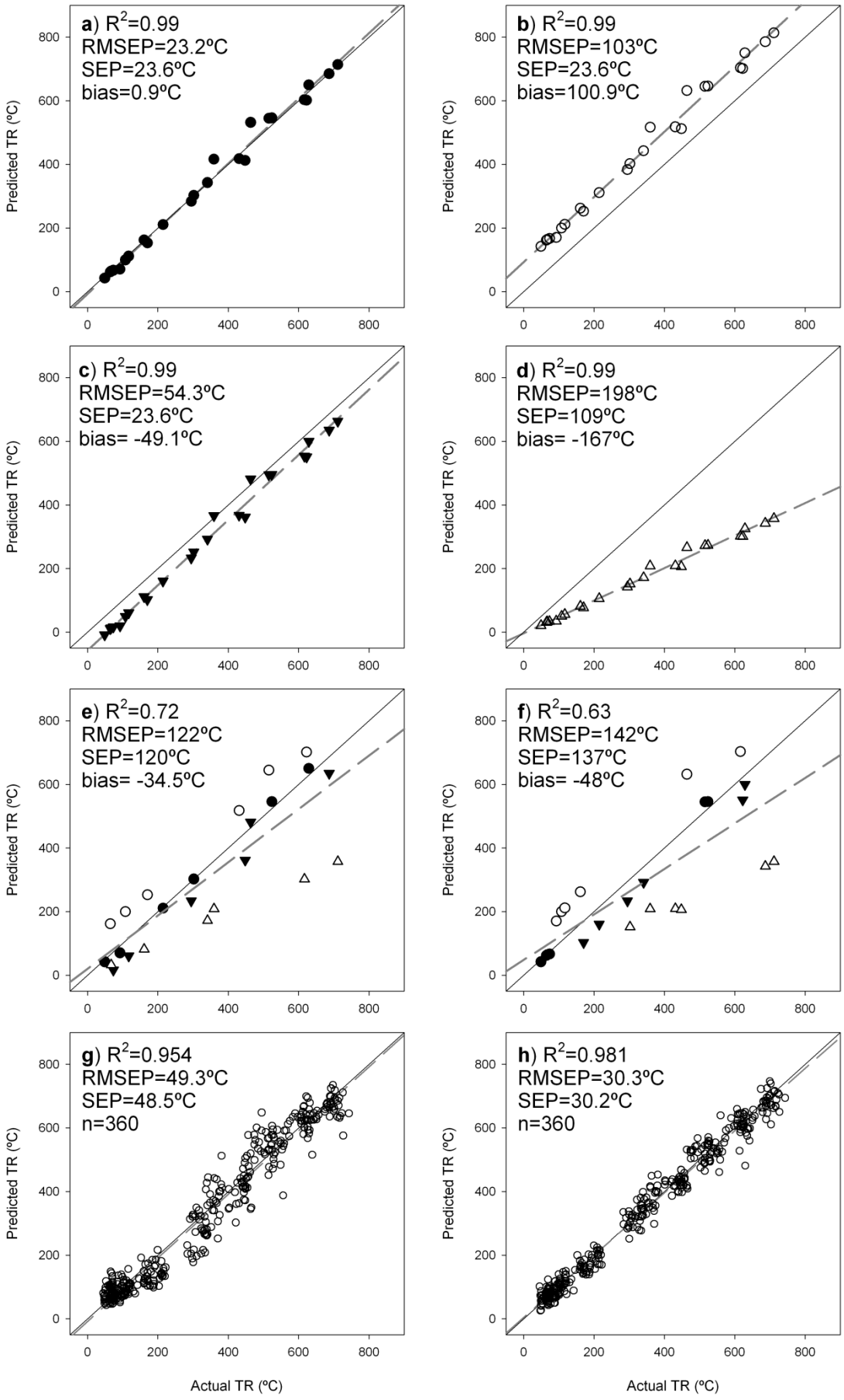
a) R2=0.99
RMSEP=23.2ºC
SEP=23.6ºC
bias=0.9ºC
b) R2=0.99
RMSEP=103ºC
SEP=23.6ºC
bias=100.9ºC
c) R2=0.99
RMSEP=54.3ºC
SEP=23.6ºC
bias= -49.1ºC
d) R2=0.99
RMSEP=198ºC
SEP=109ºC
bias= -167ºC
e) R2=0.72
RMSEP=122ºC
SEP=120ºC
bias= -34.5ºC
f) R2=0.63
RMSEP=142ºC
SEP=137ºC
bias= -48ºC
g) R2=0.954
RMSEP=49.3ºC
SEP=48.5ºC
n=360
h) R2=0.981
RMSEP=30.3ºC
SEP=30.2ºC
n=360
Predicted TR (ºC)
Actual TR (ºC)
0
200
400
600
800

Figure 6

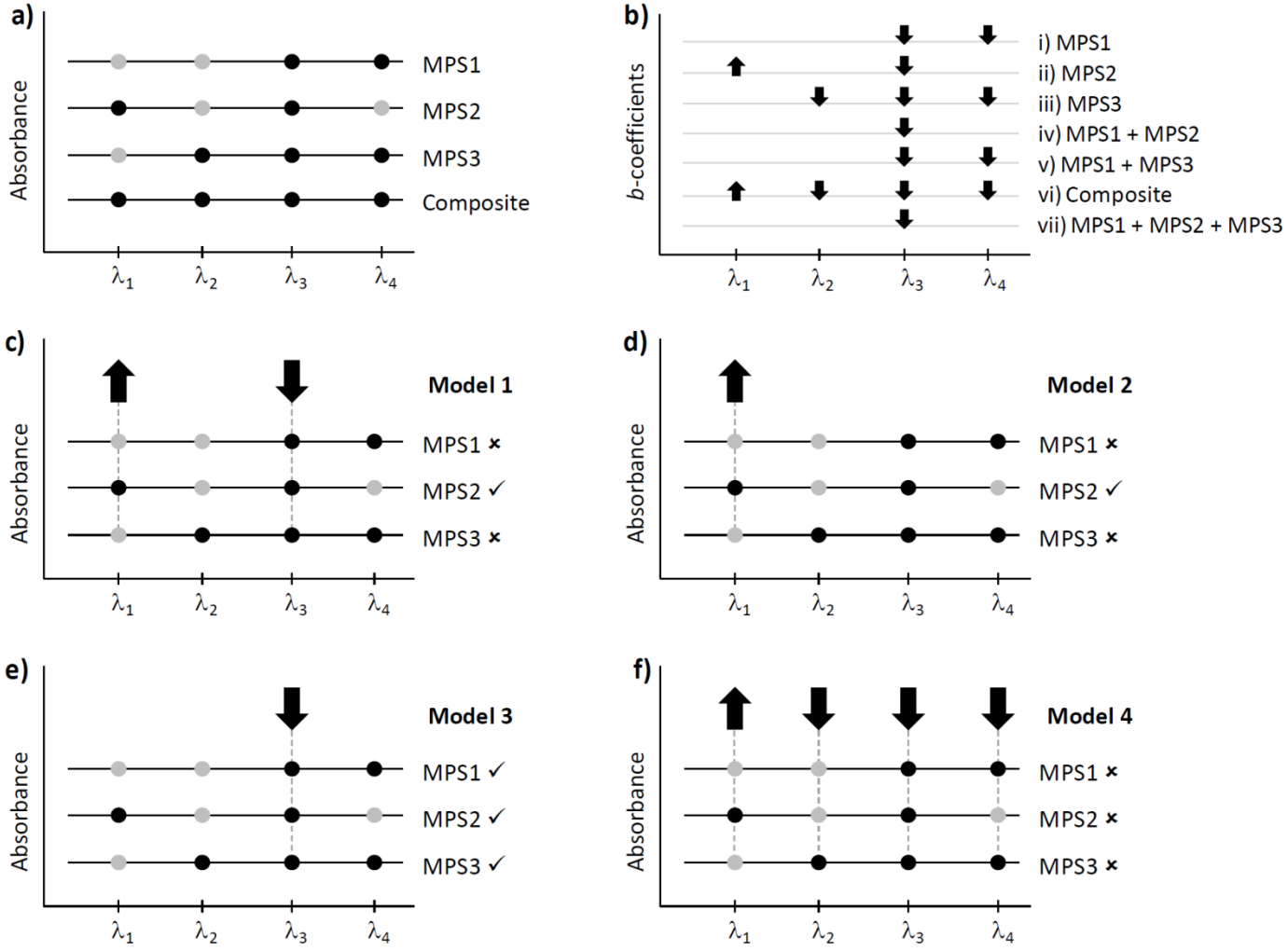
a)
Absorbance
MPS1
MPS2
MPS3
Composite
λ1 λ2 λ3 λ4
b)
b-coefficients
i) MPS1
ii) MPS2
iii) MPS3
iv) MPS1 + MPS2
v) MPS1 + MPS3
vi) Composite
vii) MPS1 + MPS2 + MPS3
λ1 λ2 λ3 λ4
c)
Absorbance
Model 1
MPS1 ×
MPS2 ✓
MPS3 ×
λ1 λ2 λ3 λ4
d)
Absorbance
Model 2
MPS1 ×
MPS2 ✓
MPS3 ×
λ1 λ2 λ3 λ4
e)
Absorbance
Model 3
MPS1 ✓
MPS2 ✓
MPS3 ✓
λ1 λ2 λ3 λ4
f)
Absorbance
Model 4
MPS1 ×
MPS2 ×
MPS3 ×
λ1 λ2 λ3 λ4

Figure 7

Table 1. Main characteristics of sites and soils characteristics (0-5cm depth)

Site [a]	UTM Coordinates [b]	T[c] (°C)	P[d] (mm)	Sand, silt, clay (%)	SOC ± SD[e] (%)	pH	$CaCO_3$ (%)
A	30SYH3080	13.8	706	33, 32, 35	7.31 ± 3.11	7.5	46.9
M	30SYH0664	18.2	302	57, 22, 21	3.60 ± 1.23	7.9	57.6
P	30SXH7749	15.8	277	31, 56, 13	4.47 ± 1.64	8.0	7.0

[a] A: Aitana; M: Maigmó; P: Pinoso

[b] UTM coordinates – WGS84 (at 1 km^2 resolution).

[c] T: Mean annual temperature

[d] P Mean annual precipitation.

[e] SD: standard deviation of the five MPS

Table 2. Description of models types in each site.

	Model type label	Size [a]	Variability [b]	Number of models	Supplementary tables [c]	Times used to predict [c]
Group *i*	24s–v1	24	1	5	#1 – #5	20
	48s–v2	48	2	10	#6 – #15	30
	72s–v3	72	3	10	#16 – #25	20
	96s–v4	96	4	5	#26 – #30	5
Group *ii*	24s–v1	24	1	5	#1 – #5	20
	24s–v2	24 *	2	10	#31 – #40	30
	24s–v3	24 *	3	10	#41 – #50	20
	24s–v4	24 *	4	5	#51 – #55	5
Group *iii*	24s–v4	24 *	4	5	#51 – #55	5
	48s–v4	48 *	4	5	#56 – #60	5
	72s–v4	72 *	4	5	#61 – #65	5
	96s–v4	96	4	5	#26 – #30	5
Group *iv*	24s–v1c	24	1	1	#66	5
Total				66 [d]		145 [d]

[a] Number of heated aliquots (number of spectra; number of standards) used to calibrate the model

[b] Number of different MPS used as sources of heated aliquots

[c] A full description of each model is shown in supplementary tables S1, S2 and S3 for Aitana, Maigmó and Pinoso sites respectively.

[d] Total number (repeated cases excluded)

[*] The Kennard-Stone algorithm was used to select the desired number of spectra (model size).

Appendices

Appendix 1: Supplementary content (Tables S1, S2 and S3):

General comments about how interpret the tables S1, S2 and S3:

1) Regarding the construction of models:

This is a brief comment about how to read tables to understand the construction of models. The left part of tables (in blue colour) identifies each model with a number (#) and also contains information about the MPS used as source of aliquots. The model type (size, variability) is also shown. Examples for three models types are described, as follows:

- Model #1:

The model #1 in Aitana site was constructed with 24 standards (24 spectra) (see table S1). All the heated aliquots used were obtained from one MPS, namely A1.
The model #1 in Maigmó site was constructed with 24 standards (24 spectra) (see table S2). All the heated aliquots used were obtained from one MPS, namely M1.
The model #1 in Pinoso site was constructed with 24 standards (24 spectra) (see table S3). All the heated aliquots used were obtained from one MPS, namely P1.

- Model #26:

The model #26 in Aitana site was constructed with 96 standards (96 spectra) (see table S1). The standards (heated aliquots) were obtained from four MPS, namely A2, A3, A4 and A5.
The model #26 in Maigmó site was constructed with 96 standards (96 spectra) (see table S2). The standards (heated aliquots) were obtained from four MPS, namely M2, M3, M4 and M5.
The model #26 in Pinoso site was constructed with 96 standards (96 spectra) (see table S3). The standards (heated aliquots) were obtained from four MPS, namely P2, P3, P4 and P5.

- Model #63:

The model #63 in Aitana site was constructed with 72 standards (72 spectra) (see table S1). These 72 standards (heated aliquots) were selected from a pool with 96 (those available from four MPS, namely A1, A2, A4 and A5). The Kennard-Stone algorithm was used to select the 72 spectra.
The model #63 in Maigmó site was constructed with 72 standards (72 spectra) (see table S2). These 72 standards (heated aliquots) were selected from a pool with 96 (those available from four MPS, namely M1, M2, M4 and M5). The Kennard-Stone algorithm was used to select the 72 spectra.
The model #63 in Pinoso site was constructed with 72 standards (72 spectra) (see table S3). These 72 standards (heated aliquots) were selected from a pool with 96 (those available from four MPS, namely P1, P2, P4 and P5). The Kennard-Stone algorithm was used to select the 72 spectra.

2) Regarding the use of models (predictions):

The right part of tables S1, S2 and S3 shows what prediction sets were used for each model (in red colour).

A total of 66 models were constructed in each site. However, in each site we have obtained 145 values of R^2, RMSEP, bias, SEP and RPD. The reason is that some models were used to predict more than one time.
For instance, the model #26 in Aitana site was used to predict in the A1 as prediction set (indicated with a cross in table S1), and therefore one value of R^2, RMSEP, bias, SEP and RPD was computed using the 24 heated aliquots from A1. Thus, this model was used only one time.
However, other models, such as model #1, were used more than one time. The model #1 in Aitana site was used to predict in the 24 heated aliquots of A2. But the model #1 was also used to predict in A3, A4 and A5. So, the model #1 was used four times, and consequently the model #1 produced four values of R^2, RMSEP, bias, SEP and RPD (different values for each case: A2, A3, A4 and A5).
In overall, the 66 models produced 145 values of R^2, RMSEP, bias, SEP and RPD (in each site).
Those values of R^2, RMSEP, bias, SEP and RPD obtained with the same model type were considered replicates, therefore, they were averaged.

Table S1. Models and predictions in Aitana site

#	MPS used as source of aliquots	Type (label)	Size	Variability	MPS used as prediction set				
					A1	A2	A3	A4	A5
#1	A1	24s-v1	24	1		×	×	×	×
#2	A2	24s-v1	24	1	×		×	×	×
#3	A3	24s-v1	24	1	×	×		×	×
#4	A4	24s-v1	24	1	×	×	×		×
#5	A5	24s-v1	24	1	×	×	×	×	
#6	A1+A2	48s-v2	48	2			×	×	×
#7	A1+A3	48s-v2	48	2		×		×	×
#8	A1+A4	48s-v2	48	2		×	×		×
#9	A1+A5	48s-v2	48	2		×	×	×	
#10	A2+A3	48s-v2	48	2	×			×	×
#11	A2+A4	48s-v2	48	2	×		×		×
#12	A2+A5	48s-v2	48	2	×		×	×	
#13	A3+A4	48s-v2	48	2	×	×			×
#14	A3+A5	48s-v2	48	2	×	×		×	
#15	A4+A5	48s-v2	48	2	×	×	×		
#16	A1+A2+A3	72s-v3	72	3				×	×
#17	A1+A2+A4	72s-v3	72	3			×		×
#18	A1+A2+A5	72s-v3	72	3			×	×	
#19	A1+A3+A4	72s-v3	72	3		×			×
#20	A1+A3+A5	72s-v3	72	3		×		×	
#21	A1+A4+A5	72s-v3	72	3		×	×		
#22	A2+A3+A4	72s-v3	72	3	×				×
#23	A2+A3+A5	72s-v3	72	3	×			×	
#24	A2+A4+A5	72s-v3	72	3	×		×		
#25	A3+A4+A5	72s-v3	72	3	×	×			
#26	A2+A3+A4+A5	96s-v4	96	4	×				
#27	A1+A3+A4+A5	96s-v4	96	4		×			
#28	A1+A2+A4+A5	96s-v4	96	4			×		
#29	A1+A2+A3+A5	96s-v4	96	4				×	
#30	A1+A2+A3+A4	96s-v4	96	4					×
#31	A1+A2	24s-v2	24	2			×	×	×
#32	A1+A3	24s-v2	24	2		×		×	×
#33	A1+A4	24s-v2	24	2		×	×		×
#34	A1+A5	24s-v2	24	2		×	×	×	
#35	A2+A3	24s-v2	24	2	×			×	×
#36	A2+A4	24s-v2	24	2	×		×		×
#37	A2+A5	24s-v2	24	2	×		×	×	
#38	A3+A4	24s-v2	24	2	×	×			×
#39	A3+A5	24s-v2	24	2	×	×		×	
#40	A4+A5	24s-v2	24	2	×	×	×		
#41	A1+A2+A3	24s-v3	24	3				×	×
#42	A1+A2+A4	24s-v3	24	3			×		×
#43	A1+A2+A5	24s-v3	24	3			×	×	
#44	A1+A3+A4	24s-v3	24	3		×			×
#45	A1+A3+A5	24s-v3	24	3		×		×	
#46	A1+A4+A5	24s-v3	24	3		×	×		
#47	A2+A3+A4	24s-v3	24	3	×				×
#48	A2+A3+A5	24s-v3	24	3	×			×	
#49	A2+A4+A5	24s-v3	24	3	×		×		
#50	A3+A4+A5	24s-v3	24	3	×	×			
#51	A2+A3+A4+A5	24s-v4	24	4	×				
#52	A1+A3+A4+A5	24s-v4	24	4		×			
#53	A1+A2+A4+A5	24s-v4	24	4			×		
#54	A1+A2+A3+A5	24s-v4	24	4				×	
#55	A1+A2+A3+A4	24s-v4	24	4					×
#56	A2+A3+A4+A5	48s-v4	48	4	×				
#57	A1+A3+A4+A5	48s-v4	48	4		×			
#58	A1+A2+A4+A5	48s-v4	48	4			×		
#59	A1+A2+A3+A5	48s-v4	48	4				×	
#60	A1+A2+A3+A4	48s-v4	48	4					×
#61	A2+A3+A4+A5	72s-v4	72	4	×				
#62	A1+A3+A4+A5	72s-v4	72	4		×			
#63	A1+A2+A4+A5	72s-v4	72	4			×		
#64	A1+A2+A3+A5	72s-v4	72	4				×	
#65	A1+A2+A3+A4	72s-v4	72	4					×
#66	Ac (composite)	24s-v1c	24	1	×	×	×	×	×

Table S2. Models and predictions in Maigmó site

#	MPS used as source of aliquots	Type (label)	Size	Variability	MPS used as prediction set				
					M1	M2	M3	M4	M5
#1	M1	24s-v1	24	1		×	×	×	×
#2	M2	24s-v1	24	1	×		×	×	×
#3	M3	24s-v1	24	1	×	×		×	×
#4	M4	24s-v1	24	1	×	×	×		×
#5	M5	24s-v1	24	1	×	×	×	×	
#6	M1+M2	48s-v2	48	2			×	×	×
#7	M1+M3	48s-v2	48	2		×		×	×
#8	M1+M4	48s-v2	48	2		×	×		×
#9	M1+M5	48s-v2	48	2		×	×	×	
#10	M2+M3	48s-v2	48	2	×			×	×
#11	M2+M4	48s-v2	48	2	×		×		×
#12	M2+M5	48s-v2	48	2	×		×	×	
#13	M3+M4	48s-v2	48	2	×	×			×
#14	M3+M5	48s-v2	48	2	×	×		×	
#15	M4+M5	48s-v2	48	2	×	×	×		
#16	M1+M2+M3	72s-v3	72	3				×	×
#17	M1+M2+M4	72s-v3	72	3			×		×
#18	M1+M2+M5	72s-v3	72	3			×	×	
#19	M1+M3+M4	72s-v3	72	3		×			×
#20	M1+M3+M5	72s-v3	72	3		×		×	
#21	M1+M4+M5	72s-v3	72	3		×	×		
#22	M2+M3+M4	72s-v3	72	3	×				×
#23	M2+M3+M5	72s-v3	72	3	×			×	
#24	M2+M4+M5	72s-v3	72	3	×		×		
#25	M3+M4+M5	72s-v3	72	3	×	×			
#26	M2+M3+M4+M5	96s-v4	96	4	×				
#27	M1+M3+M4+M5	96s-v4	96	4		×			
#28	M1+M2+M4+M5	96s-v4	96	4			×		
#29	M1+M2+M3+M5	96s-v4	96	4				×	
#30	M1+M2+M3+M4	96s-v4	96	4					×
#31	M1+M2	24s-v2	24	2			×	×	×
#32	M1+M3	24s-v2	24	2		×		×	×
#33	M1+M4	24s-v2	24	2		×	×		×
#34	M1+M5	24s-v2	24	2		×	×	×	
#35	M2+M3	24s-v2	24	2	×			×	×
#36	M2+M4	24s-v2	24	2	×		×		×
#37	M2+M5	24s-v2	24	2	×		×	×	
#38	M3+M4	24s-v2	24	2	×	×			×
#39	M3+M5	24s-v2	24	2	×	×		×	
#40	M4+M5	24s-v2	24	2	×	×	×		
#41	M1+M2+M3	24s-v3	24	3				×	×
#42	M1+M2+M4	24s-v3	24	3			×		×
#43	M1+M2+M5	24s-v3	24	3			×	×	
#44	M1+M3+M4	24s-v3	24	3		×			×
#45	M1+M3+M5	24s-v3	24	3		×		×	
#46	M1+M4+M5	24s-v3	24	3		×	×		
#47	M2+M3+M4	24s-v3	24	3	×				×
#48	M2+M3+M5	24s-v3	24	3	×			×	
#49	M2+M4+M5	24s-v3	24	3	×		×		
#50	M3+M4+M5	24s-v3	24	3	×	×			
#51	M2+M3+M4+M5	24s-v4	24	4	×				
#52	M1+M3+M4+M5	24s-v4	24	4		×			
#53	M1+M2+M4+M5	24s-v4	24	4			×		
#54	M1+M2+M3+M5	24s-v4	24	4				×	
#55	M1+M2+M3+M4	24s-v4	24	4					×
#56	M2+M3+M4+M5	48s-v4	48	4	×				
#57	M1+M3+M4+M5	48s-v4	48	4		×			
#58	M1+M2+M4+M5	48s-v4	48	4			×		
#59	M1+M2+M3+M5	48s-v4	48	4				×	
#60	M1+M2+M3+M4	48s-v4	48	4					×
#61	M2+M3+M4+M5	72s-v4	72	4	×				
#62	M1+M3+M4+M5	72s-v4	72	4		×			
#63	M1+M2+M4+M5	72s-v4	72	4			×		
#64	M1+M2+M3+M5	72s-v4	72	4				×	
#65	M1+M2+M3+M4	72s-v4	72	4					×
#66	Mc (composite)	24s-v1c	24	1	×	×	×	×	×

Table S3. Models and predictions in Pinoso site

#	MPS used as source of aliquots	Type (label)	Size	Variability	MPS used as prediction set				
					P1	P2	P3	P4	P5
#1	P1	24s-v1	24	1		×	×	×	×
#2	P2	24s-v1	24	1	×		×	×	×
#3	P3	24s-v1	24	1	×	×		×	×
#4	P4	24s-v1	24	1	×	×	×		×
#5	P5	24s-v1	24	1	×	×	×	×	
#6	P1+P2	48s-v2	48	2			×	×	×
#7	P1+P3	48s-v2	48	2		×		×	×
#8	P1+P4	48s-v2	48	2		×	×		×
#9	P1+P5	48s-v2	48	2		×	×	×	
#10	P2+P3	48s-v2	48	2	×			×	×
#11	P2+P4	48s-v2	48	2	×		×		×
#12	P2+P5	48s-v2	48	2	×		×	×	
#13	P3+P4	48s-v2	48	2	×	×			×
#14	P3+P5	48s-v2	48	2	×	×		×	
#15	P4+P5	48s-v2	48	2	×	×	×		
#16	P1+P2+P3	72s-v3	72	3				×	×
#17	P1+P2+P4	72s-v3	72	3			×		×
#18	P1+P2+P5	72s-v3	72	3			×	×	
#19	P1+P3+P4	72s-v3	72	3		×			×
#20	P1+P3+P5	72s-v3	72	3		×		×	
#21	P1+P4+P5	72s-v3	72	3		×	×		
#22	P2+P3+P4	72s-v3	72	3	×				×
#23	P2+P3+P5	72s-v3	72	3	×			×	
#24	P2+P4+P5	72s-v3	72	3	×		×		
#25	P3+P4+P5	72s-v3	72	3	×	×			
#26	P2+P3+P4+P5	96s-v4	96	4	×				
#27	P1+P3+P4+P5	96s-v4	96	4		×			
#28	P1+P2+P4+P5	96s-v4	96	4			×		
#29	P1+P2+P3+P5	96s-v4	96	4				×	
#30	P1+P2+P3+P4	96s-v4	96	4					×
#31	P1+P2	24s-v2	24	2			×	×	×
#32	P1+P3	24s-v2	24	2		×		×	×
#33	P1+P4	24s-v2	24	2		×	×		×
#34	P1+P5	24s-v2	24	2		×	×	×	
#35	P2+P3	24s-v2	24	2	×			×	×
#36	P2+P4	24s-v2	24	2	×		×		×
#37	P2+P5	24s-v2	24	2	×		×	×	
#38	P3+P4	24s-v2	24	2	×	×			×
#39	P3+P5	24s-v2	24	2	×	×		×	
#40	P4+P5	24s-v2	24	2	×	×	×		
#41	P1+P2+P3	24s-v3	24	3				×	×
#42	P1+P2+P4	24s-v3	24	3			×		×
#43	P1+P2+P5	24s-v3	24	3			×	×	
#44	P1+P3+P4	24s-v3	24	3		×			×
#45	P1+P3+P5	24s-v3	24	3		×		×	
#46	P1+P4+P5	24s-v3	24	3		×	×		
#47	P2+P3+P4	24s-v3	24	3	×				×
#48	P2+P3+P5	24s-v3	24	3	×			×	
#49	P2+P4+P5	24s-v3	24	3	×		×		
#50	P3+P4+P5	24s-v3	24	3	×	×			
#51	P2+P3+P4+P5	24s-v4	24	4	×				
#52	P1+P3+P4+P5	24s-v4	24	4		×			
#53	P1+P2+P4+P5	24s-v4	24	4			×		
#54	P1+P2+P3+P5	24s-v4	24	4				×	
#55	P1+P2+P3+P4	24s-v4	24	4					×
#56	P2+P3+P4+P5	48s-v4	48	4	×				
#57	P1+P3+P4+P5	48s-v4	48	4		×			
#58	P1+P2+P4+P5	48s-v4	48	4			×		
#59	P1+P2+P3+P5	48s-v4	48	4				×	
#60	P1+P2+P3+P4	48s-v4	48	4					×
#61	P2+P3+P4+P5	72s-v4	72	4	×				
#62	P1+P3+P4+P5	72s-v4	72	4		×			
#63	P1+P2+P4+P5	72s-v4	72	4			×		
#64	P1+P2+P3+P5	72s-v4	72	4				×	
#65	P1+P2+P3+P4	72s-v4	72	4					×
#66	Pc (composite)	24s-v1c	24	1	×	×	×	×	×

Appendix 2: Supplementary content (Figures S1 to S27):

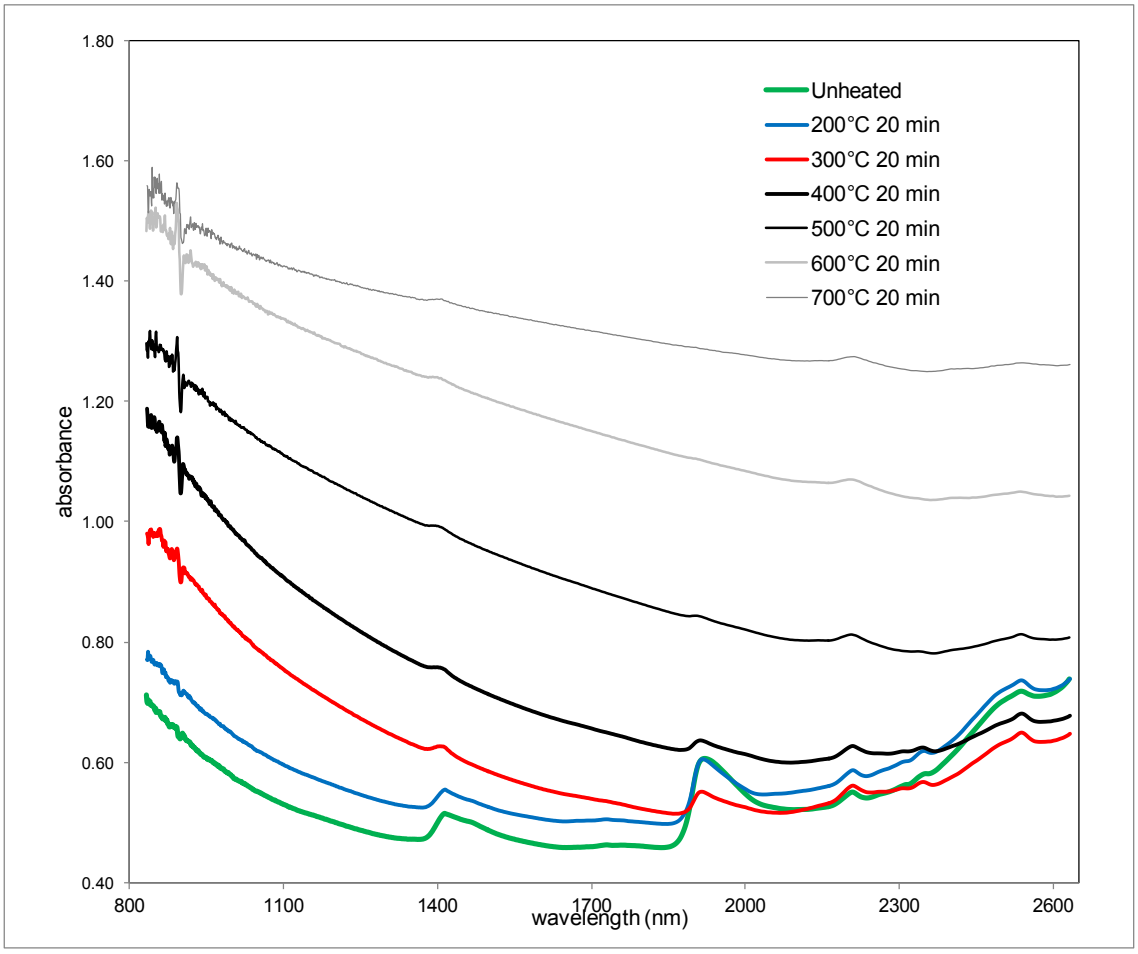

Figure S1. Representative NIR spectra of a few heated aliquots (only those heated during 20 minutes) from Aitana (A3): without preprocessing. Samples heated at 70°C and 100°C were omitted to enhance the clarity.

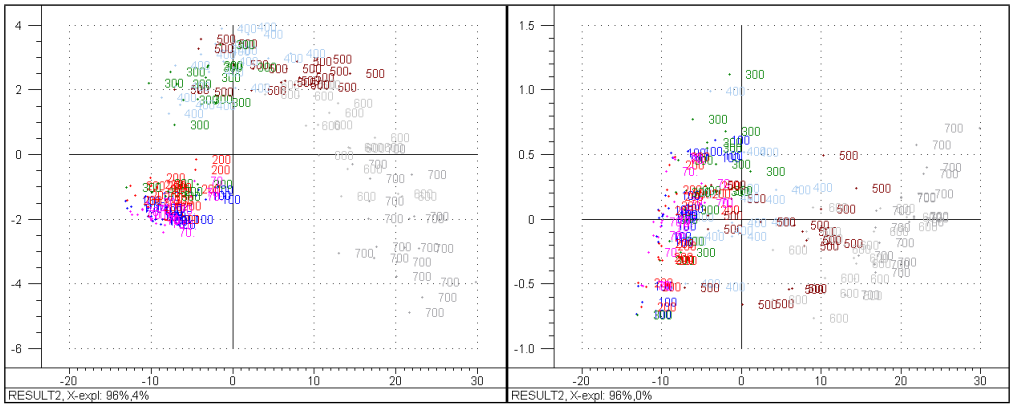

Figure S2. Principal component analysis of the NIR spectra (without preprocessing) of the 144 heated aliquots from Aitana site. Left: scores over first and second factors. Right: scores over the first and third factors.

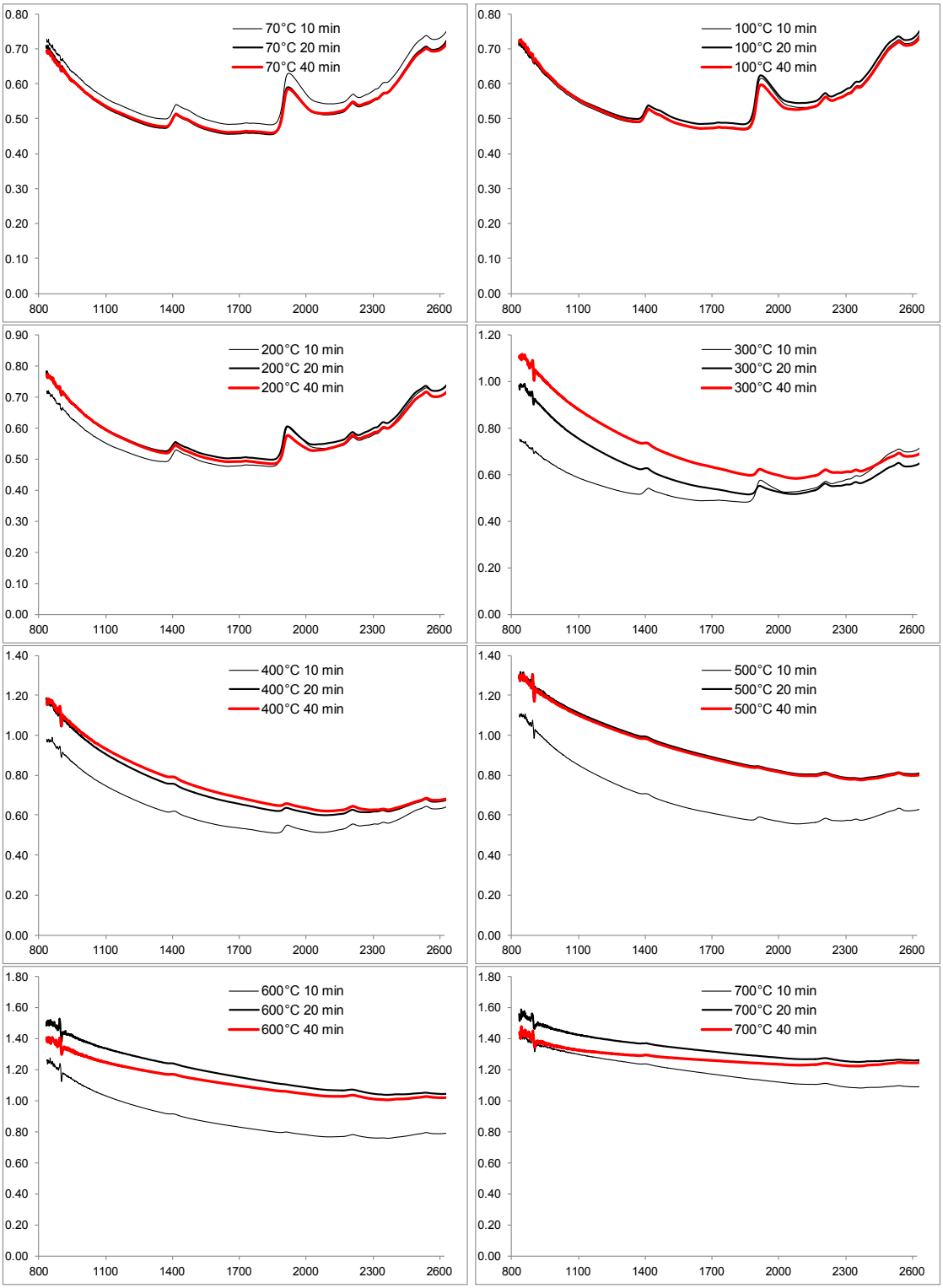

Figure S3. Representative NIR spectra of a few heated aliquots from Aitana (A3) at different exposure times. Without preprocessing.

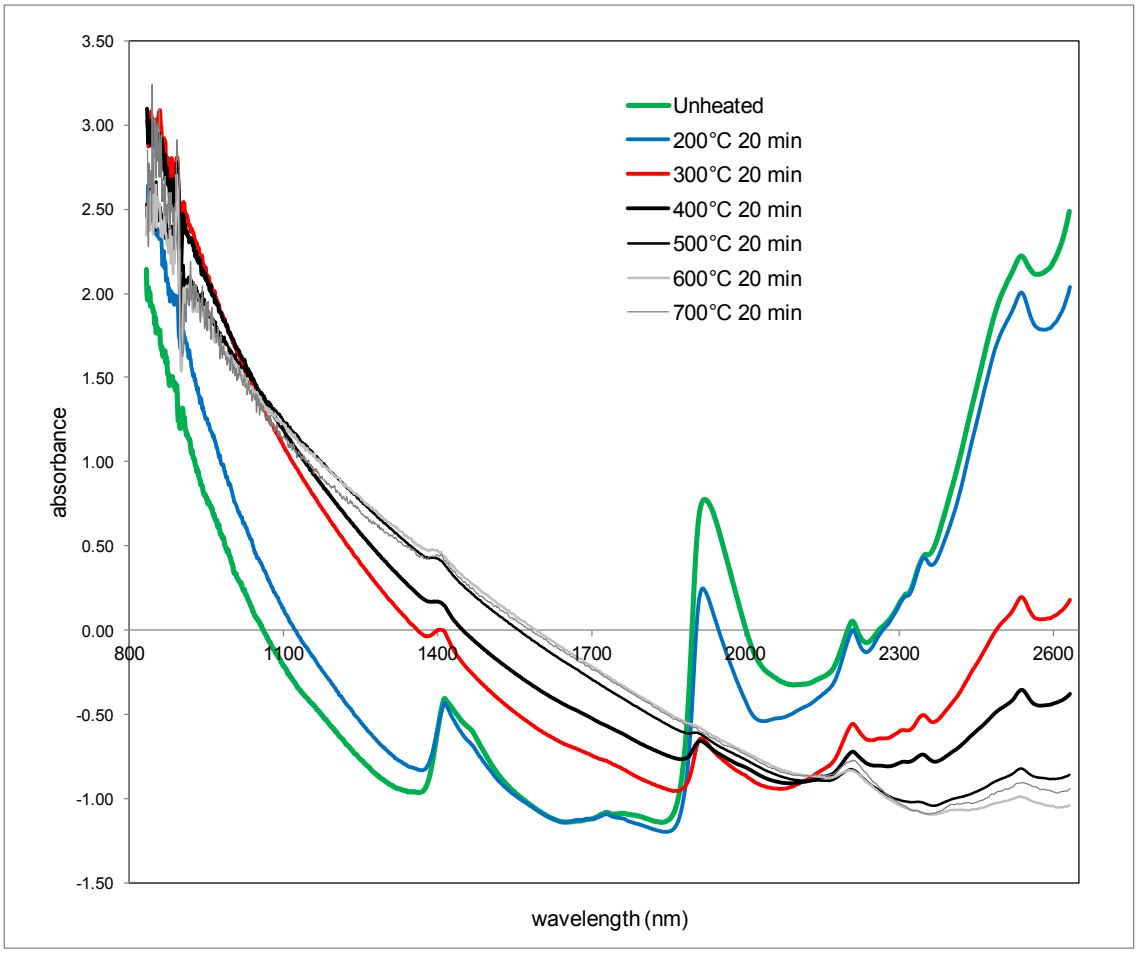

Figure S4. Representative NIR spectra of a few heated aliquots (only those heated during 20 minutes) from Aitana (A3): Standard normal variate. Samples heated at 70°C and 100°C were omitted to enhance the clarity.

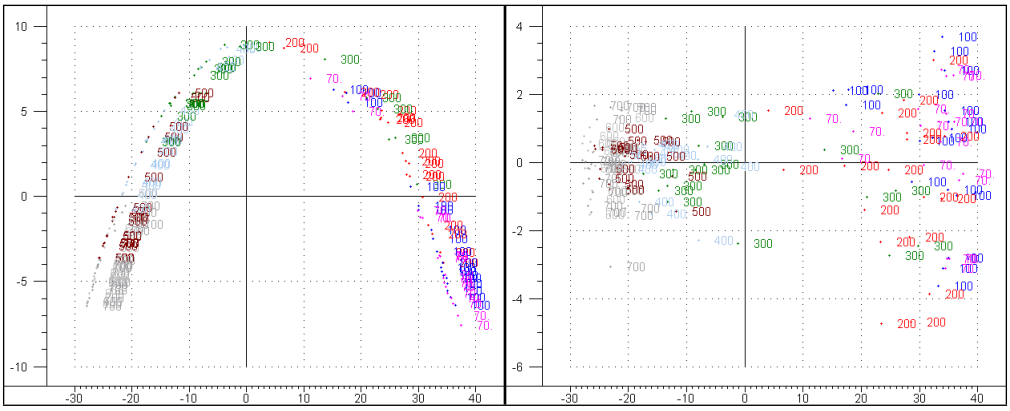

Figure S5. Principal component analysis of the NIR spectra (Standard normal variate) of the 144 heated aliquots from Aitana site. Left: scores over first and second factors. Right: scores over the first and third factors.

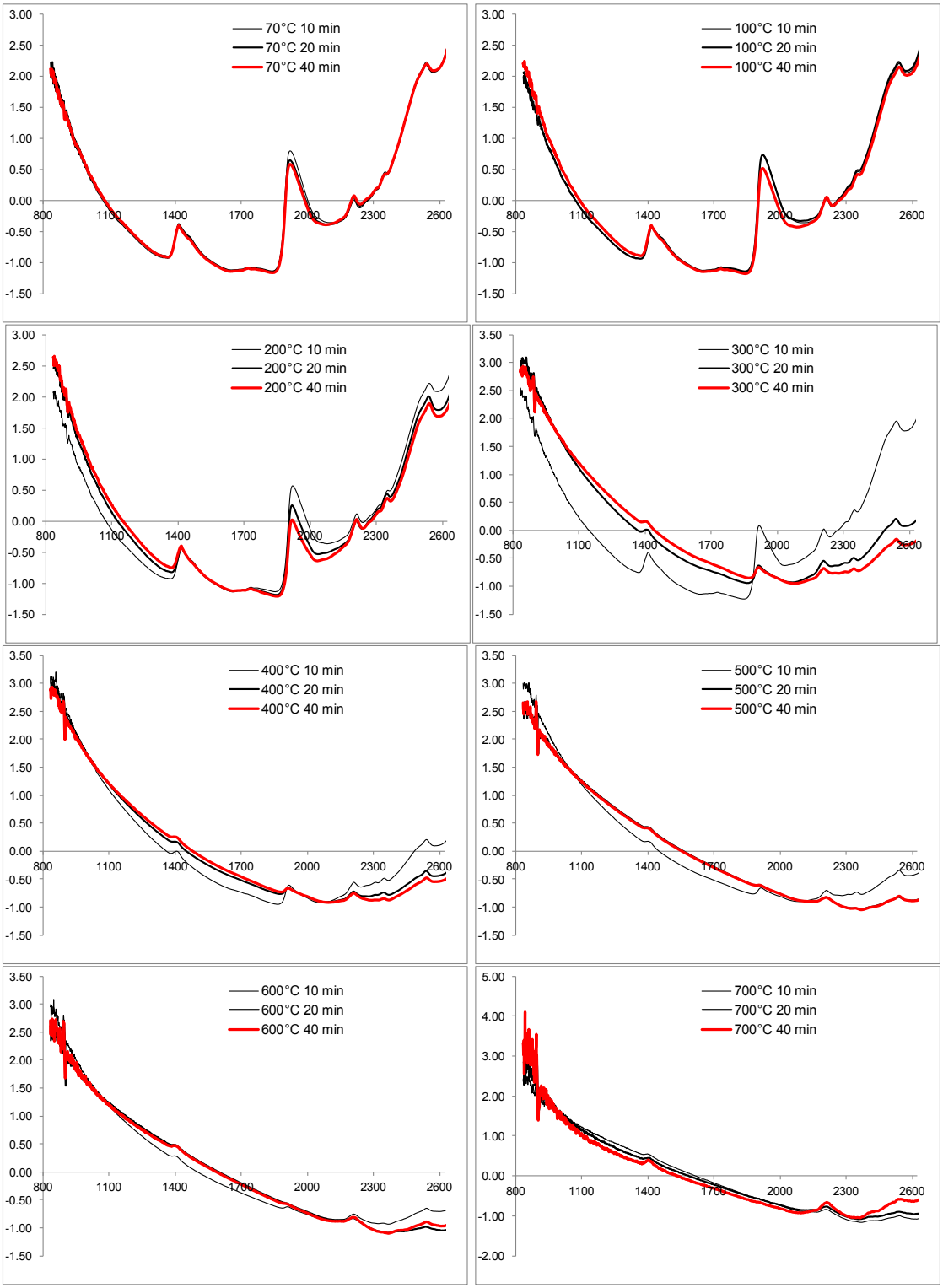

Figure S6. Representative NIR spectra of a few heated aliquots from Aitana (A3) at different exposure times. Standard normal variate.

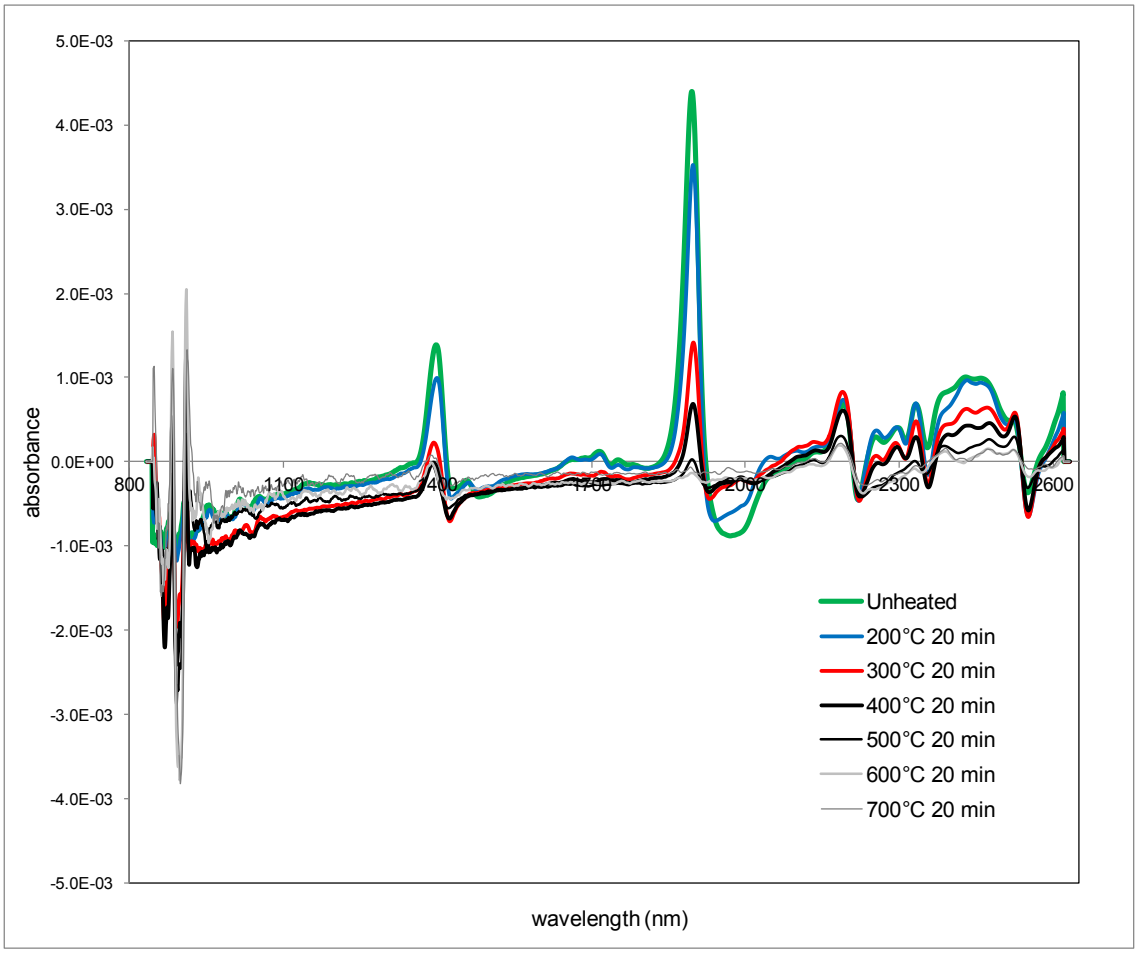

Figure S7. Representative NIR spectra of a few heated aliquots (only those heated during 20 minutes) from Aitana (A3): First derivate (25 points, Savitzky-Golay). Samples heated at 70°C and 100°C were omitted to enhance the clarity.

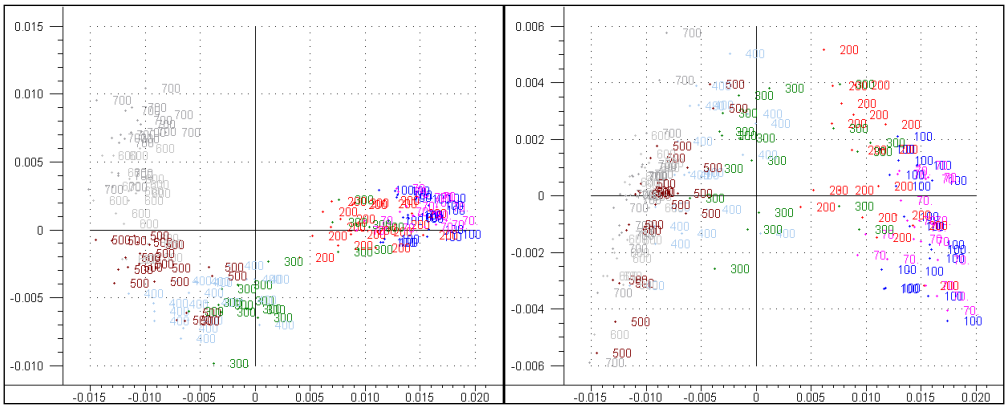

Figure S8. Principal component analysis of the NIR spectra (First derivate, 25 points, Savitzky-Golay) of the 144 heated aliquots from Aitana site. Left: scores over first and second factors. Right: scores over the first and third factors.

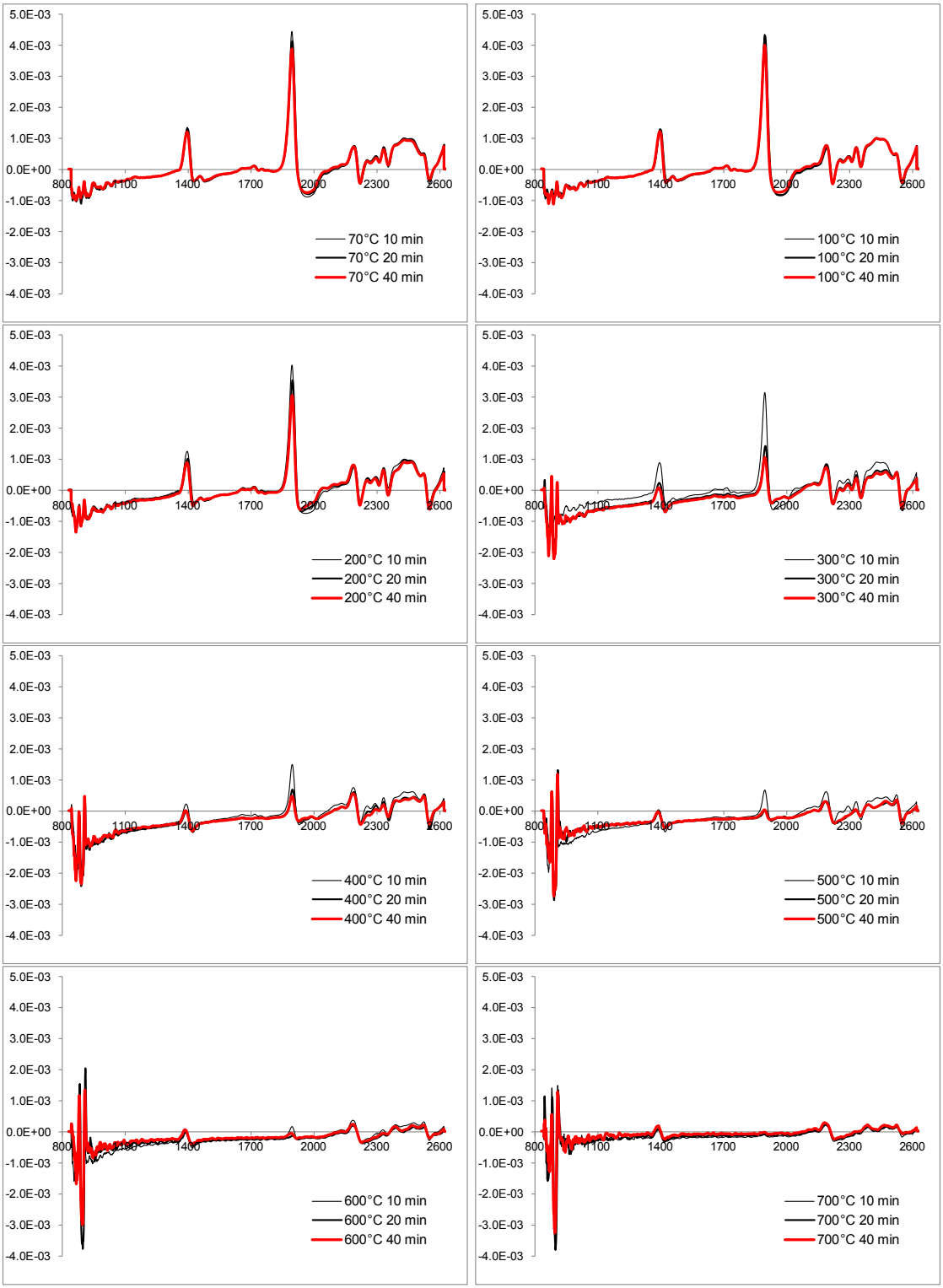

Figure S9. Representative NIR spectra of a few heated aliquots from Aitana (A3) at different exposure times. First derivate (25 points, Savitzky-Golay).

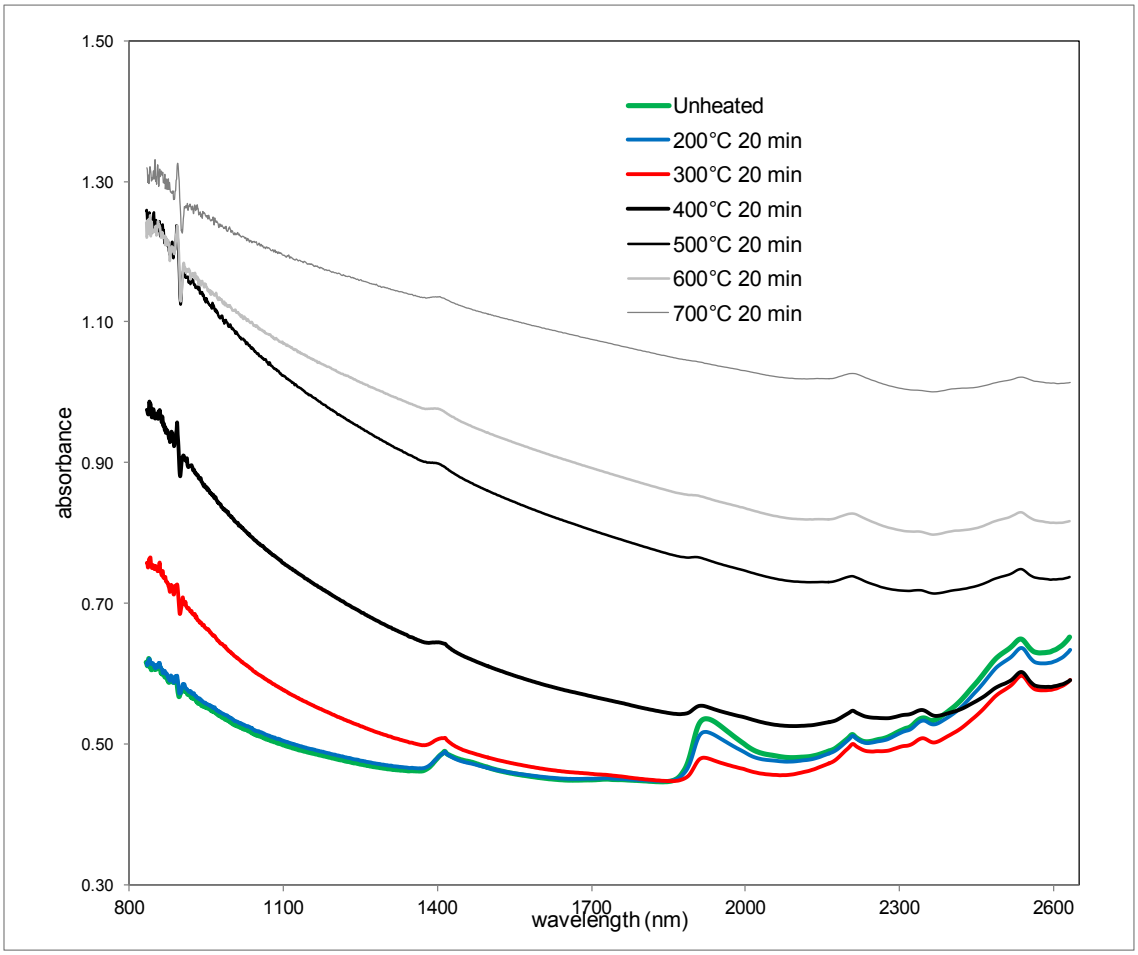

Figure S10. Representative NIR spectra of a few heated aliquots (only those heated during 20 minutes) from Maigmo (M3): Without preprocessing. Samples heated at 70°C and 100°C were omitted to enhance the clarity.

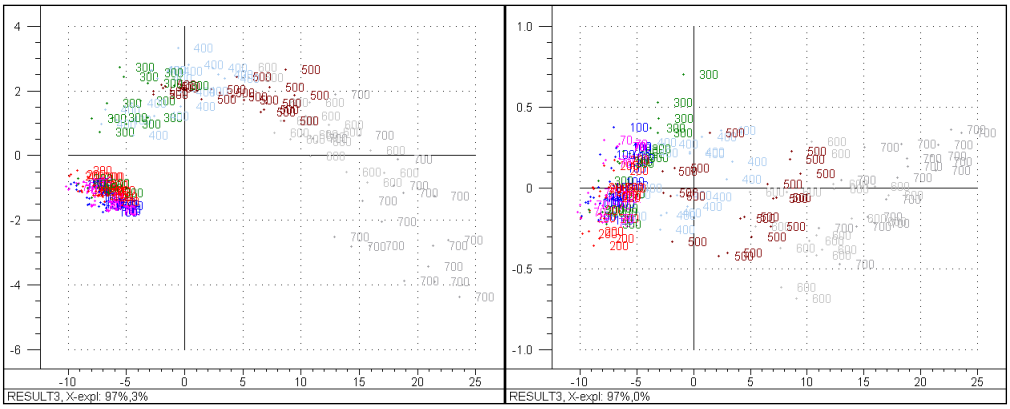

Figure S11. Principal component analysis of the NIR spectra (without preprocessing) of the 144 heated aliquots from Maigmo site. Left: scores over first and second factors. Right: scores over the first and third factors.

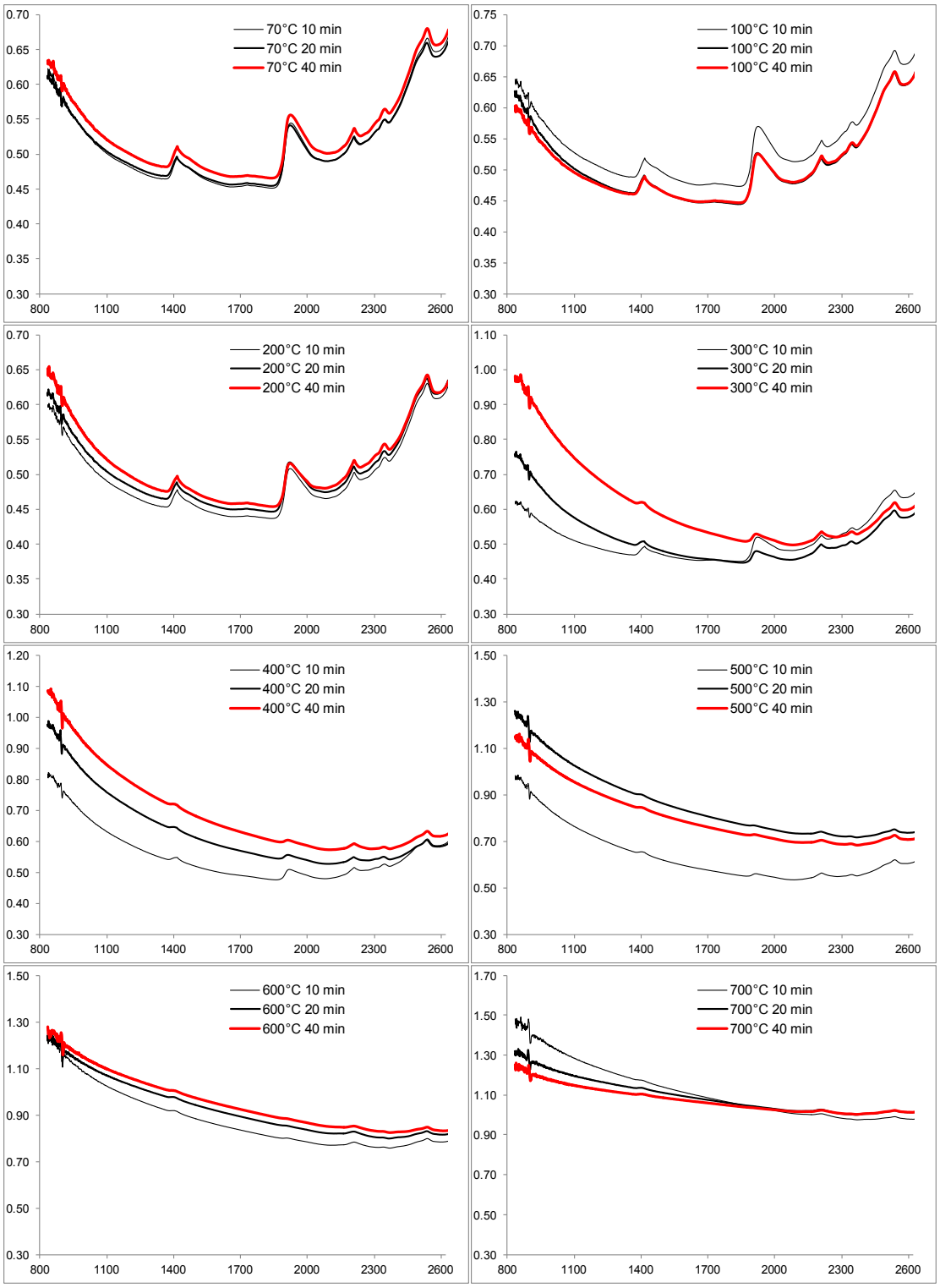

Figure S12. Representative NIR spectra of a few heated aliquots from Maigmo (M3) at different exposure times. Without preprocessing.

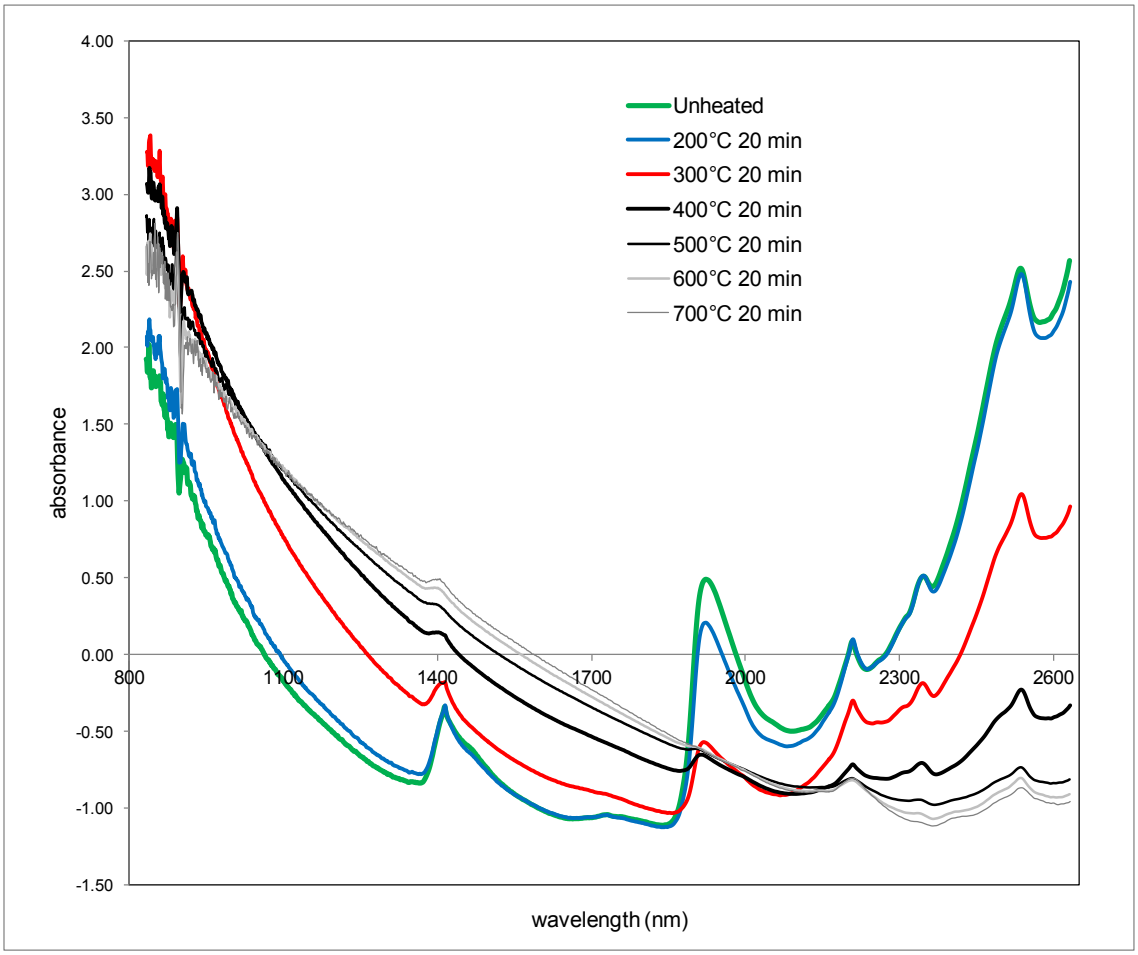

Figure S13. Representative NIR spectra of a few heated aliquots (only those heated during 20 minutes) from Maigmo (M3): Standard normal variate. Samples heated at 70°C and 100°C were omitted to enhance the clarity.

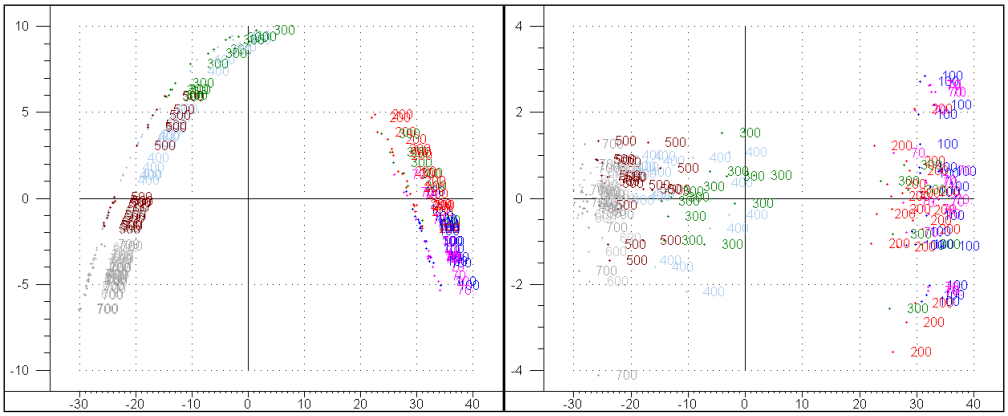

Figure S14. Principal component analysis of the NIR spectra (Standard normal variate) of the 144 heated aliquots from Maigmo site. Left: scores over first and second factors. Right: scores over the first and third factors.

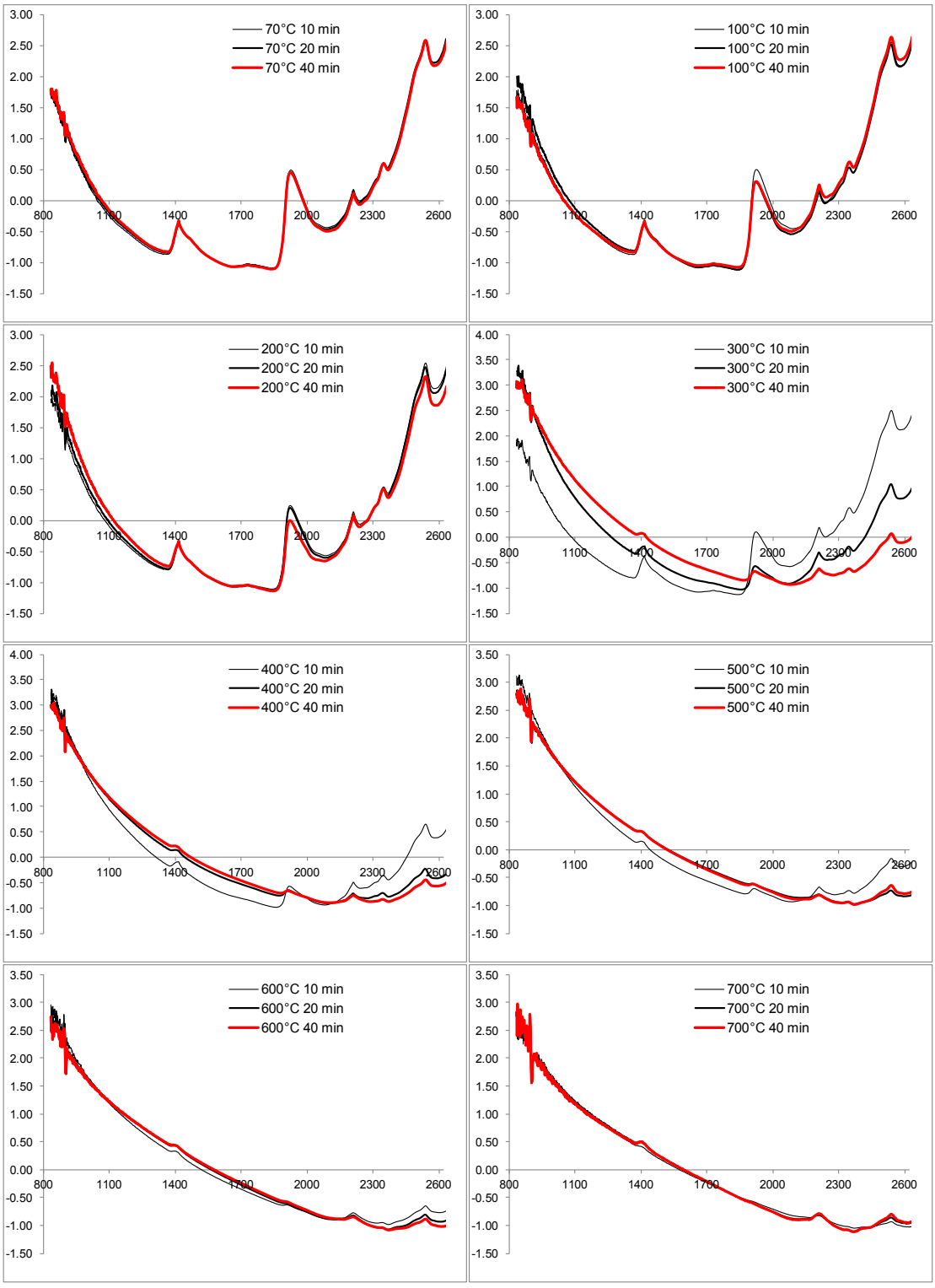

Figure S15. Representative NIR spectra of a few heated aliquots from Maigmo (M3) at different exposure times. Standard normal variate.

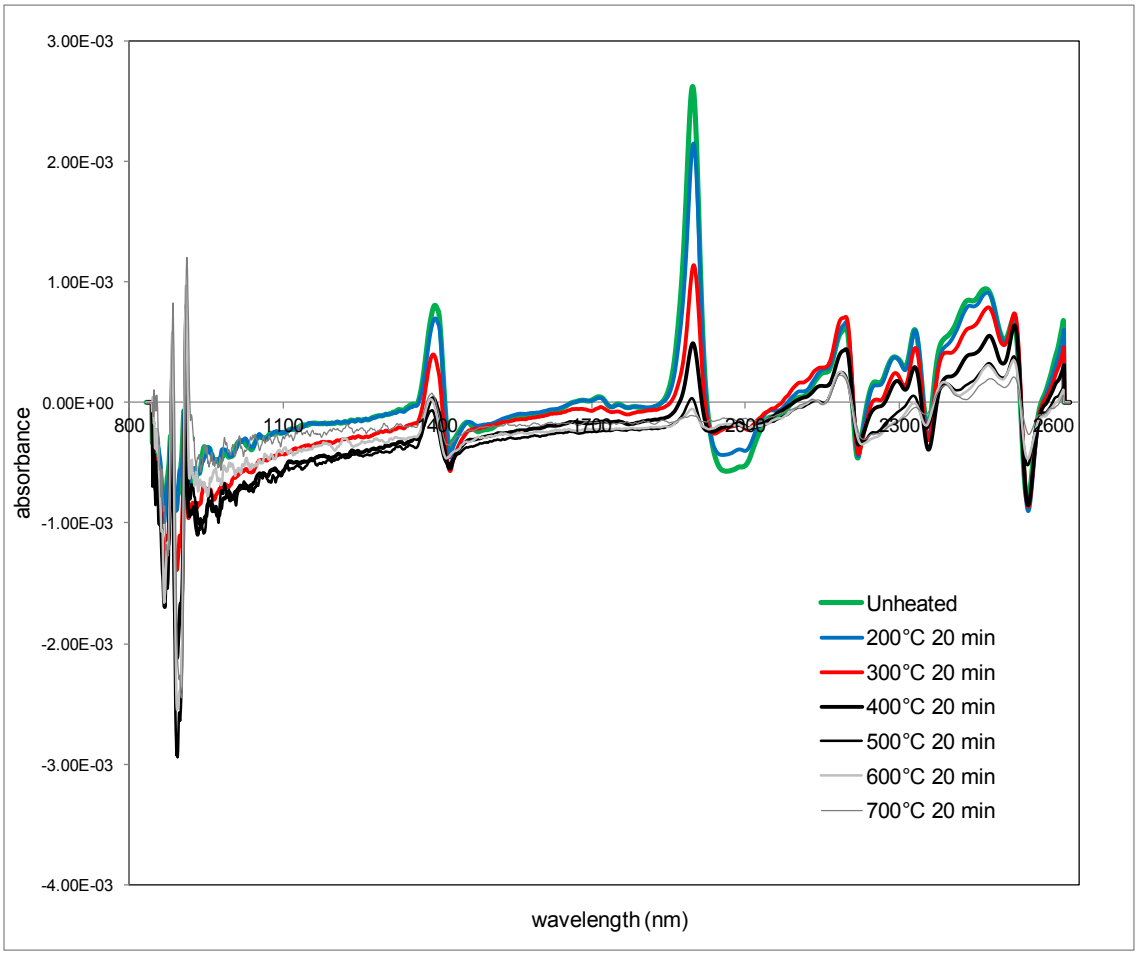

Figure S16. Representative NIR spectra of a few heated aliquots (only those heated during 20 minutes) from Maigmo (M3): First derivate (25 points, Savitzky-Golay). Samples heated at 70°C and 100°C were omitted to enhance the clarity.

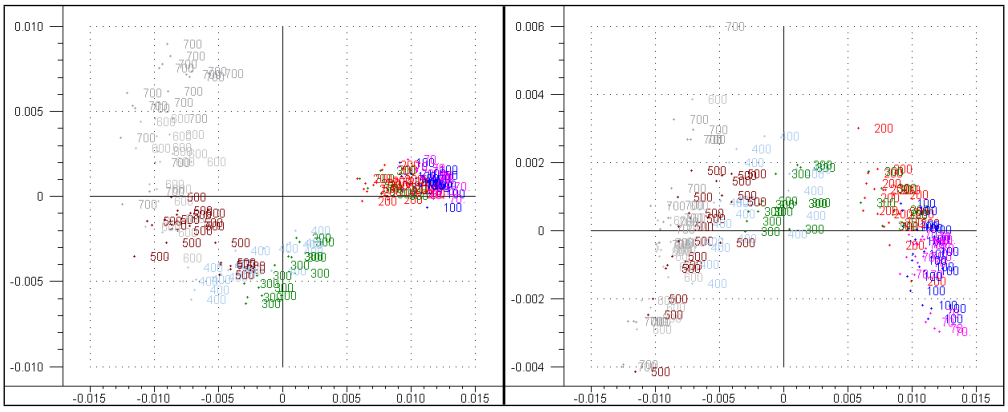

Figure S17. Principal component analysis of the NIR spectra (First derivate, 25 points, Savitzky-Golay) of the 144 heated aliquots from Maigmo site. Left: scores over first and second factors. Right: scores over the first and third factors.

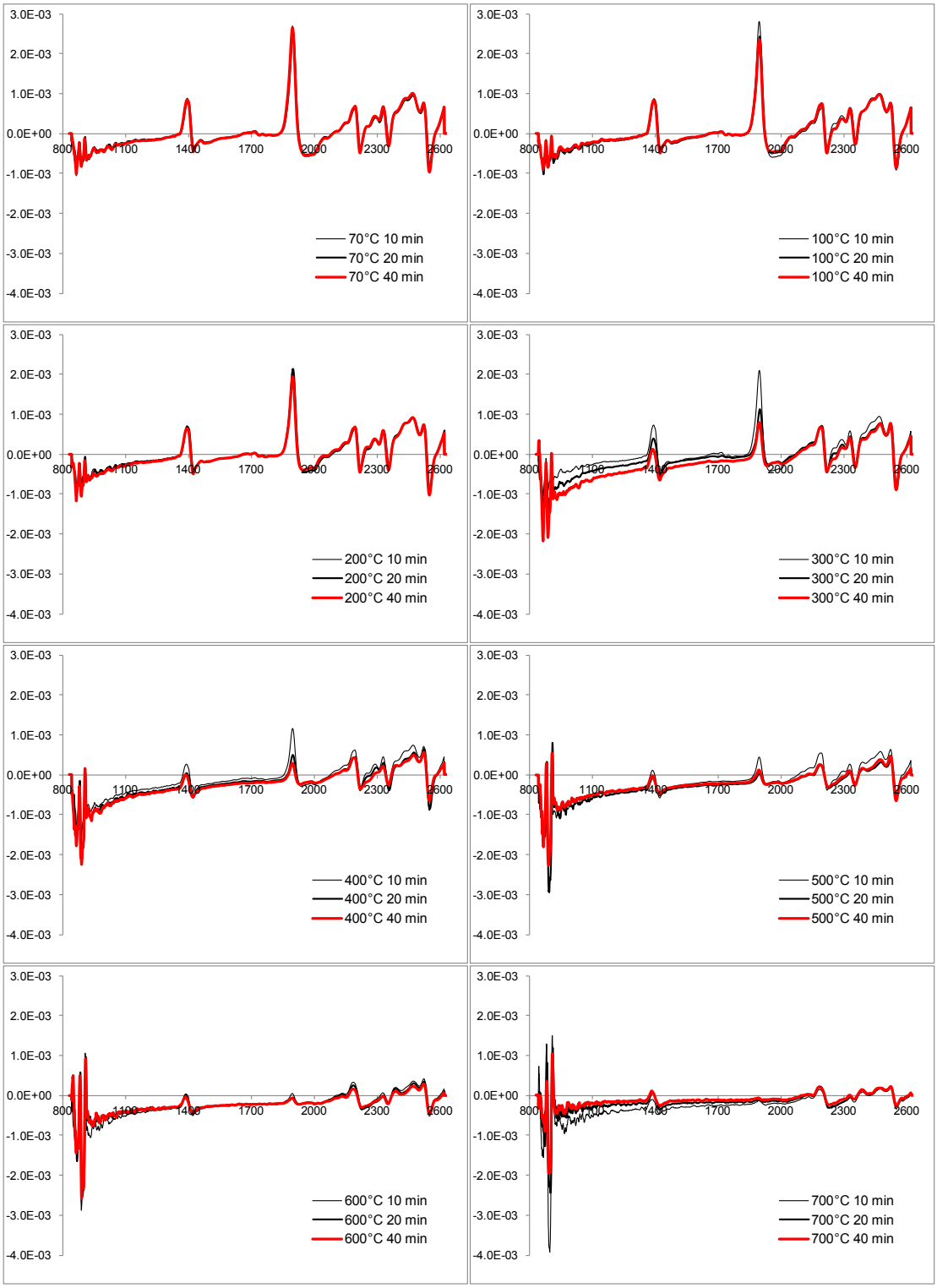

Figure S18. Representative NIR spectra of a few heated aliquots from Maigmo (M3) at different exposure times. First derivate (25 points, Savitzky-Golay).

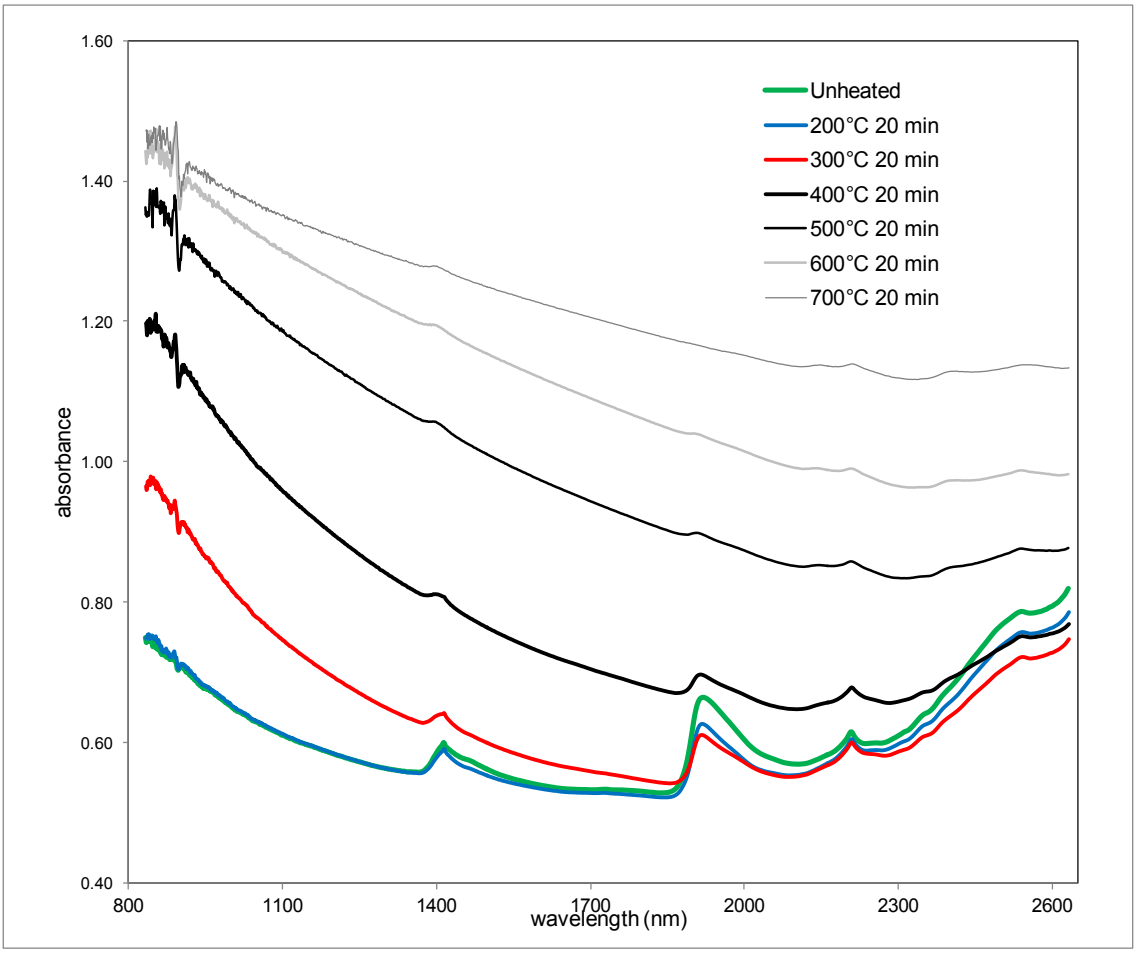

Figure S19. Representative NIR spectra of a few heated aliquots (only those heated during 20 minutes) from Pinoso (P3): Without preprocessing. Samples heated at 70ºC and 100ºC were omitted to enhance the clarity.

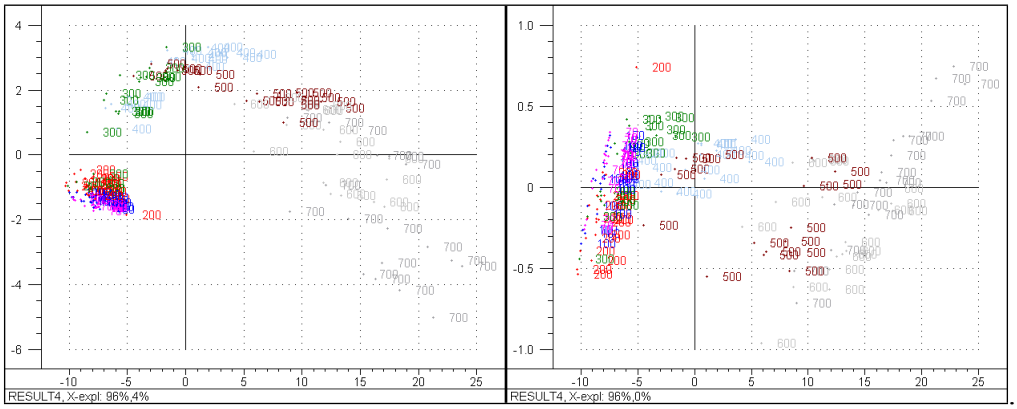

Figure S20. Principal component analysis of the NIR spectra (without preprocessing) of the 144 heated aliquots from Pinoso site. Left: scores over first and second factors. Right: scores over the first and third factors.

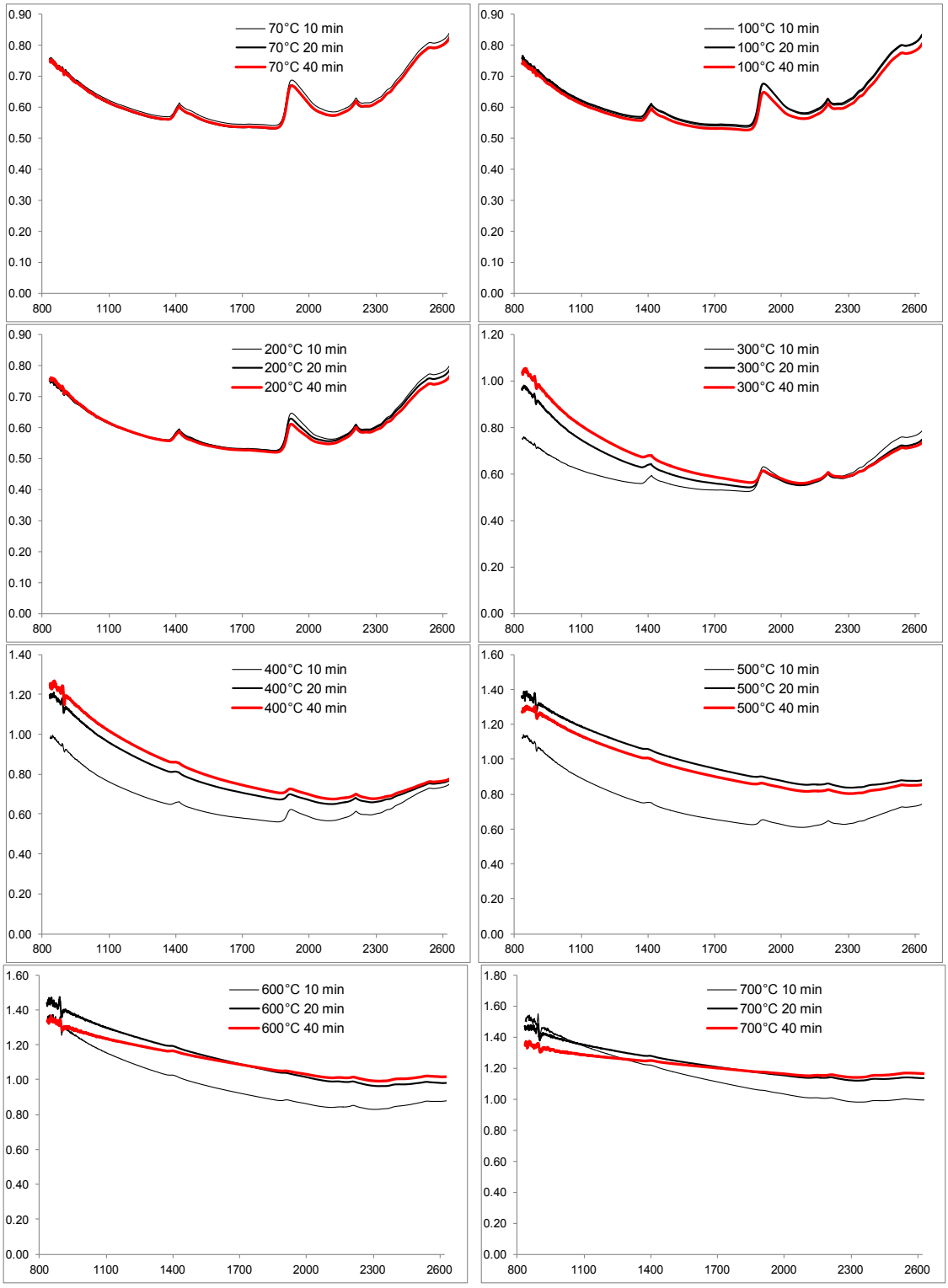

Figure S21. Representative NIR spectra of a few heated aliquots from Pinoso (P3) at different exposure times. Without preprocessing.

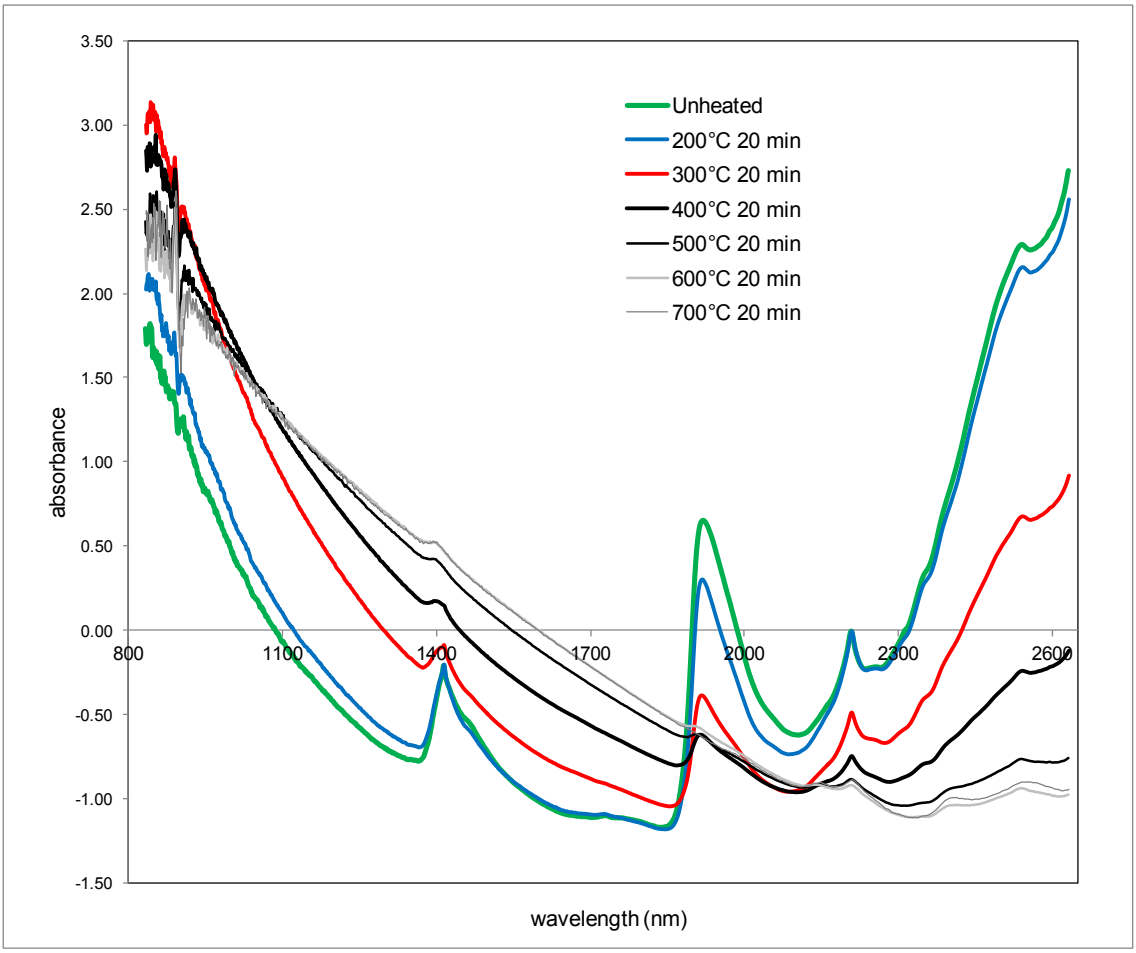

Figure S22. Representative NIR spectra of a few heated aliquots (only those heated during 20 minutes) from Pinoso (P3): Standard normal variate. Samples heated at 70°C and 100°C were omitted to enhance the clarity.

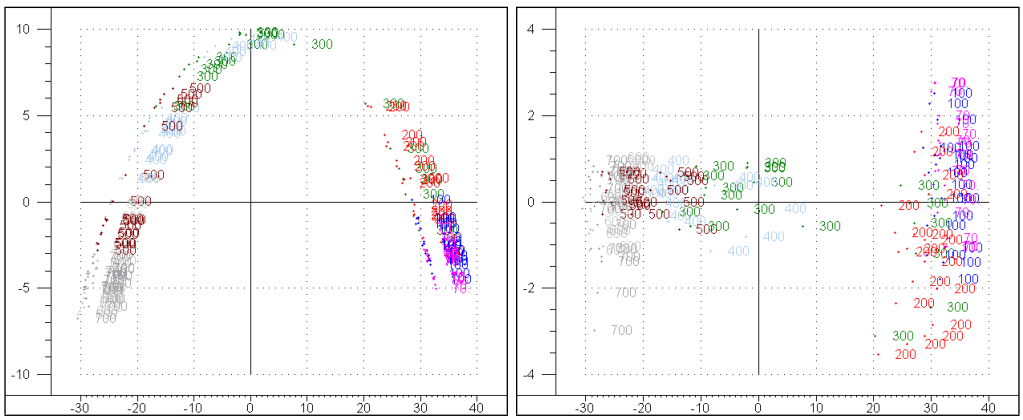

Figure S23. Principal component analysis of the NIR spectra (Standard normal variate) of the 144 heated aliquots from Pinoso site. Left: scores over first and second factors. Right: scores over the first and third factors.

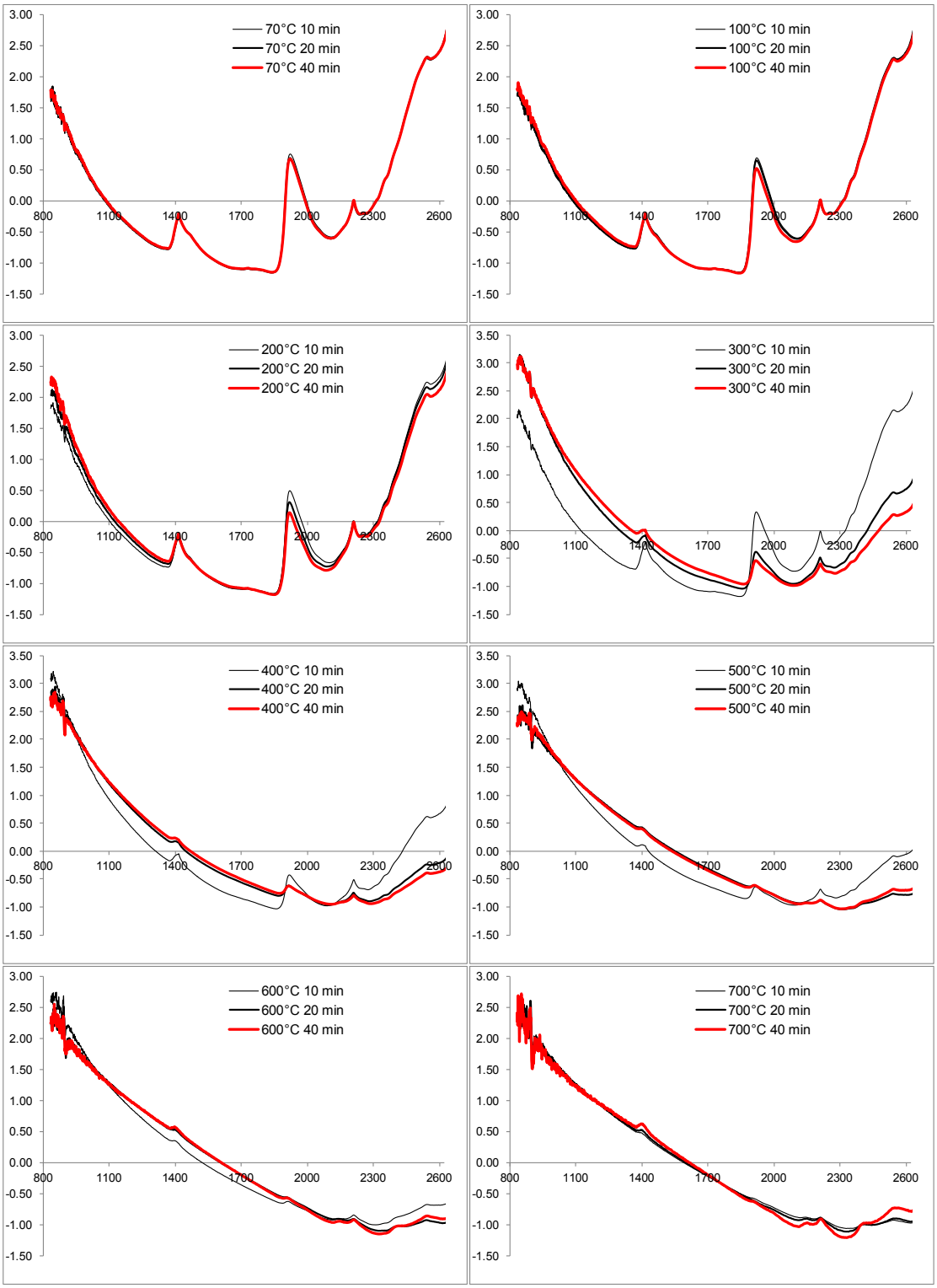

Figure S24. Representative NIR spectra of a few heated aliquots from Pinoso (P3) at different exposure times. Standard normal variate.

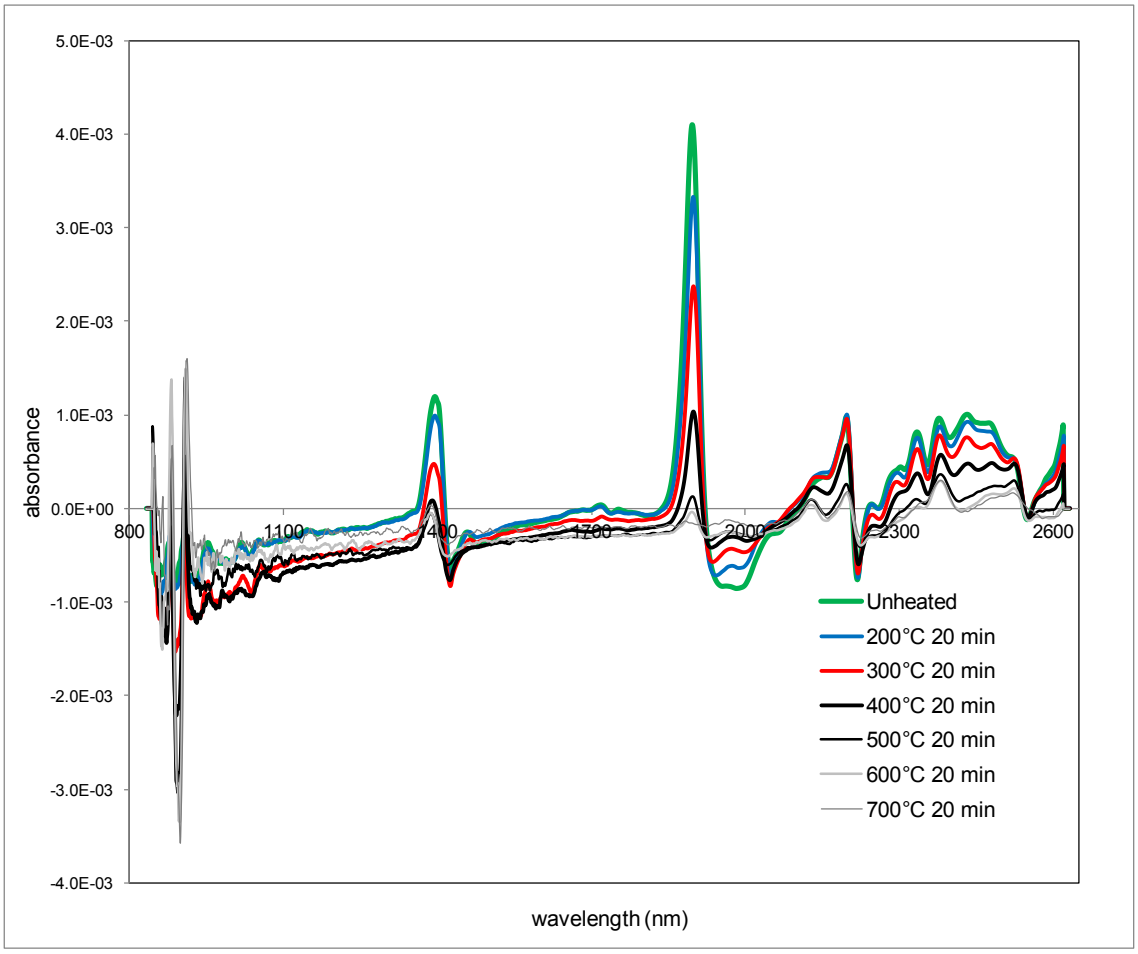

Figure S25. Representative NIR spectra of a few heated aliquots (only those heated during 20 minutes) from Pinoso (P3): First derivate (25 points, Savitzky-Golay). Samples heated at 70ºC and 100ºC were omitted to enhance the clarity.

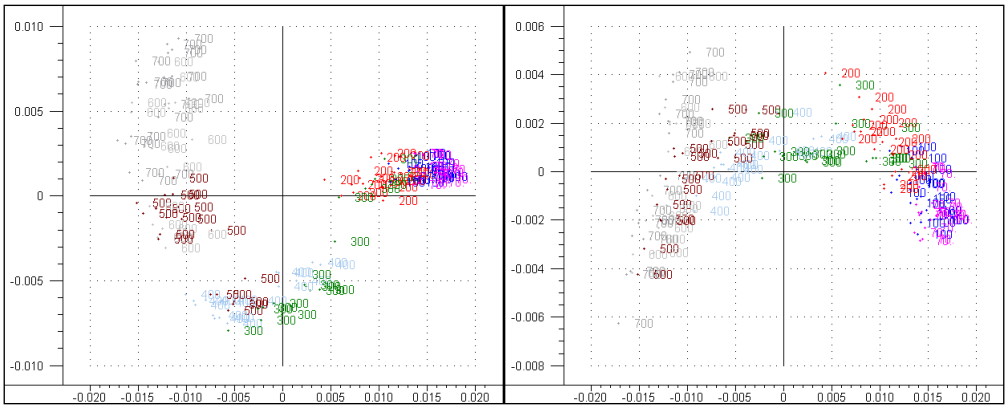

Figure S26. Principal component analysis of the NIR spectra (First derivate, 25 points, Savitzky-Golay) of the 144 heated aliquots from Pinoso site. Left: scores over first and second factors. Right: scores over the first and third factors.

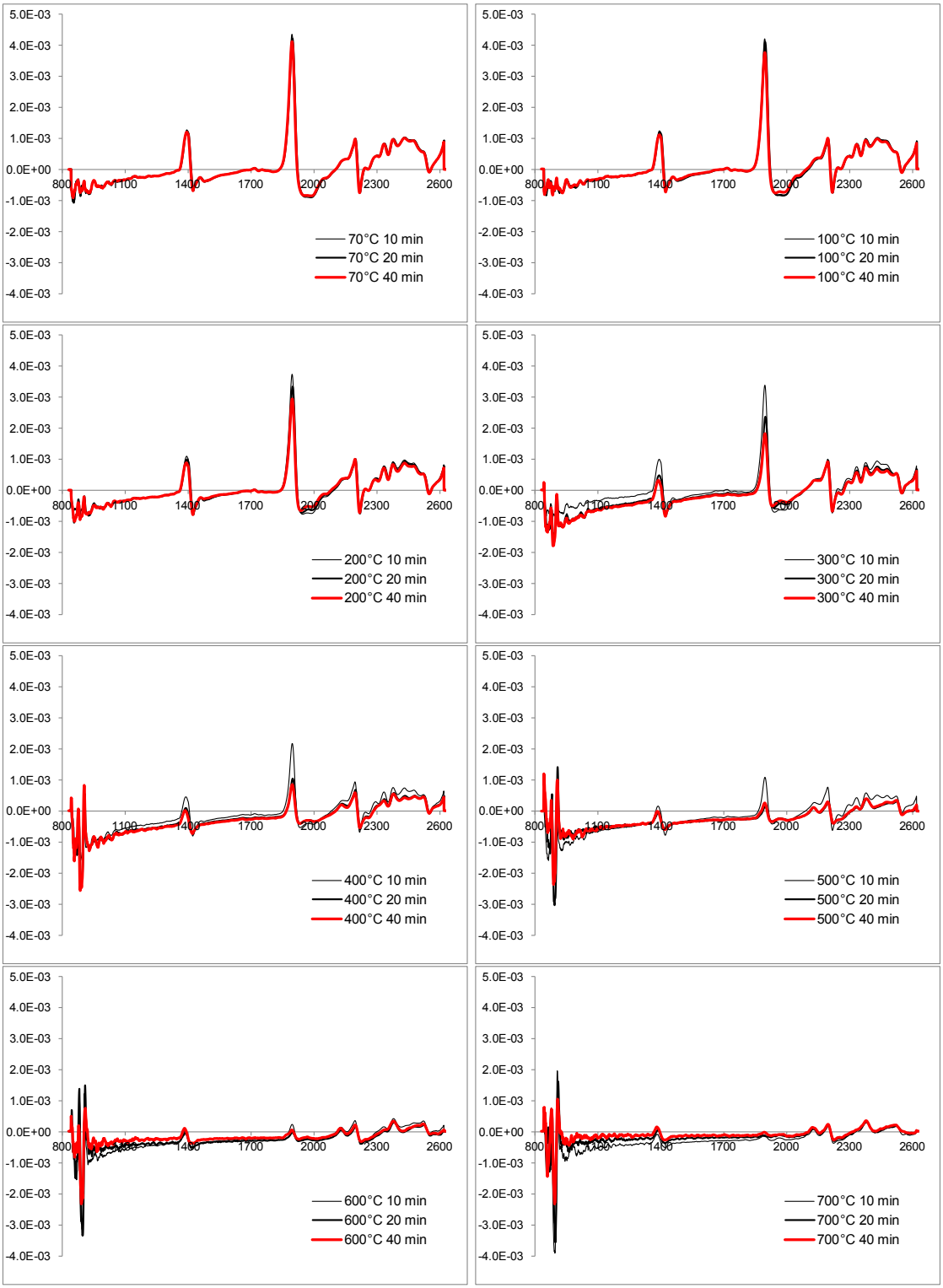

Figure S27. Representative NIR spectra of a few heated aliquots from Pinoso (P3) at different exposure times. First derivate (25 points, Savitzky-Golay).