PREDICTING WITH LIMITED DATA – INCREASING THE ACCURACY IN VIS-NIR DIFFUSE REFLECTANCE SPECTROSCOPY BY SMOTE

Christina Bogner

Ecological Modelling
BayCEER, University of Bayreuth
Dr.-Hans-Frisch-Str. 1–3
95445 Bayreuth, Germany

Anna Kühnel, Bernd Huwe

Soil Physics Group
BayCEER, University of Bayreuth
Universitätstr. 30
95447 Bayreuth, Germany

ABSTRACT

Diffuse reflectance spectroscopy is a powerful technique to predict soil properties. It can be used *in situ* to provide data inexpensively and rapidly compared to the standard laboratory measurements. Because most spectral data bases contain air-dried samples scanned in the laboratory, field spectra acquired *in situ* are either absent or rare in calibration data sets. However, when models are calibrated on air-dried spectra, prediction using field spectra are often inaccurate. We propose a framework to calibrate partial least squares models when field spectra are rare using synthetic minority oversampling technique (SMOTE). We calibrated a model to predict soil organic carbon content using air-dried spectra spiked with synthetic field spectra. The root mean-squared error of prediction decreased from 6.18 to 2.12 mg g^{-1} and R^2 increased from -0.53 to 0.82 compared to the model calibrated on air-dried spectra only.

***Index Terms*—** diffuse reflectance spectroscopy, soil, partial least squares, calibration, SMOTE

1. INTRODUCTION

Diffuse reflectance spectroscopy in the visible and near-infrared range (VIS-NIR DRS) has proved to be useful to assess various soil properties [1]. It can be employed to provide more data rapidly and inexpensively compared to classical laboratory analysis. Therefore, DRS is increasingly used for vast soil surveys in agriculture and environmental research [2, 3]. Recently, several studies have shown the applicability of VIS-NIR DRS *in situ* as a proximal soil sensing technique [4, 5].

To predict soil properties from soil spectra, a model is calibrated, often using partial least squares (PLS) regression. However, when calibration is based on air-dried spectra collected under laboratory conditions, predictions of soil properties from field spectra tend to be less accurate [4]. Usually, this decrease in accuracy is attributed to varying moisture between air-dried calibration samples and field spectra recorded with a variable moisture content. Different remediation techniques have been proposed, ranging from advanced preprocessing of the spectra [6] to "spiking" the calibration set with field spectra [4].

In our study, we adopt a slightly different view on the calibration problem. It does not only apply to the varying moisture conditions between the calibration data set and the field spectra. Indeed, it is also valid if we want to predict soil properties in a range where calibration samples are rare. Mining with rarity or learning from imbalanced data is an ongoing research topic in Machine Learning [7]. In imbalanced data sets frequent samples outnumber the rare once. Therefore, a model will be better at predicting the former and might fail for the latter.

Two different approaches exist to take care of the data imbalance: we can either adjust the model or "balance" the data. The latter approach has the advantage that we can use the usual modelling framework. Synthetic minority oversampling technique (SMOTE) is one way to balance the data. It was first proposed for classification [8] and recently for regression [9]. SMOTE oversamples the rare data by generating synthetic points and thus helps to equalize the distribution.

In this study, we propose a strategy to increase the prediction accuracy of soil properties from field spectra when they are rare in calibration. The goal of this study is to build a calibration model to predict soil organic carbon content (SOCC) from field spectra by air-dried samples spiked with synthetic field spectra.

2. MATERIAL AND METHODS

2.1. Data acquisition

The studied soil was sampled at the southern slopes of Mt. Kilimanjaro, Tanzania (3° 4′ 33″ S, 37° 21′ 12″ E) in coffee plantations. Due to favourable soil and climate in this region, extensive coffee plantations constitute a frequent form of land use. We took 31 samples for calibration at 4 different study sites. For validation, we scanned 12 field spectra at a wall of a soil pit and sampled soil material for chemical analysis at the

scanned spots. We call these validation field spectra F.

After collection, the calibration samples were dried in an oven at 45°C and sieved < 2 mm. Subsequently, they were scanned with an AgriSpec portable spectrophotometer equipped with a Contact Probe (Analytical Spectral Devices, Boulder, Colorado) in the range 350–2500 nm with 1 nm intervals. The same spectrometer was used in the field. The instrument was calibrated with a Spectralon white tile before scanning the soil samples. For the measurement, a thoroughly mixed aliquot of the sample was placed in a small cup and the surface was smoothed with a spatula. Each sample was scanned 30 times and the signal averaged to reduce the noise. In the following, we call this calibration data set L.

SOCC was measured in a CNS-Analyser by high temperature combustion with conductivity detectors.

2.2. Generating data by synthetic minority oversampling

To generate new data to spike the calibration data set L, we used SMOTE [8] and its extension for regression [9]. This algorithm consists of generating new synthetic data using existing data and is summarized below. In our case, we generated new spectra and the related SOCC using the field spectra F. The new spectra are created by calculating the difference between a field spectrum and one of its nearest neighbours and adding this difference (weighted by a random number between 0 and 1) to the field spectrum. The SOCC of the synthetic spectrum is then a weighted average between the SOCC of the field spectrum and the used nearest neighbour.

SMOTE has two parameters, namely N, the number of points to generate for each existing point (given in percent of the whole data set) and k, the number of nearest neighbours. To study the influence of these parameters we generated six different synthetic data sets S1 through S6, varying $N = 100, 200, 300$ and $k = 3, 5$.

Algorithm: SMOTE

Input: T original samples to be SMOTEd
Amount of SMOTE $N\%$
Number of nearest neighbours k
Output: $(N/100) \times T$ synthetic samples with their target values (i.e. concentrations)
if $N < 100$ **then**
  Randomize the T original samples:
  $T = (N/100) \times T$
  $N = 100$
end
$orig.s[i]$: original sample $i, i = 1, \ldots, T$
$orig.t[i]$: target value of original sample i
$new.s[j]$: synthetic sample $j, j = 1, \ldots, (N/100) \times T$
$new.t[j]$: target values of synthetic sample j
$ng \leftarrow N/100$: number of synthetic samples to compute for each original sample
Generate synthetic samples:
for i *in* 1 *to* T **do**
  $nns \leftarrow$ compute k nearest neighbours for $orig.s[i]$
  for ℓ *in* 1 *to* ng **do**
    randomly choose $x \in nns$
    $diff = x - orig.s[i]$
    $new.s[(i-1) \times ng + \ell] =$
    $orig.s[i] + \mathrm{RANDOM}(0,1) \times diff$
    $d_1 = \mathrm{DIST}(new.s, orig.s[i])$
    $d_2 = \mathrm{DIST}(new.s, x)$
    $target = \frac{d_2 \times orig.t(orig.s) + d_1 \times orig.t(x)}{d_1 + d_2}$
  end
end
return $new.t \cup new.s$

2.3. Data pretreatment and explorative analysis

We corrected each spectrum (calibration, validation and synthetic) for the offset at 1000 and 1830 nm and kept only parts with a high signal-to-noise ratio (450–2400 nm). Then, we transformed the spectra to absorbance ($\log_{10}(1/\mathrm{reflectance})$) and smoothed them using the Singular Spectrum Analysis (SSA). SSA is a non-parametric technique to decompose a signal into additive components that can be identified as the signal itself or as noise [10]. Finally, we divided each spectrum by its maximum and calculated the first derivative.

In order to assess similarities between the calibration, validation and synthetic data sets, we calculated the Principal Component Analysis (PCA) of the (uncorrected original) spectra L and F and projected the synthetic data into the space spanned by the principal components.

2.4. Partial least squares regression

We calibrated seven different PLS models. For model I we used the data set L, the spectra scanned under laboratory conditions. Model II through VII were calibrated on L spiked with synthetic spectra S1 through S6. To find the best model I through VII, we varied the number of PLS components between 1 and 15. Based on the predictions in the leave-one-out cross-validation (LOOCV) we calculated the corrected Akaike Information Criterion [11] $\mathrm{AIC_c} = n \ln(\mathrm{RMSE}^2) + 2p + \frac{2p(p+1)}{n-p-1}$, where n is the number of calibration samples, p the number of PLS components and RMSE the root mean-squared error. The latter is defined as $\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$, where $\hat{y}_i$ are the predicted and y_i the measured SOCCs. We selected the model with the smallest $\mathrm{AIC_c}$ as the most plausible.

To assess the model quality, we used the RMSE, the mean error $\mathrm{ME} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i - y_i$ and the coefficient of determination $R^2 = 1 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / \sum_{i=1}^{n}(y_i - \bar{y})^2$, where $\bar{y}$ is the mean SOCC.

2.5. Monte Carlo simulations

SMOTE has two random components because it selects spectra randomly (with replacement) among the nearest neighbours and weights the difference between spectra by a random number (between 0 and 1). To study the influence of these random steps, we generated 100 different datasets S1 through S6. Each data set was then used to spike the calibration data set L, to build a new PLS model and to predict the data set F.

3. RESULTS AND DISCUSSION

The first principal components (PCs) explain 85.4% and 11.2% of variance, respectively. We can clearly identify two distinct groups of samples: the calibration data set L and the field spectra F (Fig. 1). In other words, the data sets L and F differ. The synthetic points that were projected into the space spanned by the PCs resemble the field spectra as expected.

The distinct characteristics of the data sets L and F accord well with the difficulties to predict the data set F by using the laboratory spectra L only (Table 1 and Table 2). Although the LOOCV of model I yields a moderate RMSE and a large R^2, the validation on the data set F fails.

Spiking the calibration data set L with synthetic spectra increases the prediction accuracy of the SOCC in the data set F. Actually, the RMSE decreases and R^2 increases with increasing number of synthetic points both for the LOOCV and the validation (Table 1 and Table 2). However, the number of model parameters also increases from 2 to 7.

The Monte Carlo results show only a small variability in the interquartile range. However, some synthetic data sets in model V produced R^2 values smaller than -0.53, the value we obtain in model I on air-dried samples only. This might be due to the combination of neighbours during smoting. In general, models with 5 neighbours were more accurate than those with 3 neighbours. However, the number of neighbours had a smaller influence on the prediction accuracy than the number of synthetic points.

It is difficult to decide *a priory* how many synthetic points should be included in the calibration. Indeed, in a classification problem the goal is to approximate an equal distribution of different classes such that the rare class becomes an ordinary one. In regression, however, we do not know which features of the data make them rare. For our data, the range of SOCC in the data set L is larger than in the data set F. Therefore, we conclude that concentration is not responsible for the difference between these data sets.

Based on the Monte Carlo results we chose one synthetic data set from model VI, namely the one with the median number of model parameters and the best R^2 in the validation. Thus, the calibration data set includes 31 air-dried and 24 synthetic spectra. Compared to model I, spiking the air-dried data set L with these synthetic spectra clearly improves the prediction of the data set F (Fig. 2).

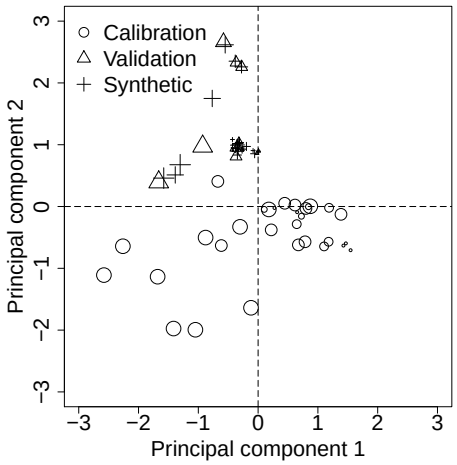

Fig. 1. Principal component analysis of calibration data set L, validation data set F and one synthetic data set S5. The symbol size was scaled according to the SOCC.

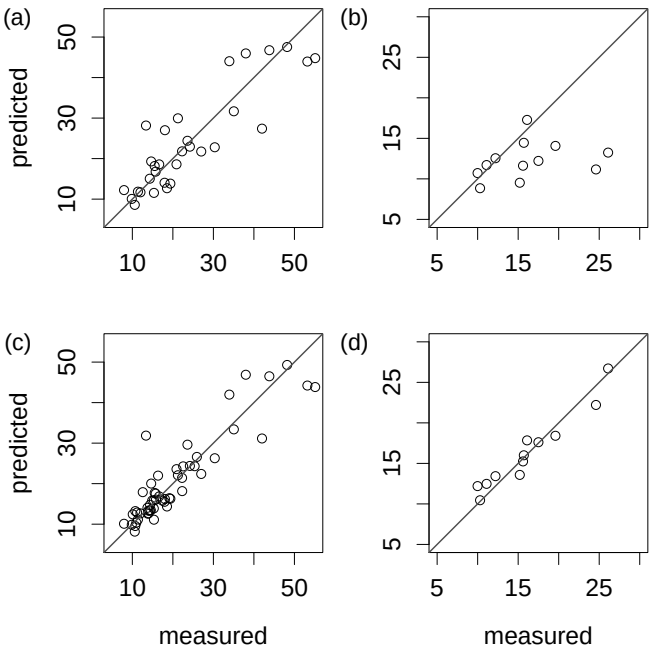

Fig. 2. Results of (a) leave-one-out cross-validation on data set L (model I), (b) validation on data set F, (c) leave-one-out cross-validation on data set L spiked with a synthetic data set (model VI) and (d) validation on data set F.

4. CONCLUSIONS

We propose a framework to predict soil properties from *in situ* acquired field spectra by spiking air-dried laboratory calibration data by synthetic ones generated from these field spectra. In general, the prediction accuracy increases when a sufficient number of synthetic points is included in the calibration. However, because it is difficult to determine this number *a priori*,

Table 1. Statistics of the PLS calibration. Median values and 25% and 75% quantiles in parenthesis.

Model	Data set(s)	$N(\%)$	k	p	RMSE (mg g^{-1})	R^2	ME (mg g^{-1})
I	L	–	–	2	6.25	0.77	−0.20
II	L and S1	100	3	5 (4; 5)	5.29 (5.18; 5.47)	0.80 (0.79; 0.81)	−0.06 (−0.10; −0.01)
III	L and S2	200	3	6 (6; 6)	4.51 (4.47; 4.56)	0.83 (0.83; 0.84)	0.07 (0.03; 0.11)
IV	L and S3	300	3	7 (6; 7)	4.01 (3.98; 4.06)	0.85 (0.84; 0.85)	0.08 (0.05; 0.11)
V	L and S4	100	5	4 (3; 5)	5.31 (5.16; 5.55)	0.80 (0.78; 0.81)	−0.02 (−0.10; 0.04)
VI	L and S5	200	5	6 (6; 6)	4.51 (4.45; 4.55)	0.83 (0.83; 0.84)	0.06 (0.01; 0.10)
VII	L and S6	300	5	6 (6; 7)	4.05 (4.02; 4.08)	0.84 (0.84; 0.85)	0.07 (0.05; 0.09)

Table 2. Statistics of the PLS validation. Median values and 25% and 75% quantiles in parenthesis.

Model	RMSE (mg g^{-1})	R^2	ME (mg g^{-1})
I	6.18	−0.53	−3.88
II	3.09 (2.82; 3.58)	0.62 (0.49; 0.68)	−0.03 (−0.53; 0.79)
III	2.00 (1.79; 2.40)	0.84 (0.77; 0.87)	0.14 (−0.01; 0.36)
IV	1.31 (1.08; 1.58)	0.93 (0.90; 0.95)	0.16 (0.06; 0.27)
V	3.06 (2.79; 3.56)	0.62 (0.49; 0.69)	−0.28 (−0.70; 0.79)
VI	2.12 (1.81; 2.39)	0.82 (0.77; 0.87)	0.24 (−0.04; 0.48)
VII	1.62 (1.29; 2.07)	0.89 (0.83; 0.93)	0.18 (0.02; 0.37)

we recommend to generate several synthetic data sets to find an appropriate model.

ACKNOWLEDGEMENTS

This study is part of the project DFG FOR 1246 "Kilimanjaro ecosystems under global change: Linking biodiversity, biotic interactions and biogeochemical ecosystem processes" and was supported by the Deutsche Forschungsgemeinschaft.

5. REFERENCES

[1] B. Stenberg and R. A. Viscarra Rossel, "Diffuse reflectance spectroscopy for high-resolution soil sensing," in *Proximal Soil Sensing*, R. A. Viscarra Rossel, A. B. McBratney, and B. Minasny, Eds., pp. 29–47. Springer, 2010.

[2] K. D. Shepherd and M. G. Walsh, "Infrared spectroscopy - enabling an evidence-based diagnostic surveillance approach to agricultural and environmental management in developing countries," *Journal of near Infrared Spectroscopy*, vol. 15, no. 1, pp. 1–19, 2007.

[3] T.-G. Vågen, K. D. Shepherd, and M. G. Walsh, "Sensing landscape level change in soil fertility following deforestation and conversion in the highlands of Madagascar using Vis-NIR spectroscopy," *Geoderma*, vol. 133, no. 3, pp. 281–294, 2006.

[4] R. A. Viscarra Rossel, S. R. Cattle, A. Ortega, and Y. Fouad, "In situ measurements of soil colour, mineral composition and clay content by vis–nir spectroscopy," *Geoderma*, vol. 150, no. 3, pp. 253–266, 2009.

[5] T. H. Waiser, C. L. S. Morgan, D. J. Brown, and C. T. Hallmark, "In situ characterization of soil clay content with visible near-infrared diffuse reflectance spectroscopy," *Soil Science Society of America Journal*, vol. 71, no. 2, pp. 389–396, 2007.

[6] B. Minasny, A. B. McBratney, V. Bellon-Maurel, J.-M. Roger, A. Gobrecht, L. Ferrand, and S. Joalland, "Removing the effect of soil moisture from nir diffuse reflectance spectra for the prediction of soil organic carbon," *Geoderma*, vol. 167, pp. 118–124, 2011.

[7] G. M. Weiss, "Mining with rarity: A unifying framework," *Sigkdd Explorations*, vol. 6, pp. 1–19, 2004.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[9] Luís Torgo, Rita P Ribeiro, Bernhard Pfahringer, and Paula Branco, "Smote for regression," in *Progress in Artificial Intelligence*, pp. 378–389. Springer, 2013.

[10] Nina Golyandina and Anatoly Zhigljavsky, *Singular spectrum analysis for time series*, Springer, 2013.

[11] N. Sugiura, "Further analysts of the data by akaike's information criterion and the finite corrections," *Communications in Statistics-Theory and Methods*, vol. 7, no. 1, pp. 13–26, 1978.