European Journal of Soil Science

European Journal of Soil Science, January 2019, **70**, 42–53 doi: 10.1111/ejss.12715

Special issue article

Rapid determination of soil classes in soil profiles using vis–NIR spectroscopy and multiple objectives mixed support vector classification

S. Chen[a,b,c], S. Li[a], W. Ma[d], W. Ji[e], D. Xu[a], Z. Shi[a,f] & G. Zhang[f]

[a]*Institute of Agricultural Remote Sensing and Information Technology Application, College of Environmental and Resource Sciences, Zhejiang University, Hangzhou, China,* [b]*INRA, Unité InfoSol, Orléans, France,* [c]*SAS, INRA, Agrocampus Ouest, Rennes, France,* [d]*Key Laboratory of Information Traceability for Agricultural Products, Ministry of Agriculture of China, Hangzhou, China,* [e]*Department of Soil and Environment, Swedish University of Agricultural Sciences, Skara, Sweden, and* [f]*State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing, China*

Summary

Visible-near infrared (vis–NIR) spectroscopy can reveal various soil properties and facilitate soil classification. However, few studies have attempted to classify vertical soil profiles that contain several genetic horizons. Here, we propose the 'multiple objectives mixed support vector classification' (MOM–SVC) method to classify soil profiles. A total of 130 soil profiles were collected from genetic horizons in Zhejiang Province, China. After laboratory analysis, soil profiles were classified according to the Chinese Soil Taxonomy system. Vis–NIR spectra were recorded from each genetic horizon of each soil profile and were then pre-processed. We performed the MOM–SVC method as follows: (i) created a support vector machine (SVM) model (one-versus-one approach) using spectral data from all soil horizons in calibration profiles, (ii) applied the SVM model on each horizon of the profile to be predicted, (iii) extracted 'votes' from each horizon and mixed (or summarized) them into the votes of each profile to be predicted and (iv) classified each profile by the majority-voting method. We also investigated whether the additional input of auxiliary soil information (e.g. moist soil colour, soil organic matter and soil texture), which could be measured easily or be well predicted by vis–NIR spectroscopy, could improve the accuracy of soil classification when combined with it. Independent validation results showed that the MOM–SVC method performed better at the soil order level than at the suborder level. Adding auxiliary soil information to the classification model improved the overall accuracy of classification at the soil order level. The proposed MOM–SVC method provides a fast objective diagnostic of soil classes for use in soil surveys and can help to update soil databases when a more objective soil classification system is developed.

Highlights

- The MOM–SVC method can be used to classify soil profiles objectively with a variety of soil horizons.
- Stratified random sampling was used to quantify prediction uncertainty in classification
- MOM–SVC can predict soil orders with greater accuracy than suborders.
- Adding auxiliary soil information into the classification model improved prediction accuracy.

Introduction

It is well known that a proper understanding of the patterns of soil distribution at various scales makes a significant contribution to sustainable soil management (McBratney *et al.,* 2000). Better knowledge about soil classes helps planners and policy makers to make informed decisions on soil management, including cultivation planning and the design of drainage systems (Pontes *et al.,* 2009). Traditionally, soil surveys combine a soil surveyor's specialized knowledge, field descriptions, laboratory analysis, and subsequent classification and mapping (Vasques *et al.,* 2014). However, with the increasing demand for precision agriculture, more detailed soil

Correspondence: Z. Shi. E-mail: shizhou@zju.edu.cn
Received 23 October 2017; revised version accepted 11 July 2018

classification maps are needed for decision making (Mouazen *et al.,* 2007; Viscarra Rossel *et al.,* 2016). Consequently, classification maps obtained from traditional soil surveys can no longer meet the growing need for high-resolution mapping; thus a labour-saving and cost-effective method is needed to fill the gap (Arrouays *et al.,* 2014; Viscarra Rossel & Bouma, 2016; Chen *et al.,* 2018). Proximal soil sensing techniques such as visible near-infrared spectroscopy (vis–NIR) might be an aid to more detailed automated soil surveys.

The vis–NIR technology can assist characterization of soil effectively, and its measurements have the advantage of being rapid, non-destructive, labour saving and inexpensive (Stenberg *et al.,* 2010; Li *et al.,* 2015; Nocita *et al.,* 2015; Ji *et al.,* 2016). Furthermore, various soil physicochemical properties can be predicted simultaneously (Chang & Laird, 2002; Ji *et al.,* 2014; Xu *et al.,* 2018). Soil information is extracted from characteristic absorption peaks at specific wavelengths of electromagnetic radiation using chemometrics. In this way, vis–NIR spectra have been used for the prediction of many soil properties, including soil organic carbon (SOC), colour, clay, pH and other soil survey related macro- and micro-constituents at various scales (Viscarra Rossel *et al.,* 2006b; Stenberg *et al.,* 2010; Shi *et al.,* 2014; Ji *et al.,* 2015; Chen *et al.,* 2016; Jia *et al.,* 2017).

It is a major challenge to combine spectral information from soil profiles when classifying soils with spectroscopic techniques. Commonly, the spectral response of soil is evaluated only at a given depth or depth by depth, which would result in an incomplete interpretation because most soil classification systems are based on multiple horizons (Vasques *et al.,* 2014). Viscarra Rossel & Webster (2011) found that vis–NIR techniques could be used to discriminate soil classes by averaging the spectra of topsoil and subsoil horizons under the Australian soil classification system. Vasques *et al.* (2014) successfully performed soil classification by combining vis–NIR spectral data from three depth intervals (0–20, 40–60 and 80–100 cm) under the Brazilian soil classification system. However, using soil data from fixed depths might not be optimal for soil classification for vis–NIR because epipedons (diagnostic horizons) are used for soil classification in most countries (e.g. USA, Brazil and China). In China, a soil class is determined by diagnostic horizons and diagnostic characteristics that are specified in the Chinese Soil Taxonomy (CST, Shi *et al.,* 2006). Diagnostic horizons refer to generalized soil interfaces such as: A horizon, AB horizon, human-cultivation-related siltigic epipedon, cumulic epipedon and so on; diagnostic subsurface horizons, which include B and E horizons, refer to soil horizons that were formed by transport, eluviation or illuviation under surface horizons. Therefore, soil genetic horizons should possibly be of more value for soil classification than soil data gained from fixed depths.

The main challenge for determining soil classes by vis–NIR data from soil genetic horizons lies in the different numbers of such horizons in profiles. A large percentage of soil profiles have A, B and C horizons, whereas some Cambosols have only A and B and some Primosols have only A and C horizons. Therefore, developing a new approach to classify soils by merging spectral data from the various genetic horizons within one profile is needed for the rapid determination of soil class.

The support vector machine (SVM) algorithm is a data mining approach for classification as well as regression (Vapnik, 1995), which is based on the structural risk minimization principle and can overcome over-fitting problems. The SVM has been used for determining soil classes in recent decades, but the main focus has been on topsoil (Kovačević *et al.,* 2010; Brungard *et al.,* 2015; Lorenzetti *et al.,* 2015; Heung *et al.,* 2016). We aimed to extend the SVM and propose a method named 'multiple objectives mixed support vector classification' (MOM–SVC) in order to mix all classification results of multiple horizons (multiple objectives) from the individual profiles. We also tested the potential of adding auxiliary soil information including moist soil colour, soil organic matter (SOM) and soil texture for modelling at the soil order and suborder levels.

Materials and methods

Study area

The study was carried out in Zhejiang Province, a southeast coastal region of China (Figure 1). It is located between 27°N–31.5°N and 118°E–123°E and covers an area of more than 105 000 km^2. The elevation ranges from 0 to 1907 m, with an ascending gradient from the southwest to the northeast. With a subtropical monsoon climate, the mean annual precipitation is almost 2000 mm, with about 70% accumulating between May and December (Teng *et al.,* 2014). The mean annual temperature is between 15 °C and 18 °C. Water resources are abundant and water levels are highly variable throughout the year. Forest cover is about 54.6% in Zhejiang, and evergreen broad-leaf trees are the dominant vegetation. The region has been cultivated with rice for thousands of years and has been gradually expanding because of the creation of polders on the seashore. In the study area, soils have developed mainly from residual, water-transported and wind-transported parent materials. According to the CST, Cambosols, Anthrosols, Primosols, Argosols and Ferrosols are the dominant soil orders in Zhejiang; they cover more than 96% of the total area. Their equivalent soil classes under the World Reference Base (WRB) soil classification system are listed in Table 1 (Gong & Zhang, 2006). Cambosols and Anthrosols are widely distributed in the study area and can be found in almost 55% of it. Primosols cover 16.3% of the study area and are mainly on alluvial plains and the highlands. Argosols and Ferrosols typically occur in the highlands.

Soil sampling and classification

Based on the spatial distribution, area proportions of soil classes from the Second National Soil Survey of China and the knowledge of soil experts, a total of 130 soil profiles were visited in the study area (Table 1). Soil samples were taken from soil genetic horizons (A, B or C). Prior to dispatch to the laboratory for physicochemical analysis, soil sample moist colour was recorded using the Munsell system. Clay, silt and sand were determined by the pipette method.

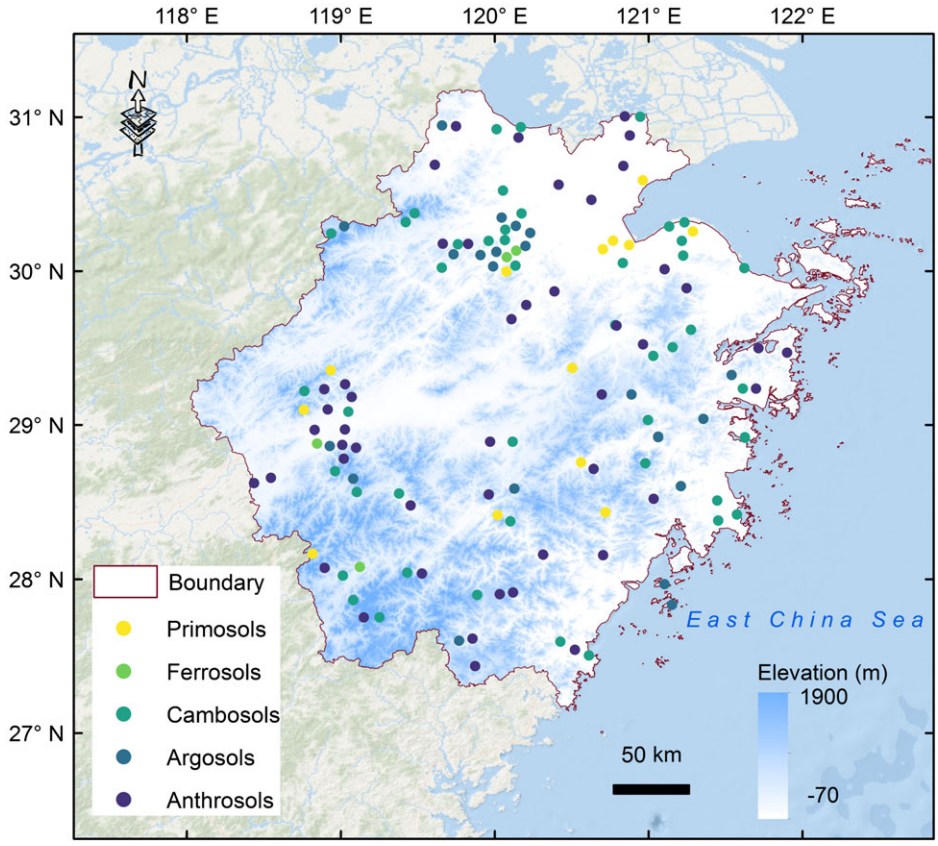

Figure 1 Locations of the study area and of the legacy measured soil profiles. Soil orders are classified using the Chinese Soil Taxonomy scheme. [Colour figure can be viewed at wileyonlinelibrary.com].

Table 1 Correlation between the Chinese Soil Taxonomy and World Reference Base (adapted from Gong & Zhang, 2006)

Chinese soil taxonomy	World reference base
Cambosols	Cambisols
Anthrosols	Anthrosols
Primosols	Fluvisols, Leptisols
Argosols	Luvisols
Ferrosols	Acrisols

Table 2 Soil orders and suborders classified for 130 soil profiles according to the Chinese Soil Taxonomy (CST)

Soil order	N_1[a]/N_2[b]	Soil suborder	N_1/N_2
Anthrosols (An)	47/113	Orthic Anthrosols (OrAn)	2/6
		Stagic Anthrosols (StAn)	45/107
Argosols (Ar)	21/46	Udic Argosols (UdAr)	21/46
Cambosols (Ca)	45/109	Aquic Cambosols (AqCa)	16/45
		Perudic Cambosols (PeCa)	6/14
		Udic Cambosols (UdCa)	23/50
Ferrosols (Fe)	4/10	Udic Ferrosols (UdFe)	4/10
Primosols (Pr)	13/24	Alluvic Primosols (AlPr)	6/12
		Anthric Primosols (AnPr)	1/2
		Orthic Primosols (OrPr)	6/10

[a]N_1 is the number of soil profiles.
[b]N_2 is the number of genetic horizons in soil profiles.

The content of SOM was determined by the H_2SO_4–$K_2Cr_2O_7$ oxidation method at 180 °C for 5 minutes. All the soil profiles were classified by soil experts according to the CST (Table 2).

To ensure that calibration data were covered fully, stratified random sampling was performed on the soil suborder levels by setting the ratio of calibration profiles to validation profiles at ~2:1. There was only one soil profile in the suborder Anthric Primosols and two profiles in Orthic Anthrosols; therefore, both were allocated to the calibration dataset. Finally, the whole dataset was divided into 89 calibration profiles and 41 validation profiles. To quantify the uncertainty induced by the random sampling procedure, we repeated the stratified random sampling 100 times. Finally, 100 classification models were obtained from calibration datasets and 100 corresponding validation datasets were used to evaluate model performance independently.

Spectroscopic measurements and pre-processing

All soil samples were air-dried, ground and sieved to less than 2 mm. Soil vis–NIR spectra were then measured using a

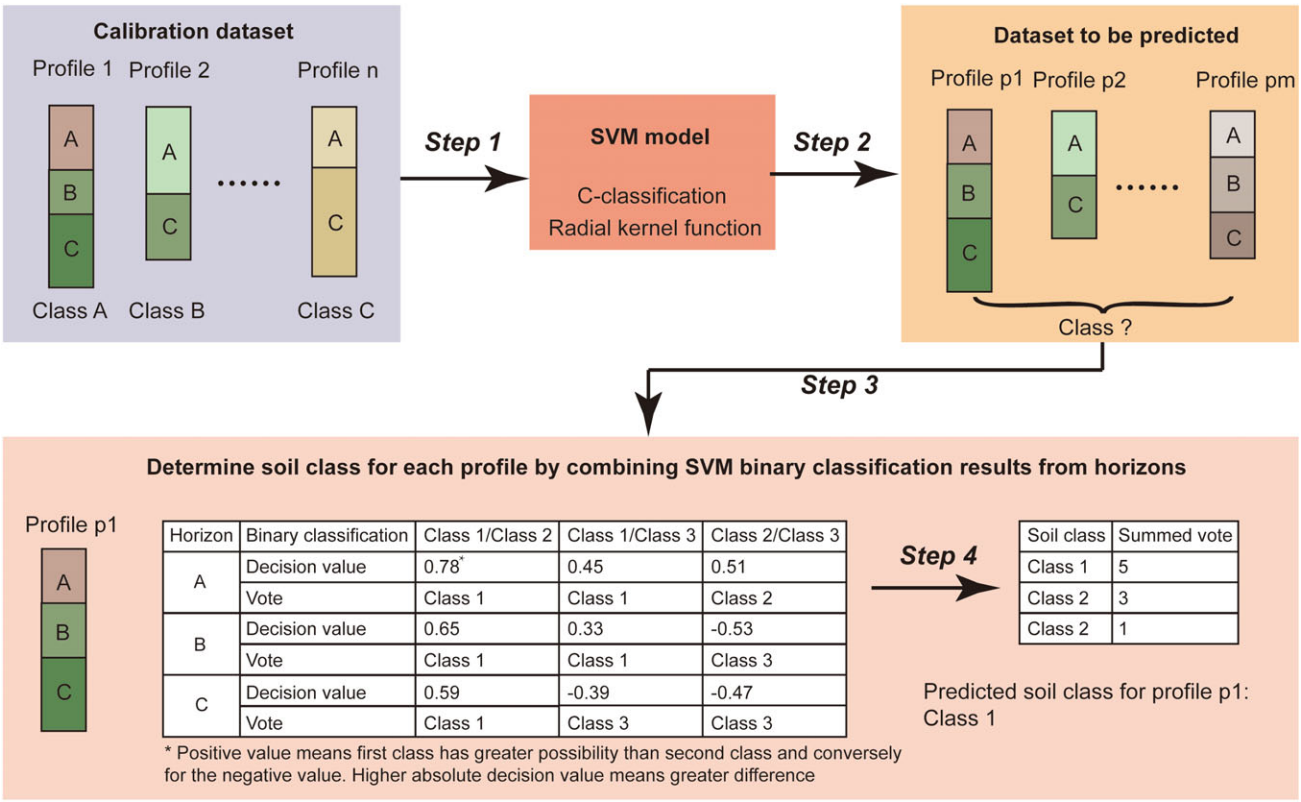

Figure 2 Workflow of the multiple objectives mixed support vector classification (MOM–SVC) model. [Colour figure can be viewed at wileyonlinelibrary.com].

FieldSpec 3 Spectrometer with a high-intensity contact probe (Analytical Spectral Devices Inc., Boulder, CO, USA). The instrument has a spectral range of between 350 and 2500 nm and a resolution of 3 nm at 700 nm and 10 nm at 1400 nm and 2100 nm with a sampling resolution of 1 nm. A Spectralon panel with 99% reflectance was used as a white reference for each measurement. For each soil sample, 10 internal replicated spectra were averaged to provide one representative spectrum, with minimized noise and maximized signal-to-noise ratio.

Spectra were reduced to 400–2450 nm (2051 bands) to eliminate noise at their edges. The spectral data were then smoothed using the Savitzky & Golay (1964) algorithm with a window size of 11 and polynomial of order 2.

Soil colour transformation

The Munsell soil colour data, which cannot be used directly in the regression models, were transformed to RGB (red, green, blue) with the method described by Viscarra Rossel *et al.* (2006a) and using the `munsell2rgb` function in the package `aqp` of R 3.1.3 (Beaudette *et al.,* 2013; R Core Team, 2014).

Multiple objectives mixed support vector classification

In this study, multi-classification problems have been solved by SVMs using 'one-versus-one'. Suppose that we have m classes and that m is greater than 2. In the one-versus-one approach, SVM train $\frac{m(m-1)}{2}$ binary classifiers based on each pair of classes. Each binary classifier gives a 'vote' to the more likely class in a pair of classes by decision values, and the majority voting method will determine the final class.

Figure 2 demonstrates the workflow of MOM–SVC, which is an extension of SVM. We prepared an SVM model using spectral data, SOM and soil texture for all soil horizons in all the profiles of the calibration set. We then applied the resulting classification model to each horizon in each profile in the prediction set. The classification of the profile was then determined from the most frequent horizon classification within the profile. Where this was not possible (i.e. there were two or more most frequent horizon classifications) the decision value within these classes was revisited to examine the SVM binary decision value ($\mathrm{dv} \in \{-1, 1\}$). If the value, say, between class one and class two (c_1-vs-c_2), was positive then the profile class was allocated a unit vote in favour of the class, c_1, or else it was allocated in favour of c_2. A larger |dv| indicates a greater distinction between c_1 and c_2; therefore, the largest |dv| from the most heavily weighted or voted soil classes determined the final soil class for the test profile.

The SVM modelling was performed with package e1071 in R 3.1.3 (Dimitriadou *et al.,* 2005; R Core Team, 2014) using C-classification with a radial kernel function. The parameters 'gamma' and 'cost' were optimized automatically by grid searching

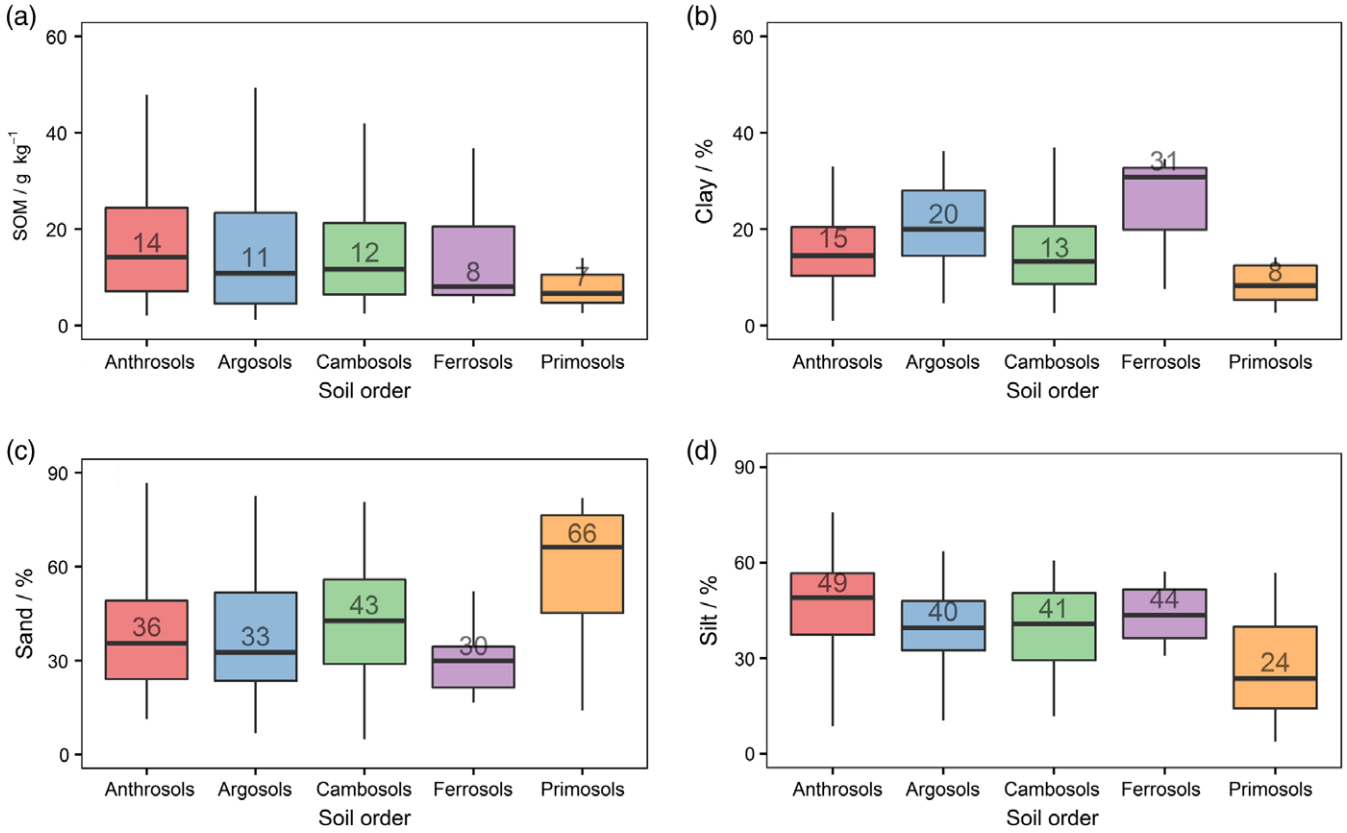

Figure 3 Boxplots of (a) soil organic matter (SOM), (b) clay, (c) silt and (d) sand for five soil orders. The median values of each soil order are indicated by the horizontal black lines and the numbers inside each box. [Colour figure can be viewed at wileyonlinelibrary.com].

(smallest RMSE) with 10-fold cross-validation. The settings of the search grid for 'gamma' were 2^{-3}, 2^{-2}, 2^{-1} and 2^{0}, whereas for 'cost' they were 2^{-1}, 2^{0}, 2^{1} and 2^{2}. To assess model uncertainty, the mean value and 90% confidence intervals ($CIs_{90\%}$) of producer accuracy were evaluated for both calibration and validation performances from 100 sets of stratified random sampling.

Results and discussions

Soil properties and vis–NIR spectral characteristics

Considerable variation in SOM, clay, silt and sand was observed among the five soil orders (Figure 3). Anthrosols had the largest median SOM contents (14 g kg^{-1}) as well as a large interquartile range (IQR) because of land management practices (e.g. fertilization, liming, irrigation and organic amendments) and their operational differences among farmers (Pan *et al.,* 2004; Chenu *et al.,* 2018). Anthrosols had large median silt content (49%) and moderate median clay (15%) and sand contents (36%). Argosols had a similar distribution of SOM to Anthrosols, whereas their clay content (20%) was much larger because their formation was accompanied by the leaching of water-soluble salts and the subsequent formation of clay minerals. Ferrosols usually occur in tropical and subtropical zones under high temperature, high precipitation and adequate drainage conditions. With considerable leaching of both silicic acid and base ions, iron and aluminum oxides were concentrated in Ferrosols, which had the largest clay content (31%) but relatively small SOM content (8 g kg^{-1}). Primosols and Cambosols were both less developed; the largest difference between them is that Primosols do not have diagnostic horizons or features, whereas Cambosols have cambic horizons. In comparison to other soil orders, Primosols have the smallest SOM (7 g kg^{-1}), clay (8%) and silt (24%) contents and the largest sand content (66%).

Figure 4 shows the spatial distribution of soil colour in RGB colour space. After transformation into this form, moist soil colour was clearly discernible. It is associated with soil composition and properties (e.g. humus, water content, iron oxide and silicon dioxide); therefore, it is a crucial reference for discriminating soil classes using horizons and diagnostic characteristics in the CST. Moist soil colour from the field provides a contribution to soil classification independent of the colour obtained from the vis–NIR spectra of processed samples in the laboratory.

Figure 5 presents field conditions, ground soil samples from two profiles of Anthrosols and Cambosols, and their corresponding vis–NIR spectra shown for three genetic soil horizons. Anthrosols had a distinct reflection peak between 750 and 800 nm in B and C horizons, which were mainly dominated by iron oxides (Stenberg *et al.,* 2010). Zhejiang province is in the lower reach area of the Yangtze River delta and it has a long history of rice cultivation;

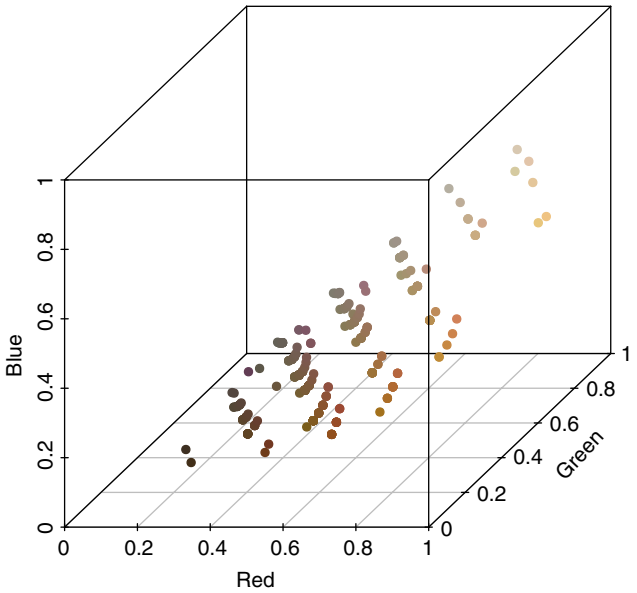

Figure 4 Soil RGB colour space. [Colour figure can be viewed at wileyonlinelibrary.com].

therefore, paddy soil accounts for most of the Anthrosols in the study area. Because these soils have been subjected to long-term dry–wet alternation, hydrous iron oxides have accumulated in the B horizon of Anthrosols. Iron oxide absorption was also found in the C horizon, which has resulted mainly from leaching from the B horizon. Cambosols had similar reflectance curves in the B and C horizons (Figure 5) because they are not yet developed sufficiently to have a large difference between them. The A horizon in Cambosols was greatly affected by biotic activities; therefore, it had more SOM than the other two horizons.

Classification performance at soil order level

The calibration accuracy of soil classification with vis–NIR as input to MOM–SVC provided an overall estimate of 0.85 (Figure 6) with lower and upper 90% confidence intervals at 0.83 and 0.87, respectively. In the CST, a series of soil properties (e.g. soil organic matter, iron-aluminum oxides, silicon dioxide, silicate, humus, soil texture, calcium carbonate, gypsum, soluble salts and pH) have been taken into consideration at the soil order level. The MOM–SVC resulted in large and stable accuracies of 0.95 (0.94, 0.97) and 0.97 (0.97, 0.97) for Anthrosols and Cambosols, with a good accuracy of 0.79 (0.79, 0.79) for Argosols. The accuracy for Ferrosols was moderate at 0.67 (0.67, 0.67) and Primosols had the smallest and most variable classification accuracy of 0.18 (0, 0.33).

The overall accuracy was 0.57 (0.54, 0.59) for a total of 41 profiles in the validation procedure using the MOM–SVC method. Cambosols and Anthrosols were classified correctly with large mean accuracies at 0.70 and 0.80, respectively, whereas mean accuracy (<0.20) was small for Argosols and Primosols. There was only one profile for the validation of Ferrosols; thus, they were always misclassified and had the smallest accuracy of 0. In contrast, Cambosols and Anthrosols had good accuracy because there were many training profiles in the calibration dataset. For Ferrosols, the training profiles accounted for only a small part of the calibration dataset. The poor predictive accuracy for Primosols was a result of its small calibration accuracy.

When information on available soil properties was added (including moist soil colour, soil organic matter and soil texture) in modelling, MOM–SVC performed better in both calibration and validation procedures (Figure 7) than the calibration results modelled with soil spectra alone. Including additional soil properties into classification slightly improved the calibration accuracy for Anthrosols, Arogosols and Primosols, but reduced the accuracy for Ferrosols. The improvement in accuracy for Argosols and Primosols probably resulted from the additional information about clay and soil organic matter, which are diagnostic characteristics for these soils. The overall accuracy improved from 0.57 to 0.68 in the validation procedure. A large increase in accuracy was observed for Cambosols, which improved from 0.70 to 0.93. Predictions for Anthrosols and Primosols were also better. The predictive accuracy for Ferrosols and Argosols remained the same. These improvements indicated that including auxiliary soil information in classification models could improve classification accuracy at the soil order level. The improvement in accuracy for Anthrosols might be because of the fact that several diagnostic horizons in Anthrosols are related to moist soil colour and SOM. For example, the SOC content in a fimic epipedon should be larger than 6 g kg^{-1}, and for a waterlogged anthrostagnic epipedon, the moist soil colour is as follows: hue ≤4, value ≤2 and chroma more yellow than 7.5 years.

For comparison, similar model accuracies for calibration (0.89) and validation (0.67) at the soil order level were obtained for Brazilian soil using vis–NIR spectra from multiple depths (Vasques *et al.,* 2014). The authors used 20 principal components from 630 reflectance bands and multinomial logistic regression.

Classification performance at the suborder level

Figure 8 illustrates the prediction accuracy of soil profiles using vis–NIR spectra at the soil suborder level. The MOM–SVC achieved overall accuracies of 0.81 (0.79, 0.83) in the calibration data and of 0.55 (0.49, 0.59) in the validation data, both of which were slightly smaller than that at the soil order level.

For the subsoil, more detailed diagnostic information was taken into consideration for soil classification, which made the classification more complex and difficult. Taking Cambosols for example, the largest difference between Aquic, Perudic and Udic Cambosols is the soil moisture regime, which controls moist soil colour, the strength of redox and the transport of soluble ions. Therefore, at the subsoil order, misclassification occurred within soil orders: in calibration, all Orthic Anthrosols were misclassified as Stagic Anthrosols; in validation, 18% of Udic Cambosols were misclassified as Aquic Cambosols. Sometimes, there might be strong interference from the soil moisture regime in classification: in validation, 71% Udic Argosols were misclassified as Udic Cambosols whereas

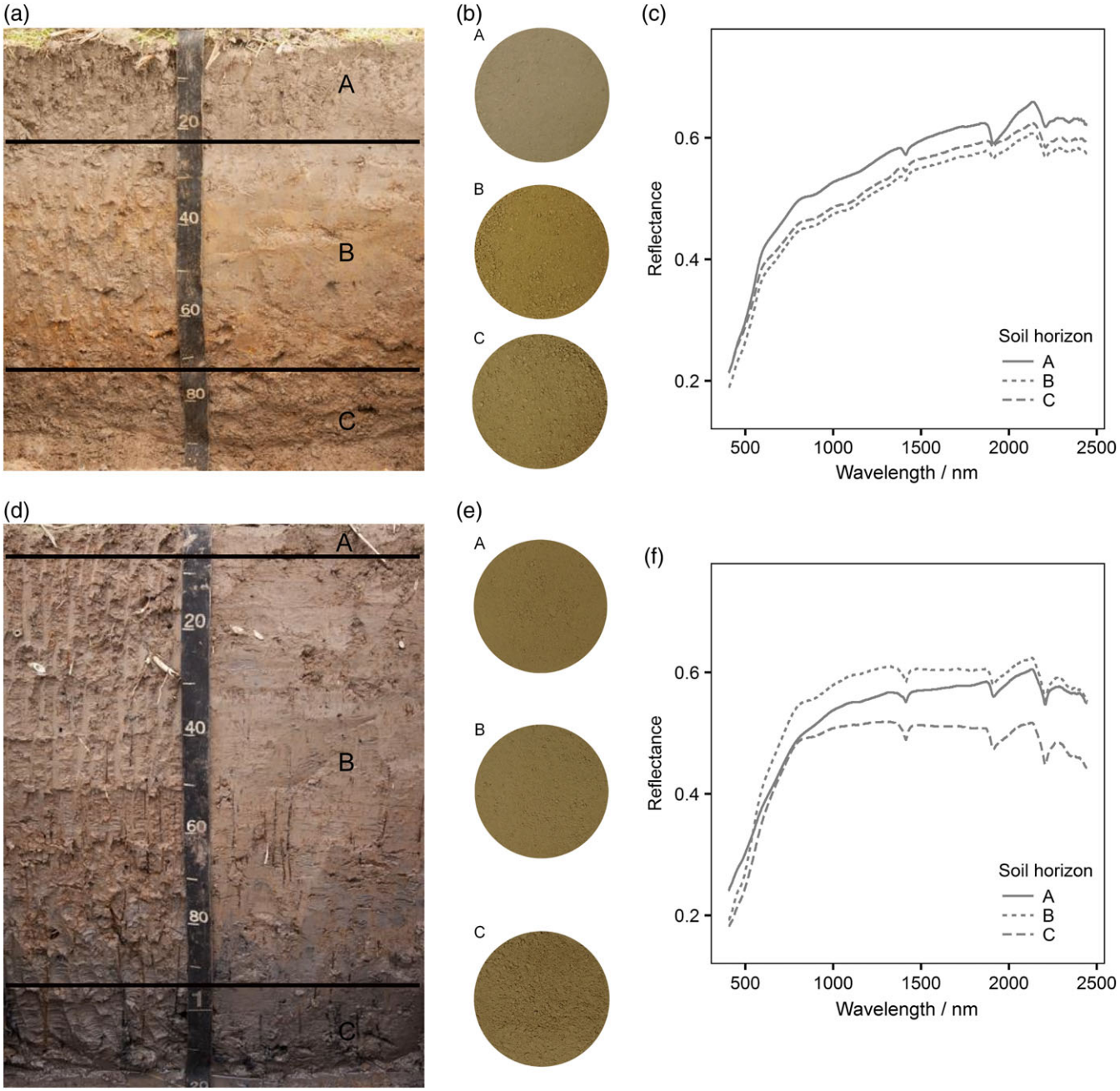

Figure 5 Images of (a, d) soil profiles of Anthrosols (upper) and Cambosols (lower), their corresponding ground soil samples (b, e) for each vertical profile and soil vis–NIR spectra in three soil horizons for each profile (c, f). [Colour figure can be viewed at wileyonlinelibrary.com].

18% Udic Cambosols and 100% Udic Ferrosols were incorrectly classified as Udic Argosols.

Figure 9 shows the model accuracy when auxiliary soil information was included in the classification at the soil suborder. It indicates that there was a small improvement with auxiliary soil information in both calibration (0.91 (0.90, 0.92)) and validation (0.61 (0.59, 0.64)) data for soil suborders. In comparison with the validation accuracy in Figure 7, performance was better or equivalent for almost all the soil suborders, except for Stagic Anthrosols. Misclassifications were also observed within the same soil order or the same soil moisture regime of different soil orders. The result also shows that adding auxiliary soil information into soil classification is of more help at the soil order level than at the suborder level because the differences in accuracy between classification with and without auxiliary soil information were 0.11 and 0.06 for the soil order and suborder, respectively.

The performance of our validation results at the soil suborder level was better than the independent validation (0.48) from the

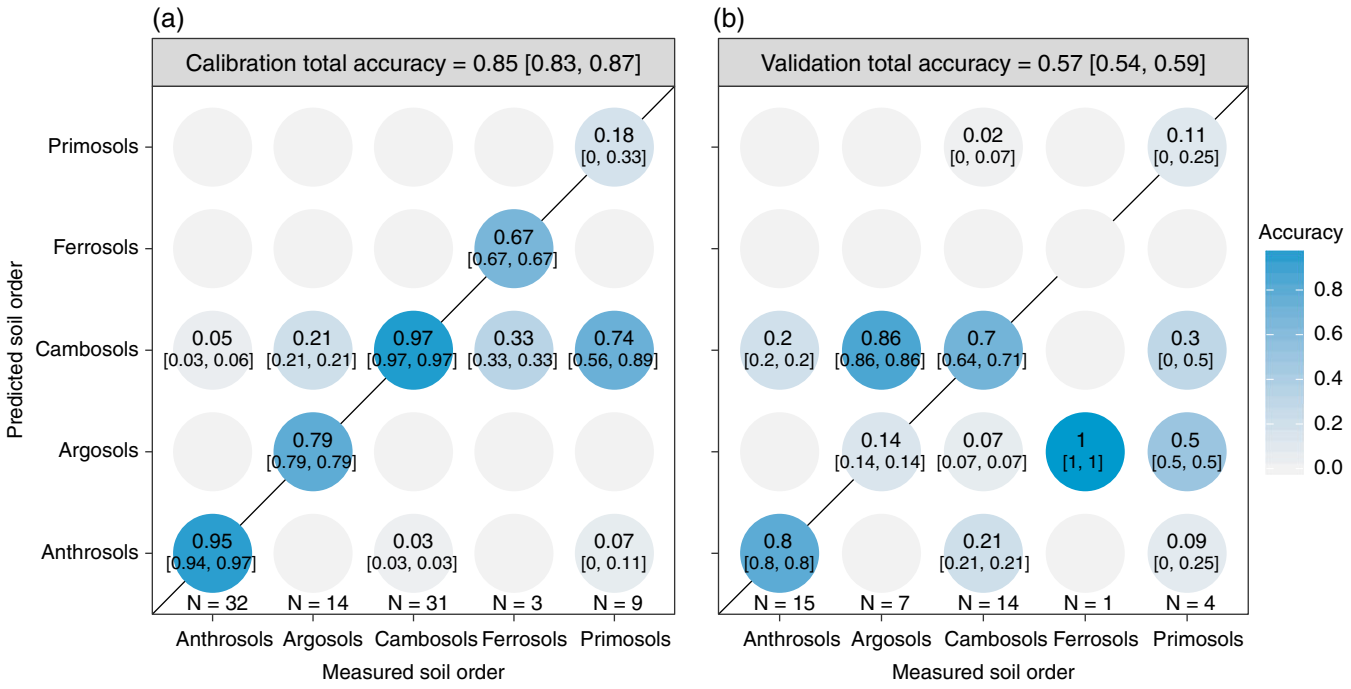

Figure 6 Accuracy matrices of the soil order classification in profiles using vis–NIR spectra in (a) calibration and (b) validation. Mean values are presented outside the square brackets, and lower and upper limits of the 90% confidence intervals are shown inside them. [Colour figure can be viewed at wileyonlinelibrary.com].

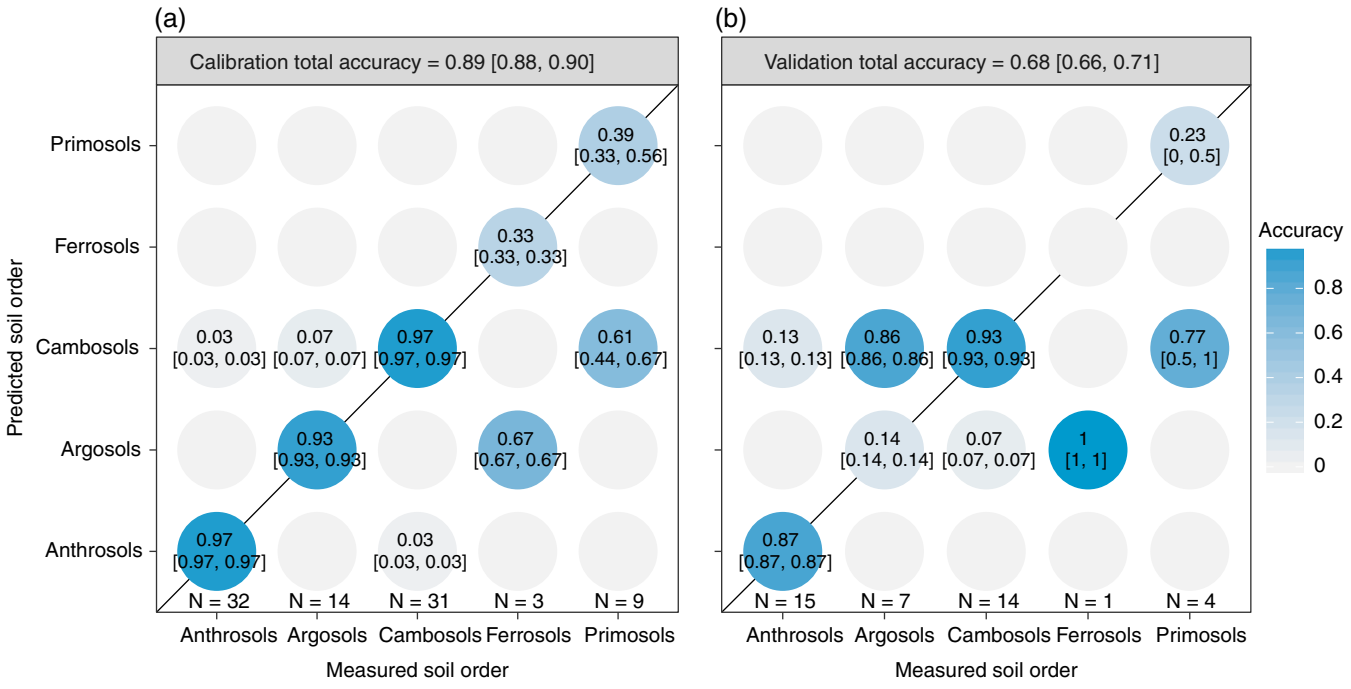

Figure 7 Accuracy matrices of the soil order classification in profiles using vis–NIR spectra and available soil properties in (a) calibration and (b) validation. Mean values are presented outside the square brackets, and lower and upper limits of the 90% confidence intervals are shown inside them. [Colour figure can be viewed at wileyonlinelibrary.com].

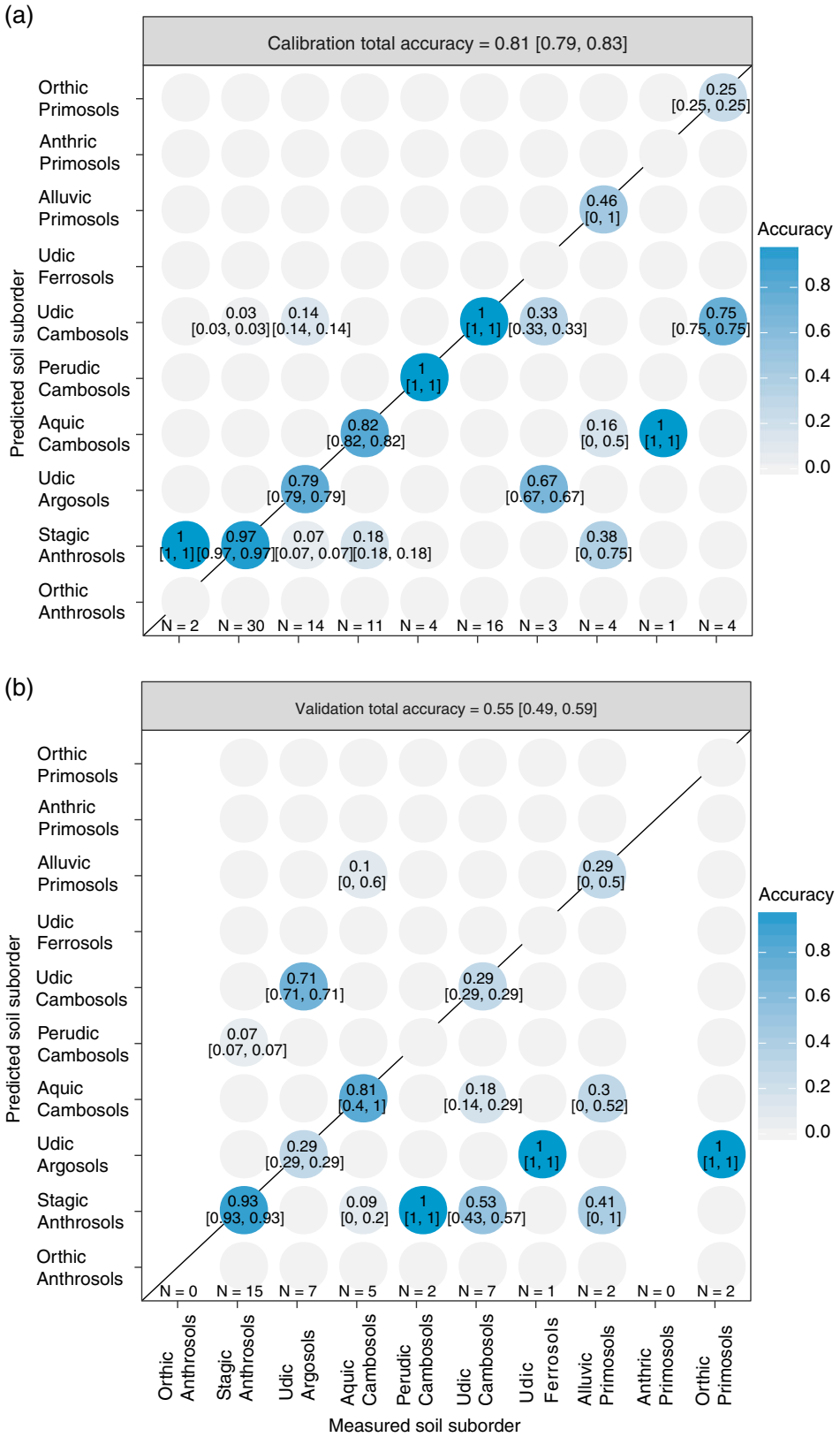

Figure 8 Accuracy matrices of the soil suborder classification using vis–NIR spectra in (a) calibration and (b) validation. Mean values are presented outside the square brackets, and lower and upper limits of the 90% confidence intervals are shown inside them. [Colour figure can be viewed at wileyonlinelibrary.com].

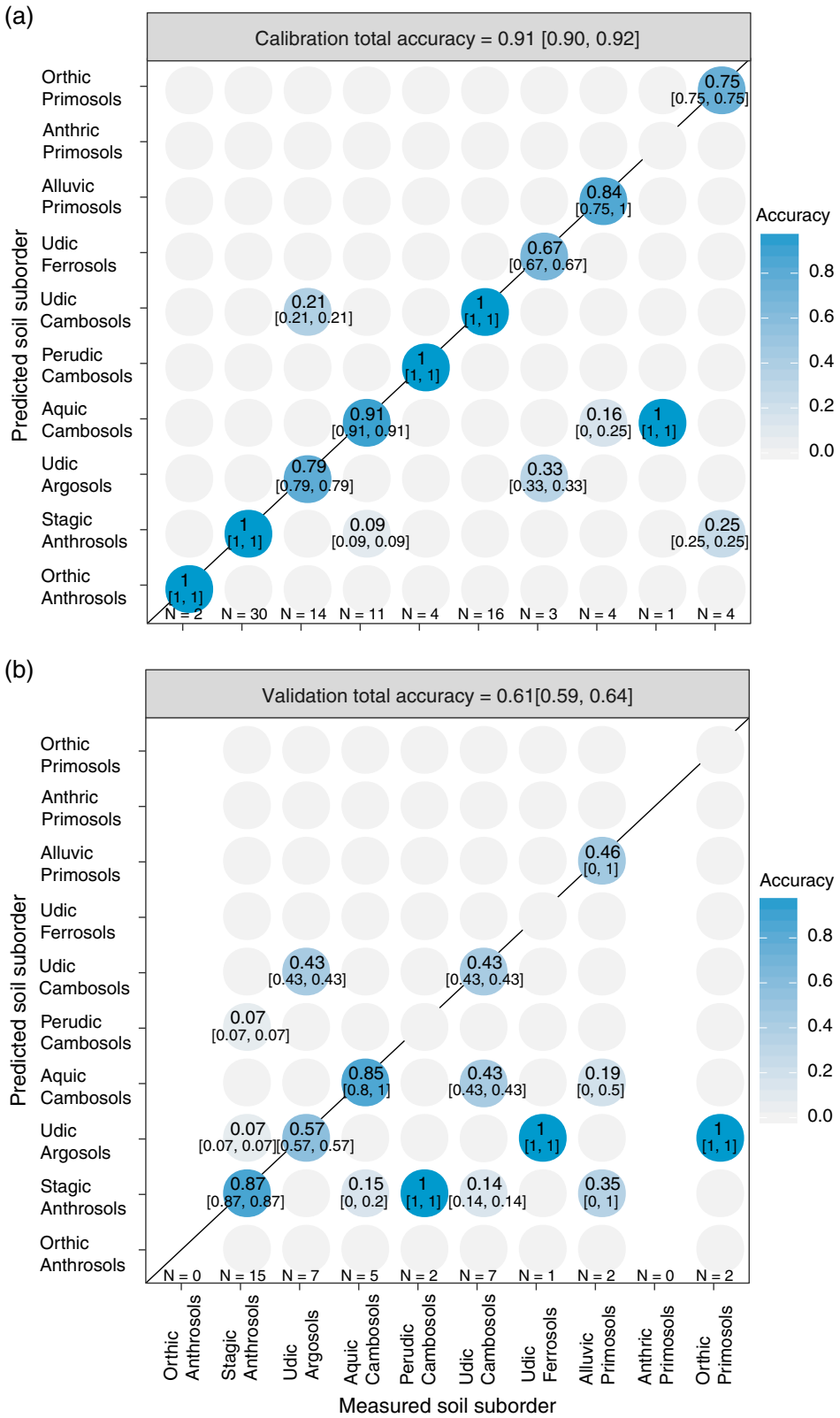

Figure 9 Accuracy matrices of the soil suborder classification using vis–NIR spectra and available soil properties in (a) calibration and (b) validation. Mean values are presented outside the square brackets, and lower and upper limits of the 90% confidence intervals are shown inside them. [Colour figure can be viewed at wileyonlinelibrary.com].

study by Vasques *et al.* (2014), whereas it was worse than the leave-one-out cross-validation (0.70–0.76) from the work of Zeng *et al.* (2016), in which vis–NIR spectra were used to classify soils by multinomial logistic regression from Anhui province, east China. The differences in performance might result from the differences in data sources, modelling approaches and validation schemes.

Our results demonstrate that the MOM–SVC method has a good ability to classify soil profiles with vis–NIR spectroscopy even when the profiles had a different number of horizons. We are not proposing that this is as yet a replacement for laboratory analysis and expert knowledge, but vis–NIR spectroscopy offers a new possibility for rapid soil classification based on legacy soil data and therefore more detailed soil class maps. Rizzo *et al.* (2016) and Teng *et al.* (2018) have already demonstrated the potential use of vis–NIR spectra in updating soil class maps from a local to national scale. Furthermore, the vis–NIR technique makes it possible to update soil databases rapidly when a more objective soil classification system is developed. Developing advanced automated proximal soil sensing platforms such as the Soil Condition Analysis System (Viscarra Rossel *et al.,* 2017) would enable us to collect more vertical measurements directly in the field at a fine depth resolution. The MOM–SVC is flexible and can deal with multi-depth data; therefore, more robust prediction results could be obtained by this method if more detailed soil profile information was available.

Conclusions

The proposed MOM–SVC method was able to model a soil database that included profiles with different numbers of genetic horizons, and it performed well in predicting soil classes of soil profiles using vis–NIR spectra. Predictive accuracy was much better at the soil order level (classification accuracy from 0.57 to 0.68) than the suborder level (classification accuracy from 0.55 to 0.61). Our results also showed that inadequate calibration data at each soil suborder is a reason for its poorer classification accuracy at the suborder level; therefore, more calibration data would be needed for a more robust soil classification. Easily available soil details, such as moist soil colour, SOM and soil texture, are important diagnostic characteristics in soil classification and including this information in the classification model improved accuracy at the soil order level, although it showed less improvement at the suborder level. More auxiliary information might be needed for better classification models at the soil suborder level.

We conclude, therefore, that, based on soil legacy data, vis–NIR spectroscopy would be a useful auxiliary technique for the rapid determination of soil classes and thus could make a contribution to updating soil classification maps at regional and even larger scales. There is a strong possibility that B horizons (the product of pedogenesis) are more important than A and C horizons in soil classification; therefore, further work should be focused on whether it might be better to assign different weights to each genetic horizon based upon expert knowledge than to apply equal weights as in the MOM–SVC method.

Acknowledgements

The National Key Research and Development Program (2017YFD0700501), and the Research Fund of State Key Laboratory of Soil and Sustainable Agriculture, Nanjing Institute of Soil Science, Chinese Academy of Sciences (No Y412201430), supported this work. Songchao Chen received the support of the China Scholarship Council for 3 years' PhD study in INRA and Argocampus Ouest (under grant agreement no 201606320211).

References

Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B.M., Hong, S.Y. *et al.* 2014. GlobalSoilMap toward a fine-resolution global grid of soil properties. *Advances in Agronomy*, **125**, 93–134.

Beaudette, D.E., Roudier, P. & O'Geen, A.T. 2013. Algorithms for quantitative pedology: a toolkit for soil scientists. *Computers & Geosciences*, **52**, 258–268.

Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A. & Edwards, T.C. 2015. Machine learning for predicting soil classes in three semi-arid landscapes. *Geoderma*, **239**, 68–83.

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

Chen, S., Peng, J., Ji, W., Zhou, Y., He, J. & Shi, Z. 2016. Study on the characterization of VNIR-MIR spectra and prediction of soil organic matter in paddy soil. *Spectroscopy and Spectral Analysis*, **36**, 1712–1716.

Chen, S., Martin, M.P., Saby, N.P., Walter, C., Angers, D.A. & Arrouays, D. 2018. Fine resolution map of top- and subsoil carbon sequestration potential in France. *Science of the Total Environment*, **630**, 389–400.

Chenu, C., Angers, D.A., Barré, P., Derrien, D., Arrouays, D. & Balesdent, J. 2018. Increasing organic stocks in agricultural soils: knowledge gaps and potential innovations. *Soil & Tillage Research*. https://doi.org/10.1016/j.still.2018.04.011.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D. & Weingessel, A. 2005. *e1071: Misc Functions of the Department of Statistics (e1071)*, *TU Wien*. R package version, 1–5.

Gong, Z. & Zhang, G. 2006. Classification systems: Chinese. In: *Encyclopedia of Soil Science, Volume 1* (ed. R. Lal), pp. 245–246. CRC Press, Boca Raton, FL.

Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E. & Schmidt, M.G. 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. *Geoderma*, **265**, 62–77.

Ji, W., Shi, Z., Huang, J. & Li, S. 2014. In situ measurement of some soil properties in paddy soil using visible and near-infrared spectroscopy. *PLoS One*, **9**, e105708.

Ji, W., Viscarra Rossel, R.A. & Shi, Z. 2015. Accounting for the effects of water and the environment on proximally sensed vis–NIR soil spectra and their calibrations. *European Journal of Soil Science*, **66**, 555–565.

Ji, W., Li, S., Chen, S., Shi, Z., Viscarra Rossel, R.A. & Mouazen, A.M. 2016. Prediction of soil attributes using the Chinese soil spectral library and standardized spectra recorded at field conditions. *Soil & Tillage Research*, **155**, 492–500.

Jia, X., Chen, S., Yang, Y., Zhou, L., Yu, W. & Shi, Z. 2017. Organic carbon prediction in soil cores using VNIR and MIR techniques in an alpine landscape. *Scientific Reports*, **7**, 2144.

Kovačević, M., Bajat, B. & Gajić, B. 2010. Soil type classification and estimation of soil properties using support vector machines. *Geoderma*, **154**, 340–347.

Li, S., Shi, Z., Chen, S., Ji, W., Zhou, L., Yu, W. *et al.* 2015. In situ measurements of organic carbon in soil profiles using vis-NIR spectroscopy on the Qinghai–Tibet plateau. *Environmental Science & Technology*, **49**, 4980–4987.

Lorenzetti, R., Barbetti, R., Fantappiè, M., L'Abate, G. & Costantini, E.A. 2015. Comparing data mining and deterministic pedology to assess the frequency of WRB reference soil groups in the legend of small scale maps. *Geoderma*, **237**, 237–245.

McBratney, A., Odeh, I., Bishop, T., Dunbar, M. & Shatar, T. 2000. An overview of pedometric techniques for use in soil survey. *Geoderma*, **97**, 293–327.

Mouazen, A.M., Maleki, M.R., Baerdemaeker, J. & Ramon, H. 2007. On-line measurement of some selected soil properties using a VIS–NIR sensor. *Soil & Tillage Research*, **93**, 13–27.

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B. *et al.* 2015. Soil spectroscopy: an alternative to wet chemistry for soil monitoring. *Advances in Agronomy*, **132**, 139–159.

Pan, G., Li, L., Wu, L. & Zhang, X. 2004. Storage and sequestration potential of topsoil organic carbon in China's paddy soils. *Global Change Biology*, **10**, 79–92.

Pontes, M., Cortez, J., Galvão, R., Pasquini, C., Araújo, M., Coelho, R. *et al.* 2009. Classification of Brazilian soils by using LIBS and variable selection in the wavelet domain. *Analytica Chimica Acta*, **642**, 12–18.

R Core Team 2014. *R: A language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Rizzo, R., Demattê, J., Lepsch, I., Gallo, B. & Fongaro, C. 2016. Digital soil mapping at local scale using a multi-depth vis–NIR spectral library and terrain attributes. *Geoderma*, **274**, 18–27.

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

Shi, X., Yu, D., Yang, G., Wang, H., Sun, W., Du, G. *et al.* 2006. Cross-reference benchmarks for translating the genetic soil classification of China into the Chinese soil taxonomy. *Pedosphere*, **16**, 147–153.

Shi, Z., Wang, Q., Peng, J., Ji, W., Liu, H., Li, X. *et al.* 2014. Development of a national VNIR soil-spectral library for soil classification and prediction of organic matter concentrations. *Science China Earth Sciences*, **57**, 1671–1680.

Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M. & Wetterlind, J. 2010. Visible and near infrared spectroscopy in soil science. *Advances in Agronomy*, **107**, 163–215.

Teng, H., Shi, Z., Ma, Z. & Li, Y. 2014. Estimating spatially downscaled rainfall by regression kriging using TRMM precipitation and elevation in Zhejiang Province, southeast China. *International Journal of Remote Sensing*, **35**, 7775–7794.

Teng, H., Viscarra Rossel, R.A., Shi, Z. & Behrens, T. 2018. Updating a national soil classification with spectroscopic predictions and digital soil mapping. *Catena*, **164**, 125–134.

Vapnik, V. 1995. *The Nature of Statistical Learning Theory*. Springer, New York, NY.

Vasques, G.M., Demattê, J.A.M., Rossel, R., Ramírez-López, L. & Terra, F.S. 2014. Soil classification using visible/near-infrared diffuse reflectance spectra from multiple depths. *Geoderma*, **223**, 73–78.

Viscarra Rossel, R.A. & Bouma, J. 2016. Soil sensing: a new paradigm for agriculture. *Agricultural Systems*, **148**, 71–74.

Viscarra Rossel, R.A. & Webster, R. 2011. Discrimination of Australian soil horizons and classes from their visible–near infrared spectra. *European Journal of Soil Science*, **62**, 637–647.

Viscarra Rossel, R.A., Minasny, B., Roudier, P. & McBratney, A.B. 2006a. Colour space models for soil science. *Geoderma*, **133**, 320–337.

Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J. & Skjemstad, J.O. 2006b. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma*, **131**, 59–75.

Viscarra Rossel, R.A., Behrens, B.-D., Brown, D.J., Demattê, J.A.M., Shepherd, K.D., Shi, Z. *et al.* 2016. A global spectral library to characterize the world's soil. *Earth-Science Reviews*, **155**, 198–230.

Viscarra Rossel, R.A., Lobsey, C.R., Sharman, C., Flick, P. & McLachlan, G. 2017. Novel proximal sensing for monitoring soil organic C stocks and condition. *Environmetal Science & Technology*, **51**, 5630–5641.

Xu, D., Ma, W., Chen, S., Jiang, Q., He, K. & Shi, Z. 2018. Assessment of important soil properties related to Chinese Soil Taxonomy based on vis–NIR reflectance spectroscopy. *Computers and Electronics in Agriculture*, **144**, 1–8.

Zeng, R., Zhang, G., Li, D., Rossiter, D.G. & Zhao, Y. 2016. How well can VNIR spectroscopy distinguish soil classes? *Biosystems Engineering*, **152**, 117–125.