# USE OF NIR SPECTROSCOPY AND LS-SVM MODEL FOR THE DISCRIMINATION OF VARIETIES OF SOIL

Zengfang Li [1] , Jiajia Yu [2] , Yong He [2,*]

[1] *Zhejiang Water Conservancy and Hydropower College,Hangzhou 310018,China*

[2] *College of Biosystems Engineering and Food Science, Zhejiang University,268 Kaixua Road, Han zhou, 310029, China*

[*] *Corresponding author, Address: College of Biosystems Engineering and Food Science,Zhejiang University, 268 Kaixuan Road, Hangzhou, 310029, China , Tel: +86-571-86971143, Fax: +86-571-86971143, Email: yhe@zju.edu.cn*

Abstract: Near-infrared spectroscopy (NIR) combined with chemomatrics method was employed in this paper. The objectives were to investigate the feasibility of using VIS/SWNIR spectroscopy to discriminate soil of different kinds, and to validate the performance of selected sensitive ebands. Spectrums in the NIR regions (4003.563-12496.67 $cm^{-1}$) were collected from 380 samples and the data was expressed as absorbance, the logarithm of the reciprocal of reflectance(log 1/R).240 samples were randomly collected as modeling, and the others were used to check the model's performance. Principal components analysis (PCA) tested the clustering of these four kinds of soil, which made a qualitative analysis for the discrimination, but for the sake of speedup the calculating time, the mathematics analysis of support vector machine classification and 10-folds cross-validation were used to model, and Based on SVM, the recognition ratio of 98% (0.2 as threshold value) was obtained. Compared with the PLS result of 79% (0.2 as threshold value), SVM classification is apparently more effective. At the same time, the sensitive bands of soil varieties were caculated,in which we found the 801-972 nm can predict well with the result of 90% from LS-SVM.The prediction results of 99% indicated that the NIR can mainly represent the characteristics of soil of different kind based on SVM model.

Keywords: NIR; chemomatrics; soil; PCA; SVM

## 1. INTRODUCTION

The prospect of meeting the world's food demand for an additional people in future is a formidable challenge, but it is also an obligatory for our studies. Higher technology is needed, so precision agriculture is one agronomic means of meeting this challenge. The concept of precion agriculture entails the use of some high-tech equipment of assessing field conditons and applying chemicals and fertilizers(Dan Ess et al.1997). Therefore, finding a faster and more efficient technology to discriminate the varieties of soil is a key point in precision farming, whereas the ways to differentiate the soil can not meet the demand of precision farming, which are lower-tech depend on the farmer self, which are too complex ,time-consuming, laborious and may indirect such as measuring soil pH in the field with an ISFET(Viscarra Rossel R A et al.2004) and measuring clay content using EMI instruments(Sudduth K A et al.2001).So it is necessitous to find a faster and cheaper technology to make the breakthrough.

Recently, the visible and short-wave near-infrared spectroscopy (VIS/SWNIR) technique ranging from 400 to 1075 nm, has been widely applied because of its non-destructive feature for biological and biomedical materials,such as discrimination of varieties of apple(He, Y. et al.2006) and other fruits , geographic classification of wine(Liu, L. et al.2006). it is suitable as an excellent detector .. In the VIS/SWNIR region, there are several advantages such as the signal exploitation is reliable, the measurement time is low and the effective of intense water bands in NIR region can be diminished (J. B. Reeves et al.1994). with reliable signal transformation, low time-consuming, and sensitive spectral response of water.

The presence of soil constituents such as iron oxides, organic matter (OM) and water content lead to overlapped peaks spectroscopy. So it show the feasible of NIR technology used in discriminate soil varieties. In addition, there is much noise and other unrelated information arising from overtones and combinations of such vibrations, rendering them much more difficult to interpret. For sake of finding the relevant quantitative information, especially the nonlinear one , it is necessary to use the suitable and effective chemometrics methods. Least-squares support vector machine In this study ,LS-SVM and PCA is used, Least-squares support vector machine (LS-SVM) is an optimized algorithm based on standard SVM by Suykens et al. (J. B. Reeves, et al.1994).

The aim of this study was to explore the feasibility of using NIR spectroscopy to discriminate different varieties of soil. A nonlinear mathematical method for rapid, nondestructive identification of soil was developed by LS-SVM. In order to validate the performance of LS-SVM,PLS is used to be a comparison analysis.

# 2. MATERIALS AND METHODS

## 2.1 Sample preparation

The soils were abtained from different region which belong different varities ( One of the Haining soil belongs to coastal saline soil, and the other Haining soil and Cixi soil belongs to paddy soil Quzhou soil belongs tored soil.) The soil samples were prepared in same capacity. for each ,and the sample number for each variety was 100(totally 400 samples).They were stored in the lab with a constant temperature of 25±1 ºC to equalize the temperature. 300 samples were randomly selected for the calibration set, while the remaining 100 samples for the prediction set.

## 2.2 Spectra measurement

For each sample, reflectance spectra were scanned by a handheld FieldSpec Pro FR (325–1075 nm)/A110070, Trademarks of Analytical Spectral Devices, Inc. (Analytical Spectral Devices, Boulder, USA) for three times each. The light source consists of a Lowell pro-lam interior light source assemble/128930 with Lowell pro-lam 14.5V Bulb/128690 tungsten halogen bulb that could be used both in visible and near infrared region. The field-of-view (FOV) of the spectroradiometer is 10°. The spectroradiometer was placed at a height of approximately 250 mm and 45° angle away from the center of sample. The light source was placed at a height of approximately 150 mm 45° angle away from the sample. The spectrum of each sample was the average of 30 successive scans with 1.5 nm intervals. Three spectra were collected for each sample and the average spectrum of these two measurements was used in the later analysis. All spectral data were stored in a computer and processed using the RS3 software for Windows (Analytical Spectral Devices, Boulder, USA) designed with a Graphical User Interface.

## 2.3 Spectral data pretreatment

The reflectance spectra were firstly transformed into ASCII format by using the ASD ViewSpecPro software (Analytical Spectral Devices, Boulder, USA). Then three spectra for each sample were averaged into one spectrum and transformed by log (1/T) into absorbance spectrum. The pretreatments were implemented by “The Unscrambler V 9.7” (CAMO PROCESS AS, OSLO, Norway). Because of the high noise at the beginning of the spectrum, 421-1075 nm was used in this study.After some trial computations, the

optimal smoothing way of moving average with 3 segments was applied to decrease the noise. Standard normal variate (SNV) was applied for light scatter correction and reducing the changes of light path length.

## 2.4 Least squares support vector machines (LS-SVM)

Support Vector Machines (SVM) is a powerful methodology for solving problems in nonlinear classification, function estimation and density estimation which has also led to many other recent developments in kernel based learning methods in general (Cristianini N. et al.2000),which have been introduced within the context of statistical learning theory and structural risk minimization.like SVM, In the LS-SVM algorithm, a non-linear mapping function $\varphi(\cdot)$ is abtained by constructing the regression model while the input data is mapped to a higher dimensional feature space. When the least-squares support vector is used as a soft testing tool, a new optimization problem is formulated in the case of SRM. And then, Lagrange function is adopted to solve this optimization problem. Based on Mercer's theory, there is an equation between kernel function $K(x_i, x_j)$ and mapping function $\varphi(\cdot)$:

$$\varphi(x_k)^T \varphi(x_l) = K(x_k, x_l) \quad (1)$$
$$k, l = 1, \ldots, N$$

The kernel function must meet Mercer theorem.The common examples of kernel function contain linear, polynomial, radial basis function (RBF) kernel and multi-layer perceptron (MLP). In this study, RBF kernel was selected. The formula is:

$$K(x_k, x_l) = \exp\left(-\left\|x_k - x_l\right\|^2 \Big/ \sigma^2\right) \quad (2)$$

Finally, the LS-SVM regression model can be obtained as:

$$y(x) = \sum_{k=1}^{N} \alpha_k K(x, x_k) + b \quad (3)$$

A good LS-SVM classification model with high prediction accuracy and stability is built based on proper parameter setting. Grid-search technique is applied to find out the optimal parameter values which include regularization parameter gam (γ) and the RBF kernel function parameter sig2 (σ2) which is the bandwidth in the common case of the RBF kernel. γ determines the trade-off between SRM and ERM, and is important to improve the generalization performance of LS-SVM model. σ2 controls the value of function classification error, and influences the number of initial eigenvalues and eigenvectors directly. Moreover, σ2 reflects the sensitivity of LS-SVM

model to the noises from input variables. All the aforementioned calculations were performed using MATLAB 2006b (The Math Works, Natick, USA).

## 3. RESULT AND DISCUSION

### 3.1 Spectral features of soil

Fig. 1 shows the whole transmission's spectra of four kinds of soil It can be finded that there are same trend in the spectra, and hav a little different in some peaks 。The baseline changes are the main differences from one spectrum to another while they are unavoidable in spectra, and different types of pretreatment have been applied to the spectra to eliminate them (R. Tsenkova, et al.1999; Y.J. Chen, et a1999). One of the pretreatment methods often used is standard normal variate (SNV). SNV was applied for light scatter correction and reducing the changes of light path length. In the spectra, the baseline changes are induced maybe by the light scattering due to the size of powder granule. Hence, we have applied SNV to the spectra.But it is still difficult to discriminate the soil .Therefore, further treatments would be needed and then the latent features of the spectra could be applied.

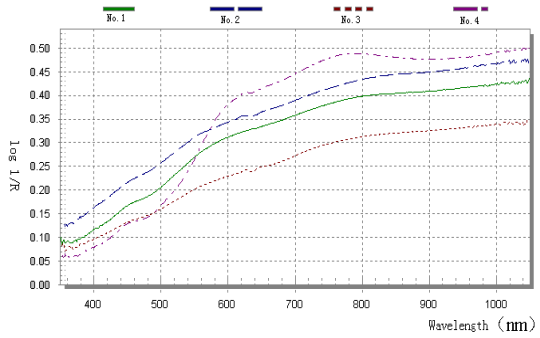


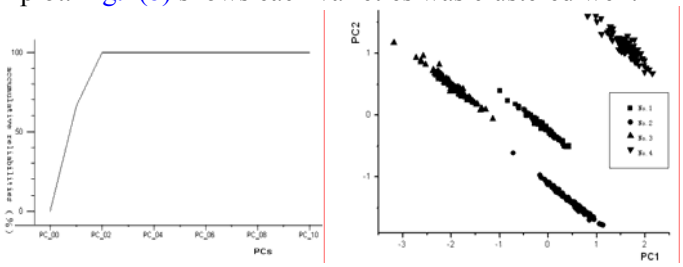

*Fig. 1*. Whole Absorbance spectra of soil

### 3.2 Principal components analysis (PCA) Clustering

PCA is a method for the re-expressing multivariate data. It allows the researcher to reorient the data so that the first few dimensions account for as much of the available information as possible. The principla components solution has the property that each component is uncorrelated with all others, which has the advantage of eliminating multicollinearity. The principal

components (PCs) plot obtained using the first primary PCs can be used for the partern recognition.

Fig. 2(a) shows the accumulative reliabilities of the PCs. It indicated that the accumulative reliabilities of the first tow PCs could explain up to 99.9% of the total variance. The third PC only interpreted an additional information which is less than 0.1% and it can not contribute so much as the aforementioned PCs. So the first three PCs were considered as the inputs of the cluster plot. Fig.2(b) shows each varieties was clustered well.




(a) (b)

*Figure 2.*PC accumulative reliabilities of soil(a) and Cluster plot with PC1xPC2 of four varieties of soil(b)

## 3.3 LS-SVM Classification Model

With the encoding and decoding of the output varieties , LS-SVM classification model can use to slove multi-class problems, The coding is defined by a codebook. The codebook is represented by a matrix where the columns represent all different classes and the rows indicate the results of the binary classifiers. The four brands with original labels [1 2 3 4] were encoded in the following codebook (using Minimum Output Coding) (Table 1):

*Table 1.*Multiple binary classifiers for output sets of LS-SVM for three Brands

| Brand No. | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Binary Classifiers | -1 | -1 | 1 | 1 |
| | -1 | 1 | -1 | 1 |

Finally, the classified results of LS-SVM are decoded again to its original form. Two models were developed using different sample set consisting of 600 spectra each. Based on the model developed, the LS-SVM model was applied to predict the 45 remaining samples after similar mathematical pretreatment to the calibration ones. LS-SVM was performed with RBF kernel. In the LS-SVM model, the determination of parameters γ and σ2 is important. In this study, these parameters were optimized with values of γ in the range of 2-1-210 and σ2 in the range of 2-215 with adequate increments by grid-search technique of “leaveoneout_lssvm” Validation. These ranges were chosen from previous studies where the magnitude of parameters to be

optimized was established. For each combination of γ and σ2 parameters.Because there are two levels in each row of binary classifiers (Table 2) and LS-SVM classification model can only classify two varieties, the LS-SVM classification model needs to process on these two rows (dims) separately. The optimal pair of (γ, σ2) and prediction results for model is shown in Table 3 and the results show a satisfying prediction precision.

*Table 3*.Prediction results for these two separate models

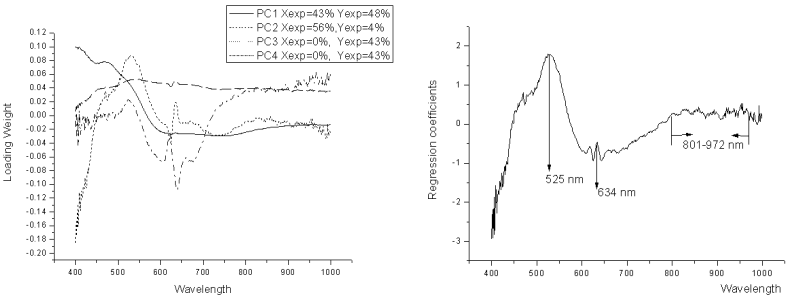
| $\gamma_1$ | $\gamma_2$ | $\sigma^2_1$ | $\sigma^2_2$ | Variety | Prediction (n=25×4) |
|---|---|---|---|---|---|
| | | | | | Recognition ratio |
| | | | | 1 | 100% |
| | | | | 2 | 96% |
| 2.3608 | 0.9477 | 5.6394 | 6.2498 | 3 | 100% |
| | | | | 4 | 100% |
| | | | | All | 99% |

## 3.4 Selections of sensitive bands

In order to obtain the sensitive wavebands for further designing of optical instruments, a calibration model based on the 400 samples of calibration set was calculated both by PLS analysis. The optimal number of LVs in PLS analysis was determined as four by full cross-validation. These wavelengths with small absolute x-loading values were less important *than those* with large ones. From loading weights of each LV, there are some sensitive bands chose: 525 nm,634 nm,801-972 nm in regression coefficients Fig. 3(a), Also, these sensitive bands can also been found in LV1-LV4 Fig. 3(b).But we still find that the peak at 525 nm and 634 nm appeared not constantly in each LV's loading weight, so more work need to do.In order to know the sensitive bands,the band from 801-972 nm(Fig.4) were used to analysis by PLS again, in which we find the importance of them are very close, Thus, we chose these wavelengths as the input of LS-SVM, Each wavelengths predict the 100 remaining samples. when it goes to LS-SVM,the results of 801-972 nm (totally 164 bands) are between 0.83 to 0.93.still very close. But the results of 525 nm and 634 nm is 25% and 20%,which can not meet to the demand.

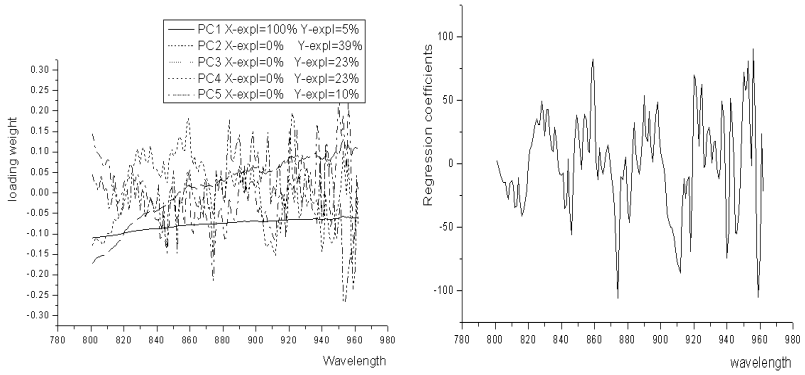
So the 801-972 nm were used to be the sensitive bands of soil's varieties, and the result of it from LS-SVM is 90%.

The prediction results indicated that the selected wavelengths of 801-972 by loading weights and regression coefficients and the result of LS-SVM analysis can reflect the main characteristics of soil of different verities. The sensitive wavelengths would be useful for the development of portable instrument or online applications to discriminate the verities of soil.




(a) (b)

*Figure 3.* Regression coefficients (a) and loading weights (b) of whole spectrum




(a) (b)

*Figure 4.* Regression coefficients (a) and loading weights (b) of 801-972 nm

## 4. CONCLUSION

NIR spectroscopy was successfully utilized for the identification of soil's varieties. An excellent recognition ratio of 100% was achieved based on LS-SVM. The results demonstrated that NIR spectroscopy has the potential ability to identify soil with different internal qualities. The sensitive bands for the discrimination would be helpful to develop portable instruments for commercial applications of adulteration detection. Therefore, more samples of different brands should be included to expand the model and improve the model generalization, specificity and accuracy.

## ACKNOWLEDGMENT

This study was supported by National Science and Technology Support Program (2006BAD10A09), Zhejiang Provincial Natural Science

Foundation of China (Y307119,Y104616),Finance aids projects from the water conservancy construction of Zhejiang Province(RC0614).

## REFERENCES

Cristianini N., Shawe-Taylor J., An Introduction to Support Vector Machines, Cambridge University Press.2000

Dan Ess ,Mark Morgan.The precision farming guide for agriculturists,John Deere Publishing.1997.Jan

He, Y., Li, X.L., Shao, Y.N.,Discrimination of Varieties of Apple Using Near Infrared Spectra Based on Principal Component Analysis and Artificial Neural Network Model. Spectroscopy and Spectral Analysis, 2006(5): 850-853

J. B. Reeves, III, Effects of water on the spectra of model compounds, J. Near Infrared Spectrosc. 1994:199-212

Liu, L., Cozzolino, D., Cynkar, W. U., Gishen, M., Colby. C. B., Geographic Classification of Spanish and Australian Tempranillo Red Wines by Visible and Near-infrared Spectroscopy Combined with Multivariate Analysis. J. Agric. Food Chem. 2006(5): 6754-6755

R. Tsenkova, et al, Near-infrared spectroscopy for dairy management: measurement of unhomogenized milk composition, J. Dairy Sci. 1999, 82: 2344-2351

Sudduth K A,Drummond S T, Kitchen N R. Accuracy issues in electromagnetic induction sensing of soil electrical conductivity for precision agriculture. Computers and Electronics in Agriculture,2001 31: 239–264

Viscarra Rossel R A; Walter C, Rapid, quantitative and spatial field measurements of soil pH using an ion sensitive field effect transistor. Geoderma, 2004:119, 9–20

Y.J. Chen, et al, Development of calibration with sample cell. compensation for determining fat content in unhomogenized raw milk by a simple NIR. transmittance method, J.Near Infrared Spectrosc, 1999, 7: 265

Z.M.G，W.A.F,elt. The conspectus of soil series in Zhejiang province[M],the press of china agricalture of science and technology,2000