4th Intercontinental Geoinformation Days

igd.mersin.edu.tr

Vis-NIR spectroscopy coupled with machine learning algorithms to predict soil gypsum in calcareous soils, southern Iran

Monire Mina [1], Mahrooz Rezaei [*2], Leila Hossein Abadi [3], Abdolmajid Sameni [1]

[1]*Shiraz University, School of Agriculture, Department of Soil Science, Shiraz, Iran*
[2]*Wageningen University, Meteorology and Air Quality Department, Wageningen, the Netherlands*
[3] *Shahid Beheshti University, Remote Sensing and GIS Center, Tehran, Iran*

Keywords
PLSR model
Savitzky-Golay filter
Spectral reflectance

Abstract

The use of soil spectral reflectance, which has been introduced as a new method in soil science, is widely used in estimating the physicochemical properties of soil. The aim of this study was to estimate the amount of gypsum in surface soils of Fars province. Based on random sampling method, 100 soil samples were collected and measured by standard method. Spectral analysis of soil samples was performed using a spectrophotometer in the range of 2500-400 nm. After this stage, various preprocessing methods were evaluated and finally the percentage of soil gypsum was modeled using two models of partial least squares regression (PLSR) and support vector regression (SVR). The results showed that the best results for estimating the percentage of soil gypsum are related to the SVR model with Preprocessing Savitzky- Golay Filter with the first derivative. Also, according to RPIQ statistics, the estimation of PLSR model for the percentage of soil gypsum in the weak class is 1.02% and for the SVR model in the moderate class is 1.54%.

1. Introduction

The use of visible-near-infrared spectroscopy has been introduced as a fast, inexpensive and non-destructive method that has a good capability in estimating different soil properties (Cambou et al., 2016). One of the most important characteristics of soil is the amount of soil gypsum. Gypsum has more solubility than carbonates and therefore, is under the influence of leaching process, and this resulted in less amount in the soil (Chaternour et al., 2020). The amount of gypsum has significant effect on soil properties such as soil water retention, aggregate stability and soil structure; More than 25% gypsum has a negative effect on plant growth and soil resilience (Smith and Robertson., 1962). Due to the cost, time and difficulty of direct measurement of soil gypsum, the use of indirect methods such as soil spectral behavior and spectroscopy has become common (Khayamim et al., 2015). So far, many studies have been done in this field, most of which have been researched on soil particle size, CaCO3 (Gomez et al., 2008), the soil organic matter (Ostovari et al., 2018) and soil moisture (Mina et al., 2021). In these studies, methods such as partial least squares regression, principal component regression, and support vector machine have been used to estimate the corelation between soil properties and spectral data (Farifteh et al., 2007). Also, studies have been reported in estimating the amount of soil gypsum using spectral reflections. Among these studies is a study by Chaternor et. al., 2020, they used PLSR and SVR models to estimate the amount of gypsum using spectral reflections of soil. Their results showed that the SVR model has a higher accuracy than the PLSR model. In another study, some chemical properties of soils in Isfahan province were estimated by spectroscopy. The properties of calcium carbonate, gypsum and organic matter were obtained with optimal accuracy with coefficients of determination of 0.45, 0.8 and 0.61, respectively (Khayamim et al., 2015). Hassani et al. (2014) estimated the properties of gypsum (RPD = 2.65), organic matter (RPD = 1.64) and calcium carbonate (RPD = 2.86) using spectral reflections. Gohari et al. (2017) used visible-near-infrared spectroscopy to estimate the amounts of gypsum, organic matter and carbonates. their research showed better result in the percentage of gypsum and organic matter in the good class, while it

* **Corresponding Author**

*(monireh.mina@gmail.com) ORCID ID xxxx – xxxx – xxxx – xxxx
(mahrooz.rezaei@wur.nl) ORCID ID xxxx – xxxx – xxxx – xxxx
(leilahosseinabadi1993@gmail.com) ORCID ID xxxx – xxxx – xxxx – xxxx
(majid.baba@gmail.com) ORCID ID

Cite this study

Mina, M., Rezaei, M., Abadi, L. H., & Sameni, A. (2022). Vis-NIR spectroscopy coupled with machine learning algorithms to predict soil gypsum in calcareous soils, southern Iran. 4th Intercontinental Geoinformation Days (IGD), 246-249, Tabriz, Iran

showed worse result in the carbonates in the weak class due to the relative deviation of the model prediction.

Due to the importance of gypsum content, this study was conducted to estimate the percentage of soil gypsum with PLSR and SVR models using spectral data by applying Savitzky -Glaye filters with the first derivative.

2. Method

2.1. Study area

The study area was Fars province which is located in the south-central region of Iran with coordinates 27° 2' to 31° 42' latitude and 50° 42' to 55° 36' longitude.

2.2. Soil sampling and soil analysis

Soil sampling was done randomly from a depth of 0 -10 cm. 100 soil samples were collected and then the samples were transferred to the laboratory, air dried and passed through a 2 mm sieve. The gypsum content was determined by acetone method (Nelson,1982).

2.3. Spectral reflectance measurement

Soil spectral data were determined using a spectroscopy device (NIRS-XDS) in the range of visible and near-infrared wavelengths (2500-400 nm). 20 g of each sample of air-dried soil with a size of less than 2 mm was placed in a special container and then 5 scans were performed on them (Figure 1, a). Due to the high noise at the beginning and end of the spectral data, the range of 449-400 and 2500-2451 nm was removed from the modeling process and then the preprocessing of the Savitzky and Glaye filters with the first derivative was applied to the spectral data of soil samples (Figure 1, B).

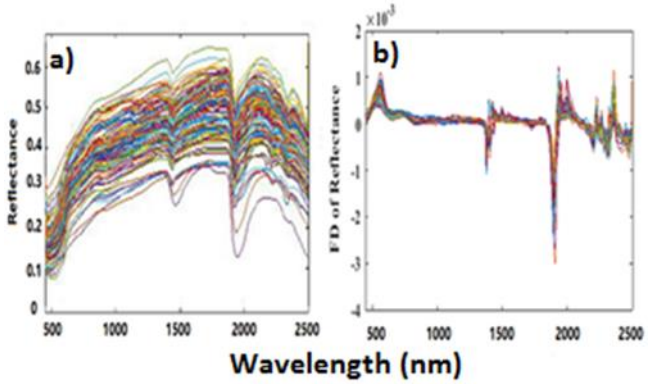

Figure 1. The a) raw and b) preprocessed spectral reflectance data.

2.4. Model evaluation

To predict the percentage of soil gypsum based on soil spectral reflectance, partial least squares regression and support vector regression were used. To evaluate the accuracy of the models, three statistical criteria including coefficient of determination (R2), root mean square error (RMSE) and ratio of performance to the interquartile range (RPIQ) were used (Mina et al., 2022).

3. Results

For the purpose of modeling, the data were randomly divided into two sets of training (70% of data) and testing (30% of data). Using t-test, no significant difference was observed between the two data sets. Table 1 shows the statistical summary of the measured soil gypsum for both train and test datasets. All soil samples had a low amount of gypsum with a mean of 0.97% and 0.99% for train and test datasets, respectively.

Table 1. Statistical analysis of the soil gypsum, minimum, maximum (Range), mean values, standard deviation (SD) and coefficient of variation (CV).

soil parameter	Gypsum	
	Train	Test
Unit	%	%
Range	0.2-3.98	0-3.90
Mean ± SD	0.97[a] ±0.64	0.99[a] ±0.68
CV (%)	65.97	68.68

a-significant difference (p< 0.05)

The values of R^2, RMSE and RPIQ from modeling in estimating soil gypsum based on soil spectral reflections are presented in Table 2. The results of Table 2 show that the SVR model can have a higher performance in estimating the amount of soil gypsum than the PLSR model. The SVR model has the highest R2 (0.85, 0.73%) and RMSE (0.22, 0.39%) in both training and testing stages, respectively. In addition to RMSE, the accuracy of the model predicted by RPIQ was also evaluated. Classification is done by Lacerda et al., 2016 into 6 classes: Very Poor with RPIQ < 1, weak with RPIQ = 1 – 1.4, Moderate with RPIQ = 1.4-1.8, Good with RPIQ = 1.8-2, Very Good with RPIQ = 2-2.5 and Excellent with RPIQ >2.5. The SVR model has a moderate performance using spectral reflectance with Savitzky- Glaye filter with the first derivative, and the PLSR model has a poor performance in estimating the amount of soil gypsum.

Table 2. Prediction result for gypsum using partial least squares regression (PLSR) and support vector regression (SVR) model.

Method	Train			Test		
Model	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ
PLSR	0.74	0.30	1.97	0.57	0.48	1.02
SVR	0.85	0.22	2.10	0.73	0.39	1.54

R^2 - coefficient of determination, RMSE - root mean square error, and RPIQ - ratio of performance to the interquartile range.

Figure 2 depicts the measured gypsum versus the predicted gypsum using the PLSR and SVR models in both train and test datasets. In both datasets, the points are well-scattered around 1:1 line for SVR compared to PLSR, which shows the better performance of the SVR model.

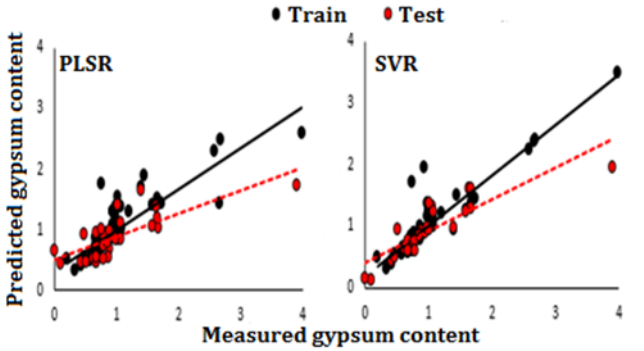

Figure 2. Scatter plots of predicted versus measured gypsum by PLSR and SVR. a) Train set (N=70), b) Test set (N=30). PLSR: Partial least squares regression; SVR: Support vector regression.

4. Discussion

The t-test showed that there was no significant difference between soil gypsum in the train and test datasets. Table 1 shows the statistical description of soil properties in two sets of training and testing. The standard deviation of the amount of gypsum in the training set is 0.64 and, in the test, set is 0.68. This clearly shows that the test dataset is a good representation of the dataset.

It has been used as an input parameter in PLSR and SVR models to estimate soil gypsum using wavelengths of the visible-near-infrared range (400-2500 nm). In predicting the percentage of gypsum by two models PLSR and SVR, the highest value of R^2 and the lowest value of RMSE were obtained in each training and test sets (Table 2). The results of this study are consistent with the study by Chaternor et al., 2020. The results clearly show that the SVR model is better for estimating soil gypsum than the PLSR model. Research has been conducted using PLSR and SVR models in 72 soil spectral samples in Iran. Their results showed that the SVR model has the highest performance in estimating soil CEC.

In another study, Khayamim et al. (2015) obtained an excellent yield (RPIQ> 2) for the amount of soil gypsum using the PLSR model.

Also, the SVR model had the shortest distance from the line (1: 1) and the best fit (Figure 2).

In general, according to Table 2 and Figure 2, the results clearly show that the performance of the SVR model is better than the PLSR model in estimating soil gypsum. Therefore, it can be concluded that SVR is a more suitable multivariate method for soil spectral data.

According to Nawar et al. 2016 research, the range of changes in the concentration of soil properties has an important role in the accuracy of the regression model and with an increase in changes and data breadth and also an increase in range, the model's accuracy estimation increases.

Also, according to the Wilding (1985), the extent of the data with the coefficient of variation (CV) in the range of > 35 is considered as a large extent. In the present study, the CV for the training and test data sets is 65.97% and 68.68 % respectively, which indicates the appropriate breadth in the Collected data and has improved the accuracy of gypsum estimation in both models.

5. Conclusion

In this study, we explored the ability of reflectance spectroscopy to estimate gypsum. Generally, the results showed that there is a correlation between the gypsum and soil spectral reflectance. Among the two predictive models, the machine learning algorithm performed better compared to the common PLSR method. Our results proved that spectral reflectance is a promising tool for efficiently assessing large areas.

Therefore, it can be said that soil spectral reflections can be used as a rapid and alternative method in soil.

To get a better view of the performances of machine learning methods in soil science studies, we recommend comparing other data mining approaches such as random forest and artificial neural networks, for the future studies.

References

Cambou, A., Cardinael, R., Kouakoua, E., Villeneuve, M., Durand, C., & Barthès, B. G. (2016). Prediction of soil organic carbon stock using visible and near infrared reflectance spectroscopy (VNIRS) in the field. *Geoderma*. *261*, 151-159.

Chatrenor, M., Landi, A., Farrokhian Firouzi, A., Noroozi, A., & Bahrami, H. A. (2020). Application of hyperspectral images in Quantification of soil gypsum in center areas of Khuzestan province prone to dust generation. *Applied Soil Research*, *8*(3), 1-13.

Farifteh, J., Van der Meer, F., Atzberger, C., & Carranza, E. J. M. (2007). Quantitative analysis of salt-affected soil reflectance spectra: A comparison of two adaptive methods (PLSR and ANN). *Remote Sensing of Environment*. *110*(1), 59-78.

Hassani, A., Bahrami, H., Noroozi, A., & Oustan, S. (2014). Visible-near infrared reflectance spectroscopy for assessment of soil properties in gypseous and calcareous soils. *Watershed Engineering and Management*, *6*(2), 125-138.

Khayamim, F., Wetterlind, J., Khademi, H., Robertson, A. J., Cano, A. F., & Stenberg, B. (2015). Using visible and near infrared spectroscopy to estimate carbonates and gypsum in soils in arid and subhumid regions of Isfahan, Iran. *Journal of Near Infrared Spectroscopy*, *23*(3), 155-165.

Lacerda, M. P., Demattê, J. A., Sato, M. V., Fongaro, C. T., Gallo, B. C., & Souza, A. B. (2016). Tropical texture determination by proximal sensing using a regional spectral library and its relationship with soil classification. *Remote Sensing*, *8*(9), 701.

Mehrabi Gohari, E., Matinfar, H. R., Jafari, A., Taghizadeh-Mehrjardi, R., & Khayamim, F. (2020). Comparing Different Statistical Models and Pre-processing Techniques for Estimation several chemical properties of the soil Using VNIR/SWIR Spectrum. *Iranian Journal of Remote Sensing & GIS*, *11*(4), 47-60.

Mina, M., Rezaei, M., Sameni, A., Moosavi, A. A., & Fallah Shamsi, R. A. S. H. I. D. (2022). Using Soil Pedotransfer and Spectrotransfer Functions to Estimate Cation Exchange Capacity in Calcareous Soils, Fars Province. *Iranian Journal of Soil and Water Research*, *52*(11), 2911-2922.

Mina, M., Rezaei, M., Sameni, A., Moosavi, A. A., & Ritsema, C. (2021). Vis-NIR spectroscopy predicts threshold velocity of wind erosion in calcareous soils. *Geoderma*, *401*, 115163.

Nawar, S., Buddenbaum, H., Hill, J., Kozak, J., & Mouazen, A. M. (2016). Estimating the soil clay content and organic matter by means of different calibration methods of vis-NIR diffuse reflectance spectroscopy. *Soil and Tillage Research*, *155*, 510-522.

Nelson, R. E. (1982). Carbonate and gypsum. In: Page, A.L. (Ed.), Methods of Soil Analysis: Part1Agronomy Handbook 9. *American Society of Agronomy and Soil Science Society of America, Madison (WI)*. 6, 181–197.

Ostovari, Y., Ghorbani-Dashtaki, S., Bahrami, H. A., Abbasi, M., Dematte, J. A. M., Arthur, E., & Panagos, P. (2018). Towards prediction of soil erodibility, SOM and CaCO3 using laboratory Vis-NIR spectra: A case study in a semi-arid region of Iran. *Geoderma*, *314*, 102-112.

Smith, P. A., & Robertson, J. E. (1962). Some factors affecting the site of alkylation of oxime salts. *Journal of the American Chemical Society*, *84*(7), 1197-1204.

Wilding, L. P. (1985). Spatial variability: its documentation, accomodation and implication to soil surveys. In *Soil spatial variability, Las Vegas NV, 30 November-1 December 1984* (pp. 166-194).