iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VIII **Month of publication:** August 2020

DOI: https://doi.org/10.22214/ijraset.2020.30940

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VIII Aug 2020- Available at www.ijraset.com

Influence of Data Pre-Processing on the Results of PCA of Soil Reflectance Spectra

Shaikh Aamir[1], Vijayashri A. Losarwar[2]
[1, 2]*P. E. S. Engineering College, Dept. of Computer Science, Aurangabad, India*

Abstract: ***Principal component analysis (PCA) is becoming a routine mathematical tool in several fields for data analysis, dimensionality and noise reduction, data compression. The new branch of data analysis called as big data analysis or data mining has emerged in which data in very large quantities, having large number of variables in the form of signal or an image has to be processed. PCA is, usually the first step in data mining and automated decision making as a feature extraction technique. The data may contain variables having different units in which they were measured. Data pre-processing plays an important role in the final results of PCA.***
We analyse the influence of different data pre-processing methods which are performed before subjecting the data to PCA. Recent development of Vis-NIR spectrometers, have provided a non-destructive, economical and fast means to study soil, liquids, vegetation, and biological samples from their reflectance spectra. Also remote sensing is increasingly used for the study of earth with the help of orbiting satellites which provide data in the form of reflectance spectra. We study the influence of various data pre-processing methods applied to soil reflectance spectra and the resulting PCA output with the help of Pearson correlation coefficient.
Keywords: Soil, PCA, data pre-processing, soil reflectance spectra, Pearson correlation Coefficient

I. INTRODUCTION

Soil is the top layer of earth's surface. It consists of organic and inorganic matter. The plants grow in soil. It is very important to know the features of soil for agricultural and industrial use. Minerals present in the soil give a characteristic texture and color to the soil. Several methods exist for the characterization of soil. Soil spectral reflectance [1] and [2] is one of the optical means for characterizing soil.
In this method reflectance is measured as a function of wavelength. Reflectance is the fraction of incident electromagnetic radiation reflected by the surface. Reflectance depends on the nature of soil hence spectral reflectance can be used for characterizing soil or to know the dominant ingredient in the soil. Soil reflectance spectra are usually depicted in the form of graphs with reflectance versus wavelength. These are spectral signatures of soil [3].

II. REVIEW OF LITERATURE

Spectral reflectance technique is a powerful non-destructive method of studying matter in organic, inorganic and biological form. It has found many applications in diverse fields [4], [5], [6], [7] and [8]. Basically, electromagnetic radiation in the range of 1000nm to 3000 nm is made incident on the sample to be studied and the reflectance is measured. The spectral reflectance arises from the stretching, bending or twisting of the chemical bonds of the minerals present in the sample. Hence, they are unique and mirror the nature of soil.
Principal component analysis (PCA) is also known as Karhunen–Loéve transform which was derived from Hotelling transform. It is used in many diverse fields [9] and [10] and [11], wherever a large amount of data is generated and one wants to extract useful information from it. PCA takes data as input and gives information about the data in the form of what are called as Principal Components. It converts correlated variables in the data to uncorrelated variables called as principal components. In fact they are linear combinations of the original variables of the data. The data may contain more dimensions than needed to understand its underlying dynamics.
The usefulness of PCA arises from its ability to reduce the dimensionality of data and reveal the dynamics of the system with the help of fewer amounts of data. The principal components are selected depending upon the largest variance content; usually it is the first principal component PC1. The succeeding principal components PC2, PC3 and so on are orthogonal to the preceding ones, and contain the remaining variance as possible. PCA has been applied successfully in the study of soil spectral reflectance [12], [13], [14], [15] and [16].

III.MATERIALS AND METHODS

Pre-processing of spectral reflectance data is an important step in order to apply multivariate techniques like the PCA. PCA gives new data that is the transformed data in the form of principal components. They represent physical variables more prominently when they are calculated from standardized data as compared to un-standardized data. The standardized spectral reflectance data is obtained by mean centring the data and normalizing them by the standard deviation.

The next step of processing the spectral reflectance data is some type of manipulation to reduce noise, maximize information content of data, so that one obtains features that truly reflect the original data. There are so many methods for this aspect of pre-processing.

Here we select three methods, namely data smoothening, data down sampling, and band data selecting and study their effects on the final results of PCA which are in the form of principal components.

Soil reflectance data files are available at the USGS repository [1]. We selected four data files for our investigation. The soil names are Acmite, Alunite, Analcime and Augite.

1) *Step 1. Obtain Mean Corrected Data:* The values in data can be used directly for further calculations and analysis. Data can also be represented as deviations from the mean or average, such data are called as mean corrected data or mean centred data or zero mean data. Corrected data = data – data mean
2) *Step 2. Standardized data:* It is also called as normalized data. It is obtained by dividing the mean-corrected data by the respective standard deviation (square root of the variance). The variances of the standardized variables are always 1, and the correlation will always lie between -1 and +1. The value will be zero if there is no linear association between the two variables, -1 if there is perfect inverse linear relationship between the two variables, and +1 for a perfect direct linear relationship between the two variables. Covariance of two standardized variables is called as correlation coefficient or Pearson product moment correlation, or simply Pearson coefficient.

Principal component analysis (PCA) in simple words means analysis of data by forming new data which is linear combination of original data variables. The steps to perform PCA are:

a) Step 1. Put data in matrix form.
b) Step 2. Standardize (normalize) data Calculate mean values of each column. Next make the data zero mean or mean-centred data. This is done by subtracting each element of the column with the mean of that column only. Thus we get a mean-centred matrix.
c) Step 3. Calculate the covariance of the matrix. From this step we get eigen-values and eigen-vectors, the Principal Components.
d) Step 4. Calculate the transformed data from the Principal Components. As the soil reflectance data is having same units there is no need for standardization of data, only the mean-centred data is sufficient for analysis. Taking the mean-centred data as a starting point for the purpose of further pre-processing of data, we employ three methods. The three methods are:

- Local average smoothening of data (smooth data
- Down sampling of data (down sampled data)
- Water absorption band elimination from data (band limited data)

We implemented the local averaging algorithm on the data for the smoothening of data. We tried employing other simple methods but they did not give satisfactory results.

The down sampling of data was done by selecting every third consecutive element from data, thus reducing the number of data points. Water strongly absorbs electromagnetic radiation of wavelength 1440 nm and 1950 nm. The data points around these regions were removed from the working data, thus giving what we call as band limited data.

3) *Step 3. Applying PCA to Pre-processed Data:* We applied PCA to data of four different soil samples which were pre-processed by three different techniques, and studied the eigen-vectors, eigen-values, variances, principal components, percentage of principal components and Pearson correlation coefficients. We calculated the Pearson correlation coefficients of a matrix formed by taking a column from original data and from the transformed data. Here we report only the Pearson correlation coefficient of principal components PC1 and PC2, as they reflect nicely the efficiency of different data pre-processing techniques. They are listed in Tables I (A) and I (B)

IV. CONCLUSIONS

The graph in Figure 1 is the reflectance spectra of the four soil samples selected for this study, along with the log (absorbance) spectra to know about the absorption lines. Their data were only mean-centred and plotted.

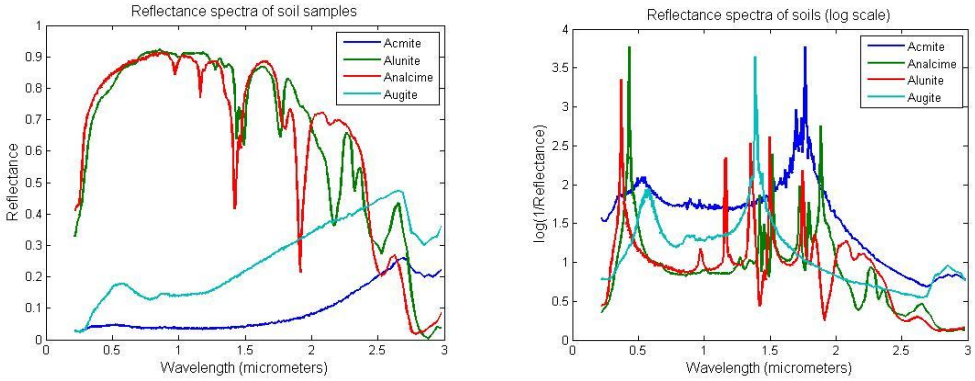


Figure 1. Reflectance and log (Absorbance) spectra of four soil samples.

The graphs of reflectance spectra obtained by local average smoothening and band limiting pre-processing methods are shown in Figure 2.

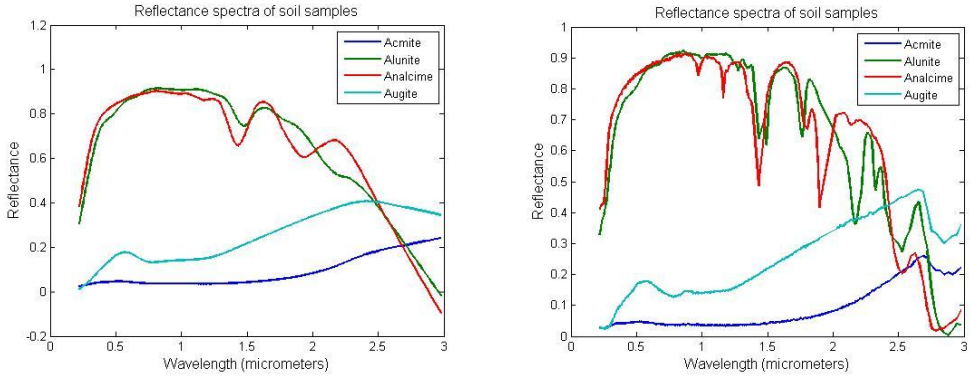


Figure 2. Smoothed (left), band limited (right) reflectance spectra of soil samples.

The results obtained in the form of Pearson correlation coefficients for the first two principal components (PC1 and PC2) are summarized in Tables I (A) and (B).

TABLE I (A)
PEARSON CORRELATION COEFFICIENTS FOR PC1

Soil Name	Smoothened data	D sampled data	Band selected data
Acmite	-0.848	-0.816	-0.832
Alunite	0.983	0.965	0.979
Analcime	0.983	0.964	0.974
Augite	-0.578	-0.563	-0.552

TABLE I (B)
PEARSON CORRELATION COEFFICIENTS FOR PC2

SOIL NAME	SMOOTHENED DATA	D SAMPLED DATA	BAND SELECTED DATA
ACMITE	0.398	0.433	0.447
ALUNITE	0.106	0.119	0.095
ANALCIME	0.053	0.033	0.061
AUGITE	0.815	0.822	0.832

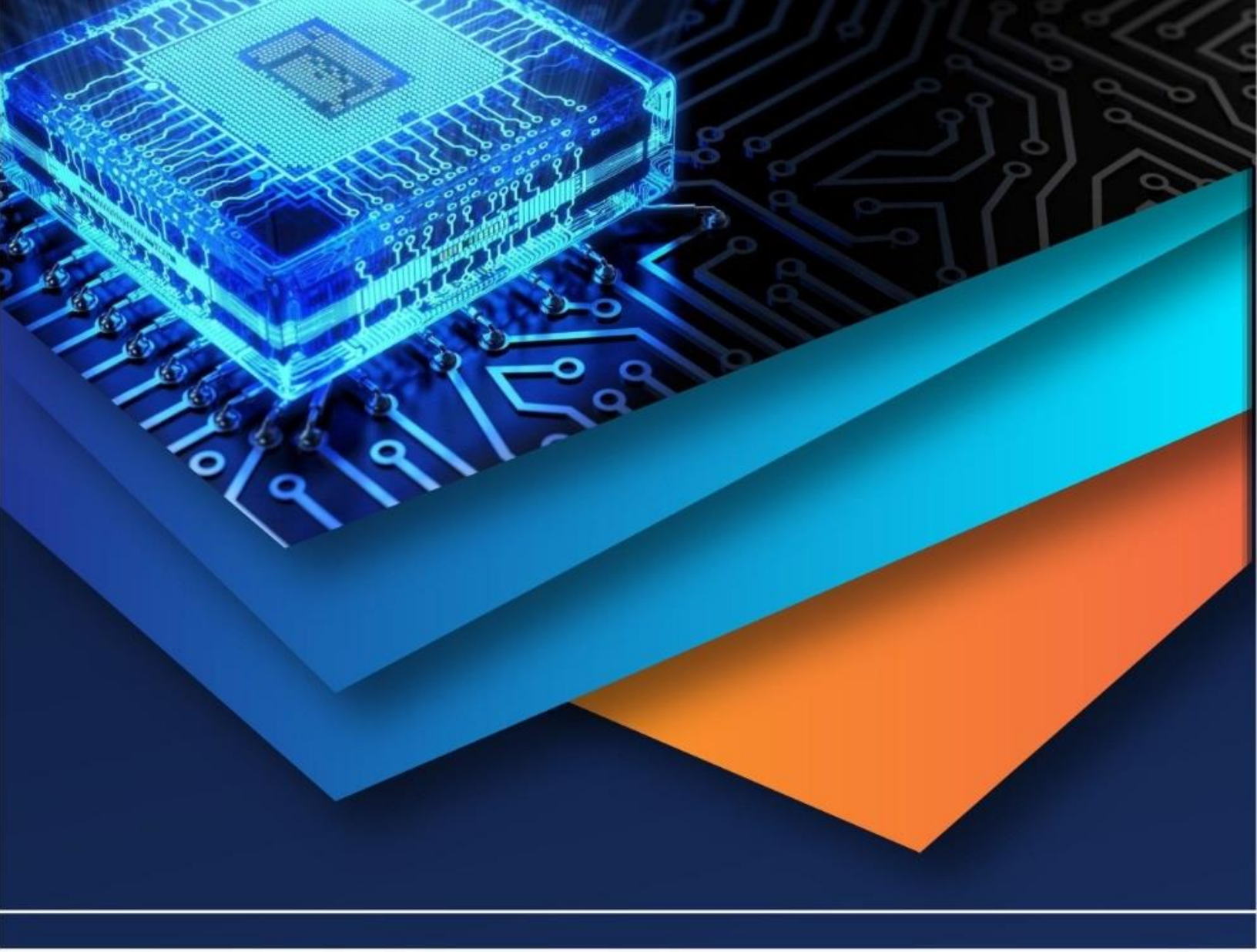
It is seen that the choice of the local average smoothening of data is the best choice. Down sampling of the reflectance spectra data reduces the information content slightly. For the two sample soils Alunite and Analcine the Pearson correlation coefficients are high indicating strong correlation between the input data to PCA and the output transformed data from PCA. It shows that pre-processing techniques give good results for some soil reflectance spectra. In future efficiency of different filter techniques will be compared with some other simple smoothening methods.

V. ACKNOWLEDGMENT

We are profoundly grateful to USGS for the soil reflectance spectra data provided on their website.

REFERENCES

[1] Bowers, S.A. & Hanks, R.J., Reflection of Radiant Energy from Soils, New York. Soil Science, 100, 130–138, 1965.
[2] Reeves III, J.B., Near- versus mid-infrared diffuse reflectance spectroscopy for soil analysis emphasizing carbon and laboratory versus on-site analysis: where are we and what needs to be done? Geoderma 158, 3–14. 2010.
[3] Jose´ Francirlei Oliveira, Michel Brossard, Pedro Rodolfo Siqueira Vendrame, Stanislas Mayi III, Edemar Joaquim Corazza, Robe´ lio Leandro Marcha˜o, Maria de Fa´tima Guimara˜es, Soil discrimination using diffuse reflectance Vis–NIR spectroscopy in a local toposequence. C. R. Geoscience 345, 446–453. 2013.
[4] Hunt, G.R., Salisbury, J.W., Lenhoff, C.J., Visible and near-infrared spectra of minerals and rocks: III. Oxides and hydroxides. Modern Geol. 2, 195–205. 1971.
[5] Foley, W.J. McIlwee, A., Lawler, I.R.,Aragones, L.,Woolnough, A. & Berding, N. Ecological applications of near-infrared spectroscopy - a tool for rapid, cost-effective prediction of the composition of plant and animal tissues and aspects of animal performance. Oecologia 116, 293-305. 1998
[6] He, Y., Huang, M., Garcı´a, A., Herna´ ndez, A., Song, H.. Prediction of soil macronutrients content using near-infrared spectroscopy. Comput. Electron. Agri. 58, 144–153. 2007
[7] Liu, Q.S., Torrent, J., Barro´ n, V., Duan, Z.Q., Bloemendal, J.. Quantification of hematite from the visible diffuse reflectance spectrum: effects of aluminum substitution and grain morphology. Clay Minerals 46, 137–147. 2011
[8] Cozzolino, D. & Moron, A. The potential of near-infrared reflectance spectroscopy to analyse soil chemical and physical characteristics. The Journal of Agricultural Science, 140, 65–71. 2003
[9] Horn, R. A. and Johnson, C. R. Matrix Analysis. England: Cambridge University Press, 1990.
[10] Manly, B. F. J.: Multivariate Statistical Methods: A Primer. London, Chapman & Hall, 1994.
[11] Chatfield, C. and A. J. Collins: Introduction to Multivariate Analysis. London, Chapman & Hall, 1980.
[12] Demattê, J.A.M., Campos, R.C., Alves, M.C., Fiorio, P.R., Nanni, M.R.. Visible–NIR reflectance: a new approach on soil evaluation. Geoderma 121, 95–112. 2004.
[13] Chang, C.W., Laird, D.A., Mausbach, M.J., Hurburgh Jr., C.R.. Near-infrared reflectance spectroscopy–Principal Components Regression analyses of soil properties. Soil Sci. Soc. Am. J. 65, 480–490. 2001.
[14] Islam, K., Singh, B., Macbratney, A. Simultaneous estimation of several soil properties by ultra-violet, visible, and near-infrared reflectance spectroscopy. Aust. J. Soil Res. 41, 1101–1114. 2003.
[15] Smith, M.O., Johnson, P.A., Adams, J.B. Quantitative determination of mineral types and abundances from reflectance spectra using principal component analysis. J. Geophys. Res. 90, C797–C804. 1985.
[16] Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J. Visible and near infrared spectroscopy in soil science. In: Sparks, D.L. (Ed.), Advances in Agronomy, 107. Academic Press, Burlington, pp. 163–215. 2010.
[17] USGS Homepage, http://speclab.cr.usgs.gov.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY