Proc. of the 9th International Symposium on Geotechnical Safety and Risk (ISGSR)
Edited by Zhongqiang Liu, Jian Dai and Kate Robinson
Published by Research Publishing, Singapore. ©2025 The Author(s), CC-BY 4.0 licence.
doi: 10.3850/GRF-25280825_isgsr-168-P127-cd

DENDROGRAM AND PRINCIPAL COMPONENT ANALYSIS APPLIED TO GEOTECHNICAL CBR DATA TO REMOVE DATA NOISE

Burt G. Look

AGTRE Pty Ltd, Brisbane, Queensland, Australia, E-mail: lookburt@agtre.au

Dendrogram analysis involves creating a hierarchical tree structure, which visually represents the relationships between different data points based on their similarities or differences. By clustering similar data points together, dendrogram analysis can help geotechnical engineers gain insights into the underlying relationships present in the data. Principal Component Analysis (PCA) identifies patterns that encode the highest variance in the data. PCA involves aggregating information inherent in multi-dimensional data, representing it with a reduced number of new variables. Dendrogram and PCA analysis application in geotechnical engineering are shown using California Bearing Ratio (CBR) test data. The relationship and groupings within the many data attributes not typically apparent are shown. Engineers typically uses only the CBR test value but the inter-relationships between the various "pieces" contributing to the CBR result is not evident. The PCA quantifies the key component of the tests where most variance occurs. A correlation matrix is used to show all analysis methods point to similar conclusion and industry practice of (incorrectly) using density ratio as the key parameter in quality control during earthworks construction.

Keywords: Dendrograms, Hierarchical cluster analysis, CBR Tests, Density ratio, Principal Component

1. Introduction

Geotechnical models require data input derived measured directly or using correlations from associated test parameters. However, regression models assume dependent and independent variables. Multicollinearity occurs when independent variables in a regression model are correlated. The predictive model may then be unreliable. Descriptive modelling to first assess trends can eliminate the multicollinearity issue. Alsanabani et al. (2025) showed by eliminating multicollinearity, principal component analysis (PCA) can provide more accurate and reliable assessments of the impacts of individual risks on pile installation.

When several tests or variables occur in geotechnical data, a multivariate analysis (MVA) is required when interrelationships among data is required. MVA considers more than one factor of independent variables that influence the variability of dependent variables. PCA and cluster analysis are two procedures used in MVA, by working on a number of variables simultaneously (StatTools, 2016). PCA is used to find correlated variables that can be combined, so that the dimension can be reduced. Cluster analysis finds similar subsets of data. Dendrograms provides a visual representation of the hierarchical clustering structure, making it easy to interpret, and identify groups or clusters within the data.

Various *in situ* modern test equipment compared well to each other, but least with the density ratio (DR), which is the de facto standard for quality control on earthworks projects. Dendrogram analysis was used to overcome this (unexpected) poor relationship problem, when poor correlations with DR for a test site in Cairns, Australia (Look, 2023) was evident. This dendrogram analysis suggests the past successes of DR now impede the implementation of more accurate and reliable technologies. Similar studies by Nazarian et al. (2014) concluded the adaptation of the modulus-based specification needs to be approached in the context of the levels of uncertainty associated with the current well-established density criteria. It was shown that achieving quality compaction (defined as achieving adequate layer modulus) is only weakly associated with achieving density.

Pavement designs commonly used the soaked California Bearing Ratio (CBR) which used is assumed well correlated with the DR measurements in earthworks quality control. The CBR is also related to the resilient modulus. Yet poor correlations often result when density versus CBR or modulus tests are compared. CBR test data is used to show the relationship and groupings within the many data attributes not typically reported. as engineers typically use only the CBR test value. However, using the tree – like data structures the inter-relationships between the various "pieces" contributing to the CBR result is evident (Look, 2021).

1.1 Multivariate Analysis

A dendrogram is a diagram that shows the hierarchical relationship between objects or groups. Its purpose is to determine the best way to allocate objects to clusters based on their similarity or dissimilarity. In hierarchical clustering, objects are successively paired together based on similarity, forming a tree-like structure (the dendrogram). By clustering similar data points together, dendrogram analysis can help geotechnical engineers gain insights into the underlying structures and trends present in their data.

Principal Component Analysis (PCA) identifies patterns (principal components) that encode the highest variance in the data. PCA involves aggregating information inherent in multi-dimensional data, representing it

with a reduced number of new variables. Unlike hierarchical clustering, PCA is not directly aimed at separating groups of samples. Dendrogram analysis focuses on grouping objects based on similarity, while PCA aims to reduce dimensionality and capture variance patterns. Both methods are unsupervised and useful for exploratory data analysis. In unsupervised methods no information about class membership or other response variables are used to obtain the graphical representation. Table 1 compares dendrogram and PCA when used in multivariate analysis. Both PCA and dendrogram analysis are illustrated using soaked CBR results and also comparing with typical correlation matrix and simple scatter plots for establishing correlations.

Table 1. Comparison between Dendrogram and PCA in multivariate analysis

Aspect	Dendrogram	Principal Component Analysis (PCA)
Definition	A visual representation of hierarchical clustering	A statistical method used to reduce the dimensionality of data
Purpose	Visualize relationships	To identify patterns and relationships in data
Type of analysis	Clustering analysis	Dimensionality reduction analysis
Output	Tree-like structure	Scatter plots and eigenvectors
Interpretation	Shows relationships and groupings within the data	Identifies principal components explaining the most variance in the data
Data requirements	Distance or similarity matrix	Numerical data

Nagoya et al. (2024) describes and uses the PCA for evaluation of the compaction property of fine aggregate. The compaction energy and degree of saturation in applying PCA to 726 observations obtained from compaction tests. Other grain size features such as uniformity coefficient and wet density indicated poor correlation.

2. Soaked CBR dendrogram and PCA analysis

CBR data points (55 No.) from a "uniform" Cooroy (CH) clay is used to illustrate the analysis using dendrograms, corelation matrix and PCA as well as simple paired graphs with correlation. The dendrogram analysis was shown in Look (2021) with clustering of 6 parameters. This showed the CBR is closely clustered to the compaction moisture content (MC) and the optimum moisture content (OMC) rather than density measurements (Figure 1).

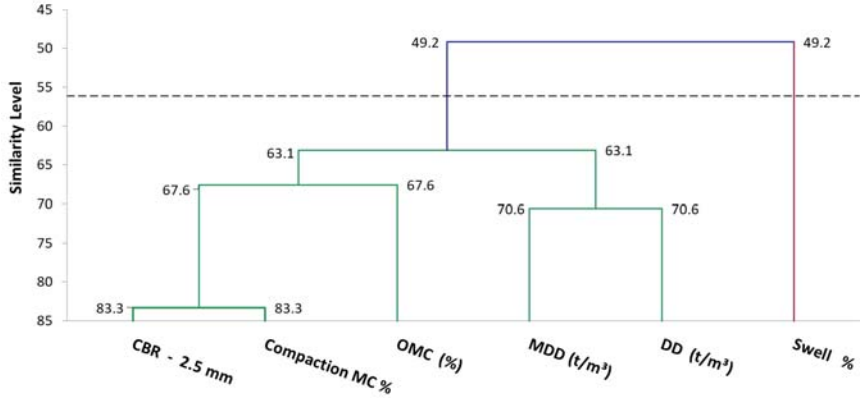

Fig. 1. Dendrogram of 6 key measurements in a 4-day soaked CBR test.

This could also have been seen in the correlation matrix (Table 1). CBR is most strongly correlated with the compaction moisture content (0.69) and least relarted with the compacted dry density (0.04). CBR is negatively correlated with swell (-0.83). This suggests that CBR for this expansive CH clay is most (negatively) correlated to the swell value after soaking for the 4 days, even more than the compaction MC. The swell is also shown to be strongly correlated to the compaction MC as was also shown with field measuremenst in Look (2021, 2023).

Figure 2 shows the PCA of the same data with 46.2%, 76.9% (46.2+30.7) and 89.4% (76.9+12.6) of total variance explained with the first, second and third components, respectively. CBR, compaction moisture content,

and swell represent the key components of the 6 variables. For this data set, the other 3 components of DD, OMC and MDD have little relationship. Yet quality control is based on the density ratio (Field dry density / MDD) and OMC compaction. This suggests the current quality control practice of focusing on Density ratio (DR) and assuming a relationship with CBR does not focus on the key variable affecting the CBR value. As the variables increases from 6 No. variables to 15 No. variables, the dendrogram analysis provides an easier approach to visualising the key relationships (Figure 3).

Table 2. Correlation Matrix for 6 No. test outputs for the 55 data soaked CBR data points

Correlation Matrix	Comp. MC %	DD (t/m³)	OMC (%)	MDD (t/m³)	CBR @2.5mm	Swell %
Comp. MC %	1.00	-0.30	0.23	-0.04	0.69	-0.85
DD (t/m³)	-0.30	1.00	-0.38	0.46	0.04	0.06
OMC (%)	0.23	-0.38	1.00	-0.34	0.40	-0.14
MDD (t/m³)	-0.04	0.46	-0.34	1.00	0.32	-0.38
CBR@2.5mm	0.69	0.04	0.40	0.32	1.00	-0.83
Swell %	-0.85	0.06	-0.14	-0.38	-0.83	1.00

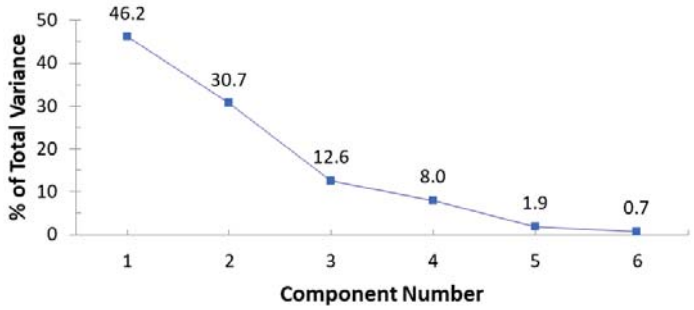

Fig. 2. Principal component analysis on scree plot which helps explain total variance.

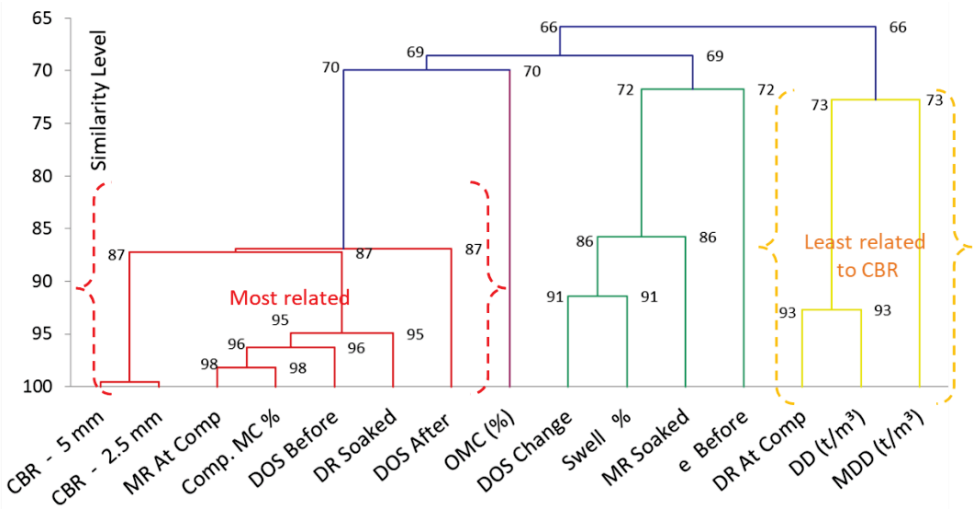

Fig. 3. Dendrogram of 15 parameters in a soaked CBR test.

The moisture related parameters (Degree of saturation, moisture ratio at compaction is most related to the CBR while, the density related parmaters (Density ratio at compaction) are least related. These insights could also have been achieved by carrying out multiple plots such as Figure 4 for the relationship of CBR with the density an moisture ratio (MC/OMC) when compacted and after soaked, respectively. Figure 4a shows that the

compacted density has little effect on the soaked CBR. Many other such graphs (not shown) are required to establish key relationships. Such parametric regresssion type analyses assume the distribution is normally distributed. The data noise and non-normal distribution is evident in Figure 4, which fails to show the many associated parameters affecting the CBR value. The factors are shown and visualised with PCA (Figure 2) and dendrogram analysis (Figure 3) with non parametric statistical tests.

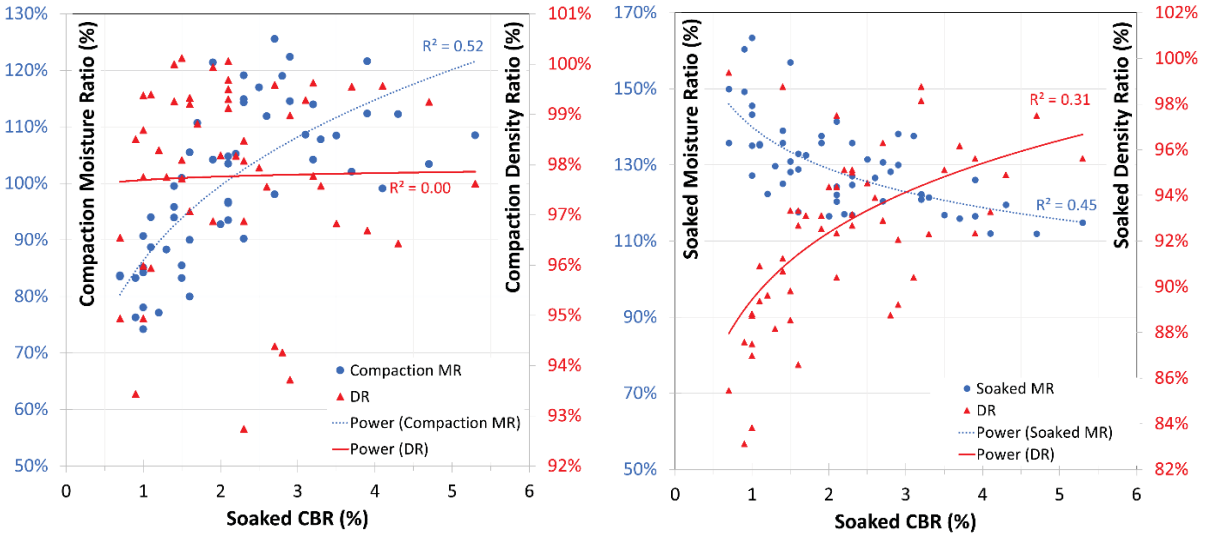

Fig. 4. (a) Soaked CBR vs compaction MR and DR (b) Soaked CBR vs after soak MR and DR

3. Conclusion

Multivariate analysis with complex data requires looking beyond paired correlations. PCA creates a low-dimensional representation of samples from a data set while preserving as much variance as possible. It identifies underlying attributes that are the most variable across all samples and cluster samples into different groups. Dendrogram analysis was used to cluster the components of the soaked CBR to provide a visual representation of the groups of the data. Correlation matrix and paired correlations were also used to show all the methods demonstrate a similar conclusion. Using PCA and dendrogram analysis allows key factors to be identified. This information can be crucial in making informed decisions to reduce project risk.

CBR is used in design and (incorrectly) assumed to be closely related to the compaction density which is used in earthworks quality control testing. These analyses show for this CH clay material the after-swell moisture ratio governs the CBR test value, and the density ratio at compaction is least related to the CBR value. The application of dendrogram analysis to field data in Look (2021 and 2023) similarly show compaction density testing may be misleading due its poor relationship with modulus. By adopting such hierarchical clustering analysis, data noise is removed and key risk factors for geotechnical modelling can be identified.

References

Alsanabani, N.M., Al-Gahtan, K. S, Alsharef, A. and Almohsen, A. S. (2025). Identifying Pile Installation Risks Using Principal Component Analysis. *ASCE Journal of Structural Design and Construction Practice*, 30(1)

Look, Burt G. (2021). An earthworks quality assurance methodology which avoids unreliable correlations. *4th International Conference on Transportation Geotechnics*, Chicago, USA. In: Tutumluer, E., Nazarian, S., Al-Qadi, I., Qamhia, I.I. (eds) Advances in Transportation Geotechnics IV. Lecture Notes in Civil Engineering, vol 166. pp 179 -192, Springer, https://doi.org/10.1007/978-3-030-77238-3_14

Look, Burt G. (2023). Earthworks testing and the density illusion. *Proceedings of the 14th Australia and New Zealand Conference on Geomechanics*, Cairns 2023 (ANZ2023)

Nazarian S, Mazari M, Abdallah I, Puppala I, Mohammad L, and Abu-Farsakh M, (2014). Modulus-Based Construction Specification for Compaction of Earthwork and Unbound Aggregate. Draft Final report, *National Cooperative Highway Research Program* NCHRP Project 10-84, Transportation Research Board

Nagoya, A., Unno, T., Sakamoto, R., Kamura, A. (2025). Feature Value Evaluation of Compaction Property of Fine Aggregate by Principal Component Analysis. In: Rujikiatkamjorn, C., Xue, J., Indraratna, B. (eds) *Proceedings of the 5th International Conference on Transportation Geotechnics (ICTG)* Sydney, Australia, 2024, Volume 6. ICTG 2024. Lecture Notes in Civil Engineering, vol 407. Springer, Singapore. https://doi.org/10.1007/978-981-97-8233-8_28

StatTools (2016). User's Guide – Statistics *Add-In for Microsoft Excel*, Version 7. Palisade Corporation