Porous Media Classification Using Multivariate Statistical Methods

M. Elmorsy, W. El-Dakhakhni, and B. Zhao

1 Introduction

Subsurface porous media characterization is important in many natural and industrial processes such as groundwater movement, oil extraction, and geologic CO_2 sequestration. Classifying the type of porous media (e.g., sandstone, carbonate) is often the first step in the characterization process, and it provides critical information regarding the physical properties of the porous media. Conventionally, trained experts classify subsurface porous media samples via laboratory analyses [22]. More recently, advances in remote sensing technologies such as laser-induced breakdown spectroscopy (LIBS) have made in-situ characterization of porous media possible, whereas computed micro-tomography (µCT) technologies have made characterization of porous media samples much more efficient [4, 14]. For example, modern desktop X-ray µCT machines are capable of scanning a rock sample in as little as a few minutes. As a result, we now have unprecedented access to three-dimensional (3D) visualizations of various subsurface materials, which are readily available in online repositories [15].

Digital classification of porous media samples is now possible via the combination of imaging, chemical analysis and multivariate statistical methods. Multivariate statistical methods analyze the common behaviour of multiple independent variables, and they include principal component analysis (PCA), soft independent modeling of class analogy (SIMCA), and partial least squares discriminant analysis (PLS-DA). They have been utilized to analyze and classify porous media samples based on their chemical composition, textural features, pore characteristics, and physical properties.

M. Elmorsy (✉) · W. El-Dakhakhni · B. Zhao
Department of Civil Engineering, McMaster University, Hamilton, ON L8S 4L7, Canada
e-mail: elmorsym@mcmaster.ca

S. Walbridge et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2021*, Lecture Notes in Civil Engineering 249,
https://doi.org/10.1007/978-981-19-1061-6_35

Multivariate statistical method based classification make in-situ classification of porous media possible, in situations where sample collection and retrieval is exorbitantly expensive or infeasible. For example, Sirven et al. [17] studied the feasibility of remotely identifying rocks on the Martian surface, using LIBS spectra data and multivariate methods including PCA, SIMCA, and PLS-DA. Their results show that SIMCA outperforms PLS-DA in discriminating materials that share similar features, however, combination of both models achieves the highest classification rate (97%). Kim et al. [9] used PCA and PLS-DA methods to classify three types of soil samples based on LIBS spectral data. Guang et al. [6] classified different rocks and soils using PLS-DA and support vector machine (SVM) algorithms based on LIBS spectra. Similarly, Xie et al. [22] used PLS-DA for soil type identification using near-infrared (NIR) spectra. Lepistö et al. [11], and Kachanubal and Udomhunsakul [8] utilized neural networks to categorize rocks into homogeneous and non-homogenous groups based on their color and textural features using RGB images. Valentín et al. [19] used Naïve Bayes classifier to classify rock textures based on 31 different combinations of 520 textural and spectral features. In order to reduce the computational cost of the classification process, they used PCA to reduce the problem dimensionality followed by a genetic algorithm to define the most statistically significant input configuration.

While LIBS spectra-based analysis has shown success in classifying rocks and soils based on their chemical composition [6], LIBS cannot provide precise information about the inner domain structure. μCT scans capture 3D information of porous media's inner structure with micron-scale precision, enabling precise characterization at the pore-scale. Adhikari et al. [1] studied the variability of CT-measured pore characteristics and physical properties of three soil samples obtained from different locations. The CT-measured pore characteristics are macroporosity, mesoporosity, number of pores, circularity and fractal dimension, while the soil physical properties are bulk density, hydraulic conductivity, sand, silt and clay content. They employed PCA to perform a redundancy analysis that reduced pore features, and soil physical properties into three principal components. Their analysis shows that the soil porosity and the sample number of pores are the most governing characteristics in constructing the principle components.

Here, we present a fast and robust data-driven model for rock classification using 3D μCT images. We find OPLS is the most efficient at extracting latent variables of domain features (e.g., porosity, convexity, etc.) from μCT images compared to other commonly used methods such as PCA, SIMCA, and PLS. In addition, our work provides quantitative insights into the homogeneity of the rock sample, and uncovers the relative influence between different domain features in rock classification, which improves our understanding of rock formation and evolution.

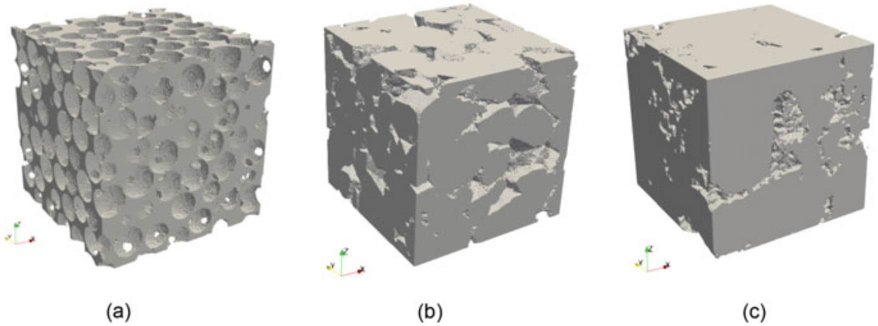

Fig. 1 3D visualization of μCT scans of **a** synthetic rock, **b** sandstone, and **c** limestone

2 Data

2.1 Description

The μCT imaging dataset used in our analysis is obtained from Muljadi et al. [13] and it is accessible via the digital rock portal [15]. Specifically, the dataset contains a 1500^3 μm^3 sandstone sample (3 μm/pixel), a 1655^3 μm^3 limestone sample (3.31 μm/pixel), and a 1000^3 μm^3 synthetic rock sample (2 μm/pixel) (Fig. 1). The synthetic rock was stochastically generated such that the pore bodies are spherical. The synthetic rock sample has homogeneous pore sizes in all spatial directions, while the sandstone and limestone samples have spatially heterogeneous pore sizes and irregular pore shapes.

2.2 Pre-processing and Feature Visualization

We perform pre-processing of the data by slicing the segmented 3D μCT scans and obtaining 500 sequential, equally-spaced 2D binary images for each sample dataset (Fig. 2). We then transform the 2D binary images of the porous media to extract the contours of the void space. The contours are used to calculate different geometric features including perimeter, area, porosity, convexity, moments, etc. (Fig. 3). Convexity is the ratio between the contour area and its convex hull area, where a convex hull of a contour is the minimum perimeter that contains the contour (e.g. a convex contour will have a convexity ratio = 1, while a concave contour will have convexity ratio <1). Convex hull descriptor is an important geometrical feature to detect shapes' similarities and it has been used for a variety of computer vision applications [7]. Similarly, moments and functions of moments are common contour-based shape features used in object recognition [12]. We calculated three groups of moments, (i) spatial moments: m01, m02, m03, m10, m11, m12, m20,

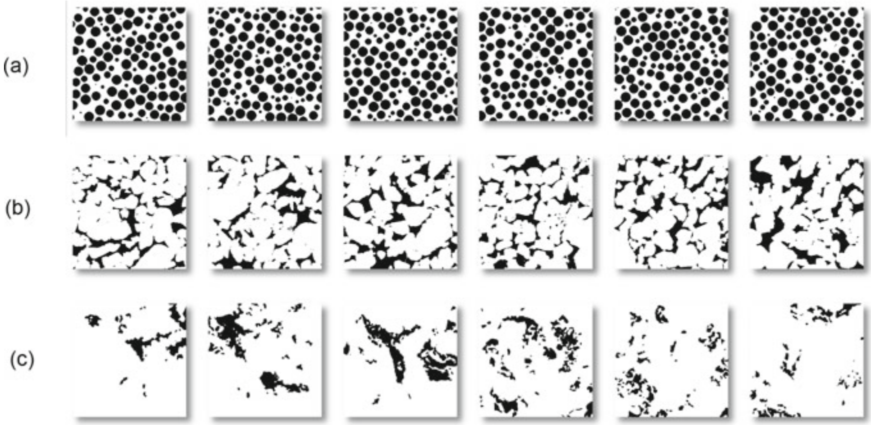

Fig. 2 Sequential 2D slices of 3D μCT scans of **a** synthetic rock, **b** sandstone, and **c** limestone. The void fractions of the porous media are shown in black, while the solid fractions are shown in white

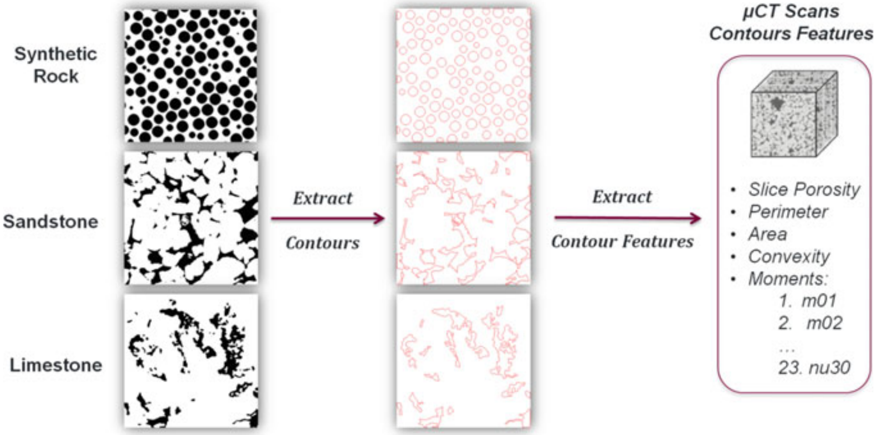

Fig. 3 We transform the 2D binary images of the porous media to extract the contours of the void space. We extract different geometric features of the contours including perimeter, area, porosity, convexity, moments, etc. using the OpenCV library in Python [3]

m21, m30, (Eq. 1), (ii) central moments: mu02, mu03, mu11, mu12, m20, mu21, mu30, (Eq. 2) and (iii) normalized central moments: nu02, nu03, nu11, nu12, nu20, nu21 and nu30 (Eq. 3). Their mathematical derivations are defined by,

Spatial moments (m_{ji}):

$$m_{ji} = \sum_x \sum_y x^j y^i I(x, y) \tag{1}$$

where I(x, y) = contour pixel intensity.

Central moments $\left(\mathrm{mu}_{\mathrm{ji}}\right)$:

$$mu_{ji} = \sum_{x}\sum_{y}(x - \overline{x})^{j}(x - \overline{y})^{i} I(x, y) \tag{2}$$

where $\overline{x}$, $\overline{y}$ are the contour centroid: $\overline{x} = \frac{\mathrm{m}_{10}}{\mathrm{m}_{00}}$, $\overline{y} = \frac{\mathrm{m}_{01}}{\mathrm{m}_{00}}$.

Normalized central moments $\left(\mathrm{nu}_{\mathrm{ji}}\right)$:

$$nu_{ji} = \frac{mu_{ji}}{m_{00}^{(i+j)/2+1}} \tag{3}$$

Then, the slice porosity for each contour was calculated, and the contour perimeter to area ratio (P/A). The final multivariate dataset consists of 3000 contours (i.e., 1000 contours for each porous media) and the extracted features. We divide this dataset into a training set and a testing set, whose sizes have a ratio of 4:1.

We develop a correlation heat map to visualize the correlation matrix between all rock features and the rock types (Fig. 4). Correlation matrix is a useful tool for exploratory data analysis, it inform us the degree and direction of correlation between data variables and their corresponding target. Correlation heat maps transform the correlation matrix information and present them in a visually appealing format. The developed correlation heat map shows that porosity and convexity are the most strongly correlated to the rock type, while area and moments (m01-3, m10-12, m20-21, m30, mu30, nu02 and nu11) exhibits intermediate correlation to the rock type. The remaining features (P/A ratio, mu03, mu12, mu20-21, mu30, nu30, nu12 and nu20-21) have weak correlation to the rock type. This illustrates that not all pore features have a statistically significant correlation to the rock type.

3 Methodology

We test three multivariate statistical methods—principal component analysis, partial least squares, and orthogonal partial least squares—to classify porous media samples using the geometric features extracted from pre-processing of the 3D μCT scans. We briefly describe each of the methods below.

Principal component analysis (PCA) is a mathematical procedure that reduces the dimensionality of large datasets by finding linear combinations of large number of correlated variables to create a smaller number of uncorrelated variables (orthogonal to each other), known as principal components, while preserving as much variability as possible [20]. PCA is commonly used as a pre-processing step to reduce dimensionality of multivariate data prior to using machine learning classification algorithms [16]. In addition to the PCA popularity as a dimensionality reduction technique, it is also a useful technique for data visualization and feature discovery [10]. Based

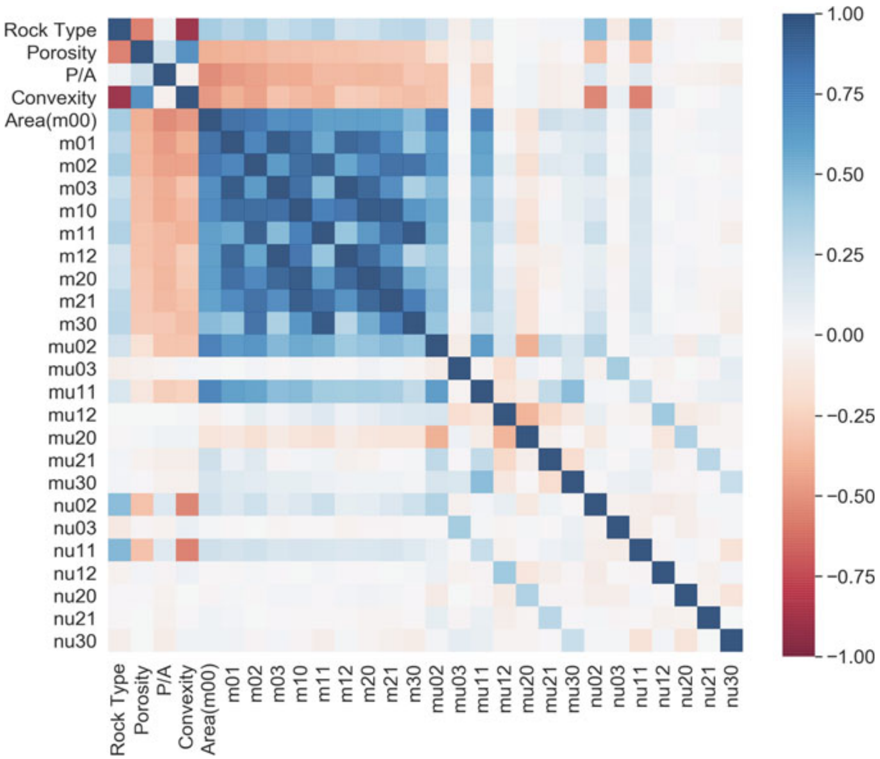

Fig. 4 Correlation heat map of the correlation matrix between the extracted rock features and rock types. The color bar displays the correlation coefficient. Dark blue indicates strong positive correlation, dark red indicates strong negative correlation and white indicates no correlation exists

on the PCA technique, the soft independent modeling of class analogy (SIMCA) method is used for complex classification tasks where a single PCA model does not encompass all of the dataset's variability. SIMCA is collection of PCA models, in which each class in the dataset has its own PCA model [21]. Here, we first develop a PCA model for each rock type, by fitting and calculating the scores and loadings of the model components. Then, we use the developed models to classify porous media samples in a supervised manner by projecting them onto each PCA model and calculating the corresponding residual. The porous media sample is classified as the rock type that yields the lowest residual and that is within its statistical limit. Partial least squares (PLS) is a mathematical method that finds latent variables of dataset variables and sequentially extracts each component. PLS differs from PCA in that it uses two blocks of data, X (variables) and Y (target), where X is used to predict Y, and Y can have multiple variables. PLS maximizes the relationship between X and Y while explaining the best variability in both X and Y, where scores and loadings are calculated for both blocks simultaneously [2]. PLS can handle multi-class datasets using one model and it use cross-validation to check the number of components. We assign the dummy variable Y to each porous media sample (synthetic rock = 0, sandstone = 1, limestone = 2) in the training process. Orthogonal PLS (OPLS) method divides variability in the X block into systematic variability and residual variability. OPLS further splits systematic variability into two parts—one part is correlated to the Y block (predictive), while the other part is uncorrelated to Y (orthogonal). By evaluating the variation explained in each PLS component, OPLS can filter the systematic variability extracted from the input block but not related to the Y block. Therefore, an effective model with reduced complexity is obtained while maintaining the predictability of the model [18]. As a result, usually, one or two components are enough to represent variation when using the OPLS technique. By the end, rock features components generated by PLS and OPLS models were used in conjunction with discriminant analysis. The multivariate data analyses were carried out using SIMCA® software (Version 16.0, Umetrics).

4 Results

4.1 PCA Analysis

PCA model was developed and R^2 and Q^2 for each component of the X matrix in the training set were calculated. R^2 is the percentage of the variance explained by the model. It indicates how well the model fits the data. Q^2 is the percentage of the variance of the training set predicted by the model according to cross validation. By increasing the number of components, the value of R^2 increases for incorporating more variability, however, the value of Q^2 has decreased after the second component. In fact, Q^2 is calculated the same way as R^2, but it is applied on validation set which is not used in fitting the model. Therefore, by increasing more components and including more variability representing more variables from the data, this leads to over-fitting due to noise. This shows that weakly correlated moment variables discussed in the data visualization section contribute negatively to Q^2 of the model. So by adding more components to explain more variability of the X matrix this weakness the model's predictive capability since it incorporate more noise. The highest Q^2 was recorded at the first component and equals to 0.31 with R^2 equals to 0.344. Then, the t_1-t_2 scores plot was developed and it shows good separation between the synthetic rock, and the two natural rocks, sandstone and limestone; however the natural rocks scores are totally overlapped (Fig. 5).

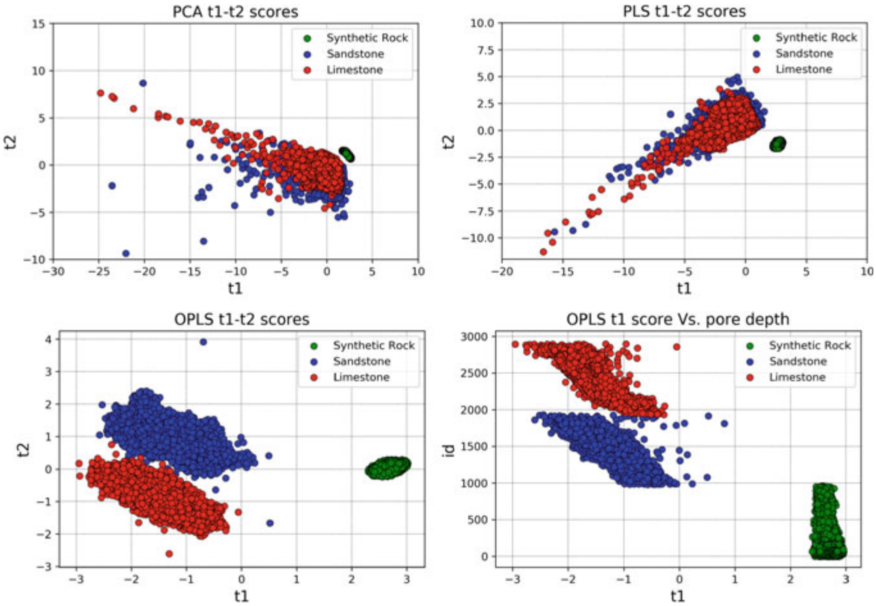

Fig. 5 Scores plot for PCA, PLS and OPLS models. The t_1-t_2 PCA scores were the most scattered and have a near complete overlap of the sandstone and limestone scores, making the separation task between these two rocks difficult; however, it is easier to separate the synthetic rock samples. Similarly, the PLS t_1-t_2 scores show similar distribution with narrower ranges and less outliers. On the other hand, the OPLS t_1-t_2 scores have a neat separation in the t_2 scores dimension, with few outliers. Finally, the OPLS t_1-id plot illustrates two information regarding the homogeneity of rocks, (i) the synthetic rock has a narrow t1 range for a given rock slice indicating close pore features values; on the other hand, real rocks t_1 scores have a wider range indicating wider pore features values range on the slice level. (ii) the synthetic rock t_1 scores almost stabilize when the pore depth increases in the X direction; unlike, real rocks t_1 scores that shift to the left (negative direction), indicating change in the features values. This concludes that, the synthetic rock has a homogeneous pore structure, whilst, real rocks are heterogeneous

4.2 SIMCA Classification Model

We perform supervised classification of porous media samples using an SIMCA model. SIMCA is collection of PCA models, each fitted for a specific rock type. In the classification process, we calculate the average orthogonal distance of the test sample to each model. The orthogonal distance is the Euclidian distance of the test sample to the PCA model of a given class. If the orthogonal distance of a new sample was found to be within the class model border (or below its statistical limit) then the sample belongs to that class, and vice versa. Similarly, to the PCA model, by increasing the number of components for each class model, we obtain a higher R^2 value, however Q^2 value decreases. The best fit models were used to classify rock samples by using two components for sandstone, limestone and synthetic rock models. The SIMCA classification recorded accuracy of 96.63% for training set and

Table 1 SIMCA model classification results on the testing set using two components

Rock type	Pores	Synthetic rock	Sandstone	Limestone	No class	Classification rate
Synthetic rock	200	189	11	0	0	0.945
Sandstone	200	0	197	1	2	0.985
Limestone	200	0	17	182	1	0.91
Total	600	189	225	183	3	0.9467

94.67% for testing set. When the incorporated PCA components were increased to three for the three models, the training and testing accuracies decreased to 96.17 and 92.83% respectively (Table 1).

4.3 PLS-DA Classification Model

PLS-DA model was developed and it was found that by increasing the incorporated components, the R^2 and Q^2 values increases. It was also noted that using two components was not enough to have a robust model since the cumulative R^2, Q^2 values for the second component were below 0.5, however, by adding the third component, the cumulative R^2 and Q^2 values has increased significantly reaching nearly 0.7 for both of them. Also, the scores of the first and second PLS components recorded overlap for the sandstone and limestone samples, as displayed Fig. 5. This emphasized that using only the first two PLS components was not enough for efficient classification. The classification results also matched that observation, since with using only the first two PLS components, total training and testing accuracies recorded 68.63 and 70.5%, respectively, with significant low limestone classification accuracies equals to 29.25 and 34% for the training and testing samples respectively. To further increase the classification accuracies, seven PLS components had to be used to achieve a strong classification, recording training and testing accuracies equals to 98.17 and 98% respectively (Tables 2 and 3).

Table 2 PLS-DA model classification results on the testing set using the first two components

Rock type	Pores	Synthetic rock	Sandstone	Limestone	No class	Classification rate
Synthetic rock	200	200	0	0	0	1
Sandstone	200	1	155	44	0	0.775
Limestone	200	2	130	68	0	0.34
Total	600	203	285	112	0	0.705

Table 3 PLS-DA model classification results on the testing set using the first seven components

Rock type	Pores	Synthetic rock	Sandstone	Limestone	No class	Classification rate
Synthetic rock	200	200	0	0	0	1
Sandstone	200	1	197	2	0	0.985
Limestone	200	0	9	191	0	0.955
Total	600	201	206	193	0	0.98

4.4 OPLS-DA Classification Model

Finally, we developed an OPLS-DA model that achieved cumulative R^2 and Q^2 equal to 0.846 and 0.839 respectively, by using only two components. In addition, the first two components had a cumulative $R^2_{(X)}$ equal to only 0.163, however their contribution to $R^2_{(Y)}$ (Predictive—Y) is equal to 1. This means that only 16.3% of variance in X matrix (rock features) was enough to predict 100% of Y (rock types), and the remaining variance was orthogonal to Y. The first component contributed to 100% of the synthetic rock samples' variance, and below 50% for the sandstone and limestone variance. However, the second component contributed only to explain the most variability in the sandstone and limestone samples. The scores of the first and second components are plotted in Fig. 5, and they showed strong separating capability. The first component could strongly separate between the synthetic rock, and the other two real rock samples, while the second component could separate between the sandstone and limestone samples. The discriminant analysis classification results recorded accuracies of 97.96% and 97.17% on the training and testing sets respectively, outperforming SIMCA and PLS-DA models when using only two principal components. In order to study the homogeneity of the samples, the t_1 scores that represent the highest X variance (pore features) that is predictive to Y (rock types) were plotted versus the contours id, where they are indexed according to their depth order in the X direction in each rock sample. It was noted that synthetic rock pores have smaller variation range of t_1 scores, and they do not change significantly with changing the pore location. This means all the synthetic rock pores have similar characteristics and hence it is a homogenous sample. On the contrary, the sandstone and the limestone had a wider variation range of t_1 scores and display a shift in t_1 scores as the contours depth increases. This means that the characteristics of the sandstone and limestone samples pores change in the three spatial directions, and thus they are heterogeneous. Similar patterns were also found in the t_2 scores values. These findings agree with our observations from visualizing the samples' slices in Fig. 2, but using t1 and t2 scores enables analyzing rock homogeneity in a quantitative way (Table 4).

Finally, we investigated the variables influence in predicting the rock types by utilizing the variable influence in projection (VIP) method. VIP method evaluate the influence of each variable in the projection used in the model by calculating VIP scores that is often used for variable selection in multivariate analysis for big

Table 4 OPLS-DA model classification results on the testing set using the first two components

Rock type	Pores	Synthetic rock	Sandstone	Limestone	No class	Classification rate
Synthetic rock	200	200	0	0	0	1
Sandstone	200	1	195	4	0	0.975
Limestone	200	0	12	188	0	0.94
Total	600	201	207	192	0	0.9717

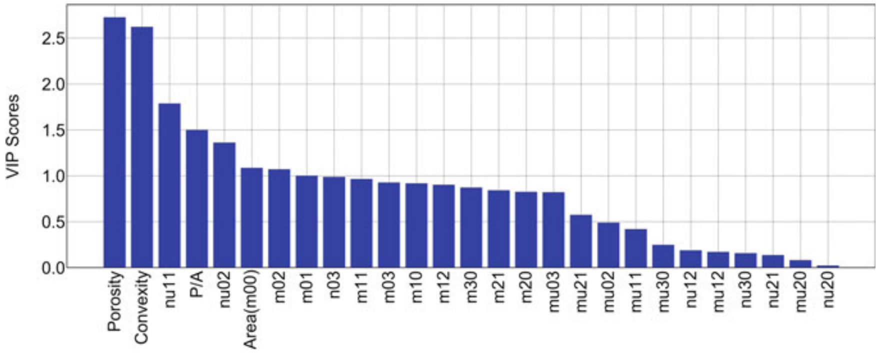

Fig. 6 Variable influence in projection (VIP)

data. Variables with higher VIP scores indicate higher influence in the projection in the model and consequently higher influence in predicting the class of a new data. Specifically, a variable with a VIP Score close to or greater than one (> = 1) can be considered influential in a given model, on the contrary, variables with VIP scores substantially less than one (<1) are less influential and might be discarded from the model. For more details about the VIP theory, the reader can refer to [5]. Here we developed the rocks variables VIP scores for the predictive components in the OPLS model, and we illustrate them in Fig. 6. It is illustrated that porosity and convexity have the highest influence on the model with VIP scores greater than 2.5 (>2.5), relatively, less influential variables are nu11 moment and P/A ratio with VIP scores greater than 1.5 (>1.5), and nu02 moment, area and mu02 moment with scores greater than one (>1). The remaining variables are found to be insignificant to the model projection by having VIP scores less than one (<1).

5 Conclusions

In this paper, we carried out a classification analysis over digital rocks dataset, including a synthetic rock and two natural rocks, sandstone and limestone. The digital rocks volumes were sliced in the X direction, and their pore structure contours were extracted, followed by calculating handcrafted features out of them. We employed four data reduction methods to perform multivariate statistical analysis, namely PCA, SIMCA, PLS-DA and OPLS-DA. The PCA model showed a poor capability to explain the whole dataset in one single model. On the other hand, the SIMCA model demonstrated higher prediction capability by building a single PCA model for each rock type. We found the PLS model to be more capable of dealing with the analyzed dataset because of its ability to deal with noise and uncorrelated variables to the sample class. However, its t1-t2 plot showed that the first 2 PLS components were not enough to separate rock samples effectively. On the other hand, the OPLS method effectively filtered the variables that did not correlate to the sample class. It resulted in only two components that predicted 100% of the Y matrix (rock type) while explaining only 16.3% of the X matrix (rock features) variance. The t1-t2 scores plot of the OPLS model showed a near clear separation between all rock types with few outliers. That emphasized the OPLS capability in filtering the uncorrelated variables to the classified classes, and consequently, it uses less information with higher prediction ability. Therefore, the OPLS-DA model achieved the highest classification rate equals to 0.9717 when using only 2 components. Finally, we utilized the VIP method to uncover the most influential rock features in classifying rock types. Variables such as porosity, convexity, P/A ratio, nu02 and nu11 moments were revealed to be the key features to distinguish rock types.

References

1. Adhikari P, Anderson SH, Udawatta RP, Kumar S (2016) Analysis of CT-measured pore characteristics of porous media relative to physical properties. Procedia Comput Sci 95:442–449
2. Barker M, Rayens W (2003) Partial least squares for discrimination. J Chemom 17(3):166–173
3. Bradski G (2000) The OpenCV library. Dr. Dobb's J Softw Tools 120:122–125
4. Díaz Pace DM, Gabriele NA, Garcimuño M, D'Angelo CA, Bertuccelli G, Bertuccelli D (2011) Analysis of minerals and rocks by laser-induced breakdown spectroscopy. Spectrosc Lett 44:399–411
5. Galindo-Prieto B, Eriksson L, Trygg J (2014) Variable influence on projection (VIP) for orthogonal projections to latent structures (OPLS). J Chemom 28:623–632
6. Guang Y, Shujun Q, Pengfei C, Yu D, Di T (2015) Rock and soil classification using PLS-DA and SVM combined with a laser-induced breakdown spectroscopy library. Plasma Sci Technol 17(8):656–663
7. Jayaram M, Fleyeh H (2016) Convex hulls in image processing: a scoping review. Am J Intell Syst 6(2):48–58
8. Kachanubal T, Udomhunsakul S (2008) Rock textures classification based on textural and spectral features. Int J Geotech Geol Eng 2(3):658–664
9. Kim G, Kwak J, Kim K-R, Lee H, Kim K-W, Yang H, Park K (2013) Rapid detection of soils contaminated with heavy metals and oils by laser induced breakdown spectroscopy (LIBS). J Hazard Mater 263:754–760
10. Landgraf AJ, Lee Y (2020) Dimensionality reduction for binary data through the projection of natural parameters. J Multivar Anal 180:104668
11. Lepistö L, Kunttu I, Autio J, Visa A (2003) Rock image classification using non-homogenous textures and spectral imaging. In: Skala V (ed) WSCG 2003, the 11th international conference in central Europe on computer graphics, visualization and computer vision 2003, short papers, 3–7 February 2003. Czech Republic, Iniversity of West Bohemia, Plzen, pp 82–86

12. Mercimek M, Gulez K, Mumcu TV (2005) Real object recognition using moment invariants. Sādhanā 30(6):765–775
13. Muljadi BP, Blunt MJ, Raeini AQ, Bijeljic B (2016) The impact of porous media heterogeneity on non-Darcy flow behaviour from pore-scale simulation. Adv Water Resour 95:329–340
14. Pak T, Archilha NL, Mantovani IF, Moreira AC, Butler IB (2019) An X-ray computed micro-tomography dataset for oil removal from carbonate porous media. Sci Data 6:190004
15. Prodanovic M, Esteva M, Hanlon M, Nanda G, Agarwal P (2015) Digital rocks portal: a sustainable platform for imaged dataset sharing, translation and automated analysis.
16. Reddy GT, Reddy MPK, Lakshmanna K, Kaluri R, Rajput DS, Srivastava G, Baker T (2020) Analysis of dimensionality reduction techniques on big data. IEEE Access 8:54776–54788
17. Sirven J-B, Sallé B, Mauchien P, Lacour J-L, Maurice S, Manhès G (2007) Feasibility study of rock identification at the surface of Mars by remote laser-induced breakdown spectroscopy and three chemometric methods. J Anal At Spectrom 22(12):1471–1568
18. Trygg J, Wold S (2002) Orthogonal projections to latent structures (O-PLS). J Chemometr J Chemometr Soc 16(3):119–128
19. Valentín M, Bom C, Albuquerque M, Albuquerque M, Faria E, Correia M, Surmas R (2017) On a method for rock classification using textural features and genetic optimization. Notas Técnicas 7(1):18–30
20. Wold S, Esbensen K, Geladi P (1987) Principal component analysis. Chemom Intell Lab Syst 2(1–3):37–52
21. Wold S, Sjöström M (1977) SIMCA: a method for analyzing chemical data in terms of similarity and analogy. Chemometr: Theor Appl 52(12):243–282
22. Xie H, Zhao J, Wang Q, Sui Y, Wang J, Yang X, Zhang X, Liang C (2015) Soil type recognition as improved by genetic algorithm-based variable selection using near infrared spectroscopy and partial least squares discriminant analysis. Sci Rep 5(1):10930