SEMI-AUTOMATIC SUPERVISED CLASSIFICATION OF MINERALS FROM X-RAY MAPPING IMAGES

Allan A. Nielsen[1], Harald Flesche[2], Rasmus Larsen[1], Johannes M. Rykkje[2] and Mogens Ramm[3]

[1]IMM, Department of Mathematical Modelling, Technical University of Denmark,
DK-2800 Lyngby, Denmark, aa@imm.dtu.dk
[2]Norsk Hydro Research Centre, N-5020 Bergen, Norway, Harald.Flesche@nho.hydro.com
[3]Norsk Hydro Technology and Competence, N-1321 Stabekk, Norway

SUMMARY

This paper addresses the problem of classifying minerals common in siliciclastic and carbonate rocks. Twelve chemical elements are mapped from thin sections by energy dispersive spectroscopy (EDS) in a scanning electron microscope (SEM). Traditional multivariate statistical methods and extensions hereof are applied to perform the classification. First, training and validation sets are grown from one or a few seed points by a method that ensures spatial and spectral closeness of observations. Spectral closeness is obtained by excluding observations that have high Mahalanobis distances to the training class mean. Spatial closeness is obtained by requesting connectivity. Second, class consistency is controlled by forcing each class into 5-10 subclasses and checking the separability of these sub-classes by means of canonical discriminant analysis. Third, class separability is checked by means of the Jeffreys-Matusita distance and the posterior probability of a class mean being classified as another class. Fourth, the actual classification is carried out based on four supervised classifiers all assuming multi-normal distributions: simple quadratic, a contextual quadratic and two hierarchical quadratic classifiers. Overall weighted misclassification rates for all quadratic classifiers are very low, for both the training (0.25%-0.33%) and validation sets (0.65%-1.13%). Finally, the number of rejected observations in routine runs is checked to control the performance of the SEM image acquisition and the classification. Although the contextual classifier performs (marginally) best on the validation set, the simple quadratic classifier is chosen in routine classifications because of the lower processing time required. This method is presently used as a routine petrographical analysis method at Norsk Hydro Research Centre.

1. INTRODUCTION

Mineral classification and quantification is traditionally done using point counting of thin sections or by x-ray diffraction (XRD). The first of these methods is very time consuming and requires a trained petrographer; the latter does not give any spatial information about the samples being analysed. Point counting also has an element of subjectivity in that a more skilled petrographer will be better at recognising rare minerals, separating cement from detrital grains, etc. A third method is to do x-ray mapping or energy dispersive spectroscopy (EDS) in a scanning electron microscope (SEM). Here, an x-ray spectrum is acquired for each pixel. By means of this spectrum the most likely mineral for each pixel can be identified using a manual or automatic classification method. Spatial information about the mineral composition can thereby be obtained by an objective and reproducible method. Earlier work in this field has used classification methods that range from lookup table to maximum likelihood classification, [2,10,13]. Earlier, long image acquisition times has made use of EDS images for mineral classification difficult.

New equipment enables acquisition of a 256$\times$256 pixels image with 12 elements mapped in 36 minutes. Each pixel is 2.4 μm$\times$2.4 μm.

This paper addresses a method for classification of EDS images. Due to space limitations, we put emphasis on the training and validation set generation only. A fuller description of the entire procedure is found in [5].

2. DATA, MINERALS AND ELEMENTS

As the aim is to use the method for standard studies of sedimentary rocks, it is important to cover the most frequently occurring minerals. Table 1 shows all minerals in the model. In some cases more than one class is needed to describe a mineral. This can be due to natural variation in the chemical composition of the mineral, such as in the biotites, the chlorites, the garnets and the siderites. There is also a case, for illite and muscovite, where it is known in advance that different minerals have the same chemical composition and therefore will be overlapping in the EDS measurements. In addition to minerals it is also important to measure the porosity in the samples. The mapped elements reflect the major components in the minerals. It is normally the K_α line that is mapped, but in some instances this is superimposed by another element's L_α line. This is the case for P and Zr, Ti and Ba, and Na and Zn. The set of mapped elements is shown in Table 1.

Albite	Chlorite 2	Gypsum	Quartz
Ankerite	Chlorite 3	Illite/Muscovite	Rutile
Apatite	Dolomite	Ilmenite	Siderite 1
Barite	Fe-calcite	Kaolin	Siderite 2
Biotite 1	Garnet 1	K-feldspar	Titanite
Biotite 2	Garnet 2	Monazite	Tourmaline
Calcite	Garnet 3	Porosity	Zincblende
Chlorite 1	Glauconite	Pyrite	Zircon

Al	Fe	Mn	S
C	K	Na (+Zn)	Si
Ca	Mg	P (+Zr)	Ti (+Ba)

Table 1: Mineral classes in the model (left) and mapped elements (right)

3. TRAINING SET GENERATION AND CONSISTENCY CHECK

Good supervised classification is contingent on good training sets. This is not always obtained with training sets drawn by human operators. This is partly due to the human inability to obtain an overview of multidimensional spaces. Another problem with training sets drawn by humans is inconsistency. Training sets need to be extracted in a consistent way across time and classes irrespective of operator and shape of image structures. Therefore we propose a new semi-automatic algorithm for generation of a set of training classes from a series of seed points. For each class the operator needs only supply one or a few observations. From these points training classes are grown in a fashion which ensures both spatial and spectral closeness. Spatial closeness is obtained by demanding that all pixels in one training class be connected with the initial seed point(s). This is a very useful condition, because most relevant phenomena appear as connected objects. The connectivity may be defined in terms of first-order, second-order neighbours etc. This definition has the advantage of being able to allow for small holes in the training sets by specifying that pixels should be higher order neighbours. This is useful for classes that occur as clusters of smaller objects in the image, and also in the case of classes that occur as thin strings or layers. Spectral closeness is achieved by making restrictions on the distance to the current mean value of the class while growing the training set. Here two distance measures are considered. One is the Euclidean spectral distance

$$D_E^2 = (\boldsymbol{x} - \boldsymbol{\mu}_i^*)^T (\boldsymbol{x} - \boldsymbol{\mu}_i^*),$$

where $\boldsymbol{x} = [x_1, \ldots, x_n]^T$ is the value observed in a pixel, and $\boldsymbol{\mu}_i^*$ is the current estimate of the class mean. The application of Euclidean distance to seed-growing is suggested in [3]. The other distance measure used is the Mahalanobis distance

$$D_M^2 = (\boldsymbol{x} - \boldsymbol{\mu}_i^*)^T \boldsymbol{\Sigma}_i^{*-1} (\boldsymbol{x} - \boldsymbol{\mu}_i^*),$$

where $\boldsymbol{\Sigma}_i^*$ is the current estimate of the class dispersion matrix. For the Euclidean distance, an upper limit for the distance should be supplied by the user. The Mahalanobis distance is χ^2-distributed with n degrees of freedom. This enables us to use a systematic threshold defined by e.g. the 0.99 quantile, which is a major advantage over the use of Euclidean distance.

If the seed growing begins with a single pixel we cannot get an estimate of the dispersion matrix. Therefore we first grow the seed point to an initial training set using the Euclidean distance method with a preset maximum distance, spectrally and spatially. From the small number of pixels thus included an estimate of the dispersion matrix is obtained. This estimate may first be used to exclude any outliers in the current set, and second, to grow the training set further from the initial training set using the Mahalanobis distance. The application of this method gives us training classes that are in good correspondence with Gaussian distributions.

Training and validation data should be checked for consistency to make sure that the multivariate data in each assumed class make up one class only. Here, a method based on a partitioning of the training and validation data for each class into five to ten sub-classes by means of an unsupervised clustering algorithm is used. First, observations called cluster seeds are selected as a first guess of the sub-class means. Second, clusters are formed by assigning observations to the nearest seed as measured by Euclidean distance. After all observations are assigned, new cluster means are calculated. This step is repeated until changes in cluster means become zero (or small). This clustering is followed by a canonical discriminant analysis, which combines the original variables into new orthogonal variables called canonical discriminant functions (CDFs) which are the best possible linear discriminators between the sub-classes into which the training and validation data have been clustered. If a scatter plot of the first two CDFs shows no outliers and no sign of grouping, the training and validation data are considered as being consistent.

To construct a training area only one pixel is chosen as the starting point. In the first step, a radius of 5 pixels is normally sufficient to get an initial estimate of the mean value and dispersion matrix of the class. For more spatially dispersed minerals, such as chlorite 2 (which is a chlorite coating on grains) and illite/muscovite it is necessary to increase the radius to 10. The Euclidean distance that is used as an acceptance limit is normally set to 30. Barite, K-feldspar, zincblende and zircon need a Euclidean distance of 50 in order to include a sufficient number of pixels. The resulting (validation) area for chlorite 2 is seen in Figure 1.

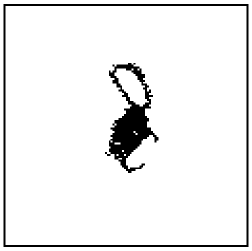

Figure 1: Validation area for chlorite 2 as grown with the new algorithm

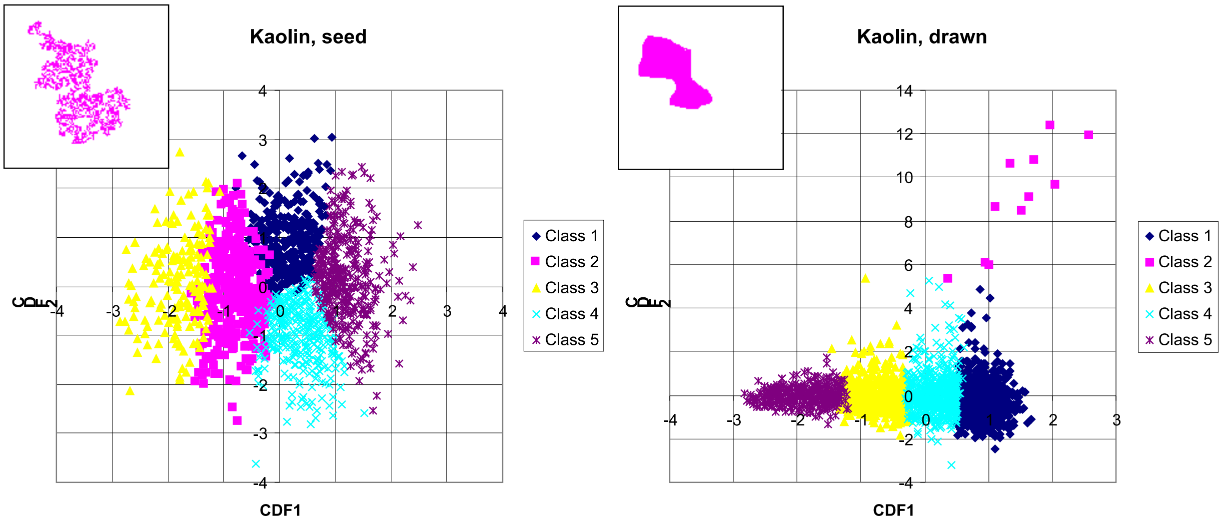

Figure 2: Comparison of seed-grown and hand-drawn training areas and scatterplots of first two CDFs for kaolin

The resulting training area for kaolin is seen in Figure 2 (left) compared with a typical hand-drawn training area (right). Scatter plots of the two first CDFs from most training areas show a consistent set, with no sub-groups and with little scatter around the main cluster. This is contrary to attempts to create training areas by painting an area in the image, which is what most classification packages encourage. Poor definition of the structure of the real class and inclusion of much noise is often the result from such an approach. Figure 2 shows these scatter plots of hand-drawn training area compared to a seed-grown training area for kaolin.

4. CLASSIFICATIONS

Simple quadratic, contextual quadratic, hierarchical and extended hierarchical quadratic classifications all assuming multi-Gaussian data are used. For descriptions of statistical methodology, see [1,4,5,6,7,8,9,11,12]. The results of the analyses are shown by an example of two classification images, Figure 3. In the centre-left part of the left image we see how quartz cement (light grey) engulf a pyrite grain (very bright). The pore space between the grains is mostly filled with clays. In the right image we see chlorite (dark gray) coating the pores (lighter gray). Overall misclassification rates for 36 images (not shown) are seen in Table 2 which also shows relative CPU time requirements. As we cannot discriminate between 1) calcite and Fe-calcite (maybe due to high acceleration voltage in the SEM favouring good resolution in the lightest elements at the expense of resolution in the heavy elements), and 2) the three garnet classes, we combine

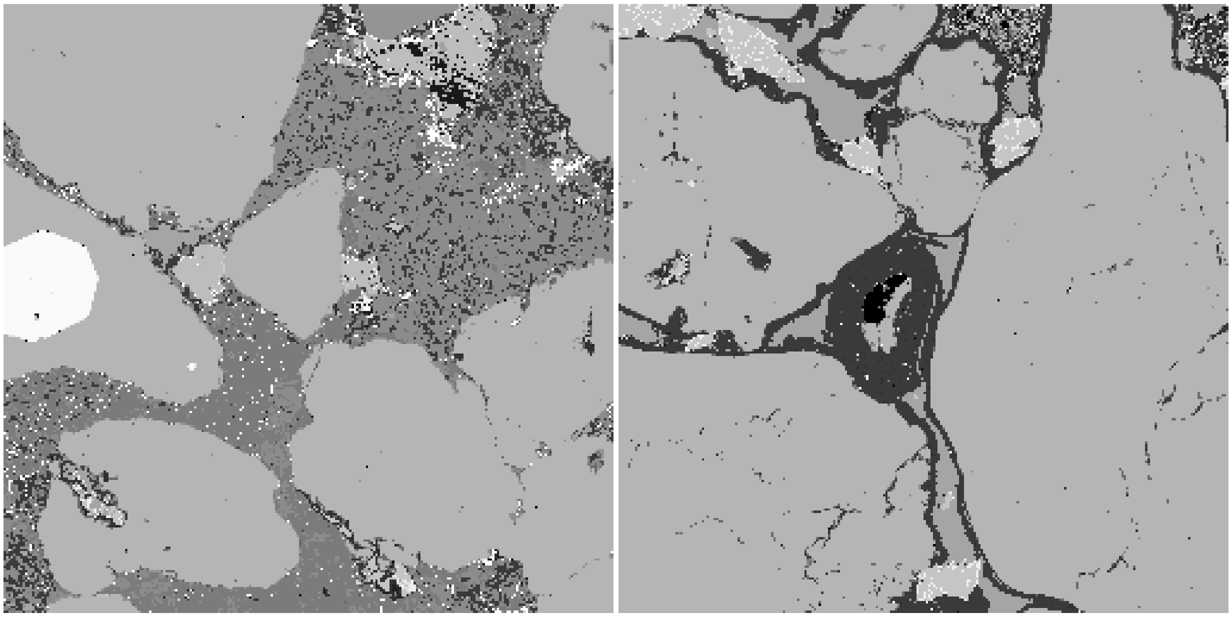

Figure 3: Visual appearance of typical classification result

calcite and Fe-calcite into one class and the three garnets into one class. Biotite in the validation set turned out to be altered and is removed from the validation set.

	Quadratic	Context. Quad.	Hierarch.	Ext. Hierarch.
Training set	0.0025	0.0033	0.0025	0.0025
Validation set	0.0106	0.0065	0.0109	0.0113
CPU time	1	5.48	0.47	3.49

Table 2: Overall misclassification rates (with calcite and Fe-calcite combined, the three garnet classes combined and biotite removed from the validation set), and relative CPU times

5. CONCLUSIONS

Based on a novel method for semi-automatic training and validation data generation very successful classifications of the most frequently occurring minerals in sedimentary rocks are obtained. Overall weighted misclassification rates for all classifiers applied are very low for both the training (0.25%-0.33%) and validation (0.65%-1.13%) data. The methods used are likely to perform with a similar success when applied to other types of data such as space or airborne remote sensing data.

ACKNOWLEDGMENTS

The authors would like to thank Professor Knut Conradsen and Dr. Bjarne K. Ersbøll, both IMM, for many good discussions on multivariate statistics and Johan Doré, also IMM, for coding part of the seed growing algorithm.

REFERENCES

1. ANDERSON, T.W. - “*An Introduction to Statistical Analysis*” (second edition). John Wiley, New York. 1984.

2. CLELLAND, W. D. and FENS, T.W. - ’Automated rock characterization with SEM/Image analysis technique’, SPE Formation Evaluation, pp. 437-443, 1991.

3. ERDAS Inc. - “*ERDAS Version 7.4*”. 1990.

4. ERSBØLL, B.K. - “*Transformations and Classifications of Remotely Sensed Data: Theory and Geological Cases*”, Ph.D. thesis, Department of Mathematical Modelling, Technical University of Denmark, 1989.

5. FLESCHE, H., NIELSEN, A.A. and LARSEN, R. - ’Supervised mineral classification with semi-automatic training and validation set generation in scanning electron microscope energy dispersive spectroscopy images of thin sections’. In prep.

6. HASLETT, J. - ’Maximum likelihood discriminant analysis on the plane using a Markovian model of spatial context’, *Pattern Recognition*, vol. 18, no. 3, pp. 287-296, 1985.

7. HJORT, N.L. and MOHN, E. - ’A comparison of some contextual methods in remote sensing classification’. In The 18th International Symposium on Remote Sensing of Environment, Paris, France, 1984.

8. HJORT, N.L., MOHN, E. and STORVIK, G. - ’Contextual classification of remotely sensed data, based on an auto-correlated model’. In H.V. Sæbø, K. Bråten, N.L. Hjort, B. Llewellyn, and E. Mohn, editors, Contextual classification of remotely sensed data: Statistical methods and development of a system, Norwegian Computing Center, Technical report no. 768, 1985.

9. JIA, X. and RICHARDS, J.A. - ’Feature reduction using a supervised hierarchical classifier’. In The 8th Australasian Remote Sensing Conference, Canberra, Australia, 1996.

10. MINNIS, M.M. - ’An automatic point-counting method for mineralogical assessment’, *AAPG Bulletin*, vol. 68, no. 6, pp. 744-752, 1984.

11. OWEN, A. ’A neighbourhood-based classifier for LANDSAT data’, *The Canadian Journal of Statistics*, vol. 12, pp. 191-200, 1984.

12. SAFARIAN, S.R. and LANDGREBE, D. - ’A survey of decision tree classifier methodology’, *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, no. 3, pp. 660-674, 1991.

13. TOVEY, N.K. and KRINSLEY, D.H. - ’Mineralogical mapping of scanning electron micrographs’, *Sedimentary Geology*, vol. 75, pp. 109-123, 1991.

14. WELCH, J.R. and SALTER, K.G. - ’A context algorithm for pattern recognition and image interpretation’, *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 1, pp. 24- 30, 1971.