Comparison of pattern-recognition techniques for classification of Silurian rocks from Lithuania based on geochemical data

Donatas Kaminskas & Björn A. Malmgren

Kaminskas, D. & Malmgren, B.A. 2003: Comparison of pattern-recognition techniques for classification of Silurian sedimentary rocks from Lithuania based on geochemical data. *Norwegian Journal of Geology*, Vol. 84 , pp. 117-124. Trondheim 2003. ISSN 029-196X.

There is no widely recognized chemical classification of sedimentary rocks. The geochemical classification of sedimentary rocks utilizes features that can be observed in hand specimens or in thin sections, such as grain size and the mineralogy of the particles and matrix. The main objective behind this paper is to compare the performance of different pattern-recognition techniques, such as artificial neural networks, linear discriminant analysis, the k-nearest neighbour technique and "soft independent modelling of class analogy" in classifying Silurian sedimentary rocks from Lithuania. The comparison was made separately for major and trace elements. To obtain an idea about the success of the various classifiers in correctly predicting samples from each of the individual rock types, error rates were also computed for each of seven petrographically established rock types. For testing the predictive power of the artificial neural networks we applied a back propagation network. Error rates were computed based on the average percentage of misclassifications. The comparison of pattern-recognition techniques used in this study indicates that not only the technique for classification should be applied with care but also that attention must be paid to the objects (rock types) being classified. It was also noticed that not only the major elements, as is usually the case, could be used in "recognition" of certain rock types. Trace elements could also successfully be used and readily handle the same tasks. The results indicate that two techniques, artificial neural networks and linear discriminant analysis, yield the lowest error rates for both major and trace elements (7-8% for the major elements and 8-14% for the trace elements). The k-nearest neighbour technique and soft independent modelling of class analogy produce considerably higher error rates for both types of elements (16-25%). Since no strict statistical assumptions, for example, multivariate normality of a data set, are required in the artificial neural networks, this procedure seems to be the optimum technique for geochemical classification of the sedimentary rocks.

Kaminskas, D., Department of Geology and Mineralogy, Faculty of Natural Sciences, Vilnius University, Ciurlionio str. 21, 2009-LT Vilnius, Lithuania (e-mail: donatas.kaminskas@gf.vu.lt); Malmgren, B.A., Department of Earth Sciences, Göteborg University, Box 460, SE-405 30 Göteborg, Sweden (e-mail: bjorn.malmgren@marine-geology.gu.se)

Introduction

There is no widely recognized chemical classification of sedimentary rocks. The geochemical classification of sedimentary rocks is not as well developed as that for igneous rocks, and most systems for sedimentary rock classification utilize features such as grain size and the mineralogy of the particles and matrix (Rollinson 1995), which can be observed in hand specimens or in thin sections. In cases where the geochemistry of particular sedimentary rocks is being studied it is always useful to find "boundaries" that demarcate certain rock types or, at least, separate them on the basis of major and/or trace elements. Separation of rock types is very important in investigations of the relationships among chemical elements, their associations in certain rock types etc.

Calcitic, dolomitic and terrigenous materials are the major constituents of most carbonate rocks and mudrocks. These constituents could aid in the establishment of a general classification. Mineralogical investigations of sedimentary rocks, combining thin sections, XRD along with major elements like Ca, Mg, Si and Al, could provide a more reliable classification. However, in many cases, determinations of major and trace element concentrations are not always accompanied by thin section or XRD studies. The handling of geochemical information (for example, geochemical variables – chemical elements) is most common nowadays. Considering the fact that the nature of most real-world data is very complex and that the relationships among variables (for example, chemical elements) are nonlinear, it is essential to employ an appropriate technique that could handle such data in a statistically "correct" fashion. Whereas multivariate-statistical approaches always produce the same result when applied to the same data set, a technique like the artificial neural network (ANN) is more like a living system in that various analyses most likely will not produce exactly the same result. ANNs have been applied to solving problems in a wide variety of fields. Applications of

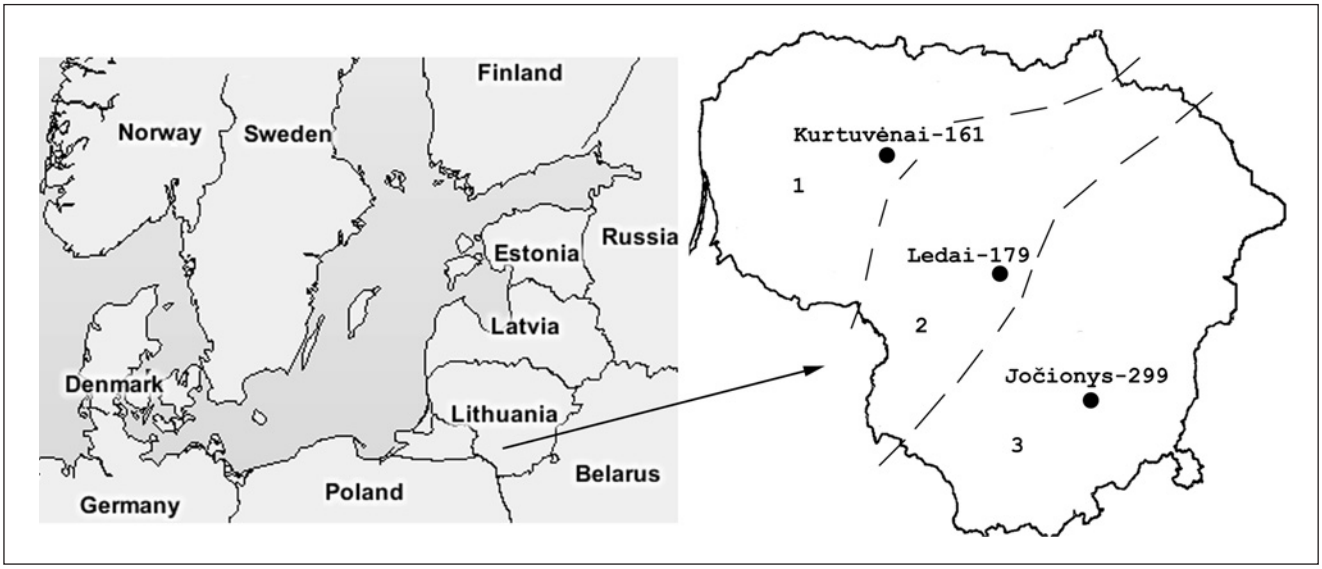

Fig. 1. Study boreholes: location and facial zonation of Wenlock (Silurian) sedimentary environments: 1 – the deepest part; 2 - intermediate zone; 3 – the shallowest part. Dotted lines mark the boundaries between sedimentary environments zones.

ANNs to the Earth sciences are still rare. They have been applied, for example, to problems of well log interpretations (Baldwin et al. 1989, 1990; Rogers et al. 1992), for identifications of linear features in satellite imagery (Penn et al. 1993), for geophysical inversion problems (Raiche 1991), for chemostratigraphy (Malmgren & Nordlund 1996), for predictions of past sea-water temperatures (Malmgren et al. 2001) and for analyses of paleovegetation data (Grieger 2002).

For comparison of the performance of different techniques applicable to classification of sedimentary rocks based on geochemical properties we explored the potential of four pattern-recognition techniques, ANNs, the k-nearest neighbour technique (k-NN), linear discriminant analysis (LDA) and "soft independent modelling of class analogy" (SIMCA). The study was based on geochemical data from three boreholes penetrating parts of the Silurian of Lithuania.

Material and methods

The three boreholes analysed here, Kurtuvėnai-161, Ledai-179 and Jočionys-299, represent different sedimentary environments of the Wenlock (Silurian) of Lithuania (Fig. 1). Initially, samples were collected to cover a wide range of rock types according to their lithology. In total, 210 samples were taken: 89 from the Kurtuvėnai-161, 69 from the Ledai-179 and 52 from the Jočionys-299 boreholes. We used 179 of these samples in this study; the remaining 31 samples were not included, since some rock types contained only a few samples. The deepest part of the sedimentary basin environment is represented by the Kurtuvėnai-161 borehole, the shallowest by the Jočionys-299 borehole and the transitional by the Ledai-179 borehole (Paškevičius 1997).

The samples were analyzed for the content of major (Si, Al, Fe, Mn, Mg, Ca, Na, K, Ti and P) elements, oxides and trace elements (V, Cr, Co, Ni, Cu, Zn, Rb, Pb, Ba, Sr, Y, Zr, Nb and Th) using XRF. Inorganic and organic carbon contents were determined using IR spectrometry. All geochemical analyses were carried out at the Geological Institute, Oslo University, Norway.

A raw semi-quantitative classification of the seven rock types was made according to the calcitic, dolomitic and terrigenous material present in the samples (Table 1). This simplified classification table is usually used for classification of carbonate rocks and mudrocks. The details of this mineralogical classification and its principles can be found in Grigelis *(ed.)* 1981.

Initially, the rock samples were classified according to the general description of the core material. Thin sections and X-ray diffraction were applied to obtain a more detailed classification. Detailed descriptions of sampling, analytical techniques and thin-section studies can be found in Kaminskas (2002).

Brief descriptions of the quantitative techniques

Artificial neural networks (ANNs)

ANNs are computer systems that have the ability to learn, using some pertinent learning algorithm, one or several output signals from a set of input signals. The

Table 1. The simplified classification table of carbonate rocks and mudrocks according calcite, dolomite and terrigenous material* percentages.

Rock type	Calcite (%)	Dolomite (%)	Terrigenous material (%)	Remarks
limestone "clayey"	65-90	0-10	10-25	
limestone "clayey" dolomitic	50-80	10-25	10-25	Terrigenous material>Dolomite
limestone dolomitic "clayey"	50-80	10-25	10-25	Dolomite>Terrigenous material
mudstone	>33.3	0-10	25-50	
mudstone dolomitic	>33.3	25-50	25-50	Calcite>Dolomite
dolostone	0-10	80-100	0-10	
dolostone "clayey"	0-10	65-90	10-25	

* - terrigenous material equals: 100-(calcite+dolomite)

objective behind the application of ANNs is to attempt at reproducing the output signal(s) from the input signals with a minimum error rate through a specific training process. ANNs have the ability to overcome problems of fuzzy and nonlinear relationships between the sets of input and output signals. The initial data-set is divided into two random portions, a training set, which is used for training the ANN, and a test set to which the trained network is applied for estimates of the error rate.

An ANN is an information-processing system inspired by the way the densely interconnected, parallel structure of the mammalian brain processes information. The ANN is composed of a great number of processing elements that are analogous to neurons and are tied together with weighted connections that are analogous to synapses. The most common type of ANNs is the multilayer perceptron, which is most often trained using the back propagation (BP) algorithm. The ANN is trained to reproduce the target variable(s) from the input variables by adaptively updating the synaptic weights that are associated with the strength of the connections. Learning in a BP network is based on the gradient-descent method, that is, the weights are adjusted so that the changes at each time will follow the steepest "downhill" direction on the error surface. The optimum weights are thus determined iteratively by optimizing certain "energy" functions as training proceeds. Comprehensive descriptions of multilayer perceptron ANN can be found in Wasserman (1989), Webb (2002) and Malmgren & Nordlund (1996, 1997).

We used the Trajan 4.0 Professional software package (www.trajan-software.demon.co.uk) in our ANN applications.

Linear discriminant analysis (LDA)

Normally, discriminant analysis amounts to establishing linear functions, representing a planar surface with p-one dimensions (p=the number of variables), that optimally distinguish two predefined groups of observations. In this study, we assigned each of the observation vectors in the various test sets to one of the predefined rock groups through computations of Mahalanobis' generalized distances between these vectors and group mean vectors (Cooley & Lohnes 1971). These generalized distance measures were then converted to mathematical probabilities of referability of a test-set observation to any of the predefined groups (Cooley & Lohnes 1971). Each of the test-set observations was subsequently assigned to the group for which the probability was highest. Malmgren & Kennett (1977) applied this procedure to a taxonomic problem in recent planktonic foraminifera.

K-nearest neighbors (k-NN)

The k-nearest neighbor is a conceptually simple technique based on the Euclidean distance between observations in multidimensional space (Kowalski & Bender 1972). The allocation of the test-set members to the training set classes is dependent upon the distances of the k shortest Euclidean distances between these sets. In the application to the current data-set, we set k equal to 3, and we monitored the distances from each of the test set members to each of the training set members. A test-set member is referred to that training-set class to which the majority (two or three) of the three closest training set members belonged, as indicated by the Euclidean distances.

Soft independent modeling of class analogy (SIMCA)

SIMCA may involve any or several of four distinct levels (Wold 1976; Wold et al. 1984). Level 1 of SIMCA is devoted to developing mathematical rules for each of a number of preset groups (termed classes in SIMCA) in the training set by fitting separate R-mode principal component models to each of them. The dimensionality of a principal component solution is determined by a cross-validation technique. In level 2, the prediction phase, these rules are used to assign new observations to any of the given classes on the basis of their degree of fit to the various class models using a distance measure. At both levels, atypical observations ("outliers"), that is, observations with a data structure that does not accord with a class model, may be identified. In this way, observations of unknown affinity that cannot be classified with any training set class may be interpreted as being referable to a yet unknown class.

Levels 3 and 4 of SIMCA are designed for quantitative predictions of one or several variables from a multivariate setup through partial least squares (PLS) models (Wold 1982). These levels of SIMCA are not used here. So far, SIMCA modeling, originally developed for data analysis in the field of chemometry, has been applied to geological problems by, for example, Griffiths (1984), Haugen et al. (1989), and Wei (1994).

Canonical Variates Analysis (CVA)

CVA is a technique for graphical representation of the interrelationships among groups of multivariate samples, like the rock groups analysed here, on the basis of plots of the group means and individual sample points along specific coordinate axes (the canonical variates; Reyment et al. 1984). These canonical variate axes are computed so as to maximize the ratio of the between-to within-group variance and to be uncorrelated in canonical variate space.

Estimates of Error Rates

The success of a classifier may be determined by computing the percentage of misclassifications, the "error rate," of predictions in a data set which is not part of the training set. Instead of relying on a single test set for estimating the performance of the various classifiers, which may be misleading (Weiss & Kapouleas 1989), we created five different random test sets from our original data. Each such test set contained 20% (37 particles) of the original observations. The remaining 80% (146 particles) was used as training sets. We then employed a cross-validation technique for estimating the ability of the classifiers to correctly predict the class referability of the test set samples (Stone 1974; Weiss & Kapouleas 1989). The error rates computed are thus average rates of misclassification (%) for the five different test sets. The training and test sets are automatically generated by the Trajan ANN software, and the same partitions were used to derive the error rates also for the LDA, k-NN, and SIMCA.

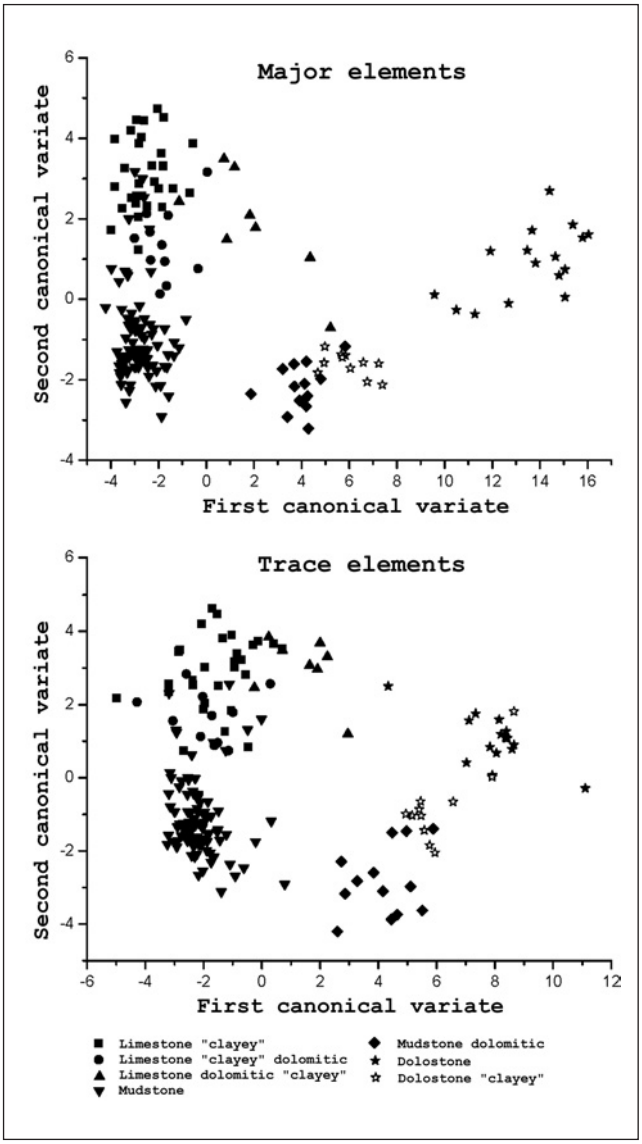

Fig. 2. Graphical illustration of the configurations of the individual samples from the seven petrographically identified rock types along the axes of the first two canonical variates for the major and trace elements. These axes account for 93.5 and 94.2%, respectively, of the variability in multivariate space.

Relative abundance data

Relative abundance data, adding up to a unit value (unity or 100%) in individual samples, have long been known to be subject to the so-called constant-sum constraint (Pearson 1897; Chayes 1960). In our applications of CVA, LDA and SIMCA to the major element data-set in which the correlation structure of the data

matrix suffers from the constant-sum constraint, which could seriously impair the outcome of these analyses, we used a centred log-ratio transformation to relieve this constraint (Aitchison 1981, 1986). Log-ratio transformations imply mathematical operations in a so-called simplex space, constituting a limited part of the original p-dimensional Cartesian space. The results of k-NN and ANNs are not affected by the constant-sum constraint, since they do not involve computations of a covariance or correlation matrix, and in the applications of these techniques the raw, untransformed data were used. The data-set used was the same for all these techniques.

Results

Canonical variates analysis

Figure 2 shows the locations of individual rock samples along the axes of the first two canonical variates for the major and trace elements, respectively. These canonical variates account for most of the variability among the group mean vectors (93.5 and 94.2%, respectively). For the major elements, the dolostone samples can be clearly distinguished from the other rock samples along the first canonical variate axis. The mudstone dolomitic and dolostone "clayey" samples are likewise separated from most of the other samples along the first axis but display some slight overlap with the limestone dolomitic "clayey" samples. Along the second axis, most of the mudstones are separated from the limestone "clayey", limestone "clayey" dolomitic, limestone dolomitic "clayey" and mudstone samples, but some of the mudstone samples overlap with these other rock groups. The limestone "clayey", limestone "clayey" dolomitic and limestone dolomitic "clayey" samples cannot be unequivocally differentiated on the basis of their geochemical compositions.

The dolostones cannot be as clearly separated from the dolostone "clayey" samples on the basis of the trace elements as in the case of the major elements. For the trace elements the dolostone "clayey" also display a considerable overlap with the mudstone dolomitic samples. The limestone "clayey", limestone "clayey" dolomitic, limestone dolomitic "clayey" and mudstone samples show overlaps similar to the situation for the major elements, even though most of the mudstone and limestone dolomitic "clayey" samples are also distinguishable from the limestone "clayey" and limestone dolomitic "clayey" samples by their trace elements.

Considering the lack of discrete subclusters for both the major and trace elements it is relevant to ask the question of how well the ANN and statistically based pattern recognition techniques are able to distinguish these various rock types.

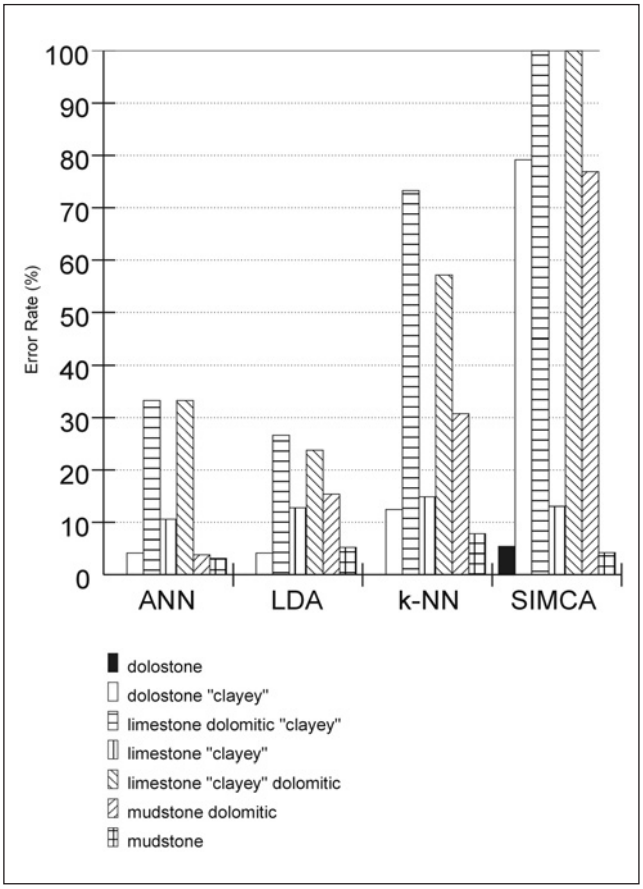

Fig. 3. Error rates (percentages of misclassification in the various rock types) for the artificial neural network (ANN), linear discriminant analysis (LDA), the k-nearest neighbours (k-NN) and soft independent modeling of class analogy (SIMCA) using major elements

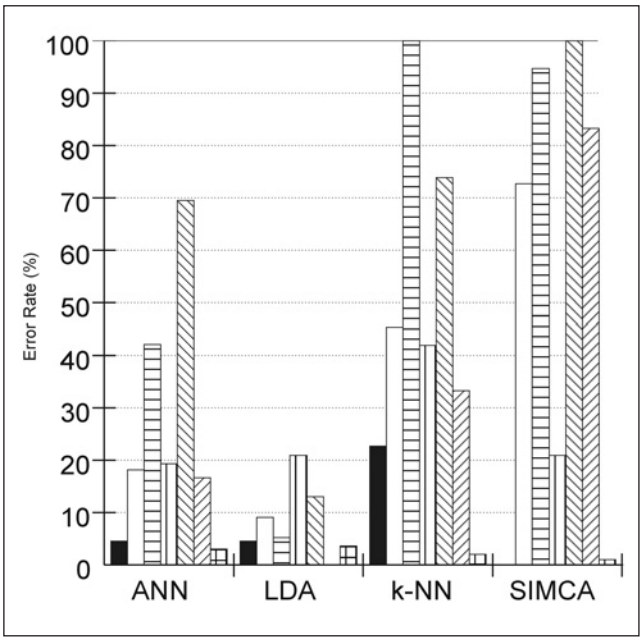

Fig. 4. Error rates (percentages of misclassification in the various rock types) for the artificial neural network (ANN), linear discriminant analysis (LDA), the k-nearest neighbours (k-NN) and soft independent modeling of class analogy (SIMCA) using trace elements. For legend, please refer to Fig. 3.

Table 2. Error rates in each of five test set partitions, average rates in the test sets, and 95% confidence intervals for the average error rates for the techniques discussed in this paper using major elements.

	Error rates, %						
	Set1	Set2	Set3	Set4	Set5	Average	95%
ANN	8.3	4.2	5.6	9.7	6.9	**6.9**	±2.7
LDA	8.3	6.9	6.9	11.1	6.9	**8.1**	±2.3
k-NN	12.5	23.6	13.9	15.3	12.5	**15.6**	±5.8
SIMCA	19.4	27.8	27.8	20.8	27.8	**24.7**	±5.2

Table 3. Error rates in each of five test set partitions, average rates in the test sets, and 95% confidence intervals for the average error rates for the techniques discussed in this paper using trace elements.

	Error rates, %						
	Set1	Set2	Set3	Set4	Set5	Average	95%
ANN	13.9	13.9	11.1	15.3	15.3	**13.9**	±2.1
LDA	6.5	9.7	5.6	9.7	8.3	**8.0**	±2.3
k-NN	19.4	25.0	29.2	19.4	27.8	**24.2**	±5.7
SIMCA	20.8	23.6	31.9	25.0	19.4	**24.2**	±6.1

Major and trace elements

Individual and average error rate percentages for ANN, LDA, k-NN and SIMCA are given in Tables 2 and 3 for major and trace elements, respectively. For the major elements two techniques, k-NN and SIMCA, show the highest mean error rates (15.6 and 24.7%, respectively; Table 2). The error rates for the individual test sets range between 12.5 and 23.6% for k-NN and between 19.4 and 27.8% for SIMCA.

The lowest mean error rates for the major elements are achieved applying ANN (6.9%) and LDA (8.1%). The overall better performances of ANN and LDA as compared to k-NN and SIMCA are also indicated by the error rates for individual test sets, which range between 4.2 and 9.7% for ANN and between 6.9 and 11.1% for LDA. Hence, in terms of average error rates ANN and LDA provide better results than k-NN and SIMCA (Table 2).

For the trace elements the results are similar: again the highest mean error rates are for k-NN (24.2%) and SIMCA (24.2%; Table 3). For these techniques the error rates for the individual test sets fluctuate between 19.4 and 29.2% for k-NN and between 19.4 and 31.9% for SIMCA. Similar to the major elements, the lowest mean error rates are obtained for ANN and LDA (13.9 and 8.0% respectively). Individual errors are also clearly much improved for the ANN (11.1-15.3%) and LDA (5.6-9.7%) compared to k-NN and SIMCA.

A comparison of the performance of the techniques for major and trace elements indicates that two of them, LDA and SIMCA, produce similar results for both types of elements (about 8 and 24-25%, respectively). On the other hand, the ANN and k-NN yield considerably lower error rates for the major elements (about 7 and 16%) as compared to the trace elements (about 14 and 24%, respectively).

From these results we conclude that k-NN and SIMCA do not handle the partitioning of the rock types in an optimal fashion. The LDA and ANN give a better result than the k-NN and SIMCA models whether major or trace elements are used as variables.

Error rates for individual rock types

Error rates were also computed for each of seven rock types. The results, presented in Figs. 3 and 4, are evaluated for both major and trace elements.

The error rates in limestone varieties (dolomitic "clayey" and "clayey" dolomitic) range from 33.3 to 100% (Fig. 3). In contrast, the lowest error rate is obtained for dolostone (0.0-5.4%). Error rates in "clayey" limestone and mudstone range from 10.6 to 14.9% and from 3.2 to 7.9% respectively. This may indicate that the variation of the major elements in these rock types is quite stable. It suggests that the terrigenous ("clay") component was not that much affected by post-depositional events. However, the major elemental composi-

tion of the sedimentary rocks largely depends on the two major components - carbonate and terrigenous (not discussed in this paper). From Figure 3 it is apparent that the most successful rock type classification for the major element data could be achieved by applying the ANN and LDA techniques. The overall error rate in individual classes (rock types) ranges from 0.0 to 33.3% (ANN) and from 0.0 to 26.7% (LDA). The k-NN technique as well as SIMCA fail to provide good classification (Fig. 3).

Most trace elements (Zr, V, Rb, Th, U etc) usually show high correlation with the terrigenous material, and only Sr is related to carbonate content in the sedimentary rocks studied (Kaminskas 2001a, 2001b, 2002). The error rate bars presented in Figure 4 suggest that neither k-NN nor SIMCA yield satisfactory results in predicting rock types. Almost the same pattern of the error rate bars for the trace elements (Fig. 4) is recognized for the major elements (Fig. 3) as well. In the mudstone the lowest error rate ranges from 0.5 to 3.6%. Dolostone could be "recognized" with an error of 4.6% by ANN, and 22.7% by k-NN (Fig. 4). Accordingly, for the trace element data only the ANN and LDA managed to classify all the rock types studied, except limestone dolomitic "clayey" and limestone "clayey" dolomitic. The k-NN and SIMCA show extremely poor results when handling trace element data.

The use of trace elements in classifying sedimentary rocks is a very important feature. ANN and LDA demonstrate that even if only trace element data are available a reasonably good classification is still possible.

Discussion and conclusions

We applied ANN, k-NN, LDA and SIMCA in order to analyse the rock classification based on mineralogical information and to compare the performance of these techniques. It was important to assess which of the techniques that would enable minimum misclassification. The results demonstrate that classification is not only dependent on the trace or major elements that were taken into account (not discussed here), but that they are also dependent on the statistical pattern-recognition technique that is chosen.

The comparison of the statistical pattern-recognition techniques used in this study has shown that not only the technique for classification should be used with care but also that attention must be paid to the objects (rock types) being classified. It was also noticed that not only the major elements, as is usually the case, could be used in the "recognition" of certain rock type. Trace elements could also be used successfully and readily handle the same tasks.

Prior to applying a statistical pattern-recognition technique the mineralogical composition of the formation should be investigated. As described above both ANN and LDA could be successfully applied for carbonaceous sedimentary rock classification, except for some limestone types. Moreover, ANN is a more flexible technique, and results could be improved by additional "learning" and modification of the network configuration.

Several conclusions can be made from the results obtained in this study. Unfortunately, the comparison of the pattern-recognition techniques has demonstrated that classification success depends not only on variables involved in calculations but also on the pattern recognition technique chosen. Since no strict statistical assumptions, e.g. multivariate normality of the datasets, is required in the artificial neural networks this procedure seems to be one of the optimum techniques for geochemical classification of the sedimentary rocks. Linear discriminant analysis could also be utilized, but the results obtained by this technique are very dependent on the data structure (multivariate normality of the datasets). The k-nearest neighbour technique and soft independent modelling of class analogy were not found to produce optimum error rates when compared to artificial neural networks and linear discriminant analysis. The conclusion is that the application of the artificial neural network technique for distinguishing Silurian rock types on the basis of geochemical data seems to be the best choice among the techniques compared.

Acknowledgments: - The senior author would like to thank the Research Council of Norway and Swedish Royal Academy of Sciences for financial support. Prof. Dr. N. Spjeldnaes, Dr. B.L. Berg, Dr. T. Winje and M. Naoroz from Institute of Geology, Oslo University (Norway) deserve warm compliments for their advice and support.

References

Aitchison, J. 1981: A new approach to null correlations of proportions. *Journal of International Mathematical Geology 13,* 175-189.

Aitchison, J. 1986: *The statistical analysis of compositional data.* Chapman & Hall, New York, 416 pp.

Baldwin, J.L., Otte, D.N. & Wheatley, C.L. 1989: Computer emulation of human mental process: application of neural network simulations to problems in well log interpretation. *Society of Petroleum Engineering 19619,* 481-493.

Baldwin, J.L., Bateman, R.M. & Wheatley, C.L. 1990: Application of neural network to the problem of mineral identification from well logs. *Log Analysis 3,* 279-293.

Chayes, F. 1960: On correlations between variables of constant sum. *Journal of Geophysical Research 65,* 4185-4193.

Cooley, W.W. & Lohnes, P.R. 1971: *Multivariate Data Analysis.* John Wiley, New York, 364 pp.

Grieger, B. 2002: Interpolating paleovegetation data with an artificial neural network. *Global and Planetary Change 34,* 199-208.

Griffiths, C.M. 1984: A pattern recognition approach to the derivation of geological information from drill process monitoring. Underwater Technology, *Winter,* 2-13.

Grigelis, A (ed.) 1981: *Metodicheskiye rekomendacii po sostavleniu legend krupnomashtabnych geologicheskich kart Pribaltiki. Kolektiv avtorov.* Tallinn. 237 pp. (in Russian).

Haugen, J.-E., Sejrup, H.-P., & Vogt, N.B. 1989: Chemotaxonomy of Quaternary benthic foraminifera using amino acids. Journal of Foraminiferal Research 19, 38-51.

Kaminskas, D. 2001a: Geochemical pecularities of the Upper Llandovery and Wenlock (Silurian) rocks in Kurtuvėnai-161 borehole (NW Lithuania). *Geologija 33,* 3-9.

Kaminskas, D. 2001b: Geochemical pecularities of the Wenlock (Lower Silurian) rocks in Ledai-179 and Jočionys-299 boreholes (E. Lithuania). *Geologija* 35, 3-14.

Kaminskas, D. 2002: Geochemistry of Wenlock (Silurian) rocks of Lithuania. Doctoral dissertation, Vilnius University, 145 pp.

Kowalski, B.R. & Bender, C.F. 1972: The K-nearest neighbour classification rule (pattern recognition) applied to nuclear magnetic resonance spectral interpretation. *Analytical Chemistry 44,* 1405-1411.

Malmgren, B.A. & Kennett, J.P. 1977: Biometric differentiation between Recent *Globigerina bulloides and Globigerina falconensis* in the southern Indian Ocean. *Journal of Foraminiferal Research 7,* 130-148.

Malmgren, B.A. & Nordlund, U. 1996: Application of artificial neural networks to chemostratigraphy. *Paleoceanography 11,* 505-512.

Malmgren, B.A. & Nordlund, U. 1997: Application of artificial neural networks to paleoceanographic data. *Palaeogeography, Palaeoclimatology, Palaeoecology 136,* 359-373.

Malmgren, B.A., Kucera, M., Nyberg, J. & Waelbroeck, C. 2001: Comparison of statistical and artificial neural network techniques for estimating past sea surface temperatures from planktonic foraminifer census data. *Paleoceanography 16,* 520-530.

Paškevičius, J. 1997: *Geology of the Baltic Republics.* Vilnius university & Geological survey of Lithuania, Vilnius, 387 pp.

Pearson, K. 1879: Mathematical contributions to the theory of evolution: On a form of spurious correlation which may arise when indices are used in the measurements of organs. *Proceedings of Royal Society 60,* 489-498.

Penn, B.S., Gordon, A.J. & Wendlandt, R.F. 1993: Using neural networks to locate edges and linear features in satellite images. *Computing Geoscience 19,* 1545-1565.

Raiche, A. 1991: A pattern recognition approach to geophysical inversion using neural nets. *International journal of Geophysics 105,* 629-648.

Rao, C.R. 1970: *Advanced Statistical Methods in Biometric Research.* Hafner, Darien, Conn., 390 pp.

Reyment, R.A., Blackith, R.E. & Campbell, N.A. 1984. *Multivariate Morphometrics, 2nd* Edition. Academic Press, London, 233 pp.

Rogers, S.J., Fang J.H., Karr C.L. & Stanley D.A. 1992. Determination of lithology from well logs using a neural network. *American Association of Petroleum Geologists Bulletin 76,* 731-739.

Rollinson, H. 1995: *Using geochemical data: evaluation, presentation, interpretation.* Longman Group UK Limited, 352 pp.

Stone, M. 1974: Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society 36,* 111-147.

Weiss, S.M. & Kapouleas, I. 1989: An empirical comparison of pattern recognition, neural nets, and machine learning classification methods. In Sridharan, N.S. (ed): *Proceedings of the 11th International Joint Conference on Artificial Intelligence.* Kaufmann, California, 781-787

Wold, S. 1976: Pattern recognition by means of disjoint principal component models. *Pattern Recognition 8,* 127-139.

Wasserman, P.D. 1989: *Neural computing – theory and practice.* Van Nostrand Reynold, New York, 230 pp.

Webb, A. 2002: *Statistical pattern recognition. 2nd* Edition. Wiley & Sons, Chichester, U.K., 496 pp.

Wei, K.-Y. 1994: Statistical pattern recognition in paleontology using SIMCA-MACUP. *Journal of Paleontology 68,* 689-703.

Wold, H. 1982: Soft modeling: The basic design and some extensions. In Jöreskog, H & Wold, H. (eds): *Systems Under Direct Observation.* North-Holland, New York, 1-53.

Wold, S., Albano, C. Dunn III, W.J., Esbensen, K., Hellberg, S., Johansson, E., Lindberg, W. & Sjöström, M. 1984: Modelling data tables by principal components and PLS: Class patterns and quantitative predictive relations. *Analusis 12,* 477-485.