Vol. 112 No. 1 March 15, 1979
(Computer Techniques and Optimization, Vol. 3 No. 1)
ISSN 0378-4304

ANALYTICA CHIMICA ACTA

International journal devoted to all branches of analytical chemistry

COMPUTER TECHNIQUES AND OPTIMIZATION

ELSEVIER SCIENTIFIC PUBLISHING COMPANY

ANALYTICA CHIMICA ACTA

International journal devoted to all branches of analytical chemistry
Revue internationale consacrée à tous les domaines de la chimie analytique
Internationale Zeitschrift für alle Gebiete der analytischen Chemie

PUBLICATION SCHEDULE FOR 1979 (incorporating the section on Computer Techniques and Optimization).

	J	F	M	A	M	J	J	A	S	O	N	D
Analytica Chimica Acta	104/1	104/2	105	106/1	106/2	107	108	109/1	109/2	110/1	110/2	111
Section on Computer Techniques and Optimization			112/1			112/2			112/3			112/4

Scope. *Analytica Chimica Acta* publishes original papers, short communications, and reviews dealing with every aspect of modern chemical analysis, both fundamental and applied. The section on *Computer Techniques and Optimization* is devoted to new developments in chemical analysis by the application of computer techniques and by interdisciplinary approaches, including statistics, systems theory and operation research. The section deals with the following topics: Computerized acquisition, processing and evaluation of data. Computerized methods for the interpretation of analytical data including chemometrics, cluster analysis, and pattern recognition. Storage and retrieval systems. Optimization procedures and their application. Automated analysis for industrial processes and quality control. Organizational problems.

Submission of Papers. Manuscripts (three copies) should be submitted to:
for *Analytica Chimica Acta:* Dr. A. M. G. Macdonald, Department of Chemistry, The University, P.O. Box 363, Birmingham B15 2TT, England;
for the section on *Computer Techniques and Optimization:* Dr. J. T. Clerc, Universität Bern, Pharmazeutisches Institut, Sahlistrasse 10, CH-3012 Bern, Switzerland.

Information for Authors. Papers in English, French and German are published. There are no page charges. Manuscripts should conform in layout and style to the papers published in this Volume. Authors should consult Vol. 102, p. 253 for detailed information. Reprints of this information are available from the Editors or from: Elsevier Editorial Services Ltd., Mayfield House, 256 Banbury Road, Oxford OX2 7DE (Great Britain).

Reprints. Fifty reprints will be supplied free of charge. Additional reprints (minimum 100) can be ordered. An order form containing price quotations will be sent to the authors together with the proofs of their article.

Advertisements. Advertisement rates are available from the publisher.

Subscriptions. Subscriptions should be sent to: Elsevier Scientific Publishing Company, P.O. Box 211, 1000 AE Amsterdam, The Netherlands. The section on *Computer Techniques and Optimization* can be subscribed to separately.

Publication. *Analytica Chimica Acta* (including the section on *Computer Techniques and Optimization*) appears in 9 volumes in 1979. The subscription for 1979 (Vols. 104–112) is Dfl. 1179.00 plus Dfl. 135.00 (postage) (Total approx. U.S. $641.00). The subscription for the *Computer Techniques and Optimization* section only (Vol. 112) is Dfl. 131.00 plus Dfl. 15.00 (postage) (Total approx. U.S. $71.00). Journals are sent automatically by air mail to the U.S.A. and Canada at no extra cost and to Japan, Australia and New Zealand for a small additional postal charge. All earlier volumes (Vols. 1–95) except Vols. 23 and 28 are available at Dfl. 144.00 (U.S. $72.00), plus Dfl. 10.00 (U.S. $5.00) postage and handling, per volume.

Claims for issues not received should be made within three months of publication of the issue, otherwise they cannot be honoured free of charge.

Customers in the U.S.A. and Canada who wish to obtain additional bibliographic information on this and other Elsevier journals should contact Elsevier/North Holland Inc., Journal Information Center, 52, Vanderbilt Avenue, New York, NY 10017. Tel: (212) 867-9040.

ANALYTICA CHIMICA ACTA
VOL. 112 (1979)
(Computer Techniques and Optimization, Vol. 3)

ANALYTICA CHIMICA ACTA

International journal devoted to all branches of analytical chemistry

COMPUTER TECHNIQUES AND OPTIMIZATION

VOL. 3 1979

ELSEVIER SCIENTIFIC PUBLISHING COMPANY

Anal. Chim. Acta, Vol. 112 (1979)

Printed in The Netherlands

Analytica Chimica Acta, 112 (1979) 1—9
Computer Techniques and Optimization

A COMPUTERIZED SYSTEM FOR DETERMINING SECONDARY ION ENERGY SPECTRA

M. A. RUDAT** and G. H. MORRISON*

Department of Chemistry, Cornell University, Ithaca, N.Y. 14853 (U.S.A.)

(Received 18th August 1978)

SUMMARY

A computer program and computer-directed high-voltage sample potential controller which allow a rapid determination of the energy spectrum of an ion species in a secondary ion mass spectrometer are presented. An electrostatic sector energy analyzer on the mass spectrometer provides high energy resolution. The spectra are obtained by stepping the sample accelerating potential by small voltage steps and determining the detector pulse count rate at each energy. The computer program controls the sample potential, collects the data, makes corrections for ion optical and geometric effects, calculates several important parameters of the energy distribution, and outputs the corrected and uncorrected results to line printer and a plotter. The system is fast, accurate, and very easy to operate. It should be necessary only to change the ion optical correction procedure in order to use the system with other mass spectrometers.

Secondary ion mass spectrometry (s.i.m.s.) is a powerful analytical technique that can reveal surface compositional details down to micrometer sizes [1, 2], analyze thin films [2], and provide composition versus depth profiles [2, 3]. The commonest approach to analysis with this technique is to monitor changes in the mass spectrum or changes in individual mass line intensities. As a result, one of the most important sources of information is frequently overlooked: the energy spectrum of individual secondary-ion mass lines. The energy spectrum can provide a wealth of information: its shape can provide clues about the surface binding energy and the ionization mechanism of the ion, and changes in the shape when the ion bombardment conditions are varied provide important information about alterations in the surface environment (such as oxidation).

A completely different type of information is also available from the energy spectrum of secondary ions transmitted through the ion optics of the spectrometer: the effect of the optics on the signals in a mass spectrum. Discrimination against higher energy secondary ions in the ion optics results in an increased transmission for ions whose energy spectra show narrow, low-energy distributions and lower transmission for ions with broad energy

** Present address: Central Research and Development Dept., E.I. du Pont de Nemours & Co., Wilmington, Dela. 19898 U.S.A.

spectra. If a detailed description of the ion optical effects is known, then it is possible to reconstruct the appearance of the original ("true") energy spectrum from the energy spectrum seen after passage through the ion optics. A comparison of the uncorrected spectrum over the range of the normal energy window for the spectrometer with the "true" energy spectrum yields a direct measure of the energy discrimination of the instrument for the ion of interest. Then, in order to construct a mass spectrum which is representative of the secondary ion population at the sample surface, the individual signals recorded must be corrected for the energy discrimination as well as detector discrimination [4, 5] effects.

In this paper a computer-controlled system for obtaining energy spectra, correcting them for the ion optical and geometrical effects of a commercial instrument, evaluating them for several important spectral parameters including the discrimination effect, and plotting and printing out the data is described. The program provides a fast, easy, and accurate method for determining energy spectra with s.i.m.s. instrumentation, and should be readily adaptable to any instrument for which the ion optical characteristics are well known. The system has been applied to the evaluation of energy discrimination effects [6], the monitoring of energy spectra during the oxidation and nitridation of samples [7], the evaluation of the effect of different matrices on secondary-ion energy spectra [8], and the compilation and interpretation of a large number of energy spectra from pure materials [9]. Several typical spectra obtained during these experiments are shown to illustrate the results obtained with the system.

INSTRUMENTATION

A schematic diagram of the CAMECA IMS-300 ion microanalyzer is shown in Fig. 1. Primary ions created in the duoplasmatron ion gun are accelerated and focused onto the target, striking it at an energy of 5.5 keV or 14.5 keV, depending on whether the primary ion and secondary ion polarities are alike or opposite. The secondary ions are accelerated away from the sample by placing a potential on the target of ± 4.5 kV, depending on the desired

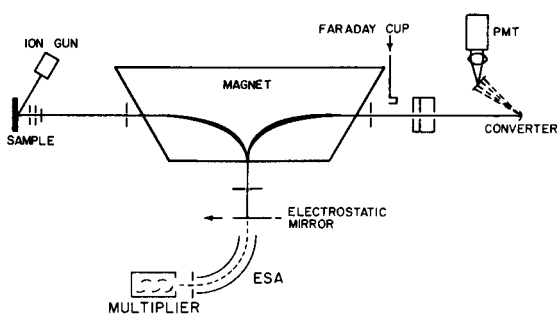


Fig. 1. Schematic diagram of the CAMECA IMS-300.

secondary ion polarity. The immersion lens focuses the ions into the mass spectrometer while discriminating against high energy ions (see below). In normal operation, the ions are reflected by the electrostatic mirror after one pass through the magnetic sector, and undergo a second deflection by the magnet. The mirror sets the energy window of the instrument, and is usually operated so that a 0—15 eV energy window is presented to the ions. Following the second magnetic deflection, the ions enter the detector section where they are accelerated into an ion-to-electron converter. Electrons from the converter can be detected in three ways: electronically, through a scintillator—photomultiplier tube combination; photographically, by using electron-sensitive film; or optically, by using a fluorescent screen and binocular viewing.

To obtain energy spectra, the spherical condenser electrostatic analyzer (ESA) must be used. The electrostatic mirror is displaced and the ion beam enters the ESA, being detected after energy analysis by an electron multiplier. The entrance slits to the ESA determine the energy bandwidth accepted, and the exit slits determine the mass resolution. Energy spectra are obtained by energizing the ESA sector plates to voltages determined by the curvature of the sector, setting the magnet current so that the secondary ion mass line of interest is lined up with the entrance slits for the ESA, and scanning the sample potential. Only ions having energies falling within the bandwidth set by the ESA entrance slit strike the detector; by changing the sample potential, the nominal center of that window is changed.

A computer is interfaced with the IMS-300 so that the sample potential, magnetic sector and detector systems can be controlled by the computer [10].

Computer-controlled hardware

The parts of the computer system which are important for this work are shown schematically linked together in Fig. 2. The Digital Equipment Corporation (DEC) PDP 11/20 with a 24K core is interfaced through standard DEC modules to a 1.2 M word RKO3 removable disk on which all programs are stored, a DEC GT-40 graphics terminal (used only as a teletype for this

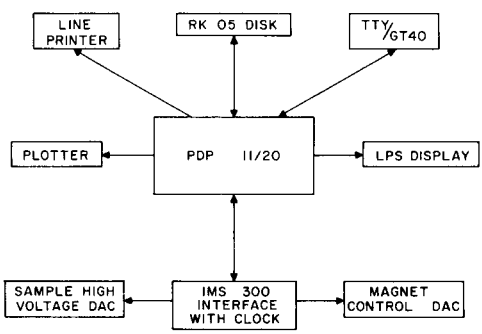


Fig. 2. Interaction diagram of the computer system and some of the peripherals and CAMECA interface.

program), a Houston Instruments Complot Plotter, a Xerox-Versatec line printer/plotter, and a DEC Laboratory Peripheral System (LPS-11) which provides a digital LED display. A special interface system provided with a high-speed clock [10] links the computer to a Nuclide DAC-16 digital-to-analog converter which controls the magnet current of the mass spectrometer [10], and to the sample voltage controller used by the present program.

Figure 3 is a schematic diagram of the sample voltage controller. The DAC is a 10-bit device (Analog Devices) and drives a high-voltage operational amplifier with a nominal range of ±100 V, "floated" on the normal sample potential of 4500 V. Calibration of the system has shown the resolution to be ca. 0.192 V/DAC step, with a range of ±96 V. A voltage offset of + 1.2 V when the DAC is set at zero was a problem in setting up the ESA for an energy-spectrum run, as well as in normal operation of the CAMECA. This was overcome by connecting a high voltage relay to the power switch, so that the DAC can be easily bypassed. The offset was inconsequential for the determination of energy spectra, since the least-squares fit to the calibration line allows an accurate value of the sample potential to be calculated.

BASIS OF THE COMPUTATIONS

The equations of Slodzian [11] are assumed to describe accurately the effects of the ion optics of the ion microanalyzer on the energy spectra.

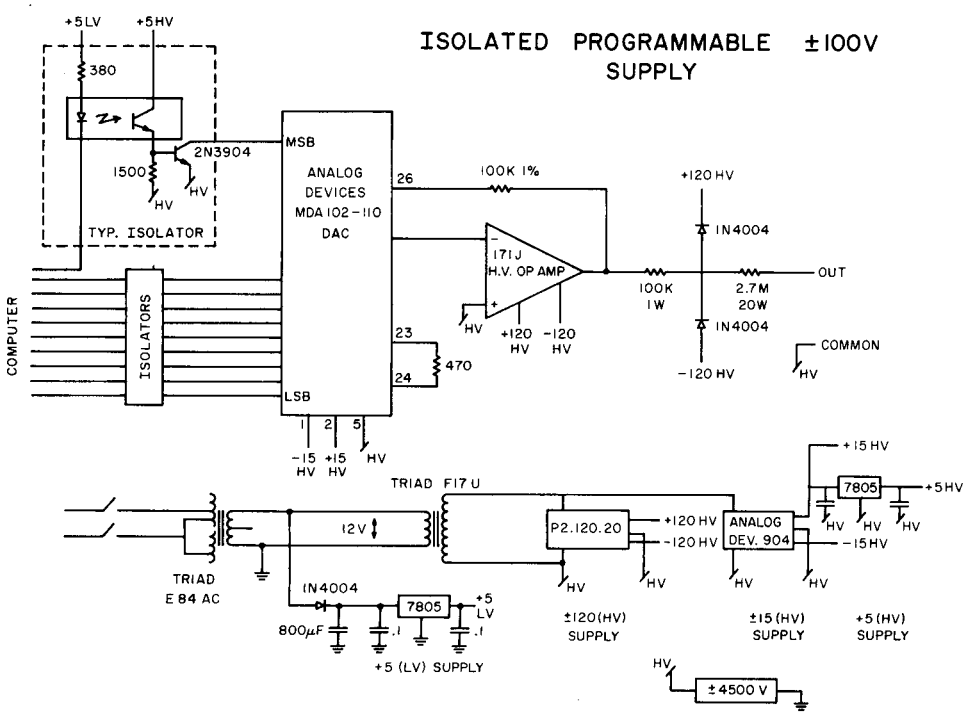


Fig. 3. Schematic diagram of the computer-controlled sample voltage controller.

These equations describe the action of the immersion lens on secondary ions, based on the initial energy and angle of emission of the ions. Since the primary purpose in the design of the ion microanalyzer was to create an ion microscope with high lateral resolution at the sample [1], the immersion lens was designed to discriminate against ions whose lateral velocity (i.e., parallel to the sample surface) would lead to a reduction in resolution. The highest lateral velocity that ions may have and pass through the lens is controlled by the size of the contrast aperture. Figure 4 shows the maximum angle of acceptance from the surface normal versus ion energy for a 400-μm contrast aperture. All ions below ca. 0.5 eV initial energy are accepted (α_m = 90°), whereas only ions emitted at an angle within a cone ca. 15° from the surface normal will be accepted if the initial energy is 7 eV. Equation 1 is the analytical expression of this curve [11]:

$$\sin \alpha_m = a/4D(E/V + E)^{\frac{1}{2}} \tag{1}$$

where α_m is the maximum angle of acceptance from the surface normal, a is the radius of the contrast aperture, D is the distance between the target surface and the front plate of the immersion lens (5 mm), E is the initial energy of the secondary ion, and V is the accelerating potential placed on the target (4500 V nominal).

It is clear that if the angular distribution is not cosinusoidal, then the relationship between the true energy spectrum and the energy spectrum of ions leaving the immersion lens will be affected by the angular distribution. From polycrystalline samples bombarded at various angles of incidence the angular distribution is overcosine [11—16]. An analytical expression to describe the results of Hennequin's experiments [13] has been found which describes those results better than the expression proposed in that paper:

$$f = \cos \phi \{-(1 + \sin \theta) \cos [1.5(-90 - \theta/2)] + 2 \sin \theta |\cos \psi| \cos [1.5(\phi - \theta/2)]\} \tag{2}$$

where ϕ is the angle of emission of the secondary ion measured from the surface normal, θ is the primary ion angle of incidence measured from the surface normal, and ψ is the angle of emission measured from the plane formed by the path of the incident ion and the surface normal, all measured in degrees. For a given θ, it is necessary to normalize the three-dimensional integral to unity (i.e., no more than 100% of the emitted ions can be collected). Figure 5 illustrates the appearance of this function for $\theta = 60°$ and $\psi = 0°$. For $\psi = 90°$ (the cross-section of the distribution), a cosinusoidal graph is obtained, as found experimentally [13]. At $\psi = 0$, the function is basically $\cos^2\phi$, also experimentally verified [12, 13, 15].

THE PROGRAM

In its present form designed for rapid calculation the main program relies on the standardization of certain instrumental operating parameters to

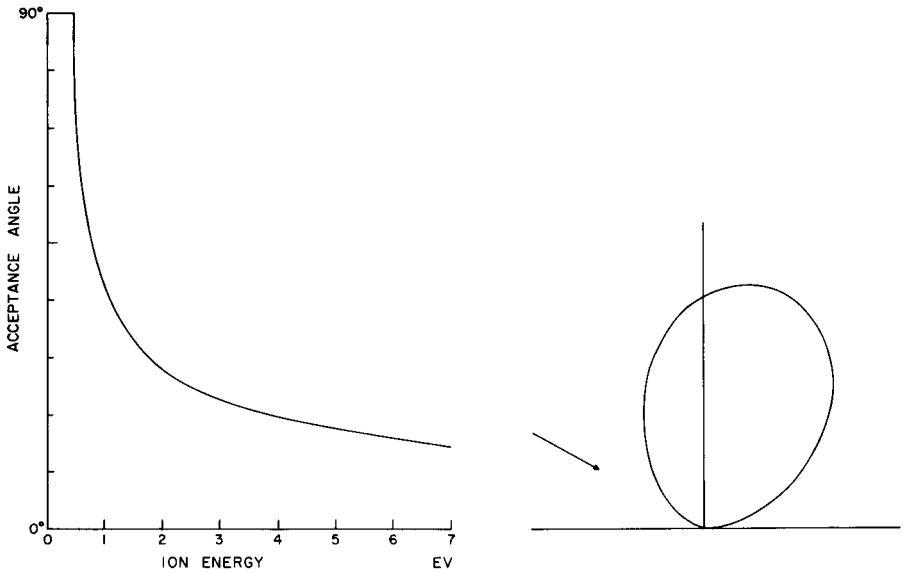


Fig. 4. Maximum angle of acceptance (α_m) from the surface normal versus initial secondary ion energy for a contrast aperture of 400 μm. The value of α_m determines the solid angle within which the secondary ion must be ejected initially in order to be collected and transmitted by the immersion lens.

Fig. 5. Appearance of the angular distribution function given by eqn. 2 for $\theta = 60°$ and $\psi = 0°$.

reduce the number of calculations: (1) the angle of incidence, θ, in eqn. (2), is fixed, and is determined by the mode of operation of the instrument (i.e., whether the ion beams are of the opposite or same polarity); (2) the size of the contrast aperture is fixed at 300-μm diameter; and (3) the magnitude of the sample voltage step per point is fixed at 0.384 V/point (two DAC increments).

Combination of these three parameters allows the use of two tables of correction factors based on eqns. (1) and (2), one each for the two modes of operation. The construction of the corrected spectrum therefore requires only a few seconds, rather than the 20 min of computational time needed when these parameters are allowed to be variable.

The user determines all of the parameters of the energy spectrum scan (except the voltage step size): the pulse counting time; the number of such independent pulse counting times to be averaged together for the data point at each voltage step; the "wait time" between the pulse counting periods at a given voltage step; the starting value and scanning step rate of the magnet DAC (usually not used); the waiting time between sample voltage steps; the starting sample—voltage—DAC setting; and the number of voltage steps to be done. Questions about the relative polarity of the ion beams and the absolute polarity of the secondary ion beam determine which reference table is used and which direction to scan the sample voltage (positive or negative).

Assembler-level programs read the clock, count the pulses from the electron multiplier, convert the data from double-precision integer to floating-point format, and output the present setting of the sample—voltage—DAC through the LED display (0 to ± 512). The FORTRAN main program runs simultaneously, retrieving the data.

After data acquisition, all data handling is done by FORTRAN level programming. The zero of the energy scale is determined by finding the voltage at which the derivative of the raw data exceeds 1% of the maximum intensity of the raw data. The absolute value of the sample voltage to which this corresponds is saved and output with the results later. Because of the method normally used to set up the experiments, the maximum in the uncorrected energy distribution occurs at about 0 V, so that the assigned value of zero is at a "negative energy".

The correction function as given in the stored tables is then applied point-by-point, and several parameters of interest are simultaneously calculated. These parameters include: the total integrated intensity of the uncorrected and corrected energy spectra; the integrated intensity of the uncorrected spectrum in the energy window 0—15 eV; and the average energy of the corrected and uncorrected spectra, given by

$$\overline{E} = \sum I_i E_i / \sum I_i \tag{3}$$

where I_i and E_i are the intensity and energy at step i. In addition, a least-squares fit of a power function of the form E^{-n} is performed using all data above a cutoff energy which is determined by the energy of the most intense point in the corrected spectrum. The low energy cutoff is 2 eV, 5 eV, or 10 eV above the most probable energy of the distribution, if the most probable energy falls in the range of 0—5 eV, 5—10 eV, or > 10 eV, respectively. A high energy cutoff is implemented if the energy distribution goes to zero before 10 eV. The correlation coefficient for the log—log fit is also calculated.

A normalized linear plot is automatically produced on the Complot plotter, with the title of the plot being input by the operator, and a preset energy scale of 0—100 eV. The uncorrected spectrum can also be output on the same graph as the corrected spectrum if desired, and is on the same scale as the corrected spectrum.

The line printer output consists of: (1) the title of the run; (2) all input parameters; (3) the absolute voltage which was selected as the zero of energy; (4) the uncorrected data and its derivative; (5) the corrected data in normal and normalized forms; (6) the integrated intensities for the uncorrected and corrected spectra; (7) the integrated intensity of the uncorrected spectrum up to 15 eV; (8) the ratio of (7) to the corrected spectrum integrated intensity (the bandpass efficiency); (9) the average energies of the uncorrected energy spectrum at energies of −1.5 eV, 25 eV, and 65 eV, normalized to the maximum intensity in the spectrum; and (10) the power fit of the high energy part of the spectrum (E^{-n}) with the correlation coefficient of the fit.

APPLICATION OF THE SYSTEM

Energy spectra of Nb^+ and NbO^+ from the bombardment of pure niobium by 5.5-keV O_2^+ in an oxygen atmosphere at 1×10^{-6} torr are shown in Fig. 6 in both corrected and uncorrected forms. The energy window of the ESA was 3 eV.

For many of the spectra recorded, the noise level equivalent to that exhibited by the spectra in Fig. 6 has been routinely accomplished. Obviously, ions with lower signal levels would have correspondingly higher noise levels, but the point density (300 points per run, 0.38-V steps) is high enough to allow an excellent definition of the spectral shapes. The spectra obtained with this system are very similar to those obtained in other work [9], suggesting that in spite of the inherent assumptions the spectra are reasonably accurate.

For these spectra, the parameters were: 100 ms/count period, two count periods per data point, 5 ms between count periods, and 25 ms between voltage steps. The zero-set values were −5.50 eV for Nb^+ and −4.73 eV for NbO^+. The calculated values of the spectral parameters (uncorrected for the extension of the spectra beyond 100 eV) for Nb^+ and NbO^+, respectively, were: bandpass efficiencies, 0.1293 and 0.2111; average energies of the uncorrected energy distributions 14.0 eV and 7.9 eV; average energies of the corrected ("true") energy distributions, 18.9 eV and 9.8 eV; and high-energy power dependences and correlation coefficients, $E^{-2.44}$ (0.9612) and $E^{-3.64}$ (0.9744). The most probable energies of the corrected spectra and full widths at half maximum of the corrected spectra were measured from the spectra plots to be 7.3 eV and 35.0 eV for Nb^+, and 5.4 eV and 11.0 eV for NbO^+.

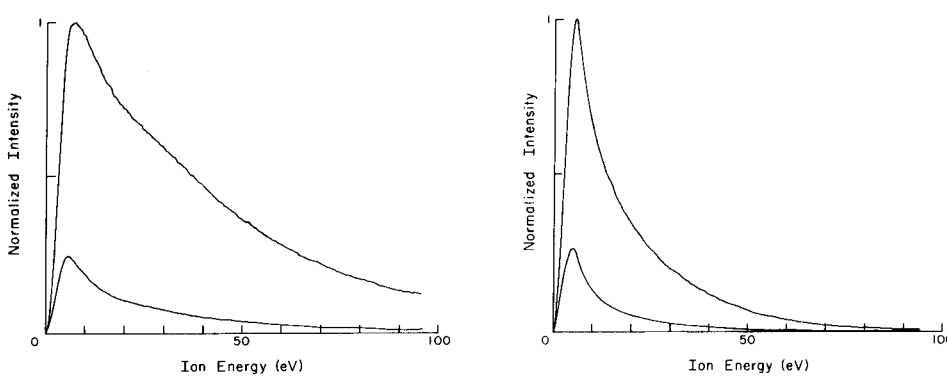


Fig. 6. Energy spectra of (left) Nb^+ and NbO^+ in corrected (upper curves) and uncorrected forms. Corrected spectra approximate the true energy distributions.

CONCLUSION

The fast, easy-to-use, accurate system developed is suitable for determining the energy of secondary ions sputtered from surfaces in a mass spectrometer. The use of a computer program allows the data to be corrected "instantly" for ion-optical effects, and also immediately provides several important parameters of the distributions which would otherwise be difficult and tedious to calculate. The approach used can readily be adapted to other s.i.m.s. instruments for which energy spectra can be obtained by scanning the sample accelerating or bias potential; the correction procedure must be modified for the ion-optical effects of each type of instrument.

This work was supported by the National Science Foundation under Grant No. CHE-77-04405 and through the Cornell Materials Science Center.

REFERENCES

1 G. H. Morrison and G. Slodzian, Anal. Chem., 47 (1975) 932A.
2 C. A. Evans, Anal. Chem., 47 (1975) 855A.
3 H. Liebl, J. Vac. Sci. Technol., 12 (1975) 385.
4 M. A. Rudat and G. H. Morrison, Int. J. Mass Spectrom. Ion Phys., 27 (1978) 249.
5 M. A. Rudat and G. H. Morrison, Int. J. Mass Spectrom. Ion Phys., in press.
6 M. A. Rudat and G. H. Morrison, Int. J. Mass Spectrom. Ion Phys., in press.
7 M. A. Rudat and G. H. Morrison, Int. J. Mass Spectrom. Ion Phys., in press.
8 M. A. Rudat and G. H. Morrison, Int. J. Mass Spectrom. Ion Phys., in press.
9 M. A. Rudat and G. H. Morrison, Surf. Sci., in press.
10 B. K. Furman and G. H. Morrison, to be published.
11 G. Slodzian, in: Secondary Ion Mass Spectrometry, Ed. K. F. J. Heinrich and D. E. Newbury, Special Publication No. 427, National Bureau of Standards, Wash., D.C., 1975, p. 33.
12 K. Rödelsperger and A. Scharmann, Nucl. Instrum. Methods, 132 (1976) 355.
13 J.-F. Hennequin, J. Phys. Paris 29 (1968) 957.
14 B. Emmoth and M. Braun, Proc. VII Int. Vac. Congr. & III Int. Conf. Solid Surfaces (Vienna 1977), p. 1465.
15 G. A. van der Schootbrugge, A. G. J. de Wit, and J. M. Fluit, Nucl. Instr. Meth., 132 (1976) 321.
16 H. Kerkow and M. Trapp, Int. J. Mass Spectrom. Ion Phys., 13 (1974) 113.

Analytica Chimica Acta, 112 (1979) 11—30
Computer Techniques and Optimization

A COMPARISON OF FIVE PATTERN RECOGNITION METHODS BASED ON THE CLASSIFICATION RESULTS FROM SIX REAL DATA BASES

MICHAEL SJÖSTRÖM

Research Group for Chemometrics, Department of Chemistry, Umeå University, S-901 87 Umeå (Sweden)

BRUCE R. KOWALSKI

Laboratory for Chemometrics, Department of Chemistry, University of Washington, Seattle, Washington 98195 (U.S.A.)

(Received 11 August 1978)

SUMMARY

Several pattern recognition methods are compared, including the Bayesian classification rule, linear discriminant analysis, the K-nearest neighbour rule, the linear learning machine for multicategory data, and soft independent modelling of class analogy. Several preprocessing methods are discussed in connection with these methods. Six real data bases previously described in the literature are investigated with these methods and the advantages and limitations of the preprocessing and classification methods are discussed.

Multivariate data analysis in the form of pattern recognition has proved a versatile tool for extracting information from data bases in chemistry [1, 2]. The reasons can be traced to several factors of which the following are considered the most important. First, there is the large amount of data generated by modern spectroscopic methods, data often of multivariate type and from which classification problems can be formulated. Secondly, the complex nature of systems studied in many branches of chemistry is such that models and methods for data interpretation which are applicable to systems of lower complexity, e.g. in physics, are often of little value. Thirdly, suitable program packages for data analysis have been developed where the pattern recognition methods included have been adapted for chemical purposes.

A pattern recognition problem can be formulated in the following way. For N objects, it is known or can be assumed that each object belongs to one of Q classes. An object is characterized by a data vector y_i with M elements, where each element in the vector consists of a measurement of a variable. The objects with known class membership belong to a training set from which information is extracted to permit separation of the classes from each other. In some applications this information is used to classify a set of T additional objects (a test set) with known data vectors but with unknown class membership (Fig. 1). The possibility that some of the objects in the

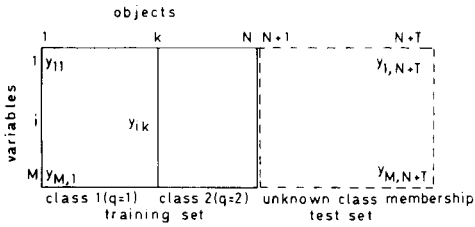


Fig. 1. Observation matrix for a classification problem.

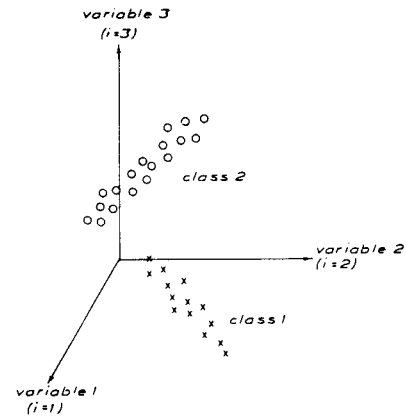


Fig. 2. A graphical representation of a two-class problem with three variables.

training set or test sets belong to none of the classes must also be considered in many applications. It should be noted that for many classification problems there is no prior information of the variable relevance of the variables included. In fact, one of the more urgent purposes of a classification approach is to elucidate differences between the classes in the training set and to determine which variables contain information about these differences.

A pattern recognition problem with three variables ($M = 3$) can be illustrated by a three-dimensional plot where each object is represented by a point in this space. Figure 2 shows such a visualization of a three-variable problem with two classes. In the same way, each object in a problem where $M > 3$ can be thought of as a point in an M-dimensional space.

In the literature, technical and computational aspects of pattern recognition methods have been the principal topics discussed. The performance of a method is often demonstrated by a single example, sometimes only on a synthetic data base and without comparisons with other methods. These limitations have recently been discussed and emphasized by Kanal [3]. Consequently, it was considered desirable to compare the information that can be gained from a number of real data bases when different preprocessing and classification methods are used.

From pattern recognition problems in the literature, six real data bases were selected for investigation by the following methods: the Bayesian classification rule (BAYES), linear discriminant analysis (LDA), the K-nearest neighbour rule (KNN), the linear learning machine for multicategory data (LLM), and the soft independent modelling of class analogy (SIMCA). These well-documented methods are probably the most frequently used in chemical applications, and they are easily available through the ARTHUR [4, 5], SPSS [6] and SIMCA [7] packages.

EXPERIMENTAL

The calculations based on the ARTHUR and SPSS packages were done at the University of Washington Academic Computer Center and the calculations with SIMCA (version SIMCA-2T) at Umeå University Computer Center.

The computational strategies for solving a problem involved: (i) formulation of the classification problem; (ii) preprocessing of data by scaling and/or transformation of data and reduction of the dimensionality of the data; (iii) two- or three-dimensional graphic representation of the data; (iv) fitting of the original or the preprocessed data to the chosen classification method; (v) interpretation of the results.

PREPROCESSING METHODS

Scaling and transformation of the variables

Often the variances of the variables differ considerably and prior information is not available about the relevance of the variables. A reasonable approach to this problem is to give each variable the same weight in the initial stage of the analysis. This is done by so-called regularization or autoscaling of the variables, so that they all have unit variance and zero mean over the whole data set. If the distribution of a variable is skewed, a transformation of the data is recommended to give a distribution closer to the normal distribution. The scaling and transformation of the data influence the classification results in different ways for different methods. Three of the methods used — BAYES, LDA and LLM — are not affected by autoscaling.

Variable or feature reduction

The exclusion of data which contain little or no class-separating information is often crucial for a successful classification. One approach to this problem is to generate from the original M-space (M variables) a new orthogonal M-space by solving the eigenvalue problem $R\mathbf{v}_m = \lambda_m \mathbf{v}_m$, where λ_m are the eigenvalues, and $\mathbf{v}_m$ the eigenvectors of R, the correlation matrix. A reduction of the dimensionality of the data can then easily be accomplished by exclusion of the eigenvectors corresponding to the a smallest eigenvalues (eigenvector reduction) and a new $(M-a)$ space is obtained.

Another solution is the SELECT method [8]. This preprocessing method first selects the most discriminating variable (feature), where the Fisher weight or variance weight can be chosen as the discriminating criterion. This corresponds to finding the variable (feature) which differs the most between the classes in terms of a standardized average. The remaining variables are then decorrelated from the first chosen and reweighted, and the feature which in this step gets the highest weight is chosen as the second feature. Features are selected until a previously specified number of features are selected or a predetermined weight is reached. The method gives orthogonal features where the first remains a single variable, while the others are combinations of several variables.

For LDA in the SPSS program, five stepwise variable selection methods are described based on five different variable reduction criteria (see the SPSS manual [6]). These variable reduction methods were tested here with the inclusion levels $P = 0.50$ (default values in the SPSS program).

In SIMCA, measures of the discriminating and modelling powers for the variables are given. The initial classification with all variables included is thus usually followed by a re-computation where variables with low discriminating and modelling powers are deleted.

The variable reduction methods in the SPSS program have been used only with LDA, and the SIMCA variable reduction method only with SIMCA. The eigenvector reduction method and SELECT are used with BAYES, LLM and KNN.

CLASSIFICATION METHODS

The pattern recognition methods used are described briefly below, to emphasize the quite different strategies involved in solving problems.

The Bayesian classification

For each class, and also over all objects in the training set, the frequency distribution of each variable is determined [5, 9]. A probability measure describing the fit of an object to a class can then be estimated from how well the elements of the data vector of the object fit the frequency distributions of the class. The probability for an object is then calculated for each class. The object is then considered to belong to the class with the highest probability. An orthogonal presentation of the variables is recommended because the method assumes the variables to be independent. As an option in the ARTHUR package [4, 5] the estimated distributions can be smoothed by Gaussian function approximations or by cubic spline functions.

Linear discriminant analysis

For a two-class problem, a class-separating or discriminating function is determined by a linear combination of the variable vectors y_i:

$$D_k = \sum_{i=1}^{M} d_i y_{ik}$$

where D_k is the discriminating score for object k (Fig. 3). The weights d_i are determined in such a way that they will exhibit the largest ratio of variances between the two groups relative to that within the groups. The maximum numbers of discriminating functions L are $Q - 1$ when the classes are less than or equal to the numbers of variables ($Q \leqslant M$) and M when $Q > M$.

In the SPSS program, classification information for an object is given as a probability measure of class membership, denoted as $P(G/X)$, where the pooled probability over the classes is unity. The probability measure given, denoted as $P(X/G)$, expresses the probability that an object will be that far away from the discriminant score centroid for a class.

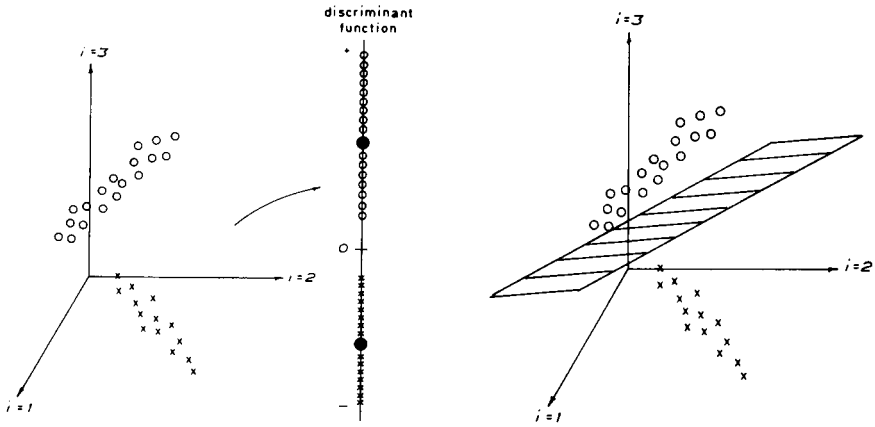


Fig. 3. With LDA the original M-space is reduced to an L-dimensional space where L is the number of discriminating functions. Objects with different signs on their scores on the discriminant function belong to different classes. For a two-class problem as in this case, $L = 1$. The class centroids are denoted by •.

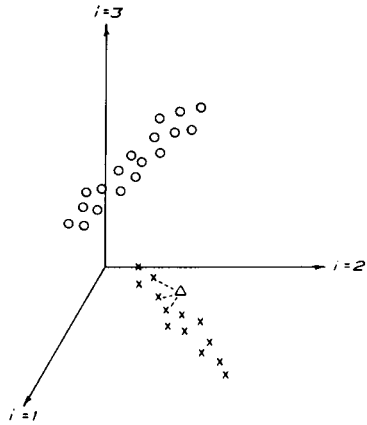
Fig. 4. With LLM the classes are separated by an $M - 1$ dimensional hyperplane.

Linear learning machine for multicategory data

With this method $(Q - 1)$ hyperplanes with the dimensionality $M - 1$ are determined by a feedback procedure in such a way that the objects for different classes in the training set fall, as far as possible, on different sides of the hyperplanes [10] (Fig. 4). The method always gives complete separation of the classes in the training set if the number of variables is larger than the number of objects, whether or not the variables contain class-separating information. This limitation has recently been discussed [11]. In brief, for a two-class problem, the number of variables (or features) should be less than one fourth of the number of objects in the training set for small data sets ($n \leqslant 20$), and less than one third of this number for larger data sets to prevent doubtful classifications.

K-*nearest neighbour rule*

In the M-dimensional space, the class membership of the K closest neighbours to an object are determined. Normally, the Euclidian distance is used as a measure of the closeness of the objects, although other measurements of distance are also available. An object is then assigned to the class to which the majority of its K closest neighbours belongs [12] (Fig. 5). In the KNN routine in the ARTHUR program, the $K = 1,3 \ldots 10$ closest neighbours are calculated for each object. The training set can then be used to determine the K value giving the best prediction of the training set. In the same way as in the training set, a test-set object is classified in the class to which most of its K closest neighbours belong. This method is heavily dependent on the scaling of the variables, when distances are used as criteria of similarity. In



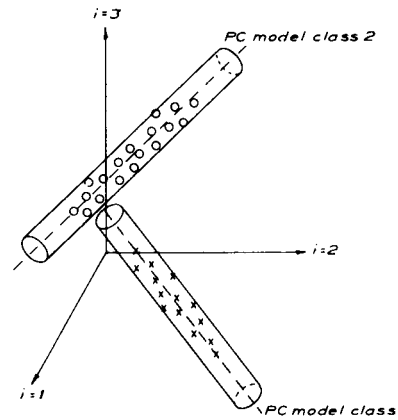


Fig. 5. In KNN the class membership of an object is determined by the majority class membership of its K closest neighbours.

Fig. 6. In SIMCA each class is described by a disjointed PC model. In this case $A = 1$ for both class 1 and 2 in eqn. (1). An object is considered to show a class-typical behaviour within the volume defined by eqn. (2).

cases where the transformations of the variables are made with the aim of optimizing the separation of the classes in the training set, the same restrictions of the objects/features ratio are applicable as for LLM. Therefore, for KNN, only those results derived from autoscaled data are presented.

Simca

A principal component model is fitted to the data $y_{ik}^{(q)}$ for a class q in the training set according to [13]:

$$y_{ik}^{9} = \alpha_i^{9} + \sum_{a=1}^{A_q} \beta_{ia}^{(q)} \theta_{ka}^{(q)} + \epsilon_{ik}^{(q)} \qquad (1)$$

Consequently each class q is represented by a disjointed principal-component model characterized by the parameters $\alpha_i^{(q)}$, $\beta_{ia}^{(q)}$ dependent only on the variables ($i = 1,2 \ldots M$), and $\theta_{ka}^{(q)}$ dependent only on the objects and where $\epsilon_{ik}^{(q)}$ are the residuals. The number of significant terms A_q for a class in eqn. (1) is determined by the so-called cross-validation technique [14]. The classification of an object p is then accomplished by fitting the data vector of the object to the class model (eqn. 1) with the parameters $\alpha_i^{(q)}$ and $\beta_{ia}^{(q)}$ for the different classes q. This is followed by comparison of the class residual variances $d_p^{(q)}$ obtained for the object with the typical residual variances $S_0^{(q)}$ for each class by means of F-tests:

$$F = d_p^{(q)2} / S_0^{(q)2} \; (M - A_q \text{ and } (N_q - A_q - 1)(M - A_q) \text{ degrees of freedom}) \qquad (2)$$

The object then belongs to the class for which the smallest F value is obtained. If an object p, when fitted to a class model, gets θ_a values outside the normal range, i.e. outside $\theta_{a,\,\mathrm{lim}}^{(q)}$, then the distance $d_p^{(q)}$ is calculated instead as the distance between the $\theta_{a,\,\mathrm{lim}}^{(q)}$ and the object p. (For how $\theta_{a,\,\mathrm{lim}}^{(q)}$ is defined and for further details, see [15]). This means that the confidence interval for the class will describe a hypervolume in the M-space (Fig. 6).

In addition, for an object to be considered uniquely classified, the ratio of the next smallest residual variance $d_p^{(r)2}$ and the smallest residual variance $d_p^{(q)}$ must be larger than a critical F-value:

$$F = d_p^{(r)2}/d_p^{(q)2} \quad (M - A_r \text{ and } M - A_q \text{ degrees of freedom}) \tag{3}$$

Measures of the modelling as well as the discriminating powers for the variables are also calculated [15]. SIMCA does not disregard the possibility that an object can belong to none of the classes, or the possibility that an object can belong to more than one class.

DATA BASES

The data bases were chosen to be representative of a large variety of chemical problems. Data bases with few classes and a data base with as many as 40 classes were selected. Classification problems with a few objects in each class, as well as problems with a larger number of objects in the classes, are presented. Problems with few, as well as numerous, variables compared with the number of objects in the training set are also included. Only complete data bases with continuous variables (no missing data) are used. The data bases and the classification problems formulated from them are described briefly below; extensive presentations are available in the cited literature.

Trace element composition of archeological artifacts (ARCH)

On 45 obsidian quarry samples known to come from four different quarries north of San Francisco, ten trace elements were measured. In addition, these measurements were made on 29 obsidian artifacts for which little information of their origin was available. The aim of an earlier pattern recognition approach [16] was to establish to which of the four classes these unknown objects belonged, if any. Consequently, the objects from each of the well-known locations formed the training set with four classes, and the unclassified objects formed a test set.

Cis *and* trans α, β-*unsaturated carbonyl compounds (CARBONYL)*

Ketones and aldehydes with olefinic double bonds in α, β-position can exist in planar *cis* or *trans* conformers. For sterically hindered compounds, twisted conformers are also possible. The training set for this data base consists of one class with seven sterically hindered compounds and one class

of six *trans* compounds. The seven measured variables are the frequencies for two i.r. and two u.v. absorption bands, and the absorption intensities of three of these bands. The test set contains three compounds known to be *cis*, and twelve compounds with unknown class membership. For seven of these, two different assignments are given. The classification problem was formulated to establish if there were any difference between the two classes, and to which classes the test-set objects belonged. A classification of the three *cis* compounds in the test set in the class with sterically hindered compounds is strong evidence that this class consists of *cis* compounds. The data were taken from Mecke and Noack [17] and the editing of the data and the formulation of the classification problem from Wold and Sjöström [18].

Trace element study of blood samples from welders (WELDERS)

Concentrations of 17 trace elements in blood samples from welders [19] were measured by spark-source mass spectrometry. The blood samples were divided in four classes with 23, 7, 28 and 23 samples in each class, representing welders using four different welding techniques. A fifth class was formed by a control group of 68 persons not involved in welding. Of interest are possible dissimilarities between the blood heavy metal compositions of the different classes and the control group.

Trace element composition of oil spills (OIL)

This data base consists of 40 classes with 10 objects in each class [20]. The objects in a class are represented by one oil sample from a special oil field and 9 different artificial weatherings performed on this oil sample. Originally, 22 trace elements were measured for each object, but actual concentrations of only seven of these are available for all samples, so that only the latter data were used. The aim of the investigation was to determine if the information on the trace element composition could be used to make accurate classifications of oil spills of unknown origin.

Taxonomy of iris species (IRIS)

Fisher's classical iris data [21] are often used for model studies. Thus it was considered valuable to include this non-chemical classification problem here. Three species of iris form the three classes. For 50 flowers (objects) for each class, there are four measured variables (sepal length and width, petal length and width); 25 objects randomly selected from each class form the test set. This means that 25 objects from each class form the training set and that the test set contains 75 objects. The editing of the data is the same as described earlier [13].

^{13}C-n.m.r. spectra of exo *and* endo *substituted norbornanes (NMR)*

This classification problem concerns ^{13}C-n.m.r. spectra of eight *exo* and seven *endo* 2-substituted norbornanes, where the *exo* compounds form one

class and the *endo* compounds another class [22]. The variables are the seven relative shift differences between the actual structure and the unsubstituted norbornane. In addition, the data base consists of a test set of 28 *exo* and *endo* compounds which are, from a chemist's point of view, related to the compounds in the training set. For all the investigated compounds, it is known which compounds are *exo* and which are *endo*. Thus the aim of the pattern recognition approach was not an *exo/endo* classification but merely to establish: (a) if ^{13}C-n.m.r. shifts contain information on whether a compound is an *exo* or *endo* compound; (b) which variables (carbon atoms in the molecular framework) contain such information; (c) if the ^{13}C-n.m.r. shifts for the compounds show a behaviour analogous to that of the training set; (d) if a classification approach could give indications of erroneous assignments.

RESULTS

ARCH

For the training set, all methods solved this classification problem with a 100% correct classification. In contrast, the 29 objects in the test set showed a considerable scatter in the classification from one method to another. The classifications of the test set for the different methods are collected in Table 1. For BAYES and LLM some different preprocessing approaches are also presented. For BAYES the classification is given with and without orthogonalization and with spline and gauss smoothing of the histograms. In the use of LLM a reduction to three features has to precede the classification, since there are few objects in three of the classes. This feature reduction is done with SELECT and with eigenvector reduction.

For all BAYES calculations, the training set is 100% correctly classified. This means that there are no criteria for a preference for one BAYES approach over another, which causes a real dilemma in the test set, when 14 out of 29 objects are not consistently classified. The same problem is present for the LLM calculations where the training set is 100% correctly classified, while for the 29 objects in the test set the classification results differ for 8 objects in the three calculations. In contrast, LDA, SIMCA and KNN showed very similar classification results; only object 47 is classified differently by these three methods. With SIMCA some objects are classified as outliers (47, 50, 51, 55 and 73). According to SIMCA, these outliers are not significantly closer to any particular class, which shows that their class membership information is low, and thus explains the startling scatter in the classifications obtained by the different BAYES and LLM approaches for these objects.

WELDERS

For this data base which lacks a test set BAYES gave a superior classification (Table 2). This may indicate an over-optimistic number of correctly classified objects. To check for such overfitting, the following experiment

TABLE 1

A comparison of methods for the ARCH test set

Obj. no.	BAYES[a,b]			BAYES[b,c]			LDA[d]	KNN[e] $N=3$	LLM[e]			SIMCA[e,f,g] $A=0$
	Direct	Gauss	Spline	Direct	Gauss	Spline			SELECT		eigenvector reduction	
									Variance weight	Fisher weight		
46	1(0.99)	1(0.97)	1(0.98)	3(0.41)	3(0.78)	1(0.41)	1	1	1	1	1	1
47	3(0.65)	2(0.35)	2(0.49)	4(0.96)	4(0.39)	4(0.92)	1—	3	1	1	2	(2)
48	1(0.99)	1(0.94)	1(0.98)	1(0.92)	1(0.81)	1(0.91)	1	1	1	1	1	1
49	3(0.62)	1(0.49)	1(0.55)	1(0.84)	1(0.69)	1(0.79)	1	1	1	1	1	1
50	1(0.93)	1(0.70)	1(0.81)	4(0.73)	1(0.41)	2(0.38)	1	1	1	1	2	(1)
51	3(0.93)	2(0.81)	3(0.71)	4(0.81)	2(0.46)	4(0.53)	2—	2	3	2	3	(2)
52	2(0.94)	2(0.97)	2(0.96)	2(0.42)	2(0.64)	2(0.57)	2—	2	3	2	2	2
53	2(0.97)	2(0.98)	2(0.98)	2(0.79)	2(0.79)	2(0.78)	2	2	2	2	2	2
54	2(0.52)	2(0.94)	2(0.79)	2(0.43)	2(0.63)	2(0.58)	2—	2	3	2	2	2
55	3(0.56)	2(0.78)	3(0.55)	4(0.53)	2(0.41)	3(0.45)	2—	2	3	2	3	(2)
56	3(1.00)	3(0.99)	3(1.00)	2(0.70)	3(0.75)	2(0.40)	3	3	3	3	3	3
57	3(1.00)	3(0.95)	3(0.99)	3(0.93)	3(0.68)	3(0.82)	3	3	3	3	3	3
58	3(1.00)	3(0.98)	3(1.00)	3(0.48)	3(0.71)	3(0.67)	3	3	3	3	3	3
59	3(1.00)	3(0.96)	3(0.99)	3(0.49)	3(0.65)	3(0.64)	3	3	3	3	3	3
60	3(1.00)	3(0.97)	3(1.00)	3(0.86)	3(0.76)	3(0.71)	3	3	3	3	3	3
61	3(0.99)	3(0.94)	3(0.99)	3(0.74)	3(0.71)	3(0.62)	3—	3	1	1	3	3
62	3(0.99)	3(0.95)	3(0.97)	4(0.50)	3(0.59)	3(0.38)	3	3	3	3	3	3
63	3(1.00)	3(0.93)	3(0.97)	3(0.69)	3(0.70)	4(0.61)	3	3	3	3	3	3
64	3(1.00)	3(0.93)	3(0.99)	3(0.81)	3(0.69)	3(0.72)	3	3	3	3	3	3
65	3(1.00)	3(0.96)	3(1.00)	4(0.76)	3(0.65)	3(0.63)	3	3	3	3	3	3
66	3(0.99)	3(0.91)	3(0.98)	4(0.47)	3(0.56)	3(0.55)	3	3	3	3	3	3
67	3(0.98)	3(0.78)	3(0.96)	4(0.59)	3(0.64)	3(0.41)	3	3	3	3	3	3
68	3(1.00)	3(0.98)	3(0.99)	4(0.49)	3(0.79)	3(0.42)	3	3	3	3	3	3
69	3(0.99)	3(0.98)	3(0.99)	2(0.42)	3(0.69)	3(0.38)	3	3	3	3	3	3
70	3(0.99)	3(0.93)	3(0.98)	3(0.51)	3(0.66)	3(0.62)	3	3	4	1	3	3
71	3(1.00)	3(0.97)	3(0.98)	3(0.49)	3(0.62)	3(0.63)	3	3	3	3	3	3
72	3(1.00)	3(0.93)	3(0.98)	3(0.80)	3(0.72)	3(0.77)	3	3	3	3	3	3
73	3(0.99)	3(0.98)	3(0.97)	4(0.94)	3(0.58)	4(0.79)	3	3	3	3	3	(3)
74	3(0.99)	3(0.97)	3(0.98)	3(0.35)	3(0.52)	3(0.58)	3	3	3	3	3	3

[a]Original variables. [b]BAYES weights are given in parentheses. [c]Ten orthogonal features. [d]For objects where the probability measure $P(X/G) < 0.0005$, the class number is followed by a minus sign. For all objects in the training and test sets, the class membership probability $P(G/X) = 1.000$. [e]Autoscaled data over the training set. [f]Variable 9 is excluded because of low modelling and discriminating powers. [g]For SIMCA the classification result for an object is given in the following way: (i) the class number for the closest class is given; (ii) if the F-test (eqn. 2) is not fulfilled the class number is given in parentheses; (iii) if the F-test defined by eqn. (3) is fulfilled, the class number is underlined. This means that an object denoted $\underline{q}$ is uniquely classified to class q, and an object with the class number in parentheses is considered to be an outlier.

was done. From the data base 30 objects were picked by using random sampling numbers to form a separate test set, and the classification was repeated. With BAYES, 15 out of these 30 objects were incorrectly classified. However, the classification of the training set was still high (92%). A χ^2-test showed that the test-set classification was not consistent with the result expected from the training set (Table 2). This result confirmed the suspicion of an overfitting of the training set with BAYES. LDA, KNN and SIMCA were checked in the same way and with the same test set. For the two latter methods the results from the training and test sets were consistent with each other; however, the result for LDA also indicated an overfitting (Table 2).

CARBONYL

For all the test-set classifications presented in Table 3, the training sets are correctly classified except for KNN, $N = 3$, for which two objects were classified incorrectly. It is noticeable that for all methods the *cis* compounds (33—35) are classified in class two, i.e. with the sterically hindered compounds. For objects 14—32, there was no advance information of their class membership when twisted conformers are quite possible: such compounds can be expected not to show typical *cis* or *trans* behaviour. For the different classifications, inspection of Table 3 also reveals discrepancies in the classification for most of the compounds 14—32. The BAYES weights for these compounds are in most cases much smaller than those in the training-set and the low $P(X/G)$ values in LDA for most of the objects show that they are far away from the class centroid. SIMCA indicates that these objects are closer to the *trans* class (class one) but only two of them are uniquely classified to this class.

TABLE 2

Classification results for the WELDERS data with and without a test set[a]. The percentage of objects classified correctly is given

Method	Training set 149 objects	Training set 119 objects	Test set 30 objects	χ^{2b}
BAYES[c]	87[d]	92	50	91.0
LDA	42	46	23	6.6
KNN ($N = 9$)	39	35	40	0.6
LLM[e]	0	0	0	
SIMCA ($A = 1$)	36[f]	42	37	0.6

[a]All calculations on $\ln(1 + y)$ transformed data followed by autoscaling. [b]Null hypothesis: the classification result of the test set is consistent with that of the training set. $\chi^2_{\text{crit. } 0.05} = 3.8$. [c]17 orthogonal features. [d]Calculations without orthogonalization gave 62% correct classification and after gauss or spline smoothing 42% and 82% correct classification. [e]3 features chosen by eigenvector reduction or by SELECT. [f]None of these was uniquely classified, showing the very small separation between the classes.

OIL

When the variable distributions were skew, the originally observed y-values were transformed to $\ln(1 + y)$ followed by autoscaling. BAYES, LDA, KNN and SIMCA all gave meaningful classifications; SIMCA and BAYES gave the highest rate of correctly classified objects (Table 4). No separation of the

TABLE 3

A method comparison of the test-set classification for the CARBONYL data

Obj. no.	BAYES[a]	BAYES[b]		LDA[d]	KNN[e]		LLM[e,f]		SIMCA[e,h,i] $A = 2$
	Direct	Direct	Gauss[c]		$N = 1$	$N = 3$	Eigenvector reduction	SELECT	
14	1(0.96)	1(0.51)	1(0.77)	2—	1	1	1	1	(1)
15	1(0.79)	1(0.52)	1(0.80)	1	1	1	1	1	1
16	1(0.99)	2(0.79)	1(0.84)	1—	1	1	1	1	1
17	1(0.96)	1(0.52)	1(0.84)	1—	1	2	1	1	(1)
18	2(0.50)	2(0.83)	1(0.50)	2—	1	2	1	1	(2)
19	2(0.53)	1(0.78)	1(0.54)	1—	1	1	1	1	(1)
20	1(0.56)	2(0.83)	1(0.52)	2—	1	2	1	2	(2)
21	1(0.54)	2(0.99)	2(0.57)	2—	2	2	1	2	(2)
22	1(0.80)	2(0.94)	1(0.79)	1	1	1	1	1	(1)
23	1(0.95)	1(0.94)	1(0.71)	1—	1	1	1	1	(1)
24	1(0.94)	2(0.92)	1(0.79)	1—	1	1	1	1	(1)
25	1(0.53)	1(0.55)	1(0.58)	1—	2	2	1	1	(1)
26	2(0.50)	2(0.76)	1(0.79)	1	1	2	1	1	(1)
27	1(0.99)	1(0.95)	1(0.84)	1—	1	1	1	1	(1)
28	1(0.99)	1(0.52)	1(0.78)	1	1	1	1	1	(1)
29	1(0.53)	2(0.80)	1(0.55)	1—	2	2	1	1	(1)
30	1(0.50)	1(0.55)	1(0.82)	1	1	2	1	1	(1)
31	1(0.94)	1(0.79)	1(0.84)	1—	1	1	1	1	(1)
32	1(0.94)	1(0.80)	1(0.83)	1—	1	1	1	2	(1)
33	2(0.99)	2(0.99)	2(0.65)	2	2	2	2	2	2
34	2(0.94)	2(0.99)	2(0.66)	2	2	2	2	2	2
35	2(0.99)	2(0.96)	2(0.59)	2	2	2	2	2	(2)

[a]7 original variables. [b]7 orthogonal features. [c]Spline approximation gave the same result. [d]See footnote d, Table 1. [e]Autoscaled data. [f]2 features. [g]Variance and Fisher weights gave the same results. [h]6 variables; variable 3 excluded because of low modelling and discriminating powers. [i]For the presentation of the SIMCA classification, see footnote g, Table 1.

TABLE 4

A comparison of the classification results for the OIL data[a]

Method	Correctly classified objects (%)
BAYES[b]	92
LDA	88
KNN, $N = 1$[c]	76
SIMCA ($A = 2$)[d]	94

[a]All calculations on $\ln(1 + y)$ transformed data. [b]Orthogonal features. [c]$N = 3$, 74% correctly classified. [d]15% were uniquely classified. None of the incorrectly classified objects was uniquely classified to another class.

classes was obtained with LLM. A comparison of the results from BAYES, LDA and SIMCA classifications with 6, 13 and 8% incorrectly classified objects showed that 20% of all objects were classified incorrectly in at least one of these cases. In addition, when $N = 1$ and $N = 3$ in KNN, 24 and 26% of the objects were classified incorrectly. However, 31% of the objects were classified incorrectly in at least one of these KNN classifications. These results show that the classes overlap partly. This is explicitly indicated by the SIMCA classification where just 15% of the objects were uniquely classified. The small difference in BAYES weights, and the class membership probability measure $P(G/X)$ in LDA for most objects between the closest and next closest class also reflected this fact.

IRIS

All the methods classified class one correctly, and small differences in the classification of the other two classes were found with only a few objects misclassified (Table 5). For the methods that give some kind of class membership probability measurement (BAYES, LDA and SIMCA), these measures were compared for the objects classified incorrectly in one classification from each of these methods (Table 6). Of these objects, object 134 is classified incorrectly in all three cases. For the other objects classified incorrectly with BAYES, none is classified incorrectly with LDA, and where the class membership probabilities $P(G/X) > 0.70$. In SIMCA all these objects except object 135 were within the 95% confidence interval for the correct class. The objects classified incorrectly in LDA (except 134) were also classified incorrectly by SIMCA, but not with BAYES. The additional objects classified incorrectly with SIMCA (69 and 132) had high BAYES weights and high class membership probabilities in LDA. Thus for these incorrectly classified objects it is hard to find a major feature in common for the probability measures given for the different methods. However, the BAYES weights and the $P(X/G)$ values in LDA tend to be lower than the typical values for the objects classified correctly for all methods. It is also noticeable that, among all objects classified incorrectly in BAYES and LDA, none is uniquely classified in SIMCA, a finding that also holds for all objects classified incorrectly in at least one of the classifications presented in Table 5. The SIMCA and LDA classifications are also more similar to each other than to the BAYES classifications.

NMR

All methods gave correct classification of the training set with the seven original variables. In contrast, all methods gave one or more incorrectly classified objects for the test set: objects 21, 27, 31 and 40—43 were the most frequently misclassified.

For this special problem, SIMCA is attractive because it gives direct information of the variables irrelevant to the classification problem. It also reveals those objects which do not show typical class membership behaviour. For SIMCA, the variables 1—3 and 5 had low discriminating powers [22].

TABLE 5

A method comparison for the IRIS data. Objects numbers for incorrectly classified objects are given in parentheses

Method	Training set	Test set
BAYES[a]	96%(73, 107, 120)	93%(78, 127, 134, 135, 139)
BAYES[c]	100%[b]	91%(78, 88, 89, 134, 139, 147, 150)
KNN, $N = 3$[a]	97%(107, 120)	93%(84, 134, 135, 139, 150)
LDA[d]	99%(71)	96%(84, 130, 134)
LLM[a]	100%	92%(42, 126, 130, 132, 134, 135)
SIMCA ($A = 2$)[a]	97%(69, 71)	95%(84, 130, 132, 134)

[a]Autoscaled data. [b]Spline smoothing gave the same results. For gauss smoothing, object 73 is correctly classified but 53 is incorrectly classified. [c]4 orthogonal features. [d]Original variables.

TABLE 6

A comparison of the class membership probability measures given in a BAYES, a LDA and a SIMCA classification for the IRIS data. Only the objects incorrectly classified in at least one of the classifications are given. The probability measures are underlined for incorrectly classified objects

Obj. no.	BAYES[a]	LDA[a]		SIMCA[b]	
	weights[c]	$P(G/X)$[c]	$P(X/G)$[d]	$F^e(2)$	$F^f(3)$
Training set					
69	0.80 0.20	0.96 0.04	0.07	3.5	
71	0.67 0.35	0.57 0.43	0.03	3.7	
73	0.50 0.50	0.92 0.08	0.11	2.6	1.1
107	0.58 0.42	0.98 0.02	0.36	0.6	17.0
120	0.70 0.30	0.75 0.25	0.02	0.7	8.0
Test set					
78	0.72 0.28	0.88 0.12	0.20	1.6	
84	0.64 0.36	0.73 0.27	0.11	4.7	
127	0.53 0.47	0.81 0.19	0.15	1.5	3.5
130	0.99 0.01	0.50 0.50	0.06	4.9	
132	1.00 0.00	0.99 0.01	0.25	3.3	
134	0.51 0.48	0.93 0.07	0.22	4.7	
135	0.60 0.40	0.71 0.21	0.05	4.8	1.5
139	0.65 0.35	0.74 0.26	0.09	1.9	2.2
Typical[g]	>0.95	>0.95	>0.10	<3.2	>20.0

[a]Original variables. [b]Autoscaled variables. [c]BAYES weights and $P(G/X)$ values (probabilities for class membership) for the closest and next closest class. [d]Probability of the closest object being that distance from the class centroid. [e]Test for class-typical behaviour (eqn. 2). [f]Test (eqn. 3) showing how close an object is to the model of its next closest class compared with that of its closest class. The F-value is given only if the object is correctly classified. [g]Typical values for objects classified correctly for all methods and uniquely in SIMCA.

TABLE 7

A method comparison of the test-set classification for the NMR data

Obj. no.	Known class	BAYES[a]	KNN[a] $N = 1$	LDA[b]	LLM[a]	SIMCA[c, d] $A = 1$
16	1	1(1.00)	1	1(0.03)	1	1
17	2	2(0.96)	2	2(0.00)	2	2
18	1	1(1.00)	1	1(0.06)	1	1
19	2	2(0.54)	2	2(0.29)	2	2
20	1	1(1.00)	1	1(0.38)	1	1
21[e]	2	1(1.00)	1	1(0.00)	1	1
21[f]	2	1(0.51)	1	1(0.00)	1	2
22	1	1(1.00)	1	1(0.02)	1	1
23[e]	2	2(0.96)	1	2(0.03)	1	1
23[f]	2	2(1.00)	2	2(0.00)	2	2
24	1	1(1.00)	1	1(0.11)	1	1
25	2	2(0.96)	2	2(0.47)	2	2
26	1	1(1.00)	1	1(0.51)	1	1
27	2	1(0.57)	1	1(0.00)	1	2
28	1	1(1.00)	1	1(0.00)	1	1
29	2	2(0.96)	2	2(0.66)	2	2
30	1	1(0.55)	1	1(0.00)	1	1
31	2	2(0.98)	2	1(0.00)	2	2
32	1	1(1.00)	1	1(0.00)	1	1
33	2	1(0.56)	2	1(0.57)	1	2
34	1	1(1.00)	1	2(0.00)	1	1
35[e]	2	2(0.60)	2	1(0.00)	2	2
35[f]	2	2(0.97)	2	2(0.00)	2	2
36	1	1(1.00)	1	2(0.07)	1	1
37	2	1(0.96)	2	1(0.02)	2	2
38	1	1(1.00)	1	1(0.33)	1	1
39	2	2(0.54)	2	2(0.90)	2	2
40	1	1(0.96)	1	1(0.00)	1	2
41	2	2(0.60)	2	1(0.01)	2	2
42	1	1(0.96)	2	1(0.00)	1	2
43	2	2(0.61)	1	1(0.66)	1	1

[a]The three most discriminating features with SELECT after autoscaling. [b]Seven original variables. The $P(G/X)$ values are given in parentheses. For all objects the class membership probability $P(X/G) = 1.00$. With six variables the same poor prediction of the test set was obtained. [c]3 variables. [d]For an explanation of the presentation of the SIMCA classification, see Table 1 footnote g. [e]Incorrect assignment. [f]Correct assignment.

It is illuminating that if only these four low-discriminating variables were used in a reclassification, only a 73% correct classification of the training set was obtained; only 20% of the objects were uniquely classified, showing that these variables contained very little class membership information. A classification with the four low discriminating variables excluded gave a 100% classification of the training-set, but in the test-set objects 21, 23, 31, 35 and 40—43 were outliers or were classified to the wrong class (Table 7).

An examination of the residuals for these objects showed that for the compounds 21, 23 and 35 the ^{13}C-n.m.r. spectra were not correctly assigned. After a reassignment of these compounds, they were classified correctly [22]. Compound 31 showed large residuals for one of the variables, indicating behaviour inconsistent with the training set. The compounds 40—43 similarly showed large residuals for one of the variables (the shift for the C_7 carbon). This special behaviour has been noticed for other compounds which, like compounds 40—43, have methyl substituents on the C_7 carbon [23].

With LDA and with the original seven variables, the misclassifications are numerous in the test set and for most of the objects the $P(G/X)$ values were low (see Table 7). Of the incorrectly assigned objects detected with SIMCA, object 21 was incorrectly and object 23 correctly classified with the incorrect as well as the correct assignments. Object 35 was only correctly classified with the correct assignment. Objects 40—43 were all classified in class 1. The variable reduction methods available in LDA were also used and they all excluded variable 7, but no improvement of the test-set classification was thereby obtained.

A feature reduction with SELECT gave (with both variance and Fisher weights) variable 7 the highest weight followed by two features mainly derived from variables 6 and 4. These three features were used in a reclassification with BAYES, LLM and KNN. With the different assignments of 21 and 35, LLM classified object 21 in both cases to the incorrect class and 35 to the correct class, thus giving no guide-lines for incorrect assignments. With BAYES the objects 23 and 35 were classified correctly with incorrect as well as correct assignments. The objects 40—43 were all classified to the correct class. For KNN, $N = 1$, object 23 was classified incorrectly with incorrect assignment and correctly with correct assignment. For the objects 21 and 35 no conclusions about their assignments could be drawn.

From these results it is obvious that the far-reaching conclusions drawn from the SIMCA computations are due to an effective variable reduction method, combined with the ability to deal with outliers and to study individual residuals, conclusions that cannot easily be drawn from the versions of BAYES LDA, KNN and LLM used here.

Feature reduction with SELECT and eigenvector reduction

For the classification problems where the classes in the training set are well-separated, i.e. the ARCH, NMR and CARBONYL data, feature reduction with SELECT or eigenvector reduction can be done with a retained 100% correct classification. For example, for the CARBONYL data a reduction from 7 to 4 or even 2 features with SELECT or eigenvector reduction still gives a 100% correct classification of the training set with BAYES, KNN and LLM.

For problems where the classes are closer to each other in the M-space, resulting in incorrectly classified objects in the training set, e.g. the IRIS, OIL and WELDERS data, no general improvements of the classifications were found after feature reduction. This is exemplified (Table 8) by systematic investigation of the IRIS data.

TABLE 8

Feature reduction of the IRIS data. The percentage of correctly classified objects is given in each case.

	BAYES	KNN, $N = 3$	LLM
4 orthogonal features	100, 93	96, 93	100, 92
2 orthogonal features	92, 91	91, 92	no separation
2 features with SELECT[a]	97, 93[b]	97, 92	no separation

[a]Variance weight, no improvement with Fisher weight.
[b]No improvement with gauss or spline smoothing of the histograms.

Variable reduction with SIMCA and SPSS

When SIMCA was used, all variables contained class-separating information for the ARCH, OIL and IRIS data. For the WELDERS data all variables showed low discriminating and modelling powers, reflecting the poor information on class membership contained in this data base. For the CARBONYL data, only one of the seven variables was excluded (variable 3) and for the NMR data, four of the variables (1—3 and 5) were excluded because of low discriminating powers. The little class-separating information in these was verified in a separate classification with only these four variables.

The variable reduction methods in SPSS used together with LDA gave in just one case a variable reduction on the probability level $P = 0.50$. This occurred in the NMR data where variable 7 was excluded. This variable is one of the three most discriminating variables according to the SIMCA and SELECT methods. Instead, variables 1—3 and 5, which have low discriminating power according to SIMCA and SELECT, were included.

In LDA, the absolute values of the standardized discriminant coefficient for a variable is a measure of the contribution from the variables to the discriminant function [5]. For example, for the NMR data the variables 3—6 had the highest coefficients. Thus these coefficients as a measure of the variable relevance also give a quite different result compared with the SELECT weights and the discriminating powers for the variables given in SIMCA.

Scaling

The effects of the scaling of the variables were not systematically investigated. BAYES, LDA and LLM are not affected by autoscaling. Since there were restrictions, such as the limitations on the objects/features ratio when scaling was used to optimize the classification for KNN, it was decided to present all KNN calculations on autoscaled data. For SIMCA, there are no such restrictions, but the classification results are in most cases retained or improved by autoscaling.

Smoothing and orthogonalization in BAYES

In no case did smoothing of the histograms and orthogonalization of the variables give dramatic improvements of the classification results of the training set. For the classification of the test sets, large discrepancies between different BAYES approaches are prevalent (see, e.g. Tables 1 and 3). But from this investigation it is difficult to give any particular approach preference.

DISCUSSION

On the whole, the five pattern recognition methods used performed well. They are well suited to solve classification problems when only classification is of interest and when the classes are well-separated, as in the training-set classification of the NMR, CARBONYL and ARCH data. However, some reservations are in order. BAYES, even if optimal as a classifier when the class distributions are exactly known, seems to be less reliable when used in the way described in this investigation. When the class distributions are unknown, they are estimated from the training set, i.e. the objects to be classified are also used to determine the class distributions. This often leads to an over-optimistic classification of the training set, e.g. for the WELDERS data. Too few objects in the classes lead to poor estimates of the real class distributions, increasing the risk of incorrectly classified objects in the test set. Thus, for chemical applications where the class distributions are rarely known and small data sets are common, BAYES should be used with caution. Application of LDA to the WELDERS data also revealed an over-optimistic test-set classification. However, the overfitting in this case seemed to be less pronounced, as LDA and SIMCA gave a very similar classification of the test sets for the ARCH, CARBONYL and IRIS data.

Futhermore, LLM and LDA should not be used when the ratio of objects to features is less than 3. Another drawback with LLM is that the feedback procedure for the same data set can converge to different hyperplanes [24].

A very poor prediction ability for LLM has also been found by Weisel and Fasching [25] using the well known leave-one-out method: one object at a time was excluded from the training set where the objects were 100% correctly classified. When the excluded objects were treated as test-set objects, only 68% of them were correctly classified.

When the classes are closer to each other, or partly overlap in the M-space, the scatter in the classifications from one method to another can be wide. Thus, to trust the classification results from only a single method in such cases might be inadvisable since the different methods rely on different classification strategies. This dilemma can be partly circumvented by studying the classification results from different methods, not with the aim of finding the "best" classification, but with the aim of understanding the data structure and detecting the objects for which the class-separating information is low. The IRIS, OIL and WELDERS data are typical examples where the classes overlap one another and where the methods sometimes give different classification results.

To solve a problem just as a classification problem or as a so-called level-one problem [26] is rarely the sole aim of a chemical pattern recognition method. Of interest in most chemical applications, like the ARCH, CARBONYL and NMR problems, is to obtain information on which objects show atypical class behaviour, i.e. outliers must be considered. This kind of problem (a level-two problem [26]) demands that the method be capable of describing each class as a closed structure in the M-space. Of the methods tested here, only SIMCA fulfils this demand when eqns. (2) and (3) describe each class separately as a closed hypervolume in the M-space. Neither the probability measures $P(G/X)$ and $P(X/G)$ in LDA nor the BAYES weights can describe a closed structure in the M-space. Thus the large discrepancies for the class membership probabilities from method to method found for some of the objects in, e.g., the ARCH, CARBONYL and IRIS data are not surprising. For problems on this level, BAYES, LDA and LLM cannot be recommended when the versions of these methods used are not designed for problems where outliers must be considered. KNN can provide some information of outliers if the actual nearest-neighbours distances of a test-set object are compared with the typical nearest-neighbours distances of the training set, although stringent tests of significance are not described. Also the use of display methods [27] can give some guidance on outliers.

This investigation emphasizes that methods which can deal with outliers are desirable in many chemical applications, since it can rarely be stated in advance that there are no outliers among the objects in the training or test sets. To our knowledge only one method apart from SIMCA, namely the entropy minimax method [28], can deal with level-two problems. However, it is probable that some of the existing level-one methods, e.g. KNN, can be further developed as suggested above to deal with outliers.

This work was supported by the Swedish Natural Science Research Council (NFR). M. S. thanks the NFR for a travel grant. We thank Dr. Alice Harper and M. da Koven for helpful discussions.

REFERENCES

1 B. R. Kowalski in C. E. Klopfenstein and C. L. Wilkins (Eds.), Computers in Chemical and Biological Research, Vol. 2, Academic Press, New York, 1974.
2 B. R. Kowalski, Anal. Chem., 47 (1975) 1152A.
3 L. Kanal, IEEE Trans. Inform. Theory, IT-20 (1974) 697.
4 D. L. Duewer, J. R. Koskinen and B. R. Kowalski, ARTHUR; (available from B. R. Kowalksi).
5 A. M. Harper, D. L. Duewer and B. R. Kowalski, in B. R. Kowalski (Ed.), Chemometrics, Theory and Practice, Am. Chem. Soc. Symp. Ser., No. 52, 1977.
6 N. H. Nie, C. H. Hull, J. G. Jenkins, K. Steinbrenner and D. H. Brent, SPSS: Statistical Package for Social Sciences, McGraw-Hill, New York, 1975.
7 S. Wold, SIMCA-2T manual (available from S. Wold).
8 B. R. Kowalski and C. F. Bender, Pattern Recognition, 8 (1976) 1.
9 G. F. Box and G. C. Tiao, Bayesian Interference in Statistical Analysis, Addison-Wesley, New York, 1973.

10 N. B. Nilsson, Learning Machines, McGraw-Hill, New York, 1965.
11 N. A. B. Gray, Anal. Chem., 48 (1976) 2265.
12 T. H. Cover and P. E. Hart, IEEE Trans. Inform. Theory, IT-13 (1967) 21.
13 S. Wold, Pattern Recognition, 8 (1976) 127.
14 S. Wold, Technometrics, (1978) in press.
15 S. Wold and M. Sjöström, in B. R. Kowalski (Ed.), Chemometrics, Theory and Practice, Am. Chem. Soc. Symp. Ser. No. 52, 1977.
16 B. R. Kowalski, T. F. Schatzki and F. H. Stross, Anal. Chem., 44 (1972) 2176.
17 R. Mecke and K. Noack, Chem. Ber., 93 (1960) 210.
18 S. Wold and M. Sjöström, in N. B. Chapman and J. Shorter (Eds.), Correlation Analysis in Chemistry, Plenum Press, New York, 1978.
19 U. Ulfvarsson and S. Wold, Scand. J. Work, Environ. Health, 3 (1977) 183.
20 D. L. Duewer, B. R. Kowalski and T. F. Schatzki, Anal. Chem., 47 (1975) 1573.
21 R. A. Fisher, Ann. Eugenetics, 7 (1936) 179.
22 M. Sjöström and E. Edlund, J. Magn. Reson., 25 (1977) 285.
23 J. B. Stothers, C. T. Tan and K. C. Teo, J. Magn. Reson., 20 (1975) 570.
24 J. R. McGill and B. R. Kowalski, J. Chem. Inf. Comp. Sci., 18 (1978) 52.
25 C. P. Weisel and J. L. Fasching, Anal. Chem., 49 (1977) 2114.
26 C. Albano, W. Dunn, U. Edlund, E. Johansson, B. Norden, M. Sjöström and S. Wold, Anal. Chim. Acta, 103 (1978) 429. (Computers and Optimization in Analytical Chemistry, Amsterdam, April 1978.)
27 B. R. Kowalski and C. F. Bender, J. Am. Chem. Soc., 95 (1973) 686.
28 A. R. C. Wong and T. S. Liu, IEEE Trans. Comp., 24 (1975) 158.

Analytica Chimica Acta, 112 (1979) 31—44
Computer Techniques and Optimization

CALCULATION OF ADSORPTION-RELATED PARAMETERS FROM A.C. POLAROGRAPHIC DATA: Basis and Computer Programs

TIMOTHY E. CUMMINGS,** MICHAEL KATZ*** and PHILIP J. ELVING*
University of Michigan, Ann Arbor, Michigan, 48109 (U.S.A.)
(Received 13th November 1978)

SUMMARY

Phase-selective alternating current polarography can be advantageously used for the observation and quantitative description of adsorption at a solution—electrode interface; for example, in the absence of a faradaic process, the quadrature current component is simply related to the interfacial differential capacity. Such a.c. measurements are especially advantageous for the occasional investigation of adsorption. The basis of using such measurements is considered; the data analysis is examined; specifics for computer calculation of differential capacity, surface charge density, and relative surface excess, and the requirements for data smoothing are described. The computer programs developed are sufficiently general for handling special situations.

The thermodynamic properties of and conceptual models for the electrical double layer have been reviewed [1—6], as have experimental and theoretical progress in double-layer research and remaining problems [7, 8], the experimental measurement of adsorption at electrodes [9], and the adsorption of organic compounds [10—12]. However, most electrochemists seem to avoid detailed adsorption studies, probably because of (a) the tedium of employing either a capillary electrometer to obtain interfacial surface tension data or an a.c. impedance bridge to obtain differential capacitance data, (b) the difficulty involved in precise and accurate determinations of the potential of zero charge (p.z.c.) [13] and of surface tension at the p.z.c., and (c) the quite lengthy analysis of experimental double-layer data.

Although digital computers have made the data analysis less arduous, computer program development can be very costly in terms of man-hours and computer expense, particularly for one with limited experience in programming. Mohilner and Mohilner [14] have discussed the basis for curve fitting and data smoothing with emphasis on electrocapillary data analysis;

**Present address: Department of Chemistry, University of Miami, Coral Gables, Florida, 33124, U.S.A.
***Present address: Diamond Shamrock Electrosearch S.A., 3 Rue de Trainex, 1227 Carouge, Switzerland.

no main program for combining the subroutines to form a complete program is given and the smoothing routines may be more sophisticated than necessary for capacitance-data analysis. MacDonald [15] has described a program for the determination, based on capacitance data, of relative surface excesses of uncharged species, which totally desorb at one potential extreme before the onset of faradaic activity; the program is not mathematically flexible in terms of handling special situations.

In an attempt to minimize the difficulties indicated, various methods of data analysis, based on alternating current (a.c.) polarography as a means of data acquisition, have been compared with the aim of developing the most efficient combination of data acquisition and analysis which would meet the needs of routine adsorption studies.

In the absence of faradaic current, the quadrature current component obtainable on phase-selective a.c. polarography (as well as the total a.c.) is capacitive and directly related to the differential capacity. As a.c. polarography is basically an automated non-nulling version of a.c. bridge measurement, it should have the advantages of the bridge method except for the slight loss in precision inherent in techniques which do not employ null detection. Phase-selective a.c. polarographs are commercially available and are becoming quite common.

Computer programs were developed for analyzing data on uncharged adsorbates; these are sufficiently general for handling most special situations. Because of space limitation, the programs are not given; program listings or decks can be made available. Because some investigators may desire to develop comparable programs to meet special requirements or may find the programming language employed (FORTRAN IV) unacceptable, precautions are noted which must be observed in efficient and successful programming of adsorption data analysis.

MATHEMATICAL BASIS

The physical situation involves an electrochemical cell, in which the solution to be investigated contains solvent, supporting electrolyte (including buffer if necessary), and the uncharged adsorbate of interest with only the absorbate concentration being varied. The applied potential is varied over the range of interest and capacitance data are acquired in the form of capacity-current magnitudes, e.g., a.c. polarographic quadrature-current components.

The capacity current, $I_{c'}$ is converted to capacitance, C', and then to differential capacitance, C:

$$C' = I_c/(2\ \pi\ f\ U_{ac});\ C = C'/A$$

where f is the frequency in Hz of the applied alternating voltage, U_{ac} is its amplitude in V (peak or r.m.s. depending on the instrumental current display utilized), and A is the electrode area.

The relations between C and electrode charge, q, and between q and

interfacial surface tension, γ, are given [3] by

$$q = q' + \int_{E'}^{E} C \, dE \text{ and } \gamma = \gamma' - \int_{E'}^{E} q \, dE \tag{1}$$

where q' is the charge at any arbitrary starting potential E', and where γ' is the surface tension at E'. The relation between relative surface excess, Γ, and γ is given [3] by

$$\Gamma = -(\partial\gamma/\partial\mu)_E \tag{2}$$

A computer must work with these five equations to calculate relative surface excesses from a.c. capacity current data.

The commonest choice of E' is the p.z.c.: $E' = E_z$. Then, $\gamma' = \gamma_z$ and, by definition, $q' = 0$. Equations (1) are then written as

$$q = \int_{E_z}^{E} C \, dE \text{ and } \gamma = \gamma_z - \int_{E_z}^{E} q \, dE \tag{3}$$

To apply eqns. (2) and (3) it is necessary to obtain capacitance data for each adsorbate concentration involved, and to determine E_z and γ_z at each concentration. In a second possibility [15, 16], if a potential region exists in which complete desorption has occurred for all investigated concentrations and faradaic activity is absent, then $\Gamma = 0$ within this potential region and, from eqn. (2) $(\partial\gamma/\partial\mu)_E = 0$, Additionally, within the potential region of complete desorption, $(\partial q/\partial\mu)_E = 0$. If integration of eqns. (1) is begun within that region, then $E' = E_l$ $q' = q_l$, and $\gamma' = \gamma_l$ (subscript l refers to a specific potential at which desorption is complete at all concentrations); by defining a parameter $q^* = q - q_l$, eqns. (1) may be rewritten as

$$q^* = \int_{E_l}^{E} C \, dE \tag{4}$$

$$\gamma = \gamma_l - \int_{E}^{E} q^* \, dE - \int_{E_l}^{E} q_l \, dE \tag{5}$$

Based on eqns. (2) and (5), Γ is given by

$$\Gamma = -(\partial\gamma_l/\partial\mu)_E + \left[(\partial/\partial\mu)\int_{E_l}^{E} q^* \, dE\right]_E + \left[(\partial/\partial\mu)\int_{E_l}^{E} q_l \, dE\right]_E \tag{6}$$

Since γ_l and q_l are independent of adsorbate concentration (see above), eqn. (6) reduces to

$$\Gamma = \left[(\partial/\partial\mu)\int_{E_l}^{E} q^* \, dE\right]_E \tag{7}$$

If a "relative" surface tension is defined as $\gamma^* = \int_{E_l}^{E} q^* \, dE$, then eqn. (7) may be rewritten as

$$\Gamma = (\partial\gamma^*/\partial\mu)_E \tag{8}$$

It can be shown that, within the potential region for which $(\partial\gamma/\partial\mu)_E = 0$, $(\partial C/\partial\mu)_E = 0$; hence, coincidence of the capacitance curves, independent of adsorbate concentration, is an experimentally observable verification for the applicability of eqns. (4) and (8). Coincidence of the capacitance curves at one potential should not be adjudged as confirming the applicability; coincidence of all capacitance curves over a finite potential range is necessary.

PROGRAMS DEVELOPED

The general formula involved in a least-squares fit, $y = \Sigma_{i=0}^{n} \pi_i x^i$, is a power-series relation between independent variable, x, and dependent variable, y, with the coefficients, π_i, to be determined. The relations involved (capacitance and charge as a function of potential) in the case of a quadratic polynomial are as follows (the π coefficients differ for the parameter involved):

$$C = \pi_0 + \pi_1 E + \pi_2 E^2 \tag{9}$$

$$q = \pi_0 + \pi_1 E + \pi_2 E^2 \tag{10}$$

Because the chemical potential is generally unknown, it is necessary to express eqns. (2) and (8) in terms of activity, a, or concentration, c, through the equation $\mu = \mu^0 + RT \ln a$. At constant ionic strength, the activity coefficient of the adsorbate is independent of adsorbate concentration; hence, $\mathrm{d}(\ln a) = \mathrm{d}(\ln c)$, and $\mathrm{d}\mu = RT\,\mathrm{d}(\ln c)$. Substituting this equation into eqn. (2) yields

$$\Gamma = -(1/RT)\,[\partial\gamma/\partial(\ln c)]_E \tag{11}$$

Thus, the polynomial necessary for least-squares analysis is

$$\gamma = \pi_0 + \pi_1 \ln c + \pi_2 (\ln c)^2 \tag{12}$$

Equations (9), (10) and (12) are also the basis for all least-squares curve smoothing, interpolating, integrating and differentiating involving C', q^* and γ^*, respectively.

Procedures for performing a least-squares fit can be found in numerical analysis books [17]. Figures 1 and 2 give general flow diagrams for setting up the least-squares-matrix equation and for its solution by the Gauss—Jordan method of upper triangulation.

Smoothing and interpolating program

Some smoothing of a.c. polarographic capacity-current data is advisable because these data lack the precision of a.c. impedance-bridge data and small errors may occur in reading the x—y recorder plots. Additionally, within potential regions of smooth capacitance variation, data can be supplied to the program at large potential increments and more closely-spaced interpolated values can be calculated from the smoothing functions, saving considerable man-hours of reading recorder plots and key-punching data.

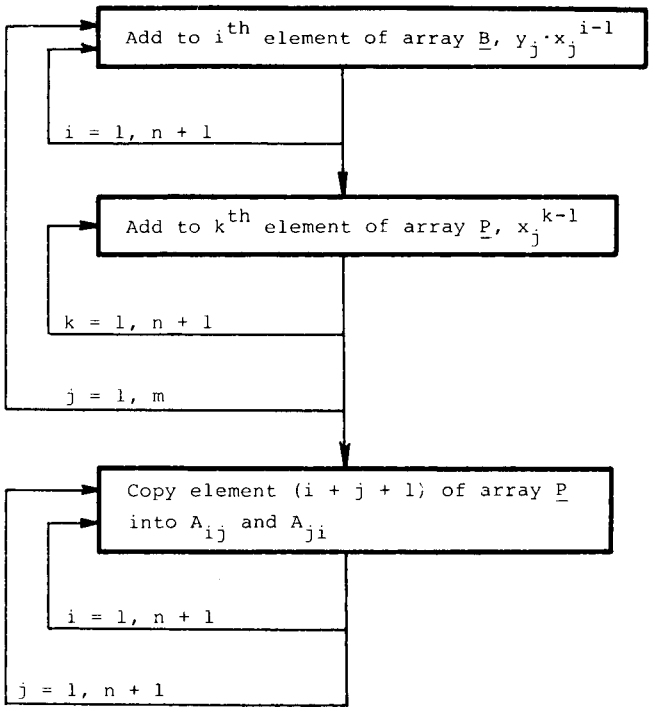


Fig. 1. Procedure for assembling the least-squares matrix, $A\pi = B$.

Because a seriously faulty point can bias a curve fit, a computer-controlled check of each data point is recommended before curve fitting. The procedure described (Fig. 3) separately checks all data points for each concentration except the first and last. Because the computer cannot distinguish fine structure in the data from faulty data points, only regions showing an apparent minimum or maximum ("minimax") in the data can be tested; therefore, as the program proceeds sequentially through the data, any point, C'_i, for which the criterion $C'_{i-1} \leqslant C'_i \leqslant C'_{i+1}$ or $C'_{i-1} \geqslant C'_i \geqslant C'_{i+1}$ is satisfied, is assumed correct, and the program proceeds to check data point C'_{i+1}. If the criterion mentioned is not satisfied, an apparent "minimax" at E_i is indicated. Since a true minimax shows a decreasing slope as the minimum or maximum point is approached, then the criteria

$$\Delta_i/\Delta_{i-1} > 0; |\Delta_{i-1}| > |\Delta_i|; \Delta_{i+2}/\Delta_{i+1} > 0; |\Delta_{i+2}| > |\Delta_{i+1}| \qquad (13)$$

(where Δ is the slope of a straight line connecting the point indicated and the preceding point) can be used to determine whether a true minimax occurs at E_i. If these conditions are satisfied, C'_i is assumed to be correct. A faulty point at E_i is easily corrected by calculation of a capacitance value at E_i from a curve fit to the adjacent points, (E_{i-1}, C'_{i-1}) and (E_{i+1}, C'_{i+1}). If the quadratic fit is used, the third point required can be determined from

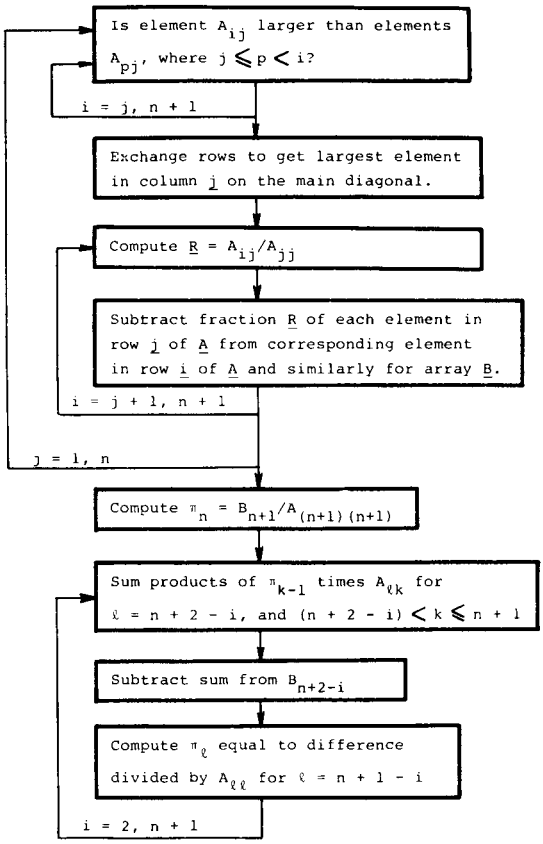


Fig. 2. Solution of n^{th} degree polynomial least-squares matrix equation, $A\pi = B$, for values of π_i.

$|\Delta_{i-1}| - |\bar{\Delta}| = K$; $|\Delta_{i-2}| - |\bar{\Delta}| = J$, where $\bar{\Delta}$ is the slope of a line connecting (E_{i-1}, C'_{i-1}) and (E_{i+1}, C'_{i+1}). If $K < J$, the third point is (E_{i-2}, C'_{i-2}); otherwise, the third point is (E_{i+2}, C'_{i+2}).

Next, the data are smoothed and interpolated by using a least-squares moving polynomial fit to eqn. (9), i.e., m data points are fitted k points at a time ($k < m$) with each fit corresponding to a progression of l points farther into the data ($l < k$); the options available are the values of k, l, and n (the degree of the fit). Figure 4 shows various possibilities for a moving quadratic polynomial fitting four points at a time. Low-order polynomials are preferable, particularly the quadratic case ($n = 2$), and one degree of smoothing is generally sufficient, so that $k = n + 2$ [14]. For optimum smoothing, l should be 1. With the exception of data points 1, 2, $m - 1$, and m, all points are fitted by more than one polynomial (two for $k = 4$) (cf. curve C, Fig. 4); averages of the values generated by each fit add to the smoothness of the data.

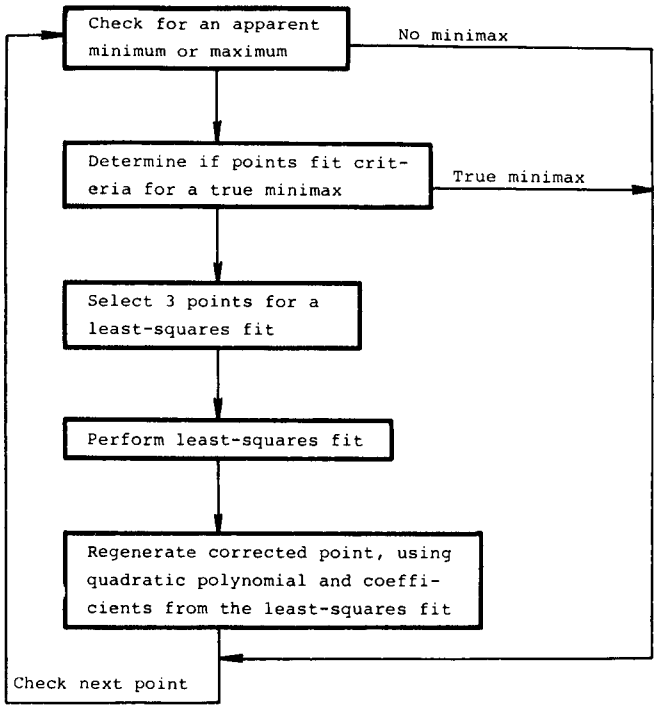


Fig. 3. Check of raw capacitance data to isolate and to correct faulty points.

For each fit, it is acceptable to generate smoothed capacitance data at any potential within the potential region defined by the fitted data points; however, because one or more degrees of smoothing are permitted, the polynomial need not perfectly follow the data points, and the data trend beyond the limits of the points being fitted has no effect on the curve fit. The combination of these two facts indicates that values generated near the potential range extremes may not tend smoothly toward the data outside this range. Therefore, the region in which values are to be generated should be restricted to at least one-half potential increment (magnitude between measured values) from potential range ends for an individual polynomial fit. A flow diagram for smoothing, which employs this restriction and the condition $l = 1$, is shown in Fig. 5.

Once the data for each concentration have been smoothed, printed, punched and plotted, the adsorption parameters can be calculated.

Adsorption parameter program

The program for analyzing the smoothed capacitance data, based on eqns. (2) and (3) (version CP-I) or eqns. (4), (7) and (8) (version CP-II), performs both numerical integration and differentiation and if desired, attempts an adsorption isotherm assignment based on the computed Γ—c profiles. The procedure is outlined in Fig. 6.

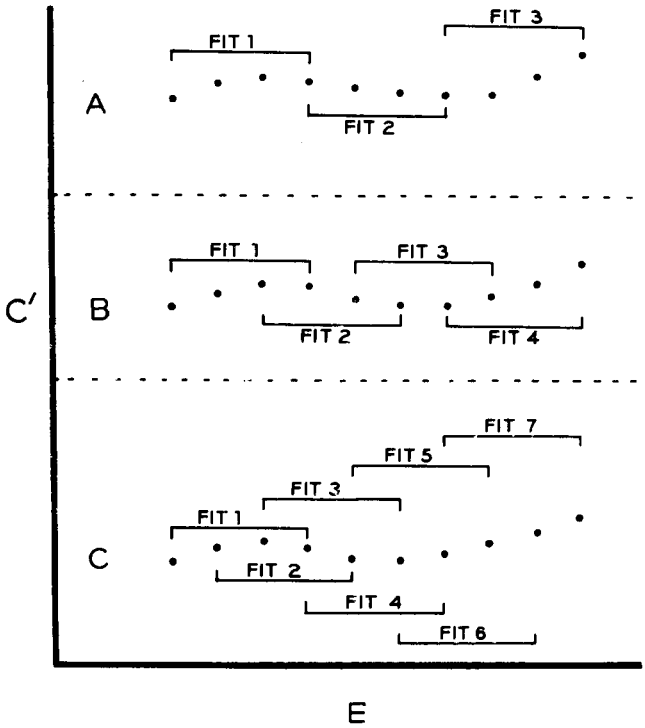


Fig. 4. Overlap of curve-fitting intervals depending on the number of points advanced for sequential fits, by using a four-point fit. A: advance three points. B: advance two points. C: advance one point.

Two alternative procedures are available for numerical integration: Simpson's Rule [17] or the least-squares technique selected which involves fitting a moving polynomial with subsequent integration of the fitted polynomial. Integration of eqn. (9) as the moving polynomial for calculation of q yields

$$\int_{E_i}^{E_{i+1}} C \, dE = \pi_0 (E_{i+1} - E_i) + \pi_1 ((E^2_{i+1} - E^2_i)/2) + \pi_2 ((E^3_{i+1} - E^3_i)/3) \quad (14)$$

whose left-hand side defines the difference in q between E_i and E_{i+1}, so that it can be rewritten as

$$q_{i+1} - q_i = \pi_0 (E_{i+1} - E_i) + \pi_1 (E^2_{i+1} - E^2_i)/2 + \pi_2 (E^3_{i+1} - E^3_i)/3 \quad (15)$$

The coefficients, π_1, determined in the moving-polynomial fit of eqn. (9) are used in eqn. (15). A similar procedure is used to integrate the q—E curves to determine γ.

To minimize integration errors, a second-degree polynomial is used without smoothing. Since three points are fitted, two integrals are calculated for each fit. The polynomial is advanced one point at a time; where the

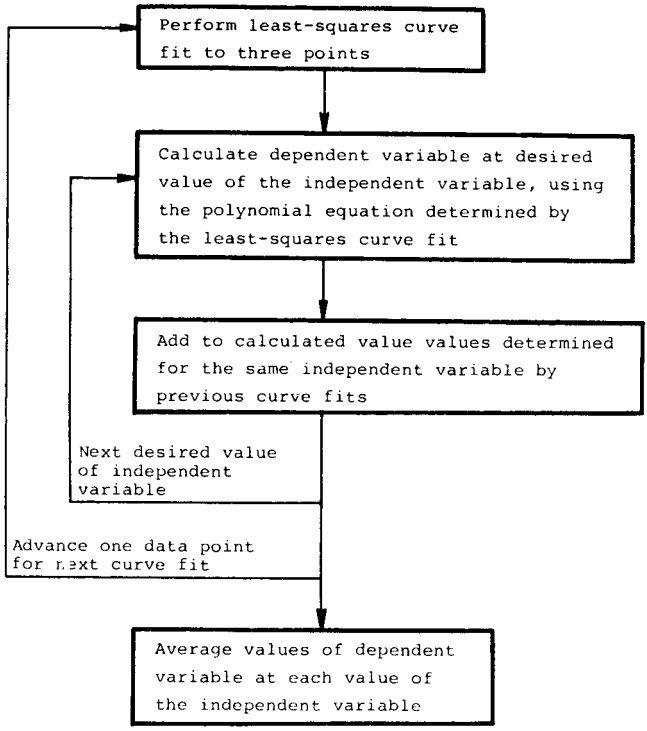


Fig. 5. Smoothing and interpolating procedure.

polynomials overlap, the integrals are averaged. Figure 7 shows a flow diagram for the integration procedure.

Starting values for q (q_1) and γ (γ_1) are required in applying eqn. (15) or the analogous equation for γ. If eqns. (2) and (3) are used, $q_1 = 0$ and $\gamma_1 = \gamma_z$. If eqns. (4), (7) and (8) are used, it is possible to use $q_1^* = 0$ and $\gamma_1^* = 0$, since the values q^* and γ^* are relative and are not concentration-dependent at E_l.

The major difference between the method based on q and γ, and that based on q^* and γ^*, is that the latter permits integration to begin at one potential extreme, whereas the former generally requires integration to begin at a potential which is internal to the potential range. Although this is no problem with a compiler language such as PL1, it presents a problem in FORTRAN IV because data are generally stored in such an order that the corresponding potential sequentially increases or decreases. Thus, for FORTRAN IV programs, the data array to be integrated must be separated into two arrays, properly ordered for integration. It is obvious from the equation

$$\int_{E_l}^{E_i} C\,dE = \int_{E_1}^{E_i} C\,dE - \int_{E_1}^{E_l} C\,dE \qquad (16)$$

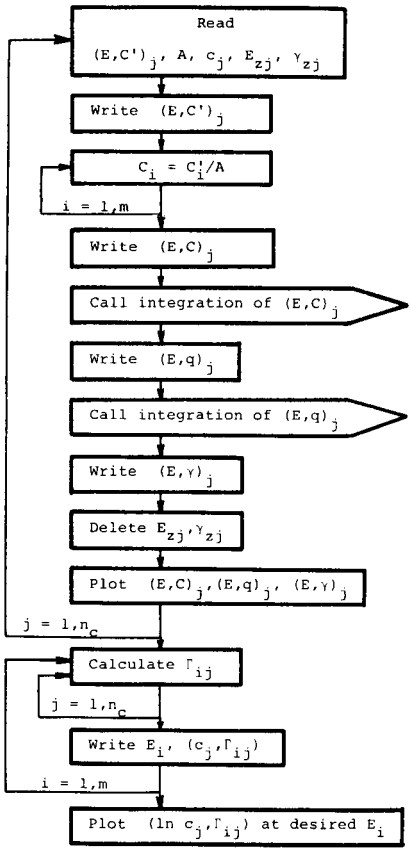


Fig. 6. Main routine procedure, where n_c is the number of concentrations and m is the number of data points per concentration.

that the method based on q^* and γ^* never requires reordering of the array. After γ or γ^* has been computed as a function of potential and concentration, the relative surface excess can be calculated, by using a moving polynomial curve fitted to eqn. (12), whose derivative is

$$\partial\gamma/\partial\,(\ln c) = \pi_1 + 2\,\pi_2 \ln c \qquad (17)$$

From eqns. (11) and (17), it follows that

$$\Gamma = -\,(1/RT)\,(\pi_1 + 2\,\pi_2 \ln c) \qquad (18)$$

For calculations based on γ^*, the negative sign in eqn. (18) is deleted. The blank data are not used in the calculations. Because eqn. (18) is a derivative of the moving polynomial equation, it is unwise to compute Γ at the edge of a fit; hence Γ is not computed for the highest and lowest substrate concentrations.

Adsorption isotherm assignment generally begins by testing for a fit to

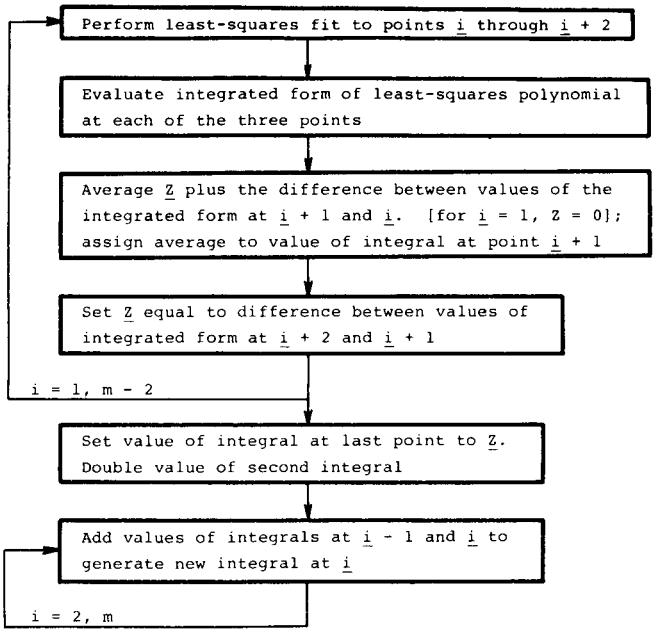


Fig. 7. Integration with the least-squares technique.

either a Langmuir isotherm (eqn. 19) or a Frumkin isotherm (eqn. 20)

$$Kc = \theta/(1-\theta) \tag{19}$$

$$Kc = [\theta/(1-\theta)] \exp(-2a\theta/RT) \tag{20}$$

where it is assumed that $a = c$, Γ_s is the saturation coverage, $\theta = \Gamma/\Gamma_s$, a is the interaction factor, $K = \exp(-\Delta\bar{G}^0/RT)$, and $\Delta\bar{G}^0$ is the standard free energy of adsorption. Equation (20) can be written as $K'c = \theta/(1-\theta)$, where $K' = \exp[-(\Delta\bar{G}^0 - 2a\theta)/RT]$. The adsorption isotherm fit is made here at constant potential. Γ—c profiles are tested one potential at a time as follows: (1) supply Γ_s; (2) calculate θ and $\theta/(1-\theta)$ from Γ at each value of c; (3) calculate K (hence, $\Delta\bar{G}^0$) at each c from eqn. (19); (4) check that $\Delta\bar{G}^0$ is relatively independent of c ("relatively" is defined in terms of accuracy of the data; a systematic change in $\Delta\bar{G}^0$ with c probably means that $\Delta\bar{G}^0$ is not independent of c); (5) if $\Delta\bar{G}^0$ is dependent on c, perform a linear regression on $\Delta\bar{G}^0$ vs. θ. The slope will be equal to $-2a$, and the intercept will be the "true" value of $\Delta\bar{G}^0$ for a Frumkin isotherm. Figure 8 shows a flow diagram for computing Γ and for testing for either a Langmuir or a Frumkin isotherm fit.

Surface excess at constant electrode charge

Because of the simplicity of computing Γ at constant potential, adsorption isotherm fits are made at constant electrode potential. Since there is considerable

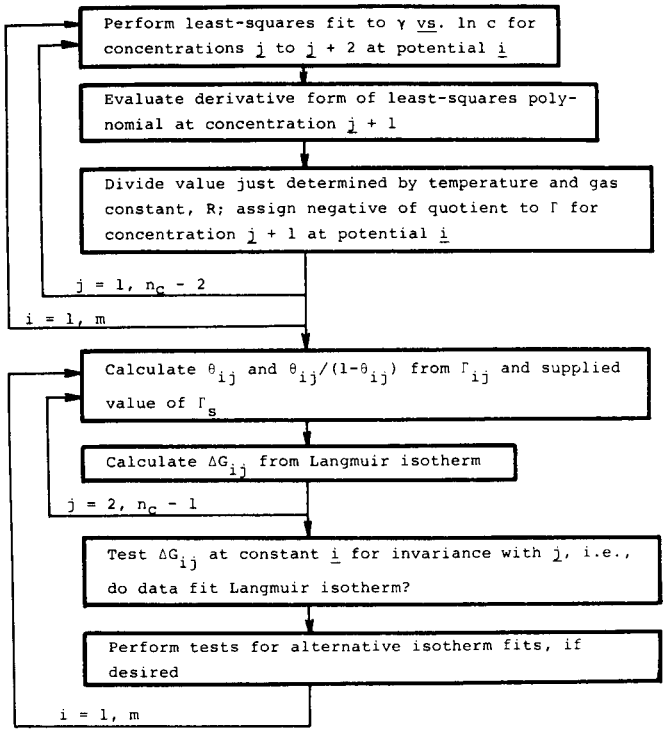


Fig. 8. Calculation of Γ at constant electrode potential and test for isotherm assignment.

controversy (cf. refs. 7 and 8) as to whether adsorption of a neutral species should show isotherm congruency at constant potential or constant charge, Γ—c profiles should be computed at constant charge and isotherm assignment under those conditions attempted.

Surface tension and charge data (available as punched output from the previous program) are supplied along with the values of q at which Γ is to be computed. The program sequentially deals with the q values; for each concentration, it scans the q—E profile to determine in what potential region q occurs. If the value of q is a data point, the corresponding value of E is noted; otherwise, a quadratic least-squares fit of q vs. E, based on eqn. (10), is performed on the three (E,q) points having values of q closest to that requested. Since eqn. (10) is a quadratic polynomial, the potential at which the desired q occurs can be calculated from

$$E = \{-\pi_1 \pm [\pi_1^2 + 4\,\pi_2\,(q - \pi_0)]^{\frac{1}{2}}\}/2\pi_2 \qquad (21)$$

Since there are two possible solutions for E, a test must be performed to determine which one lies within the potential range of the points fitted. For the concentration being considered and those immediately higher and lower, the γ—E profiles are scanned to find the region which includes the value of

E just determined. For each concentration, the three points (E,γ) closest to the desired E are fitted to the equation $\gamma = \pi_0 + \pi_1 E + \pi_2 E^2$; γ at the desired E is computed by interpolation for each of the three concentrations. The three points (γ, ln c) are used to determine Γ at the central value of c, as previously described.

Once values of Γ have been computed at all concentrations, the adsorption-isotherm tests previously described are made. Figure 9 shows a flow diagram for computing Γ at constant charge.

Charged adsorbate modification

The thermodynamic derivation for the relative surface excess of a charged species yields a relationship identical to eqn. (2), except that (1) the derivative is evaluated at constant potential versus a reference electrode reversible to the counter-ion of the charged adsorbate, or (2) a constant potential reference electrode is used and a theoretically computed correction is applied to reflect potentials referred to an electrode reversible to a solution ion.

If the adsorbate ion is sufficiently strongly adsorbed that it displaces any other adsorbed ions, which would be the usual situation of interest, the terms involving surface excesses will be negligible except that involving the adsorbate-ion of interest. If the theoretical reference electrode chosen is reversible to a supporting electrolyte-ion whose concentration remains constant as that of the adsorbate-ion is varied, the theoretical reference electrode potential is constant. Under these conditions, no potential-scale correction is required.

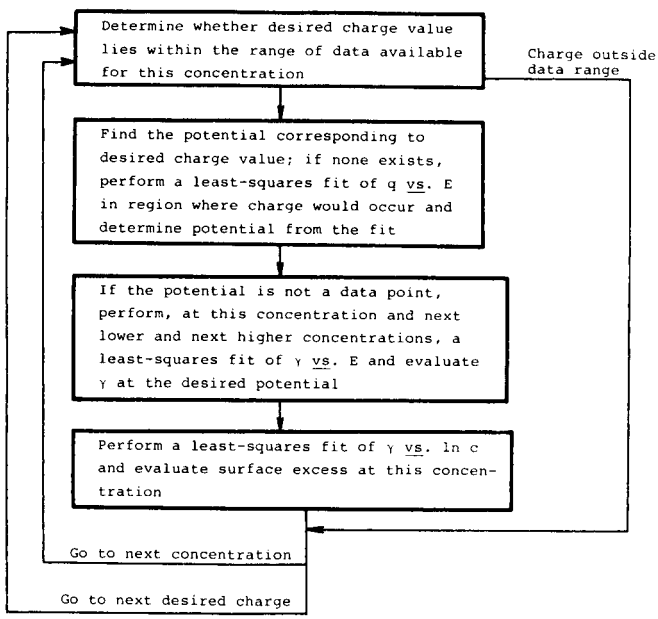


Fig. 9. Evaluation of surface excess at constant electrode charge.

Application

The experimental application of the program is discussed and the two forms of data analysis, i.e., eqns (2) and (3) (version CP-I and eqns. (4) and (8) (version CP-II), are compared in a subsequent paper [18]; the chemical system investigated is adenine in aqueous media at a dropping mercury electrode.

The authors thank the National Science Foundation, which helped support the work described.

REFERENCES

1 D. C. Grahame, Chem. Rev., 41 (1947) 441.
2 R. Parsons, in P. Delahay and C. W. Tobias (Eds.), Advances in Electrochemistry and Electrochemical Engineering, Vol. 1, Interscience, New York, 1961, pp. 1—64.
3 P. Delahay, Double Layer and Electrode Kinetics, Interscience, New York, 1965.
4 D. M. Mohilner, in A. J. Bard (Ed.), Electroanalytical Chemistry, Vol. 1, Marcel Dekker, New York, 1966, pp. 241—409.
5 E. Gileadi, (Ed.), Electrosorption, Plenum Press, New York, 1967.
6 C. A. Barlow, Jr., in H. Eyring, D. Henderson and W. Jost (Eds.), Physical Chemistry — An Advanced Treatise, Vol. IXA, Academic Press, New York, 1970, pp. 167—246.
7 R. Parsons, Pure Appl. Chem., 18 (1968) 91.
8 R. Payne, J. Electroanal. Chem., 41 (1973) 277.
9 H. H. Bauer, P. J. Herman and P. J. Elving, in J. O'M. Bockris and B. E. Conway (Eds.), Modern Aspects of Electrochemsitry, Vol. 7, Plenum Press, New York, 1972, pp. 143—197.
10 A. N. Frumkin and B. B. Damaskin, in J. O'M. Bockris and B. E. Conway (Eds.), Modern Aspects of Electrochemistry, Vol. 3, Butterworths, London, 1964, pp. 149—223.
11 B. B. Damaskin, O. A. Petrii and V. V. Batrakov, Adsorption of Organic Compounds on Electrodes, Plenum Press, New York, 1971.
12 S. Trasatti, J. Electroanal. Chem., 53 (1974) 335.
13 R. S. Perkins and T. N. Anderson, in J. O'M. Bockris and B. E. Conway (Eds.), Modern Aspects of Electrochemistry, Vol. 5, Plenum Press, New York, 1969, pp. 203—290.
14 P. R. Mohilner and D. M. Mohilner, in J. S. Mattson, H. B. Mark, Jr. and H. C. MacDonald, Jr. (Eds.), Electrochemistry, Marcel Dekker, New York, 1972, pp. 3—44.
15 H. C. MacDonald, in J. S. Mattson, H. B. Mark, Jr. and H. C. MacDonald, Jr. (Eds.), Electrochemistry, Marcel Dekker, New York, 1972, pp. 45—59; H. C. MacDonald, Ph.D. Thesis, The University of Michigan, Ann Arbor, 1969.
16 D. C. Grahame, E. M. Coffin, J. I. Cummings and M. A. Poth, J. Am. Chem. Soc., 74 (1952) 1207.
17 D. G. Moursund and C. S. Duris, Elementary Theory and Application of Numerical Analysis, McGraw-Hill, New York, 1967.
18 M. Katz, T. E. Cummings and P. J. Elving, submitted for publication.

Analytica Chimica Acta, 112 (1979) 45—54
Computer Techniques and Optimization

COMPUTER AUTOMATION OF POTENTIOMETRIC ANALYSIS WITH ION-SELECTIVE ELECTRODES

J. SLANINA*, F. BAKKER, J. J. MÖLS, J. E. ORDELMAN and A. G. M. BRUYN-HES
Netherlands Energy Research Foundation ECN, Petten (The Netherlands)
(Received 12th October 1978)

SUMMARY

A PDP-11 computer system, suitable for control of many types of wet chemical analysis, is applied to the automation of ion-selective electrodes. Direct measurements as well as standard addition methods can be used. The electrodes are calibrated automatically, the calibration being repeated until satisfactory results are obtained. Samples and standards are then analysed and the system is recalibrated if the results of a standard are not within preset limits. Up to 300 samples can be analysed in one run. The accuracy is typically better than 5%; 15—25 analyses per hour are possible in routine work.

Recently, a computer system was developed for automation and control of wet chemical techniques [1], such as spectrophotometric measurements [2], titrations and potentiometry with ion-selective electrodes [1], but the system was not used as a real automat. The calibration of the detection system and quality control of the sample analyses were not automated. In this paper, the auxiliary apparatus and programs are described for a system where calibration of the electrodes, checks on that calibration, analysis of samples and quality control of the measurements are performed automatically.

Basically the computer system has remained the same during the 2 years it has been used. A number of input and output functions were added to the system. The most important additions are: (a) a 4-channel, 12-bit, analog/digital converter, as it was found necessary to connect instruments not equipped with a digital voltmeter to the computer; (b) 4 digital/analog converters to obtain analog signals for display on recorders, control of furnaces, etc.; and (c) a pulse counter.

Several commercial sampler systems were tested but they did not meet the demands of reliability and accuracy necessary for this system; accordingly, a pneumatic sample transfer system and a pneumatic sampler diluter were developed.

EXPERIMENTAL

Apparatus

The computer system (Fig. 1) comprises a PDP 11-03 LSI with 24K memory equipped with interfaces for read-out of digital voltmeters, input

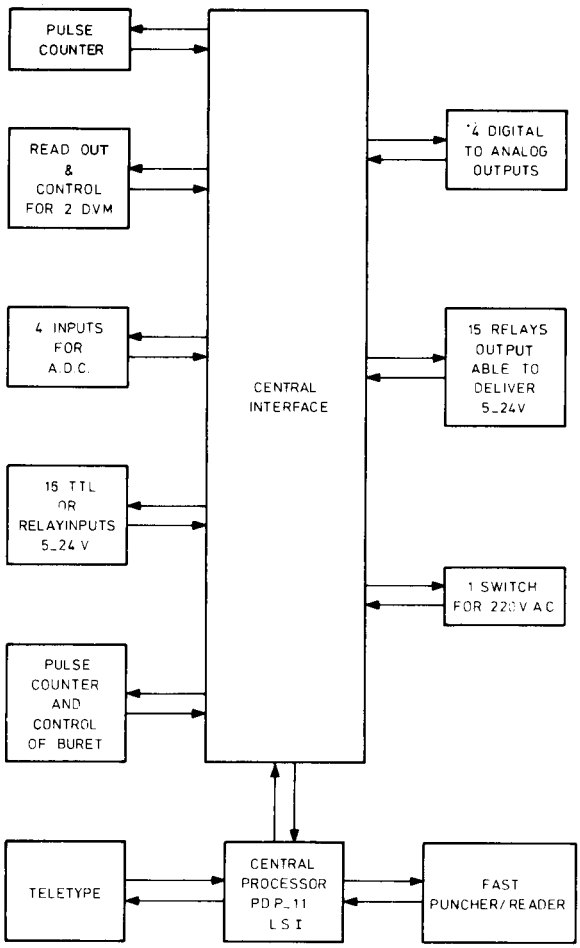
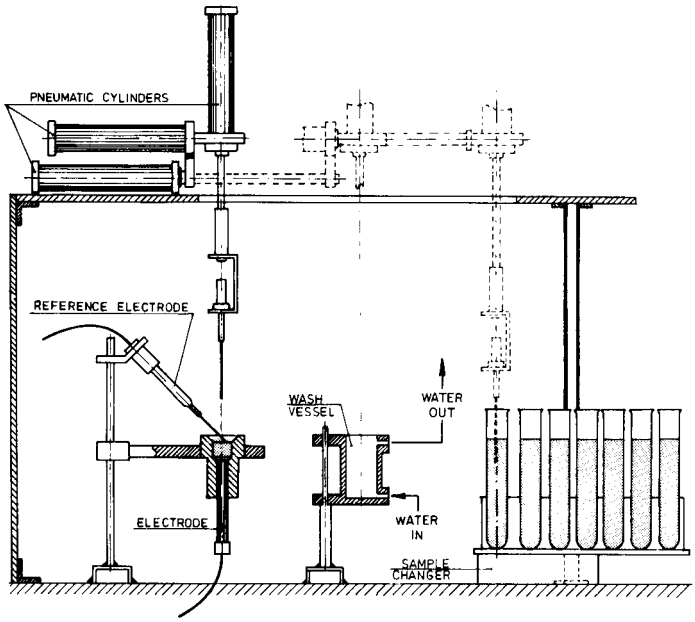


Fig. 1. Computer system.

and output relays for the control sample changer, samplers, diluters, etc., control of burets, pulse counting, an A/D converter and four D/A converters.

The sample changer is a Gilson Sc 6 model with a capacity of 300 sample tubes. The sample transfer system was designed at ECN (Fig. 2). The horizontal and vertical movements are made with pneumatic cylinders (Kunke, B.R.D.). Three positions are available: sample, wash and electrode vessel. The sampler-diluter was also designed at ECN (Fig. 3); it consists of two identical units. The piston of a Metrohm 1-ml buret is connected to a pneumatic cylinder (Kunke, B.R.D.); the buret is equipped with a Flare-Fit connector and a 3-way Flare-Fit pneumatic valve (Durrum Chromatographic, California). The valve and cylinder are under direct computer control. The units can be used as independent samplers or together as a sampler-diluter.

The amplifier was designed at ECN. Two high-impedance ($> 10^{13}\ \Omega$)



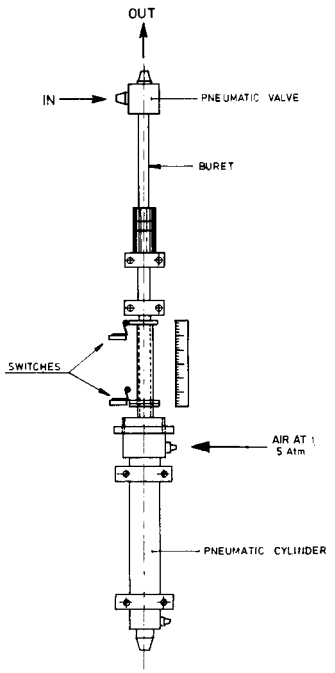


Fig. 2. Sample transfer system.

Fig. 3. Sample diluter.

inputs and an input for a reference electrode are available. The gain is variable between 0.5 and 20; the offset is ±2 V. The computer selects the inputs.

The burets used were of 5-ml and 1-ml volume (Mettler DV-10 and Metrohm 538). For the 0.5-ml buret a Metrohm 457 buret was converted to a motor buret by adding a stepping motor, a stepping motor drive and a Flare-Fit 3-way valve.

The indicating electrodes used were a hydrogen-ion electrode (Metrohm type EA 147), a fluoride electrode (Metrohm type EA 306-F, and Orion type 94-03), a chloride electrode (Ionel type SL-01 Mount Hope, Canada), and an ammonia electrode (EIL type 8002-8 gas sensor). The reference electrode was an Ingold type M 3480 double-junction electrode.

Electrode vessels were designed at ECN with special adaptations for samples of 0.5—2 ml (Figs. 4 and 5).

The proper functioning of the sample transfer unit and the sampler-diluter is checked by the computer by means of microswitches indicating the position of all pneumatic cylinders. Each step of the sample changer activates a micro-switch which is monitored by the computer. The status of the DVM is checked at every read-out. The computer receives information when the DVM does not convert or is out of range, which happens if no solution is

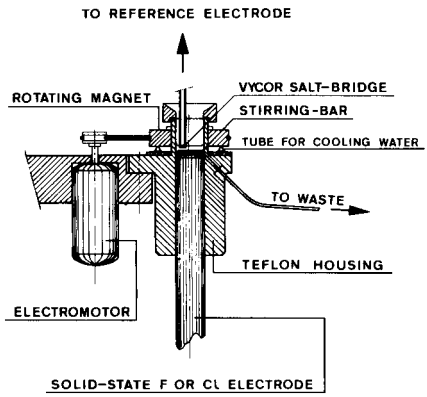


Fig. 4. Fluoride and chloride cell.

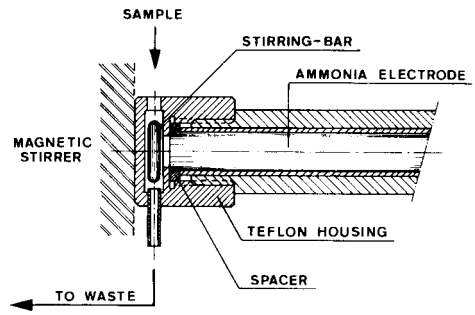


Fig. 5. Ammonia cell in horizontal position.

transported to the electrode vessel so that there is no contact between reference and measurement electrodes.

The status of the burets is monitored before every addition and after each filling. If filling the buret takes too long, the computer gives a warning. Malfunctioning in any of the devices brings the program to a stop.

All the instrumentation is connected via a 220-V relay to the mains so that the computer can switch them off. The general set-up of the system is given in Fig. 6. All details about the devices designed at ECN are available on request from the authors.

Calibration of the electrodes

The flow chart for the calibration procedure is shown in Fig. 7. Parts of the procedure are given in detail in Figs. 8 and 9. The computer starts by asking a number of questions about: (a) the electrode required for the determinations; (b) the mode of measurement, i.e. direct measurement or standard addition or both; (c) the desired range (if the range is near the detection limit, the concentration above which the calibration curve is straight must be specified); (d) the volume of the buret and the concentration of the titrant

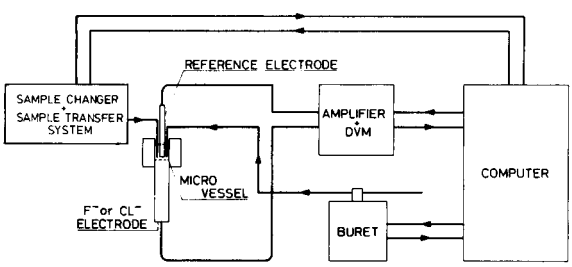


Fig. 6. General set-up of measurement system.

used for standard addition; (e) the volume of sample and reagents; (f) the desired accuracy, etc.

Some electrodes have no selectivity or sensitivity problems in the range of applications required here and these are used in the direct measurement mode, e.g. the gas-sensing ammonia electrode and the glass electrode. These electrodes are calibrated by direct measurement of standards or buffers. The computer asks for 3 standards in the specified range and calculates the E^0 value and slope from the results. The measurements are repeated and if the second calibration is in accordance with the first, the calibration is accepted. If not, the calibration is repeated. A maximum of 4 calibrations are performed. If results are not consistent after 4 calibrations, the computer asks for assistance.

All other measurements are done by the standard addition method, and the electrodes (e.g. Cl^-, Br^-, F^-) are calibrated by making additions of the ion of interest to a sample of double-demineralized water. The computer calculates and performs 20 additions in the range of interest, and the E^0 value and slope of the electrode are calculated from these 20 data points by linear regression.

Calibration by means of multiple standard additions has two advantages. First, a blank can be detected and computed by testing the differences between the observed values and the computed line as a function of discrete values of a blank, based on the points of the straight part of the calibration curve. In practice, if the demineralized water or the reagents are contaminated with the ion being determined, the calibration graph will show curvature until the concentration exceeds 100 times the blank concentration. The computer is given the concentration above which a straight calibration graph is normally obtained in the absence of a blank; it then tests which value of the blank gives a straight calibration graph with the highest correlation coefficient and subtracts this blank from all subsequent results. The second advantage is that the curvature of the calibration graph near the detection limit of the electrode is computed by means of a second-order approximation. Accurate measurements are thus possible at very low concentrations. As described for the direct method, calibrations are repeated and assistance is asked if inconsistent results are obtained after 4 calibrations.

Check of the calibration

As the electrodes are calibrated with one standard solution, or one set of solutions, it is necessary to check the calibration by means of separately prepared standards. The flow chart is given in Fig. 10. After completion of the calibration, the computer asks for the number of standards and their concentration. The standards are analyzed and the results are compared with the given values. The accuracy chosen at the start of the procedure is the value against which the differences are judged. If the difference exceeds the desired accuracy, the standard is analyzed again. The computer asks for help if the second determination is also faulty. Either the calibration procedure is started again or the standards are checked and renewed.

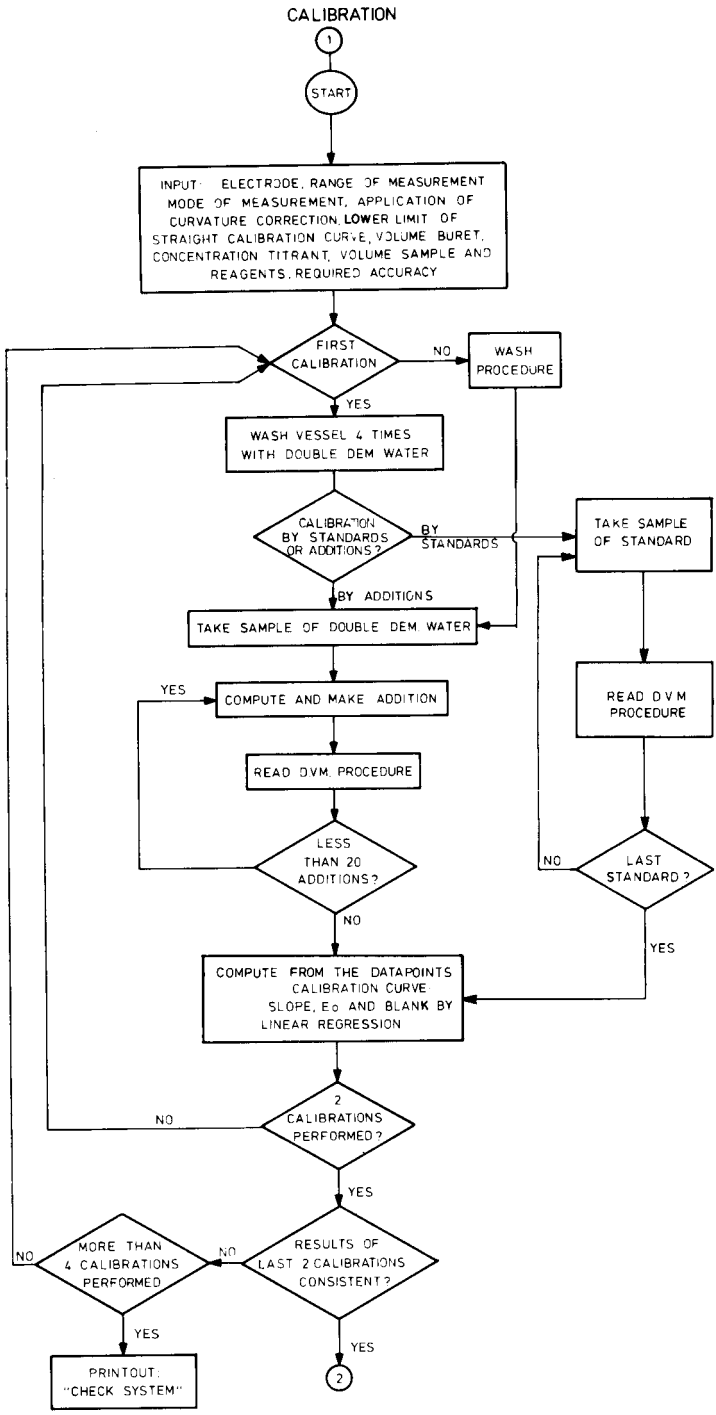


Fig. 7. Flow chart for the calibration procedure.

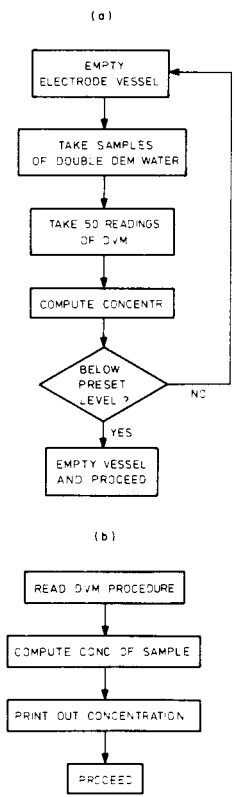


Fig. 8. (a) Wash procedure; (b) direct measurement method.

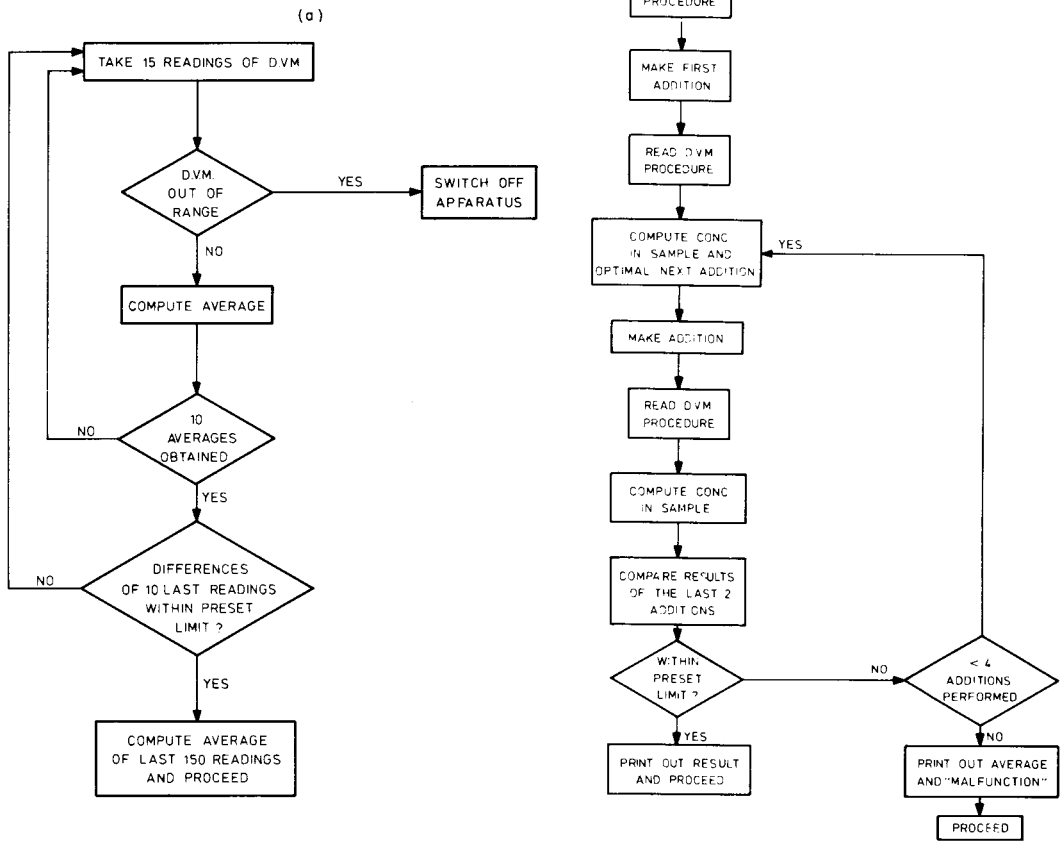


Fig. 9. Procedures for (a) reading the DVM and (b) standard addition.

Analysis of samples

The flow charts are shown in Fig. 11. When the check of the calibration is complete, the computer asks the operator for the number of samples and their identification. The operator is asked to put standards of specified concentrations on each 5th and 10th position of the sample changer. For pH measurements, buffers are put in these positions. The analysis of samples and standards is started. In the standard addition mode, the computer compares the observed concentration with the given value, repeats the measurement if the difference exceeds the desired accuracy, and recalibrates the electrode if the outcome of the second measurement is again unsatisfactory. In the direct measurement mode, the computer recalibrates the electrode by means of the results on the standards. For pH measurements, the standard addition method is used, but the electrode is recalibrated with buffers.

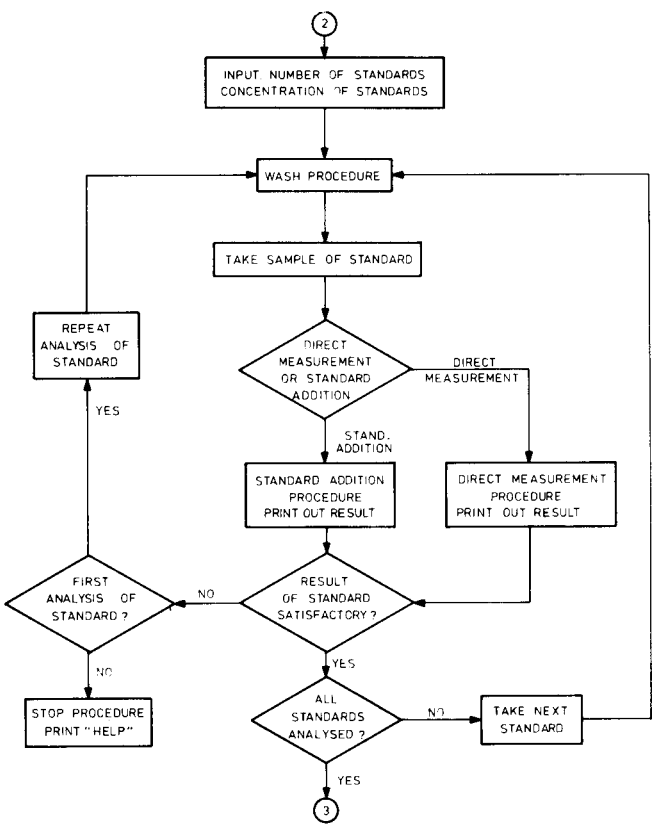


Fig. 10. Calibration check.

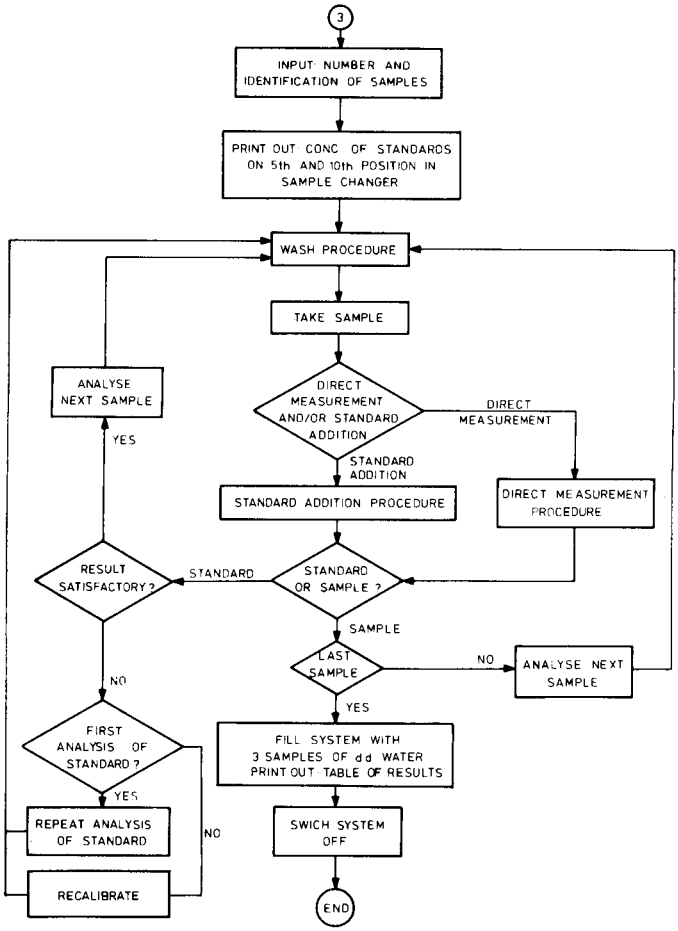


Fig. 11. Analysis of samples and standards.

The measurements are terminated when all samples have been analyzed, or when malfunctioning is detected in the sample transfer, sample changer, sampler-diluter or the DVM, or when the results of the recalibrations are inconsistent.

RESULTS AND CONCLUSION

The accuracy of the system was tested by measurement of standards with the results given in Table 1. Table 2 shows the results of repeated analyses of one rain-water sample.

The system is capable of analyzing long series of samples without human assistance. The automatic quality control and recalibration procedures ensure that the results are reliable. In general, the computer system gives

more accurate results than manual methods, especially at low concentrations where the electrodes are very slow. The system will perform 15 analyses per hour in the standard addition mode, and 25 analyses per hour if direct measurement is used.

TABLE 1

Test of accuracy

Species	Mode of measurement	Number of detns.	Concentration[a]		R.s.d. (%)
			Given	Found	
F^-	Standard addition	30	300	300	2.1
		30	200	201.5	3.0
		30	100	100.7	3.5
		30	50	51.3	4.8
		30	25	26.6	6.0
Cl^-	Standard addition	58	20	20.1	3.4
		58	10	10.1	3.6
		51	5	4.90	4.7
		66	2	1.99	4.1
		71	1	0.98	4.2
NH_4^+	Direct measurement	8	20	19.8	0.7
		12	10	9.95	1.7
		22	5	4.96	1.8
		22	2	2.04	3.7
H^+	Standard addition	12	1	1.03	2.0
		12	0.1	0.095	1.3

[a] All concentrations are given in ppm except for fluoride which is given in ppb.

TABLE 2

Replicate analysis of a rain-water sample

Species	Number of detns.	Average found (ppm)	S.d. (ppm)
F^-	14	0.031	0.0029
Cl^-	16	16.8	0.4
NH_4^+	13	2.20	0.03
H^+	15	0.097	0.002

REFERENCES

1 J. Slanina, F. Bakker, C. Lautenbag, W. A. Lingerak and T. Sier, Mikrochim. Acta, (1978) 519.
2 J. Slanina, F. Bakker, A. G. M. Bruyn-Hess and J. J. Möls, Fresenius Z. Anal. Chem., 289 (1978) 38.

Analytica Chimica Acta, 112 (1979) 55—63
Computer Techniques and Optimization

COMPUTER-ASSISTED MEASUREMENT OF ION-DIFFUSION COEFFICIENTS BY USE OF THE COTTRELL EQUATION

A. YAMADA, Y. KATO, T. YOSHIKUNI, Y. TANAKA and N. TANAKA*
Department of Chemistry, Faculty of Science, Tohoku University, Sendai 980 (Japan)
(Received 18th September 1978)

SUMMARY

A computer-assisted technique based on application of the Cottrell equation is examined for the measurement of ion-diffusion coefficients. It is particularly useful for the investigation of unstable species because of the simple and rapid processing of data. Diffusion coefficients of chromium(III) and cobalt(III) complex ions are given.

The diffusion coefficients of electroactive species in solution give important information in electroanalytical chemistry. The procedures for obtaining diffusion coefficients in solutions are well documented [1—6]. Often, however, the complexity of the measurement system precludes easy evaluation, so that measurements become time-consuming. A simple and rapid method for the determination of diffusion coefficients can be based on the Cottrell diffusion cell [7]. This fundamental technique has been used in this work with computer-aided data acquisition, and applied to the determination of diffusion coefficients of unstable cobalt(III) complexes.

FUNDAMENTAL EQUATION FOR DETERMINATION OF DIFFUSION COEFFICIENTS

The diffusion-controlled current—time and coulomb—time curves at constant potential are represented by the Cottrell equation

$$I_d = nFA\, D^{\frac{1}{2}} c\, (\pi t)^{-\frac{1}{2}} \quad (1)$$

and its integrated form

$$Q_d = 2nFA\, D^{\frac{1}{2}} c\, t^{\frac{1}{2}}\, \pi^{-\frac{1}{2}} \quad (2)$$

respectively, where I_d is the diffusion-controlled current; F, the Faraday constant; n, the number of electrons involved; A, the surface area of the electrode; t, the time; Q_d, the amount of electricity consumed for electrolysis; and D and c are the diffusion coefficient and the concentration of the electroactive species, respectively.

If the current is plotted against the reciprocal of the square root of time, or the number of coulombs against the square root of time, a straight line should be obtained. A diffusion coefficient can then be computed from the

slope obtained. However, the Cottrell equation is derived under the following conditions: (1) a simple process of mass transport prevails; (2) the solution is not stirred; (3) a high concentration of supporting electrolyte is present in solution, so that migration effects can be neglected; (4) conditions of semi-infinite linear diffusion are achieved. Diffusion coefficients can be obtained for both inorganic and organic substances from measurements over a short range of time (less than a few tenths of a second under normal conditions), provided that the current is controlled by diffusion of the substance to be determined.

For evaluation of the diffusion coefficient from these relationships, two methods are considered. One is an absolute method and the other is a comparative method which requires a reference substance.

The absolute method

This method is based on the direct application of eqn. (1) or (2) to the determination of the diffusion coefficient. To calculate the diffusion coefficient D, it is necessary to know accurately the area of the electrode surface. The diffusion coefficient obtained by this method may be used in the comparative method as a standard.

The comparative method

This method involves measuring the current or the amount of electricity for a reference substance at the same time as the measurements for the sample and using the ratio of these measured quantities of the reference substance and the sample. Diffusion coefficients are then determined from the relation

$$D^{\frac{1}{2}} = (\alpha/\alpha_r)\,(c_r/c)\,D_r^{\frac{1}{2}} \qquad (3)$$

where α represents I_d or Q_d itself, or the slope of the plot of I_d vs. $t^{-\frac{1}{2}}$ or Q_d vs. $t^{\frac{1}{2}}$. Subscript r means a reference substance.

SYSTEM DESCRIPTION

The electrochemical system used in this work was essentially the same as that described previously [8], except for some minor modifications. The computer-interfaceable potentiostat and the integrator were specially designed for the present study.

The computer-assisted electrochemical system for the determination of diffusion coefficients was developed on the Real-Time Data Analysis System (RETDAS) [9], which is built around a CUP TACC 1200M (NOVA 01) with a 24 K-word core memory. A schematic diagram of the RETDAS system is shown in Fig. 1. The data acquisition system was named ELZA, the abbreviation of *E*lectrochemical Data *L*ogging Program System with *Z*ero-level Current *A*uto-correction Function. The ELZA system (Fig. 2) is composed of two subsets ELZAL and ELZAP with two support programs. The subset ELZAL was developed for real-time control and data acquisition, and is composed of a group of modules.

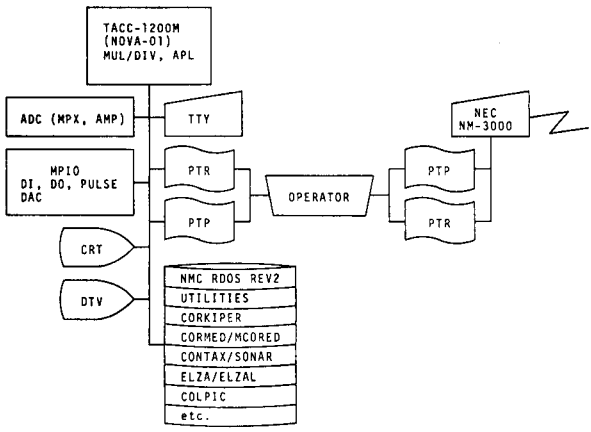


Fig. 1. Block diagram of Real Time Data Analysis System (RETDAS).

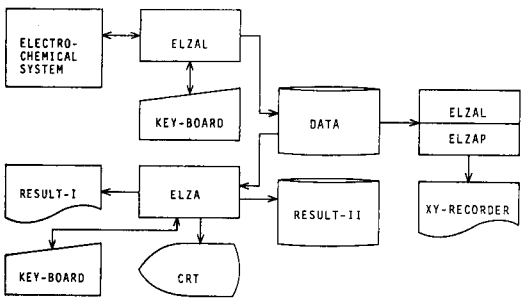


Fig. 2. Configuration of the system ELZA.

EXPERIMENTAL

Reagents

Cadmium(II) nitrate was prepared by dissolving a known amount of the metal (99.999%) in reagent-grade nitric acid. Reagent-grade lead(II) nitrate was recrystallized twice from redistilled water containing a small amount of nitric acid and dried in vacuum at 50°C. Cobalt(III) and chromium(III) complexes were prepared by methods given in the literature; their purity was confirmed by elemental analysis. All other chemicals used were of analytical-reagent grade. A small amount of gelatin or polyoxyethylene lauryl ether (LEO; mean molecular weight, 862) was used as a maximum suppressor in the potentiostatic measurement. Redistilled water was used throughout.

Electrode system

A computer-interfaceable potentiostat with a current follower was used. The output signal of the current follower was interfaced to the RETDAS. When the current was integrated, the integrator was connected to the output of the current follower.

The electrolytic cell system has already been described [8]. A slowly

dropping mercury electrode with a drop time of 60 s (mercury flow rate, 0.213 mg s^{-1}) and a normal dropping mercury electrode with a drop time of 4 s (mercury flow rate, 1.16 mg s^{-1}) were used.

Dissolved oxygen was removed by bubbling purified nitrogen gas through the solution. All measurements were made at 25 ± 0.1 °C.

Procedures

Parameters which are required for the experiment to proceed are input through the communication module in a question/answer mode. An example of communication between the operator and the subset ELZAL is given in Table 1. There are two types of control mode, manual and automatic. Under manual control, measurements are made stepwise by the operator; under automatic control, data acquisition is repeated according to the input repeat number

Data acquired in core memory by the ELZAL are displayed on the graphic display, and transferred to an appropriate disk file as directed by the operator. The disk file to which such data are transferred is designed to hold all information on the experimental data and to initiate the further processing of those data.

TABLE 1

An example of the application of subset ELZAL

ELZAL $TTI/I $TTO/O	
DATA FILE NAME ?	D110102.DT
A/D CHANNEL ?	0
GAIN OF AMP ?	1.988E4
TRIG. MODE (+, —,E)	—
TRIG. LEVEL (MICRO AMP.)	—100
FREQUENCY CODE ?	11
ELECTRODE SYSTEM ?	CR(III)CYDTA / CR(II)CYDTA (HG)
NUMBER OF ELECTRON ?	1
CONCENTRATION (MOL/L) ?	1.000E-3
MEDIUM ?	NACL 1 M, AC-BUF(PH5) 0.1M, LEO 2MMM
TEMPERATURE (DEG.C.) ?	25
REFERENCE ELECTRODE ?	SCE
INITIAL POTENTIAL (V) ?	—0.98
STEP POTENTIAL (V) ?	—1.40
REPEAT NUMBER ?	2
AUTO CONTROL ?	NO
DELAY TIME (SEC) ?	40
IS ELECTRODE SDME ?	YES
FLOW RATE (G/SEC) ?	2.2887E-4
E(STEP) = —1.40	
*** PAUS ***	
*** PAUS ***	
LOGGING OK, STORE ? YES	
STORE OK, CONTINUE ? NO	

In a series of determinations of the diffusion coefficients of various cobalt(III) complexes, the electrochemical procedure was as follows. After measurements of the residual current of 40 ml of supporting electrolyte solution, 200 μl of a solution of the reference substance (200 mM $[Co(NH_3)_6]Cl_3$) was injected. As soon as the current from the reference substance had been measured, a solid sample of the cobalt complex under test was added to the solution. The total current was then measured. The difference between the total current and the current from the reference substance was used to obtain α in eqn. (3), and the current from the reference substance after correction for the residual current was used to obtain α_r.

RESULTS AND DISCUSSION

Diffusion coefficients of several species were determined as a test of the proposed potentiostatic method. Figure 3(a) shows an example of the original current—time curve obtained with 1 mM cadmium(II) in 0.9 M $NaClO_4$ at 25°C. Figure 3(b,c) shows examples of the relationship between current and the reciprocal of square root of time and between the logarithm

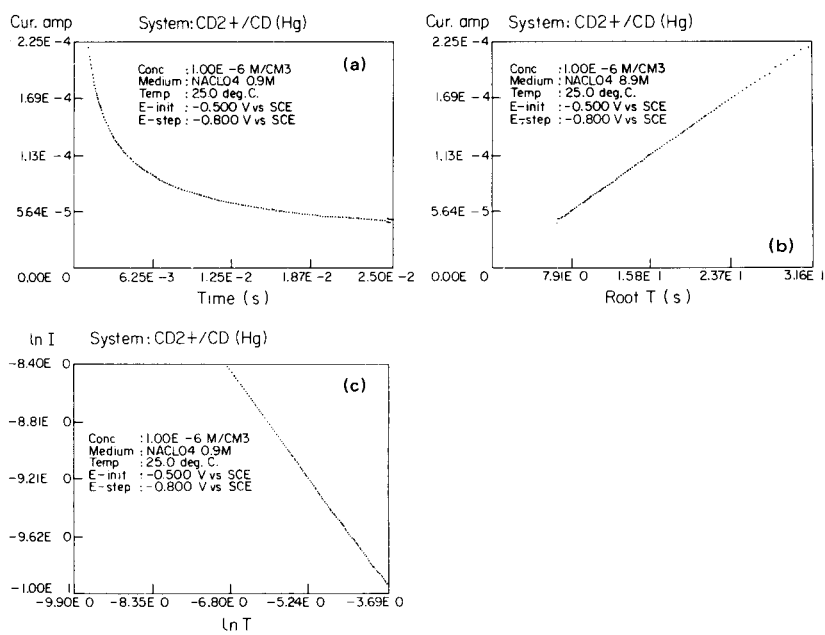


Fig. 3. Examples of data processing for 1 mM cadmium(II) obtained in 0.9 M $NaClO_4$ at 25°C: (a) I_d vs. t; (b) I_d vs. $t^{\frac{1}{2}}$; (c) ln I_d vs. ln t. The initial potential was —0.500 V vs. SCE and the stepwise potential increase was —0.800 V vs. SCE.

of the current and the logarithm of time, respectively. The linearities are clearly satisfactory; the slope of Fig. 3(c) was found to be −0.5 within experimental error. These results indicate that the correction for the residual current is satisfactory and that the Cottrell equation is applicable. Diffusion coefficients were calculated from eqn. (1). The surface area of the electrode was carefully calculated [10] from the flow rate of mercury and the time. More than 10 measurements were averaged automatically.

The results of these determinations are summarized in Table 2. Literature values, where available, are also given. The addition of a little gelatin or polyoxyethylene lauryl ether (see Experimental) did not affect the cathodic limiting current in the potentiostatic measurement. The presence of large amounts of surface-active substances tends to distort polarographic waves [11, 12]. The concentrations used here were carefully selected and are considered to have no effect on the diffusion coefficients. There are no significant differences between the values obtained in this work and those

TABLE 2

Diffusion coefficients (D) determined by the Cottrell equation at 25°C

Ions	Solution[a]	D ($\times 10^{-6}$ cm^2 s^{-1})		
		This study	Other data	
Cd^{2+}	0.5M $NaClO_4$ + 1mM $HClO_4$	6.97 ± 0.03[b]	6.82	[3]
	1.0M $NaClO_4$ + 1mM $HClO_4$	6.82 ± 0.02	6.74	[3]
	1.0M $NaNO_3$	6.87 ± 0.80	6.80	[3]
Pb^{2+}	0.1M $NaNO_3$[c]	10.6 ± 0.05	9.54 ± 0.05	[5]
	0.1M NaOAc[c,d]	8.61 ± 0.04		
$Co(NH_3)_6^{3+}$	0.1M NaOAc[c,d]	8.06 ± 0.04	8.08 ± 0.03	[6]
	1.0M NaCl + 0.1M NaOAc	8.71 ± 0.05		
$Cr(DCTA)^-$ [e]	0.4M NaCl + 0.04M NaOAc	5.55 ± 0.06	5.5	[14]
	1.0M NaCl + 0.1M NaOAc	5.05 ± 0.05		
	2.0M NaCl + 0.2M NaOAc	4.24 ± 0.04		
	0.4M $NaClO_4$ + 0.04M NaOAc	5.45 ± 0.03		
	1.0M $NaClO_4$ + 0.1M NaOAc	4.89 ± 0.04		
	2.0M $NaClO_4$ + 0.2M NaOAc	4.19 ± 0.04		
$Cr(en)_3^{3+}$ [f]	0.4M NaCl + 0.04M NaOAc	6.55 ± 0.04		
	1.0M NaCl + 0.1M NaOAc	6.17 ± 0.09	6.15	[15]
	2.0M NaCl + 0.2M NaOAc	5.65 ± 0.08		
	0.4M $NaClO_4$ + 0.04M NaOAc	6.36 ± 0.04		
	1.0M $NaClO_4$ + 0.1M NaOAc	6.12 ± 0.07		
	2.0M $NaClO_4$ + 0.2M NaOAc	5.45 ± 0.03		

[a]Unless otherwise stated, all solutions contained 2 μM LEO. [b]Standard deviation. [c]0.005% gelatin was used instead of 2 μM LEO. [d]Sodium acetate buffer pH 5. [e]Trans-1,2-diaminocyclohexanetetraacetate. [f]Ethylenediamine.

obtained by other electrochemical methods. In some cases, the precision reported for diffusion coefficients determined by various methods has ranged from 5—20% [2].

As the diffusion coefficient for hexa aminecobalt(III) ions agreed with that determined precisely by assuming spherical diffusion [6], this value was used as the reference in a series of experiments with cobalt(III) complexes. Table 3 summarizes the diffusion coefficients found for various cobalt(III) complexes determined by application of eqn. (3).

According to the Stokes—Einstein equation $D = kT/6\pi r\eta$ (where k is the Boltzman constant, η the viscosity, T the temperature, and r the ionic radius), a diffusion coefficient increases with a decrease in ion size. Since the size of the complex depends on that of ligand, the magnitude of the diffusion coefficient reflects the bulkiness of the ligand. The results in Table 3 show an interesting tendency of the relationship between the diffusion coefficients and structural characteristics of the ligand.

When the ligands are arranged in order of increasing size (Br > Cl > F or tn > pn > en), it can be seen that the diffusion coefficients of the complexes

TABLE 3

Diffusion coefficients (D) of cobalt(III) complexes obtained in 0.1 M sodium acetate buffer (pH 5.0) containing 0.005% gelatin at 25°C

Entry	Complex[a]	D ($\times 10^{-6}$ cm^2 s^{-1})	
1	*Trans*-$[Co(NO_2)_2(en)_2]NO_3$	[16]	8.00
2	*Trans*-$[CoCl(NO_2)(en)_2]NO_3$	[17]	8.10
3	*Trans*-$[CoF_2(en)_2]NO_3$	[18]	8.64
4	*Trans*-$[CoCl_2(en)_2]NO_3$	[19]	8.56
5	Cis-$[CoCl_2(en)_2]NO_3$	[20]	8.13
6	*Trans*-$[CoBr_2(en)_2]NO_3$	[21]	8.32
7	$[Co(CO_3)(en)_2]ClO_4$	[22]	7.85
8	$[Co(CO_3)(pn)_2]NO_3$	[23]	6.21
9	$[Co(CO_3)(tn)_2]ClO_4$	[24]	4.76
10	$K[Co(CO_3)_2(en)]H_2O$	[25]	6.74
11	$[Co(ox)(en)_2]NO_3$	[26]	7.20
12	$[Co(en)_3]Cl_3 \cdot 3H_2O$	[27]	6.67
13	$K_3[Co(ox)_3] \cdot 2.5H_2O$	[28]	6.38
14	*Cis*-$[Co(H_2O)_2(NH_3)_4](ClO_4)_3$	[29]	7.11
15	$[Co(CO_3)(NH_3)_4]NO_3 \cdot 0.5H_2O$	[30]	9.04
16	$[CoCl(NH_3)_5](NO_3)_2$	[31]	9.13
17	$[CoBr(NH_3)_5](NO_3)_2$	[32]	9.11
18	$[Co(NO_3)(NH_3)_5](NO_3)_2$	[33]	8.46
19	$[Co(CO_3)(NH_3)_5]NO_3 \cdot 0.5H_2O$	[34]	7.09
20	$[Co(H_2O)(NH_3)_5](ClO_4)_3$	[35]	7.76
21	$NH_4[Co(ox)_2(NH_3)_2]H_2O$	[36]	7.58
22	$[Co(NH_3)_6]Cl_3$	[37]	8.06

[a]en, ethylenediamine; pn, propylenediamine; tn, trimethylenediamine; ox, oxalate.

increase in the order, trans-$[CoBr_2(en)_2]^+$ < trans-$[CoCl_2(en)_2]^+$ < trans-$[CoF_2(en)_2]^+$, or $[Co(CO_3)(tn)_2]^+$ < $[Co(CO_3)(pn)_2]^+$ < $[Co(CO_3)(en)_2]^+$. A similar correlation can also be seen between the diffusion coefficients and the size of the ligand with other complexes, the order of the diffusion coefficient of different pairs of complexes being trans-$[CoCl(NO_2)(en)_2]^+$ > trans-$[Co(NO_2)_2(en)_2]^+$; $[Co(CO_3)(NH_3)_4]^+$ > $[Co(CO_3)(en)_2]^+$; $[Co(ox)_2(NH_3)_2]^-$ > $[Co(ox)_3]^{3-}$.

The diffusion coefficient may also be calculated from the Nernst—Einstein equation, $D = RT\Lambda/zF^2$ (where Λ is the equivalent conductivity and other terms have their conventional meaning) if conductivity data for the individual ions are available. If Λ is known for two ions, an approximate ratio of the diffusion coefficients may be obtained by taking account of the valencies. This was done, by using data for the equivalent conductivities of several cobalt complexes at infinite dilution [13]. Calculated values of the ratio of the diffusion coefficients are 1.05 for Pb^{2+}: $[Co(NH_3)_6]^{3+}$, 0.75 for $[Co(en)_3]^{3+}$: $[Co(NH_3)_6]^{3+}$, 1.06 for $[CoCl(NH_3)_5]^{2+}$: $[Co(NH_3)_6]^{3+}$, and 1.06 for $[Co(H_2O)(NH_3)_5]^{3+}$: $[Co(NH_3)_6]^{3+}$, while the values observed were 1.07, 0.83, 1.13 and 0.96, respectively. The calculated values approximately agree with the experimental ones.

In summary, computer-assisted measurement with application of the Cottrell equation is a simple and rapid technique for the determination of diffusion coefficients within reasonable limits of reproducibility. This technique may be particularly useful for the investigation of unstable species.

The authors thank the Ministry of Education for financial support.

REFERENCES

1 R. H. Stokes, L. A. Wools and R. Mills, J. Phys. Chem., 61 (1957) 1634.
2 R. N. Adams, Electrochemistry at Solid Electrodes, Dekker, New York, 1969, p. 214 et seq.
3 J. A. Bolzan, J. Electroanal. Chem., 59 (1975) 303.
4 A. C. Quano and J. A. Carothers, J. Phys. Chem., 79 (1975) 1314.
5 N. C. Fawcett and R. D. Caton, Jr., Anal. Chem., 48 (1976) 600.
6 H. Ikeuchi, Y. Fujita, K. Iwai and G. P. Sato, Bull. Chem. Soc. Jpn., 49 (1976) 1883.
7 F. G. Cottrell, Z. Phys. Chem., 42 (1902) 385.
8 A. Yamada and N. Tanaka, Sci. Repts. Tohoku Univ. Ser. I, 52 (1969) 73.
9 A. Yamada, M. Sakata, Y. Kato and N. Tanaka, Sci. Repts. Tohoku Univ. Ser. I, 58 (1975) 94.
10 T. E. Cummings and P. J. Elving, Anal. Chem., 50 (1978) 480.
11 R. Tamamushi, S. Yamamoto, A. Takahashi and N. Tanaka, Anal. Chim. Acta, 20 (1959) 486.
12 N. Tanaka, R. Tamamushi and A. Takahashi, Collect. Czech. Chem. Commun. 25 (1960) 3016.
13 Chemical Society of Japan, Handbook in Chemistry, Maruzen, Tokyo, 1969.
14 N. Tanaka and A. Yamada, Bull. Chem. Soc. Jpn., 39 (1966) 920.
15 N. Tanaka, T. Tomita and A. Yamada, Bull. Chem. Soc. Jpn., 45 (1972) 940.
16 A. Werner, Ber., 44 (1911) 2450.
17 A. Werner, Justus Liebigs Ann. Chem., 386 (1912) 251.

18 W. R. Matoush and F. Basolo, J. Am. Chem. Soc., 78 (1956) 3972.
19 A. Wold and J. K. Ruff (Eds.), Inorganic Syntheses, Vol. 14, McGraw-Hill, New York, 1973, p. 68.
20 A. Wold and J. K. Ruff (Eds.), Inorganic Syntheses, Vol. 14, McGraw-Hill, New York, 1973, p. 70.
21 A. Werner, L. Gerb, S. Lorie and J. Rapaport, Ann., 386 (1912) 111.
22 P. Pfeiffer, S. Golther and O. Angern, Ber., 60 (1927) 308.
23 F. P. Dwyer and T. E. MacDermott, Inorg. Chem., 2 (1963) 871.
24 Y. Kojima, Bull. Chem. Soc. Jpn., 48 (1975) 2033.
25 M. Shibata, Nippon Kagaku Zasshi, 87 (1966) 771.
26 W. Schramm, Z. Anorg. Chem., 180 (1929) 167.
27 S. M. Jørgensen, J. Prakt. Chem., 39 (1889) 8.
28 S. P. L. Sorensen, Z. Anorg. Chem., 11 (1896) 1.
29 R. G. Jørgensen, Z. Anorg. Chem., 2 (1892) 294.
30 S. M. Jørgensen, Z. Anorg. Chem., 2 (1892) 279.
31 S. Y. Tyree, Jr., (Ed.), Inorganic Syntheses, Vol. 9, McGraw-Hill, New York, 1967, p. 160.
32 H. S. Booth, (Ed.), Inorganic Syntheses, Vol. 1, McGraw-Hill, New York, 1939, p. 186.
33 S. M. Jørgensen, Z. Anorg. Chem., 17 (1898) 463.
34 J. C. Bailar, Jr., (Ed.), Inorganic Syntheses, Vol. 4, McGraw-Hill, New York, 1953, p. 171.
35 A. Benrath and A. Miens, Z. Anorg. Chem., 177 (1929) 289.
36 E. H. Riesenfeld and R. Klement, Z. Anorg. Chem., 124 (1922) 1.
37 W. C. Fernelius, (Ed.), Inorganic Syntheses, Vol. 2, McGraw-Hill, New York, 1946, p. 216.

Analytica Chimica Acta, 112 (1979) 65—73
Computer Techniques and Optimization

THE LEARNING MACHINE IN QUANTITATIVE CHEMICAL ANALYSIS
Part 2. Potentiometric Titrations of Mixtures of three Bases

M. BOS

Department of Chemical Technology, Twente University of Technology, Enschede (The Netherlands)

(Received 28th June 1978)

SUMMARY

Computer-calculated curves for the titration of mixtures of one strong base and two weak bases are used in the training and testing of a linear learning machine. The results indicate that multicategory classifiers can be calculated from a computer-generated training set of titration curves in which a random error of ±0.01 unit in the pH values is introduced. The relative error in the predictions for concentrations of bases not included in the training set was of the order of ± 1% for concentration ratios up to 10:1 when ΔpK_b for the weak bases exceeded 1 pK unit and for $K_{b_1} \leq 5 \times 10^{-4}$ and $K_{b_2} \geq 10^{-9}$. Calculation of the first derivative of the volume of titrant versus pH curves as a preprocessing step was necessary to obtain this accuracy for the weak bases, whereas the volume of titrant versus pH curves had to be used directly in the determination of the strong base. Predictions of concentrations of actual samples were in agreement with the computer-calculated results.

Multicomponent titrations are of great interest, especially for the accurate determination of mixtures of acids and bases. The information wanted can be found directly from the titration curves in favourable cases only. In many unfavourable cases, multiparametric curve fitting can be applied to evaluate the titration curves [1–3]. This method is very suitable for separate samples of different kinds, and requires only the titration of the sample itself followed by the calculations. The time needed for these calculations cannot be neglected compared to the time required for the titration, especially when the sample contains two or more components with unknown dissociation constants.

It has been shown that learning-machine methods can be applied successfully to quantitative analytical problems [4, 5]. In comparison to curve-fitting methods, evaluation of results is very fast. The major disadvantage of the learning-machine method is the large investment in preparation of standards and computer time for training. Analytical methods designed for process control generally deal with large numbers of samples of one kind. For these methods, evaluation of the results by means of the learning machine pays, and may perhaps be the only solution in situations where the total time of analysis must be minimized.

In this paper, the limits of the learning-machine method applied to the

titration of a mixture of one strong base and two weak bases are investigated with the use of theoretically calculated potentiometric titration curves.

THEORY

Multicategory classifiers w that can be used to calculate quantitative results can be obtained by solving the equation $Aw = b$, where A is the two-dimensional data matrix of which the rows are constituted by the patterns of the training set, and b is a column vector, its elements being the quantitative property of interest. The mathematics needed to solve this equation by using singular value decomposition (SVD) has been described [5].

For evaluation of potentiometric acid—base titrations, the input patterns for the learning machine can be the titration curves presented as the volumes of titrant added to the sample at a set of fixed preselected pH values. The titration curve for a mixture of the strong base sodium hydroxide and two weak bases B1 and B2 titrated with a strong acid like hydrochloric acid can be described by

$$V = \left[\frac{a_{H^+}\,VNUL}{f_{H^+}} - \frac{K_w\,VNUL}{a_{H^+}\,f_{OH^-}} + CNAOH\;VNUL + \frac{CB1\;VNUL\,K_1\,a_{H^+}}{(K_1\,a_{H^+} + f_{B1H^+}K_w)} + \frac{CB2\;VNUL\,K_2\,a_{H^+}}{(K_2\,a_{H^+} + f_{B2H^+}K_w)}\right] \Big/ \left[\frac{K_w}{a_{H^+}\,f_{OH^-}} - \frac{a_{H^+}}{f_{H^+}} + T\right] \quad (1)$$

The symbols used are defined in Table 1.

In practice, titration data will not exactly fit the equation $Aw = b$ but will contain a certain amount of noise. More realistic titration curves can therefore be obtained by adding artificial noise to the data obtained from eqn. (1). For low values of $CNAOH$, negative values of V may be obtained in calculations with eqn. (1) at high pH values. In practice, this means that extra NaOH must be added to the sample to obtain the starting value of the fixed pH-value set for which the V values of the patterns are calculated. For the learning machine input, these negative values are used as easily as the positive ones.

TABLE 1

Glossary of symbols

a	Activity
f	Activity coefficient
$VNUL$	Volume at start of titration
V	Volume of titrant added
K_w	Dissociation constant of water
K_1	Dissociation constant of base B1
K_2	Dissociation constant of base B2
$CB1$	Stoichiometric concentration of base B1 at start of titration
$CB2$	Stoichiometric concentration of base B2 at start of titration
$CNAOH$	Stoichiometric concentration of NaOH at start of titration
T	Molarity of titrant used

Preprocessing of the input data for a learning machine often improves the results. One form of preprocessing is the use of the first derivative of the data. The derivative $dV/d(pa_{H^+})$ is given by

$$\frac{dV}{d(pa_{H^+})} = \frac{2.30\, a_{H^+}}{\{K_w/(a_{H^+} f_{OH^-}) - a_{H^+}/f_{H^+} + T\}^2}\left[\left\{\frac{a_{H^+}\, VNUL}{f_{H^+}} - \frac{K_w\, VNUL}{a_{H^+} f_{OH^-}} + \frac{CB1\, VNUL\, K_1\, a_{H^+}}{K_1\, a_{H^+} + f_{B1H^+} K_w} + \frac{CB2\, VNUL\, K_2\, a_{H^+}}{K_2\, a_{H^+} + f_{B2H^+} K_w} + CNAOH\, VNUL\right\}\left\{-\frac{1}{f_{H^+}} - \frac{K_w f_{OH^-}}{(a_{H^+} f_{OH^-})^2}\right\} - \left\{\frac{K_w}{a_{H^+} f_{OH^-}} - \frac{a_{H^+}}{f_{H^+}} + T\right\}\left\{\frac{VNUL}{f_{H^+}} + \frac{K_w\, VNUL}{f_{OH^-}\,(a_{H^+})^2} + \frac{CB1\, VNUL\, K_1\, f_{B1H^+} K_w}{(K_1\, a_{H^+} + f_{B1H^+} K_w)^2} + \frac{CB2\, VNUL\, K_2\, f_{B2H^+} K_w}{(K_2\, a_{H^+} + f_{B2H^+} K_w)^2}\right\}\right] \quad (2)$$

With the use of this preprocessing algorithm the input patterns for the learning machine become the calculated $dV/d(pa_{H^+})$ values at the fixed pH values.

EXPERIMENTAL

Chemicals

All chemicals were of analytical grade. Potassium chloride, sodium acetate, pyridine (all Merck) and tris(hydroxymethyl)-aminomethane (Tris; Fluka) were used as received. Hydrochloric acid, acetic acid and sodium hydroxide solutions were prepared in 1 M potassium chloride from Merck Titrisol ampoules by adding the calculated amount of potassium chloride and diluting to the specified volume with carbon dioxide-free double-distilled water. The titers of the Tris, sodium acetate and pyridine solutions were calculated from the weight of compound added to a 1 M potassium chloride solution.

Equipment and procedures

All titrations were done in 1 M potassium chloride solutions. The computerized titration system and the calibration procedure for the glass-calomel electrode set were the same as described earlier [3]. The software for the control of the titration was changed to measure the volume of titrant used at a set of pH values at fixed intervals.

RESULTS

Computer-calculated titrations

The patterns V were calculated by means of eqn. (1) for 49 points, $V(1)$ being the volume of titrant needed to obtain pH 12, $V(2)$ for pH 11.8, $V(3)$ for pH 11.6, etc. to $V(49)$ for pH 2.4. $V(50)$ was taken as equal to 100.00. Activity coefficients in eqn. (1) were taken as 1.0, the value 10^{-14} was used for K_w, and the starting volume $VNUL$ was taken as 50.00. A titer of 0.100 was used in all calculations.

As a training set, 200 titration curves were used. In this set there were 50 curves for each of the single components, at concentrations of 0.0, 1.0×10^{-4}, 2.0×10^{-4}, 3.0×10^{-4}, ... , 4.9×10^{-3} M. The last 50 members of the training set were the titration curves for mixtures of the three components: strong base, weak base 1 and weak base 2 in equal concentrations of 0.0, 1.0×10^{-4}, 2.0×10^{-4}, ... , 4.9×10^{-3} M.

Multicategory classifiers for each of the three components of the sample were calculated by using 5 non-zero singular values [5]. The concentration values multiplied by 1000.0 were taken as the column vector **b**.

The predictive ability of the classifiers was tested by using a test set of titration curves calculated in the same way as the training set but with different values for the concentrations and concentration ratios. Concentrations outside the training range were included in the test set. Progressively more difficult cases were constructed by decreasing the difference in pK_b of the weak bases and shifting the pK_b values either to the very low end of the scale $pK_{b1} = 0$ or to the high end $pK_{b2} = 12.7$. For the sets of K_b values $10^{-4}/10^{-8}$, $2 \times 10^{-4}/10^{-4}$, $10^{-12}/10^{-13}$, $10^{-1}/10^{-2}$, the results predicted with the classifiers were equal to the concentrations used to calculate the titration curves within the precision of the computer presentation of the figures (5 digits). Deviations between predicted and actual concentrations did occur for the K_b combinations 1/0.1, $1.1 \times 10^{-4}/10^{-4}$, $2.0 \times 10^{-13}/10^{-13}$ (Table 2). It should, however, be noted that even in most of these cases the deviations were very small. Another point worth noting is that predictions well outside the concentration range of the training set are still quite good.

In practice, however, the measurements will not be as accurate as the computer representation of the data obtained with eqn. (2). The influence of the accuracy of the data on the results was studied by introducing random but equally distributed deviations in the pH values used to calculate $V(1)$ up to

TABLE 2

Prediction efficiency of weight vectors trained and tested on noise-free titration curves

Sample composition ($\times 10^{-3}$ M)			Prediction $K_{b1} = 1$, $K_{b2} = 0.1$ Relative error (%)			Prediction $K_{b1} = 1.1 \times 10^{-4}$, $K_{b2} = 1.0 \times 10^{-4}$ Relative error (%)			Prediction $K_{b1} = 2 \times 10$, $K_{b2} = 1 \times 10$ Relative error (%)		
*C*NAOH	*C*B1	*C*B2	*C*NAOH	*C*B1	*C*B2	*C*NAOH	*C*B1	*C*B2	*C*NAOH	*C*B1	*C*
10.0000	10.0000	10.0000	−0.1	+0.1	−0.0	0.0	0.0	0.0	0.0	0.0	0.
4.0000	4.0000	4.0000	−0.2	+0.2	−0.0	0.0	0.0	0.0	0.0	0.0	0.
1.0000	1.0000	1.0000	−0.6	+0.6	−0.1	0.0	0.0	0.0	0.0	0.0	0.
0.4000	0.4000	0.4000	−1.5	+1.6	−0.2	0.0	0.0	0.0	0.0	0.0	0.
0.1000	0.1000	0.1000	−5.8	+6.3	−0.7	0.0	0.0	0.0	0.0	−0.1	0.
0.0400	0.0400	0.0400	−14.2	+15.5	−1.8	0.0	0.0	0.0	0.0	0.0	0.
0.0100	0.0100	0.0100	−58.0	+65.0	−7.0	0.0	0.0	0.0	0.0	0.0	−2.
0.1000	10.0000	10.0000	−3.8	+0.1	−0.0	0.0	+0.002	−0.002	0.0	0.0	0.
10.0000	0.1000	10.0000	−0.0	+5.9	−0.0	0.0	0.0	0.0	0.0	−0.5	0.
10.0000	10.0000	0.1000	−0.0	+0.1	−0.5	0.0	0.0	+0.1	0.0	0.0	+0.
0.0100	10.0000	10.0000	−51.0	+0.1	−0.0	0.0	+0.002	−0.002	0.0	0.0	0.
10.0000	0.0100	10.0000	−0.1	+50.0	−0.0	0.0	0.0	+0.001	0.0	−6.0	0.
10.0000	10.0000	0.0100	−0.1	+0.0	−8.0	0.0	−0.0001	0.00	0.0	0.0	+9.

3

on efficiency of weight vectors trained on noise-free data and tested on titration curves with a ›f 0.01 pH unit in the measurements

omposition l)		Prediction $K_{b_1} = 2 \times 10^{-4}$, $K_{b_2} = 1 \times 10^{-4}$ Relative error (%)			Prediction $K_{b_1} = 5 \times 10^{-4}$, $K_{b_2} = 1 \times 10^{-4}$ Relative error (%)			Prediction $K_{b_1} = 10^{-4}$, $K_{b_2} = 10^{-8}$ Relative error (%)		
CB1	*CB2*	*CNAOH*	*CB1*	*CB2*	*CNAOH*	*CB1*	*CB2*	*CNAOH*	*CB1*	*CB2*
10.0000	10.0000	−0.5	+2.8	−1.5	−1.4	−2.0	−2.7	−0.1	−0.5	−0.4
4.0000	4.0000	−1.1	+7.3	−4.4	−3.1	−4.8	−6.1	−0.2	−1.3	+1.3
1.0000	1.0000	−3.8	+29.8	−18.6	−11.9	−18.9	−23.5	−0.6	−5.0	+6.8
0.4000	0.4000	−12.6	−63.3	−55.8	−34.9	−21.1	−1.9	+3.6	+16.8	−0.2
0.1000	0.1000	−50.4	−253.7	−226.1	−139.0	+241.3	−7.0	+14.3	+66.9	+39.1
0.0400	0.0400	−90.5	−1011.5	+667.0	−347.2	416.8	372.5	−43.0	+198.5	−85.5
0.0100	0.0100	−362.0	−3837.0	2660.0	−1389.0	1668.0	1487.0	−172.0	+793.0	+206.0
10.0000	10.0000	+45.0	+3.3	+1.4	+86.3	+3.7	+1.0	−18.3	−0.4	0.3
0.1000	10.0000	+0.5	+429.6	+1.7	+0.9	+408.6	+1.1	−0.1	−40.4	−0.6
10.0000	0.1000	+0.5	+4.4	529.0	+0.9	+4.3	+109.4	−0.2	−0.5	+53.4
10.0000	10.0000	−644.0	+3.5	−1.6	−1251.0	−4.7	−0.5	−113.0	0.4	−0.4
0.0100	10.0000	−0.7	+3875.0	−2.5	−1.3	−4601.0	−1.2	−0.1	+492.9	−0.8
10.0000	0.0100	−0.7	+3.9	−2631.0	−1.4	−4.6	−1439.0	−0.1	+0.5	−3750.0

V(49). Table 3 shows the results for the case in which training was done with a training set of 200 "perfect" members and in which predictive efficiency was calculated on a test set with a spread of ±0.01 pH unit in the measurements. It can be seen that the predictive efficiency of the classifiers trained with "perfect" titration curves is rather poor for "noisy" titration curves.

A much better performance in the prediction of titration curves with random errors is shown by classifiers trained on titration curves which also contained random errors in the pH values (Table 4). Table 5 shows that the use of the derivative preprocessing algorithm of eqn. (2) improves the results for the weak bases but degrades the results for the strong base.

Real titrations

The training of classifiers to predict results for real titrations of mixtures of sodium hydroxide, Tris and sodium acetate, or of sodium hydroxide, Tris and pyridine, or of sodium hydroxide, pyridine and sodium acetate was done with patterns, calculated as described above, with a spread of ±0.01 pH unit. A different set of concentrations of the components was used, i.e. 0.0, 2×10^{-4}, 4×10^{-4}, 6×10^{-4}, ..., 10^{-3} M. In the calculations, the following constants were used for the 1 M KCl medium: $K = 6.81 \times 10^{-15}$, $f_{OH^-} = 0.68$, $f_{B1H^+} = 1.0$, $f_{B2H^+} = 1.0$, K_b (pyridine) $= 2.87 \times 10^{-9}$, $f_{H^+} = 0.87$, K_b (acetate) $= 2.81 \times 10^{-10}$ and K_b (Tris) $= 1.98 \times 10^{-6}$. The K_b values in 1 M KCl were determined from separate experiments by curve-fitting the titration data as described earlier [3].

The predictive ability of the classifiers was tested on actual titration curves recorded by the computerized titration system as arrays of values comprising the amount of 0.1000 M hydrochloric acid titrant used at pH values of 12.0, 11.8, 11.6, etc. down to pH 2.4. The starting volume for all samples was 50.0 ml. The results for the three 3-component mixtures are given in Table 6. Table 7

TABLE 4

Prediction efficiency of weight vectors trained and tested on titration curves with a spread of 0.01 pH u

Sample composition ($\times 10^{-3}$ M)			Prediction $K_{b1} = 10^{-1}$, $K_{b2} = 10^{-2}$ Relative error (%)			Prediction $K_{b1} = 10^{-2}$, $K_{b2} = 10^{-3}$ Relative error (%)			Prediction $K_{b1} = 1 \times 10^{-}$, $K_{b2} = 5 \times 10^{-}$ Relative error (%)		
*C*NAOH	*C*B1	*C*B2	*C*NAOH	*C*B1	*C*B2	*C*NAOH	*C*B1	*C*B2	*C*NAOH	*C*B1	*C*B
10.0000	10.0000	10.0000	+2.9	−1.9	−0.6	+5.4	−11.7	+0.4	+0.1	+3.5	−
4.0000	4.0000	4.0000	−1.8	−0.7	+0.9	−5.2	−7.1	+2.7	−1.0	+0.9	−
1.0000	1.0000	1.0000	+16.1	−13.4	+3.5	+35.0	+19.0	+9.2	+4.0	−7.4	−
0.4000	0.4000	0.4000	−32.2	+9.1	−47.5	−50.0	16.9	+28.2	−14.1	−16.8	−1
0.1000	0.1000	0.1000	−64.8	0	+83.5	−197.8	−4.0	+15.6	−14.9	+10.9	+
0.0400	0.0400	0.0400	+102.5	−235.5	−36.3	−704.3	+946.0	−157.8	−39.8	+328.3	+7
0.0100	0.0100	0.0100	+1968	+1376	+512	−36.0	−4343.3	−732.0	+771.0	+1189.0	−67
0.1000	10.0000	10.0000	+4413	−47.3	+4.6	+1008.7	−0.2	+2.3	+0.4	−0.4	−
10.0000	0.1000	10.0000	−47.7	+5452	−9.8	−0.1	−327.3	+1.0	−0.0	+82.6	
10.0000	10.0000	0.1000	+5.1	−8.5	+260.9	+0.5	+4.6	+87.5	−0.4	+0.4	−2
0.0100	10.0000	10.0000	>>>	−47.3	+5.9	+1308.0	−5.7	−0.5	−825.0	+2.9	+
10.0000	0.0100	10.0000	−46.2	>>>	−9.6	−2.3	−5175.0	+0.2	+0.4	4429.0	+
10.0000	10.0000	0.0100	+7.1	−10.0	+1828	+1.1	−9.9	+768.0	+0.3	+0.3	−75
1.0000	10.0000	10.0000	+389.3	−43.7	−0.3	+47.2	0.0	+1.3	−0.7	+0.8	−
10.0000	1.0000	10.0000	−43.6	+509.4	−6.1	+1.2	+10.8	+0.3	−0.1	−3.2	+
10.0000	10.0000	1.0000	+5.3	−7.1	28.5	+4.3	−2.8	10.6	−0.1	+1.0	+

TABLE 5

Prediction efficiency of weight vectors trained and tested on derivatives of titration curves with a spread

Sample composition ($\times 10^{-3}$ M)			Prediction $K_{b1} = 10^{-2}$, $K_{b2} = 10^{-3}$ Relative error (%)			Prediction $K_{b1} = 10^{-3}$, $K_{b2} = 10^{-4}$ Relative error (%)			Prediction $K_{b1} = 5.0 \times 10$, $K_{b2} = 1.1 \times 10$ Relative error (%)		
*C*NAOH	*C*B1	*C*B2	*C*NAOH	*C*B1	*C*B2	*C*NAOH	*C*B1	*C*B2	*C*NAOH	*C*B1	*C*B
10.0000	10.0000	10.0000	−8.0	−1.4	+0.6	−11.4	+0.7	−0.5	−11.6	+0.9	−
4.0000	4.0000	4.0000	+5.0	−3.6	+1.3	+11.3	+0.3	−0.2	+11.4	+0.4	−
1.0000	1.0000	1.0000	+11.7	+10.8	−0.4	−51.7	−1.3	−1.3	−52.2	−1.5	−
0.4000	0.4000	0.4000	+3.0	+5.4	+6.8	+119.7	+17.5	+4.6	+134.6	+16.6	+
0.1000	0.1000	0.1000	+76.5	+133.3	+101.6	+298.5	−36.5	−31.4	+285.0	−49.6	−2
0.0400	0.0400	0.0400	>>>	−187.2	−289.5	−587.0	+46.0	−63.0	<<<	−33.0	−7
0.0100	0.0100	0.0100	>>>	<<<	+465.0	<<<	−134.0	−37.0	<<<	−47.0	+5
0.1000	10.0000	10.0000	>>>	−1.25	+1.1	+219.4	−0.5	+0.5	+181.3	−0.4	+
10.0000	0.1000	10.0000	−18.8	+702.3	−0.9	−9.6	+53.5	+0.3	−9.3	+52.0	+
10.0000	10.0000	0.1000	−23.7	+7.8	+141.7	−9.6	+2.0	+69.0	−8.3	+1.7	−11
0.0100	10.0000	10.0000	<<<	−3.5	+1.9	>>>	+0.1	+0.6	>>>	+0.4	+
10.0000	0.0100	10.0000	−26.9	>>>	−0.5	−19.1	+37.0	+0.1	−18.5	−24.0	+
10.0000	10.0000	0.0100	−17.5	+5.4	−308.0	−11.9	+0.9	−374.0	−11.3	+0.9	−36
1.0000	10.0000	10.0000	+261.7	−2.1	+2.4	+64.0	+0.4	−0.1	+46.7	+0.8	−
10.0000	1.0000	10.0000	−18.6	+32.6	−0.4	−12.9	+4.9	−0.5	−12.5	+4.6	−
10.0000	10.0000	1.0000	−35.3	+6.6	−1.9	−18.8	+1.4	−6.7	−17.1	+1.2	−

shows that the use of the derivative preprocessing algorithm of eqn. (2) also improves the results for the weak bases in actual titrations. In this case, the recorded titration curves were differentiated by means of a 5-point cubic first-derivative convolution procedure described by Savitzky and Golay [6].

he measurements

iction $K_{b1} = 10^{-4}$ $K_{b2} = 10^{-8}$ tive error (%)			Prediction $K_{b1} = 10^{-11}$ $K_{b2} = 10^{-12}$ Relative error (%)			Prediction $K_{b1} = 10^{-10}$ $K_{b2} = 10^{-11}$ Relative error (%)			Prediction $K_{b1} = 10^{-9}$ $K_{b2} = 10^{-10}$ Relative error (%)		
ıOH	CB1	CB2	CNAOH	CB1	CB2	CNAOH	CB1	CB2	CNAOH	CB1	CB2
4	+0.1	+0.2	+0.1	−0.1	+4.4	+0.1	−0.6	−0.4	+0.0	+0.3	−0.4
4	−0.0	−0.0	−0.0	+1.1	+15.4	−0.0	−1.1	−0.6	−0.1	−0.6	−0.3
6	−0.4	+0.3	+0.2	−1.4	+19.2	+0.3	+2.3	−5.1	+0.3	+0.5	−1.8
6	−3.3	−1.0	−1.1	−25.7	+5.0	−1.2	+4.1	−2.6	−1.5	+2.2	−7.9
2	+7.0	−0.9	−0.1	+15.2	−24.8	−0.3	−4.1	+39.9	−0.4	+6.4	−2.9
2	+70.5	+9.8	+3.0	+301.0	−576.7	−6.0	+114.8	123.0	−6.3	−9.0	−105.5
0	+232.0	−20.0	+30.0	−480.0	+2135.0	+55.0	−284.0	+1151.0	+64.0	−78.0	123.0
7	−0.3	+0.1	−0.8	+0.4	+1.8	+0.7	−0.3	+3.0	−3.5	+0.2	+0.4
1	−3.1	+0.1	−0.0	+203.3	+3.0	+0.0	−35.2	−1.5	+0.0	+29.4	−0.1
1	−0.2	−6.5	−0.0	−0.4	+245.4	−0.0	+0.5	−10.5	−0.0	+0.2	−34.3
0	+0.1	+0.2	+37.0	+0.9	+1.2	−131.0	0.7	−0.6	−62.0	+0.3	+0.5
3	+29.0	+0.0	−0.0	+1085.0	−3.5	+0.0	+511.0	0.9	+0.1	−365.0	+0.5
1	+0.2	−7.0	−0.0	−0.3	>>>	+0.0	+0.4	+577.0	+0.0	+0.3	−67.0
5	−0.1	−0.1	−0.0	+1.0	+3.8	+0.5	+0.7	−0.4	+0.5	+1.0	−0.3
1	−2.6	+0.0	−0.0	+7.0	+0.2	−10.0	+0.6	+2.2	+0.0	+1.5	+0.1
1	0.1	−0.9	0.0	−2.6	−45.3	−0.0	+0.4	−19.5	−0.0	−0.1	−0.8

.01 pH unit in the measurements

iction $K_{b1} = 10^{-9}$ $K_{b2} = 10^{-10}$ tive error (%)			Prediction $K_{b1} = 10^{-10}$ $K_{b2} = 10^{-11}$ Relative error (%)			Prediction $K_{b1} = 10^{-11}$ $K_{b2} = 10^{-12}$ Relative error (%)		
ıOH	CB1	CB2	CNAOH	CB1	CB2	CNAOH	CB1	CB2
5	−0.6	+0.1	−16.1	−0.1	+1.6	−19.6	+0.2	−3.0
8	+0.2	−0.0	+11.7	−0.1	−0.6	+8.8	+2.4	−2.3
6	+0.4	−2.5	−69.3	−1.8	+3.7	−89.0	−0.7	+32.3
2	−2.1	−1.5	+187.9	−10.8	+25.5	+213.6	+2.0	+60.0
2	+5.3	−23.0	+386.3	−22.3	+44.6	+502.6	+186.1	−89.0
>	+2.3	+21.0	+421.0	+109.8	−257.5	+42.5	−308.5	−159.0
<	+10.0	−29.0	>>>	+135.0	>>>	<<<	−822.0	<<<
>	+0.0	+0.1	+107.9	−0.3	−0.5	+43.0	−0.3	−5.1
8	+9.7	+0.2	−10.7	−80.1	+1.2	−13.3	−14.9	+2.4
4	−0.3	−8.4	−7.9	−0.6	+63.3	−9.0	−1.4	>>>
>	−0.4	−0.4	>>>	−0.2	−1.1	>>>	−1.0	−4.5
7	−151.0	+0.1	−19.7	+170.0	+1.4	−21.9	>>>	−7.5
6	+0.1	−452.0	−13.8	−0.2	−189.0	−17.7	−0.8	>>>
0	+0.5	−0.3	+11.0	−0.1	−0.6	+4.3	−0.9	+1.8
5	+0.6	+0.4	−16.6	+1.4	+0.4	−17.8	+9.3	+3.3
9	−0.5	+3.4	−8.3	−0.4	+10.8	−15.6	−0.7	+42.2

DISCUSSION

The results clearly indicate that neither the titration curves themselves nor their derivatives can be used alone to operate the learning machine method to its full extent. The most accurate results are obtained by training the weight vector for the strong base content on the titration curves themselves, whereas for the training of the classifiers for the two weak bases it is advantageous to use the derivatives of the titration curves.

TABLE 6

Prediction efficiency of weight vectors trained on titration curves with a spread of 0.01 pH unit in the measurements for actual titrations

Sample	NaOH added ($\times 10^{-3}$ M)	NaOH found ($\times 10^{-3}$ M)	Compound 2	Added ($\times 10^{-3}$ M)	Found ($\times 10^{-3}$ M)	Compound 3	Added ($\times 10^{-3}$ M)	Found ($\times 10^{-3}$ M)
1	10.00	9.86	Tris	8.00	7.95	Sodium acetate	—	—0.13
2	10.00	9.95		8.00	8.08		—	—0.17
3	10.00	9.83		—	0.00		8.00	7.51
4	10.00	10.14		4.00	3.96		4.00	3.75
5	10.00	10.11		4.00	3.98		4.00	3.71
6	9.60	9.61		—	—0.06		8.00	7.57
7	10.00	9.95		—	—0.01		8.00	7.46
8	10.00	9.97		—	—0.01		8.00	7.43
9	10.00	9.96		—	—0.01		8.00	7.49
10	10.00	9.91		4.00	3.98		8.00	7.36
11	10.00	9.95		4.00	3.96		8.00	7.42
12	10.00	9.98		8.00	7.92		4.00	3.65
13	10.00	9.97		8.00	7.92		4.00	3.65
14	10.00	9.99		8.00	7.91		2.00	1.73
15	10.00	9.96		8.00	7.95		2.00	1.74
16	10.00	9.90		4.00	3.99	Pyridine	7.88	7.65
17	10.00	9.86		4.00	4.01		7.88	7.65
18	10.00	9.86		4.00	4.00		3.94	3.73
19	10.00	9.83		4.00	4.01		3.94	3.78
20	10.00	9.86	Pyridine	3.94	4.10	Sodium acetate	8.00	7.22
21	10.00	9.89		3.94	4.02		8.00	7.26
22	10.00	9.83		1.97	2.25		8.00	7.17

TABLE 7

Prediction efficiency of weight vectors trained on titration curves with a spread of 0.01 pH unit in the measurement for actual titrations after preprocessing

Sample	NaOH added ($\times 10^{-3}$ M)	NaOH found ($\times 10^{-3}$ M)	Compound 2	Added ($\times 10^{-3}$ M)	Found ($\times 10^{-3}$ M)	Compound 3	Added ($\times 10^{-3}$ M)	Found ($\times 10^{-3}$ M
1	10.00	220	Tris	8.00	7.98	Sodium acetate	—	0.24
2	10.00	200		8.00	7.94		—	0.08
3	10.00	195		—	0.05		8.00	7.75
4	10.00	194		4.00	4.04		4.00	4.07
5	10.00	195		4.00	4.03		4.00	4.03
6	9.60	187		—	0.13		8.00	7.86
7	10.00	191		—	0.00		8.00	7.76
8	10.00	193		—	0.06		8.00	7.71
9	10.00	191		—	0.07		8.00	7.78
10	10.00	196		4.00	4.01		8.00	7.69
11	10.00	195		4.00	4.03		8.00	7.75
12	10.00	196		8.00	7.99		4.00	4.01
13	10.00	196		8.00	7.99		4.00	4.03
14	10.00	192		8.00	7.98		2.00	2.13
15	10.00	194		8.00	7.99		2.00	2.12
16	10.00	194		4.00	4.08	Pyridine	7.88	7.79
17	10.00	195		4.00	4.06		7.88	7.79
18	10.00	192		4.00	4.06		3.94	3.90
19	10.00	194		4.00	4.02		3.94	3.93
20	10.00	195	Pyridine	3.94	3.86	Sodium acetate	8.00	7.68
21	10.00	196		3.94	3.77		8.00	7.70
22	10.00	195		1.97	2.04		8.00	7.58

The accuracy in the pH measurements that can be maintained in practice during acid—base titrations is of the order of ±0.01 pH unit. Improvement in this accuracy would greatly improve the results of the learning machine method applied to titration curves. In practice, the pH range of the titration can be adjusted to that region of the pH scale where the protonation of the compounds of interest changes.

The experimental results show that a set of theoretically calculated titration curves can be used successfully for prediction of actual samples. This circumvents the major disadvantage of a learning-machine method, viz. the lengthy gathering of the training set. A point of further study should be the influence of the upper and lower bounds of the pH range of the titration on the limits of the pK values of the weak bases that can be titrated successfully.

The author thanks Ms. B. Verbeeten-van Hettema for preparing the manuscript, Drs. J. H. H. G. van Willigen for mathematical help, and Prof. dr. ir. E. A. M. F. Dahmen for his interest in this work.

REFERENCES

1 D. M. Barry, L. Meites and B. H. Campbell, Anal. Chim. Acta, 69 (1974) 143.
2 F. Ingman, A. Johansson, S. Johansson and R. Karlsson, Anal. Chim. Acta, 64 (1973) 113.
3 M. Bos, Anal. Chim. Acta, 90 (1977) 61.
4 B. R. Kowalski, P. C. Jurs, T. L. Isenhour and C. N. Reilley, Anal. Chem., 41 (1969) 695.
5 M. Bos and G. Jasink, Anal. Chim. Acta, 103 (1978) 151.
6 A. Savitzky and M. J. E. Golay, Anal. Chem., 36 (1964) 1627.

Analytica Chimica Acta, 112 (1979) 75—82
Computer Techniques and Optimization

LEISTUNGSFÄHIGE MATRIXKORREKTUR IN DER RÖNTGENSPEKTROMETRIE

ROLF PLESCH* und BERTHOLD THIELE

Siemens Aktiengesellschaft, Bereich Meß- und Prozeßtechnik, D-75 Karlsruhe (Bundesrepublik Deutschland)

(Eingegangen den 18 Juni 1978)

SUMMARY

Efficient matrix correction in x-ray fluorescence spectrometry

The SPECTRA 310 program for matrix correction in x-ray fluorescence spectrometry allows concentrations as well as intensities of the matrix elements to be included in the correction factors. Alongside an existing testing system, this allows the efficiency of the correction method to be estimated and the quality of standards to be assessed. A program for trace analysis improves general analysis in the trace range, and a test program allows conclusions to be made about the statistical precision of the analytical equipment used.

ZUSAMMENFASSUNG

Das Programmsystem SPECTRA 310 für die Matrixkorrektur in der Röntgenspektrometrie enthält die Möglichkeiten, sowohl die Konzentrationen als auch die Intensitäten der Matrixelemente als Korrekturgrößen zu verwenden. Daneben existieren Prüfprogramme, die es dem Praktiker erlauben, die Effektivität seiner Korrekturmaßnahmen zu erkennen und die Qualität seiner Standards zu beurteilen. Ein Spurenprogramm verbessert die Analytik im Spurenbereich und ein Testprogramm liefert Rückschlüsse auf die statistisch einwandfreie Funktion der Analyseneinrichtung.

Unter der Matrixkorrektur versteht man die Möglichkeit, die Einflüsse der von Probe zu Probe in ihrem Gehalt wechselnden Matrixelemente so weit als möglich auszuschalten und damit den Analysenfehler zu verringern.

Da die Stärke der Röntgenspektrometrie in der Analyse von Multielementproben liegt, wobei jedes Probenelement die gemessene Intensität eines jeden anderen beeinflussen kann, ergeben sich umfangreiche Korrekturrechnungen, die automatisch von eingebauten Geräterechnern vorgenommen werden. Als Ausgangspunkt dienen mathematische Ansätze, die auf den physikalischen Grundlagen der Röntgenanalyse beruhen und in der Lage sind, allen in den Proben auftretenden Wechselwirkungen der Elemente gerecht zu werden. Welche Möglichkeiten ein modernes Korrektursystem dem Analytiker bietet, sei am Beispiel des Systems SPECTRA 310 (System der Fa. Siemens AG) im folgenden erläutert. Die Theorie der Ansätze wurde bereits an anderer Stelle behandelt [1, 2]. Hier soll gezeigt werden, welche Eigenschaften das System

in der Praxis besitzt und wie es dazu beiträgt, einen möglichst geringen Analysenfehler zu erhalten, wobei von der Theorie nur so weit als nötig Gerbrauch gemacht wird.

Die Forderungen an ein Korrektursystem lassen sich in zwei Begriffen zusammenfassen: es muß wirksam und flexibel sein. Die Korrektur darf nur einen sehr geringen Rest an unkorrigierten Einflüssen zurücklassen und sie muß vom Analytiker schnell und einfach an die Gegebenheiten seiner Proben angepaßt werden können. Beides trifft für das hier behandelte System zu.

Die Elemente einer Probe, soweit sie der Röntgenanalyse überhaupt zugänglich sind, können durch zwei Größen charakterisiert werden: durch ihren Massengehalt (Konzentration) und die von ihnen ausgehende Röntgenintensität. Beide Größen können zur Grundlage der Matrixkorrektur gemacht werden, weshalb SPECTRA 310 auch über zwei grundlegende mathematische Ansätze verfügt, die künftig als Konzentrations- und Intensitätsansatz bezeichnet werden sollen.

KONZENTRATIONSANSATZ

Der vollständige Ansatz lautet:

$$C_i = (a_{i0} + a_{i1} I_i + a_{i2} I_i^2) \left(1 + \sum k_{ij} C_j + \frac{\Sigma B_{ij} C_j}{1 + C_i}\right) \qquad (1)$$

Er stellt in der ersten Klammer den Massengehalt C_i des Analysenelements i als Funktion der Intensität I_i (Brutto- oder Nettointensität) dar, während in der zweiten Klammer die Interelementeinflüsse der Matrixelemente j mit ihren Massengehalten C_j zusammengefaßt sind.

Das Korrekturprogramm hat zwei Aufgaben. Es muß zunächst einmal die Koeffizienten a_{i0}, a_{i1}, a_{i2}, k_{ij} und B_{ij} berechnen, wobei die beiden letzteren einen für jedes Matrixelement verschiedenen Wert besitzen. Nach ihrer Berechnung dient Gl. (1) dann als Eichfunktion für die Analyse der unbekannten Proben. Da C_i auf beiden Seiten auftritt, ist hierzu eine Iterationsrechnung notwendig, die aber bereits nach wenigen Schritten konvergiert.

Scheinbare Konzentration

Unter diesem Begriff versteht man diejenige Konzentration, die von der ersten Klammer des Ansatzes (1) dargestellt wird, wenn man den Intensitätswert I_i eines bestimmten Standards einführt und die Koeffizienten k_{ij} und B_{ij} der zweiten Klammer zu Null annimmt. Man erhält dann in Gestalt der ersten Klammer eine Eichfunktion für die scheinbare Konzentration, die alle Matrixeffekte vorläufig ignoriert und deren Koeffizienten a_{i0}, a_{i1} und a_{i2} mittels Regressionsrechnung berechnet werden, wozu man die Werte C_i und I_i der Standards verwendet.

Der Koeffizient a_{i2} kann in vielen Fällen zu Null angenommen werden, wenn mit genügender Genauigkeit ein linearer Zusammenhang zwischen der Intensität und der scheinbaren Konzentration vorhanden ist. (Hinweis: Der

Begriff der Konzentration ist hier immer als Massengehalt zu verstehen. Einer Konzentration von beispielsweise 10% entspricht ein Massengehalt von 0,1.) Das Programm erlaubt es daher, wahlweise a_{i2} wegzulassen.

Der Koeffizient a_{i0} stellt die mittlere Höhe des Untergrunds über alle Standards dar, wenn man als I_i die Bruttointensität einführt. Werden Nettointensitäten verwendet, so könnte a_{i0} theoretisch gleich Null sein. Zur Ermittlung der Nettointensität muß der Untergrund gemessen werden, was in der Praxis, vor allem bei einem Spektrum mit mehreren Linien, nicht immer fehlerfrei erfolgen kann. Der Koeffizient a_{i0} dient in diesem Falle dazu, diese Fehler aufzufangen und als Mittelwert auszukorrigieren. Er kann daher nicht aus dem Programm entfernt werden.

Koeffizienten k_{ij} *und* B_{ij}

Die zweite Klammer in Ansatz (1) hat die Aufgabe, die eigentliche Matrixkorrektur mit Hilfe der Koeffizienten k_{ij} und B_{ij} zu übernehmen. Sie verwandelt die scheinbare Konzentration der ersten Klammer in die wirkliche Konzentration, die den Analytiker interessiert. Daß dies naturgemäß nicht völlig ohne Fehler geschehen kann, wird noch Gegenstand einer weiteren Betrachtung sein. Nachdem die Koeffizienten der scheinbaren Konzentration durch Regressionsrechnung bestimmt worden sind, erhält man aus Ansatz (1) lineare Gleichungen, aus denen der Rechner automatisch die Koeffizienten k_{ij} and B_{ij} bestimmt. Er greift aus der Gesamtzahl n der Standards einzelne Gruppen heraus, deren Umfang der Zahl der zu bestimmenden Koeffizienten entspricht, so daß jeweils ein vollständig bestimmtes lineares System auftritt. Jeder Gruppe entspricht für jedes Matrixelement ein Koeffizientenpaar k_{ij} und B_{ij} und für die endgültige Eichfunktion nach Ansatz (1) wird denn das Mittel aus den Gruppen verwendet. Fällt ein Gruppenergebnis aus dem Rahmen der übrigen, so kann es vom Rechner mittels eines Prüfprogramms ausgeschaltet werden. Die Schranke ist wählbar und damit dem jeweiligen Analysenproblem anpaßbar.

Die Koeffizienten haben verschiedene Aufgaben im Rahmen der Matrixkorrektur. Die absorptiven Einflüsse der Matrixelemente auf die Strahlung des Analyten werden durch die k_{ij} beschrieben. Wie man theoretisch zeigen kann [1], haben sie einen physikalischen Charakter, was bedeutet, daß sie innerhalb eines bestimmten Analysenproblems, z.B. Stahl, Zement usw., einen weitgehend konstanten Wert besitzen, und daher auch archiviert werden können.

Die k_{ij} werden aber nicht allein durch den absorptiven Einfluß der Matrixelemente auf den Analyten bestimmt, sondern sie enthalten auch einen Beitrag, der durch die Absorption der anregenden Strahlung im Analyten und den Matrixelementen bestimmt wird [1]. Derjenige Anteil der Primärstrahlung, der die Anregung bewirkt, ist Teil eines kontinuierlichen Spektrums und daher ebenfalls kontinuierlich über einen bestimmten Energiebereich verteilt. Diese Verteilung hängt aber von der Konzentration C_i des Analyten ab, da er die anregende Strahlung besonders stark absorbiert. Eine Änderung des anregenden Spektrums aber bedeutet auch eine Änderung der k_{ij}, so daß diese Konstanten über einen großen Konzentrationsbereich des Analyten nicht als völlig konstant betrachtet werden können.

In sehr vielen Fällen der täglichen Praxis kann diese Inkonstanz unbeachtet bleiben. Um aber für alle Fälle gerüstet zu sein, enthält SPECTRA 310 auch die Möglichkeit, die Inkonstanz der k_{ij} mit folgendem Ansatz auszukorrigieren, der von Tertian [3] angegeben wurde,

$$k_{ij} = a_{ij} + \beta_{ij} C_i \qquad (2)$$

Mit dem Ansatz (2) können die k_{ij}, falls notwendig, noch verfeinert werden. Seine Koeffizienten werden durch lineare Regression bestimmt.

Die Koeffizienten B_{ij} wurden aufgrund empirischer Ergebnisse von Rasberry und Heinrich [4] in die Korrekturrechnung eingeführt. Sie berücksichtigen die Erhöhung der Intensität des Analyten durch die Sekundärfluoreszenz, die der Analyt von den Strahlungen solcher Matrixelemente erfährt, die im periodischen System über ihm stehen. Nach Rasberry und Heinrich sollen für Matrixelemente, die keine Sekundärfluoreszenz hervorrufen können, nur die k_{ij} und im anderen Falle nur die B_{ij} verwendet werden. Eingehende Untersuchungen haben aber gezeigt [5], daß es Fälle gibt, wo mit gleichzeitiger Anwendung beider Koeffizienten auf sekundäranregende Matrixelemente eine Verbesserung der Analysengenauigkeit erzielt werden kann. Daher bietet SPECTRA 310 die Möglichkeit, entweder nur einen der Koeffizienten oder gleichzeitig beide anzuwenden. Bei Fehlen von Sekundäranregung bleibt B_{ij} für das betreffende Matrixelement außer Betracht.

Untergrundskorrektur

Der Ansatz (1) arbeitet, wie bereits erwähnt, mit einem mittleren Untergrund a_{i0}, der sich aus der Regressionsrechnung ergibt. Wenn der Untergrund gegenüber der Linienintensität vernachlässigbar ist und keine Linienüberlagerungen vorliegen, führt dieses Verfahren zu guten Ergebnissen. Sind die Konzentrationen des Analyten aber klein, so machen sich die Schwankungen des Untergrunds von Probe zu Probe bemerkbar, die durch Unterschiede im Absorptions- und Streuvermögen der Proben hervorgerufen werden. Ebenso können keine Linienüberlagerungen vom Ansatz (1) auskorrigiert werden.

Diese Nachteile lassen sich dadurch beseitigen, daß man die beiden Klammern des Ansatzes (1) ausmultipliziert. Man erhält dann mit teilweise neuen Koeffizienten, wenn man das wenig benötigte Glied mit I_i^2 wegläßt:

$$C_i = a_{i0} + \Sigma A_{ij} C_j + a_{i1} I_i + I_i \Sigma D_{ij} C_j + (1/1 + C_i)(\Sigma E_{ij} C_j + I_i \Sigma F_{ij} C_j) \qquad (3)$$

Die Koeffizienten des Ansatzes (3) werden nunmehr in einem Zuge berechnet. Der Untergrund wird wieder durch seinen Mittelwert a_{i0} repräsentiert. Darüberhinaus aber sorgt der zweite Term mit A_{ij}, der nur von den Konzentrationen der Matrixelemente abhängt, dafür, daß noch automatisch eine matrixabhängige Korrektur von a_{i0} erfolgt. Gleichzeitig ist dieser Term auch in der Lage, Linienüberlagerungen auszukorrigieren, die letztlich nichts anderes als eine Erhöhung des Untergrunds unterhalb der Linie darstellen.

Theoretische k_{ij}

Wenn die Meßwerte der Standards für die Bestimmung der k_{ij}-Werte verwendet werden, spricht man von empirischen Koeffizienten. Da die k_{ij}, wie schon erwähnt, eine physikalische Bedeutung haben, können sie auch unabhängig von Standards mit Hilfe physikalischer Parameter berechnet werden, die man aus Tabellen in der Literatur entnimmt. Weiterhin wird noch die Kenntnis des anregenden Spektrums benötigt. Die so erhaltenen Werte werden als theoretische Koeffizienten bezeichnet [6], was nicht ganz logisch ist, weil auch die Tabellenwerte der verwendeten Parameter experimentell bestimmt worden sind und daher auch von Autor zu Autor etwas schwanken. Die theoretischen Koeffizienten sind deshalb mit einem systematischen Fehler behaftet, der sich auch im Analysenfehler niederschlägt [7, 8]. Sie stellen aber auf der anderen Seite ein wertvolles Hilfsmittel dar, wenn nur wenige Standards verfügbar sind und weitere entweder nicht oder nur unter großen Schwierigkeiten beschafft werden können [9].

Um auch theoretische Koeffizienten oder solche anderer Herkunft verwenden zu können, enthält SPECTRA 310 die Möglichkeit, fremdbestimmte Koeffizienten in das Korrekturverfahren einzuführen.

INTENSITÄTSANSATZ

Um den Konzentrationsansatz (1) verwenden zu können, ist es notwendig, in den Standards die Konzentrationen aller Elemente zu kennen, die entweder als Analyt oder als Matrixelemente von Wichtigkeit sind. Mitunter sind nicht alle Elemente analysiert, die als Matrixelemente von Wichtigkeit für die Korrektur sein können, weil sie nicht als Analyt in Frage kommen. Oder der Analytiker muß während der Analyse erkennen, daß er eine Verbesserung durch die Einführung von Matrixelementen erzielen könnte, deren Konzentrationen nicht bekannt sind. In solchen Fällen führt Ansatz (1) nicht mehr zum Ziel. Das Gleiche gilt, wenn Matrixelemente schlecht analysiert sind, so daß ihre fehlerhaften Konzentrationen bei der Verwendung in Ansatz (1) zu ungenügenden Ergebnissen führen.

In allen diesen Fällen leistet ein Intensitätsansatz gute Dienste:

$$C_i = a_{i0} + \Sigma\, b_{ij}I_j + a_{i1}I_i + a_{i2}I_i^2 + I_i \,\Sigma\, m_{ij}I_j \qquad (4)$$

Die Koeffizienten des Ansatzes (4) werden durch multiple Regressionsrechnung bestimmt. Sie haben keinen physikalischen Charakter mehr, sondern sind reine Rechengrößen, die vom Rechner so bestimmt werden, daß die Standardabweichung der Eichung nach Gl. (5) zu einem Minimum wird. Sie gelten nur innerhalb des von den Standards bestrichenen Konzentrationsbereichs aller Probenelemente, eine Extrapolation darüberhinaus ist nicht statthaft.

Mit Hilfe des Ansatzes (4) ist es möglich, über ihre gemessene Intensität I_j auch solche Matrixelemente in die Korrektur einzuführen, deren Konzentration nicht oder nicht genau genug bekannt ist. Wie man erkennt, ist auch

im Ansatz (4) dem mittleren Untergrund a_{i0} ein Korrekturglied mit den Koeffizienten b_{ij} beigefügt, das nur von den Matrixintensitäten abhängt und auch die Korrektur von Linienüberlagerungen ausführen kann. Auch hier können nicht benötigte Koeffizienten weggelassen werden.

Die absorptiven Einflüsse der Matrixelemente und die Sekundärfluoreszenz werden durch dieselben Koeffizienten m_{ij} erfaßt. Da die Intensität des Analyten durch die Absorption in den Matrixelementen erniedrigt und durch die Sekundärfluoreszenz erhöht wird, kann man die letztere als negative Absorption auffassen und ihre Korrektur in erster Näherung ebenfalls den m_{ij} übertragen.

ANALYSENFEHLER UND SPURENANALYSE

Wenn auf die beschriebene Weise die Koeffizienten der Ansätze (1), (3) oder (4) bestimmt sind, können mit ihrer Hilfe die Konzentrationen unbekannter Proben vom Rechner automatisch berechnet werden, wenn man mit ihren gemessenen Intensitäten in die Ansätze eingeht.

Die entscheidende Größe, an der sich die Arbeit des Analytikers ausrichtet, ist der Analysenfehler, oder anders gesagt, die Standardabweichung s_x, die bei der Analyse der unbekannten Proben auftritt.

Auf direkte Weise läßt sich s_x nicht ermitteln, wohl aber über die Standardabweichung s, die für das Eichverfahren mit den Standards gilt.

Für die Eich-Standardabweichung s (in der Literatur auch oft als Reststreuung bezeichnet) kann man schreiben [10]:

$$s^2 = \Sigma_n \, (C_i - C_{ib})^2/(n-p) \qquad (5)$$

C_i = gegebene Konzentration der Standards; C_{ib} = nach Bestimmung der Koeffizienten nach Ansatz (1), (3) oder (4) berechnete Konzentration der Standards; n = Anzahl der verwendeten Standards; p = Anzahl aller berechneten Koeffizienten; $n-p$ = Anzahl der Freiheitsgrade für die vorliegende Eichung.

Man kann zeigen, daß der Analysenfehler s_x in guter Näherung durch s nach Gl. (5) repräsentiert wird [10]. Die Werte von s werden vom Programm ausgedruckt und können zur Beurteilung des Eichverfahrens dienen. Der Wert von s nach Gl. (5) wird stark von der Zahl $n-p$ der Freiheitsgrade beeinflußt, die möglichst groß sein sollte. Dies kann man dadurch erreichen, daß man p möglichst klein macht, oder mit anderen Worten, die Anzahl p der Koeffizienten möglichst beschränkt, indem nur die notwendigen Matrixelemente zur Korrektur herangezogen werden. SPECTRA 310 enthält daher ein Prüfprogramm, das eine Beurteilung der Matrixelemente in Bezug auf ihre Wichtigkeit für die Korrektur erlaubt [1]. Sie werden in dieser Reihenfolge in die Ansätze eingeführt und die Korrektur mit demjenigen Matrixelement beendet, dessen Nachfolger keine merkliche Verbesserung des mit ihm durchgeführten Verfahrens mehr nach Gl. (5) erkennen läßt.

Eine Überkorrektur mit zu vielen Matrixelementen würde wieder zu

einer Verschlechterung des Analysenfehlers führen, weil ein Teil von ihnen die Abweichungsquadrate im Zähler der Gl. (5) nicht mehr verkleinert, wohl aber die Anzahl der Freiheitsgrade im Nenner und somit zu einer Vergrößerung von s führt.

Weiterhin muß dafür Sorge getragen werden, daß bei der Eichung keine Meßwerte verwendet werden, die offensichtlich aus der Reihe der übrigen herausfallen. Mit Hilfe eines Prüfprogrammes [1] mit wählbarer Schranke können daher zu große standardisierte Abweichungen $(C_i - C_{ib})/s$ ausgeschaltet werden, wobei aber zunächst noch nichts darüber gesagt werden kann, ob es sich um einen ungeeigneten Standard oder lediglich um einen falschen Meßwert der Analyseneinrichtung handelt. Angesichts des hohen Arbeitsaufwandes, der mit der Herstellung der Standards verbunden ist, kommt dieser Entscheidung besondere Bedeutung zu. Deshalb enthält SPECTRA 310 noch ein weiteres Prüfprogramm, das aufgrund physikalischer Kriterien die Beurteilung der Standards gestattet [1].

Dieses Prüfprogramm erlaubt es, ungeeignete Standards aus dem Eichverfahren zu entfernen, um eine möglichst niedrige Eich-Standardabweichung s nach Gl. (5) zu erhalten. Die statistisch einwandfreie Funktion der Analyseneinrichtung kann mit Hilfe eines Testprogramms überprüft werden [11].

Man darf sich aber nicht dazu verleiten lassen, zu viele Standards als ungeeignet zu entfernen. Man kann dann zwar einen guten Wert von s erreichen, aber die verbliebenen Standards sind unter Umständen nicht mehr repräsentativ für die unbekannten Proben. Statistisch gesehen, gehören die Standards und die unbekannten Proben nicht mehr derselben Grundgesamtheit an, womit ein wesentliches Erfordernis der Analytik verletzt wird.

Es empfiehlt sich, dieser Gefahr auf folgende Weise zu entgehen: die vorhandenen n Standards werden in zwei Gruppen vom Umfang n_1 und n_2 unterteilt. Die erste Gruppe wird für die Koeffizientenbestimmung nach einem der Ansätze (1), (3) oder (4) verwendet. Anschließend erfolgt die Analyse der zweiten Gruppe im Sinne unbekannter Proben. Bei richtiger Korrektur müssen die Standardabweichungen s_1 und s_2 beider Gruppen nach Gl. (5) etwa übereinstimmen. Für die Berechnung von s_1 muß $n_1 - p$ im Nenner der Gl. (5) eingesetzt werden, während bei s_2 lediglich n_2 im Nenner erscheint, da mit Hilfe der zweiten Gruppe keine Koeffizienten bestimmt wurden ($p = 0$). Wenn nach dem Ausscheiden von Standards aus der ersten Gruppe s_1 wesentlich kleiner als s_2 ausfällt, dann hat man zu viele Standards ausgeschieden. In Zweifelsfällen kann man s_1 und s_2 mit Hilfe des F-Tests miteinander vergleichen [12—14]. Die Freiheitsgrade sind dabei $n_1 - p$ und n_2.

Es ist erforderlich, daß immer mehr Standards vorhanden sind als Koeffizienten bestimmt werden müssen. Nach einer Faustregel soll $n = 2p$ sein, bei guten Standards auch etwas weniger. Dasselbe gilt, wenn man die erwähnte Aufteilung der Standards in zwei Gruppen vornimmt. Es sei noch erwähnt, daß alle genannten Prüfprogramme auf alle drei Ansätze angewandt werden können.

Wie schon erwähnt, enthalten die Ansätze (3) und (4) die Möglichkeit, Schwankungen des Untergrundes von Probe zu Probe automatisch auszukorrigieren, wenn man mit Bruttointensitäten in die Ansätze eingeht. Bei sehr kleinen Konzentrationen des Analyten, insbesondere im Spurenbereich, können aber die Schwankungen des Untergrunds so groß werden, daß die Verwendung der Nettointensität einebeträchtliche Verringerung von s zur Folge hat. Mit Hilfe des Spurenprogramms ist die automatische Berechnung des Untergrunds und seine Subtraktion von der Bruttointensität möglich, wobei die Messung des Untergrunds wahlweise beidseitig der Röntgenlinie oder nur auf einer Seite erfolgen kann. Sein Wert unter der Linie wird daraus vom Rechner berechnet. Er kann auch den Quotienten aus Intensität und Untergrund bilden, dessen Verwendung in den Ansätzen (3) und (4) in vielen Fällen aus physikalischen Gründen [15] zu einer weiteren Verbesserung der Analyse im Sinne einer Verkleinerung von s führt.

Schlußbemerkung

Die Vielfalt der in SPECTRA 310 vorhandenen Möglichkeiten läßt die Frage aufkommen, welchen Ansatz beispielsweise man am besten bei welchem Analysenproblem verwendet. Es ist aber nicht möglich, eine allgemeingültige Antwort zu geben, weil vor allem die Art und Güte der Standards über das Ausmaß und die Methode der Matrixkorrektur entscheiden, wobei immer die Standardabweichung nach Gl. (5) als Entscheidungskriterium dient.

Das beschriebene System erlaubt es dem Praktiker, auf empirische Weise schnell und sicher die für ihn günstigste Methode aufzufinden, wobei ihm die hier behandelten Prinzipien eine nützliche Hilfe sein können. Letztlich liegt es aber an ihm, das reichhaltige Instrumentarium möglichst umfassend für seine Zwecke dienstbar zu machen. Die Möglichkeiten dazu sind gegeben.

LITERATUR

1 R. Plesch und B. Thiele, Siemens Analysentechnische Mitteilung, Nr. 196 (1977) Best. Nr. E 632/113.
2 R. Plesch, Siemens Z., 49 (1975) 657.
3 R. Tertian, X-Ray Spectrom., 2 (1973) 95.
4 S. D. Rasberry und K. F. J. Heinrich, Anal. Chem., 46 (1974) 81.
5 R. Plesch, Siemens Z., 51 (1977) 841.
6 W. K. De Jongh, X-Ray Spectrom., 2 (1973) 151.
7 J. Kramer, H. Ebel und F. Tschismarov, X-Ray Spectrom., 6 (1977) 30.
8 R. Plesch, Fresenius Z. Anal. Chem., 292 (1978) 378.
9 R. Jenkins, J. F. Croke, R. L. Niemann und R. G. Westberg, Norelco Rep., 23 (1976) 32.
10 R. Plesch, G-I-T Fachz. Lab., 17 (1973) 677.
11 R. Plesch, Arch. Tech. Messen. Lfg., 437 (1972) R85.
12 Graf, Henning und Stange, Formeln und Tabellen der mathematischen Statistik, Springer-Verlag, Berlin, 1966.
13 L. Sachs, Angewandte Statistik, Springer-Verlag, Berlin, 1974.
14 R. Kaiser und G. Gottschalk, Elementare Tests zur Beurteilung von Meßdaten, Bibliographisches Institut, 1972.
15 R. Plesch, Siemens Analysentechnische Mitteilungen, Nr. 91 (1974) Best. Nr. E 632/026.

Analytica Chimica Acta, 112 (1979) 83—93
Computer Techniques and Optimization

INFORMATION THEORY APPLIED TO FEATURE SELECTION OF BINARY-CODED INFRARED SPECTRA FOR AUTOMATED INTERPRETATION BY RETRIEVAL OF REFERENCE DATA

P. F. DUPUIS**, P. CLEIJ, H. A. VAN 'T KLOOSTER* and A. DIJKSTRA

Analytisch Chemisch Laboratorium, Rijksuniversiteit Utrecht, Croesestraat 77A, Utrecht (The Netherlands)

(Received 4th July 1978)

SUMMARY

A method is described for feature selection from infrared spectra, intended for identification of organic compounds by computer-aided retrieval of reference data contained in small files. Complete discrimination of the binary-coded spectra is achieved by selecting a minimum number of spectral features; the information content is used as the selection criterion. The selection procedure is applied to five data sets (saturated and unsaturated hydrocarbons, alcohols, ethers and aldehydes/ketones) involving some 400 spectra. Each spectrum is uniquely coded by using about 10% of the 140 spectral features (binary-coded peak positions) available originally. For the intensity, a threshold of 50% appears to be applicable in some cases. For coding the frequency or wavelength parameter, wavenumbers (cm^{-1}) are preferred to wavelengths (μm). The method takes into account the a priori probabilities of spectral features and their correlations. Results of a retrieval program for a few "unknown" spectra are given.

Infrared spectrometry is an indispensable tool for identification of organic compounds. As in other areas of analytical chemistry, the vast numbers of spectra now available have emphasized the importance of computer-aided techniques, particularly for data processing and interpretation. Retrieval of reference spectra is one of the methods being used for automated identification [1].

In previous papers, the application of information theory to the coding of infrared spectra for retrieval purposes has been reported [2, 3]. Shannon's formula was used to calculate the information contents of retrieval procedures involving large files and different coding methods. It was shown that only a part of the spectral features contributes to the information content. It also appeared that errors occurring in the coded spectra and correlations between spectral features considerably diminish the information content and thus the applicability of infrared data collections, such as the (binary-coded) ASTM Infrared Spectral Index for retrieval purposes [2].

The application of numerical taxonomy and information theory to the selection of features from infrared spectra contained in files has recently been

**Present address: Dow Chemical (Nederland) B.V., Terneuzen, The Netherlands.

reported [4]. Numerical taxonomy was applied to the classification of peak positions with the correlation coefficient as a criterion of similarity. Features with relatively high calculated information contents were selected from groups of highly correlated peak positions. A file of 5100 spectra of various compounds, taken from the ASTM Index, was uniquely coded to the extent of 97.7% of the spectra, by selecting 40 out of 140 features (binary-coded peak positions); 99% of 395 spectra of hydrocarbons, alcohols, ethers and carbonyl compounds were uniquely coded by using 27 selected features.

This paper describes a method for unique coding of infrared spectra contained in small files with a minimum number of features; the information content is used as the selection criterion. The features are selected out of 140 possible peak positions, expressed in terms of the wavelength or the wavenumber, at intervals of 0.1 μm or 25 cm^{-1}, respectively. The method (called TREE) is applied to five data sets, each containing spectra of one type of compound (saturated and unsaturated hydrocarbons, alcohols, ethers and aldehydes/ketones). The underlying considerations on information theory are elucidated in the following paragraphs.

Uncertainty and information

Identification procedures involve gathering information, which can be considered as reducing uncertainty. Therefore, decrease of uncertainty with regard to the identity of the unknown compound (in this case, a pure organic compound) is used as a measure of the information obtained from the analysis. If it is assumed that, before analysis is carried out, there are n possibilities with equal probabilities for the identity of the unknown compound X, the uncertainty before analysis, $H(X)$, is, according to Shannon [5], given by:

$$H(X) = \mathrm{ld}(n) \tag{1}$$

with ld = $\log_2$. In the ideal case, the uncertainty will be reduced to zero and the information obtained will equal $H(X)$. Thus $H(X)$ can also be regarded as the "missing" or "required" information (for unambiguous identification).

In this study, infrared spectra are binary coded, thus for each feature F_j (peak position; j = 1—140) only two signals are possible: Y_{0_j} and Y_{1_j}, representing peak absent and peak present, respectively. When a signal Y_{k_j} is measured and provided that there are no errors, the uncertainty about the unknown identity will be reduced to a value

$$H(X/Y_{k_j}) = \mathrm{ld}(n_{k_j}) \tag{2}$$

where n_{k_j} is the number of compounds giving a signal Y_{k_j} on analysis. The information gained is then defined by

$$I(X/Y_{k_j}) = -\Delta H(X/Y_{k_j}) = H(X) - H(X/Y_{k_j}) \tag{3}$$

Substitution of eqns. (1) and (2) into eqn. (3) gives

$$I(X/Y_{k_j}) = \mathrm{ld}(n) - \mathrm{ld}(n_{k_j}) = -\mathrm{ld}(n_{k_j}/n) \tag{4}$$

The information content $I(F_j)$ of a feature (F_j) will be defined as the "expected value" of the information to be obtained:

$$I(F_j) = I(X/Y_j) = \sum_{k_j = 0,\,1} p(Y_{k_j}) \,.\, I(X/Y_{k_j}) \tag{5}$$

where $p(Y_{k_j})$ is the probability of measuring a signal Y_{k_j}, when analysing the unknown compound. This probability can be calculated from

$$p(Y_{k_j}) = n_{k_j}/n \tag{6}$$

Substitution of eqns. (4) and (6) into eqn. (5) yields

$$I(F_j) = -\sum_{k = 0,\,1} (n_{k_j}/n)\,\mathrm{ld}(n_{k_j}/n) \tag{7}$$

If unknown and reference spectra are represented by series of features F_1, F_2,.....F_m, the information content of a retrieval procedure which searches and compares these coded spectra is given by

$$I(F_1, F_2,.....F_m) = -\sum_{k_1 = 0,\,1} \sum_{k_2 = 0,\,1} \cdots\cdots \sum_{k_m = 0,\,1} (n_{k_1, k_2,...k_m}/n)\,\mathrm{ld}(n_{k_1, k_2,...k_m}/n) \tag{8}$$

where $n_{k_1, k_2,...k_m}$ is the number of compounds in the reference file (= a priori possible identities) having a coded spectrum $(Y_{k_1}, Y_{k_2},.....Y_{k_m})$. If the features (F_j) are not correlated, then eqns. (9) and (10) are valid:

$$n_{k_1, k_2,...k_m}/n = (n_{k_1}/n)(n_{k_2}/n) \cdots\cdots (n_{k_m}/n) \tag{9}$$

$$I(F_1, F_2,.....F_m) = \sum_{j = 1}^{m} I(F_j) \tag{10}$$

In most cases, however, correlations between features do occur and the information content becomes:

$$I(F_1, F_2,.....F_m) < \sum_{j = 1}^{m} I(F_j) \tag{11}$$

Accordingly, unique coding of all spectra in a file is possible only if the information content $I(F_1, F_2,.....F_m)$ equals the a priori uncertainty $H(X)$ (eqn. 1).

Feature selection by the TREE procedure

Efficient coding involves unique coding of the spectra representing the compounds of interest, with a minimum number of features and consequently a minimum use of computer bits for the storage of the spectra. In principle the optimal combination of peak positions can be determined by calculating the information contents for all possible combinations of m out of 140 peak positions. The number N of these combinations is given by $N = \Sigma_{m = 1}^{140} \binom{140}{m}$. Although in practice iterative methods should greatly reduce this number, the remaining number of calculations will be too large to be handled by a computer. Therefore approximate methods must be used. Suitable approximations are provided by the selection procedure which we have called TREE.

In the TREE procedure, the information content of each of the 140 peak positions is calculated by using eqn. (7) and the one wih the highest information is selected. Next, the information contents of all 139 combinations of the first selected peak position with the remaining ones are calculated from eqn. (8). A second peak position is then selected, corresponding to the highest increment of the total amount of information. The third and following peak positions are selected in the same way. The procedure stops when the maximum increment of the information content equals zero, or when all spectra are uniquely coded. In this way correlations between peak positions are implicitly taken into account.

A flow chart of TREE is given in Fig. 1. The following example illustrates the TREE performance. Consider 6 compounds with their binary-coded spectra, each consisting of 5 features (see Table 1). From these data TREE selects 3 features as shown in Table 2.

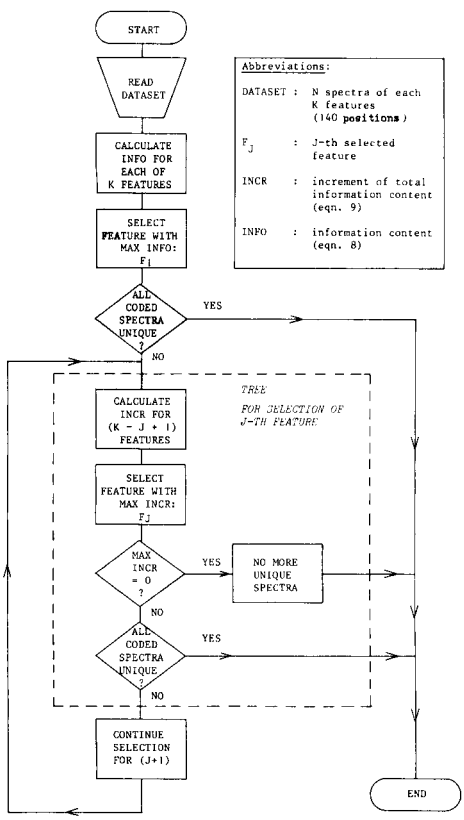


Fig. 1. Flow chart of the TREE feature selection procedure.

TABLE 1

Binary-coded five-feature spectra of 6 compounds (hypothetical)

Compound	Feature				
	F_1	F_2	F_3	F_4	F_5
X_1	0	1	0	1	0
X_2	1	0	1	0	0
X_3	1	1	1	1	0
X_4	0	0	0	0	0
X_5	1	0	1	1	1
X_6	0	0	1	1	0

TABLE 2

Selection of features by TREE from data of Table 1

Step	Calculated information contents (eqns. 7 and 8) (bits)	Maximum information content (bits)	Maximum increment of total information content (bits)	Selected features	All spectra uniquely coded?
1	$I(1) = 1.00$ $I(2) = 0.92$ $I(3) = 0.92$ $I(4) = 0.92$ $I(5) = 0.65$	1.00	1.00	1	NO
2	$I(1, 2) = 1.92$ $I(1, 3) = 1.46$ $I(1, 4) = 1.92$[a] $I(1, 5) = 1.46$	1.92	0.92	1, 2	NO
3	$I(1, 2, 3) = 2.25$ $I(1, 2, 4) = 2.58$ $I(1, 2, 5) = 2.25$	2.58	0.66	1, 2, 4	YES
4	Procedure stops				

[a]In the algorithm used this value is discarded, as it does not exceed $I(1, 2)$.

EXPERIMENTAL

Data sets

The six data sets used are listed in Table 3. The infrared spectra were recorded and encoded by a standardized procedure, as described previously [3]. The spectral range was coded by using the wavenumber (WN code) and the wavelength (WL code); each resulted in 140 peak positions at intervals of 25 cm^{-1} and 0.1 μm, respectively. For the intensity threshold (IT), five different values (3, 5, 10, 25 and 50%) were used; this eventually resulted in 60 work files.

TABLE 3

Data sets used

Set	Number of spectra	Type of compounds
CHS	48	Saturated hydrocarbons
CHU	112	Unsaturated hydrocarbons
CH	160	Hydrocarbons (CHS + CHU)
ALC	100	Alcohols
ETH	66	Ethers
CARB	41	Aldehydes/ketones

Computer program

The computer program for TREE was written in Fortran IV and required about 60000 (octal) words of 60 bits. For testing and running, the CDC 73/26 computer of the Academic Computer Center of this University (ACCU) was used. The computer time T (in seconds) of a run depends mainly on the number (n) of spectra in the file and, with the sorting algorithm used, approximately satisfies the equation $T = 4.0 \times 10^{-3} n^2 + 0.25 n$. The margin in T is about 15%, because of other variables. This equation was checked for n values smaller than 400. The second-order term is due to the presence of sorting procedures in the program, which are proportional to the square of the number of elements to be sorted. Thus the sorting procedure is the main limiting factor. Fast sorting routines, such as "heapsort" or "quicksort" [13], should extend the possibilities of the method to larger files. Although the computer time required for running the TREE program is substantial, it should be emphasized that in practice a feature selection procedure will be carried out only occasionally. The purpose is to provide not only a reliable but also a fast retrieval procedure, which is made possible by using a highly compressed code.

RESULTS AND DISCUSSION

The TREE procedure was applied to each of the 60 work files. An example of the results of the selection and coding procedures is illustrated in Fig. 2 for *trans*-4-octene. Out of the 140 binary-coded peak positions (WN code, IT = 10%, line b), TREE selected 12 features (line c), resulting in the reduced binary-coded spectrum given in line (d).

Table 4 shows the results of TREE for data set CHU (112 spectra of unsaturated hydrocarbons). When the WN code and an intensity threshold of 10% are used, the 112 spectra are all uniquely coded with 12 selected features. The total information content of the 12 selected features, calculated from eqn. (8), amounts to 6.81 bits, whereas addition of the information contents of the individual features, calculated from eqn. (7), yields 11.28 bits. The difference of 4.47 bits must be considered as due to correlation between the selected features. Table 5 shows the results obtained for all the data sets, WN- and WL-coded and with different intensity thresholds.

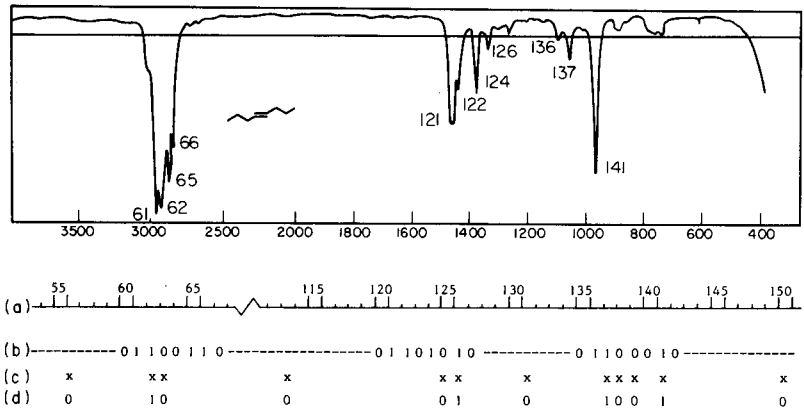


Fig. 2. The recorded infrared spectrum for *trans*-4-octene with the results of feature selection and coding procedures. Peak position numbers, corresponding to scale (a), are indicated on the spectrum. Line (b) shows the binary-coded peak positions (WN code) for the 10% intensity threshold indicated by the horizontal line on the spectrum; the dashes indicate zeros. Line (c) shows the features selected by TREE and line (d) the spectrum coded by TREE.

TABLE 4

Results of feature selection program TREE for the CHU data set (112 spectra of unsaturated hydrocarbons)

Information content $I(F_1.....F_m)$ from eqn. (8) and number of uniquely coded spectra (n) for increasing number of selected features (m)			Information content $I(F_j)$ from eqn. (7), code and interval of spectral feature and sequence number of selected features (F_j)			
m	$I(F_1.....F_m)$ (bits)	n	F_j	Code	Interval (cm^{-1})	$I(F_j)$ (bits)
1	1.00	0	1	63	2925—2901	1.00
2	1.99	0	2	125	1375—1351	0.99
3	2.94	0	3	139	1025—1001	0.99
4	3.87	0	4	138	1050—1026	0.99
5	4.77	6	5	56	3100—3076	0.94
6	5.49	15	6	141	975—951	0.92
7	6.05	42	7	131	1225—1201	0.92
8	6.40	70	8	137	1075—1051	0.97
9	6.61	90	9	150	750—726	0.90
10	6.72	102	10	126	1350—1326	0.95
11	6.77	108	11	62	2950—2926	1.00
12	6.81	112	12	113	1675—1651	0.71

Spectroscopic relevance of the selected features

It should be realized that the signals 1 (peak present) and 0 (peak absent) are equivalent in feature selection. At first sight, this chemometric approach differs from interpretation by an experienced spectroscopist, who looks

TABLE 5

Results of the feature selection program TREE. Percentages of uniquely coded reduced spectra (UCRS) and numbers of selected features (m) for all data sets, WN- and WL-coded with 5 different intensity thresholds.

Set	No. of spectra	Intensity threshold (%)	WN-code (cm^{-1})		WL-code (μm)	
			m	UCRS (%)	m	UCRS (%)
CHS	48	3	9	100	11	100
		5	9	96	11	96
		10	13	83	19	81
		25	9	50	9	23
		50	10	44	7	17
CHU	112	3	11	100	12	100
		5	11	100	13	100
		10	12	100	15	100
		25	15	100	22	96
		50	21	94	29	91
CH	160	3	12	100	16	100
		5	13	99	16	100
		10	17	95	25	93
		25	21	85	25	74
		50	23	78	33	66
ALC	100	3	10	100	13	100
		5	11	100	13	100
		10	10	100	12	100
		25	11	100	14	100
		50	13	100	17	100
ETH	66	3	9	100	10	100
		5	9	100	10	100
		10	10	100	10	100
		25	11	100	12	100
		50	11	94	13	94
CARB	41	3	8	100	9	100
		5	8	100	9	100
		10	8	100	9	100
		25	9	100	11	100
		50	12	100	14	95

primarily at the peaks present, though inferences are also made from the absence of peaks and spectra are interpreted on a retrieval basis from human memory as well as literature data. However, a distinction is usually made between features indicating functional groups and those in the fingerprint area ($< 1400\ cm^{-1}$); such discrimination is not involved in the TREE procedur Otherwise, it is noticeable that a considerable part of the selected features lies in the fingerprint area, as is illustrated by Table 4 for data set CHU.

Wavenumber and wavelength codes and intensity threshold

Data sets with WN-coded spectra require fewer features for unique coding than the corresponding WL-coded files (Table 5), because of the different numbers of peak positions per interval resulting from the use of the wavenumber or wavelength scale; particularly in the region around 3000 cm^{-1} more peak positions occur when the WN code is used.

The spectra of alcohols and aldehydes/ketones are all uniquely coded at an intensity threshold (IT) of 50% with the WN code. The complete data sets for unsaturated carbons (CHU) and ethers (ETH) are uniquely coded at an IT of 25%. The data sets for saturated hydrocarbons (CHS) and saturated and unsaturated hydrocarbons (CH) require an IT of 3% in order to obtain 100% uniquely coded spectra. These exceptional results are caused by the low number of intense peaks occurring in the spectra of saturated hydrocarbons.

Effects of errors in TREE-coded spectra on retrieval results

To illustrate the performance of the TREE procedure, a straightforward retrieval program was written to indicate the effect of errors in the coded spectra. The retrieval program was run on only one data set, viz. the file of 160 hydrocarbon spectra, which was considered to provide a fairly good illustration because the infrared spectra of hydrocarbons show considerable similarity. The number of "bit-mismatches" (mismatches of corresponding binary-coded features) between the TREE-coded "unknown" and the reference spectra was used as a match criterion; this was tested by using a logic XOR function. (The number of bit-mismatches can be considered as a measure of distance between two spectra). The maximum allowed number of bit-mismatches was taken as a variable, which was, at the start of the program, set to the number of selected features resulting from TREE.

Retrieval was done for eight infrared spectra which were not present in the reference data set (CH), but represented eight arbitrarily chosen compounds contained in the reference file. These "unknown" spectra were TREE-coded manually. Table 6 gives the results, showing the seven best matches for each "unknown" spectrum/compound. It appears that the correct answer is among the first 1—6 candidates, i.e. those with the lowest one or two numbers of bit-mismatches, with an average of 1.75. This indicates an error of about 5% in the coded unknown and/or reference spectra [2].

The first five candidates in each category indicate a certain classification power; work aimed at such classification is in progress. With respect to identification, the discriminating power of the procedure is diminished by minor errors. It should be stressed that this paper presents a method for the selection of the minimum number of spectral features, rather than an elaborate retrieval system of which feature selection is considered as a separate step. Therefore, none of the techniques described in the literature, e.g. windows [6] or weight factors [7], was employed; and extensive evaluation with a large test set of spectra based on criteria such as accuracy/precision/performance [8], recall/

TABLE 6

Results of retrieval program for 8 alternative spectra of reference compounds as unknowns. (CH data set (160 spectra of hydrocarbons); WN code; 10% intensity threshold; 17 selected features)

"Unknown"	7 best matches	No. of bit-mismatches	"Unknown"	7 best matches	No bit ma
1-Heptene	1-Heptene	1	*o*-Xylene	Toluene	3
	2-Octene	2		*o*-Xylene	3
	1-Heptyne	2		1,2-Dimethylcyclohexane	4
	2-Methyl-2-pentene	2		Isopropylbenzene	4
	1-Hexene	2		*p*-Xylene	4
	n-Pentane	3		3-Methyl-2,4-pentadiene	4
	Cyclohexane	3		1-Phenylisobutene	4
n-Octane	n-Octane	0	1-Heptyne	1-Heptyne	0
	n-Hexadecane	0		1,3-Octadiyne	1
	n-Hexane	1		2-Methyl-2-pentene	2
	Cyclohexadecane	1		1-Hexene	2
	n-Decane	1		n-Pentane	3
	Methylcyclopentane	2		Cyclohexane	3
	1,2-Dimethylcyclohexane (c + t)	2		Methylcyclopentane	3
t-Butyl-benzene	t-Butylbenzene	1	*m*-Xylene	1,3,5-Trimethylbenzene	2
	Methylcyclopentane	2		2,4,4-Trimethylpentane	3
	1,3,5-Trimethylbenzene	2		*m*-Xylene	3
	2,3,3-Trimethyl-1-butene	3		1,3-Dimethylcyclopentane (*c* + *t*)	3
	1,3-Dimethylcyclopentane (*c* + *t*)	3		1,3-Dimethylcyclopentane (*cis*)	3
	1,2-Dimethylcyclohexane	3		Pentaheptacontane	3
	1-Methyl-3-isopropylcyclopentane	3		Cyclohexadecane	4
Cyclo-hexane	Cyclohexane	2	Cyclo-hexadecane	1,3-Pentadiyne	3
	Decalin	2		Cyclohexadecane	4
	Benzene	2		Cyclooctane	4
	1,9-Decadiyne	2		t-1,3-Dimethylcyclohexane	4
	2,2-Dimethylhexane	3		1,3,5-Trimethylbenzene	4
	Cyclopentene	3		Diphenylmethane	4
	1-Pentyne	3		Adamantane	5

confusion [9], or recall/reliability [10, 11], was not considered to be relevant within the scope of this paper. The results of the straightforward retrieval program are presented mainly as an indication of the effect of errors in TREE-coded spectra on such results.

The effect of errors in binary-coded infrared spectra on information content and retrieval results has already been reported [2]. An analogous study with binary-coded mass spectra has shown [12] that the efficiency of a retrieval system is determined far more by the extent to which errors occur in the (coded) spectra involved than by the matching criterion used, even if the latter takes some account of errors. There is no doubt that any effort at evaluation, optimization or development of retrieval systems must deal explicitly with these errors.

Conclusions

The information content is a useful criterion for feature selection of binary-coded infrared spectra. Unique coding of all the reference spectra in each of the data sets considered is possible by using about 10% of the available 140 peak positions. As the method takes into account the a priori probabilities of the peak positions and their correlations, the selection obtained is specific for the file considered.

The influence of the intensity threshold is similar for the data sets of unsaturated hydrocarbons, alcohols, ethers and aldehydes/ketones; values of up to 25% scarcely affect the number of selected features required for 100% uniquely coded spectra. An intensity threshold as high as 50% can safely be applied to the spectra of alcohols and aldehydes/ketones. However, there are considerable difficulties for the files containing saturated hydrocarbons, and only an intensity threshold of 3% gives meaningful results. All the data support the conclusion that use of the wavenumber scale is preferable to the wavelength scale in the coding procedure. A significant part of the selected features is situated in the fingerprint area.

The results indicate that automated identification by retrieval of spectra reduced and coded by the TREE procedure is successful if there are few errors in the coded unknown and/or reference spectra.

The authors are indebted to Dr. J. H. van der Maas, Mr. T. Visser and Mr. B. Lutz for kind supply of the data sets and valuable discussions.

REFERENCES

1 R. S. McDonald, Anal. Chem., 50 (1978) 282R.
2 P. F. Dupuis and A. Dijkstra, Fresenius Z. Anal. Chem., 290 (1978) 357.
3 P. F. Dupuis, J. H. van der Maas and A. Dijkstra, Fresenius Z. Anal. Chem., 291 (1978) 27.
4 F. H. Heite, P. F. Dupuis, H. A. van 't Klooster and A. Dijkstra, Anal. Chim. Acta, 103 (1978) 313.
5 C. E. Shannon and W. Weaver, The Mathematical Theory of Communication, University of Illinois Press, Urbana, Ill., 1949.
6 Sirch III — Instruction Manual, American Society for Testing and Materials, 1969.
7 P. R. Naegeli and J. T. Clerc, Anal. Chem., 46 (1974) 739A.
8 D. S. Erley, Appl. Spectrosc., 25 (1971) 200.
9 S. L. Grotch, Anal. Chem., 45 (1973) 3.
10 G. M. Pesyna, R. Venkataraghavan, M. E. Dayringer and F. W. McLafferty, Anal. Chem., 48 (1976) 1362.
11 F. W. McLafferty, Anal. Chem., 49 (1977) 1443.
12 G. van Marlen, A. Dijkstra and H. A. van 't Klooster, Anal. Chem., in press.
13 D. E. Knuth, The Art of Computer Programming, Vol. 3. Sorting and Searching, Addison Wesley, Reading, Mass., 1973.

Analytica Chimica Acta, 112 (1979) 95
Computer Techniques and Optimization

Announcement

Computer-based Analytical Chemistry — 20th FECHEM Conference

Portorož, Yugoslavia, September 24—28, 1979

This FECHEM Conference will be held during 24—28th September, 1979, at Portorož, Yugoslavia. The main topics will be: fundamentals of computerization of analytical laboratories; principles and problems of computer-based instruments and networks; analytical information systems (storage, retrieval and computer-based interpretation of instrumental data); special topics in computer-based analytical procedures.

Further information from: Dr. Jure Zupan, Boris Kidrič Institute of Chemistry, P.O. Box 380, 61001 Ljubljana, Yugoslavia.

nouncing two new volumes in the series:

udies in Environmental Science

ume 4

)TENTIAL INDUSTRIAL CARCINOGENS AND UTAGENS

/RENCE FISHBEIN, *National Center for Toxicological Research, Jefferson, AR, U.S.A.*

s work provides detailed information on reported industrial carcinogens and mutagens l arranges them by structural categories in order to highlight their potential risks l to help predict the hazards of new agents considered for introduction into the ironment. It includes information on such topics as: the synthesis of these agents, ure of their trace impurities, environmental occurrence, chemical and biological ctivity, TLV's and MAC's, test systems, combination effects in chemical carcinogenesis, demiology, and risk-assessment.

s volume will therefore be of great interest to scientists involved in toxicology, cinogenesis and mutagenesis studies, genetics, and environmental health. In addition, /ill provide valuable assistance to officials working in public health and environmental tection agencies.

. 1979 x + 534 pages US $66.75/Dfl. 150.00 ISBN 0-444-41777-X

ume 2

IR POLLUTION REFERENCE MEASUREMENT ETHODS AND SYSTEMS

ceedings of the International Workshop, Bilthoven, December 12-16, 1977

anized by The National Institute of Public Health, Bilthoven, The Netherlands **sponsored by The World Health Organization. T. SCHNEIDER,** *The National titute of Public Health, The Netherlands,* **H. W. DE KONING,** *WHO, Geneva, itzerland, and* **L. J. BRASSER,** *TNO, The Netherlands* (Editors).

particularly valuable feature of this work is the presentation of recommendations l follow-up projects, including international projects that will contain and apply reference principles discussed during the workshop. The book will serve as an to-date review of the status of Air Pollution Reference Methods and Systems for hnicians involved in air pollution and will also provide useful background information those involved in air pollution activities in general. It is hoped that this work l stimulate greater international cooperation in the development of good reference tems.

c. 1978 vii + 168 pages US $35.50/Dfl. 80.00 ISBN 0-444-41764-8

ELSEVIER

P.O. Box 211,
1000 AE Amsterdam
The Netherlands

52 Vanderbilt Ave
New York, N.Y. 10017

7125

Dutch guilder price is definitive. US $ prices are subject to exchange rate fluctuations.

VOL. 112, No. 1 ANALYTICA CHIMICA ACTA March 15, 1979

(Computer Techniques and Optimization, Vol. 3, No. 1)

CONTENTS

A computerized system for determining secondary ion energy spectra
M. A. Rudat and G. H. Morrison (Ithaca, NY, U.S.A.) 1
A comparison of five pattern recognition methods based on the classification results from six real data bases
M. Sjöström (Umeå, Sweden) and B. R. Kowalski (Seattle, WA, U.S.A.) 11
Calculation of adsorption-related parameters from a.c. polarographic data: basis and computer programs
T. E. Cummings, M. Katz and P. J. Elving (Ann Arbor, MI, U.S.A.) 31
Computer automation of potentiometric analysis with ion-selective electrodes
J. Slanina, F. Bakker, J. J. Möls, J. E. Ordelman and A. G. M. Bruyn-Hes (Petten, The Netherlands) 45
Computer-assisted measurement of ion-diffusion coefficients by use of the Cottrell equation
A. Yamada, Y. Kato, T. Yoshikuni, Y. Tanaka and N. Tanaka (Sendai, Japan) 55
The learning machine in quantitative chemical analysis. Part 2. Potentiometric titrations of mixtures of three bases
M. Bos (Enschede, The Netherlands) 65
Leistungsfähige Matrixkorrektur in der Rontgenspektrometrie
R. Plesch und B. Thiele (Karlsruhe, B.R.D.) 75
Information theory applied to feature selection of binary-coded infrared spectra for automated interpretation by retrieval of referece data
P. F. Dupuis, P. Cleij, H. A. van 't Klooster and A. Dijkstra (Utrecht, The Netherlands) 83

Announcements 95

Printed in The Netherlands