Vol. 150 No. 1, June 1, 1983 ISSN 0003-2670

INTERNATIONAL CONFERENCE ON CHEMOMETRICS IN ANALYTICAL CHEMISTRY, PETTEN, THE NETHERLANDS, SEPTEMBER 15–17, 1982

ANALYTICA CHIMICA ACTA

International journal devoted to all branches of analytical chemistry

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 1983

	J	F	M	A	M	J	J	A	S	O	N	D
Analytica Chimica Acta	145	146	147	148	149	150/1 150/2	151/1	151/2	152	153	154	155

Scope. *Analytica Chimica Acta* publishes original papers, short communications, and reviews dealing with eve aspect of modern chemical analysis, both fundamental and applied.

Submission of Papers. Manuscripts (three copies) should be submitted as designated below for rapid and efficie handling:

Papers from the Americas to: Professor Harry L. Pardue, Department of Chemistry, Purdue University, West Lafayet IN 47907, U.S.A.

Papers from all other countries to: Dr. A. M. G. Macdonald, Department of Chemistry, The University, P.O. Box 3(Birmingham B15 2TT, England. Papers dealing particularly with computer techniques to: Professor J. T. Cle Universität Bern, Pharmazeutisches Institut, Baltzerstrasse 5, CH-3012 Bern, Switzerland.

Submission of an article is understood to imply that the article is original and unpublished and is not being consider for publication elsewhere. Upon acceptance of an article by the journal, authors resident in the U.S.A. will be asked transfer the copyright of the article to the publisher. This transfer will ensure the widest dissemination of informati under the U.S. Copyright Law.

Information for Authors. Papers in English, French and German are published. There are no page charg Manuscripts should conform in layout and style to the papers published in this Volume. Authors should cons Vol. 132, p. 239 for detailed information. Reprints of this information are available from the Editors or from: Elsev Editorial Services Ltd., Mayfield House, 256 Banbury Road, Oxford OX2 7DH (Great Britain).

Reprints. Fifty reprints will be supplied free of charge. Additional reprints (minimum 100) can be ordered. An orc 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 Science Publishers B.V., P.O. Box 211, 1000 A Amsterdam, The Netherlands.

Publication. *Analytica Chimica Acta* appears in 11 volumes in 1983. The subscription for 1983 (Vols. 145–155) Dfl. 1980.00 plus Dfl. 220.00 (postage) (total approx. U.S. $880.00). Journals are sent automatically by airmail to t U.S.A. and Canada at no extra cost and to Japan, Australia and New Zealand for a small additional postal charge. A earlier volumes (Vols. 1–144) except Vols. 23 and 28 are available at Dfl. 182.00 (U.S. $72.80), plus Dfl. 14.00 (U $5.60) postage and handling, per volume.

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

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

ANALYTICA CHIMICA ACTA
VOL. 150 (1983)

ANALYTICA CHIMICA ACTA

International journal devoted to all branches of analytical chemistry

ELSEVIER Amsterdam–Oxford–New York

0003-2670/83/$03.00

Printed in The Netherlands.

SPECIAL ISSUE

CHEMOMETRICS IN ANALYTICAL CHEMISTRY

Proceedings of a Conference held in Petten, The Netherlands, September 15–17, 1982

Analytica Chimica Acta, 150 (1983) 1—10
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

SOME USEFUL EXTENSIONS OF THE STANDARD MODEL FOR PROBABILISTIC SUPERVISED PATTERN RECOGNITION

J. D. F. HABBEMA

Institute of Public Health and Social Medicine, Medical Faculty, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam (The Netherlands)

(Received 17th September 1982)

SUMMARY

The standard probabilistic supervised pattern recognition model is described briefly. Extensions and adaptations of this standard model are proposed. These are discussed under five headings: the classification rule, measurement of performance, feature vector, training sets, the actual pattern class of an object. The aim is to suggest answers to questions and problems encountered in applications. Many of the extensions and adaptations are easy to incorporate in any probabilistic supervised pattern recognition approach.

Recognition of patterns is a fundamental human activity. In its scientific aspects, many disciplines, including psychology, biology, statistics, medicine, artificial intelligence and chemistry, contribute to the subject. Supervised pattern recognition or discriminant analysis has been widely studied in the area of electrical engineering, although some fundamental contributions have also been made by research statisticians. Unsupervised pattern recognition (or cluster analysis) owes much to biological research, where it is studied as numerical taxonomy. Some excellent textbooks appeared in the early seventies; those by Duda and Hart [1] and Meisel [2] are recommended. A statistical textbook has recently been published [3]. Early applications in analytical chemistry have been described by Jurs and Isenhour [4] and Kowalski [5]; Varmuza [6] has given a more recent overview of such applications. Solberg [7] has reviewed applications in clinical chemistry.

The present paper is restricted to the probabilistic branch of supervised pattern recognition. It may seem rather limited to discuss probabilistic supervised pattern recognition in isolation from other chemometric methods. In the main chemometric programs, ARTHUR [8] and SIMCA [9], supervised pattern recognition is only one, albeit important, procedure in the general statistical analysis of a problem. A justification is that many research problems appear in a relatively uncomplicated way as probabilistic pattern recognition problems, especially in clinical chemistry. Moreover, many of the suggestions made below are equally relevant to users of the more general approaches [8, 9], at least when their procedures for probabilistic supervised pattern recognition are reached; some of the problems discussed have been recognized and examined previously, especially in the SIMCA system [9].

0003-2670/83/$03.00

Notwithstanding the impressive literature on probabilistic supervised pattern recognition, the actual use of the methods in applications can be improved. The present paper is intended to stimulate such improvement. Many of the suggestions made below are quite easy to implement in existing approaches. Some of the suggestions are relevant in many applications, whereas others are only useful in special situations.

The standard probabilistic supervised pattern recognition model

The probabilistic supervised pattern recognition problem in its simplest formulation can be described as follows [1—3] (synonyms of basic terms are given in parentheses). Objects (elements, individuals, compounds, samples, patients, etc.) belong to one of a number of classes (pattern classes, populations, groups, categories). A feature vector (data, pattern, parameter, measurement, vector, etc.) is measured on each object. Feature vectors of the objects of training sets (learning, design, reference, or construction sets) are available for each of the classes. These training sets are used for constructing a classification rule (classifier, discriminant rule, decision rule, prediction rule, allocation rule). The rule is applied to an object in two steps. First, the feature vector is used to calculate a probability value for each class, indicating the likelihood that the object belongs to the class. Then the object is allocated to the pattern class with the highest probability. Some methods bypass these steps, and some do not even use a probability model, but these methods are not considered here. The quality (performance, discriminatory ability) of the classification rule is measured by the estimated rates of misclassification when the rule is used.

EXTENSIONS AND ADAPTATIONS OF THE STANDARD MODEL

Each application of supervised pattern recognition is based more or less on the use of the standard model. But each application also has its own peculiarities which may require extensions and adaptations. The extensions discussed derive from experience in problems of diagnosis, prognosis and therapy choice (including clinical chemistry). But many of the suggestions should be more generally useful, especially when the pattern recognition is done in the context of decision or consultation. The proposals are outlined according to the basic elements of the standard pattern recognition model; they are roughly ordered according to their decreasing general usefulness and increasing difficulty of implementation.

The classification rule

The form of the probabilistic classification rule depends primarily on the assumptions made in modelling the variability of the feature vectors within each of the classes. Readily available computer packages use, for example, assumptions of independence [8, 10], normal distributions [9, 11, 12], and kernel distributions [13]. The last-mentioned package [13] also contains

a program that is specially tailored for mixed discrete/continuous feature vectors. Indeed, much research work on classification rules is concerned with this statistical modelling aspect of the pattern recognition problem [1—3]. The performance of different approaches has been compared [14—17]. The suggestions made below are not concerned with this problem of deriving appropriate probabilistic classification rules. They are proposals for making more sophisticated use of the resulting probability vector with greater nuance than simple allocation of the object to the class with the greatest probability (or, in the context of decision theory with differential weighing of different types of error, to the pattern class with minimum expected loss).

Doubt classification. The idea of forced allocation to one of the pattern classes is not appropriate in many applications, and certainly not in most medical applications. The concept of classifying an object as doubtful is useful. An object is classified as a case of doubt when there is too much uncertainty, in some sense, to make a forced allocation to the most probable pattern class. The theory [18, 19] and implementation in computer programs [10, 13] have been described.

Classification as atypical. When the feature vector of an object does not fit well in any of the classes, the object should be classified as atypical or as an outlier. This may indicate a new unsuspected class, or a measurement error. A well-known outlier criterion for normally distributed feature vectors is the Mahalanobis distance [11, 12] which has been used also in a modified form [9]. A more general atypicality index has been applied in pattern recognition [20] and a related measure has been implemented [13].

Exclusion classification. It can be very useful to know for an object which classes can be excluded from further consideration. A class may be excluded because the probability assigned to a class is very small, or because the feature vector is atypical for the class, or both. The exclusion concept has been discussed [21].

The above three concepts may be combined into a generalized doubt-exclusion classification decision for an object, by indicating which classes are to be excluded from further consideration, and which classes are to be retained. The usual forced classification will arise within this generalized decision as the special case where all but one class have been excluded; an atypical classification is made when all pattern classes are excluded; and the "pure" doubt classification is made when none of the pattern classes is excluded.

Measurement of performance

The performance of probabilistic classification rules is usually measured by estimating the error rate (i.e., the percentage of misclassified cases). Discussion has traditionally centred around the problem of estimating the error rate: by means of independent evaluation sets or test sets, by resubstitution of the training objects, or by leave-current-object-out methods [2, 3]. The following suggestions are all concerned with introducing more

complete methods for evaluation of the classification rule than the error rate (or, in the context of decision theory, the expected loss).

Continuous scoring rules. The numerical values of the probabilities assigned to the classes are important in many applications, e.g., in medical decision-making. The above suggestions on classification were based on making more complete use of these probability values. The same applies to evaluation of the performance of the rule. Calculation of the error rate only is very crude exercise indeed, and therefore unsatisfactory, although it should be noted that the error rate is about the most sophisticated means of evaluation for a non-probabilistic classification rule.

It is suggested here that probabilistic rules should be evaluated by means of scoring rules which depend in a continuous way on the assigned probabilities. With continuous scoring rules, a higher probability value for the actual class always receives a better score (the error rate is a discontinuous 0—1 scoring rule in this terminology. Continuous scoring rules were initially developed in weather forecasting [22], and were later adapted and extended for use in medical diagnoses and prognoses [23, 24]. They were implemented in the pattern recognition program [10].

Reliability of the assigned probabilities. Continuous scoring rules are intended for measuring the discriminatory power of a classification rule. Users of probabilistic rules are sometimes also interested in the reliability of the assigned probabilities. This is quite distinct from the question of good discrimination for the following reasons. A probabilistic assertion which refers to a particular problem (e.g., "you can be 80% sure that this patient has a benign tumour") can be judged neither right nor wrong in isolation. Also, it is usually impossible to collect 100 cases with exactly the same feature vector and verify that, within sampling fluctuation, 80% of them have benign tumours. Consequently, the evaluator must look at the sample available to him in its entirety and hypothesize that whenever an event is assigned a probability p it will occur with frequency p. When this hypothesis of perfect reliability is taken as the reference point or null hypothesis, measures can be developed for detecting deviations from such perfect reliability [25]. Evaluation of reliability may further be used for calibrating the rule towards reliable probability values. Applications of the reliability concept to diagnostic problems have been described [14, 26].

Scoring rules based on variable loss. When the pattern recognition problem is placed in the context of decision theory, possible decisions (e.g., treatment in a medical context) should be specified and utilities or losses involved with wrong decisions quantified. However, in many pattern recognition problems, the decision situation is not uniquely specified. An example is a central electrocardiogram (ECG) interpretation system, which is utilized for patients to whom wrong decisions may have vastly divergent practical and therapeutic consequences. Such situations frequently occur in consulting problems in weather forecasting and clinical chemistry. The losses involved vary from one problem to the next, and "fixed" loss values are not always appropriate.

Therefore, scoring rules based on the so-called variable-loss decision theory have been developed [24, 27].

Leave-current-object-out. A brief general remark seems in order. Whenever the measurement of performance is based on the training sets, the probabilities for the pattern classes, and the evaluation measures based on these probabilities, should be calculated by using the leave-current-object-out device or modifications thereof. Various computer programs [9—11, 13] do this. Overoptimism about performance may easily result when resubstitution methods are used.

The feature vector

The extraction of useful features is the most crucial problem in pattern recognition. The choice of feature defines the limits to the possibilities of discriminating between classes. Feature extraction is, however, very problem-specific and falls outside the scope of this paper. The same applies to feature preprocessing, including the definition of new features, of calculating principal components, etc. Powerful general computer packages for feature preprocessing are available [11, 12] and the ARTHUR [8] and SIMCA [9] packages contain many useful extras for problems in analytical chemistry. A third problem that should be mentioned is measurement accuracy, and its impact on classification performance [8, 28]; a general review of precision requirements in quality control is available [29]. Measurement or coding errors will sometimes be detected by atypical or outlier criteria (see above). The discussion given in the following paragraph does not address these problems; it is assumed that a well-defined feature vector is available, and that it is used for further evaluation.

Missing values. In many pattern recognition applications, feature values are sometimes missing on objects. Missing values can occur for various reasons. The least difficult situation occurs when missing values occur at random, i.e., unrelated to the values of the other features, and also unrelated to the actual class of the object. Sometimes, the fact of being missing is informative about the actual class of the object (e.g., in diagnoses or prognoses, when ocular reactions cannot be examined because of swollen eyelids). In this case, "missing" should be regarded as a measurement value like any other, and evaluated as such. Extremely tricky is the confounding case; e.g., in a diagnostic context, a feature was not measured because the actual disease of the patient was considered as known or because some diseases were already excluded by the physician involved. The random and the informative situations can be resolved by several methods for handling missing data. When a confounding case is evaluated as if the missing values were random or informative, then misleading argument in a circle may result. The handling of missing data has been discussed [15, 30]. Procedures for missing data are implemented in many computer packages, usually for the random case.

Feature selection. Selection of a subset of informative features is frequently required because uninformative features produce noise, and because of

redundancy and/or cost considerations. The literature on selection of features is extensive [1—3]. The emphasis is usually on the selection procedure (e.g., exhaustive search, stepwise forward, stepwise backward, and assorted modification). Less attention has been paid to the criterion by which to decide which subset is best; criteria have long been used which are unrelated to classification performance. This has been criticized [31], and performance-related selection criteria have been implemented [10].

Sequential classification. Sequential measurement or use of features is theoretically even more efficient than selection of an informative subset. A fixed subset of features disregards the possibility of fewer features being needed on some objects, and more features on other objects.

A sequential choice of features (or groups of features) allows for such inter-object differences, by making use of the exact feature values already observed for each particular object, and by stopping the measurement of more features as soon as no more worthwhile improvement can be expected for the object at hand. Excellent research in sequential pattern recognition has been done by Fu, but there has been comparatively little progress since his 1968 book [32], perhaps because the framework of statistical pattern recognition is too rigid for the sequential approach. Real breakthroughs would be expected from a combined use of supervised pattern recognition with more general artificial intelligence methods. Technology based on theoretical research may ultimately result in automated sequential analyses in the clinical laboratory, at least for common problems such as the analysis of liver function.

Training sets

Training sets play a decisive role in pattern recognition: they constitute the sole body of information according to which the statistical distribution of feature vectors is modelled in each class. The training sets must be aselect or representative of the problem at hand. They should not contain labelling errors, i.e., they should not be wrongly classified; of course, most such labelling errors will be detected by the atypical criterion or another criterion. The impact of labelling errors on performance has been discussed [33]. Four other important problems concerning training sets will now be discussed.

Updating of training sets. No classification rule is meant for eternity. As soon as new objects with known pattern classes become available, the classification rule should be updated. Although the importance of updating is well recognized, and the problem is theoretically tractable, there are (nearly) no computer programs for pattern recognition that are especially tailored for continuous updating of classification rules.

Uncertain reference classification. Sometimes, training objects can be classified only with less than 100% certainty. This may occur in diagnostic problems, if a final diagnosis is far from certain. It would be bad policy to delete objects with uncertain reference classification from the training sets; serious biases and over-optimistic estimates of performance would result. The problem, not an easy one to solve satisfactorily, has been discussed [34].

Secular changes. Changes in the feature distribution of the classes are undoubtedly caused by changes in instruments, personnel, etc. Such changes may be anticipated in a general way by giving recent training objects more weight than old objects in deriving the classification rule. The rules for such weighting or discounting depend very much on the problem at hand: general guidelines are of little value. Approaches to detecting time-trends are described by Massart et al. [35]; this text should also be consulted for the general quantitative context of the problems discussed here.

Transfer of a classification rule. The careful collection of training objects and the derivation of a classification rule require a lot of time, money, and energy. It is therefore tempting to try to transfer a classification rule to other places where the same pattern recognition problem is studied. There are, however, many sources of bias involved in such a transfer. In general, calibration methods are needed for transferring feature values. And as soon as training objects are available in the new place, they should be used for checking and updating the transferred classification rule. The training objects from the new place should be given more weight than the original training objects. A statistical analysis of the transfer problem has been reported [36] and a medical application has been described [37].

The classes

The classes in pattern recognition problems are usually unordered. When there is ordering of the classes, or when the pattern class is more like a continuous response, regression methods may be more appropriate. But it must be remembered that standard models in regression analysis are based on much more restricted assumptions about the statistical relation between feature values and classes than pattern recognition. The evaluation of classification rules with ordered classes has been discussed [24].

In the usual supervised pattern recognition model, it is assumed that an object belongs exactly to one of the classes. This assumption is not always satisfied. Sometimes an object belongs to none of the classes; this possibility has been outlined above in considering atypical classification. The object may also belong to more than one of the classes (see below). Another (implicit) assumption is that each new object has the same a priori chance of belonging to the objects which are classified just before or just afterwards. The generality of this assumption can be challenged.

Multi-class membership. The possibility that an object belongs to more than one class arises in differential diagnostic problems, for example, where a patient may have several diseases. One obvious way of dealing with this problem is by defining new classes, consisting of objects belonging to some of the original classes simultaneously. This is a useful approach when the combinations are few and relatively infrequent. More general models for combinations of classes have been described [38, 39].

Compound classification. There may be situations where there is a dependence between the actual classes of consecutive objects. General models for

this problem have been studied under the subject heading of compound classification, especially for Markov dependences between the actual class of an object and its precessor [40] and in the context of chromosome classification for combinatorial dependencies between the actual classes of a group of objects [41].

DISCUSSION

Sixteen extensions and adaptations of the standard probabilistic supervised pattern recognition problem have been described, and about ten more topics have been discussed in introducing the five subgroups into which the sixteen subjects have been classified. Research workers involved in applications of probabilistic supervised pattern recognition can undoubtedly call attention to even more problems in application of the standard model. It would be tempting to try to recommend some pattern recognition computer programs because they can deal with a considerable number of the issues raised, and to discourage the use of other programs which cannot, but this would not be fair. Special-purpose programs [10, 13] may have paid more attention to topics that are especially relevant to supervised pattern recognition than general chemometric programs for analysing class structure [8, 9], or programs embedded in multipurpose statistical packages [11, 12]. Moreover, there are other aspects which are also relevant to choice of a computer program; costs, ease of implementation, user friendliness of the control language and error messages, quality of the manual, output and display options, etc., can be important. Recommendations should be based on a judgment of all these aspects together.

With respect to their implementation, the sixteen subjects can be assigned roughly to three groups. The first group consists of extensions of the use and evaluation of the probabilistic classification rule. These are the doubt and exclusion classification, the evaluation of the quality of the rule (measurement of performance), and the criteria used for feature selection. All these extensions are based only on the probabilities assigned to the classes, and apply in exactly the same way to all probabilistic supervised pattern recognition approaches. A series of published papers [23—25, 42] is a useful starting point for implementation.

A second group of extensions (classification as atypical, feature selection, leave-current-object-out, and updating of training sets) depends on the methods used for estimating the probability distributions in the classes. The assessment of atypicality or detection of outliers belongs to this group. In addition to the reference already mentioned, there is available a general statistical textbook on outlier detection [43]; the development of atypicality indices in a diagnostic context has been discussed [44]. The treatment of missing values also depends on the method used for estimating the probability distributions. Some methods (e.g., those based on the independence model and on normal distributions) allow relatively easy techniques to be used.

Other methods (e.g., kernel methods) present difficulties or call for excessive amounts of computation in handling missing data [15, 30, 45]. Also the leave-current-object-out device and its modifications and the continuous updating of training sets are much easier to implement for some methods of density estimation (especially kernel methods) than for other methods.

Five subjects remain. Although sometimes relevant, these are of less general usefulness than those discussed before, namely uncertain reference classification, secular changes, transfer of a classification rule, multi-class membership, and compound classification. There is another subject, sequential classification, that is of paramount importance, but requires fundamental changes in the standard model of probabilistic supervised pattern recognition. The generalized doubt-exclusion classification which was introduced after consideration of exclusion classification can be regarded as a useful concept when the classification rule is used as a step in an "informal" sequential approach.

Dr. J. Hilden and Dr. D. Coomans are thanked for their comments on an earlier draft of this paper.

REFERENCES

1 R. O. Duda and P. E. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973.
2 W. S. Meisel, Computer-oriented Approaches to Pattern Recognition, Academic Press, New York, 1972.
3 D. J. Hand, Discrimination and Classification, Wiley, New York, 1981.
4 P. C. Jurs and T. L. Isenhour, Chemical Applications of Pattern Recognition, Wiley, New York, 1975.
5 B. R. Kowalski, Anal. Chem., 47 (1975) 1152A.
6 K. Varmuza, Anal. Chim. Acta, 122 (1980) 227.
7 H. E. Solberg, CRC Crit. Rev. Clin. Lab. Sci., (1978) 209.
8 A. M. Harper, D. L. Duewer, B. R. Kowalski, and J. L. Fasching, in B. R. Kowalski (Ed.), Chemometrics: Theory and Application, American Chemical Society, Washington, 1977.
9 S. Wold and M. Sjöström, in B. R. Kowalski (Ed.), Chemometrics: Theory and Application, American Chemical Society, Washington, 1977.
10 J. D. F. Habbema and G. J. Gelpke, Comput. Programs Biomed., 13 (1981) 251.
11 W. J. Dixon (Ed.), BMDP, Univ. of California Press, Berkeley, CA, 1975.
12 N. H. Nie et al., Statistical Package for the Social Sciences, McGraw-Hill, New York, 1975.
13 J. Hermans and J. D. F. Habbema, ALLOC80 Manual, 1982, available from J. D. F. Habbema.
14 D. Coomans, Ph.D. Thesis, Vrije Universiteit, Brussel, 1982.
15 D. M. Titterington, G. D. Murray, L. S. Murray, D. J. Spiegelhalter, A. M. Skene, J. D. F. Habbema and G. J. Gelpke, J. R. Stat. Soc. A, 144 (1981) 145.
16 J. Remme, J. D. F. Habbema and J. Hermans, J. Stat. Comput. Simul., 11 (1980) 87.
17 M. Sjöström and B. R. Kowalski, Anal. Chim. Acta, 112 (1979) 11.
18 C. K. Chow, IEEE Trans. Inform. Theory, IT16 (1970) 41.
19 J. D. F. Habbema, J. Hermans and A. T. van der Burgt, Biometrika, 61 (1974) 313.
20 J. Aitchison, J. D. F. Habbema and J. W. Kay, Appl. Stat., 26 (1977) 15.

21 J. D. F. Habbema, J. Hilden and B. Bjerregaard, Meth. Inform. Med., 17 (1978) 217.
22 A. H. Murphy, in A. H. Murphy and D. W. Williamson (Eds.), Weather Forecasting and Weather Forecasts: Models, Systems and Users, National Center for Atmospheric Research, Boulder, CO, 1977.
23 J. Hilden, J. D. F. Habbema and B. Bjerregaard, Meth. Inform. Med., 17 (1978) 238.
24 J. D. F. Habbema and J. Hilden, Meth. Inform. Med., 20 (1981) 80.
25 J. Hilden, J. D. F. Habbema and B. Bjerregaard, Meth. Inform. Med., 17 (1978) 227.
26 J. Hilden and J. D. F. Habbema, in A. Alperovitch, F. T. de Dombal and F. Grémy (Eds.), Evaluation of Efficacy of Medical Action, North-Holland, Amsterdam, 1979.
27 J. Pearl, Int. J. Man—Machine Studies, 10 (1978) 175.
28 J. Aitchison and I. J. Lauder, Biometrika, 66 (1979) 475.
29 M. Hørder (Ed.), Assessing Quality Requirements in Clinical Chemistry, Scand. J. Clin. Lab. Invest., Suppl., 155 (1980).
30 J. Kittler, IEEE Journal on Computers and Digital Techniques, 1 (1978) 53.
31 J. D. F. Habbema and J. Hermans, Technometrics, 19 (1977) 487.
32 K. S. Fu, Sequential Methods in Pattern Recognition, Academic Press, New York, 1968.
33 P. A. Lachenbruch, Technometrics, 8 (1966) 657.
34 J. Aitchison and C. B. Begg, Biometrika, 63 (1976) 1.
35 D. L. Massart, A. Dijkstra and L. Kaufman, Evaluation and Optimization of Laboratory Methods and Analytical Procedures, Elsevier, Amsterdam, 1978.
36 J. Aitchison, Biometrika, 64 (1977) 461.
37 B. Bjerregaard, S. Brynitz, J. Holst-Christensen, E. Kalaja, J. Lund-Christensen, J. Hilden, F. T. de Dombal and J. C. Horrocks, in F. T. de Dombal and F. Grémy (Eds.), Decision Making and Medical Care: Can Information Science Help?, North-Holland, Amsterdam, 1976.
38 J. D. F. Habbema, in F. T. de Dombal and F. Grémy (Eds.), Decision Making and Medical Care: Can Information Science Help?, North-Holland, Amsterdam, 1976.
39 M. Ben-Bassat, IEEE Transactions on Systems, Man, and Cybernetics, 10 (1980) 331.
40 K. Abend, Proc. NEC, 22 (1966) 777.
41 J. D. F. Habbema, Biometrics, 35 (1979) 103.
42 J. D. F. Habbema, J. Hilden and B. Bjerregaard, Meth. Inform. Med., 20 (1981) 97.
43 V. Barnett and L. Lewis, Outliers in Statistical Data, Wiley, New York, 1978.
44 J. Hilden and B. Bjerregaard, in F. T. de Dombal and F. Grémy (Eds.), Decision Making and Medical Care: Can Information Science Help?, North-Holland, Amsterdam, 1976.
45 L. S. Chan, J. A. Gilman and O. J. Dunn, J. Am. Stat. Assoc., 71 (1976) 842.

Analytica Chimica Acta, 150 (1983) 11—21
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

PROBLEMS IN THE APPLICATION OF ARTIFICIAL INTELLIGENCE IN ANALYTICAL CHEMISTRY

Z. HIPPE

Department of Physical and Computer Chemistry, I. Łukasiewicz Technical University, 35-959 Rzeszów (Poland)

(Received 17th September 1982)

SUMMARY

Recent applications of artificial intelligence in analytical chemistry are discussed. Particular attention is given to the approaches based on problem-reduction search methods, algorithmic and heuristic, in elucidation of the structure of complex organic compounds by sophisticated computer program systems. Some likely future developments of the use of artificial intelligence in analytical chemistry are outlined.

In a book on problem solving by artificial intelligence, Nilsson [1] wrote: ... "many human activities such as solving puzzles, playing games, doing mathematics, and even driving a car, are said to demand 'intelligence'. If computers could perform such tasks as these, then presumably these computers (together with their programs) would possess some degree of artificial intelligence". Deeper consideration of this fragment suggests that a proper definition and understanding of artificial intelligence requires first an accurate definition of intelligence.

INTELLIGENCE AND ARTIFICIAL INTELLIGENCE

Intelligence may be identified with wise activity, ability of perception and giving opinions. The German psychologist W. Stern, the originator of the idea of the intelligence quotient, considered that intelligence is the power of being adaptable to new (unknown) tasks or life conditions. Accordingly, some forms of intelligence are observed even for animals [2]. It has not been proven whether intelligence is a set of various features giving a total effect of intelligent behaviour, or is a distinct, indivisible attribute. Indubitably, the difficulties in defining intelligence have caused disagreement about the meaning of artificial intelligence. This notion was created in the 1920s, probably in science fiction, but for 15 years it has been a fully fledged scientific expression. For the present purpose, the following working definitions may be acceptable: intelligence is the assembley of intellectual abilities, enabling acquired knowledge to be used efficiently in the solution of new (unknown), theoretical and/or practical problems; artificial intelligence is the application of mathematical techniques and/or mathematical logic in research

0003-2670/83/$03.00

devoted to all aspects of intelligence, done by technical or theoretical means [3]. (At one time, the term artificial intelligence covered even algorithmic programming languages.)

It should be emphasized that this definition encompasses not only the technical realization of artificial intelligence by the computer, but also gives the necessary margin for theoretical approaches. It is more general than the original definition given by Feigenbaum and Feldman [4] who stated that "in artificial intelligence, research and programming are directed toward the development of computer systems that exhibit behaviour which, if observed in humans, would be likely to lead one to say that the individual is showing intelligence". The most distinctive features of artificial intelligence seem to be the ability to use acquired knowledge (stored in the computer memory) and the ability to recognize and generalize problems.

The most common research areas in artificial intelligence reported by various authors [1, 5—7], are: (1) theorem proving, (2) game playing, (3) pattern recognition, (4) sophisticated information storage and retrieval systems, (5) problem solving, (6) processing sensory data, especially visual images and speech sounds, and (7) processing natural languages. All these general problems are recognized as requiring intellectual effort, hence the designation of artificial intelligence. All these items except (1) and (2) have proved useful in analytical chemistry, although most attention has been given to items (3)—(5). Theorem proving and game playing are beyond the scope of this paper; interesting references are available [1, 8, 9].

Pattern recognition

Pattern recognition methods were not initially considered as part of artificial intelligence. However, these methods are clearly semantic models (in terms of chemical information theory [10]) where the basic features are concepts and interconnections between them. Recently, pattern recognition methods have become among the most important applications of artificial intelligence in analytical chemistry. These methods have been discussed in comprehensive monographs, books and papers [11—17].

Information storage and retrieval systems

The importance of sophisticated information storage and retrieval systems in analytical chemistry is obvious. Interesting reviews on this subject have been published by several authors [18—22]. Another useful retrieval system, rarely cited in the analytical literature, is concerned with standard reference data banks of properties of the chemical elements, pure substances and mixtures [23]. A valuable summary of 53 interactive physicochemical numerical data systems, with computer programs for handling technical data, has been given by Hilsenrath [24].

It may not be immediately obvious why data banks and library search systems should be included in artificial intelligence. However, the word intelligence is derived from the Latin intellegere (to understand) and is con-

sidered in current artificial intelligence theory in terms of the organization and storage of large bodies of knowledge, i.e., facts and theories relevant to the given domain.

PROBLEM SOLVING

Problem solving (item 5 above) is certainly the most important and successful field of application of artificial intelligence in analytical chemistry. Of course, any computational task can be regarded as a problem to be solved. A more exact definition, that excludes routine computation, is needed for the present purpose. Analytical chemistry involves the solution of many various problems, e.g., selection of the most appropriate analytical method, automatic optimization of the main setting of an analytical instrument to obtain precise and reliable results in a reasonable time [25], correct numerical resolution of overlapped peaks [26], intelligent fraction collection in liquid chromatography [27]. Such problems can, however, be solved without any advanced artificial intelligence techniques. A survey of recent literature suggests that the weighty area of analytical problem solving is computer-assisted structure elucidation of complex molecules. Such elucidation may be regarded as an example of complicated qualitative analysis, but it is also to some extent quantitative analysis, considering the number of particular chemical groups (substructure) present in the molecule, etc. Besides its cognitive significance, structural identification may play an important role in checking and monitoring chemical experiments.

Computers were first applied to automate structural investigations at Stanford University in 1969 [28], in the interpretation of mass spectra. Since then, the original DENDRAL system has been under constant development [29—33]. The success of this approach has undoubtedly influenced many other research centers, where in turn other systems have been developed, e.g., the CASE system of Munk and co-workers [34, 35], the CHEMICS system of Sasaki et al. [36, 37], the SCANSPEC [38] and SEAC [39, 40] systems of Hippe and co-workers, and the STREC system of Gribov et al. [41, 42]. These systems have certain common features in terms of artificial intelligence.

An examination of problem-solving methods in artificial intelligence indicates that many involve the notion of trial-and-error search, i.e., problems are solved by searching for a solution in a space of possible solutions. In structure elucidation, it is usual to obtain a set of candidate structures (so-called informational isomers) that may be generated from substructures detected during spectral analysis of the molecule. However, various heuristics, e.g., a special data base dedicated to the system or even simple ad-hoc programming shortcuts can be applied to reduce the solution space, thus decreasing the machine time needed for identification of sub-units. Generally speaking, there are two cases: (i) an algorithmic search of the entire space of solutions, which always guarantees the solution, or (ii) an heuristic

search, which covers only those variants that indicate the highest probability of success. The latter approach does not always guarantee a solution. It is obvious that the heuristic search is closer to the operation of the human brain.

The overall strategy of computer-assisted elucidation of chemical structures consists of five formal steps [19].

(1) Correlation. This infers possible structural fragments, by means of one or any combination of spectral methods: m.s., ^{13}C-n.m.r., ^{1}H-n.m.r., i.r., Raman, u.v. (This order is related to decreasing entropy of the information of the given spectrum.)

(2) Consistency test. This test selects from the set of structural fragments found by correlation, those substructures that are internally consistent.

(3) Structural assembly. This involves combination of the partial structures (substructures) into a meaningful total structure (tentative or candidate structure).

(4) Spectrum prediction. Selected spectral features (or the whole spectrum) are predicted for the candidate structure.

(5) Spectra comparison. Predicted and experimental spectra are compared. If they agree, the candidate structure may be correct. Return to step (3) then generates another tentative structure; if no more candidate structures can be generated, the program ceases.

The versatility of such a strategy depends strongly on the efficiency of the algorithms applied in consecutive steps and on the general organization of the program system itself.

Correlation

Correlation (step 1) means the automatic interpretation (by the computer) of a given spectrum, in order to detect as precisely as possible the structural fragments that go to form the molecule. This interpretation is done by means of identification algorithms, which can be roughly classified into three groups (Table 1).

TABLE 1

Basic types of identification algorithms and their properties

Attribute	Type of algorithm		
	I AND/OR tree	II Network algorithm	III Matrix algorithm
Internal structure	Clear	Very complex	Fairly clear
Programming effort	Onerous	Extremely high	Slight
Machine time for identification of fragments	Short	Very short	Fairly short

The AND/OR tree identification algorithms (type I) are most popular, presumably because of their clear internal structure and quite efficient (short) machine time for identification of structural fragments. The main disadvantage (high programming effort) may be avoided by automatic program-writing procedures [43]. The process requires some human guidance, but only a small portion of the program (the final stage which consists of commands for printing of labels) need be written by the human programmer. Automatic program-writing procedures were later improved by introducing novel logic [44], in which the logical state of truth is numbered as 1, whereas falsehood is denoted by 2. In this way, a family of keys generated for a tree of specified order may be interpreted as labels (e.g., as line numbers in such languages as FORTRAN or BASIC).

The algorithms of type II (Table 1) resemble networks (AND/OR predicates with loops, etc.) with extended logical interconnections. These algorithms are analogous to human reasoning during interpretation of an unknown spectrum. The internal structure of such algorithms, which are usually extremely complex, is distinctly different for different spectroscopic techniques, but they are very fast. In matrix algorithms (type III), the correlation parameters (group frequencies, chemical shifts, etc.) are stored as arrays (in the sense of computer science). This type of identification algorithm is probably better adjusted for computer processing, although such algorithms usually act as filters for screening substructures when too many have been detected. Of course, this division of identification algorithms is one of convenience; several known examples may belong simultaneously to more than one type.

The effectiveness of identification algorithms is certainly a very complicated matter, and is connected with: (i) the selection of the type and number of identified substructures, (ii) the spectral parameters chosen for interpretation of a spectrum, (iii) the quality of the spectrum–structure correlations available, and (iv) the characteristic features of the spectroscopic technique used [45].

A detailed inspection of the relevant literature has shown that nothing has been published about the dependence of the versatility of identification on the number and/or type of substructures selected by the algorithm. Intuitively, it can be assumed that the number of substructures recognized should be as low as possible, yet sufficiently high to generate any structure within the scope of the system. The spectral parameters taken into account in the empirical interpretation of a spectrum are closely connected with the spectroscopic technique selected. Usually, at least two parameters are used to describe an absorption band: its location and intensity. But in many cases (see, e.g. [36, 38, 41]), additional parameters of spectral bands are also used, such as half-bandwidth, coupling constant. The crucial factor is the band intensity, because it strongly depends (particularly in infrared) on the sampling technique and concentration of the test substance. This problem may be bypassed by self-normalization [46] which relies on standardization of each band intensity (if Beer's law is obeyed) against the strongest band,

assigned to 100%. Another critical point is the quality of the accessible spectrum—structure correlations. Usually, a comparison of any correlation table shows how divergent is the available information about a particular group frequency (or chemical shift, etc.). It seems that only statistical elaboration of group frequencies (or chemical shifts) based on a large set of data may yield more reliable correlations. The characteristic features of the spectral methods are important. Thus, mass spectra are most informative, followed by ^{13}C-n.m.r., ^{1}H-n.m.r., infrared, Raman and u.v. spectra. Far-infrared spectroscopy suffers from a lack of simple empirical rules for extraction of all the valuable structural information hidden in the spectrum.

It is very difficult to evaluate and to compare the efficiency of all the types of identification algorithms mentioned, because of their fuzzy nature and particularly because they reflect the skill and experience of the algorithm designer. Recently, a more efficient self-adaptive i.r. matrix algorithm has been reported [47]. This enables 30 categories (substructures) to be detected with an efficiency of about 99.5% for correct recognition; the percentage of substructures recognized erroneously (in excess) was comparatively low (130% compared to 500—900% for other algorithms [45]). It should be emphasized that high efficiency of an identification algorithm is usually accompanied by a high percentage of substructures recognized in excess. This arises from the ambiguity of empirical interpretation of any type of spectrum (a given signal may well represent, with equal probability, many different substructures). However, it is necessary to detect all the substructures present in the molecule under examination, otherwise the proper structure cannot be generated in the further steps of the operation.

Consistency tests

The consistency test (step 2) usually takes advantage of cross-correlations of the results obtained by means of different spectral techniques. The results may confirm, complete, or mutually eliminate each other, depending on the power of identification of a given method (e.g., a substructure CH_3—C=O suggested by ^{1}H-n.m.r. must be eliminated if there is carbonyl frequency in the i.r. spectrum). When cross-correlations are made to one spectral technique only, the main problem, treated with unusual reticence by most authors, is that the consistency test must not use those spectral features that have been selected throughout the identification algorithms.

Besides cross-correlations, the consistency test may apply many ad-hoc programming shortcuts based on stored chemical theory, chemical stability, etc. Also, the results of identification can be influenced interactively by forcing the incorporation of substructures known to be present or the elimination of substructures (or heteroatoms) known to be absent. All those procedures have the sensible goal of diminishing as far as possible the set of substructures recognized, provided of course that none of the chemical groups actually present in the molecule investigated is eliminated. Decreasing the number of detected substructures has a significant influence on the subsequent structural assembly (see Table 2).

TABLE 2

Structural isomers generated by SCANGEN [48] for the empirical formula $C_6H_{10}O_3$ and different substructures specified by the user. The compound was $CH_3—CH_2—\overset{\overset{O}{\|}}{C}—O—\overset{\overset{O}{\|}}{C}—CH_2—CH_3$

Empirical formula
$C_6H_{10}O_3$

Substructures given
CH_2
CH_3
:C=O
O

Structural isomers
$CH_3—O—:C=O—:C=O—CH_2—CH_2—CH_3$
$CH_3—:C=O—O—:C=O—CH_2—CH_2—CH_3$
$CH_3—:C=O—:C=O—O—CH_2—CH_2—CH_3$
$CH_3—:C=O—:C=O—CH_2—O—CH_2—CH_3$
$CH_3—:C=O—:C=O—CH_2—CH_2—O—CH_3$
$CH_3—O—:C=O—CH_2—:C=O—CH_2—CH_3$
$CH_3—:C=O—O—CH_2—:C=O—CH_2—CH_3$
$CH_3—:C=O—CH_2—O—:C=O—CH_2—CH_3$
$CH_3—:C=O—CH_2—:C=O—O—CH_2—CH_3$
$CH_3—:C=O—CH_2—:C=O—CH_2—O—CH_3$
$CH_3—O—:C=O—CH_2—CH_2—:C=O—CH_3$
$CH_3—:C=O—O—CH_2—CH_2—:C=O—CH_3$
$CH_3—:C=O—CH_2—O—CH_2—:C=O—CH_3$
$CH_3—O—CH_2—:C=O—:C=O—CH_2—CH_3$
$CH_3—CH_2—O—:C=O—:C=O—CH_2—CH_3$
$CH_3—CH_2—:C=O—O—:C=O—CH_2—CH_3$
CYKLO— $O—O—:C=O—CH_2—CH_2—CH_2—CH_2—CH_2—$
CYKLO— $O—:C=O—O—CH_2—CH_2—CH_2—CH_2—CH_2—$
CYKLO— $O—:C=O—CH_2—O—CH_2—CH_2—CH_2—CH_2—$
CYKLO— $O—:C=O—CH_2—CH_2—O—CH_2—CH_2—CH_2—$
CYKLO— $O—:C=O—CH_2—CH_2—CH_2—O—CH_2—CH_2—$
CYKLO— $O—:C=O—CH_2—CH_2—CH_2—CH_2—O—CH_2—$
CYKLO— $:C=O—CH_2—O—O—CH_2—CH_2—CH_2—CH_2—$
CYKLO— $:C=O—CH_2—O—CH_2—O—CH_2—CH_2—CH_2—$
CYKLO— $:C=O—CH_2—O—CH_2—CH_2—O—CH_2—CH_2—$
CYKLO— $:C=O—CH_2—O—CH_2—CH_2—CH_2—O—CH_2—$
CYKLO— $:C=O—CH_2—CH_2—O—O—CH_2—CH_2—CH_2—$
CYKLO— $:C=O—CH_2—CH_2—O—CH_2—O—CH_2—CH_2—$

Empirical formula
$C_6H_{10}O_3$

Substructures given
CH_2
CH_3
CO—O
:C=O

Structural isomers
$CH_3—:C=O—CO—O—CH_2—CH_2—CH_3$
$CH_3—CO—O—:C=O—CH_2—CH_2—CH_3$
$CH_3—CO—O—CH_2—:C=O—CH_2—CH_3$
$CH_3—CO—O—CH_2—CH_2—:C=O—CH_3$
$CH_3—:C=O—CH_2—CO—O—CH_2—CH_3$
$CH_3—CH_2—:C=O—CO—O—CH_2—CH_3$

Empirical formula
$C_6H_{10}O_3$

Substructures given
CH_2
CH_3
CO—O—CO

Structural isomers
$CH_3—CO—O—CO—CH_2—CH_2—CH_3$
$CH_3—CH_2—CO—O—CO—CH_2—CH_3$

Structural assembly, spectrum prediction and comparison of spectra

Structural assembly relies on generating all possible structures compatible with the substructures detected, with the empirical formula, and with the stored theory of valency and chemical stability [30]. This operation may be organized in very different ways; one of the most sophisticated approaches was applied in the DENDRAL system, and now in the GENOA system [32]. The action of the structure-generating program may be simplified to a large extent by application of a novel mathematical operation, called vector division [40]. However, the importance of cross-correlation and reduction of the number of substructures applied in the assembly process should be stressed: for a molecule having the empirical formula $C_6H_{10}O_3$, the number of informational isomers generated was 28, 6 or 2, depending on the input data [48] (Table 2).

Spectrum prediction and spectra comparison may together be treated as another consistency test. Although not all known program systems for structure elucidation by artificial intelligence methods contain such checking algorithms, the basic task is to remove redundant candidate structures of low probability from the computer memory. The rule that must be obeyed is simple: the structure giving the best match between the prediction and the experimental spectrum is selected as the answer. This procedure of generating simulated spectra for candidate structures (or even macrofragments) and comparing them with input spectrum, was probably originated by Gribov et al. [41]. At the moment, the main limitation of this approach is connected with machine time (and so the cost of computations), but in the near future, especially when supercomputers become more widespread, it will become a necessary routine operation.

The above general strategy of computer-assisted structure elucidation may be somewhat changed with respect to the sequence of particular steps. It would be possible to use, as a first step, the information about the empirical formula to generate all substructures or macrofragments that match the chemical formula in respect of any atom. Then, in the next step, the identification algorithms would act as filters to screen out excessive substructures (as, e.g., in CHEMICS). It is very difficult to predict what influence this internal re-organization of the strategy would have on the versatility and efficiency of structure elucidation. It should be noted, however, that starting with generation of all plausible substructures matched to the empirical formula, would probably limit to some extent the selectivity of the strategy and enlarge the machine time.

In summary, the actual performance of program systems for structure elucidation by artificial intelligence at the present time is generally at the level of performance of a post-doctoral spectroscopist. Their performance is good, not because they know more than an experienced spectroscopist, but because they use most of the rules applied by a spectroscopist to solve structure problems, because they apply the same set of rules to every problem, and because they apply systematically the whole set of rules each time, without mistakes or loss of memory [30].

This problem, which may seem rather futuristic, has already found practical application in analytical chemistry. In 1979, it was revealed that in one American firm the A/D conversion of infrared spectra had been done routinely by photography of the spectrum image, with the camera coupled directly to the computer doing the actual conversion by an artificial intelligence program. Also, automatic perception of molecular formulae has been reported [22]. Processing of sensory data may therefore be the most progressive way of creating encyclopedic data bases of well-indexed facts.

Discussion of future developments in processing natural languages should be preceded by consideration of the level of maturity attained in some analytical problems solved by artificial intelligence methods. The current status in analytical chemistry is entirely typical of the situation in artificial intelligence as a whole, with the focus of attention moving from level I to level II (Table 3). Thus, practical applications of artificial intelligence in analytical chemistry were first dominated by an interest in behaviour that was ostensibly intelligent, e.g., simple problem solving in structure elucidation. At present, research in artificial intelligence is dominated by problems such as natural language processing. Thus, research in artificial intelligence seems to have regressed from activities that are impressively adult to those that can be mastered by a 3-year-old [7]. But it might be possible to program a computer so that it could understand what it read, thus stumbling directly into the true preserve of artificial intelligence, i.e., the design and implementation of computer programs which understand [30].

Some experimental computer programs are already known to read paragraphs of text (one sentence at a time) and generate memory representations of what was read. The questions are also processed and answered one at a

TABLE 3

Levels of artificial intelligence solutions in analytical chemistry, based on Michie and Buchanan [30]

Level	Action	Example
I	Transfer by analyst and programmer of algorithmic knowledge into program	Program interprets molecular spectra of pure substances
II	Generation by computer of descriptions and action scheme sequences to bridge gaps in book knowledge	Program interprets molecular spectra of mixtures, without prior separation
III	Acquisition by the computer of algorithmic knowledge by reading books	Program copes with new spectra/structure correlations by looking up chemistry text

time. One such program (SAM), designed at Yale University [7] started in 1975 and has been continuously expanded. However, chemical texts are difficult to process because they contain essentially three different kinds of information, numerical, conceptual, and structural; the order of difficulty increases in that order. Thus, level III (Table 3) remains a distant goal for spectroscopy. To attain it, the broad strategy must include the following steps: (1) perception, i.e., recognition of the problem with a proposed theoretical explanation; (2) design (i.e., development of a computer program system; (3) evaluation, i.e., checking whether the proper goal was achieved; and (4) learning, i.e., concluding why the system fails in some cases. This is a cyclic process: after step (4), it is necessary to go back to step (2) to incorporate what was learned in the evaluation in step (3). Thus, the computer is used as a research tool: without it, the truth can be guessed; with it, in artificial intelligence systems, the guessing remains but at least the errors are clear.

REFERENCES

1 N. J. Nilsson, Problem Solving in Artificial Intelligence, McGraw-Hill, New York, 1971, p. 1.
2 M. Hołyński, Informatyka, 13(4) (1978) 15.
3 A. Szewczyk, Informatyka, 14(3) (1979) 20.
4 E. A. Feigenbaum and J. Feldman (Eds.), Computer and Thought, McGraw-Hill, New York, 1963.
5 H. Gelernter, Proc. Int. Conf. Inform. Proc., UNESCO, Paris, 1959.
6 M. Minsky, in E. Feigenbaum and J. Feldman (Eds.), Computer and Thought, McGraw-Hill, New York, 1963, pp. 406—450.
7 W. G. Lehnert, The Process of Question Answering. A Computer Simulation of Cognition, L. Erlbaum Publ., Hillsdale, NJ, 1978.
8 J. R. Slagle, Artificial Intelligence. The Heuristic Programming Approach, McGraw-Hill, New York, 1971.
9 E. B. Hunt, Artificial Intelligence, Academic Press, New York, 1975.
10 Z. Hippe, Progr. Org. Coat., 5 (1977) 219.
11 B. R. Kowalski, in C. E. Klopfenstein and C. L. Wilkins (Eds.), Computers in Chemical and Biochemical Research, Academic Press, New York, 1974, p. 1.
12 P. C. Jurs and T. L. Isenhour, Chemical Applications of Pattern Recognition, Wiley, New York, 1975.
13 K. Varmuza, Pattern Recognition in Chemistry, Springer-Verlag, Berlin, 1980.
14 D. Coomans, D. L. Massart, I. Broeckaert and A. Tassin, Anal. Chim. Acta, 133 (1981) 215.
15 D. Coomans and D. L. Massart, Anal. Chim. Acta, 133 (1981) 225.
16 D. Coomans, M. Derde, D. L. Massart and I. Broeckaert, Anal. Chim. Acta, 133 (1981) 241.
17 S. Wold and E. Johansson, Anal. Chim. Acta, 133 (1981) 251.
18 J. T. Clerc and F. Erni, Fortschr. Chem. Forsch., 39 (1973) 91.
19 J. T. Clerc and H. Koenitzer, in Z. Hippe (Ed.), Data Processing in Chemistry, PWN-Elsevier, Warsaw, 1981, p. 151.
20 R. P. Young, in R. A. G. Carrington (Ed.), Computer for Spectroscopists, A. Hilger, London, 1974, p. 238.
21 J. Zupan, 2nd FECHEM Conference COBAC-II, Munich, 1982.
22 R. Hippe and Z. Hippe, Appl. Spectrosc. Rev. 16 (1980) 135.

23 J. Hilsenrath and B. Breen, OMNIDATA, An Interactive System for Data Retrieval, Statistical and Graphical Analysis and Data-Base Management, N.B.S. Handbook 125, Washington, DC, 1978.
24 J. Hilsenrath, Summary of On-Line or Interactive Physico-chemical Numerical Data Systems, N.B.S., Washington, DC, 1980.
25 Z. Hippe, A. Bierowska and T. Pietryga, Anal. Chim. Acta, 122 (1980) 279.
26 A. Bierowska and T. Pietryga, Proc. Int. Conf. Data Processing in Chemistry, Techn. University Publ., Rzeszów, 1979, p. 111.
27 L. Garpe, H. Lundin and J. Sjödahl, Int. Lab., 14(4) (1982) 62.
28 J. Lederberg, G. L. Sutherland, B. G. Buchanan, E. A. Feigenbaum, A. U. Robertson, A. M. Duffield and C. Djerassi, J. Am. Chem. Soc., 91 (1969) 2973.
29 A. B. Delfino and A. Buchs, Fortschr. Chem. Forsch., 39 (1973) 109.
30 D. Michie and G. B. Buchanan, in R. A. G. Carrington (Ed.), Computer for Spectroscopists, A. Hilger, London, 1974, p. 114.
31 D. H. Smith, L. M. Masinter and N. S. Sridharan, in W. Todd Wipke, S. R. Heller, R. J. Feldman and E. Hyde (Eds.), Computer Representation and Manipulation of Chemical Information, Wiley, New York, 1974, p. 287.
32 R. K. Lindsay, G. B. Buchanan, E. A. Feigenbaum and J. Lederberg, Applications of Artificial Intelligence for Organic Chemistry—The Dendral Project, McGraw-Hill, New York, 1980.
33 D. H. Smith, N. A. B. Gray, J. G. Nourse and C. W. Crandell, Anal. Chim. Acta, 133 (1981) 471.
34 C. A. Shelley, H. B. Woodruff, C. R. Snelling and M. E. Munk, in D. H. Smith (Ed.), Computer-Assisted Structure Elucidation, ACS Symp. Ser. 54, American Chemical Society, Washington, DC, 1977, p. 92.
35 C. A. Shelley and M. E. Munk, Anal. Chim. Acta, 133 (1981) 507.
36 T. Yamasaki, H. Abe, Y. Kudo and S. Sasaki, in D. H. Smith (Ed.), Computer-Assisted Structure Elucidation, ACS Symp. Ser. 54, American Chemical Society, Washington, DC, 1977, p. 108.
37 S. Sasaki, H. Abe, I. Fujiwara and T. Yamasaki, in Z. Hippe (Ed.), Data Processing in Chemistry, PWN-Elsevier, Warsaw, 1981, p. 186.
38 Z. Hippe, R. Hippe, J. Duliban and B. Dębska, Techn. Univ. Res. Project, Rzeszów, 1981.
39 B. Dębska, B. Guzowska-Świder and J. Duliban, Proc. Int. Conf. Data Processing in Chemistry, Techn. University Publ., Rzeszów, 1979, p. 71.
40 B. Dębska, J. Duliban, B. Guzowska-Świder and Z. Hippe, Anal. Chim. Acta, 133 (1981) 303.
41 L. A. Gribov, M. E. Elyashberg and V. V. Serov, Anal. Chim. Acta, 95 (1977) 75.
42 L. A. Gribov, Anal. Chim. Acta, 122 (1980) 249.
43 Z. Hippe and B. Dębska, Bull. Acad. Polon. Sci., Ser. Sci. Chim., 22 (1974) 551.
44 Z. Hippe, Proc. Int. Conf. Data Processing in Chemistry, Techn. University Publ., Rzeszów, 1979, p. 43.
45 S. Sasaki, H. Abe, I. Fujiwara, T. Yamasaki, Z. Hippe, B. Dębska, J. Duliban and B. Guzowska-Świder, Chemia Analityczna, in press.
46 Z. Hippe, R. Hippe and J. Duliban, Fresenius Z. Anal. Chem., 311 (1982) 440.
47 R. Hippe, Z. Hippe and J. Duliban, Techn. Univ. Publ., Rzeszów, 1982.
48 K. Kłoda, Techn. Univ., Master's thesis, Rzeszów, 1982.

Analytica Chimica Acta, 150 (1983) 23—36
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

REPRODUCIBILITY AS THE BASIS OF A SIMILARITY INDEX FOR CONTINUOUS VARIABLES IN STRAIGHTFORWARD LIBRARY SEARCH METHODS

P. CLEIJ and H. A. VAN 'T KLOOSTER*

State University of Utrecht, Laboratory for Analytical Chemistry, Research Group of Chemometrics, Croesestraat 77[a], 3522 AD Utrecht (The Netherlands)

J. C. VAN HOUWELINGEN

State University of Utrecht, Institute for Mathematical Statistics, Budapestlaan 6, 3584 CD Utrecht (The Netherlands)

(Received 17th September 1982)

SUMMARY

Straightforward library search methods, aiming at identification of (organic) compounds and based on comparison of analytical data for continuous variables, are considered with respect to a definition of the similarity of data. In the context used, the main object of such a search method is simply the retrieval of the reference date of the unknown compound. The proposed similarity index has the form of a significance probability (*P* value), a quantity originating from the general theory of hypothesis testing, and can be calculated from a statistical model of the reproducibility of the quantities used for comparison. The index is defined in general terms, but is intended for applicability to library search methods for different types, or combinations, of analytical data. It is primarily designed for use in situations where the application of very large data bases suffers from the generally low (interlaboratory) reproducibility of the data.

Computer-aided interpretation of spectroscopic and other results by retrieval of reference data (library or file search) is of growing importance for the identification of organic compounds. Recent reviews cover a variety of methods and systems, mainly involving mass, infrared and n.m.r. spectra and also combinations thereof [1—11].

In library search methods, two types can generally be distinguished [6, 7]; they are described here as straightforward and interpretative methods. The main object of straightforward library search methods is to retrieve the reference data of the unknown compound, which, for useful application, generally requires that the data base contains the relevant data. Well known examples of such systems are the Biemann/MIT search [6, 12] and the PBM system [6, 13] for mass spectra. An interpretative search is primarily designed for the retrieval of reference data of compounds similar to the unknown, and is especially useful in cases where the unknown has no reference in the data base. Examples of such systems are STIRS [6, 14] and SISCOM [7, 15] for

0003-2670/83/$03.00

mass spectra, and the Clerc search [16, 17] for ^{13}C-n.m.r. spectra. Of these systems, STIRS automatically interprets the retrieval results (for multiple data classes) in terms of the substructures present in the unknown molecule.

The difference between straightfoward and interpretative library search methods is not always sharp [6]. In general, a straightforward system can also, to some extent, retrieve compounds similar to the unknown, while an interpretative system is often required to perform as well as a straightforward method. In this paper, attention is restricted to straightforward library search methods.

One of the factors determining the performance of a straightforward library search method is the matching criterion applied. The influence of this factor is especially important when the data base consists of data from various sources, so that one has to consider the poor interlaboratory reproducibility of the data involved [13, 18—21]. This applies to all large (mostly commercially available) libraries of molecular spectra. With respect to the matching criterion, improvements of library search methods are to be expected from statistical concepts, formulated in probability theory, information theory and hypothesis testing [13, 19—23].

In this paper, a new approach to designing matching criteria is based on hypothesis testing and is restricted to data of continuous variables (e.g., retention indices, spectral peak heights, and chemical shifts). This approach starts from situations in which large data bases are used.

The theory of hypothesis testing, with respect to library search, is described in general terms, which implies that the proposed matching criterion should be universally applicable to different types of data, including mixed data from different analytical techniques. Moreover, this approach allows some uniformity in matching criteria for use in various library search systems. The reproducibility of the search data (unknown and reference data) plays an important role in the definition of the proposed matching criterion, which takes the form of a similarity index.

GENERAL CONCEPTS

Library search methods for large data bases are applied in various analytical situations, involving different amounts of prior information about the identity X of the unknown compound. If $X_1, X_2 \ldots X_N$ is defined as the set of all possible (organic) compounds, this prior information can be considered as a set of probabilities $p(X_1), p(X_2) \ldots p(X_N)$, where $p(X_j)$ is the probability, before evaluation (searching), of $X = X_j$ [24]. If, for instance, the sample is known to be a hydrocarbon, all probabilities of non-hydrocarbons are zero.

As it is impossible (or at least impractical) to design library search methods that take all possible specific situations into account, the starting point here is a situation in which all compounds have equal a priori probability, i.e., the situation in which $p(X_j) = 1/N$ for all j. This choice seems the best compromise for a library search system intended for use in very different situ-

ations. Further, it is assumed that the number of possible identities N is very large (indeed infinite, in principle). In this application of the theory of hypothesis testing, a reference compound R is considered; compound R is one of a series $R_1, R_2 \ldots, R_M$, for which reference data are available. As a real data base contains data for a relatively small part of all possible compounds, the set $[R_k]_{k=1\ldots M}$ should be considered as a subset of the set $[X_j]_{j=1\ldots N}$. The comparison, during the search, of X with R is considered as a comparison of values for a set of feature quantities, $q_1, q_2 \ldots q_n$. These feature quantities may relate to one-dimensional techniques (e.g., melting point or retention index) and/or to characteristics (two-dimensional) of molecular spectra (e.g., peak heights and peak positions). The important restriction in this approach is that the feature quantities should behave as continuous variables. This excludes, for instance, binary quantities, i.e., quantities that can only take the value 0 (feature absent) or 1 (feature present) and quantities such as the mass number of the most intense peak in a mass spectrum.

The actual values of the feature quantities are represented by $Q_1^x, Q_2^x \ldots Q_n^x$ for X (the unknown data) and by $Q_1^r, Q_2^r \ldots Q_n^r$ for R (the reference data). Both the unknown and reference data are considered to be the results of single measurements (i.e., no averaging). The distribution of q values for repeated measurements on the same compound is essential in a theory concerning similarity of unknown and reference data. This distribution, for compound X_j represented by the probability function $p_j[q_i]$, describes the reproducibility of the features. In the case of a library search with large data bases, one should think of a distribution of measurements on different types of instruments, performed under various experimental conditions (interlaboratory reproducibility). Another relevant probability function is

$$p_X[q_i] = N^{-1} \sum_{j=1}^{N} p_j[q_i] \qquad (1)$$

determining the probability of observing a combination of values $q_1, q_2 \ldots q_n$ for the respective features in case of a measurement of X. If it is assumed that the reference compounds form an arbitrary subset of the set of all compounds, the function $p_X[q_i]$ describes the distribution of q values in the data base [25, 26].

The function of a matching criterion

The main function of a matching criterion in a straightforward library search method is to enable the reference compounds to be separated into two classes: compounds that could be and that cannot be the unknown. It is, of course, generally not realistic to demand 100% certainty that the correct reference compound is never classified as "cannot be". A certain (small) risk of such wrong classification must be accepted. A straightforward library search method should therefore be considered primarily as a means of

reducing the number of possibilities for the identity of the unknown compound, rather than as a tool for unambiguous identification.

This approach requires a matching criterion, which allows the specification of a threshold value for separation of the two classes of reference compounds. The precise value of this threshold should depend on the acceptable risk of misclassifying the correct reference compound. A library search system based on this principle should retrieve all references of the "could be" class, i.e., all references with a matching value above (for a similarity index) or below (for a dissimilarity index) the threshold value, rather than, for instance, the 5 or 10 "best" matches (which also may be very bad matches).

LIBRARY SEARCH AS HYPOTHESIS TESTING

The general theory of hypothesis testing starts from some null hypothesis, H_0, the truth or falsity of which has to be established [27—29]. Every comparison of unknown and reference data during a straightforward library search should, in principle, provide information on whether or not the unknown and reference compound are identical. In terms of hypothesis testing, this is equivalent to establishing the truth or falsity of the (null) hypothesis that the unknown and reference compound are identical. This is written formally as H_0: $X = R$. A procedure for testing this hypothesis cannot be complete without defining an alternative hypothesis, H_1, which should be true if H_0 is false; thus H_1: $X \neq R$. In other words, under the alternative hypothesis, X is any of the compounds $X_1, X_2 \ldots X_N$, except R. Naturally, for hypothesis testing, a set of test quantities is required. For this purpose "difference quantities" are selected: $\Delta q_1, \Delta q_2 \ldots \Delta q_n$, defined by $\Delta q_i = q_i^x - q_i^r$ (for $i = 1 \ldots n$), where q_i^x and q_i^r refer to the values of the feature quantity q_i for X and R, respectively.

To proceed with a test, it is assumed that the probability density function $p_0[\Delta q_i]$, measuring the probability of observing a combination of Δq values $(\Delta q_1, \Delta q_2 \ldots \Delta q_n)$ when H_0 is true, is known. This function is called the reproducibility function, as it represents the distribution of the sums of two (measuring) errors, i.e., the error of q_i^x and the error of q_i^r. This description of reproducibility in terms of the difference between two measured values differs from the usual description in terms of single errors, i.e., the difference between a measured and a "true" value (compare with the function $p_j[q_i]$).

There are two (equivalent) ways to test the null hypothesis. The first is to define a critical region R_c in the space of difference quantities at a significance level α by requiring that

$$P(\Delta \in R_c | H_0) = \alpha \tag{2}$$

i.e., by requiring, when H_0 is true, that the probability of observing a point $\Delta = (\Delta q_1, \Delta q_2 \ldots \Delta q_n)$ inside R_c is equal to α. If Δ is found inside R_c, the null hypothesis is rejected at a significance level α, otherwise H_0 is accepted. In this context, "accepted" does not necessarily mean that H_0 is true but

that it cannot be rejected on the available evidence. The significance levels usually applied are 5, 1 and 0.1% [28]. Of course, the null hypothesis might be rejected even if it is true. The probability of such an error (an errror of the first kind) is equal to α. Also H_0 may be accepted even if it is actually false. This represents an error of the second kind (the term "error" is less appropriate here, in view of the above intepretation of "accepted"). The expression for its probability of occurrence

$$P(\Delta \in R_c|H_1) = \beta \tag{3}$$

requires the probability density function $p_1[\Delta q_i]$ that measures the probability of observing a combination of Δq values (Δq_1, Δq_2 . . . Δq_n) when H_1 is true. The quantity $(1 - \beta)$ is called the power of the test and represents the probability of rejecting the null hypothesis when it is false. The test is particularly useful when, for a given significance level α, the power of the test is maximized. In terms of library search, this corresponds to a situation in which the probability of a correct outcome of the test is maximal when the unknown and reference compound are not identical.

One factor that determines the power of the test is the exact shape of the critical region R_c. The requirement given by Eqn. (2) does not by any means uniquely determine R_c. The optimal shape of the critical region (i.e., the shape that maximizes the power of the test) is given by the theorem of Neyman and Pearson [27]. According to this theorem, a test of the null hypothesis is most powerful when the critical region R_c is that region in the space of difference quantities for which the following condition holds

$$(p_0[\Delta q_i])/(p_1[\Delta q_i]) \leqslant c \tag{4}$$

where c is a constant, depending on the level of significance.

The second way of testing the null hypothesis is by calculating a significance probability or P value [28, 29] for the observed point in the space of difference quantities (Δ_{obs}). Here it is assumed that the critical region $R_c(\alpha')$ is defined for all values of the significance level α', in such a way that two arbitrary critical regions, with different values of α', are nested and have boundaries with no points in common. The significance probability can then be defined as the α' value of the critical region with Δ_{obs} lying on the boundary (Fig. 1). If the calculated significance probability is lower than the level of significance α actually applied, the null hypothesis is rejected; otherwise it is accepted.

As with the first method of testing a null hypothesis, the probability of occurrence of an error of the first kind is α:

$$P(P \text{ value} < \alpha|H_0) = P(\Delta \in R_c(\alpha)|H_0) = \alpha \tag{5}$$

Analogously the probability of occurrence of an error of the second kind is

$$P(P \text{ value} > \alpha|H_1) = P(\Delta \in R_c(\alpha)|H_1) = \beta \tag{6}$$

If the critical regions $R_c(\alpha')$ are defined according to Neyman and Pearson,

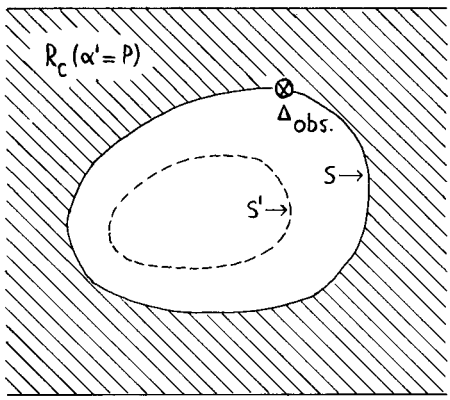


Fig. 1. Schematic representation of the (hyper)space of difference quantities with the observed point (Δ_{obs}), the critical region (shaded area, $R_c(\alpha' = P)$) for a significance level α', equal to the significance probability P and its boundary (S). The boundary of another critical region (with a higher value of α') is represented by S'.

the test obtained has maximum power for all values of the significance level. The corresponding P value is given by

$$P \text{ value} = \int_{\Delta q_1} \int_{\Delta q_2} \cdots \int_{\Delta q_n} p_0[\Delta q_i] \, d\Delta q_1 \, d\Delta q_2 \ldots d\Delta q_n \qquad (7)$$
$$R[\Delta Q_i]$$

where $R[\Delta Q_i]$ is the region in the space of difference quantities, for which the following condition holds

$$p_0[\Delta q_i]/p_1[\Delta q_i] \leqslant p_0[\Delta Q_i]/p_1[\Delta Q_i] \qquad (8)$$

where $[\Delta Q_i]$ is the value actually observed for the respective difference quantities.

The advantage of the second method of testing is that it yields a value for the significance probability, which provides more information about the truth or falsity of the null hypothesis, than just saying that Δ_{obs} is found inside or outside the critical region for a certain level of significance. For instance, with a P value of, say, 3% the possibility of the unknown and reference compound being different must normally be taken much more seriously than when P = 95%, although in both cases the null hypothesis might be accepted.

The above role of hypothesis testing in evaluating unknown and reference data strongly suggests the use of significance probability as the similarity index for straightforward library search systems. The required definition of the critical regions should obviously be based on the theorem of Neyman and Pearson. However, the resulting significance probability (see Eqn. 7) is difficult to apply, especially because the integration area, defined by Eqn. (8), depends on the shape of the probability distribution of the difference quantities under the alternative hypothesis, $p_1[\Delta q_i]$. Therefore this function and the integration area must be subjected to some approximations.

The function $p_1[\Delta q_i]$ can be expressed by an integral of the convolution type:

$$p_1[\Delta q_i] = \int_{q_1} \int_{q_2} \ldots \int_{q_n=-\infty}^{\infty} p_R[q_i]\, p_{X \neq R}[q_i + \Delta q_i]\, dq_1\, dq_2 \ldots dq_n \qquad (9)$$

Here the probability density function $p_{X \neq R}[q_i]$ measures, for any datum point of the unknown compound, the probability of observing a point $(q_1, q_2 \ldots q_n)$ in the space of feature quantities, on the assumption that $X \neq R$. This function can be expressed by

$$p_{X \neq R}[q_i] = [1/(N-1)]\, \{\sum_{j=1}^{N} p_j[q_i] - p_R[q_i]\} \qquad (10)$$

where $p_R[q_i]$ is the function $p_j[q_i]$ for $X_j = R$.

The assumption that the total number of a priori possible identities of the unknown compound is infinitely large in principle, implies that the approximation $p_{X \neq R}[q_i] \approx p_X[q_i]$ is permissible. This leads to

$$p_1[\Delta q_i] \approx \int_{q_1} \int_{q_2} \ldots \int_{q_n=-\infty}^{\infty} p_R[q_i]\, p_X[q_i + \Delta q_i]\, dq_1\, dq_2 \ldots dq_n \qquad (11)$$

This invokes the important assumption that the function $p_X[q_i]$ is relatively flat (constant) within regions of the q space with dimensions of the order of the measuring errors. In other words, it is assumed that the joint distribution of q values for all compounds is relatively broad (in all directions of the q space) compared to the distribution(s) of the measuring errors; and that this joint distribution is relatively smooth (i.e., without sharp peaks and dips). This assumption of flatness will fit reality to a large extent for a (single) feature such as a melting point or retention index, but less so for features such as the intensity of a mass spectral peak, because the (interlaboratory) reproducibilty of mass spectra is rather poor.

According to the assumption of flatness, a further approximation is allowed, i.e.,

$$p_1[\Delta q_i] \approx p_X[\mu_i^r + \Delta q_i] \qquad (12)$$

where μ_i^r is the expected value of q_i for reference compound R. From this approximation and the assumption of flatness, it follows that the probability density $p_1[\Delta q_i]$ is about constant in the region around the origin. This implies that, if the value of α' is not too small, the condition for the optimal shape of a critical region (see Eqn. (4)) can be approximated by $p_0[\Delta q_i] \leqslant c'$, where $c' = c\, p_X[\mu_i^r]$ is a constant that depends on the selected level of significance.

On the basis of this approximative condition of Neyman and Pearson for the optimal shape of a critical region, a significance probability can be defined and can be used as a similarity index in straightforward library search systems. This similarity index, represented here by S_I is given by

$$S_I = \int_{\Delta q_1} \int_{\Delta q_2} \ldots\ldots \int_{\Delta q_n} p_0[\Delta q_i]\, d\Delta q_1\, d\Delta q_2 \ldots d\Delta q_n \qquad (13)$$

$$R[\Delta Q_i]$$

where $R[\Delta Q_i]$ is the region in the space of difference quantities, for which $p_0[\Delta q_i] \leqslant p_0[\Delta Q_i]$, where $[\Delta Q_i]$ represents the actual values of the respective difference quantities.

The matching criterion thus formulated meets the requirements described above. If all references with an S_I value above a certain threshold value Th are retrieved, each search will yield a set of compounds that could be identical to the unknown (and for which H_0 is accepted at a significance level Th. All unretrieved reference compounds will belong to the class of compounds that cannot be identical to the unknown (H_0 rejected), when the accepted risk of the correct reference being misclassified (see Eqn. (5)) is

$$P(S_I < Th \,|\, H_0) = Th \tag{14}$$

Naturally, Th should have a low value, comparable to the values usually applied to a level of significance.

From its definition, S_I is determined by the reproducibility function. Hence, before the proposed similarity index can be applied, the reproducibility function must be developed for a particular kind of data. However, the reproducibility function was formulated for an arbitrary reference compound R, which implies essentially that such functions should be established for all (reference) compounds. In principle, the reproducibility function for compound R can be derived from the function $p_R[q_i]$ by means of the expression

$$p_0[\Delta q_i] \approx \int_{q_1} \int_{q_2} \ldots \int_{q_n=-\infty}^{\infty} p_R[q_i] p_R[q_i + \Delta q_i] \, dq_1 \, dq_2 \ldots dq_n \tag{15}$$

Unfortunately, often only one and sometimes not more than a few samples of the distribution $p_R[q_i]$ are available for the various reference compounds (from the data base and possibly from the literature or other sources). Thus the problem of finding adequate reproducibility functions cannot be solved in this way.

However, if some principal equivalent can be assumed for the reproducibility functions for the various (reference) compounds, then one joint reproducibility function can be formulated; in addition to some global parameters, this joint function may include compound-dependent parameters with (approximately) known values for every reference compound. A joint reproducibility function can be established, given a set of (reference) compounds for which only two measurements per compound are available. Each pair of measurements will provide one set of Δq values (one value for each Δq_i value, with $i = 1 \ldots n$). Fitting the sets of Δq values by an appropriate distribution function yields an estimate of the joint function with empirically determined global parameters. Pairs of measurements for a particular compound are easily obtained, when (at least) two alternative references (from different sources) for a set of compounds are available in the data base.

As an example, the case of a library search with a single retention index may be considered. It is assumed that the distributions of measuring errors

are of the same type for every compound. Here, the distribution is assumed to be normal, yet with a variance that is a linear function of the magnitude of the (true) retention index of the compound involved. The difference of the retention index for the unknown and the reference compound, ΔR_I, is chosen as the (only) difference quantity. The joint reproducibility function can then be written as

$$p_0(\Delta R_I) = ((2\pi)^{1/2}\ \sigma_k)^{-1}\ \exp(-\Delta R_I^2/2\sigma_k^2) \qquad (16)$$

where σ_k^2 is the variance of ΔR_I for reference compound R_k, given by $\sigma_k^2 = K_1 + K_2\ R_{I(k)}$. Here $R_{I(k)}$ is the compound-dependent parameter in the form of the true value of the retention index for reference compound R_k (a value to be approximated by the reference value); K_1 and K_2 are the global parameters which may be calculated from pairs of measured R_I values for a set of compounds with varying values of the (true) retention index.

With regard to molecular spectra, another important item concerns the often complex nature of an adequate reproducibility function. This complexity is caused particularly by systematic errors (e.g., normalization effects) and the use of distribution functions. It can easily lead to complex search algorithms, consuming excessive computer time. Three solutions to the problem have been found. The first starts from a not very accurate description of the (interlaboratory) reproducibility, on the basis of which is formulated a global model, including the main effects. The second solution implies a sharp preselection (possibly done in more than one step), resulting in a minimum number of references for which the similarity index is actually calculated. A third means of reducing computer time, is to allow some approximation in calculating the values of the similarity index.

PROPERTIES OF THE SIMILARITY INDEX

Formulation of the similarity index (S_I) as a probability implies that its value may vary from 0 to 1 (0 to 100%) similarity. Another property concerns the recall [30] for a library search method, i.e., the probability of retrieving the correct reference. From Eqn. (14), this recall for a search method when S_I is used as matching criterion and Th as the retrieval threshold, is given by Recall = $1 - Th$. Another implication of Eqn. (14) is that the probability distribution of S_I values, calculated for two (arbitrary) data sets of a same compound, is given by

$$p(S_I|H_0) = 1 \qquad (\text{for } 0 \leqslant S_I \leqslant 1) \qquad (17)$$

Hence, when two sets of q values are selected at random for a (large) number of compounds, and the S_I values of these data sets are calculated, a rectangular distribution of S_I values is obtained. For library search systems, this property means that the S_I values for correct matches are evenly spread over the range 0—100% similarity. Evidently, this property holds for any kind of analytical data, including mixed data resulting from different measuring techniques.

This allows a more or less uniform evaluation of matching values obtained in library search methods based on different kinds of data.

The distribution of S_I values for mismatches will have a (very) high density near zero with a (generally) decreasing density for increasing S_I values. For most mismatches, extremely low similarities, say $<0.01\%$, may be expected. The precise shape of the S_I distribution for mismatches, however, is determined by the kind of search data (feature quantities) used. For instance, a library search with a single retention index will produce considerably more mismatches with high S_I values than will a library search with mass spectra; the discriminating power of a single retention index is, of course, less than that of a mass spectrum.

A univariate example

For a library search with a univariate quantity (e.g., a retention index), the calculation of an S_I value is illustrated in Fig. 2. Only one feature quantity and one difference quantity are involved. Here Δq_1 is obtained by simple subtraction of two retention indices, normally distributed errors in the indices are assumed, and ΔQ_1 is defined as the actual value of Δq_1. Then the sum of the shaded areas in Fig. 2 is equal to the value of S_I.

A bivariate example

A second example shows the behaviour of S_I in a two-dimensional case. The situation considered is a spectrum characterized by two feature quantities

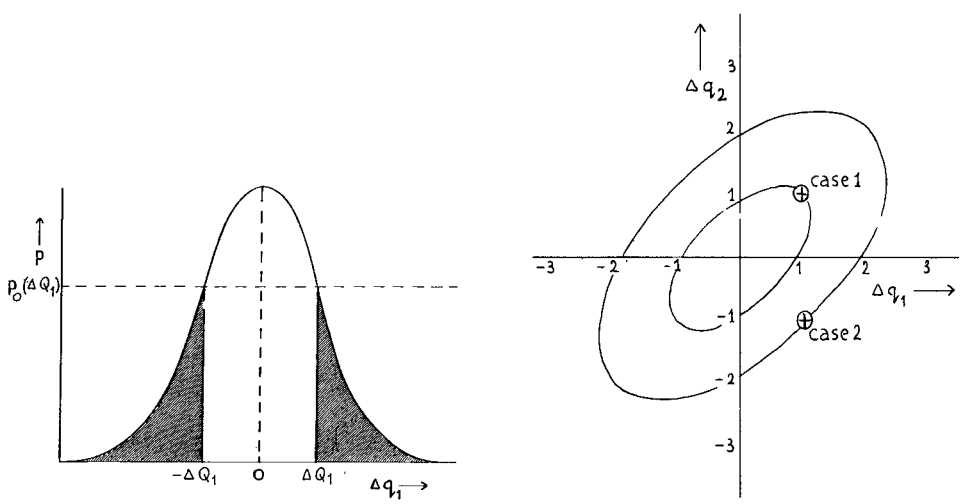


Fig. 2. Illustration of the similarity index S_I for a one-dimensional case (S_I = sum of shaded areas).

Fig. 3. Plot of a bivariate normal probability function with non-zero correlation for two difference quantities Δq_1 and Δq_2. The ellipses represent points with probability densities equal to the densities at point (1, 1) or point (1, −1).

(e.g., two relative intensities of mass spectral peaks at certain masses). The reproducibility function, $p_0(\Delta q_1, \Delta q_2)$, is assumed to be a bivariate normal probability function, with non-zero correlation caused by some kind of normalization effect (Fig. 3). This means that the (total) difference Δq can be considered as being partly systematic and partly random. Two cases, the first with $\Delta Q_1 = 1$ and $\Delta Q_2 = 1$ and the second with $\Delta Q_1 = 1$ and $\Delta Q_2 = -1$ are compared in Fig. 3.

When a simple match criterion is applied, e.g., the Euclidic distance to the origin, $(\Sigma\ \Delta^2 q_i)^{1/2}$, or the average absolute difference, $(\Sigma|\Delta q_i|)/n$, then both cases will be treated identically. However, as the differences (errors) are partly systematic, case 2 obviously represents a situation of lower similarity than case 1. When the proposed similarity index, S_I, is applied, this difference in similarity will be expressed in the calculated similarity values. The values of S_I are obtained by integrating $p_0(\Delta q_1, \Delta q_2)$ over the area for which the probability density is less than p_0 (1, 1), or p_0 (1, −1), which implies integration over an area outside the ellipses (Fig. 3). Obviously, for case 2 the lowest S_I value is calculated. This example illustrates how systematic errors, in combination with random errors, can be taken into account by applying the similarity index as the matching criterion.

CHOICE OF FEATURE QUANTITIES

Apart from the definition of a matching criterion, a vital factor in the performance of a library search system is the selection of feature quantities. This is especially relevant when numerous possibilities are available for selection, as in the case of molecular spectra. This problem is directly connected to the question of how to characterize a reference spectrum by a minimum number of computer bits retaining a maximum amount of relevant information, i.e., the problem of reduction and coding of spectral data (see, e.g., [12, 13, 31—36]). Although this subject is somewhat beyond the scope of this paper, there is a straightforward connection between feature selection and the theory of hypothesis testing and the similarity index S_I.

For optimal selectivity of the similarity index, it is obvious that the features chosen must provide a minimum probability of a false positive result, i.e., a mismatch with $S_I > Th$. In terms of hypothesis testing, this means that the probability of an error of the second kind, β, should be minimized (i.e., the power of the test must be maximized).

The exact expression of β for a library search with the similarity index S_I as the matching criterion, is

$$\beta = \int_{\Delta q_1} \int_{\Delta q_2} \cdots\cdots \int_{R(Th)} \int_{\Delta q_n} p_1[\Delta q_i]\ \mathrm{d}\Delta q_1\, \mathrm{d}\Delta q_2 \ldots \mathrm{d}\Delta q_n \qquad (18)$$

where $R(Th)$ is the region in the space of difference quantities, defined by $p_0[\Delta q_i] \leqslant p_0[\Delta Q_i]$, for which $S_I = Th$. From the approximation (12), β can be written as

$$\beta \approx \int_{\Delta q_1} \int_{\Delta q_2} \ldots\ldots \int_{R(Th)} \int_{\Delta q_n} p_X[\mu_i^r + \Delta q_i] \, d\Delta q_1 \, d\Delta q_2 \ldots d\Delta q_n \qquad (19)$$

According to the assumption of flatness, the probability density p_X is almost constant within the region $R(Th)$ for Th values that are not too small. For such values, therefore, β can be further approximated by

$$\beta \approx p_X[\mu_i^r] \cdot \int_{\Delta q_1} \int_{\Delta q_2} \ldots\ldots \int_{R(Th)} \int_{\Delta q_n} d\Delta q_1 \, d\Delta q_2 \ldots d\Delta q_n = p_X[\mu_i^r] \; V \qquad (20)$$

where V is the volume of the region $R(Th)$. By again using the assumption of flatness, the following approximation of β is obtained

$$\beta \approx p_X[Q_i^r] \; V \qquad (21)$$

This expression shows that β is determined by two factors: (1) the probability for the unknown compound (one of the compounds $[X_j]$) of measuring the reference values $Q_1^r, Q_2^r \ldots Q_n^r$ for the particular feature quantity, and (2) the volume V, which (apart from Th) depends on the shape of the reproducibility function (i.e., on the distribution of the measuring errors).

According to the first factor, the preferred feature quantities are those for which the reference data form a relatively unique (in the sense of a low probability of occurrence) combination of values. The second factor V can be minimized by selecting feature quantities with relatively small measurement errors. Moreover, the existence of correlations between errors for the different feature quantities favours a low value of β, as such correlations also reduce the volume V.

In this way some general, though rather obvious, indications are possible concerning the choice of appropriate features for the reduction and coding of spectra. However, a rigorous and quantitative application of the concept of minimizing the error of the second kind, is a difficult and rather intractable matter, especially because of correlations between the occurrences of features in spectra (influencing the function $p_X[q_i]$) and probably also because of correlations between the measurement errors for different features (influencing the volume V).

Applications

Based on the principles described above, library search algorithms for mass and ^{13}C-n.m.r. spectra were developed. A first version of what is calculated the "Mass Spectral Reproducibility-based Retrieval System" (MSRR) is operational; for the ^{13}C-n.m.r. library search system, a preliminary version has been completed. Results, obtained in testing and evaluation of both systems, are quite promising. Detailed descriptions of the algorithms developed and test results will be reported later from this laboratory.

CONCLUSIONS

A matching criterion in straightforward library search methods should permit classification of reference compounds as compounds that may be and compounds that cannot be identical to the unknown. With this starting point, the theory of hypothesis testing has proved to be very useful in quantifying the similarity of analytical data for continuous variables. Development of this theory has led to the main conclusion that a model for the reproducibility of the search data is essential in this quantification. It has resulted in a similarity index which evaluates the differences between unknown and reference data in comparison with the expected differences between sets of data measured for the same compound. The index is optimized for a situation with numerous a priori possibilities for the unknown compound, a situation which applies to the use of large data bases. When only a few a priori possibilities are relevant and a small data base is being used, the retrieval performance may deviate from the (theoretical) optimum. An adequate description of the reproducibility for a particular type of data generally involves a relatively complex (main) search algorithm. Successful application of the proposed similarity index therefore requires efficient search algorithms, including sharp preselections.

REFERENCES

1 B. R. Kowalski, Anal. Chem., 52 (1980) 113R.
2 I. E. Frank and B. R. Kowalski, Anal. Chem., 54 (1982) 232R.
3 R. E. Dessy and M. K. Starling, Anal. Chem., 51 (1979) 924A.
4 J. Zupan, Anal. Chim. Acta, 103 (1978) 273.
5 D. P. Martinson, Appl. Spectrosc., 35 (1981) 255.
6 F. W. McLafferty and R. Venkataraghavan, J. Chromatogr. Sci., 17 (1979) 24.
7 D. Henneberg, in A. Quayle (Ed.), Advances in Mass Spectrometry, Vol. 8B, Heyden, London, 1980, p. 1511.
8 S. R. Heller, A. McCormick and T. Sargent, in G. R. Waller and O. C. Dermer (Eds.), Biomedical Applications of Mass Spectrometry (First Suppl. Vol.), Wiley-Interscience, New York, 1980, p. 103.
9 J. R. Chapman, Computers in Mass Spectrometry, Academic Press, London, 1978, p. 101.
10 G. T. Rasmussen and T. L. Isenhour, Appl. Spectrosc., 33 (1979) 371.
11 R. S. McDonald, Anal. Chem., 54 (1982) 1250.
12 H. S. Hertz, R. A. Hites and K. Biemann, Anal. Chem., 43 (1971) 681.
13 G. M. Pesyna, R. Venkataraghavan, H. E. Dayringer and F. W. McLafferty, Anal. Chem., 48 (1976) 1362.
14 K. S. Haraki, R. Venkataraghavan and F. W. McLafferty, Anal. Chem., 52 (1981) 386.
15 H. Damen, D. Henneberg and B. Weimann, Anal. Chim. Acta, 103 (1978) 289.
16 R. S. Schwarzenbach, J. Meili, H. Könitzer and J. T. Clerc, Org. Magn. Reson., 8 (1976) 11.
17 D. L. Dalrymple, C. L. Wilkins, G. W. A. Milne and S. R. Heller, Org. Magn. Reson., 11 (1978) 535.
18 P. R. Naegeli and J. T. Clerc, Anal. Chem., 46 (1974) 739A.
19 S. L. Grotch, Anal. Chem., 47 (1975) 1285.

20 P. F. Dupuis and A. Dijkstra, Fresenius Z. Anal. Chem., 290 (1978) 357.
21 G. van Marlen, A. Dijkstra and H. A. van 't Klooster, Anal. Chem., 51 (1979) 420.
22 F. W. McLafferty, R. H. Hertel and R. D. Villwock, Org. Mass Spectrom., 9 (1974) 690.
23 F. Erni and J. T. Clerc, Helv. Chim. Acta, 55 (1972) 489.
24 P. Cleij and A. Dijkstra, Fresenius Z. Anal. Chem., 298 (1979) 97.
25 G. M. Pesyna, F. W. McLafferty, R. Venkataraghavan and H. E. Dayringer, Anal. Chem., 47 (1975) 1161.
26 P. F. Dupuis and A. Dijkstra, Anal. Chem., 47 (1975) 379.
27 S. Brandt, Statistical and Computational Methods in Data Analysis, North-Holland, Amsterdam, 1978, p. 112.
28 J. R. Green and D. Margerison, Statistical Treatment of Experimental Data, Elsevier, Amsterdam, 1978, p. 158.
29 H. L. Alder and E. B. Roessler, Introduction to Probability and Statistics, W. H. Freeman, San Francisco, 1977, p. 150.
30 F. W. McLafferty, Anal. Chem., 49 (1977) 1442.
31 S. L. Grotch, Anal. Chem., 42 (1970) 1214.
32 J. T. Clerc, F. Erni, C. Jost, J. Meili, P. Naegeli and B. Schwarzenbach, Fresenius Z. Anal. Chem., 264 (1973) 192.
33 G. van Marlen and A. Dijkstra, Anal. Chem., 48 (1976) 595.
34 E. G. de Jong, J. van Bekkum, H. A. van 't Klooster and J. Freudenthal, in N. R. Daly (Ed.), Advances in Mass Spectrometry, Vol. 7B, Heyden, London, 1978, p. 1091.
35 R. Buechli, J. T. Clerc, C. Jost, H. Koenitzer and D. Wegman, Anal. Chim. Acta, 103 (1978) 21.
36 P. F. Dupuis, P. Cleij, H. A. van 't Klooster and A. Dijkstra, Anal. Chim. Acta, 112 (1979) 83.

Analytica Chimica Acta, 150 (1983) 37—44
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

OPTIMIZATION OF SEARCH ALGORITHMS FOR A MASS SPECTRA LIBRARY

L. DOMOKOS*[a], D. HENNEBERG and B. WEIMANN
Max-Planck Institut für Kohlenforschung, Mülheim/Ruhr (German Federal Republic)
(Received 17th September 1982)

SUMMARY

The SISCOM mass spectra library search is mainly an interpretative system producing a "hit list" of similar spectra based on six comparison factors. This paper deals with extension of the system; the aim is exact identification (retrieval) of those reference spectra in the SISCOM hit list that correspond to the unknown compounds or components of the mixture. Thus, instead of a similarity measure, a decision (retrieval) function is needed to establish the identity of reference and unknown compounds by comparison of their spectra. To facilitate estimation of the weightings of the different variables in the retrieval function, pattern recognition algorithms were applied. Numerous statistical evaluations of three different library collections were made to check the quality of data bases and to derive appropriate variables for the retrieval function.

The SISCOM mass spectra library search system is part of a software system which has been developed at the Max-Planck Institut für Kohlenforschung [1]. The SISCOM program itself conducts the library search [2]. For a given unknown spectrum, it produces a hit list of similar compounds ordered according to a similarity measure. The latter is calculated from six comparison factors, such as the number of common characteristics in unknown and reference, the number and summed intensities of characteristics remaining in the unknown or reference, and correlation of common characteristics. These factors, derived from comparison of the encoded spectra, are also contained in the hit list in order to facilitate the recognition of different types of spectral similarities. Five years of routine application of SISCOM has proved its usefulness in qualitative evaluation of mass spectra.

Two basically different types of search system can be distinguished, interpretative and retrieval systems [3], which are designed to find structurally similar and identical compounds, respectively. Originally, SISCOM was designed especially for use in finding similar spectra; some basic elements necessary for a retrieval system, however, are already incorporated into SISCOM.

---

[a]On leave from the Technical University Budapest, Institute for General and Analytical Chemistry, Hungary.

0003-2670/83/$03.00

If exact reference spectra of compounds identical with the unknown or its components are available, then a good working interpretative system ranks them as the most similar ones. This is normally the case with SISCOM. In many cases, however, the references corresponding to the unknown are more or less distorted, thus the similarity search does not necessarily rank them as the most similar ones. The particular task of a retrieval search is to pick out, if possible, only the identical reference compounds. Thus all reference spectra not certainly identical to the unknown should be discarded, and the remaining, probably identical ones should be ranked according to an identity-orientated measure (retrieval function). If the corresponding reference spectra are missing, then of course the retrieval search might be confusing. To recognize such cases, the value of the retrieval function and that of the features can be helpful.

The target of the present work was to improve the retrieval capabilities of SISCOM by introducing a search algorithm tailored to the retrieval of identical compounds. This retrieval is intended for application to the ≥ 150 spectra of the hit list delivered by SISCOM. It involves selection and re-ranking of the list according to the retrieval function. To handle only the spectra of the hit list is not a real restriction, because the spectra of interest are practically always present somewhere in the list. The problem is to find the variables and the form of a satisfactory retrieval function.

DATA BASES

It is well known that there are several difficulties in designing a satisfactory retrieval system. Some of these are: (1) the unknown spectrum is incorrect; (2) the reference spectrum is not present in the library; (3) the reference spectrum is incorrect; (4) the reference spectrum in the library is correct, but is different because of the influence of measuring parameters (e.g., ion source temperature and pressure, pressure changes across g.c. peaks); (5) the unknown is a mixture. Because of such unreliabilities, great care is needed in applying even apparently good, identity-orientated comparison factors. Several earlier efforts have proved, for instance, that search algorithms comparing only the intensities do not provide satisfactory retrieval systems.

To establish the quality of data and the correctness or incorrectness of spectra, an extensive investigation was conducted on the applicability of quality factors reported earlier [4], and on some additional ones. As the data base, three larger mass spectra collections were used labelled E, C and L. These collections are also used with SISCOM. The collections are:

E, in-house library of ca. 6000 spectra;
C, EPA/NIH Mass Spectral Data Base of ca. 38 800 spectra;
L, collection of ca. 41 400 spectra with duplicates, compiled at Cornell.

Investigation of the correlation between the number of atoms in the molecule and the number of peaks in the spectrum, claimed to be a significant quality

factor [4], showed that it cannot be used reliably for quality control. Figure 1 represents a plot of number of peaks (nP) vs. number of atoms (nA). There is no need to emphasize that the correlation is far too bad to be used for any assessment of spectrum quality. Even for selected classes of similar compounds, no significant correlations could be observed.

Further statistics and a comparison of duplicate spectra indicated a considerable number of bad spectra in the collections. Table 1 shows the number of spectra with significant peaks beyond the molecular peak M and its isotopes, exactly beyond $m/z = M + 3 + 2$ (Cl + Br) + 0.5 (Si + S), as suggested earlier [4]. Table 2 contains the number of spectra in which the base peak does not have any isotopic peak, indicating either a bad dynamic range of intensity values or an incomplete spectrum. Spectra with cut-off intensities (Table 3) contain incorrect relative intensities leading to failures in encoded spectra. Table 4 summarizes how the lowest m/z is distributed, and shows that for many spectra the lower m/z range had not been stored. These few examples illustrate that a large number of spectra are of mediocre quality. Thus a retrieval system has to be constructed to handle even such unreliable references as far as possible.

TABLE 1

Number of spectra with peaks beyond the molecular peak and its isotopes with at least 1%, 5% and 30% intensity

Data base	Number of spectra		
	1%	5%	30%
E	232(4%)	33(0.6%)	1(0%)
C	6272(16%)	1636(4%)	339(0.9%)
L	6886(17%)	1673(4%)	272(0.6%)

TABLE 2

Base peak without isotope peak

Data base	E	C	L
Number of spectra	8 (0.1%)	4095 (10.6%)	3868 (9.3%)

TABLE 3

Number of spectra with cut-off intensities

Data base	E	C	L
Number of spectra	0 (0%)	1473 (3.9%)	1481 (3.6%)

TABLE 4

Distribution of the lowest m/z values

Data base	Percentages with lowest m/z greater than			
	29	40	50	60
E	1	0.2	0	0
C	53	33	13	8
L	54	31	13	9

FEATURE SELECTION

A crucial point in the construction of a retrieval function is the quantitative and qualitative selection of its variables (features). A feature is a function of the pair of spectra considered, unknown and reference, describing some kind of comparison between them. The six SISCOM comparison factors can be considered as possible features, but they were designed basically to select not only identical but also structurally similar compounds. Thus the retrieval function requires further, more identity-oriented features. One group of features was developed to describe the intensity relations of several relevant peaks of unknown and reference, respectively (e.g., molecular peaks, base peaks). Another group of features was developed to compensate for tilted spectra measured at the different positions of a g.c. peak. Further features are mathematical functions of the above-mentioned features (e.g., product, quotient).

The features, denoted by $x_1, x_2, \ldots, x_k$, can be represented by a pattern vector $\underline{x}$. The set of pattern vectors belonging to the same unknown U and to the various references is denoted by $\mathbf{R}^U$. $\mathbf{R}^U$ can be divided into two subsets, "good" ($\mathbf{G}^U$) and "wrong" ($\mathbf{W}^U$). $\mathbf{G}^U$ consists of a vector, or vectors, corresponding to the references which are identical with the unknown or with its components in admixture. $\mathbf{W}^U$ consists of vectors corresponding to spectra of substances that are not present in the unknown compound. Clearly, $\mathbf{R}^U = \mathbf{G}^U + \mathbf{W}^U$. The set of references corresponding to $\mathbf{R}^U$ is delivered here by the SISCOM hit list. It is the task of the retrieval process to find $\mathbf{G}^U$, $\mathbf{G}^U \subset \mathbf{R}^U$. $\mathbf{G}^U$ can be empty, if the unknown is not stored in the reference collection.

Finding $\mathbf{G}^U$ is equivalent to constructing a retrieval function $f(\underline{x})$, so that for all U

$$f(\underline{x}) > 0 \quad \text{if } \underline{x} \in \mathbf{G}^U; \qquad f(\underline{x}) < 0 \quad \text{if } \underline{x} \in \mathbf{W}^U \tag{1}$$

or the less restrictive condition

$$f(\underline{x}) > f(\underline{y}) \quad \text{for all } \underline{x} \in \mathbf{G}^U \text{ and } \underline{y} \in \mathbf{W}^U \tag{2}$$

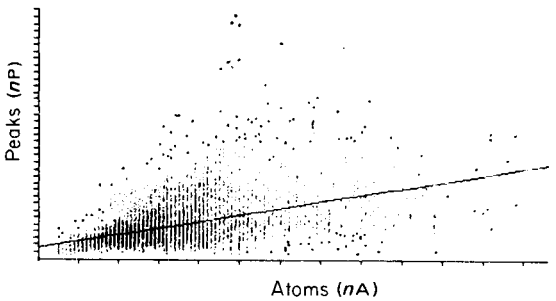


Fig. 1. Plot of the number of peaks vs. number of atoms. Correlation coefficient, $r = 0.46$. Library E, 5727 spectra, $nP = 0.87\,nA + 16.05; 1 < nP < 361$.

which means that $f(\underline{x})$ takes its highest value for pattern vectors belonging to G^U. In case (2), in contrast to case (1), $f(\underline{x})$ does not determine the classes exactly, but produces the patterns with highest $f(\underline{x})$ as candidates for G^U. Consequently, in the absence of the corresponding reference(s), the reference spectra with highest $f(\underline{x})$ must be wrong candidates, but the low highest value might indicate the absence of identical reference spectra. Unfortunately, the construction of $f(\underline{x})$ satisfying case (1) or (2) is difficult, if not impossible, especially with the quality of spectra collections described above. The most to be expected is a function $f(\underline{x})$ which satisfies case (1) or (2) with high probability.

A further simplification is to suppose that $f(\underline{x})$ is a linear function of $\underline{x}$, that is $f(\underline{x}) = \underline{c}'\underline{x}$, where $\underline{c}$ is a vector of weighting coefficients $c_1, c_2, \ldots, c_k$, called a decision vector. Of course, many nonlinear functions of $\underline{x}$ can be transformed to linear functions, by introducing new features for the nonlinear terms.

PATTERN RECOGNITION

With some skill and understanding of the meaning of the features x_i, fairly good coefficients c_i can be found empirically, but this is a time-consuming iterative process. In addition to such empirical considerations, pattern recognition algorithms can be applied for estimating the decision vector $\underline{c}$. The aim of the present work was to achieve a more objective estimate of $\underline{c}$, and to understand better the contribution of the features. The results should be helpful in creating a useful retrieval function. Pattern recognition has recently become a widely used method in chemometrics [5].

To develop and test the decision vector $\underline{c}$, a training set R was built up. Fifty known pure and mixed compounds were taken as unknowns; the spectra were carefully selected and recorded during earlier practice, representing interesting and typical difficulties. Most of these "problem spectra" were used earlier in the development of the former SISCOM. Training set R

consisted of pattern vectors which belonged to these 50 unknowns and to their corresponding reference spectra. For each unknown i ($i = 1, 2, \ldots, 50$), the first 50 spectra of the SISCOM hit list were retained as references for constructing the pattern set $\mathbf{R}^i$, so that the whole training set $\mathbf{R}$,

$$\mathbf{R} = \bigcup_{i=1}^{50} \mathbf{R}^i,$$

consisted of 2500 pattern vectors. Then $\mathbf{R}$ was divided into the two a priori known classes.

$$\mathbf{G} = \bigcup_{i=1}^{50} \mathbf{G}^i \quad \text{and} \quad \mathbf{W} = \bigcup_{i=1}^{50} \mathbf{W}^i = \mathbf{R} - \mathbf{G}$$

where $\mathbf{G}$ consisted of 92 vectors, $\mathbf{W}$ of the remaining 2408.

The pattern vectors were 30-dimensional, i.e., 30 features were computed, about 20 of them from the complete spectra or from the encoded ones, and the rest as some function of these. The features and their ability to separate the classes $\mathbf{G}$ and $\mathbf{W}$ were studied by looking at their distributions and by using them in the pattern recognition processes. If necessary, features were changed or omitted or new ones were included. For the pattern recognition algorithms, only a few of these features were selected, usually 5—14.

To select and transform the features, and to conduct the training procedure, an interactive program was written. The pattern recognition methods that were applied were the binary classification by error correction feedback and by TLU algorithms, slightly modified for this special problem.

As expected, the attempt to train a reasonable decision vector $\underline{c}$ satisfying condition (1) failed. Condition (1) turned out to be too strict, because the scores of the features were very different within the different $\mathbf{R}^i$ pattern sets.

To train according to the less rigorous condition (2), transformation of the training set was necessary. From condition (2), the following expression is valid:

$$\underline{c}'(\underline{x} - \underline{y}) > 0, \qquad \underline{x} \in \mathbf{G}^i, \underline{y} \in \mathbf{W}^i, i = 1, 2, \ldots \tag{3}$$

Thus, for all $\mathbf{R}^i$, a new set $\mathbf{D}$ of pattern vectors $\underline{z} = \underline{x} - \underline{y}$ was created for all pairs $\underline{x} \in \mathbf{G}^i$, $\underline{y} \in \mathbf{W}^i$, $i = 1, 2, \ldots$. Then training was needed for the vector $\underline{c}$ such that

$$\underline{c}'\underline{z} > 0 \qquad \text{for all } \underline{z} \in \mathbf{D} \tag{4}$$

The cost of this simplification is that the increased training set $\mathbf{D}$ requires more computing time. Condition (4) means the presence of only one category. Thus a hyperplane has to be found such that all $\underline{z} \in \mathbf{D}$ are on the same side of that plane. If an algorithm needs two categories, then $\mathbf{D}$ should be divided into two roughly equal parts $\mathbf{D}'$ and $\mathbf{D}''$ ($\mathbf{D} = \mathbf{D}' + \mathbf{D}''$) and the signs of all pattern vectors in $\mathbf{D}''$ must be changed. Thus the one-category case (4) is equivalent to the following two-category case

$$\underline{c}'\underline{z} > 0 \quad \text{if} \quad \underline{z} \in D'; \quad \underline{c}'\underline{z} < 0 \quad \text{if} \quad \underline{z} \in D'' \tag{5}$$

The problem is then to find a hyperplane ($\underline{c}$), if it exists, to separate the two categories represented by the vectors of D′ and D″, respectively. To train such a vector $\underline{c}$, error correction feedback and TLU methods were used.

The attempt to obtain a decision vector $\underline{c}$ satisfying condition (4) or (5) with each $\underline{z} \in D$, without further restriction of D, failed. Fairly good vectors $\underline{c}$, however, could be obtained, which classified 98—99.5% of pattern vectors $\underline{z}$ correctly. Depending on the starting value of $\underline{c}$, on the number of iterations, and on the applied algorithm, the attained $\underline{c}$ vectors were more or less different but they gave almost the same performance in separation.

RESULTS AND CONCLUSIONS

Based on the results obtained by pattern recognition and on spectroscopic experiences, three different $\underline{c}$ decision vectors were selected. They were applied for several weeks and tested as retrieval functions within the routine evaluation of mass spectra.

Each of the three corresponding retrieval functions classified more than 95% of the training set R correctly; as mentioned above, this set comprised mainly spectra with complications. The test set included more than 200 spectra, which were selected partly from earlier measurements and partly from daily measurements. In almost all of these cases, the retrieval was successful. In the case of mixtures, the system gave successful results; usually, not just one but several components were identified. There were also several spectra where the retrieval failed. More thorough investigation of these cases showed that these failures were caused by strongly deflected, tilted, noisy or bad reference spectra and by difficulties in identifying all the components of mixtures. As a necessary consequence of this retrieval function, suited to detect also components of mixtures, it can happen that a reference spectrum is covered completely by the unknown spectrum, and so is identified as a component even if it is a very different compound from the unknown.

These first efforts have shown the possibility of developing a successful retrieval system in this way, although a system with 100% success cannot be expected. Further improvements can be expected from extension of the training set by new representative spectra and from the introduction of several new or transformed features, providing a useful aid for the final assessment of spectra.

The programs were written in Fortran and run on the VAX computer system [6] of the Institute.

REFERENCES

1 D. Henneberg, K. Casper, B. Weimann and E. Ziegler, Advances in Mass Spectrometry, Vol. 6., Applied Science Publishers, Barking, Essex, 1970, p. 477.
2 H. Damen, D. Henneberg and B. Weimann, Anal. Chim. Acta, 103 (1978) 208.

3 F. W. McLafferty, Pure Appl. Chem., 50 (1978) 197.
4 D. D. Speck, R. Venkataraghavan and F. W. McLafferty, Org. Mass Spectrom., 13 (1978) 209.
5 K. Varmuza, Pattern Recognition in Chemistry, Springer-Verlag, Berlin, 1980.
6 E. Ziegler, Anal. Chim. Acta, 144 (1982) 1.

Analytica Chimica Acta, 150 (1983) 45—52
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

EVALUATION OF FIELD-DESORPTION AND FAST ATOM-BOMBARDMENT MASS SPECTROMETRIC PROFILES BY PATTERN RECOGNITION TECHNIQUES

J. VAN DER GREEF*, A. C. TAS, J. BOUWMAN, M. C. TEN NOEVER DE BRAUW and W. H. P. SCHREURS

CIVO Institutes TNO, Utrechtseweg 48, P.O. Box 360, 3700 AJ Zeist (The Netherlands)

(Received 17th September 1982)

SUMMARY

Field-desorption and fast atom-bombardment mass spectrometry are promising techniques for generating profiles of complex mixture of non-volatile compounds. This was concluded from results obtained with urine and wine samples. Pattern recognition techniques, such as non-linear mapping, K-nearest neighbours and principal component analysis, in combination with Fisher weighting, can be applied to optimize the experiments and to evaluate the data sets.

Field-desorption mass spectrometry (f.d.m.s.) is a well established technique for the analysis of non-volatile, thermally labile compounds. Because it is regarded as a "soft ionization" technique [1], f.d.m.s. is very suitable for application to complex mixtures, as has been shown in identifying metabolites of xenobiotics in urine samples of rats [2]. In that study, even minor metabolites could be traced by applying feature-selection methods. However, the great variations occurring in human urine samples necessitate a large number of measurements to establish significant small differences between different classes. This again leads to large data sets, especially if broad mass ranges are studied. Despite the problems involved, it has been reported that field ionization can be used in diagnosing viral infections and infectious hepatitis by measuring the volatile compounds in human urine extracts [3, 4]. In the present study, the application of f.d.m.s. to profile evaluation for non-volatile compounds was investigated. Because profile evaluation demands a high sample throughput for complex matrices, the recently introduced technique of fast atom-bombardment mass spectrometry (f.a.b.m.s.) [5] was also tested; it is simple and can easily be automated.

EXPERIMENTAL

Samples

Urine samples. Duplicate urine samples from eight men and six women were desalted as follows. The urine samples were adjusted with hydrochloric

0003-2670/83/$03.00

acid to pH 2. Then about 500 mg of XAD-4 resin (200—1000 mesh) was added and after 15 min the solution was removed by filtration (pore size 40). The resin was washed five times with twice-distilled water after which it was added to 1 ml of methanol. About 1 μl of the resulting solution was used for f.d.m.s. analysis. This desalting procedure also removed most of the urea and creatinine present in the urine samples.

Wine samples. A portion (1 ml) was taken from 11 Bordeaux and 11 Rhône wines. Water, alcohol and other volatiles were removed by storing the samples at 90°C in an oven overnight. The resulting slurry was mixed with glycerol and subjected to f.a.b.m.s.

Field-desorption mass spectrometry

The f.d. mass spectra were obtained with a Varian MAT-731 double-focussing mass spectrometer equipped with a combined e.i./f.i./f.d. source. Emitters were prepared by the method of Rabrenovic et al. [6] by using indene as activation substance. The needle length was typically 40—80 μm and the syringe technique [7] was used to load the samples on the emitter. The source temperature was kept at 80°C and the total ion current was controlled manually in such a way that a constant total ion current was obtained. The mass range 80—500 was scanned repetitively at a scan speed of 8 s/decade. The mass scale was calibrated by using perfluorotributylamine as the standard in the field-ionization mode. For each sample, a urine profile was obtained by summation of all spectra (40—80) during one f.d.m.s. experiment. An averaged urine profile was obtained by adding duplicate urine profiles.

Fast atom-bombardment mass spectrometry

A Finnigan MAT-212 double-focussing mass spectrometer with reversed geometry was equipped with a combined c.i./e.i. source. The direct probe inlet system was removed and a fast argon atom gun (Ion Tech) was mounted; the other inlet system was used for the introduction of a Finnigan MAT f.a.b. probe. The fast argon beam was generated from argon ions with a kinetic energy of 5—8 kV by resonant charge exchange. The source pressure was typically 4×10^{-3} Pa; glycerol was used as matrix on the target probe. The mass range 80—800 was scanned repetitively at a rate of 1.5 s/decade.

Data acquisition and processing

The f.d. and f.a.b. data were accumulated with a Varian SS-100 and a Finnigan MAT-SS200 system, respectively. The latter comprises a PDP-11/34 computer, two RK07 disc drives and two CDC hawk disc units, a magnetic tape unit, a plotter, a printer and three terminals.

The standard Finnigan MAT software was used for data acquisition and spectra averaging. Fortran programs were written for spectra normalization and feature selection. The results were evaluated with the pattern recognition package ARTHUR (University of Seattle, Laboratory for Chemometrics), implemented in the Finnigan MAT-SS200 data system.

RESULTS

Field-desorption mass spectrometry

Feature selection. Two categories were defined: class 1 (8 duplicate urine samples of men) and class 2 (6 duplicate urine samples of women). Large individual variations are observed in the f.d. profiles. Therefore, conclusions cannot be drawn on the basis of the averaged f.d. profiles of both classes alone (Fig. 1). It is also evident that a large number of strongly correlated peaks are present in the f.d. profiles. There is not only a high correlation between the isotope peaks but also between peaks and their cluster ions. For example, m/z = 114 (probably the $[M + H]^+$ peak of creatinine) and m/z = 180 (probably the $[M + H]^+$ peak of hippuric acid) form a cluster ion at m/z = 293 $[113 + 179 + H]^+$ and m/z = 114 and m/z = 265 form a cluster at m/z = 378 $[113 + 264 + H]^+$. In order to select from this data set the peaks that are best suited to distinguish the two classes, two procedures were tested. For each mass, a Fisher index [8] and a Coomans index [8] were calculated by means of the following formulae:

$$\text{Fisher index} = (m1i - m2i)^2/(v1i + v2i)$$

$$\text{Coomans index} = |m1i - m2i|/[(v1i)^{1/2} + (v2i)^{1/2}]$$

In these formulae $m1i$ and $m2i$ are the means of mass i in the two classes, $v1i$ and $v2i$ represent the variances of mass i in the two classes.

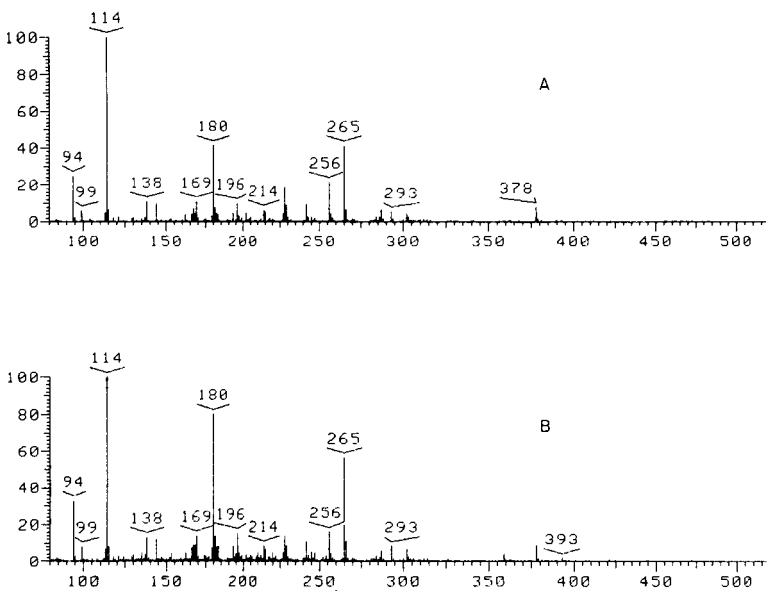


Fig. 1. Average f.d.m.s. profiles of urine samples from men (A) and women (B).

As expected with a binary classification problem, no significant difference was observed between the two methods used. Only minor differences are found if more than 40 masses are selected. The masses causing these differences are of no importance to the differentiation of the classes because of their extremely low weight factor. Fisher weighting was chosen as the selection procedure in this study.

Normalization of the spectra. The above-mentioned feature selection tests were applied after normalization of the spectra by correction for the total ion current. A better method, however, was found to be normalization with a total ion current calculated by discarding the larger peaks, because variances in these peaks are not spread over all the other ones. First, a raw total ion current was calculated. Second, a refined total ion current was obtained by omitting peaks above a certain percentage level of the raw total ion current. For several levels, the sum of the Fisher ratios (FISHTOT) of the ten peaks with the highest Fisher ratio was used as criterion for the normalization procedure. The relation between FISHTOT and the correction level is given in Fig. 2.

Above 20% of the raw total ion current all peaks are included in the correction. Therefore, no difference is found in the level range 20—100%. A clear optimum is observed around 4% and in that case peaks at *m/z* 94, 114, 138, 180 and 265 are excluded by the calculation of the refined total ion current. At levels below 4% too many peaks are omitted and consequently FISHTOT decreases again. At levels above 4%, the variances of the larger peaks are spread over all others which results again in a decrease of the FISHTOT value.

Pattern recognition

Before the software package ARTHUR was applied, the raw field-desorption data were transformed in the following way: masses were rounded off to the nearest integer, peak intensities were normalized (see above) and 50 features were selected from the 420 by Fisher weighting. Then by means of the ARTHUR program, autoscaling [7] was done followed by selection of

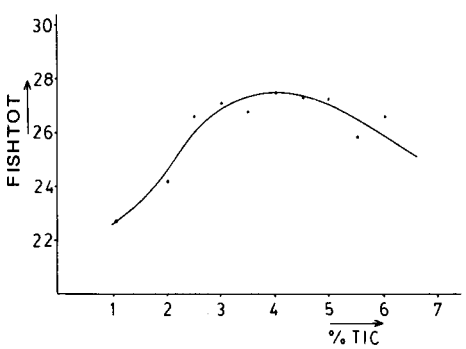


Fig. 2. Relation between the sum of 10 Fisher weights (FISHTOT) and the level above which peaks are excluded from the normalization factor.

the ten masses with the highest weighting by using the GRAB subroutine. This subroutine selects features, making use of Fisher weighting and decorrelation. Nonlinear mapping was used to get an insight into the reproducibility of the f.d.m.s. profiles. First, a non-linear map was made, by using all duplicate measurements; then the duplicate measurements were averaged and again subjected to non-linear mapping. The results given in Fig. 3 show that the two classes (urine samples from men and women) can be distinguished in the data set. Furthermore, improvement is observed after the averaging procedure. This means that the variation in the f.d. mass spectra, although much smaller than the individual variation within the samples, still contributes significantly to the total variation in this experiment.

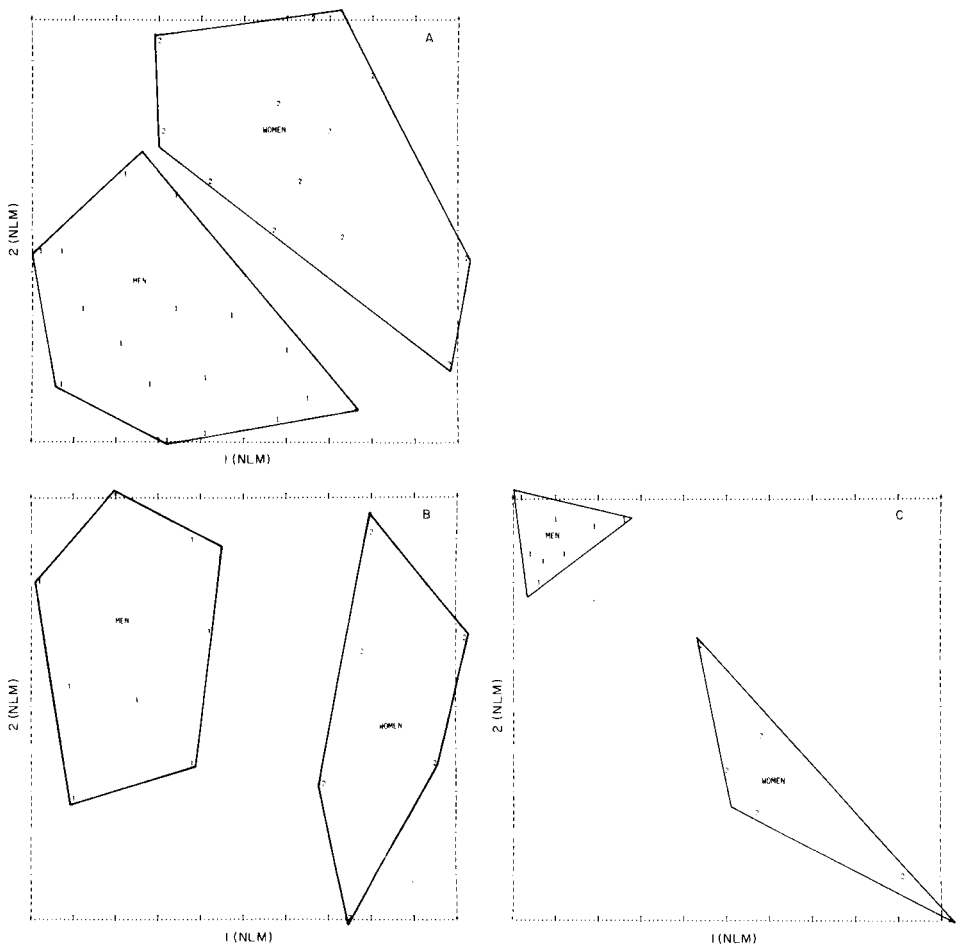


Fig. 3. Non-linear maps of f.d.m.s. profiles of urine samples from men and women (A), after averaging of the duplicate measurements (B) and after introduction of weighting factors (C).

In the next test, prior to non-linear mapping and after averaging of the duplicate measurements, the features were multiplied by their weights, as has been done for pyrolysis m.s. data [9]. This clearly improves the differentiation between the classes, as can be seen from Fig. 3.

In addition to non-linear mapping, the K nearest neighbour (KNN) method was used to classify the objects in this study. This method uses the Euclidean distance in an N-dimensional space to classify the objects. For the 1-NN method, only the closest neighbour is considered, and for the 2-NN, 3-NN, KNN methods, the closest 2, 3 or K neighbours. If an equal number of neighbours of both classes is found, an object is classified by considering the sum of distances to both classes. For the urine samples up to 6-NN, a 100% correct classification was found. Furthermore, the results of the KNN method provide a useful check on the correctness of the non-linear mapping results. The masses m/z = 93, 120 and 127 showed the highest Fisher index; even with a combination of only two masses, a complete separation of both classes could be obtained.

Fast atom-bombardment mass spectrometry

Two classes were defined: class 1 comprised Bordeaux wines and class 2 comprised Rhône wines. For each class, averaged f.a.b. spectra were used, and normalization, feature selection and autoscaling of the data were done as described above for the f.d.m.s. data. Figure 4 shows a typical f.a.b. profile of a wine from each class. In these profiles, prominent peaks are those of glycerol which was used as solvent in the experiments, e.g., the peaks at

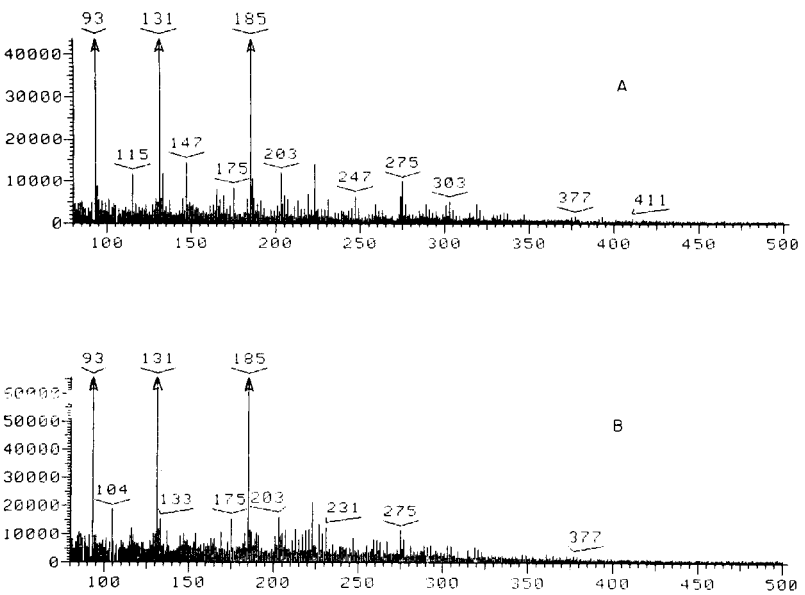


Fig. 4. F.a.b.m.s. profiles of (A) Bordeaux wine and (B) Rhône wine.

m/z = 93 $[M + H]^+$, 115 $[M + Na]^+$, 131 $[M + K]^+$, 185 $[2M + H]^+$, 223 $[2M + K]^+$ and 277 $[3M + H]^+$. Although these peaks interfere with those from the wine samples, they might sometimes contain interesting information. It was found that the wines produced a higher $[M + K]^+$ than a $[M + Na]^+$ peak of glycerol, which may give information about the sodium and potassium content of the wines. However, glycerol also leads to the formation of cluster ions of the type $[M + \text{glycerol} + H]^+$ and subsequently introduces more correlation and interference in the data set. Two approaches were made: (i) a non-linear map was constructed in the same way as described for the f.d.m.s. of urine with weighted data; (ii) principal component analysis was used.

Figure 5 shows the non-linear map and a plot of the first and second principal components. The non-linear map indicates the presence of two clusters, a condensed cluster of the Rhône wines and a broad cluster of the Bordeaux wines. This is also reflected in the KNN results, because the Rhône wines are correctly classified for 1-NN to 10-NN, the only exception being the 1-NN result of object 14 which is influenced by the presence of object 7 of class 1. The latter object together with object 6 are situated closely to the cluster of class 2. The other objects of class 1 are correctly classified up to 6-NN. This also shows that class 1 consists of a less condensed cluster than that of class 2.

From the principal component analysis of the data set, it is also clear (Fig. 5) that classes 1 and 2 can be described by two clusters and that object 7 is an outlier. Fisher weighting showed the masses 300, 240 and 211 to be the most important ones for differentiating between the two classes. A feature/feature plot of m/z = 300 and 240 revealed an almost complete separation of both classes. Only one object, object 7, was situated in the cluster of class 2, as was also found by the other techniques.

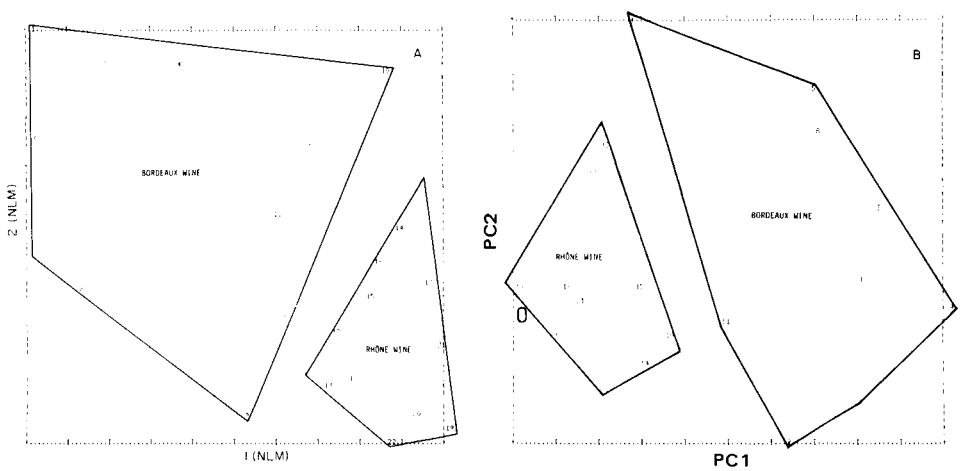


Fig. 5. Non-linear map (A) and principal component plot (B) of the f.a.b.m.s. profiles of Bordeaux and Rhône wines.

CONCLUSIONS

Field-desorption and fast atom-bombardment mass spectrometry seem to be useful as techniques for the generation of patterns from non-volatile matrices. Both techniques suffer from the generation of cluster ions and f.a.b. in particular suffers from background solvent peaks and the increased fragmentation of compounds obscuring the profile obtained. However, f.a.b. has the great advantage over f.d.m.s. that the technique is simple and capable of handling a large sample throughput.

Non-linear mapping in combination with the KNN method seems to be a promising technique in that it gives an insight into the structure of the data set. However, principal component analysis gives more information about the features which contribute most to the separation of the two classes. In cases in which the major variance in the data set is not the variance of interest, a feature-selection procedure prior to principal component analysis might be desirable. Fisher weighting is very useful in this respect and can also be applied as a criterion for the normalization method used. A standard approach for profile analysis in the future might be as follows. First, a limited number of objects (10—20/class) is measured and the influence of the experimental parameters on the cluster size is studied. Second, several objects are measured and classified by the procedure developed, which then provides information about the reliability of the method. Evaluating the classification procedure in this way can be followed by the identification of the relevant masses with the aid of high-resolution mass spectrometry.

REFERENCES

1 H.-R. Schulten, Int. J. Mass Spectrom. Ion Phys., 32 (1979) 97.
2 J. van der Greef and D. C. Leegwater, Biomed. Mass Spectrom., 10 (1983) 1.
3 M. Anbar, R. L. Dyer and M. E. Scolnick, Clin. Chem., 22 (1976) 1503.
4 R. Abbot, M. Anbar, H. Faden, J. McReynolds, W. Rieth, M. Scanlon, L. Verkh and B. Wolff, Clin. Chem., 26 (1980) 1443.
5 M. Barber, R. S. Bordoli, G. J. Elliott, R. D. Sedgwick and A. N. Tyler, Anal. Chem., 54 (1982) 645A.
6 M. Rabrenovic, T. Ast and V. Kramer, Int. J. Mass Spectrom. Ion. Phys., 37 (1981) 297.
7 H. D. Beckey, A. Heindrichs and H. U. Winkler, Int. J. Mass Spectrom. Ion. Phys., 3 (1970) 9.
8 K. Varmuza, Pattern Recognition in Chemistry, Springer-Verlag, Berlin, 1980.
9 W. Eshuis, P. G. Kistemaker and H. L. C. Meuzelaar, in C. E. R. Jones and C. A. Cramers (Eds.), Analytical Pyrolysis, Elsevier, Amsterdam, 1977, p. 151.

Analytica Chimica Acta, 150 (1983) 53—59
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

CLASSIFICATION OF ORGANIC COMPOUNDS BY INFRARED SPECTROSCOPY WITH PATTERN RECOGNITION AND INFORMATION THEORY

J. C. W. G. BINK and H. A. VAN 'T KLOOSTER*

State University of Utrecht, Laboratory for Analytical Chemistry, Chemometrics Research Group, Croesestraat 77A, 3522 AD Utrecht (The Netherlands)

(Received 17th September 1982)

SUMMARY

A structure correlation model for infrared spectra based on evaluation by the principal components technique is described. Classification is effected by the multiple linear regression method. As an optimization criterion for the classification, the equivocation parameter from information theory is used. The algorithms applied originate partly from Wold's SIMCA program package. From a data set of 549 infrared spectra, a class of some 40 spectra of compounds containing a t-butyl group was separated. Several methods for preprocessing of the data were compared. The best classification results were obtained, based on frequency data and rescaled intensity data, with weighting of variables for calculation of the model. A confidence measure of 96% was achieved.

In recent years, many workers have investigated the computer-assisted interpretation of infrared spectral data, for identification or classification of organic compounds. Studies on the application of information theory to such data have been summarized and discussed by Cleij and Dijkstra [1]. Most systems concerned with classification or substructure identification make use of Boolean logic, branching trees or heuristic algorithms [2—9]. Substructures are usually identified by a decision range or a threshold value for every selected variable, resulting in a yes/no decision with regard to the presence of a substructure. In this approach, mutual interferences of several functionalities is to be expected because of the enormous number of possibilities, and such interferences are difficult to predict. When new classes are added to the system, the updates or extensions that have to be made can be considerable, especially for branching trees. Also, extension of the system with more functional groups can cause complex combinatorial problems at higher interpretation levels, because of the decrease in the discriminatory power of the classification algorithm.

The investigation reported in this paper is based on the following ideas. First, evaluation of the structure—spectral feature correlation model should be quantitative in nature. Secondly, the classification system for multifunctional compounds should be easy to expand, which means that all (sub)-structures have to be modelled and recognized independently. Thirdly,

0003-2670/83/$03.00

the modelling method should also be suitable for treatment of small data sets or embedded data structures. For these reasons, the applicability of the SIMCA method [10, 11] was investigated; selected frequencies and/or peak intensities were used as variables. As the optimization criterion for the classification, "equivocation" (a parameter in information theory) [1] is shown to be appropriate.

EXPERIMENTAL

The data set used contains infrared spectra of 549 compounds (Table 1). The spectra were recorded on a double-beam infrared spectrometer, in the following standardized way: (1) the baseline was adjusted to (95 ± 2)% transmittance (T) for the entire spectral range; (2) the strongest band of the spectrum was adjusted to (5 ± 2)% T by varying the cell thickness; (3) the wavelength scale was frequently calibrated by using indene.

All compounds had a purity of over 98%, checked by gas—liquid chromatography. The error of the measurement of the transmittance was about 2%, whereas that of the wavenumber was 5 cm^{-1} for the range 4000—2000 cm^{-1} and 2 cm^{-1} for the range 2000—600 cm^{-1}. From every spectrum, band positions and transmittance values of the band maxima were available.

Computer programs for preprocessing of the data and classification were written in Pascal, while the programs to model the data were written in Fortran-IV and originate from the SIMCA package [11]. All programs were run on a CDC-CYBER-175 computer, and required a maximum of 64 K of 60-bit words of memory.

Preprocessing of spectral data

The data set used has two drawbacks, inherent to infrared spectra. First, the additivity of peaks causes shoulders or even missing band maxima, which can make raw intensity data less useful. Secondly, the particular geometry of some molecules may cause the disappearance of peaks which would generally be expected to be present. The first problem can be solved by deconvolution of all spectral data, but this could require much computing time. The second problem can be solved by dividing the class into a number of subclasses, each of which is represented by a specific set of peaks. With a limited data set,

TABLE 1

Data set used to classify t-butyl compounds

Code	Compound classes	Number of spectra
C1	t-Butyl	43
R1	i-Propyl, not t-butyl	36
R2	Ethyl, not t-butyl or i-propyl	160
R3	Remnants	320

however, most subclasses would be too small to model properly. Therefore, an attempt was made to solve both problems by allowing substitution of a limited number of actual or missing values by calculated mean values. This was necessary for about a third of the spectra in the class being considered. Two substitutions were allowed during processing, one in the high and one in the low frequency range. Mean band positions were selected at approximately 1365, 1392, 1464, 1479, 2872, 2908, 2936 and 2966 cm^{-1}. Split distributions (2872 cm^{-1}) or distributions with a low abundance of bands (2908, 2936 cm^{-1}) were contracted around an arbitrary value to form a new distribution. This is illustrated in Fig. 1.

Because of differences in the scaling of the peak intensities, which is dependent on the strongest band, the intensity values are randomly distributed. In order to reduce the variance of the intensity distributions, a rescaling factor r_j was derived as follows. For an absorbance A_{ij} at position i of spectrum j, the Lambert—Beer law requires

$$A_{ij} = f_j \, \epsilon_{ij} = f_j \, (\bar{\epsilon}_i + \Delta\epsilon_{ij}) \tag{1}$$

with $\bar{\epsilon}_i = \Sigma_j^N \epsilon_{ij}/N$. Here, f_j is a product of concentration and cell pathlength, ϵ_{ij} is the molar absorptivity and N is the total number of spectra of the class

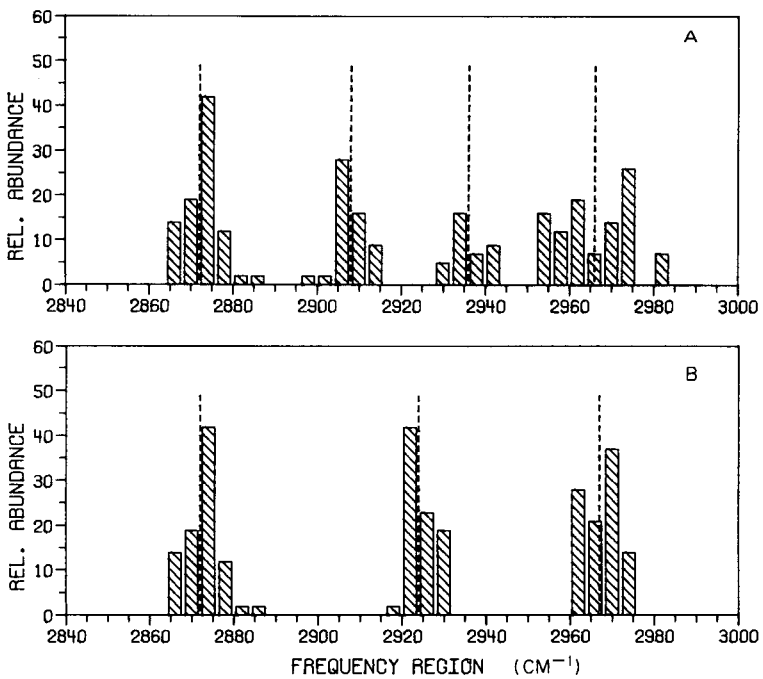


Fig. 1. Raw frequency distributions (A) and composed frequency distributions (B) of t-butyl compounds in the high frequency region.

being processed. For the mean absorbance $\bar{A}_i$ of peak i, it follows that

$$\bar{A}_i = \sum_j^N A_{ij}/N = \bar{f}\,\bar{\epsilon}_i + \sum_j^N (f_j \Delta\epsilon_{ij})/N \approx \bar{f}\,\bar{\epsilon}_i \tag{2}$$

with $\bar{f} = \sum_j^N f_j/N$.

Minimizing $\sum_i (\bar{A}_i - r_j A_{ij})^2$ yields

$$r_j = \sum_i \bar{A}_i \cdot A_{ij} / \sum_i A_{ij}^2 \tag{3}$$

If $\Delta\epsilon_{ij}$ is assumed to be random and small compared to ϵ_{ij}, then Eqns. (1–3) yield $r_j \approx \bar{f}/f_j$. By rescaling all absorbances A_i of spectrum j by a factor r_j (calculated from Eqn. 3), the variance of the intensity distributions is reduced by approximately 25%.

Modelling and classification

After preprocessing of the data and autoscaling of variables X_i, the principal components model for the t-butyl structure was calculated from

$$X_{ik} = \sum_{a=1}^{A} \beta_{ia}\,\Theta_{ak} + \epsilon_{ik} \tag{4}$$

where i is the index of the variables, k is the index of the objects (spectra), a is the index of the eigenvectors and A is the number of significant eigenvectors Θ. The residuals ϵ give an indication of the degree of imperfection of the class model [10, 11]. The typical deviation S_m from the model can be estimated from the formula

$$S_m^2 = \sum_i^M \sum_k^N \epsilon_{ik}^2/(M-A)(N-A-1) \tag{5}$$

where M is the number of variables and N is the number of "class" spectra.

Then all "non-class" spectra are fitted to the class model and each deviation S_p of spectrum p is calculated from

$$S_p^2 = \sum_i^M \epsilon_{ip}^2/(M-A) \tag{6}$$

The F ratio of S_p and S_m is an indication of the goodness of fit of each spectrum to the class model, and can be used for classification purposes.

Then a limiting F value is chosen empirically, such that classification results are optimal. The optimization criterion used was the "equivocation" (E), i.e., the expected value of the uncertainty after classification [1]. This is a general measure of the system quality; its relation with the information content (I) is

$$E = H(X) - I \tag{7}$$

where X is the class to which the unknown compound belongs, $H(X)$ is the uncertainty with respect to X before classification and I is the information

content of the classification procedure (H and I both in bits). Because $I_{max} = H(X)$, the minimum value of E (zero) is independent of the a priori probabilities. The equivocation can be calculated from

$$E = [-\sum_{k}^{2} p(Y_k)] [\sum_{i}^{2} p(X_i/Y_k)\, ^2\log p(X_i/Y_k)] \quad (8)$$

where $p(Y_k)$ is the overall probability that an unknown compound is classified in class X_k, and $p(X_i/Y_k)$ represents the (a posteriori) probability that an unknown classified as a member of class Y_k, actually belongs to class X_i. In the present case, only two classes are considered: t-butyl and non-t-butyl compounds.

For the optimal equivocation, the "confidence measure (CM)" is also calculated, which is easier to interpret [12]:

$$CM \approx 0.5 \sum_{i}^{2} p(X_i/Y_i) \quad (9)$$

In Eqns. (8) and (9), the a priori probability $p(X)$ of an unknown belonging to one of the classes (t-butyl, non-t-butyl) is assumed to be 0.5, which gives an estimate of the worst classification results, as shown in Fig. 2.

RESULTS AND DISCUSSION

Different methods of selecting and preprocessing variables were compared. The classification results are given in Table 2. The band positions are expressed in cm^{-1} and the intensities refer to absorbances. Tests 1 and 2 show that both frequency data and intensity data contain a considerable amount of information. From test 3, however, it can be deduced that frequency and intensity data are highly correlated, because the combination gives only slightly better results of classification. Furthermore, comparison of test 1 with tests 3 and 4 shows that only the inclusion of rescaled intensity data has a positive effect on classification. Inclusion of raw intensity data shows partly the opposite effect.

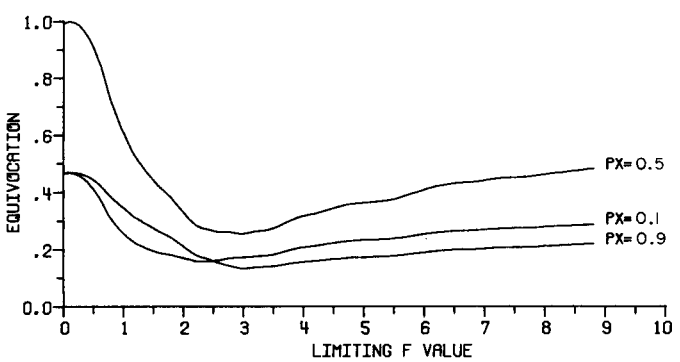


Fig. 2. The equivocation parameter as a function of the limiting F ratio at several a priori probabilities $p(X)$ of the t-butyl group abundance.

TABLE 2

Descriptions and performances of several tests. Absorbances are rescaled and variables are not weighted unless stated otherwise

Test No.	Frequency variables	Intensity variables	No. of components	Equivocation	Confidence measure (%)	Misclassifications from			
						C1	R1	R2	R3
1	7	—	2	0.313	94	1	19	16	14
2	—	7	2	0.432	91	1	24	41	24
3	7_a	7_a	2	0.278	95	1	18	13	7
4	7^a	7^a	2	0.323	94	1	21	21	10
5	4_b	5_b	3	0.299	95	1	20	17	8
6	7^b	7^b	3	0.251	96	1	16	10	6

[a]Raw. [b]Weighted.

In the final modelling stage, two methods are available to optimize the class model. In one method, attempts are made to delete variables with low modelling power and low discriminatory power. In the other method, all variables are reweighted according to their goodness of fit to a particular class model [10, 11]. Both methods were investigated. From tests 5 and 6 it seems that, for this data set, reweighting of variables gives better results than deletion of variables. Deletion of variables even has a negative effect on classification. This can be explained by assuming that variables globally showing a low discriminatory power and modelling power, are indispensable for distinguishing a few non-class compounds from the class compounds. However, both methods reduce the influence of noise; the number of significant eigenvectors, which describe the variance of the data, increases from two for tests 1—4 to three for tests 5 and 6.

In conclusion, the classification results are encouraging, except for the rather high interference percentage of group R1, the isopropyl-containing compounds. Of course, the similarity of classes R1 and C1 is emphasized by allowing substitutions for missing data. This can be avoided by introducing a number of subclasses, each with its own set of peaks. Another advantage would be that smaller groups of more analogous compounds would certainly improve the modelling strategy. This last option is also suggested after scanning the calculated F ratios, which are rather high for most members of the small subclasses $(CH_3)_3$—C—O or $(CH_3)_3$—C—N, the others containing $(CH_3)_3$—C—C.

In a larger classification system, however, modelling of a larger group of compounds for preselection purposes can reduce computing time considerably. Furthermore, the limiting F ratio can always be chosen with different criteria, such as recall and/or reliability threshold values [13], to serve the purposes of the user. A general optimization criterion (the equivocation parameter) which gives an equal weight to both a positive and a negative prediction was selected here, and the worst result of prediction possible in this case was estimated. However, for use in a structure-generating program, for example, it is desirable to detect all functional groups that are present.

Accordingly, the models have to be made very flexible (high recall), with a minimal loss of prediction power.

The authors are indebted to Drs. P. Cleij, Prof. J. H. van der Maas and Mr. T. Visser for valuable discussions.

REFERENCES

1 P. Cleij and A. Dijkstra, Anal. Chim. Acta, 133 (1981) 19.
2 N. A. B. Gray, Anal. Chem., 47 (1975) 2426.
3 L. A. Gribov, Anal. Chim. Acta, 122 (1980) 249.
4 M. Farkas, J. Markos, P. Szepesváry, I. Bartha, G. Szalontai and Z. Simon, Anal. Chim. Acta, 133 (1981) 19.
5 T. Visser and J. H. van der Maas, Anal. Chim. Acta, 133 (1981) 19.
6 H. Abe, T. Yamasaki, I. Fujiwara and S. Sasaki, Anal. Chim. Acta, 133 (1981) 499.
7 H. B. Woodruff and G. M. Smith, Anal. Chim. Acta, 133 (1981) 545.
8 J. Zupan, Anal. Chim. Acta, 139 (1982) 143.
9 R. Tsao and W. L. Switzer, Anal. Chim. Acta, 134 (1982) 111.
10 S. Wold, Pattern Recognition, 8 (1976) 127.
11 S. Wold and M. Sjöström in B. R. Kowalski (Ed.), Chemometrics: Theory and Application, ACS Symp. Ser. 52, Washington, DC, 1977, p. 243.
12 J. A. Richards and A. G. Griffiths, Anal. Chem., 51 (1979) 1358.
13 F. W. McLafferty, Anal. Chem., 49 (1977).

Analytica Chimica Acta, 150 (1983) 61—70
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

A MULTIVARIATE CALIBRATION PROBLEM IN ANALYTICAL CHEMISTRY SOLVED BY PARTIAL LEAST-SQUARES MODELS IN LATENT VARIABLES

MICHAEL SJÖSTRÖM and SVANTE WOLD*

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

WALTER LINDBERG and JAN-ÅKE PERSSON

Department of Analytical Chemistry, Institute of Chemistry, Umeå University, S-901 87 Umeå (Sweden)

HARALD MARTENS

Norwegian Institute for Food Research, As (Norway)

(Received 26th November 1982)

SUMMARY

The use of partial least squares in latent variables (PLS) for multivariate calibration problems is described. The application is the simultanous determination of ligninsulfonate, humic acid and an optical whitener, from their severely overlapping fluorescence spectra. The predictive performance of the resulting calibration model is tested with a separate set of samples. The PLS method also identifies samples which do not fit the calibration model. The PLS method is compared with principal components analysis combined with multiple regression.

For quantitative analysis of complex samples, fast and cheap spectroscopic methods are preferable to the slow and expensive "wet chemical" approach. However, a disadvantage of the spectroscopic methods is the difficulty of finding frequency regions where the constituents of interest selectively absorb or emit light. This problem can be dealt with by measuring P separate frequencies, of a set of N spectra with known composition. Then some method of data evaluation is applied to find a model which combines the P measurements in $\mathbf{X}$ to give as good a prediction of $\mathbf{Y}$ as possible. Here, the $N \times P$ matrix describing the spectra is denoted by $\mathbf{X}$, and the vector (or matrix) describing the known compositions by y (or $\mathbf{Y}$). In the latter case, Q different properties or constituents of the samples are measured. The calibration model is then used to predict the compositions of new samples n' from their spectra. The data for a calibration problem can be organized as in Fig. 1.

Traditional methods

Usually the calibration problem has been solved by applying multiple regression by a linear model of the vector y in $\mathbf{X}$. With Q y-variables, Q

0003-2670/83/$03.00

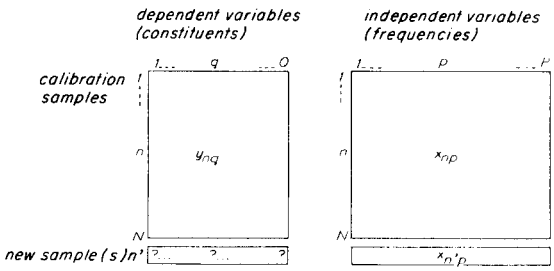


Fig. 1. Organization of data for a multivariate calibration problem.

separate multiple regressions are made, one for each separate y. However, spectroscopic data are not very suitable for ordinary multiple regression where the moment matrix $\mathbf{X}'\mathbf{X}$ is inverted. Because the variables are usually highly correlated, the inversion of $\mathbf{X}'\mathbf{X}$ will result in an almost singular matrix and no statistically stable solution is obtained. The problem can be dealt with by adding a small constant to the diagonal elements in the moment matrix before its inversion, the so-called ridge regression. An alternative is stepwise multiple regression, giving a reduced set of less highly correlated variables [1].

Another way of reducing the dimensionality of the data is to apply principal components analysis to X. Principal components analysis [2] will give a set of uncorrelated new variables (object scores). The object scores can then be used instead of the original variables in multiple regression as shown, for example, by Ho et al. [3]. In the following paragraphs, PC/MR is used to denote principal components analysis combined with multiple regression.

Present approach

A new approach to the multivariate calibration problem is described here. The method is called the partial least-squares model in latent variables (PLS) and was developed by H. Wold and coworkers [4—6]. This method has some features in common with PC/MR because X is described by a principal components type of model, combined with a regression relation between the object scores and Y. However, in PLS the information in Y is also used in the estimation of the object scores for X, which is not the case for PC/MR. Furthermore, the method works in one step and more than one constituent variable (y) at a time can be treated. In this paper, the PLS algorithm described has two blocks, suitable for multivariate calibration problems.

The PLS method will be illustrated by the simultaneous determination of the concentrations of ligninsulfonate, humic acid and a detergent containing a whitener, by molecular fluorescence [7]. Ligninsulfonate is one of the major pollutants emitted from pulp mills into sea water. An attractive method to determine ligninsulfonate is molecular fluorescence, which offers high sensitivity. However, with this method, spectral overlap is a serious limitation, because humic acids and an optical whitener emit energy in the frequency region of interest. Furthermore, the spectra of the pure constituents do

not add up to the expected measured spectrum of a mixture of the three constituents. There are deviations in both intensity and shape. This non-additive behavior places severe demands on the statistical calibration method.

EXPERIMENTAL

The emission spectra of 16 mixtures of the three constituents were recorded between 320 and 540 nm. The emission intensities at 27 equally distributed wavelengths were used. In this way, a 16 × 27 ($N \times P$) matrix (X) was formed, which described the emission at P frequencies of the N spectra. The concentrations of the three constituents for the N spectra form a 16 × 3 ($N \times Q$) matrix (Y).

Nine additional mixtures were prepared and their spectra were recorded and digitized as before. These samples were not used to calibrate the model, but for testing the predictive properties of the PLS model from Y and X.

All calculations were done with the SIMCA-3B Basic program for 8-bit microcomputers. The emission spectra were measured on a Perkin-Elmer Model 512 double-beam fluorescence spectrometer.

THE PLS METHOD AND ALGORITHM

In the PLS model, the variation in X is explained in terms of the following model (d, e and f are residuals):

$$y_{nq} = \bar{y}_q + \sum_{a=1}^{A} b_{aq} u_{na} + f_{nq} \tag{1}$$

$$x_{np} = \bar{x}_p + \sum_{a=1}^{A} b_{ap} t_{na} + e_{np} \tag{2}$$

$$u_{na} = c_a t_{na} + d_{na} \text{ (for each } a = 1 \ldots A) \tag{3}$$

A geometrical illustration of the PLS method is given in Fig. 2.

Algorithm

(i) When no information about the relative importance of the different y and x variables is available, an initial scaling to variance one is recommended. This is accomplished by dividing each variable in the X and Y blocks by a scaling factor which is equal to one divided by the standard deviation of the variable. Henceforth, X and Y refer to the scaled data.

(ii) The X and Y matrices are centered by subtracting the averages $\bar{y}_q$ and $\bar{x}_p$ for each of the Q columns in case of Y and for the P columns in X:

$$\bar{y}_q = \sum_n y_{nq}/N \tag{4}$$

$$\bar{x}_p = \sum_n x_{np}/N \tag{5}$$

The zero dimension residuals e_{np} and f_{nq} are then given by

$$f_{nq} = y_{nq} - \bar{y}_q \tag{6}$$

$$e_{np} = x_{np} - \bar{x}_p \tag{7}$$

The subsequent steps (iii)—(viii) provide iterative calculation of the latent variable u_{na} for $a = 1$.

(iii) Starting values for the first iteration of u_{na} are set by the first column in **F**:

$$u_{na} = v f_{nl} \tag{8}$$

Here v is a normalization factor giving u_{na} unit length.

(iv) Calculation of weights w_{ap} for the **X** block are calculated

$$w_{ap} = \sum_n e_{np} u_{na} \tag{9}$$

(v) Latent variable t_{na} is calculated for the **X** block:

$$t_{na} = v \sum_p w_{ap} e_{np} \tag{10}$$

Here v normalizes t to unit length.

(vi) Weights b_{aq} for the **Y** block are calculated from

$$b_{aq} = \sum_n f_{nq} t_{na} \tag{11}$$

(vii) The new latent variable u_{na} is then formed from

$$u_{na} = v \sum_q b_{aq} f_{nq} \tag{12}$$

Here v normalizes u to unit length.

(viii) If the new u_{na} in step (vii) all differ less than one part per million from the u_{na} in the previous round, convergence is reached. Then continue with step (ix), otherwise go back to (iv) for a new round.

(ix) Loadings b_{ap} are done for the **X** block:

$$b_{ap} = \sum_n e_{np} t_{na} \tag{13}$$

(x) The inner relation is

$$c_a = \sum_n u_{na} t_{na} \tag{14}$$

(xi) The new residuals in the **X** and **Y** blocks are formed from

$$e_{np} = e_{np} - b_{ap} t_{na} \tag{15}$$

$$f_{nq} = f_{aq} - c_a b_{aq} t_{na} \tag{16}$$

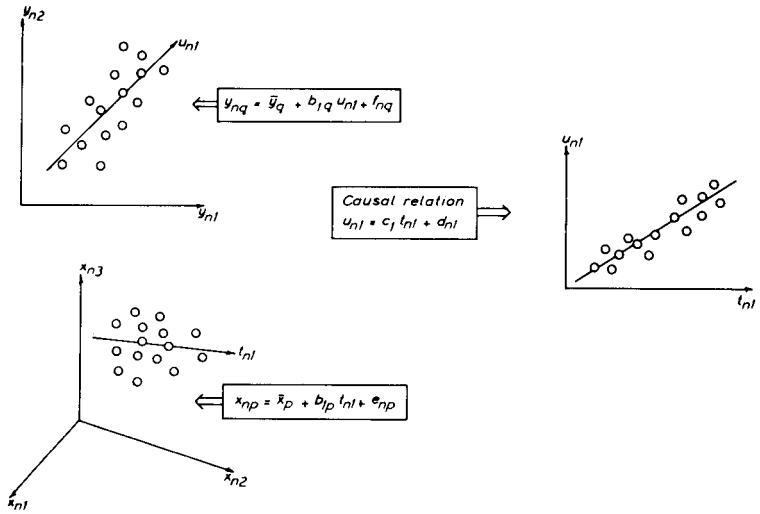


Fig. 2. A geometrical illustration of a PLS model with one cross-term ($A = 1$) for a problem with two constituents (y_{n1} and y_{n2}) and three variables ($x_{n1} - x_{n3}$). In practice, the number of variables in the X block should be much larger and usually $A \gg 1$.

If the last dimension a is deemed insignificant according to, for example, cross-validation (see below), the program is terminated. Otherwise it is continued with an additional dimension ($a + 1$), by going back to step (iii).

The number of dimensions, A. Cross-validation [8] is a method of determining if a cross-term $b_{ap}t_{na}$ is significant or not. With this method, the predictive ability of the ath cross-term is investigated by first deleting, say, one quarter of the calibration samples. The y values of the deleted samples are then predicted and the sum of squared difference (SS) between the observed and calculated y values is calculated. Another quarter of the samples is then kept out, a second SS is calculated, etc., until all calibration samples have been kept out just once.

If the sum of the four partials SS is smaller than the SS of Y after $a - 1$ dimensions, then the ath cross-term contains predictive information and is deemed to be statistically significant.

Prediction of new samples

Once a calibration model has been established, the comparison of a new sample n' can be determined as follows. The spectrum is digitized, centered and scaled in the same way as for the calibration set, giving a vector with the elements x_p ($p = 1 \ldots 27$).

The t_a parameters and the residuals are calculated by fitting x_p to b_{ap}, ($a = 1 \ldots A$) by multiple regression:

$$x_p = \sum_{a=1}^{A} t_a b_{ap} + e_p \qquad (17)$$

Equations (1—3) indicate that y_q can be estimated from

$$y_q = \bar{y}_q + \sum_{a=1}^{A} b_{aq} c_a t_a \tag{18}$$

where b_{aq} and c_a are known from the calibration set and t_a from Eqn. (17). Another way, giving slightly different values of y_q, is to use the weights w and calculate t from Eqn. (10).

The predicted values of y are now in scaled and centered form and can be transformed back to the original coordinates by applying the reverse centering and scaling as in the calibration set.

Identification of test samples giving poor prediction

The residual variance for sample n' is given by

$$s_{n'}^2 = \sum_{p} e_p^2/(P-A) \tag{19a}$$

where the residuals e_p are calculated in Eqn. (17). Similarly, the residual variances for a calibration sample n are given by

$$s_n^2 = \sum_{p} e_p^2 \, (N/(P-A)(N-A-1)) \tag{19b}$$

The difference between these equations is due to the different number of degrees of freedom for the calibration samples and the test samples. With an approximate F-test

$$F = s_{n'}^2/S_x^2 \text{ with } (P-A) \text{ and } (P-A)(N-A-1) \text{ degrees of freedom} \tag{20}$$

the variance from Eqn. (19a) can be compared with the overall variance for the calibration set in the X-block, S_x^2:

$$S_x^2 = \sum_{np} e_{np}^2/(P-A)(N-A-1) \tag{21}$$

If this approximate F-test shows that the residual variance of sample n' is significantly larger than the variance of the calibration set, this is an indication not to rely on the predicted y_q values estimated from Eqn. (18). The measurements on sample n' do not fit the model, thus giving a poor estimate of t_a used in the prediction step. However, the approximate nature of this F-test should be stressed. The calibration samples are favored compared to validation samples, and this F-test will give an excessively narrow confidence interval for the validation set. The prediction errors can be expected to increase with decreasing fit of a spectrum to the PLS model. This means that, for large F values, large prediction errors are expected for y, but for F values close to the critical F value, the increase in the prediction errors of y compared to the prediction errors for the calibration set will be small if any. Figure 3 illustrates the procedure for identifying a sample that will give an inferior prediction of y, compared to the calibration samples.

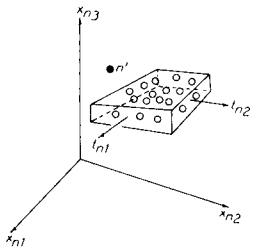


Fig. 3. Illustration of the identification of a test sample not qualified for prediction of y values. With three variables in X, the observations in the calibration set are described as points in a three-dimensional space. With two significant cross-terms, the t_{n1} and t_{n2} vectors describe two new dimensions in this space. The confidence limits from the approximate F-test can be thought of as an envelope containing the calibration samples. If a sample n' lies far outside the confidence limit of the model, this means that the spectrum of n' is unsatisfactorily described by the calibration model ($\bar{x}_p$, b_{1p} and b_{2p}). The resulting $t_{n'1}$ and $t_{n'2}$ will give poor predictions of y.

Principal components analysis with multiple regression

In this approach X is described by a principal component model:

$$x_{np} = \bar{x}_p + \sum_{a=1}^{A} b_{ap} t_{na} + e_{np} \tag{22}$$

In this case the number of significant components is determined by cross-validation. Then for the calibration set the t_{na} vectors are fitted to one y_n vector at a time by multiple regression:

$$y_n = c' + \sum_{a=1}^{A} b'_a t_{na} + e'_n \tag{23}$$

For a validation sample n', its data vector x_p is fitted to the b_{ap} values from Eqn. (22), to give the t_{na} values as in Eqn. (17). By introducing this value in Eqn. (23), where now the b'_a parameters are known, y values can be calculated.

RESULTS AND DISCUSSION

The composition of the data set is given in Table 1 together with the calculated compositions from a PLS model with seven components (A = 7 in Eqns. 1—3). Samples 1—16 are used to calibrate the model and samples 17—25 form a validation set, but their actual compositions are known. Thus the predicted values for samples 17—25 are calculated from Eqns. (17) and (18) with the input of the calibration model. The s_n value is the fit of a spectrum to the X-block part of the model (see Eqns. 19a or b). The F values are found from Eqn. (20).

The validation samples 17, 19, 20 and 21 fit the calibration model well and the sample compositions are predicted just as well as in the calibration

TABLE 1

Sample composition of the calibration and validation set and the prediction errors of a seven-component PLS model

Sample[a]	Composition[b]			Prediction errors[c]			Fit	
	y_1	y_2	y_3	d_1	d_2	d_3	s_n^d	F^e
1	3.011	0	0	0.041	0.028	2.39	0.0042	0.14
2	0	0.401	0	0.067	0.057	0.21	0.0054	0.24
3	0	0	90.63	0.030	0.003	3.67	0.0106	0.91
4	1.482	0.158	40.00	0.098	0.024	0.22	0.0123	1.27
5	1.116	0.410	30.45	0.081	0.020	12.3	0.0125	1.23
6	3.397	0.303	50.82	0.032	0.036	0.41	0.0155	1.95
7	2.428	0.298	70.59	0.041	0.019	0.16	0.0106	0.91
8	4.024	0.115	89.39	0.042	0.053	2.55	0.0094	0.72
9	2.275	0.504	81.75	0.087	0.027	6.60	0.0131	1.40
10	0.959	0.145	101.10	0.018	0.071	2.37	0.0166	1.09
11	3.190	0.253	120.00	0.085	0.012	10.24	0.0149	1.80
12	4.132	0.569	117.70	0.119	0.001	1.70	0.0129	1.35
13	2.160	0.436	27.59	0.019	0.031	3.95	0.0081	0.53
14	3.094	0.247	61.71	0.012	0.057	12.38	0.0126	1.29
15	1.604	0.286	108.80	0.075	0.025	3.40	0.0087	0.62
16	3.162	0.701	60.00	0.037	0.039	2.55	0.0086	0.61
17	2.443	0.289	80.22	0.063	0.018	0.82	0.0141	1.62
18	4.078	0.361	88.52	0.074	0.054	14.37	0.0377	11.6
19	1.065	0.234	69.23	0.084	0.009	1.22	0.0082	0.54
20	3.317	0.123	40.13	0.065	0.067	16.26	0.0136	1.50
21	0.998	0.416	30.74	0.089	0.070	3.0	0.0100	0.81
22	2.983	0.403	120.0	0.083	0.103	2.56	0.0155	1.95
23	5.132	0.229	49.05	0.334	0.041	14.64	0.0243	4.80
24	5.058	0.000	51.06	0.181	0.008	12.24	0.0195	3.10
25	0.0	0.735	99.57	0.443	0.755	31.63	0.231	433.0

[a]Samples 1—16 form the calibration set and samples 17—25 are the validation set. [b]y_1—y_3 are the sample compositions ($\mu g\ ml^{-1}$) for humic acid, ligninsulfonate and the whitener, respectively. [c]Prediction errors are the absolute values of the difference between the actual composition and the calculated value. For samples 1—16 the calculated values are from Eqn. (1) and for samples 17—25 from Eqn. (18). [d]The fit of each sample to the model is given by s_n (Eqn. 19a, b). [e]The F-test is according to Eqn. (20): $F_{0.01,\ crit.} = 2.0$.

set. Slightly enhanced F values are observed for samples 18, 23 and 24; for samples 23 and 24, the prediction for ligninsulfonate is slightly worse. Sample 25 has a high F value ($F = 443$) and the prediction errors are large for all three constituents compared to the prediction errors of the calibration set.

Thus the ability of the PLS method to predict the compositions of the validation samples from their highly overlapped spectra is demonstrated. Further, the PLS method also detects samples which will give poor predictions.

For the present data set, multiple regression methods are inferior because the number of variables is large compared to the number of samples. Step-

wise multiple regression might be used in this problem if about two thirds of the variables were deleted, which would lead to loss of information. It should also be noted that multiple regression methods, in contrast to PLS, will give no indication of the strange behavior of the spectrum of sample 25.

The individual spectra for specific concentrations do not add up to the spectrum of the mixture with the same concentrations of the constituents. This means that multiple regression with the three individual spectra as independent variables also gives poor predictions of y.

Comparison with the PC/MR method

The PLS method is compared with the PC/MR approach, and the prediction errors for the validation set are presented, in Table 2. Principal components analysis combined with cross-validation of X shows that eight components are needed to describe the systematic information in the data; this is one component more than in the PLS method. Thus the results for the two methods are presented with both seven and eight components.

The PLS method gives slightly better prediction for ligninsulfonate and the whitener compared to PC/MR with both $A = 7$ and $A = 8$. It should be noted that these are the constituents that contribute less to the spectra. This comparison with the PC/MR method shows that PLS makes a good prediction of the constituents that make a small contribution to the overall fluorescence spectra, at the cost of a very slight decrease of the precision of the most strongly-emitting compound.

Conclusion

The PLS approach has some obvious advantages over the traditional approach. First, the information in Y is used; hence if X contains a structure which has predictive relevance for Y, this will appear in the PLS solution, which is not necessarily the case for the PC/MR method, multiple regression or ridge regression. Secondly, once the PLS model has been determined, it is possible to classify a new sample as similar to the calibration set or not. This means that the information is obtained whether or not the calibration set is qualified to determine the composition of a particular new sample. Such information is not given by the multiple or ridge regression techniques.

TABLE 2

Comparison of the predictive ability of the PLS and PC/MR methods for the validation set. The predictive ability is expressed as the sum of the squared prediction errors ($\Sigma_n (y_n - y_{n,pred.})^2$)

Constituent	PLS ($A = 7$)	PC/MR ($A = 7$)	PLS ($A = 8$)	PC/MR ($A = 8$)
Humic acid	0.376	0.335	0.419	0.347
Ligninsulfonate	0.595	0.669	0.492	0.621
Whitener	1.85×10^3	3.1×10^3	1.88×10^3	1.92×10^3

Finally, the PLS algorithm is easy to program for microcomputers and calibration problems can be solved for large data sets in a fraction of the time needed for the MR approach.

Grants from the Swedish Natural Science Research Council (NFR) and the Swedish Board of Research Councils are gratefully acknowledged.

REFERENCES

1 N. R. Draper and H. Smith, Applied Regression Analysis, 2nd edn., Wiley, New York, 1981.
2 See, e.g., D. L. Massart, A. Dijkstra and L. Kaufman, Evaluation and Optimization of Laboratory Methods and Analytical Procedures, Elsevier, Amsterdam, 1980, Ch. 19.
3 C.-N. Ho, G. D. Christian and E. R. Davidson, Anal. Chem., 52 (1980) 1071.
4 H. Wold, in J. Gani (Ed.), Perspectives in Probability and Statistics. Papers in honour of M. S. Bartlett, Academic Press, London, 1975.
5 H. Wold, in K. G. Jöreskog and H. Wold (Eds.), Systems under Indirect Observation, North-Holland, Amsterdam, 1982.
6 S. Wold, H. Wold, W. J. Dunn III, A. Ruhe, Report UMINF, 83 (1980).
7 W. Lindberg and J-Å. Persson, Anal. Chem., 55 (1983) in press.
8 S. Wold, Technometrics, 20 (1978) 379.

Analytica Chimica Acta, 150 (1983) 71—86
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

INTERACTIVE CALIBRATION BY A RECURSIVE GENERALIZED STANDARD ADDITION METHOD

BERNARD VANDEGINSTE*, JO KLAESSENS and GERRIT KATEMAN

Department of Analytical Chemistry, University of Nijmegen, 6525 ED Nijmegen (The Netherlands)

(Received 6th December 1982)

SUMMARY

The generalized standard addition method (GSAM) has recently been introduced for calibration to avoid matrix effects and many interferences. In GSAM, all additions and measurements precede the calculations of the K matrix (sensitivities of all analytes at all sensors) and the analyte concentrations. A recursive version of GSAM is presented here. Some new powerful properties are added to the method: (1) on-line checking of the validity of the linear model during the data acquisition stage, (2) simultaneous evaluation of the value and precision of the elements of the K matrix during the additions, and (3) access to an on-line interactive calibration, terminating the additions when the precision is within the desired limits or a deficiency in the model has been detected. The method is demonstrated for spectrophotometric measurements of copper and nickel in the presence of EDTA.

The quality of the calibration of most analytical methods fixes the quality of the final result. In most cases, there is no absolute quantitative relationship between analyte concentration and response (or reading), so that the response obtained for an unknown sample has to be compared with the responses obtained for a number of calibration samples. The problems associated with calibration are well known: the selectivity of many analytical sensors is poor, and many constituents present in the sample may affect the response, i.e., interfere. A second serious problem associated with calibration is the varying sensitivity of the sensor with changing matrix in the sample, i.e., the matrix effect. Of course, calibration methods which handle these problems adequately have been available for many years. Interferences may be compensated by applying a multi-component analysis. Matrix effects may be avoided by using the standard addition method.

Until very recently, systems where interferences and matrix effects occur simultaneously could not be calibrated in a direct way. The problem has been solved by combining the multicomponent analysis and the standard addition method [1, 2] into a widely applicable method, the generalized standard addition method (GSAM). The merit of GSAM is not only the assessment of calibration in a more general way, but also the provision of a quantitative relationship between the precision of the analytical result and

0003-2670/83/$03.00

the measurement error, in relation to the design of the calibration. Therefore, GSAM constitutes a firm step toward a self-calibrating system, which will become of full value when the measured response is placed in a feedback loop with the progress of the experimental design. This means that each new step in the calibration stage is decided on the basis of all responses collected previously.

In its present form, GSAM does not provide this feature. All additions precede the calculations of the sensitivity constants (K matrix), the analyte concentrations and the propagated error in the result. This implies that there is no check that the underlying model remains valid during the additions. Equally, no indication is obtained on the necessity of further additions, taking into account the desired precision of the analytical result. For this purpose, an algorithm is needed which returns improved least-squares estimates of the model parameters (e.g., sensitivity coefficient), every time a new measurement has been acquired. Such a recursive algorithm opens the way to achieve a feedback between response and experimental design, which is essential in the development of self-calibrating systems. Very recently, the first applications of recursive algorithms in analytical chemistry have been reported [3—6] as an alternative to the calculation of the analyte concentrations in a multicomponent procedure.

In this paper, a recursive version of the GSAM is presented, which adds some new powerful properties to the method: (1) on-line control of the validity of the model (linear) during the calibration stage; (2) simultaneous evaluation of the value and precision of the elements in the matrix of the sensitivity coefficients, during the additions; (3) access to an on-line interactive calibration, terminating the additions, when either the precision is within the desired limits, or a deficiency in the model has been detected.

THEORY

For a clear understanding of the recursive GSAM, the underlying principles will be discussed stepwise by an outline of multicomponent analysis, the standard addition method, the generalized standard addition method (GSAM), and the recursive least-squares method, leading to the recursive generalized standard addition method (RGSAM).

Formulation of the analytical problem

The problem is to determine the concentrations of NA analytes in a particular sample. Each of the NA analytes contributes to the response measured at a particular sensor. *On the assumption that the responses are additive*, the response at sensor i can be expressed as

$$R_i = \sum_{j=1}^{NA} k_{ij} c_j + e_i \qquad (1)$$

where c_j are the unknown concentrations, e_i is the measurement noise, and k_{ij} is the sensitivity coefficient of analyte j at sensor i.

At least NA independent measurements are required to determine all NA concentrations. The analytical problem then can be formulated as follows: given a measurement vector $\mathbf{R}$ of NS ($NS \geqslant NA$) independent sensors, and a matrix $\overline{\mathbf{K}}$ of sensitivity coefficients, find the best estimates $\hat{\mathbf{c}}$ of the true concentrations $\mathbf{c}$ from the response equation: $\mathbf{R} = \overline{\mathbf{K}}\,\mathbf{c} + \mathbf{e}$, where $\mathbf{R}$ is a $NS \times 1$ column vector of NS responses, $\mathbf{c}$ is a $NA \times 1$ column vector of the unknown concentrations, $\mathbf{e}$ is a $NS \times 1$ column vector of the measurement noise, and $\overline{\mathbf{K}}$ is a $NS \times NA$ matrix of the sensitivity coefficients.

Solution 1: Multicomponent analysis

Under the very restricted condition that the sensitivity coefficients $\overline{\mathbf{K}}$ are independent of the matrix and the concentrations c, the solution of the analytical problem can be found simply by an ordinary multicomponent analysis. For $NS \geqslant NA$, the number of equations exceeds the number of unknowns. A general solution method for such an over-determined system of equations is the least-squares method that minimizes $\Sigma\, e_i^2$, where e_i represents the difference between the actual response R_i and the estimated response $\hat{R}_i$, given by Eqn. (1). The solution is given by the well known generalized inverse

$$\hat{\mathbf{c}} = (\overline{\mathbf{K}}^{\mathrm{T}}\overline{\mathbf{K}})^{-1}\,\overline{\mathbf{K}}^{\mathrm{T}}\mathbf{R} \tag{2}$$

The remaining problem is to quantify $\overline{\mathbf{K}}$ by a calibration procedure, which in view of the afore-mentioned restrictions can be kept very simple.

The condition of the matrix $\overline{\mathbf{K}}$ [2] is the factor with which the measurement error may be amplified in the analytical result, where

$$||\Delta c||/||c|| \leqslant \mathrm{cond}(\overline{\mathbf{K}})[(||\Delta K||/||K||) + (||\Delta R||/||R||)] \tag{3}$$

and $||c||$ and $||K||$ are the norms of the vector $\mathbf{c}$ and the $\mathbf{K}$ matrix, respectively. Although the ordinary least-squares method seems very attractive at first sight, a serious drawback becomes apparent when an extra measurement is made at a $(NS + 1)$th sensor. Of course, the solution of the problem is still given by Eqn. (2), but with one important distinction. Prior information is now available because an estimate $\hat{\mathbf{c}}$ of the true concentration c is available. However, Eqn. (2) does not provide any means to use that information. Although algorithms which do take into account this prior information have been in use in other disciplines for some time, their applications in analytical chemistry are more recent [3, 6].

The general form of a recursive algorithm is:

new estimate $\hat{\mathbf{c}}$ = (known function of) old estimate + correction

In terms of the above analytical problem, the "old" estimate is based on NS measurements, and the new one on $(NS + 1)$ measurements. The value of the correction term depends on the information obtained from the last, $(NS + 1)$th, measurement. Poulisse [3] clearly demonstrated that complex multicomponent systems can be solved elegantly in this way, and explained the

mathematical background of recursive estimation methods, which will be summarized below in discussion of the recursive generalized standard addition method.

Solution 2: Standard addition method

Many multicomponent analyses are not easily calibrated, because the sensitivity coefficients of the analytes depend on the constitution of the sample. In practice, the preparation of good standards, closely matched to the sample, is often very difficult or impossible, and this problem is often circumvented by calibrating directly in the sample, following the standard addition method. The response for a sample with NA analytes at sensor i is

$$R_i = \sum_{j=1}^{NA} k'_{ij} c_j + e_i \tag{4}$$

When analyte 1 is added to the sample (under the condition of constant volume), the change in the response is

$$\Delta R_i = \sum_{j=1}^{NA} k'_{ij} c_j + k'_{i1} \Delta c_1 + e_i \tag{5}$$

Further additions and calculation of the regression coefficient between ΔR_i and Δc_1 gives an estimate $\hat{k}'_{i1}$ of k'_{i1}. From Eqn. (4), however, it is clear that the concentration c_1 of the first analyte in the sample can be calculated only if the other analytes do not contribute to the response ($k'_{ij(j \neq 1)} = 0$) and if there is no background. This means that the ordinary standard addition method yields reliable results only when the sensor is specific for the analyte measured.

Solution 3: Generalized standard addition method

Difficulties arising from the lack of selectivity of the sensor when the standard addition method is applied, can be overcome by adding standards of all NA analytes (sequentially or simultaneously) to the sample; this allows calculation of all sensitivity coefficients, k'_{ij} ($j = 1, NA$). Because the concentrations of NA analytes are determined, responses at $NS \geqslant NA$ sensors should be measured in order to obtain a solvable system. This is the basic principle of the generalized standard addition method (GSAM), developed by Saxberg and Kowalski [1]. The method consists basically of two steps: (1) calibration by adding standard solutions of all analytes and calculation of the sensitivity coefficients of all analytes at all sensors (**K** matrix); (2) calculation of the analyte concentrations from the initial responses of the sample measured at all sensors. Details of the method and the mathematical expressions have been given [1, 2].

The generalized standard addition method is a powerful calibration tool. Many interferences and matrix effects can be handled by GSAM, so that

complex sample preparations can be avoided. In order to minimize the number of additions, and to ensure the validity of the calibration, the additions should be stopped as soon as the sensitivity coefficients are known with sufficient precision, or when the assumed linear model between concentration and response is no longer valid. This requires coupling of the experimental design with the outcome of the measurements during the calibration. The present form of GSAM does not provide these possibilities because results are evaluated only after completion of all additions.

Solution 4: Recursive generalized standard addition method

The problem of the calibration at one sensor (i) will be discussed first. For a sample with NA analytes, the initial response (before the additions) is

$$R_{i,s} = k_{i,1}c_{1,s} + k_{i,2}c_{2,s} + \cdots + k_{i,NA}c_{NA,s} + e_i \tag{6}$$

Because the numbers of moles are additive, instead of concentrations, volume-corrected responses and concentrations are introduced [2]: $Q_{i,s} = R_{i,s}V_0$ and $n_{i,s} = c_{i,s}V_0$, where V_0 is the initial volume. Equation (6) then becomes

$$Q_{i,s} = k_{i,1}n_{1,s} + k_{i,2}n_{2,s} + \cdots + k_{i,NA}n_{NA,s} + e_i$$

or in matrix notation

$$Q_{i,s} = \mathbf{n}_s^T \cdot \mathbf{k}_i + e_i$$

The vector of the sensitivity coefficients, $\mathbf{k}_i$, will be determined in a recursive calibration procedure.

If all the analytes are added simultaneously in a single addition with a total volume Δv, and the added number of moles of analyte i is $\Delta n_{1,i}$ then the change of the response at sensor i after the addition is

$$\Delta Q_i = R_i(V_0 + v) - Q_{i,s} = (\Delta n_{1,1}\ \Delta n_{1,2} \cdots \Delta n_{1,NA}) \begin{pmatrix} k_{i,1} \\ k_{i,2} \\ \cdot \\ \cdot \\ k_{i,NA} \end{pmatrix} = \Delta \mathbf{n}_1^T \mathbf{k} + e_i$$

In order to start the recursive estimation procedure for $\mathbf{k}$, it is necessary to initialize two parameters: (1) an estimate of the sensitivity coefficients, e.g., zero or $\hat{\mathbf{k}}(0)$; and (2) an estimate of the precision of the estimates of the sensitivity coefficients. These estimates are the diagonal elements of the matrix $\overline{\mathbf{P}}(0)$. The off-diagonal elements are set to zero. With $\Delta Q_i(1)$ in hand, a first estimate of the sensitivity coefficients can be obtained by applying the following recursive filter [3]:

$$\hat{\mathbf{k}}(1) = \hat{\mathbf{k}}(0) + \mathbf{g}(1)\,[\Delta Q_i(1) - \Delta \mathbf{n}^T(1)\,\hat{\mathbf{k}}(0)]$$

$$\mathbf{g}(1) = \overline{\mathbf{P}}(0)\,\Delta \mathbf{n}(1)\,[e(1) + \Delta \mathbf{n}^T(1)\,\overline{\mathbf{P}}(0)\,\Delta \mathbf{n}(1)]^{-1}$$

$$\overline{\mathbf{P}}(1) = \overline{\mathbf{P}}(0) - \mathbf{g}(1)\,\Delta \mathbf{n}^T \overline{\mathbf{P}}(0) \tag{7}$$

where $\hat{\mathbf{k}}$ is a $NA \times 1$ vector of the sensitivity coefficients, $\Delta\mathbf{n}$ is a $NA \times 1$ vector of added moles, $\mathbf{g}$ is a $NA \times 1$ gain vector, $\mathbf{P}$ is a $NA \times NA$ variance–covariance matrix, and $e(1)$ is a scalar for weighting the responses. The expression $\Delta\mathbf{n}^{\mathbf{T}}(1)\,\hat{\mathbf{k}}(0)$ is simply a predictor of the response $\Delta\hat{Q}_i(1)$. The difference $\Delta Q_i(1) - \Delta\hat{Q}_i(1)$ thus represents the difference between the actual measured response (volume-corrected) and the predicted response, based on the last $\hat{\mathbf{k}}$ estimate. This term is called the innovation (*In*).

After the second addition, the estimate $\hat{\mathbf{k}}$ is updated by using Eqn. (7) where $\hat{\mathbf{k}}(0)$ and $\mathbf{P}(0)$ are replaced by $\hat{\mathbf{k}}(1)$ and $\mathbf{P}(1)$, respectively, and $\hat{\mathbf{k}}(2)$, $\mathbf{g}(2)$ and $\mathbf{P}(2)$ are calculated. It is clear that when the algorithm converges to the correct value of $\mathbf{k}$, the average innovation should approach zero, with a standard deviation equal to the standard error of the measurements. Therefore, monitoring of the innovation during the additions will provide valuable information for deciding on the progress of the additions.

First, as soon as the estimates of the $\mathbf{k}$ values converge to a steady state, no further improvement of the $\mathbf{k}$ values obtained can be expected, and additions at that sensor can reasonably be stopped. Expressed in the mathematical terms of Eqn. (7), this means that the value of the gain factor ($\mathbf{g}$) by which the innovation (*In*) is multiplied for correcting the old estimate, becomes relatively small. The reason is that the variance–covariance ($\mathbf{P}$) matrix of the errors of estimation during the additions becomes small in comparison with the measurement error, i.e., a low weight is assigned to a new measurement. Another criterion for stopping the additions may be found by monitoring the diagonal elements of the $\mathbf{P}$ matrix which indicate the precision of the estimated values after each addition. When the precision has reached the desired limits (to be calculated from evaluation of the error propagation), further additions are not necessary. The criterion to stop additions because $\mathbf{k}$ values have reached sufficient accuracy is not critical. As long as the underlying linear model remains valid, a few extra additions have no negative effect on the analytical result. The performance of this stop criterion is still under investigation; details and results will be reported later.

Secondly, a warning system for a faulty model (deviation from linearity) can be achieved either by a sequential test on *In* or by testing the average *In* value. For the sequential test, when *In* reaches values that fall outside the range expected on the basis of both the measurement error and the confidence limits of the $\mathbf{k}$ value available so far, this indicates that the model may be in error. Confidence limits for the innovation (*In*) can be calculated as follows:

$$In(n) = \Delta Q_i(n+1) - \mathbf{k}(n)\,\Delta\mathbf{n}(n+1) = \Delta Q_i(n+1) - \sum_{j=1}^{NA} \hat{k}_{ij}\Delta n_j$$

$$s^2[In(n)] = s^2_{\Delta q} + \sum_{j=1}^{NA} s^2(k_{ij})\,\Delta n_j \text{ (when the additions are error-free)}$$

$$s^2(In(n)) = s^2_{\Delta q} + \sum_{j=1}^{NA} P(j, j)\Delta n_j$$

where $s^2_{\Delta q} = s^2_R [V(n)^2 R(n)^2 + v(0)^2 R(0)^2]$ (8)

The following decision tree has been designed.

$In(n) \geqslant 2.6\ s(In(n))$: this is an indication that the measurements deviate from the model. Do one more addition: if the innovation is still too high, stop the additions.

$In(n) < 2.6\ s(In(n))$: make another addition.

The test on the averaged *In* value is very similar to detecting drift in a random time series by examining the cumulative sum (cusum) of the data. In the situation where the model is valid, and the correct k-value is estimated, the residuals should behave as a random series, and therefore the cusum should fluctuate about a straight horizontal line. In contrast, when the model is invalid, the residuals between model and response (i.e., the innovation) will keep equal signs. Consequently, in a plot of the running sum of the *In* values over the additions, a defective model will be detected by a positive or negative drift of the cumulated innovation from its stationary value (see Fig. 2B below). Widely used methods for deciding whether or not a trend is significant, are the V-mask [7], and Trigg's monitoring technique [8]. The latter was preferred here because of its easy implementation and short warning time. The ratio (T_n) of the moving averages of the *In* value and absolute value of the innovation are monitored, where $e(n) = In(n + 1) - C(n)$, $C(n) = \alpha(In(n)) + (1 - \alpha)C(n - 1)$, $\bar{e}(n) = \alpha e(n) + (1 - \alpha)\bar{e}(n - 1)$, $MAD(n) = \alpha e(n) + (1 - \alpha)MAD(n - 1)$, and $T_n = \bar{e}(n)/MAD(n)$ with $\alpha = 0.1$ or 0.2.

Initial values to start the monitoring procedure are: $e(0) = 0$, $\bar{e}(0) = 0$, and $MAD = (2/\pi)^{1/2} s_{\Delta q}$, where $s^2_{\Delta q}$ is given by Eqn. (8). Decisions on the trend are made on the value of T_n, which oscillates between −1 and +1. The more T_n differs from zero, the more significant is the trend [8].

The calibration procedure is easily extended to more sensors. A recursive filter is assigned to each sensor. During the additions (several analytes simultaneously or separately), all filters are run in parallel. A system which is calibrated at *NS* sensors therefore requires *NS* independent recursive filters. In many instances the model for the response (Eqn. 6) has to be completed with a term describing the background at a particular sensor:

$$R_{i,s} = k_{1,i}c_{1,s} + k_{i,2}c_{2,s} + \cdots + k_{i,NA}c_{NA,s} + R_{i,b} + e_i \quad (9)$$

where $R_{i,b}$ is the portion of background in the signal $R_{i,s}$. When the background ($R_{i,b}$) remains constant during the additions of the standards (e.g., a mismatch between the cells in an u.v.—visible spectrophotometer), the contribution of the background in the response can be estimated by starting the calibration procedure with simple dilutions.

If Eqn. (9) is rewritten for volume-corrected responses, then

$$V_0 R_{i,s} = k_{1,i}V_0 c_{1,s} + k_{i,2}V_0 c_{2,s} + \cdots + k_{i,NA}V_0 c_{NA,s} + V_0 R_{i,b} + e_i$$

or

$$Q_{i,s} = k_{1,i}n_{1,s} + k_{i,2}n_{2,s} + \cdots + k_{i,NA}n_{NA,s} + Q_{i,b} + e_i$$

Because the number of moles ($n_{j,s}$) of the analytes in the sample remains unchanged ($\Delta n_j = 0$) on dilution, the change of the volume-corrected response (ΔQ_i) is

$$\Delta Q_i = R_{i,s}(V_0 + \Delta v) - Q_{i,s} = (V_0 + \Delta v)R_{i,b} - V_0R_{i,b}$$

$$\Delta Q_i = \qquad\qquad = \Delta vR_{i,b} \tag{10}$$

which demonstrates that $R_{i,b}$ can be estimated by measuring ΔQ as a function of Δv, the added volume of dilutent. During the calibration step, the responses should be corrected for background before volume correction: ΔQ after the nth addition (total volume V_n) is

$$\Delta Q_{i,n} = V_n(R_{i,n} - R_{i,b}) - (V_0)(R_{i,s} - R_{i,b})$$

$$\text{or } \Delta Q_{i,n} = (V_n - V_0)\,R_{i,b} + V_nR_{i,n} - V_0R_{i,s} \tag{11}$$

The whole calibration procedure can be summarized as follows.

(1) Measure the initial responses ($R_{i,s}$) of a sample aliquot at all candidate sensors.

(2) Dilute the sample and estimate $R_{i,b}$ from $\Delta Q_i = f(\Delta v)$ (Eqn. 9) by means of a recursive filter.

(3) Start the additions (one or more analytes at a time), measure the responses at the selected sensors, and calculate the volume-corrected and background-corrected responses. There are two options:

(a) total GSAM, $\Delta Q = Q_{\text{after addition}} - Q_{\text{sample}}$
(b) incremental GSAM, $\Delta Q = Q_{\text{after addition}} - Q_{\text{before addition}}$

(4) Estimate the K values of all analytes at all sensors, by using a recursive filter for each sensor.

(5) Evaluate the convergence of K:
convergence is slow? $\overset{\text{yes}}{\rightarrow}$ STOP (no further improvement may be expected)
↓no
"*In*" drifting off-zero? $\overset{\text{yes}}{\rightarrow}$ STOP (model is defective)
↓no

(6) Return to (3)

(7) Calculate the analyte concentrations.

EXPERIMENTAL

The recursive estimation procedure and the performance of the proposed stop criteria were demonstrated experimentally on a simulated system and on a u.v.—visible spectrophotometric determination of copper and nickel in a model system.

Simulated calibrations

The simulation model designed generates responses (R_i) at a set of sensors with adjustable characteristics: sensitivity, standard deviation of the noise, linearity and background. Calibration options are total GSAM and incremental GSAM, both in a separate mode and a joint mode. In the separate mode, analytes are added one at a time, the additions of each analyte being started in other aliquots of sample. In the joint mode, analytes are added in

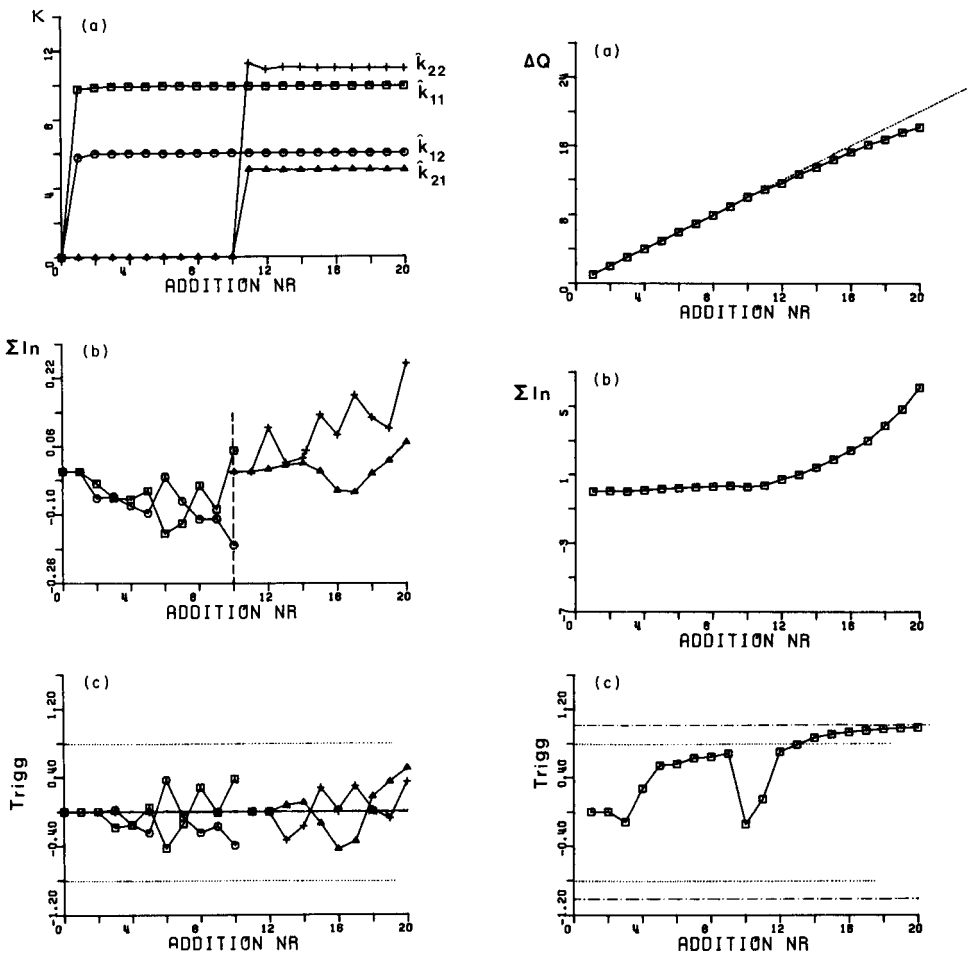


Fig. 1. Recursive total GSAM in the separate mode with 2 analytes and 2 sensors. $K = \begin{pmatrix} 10 & 6 \\ 5 & 11 \end{pmatrix}$, v_0 = 5 ml of concentrated standard solutions (c_1 = 0.1 mol l^{-1} and c_2 = 0.1 mol l^{-1}). (a) Estimation of K; (b) cumulative innovation (Σ In); (c) Trigg's values.

Fig. 2. Detection on non-linearities in recursive total GSAM for 1 analyte and 1 sensor with K = 10 l mol^{-1} cm^{-1}. (a) A non-linear calibration curve, $\Delta Q = K\,\Delta n(1 - a\,i^b)$, $a = 2 \times 10^{-4}$, $b = 2$; (b) cumulative innovation; (c) Trigg's values.

the same sample aliquot [2]. The general course of the estimation process of the sensitivity coefficients is demonstrated on a system of two analytes measured at two sensors (Fig. 1a). The addition scheme consisted of a sequence of 20 additions per analyte. All recursive estimators are started with $\overline{\mathbf{P}} = \overline{\mathbf{I}} \cdot 10^6$ and $\hat{\mathbf{k}} = 0$. The general behaviour of the estimation process of $\mathbf{k}$ is an initially fast decrease of the diagonal terms of $\overline{\mathbf{P}}$ during the first additions causing a stabilization of $\hat{\mathbf{k}}$ which in course of the estimation process becomes less influenced by the fluctuations in the readings (Table 1).

The innovation (Eqn. 7), which is the difference between the estimated next response (on the basis of the last estimate of the $\mathbf{k}$ value) and the actually measured response, very rapidly reaches values within the range expected on the basis of the standard deviation of the response ($s_{\Delta Q}$). In this ideal situation of the absence of anomalies, the cumulative sum of the innovation fluctuates about the zero baseline, and the Trigg's values are bounded by values close to zero (Fig. 1c), as a confirmation that drift is absent.

A primary source of errors when calibration is done according to the standard addition method is the lack of information about the background. At first sight, it might be assumed that the estimates of the $\mathbf{k}$ values would not be affected by the presence of background. Equation (11), however, shows that the volume correction of the response results in the estimated k value depending on the background level. That the proposed method of starting the calibration with dilutions in order to estimate the background is adequate is demonstrated in Table 2, where apparent calibration failures are completely corrected.

Defective models are another important source of errors. In many instances, calibration graphs may bend towards the concentration axis, and may produce wrong results when the validity of the supposed underlying model has not

TABLE 1

Estimated sensitivity coefficient ($\hat{k}$), estimation error ($s_{\hat{k}}^2$), volume-corrected response (ΔQ), standard deviation of the volume-corrected response ($s_{\Delta Q}$), innovation (In), cumulative innovation ($\Sigma\, In$) and gain (g) during the additions.
[System: k = 10 l mol^{-1} cm^{-1}, $N(0,1\%)$]

Addition	ΔQ	$s_{\Delta Q}$	$\hat{k}$	$s_{\hat{k}}^2$	In	$\Sigma\, In$	g
1	1.028	—	10.26	0.126	—1.023	—	—9.98
2	2.069	0.165	10.32	0.036	—0.018	—0.018	—3.58
3	2.962	0.152	10.08	0.016	0.134	0.116	—1.78
4	4.029	0.159	10.08	0.010	0.004	0.120	—1.00
5	5.034	0.175	10.07	0.007	0.005	0.125	—0.64
6	5.931	0.186	10.02	0.005	0.114	0.239	—0.45
7	7.084	0.205	10.04	0.004	—0.067	0.172	—0.31
8	8.082	0.223	10.06	0.003	—0.046	0.126	—0.24
9	8.959	0.238	10.04	0.002	0.090	0.217	—0.19
10	10.09	0.259	10.05	0.002	—0.051	0.165	—0.15

TABLE 2

Estimation of background levels by dilution.
(System: k = 10 l mol^{-1} cm^{-1}, $N(0,1\%)$, initial conc. c_0 = 0.01 mol l^{-1}, initial volume v_0 = 5 ml, standard conc. c_s = 0.1 mol l^{-1}, volume additions Δv = 1 ml. Design: 10 dilutions (Δv = 1 ml), 20 additions)

Background $R_{0,b}$ (absorbance)	Estimated background $\hat{R}_{0,b}$	Estimated $\hat{k}$	
		Corrected	Uncorrected
0.020	0.021	9.95	10.19
0.040	0.042	9.97	10.41
0.060	0.060	10.01	10.61
0.080	0.083	9.91	10.80
0.100	0.099	9.98	11.03
0.120	0.118	10.00	11.17
0.140	0.138	10.04	11.40
0.200	0.198	10.00	11.94

been checked. The power of monitoring the innovation by means of Trigg's method for detecting a defective model was investigated on two model systems: (1) a system where the k value at a given addition (i) is artificially decreased by a given factor; (2) a system where the deviation from linearity of the k value gradually increases with the sequence of the additions according to the model $\Delta Q = k\ \Delta n(1 - a\ i^b)$, where a and b are constants and i stands for the addition number. Figure 2(a) shows the effect obtained for a particular case ($a = 2 \times 10^{-4}$, $b = 2$), with the corresponding shape of the cumulative innovation and associated Trigg values. The drift of the cumulative innovation introduced by the model is apparent. Of course, the detection of deviations from linearity depends on both the magnitude of that deviation and the noise level. Table 3 lists how well deviations from linearity are detected for various magnitudes of the deviation. It demonstrates the better sensitivity of Trigg's monitoring technique over the sequential method for the detection of minor deviations. Table 3 also lists the effect of stopping the additions after the detection of a non-linearity. A significantly better sensitivity coefficient is obtained, even for very minor deviations from linearity.

Measurements on a Cu/Ni model system

The concentrations of copper and nickel in a sample containing 0.010 mol l^{-1} copper(II) nitrate, 0.010 mol l^{-1} nickel nitrate and 0.022 mol l^{-1} EDTA, in a phosphate-buffered solution at pH 7.1 were determined by using a GSAM design in the separate mode. Responses (absorbances) were measured at 378, 588, 625, 735 and 899 nm. Two wavelengths (899 and 625 nm) were chosen on a slope of the spectra for copper and nickel, in order to generate non-linear calibration graphs. The standard solutions for the additions were: I, 0.100 mol l^{-1} copper nitrate + 0.11 mol l^{-1} EDTA; II, 0.100 mol l^{-1}

TABLE 3

Detection of a defective model (non-linearity)[a]
Model I: $\Delta Q = k\ \Delta n(1 - a\ i^b)$ (see text), k = 10 l mol^{-1} cm^{-1}, $N(0,1\%)$, 20 additions

Method	*Average warning time*, i *(20 runs)*				
	$b = 1$			$b = 2$	
	$a = 5 \times 10^{-4}$	10^{-3}	2×10^{-3}	10^{-4}	2×10^{-4}
Sequential	i 20(0)	17.8(2.6)	12.4(1.3)	20(0)	17.7(2.8)
	ai^b —	0.018	0.025	—	0.063
Trigg	i 18.8(2.7)	14.1(3.5)	11.0(1.8)	18.1(3.9)	14.1(4.1)
	ai^b 0.0094	0.014	0.022	0.033	0.040
	Average k *value (20 runs)*				
Sequential	9.91(0.02)	9.87(0.05)	9.87(0.01)	9.88(0.03)	9.79(0.03)
Trigg	9.92(0.02)	9.90(0.04)	9.88(0.04)	9.88(0.03)	9.83(0.05)
No warning	9.91(0.02)	9.83(0.03)	9.66(0.03)	9.88(0.03)	9.76(0.03)

Model II: $i \leqslant 10\ \Delta Q = k\ \Delta n$, $i > 10\ \Delta Q = k(1 - a)\ \Delta n$, $N(0,1\%)$, k = 10 mol l^{-1} cm^{-1}

Method	*Average warning time*, i *(20 runs)*		
	$a = 0.01$	0.02	0.03
Sequential	19.5(1.8)	17.2(3.8)	10.8(2.3)
Trigg	17.1(4.8)	14.6(3.9)	12.5(2.4)
	Average k *value (20 runs)*		
Sequential	9.93(0.03)	9.90(0.04)	9.96(0.05)
Trigg	9.93(0.02)	9.94(0.06)	9.93(0.05)
No warning	9.92(0.03)	9.88(0.02)	9.81(0.02)

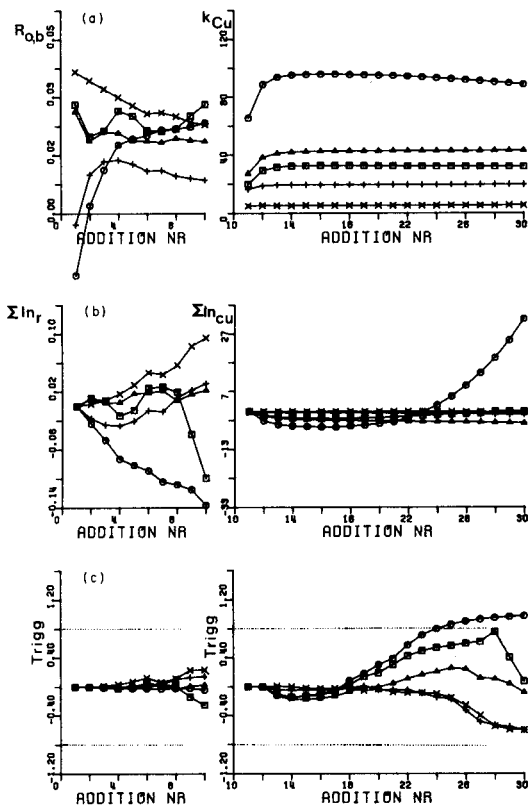
[a]The numbers in parentheses are the standard deviations from 20 runs.

nickel nitrate + 0.11 mol l^{-1} EDTA. Both solutions were buffered with the phosphate buffer at pH 7.1.

The design of the GSAM in the separate mode was as follows: (1) sample aliquot 1 had an initial volume of 22 ml and ten 0.5-ml additions of distilled water were made followed by twenty 0.5-ml additions of solution I; (2) sample aliquot 2 was similar to aliquot 1, but twenty 0.5-ml additions of solution II were made. All absorbances were measured with a Varian model 1800 u.v.—visible spectrophotometer (slit width 0.5 nm).

Figure 3(a, b) displays the course of the estimates of the background level (absorbance) and the estimates of the sensitivity coefficients of copper and nickel at the five wavelengths. As expected for such a spectrophotometric determination, the monitoring of the cumulative innovation confirms that the linear model is obeyed at most wavelengths, except there is some doubt for $k_{Cu,735}$ and $k_{Ni,899}$. Quantitative evaluation of the innovation by monitoring Trigg's values showed a significant curvature for $k_{Cu,735}$ only at the 12th addition.

Under the operating conditions, the addition procedure of both analytes would have been stopped much earlier if the tentatively-proposed procedure for detecting a steady state of the k estimates had been applied. When the magnitude (norm) of the changes of the k values, $||k_{i+1} - k_i||$, is monitored at all sensors during the additions (Table 4), it is clear that $||k_{i+1} - k_i||$ for nickel reaches a low and steady value at about the 8th addition; so does $||k_{i+1} - k_i||$ for copper but the value is higher. The calibration results and calculated analyte concentrations are summarized in Table 5. The condition of the matrix of sensitivity coefficients of the Cu/Ni system, Cond($\overline{\mathbf{K}}$) [2], equals 5.08. According to Eqn. (3), this means that in the worst case, the relative error in the initial responses and **k** values is amplified by a factor of 5 in the analytical result. On the assumption that the initial responses were measured within a 1% relative precision, $||\Delta R||/||R||$ is 0.01. The square roots of the eigenvalues of the positive definite matrix $\overline{\mathbf{K}}^T\overline{\mathbf{K}}$ are usually called the singular values of $\overline{\mathbf{K}}$. The Euclidian norm $||K||$ of a matrix $\overline{\mathbf{K}}$ is defined as the biggest singular value of $\overline{\mathbf{K}}$. For the $\overline{\mathbf{K}}$ matrix and $\overline{\mathbf{K}}$ matrices listed in Table 5, one finds $||\Delta K||/||K|| = 1.39/109.8 = 0.012$. Consequently, $||\Delta c||/||c||$ 5.08 (0.01 + 0.012) = 0.115. The analyte concentrations (Table



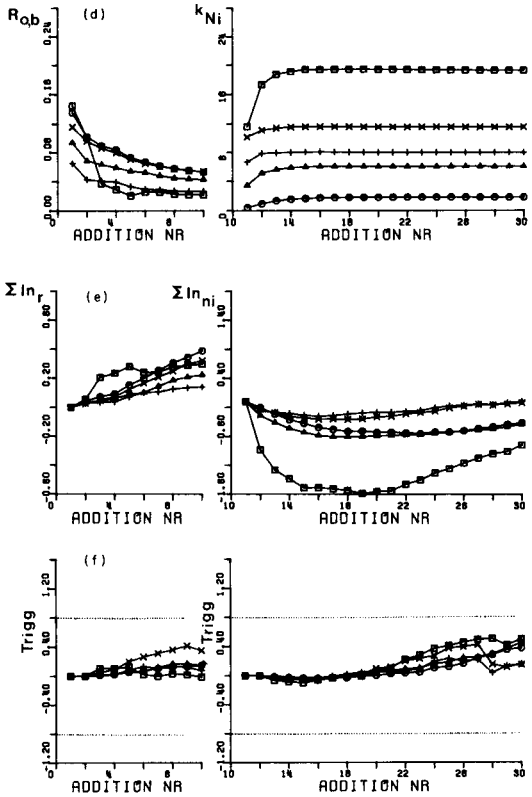


Fig. 3. Recursive GSAM in the separate mode with 2 analytes (Cu/Ni) and 5 sensors. v_0 = 4 ml; v = 1 ml; added standard solutions, c_{Cu} = 0.1 mol l^{-1} and c_{Ni} = 0.1 mol l^{-1}. (a) Estimation of background levels, $R_{0,b}$ (absorbance) and k_{Cu} at wavelengths of (□) 999, (○) 735, (▲) 625, (+) 588, (*) 378 nm; (b) same as (a) but for nickel; (c) cumulative innovation during the estimation of background (Σ In_r) and k values (Σ In) for copper; (d) same as (c) but for nickel; (e) Trigg's values during the estimation of background and k values for copper; (f) same as (e) but for nickel.

TABLE 4

Norm of the changes of the estimated k values of the Cu—EDTA and Ni--EDTA complexes during the additions (i = 1—20) at 378, 588, 625, 735 and 899 nm

Addition	$\|\|k_{i+1} - k_i\|\|_{Cu}$	$\|\|k_{i+1} - k_i\|\|_{Ni}$	Addition	$\|\|k_{i+1} - k_i\|\|_{Cu}$	$\|\|k_{i+1} - k_i\|\|_{Ni}$
1	75.6	17.1	11	0.24	0.031
2	27.4	6.2	12	0.48	0.011
3	6.4	1.5	13	0.59	0.045
4	1.9	0.56	14	0.59	0.021
5	0.54	0.32	15	0.65	0.035
6	0.29	0.12	16	0.72	0.024
7	0.14	0.04	17	0.73	0.023
8	0.30	0.038	18	0.77	0.013
9	0.22	0.044	19	0.83	0.012
10	0.27	0.031	20	0.89	0.014

TABLE 5

Calibration of the Cu/Ni model system

Wavelength (λ)		899	735	625	588	378	
$\bar{K}$	Cu	31.99	88.86	42.98	19.59	5.356	$(\lambda_1)^{1/2} = 109.8 = \|\|K\|\|$
	Ni	19.26	1.831	6.035	8.008	11.56	$(\lambda_2)^{1/2} = 21.6$
							$\mathrm{Cond}(\bar{K}) = 5.08$
$\bar{s}_K$	Cu	0.380	0.90	0.36	0.20	0.123	
	Ni	0.385	0.677	0.344	0.203	0.151	$(\lambda_1)^{1/2} = 1.39 = \|\|\Delta K\|\|$
$\mathbf{R}_0$	Aliquot 1	0.495	0.948	0.458	0.258	0.149	
	Aliquot 2	0.528	0.938	0.457	0.264	0.153	

TABLE 6

Analyte concentrations calculated from initial responses

Analyte	Estimated conc. [ĉ]		
	Aliquot 1	Aliquot 2	True conc.
Cu	0.010	0.00852	0.010
Ni	0.00977	0.0101	0.010
$\|\|\Delta c\|\|$	2.3×10^{-4}	1.48×10^{-3}	$\|\|c\|\| = (2 \times 0.01^2)^{1/2} = 1.41 \times 10^{-2}$
$\|\|\Delta c\|\|/\|\|c\|\|$	0.016	0.105	

6) calculated from the initial responses of both sample aliquots are indeed within the expected relative error, as is indicated by the $||\Delta c||/||c||$ values obtained for both estimations.

Conclusions

The proposed modification of the generalized standard addition method (GSAM) to give a recursive version by means of the Kalman filter provides a novel approach to calibration. The "real-time" conversion of the response obtained for a given measurement to new system information (e.g., sensitivity coefficients) allows the further design of the measurements to be decided. At every addition, estimates of the sensitivity coefficients and their precision and of the validity of the underlying model, are available. Feedback of such information to the experimental design is an indispensable feature for the design of self-calibrating (and optimizing) analytical procedures, which will be the precursors of intelligent analyzers.

REFERENCES

1 Bo E. H. Saxberg and B. R. Kowalski, Anal. Chem., 51 (1979) 1031.
2 C. Jochum, P. Jochum and B. R. Kowalski, Anal. Chem., 53 (1981) 8592.

3 H. N. J. Poulisse, Anal. Chim. Acta, 112 (1979) 361.
4 C. B. M. Didden and H. N. J. Poulisse, Anal. Lett., 13(A11) (1980) 921.
5 H. N. J. Poulisse and P. Engelen, Anal. Lett., 13(A14) (1980) 1211.
6 P. F. Seelig and H. N. Blount, Anal. Chem., 51 (1979) 327.
7 G. A. Barnard, J. R. Stat. Soc., Part B, 21 (1959) 239.
8 D. L. Massart, A. Dijkstra and L. Kaufman, Evaluation and Optimization of Laboratory Methods and Analytical Procedures, Elsevier, Amsterdam, 1978.

Analytica Chimica Acta, 150 (1983) 87—103
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

RELATIONSHIPS BETWEEN CHROMATOGRAPHIC RESPONSE FUNCTIONS AND PERFORMANCE CHARACTERISTICS

WOLFHARD WEGSCHEIDER*, ERNST P. LANKMAYR and MATTHIAS OTTO[a]

Institute for Analytical Chemistry, Micro- and Radiochemistry, Technical University Graz, A-8010 Graz, Technikerstraße 4 (Austria)

(Received 11th November 1982)

SUMMARY

Numerous objective functions used with formal strategies of experimental optimization are reviewed. It is shown that several of them give no meaningful indication of the quality of separation unless various other chromatographic parameters, e.g., number of plates and asymmetry, are specified. Another group of chromatographic response functions is characterized by the need for operator-defined weighting factors that not only establish the exact location of the optimal conditions but also limit the usefulness of the optimization in general. Performance characteristics, like accuracy, precision and separation time, can be related to numerical values of the response functions. Important chromatographic situations for the separation of two and ten components are simulated. It is concluded that only functions derived from Kaiser's peak separation number are directly dependent on performance characteristics.

Recent research on formal strategies for improving analytical methods has been extended to techniques such as spectrophotometry [1], atomic absorption spectrometry [2], inductively-coupled plasma emission spectrometry [3], flow injection analysis [4], energy-dispersive x-ray fluorescence spectrometry [5], electrochemistry [6], n.m.r. [7] and chromatography [8, 9]. In all cases, the examination of optimal experimental conditions is preceded by an implicit or explicit selection of a response function; this is relatively straightforward in simple systems, but the response functions become more complex when several components are measured simultaneously.

In chromatographic analysis where the number of components that can be separated and detected by any particular system may vary greatly, the problem appears to be more confusing than in many other cases. Inadequate performance of a known chromatographic response function (CRF) seems to have spurred many researchers investigating the optimization of an experimental variable to devise another CRF on an ad hoc basis. Multimodal response surfaces have kept much of the attention focussed on the strategy of globally estimating the optimum and thus neglecting the vital question

[a]Permanent address: Analytical Centre, Department of Chemistry, Karl-Marx-University, Leipzig, German Democratic Republic.

0003-2670/83/$03.00

of defining the CRF on a rational, unambiguous basis free of operator bias.

The present work selects some of the previously published response functions subject to the constraints that no estimates of quantities such as separation time, number of peaks and obtainable performance are required prior to optimization, and that the evaluation of the response functions appears feasible from complex chromatograms at least to a first approximation. While the first requirement can be judged on an absolute basis, the second could introduce bias. It will be seen, however, that this poses no serious restriction on the validity of the conclusions. Performance of these CRFs has to be seen in the light of performance characteristics in analytical chemistry in general. Although no one list of performance characteristics resembles another in all its features, there is little dissent that the most important features are accuracy, precision, selectivity, time needed for the analysis, and cost [10, 11, 13]. The first four are addressed below in their relation to the CRFs. The cost criterion is studied only insofar as it is related to time. The CRFs chosen are tested, for a couple of typical situations involving overlapped peaks, against a standard integration procedure for obtaining an estimate of accuracy. Precision is judged from noisy chromatograms, and separation time is defined as the time from the injection of a sample to the time when the detector signal indicating the last species has dropped to 5% of its peak value for this species.

THEORETICAL CONSIDERATIONS CONCERNING THE PERFORMANCE OF CHROMATOGRAPHIC RESPONSE FUNCTIONS

The large number of suggested chromatographic response functions made a preselection prior to the actual assessment of the merits of the CRFs imperative. A limited, yet representative collection of the CRFs proposed for multicomponent separations is given together with the area of their original application in Table 1. To facilitate a general discussion of common and unique features of the CRFs, the functions have been recast in common terms; those not expressed in a mathematical form in the original papers (CRF 1 and 6), have been converted to such a form. No attempt is made to list all applications of each CRF; only a characteristic case is given. Table 1 shows that considerable effort has gone into optimization problems in chromatography, and in each reference a specific narrow problem has been tackled and frequently also been solved without leading to a general solution of the optimization problem, not even to a consensus on chromatographic goals as expressed in numerical terms by a CRF. With respect to formal optimization strategies, the Simplex search predominates despite its drawbacks in multimodal situations; the probable reason is that only in rare cases can complex "real" samples be modelled for the behaviour of all their detectable components in any chromatographic system, and so the experimental conditions giving the globally optimal separation cannot be predicted

TABLE 1

Important chromatographic response functions with typical applications

No.	Chromatographic response function	System	Experimental variable(s)	Location of optimum by	Ref.
1	$\sup_R \{\min \alpha_{ij}\}$	G.l.c.	Composition of binary stationary phases	Semi-empirical and graphical solution	12
2	$1/t_{95} \sum_i \ln 1/(\Omega_{i-1} + \Omega_{i+1})$	Ion exchange	Ternary mobile phase	Simplex search	9
3	$\sum_j \ln f_j/g_j$	G.l.c.		Simplex search	8
4	$\sum_j \ln (P_j/\tau_0) + \gamma(\tau_M - t_L)$	H.p.l.c.	Parameters describing gradient (4) and flow rate	Simplex search	13
5	$\sum_j (\beta_j \ln R_{sj}/\rho_{oj}) + \gamma(\tau_M - t_L)$	H.p.l.c.	Composition of quarternary mobile phase	Factorial design, empirical model and numerical solution	14
6	$\epsilon_R \{R_{s,\min} > \rho_0\}$	H.p.l.c.	Composition of quarternary mobile phase	Factorial design, empirical model and numerical solution	14
7	$\sum_j R_{sj} + L^{\xi} - \gamma\lvert\tau_M - t_L\rvert - \beta(\tau_0 - t_1)$	H.p.l.c.	Composition of ternary mobile phase, flow rate and parameters describing gradient	Simplex search	15
8	$1/t_{95} \prod_j f_i/(g_j + 2n_j)$	H.p.l.c.	Concentration of organic modifier, pH and buffer concentration	Simplex search	16

in a general way. For samples of medium complexity, approaches to locate global optima have been developed [17–19] to include one or more powerful experimental variables.

Response function 1 selects those conditions as optimal that maximize the selectivity factor $\alpha_{ij} = k'_j/k'_i$ of the solute pair i and j showing the lowest α compared to all the other solute pairs. As the chromatographic conditions change, the solute pair with the smallest α_{ij} may also change. This has been quite a popular criterion as judged from a recent review [20] and will thus be considered in more detail below. Response function 2 [9] sums over all peaks logarithmically weighted for the fractional overlap Ω of neighboring peaks. This value is then normalized to the time needed to elute the last peak to 95% completion. In the absence of a good description of peak shape, it is cumbersome to evaluate and it drives the optimum to very short elution times. For lack of power to recognize components qualitatively, this CRF failed to reflect expectations [9] and is not considered further. Response function 3 [8] represents a multicomponent extension of the "peak separation" number suggested by Kaiser [21] and ranges from minus infinity, if any one pair of solutes shows serious overlap, to zero, if all components are separated to the baseline. For efficient use of this criterion, the number of peaks recorded should either be known in advance or kept track of during optimization, as there is no indication from the CRF whether many substances are fairly well separated or a few are poorly separated. This problem was first recognized by Spencer and Rogers [22]. On devising CRF 4,

Watson and Carr [13] pointed out the relationship of the "peak separation" number and the resolution R_s. In amending CRF 3, these authors [13] postulated that (a) incomplete separations may be sufficient in some cases, (b) an adequate estimate of separation time be available, and (c) there be a sensible way to choose the factor γ to combine the goals of good separation and short separation time. In this and subsequent CRFs, Greek letters are used to indicate factors that can be chosen by the operator at will or according to past experience. Thus π_0 is the desired peak separation, γ a weighting factor and τ_M the maximum elution time permitted. Although CRF 4 is given in the same form as in the original reference [13], the reasoning given in that paper suggests that the authors have set $\gamma \neq 0$ only for $t_L > \tau_M$. In this way, the second term serves only as a penalty function. As the usefulness of CRF 4 seems to be dominated by the proper choice of arbitrary factors, this CRF was not amenable to the design of the present study.

Response function 5 is a logical extension of CRF 4 if fairly symmetrical peaks are expected; the "peak separation" number is replaced by the resolution R_s and normalized to a desired resolution ρ_{oj}. Both ρ and β are then allowed to vary for each pair of solutes. The increasing numbers of user-selected factors as well as the substitution of P_j by $R_{s_{ij}}$ led to difficult situations that were not further exploited by the original authors who, in the same paper [14], suggested what they called "overlapping resolution mapping". This approach, previously explored by Rautela et al. [23] for a clinical analyzer, consists of modelling the response (here the k' values) by an empirical function calculating resolutions for each pair of solutes and indicating graphically the region in space where a preselected minimal resolution is achieved for all pairs. Subject to model bias, this then leaves the operator with a residual region in space where other criteria have to be considered if a clear-cut decision for a particular set of conditions is to be taken. Some of the ambiguities of this approach have been pointed out [24]. As an analogue to CRF 1, but in terms of R_s instead of α, this criterion was considered suitable for the present study. No previous selection of a limiting ρ_0 value has to be made. Although CRF 7 has been applied successfully for optimizing several experimental variables, the need for selecting not only the maximum time permitted (τ_M) and its weighting factor (γ), but also the minimal time permitted for the first component to elute (τ_0) plus its weighting factor (β), along with an exponent (ξ) for rewarding a larger number of peaks detected (L) precluded a realistic assessment of the merits of this function in a general situation.

Response function 8 is extended from the basic concept of CRF 3, while accounting for poor signal-to-noise ratios. The noise level n_j leads to serious penalization even for baseline-separation ($g_j \simeq f_j$) if small peaks are recorded ($g_j \simeq 2\ n_j$). The factor 2 was chosen because peaks exceeding a level of $2\ n_j$ are often considered as detected in chromatography. The time normalization factor $1/t_L$ is to avoid long elution times. As for severely overlapped peaks, no minima are observed between two peaks; the number of peaks detected

has to be counted as in CRF 3, to multiply appropriately the functional value by zero, the lower limiting value of this CRF [16].

Although not always included directly in the CRF, all strategies but CRF 1 account for total elution time in one way or another. A summary of the approaches is given in Table 2. When time considerations are included in the CRF, three out of five CRFs (4, 5, 7) require the estimation of additional factors. When a strict time constraint is imposed [8], it cannot be ruled out that conditions of a superb separation may be missed just by setting this constraint slightly too low. This constraint was not rigorously set by Glajch et al. [14] because the adjustment of k' values is only approximate. Additionally, it is restricted to the particular experimental variable (organic eluent composition) studied and no generalization to other variables is presently available.

SIMULATION STUDIES FOR THE EVALUATION OF SELECTED CHROMATOGRAPHIC RESPONSE FUNCTIONS

The preselection described above led to consideration of only CRFs 1, 2, 3, 6 and 8. The primary criterion of the preselection was the absence of variable, user-selected factors that made the outcome of the optimization result subject to the prior experience of the operator. For all the CRFs, their basic utility as objective functions in optimization schemes has been demonstrated in actual chromatographic experiments without establishing the general advantage of any one of them. Moreover, their functioning in other than the published chromatographic situation has hardly been considered.

For comparison purposes, simulation experiments were done for a couple of simple and more complex optimization problems; the CRFs were calculated, random and systematic errors and total elution time were evaluated, and the values were compared to the recorder trace of the chromatogram. The peak shape was approximated by the exponentially-modified Gaussian equation [25–27] in the form given by Barber and Carr [28]:

$$\frac{h(t)}{h_{max}} = \left(\frac{\pi}{2}\right)^{1/2} * \left(\frac{\sigma}{\tau}\right) * \exp\left\{-\frac{\sigma}{2\tau}\left(\frac{t-t_R}{\sigma} - \frac{\sigma}{\tau}\right)\right\} * \operatorname{erf}\left\{\frac{1}{2^{1/2}}\left(\frac{t-t_R}{\sigma} - \frac{\sigma}{\tau}\right)\right\}$$

TABLE 2

Approaches to the consideration of total elution time

Consideration of elution time	CRF 1	2	3	4	5	6	7	8
In the CRF	no	yes	no	yes	yes	no	yes	yes
Other means	no	no	time constraint	no	no	k' constraint	no	no

where $h(t)$ is the height of the chromatographic peak as a function of time, h_{max} is the maximum height that the peak would have if it were a perfect Gaussian, τ is the time constant of the exponential peak modifier and t_R is the retention time of the peak (first moment). The dependence of peak dispersion (σ) as a function of retention time was taken as $\sigma = t_R(1/N)^{1/2}$. The noise structure imposed on the chromatogram for studying the sensitivity of the CRFs to different noise levels assumed weak stationarity of the process. To keep matters simple, in complying with the requirements for a parsimonious model [29], a first-order autoregressive process of the exponential type was implied for noise simulations [30].

For realistic assessment of accuracy and precision, the data reduction scheme is a variable that has a powerful influence on the results. No particular peak shape was implied at this stage, thus ruling out all quantification algorithms employing deconvolution. Also, no assumptions were made as to the availability of on-line identification of overlapping peaks, which is now gaining increasing acceptance with multi-dimensional detectors. As most of the current quantitative measurements following chromatographic separations are made on commercial systems, it was deemed acceptable to adopt the logic of a commercial integrator and to make its results on well-separated peaks the standard against which the results obtained on overlapped peaks are compared. The sampling frequency is defined as one tenth full-width-at-half-maximum of the peak, a quantity to be specified by the operator and kept constant throughout this study (5 s). The noise level was measured from the chromatogram by giving a lead time of 100 s prior to injection and measuring the noise by subtracting the five lowest from the five highest signals recorded within this interval. The dependence of retention on an (hypothetic) experimental variable for the three cases of two-component separations is depicted in Fig. 1. Cases A and B, differing by the relative

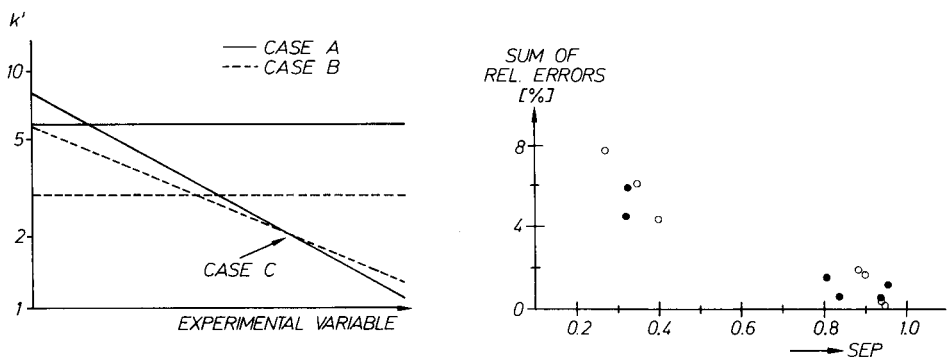


Fig. 1. Three two-component systems characterized by plotting ln k' vs. an arbitrary experimental variable.

Fig. 2. Variation of the sum of absolute systematic errors for both peaks with peak separation (SEP): (○) case A; (●) case B.

retention of complete overlap, were considered for closer inspection of the basic relation of the CRFs to accuracy. As such, CRF 1 gives no indication at all whether its value is sufficient for obtaining a certain degree of accuracy; in case A, only the high-valued end (right side of Fig. 1), and in case B both extremes, of the experimental variable would indicate a very good separation. Relations to accuracy can only be established if certain assumptions are made about the experimental peak shape, the relative size of the peaks and the plate number N of the column. With respect to separation time, the high-valued end gives shorter times; neither this nor a relation to precision can be established from CRF 1. Response function 3, in contrast, is sensitive to relative height (see below), to peak shape (asymmetry) and to plate number N. Separation time and precision cannot be reflected by CRF 3, thus in optimization the shortest time consistent with a good separation could be chosen.

Response function 6 is somewhat analogous to CRF 1: without arbitrary constraint on the maximum of the least separated pair, it gives the highest value in cases A and B on the high end of the experimental variable. There is a relationship between the size of R_s and the achievable accuracy, as long as peak asymmetry is negligible. In multicomponent separations, however, this criterion offers only an estimate of accuracy of the least resolved pair. Precision and total separation time are not reflected by this CRF.

Response function 8 is sensitive to the degradation of accuracy caused by overlap, but by incorporating aspects concerning the total separation time ($1/t_{95}$) and the noise level (n) this relationship is not obvious. To simplify discussion, it is assumed here that the total separation time is constant (an assumption exactly true in cases A and B beyond the reversal of elution order) and that n is negligible, as found for very high S/N ratios. Twenty chromatograms were simulated, equally spaced in the domain of the experimental variable, the peak area was calculated (as outlined in the Waters Associates Data Module manual), and the CRF was recorded for $n = 0$ and $t_{95} = 1$ (this abridged form of CRF 8 was called SEP and is the antilog of CRF 3). Similarly, the two peaks were generated one at a time and the integral was taken as a reference for assessing accuracy. If the sum of the absolute systematic errors is plotted against SEP, the curvilinear relation given in Fig. 2 is obtained. For equally intense peaks, the decrease of the value of SEP is much more severe than the increase in the systematic error associated with complete separation.

Case C (Fig. 1) was constructed to study the influence of the $1/t_{95}$ factor in CRF 8. Figure 3 shows the dependence of the functional value of SEP, t_{95} and the full form of CRF 8 on the value of the experimental variable, most easily identified as elution strength (Φ). In the region of elution strength 0.6—0.75, no minimum is observed in the recorder trace between the two peaks ($N = 7000$, $\tau/\sigma = 0.3$), thus SEP and CRF 8 become zero. Separation is practically complete below $\Phi = 0.45$ and at $\Phi = 1.0$. It is for the shorter elution time that CRF 8 takes on a significantly higher value for $\Phi = 1.0$ than

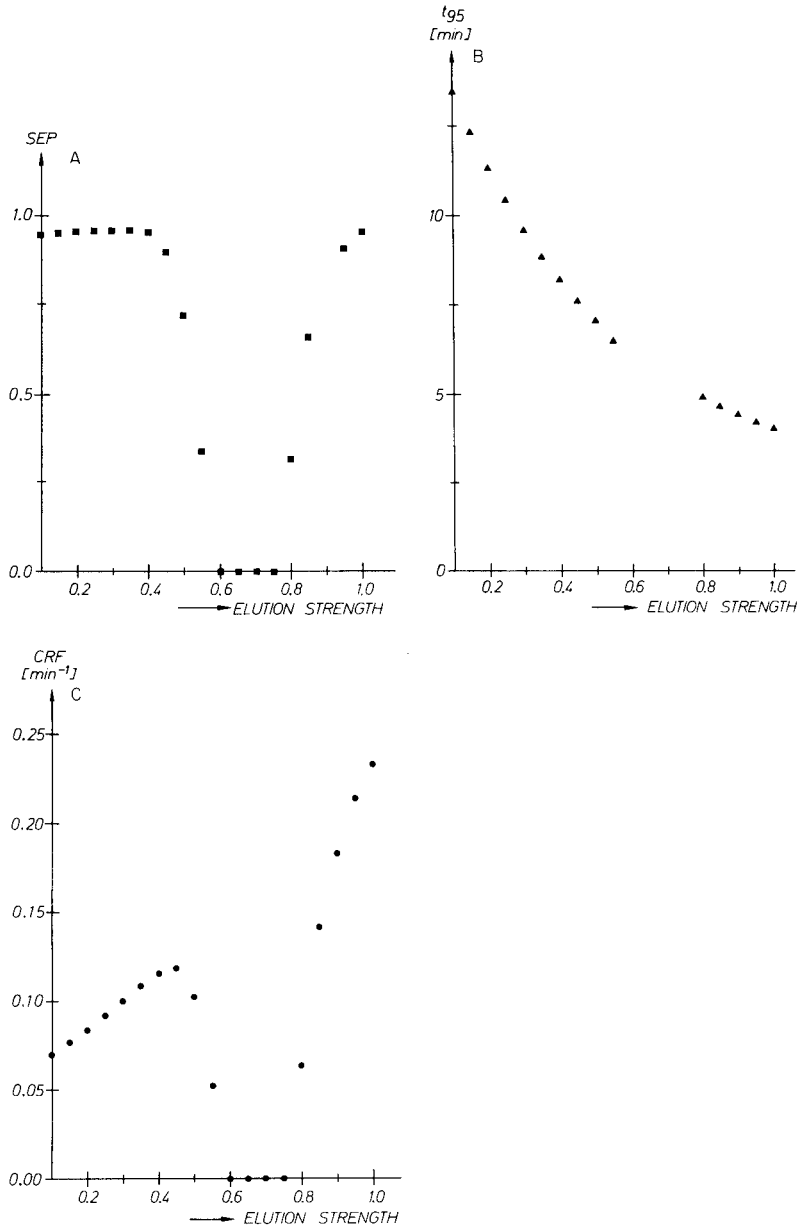


Fig. 3. The variation of SEP (A), separation time (t_{95}) (B) and CRF 8 (C) for the two-component case C (Fig. 1). N = 7000; τ/σ = 0.3.

for 0.45. From the behaviour for $0.40 < \Phi < 0.60$, it is seen that a decrease in separation time can only offset the corresponding decrease in resolution in a very limited way ($\Phi = 0.45$ as opposed to $\Phi = 0.40$ for complete separation). By comparing the SEP value for $\Phi = 0.40$ (0.85) with the relation of Fig. 2, no significant deterioration of accuracy is to be expected. Because of the reduced separation time, CRF 8 takes on an equal value in the region of $\Phi = 0.83$ in spite of the reduced resolution (SEP $\approx$ 0.6); again an inspection of Fig. 2 reveals no severe loss of accuracy. Figure 4 shows the difference in resolution between two sets of experimental conditions giving the same value of CRF 8. In view of experience with the time-normalized CRF 2 [9], this constitutes a significant advantage.

While for equally intense signals the relative error for each of the peaks is similar in size, variations in the relative size lead to grossly disparate relative errors of the two components. For peaks of varying intensity, experimental conditions of moderate to severe overlap (case C; $\Phi = 0.50, 0.52, 0.54$) were simulated; the later eluting component was varied in peak area between 0.1 and 1.0 of the first component. The plots in Fig. 5 show the important results: the highest level of SEP is reached for peaks of equal intensity, this level being dependent on relative retention (α_{ij}). The initial sharp increase of the SEP coincides with the steep improvement of relative accuracy of the smaller peak (solid line). The larger peak shows a strong dependence only under conditions where the smaller peak is not detected (SEP = 0).

In multicomponent situations, this automatically favors the resolution of

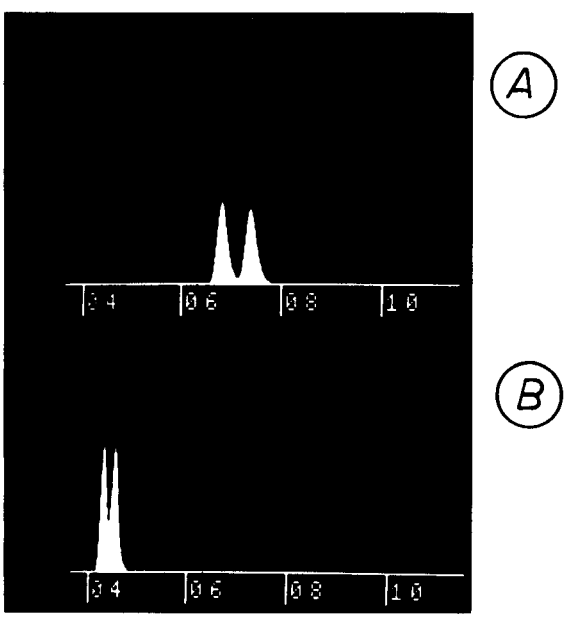


Fig. 4. Comparison of chromatograms giving the same numerical value for CRF 8. (A) $\Phi = 0.45$; (B) $\Phi = 0.83$. $\tau/\sigma = 0.3$, $N = 7000$. Numbers on the abscissa give the time in minutes.

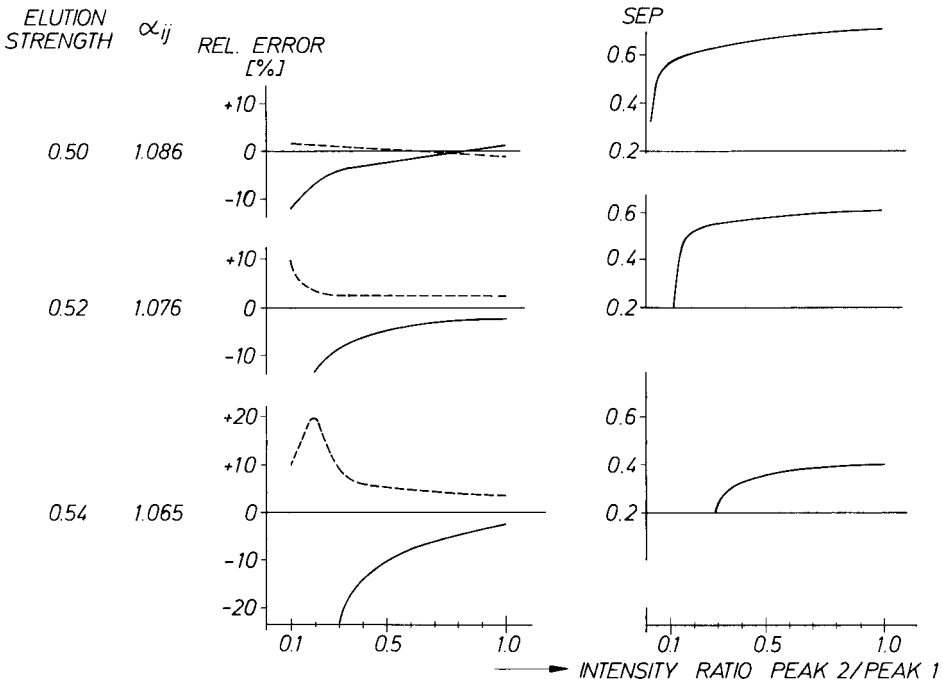


Fig. 5. Influence of relative peak height on the accuracy of the smaller, later eluting component (solid line), of the larger, earlier eluting component (dotted line) and on the numerical value of the peak separation (SEP) for case C (N = 7000, τ/σ = 0.3).

small peaks, a step taken by any experienced chromatographer doing trace analysis in order to keep the relative accuracy for all components as constant as possible. For small peaks eluting after large ones, the bias is always negative, but positive bias is frequently observed for minor components eluting before major components (Fig. 6). In this case, SEP runs through a maximum and then slowly decreases as the relative errors of both the large peak and the small peak increase again. This feature is a direct consequence of peak asymmetry (here, τ/σ = 0.3) because only the larger later peak is affected by the tailing of the first peak. Of the preselected CRFs, only CRF 3 is sensitive to variations in area ratios while CRFs 1 and 6 do not address the issue of detection. These CRFs can be regarded as optimization vehicles for separations only. In principle, CRF 2 is also sensitive to the relative size of the signals, but this CRF has other drawbacks given earlier. Any attempt to incorporate a measure of precision into a CRF has to be restricted to those components of variance that are due to the signal detection process unless repetitive experiments are conducted. Thus, in principle, only the S/N consideration will be amenable to incorporation into any CRF. Again case C was chosen for the demonstration of the effect of noise on CRF 8. From Fig. 7, it can be seen that an increase in the noise level has a particularly severe effect

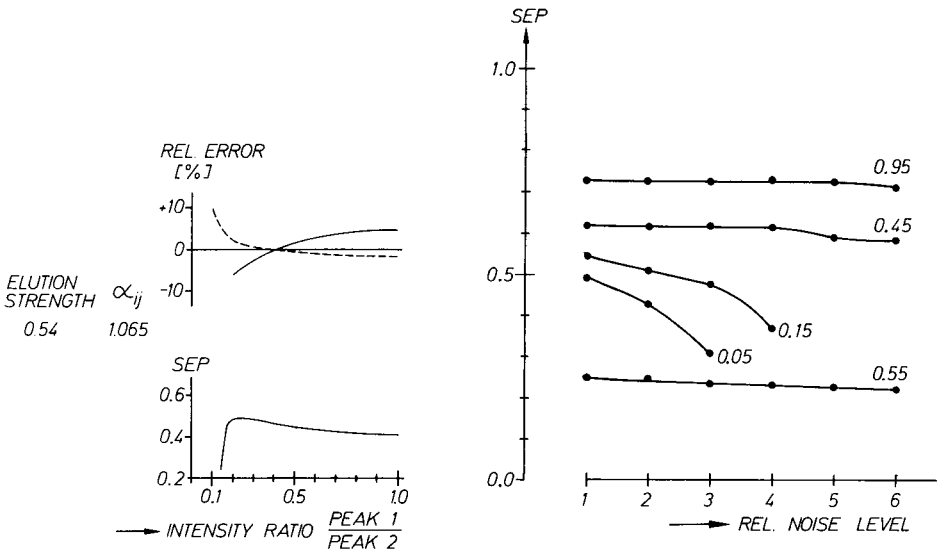


Fig. 6. Influence of relative peak height on: (—) the accuracy of the smaller, earlier eluting component; (---) the accuracy of the larger eluting component; (SEP) the numerical value of the peak separation. All other conditions as in Fig. 5.

Fig. 7. Decrease of SEP with increasing noise levels for various values of elution strength. All conditions as in Fig. 5. Values of elution strength are given on the lines.

on the value of SEP for low S/N ratios such as are found at very long elution times. Table 3 shows the S/N levels at different elution strengths; precision data from 10 repetitive simulations are included.

So far only the simplest separations involving two components have been considered. These results can, however, be generalized to a case involving n components. Such a situation was typified by a case of 10 components with the elution characteristics given in Fig. 8. The experimental variable is again called elution strength, to facilitate the discussion. This does not imply that this is the only experimental variable to which the results pertain, nor are the conclusions restricted to a one-dimensional optimization problem. The case

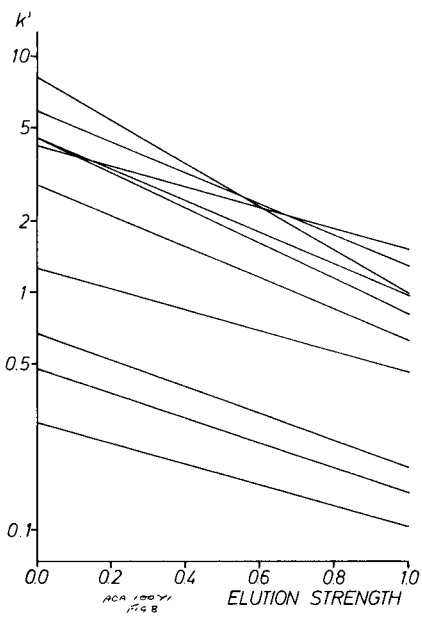
TABLE 3

Relation between S/N at a relative noise level of 1 (see Fig. 7), elution strength and relative standard deviation

S/N	80	97	147	174	265
Φ	0.05	0.15	0.45	0.55	0.95
RSD[a]	1.80	0.86	0.98	0.67	—[b]

[a]From 10 repetitive simulations. [b]Not measured.

was constructed to include several problem areas which forced the CRF to account (a) for resolution and separation time, (b) for late eluting species at elution strength values of approximately 0.15, 0.60 and 1.0, and (c) for fast eluting species, perhaps not so obvious from Fig. 8, at high values of elution strength. The peak separation (SEP) and therefore CRF 3 give a large maximum at $\Phi = 0.45$. (Again, negligible noise was assumed.) Response function 8 (Fig. 9C) hardly affects this optimum, despite the $1/t_{95}$ term. The evaluation of CRF 1 and CRF 6 is given in Fig. 10. While Fig. 10A represents the exact application of CRF 1, the asterisks in parts B and C denote the fact that α and R_s were calculated from the chromatogram by using the mode rather than the second-moment for the estimation of retention times. Comparison of Fig. 10A and B shows no striking difference. Although no indication is given of which pair of solutes is responsible for the minimal values, this pair not only changes as a function of elution strength, but also differs between CRF 1 and CRF 6. The "global" optimum is found at $\Phi = 0.50$ for CRF 6 and at $\Phi = 0.90$ for CRF 1. For closer inspection of the region $0.80 \leqslant \Phi \leqslant 1.0$, the grid was reduced to 0.02: this gave a (local) optimum for CRF 1 at $\Phi = 0.92$ and for CRF 6 at $\Phi = 0.84$. Chromatograms for these Φ values and for the optimum of CRF 8 at $\Phi = 0.45$ are



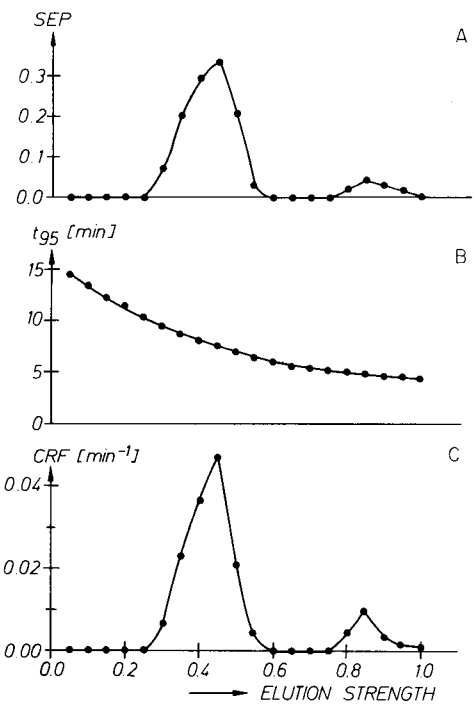


Fig. 8. Capacity factors for the 10 components.

Fig. 9. Values of the peak separation for the case given in Fig. 8: (A) antilog of CRF 3; (B) total elution time; (C) CRF 8. $N = 7000$, $\tau/\sigma = 0.3$.

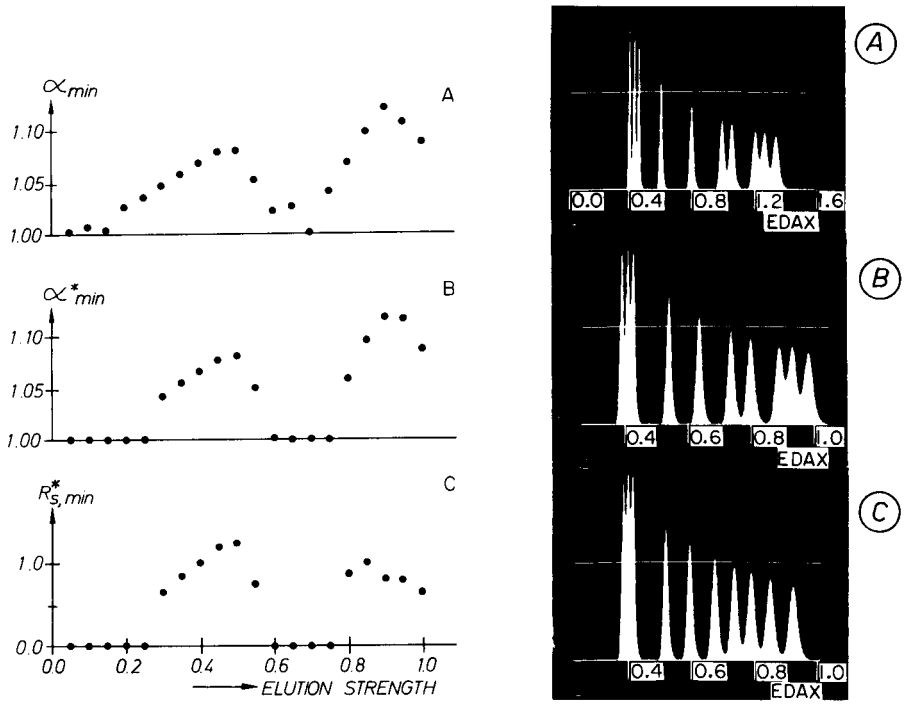


Fig. 10. Values for the case given in Fig. 8: (A) CRF 1; (B) simplified CRF 1; (C) simplified CRF 6. Other conditions are as stated for Fig. 9.

Fig. 11. Chromatograms of a 10-component mixture: (A) $\Phi = 0.45$ (global optimum of CRF 8); (B) $\Phi = 0.84$ (local optimum of CRFs 6 and 8); (C) $\Phi = 0.92$ (global optimum of CRF 1). (The x-axis is labeled in units of minutes.)

given in Fig. 11. It is readily seen that: (a) the incorporation of the time factor $1/t_{95}$ into CRF 8 has no dominant effect as the optimum is found for longer elution times than in CRF 1 and CRF 6 even though no consideration is given to time in these functions; (b) CRF 1 is not suitable at all if quickly-eluting components are present; and (c) CRFs 1 and 6 tend to guide towards optima that are characterized by grossly varying α values (or R_s values, respectively) for different pairs of solutes, while CRF 8 leads to more homogeneous performance, keeping the relative accuracy as uniform as possible.

To elaborate on the last point, the region of the global optimum with $0.30 \leqslant \Phi \leqslant 0.55$ (for CRF 6 and CRF 8) was scanned for uniformity; the results are given in Fig. 12. Both the locally best α_{min} and the globally best $R_{s,min}$ are found for $\Phi = 0.50$, contrasted by the best value of CRF 8 (as well as the best SEP and CRF 3) for $\Phi = 0.45$. As the time scale for different separation procedures varies widely, only SEP can be directly compared to CRFs 1 and 6. Because the actual value of SEP depends on the number of peaks detected, the numbers given on the right-hand side of Fig. 12 give an equivalent to the geometric mean of the separation term: this number is independent of the number of components and can be used to give a basis

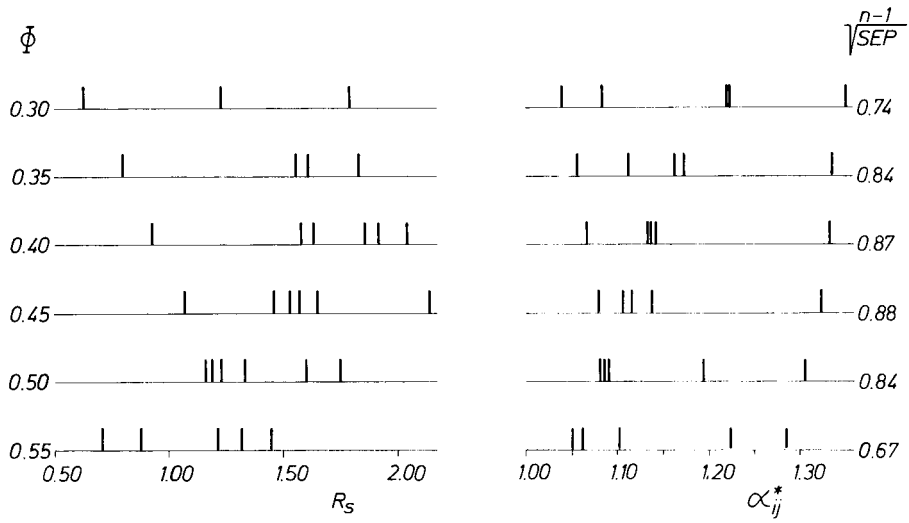


Fig. 12. Plots of the minimal α and R_s values between Φ = 0.30 and Φ = 0.55 along with the mean separation number (see text for discussion).

for comparison to Fig. 2. The more homogeneous distribution of overlap for Φ = 0.45 is easily seen: although the very worst α and R_s values are better for Φ = 0.50, there are three other pairs in terms of R_s and two other pairs in terms of α that are almost as bad. At Φ = 0.45, there is some sacrifice of resolution (and selectivity) for the worst pair, but the separation for all other components is considerably improved. Yet, complete overlap of any one pair cannot be obscured by CRF 8 because of the non-linear dependence of SEP on R_s [13].

In analytical systems with little selectivity, it was recently shown that for linear calibration curves the condition number of the calibration matrix [31, 32] limits the upper bounds of errors caused by the data reduction routine together with errors in calibration and measurements. Certain calibration and integration techniques permit the extension of this concept to chomatographic problems [33]. Table 4 gives the results for selected chromatograms; clearly, there is little difference between the three experimental

TABLE 4

Error amplification factor, cond (K), for three experimental situations of Fig. 8. (See text for discussion)

	Cond (K) for Φ values of		
	0.45	0.84	0.92
First three peaks	1.09	1.20	1.25
All ten peaks	1.06	1.04	1.09

conditions. If only the three components eluting first are considered, degradation of the condition number is noticeable, indicating loss in selectivity. Further work is needed to explore the usefulness of both the linear model for chromatographic calibration and of the condition number as a figure-of-merit in chromatography.

So far, it has been tacitly assumed that most chromatographic systems produce fairly symmetric peaks for which the number of plates, N, is a meaningful measure of column performance. For asymmetric peaks, the value of N obtained depends greatly on the method of measurement [28], a situation that totally invalidates the meaning of any numerical value assigned to R_s. In this situation, an empirical measure of resolution not depending on a particular peak shape and therefore not biased by deviations from this model should be vastly superior to R_s. The Φ region between 0.8 and 1.0 was evaluated for SEP given three different τ/σ values (0.3, 0.5, 1.0): the degradation of the resolution is strongly reflected by the value for SEP affecting CRF 8 (Fig. 13). This is also true for single skewed peaks among a series of fairly symmetric ones, as is frequently observed if special interactions with the stationary phase are operative.

CONCLUSION

Chromatographic response functions having a powerful influence on the success of a formal optimization strategy generally do not reflect important analytical performance characteristics. Many of these functions require good estimates of additional factors (e.g., resolution and separation time) from the operator. This may restrict the successful use of optimization in chromatography to the experienced operator. For those CRFs not requiring outside information, the detection process is frequently neglected: only two of them are sensitive to relative signal height, a quantity that ultimately limits the degree of relative accuracy achievable. Only one (CRF 8) addresses the problem of noise.

Response functions defined by the hardest-to-resolve pair of components only (CRF 1 and 6) tend to do a poor job of balancing one poor separation

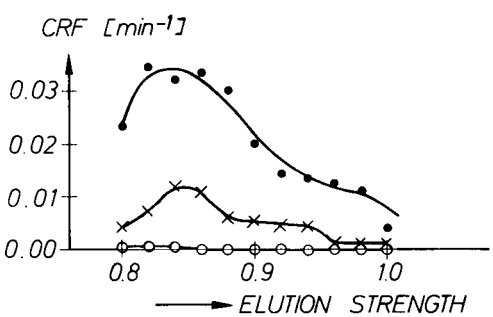


Fig. 13. Sensitivity of CRF 8 to peak asymmetry (see text for discussion). τ/σ values: (•) 0.3; (×) 0.5; (○) 1.0.

against a (possibly) large number of excellent separations. Alternatively, the optimum for these CRFs may coincide with a separation that is not much better than the worst for many other components.

It is concluded from this study that many CRFs do not consider detection but only separation and those generally do not reflect accuracy. In this respect, CRFs derived from Kaiser's peak separation perform best. For the future, however, it is very likely that this criterion will prove too stringent, as multidimensional detectors and improved data-reduction techniques are introduced in chromatography [34].

Computer resources were provided through grant no. 3543 from the Fonds zur Förderung der wissenschaftlichen Forschung, Vienna.

REFERENCES

1 S. N. Deming and S. L. Morgan, Anal. Chem., 46 (1974) 1170.
2 L. R. Parker, Jr., S. L. Morgan and S. N. Deming, Appl. Spectrosc., 29 (1975) 429.
3 J. J. Leary, A. E. Brookes, A. F. Dorrzapf, Jr. and D. W. Golightly, Appl. Spectrosc., 36 (1982) 37.
4 T. J. Sly, D. Betteridge, N. G. Courtney and D. G. Porter, Euroanalysis IV, Helsinki, 1981, Abstract no. 171.
5 W. Wegscheider, B. B. Jablonski and D. E. Leyden, in G. J. McCarthy, C. S. Barret, D. E. Leyden, J. B. Newkirk and C. O. Rund (Eds.), Advances in X-Ray Analysis, Vol. 22, Plenum, New York, 1979, p. 433.
6 L. B. Anderson and R. J. Laub, J. Electroanal. Chem., 122 (1981) 359.
7 R. J. Laub, A. Pelter and J. H. Purnell, Anal. Chem., 51 (1979) 1878.
8 S. L. Morgan and S. N. Deming, J. Chromatogr., 112 (1975) 267.
9 R. Smits, C. Vanroelen and D. L. Massart, Fresenius Z. Anal. Chem., 273 (1975) 1.
10 G. H. Morrison and R. K. Skogerboe, in G. H. Morrison (Ed.), Trace Analysis, Interscience, New York, 1965, p. 2.
11 A. L. Wilson, Talanta, 17 (1970) 21, 31; 20 (1973) 725.
12 R. J. Laub and J. H. Purnell, J. Chromatogr., 112 (1975) 71.
13 M. W. Watson and P. W. Carr, Anal. Chem., 51 (1979) 1835.
14 J. L. Glajch, J. J. Kirkland, K. M. Squire and J. M. Minor, J. Chromatogr., 199 (1980) 57.
15 J. C. Berridge, J. Chromatogr., 244 (1982) 1.
16 W. Wegscheider, E. P. Lankmayr and K. W. Budna, Chromatographia, 15 (1982) 498.
17 S. N. Deming and M. L. H. Turoff, Anal. Chem., 50 (1978) 546.
18 B. Sachok, R. C. Kong and S. N. Deming, J. Chromatogr., 199 (1980) 317.
19 B. Sachok, J. J. Stranahan and S. N. Deming, Anal. Chem., 53 (1981) 70.
20 R. J. Laub, Intern. Lab., May/June 1981, p. 16.
21 R. E. Kaiser, Gas Chromatographie, Geest and Portig, Leipzig, 1960.
22 W. A. Spencer and L. B. Rogers, Anal. Chem., 52 (1980) 950.
23 G. S. Rautela, R. D. Snee and W. K. Miller, Clin. Chem., 25 (1979) 1954.
24 P. E. Antle, Chromatographia, 15 (1982) 277.
25 H. M. Gladney, B. F. Dowden and J. D. Swalen, Anal. Chem., 41 (1969) 883.
26 E. Grushka, Anal. Chem., 44 (1972) 1733.
27 P. T. Kissinger, L. J. Felice, D. J. Miner, C. R. Preddy and R. E. Shoup, in D. M. Hercules, J. M. Hiefje, L. R. Suyden and M. A. Evenson (Eds.), Contemporary Topics in Analytical and Clinical Chemistry, Vol. 2, Plenum, New York, 1978.
28 W. E. Barber and P. W. Carr, Anal. Chem., 53 (1981) 1939.

29 G. E. P. Box, W. G. Hunter and J. S. Hunter, Statistics for Experimenters, Wiley, New York, 1978.
30 T. H. Naylor, J. L. Balintly, D. S. Burdick and K. Chu, Computer Simulation Techniques, Wiley, New York, 1968.
31 C. Jochum, P. Jochum and B. R. Kowalski, Anal. Chem., 53 (1981) 85.
32 A. Björck and G. Dahlquist, Numerische Methoden, Oldenburg, München, 1979.
33 W. Wegscheider and E. P. Lankmayr, unpublished work, Techn. Univ. Graz, 1982.
34 M. A. Sharaf and B. R. Kowalski, Anal. Chem., 54 (1982) 1291.

Analytica Chimica Acta, 150 (1983) 105—114
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

THE USE OF REGRESSION AND STATISTICAL METHODS TO ESTABLISH CALIBRATION GRAPHS IN CHROMATOGRAPHY

DAVID A. KURTZ

Pesticide Research Laboratory, Department of Entomology, The Pennsylvania State University, University Park, PA 16802 (U.S.A.)

(Received 17th September 1982)

SUMMARY

Calibration graphs are normally calculated by regression techniques, and the confidence and prediction intervals indicating the range of amounts are included with a prescribed probability. Logarithmic transformation of data helps to compensate for the change in variance over the regressed line. Examples are given from practical data on the best performance bandwidth for such intervals and the effect that curvature and change of instrumental sensitivity have on the bandwidths. The efficiency of the chromatographic process with different instruments and injection techniques can be compared by observation of the confidence and prediction bandwidths.

Performance characteristics for analytical work have always been in a state of evolution. In most early work, summarized by Wilson [1], emphasis was placed on the physical processes of sample handling, standards, and equipment. In recent times, more emphasis has been placed on how well the measured data represent the real situation. The performance characteristics of analytical methods should include both random and systematic error estimation as well as a mathematical expression of the calibration graph. Further, an estimate of the precision of the analytical procedure for at least one concentration level is needed, as well as within-batch and between-batch precision [1].

For most instrumental procedures, the preparation of a meaningful calibration graph is as essential a part of the whole analysis as the sampling, extraction, or concentration step. The calibration graph is the response of the instrument to a material charge over a given range of charges, and must be drawn to represent the data. Given a series of calibration points, an experienced worker can plot a graph that is highly representative of the data, but this gives little additional information about the quality of the data. When all the calibration points are displayed and plotted along with the estimated line, some indication of the quality of the data is immediately apparent for each measurement run.

The use of a statistically determined calibration graph, a regressed line based on least-squares estimation, improves the quality of the estimation [2]. With the appropriate calculations, the line best representing the pattern

0003-2670/83/$03.00

of the points can be drawn and the process of constructing the regressed line provides some information on the quality of the estimation. The correlation coefficient, r (or its more easily interpreted expression, r^2) is often quoted as the indicator of quality. However, in constructing chromatographic calibration graphs the r^2 value is invariably 98% or better. One should then examine the fit of the data to the proposed regression model. Lack of fit indicates curvature from the straight-line model and interactions from models containing more than one independent variable [4].

The concept of a confidence interval about an analytical calibration line was introduced about 1957 [4]. The confidence region around the regression line is defined as the band within which the true line lies with a given confidence. When the regression line is used to predict the value of unknowns which were not part of the original calibration set, the $100(1-\alpha)\%$ prediction interval is the band within which, at a level of confidence of $(1-\alpha)100\%$, the individual points will be included. Schwartz [5] used the prediction interval for calculating the interval of sample confidence about a segmented portion of a calibration graph, as did Mitchell et al. [6]. The problem of non-uniform variance, inherent in a least-squares estimation of a calibration graph was later discussed by Schwartz [7], Garden et al. [8], and Agterdenbos [9]. In the latter two references, weighting procedures were proposed that involved compensation of the unequal variance by weighting the data by the inverse of the variance at each level. This procedure requires an estimate of the standard deviations at each level and an appropriate weighted least-squares computer program. A simpler approach is to transform the data by taking the logarithm (or other appropriate transformation) of the values before the least-squares estimation is used to calculate the calibration graph. Agterdenbos [9] claimed that a log–log transformation provides a constant variance of y along the regressed line. Although this is not exactly true, it partially solves this problem. This paper provides some real laboratory data to examine the role of log–log regression and the computation of the confidence and prediction intervals in the analytical application of these concepts.

EXPERIMENTAL

Chromatographic data

All data presented here were obtained in daily analytical work. The data were selected to illustrate certain points of the discussion.

The samples were prepared by extraction, clean-up, and concentration, by the usual procedures of trace work [10]. The samples were all subjected to gas–liquid chromatography (g.c.), though liquid chromatographic processes would give the same type of data. The g.c. column was first conditioned with standard and/or sample material to avoid the adsorption problems often found in such trace work. Daily g.c. runs were made on a round-the-clock basis by using a Varian autosampler. The gas chromatographs used

were a Tracor 220 and a Varian 3700 both with linearized electron-capture detection. Recording and peak integration were done by a Varian Vista chromatographic data system.

The regression calibration line and the prediction of the sample concentrations from the line were calculated by MINITAB programming [11]. A typical output is shown in Table 1.

Computation of the confidence interval

Computation of the confidence interval about a calibration graph is fully covered in the literature [2]. The regression equation of y on x is $y = \bar{y} + b(x - \bar{x}) +$ error, where y is the instrumental response, x is the known standard amount producing the response, $\bar{x}$ is the average of the x values, b is the slope of the calibration line, and $\bar{y}$ is the average of the y values. To estimate x for an unknown specimen, $\hat{x}$, the equation to be solved is $\hat{x} = \bar{x} + (y - \bar{y})/b$, where y is the instrumental response for the unknown, and b, $\bar{x}$, and $\bar{y}$ are estimated from the calibration points. To find confidence

TABLE 1

Data in a typical MINITAB printout

THE REGRESSION EQUATION IS
RESPONSE = 1.60 + 1.01 AMOUNT

COLUMN	COEFFICIENT	ST. DEV. OF COEF.	T-RATIO = COEF/S.D.
	1.60288	0.01300	123.26
AMOUNT	1.00700	0.01631	61.75

S = 0.03406

R-SQUARED = 99.8 PERCENT
R-SQUARED = 99.8 PERCENT, ADJUSTED FOR D.F.

ANALYSIS OF VARIANCE

DUE TO	DF	SS	MS = SS/DF
REGRESSION	1	4.4237	4.4237
RESIDUAL	6	0.0070	0.0012
TOTAL	7	4.4307	

ROW	AMOUNT	Y RESPONSE	PRED. Y VALUE	ST. DEV. PRED. Y	RESIDUAL	ST.RES.
1	−1.30	0.3365	0.2928	0.0203	0.0437	1.60
2	−0.60	0.9666	0.9966	0.0130	−0.0300	−0.95
3	−0.00	1.5670	1.6029	0.0130	−0.0359	−1.14
4	0.70	2.3263	2.3067	0.0203	0.0196	0.72
5	−1.30	0.2989	0.2928	0.0203	0.0061	0.22
6	−0.60	0.9854	0.9966	0.0130	−0.0112	−0.36
7	−0.00	1.5729	1.6029	0.0130	−0.0300	−0.95
8	0.70	2.3444	2.3067	0.0203	0.0376	1.38

DURBIN-WATSON STATISTIC = 2.00

LACK OF FIT TEST
POSSIBLE CURVATURE IN VARIABLE RESPONSE (P = 0.047)
POSSIBLE LACK OF FIT AT OUTER X-VALUES (P = 0.017)
OVERALL LACK OF FIT TEST IS SIGNIFICANT AT P = 0.017

intervals for $\hat{x}$, i.e., the prediction interval, one starts from the prediction limits of y, given x:

$$y = \bar{y} + bu \pm ts(y \cdot x)[1 + 1/n + \hat{u}^2/\Sigma(x - \bar{x})^2]^{1/2} \quad (1)$$

where u is $(x - \bar{x})$, t is the Student's t value at a given probability level (usually 5%) for $(n - 2)df$, and n is the number of calibration points. When Eqn. (1) is solved as a quadratic equation in x for a given y, the roots, which can be expressed in the following form, provide an approximate prediction interval for the true x:

$$x = [\hat{x} \pm [ts(y \cdot x)/b] \{[(n + 1)/n](1 - c^2) + \hat{u}^2/\Sigma(x_i - \bar{x})^2\}^{1/2}]/(1 - c^2) \quad (2)$$

where $c^2 = t^2 s(b)^2/b^2 = (1/\Sigma(x_i - \bar{x})^2)(ts(y \cdot x)/b)^2$.
The confidence interval of the known standards without the prediction of new values of x reduces Eqn. (2) to

$$x = \{x \pm [ts(y \cdot x)/b][(1/n)(1 - c^2) + \hat{u}^2/\Sigma(x_i - \bar{x})^2]^{1/2}\}/(1 - c^2) \quad (3)$$

In most cases of calibration work and especially in computing the regressions from the logarithms of the values, the denominator of both Eqns. (2) and (3) is close to unity and can be omitted from the calculations. The log—log model equation and the log—log prediction interval are the same as the conventional equations given above for y and $\hat{x}$ except that the values for x and y are transformed to the logarithms before the other calculations are done.

A suitable indicator for the spread of the interval is the bandwidth, defined by Mitchell et al. [6] as follows:

Bandwidth, % = [(upper limit — lower limit)100] / 2 (predicted amount)

DISCUSSION

Theoretical aspects in the use of regression and confidence limits in the preparation of calibration graphs have been worked out by numerous authors [5, 6, 9]. Various models, weighting systems, and segmented calibration graphs have been proposed but these techniques are seldom used in research and industry. Various chromatographic procedures are capable of providing responses that are linear over several decades of concentration, and this can lead to problems in handling the entire range of numbers in calculating the least-squares regression. The large numbers at the high end of the line dominate the best fit, producing imprecision at low concentrations. As trace concentrations are often of most interest, the calculation proves to be unsuitable. The use of weighted least squares has been proposed by Garden et al. [8] and Agterdenbos [9] to resolve the problem caused by the non-constant variance along the regression line.

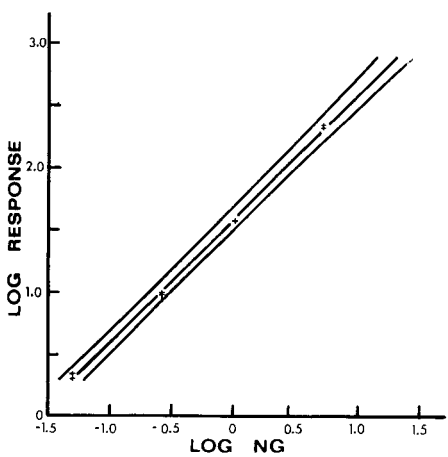
Another solution is to transform the data logarithmically. Useful data are obtained, and the confidence limits, when calculated, are quite different from non-transformed data. Table 2 shows the differences in the calculated predicted interval bandwidth between non-transformed and transformed data.

TABLE 2

Effect on the prediction interval of logarithmic transformation of data showing linearity (n = 6, lack of fit $p > 0.1$; fenvalerate determination by e.c./g.c.)

Level (ng)	Prediction interval bandwidth (%)	
	Non-transformed	Transformed
0.05	40	14.0
0.23 (mean)	—	11.0
0.43 (mean)	4.5	—
1.0	2.2	13.5

Figure 1 shows a plot of data calculated by first taking the logs of both the amount on the column and the response, calculating the prediction interval, and plotting the log—log data on rectangular coordinates. Response data were obtained at four levels in duplicate, at the start and end of the chromatographic run. The same data were subjected to calculation without the log--log transformation and are plotted in Fig. 2 on rectangular coordinates. Little differences are seen. However, if the transformed data are plotted on the same scale as the non-transformed data (Fig. 3), the differences are obvious. Non-transformed data will be effective in narrow-range calibration graphs and will be most effective at the higher levels. Transformed data will give even bandwidth over the entire range and will be effective over quite a broad range. If the data form a curved plot, the differences in the two methods are even more pronounced (Table 3). The bandwidth can be reduced as shown in Table 3 by removing the data points at the high end of the curve (5 ng) where curvature was noticed. This gave



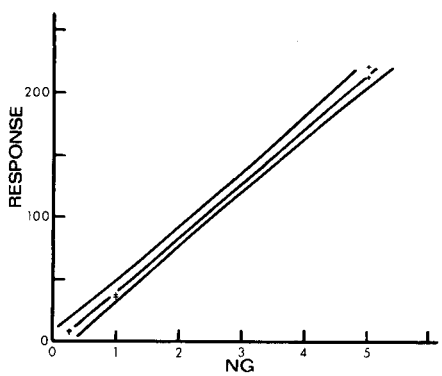


Fig. 1. Plot of regression line and 95% predicted interval from log—log transformed data.

Fig. 2. Plot of regression line and 95% predicted interval from the non-transformed data used for Fig. 1.

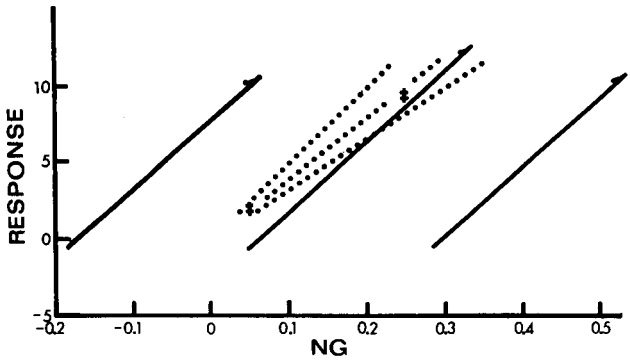


Fig. 3. Plot of regression line and 95% predicted interval from both log—log transformed (····) and non-transformed) (—) data at lowest portion of data.

TABLE 3

Effect on the prediction interval of logarithmic transformation of data showing lack of fit (n = 8, lack of fit p = 0.02; fenvalerate determination by e.c./g.c.)

Level (ng)	Prediction interval bandwidth (%)	
	Non-transformed	Transformed
0.05	450	22.0
0.50 (mean)		20.0
1.63 (mean)	14	
5.0	5.1	22.0

the data for Table 2. As expected, the regression performance is greatly improved.

Setting up a routine quantitative chromatographic system requires the establishment and regular checking of the calibration graphs for the compound(s) of interest. The example used here involves the determination of the pesticide fenvalerate in chicken eggs and tissues. In this fenvalerate project, an optimal calibration graph was eventually found. This graph was linear for 0.05—20.0 ng of fenvalerate applied to the column, and was repeatable between the start and end of daily runs. Information about a typical best-performance run is given in Table 4. The mean bandwidth was found to be about 5% for the confidence interval and 15% for the prediction interval. These bandwidth figures depend strongly on the standard deviation of the response about the regression line, and to a lesser extent on the t-value or the number of data points. For comparison, non-transformed data were also used to calculate the confidence and predicted intervals and their bandwidths; the results are again presented in Table 4. The effect of non-constant variance along the regression line is obvious.

TABLE 4

Comparison of confidence and prediction intervals with and without logarithmic transformation in an optimal situation with linearized e.c./g.c. and autosampling (n = 10, 95% confidence level; range of area response was 1.0—880 area units)

Level (ng)	Confidence interval		Prediction interval	
	Range (ng)	Bandwidth (%)[b]	Range (ng)	Bandwidth (%)[b]
With transformation[c]				
0.05	0.046—0.054	8.4	0.042—0.060	17.3
1.05[a]	1.00—1.10	4.8	0.89—1.23	15.8
20.0	18.4—21.7	8.1	16.8—23.0	17.2
Without transformation[d]				
0.05	0.14—0.24	380	—0.48—0.58	1060
1.05	0.87—1.22	17	0.52—1.6	50
5.3[a]	5.1—5.4	3	4.7—5.8	10
20.0	19.7—20.3	1.7	19.4—20.6	3

[a]Mean level applied to column. [b]Calculated from Eqn. (3). [c]No lack of fit at p = 0.1. [d]Lack of fit at p = 0.01.

Different instruments and ways of using them can be compared by observing the bandwidth. For this example, chlorothalonil was determined by e.c./g.c. In this case, solutions were injected manually and the electrometer used to amplify the signal produced a d.c. or non-linearized signal. The linear region of this system covers only one order of concentration. The data given in Table 5 show that this system was less efficient than the linearized signal/autosampler system discussed above.

In routine chromatography, calibration data sets are usually obtained at the start and end of the day's run as well as at intermediate points. The sensitivity of the chromatograph during the run can be checked by using an F-test to compare the first and last or other standard sets. This test checks the hypothesis that the additional contributions of the second set of data to the slope and intercept are significant. Such F-tests are very helpful in deciding the value of a particular chromatographic run.

TABLE 5

Minimum prediction interval bandwidth for manual injection and a non-linearized signal in the determination of chlorthalonil

Run	No. of data points	Lack of fit probability	Prediction interval bandwidth (%)
1	7	p = 0.05	43
2	6	p > 0.1	67

Such changes in sensitivity can be detected by observing the width of the confidence band. Some data were taken from the fenvalerate work and modified slightly so as to produce different F-test statistics with essentially the same group of data points. Table 6 shows the results of this test. The fenvalerate data were logarithmically transformed as usual. Three sets of F-test statistics were determined from these three groups of data, one less than, one about equal to, and one greater than the critical value of the F-distribution for the required degrees of freedom at $\alpha = 0.05$. The table shows the bandwidth percentage at three levels of response. The calculations show that the width is essentially the same if the F-test indicates no difference in the data. When the F-test showed a difference, the confidence band was wider than expected, in this case 50% wider for moderately different calibration points.

The linearity of the calibration graph also affects the bandwidth of the confidence and predicted bands, curves giving greater widths than straight lines. A set of e.c./g.c. data (Table 7) was found to have a lack-of-fit probability of $p = 0.02$; for the four concentration levels (duplicate injections) the highest level seemed to deviate from linearity. As the sample data to be evaluated were at the lowest levels, the highest levels were removed from the data set and the calibration graph was recalculated as recommended by Mitchell et al. [6]; this operation will obviously lead to improved bandwidth.

A comparison can be made of the bandwidths for any operating system. Another example is the determination of captan by using flame photometric detection of sulfur. An example is presented in Table 8. For the two data

TABLE 6

Prediction interval width at various analyte response levels and at various F-test conditions for logarithmic transformations
(All curves showed no lack of fit at $p = 0.1$ and the critical value of the F-statistic for 2,6 df, $\alpha = 0.05$ is 5.14)

F-statistic	Level (ng)	Prediction interval bandwidth (%)
F-statistic less than percentage points of F-*distribution*		
0.54	0.05	35
	1.05	32
	20.0	35
F-statistic equal to percentage points of F-*distribution*		
6.2	0.05	31
	1.05	28
	20.0	30
F-statistic greater than percentage points of F-*distribution*		
14.0	0.05	49
	1.05	44
	20.0	48

TABLE 7

Prediction interval at minimum, mean, and maximum detection levels for two lack-of-fit probabilities

Lack-of-fit probability	Level (ng)	Range (ng)	Prediction bandwidth (%)
$p = 0.02$ ($N = 8$)	0.05	0.040—0.062	22
	0.50	0.41—0.61	20
	5.0	4.02—6.25	22
$p > 0.1$ ($N = 6$)	0.05	0.044—0.057	13
	0.23	0.21—0.26	11
	1.0	0.87—1.14	13.5

TABLE 8

Prediction interval bandwidth for flame photometric detection in the determination of captan

Level (ng)	Prediction interval bandwidth (%)	
	First data set[a]	Second data set[a]
13.5	15.8	7.4
49.4 (mean)	15.0	7.0
288	16.5	7.6

[a] $n = 14$.

sets given, each with 14 data points, the prediction intervals were 15 and 7%, respectively. As the amounts of captan on the column ranged from 13.5 to 288 ng, the bandwidth information for this interval indicates a close similarity to that obtained for fenvalerate by electron capture detection.

In summary, in looking at performance characteristics of calibration runs involving a wide range of standards (several orders of magnitude), the important parameters are the range and number of the standards, and the standard deviation around the regression line of y given x, the standard deviation being the most important (assuming more than 3 samples). The calculation that puts all these parameters together is the determination of the confidence interval and the prediction interval.

The author acknowledges the helpful comments of Dr. Thomas E. Kurtz, Department of Mathematics, Dartmouth College, and Dr. James L. Rosenberger, Department of Statistics, The Pennsylvania State University. This article is paper 6592 from the Pennsylvania Agricultural Experiment Station.

REFERENCES

1 A. L. Wilson, Talanta, 17 (1970) 21, 31; 20 (1973) 725; 21 (1974) 1109.
2 See, e.g., G. W. Snedecor and W. G. Cochran, Statistical Methods, 7th edn., Iowa State University Press, Ames, IA, 1980, p. 169.
3 J. Neter and W. Wasserman, Applied Linear Statistical Models, Richard D. Irwin, Inc., Homewood, IL, 1974, pp. 117—118.
4 J. Mandel and F. J. Linnig, Anal. Chem., 29 (1957) 743.
5 L. M. Schwartz, Anal. Chem., 49 (1977) 2062.
6 D. G. Mitchell, W. N. Mills and J. S. Garden, Anal. Chem., 49 (1977) 1655.
7 L. M. Schwartz, Anal. Chem., 51 (1979) 723.
8 J. S. Garden, D. G. Mitchell and W. N. Mills, Anal. Chem., 52 (1980) 2310.
9 J. Agterdenbos, Anal. Chim. Acta, 108 (1979) 315.
10 D. A. Kurtz and K. C. Kim, Pestic. Monit. J., 10 (1976) 79.
11 T. A. Ryan, Jr., MINITAB Reference Manual, Release 81.1, The Pennsylvania State University, University Park, PA 16801.

Analytica Chimica Acta, 150 (1983) 115—127
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

THE APPLICATION OF COMPUTERS IN CHEMOMETRICS AND ANALYTICAL CHEMISTRY

R. M. BELCHAMBER, D. BETTERIDGE*[a], Y. T. CHOW, T. J. SLY and A. P. WADE
Chemistry Department, University College Swansea, Swansea SA2 8PP (Gt. Britain)
(Received 25th November 1982)

SUMMARY

A selective survey of applications of computers and chemometrics in analytical chemistry is given. These range from the relatively simple control and computational functions of a microprocessor dedicated to an automatic titrator to a technique employing complex data processing for image processing, computerised tomography. Intermediate in terms of complexity are automated self-optimising flow-injection systems and the extraction of analytical information from acoustic emissions. The relative costs of options for computerising a process are considered and the important role of the instrument companies is emphasised.

The proceedings of a 1971 meeting on The Applications of Computer Techniques in Chemical Research [1] make fascinating reading today. For example, J. M. Skinner [2] wrote: "the current prediction is, that by the end of the seventies, a chemical laboratory without a computer will be hopelessly lost", and the apparatus which was described for data handling of gas chromatography, in one of the best equipped laboratories in the world, involved outputting data onto magnetic tape, then onto punched paper tape for submission to the mainframe computer, which had a 24-h turn-round [3]. The developments in solid-state electronics have been so revolutionary that laboratories have been radically transformed during the last decade and the process is continuing. Skinner's prediction has been fulfilled to an extent which was barely envisaged in 1971.

The parallel conceptual changes required of the chemist are taking place more slowly. The problems tackled and chemical opportunities opened up by computers which were discussed at that 1971 meeting are still very live issues. It is a cardinal mistake to assume that computing in chemistry really began with the spread of microprocessors and microcomputers in the 1970s. The pioneering work was done much earlier and still may be read with profit, e.g., in the series edited by Mattson, Marks and MacDonald [4, 5]. The availability of massive cheap computing capability in the laboratory, has

---

[a]Present address: B.P. Research Centre, Chertsey Rd., Sunbury-on-Thames, Middlesex TW16 7LM, Great Britain.

0003-2670/83/$03.00

had and is having a major impact on analytical chemistry and instrumentation. Some of this [6] and chemometrics [7] have been reviewed comprehensively recently. The object of this paper is to discuss and illustrate ways in which computers may be used to advantage in the analytical laboratory.

Anyone who has lectured on the state of the art in microcomputers over the last few years must have felt uneasy, knowing that the lecture was going out of date as it was being delivered. One used to feel reasonably confident with the statement that the practical limit in microcomputer development was a commercial one: it was necessary to make a lot of chips, to make any design worthwhile. However, developments in electronics have taken place which make it profitable to produce a custom-built computer on one or two boards. A lot of the most useful, economic and easiest analytical applications arise from the dedication of an ultramicro computer to a specific task. The temptation to use the full resources of a modern "micro" computer and to over-elaborate a solution to a problem should be avoided, but it is, of course, necessary to use the right size computer for the task in hand.

It is fun playing computers, but of limited value to the laboratory unless you are prepared to do the documentation. The instrument manufacturers are constrained to sell a reproducible working product, and have met the challenge of microprocessors very successfully.

A computer is not only for data processing, but exists for control and self-validation of analytical instrumentation. The computing can be easy, but the chemistry is hard. Computer specialists have done a lot of the hard work on computers, and their results are available for exploitation. How to use the power of the computer to do better chemistry is the problem for chemists. In the past, analysts have constrained the chemistry to give a single-point measurement, e.g., a colour change at the equivalence point in a titration, or an absorbance measurement at a single wavelength. It is now possible to extract much more information from an analysis, but it requires a radical rethink of the method.

Methods in which the assistance of the computer is beneficial will be considered before those which would never have been attempted without the availability of powerful computers.

SMALL COMPUTERS FOR CONTROL AND DATA PROCESSING

Titrators

Work with microprocessors began in this laboratory with a team consisting of an academic and industrial chemist, an electrical engineer and a post-graduate student. The interdisciplinary aspect proved most beneficial to the development of the project and this aspect has been strongly emphasised in other accounts of the development of computer-controlled instrumentation [8, 9]. The first project was the design and construction of an automatic titrator [10]. The importance of the technique and its repetitive nature warranted automation, and the level of control and data processing required was

within the range of the rudimentary microprocessor then available (Intel 8008). Similar thoughts had obviously occurred to others, for Christiansen et al. [11] published their work at the same time and since then there has been a steady stream of papers on the same theme [6]. Three approaches have emerged, and they are representative of many other computerised systems.

The first approach is to take a standard working system and add on a computer, which can take over some of the control functions from microswitches and mechanical devices, then arrange for the data to be stored and displayed more conveniently than is possible with a pen recorder, and finally do some calculations. This, in essence, was the approach of Christiansen et al. [11]. It has the advantage of speed of development, and ease of use, for the product is recognisable as an up-dated and improved version of a familiar apparatus. The disadvantages of the approach are that it does little to advance conceptual development and is not likely to lead to major reduction in cost. It is a reliable but conservative approach.

A second approach [6, 10] is to dedicate the computer and titrator to a specific problem, e.g., the sequential titration of adipic and boric acids for the control of a plant process for the manufacture of nylon. Because the task was narrowly defined, potentially useful combinations of chemical, computational and electrical engineering were explored fairly quickly. Thus it was found that by using a differential method of calculating the equivalence point, it was possible to dispense with an accurate and carefully calibrated pH meter and to add the titrant at a fixed rate. By adding a fixed volume of titrant, known to be sufficient for the titrations, it was sufficient to scan the differential data for the two highest peaks and to interpolate accurately to obtain the equivalence points. These factors combined to give a rapid development time ($8\frac{1}{2}$ student months from start to tested industrial laboratory prototype), and economy in the cost of the instrumentation and development of computational programs. The material cost of the development was covered in the first year by the saving in mannitol which had been necessary to titrate the boric acid by the manual procedure. These economies in time and cost were highlighted when the titrator was converted to a general-purpose apparatus, for then the peak detection routine became more complex, and control features and validation checks were added. The time taken to make these changes was not much different from the initial development. By contrast, the time taken to replace the 8008 microprocessor with a PET microcomputer, and to use a more sophisticated program in Basic was 2 days' work. The disadvantage of the dedicated approach is that the product may be revolutionary in concept but time-consuming to convert to a general form, and so more difficult to market.

The third approach, that taken by Wu and Malmstadt [12], is to devise a general-purpose apparatus around a microcomputer. They designed and constructed a titrator, fully computer-controlled, which can be used for potentiometric, coulometric and photometric titrations, with a large number of computational options as well. Each form of titration requires

different chemistry, different engineering and different forms of calculation, and difficulties increase as the number of variables increases. Nevertheless, an elegant general-purpose solution to a problem has a distinct appeal to the intellect and gives a feeling of reassurance to the customer. Such a system is now marketed by Mettler. The final cost of the computer components is almost negligible but the development costs are great.

The general conclusions to be drawn from the application of microcomputers to automatic titrators are that there are distinct advantages in developing a dedicated system in the laboratory, but that general-purpose apparatus is best purchased. The development of an automated instrument requires good specification of the chemical, computational and engineering requirements and limitations. The successful instrument will be one in which the chemistry is rethought so as to take full advantage of the computational and engineering possibilities.

Flow-injection analysis and computer-aided optimisation

All of the above points need to be taken into account in the computerisation of flow-injection analysis (f.i.a.). This is an attractive and apparently simple method of continuous flow analysis in which the sample is injected into a flowing stream of reagent and the reaction product is detected downstream [13].

In fact, the system is rather complicated especially if its analytical potential is to be exploited to the full. First, the physical dispersion of the sample is critical. The sample must be allowed to react with the reagent in the carrier stream without the reaction product becoming too dilute for detection. Alternatively, for some applications in which the flow-injection system is used as a sample inlet for an instrument such as a pH meter or an atomic absorption spectrometer, the sample should scarcely be dispersed at all. Secondly, the chemistry takes place in what is essentially a mobile three-dimensional reactor in which the concentration of reactants is continually changing. The effects of chemical kinetics as well as physical dispersion must be taken into account, and if one is interested in exploiting the chemical information across the sample—carrier interface, thermodynamic equilibrium data may be required. Thirdly, there are instumental and computational factors. There is a need for close control of flow rate, sample injection, temperature, etc., and the speed of analysis is such that a data sampling rate of 10—100 Hz is necessary.

For the computer enthusiast, the paramount problems are how to exploit the chemical information to the full and how to control the experimental parameters to give optimum performance. The trivial solution is the add-on computer. Peaks in f.i.a. are similar to those in chromatography and the same peak detection algorithms are applicable. The data sampling rate is such that the interface must either be a microprocessor or microcomputer programmed in machine language to do the data collection and processing, or a microprocessor programmed to collect data and output it onto a micro-

computer for processing. A PET microcomputer programmed in Basic is not quite fast enough for all f.i.a. applications. Nevertheless, the add-on approach is easy to implement and gives reasonable results, but more careful integration of the computer with the flow-injection system leads to better and possibly more novel results. During the course of a lengthy study, it was found that by controlling the temperature to ± 0.02°C, measuring the flow rate, and employing a superior peak smoothing and detection algorithm, the precision of the method is improved almost by an order of magnitude [14]. The better precision permits a better evaluation of the factors which contribute to the imprecision of the method and, most importantly, opens up the possibility of making significant use of the data obtainable from the pH and concentration gradients across the interface between sample and carrier. These benefits accrue from using a computer in a control mode, where it can carry out boring, routine check measurements more frequently and more accurately than a human.

There is to date no completely satisfactory theoretical treatment of f.i.a., although several attempts have been made. Most [13, 15, 16] employ a model of physical dispersion which is based on the original work of Taylor [17] and/or standard chemical engineering reactor theory [18, 19]. They are useful, but fail to take into account chemical interactions, kinetic considerations [20] and all the physical factors [21].

Accordingly, empirical procedures have been used to select optimum conditions; with such a simple experimental rig, it is comparatively easy to alter variables. However, it is preferable to use a more rigorous optimisation procedure, and a version of the modified simplex method was therefore applied to f.i.a. A bonus of this approach is that because of its flexibility, f.i.a. provides a good experimental test of the simplex method.

The computational method employed here is an extension of the modified simplex method of Nelder and Mead [22, 23]. The changes made to the algorithm incorporate some suggestions of Routh et al. [24] and Ryan et al. [25]. Compared to the Nelder and Mead version, there are: (a) weighting of the centroid towards the better (n-1) points; (b) a Lagrange interpolative fit whenever a maximum is known to lie within the current simplex (subject to certain conditions); (c) a routine aid in the detection and avoidance of the problems caused by two or more maxima being present; (d) aids to help prevent loss of dimensionality; (e) prevention of simple oscillation setting in; and (f) improved treatment of boundary conditions. The details of the changes are relatively straightforward. That they were considered worth trying is a reflection on the simplex method, which is robust but not perfect.

The program was first implemented on a PET microcomputer and operated with manual input of data. This stage was achieved rapidly and the initial results were most satisfactory. Optimum or near optimum was reached with 4 or 5 variable flow-injection systems within ten experiments, although a further 10—20 were carried out in the process of fine-tuning which characterises the modified simplex method. The variables were pH, reagent concen-

tration, flow rate, tube length and sample size. A comparison with conventional univariate optimisation shows a considerable saving in experimental time (Fig. 1). The procedure is quick, for it takes little time to type in 4 or 5 data points and to get a new set of experimental parameters printed out. Further it is widely applicable. The procedure has been applied to the optimisation of conditions for the measurement of the catalytic wave from the polarographic reduction of uranium as a function of the concentration of hydrochloric and nitric acids (Fig. 2) [26].

The power of the modified simplex method is impressive but there are niggling doubts. First, useful chemical information is normally obtained from the results achieved by the more lengthy conventional investigations; such information is not duplicated in simplex optimisation, but the multivariate nature of the simplex results provides a viewpoint not easily obtained from conventional procedures. The problem of validation of the simplex modification, i.e., proof that the point reached is the true optimum, is more difficult. One can use mathematical functions, but they are not easy to visualise. Here, a map is preferred; this gives a convenient quantified representation of a three-dimensional surface in which it is easy to visualise peaks, ridges and valleys and to start the simplex going from different points. As a procedure for checking the best size of simplex, the probability of reaching the true optimum, and the speed of convergence, it is admirable. It enabled improvements to be made and errors to be detected. However, when the flow-injection system was investigated with five variables, the program which functioned so well for two variables was less good. It appears there may be

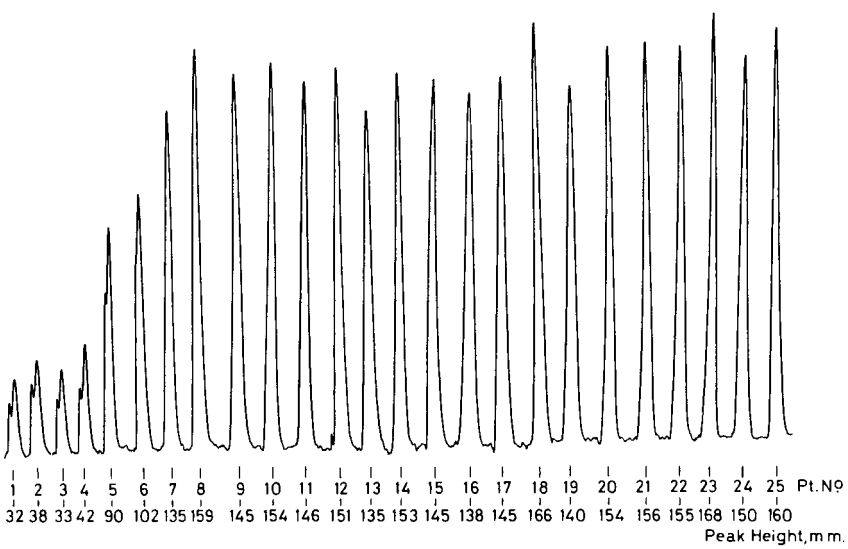


Fig. 1. Modified simplex optimisation of a flow-injection system for the determination of isoprenaline. The four variables are pH, reagent, hexacyanoferrate(III) concentration, tube length and flow rate. A sequence of only 25 experiments was needed.

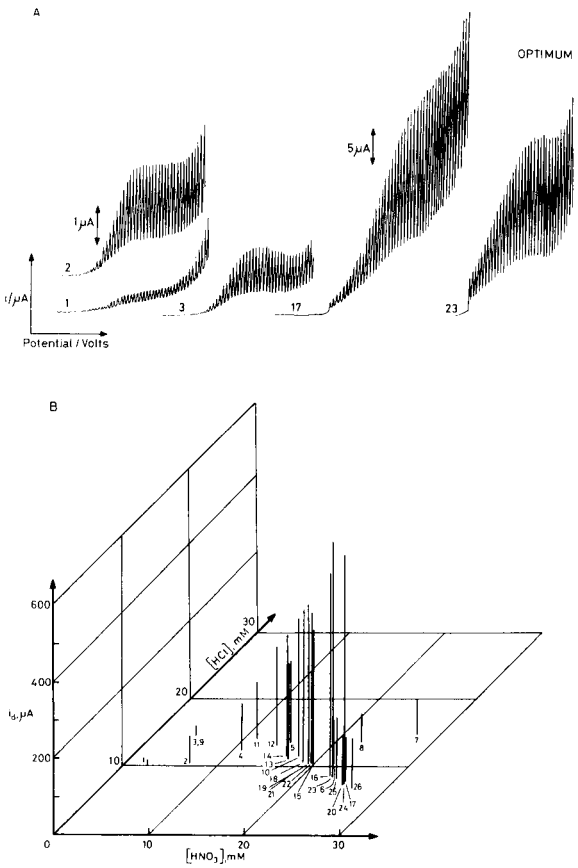


Fig. 2. Optimisation of catalytic wave from polarographic reduction of uranium(VI). A, Experimental curves, numbered by experiment, for optimisation of peak height and separation showing the merging of the catalytic wave with the hydrogen wave (17); the response function which carries penalty for bad curve shape is $i_d \cos \theta$, where θ is the angle made by the top of the wave with horizontal. B, The response as a function of concentrations of nitric and hydrochloric acids.

an optimum number of variables for a given form of the simplex program. One of the advantages of f.i.a. is that it is comparatively straightforward to operate similar systems with 3—11 variables. Work is in progress to establish by experiment the validity of the various simplex algorithms.

Another possibility is to conduct the optimisation automatically under the control of the computer. The feasibility of the approach has been demonstrated by Stieg and Nieman [27] who optimised a chemiluminescent reaction. The development of a computerised self-optimising flow-injection system [28] took about two student years. The difficulties lay in the design of a truly automated flow-injection system, the principal physical variables

of which could be altered at will by the computer. It is often awkward to automate steps which manually present no difficulty. For example, in f.i.a. the length of polypropylene tubing can be adjusted rapidly with a pair of scissors, but such a procedure is not amenable to automation. Preference was therefore given to a multiple array of detectors, which consist basically of a light-emitting diode and phototransistor pair mounted externally and transversely to the carrier stream. These are cheap and easy to fabricate so that an array spaced at suitable intervals is not difficult to arrange [28]. Then the length between injection and detection can be varied by selecting the appropriate detector. Mixing and dilution of reagents to appropriate concentrations is achieved by individually-controlled peristaltic pumps with stepper motors. There is, of course, the problem of air bubbles being mistaken for major peaks and noise being mistaken for minor peaks. In an automated regime, such glitches are potentially serious because they influence the path of the simplex. These problems are soluble, and the apparatus functions to give an automated 4-variable optimisation of a flow-injection system. The initial results are impressive. The total time taken for optimisation is shorter than in the manual system, because the rate of data transfer is faster, and the resetting of the experimental parameters is achieved much more rapidly. In addition, the operator time is negligible.

In the short term, there is no way the total development cost can be justified, except as an educational exercise. However, the system is transferable to other flow-injection or continuous flow systems. Already, such routines are being incorporated in commercial liquid chromatographs and they should prove of great benefit to the analyst.

Further developments that can be foreseen in the f.i.a./chemometrics area include the on-line pattern recognition of peak shapes in order to obtain information of sufficient quality from multicomponent mixtures to serve as a basis for production control, and the simulation of the system in order to obtain relationships which will permit better prediction of behaviour and selection of operating conditions. Thus f.i.a. is seen to be a versatile method which both benefits from computerisation and chemometrics and in turn can provide a good test bed for various chemometric procedures.

LARGER COMPUTERS FOR DATA PROCESSING

Data processing by computer is a vast subject area. Just two illustrative examples will be considered here. The first shows how the application of chemometrics and signal evaluation can help to reduce a complex problem to manageable size. The second involves a commercial development of an infrared spectrometer in which advanced data processing is made conveniently available to the analyst.

Acoustic emissions from stressed polymers and chemical reactions

Materials under stress often emit sound and frequently this is used as a criterion of mechanical failure, e.g., the cracking of timber. More recently,

the method has become adopted as a standard test for the monitoring of stress in materials, especially metals [29, 30]. The advances in sonar technology which have made available sensitive piezoelectric transducers are responsible for the development of acoustic emission tests. However, at present, these are fairly rudimentary, and the origins of the emissions are not well understood.

The problem of analysis is easily stated. Emissions of relatively short duration (10—100 μs) are given out for several minutes. Signals from stressed solids [31] and chemical reactions are similar. Thus there is a vast amount of data from each experiment. The commonly employed test method is to take the rate of emission of noise as a criterion of incipient mechanical failure.

The objectives of this study were to establish if a good correlation could be made between acoustic emission data and mechanical performance for polymers and composites and to try to obtain further insight into the origins of acoustic emissions. Given the complexity of the problem, a good testable model is extremely difficult to devise. Accordingly, it was necessary to interpret the data in such a way that useful working relationships could be established and that the theoretical problems might be reduced to more manageable proportions.

As a first step, the signals were examined for frequency, autocorrelation techniques were applied, and amplitude distributions were calculated [32]. Such operations have been done by other investigators, and on their own do not yield much useful information. Nevertheless, it is worth noting that these operations require only a few seconds on a MINC II minicomputer which has the fast Fourier transform readily available. The second stage involves application of pattern recognition techniques to the results from the first stage. The procedure followed so far has been a compound classifier method described by Batchelor [33]. This is not the best approach [34, 35] but it was straightforward to implement. Some representative results are shown in Fig. 3.

From the outset, the data have been examined by using only three variables. This has proved most informative, despite the lack of statistical validity, by providing pointers to useful experiments and tests for worthwhile parameters. Had the application of pattern recognition techniques been delayed until a satisfactory number of results had been obtained, much time would have been lost, and it is likely that pattern recognition would have shown that the wrong set of experiments had been chosen. As it is, the study has shown that acoustic emissions from polymers, composites and chemical reactions do fall into distinctive clusters and that they can form the basis of identification. The next step is to use a supervised learning method to complement the unsupervised learning approach [34, 35] so far followed and to collect acoustic and mechanical data to establish whether satisfactory working relationships can be established. There is no doubt that the multivariate approach has helped to make an immensely complicated problem of

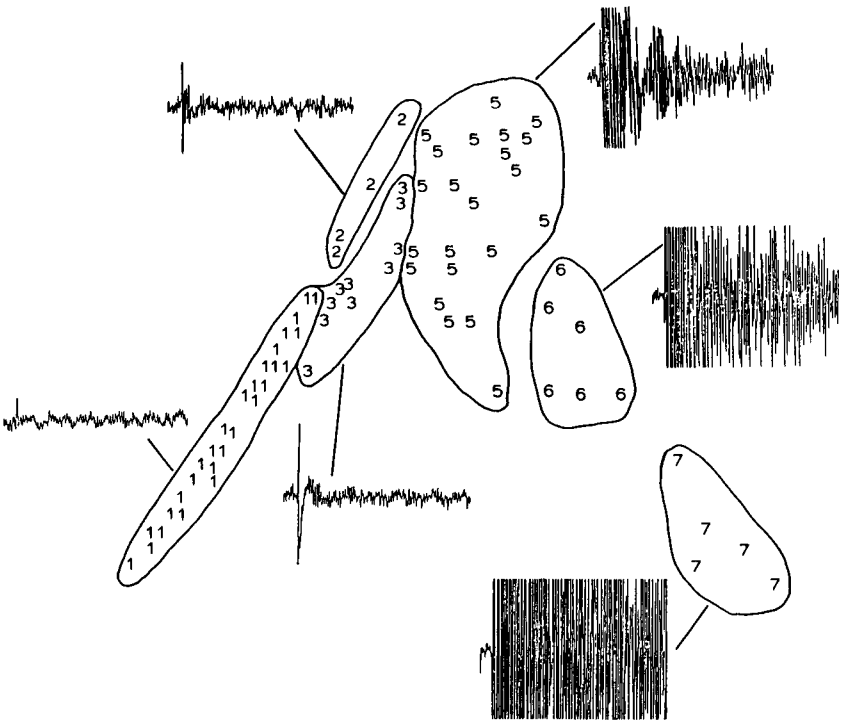


Fig. 3. Cluster analysis of 120 acoustic emissions from a glass-reinforced polypropylene composite.

data evaluation much more tractable. This approach is also much more efficient, for in the initial stages of the work, it took 3 days to evaluate one experiment. Now, at the supervised learning stage, the classification takes less than a minute, which opens up the possibilities of the acoustic emission experiment being done on-line with the standard mechanical tests.

Infrared spectrometry

By contrast to the above research problem, one may contemplate the problem faced by Perkin-Elmer in designing a modern infrared spectrometer to take full advantage of the reliability and stability of modern electronics, and to incorporate basic data-processing routines. The consequence is that additives in the concentration range of 20—50 ppm can be identified and determined with reasonable precision (5%) even though their existence in the spectrum of the major compound could not be detected visually. Additionally, it incorporates what is effectively on-line pattern recognition. As a large library search for peak matching would obviously be impossible within the confines of a single instrument, it was decided to identify the structural unit present and then to fit peaks to those compounds which were suggested by those units. Altogether 896 possible structural units have been identified

and classified on the basis of the absence or presence of i.r. peaks in a given spectral region. These can then be further classified into 48 major functional groups. This classification takes 8 s, and can be displayed with a score given to represent the probability of success. The next stage is to use this information and other basic data to search a library to make a compound identification. The development costs of the computer programming alone are reported to be equivalent to 9.25 man years. Thus although chemometric procedures may be incorporated in commercial instrumentation, the manufacturer must be convinced of the need and the reliability of the algorithm before converting the current state of the art into push-button facility.

SYSTEMS IMPOSSIBLE WITHOUT A COMPUTER

Rapid scanning u.v./visible spectrometer

Multi-element sensors, such as vidicon arrays, have been known for some time [36]. In a spectrometer, they make it unnecessary to scan wavelengths mechanically, because the light beam passes to a diffraction grating and then to the multielement detector, which typically has 250 individual sensors. Thus a spectrum of 250 points is obtained; in the visible region this corresponds to a resolution of about 2 nm. The replacement of moving parts by a fast electronic scan means that a spectrum can be obtained in ca. 10 ms every second. In a Hewlett-Packard instrument, data can be collected continuously at several wavelengths or the whole spectrum can be recorded. The spectra are digital and so mixtures may be resolved and their components quantified, and differential spectra are readily obtained. In many respects, this is a revolutionary step in spectrophotometry, made possible only by the availability of microcomputers.

Computerised tomography

Computerised tomography is a form of non-destructive imaging in which a cross-section of the sample is displayed [37, 38]. It has been most useful in medicine for obtaining rapidly the equivalent of sections of vital organs non-destructively. However, the principles involved are general and the technique has many analytical applications.

The basic idea is illustrated in Fig. 4. The absorption of radiation by the sample at well-defined positions is recorded, as a measure of the amount and density of the sample along the path of the radiation. Over 100 measurements may be made in one plane, although low-resolution images may be obtained with less. The process can be repeated for different sample heights. Then, by a rather complicated computational process, the absorption data can be converted to an image of the cross-section of the sample. In Fig. 4, solids are used to illustrate the principle, but in reality shading [39] can be introduced so that the cross-sectional image clearly reveals variations in concentration. Recently, n.m.r. has been used as the basis of tomography in medical applications. The n.m.r. shift and intensity can be related to specific

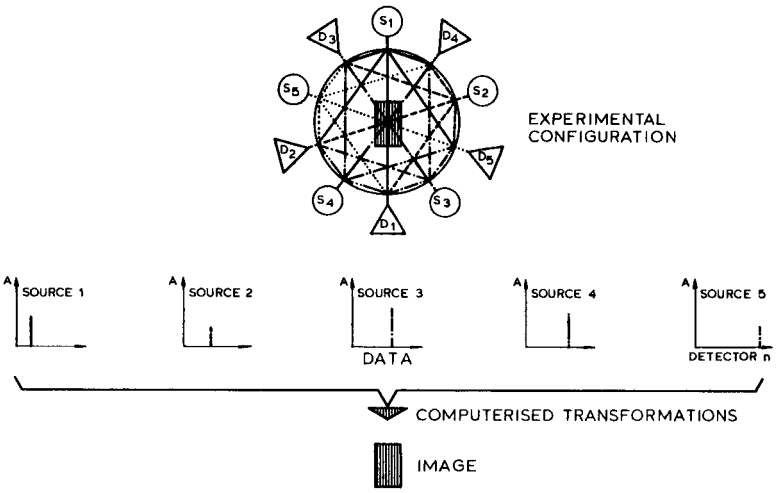


Fig. 4. Principle of computerised tomography.

tissues so that the computed image reveals the presence and position of tumours. The sequence of cross-sections then reveals precisely the geometrical position and dimensions of the growth [40]. Currently, the examination of one relevant section of the body (e.g., brain, heart, liver) takes about 5 min.

The procedure is a general one and is applicable to all types of radiation, and may be extended to emission as well. This may be of advantage, for instance, in determining the distribution of radioactive isotopes in nuclear fuel rods. However, the computational procedure is different for emission and some forms of radiation may require that factors such as scattering and counting statistics are considered, with further complication of the computations.

So far there have been comparatively few applications to the analysis of materials. Pierce et al. [41] have shown that with relatively rudimentary apparatus, based on a γ-ray source and two Z380 microcomputers for data collection and data processing, mixtures of materials can be evaluated quite successfully [41].

Conclusions

This survey has been selective, and many major applications of computers to analytical chemistry have been ignored. However, it shows that computers and chemometrics can be incorporated into analytical instrumentation with benefits in terms of speed and reliability. The cost of developing such instrumentation may be heavy, particularly for unusual designs. Instrument brochures contain details of mechanical, optical and electronic layout and specifications but rarely, if ever, indicate what algorithms form the basis of the computational routines incorporated in the instrument, however unusual they may be. Chemometric procedures will increasingly be incorporated as standard over the next decade.

It is a pleasure to acknowledge the assistance of Perkin-Elmer, Hewlett-Packard, Vacuum Generators, J. Huddleston and T. B. Pierce (A.E.R.E., Harwell) in making material available for the lectures on which this paper is based.

REFERENCES

1 P. Hepple (Ed.), The Application of Computer Techniques in Chemical Research, Institute of Petroleum, London, 1972.
2 J. M. Skinner, in P. Hepple (Ed.), The Application of Computer Techniques in Chemical Research, Institute of Petroleum, London, 1972, p. 253.
3 R. L. Erskine and N. G. M. Taggert, in P. Hepple (Ed.), The Application of Computer Techniques in Chemical Research, Institute of Petroleum, London, 1972, p. 29.
4 J. S. Mattson, H. B. Marks, Jr. and H. C. MacDonald, Jr. (Eds.), Computer Fundamentals for Chemists, Dekker, New York, 1973.
5 J. S. Mattson, H. B. Marks, Jr. and H. C. MacDonald, Jr. (Eds.), Computer Assisted Instruction in Chemistry, Dekker, New York, 1974.
6 D. Betteridge and T. B. Goad, Analyst, 106 (1981) 257.
7 I. D. Frank and B. R. Kowalski, Anal. Chem., 54 (1982) 232R.
8 T. W. Barnard, Anal. Chem., 51 (1979) 1172A.
9 J. W. Frazer, Anal. Chem., 52 (1980) 1205A.
10 D. Betteridge, E. L. Dagless, P. David, D. R. Deans, G. E. Penketh and P. Shawcross, Analyst, 101 (1976) 409.
11 T. F. Christiansen, J. E. Busch and S. C. Krogh, Anal. Chem., 48 (1976) 1051.
12 A. H. B. Wu and H. V. Malmstadt, Anal. Chem., 50 (1978) 2090.
13 J. Růžička and E. H. Hansen, Flow Injection Analysis, Wiley, New York, 1981.
14 D. Betteridge, K. C. Cheng, P. David, E. L. Dagless, T. B. Goad, D. R. Deans, D. A. Newton and T. B. Pierce, Analyst, 108 (1983) 1, 17.
15 J. F. Vanderslice, K. K. Stewart, A. G. Rosenfeld and D. J. Higgs, Talanta, 28 (1982) 11.
16 R. Tijssen, Anal. Chim. Acta, 114 (1980) 71.
17 G. Taylor, Proc. R. Soc. (London), Ser. A., 219 (1953) 186.
18 O. Lievenspiel, Chemical Reaction Engineering, 2nd edn., Wiley, New York, 1972.
19 J. M. Smith, Chemical Engineering Kinetics, 2nd edn., McGraw-Hill, Tokyo, 1970.
20 H. L. Pardue and B. Fields, Anal. Chim. Acta, 124 (1981) 39, 65.
21 J. G. Atwood and M. J. E. Golay, J. Chromatogr., 218 (1981) 97.
22 J. A. Nelder and R. Mead, Comput. J., 7 (1965) 308.
23 J. L. Morgan and S. N. Deming, Anal. Chem., 46 (1974) 1170.
24 M. N. Routh, P. A. Swartz and M. B. Denton, Anal. Chem., 49 (1977) 1422.
25 P. B. Ryan, R. L. Barr and H. D. Todd, Anal. Chem., 52 (1980) 1460.
26 D. Betteridge, A. P. Wade, E. A. Neves and I. Gutz, Proc. III Simp. Brasil Electroquim. Electroanalitica, San Carlos, 1982.
27 S. Stieg and T. A. Nieman, Anal. Chem., 52 (1980) 800.
28 T. J. Sly, D. Betteridge, D. Wibberley and D. G. Porter, J. Automatic Chem., 4 (1982) 186.
29 R. V. Williams, Acoustic Emission, Hilger, Bristol, 1980.
30 J. T. Brock, Mechanical Vibration and Shock Measurements, Brüel and Kjaer, Copenhagen, 1980, Ch. 9.
31 D. Betteridge, J. V. Cridland, T. Lilley, M. E. A. Cudby and D. G. Wood, Polymer, 23 (1982) 178.
32 D. Betteridge, M. T. Joslin and T. Lilley, Anal. Chem., 63 (1981) 1064.
33 B. G. Batchelor, Practical Approach to Pattern Recognition, Plenum, London, 1976.
34 K. Varmuza, Pattern Recognition in Chemistry, Springer-Verlag, Heidelberg, 1980.
35 L. Kryger, Talanta, 28 (1981) 871.

36 H. L. Pardue, in J. K. Foreman and P. B. Stockwell (Eds.), Topics in Automatic Chemical Analysis, Horwood, Chichester, 1979, Ch. 6.
37 G. T. Herman, Image Reconstruction from Projections, Academic Press, New York, 1980.
38 E. L. Hall, Computer Image Processing and Recognition, Academic Press, New York, 1979.
39 M. G. Moran and B. R. Kowalski, Anal. Chem., 51 (1979) 776A.
40 R. Gordon, G. T. Herman and S. A. Johnson, Sci., Am., 233 (1975) 56.
41 T. B. Pierce, J. Huddleston and I. G. Hutchinson, J. Radioanal. Chem., in press.

Analytica Chimica Acta, 150 (1983) 129—143
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

PRINCIPAL COMPONENTS AND PARTIAL LEAST-SQUARES ANALYSIS OF THE GEOCHEMISTRY OF VOLCANIC ROCKS FROM THE AEOLIAN ARCHIPELAGO

M. LAURA BISANI and DOMENICO FARAONE
Department of Earth Sciences, University of Perugia (Italy)
SERGIO CLEMENTI*
Department of Chemistry, University of Perugia, 06100-Perugia (Italy)
KIM H. ESBENSEN and SVANTE WOLD
Research Group for Chemometrics, Institute of Chemistry, University of Umeå (Sweden)
(Received 6th October 1982)

SUMMARY

A data set from total chemical analysis of 53 volcanic rocks from five islands of the Aeolian Archipelago (11 major and 13 trace elements) was examined by the multivariate SIMCA and PLS data analysis programs, in an attempt to evaluate the geochemical potential of the methods. The results outline: (a) the optimal within-islands classification and discrimination variables; and (b) the relative information content of the blocks of major elements and trace elements. A measure of the common systematic information of these two blocks is also obtained. The study illustrates blockwise, related pattern recognition of the level-4 type.

Chemometric techniques have made an impact in geological and geochemical sciences; numerous examples of individual studies are available, and the commonly used methods of data handling have been reviewed [1, 2]. Methods such as factor analysis or principal component analysis appear to be of particular interest for these multivariate disciplines. This paper is intended to show how these methods enable volcanic rock samples from the Aeolian Archipelago to be classified correctly according to the islands where they were collected. In addition, by combination of the traditional principal components analysis (PCA) with partial least-squares analysis (PLSA), it is possible to evaluate the information contents of the major elements composition and of the trace elements composition, so that their relative usefulness can be judged.

The data set was taken from a series of papers on the Aeolian Archipelago published recently [3]. The data matrix consists of 53 geochemical samples (objects) for which the percentage contents of eleven major elements (as oxides) and the $\mu g\ g^{-1}$ contents of thirteen trace elements (24 variables) were available. The data set is summarized in Table 1; the entire data set is available

0003-2670/83/$03.00

TABLE 1

Summary of data set (N = 53 objects)

Variable No.	Range of content[a]	Average	S.d.[b]	R.s.d.[b]
1. SiO_2	47.9—74.3	55.64	5.37	0.10
2. TiO_2	0.09—0.90	0.68	0.18	0.27
3. Al_2O_3	13.0—19.6	16.90	1.51	0.09
4. Fe_2O_3	0.01—6.43	4.09	1.17	0.29
5. FeO	0.95—6.50	3.21	1.40	0.44
6. MnO	0.01—0.20	0.14	0.04	0.28
7. MgO	0.33—10.91	4.63	2.18	0.47
8. CaO	0.84—12.60	7.91	2.51	0.32
9. Na_2O	2.00—4.20	3.08	0.61	0.20
10. K_2O	0.80—5.80	2.40	1.46	0.61
11. P_2O_5	0.01—0.52	0.30	0.15	0.49
12. Rb	14—350	76.1	75	0.99
13. Sr	21—1400	703	249	0.36
14. Ba	167—1318	545	241	0.44
15. Y	10—41	22.6	7.2	0.32
16. Zr	19—205	110	50.8	0.46
17. V	30—350	205	75.2	0.37
18. Cr	7—298	58.3	67	1.15
19. Co	4—36	19.1	8.7	0.46
20. Ni	3—104	27.1	23	0.83
21. Cu	20—159	99.8	42	0.42
22. Zn	51—78	68.4	9.9	0.15
23. La	11—58	33.3	13	0.39
24. Ce	26—132	66.6	30	0.46

[a]Given as % for 1—11, and $\mu g\ g^{-1}$ for 12—24.
[b]Standard deviation, and relative standard deviation expressed as s.d. divided by average (%), correctly called coefficient of variance.

from the authors on request. The data were evaluated by the SIMCA/MACUP package [4, 5].

The archipelago (located a few miles north of Sicily, Italy) is formed by seven major islands: Alicudi, Filicudi, Salina, Lipari, Vulcano, Panarea and Stromboli listed from west to east. Unfortunately, data on trace elements were not available for the latter two islands. The archipelago is often considered an island arc, according to plate tectonics [6, 7], though this is debatable [8]. The geological significance of the present work will be reported elsewhere.

PRINCIPAL COMPONENTS ANALYSIS (SIMCA)

The SIMCA method, based on disjoint principal component models for each class of homogeneous objects, has been described in detail [4, 5]. The data matrix contains the members y_{ik} where index i is used for the chemical

elements (variables) and index k for the rock samples (objects). Each member of the data set is described by

$$y_{ik} = \alpha_i + \sum_{a=1}^{A} \beta_{ia} \vartheta_{ak} + \epsilon_{ik} \tag{1}$$

where the number A of significant cross-terms and the parameters α_i, β_{ia} and ϑ_{ak} are estimated by minimizing the squared residuals ϵ_{ik}. In this method, α_i and β_{ia} are constants which only depend upon the variables and ϑ_{ak} depends on the samples. The deviations from the model are expressed by the residuals ϵ_{ik}.

Before statistical analysis, the data are autoscaled, i.e., the variables are given the same variance, fixed to unity. With this scaling, all variables are given equal importance in principal components analysis (PCA). The analysis then proceeds by model expansions, i.e., finding the correct dimensionality, A. First, a model with $A = 0$ is fitted to the data, which means that each chemical element is described by its mean value α_i. Then the α_i value for each variable is subtracted from the matrix members y_{ik}, thus giving residuals of dimension zero. A straight line is subsequently fitted to these residuals whereby the $\beta_{i1}\vartheta_{1k}$ term is estimated. Whether this first dimension is significant or not (i.e., whether or not the residuals contain systematic information) is established by cross-validation [9]. The new residuals for this $A = 1$ model are calculated by subtracting the term $\beta_{i1}\vartheta_{1k}$. If the new residuals still contain systematic information, additional $\beta_i\vartheta_k$ terms are then estimated one after the other, until the residuals contain only noise.

Once a definite model has been established, the total residual standard deviation, s_A, for A significant components

$$s_A^2 = \sum_{i=1}^{M} \sum_{k=1}^{N} \epsilon_{ik}^2/(N - A - 1)(M - A) \tag{2}$$

can be used to judge if an individual sample belongs to the class or not. Each data vector y_{ip} can be fitted to the class parameters by multiple regression:

$$y_{ip} - \alpha_i = \sum_{a=1}^{A} t_{ap} \vartheta_{ak} + e_{ip} \tag{3}$$

How well the data vector fits the model is expressed by its residual standard deviation, s_p:

$$s_p^2 = \sum_{i=1}^{M} e_{ip}^2/(M - A) \tag{4}$$

Comparison of s_p^2 with s_A^2 by an F-test ($F = s_p^2/s_A^2$) makes it possible to decide if the datum point lies within the confidence region of the class or has to be considered as an outlier.

Results

Preliminary screening of the whole matrix makes it possible to evaluate the overall inter-island sample homogeneity. According to the criterion expressed by the *F*-test, three points (objects 24, 29 and 41) were found to be outliers. Moreover, inhomogeneities in the data set were also observed in a plot of the first against the second component. Four samples (objects 7, 8, 19, 20), representing the persilicic rocks, lie in a different region of the 24-dimensional space. When these points are included in the full analysis, they indeed comply with the model according to the previous distance criterion, but they are excluded in the present evaluation.

The polished data set therefore consists of 46 objects, 6 of which belong to the island of Vulcano (1—6), 7 to Salina (9—15), 3 to Lipari (16—18), 19 to Alicudi (21—23, 25—28, 30—40, 42), and 11 to Filicudi (43—53). The global p.c.a. was repeated on this set, with the results reported in Tables 2 and 3. The data were again autoscaled, except for variables 2, 6 and 11 which were left unchanged because of their very low standard deviation (Table 1).

Four significant components are implied (cross-validation); these account for a total data variance of 78%. The fraction of variance explained by each component is given in Table 2, whereas Table 3 lists the numerical values of w_i, α_i, β_{ia} and ψ_i (the modelling power [4]):

$$\psi_i = 1 - s_i(A = A)/s_i\ (A = 0) \tag{5}$$

The four-dimensional hyperplane representing this model cannot be displayed directly, but only through "windows" implying projection into two dimensions. The two most interesting such plots (ϑ_1 vs. ϑ_2 and ϑ_2 vs. ϑ_3) are shown in Figs. 1 and 2. The plots clearly indicate the existence of at least four different classes in the data set: one includes the rocks of Vulcano, another the rocks of Salina and Filicudi, whereas the rocks of Alicudi are represented by two well separated groups.

The SIMCA method is based on fitting disjoint principal component models to each group of homogeneous objects. The subsequent step is to fit a separate model to the rock series from each island. It is known from parallel studies on larger data sets, where only the major element contents are available, that the rocks of each island constitute more refined sets of subclasses. Because the number of available samples in this study is too small for an analogous detailed evaluation (e.g., only six rocks from Vulcano to

TABLE 2

Residual standard deviation (s_A) and fraction of variance (%V) explained by model expansion on the whole data set ($N = 46$)

A	0	1	2	3	4
s_A	0.93	0.73	0.62	0.51	0.44
V (%)	0	38	56	70	78

TABLE 3

Statistical results of p.c.a. on the whole data set ($N = 46$)

Variable	w_i[a]	α_i	β_{i1}	β_{i2}	β_{i3}	β_{i4}	ψ_i
1 (Si)	0.265	14.48	0.25	−0.31	−0.05	−0.15	0.75
2 (Ti)	1	0.71	−0.01	0.01	0.01	0.03	0.02
3 (Al)	0.770	13.17	0.02	0.04	−0.42	0.45	0.53
4 (Fe^3)	0.987	4.17	−0.16	−0.05	0.14	0.56	0.51
5 (Fe^2)	0.741	2.52	−0.13	0.34	−0.18	−0.34	0.63
6 (Mn)	1	0.15	−0.01	0.01	−0.01	−0.01	0.25
7 (Mg)	0.491	2.41	−0.24	0.02	0.36	−0.06	0.62
8 (Ca)	0.512	4.28	−0.31	0.18	−0.03	−0.01	0.73
9 (Na)	1.716	5.16	0.27	−0.01	0.04	0.17	0.42
10 (K)	0.965	2.02	0.28	0.22	0.16	−0.03	0.75
11 (P)	1	0.31	0.01	0.01	0.02	0.03	0.06
12 (Rb)	0.024	1.44	0.27	0.21	0.13	−0.02	0.60
13 (Sr)	0.0046	3.38	0.08	0.46	0.13	0.09	0.68
14 (Ba)	0.0043	2.43	0.27	0.24	0.22	0.03	0.77
15 (Y)	0.184	3.97	0.01	−0.36	0.03	0.23	0.45
16 (Zr)	0.024	2.41	0.26	−0.13	0.19	0.10	0.49
17 (V)	0.018	3.92	−0.29	0.06	0.03	0.24	0.69
18 (Cr)	0.014	0.87	−0.19	−0.10	0.45	−0.13	0.70
19 (Co)	0.132	2.77	−0.20	0.06	−0.04	0.13	0.17
20 (Ni)	0.042	1.17	−0.20	−0.13	0.43	−0.01	0.68
21 (Cu)	0.026	2.76	−0.14	0.34	0.06	0.20	0.60
22 (Zn)	0.109	7.63	−0.12	0.27	0.04	0.23	0.35
23 (La)	0.081	2.59	0.21	0.01	0.30	0.22	0.43
24 (Ce)	0.043	2.63	0.29	0.14	0.08	0.09	0.75

[a] w_i is the inverse of the variable standard deviation (weight used to normalize the data).

represent five different classes [10]), the present evaluation is limited; the rocks from each island are considered as a homogeneous set.

The results of the disjoint analysis are reported in Table 4. The models, here autoscaled within each class, are calculated for six separated subsets. They are all reported with two significant components, which explains 56–80% of the total data variance; in two cases, the second component is relatively weak. (An exception is the model for the island of Lipari, where only the first component is considered, owing to the presence of three points only; it is not representative of any $A = 1$ model.) The resulting parameters α, β and ϑ are not reported as they are mostly meaningful only for the subsequent classification computations on this relatively small data set.

Classification

Once the parameters in the class models have been established, all objects can be fitted to these models by Eqn. (3). If the objects of one class, say V, are all fitted to the model of another class, say S, the resulting residuals e_{ip} indicate how far the objects of class V lie from class S. Specifically, the

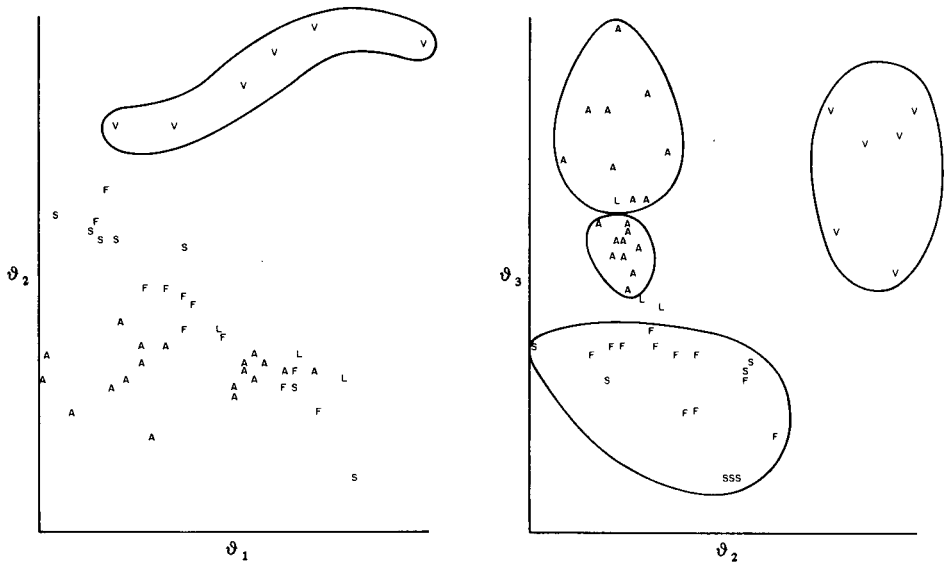


Fig. 1. Plot of the first component, ϑ_1, against the second component, ϑ_2, obtained by PCA of the whole set, showing the separation of Vulcano from the other islands. Rock samples are identified by the initial letter of each island name.

Fig. 2. Plot of the second against the third component obtained by PCA of the whole set, showing the separation of Vulcano (V), Alicudi I (A, upper), Alicudi II (A, lower), Filicudi and Salina (F, S).

TABLE 4

Residual standard deviations and fraction of variance explained by model expansion for the disjoint class models

	s_0	s_1	$\%V_1$	s_2	$\%V_2$
Vulcano (n = 6)	0.93	0.63	54	0.53	68
Salina (n = 7)	0.93	0.53	68	0.43	79
Filicudi (n = 11)	0.93	0.47	74	0.42	80
Alicudi I (n = 10)	0.87	0.68	39	0.54	61
Alicudi II (n = 9)	0.87	0.65	44	0.58	56
Lipari (n = 3)	0.92	0.76	32		

standard deviation of these residuals can be calculated for each variable by using the equation

$$s_{i,V}^{(S)2} = \sum_{p=1}^{N} e_{ip}^{(S)2}/N_V \tag{6}$$

By means of these, and the standard deviations obtained when the objects are fitted to their own class models, the discrimination power $d_i(V, S)$ for variable i between classes V and S is obtained by

$$d_i(V, S) = [s_{i,v}^{(S)2} + s_{i,S}^{(V)2}] / [s_{i,V}^{(V)2} + s_{i,S}^{(S)2}] \quad (7)$$

When the discrimination power is pooled over all variables (Eqn. 7), the distance between classes V and S, $D_{V,S}$ can be calculated:

$$D_{V,S} = \left\{ \left[\sum_{i=1}^{M} (s_{i,V}^{(S)2} + s_{i,S}^{(V)2}) \right] \Big/ \left[\sum_{i=1}^{M} (s_{i,V}^{(V)2} + s_{i,S}^{(S)2}) \right] \right\}^{1/2} - 1 \quad (8)$$

Values of D lower than one indicate poor resolution, while values larger than 2—3 correspond to good resolution. By fitting each object to each separated class, it is possible to calculate all the interclass distances. These distances (between classes for each island), calculated by Eqn. (8), are given in Table 5 in the form of a distance matrix. It is clear that Salina and Filicudi constitute the closest pair, while Filicudi and the first group of Alicudi (AI, older rocks) form the most distant pair. Moreover, Vulcano and the second group of Alicudi (AII, younger rocks) also exhibit quite a large interclass distance from any other class. The d_i values are not calculated, in view of the rough homogeneity criterion and the small class sizes used in this preliminary work.

PARTIAL LEAST-SQUARES ANALYSIS (MACUP)

While PCA gives equal weight to all variables, other statistical approaches have to be used to test the existence of relationships between one or more dependent variables and a group of explanatory variables. The former case is usually treated in chemistry by multiple regression techniques [11]. However, the multiple regression method assumes that all the descriptor variables are independent, error-free, and 100% relevant to the problem. The PLS technique was recently developed [5, 12, 13] as an alternative method to handle such problems, but where the relevance of each "independent" variable results from the statistical analysis.

When the problem under investigation does not involve a single dependent variable, there are in fact two blocks of variables, and it becomes possible to define a dependent variable/object matrix **Y** and an independent matrix **X** [5, 12, 13]. The question is whether or not the members of the **Y** matrix can be described as a simple function of the members of the **X** matrix. In

TABLE 5

Interclass distance matrix

	S	*F*	*AI*	*AII*	*L*
Vulcano	5.8	10.6	7.3	12.8	7.0
Salina		2.7	3.1	9.1	4.0
Filicudi			17.4	7.2	3.8
Alicudi I				4.6	15.7
Alicudi II					8.0

general, this problem is handled by computing principal component models for each of the two matrices, followed by establishment of any linear relationships between the principal components of these two blocks, respectively.

Instead of this two-step procedure, it is now possible to make a single analysis in which the two steps are achieved simultaneously. This method is called PLS2, and current experience shows that it is computationally much faster than PCA followed by multiple regression and that it leads to a better prediction of the Y members [4, 12, 13].

The PLS2 method gives a description of the X matrix by one principal component-like model (Eqn. 9), a description of the Y matrix by an analogous model (Eqn. 10), and predictive relations between the latent variables ξ and η (Eqn. 11).

$$x_{ik} = \alpha_i + \sum_{a=1}^{A} \beta_{ia}\xi_{ak} + \epsilon_{ik} \tag{9}$$

$$y_{jk} = \alpha_j + \sum_{a=1}^{A} \beta_{ja}\eta_{ak} + \epsilon_{jk} \tag{10}$$

$$\eta_{ak} = \rho_a \xi_{ak} + \epsilon_k \tag{11}$$

where ρ is a proportionality coefficient for each dimension.

The algorithm used in this MACUP method, presented in detail earlier [5], is iterative for each dimension as in PCA. It consists of finding the latent variables of the X matrix (ξ_k) from starting values of η_k and the X elements, and then recomputing the latent variables of the Y matrix (η_k) from the Y elements and the ξ_k values until the process converges. The meanings of β and ξ correspond to β and ϑ in PCA and can therefore be used in the same way in, for instance, classification. However, it should be pointed out that classification is based on the X block variables only.

Results

For the present geochemical problem, where the major element and the trace element contents are available for a number of rock samples, the question arises of whether it is possible to describe, and therefore predict, the trace element content from the major element content or vice versa. The importance of this problem for the magnitude of the analytical work in geochemistry seems obvious. In other words, this may provide a means of evaluating to what extent the major and trace elements contain the same information, namely, if one group contains some intrinsic information which cannot be related to the other block.

Accordingly, the principal components method was used independently on the major element and trace element blocks, and two PLS2 analyses were applied, to consider the dependence of trace on major elements and vice versa; the raw data were normalized as before. The results made it possible to

evaluate: (a) the variance explained by principal component models within each variable block; (b) the classification ability of each block; (c) how much of the variance of each block can be described as a function of the other; (d) the fraction of variance of each block which cannot be described in terms of the other, thus indicating the relative within-block information contents. Details are reported in Tables 6 and 7 and illustrated in Figs. 3—7.

The β values obtained from PCA for the two subdata sets are listed in Table 6 (the ϑ, ξ, η parameters and the β from PLS2 are not given here). Table 7 shows the fraction of variance explained by the principal component models for the major element block and the trace element block. Moreover, the fraction of variance explained as a function of the other block in the PLS2 method is indicated. The results are summarized in Fig. 3. It is clear that the two blocks contain mainly the same information up to the second dimension. However, there are four significant components and 15—18% of the variance explained by PCA for each block cannot be accounted for by the PLS2 description as a function of the other block. Consequently, both the trace elements and the major elements provide significant information of their own that renders particularly appropriate the determination of both of them in volcanic rocks. It is possible, however, to exclude a number of "redundant variables" by careful evaluation of the systematics of β-patterns (see below).

The trace element block has a better classification ability, as illustrated by Figs. 4 and 5. While the principal component projections obtained by PCA on the major elements do not show a clear separation between the individual islands (the ϑ_2 vs. ϑ_4 plot is given in Fig. 4 as the best illustration), the trace element block provides a separation almost as good as the whole set (ϑ_2 vs. ϑ_3 plot in Fig. 5). This is a situation typical of geochemical classification problems involving basaltic rocks [14].

TABLE 6

List of β values for the separated models for the major and trace element blocks

Variable	β_{i1}	β_{i2}	β_{i3}	β_{i4}	Variable	β_{i1}	β_{i2}	β_{i3}	β_{i4}
(a) Major elements					*(b) Trace elements*				
1 (Si)	−0.44	−0.11	0.12	−0.50	12 (Rb)	0.39	0.09	0.11	0.12
2 (Ti)	0.01	−0.01	−0.01	0.04	13 (Sr)	0.22	0.45	0.14	0.04
3 (Al)	−0.06	−0.07	−0.85	0.05	14 (Ba)	0.39	0.13	0.21	0.09
4 (Fe^3)	0.24	−0.65	−0.07	0.38	15 (Y)	−0.09	−0.46	0.02	−0.04
5 (Fe^2)	0.27	0.68	−0.15	0.06	16 (Zr)	0.31	−0.30	0.24	−0.20
6 (Mn)	0.01	0.01	−0.01	0.01	17 (V)	−0.33	0.17	0.11	−0.21
7 (Mg)	0.38	−0.14	0.43	0.03	18 (Cr)	−0.25	−0.03	0.57	0.27
8 (Ca)	0.49	0.09	−0.09	0.11	19 (Co)	−0.20	0.13	0.13	−0.75
9 (Na)	−0.42	−0.03	−0.02	0.38	20 (Ni)	−0.26	−0.06	0.58	0.12
10 (K)	−0.34	0.25	0.21	0.66	21 (Cu)	−0.06	0.47	0.13	−0.04
11 (P)	−0.01	−0.01	0.01	0.06	22 (Zr)	−0.04	0.42	0.14	−0.28
					23 (La)	0.29	−0.16	0.36	−0.39
					24 (Ce)	0.41	−0.02	0.05	−0.10

TABLE 7

Comparison of standard deviations and variance explained for the major and the trace element block evaluated by PCA and by PLS2

	Major elements					Trace elements				
	PCA		s_y	PLS2		PCA		s_y	PLS2	
	s_A	%V		%V	ρ	s_A	%V		%V	ρ
$A = 0$	0.85	0	0.85	0	1	1	0	0.89	0	1
$A = 1$	0.65	42	0.68	36	0.87	0.80	36	0.74	31	0.92
$A = 2$	0.56	57	0.60	50	0.73	0.66	56	0.61	53	1.37
$A = 3$	0.44	73	0.51	64	0.68	0.51	74	0.54	63	0.95
$A = 4$	0.33	85	0.49	67	0.79	0.44	82	0.51	67	0.52

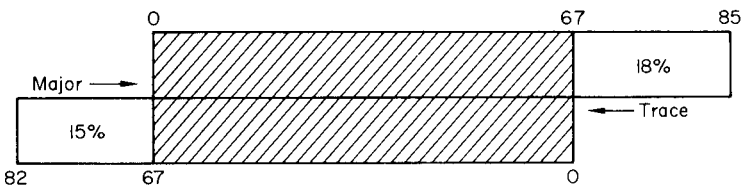


Fig. 3. Information contained in the major (upper row) and trace (lower row) element blocks.

The presence of different classes in the data set can also be indicated directly from the PLS2 procedure. Figures 6 and 7, representing respectively the plots of the second and the third components of the trace element block (η) described by the major element block (ξ), indicate a result similar to that obtained by PCA. In this example, as in Figs. 3, 5 and 6 above, it is the higher-order components that discriminate best between different classes. The significance of this is discussed later.

The PLS2 technique provides other important information as well, however. Four components are significant (Table 6). This means that four "orthogonal" compound chemical "effects" are operating jointly to produce the systematic ca. 80% of the total variance of the data that have been modelled with these components. The implication is that effects are transferred from the block of major elements to the block of trace elements (or vice versa) in the geological process(es) that is (are) responsible for the data distribution.

It is also possible to run PLS2 for each individual island after using the same scaling as for the whole set. Each model has only one significant component except the two Alicudi groups, which are better described by zero or a two-component model, both being rather weak, however. The results of this evaluation are reported in Table 8, which lists the α and β_1 values for the six disjoint classes. While the α values are obviously the same as in PCA, the information contained in the β values is somewhat easier to interpret when

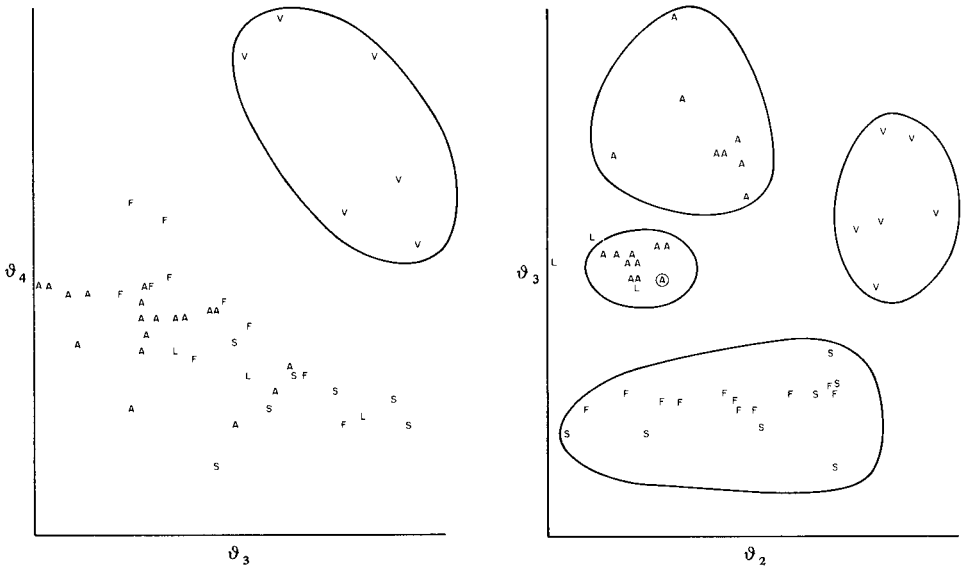


Fig. 4. Plot of the third against the fourth component obtained by PCA of the major elements matrix. Only the samples of Vulcano are easily separated.

Fig. 5. Plot of the second against the third component obtained by PCA of the trace elements matrix. The separation is as good as in Fig. 2, except for the circled point (object 27).

related to the disjoint modelling, compared to the compound discrimination power discussed above for the global analysis. The β values, taken blockwise, reflect the relative variable interrelations between any two classes, whereas the α values reflect similarities or differences related to the average chemical composition of each class (island). Thus Vulcano has higher contents of Na, K, Sr and Ba, Salina has lower contents of Ti, P, Zr and rare earths, the older rocks of Alicudi show very high Ni, Cr, V and Mg and low Rb contents, etc.

The evaluation of the β_1 values, referring here to description of the trace elements in terms of the major elements, shows which elements behave similarly or differently as modelled within each class. As an example, for the class pairs of Vulcano and Salina, the most relevant differences are in the Al, Sr, Co, Cu and Zn contents, despite their total concentrations, which often cannot be distinguished (cf. Table 8). This illustrates the superior modelling/classification/discrimination abilities related to the covariance/correlations between a multivariate array, as opposed to the simple comparison of average vectors. This type of comparison allows an evaluation of the intrinsic variable-interrelationship characteristics of each class (island), and so provides further clues about the nature of the geochemical processes involved in the genesis of the volcanic rock suite (e.g., depth of magma genesis, nature of fractionation en route before extrusion, assimilation by country rocks, etc.).

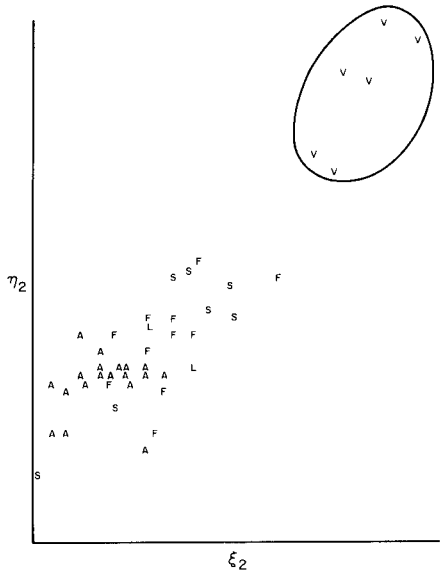


Fig. 6. Plot of the second component of the X block (ξ_2) against that of the Y block (η_2), obtained by the PLS2 method, showing the separation of Vulcano.

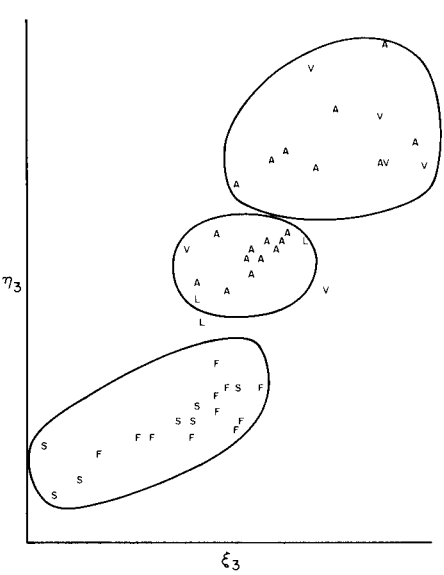


Fig. 7. Plot of the third component of the X block (ξ_3) against that of the Y block (η_3), obtained by the PLS2 method, showing the separation of the other three subclasses.

The data of Table 8 can be used to construct plots for each pair of classes, summarizing the systematics of the "transfer" relations between the two blocks of variables. Figure 8 illustrates these major element/trace element systematics for Vulcano vs. Salina by plotting the β_1 values: all elements outside the range $2 \times 2^{1/2} \times s^2_{\text{pool}}$ around the 1,1 line behave dissimilarly (see below). This type of plot is an example of level-four pattern recognition [15]. Alternatively, one can plot the respective ϑ values for the objects (rocks) to display the intrinsic relationships between the latent variables, as expressed in Eqn. (11). This gives a within-class graphic illustration of the characterizing "transfer" coefficient ρ_a (Fig. 9).

DISCUSSION

The geological implications of Table 8 will not be treated in depth here, but some types of interpretation that are possible will be given, as they are discipline-invariant. For example, it is apparent that some elements display characteristically high loadings (β values) in one or two classes only, e.g. Rb for Vulcano and Lipari only; that some elements collectively define a loading pattern characteristic of one class as opposed to others, e.g. Rb, Sr, Co, V, Cu and Zn for Vulcano; and that, in contrast, other variable sets serve to discriminate between classes, e.g., Al, Sr and Cr separate the β pattern between the islands of Vulcano and Alicudi and all others, the elements Fe^{3+}

TABLE 8

Values of α_i and β_{i1} obtained by PLS2 on disjoint classes[a]

Var.	α_V	α_S	α_F	α_{A1}	α_{A2}	α_L	β_{1V}	β_{1S}	β_{1F}	β_{iA1}	β_{iA2}	β_{1L}
1 (Si)	13.6	14.7	14.4	14.1	15.1	15.4	0.23	0.51	0.53	0.34	0.23	0.35
2 (Ti)	0.76	0.50	0.77	0.73	0.77	0.70	−0.01	−0.02	−0.01	0.03	0.01	−0.05
3 (Al)	12.5	13.8	13.6	12.6	13.4	12.8	0.48	−0.16	−0.29	0.28	0.16	−0.26
4 (Fe^3)	3.84	3.43	4.28	4.81	4.28	3.72	−0.03	−0.14	−0.48	−0.67	−0.58	0.13
5 (Fe^2)	3.41	3.42	2.49	2.05	1.78	2.46	−0.26	−0.39	−0.21	−0.13	0.04	−0.49
6 (Mn)	0.16	0.17	0.15	0.14	0.13	0.12	−0.01	−0.01	−0.01	−0.01	0.00	−0.03
7 (Mg)	2.58	2.04	1.84	3.67	2.15	1.62	−0.37	−0.26	−0.22	−0.05	−0.30	−0.26
8 (Ca)	4.50	4.54	4.40	4.79	3.53	3.37	−0.46	−0.48	−0.41	−0.48	−0.22	−0.46
9 (Na)	5.75	4.74	5.20	4.46	6.11	4.27	0.37	0.43	0.28	0.23	0.62	−0.08
10 (K)	3.72	1.43	1.94	1.27	2.01	2.76	0.40	0.24	0.27	0.22	0.22	0.51
11 (P)	0.39	0.18	0.31	0.31	0.40	0.17	0.02	0.01	−0.01	0.02	0.02	0.01
12 (Rb)	2.90	1.02	1.33	0.71	1.35	2.62	0.55	0.21	0.26	0.13	0.12	0.71
13 (Sr)	5.76	3.20	3.11	2.85	3.03	2.84	0.21	−0.07	−0.16	0.04	0.13	−0.28
14 (Ba)	4.31	1.88	1.95	1.77	2.76	2.88	0.36	0.24	0.28	0.34	0.46	0.22
15 (Y)	2.71	3.69		4.37	4.57		0.04	0.29		−0.30	−0.53	
16 (Zr)	2.48	1.47	1.96	2.17	3.25	4.46	0.40	0.32	0.29	0.30	0.33	0.11
17 (V)	3.33	3.99		4.80	3.60	2.58	−0.07	−0.43		−0.45	−0.36	−0.43
18 (Cr)	0.81	0.36	0.17	2.50	0.52	0.40	−0.22	−0.06	−0.02	−0.36	−0.29	0.01
19 (Co)	2.42	2.81	2.78			3.30	0.05	−0.45	−0.46			−0.37
20 (Ni)	0.86	0.57	0.44	2.80	1.07	0.49	−0.14	−0.12	−0.08	−0.43	−0.30	−0.01
21 (Cu)	3.71	2.71	2.61			1.55	0.10	−0.37	−0.50			0.18
22 (Zn)	8.36	7.50	7.33				0.06	−0.31	−0.42			
23 (La)	3.26	1.48	2.01	2.35	3.99		0.28	0.04	0.18	0.41	0.23	
24 (Ce)	3.65	1.99	2.48				0.44	0.26	0.27			
$s_y(A = 0)$	0.92	0.87	0.50	0.49	0.18	0.54						
$\%V(A = 1)$							0.64	0.78	0.86	0.32	0.27	0.63
$\rho(A = 1)$							0.91	0.92	0.78	1.16	0.45	0.94

[a]Underlined results are discussed in detail in the text.

and Rb between Vulcano and Lipari as opposed to all others, etc. Further interpretations of general interest can be given related to Tables 3, 6 and 8 and Figs. 8 and 9.

The β values indicate the relevance of each variable (element) in describing the principal components. The significant principal components represent the systematic part of the total data variance in their turn. Thus an outline of the most important variables with respect to delineation of the maximum systematic variance is furthered, for example, by suitable β plots, etc., which also give "maps" of the essential variable intercorrelations.

For the major element block, the most important elements are Si, Mg, Ca, Na and K, all loading with high (absolute) values on the first component (Table 6). The first component for the trace element block includes the important elements Rb, Ba, Zr, V, La and Ce. From a geochemical, petrological point of view this variable association is well known: it represents some primary elements involved in the geological process (fractional crystallization driving magmatic differentiation), that is responsible for the diversification of volcanic rock chemistries observed in the Aeolian Archipelago. This component thus represents the most similar variable-interrelations present in all individual classes; consequently, it is of little use in discriminating between

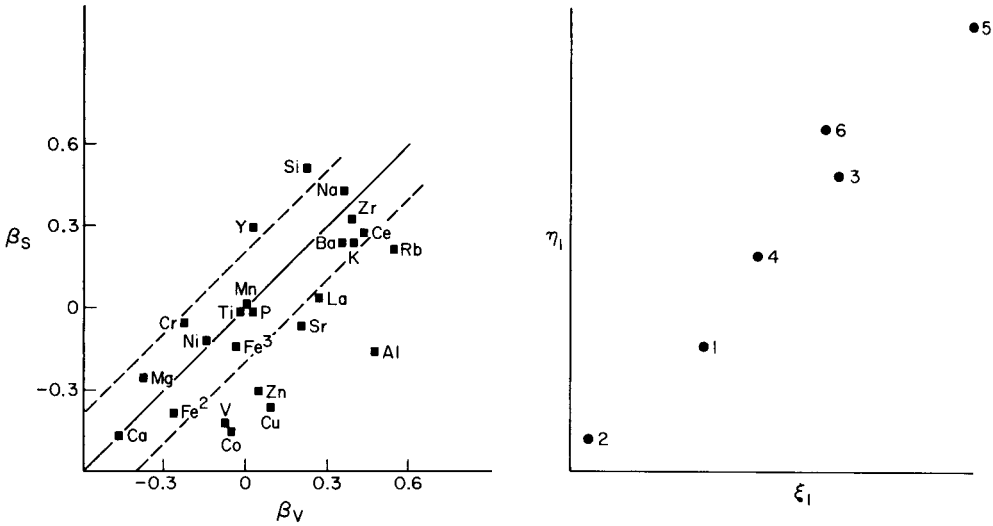


Fig. 8. Plot of the β_1 values reported in Table 9 for the islands of Vulcano and Salina. Elements falling outside the confidence region, indicated by the dashed lines, indicate a different dependence upon the major element composition in the two islands.

Fig. 9. Plot of the ξ_1 against η_1 values obtained by the PLS2 method for the class of Vulcano. The "transfer" coefficient ρ_1 is 1.08.

classes, where the second and third components are of superior value, as was shown above. Analogously, the higher components can be subjected to detailed geological interpretation of similar nature. It is very interesting in this respect that these components are orthogonal; they represent geological "effects" (mineralogical, etc.) that operate independently of each other. This constitutes one of the prime reasons for subjecting the present kind of geochemical data to principal component (or factor) analysis; interpretation is distinctly easier in the reduced (A-dimensional) genetic space, spanned by these A components. Furthermore, systematic information related to the inter-block transfer (major to trace element variables in the present case) of a highly synoptic nature can be obtained by careful inspection. These relationships are probably best displayed graphically, as in Figs. 8 and 9. This kind of interpretation is obviously not ideal with the present limited data base, but the possibilities are obvious.

Figure 9 displays the relation between the same objects (geochemical samples) when expressed as functions of their major element composition or their trace element composition. The slope (ρ_a) indicates the extent to which the latter can be derived by a one-dimensional regression on the former. Figure 8 complements this view by showing which variables are invariant with respect to interclass delineation, i.e., Ca, Mg, Fe^{2+}, Ni, Fe^{3+}, Ba, Zr, Na. These elements do not allow discrimination between classes, and are the redundant elements alluded to above.

Chemometrics can certainly be applied usefully in geochemistry. The present methods appear very promising in detecting new geochemical patterns, and therefore in assisting interpretation of geological processes of magmatic rock genesis. For big analytical programs, the present level-four pattern recognition technique would be advantageous in a pilot study of which variables provide the optimal information content relative to costs, etc.

All data information as well as specific information from the analyses reported herein are available from the authors on request.

Thanks are due to the Italian Ministry of Education (M.P.I.) for a research grant to S.C.

REFERENCES

1 J. C. Davis, Statistics and Data Analysis in Geology, Wiley, New York, 1973.
2 K. G. Jöreskog, J. E. Klovan and R. A. Raylent, Geological Factor Analysis, Elsevier, Amsterdam, 1976.
3 M. Rosi, J. Keller, H. Pichler and L. Villari, Rend. Soc. Ital. Mineral Petrol., 36 (1980) 345—489.
4 S. Wold and M. Sjöström, in B. R. Kowalski (Ed.), Chemometrics: Theory and Application, ACS Symp. Ser. 52, Washington, DC, 1977, p. 243.
5 C. Albano, G. Blomqvist, D. Coomans, W. J. Dunn, U. Edlund, B. Eliasson, S. Hellberg, E. Johansson, B. Norden, M. Sjöström, B. Söderström, H. Wold and S. Wold, in A. Höskulden et al. (Eds.), Symposium i anvent statistik, NEUCC, RECAV, RECKU, Copenhagen, 1981, p. 183.
6 A. R. Ritsema, Comm. Obs. r. Belgiques, Sér. géophis., 101 (1971) 22.
7 F. Barberi, F. Innocenti, G. Ferrara, J. Keller and L. Villari, Earth Planet. Sci. Lett., 21 (1974) 269.
8 H. Pichler, Italienische Vulkan — Gebiete III (Lipari, Vulcano, Stromboli, Tyrrenisches Meer), Borntraeger, Berlin, 1981.
9 S. Wold, Technometrics, 20 (1978) 397.
10 M. L. Bisani, S. Clementi and D. Faraone, Chim. Ind. (Milan), 64 (1982) 116.
11 N. Draper and M. Smith, Applied Regression Analysis, Wiley-Interscience, New York, 1978.
12 H. Wold, in K. G. Jöreskog and H. Wold (Eds.), Systems under Indirect Observation, Casualty-Structure Prediction, North-Holland, Amsterdam, 1981.
13 S. Wold, H. Wold, W. J. Dunn and A. Ruhe, The collinearity problem in linear and nonlinear regression. The partial least squares (PLS) approach to generalized inverses. Technical Report UMINF-83.80, Institute of Information Processing, and Chemometric Research Group Report, 1980.
14 Basaltic Volcanism. Study Project in Basaltic Volcanism on the Terrestrial Planets, Pergamon, Oxford, 1982.
15 C. Albano, W. J. Dunn, U. Edlund, E. Johansson, B. Norden, M. Sjöström and S. Wold, Anal. Chim. Acta, 103 (1978) 429.

Analytica Chimica Acta, 150 (1983) 145—152
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

SIMCA MULTIVARIATE DATA ANALYSIS OF BLUE MUSSEL COMPONENTS IN ENVIRONMENTAL POLLUTION STUDIES

OLAV M. KVALHEIM, KJELL ØYGARD[a] and OTTO GRAHL-NIELSEN*
Department of Chemistry, University of Bergen, N-5014 Bergen (Norway)
(Received 27th January 1983)

SUMMARY

Blue mussels (*Mytilus edulis*) from one pristine and one polluted location on the Norwegian coast were transferred to an aquarium. After 4 months under controlled unpolluted conditions, samples of muscle tissue and gonad tissue from ten specimens of each of the two classes of mussels were characterized by capillary gas chromatography (g.c.) after methanolysis and silylation. The g.c. patterns of the 50—60 predominant peaks representing naturally occurring components were treated by SIMCA multivariate data analysis implemented to run on a HP-85 desk-top computer. This analysis discriminated clearly between two classes of mussels for both the muscle and gonad tissue. Similarly, the g.c. patterns of the gonad tissue differentiated between male and female mussels. Multivariate data treatment of naturally occurring components might thus be an alternative to the Mussel Watch survey which is based on measurements of foreign components in the mussel tissues.

Bivalves such as oysters, clams and mussels accumulate pollutants in their tissues to concentrations that far exceed the concentrations in the ambient water. They have been used in investigations of the distribution of pollutants in coastal areas. In fact, an international Mussel Watch survey has been undertaken [1]. Here four categories of pollutants were considered: petroleum hydrocarbons, halogenated hydrocarbons, trace metals and radionuclides. Multivariate data treatment has been used on data obtained by measurement of petroleum hydrocarbons in mussel tissue to relate different polluted areas to potential sources of pollution [2].

Geographical and seasonal variations in both mussels and environments, however, makes it difficult to use the levels of pollutants found in the mussel tissue as a measure of the degree of pollution of the environment. Moreover, the identification and determination of pollutant components in mussels is usually laborious; it includes work-up of large amounts of tissue, and in many cases tissues from several animals have to be pooled for each analysis. Naturally occurring components often affect the measurements and are difficult to separate from the pollutant components before the latter

---

[a]Present address: Statoil-Lab, P.O. Box 300, N-4001 Stavanger, Norway.

0003-2670/83/$03.00

can be determined. Contamination during sampling and work-up may also pose a problem.

The purpose of the present investigation was to establish if the composition of the components naturally present in the tissue differs between mussels from pristine and polluted locations. Because of the abundance of these components, only a small amount of tissue, of the order of a few milligrams, was required for a measurement; work-up in a single vial was followed by gas chromatography. The resulting gas chromatographic patterns were subjected to multivariate data processing to reveal if the patterns reflected the difference between the environments in which the mussel had grown. The patterns were treated by the SIMCA method [3—6], implemented to run on a Hewlett-Packard desk-top computer.

EXPERIMENTAL

Samples

Unpolluted blue mussels were collected from a mussel farm at Austevoll, on the coast of Norway. Polluted mussels were collected from a dock in Bergen harbour. Austevoll is some 50 km (sea distance) south of Bergen in a relatively unfrequented fjord. To level out short-term differences between the mussels caused by their different environments (e.g., access to food, temperature, salinity, light), the mussels were transferred to an aquarium with running, uncontaminated sea water. They were kept under these conditions for 16 weeks before examination. Immediately upon retrieval of the mussels from the aquarium, their sex was determined by microscopy of the gonad tissue. Samples of muscle and gonad tissue, approximately 20 mg of each, were carefully dissected from nine female mussels from Austevoll and from ten female mussels from Bergen harbour. In addition, samples of about 20 mg of gonad tissue were dissected from six male mussels from Bergen harbour.

Each tissue sample was transferred to a 10-ml thick-walled glass tube with teflon-lined screw cap. The samples were subjected to methanolysis with 2 ml of anhydrous 2 M hydrogen chloride in methanol in the capped vials at 80°C for 18 h. The remaining pieces of tissue were removed, the HCl/methanol was evaporated under a stream of nitrogen, and the residue was trimethylsilylated for 20 min at 80°C with 50 μl of bis(trimethylsilyl)trifluoroacetamide (BSTFA). After dilution with 200 μl of n-hexane, 1 μl of the solution was injected into a Carlo Erba 4160 gas chromatograph with on-column injection; a glass capillary column (25 m long, 0.32 mm i.d.) was used with SE-52 as stationary phase and helium as carrier gas at 4 ml min^{-1}. Examples of the resulting chromatograms are shown in Fig. 1.

The reproducibility of the chromatography was excellent. In the chromatograms of muscle tissue 60 peaks were selected; beyond doubt, these represented the same components in all samples. Accordingly, for the gonad tissue samples 56 peaks were selected. The selected peaks are indicated in Fig. 1. The height of the peaks was measured in millimeters.

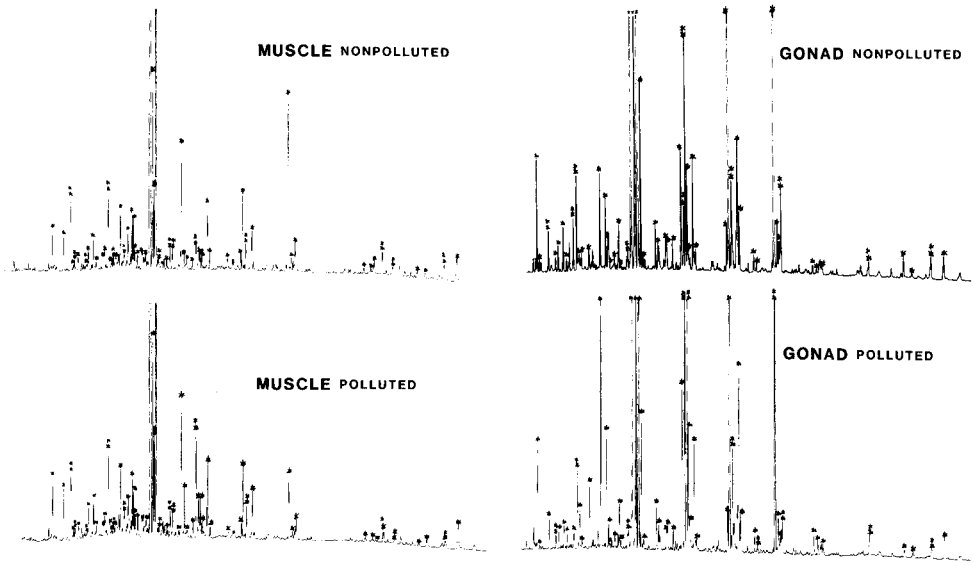


Fig. 1. Typical gas chromatograms of samples of muscle and gonad tissue. The samples were injected at 50°C, the temperature was then raised to 80°C in 1 min, kept at 80°C for 2 min, and thereafter programmed at 4°C min^{-1} to 270°C. The peaks marked with asterisks were used for the multivariate evaluation. Peaks with two asterisks were those remaining after the data sets had been polished.

The data were processed as follows. For each sample, the height of the peaks was normalized to the average peak height, to level out any differences between the samples caused by varying amounts of tissue samples and varying amounts injected on the gas chromatograph. The normalized data were used in all further computations.

Pattern recognition techniques were then applied independently on the resulting two data matrices of size 19 × 60 for the muscle tissue samples from unpolluted and polluted mussels, and of size 25 × 56 for the gonad tissue samples from unpolluted female and polluted female and male mussels. First, the data in each set were treated jointly. The data were scaled within each set by dividing them by the standard deviation over the whole set for each peak, i.e., autoscaling [7]. The largest principal components were then evaluated for each of the two sets to obtain the eigenvector projection of the data. For this purpose Wold's version of SIMCA-3B, written in Basic for the ABC-80 microcomputer [6], was implemented on a Hewlett-Packard HP-85 computer with 32 kbytes memory and a floppy-disc unit. In this version of the program, files are created dynamically with random access and with fixed record length equal to 8 bytes which is the storage needed for a numeral. This procedure reduces the storage requirement and I/O time to a minimum. However, the program is still very CPU-dependent for large data sets. The present version also gives the modelling power of a variable, the

power of a variable to discriminate between different classes of objects and the total distance between classes as defined by Albano et al. [5].

The two classes within each set were subsequently treated separately. The data were first autoscaled within each class. The significant principal components for each class were evaluated by cross-validation [4], hereby forming a class model. The distances between the two classes in each set were calculated. The residual standard deviation of each sample from the class model was found. These numbers, together with the typical standard deviation of the class, were used in an approximate *F*-test to establish if the samples were inside or outside their assigned class and to determine if they were outside the other class.

The contribution of each peak to the class models (i.e., modelling power) and its ability to discriminate between the classes (i.e., discrimination power) were calculated. The data sets were then polished by deletion of the peaks which had low modelling power, less than 0.3, for both classes in the set, and also low discrimination power, less than 3. The computations of class models were then repeated to give class distances with the reduced data sets.

RESULTS AND DISCUSSION

The eigenvector plot of the muscle tissue data is shown in Fig. 2A. The samples from unpolluted mussels are closely grouped and well separated from the polluted mussel samples. The latter samples are more scattered than the unpolluted ones. This result suggests that the individual mussels had reacted differently to pollution and that the polluted environment had been less homogeneous than the pristine environment.

When the mussels from each location were treated as separate classes, using cross-validation, both were best described by principal-component models with two components (i.e., a box). The distance between the principal-component models of the two classes was 4.9, which indicates that they are well separated in the 60-dimensional space [5]. This result substantiates the assumption based on the eigenvector plot that there is a significant difference between the patterns of organic components in the two classes of mussels.

The residual standard deviations, RSD, for all samples when fitted to both principal-component models are shown in Table 1. It is seen that the objects are well classified in their own class. The larger deviations obtained for the polluted samples when fitted to the unpolluted class model than vice versa, substantiates the information from the eigenvector plot (Fig. 2A): the unpolluted samples form a much "tighter" model than the polluted ones. For the unpolluted class, the approximate *F*-test gives a maximum allowable RSD of 1.12 when a confidence interval of $p = 0.01$ is chosen and a RSD_{max} of 1.05 for the smaller confidence interval of $p = 0.05$. When these values are compared with the sample RSDs in Table 1 it is seen that all objects in the unpolluted class fall within the smallest confidence interval. The pol-

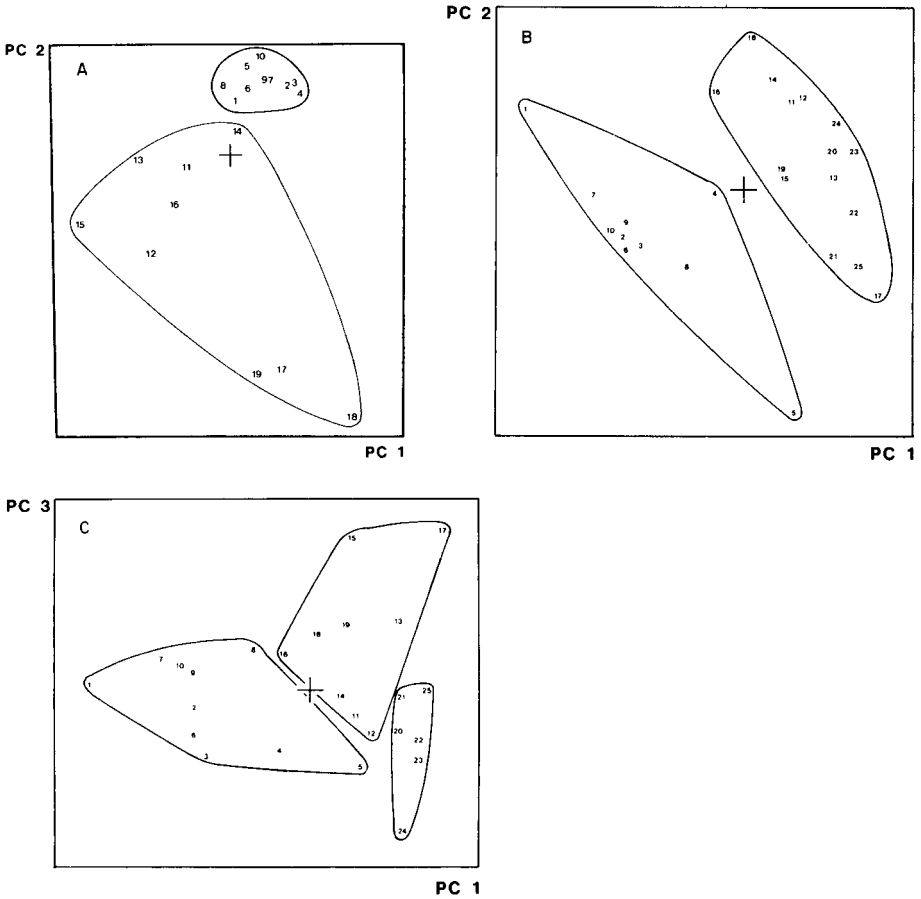


Fig. 2. Eigenvector plot: A, the two first principal components for the samples of muscle tissue, 1—10 are unpolluted and 11—19 are polluted; B, the two first principal components for the sample of gonad tissue, 1—10 are unpolluted females, 11—19 are polluted females and 20—25 are polluted males; C, the first and third principal components for the samples of gonad tissue, with sample numbers as in B.

luted class, which is more scattered according to the eigenvector plot, has RSD_{max} = 0.95 for p = 0.01 and RSD_{max} = 0.88 for p = 0.05. Here, all samples fall inside the largest confidence interval, while one sample, 19, falls outside the smallest interval.

It is thus clear that the samples of polluted and unpolluted mussels form well-described and nicely-separated models in the 60-dimensional space, as established by the 60 selected peaks in the gas chromatogram. However, it is probable that many of the peaks contribute little to the formation and separation of the class models. This suggestion was checked by calculation of modelling and discrimination power of the variables [5]. By setting the very stringent criterion on the variables that their modelling power should

TABLE 1

Distances, expressed as residual standard deviation (RSD), from the 19 samples of muscle tissue to the two class models

Class model for unpolluted samples (1—10)				Class model for polluted samples (11—19)			
Sample	RSD	Sample	RSD	Sample	RSD	Sample	RSD
1	0.68	11	3.09	11	0.51	1	1.27
2	0.89	12	4.22	12	0.86	2	1.55
3	0.84	13	2.62	13	0.49	3	1.39
4	0.78	14	2.14	14	0.85	4	1.44
5	0.97	15	3.97	15	0.82	5	1.37
6	1.02	16	3.21	16	0.71	6	1.56
7	0.77	17	6.12	17	0.71	7	1.16
8	0.91	18	9.90	18	0.65	8	1.31
9	0.71	19	8.17	19	0.94	9	1.19
10	1.05					10	1.44

be larger than 0.3 for both models and that their discrimination power should be larger than 3, only 14 peaks remained. These peaks are marked with two asterisks in the chromatograms in Fig. 1. The models formed of the two classes in the resulting 14-dimensional space are even better separated, i.e., with a distance of 9.3.

The composition of gonad tissue is obviously different from that of muscle tissue, as seen from the chromatograms in Fig. 1. Fifty-six peaks were selected for multivariate evaluation. Figure 2B shows the eigenvector plot of the first versus the second principal component resulting from 10 samples of unpolluted mussels, all female, and 15 samples of polluted mussels, 9 female and 6 male. Even though the samples of unpolluted mussels are more scattered than in the case of muscle tissue, there is a clearcut distinction between the classes. In the class of polluted mussels, the female and male mussels are not separated in this plot. However, a plot of the first principal component against the third (Fig. 2C) shows a separation. When calculated separately, both the class of unpolluted and that of polluted female mussels were characterized by two principal components (i.e., box models). The distance between the models is 3.2, indicating that the models are moderately well separated [5].

Table 2 shows the residual standard deviation of the samples when fitted to both class models, as well as the maximum allowable RSD found from the approximate F-test with two levels of significance, $p = 0.01$ and $p = 0.05$. The values show that there is no overlap between the classes, but the fit within the classes is not too good. Even for the lowest level of significance, two of the unpolluted samples (1 and 4) are on the outer limit of their class model. It is also evident from Table 2 that samples 4 and 8 in the unpolluted class are those lying closest to the polluted class model, as is visualized in the

TABLE 2

Distances, expressed as residual standard deviation (RSD), from the 19 samples of gonad tissue to the two class models, and maximum allowable RSD for two levels of significance

Class model for unpolluted samples (1—10)[a]				Class model for polluted samples (11—19)[b]			
Sample	RSD	Sample	RSD	Sample	RSD	Sample	RSD
1	1.03	11	2.31	11	0.69	1	2.51
2	0.50	12	2.43	12	0.62	2	2.18
3	0.79	13	2.11	13	0.97	3	2.14
4	1.02	14	2.98	14	0.61	4	1.40
5	0.57	15	2.37	15	0.86	5	2.86
6	0.90	16	2.71	16	0.73	6	2.32
7	0.65	17	2.89	17	0.76	7	2.01
8	0.92	18	3.93	18	0.66	8	1.72
9	0.67	19	2.31	19	0.89	9	2.03
10	0.55					10	2.04

[a] RSD_{max} is 1.02 for $p = 0.01$ and 0.94 for $p = 0.05$.
[b] RSD_{max} is 0.98 for $p = 0.01$ and 0.91 for $p = 0.05$.

eigenvector plots (Fig. 2B, C). Only eleven peaks satisfy the strict requirements of modelling power larger than 0.3 in both classes and discrimination power larger than 3. These peaks are marked with two asterisks on the chromatograms in Fig. 1B. Class models based on these peaks have a distance between the classes of only 8.4, a significantly better separation of the models than in the case of all 56 peaks.

TABLE 3

Distances, expressed as residual standard deviation (RSD), from the 15 samples of female and male gonad tissue to the class models, and maximum allowable RSD for two levels of significance

Class model for female gonad samples (11—19)[a]				Class model for male gonad samples (20—25)[b]			
Sample	RSD	Sample	RSD	Sample	RSD	Sample	RSD
11	0.69	20	1.18	20	0.61	11	2.35
12	0.62	21	2.04	21	0.69	12	1.65
13	0.97	22	1.15	22	0.51	13	2.12
14	0.61	23	0.83	23	0.86	14	2.22
15	0.86	24	1.08	24	0.89	15	3.93
16	0.73	25	2.11	25	0.71	16	3.17
17	0.76					17	2.46
18	0.66					18	3.92
19	0.89					19	2.67

[a] RSD_{max} is 0.98 for $p = 0.01$ and 0.91 for $p = 0.05$.
[b] RSD_{max} is 0.95 for $p = 0.01$ and 0.87 for $p = 0.05$.

In order to see if this multivariate processing method really discriminates between male and female mussels, as suggested in Fig. 2C, the two classes were computed against each other. The class distance was 3.0 which barely qualifies as a class separation. The RSD values given in Table 3 also show a moderate class separation. One of the male samples (23) might belong to the female class and three of the other male samples are only slightly outside the female class model. The class model for the male samples is smaller and none of the female samples lies inside it.

Conclusions

The simple technique of single vial work-up followed by gas chromatography makes it possible to obtain patterns of organic compounds from selected tissues of blue mussels. The patterns resulting from muscle tissue are distinctly different from those resulting from gonad tissue. For each type of tissue, the patterns are reproducible for mussels collected at the same location. No attempt was made to identify the chemical components responsible for the peaks in the chromatograms. It is however, reasonable that fatty acids are the most abundant compounds.

Pattern recognition by the SIMCA method shows clearcut differences between mussels from two locations. These differences are not of short-term character, for the mussels from both locations were transplanted to the same aquarium where they lived for 16 weeks before the analysis. Muscle tissue gives better distinction than gonad tissue. The dominating difference between the two locations is that one is polluted by urban outfall and discharges from ships while the other one is pristine.

Although differences in age and genetic origin were not considered, it is reasonable to believe that the difference in the conditions at the two locations is reflected by the difference in composition of the natural components from the mussels. Further studies of similarities and differences among the components naturally present in tissues of blue mussels from different unpolluted locations are desirable to make the present method useful for monitoring of pollution in coastal areas.

REFERENCES

1 The International Mussel Watch, Report of Workshop held in Barcelona, Spain, 1978, National Academy of Sciences, Washington DC.
2 P. W. Kwan and R. C. Clark, Jr., Anal. Chim. Acta, 133 (1981) 151.
3 S. Wold and M. Sjöström in B. R. Kowalski (Ed.), Chemometrics: Theory and Application, Am. Chem. Soc. Symp. Ser. 52, 1977.
4 S. Wold, Technometrics, 20 (1978) 397.
5 C. Albano, G. Blomquist, D. Coomans, W. J. Dunn IV, U. Edlund, B. Eliasson, S. Hellberg, E. Johansson, B. Norden, M. Sjøstrøm, B. Söderström, H. Wold and S. Wold, in A. Høskuldsson, K. Conradsen, B. Sloth Jensen and K. Esbensen (Eds.), Proc. Symp. Appl. Stat., Copenhagen, 1981, p. 183.
6 S. Wold, SIMCA-3B, Manual, Umeå, 1981.
7 B. R. Kowalski and C. F. Bender, J. Am. Chem. Soc., 95 (1973) 686.

Analytica Chimica Acta, 150 (1983) 153—161
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

THE APPLICATION OF PRINCIPAL COMPONENT AND FACTOR ANALYSIS PROCEDURES TO DATA FOR ELEMENT CONCENTRATIONS IN AEROSOLS FROM A REMOTE REGION

P. VAN ESPEN* and F. ADAMS

University of Antwerp (U.I.A.), Dept. of Chemistry, Universiteitsplein 1, B-2610 Wilrijk (Belgium)

(Received 11th October 1982)

SUMMARY

The concentrations of fifteen elements in air-particulate matter collected in six particle-size fractions on a Bolivian mountain form the data set. Principal component and factor analyses are applied. Varimax-rotated principal components were found to be the simplest to interpret. The calculated components may relate to different atmospheric processes; their validity was confirmed by using laser-microprobe mass spectrometry of individual particles.

Aerosol particulate material is analyzed in order to study different aspects of air pollution and atmospheric chemistry. The procedures involved tend to generate large data sets which are difficult to interpret. In the past decade, multivariate data-processing methods have become valuable tools in analytical and environmental chemistry. Hopke et al. used factor analysis to identify sources in Boston urban aerosols [1] and to characterize urban dust [2]. Flocchini et al. [3] also employed factor analysis to investigate airborne particulate matter collected in eight western states in the U.S.A. It can safely be assumed that such data sets have a very complicated structure, as many natural and man-made sources contribute to the composition.

In order to obtain a better insight into the application of the principal-component and factor data-processing techniques, the data from particulate matter collected in a remote location in the southern hemisphere were examined in the present work. Previous studies [4, 5] had shown that this site can serve a baseline-monitoring station.

EXPERIMENTAL

Aerosol sampling

Aerosol particulate material was collected near the top of the 5245-m Chacaltaya mountain, 25 km east of La Paz, Bolivia. The sampling location has been described in detail elsewhere [4].

Size-fractionated aerosol samples were obtained by using a 12.5 l min^{-1} Battelle cascade impactor [6]. This sampler separates the particulate matter

0003-2670/83/$03.00

into six fractions with equivalent aerodynamic diameters of >16 μm, 16—8 μm, 8—4 μm, 4—2 μm, 2—1 μm and 1—0.5 μm, respectively.

A total of 34 samples was collected from September 1977 to November 1978. Sampling time ranged from 7 to 14 days.

Determination of elements

All samples were examined by energy-dispersive x-ray fluorescence spectrometry [7]. Selected samples were also examined by proton-induced x-ray emission (p.i.x.e.) [8], as the detection limits are about an order of magnitude lower with this method. The concentrations of the elements obtained by the two methods agreed within 10% for most elements present at concentrations well above the detection limit.

The data set studied comprised only those elements with a concentration larger than two times the detection limit. The detection limits were calculated from the uncertainty in the computer evaluation of the x-ray spectra [9] and from the uncertainty in the blank concentrations [10]. In the cases where the concentrations were below the detection limit of the x-ray fluorescence method, the results from p.i.x.e., when available, were used. On this basis, measurable concentrations of 14 elements were retained from stages 2—6 for the 34 sampling periods. Stage 1 was not included as the number of results available was too limited.

Mathematical methods

One of the aims of multivariate techniques of data processing, such as principal component analysis and factor analysis is to reduce the dimensions of the original data set **X**, which is characterized by n variables (concentrations of elements) measured on m individuals or samples.

In the principal component technique, the original variables **X** are transformed to a new set of variables **F**, termed components [11]: $\mathbf{X} = \mathbf{A}'\mathbf{F}$, where **X** is the $(m \times n)$ matrix of the measurements, **A** is the $(n \times n)$ component loading matrix and **F** is the $(m \times n)$ matrix of m component scores on n components. The components, which are a linear function of the original variables, are calculated in such a way that the first component explains as much as possible of the variance of the original data set. The second component then explains the maximum of the remaining variance and so on. The components are mutually orthogonal (not correlated). It can be shown that the components are the eigenvectors of the determinant equation $|\mathbf{R} - \lambda\mathbf{I}| = 0$, where **R** is the $(n \times n)$ correlation matrix of the original variables, λ is the eigenvalue, and **I** is the identity matrix. The variance explained by each component is equal to its corresponding eigenvalue. Mathematically, there are as many components as there are variables, but as most of the variance of the original data is explained by the first few components, the reduction of dimensions is done by rejecting the components which contribute little to the total variance (small eigenvalue). In order to

interpret the calculated components, the correlation between the original variable and the new components is calculated from $l_{jk} = a_{jk}\ (\lambda_k)^{1/2}$, where l_{jk} is the correlation between variable j and component k.

The factor analysis model [12] is slightly different, and is written as $\mathbf{X} = \mathbf{A'F} + \mathbf{E}$ where $\mathbf{F}$ is the factor-loading matrix. The number of common factors k, is then less than the number of variables; $\mathbf{E}$ is the residual matrix, which expresses the effect of the specific factor affecting the variable together with the measurement error. In contrast to the principal components method, there is no direct solution of this factor analysis model but the loadings can be estimated by different procedures.

In the minimal-residue factor procedure (MINRES) [13], the factor loadings are estimated by minimizing the sum of squares of the off-diagonal members of the residual correlation matrix:

$$f(\mathbf{A}) = \sum_{j=1}^{n-1} \sum_{i=j+1}^{n} \left(r_{ij} - \sum_{m=1}^{k} a_{jm} a_{im} \right)^2$$

In the maximum-likelihood method of factor analysis (MAXLIK) [14], the factor loadings are estimated from the matrix equation

$$\mathbf{A'J} = \mathbf{A'V^{-1}R} - \mathbf{A'}$$

with $\mathbf{J} = \mathbf{A'V^{-1}A}$, $\mathbf{R} = \mathbf{AA'} + \mathbf{V}$; $\mathbf{V}$ is a diagonal matrix with elements $v_i = 1 - \Sigma_{j=1}^{k} a_{ik}^2$, v_i being the variance of the specific factor.

For a given matrix of factor loadings, one can generate a new and mathematically-equivalent loading matrix by the transformation $\mathbf{A^*} = \mathbf{AT}$, where $\mathbf{T}$ is any nonsingular $(k \times k)$ matrix. The aim of this transformation (or rotation) is to obtain a solution which is easier to interpret. In the normal varimax rotation [15], a simpler factor matrix is obtained by rotating the factors until a maximum is found in the criterion

$$\mathbf{V} = \sum_{j=1}^{k} \left[p \sum_{i=1}^{n} (a_{ij}^2/h_i^2)^2 - \left(\sum_{i=1}^{n} a_{ij}^2/h_i^2 \right)^2 \right] p^{-2}$$

where h_i^2 is the communality of variable i. In this rotation, the factors remain orthogonal. In contrast, the promax rotation [16] is an oblique rotation, allowing correlation between the different factors. This method attempts to find the best fit between the oblique factor-structure matrix and a target matrix. The target matrix is based on the orthogonal structure (i.e., varimax-rotated solution).

The various methods were programmed in Fortran and run on a DEC VAX 11/780 computer. The principal component, the minimal residue factor analysis program and the rotation procedures were based on routines described by Mather [12]. The maximum likelihood estimation was based on an iterative scheme given by Weber [17].

RESULTS AND DISCUSSION

Selection of method

From the mathematical definition of the different methods, it was not clear which should be selected for the description of the present data set. Therefore the data set of each stage was processed by all three methods. The factors were also orthogonally and obliquely rotated.

The three methods produced very similar factors. Table 1 gives as an example the loading for the second component or factor for the data from stage 3 as produced by principal component (PC) analysis and minimal-residue or maximum-likelihood factor analysis; columns 4 and 5 show, respectively, the orthogonal-rotated PC solution and the oblique-rotated maximum likelihood solution. The two loading vectors are nearly identical and differ from the unrotated solution only in the higher loading on those elements which were already significant in the original solution. The observations were similar for all factors obtained from the five stages.

In the further investigation, preference was given to the varimax-rotated principal component method because it was simple for computation but yielded components that could easily be interpreted.

Number of factors to be retained

A major problem in principal component or factor analysis concerns the number of factors to be retained. In the literature, various procedures have

TABLE 1

Comparison of the loadings on the second component for the 6-μm particle size fraction obtained by the different methods

Element	Component loadings				
	PC	MINRES	MAXLIK	VARIMAX PC	PROMAX MAXLIK
Al	−0.10	−0.09	0.09	0.19	0.31
Si	−0.08	−0.08	0.08	0.21	0.30
S	0.91	0.84	0.82	0.93	0.81
Cl	−0.28	−0.25	−0.12	−0.01	0.06
K	−0.07	−0.06	0.11	0.21	0.33
Ca	0.23	0.23	0.39	0.49	0.54
Ti	−0.15	−0.15	−0.01	0.14	0.22
Mn	−0.08	−0.07	0.10	0.21	0.31
Fe	−0.15	−0.15	−0.00	0.13	0.23
Cu	0.58	0.52	0.62	0.69	0.73
Zn	0.02	0.03	0.17	0.24	0.38
As	0.56	0.54	0.54	0.75	0.65
Rb	−0.18	−0.18	−0.04	0.11	0.18
Sr	−0.06	−0.05	0.10	0.12	0.31
Pb	−0.08	−0.07	0.00	0.04	0.20

been described based on accepting only those components with eigenvalues greater than one, or observing a sharp decrease in the eigenvalues, or finally relying on statistical tests. All these approaches tend to give different results. Moreover, the statistical tests are satisfactory only for large data sets. Because the principal component method is used here in a descriptive manner, the decision to include further components is made on the basis of their ease of interpretation [15]. As an example, Table 2 gives the varimax-rotated loading matrix of the data from stage 4. Five components are extracted. As will be discussed below, the first four components can be identified, whereas the fifth does not reveal any new information other than explaining some of the variance for zinc. The last column of Table 2 shows the communality for four components. Nearly 90% of the variance in the concentrations of the elements is explained by the first four components.

Interpretation of the principal component method

Once the procedure has been established, the question arises whether the mathematical components can be related to physical components in the aerosol. The data from each of the five stages were evaluated by the principal components method and the components were compared from one stage to another. Other information available about the aerosol was also taken into account. It is known that particulate matter with large diameters, as sampled by stages 1—3, is associated with dispersion processes. The small particles (stages 5 and 6) are partly derived from gas-to-particle conversion processes and originate from pollution or, as in this data set, mainly from natural processes.

TABLE 2

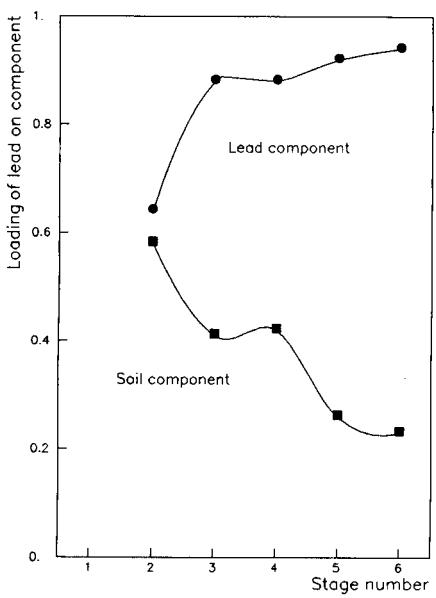
Varimax-rotated principal components obtained from the data from the 3-μm particle size fraction. The last column gives the communality for the first four components

Element	Component loadings					Comm.
	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	
Al	0.93	0.19	0.15	0.16	0.15	0.98
Si	0.93	0.20	0.14	0.21	0.08	0.98
S	−0.13	0.91	0.09	0.18	−0.29	0.90
K	0.93	0.19	0.13	0.19	0.16	0.98
Ca	0.48	0.27	0.05	0.81	0.09	0.91
Ti	0.96	0.04	0.09	0.17	0.13	0.98
Mn	0.90	0.16	0.17	0.30	0.08	0.96
Fe	0.96	0.04	0.19	0.15	0.13	0.99
Cu	0.16	0.94	0.04	−0.01	0.22	0.96
Zn	0.64	0.30	0.27	0.29	0.51	0.76
As	0.36	0.81	−0.23	0.22	0.18	0.93
Rb	0.90	−0.10	0.23	0.02	0.01	0.84
Sr	0.81	0.24	0.09	0.41	0.20	0.87
Pb	0.42	−0.04	0.89	0.05	0.05	0.97

The first component (with the largest eigenvalue) was qualitatively the same for all stages, showing high loadings for the crustal elements Al, Si, K, Ca, Ti, Mn, Fe, Rb and Sr; this points to soil-derived dispersion processes (cf. [18]). The second component was associated with sulphur and appeared in all the stages. But, as could be expected, the variance explained by this component increased towards the data for small particle sizes. Copper and zinc also gave higher loadings in this component as the particle size decreased.

A third component is characterized by a high loading for lead, but lead is also correlated with the first (soil) component for the data from coarse particles. Figure 1 shows the loading of lead on those two components, suggesting the presence of two sources of lead. The average size distribution (concentration of lead as a function of particle size) is represented in Fig. 2. Lead has indeed a bimodal size distribution typical for two distinct sources.

From the data from stages 4 and 6, a component with high loadings for only calcium was observed. This component was not found in stages 2 and 3. Again for the small particle fraction, the variance for calcium is partitioned between the first (soil) component and the specific "calcium" component. Investigation of individual aerosol particles by a laser microprobe (l.m.) mass spectrometric method established the presence of calcium sulphate particles [18]. Table 3 gives the loadings of calcium on the "soil" component and on the



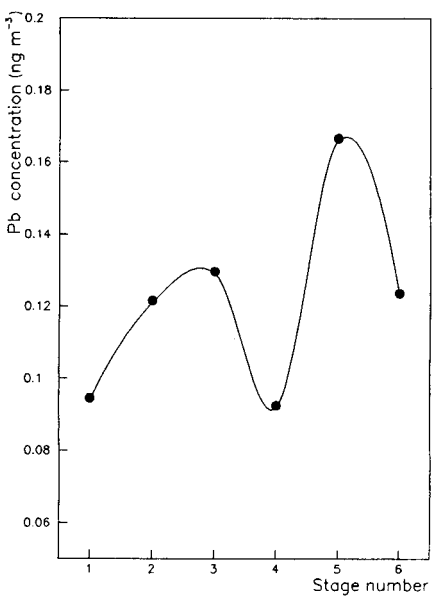


Fig. 1. The loading of lead on the "soil" component and on the "lead" component as a function of the particle size (stage number). The loading can be interpreted as a correlation between lead and the component.

Fig. 2. Average size distribution of lead in the Chacaltaya aerosol, showing a bimodal size distribution.

TABLE 3

The loading of calcium on the "soil" component and on the "calcium" component is compared with the abundance (%) by number of calcium sulphate particles in the different size fractions

Size fraction (μm)	Loading in "soil" component	Loading in "calcium" component	Abundance (%) of $CaSO_4$ in sample 18
12	0.91	—	2
6	0.77	—	23
3	0.48	0.81	35
1.5	0.51	—	9
0.75	0.32	0.88	39

"calcium" component for the different stages; the last column of this table shows the relative abundance by particle counts of calcium sulphate, determined by the l.m. method in a typical sample. In comparing these data, it should be noted that the principal component method applies to the entire data set, whereas the l.m. results apply to only one of the samples. The origin of this mostly fine particulate calcium sulphate is so far unknown.

In the aerosol with the finest particle size, a further component was identified. It showed high loadings for potassium and bromine, while the loadings for sulphur, manganese, iron and zinc were probably also significant. The data are shown in Table 4. These elements are typically associated with biological processes. This component could thus be related to forest fires and agricultural burning activities in the Amazon basin [19].

Table 5 summarizes the results of the analyses. The percentage variance in the original data set explained by the identified components is given for the five size fractions.

Conclusions

It has been shown that physically meaningful components in the aerosol can be identified by using principal component or factor analysis. The different methods produce nearly the same results. The principal components method was selected because the calculations involved are more

TABLE 4

Loadings of the elements on the "burning" component found in the aerosol of small particle size

Variable	Al	Si	S	K	Ca	Ti	Mn
Loading	0.22	0.05	0.40	0.95	0.09	0.18	0.59
Variable	Fe	Cu	Zn	As	Br	Pb	
Loading	0.33	0.01	0.40	0.12	0.79	0.01	

TABLE 5

Percentage variance in the data for the different particle size fractions explained by the identified components

Fraction (μm)	Variance explained by component				
	Soil	Sulphur	Lead	Calcium	Burning
12	64.9	12.9	14.2	—	—
6	58.9	16.5	15.8	—	—
3	57.0	19.8	7.7	8.4	—
1.5	54.6	19.6	12.9	—	—
0.75	31.3	27.3	8.9	7.7	18.5

straightforward than those in the factor procedures. The varimax-rotated loadings are the simplest to interpret.

The use of particle size fractions of the aerosol provides better confidence in the outcome of the data analysis than the use of the total collected aerosol, because the components from different fractions can be compared.

P.V.E. is indebted to the National Science Foundation of Belgium for financial support. The p.i.x.e. results were provided by Dr. W. Maenhaut, Institute of Nuclear Sciences, University of Gent. Assistance with the laser microprobe results by P. Surkyn is gratefully acknowledged. This work was financially supported by the Belgium Government through research project 80-85-10.

REFERENCES

1 P. K. Hopke, E. S. Gladney and A. G. Jones, Atmos. Environ., 10 (1976) 1015.
2 P. K. Hopke, R. E. Lamb and D. F. S. Natusch, Environ. Sci. Technol., 14 (1980) 164.
3 R. G. Flocchini, T. A. Cahill, M. L. Pitchford, R. A. Eldred, J. P. Fleeney and L. L. Ashbaugh, Atmos. Environ., 15 (1981) 2017.
4 F. Adams, R. Dams, L. Guzman and J. W. Winchester, Atmos. Environ., 11 (1977) 629.
5 F. Adams, M. Van Craen, P. Van Espen and P. Andreuzi, Atmos. Environ., 14 (1980) 879.
6 R. J. Mitchell and J. M. Pilcher, Ind. Eng. Chem., 51 (1959) 1039.
7 P. Van Espen, F. Adams and W. Maenhaut, Bull. Soc. Chem., Belg., 90 (1981) 305.
8 W. Maenhaut, A. Selen, P. Van Espen, R. Van Grieken and J. W. Winchester, Nucl. Instrum. Methods, 181 (1981) 399.
9 P. Van Espen, H. Nullens and W. Maenhaut, in P. E. Newbury (Ed.), Microbeam Analysis, San Francisco Press, 1979, 265.
10 L. A. Currie, Anal. Chem., 40 (1968) 586.
11 C. Chatfield and A. J. Collins, Introduction to Multivariate Analysis, Chapman and Hall, London, 1980.
12 P. M. Mather, Computational Methods of Multivariate Analysis in Physical Geography, Wiley, New York, 1976.

13 H. H. Harman and W. H. Jones, Psychometrika, 31 (1966) 351.
14 D. N. Lawley, Proc. R. Soc. Edinburgh, 60 (1940) 64.
15 H. H. Harman, Modern Factor Analysis, Chicago Univ. Press, Chicago, 1967.
16 A. E. Hendrickson and P. O. White, Br. J. Stat. Psychol., 17 (1964) 65.
17 E. Weber, Einfuhrung in die Factorenanalyse, Gustav Fischer Verlag, Jena, 1974.
18 P. Surkyn, J. De Waele and F. Adams, J. Environ. Anal. Chem., 13 (1983) 257.
19 See, e.g., Tellus, 31 (1979) 52.

Analytica Chimica Acta, 150 (1983) 163—170
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

PATTERN RECOGNITION AND CAPILLARY GAS CHROMATOGRAPHY IN THE ANALYSIS OF THE ORGANIC GAS PHASE OF CIGARETTE SMOKE

MILTON E. PARRISH, BENNIE W. GOOD*, MELISSA A. JELTEMA and FRANCIS S. HSU

Philip Morris U.S.A. Research Center, P.O. Box 26583, Richmond, VA 23261 (U.S.A.)

(Received 2nd November 1982)

SUMMARY

Distinguishing between several cigarette types simultaneously by glass capillary gas chromatography requires the use of multivariate statistics for data reduction, pattern extraction and ranking the importance of the chromatographic peaks. These techniques are here applied to the gas phase of the cigarette smoke. Ten cigarettes were studied, both cased and uncased of 100% bright, 100% burley, 100% oriental tobacco and blends of 33%/33%/33% and 60%/30%/10%, respectively. At least five chromatographic profiles were obtained for each different cigarette, giving a total of 51 chromatograms to serve as the data base. From each chromatogram, containing about 100 peaks, a subset of 29 peaks was selected manually. The data were evaluated by using discriminant analysis and factor analysis; the former technique produced a satisfactory separation of cigarette types into distinct groups. Results from the factor technique were successful in providing discrimination based on three factors.

Intense activity in the characterization and differentiation of various cigarette types by means of data generated by high-resolution gas chromatography has resulted in the realization that pattern recognition techniques are necessary to assist in extracting useful information and interpreting the results. The cigarette smoke system is highly complex and therefore the analysis of cigarette mainstream smoke is normally divided into three areas: the organic gas phase, the semi-volatile phase, and the particulate phase. The work presented here deals only with the organic gas phase of cigarette smoke and is defined as that portion of mainstream smoke which passes through a Cambridge filter (which allows only particles of less than 0.3 μM to pass) during standard smoking conditions (35-ml puffs of 2-s duration each minute).

The analysis of gas phase smoke by glass capillary gas chromatography (g.c.) has been reported [1]; the aim was to obtain quantitative profile differences between two cigarette samples. These profiles gave a comprehensive measure of the differences present that previously were not easily observable. Although the use of a computer was instrumental in achieving these results, pattern recognition techniques were not applied. For addressing

0003-2670/83/$03.00

the task of differentiating more than two types of cigarette, statistical evaluation is necessary because of the capacity for data reduction, pattern extraction and ranking the importance of the gas chromatographic peaks. These tools have been applied successfully to derivatized extracts of the particulate phase of cigarette smoke [2]. This paper describes their application to the data for the gas phase.

EXPERIMENTAL

The Hewlett-Packard 5830 microprocessor-controlled gas chromatograph used was equipped with a flame ionization detector and a cryogenic temperature programmer. A pyrex glass capillary column (0.5 mm i.d., 98 m long) was dynamically coated by using a liquid phase consisting of equal portions of UCON 50 HB-280-X and UCON 50-HB-5100; the mercury plug method of Schomburg et al. [3] was used after etching as described by Schieke et al. [4]. The column efficiency was about 2000 theoretical plates per meter, with a capacity factor of 5. No special deactivation was used except that provided by these surfactant-type liquid phases. A sampling valve and a cryogenic trap, both under microprocessor control, were in line with a four-port smoking machine to obtain a composite puff for injection into the gas chromatographic system [5]. The g.c. conditions have been reported [1].

Ten different cigarettes comprised of the three major tobacco types made to similar physical specifications, were studied: both cased (flavor added) and uncased (no flavors added) of 100% bright, 100% burley, 100% oriental and blends of 33%/33%/33% and 60%/30%/10%, respectively, were examined. For this report, bright, burley or oriental cigarettes are cigarettes made from the corresponding filler. At least five chromatographic profiles were obtained for each different cigarette, giving a total of 51 chromatograms to serve as the data base. Each chromatogram contained approximately 100 resolvable peaks. A subset of 29 of the more obvious or intense peaks was selected manually from every chromatogram to give an array of 51 × 29 data points (51 chromatograms each described by 29 variables). A typical chromatographic profile is shown in Fig. 1. The data were processed by using two software programs available on the DEC System 20/60 (Digital Equipment Corp., Maynard, MA): BMDP7M for discriminant analysis and BMDP4M for factor analysis [6].

RESULTS AND DISCUSSION

The discriminant and factor techniques are multivariate statistical methods which examine all the variables in the data base and provide a set of functions or factors which are hopefully fewer in number than the original set and which attempt to preserve the informational integrity of the original set. The discriminant technique selects, weighs, and linearly combines the discriminating variables in such a way that the groups are forced to be as statistically distinct as possible. This technique requires the identification of group membership prior to classification of the data groups; i.e., the data representing burley uncased cigarettes are so identified, the burley cased cigarettes are

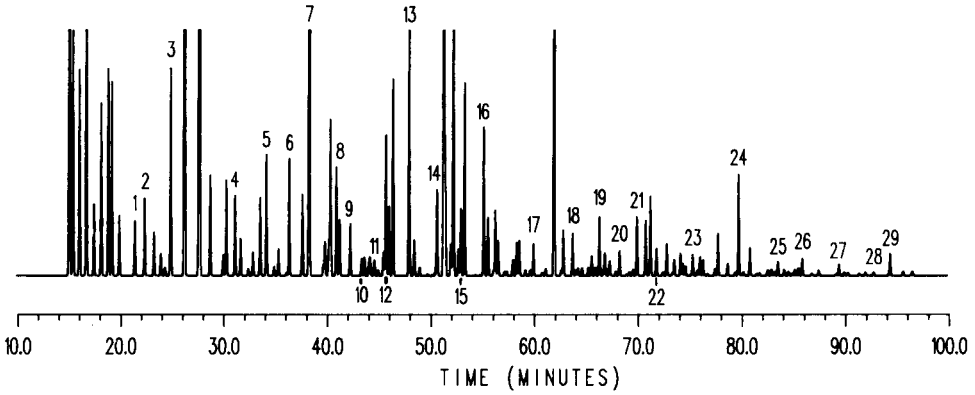


Fig. 1. A facsimile of the gas-phase profile from a burley cigarette. The profile is qualitatively similar to all the data obtained in this study. The 29 peaks selected for multivariate evaluation are numbered.

so identified, and so forth for each group. The goal is two-fold: to obtain maximum inter-group separation and maximum intra-group homogeneity. By using the BMDP7M default values for the F-ratio to enter and remove variables, five peaks were selected for their discriminating ability. The peaks are listed in Table 1. It is informative to note that the F-to-enter values are comparable, indicating that no one variable has an exaggerated power to discriminate. While the percent correctly classified was 98%, and the percent correctly classified when jack-knifed (a more critical test) was 84%, most of the misclassification occurred between cased and uncased samples of the same blend, as seen in Fig. 2. The position of the primary blend constituent forms a triangle with the 33/33/33 and 60/30/10 blends enclosed.

If cased and uncased samples are classified as one group, resulting in only five groups for discrimination, the correct classification is improved to 100% and the correct jack-knife classification to 98%; the only error is misclassification of one of the 60/30/10 samples as a 33/33/33. Eight peaks (Table 2) were entered. Of the five peaks utilized to discriminate all ten sample types, all but peak 21 was re-entered to discriminate the 5 sample types (Fig. 3).

As a final discrimination check, all the cased samples were clustered together and all the uncased samples together. When peaks 26 and 23 were entered, the F values to enter or remove were 44.9 and 9.7, respectively. Although peak 21 was not selected, its F-to-enter was the highest of the remaining data (F = 3.75). As expected, the percentage of correct classifica-

TABLE 1

Summary from the stepwise discriminant technique for separation of all ten tobacco samples

Peak entered	11	18	21	10	8
F-value[a]	39.7	24.9	34.8	14.6	6.1

[a]To enter or remove.

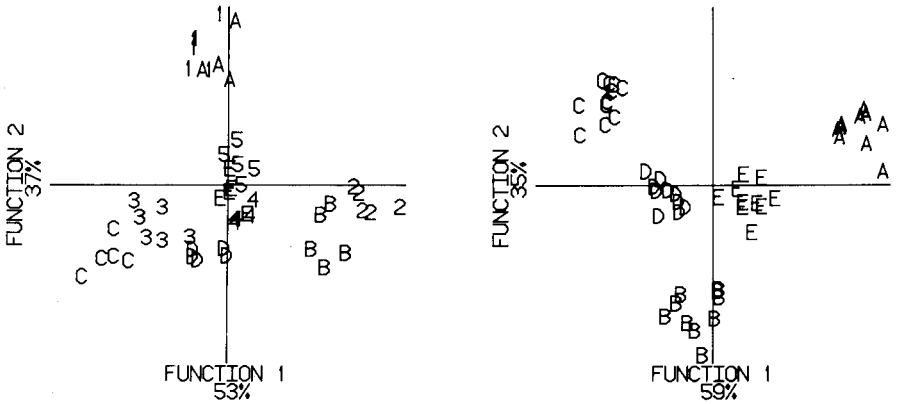


Fig. 2. A scatter plot of samples from the results of discriminant analysis when 10 groups are used. Letters are cased samples. Numbers are uncased samples. A, 1; B, 2; C, 3; D, 4; and E, 5 are bright, burley, oriental, 33/33/33 and 60/30/10 blends, respectively. The percent contribution of each discriminant function to the total dispersion is specified.

Fig. 3. A scatter plot of samples from the results of BMDP7M when no distinction between cased and uncased was provided. Sample labels: A, bright; B, burley; C, oriental; D, 33/33/33; E, 60/30/10.

TABLE 2

Stepwise discriminant technique for separation of the five tobacco blends

Peak entered	11	18	19	10	8	17	3	1
F-value[a]	83.5	43.1	40.7	11.8	6.8	6.5	4.4	6.0

[a]To enter or remove.

tion dropped to 86%, both on the classification and jack-knife. With only two groups, BMDP7M gives a histogram because there is only one discriminant function (Fig. 4). This seems to indicate that while the gas phase profiles contain information about blending, the information about casing is less obvious.

A second, more comprehensive, way of looking at the gas phase data is with factor analysis. Unlike the discriminant technique, which is a supervised learning technique where group membership is specified, the factor technique is an unsupervised learning tool. Its purpose lies not in discrimination but in establishing underlying patterns or trends in the data. The factor technique is capable of reducing the data to a smaller set of numbers that describe the original data.

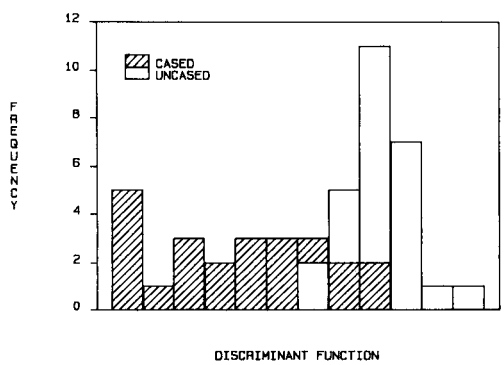
The cumulative proportion of the total variance in Table 3 shows that only three factors account for 74% of the variance in the original data. In Fig. 5, factor 1 separates bright from burley while factor 2 clearly dis-

TABLE 3

The variance explained by each factor

Factor	Variance explained	Cumulative proportion of total variance	Factor	Variance explained	Cumulative proportion of total variance
1	11.956047	0.412277	16	0.087173	0.987531
2	6.188526	0.625675	17	0.070418	0.989960
3	3.295904	0.739327	18	0.054291	0.991832
4	1.622207	0.795265	19	0.050721	0.993581
5	1.303010	0.840196	20	0.046193	0.995174
6	1.147747	0.879774	21	0.034669	0.996369
7	0.817958	0.907979	22	0.025882	0.997262
8	0.553580	0.927068	23	0.020028	0.997952
9	0.408927	0.941169	24	0.018684	0.998596
10	0.353486	0.953358	25	0.015963	0.999147
11	0.253439	0.962097	26	0.010104	0.999495
12	0.215955	0.969544	27	0.009218	0.999813
13	0.186819	0.975986	28	0.003184	0.999923
14	0.150743	0.981184	29	0.002233	1.000000
15	0.096897	0.984525			

tinguishes burley from oriental and bright. In Fig. 6, factor 3 separates cased and uncased cigarettes. It is interesting to look at the factor loadings for the first three factors in Table 4; in columns 5—7, components are asterisked if they had been utilized in the previously presented results from the discriminant technique. In addition to using peaks that were heavily loaded in one factor (e.g., 18, 1, 11, 12, 26), the discriminant technique also utilized peaks that shared information (according to the loadings in Table 4). This confirms the different purposes of the discriminant and factor tech-



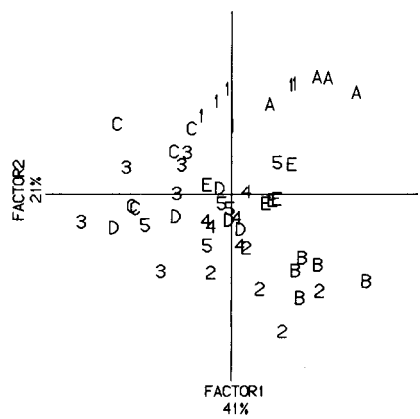


Fig. 4. A histogram of the results of BMDP7M in an attempt to discriminate between cased and uncased samples.

Fig. 5. The factor score plot of the gas phase data using factor 1 (x-axis) vs. factor 2 (y-axis). Sample labelling as for Fig. 2.

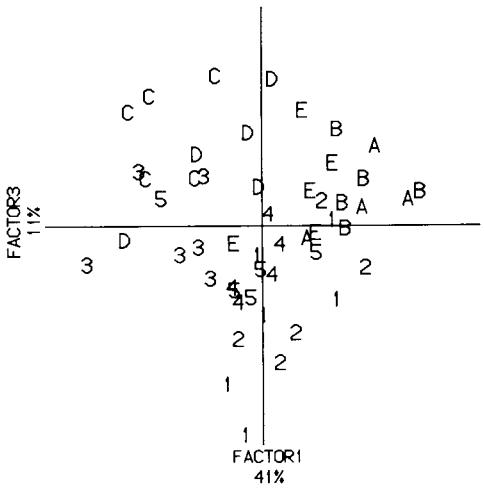
TABLE 4

Sorted rotated factor loading from BMDP4M
(The factor loading matrix has been rearranged so that the columns appear in decreasing order of variance explained by factors. The rows have been rearranged so that for each successive factor, loadings greater than 0.5000 appear first. Loadings less than 0.2500 have been replaced by zero.)

Peak	Factor 1	Factor 2	Factor 3	Used in discriminant analysis		
				10 groups	5 groups	2 groups
18	0.931	0.000	0.000	*	*	
16	0.868	0.000	0.300			
13	0.862	0.000	0.305			
15	0.836	0.395	0.000			
7	0.833	0.387	0.000			
2	0.825	0.398	0.000			
22	0.805	0.315	0.000			
1	0.800	0.000	0.000		*	
24	0.744	0.336	0.386			
6	0.738	0.000	0.000			
4	0.732	0.256	0.000			
3	0.731	0.553	0.000		*	
19	0.716	0.606	0.000		*	
8	0.703	0.421	0.000	*	*	
20	0.617	0.000	0.474			
12	0.000	0.971	0.000			
11	0.000	0.953	0.000	*	*	
10	0.000	0.944	0.000	*	*	
5	0.488	0.806	0.000			
26	0.000	0.000	0.938			*
25	0.000	0.000	0.934			
28	0.000	0.000	0.871			
21	0.573	0.000	0.656	*		
17	0.435	0.494	0.541		*	
9	0.468	0.000	0.000			
29	0.000	0.465	0.000			
27	0.000	0.000	0.000			
23	0.000	0.000	0.000			*
14	0.000	0.413	0.000			

niques. The latter tries to separate underlying patterns as well as possible, while the former is not particularly concerned with orthogonal information. The discriminant technique often chooses cross-terms, which are components that share information across several loadings, such as peaks 3, 19, 21, and 17.

To confirm this, the discriminant technique was applied again to all 10 groups, but this time with components that only had a high loading on a single factor. Peaks 18 (factor one), 10, 11 and 12 (factor two), and 25, 26, and 28 (factor three), were allowed entry. The results shown in Table 5



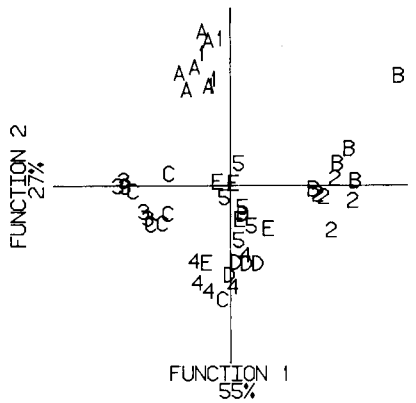


Fig. 6. The factor score plot of the gas phase data using factor 1 (x-axis) vs. factor 3 (y-axis). Sample labelling as for Fig. 2.

Fig. 7. A scatter plot of samples from the discriminant technique, when the peaks permitted for evaluation were restricted to those being principally and solely loaded on one of the first three factors from factor analysis. Sample labelling as for Fig. 2.

TABLE 5

Summary from the discriminant technique when the peaks used for input were restricted to peaks that were highly loaded on a single factor from factor analysis

Peak entered	12	18	10	26
F-value[a]	39.7	24.9	15.1	8.4

[a]To enter or remove.

indicate that if the cross-terms are not available, discrimination still occurs when at least one peak is used from each factor, although with a decrease in the percentage correctly identified (86%) and the percentage correct on jackknife classification (67%). The plot of the data obtained by using the first two discriminant functions is shown in Fig. 7, which is comparable to the results in Fig. 2, where cross-terms were available.

The success of the factor and discriminant techniques in separating the cigarette model system is remarkable, because it relies solely on the quantitative differences in the gas phase chromatographic profiles of the cigarette smoke. As stated previously [1], the raw chromatographic profiles obtained with the flame ionization detector are so complex and similar for all the cigarettes in this model system, that it was thought that discriminating information did not exist in this data set. Pattern recognition techniques clearly demonstrate their power and usefulness and now are considered essential in further attempts to understand the products more clearly.

The authors thank Anne Donathan and Howard Clark for their aid in preparation of this manuscript.

REFERENCES

1 M. E. Parrish, B. W. Good, F. S. Hsu, F. W. Hatch, D. M. Ennis, D. R. Douglas, J. H. Shelton, D. C. Watson and C. N. Reilley, Anal. Chem., 53 (1981) 826.
2 F. S. Hsu, B. W. Good, M. E. Parrish, D. A. Clabo and T. D. Crews, HRC & CC, 5 (1982) 648.
3 G. Schomburg, H. Husmann and F. Weeke, J. Chromatogr., 99 (1974) 63.
4 J. D. Schieke, N. R. Comins and V. Pretorius, J. Chromatogr., 112 (1975) 97.
5 M. E. Parrish, C. T. Higgins, D. R. Douglas and D. C. Watson, HRC & CC, 2 (1979) 551.
6 BMDP Statistical Software, Department of Biomathematics, UCLA, Los Angeles, CA 90024.

Analytica Chimica Acta, 150 (1983) 171—181
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

CHEMOMETRICS AND LIQUID CHROMATOGRAPHY IN THE STUDY OF ACUTE LYMPHOCYTIC LEUKEMIA

HUBERT A. SCOBLE[a], JAMES L. FASCHING and PHYLLIS R. BROWN*
Department of Chemistry, University of Rhode Island, Kingston, RI 02881 (U.S.A.)
(Received 11th November 1982)

SUMMARY

Parametric and nonparametric pattern-recognition techniques were used to investigate the intrinsic structure of chromatographic data obtained from blood plasma of acute lymphocytic leukemic patients. The chromatographic data base (consisting of nucleosides, bases, and other low-molecular-weight ultraviolet-absorbing compounds) was obtained by using reversed-phase high-performance liquid chromatography. When five pattern vectors obtained from twenty-four normal control and twenty-four plasma samples are used, it is possible to discriminate with excellent sensitivity and selectivity between the normal physiological and the pathological processes when both supervised and unsupervised linear learning techniques are applied.

Chromatographic profiling of physiological fluids for endogenous organic compounds has been used in several biomedical studies [1—5]. The goal of such studies has been to discriminate between pathological and physiological processes by using characteristic features derived from the chromatographic data. Several research groups have used gas chromatography and combined gas chromatography/mass spectrometry for the profiling of physiological fluids from individuals with well-established pathologies. Zlatkis et al. [6], Liebich et al. [7, 8] and Rhodes et al. [3] have examined the volatile organic distribution on the urine of patients with diabetes mellitus compared to normal individuals. Jellum [9] has used chromatography and ancillary techniques for the determination of organic acid concentrations in various physiological matrices. These analyses have generated considerable interest because the presence of abnormal metabolites or the increased concentration of others has led to the diagnosis of approximately thirty possible metabolic disorders, mostly involving organic acidurias [9].

Comparatively little research has been conducted on the effects of cancer on metabolic profiles of physiological fluids. Zlatkis et al. [4] examined the volatile organic profiles in serum of patients with leukemia, multiple myeloma, breast cancer, lymphoma and melanoma. However, in the cases studied,

---

[a]Present address: Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.

0003-2670/83/$03.00

the differences in peak area measurements for normal and pathological samples were very similar.

Using reversed-phase high-performance liquid chromatography (r.p.l.c.), Krstulović et al. [10] found increased levels of 1-methylinosine and N^2,N^2-dimethylguanosine in the sera of breast cancer patients. They also reported that these compounds are present in the sera of patients with benign breast tumors, but at significantly lower concentrations. Also using r.p.l.c., Gehrke's group [11] analyzed the urine of patients with breast cancer and leukemia and reported elevated levels of N^4-acetylcytidine in the breast cancer samples. They also reported elevated levels of N^2,N^2-dimethylguanosine, 1-methylguanosine, 1-methylinosine and pseudouridine in the urine of patients with various types of cancer [12]. Davis et al. [13] investigated urine ribonucleoside distribution patterns in patients with advanced colon cancer and found elevated levels of 1-methylinosine, 1-methylguanosine, 2-methylguanosine, adenosine and N^2,N^2-dimethylguanosine when compared to normal urine controls. In these investigations, however, the sensitivity and selectivity of the nucleoside or base marker was inadequate for clinical purposes.

In cases where distinct single or multiple markers can be used to characterize the aberrant metabolic processes, visual examination of chromatographic profile data may be sufficient for diagnostic purposes. However, when the pathological process causes subtle profile changes or when the data structure is not well understood, the use of pattern recognition techniques is beneficial.

The statistical techniques reported in these previous investigations include the use of threshold logic units [5], K-nearest neighbors [14] and Wilcoxon-test correlations [15]. In this paper, parametric and nonparametric discriminant techniques are used for the elucidation of plasma r.p.l.c. profiles from patients with acute lymphocytic leukemia and from normal subjects. The goal of this investigation was to use the inherent data structure present in these profile data to classify the normal and pathological processes, thus emphasis was placed on algorithms for category classification.

EXPERIMENTAL

Sample procurement

Acute lymphocytic leukemia plasma samples were obtained from the Cancer and Acute Leukemia Group B of Rhode Island Hospital. These samples were from male and female patients with well-established pathologies and well-documented clinical histories. Normal plasma samples were randomly selected from a larger population of ostensibly normal, healthy individuals of varying age and sex at the University of Rhode Island Health Services.

Plasma samples from the two populations were collected and processed by documented protocols [16]. Briefly, samples were processed to remove plasma proteins and other materials having molecular weights exceeding 25,000 daltons. Plasma ultrafiltrates were stored at −20°C until required.

Chromatographic system

The liquid chromatographic system used consisted of a M6000A solvent delivery system, Model U6K injector, Model 440 absorbance detector and Model 660 solvent programmer (Waters Associates, Milford, MA 01757). Other on-line detection systems include a Schoeffel FS-970-LC fluorimeter and an SF-770 spectroflow monitor with a MM-700 memory module (Kratos, Schoeffel Instruments, Westwood, NJ 07675). Peak areas were electronically integrated by using a Hewlett-Packard 3380A Integrator. Chromatographic separations were done on a Whatman Partisil 10-ODS-3 column with a column guard dry-packed with Co:Pell-ODS. A linear gradient elution (0—40%) was done with a low-strength eluent of 0.02 M potassium dihydrogenphosphate, pH 5.6 and a high-strength eluent of 60% methanol. A volumetric flow rate of 1.5 ml min^{-1} was applied at ambient temperatures.

A Data General Eclipse S/130 minicomputer operating under AOS with 388 kbytes of high-speed memory was used.

STATISTICAL EVALUATION

For the statistical evaluation, each of the 48 normal and pathological chromatographic profiles were represented as a data vector of the form

$$\mathbf{X} = (x_1, x_2, x_3, \ldots, x_{10})$$

where each component, x_i, represents integrated peak-area measurements from the profile (Fig. 1). These peak-area measurements represent all those components that could accurately be quantified from the chromatographic trace. Many of these ten features represent individual components, but where compounds coeluted or were severely merged, the total area in this time frame was treated as one measurement.

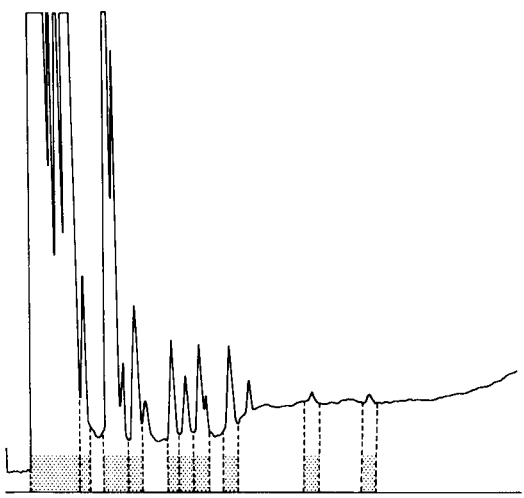


Fig. 1. Retention time intervals used for the generation of the pattern vectors are shown as the shaded portions of this representative chromatogram.

Prior to any classification or training routines, each measurement was autoscaled by means of the formula

$$x'_{ij} = [x_{ij} - u_i] / s_i \tag{1}$$

where x'_{ij} is the autoscaled ith measurement of the jth pattern, x_{ij} is the initial component, u_i the mean and s_i the standard deviation of the ith measurement in all patterns. The autoscaling procedure equalizes the initial weighting of the measurement such that the mean is 0.0 and the variance is 1.0. This removes any inadvertent weighting that might be caused by differences in magnitude of the features. Unless otherwise indicated, all statistical routines were applied to autoscaled data.

Several discriminant techniques involving supervised and unsupervised pattern recognition were utilized. The first routine involves the development of a weight vector which is used for the binary classification of the normal and pathological pattern vectors. The weight function is developed by using a linear combination of features such that the sum of squares of residuals is minimized. The weight vector is calculated from

$$y_k = \sum_{i=1}^{n} (x_{ik} - u_i) b_i + b_0 \tag{2}$$

where y_k is the predicted dependent variable, x_{ik} the autoscaled feature, u_i the mean of feature i (which for autoscaled data is zero), and b_i and b_0 are the weight or regression coefficients of the linear regression equation.

Another linear discriminant function used is based on a linear least-squares discriminant technique. In this technique, one discriminant function is developed for each category studied, by computing the least-squares discriminant through all non-members of a category (assigned a dependent variable of 0.0) to all category members (assigned a dependent variable of 1.0). Data vectors are classified into the category with the function that generates the largest value. Computationally, the discriminant function is calculated by means of Eqn. (2).

The K-nearest neighbor method is a non-parametric discriminant technique in which each pattern or chromatographic profile is represented as a point in an n-dimensional feature space. Patterns are classified as a member of the class most often represented among its K-nearest neighbors. Although various distance measures may be used to calculate interpattern or nearest-neighbor distance in the data space, the Euclidean distance was used here. This distance is defined as

$$d_{ij} = [\sum_{k=1}^{n} (x_{ik} - x_{jk})^2]^{0.5} \tag{3}$$

where x_{ik} and x_{jk} represent points in a data space of n dimensions, and each axis of the orthogonal coordinate system represents a single feature or variable.

Hierarchal Q-mode clustering, an unsupervised cluster technique, was also

applied to the data base. Conceptually, this method is quite simple and is based on the relative similarity of data vectors in the pattern space. A similarity matrix is constructed by means of the equation

$$S_{ij} = 1 - (d_{ij}/d_{max}) \tag{4}$$

where S_{ij} is the similarity index between the ith and jth data vectors and d_{ij}/d_{max} is the interpattern distance of the data vector i and the data vector j normalized by the largest interpattern distance, d_{max}, in the feature space. The data matrix is scanned for the two vectors of maximum similarity. These two data vectors are removed from the data matrix and replaced by the average of the two old vectors. The new matrix is scanned for the next greatest similarity and the procedure is repeated until all data vectors in the pattern space form a single cluster.

Another non-parametric pattern classifier that was used is based on the development of a decision or classification surface such that patterns of the same class reside on the same side of the hypersurface. In this heuristic approach, based on the linear learning machine principle, a discriminant function is calculated by first loading a weight vector with random variables and training the function as it accumulates experience in making decisions. By using negative feedback correction, the algorithm seeks a hyperplane in an $(n + 1)$ space that best separates the two categories. When a pattern of the training set is misclassified, the weight vector is adjusted by reflection of the surface about the point. This moves the surface so that after feedback training, the misclassified point is the same distance on the correct side of the decision surface as it was on the incorrect side. The discriminant function is calculated from the equation

$$\mathbf{D}_k = \sum_{i=1}^{n} (x_{ik})b_i + b_{n+1} \tag{5}$$

where $\mathbf{D}_k$ is the weight vector, x_{ik} are points in the data space, and b_i are the coefficients of the classification equation.

The polyalgorithm ARTHUR [17, 18], a compilation of pattern recognition algorithms containing the previously described routines, was used to study the intrinsic structure of the chromatographic data from the patients.

RESULTS AND DISCUSSION

Figure 2 shows representative generalized chromatographic profile data from a normal individual and an individual with acute lymphocytic leukemia.

Table 1 indicates feature—property correlations. A confidence interval (95% level) about the correlation was obtained by using a Fisher z-transform [19], while the zero probability values indicate the probability that the sample was derived from an uncorrelated parent population [20].

Many of the ten features of the pattern vector may not contribute information significant to the development of classification or decision rules, or

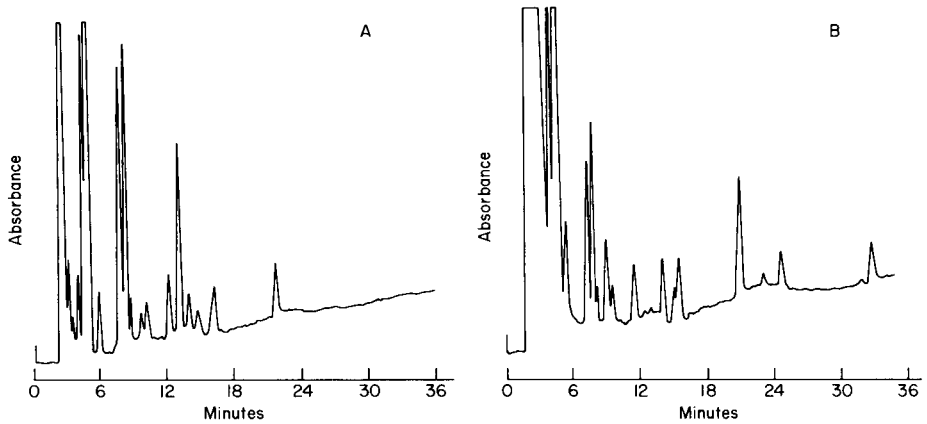


Fig. 2. A, Acute lymphocytic leukemia nucleoside and base plasma profile. B, Normal nucleoside and base plasma profile. Conditions of chromatographic separation listed in text.

may be significant but have less weight than others. Accordingly, the ordered Fisher and variance weights were examined. These preprocessing techniques weight each feature on the basis of its individual importance to the solution of a classification problem. In the first of these two criteria, the variance weighting technique, the ratio of the interclass variance of the two categories to the intraclass variance of the two categories establishes a feature weighting. The Fisher weight calculates a ratio between the square differences in the category pair means and the sums of the intraclass variances.

The weighted features are summarized in Table 2. The variance and Fisher weights indicate that the early eluting compounds, represented as feature PK1, are very important in discriminating between the two classes. This becomes obvious on visual examination of the chromatographic profiles. Other components, however, are also important in a Fisher and variance weight sense, and this is not readily apparent by visual examination.

TABLE 1

Feature property correlations (95% confidence level)

Feature	Low	Correlation	High	Zero probability
PK1	0.802	0.886	0.936	0.000
PK2	−0.529	−0.280	0.014	0.056
PK3	−0.773	−0.620	−0.400	0.000
PK4	0.100	0.382	0.607	0.008
PK5	−0.603	−0.377	−0.094	0.009
PK6	−0.678	−0.481	−0.220	0.001
PK7	−0.552	−0.309	−0.018	0.034
PK8	−0.282	0.011	0.303	0.940
PK9	−0.542	−0.297	−0.005	0.043
PK10	0.125	0.403	0.622	0.005

TABLE 2

Weighted feature measurements

Feature	Variance weight	Fisher weight	Feature	Variance weight	Fisher weight
PK1	8.337	7.031	PK4	1.278	0.2666
PK3	2.250	1.198	PK7	1.238	0.2277
PK6	1.657	0.6298	PK9	1.212	0.2032
PK10	1.402	0.3856	PK2	1.171	0.1643
PK5	1.331	0.3175	PK8	1.000	1×10^{-6}

For the training and classification processes, the error rate has been shown to be a monotonically decreasing function of the ratio (R) between the number of patterns to the number of features [21]. While ratios of $R > 3$ were found to be satisfactory for some applications, $R > 10$ are desirable. It thus becomes important to retain only those features that account for significant variance or intrinsic data structure. Initially, ten dimensional pattern vectors were used in training; however, through the use of principal component factor technique [18], the dimensionality was reduced to 5 features that accounted for approximately 88% of the total variance.

In the reduction of the feature space from a 10-coordinate system to a 5-coordinate system, a significant noise component was effectively filtered out of the data base, while the pattern vector/feature ratio was increased from 4.8 to 9.6. These five weighted features were used in the development of all discriminant functions unless indicated.

In order to ensure adequate training set size, the leave-one-out technique was employed. In this procedure, all vectors, except one, are used in the development of the classification rules or decision surface; the single vector then serves to evaluate the validity of the weight function. This procedure is repeated until each vector has served as a lone member of the test data set.

Equation (2) was used to develop a single weight vector to classify the chromatographic profile data. Initially, a dependent variable of 1.0 was assigned to the normal population, while a dependent variable of 2.0 was assigned to the pathological population. The dependent variables, referred to as disease indices, can then be calculated by using the previously generated weight vector. In this way, each pattern was classified to its respective population. Figure 3 clearly shows the bimodal distribution of the two populations; the dotted line represents the best line for separating the categorized data. At this disease index cut-off of 1.42, the sensitivity and selectivity of the developed weight function are 100%. Not surprisingly, generation of category weight vectors from Eqn. (2) gave identical results; all test patterns were classified correctly.

The 10-feature pattern vectors gave the same results as the 5-feature patterns for each of these parametric classification techniques. In the generation of the weight vectors, the noise or random element component of the data is weighted much less than those features primarily responsible for the discriminatory ability of the weight vectors.

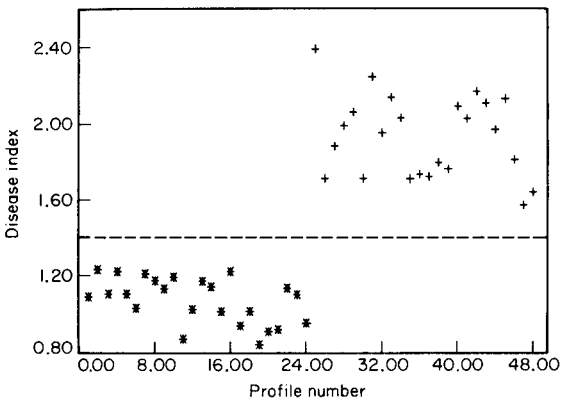


Fig. 3. Graph of disease index versus patient profile number calculated by using a multilinear least-squares fit of the features to the data property. Symbols below the dotted line represent the normal population while those above represent the pathological population.

Two non-parametric classification methods were examined, involving the computation of interpattern distances in the feature space. The underlying premise of such methods is that pattern vectors of similar structure will be resident in a local space defined by the class. Application of the K-nearest neighbor approach to the data base resulted in proper classification of all test patterns. Correct classification was independent of the number of nearest neighbors examined, and worked equally well for 3-NN as for 10-NN. Such overwhelming classification indicates that the local space about each pattern is truly representative of the global class structure.

The same interpattern distance matrix was utilized for hierarchal clustering based on a single link method. The resultant branching diagram shown in Fig. 4, clearly shows the two-class structure of the population. The results of such clustering techniques should be interpreted with caution because of the distortion that may occur in the transition from the feature space to the display space.

The use of a binary linear learning machine approach on the jack-knifed data set resulted in the misclassification of two pattern vectors yielding a final prediction rate of 95.8%. Use of the 10-feature patterns for the development of this decision surface resulted in a 73% prediction rate, which again indicated that a nondescriptive or random component is present in the 10-feature pattern vectors. The relative merits and limitations of this pattern recognition technique have been described previously [22—24].

DISCUSSION

The use of pattern recognition techniques is beneficial for extracting useful information from chromatographic profile data, especially when the intrinsic data structure is not well understood. Because it is difficult, if not

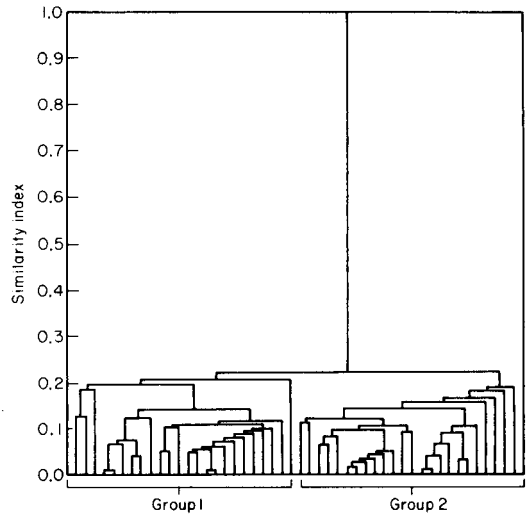


Fig. 4. Single-link clustering of interpattern distances as calculated from Eqn. (4).

impossible, to make a priori selection of classification algorithms, the comparison of results from different methodologies facilitates greater understanding of the data base. In addition, the use of preprocessing techniques such as feature normalization, feature weighting, and dimensionality reduction help to minimize the contribution from features which constitute a nondescriptive element of the profile.

All classification methodologies, as summarized in Table 3, show excellent predictive abilities with the 5-feature patterns. All algorithms generate sensitivity and selectivity values of 100%, with the exception of the linear learning machine which yields a selectivity of 91.6%. Presumably the inclusion of a dead zone about the classification surface would result in no mis-

TABLE 3

Summary of pattern classification results with 5-feature patterns and 4-feature patterns

Technique	Correct classification		Incorrect classification		% Correct	
	5-feature	4-feature	5-feature	4-feature	5-feature	4-feature
Multilinear least-squares fit of features to property	48	38	0	10	100	79.2
Multilinear least-squares discriminant	48	38	0	10	100	79.2
K-nearest neighbor	48	42	0	6	100	87.5
Binary linear learning machine	46	41	2	7	95.8	85.4

classifications because patterns falling in this bounded region would not be classified [25].

An additional advantage of this strategy is that prior identification of components is not mandatory; it is only necessary that the same components fall within a specified retention time window. The problem thus becomes that of identifying only those compounds that contribute significant variance.

While PK1 is very important in the discrimination of the two data sets, PK1 alone does not allow for the complete discrimination of the two populations. Because this retention time interval contains unresolved ionic or polar compounds that are not retained under reversed-phase conditions, the classification results were examined when this feature had been deleted from the training process (cf. Table 3). As expected, PK1 adds significantly to the development of the decision process, and the use of four feature pattern vectors (minus PK1) results in some misclassifications.

The promising results obtained demonstrate the utility of the combination of high-performance liquid chromatography with pattern recognition for the establishment of disease classifications. Research is continuing on the resolution of peak 1 by means of ion-pairing and other modifications in the mobile and stationary phases. Work is also continuing on leukemia and other neoplastic diseases to examine: (1) the biomedical correlations of nucleoside and base profiles with the specific disease, (2) the site and stage specificity of the disease, and (3) the potential of such techniques in studies involving early diagnoses and remission.

The authors thank Patricia Farnes and Barbara E. Barker, Department of Pathology, Rhode Island Hospital, for providing the leukemic plasma samples, the University of Rhode Island Health Services for providing normal plasma samples, and Mona Zakaria for help with the chromatography. We also express our gratitude to Horace F. Martin and Mitchell H. Gail for helpful discussions. This research was supported, in part, by grants from the National Cancer Institute (P.R.B.), National Science Foundation (J.L.F.) and by Gillette Company and American Hoechst Corporation Fellowships (H.A.S.).

REFERENCES

1 D. Issachar, J. F. Holland and C. C. Sweeley, Anal. Chem., 54 (1982) 32.
2 A. Zlatkis, B. S. Brazell and C. F. Poole, Clin. Chem., 27 (1981) 789.
3 G. Rhodes, M. Miller, M. L. McConnell and M. Novotny, Clin. Chem., 27 (1981) 580.
4 A. Zlatkis, C. F. Poole, B. S. Brazell, K. Y. Lee, F. Hsu and S. Singhawangcha, Analyst, 106 (1981) 352.
5 M. L. McConnell, G. Rhodes, U. Watson and M. Novotny, J. Chromatogr., 162 (1979) 495.
6 A. Zlatkis, C. F. Poole, R. S. Brazell, D. A. Bayfus and P. S. Spencer, J. Chromatogr., 182 (1980) 137.
7 H. M. Liebich and G. Hugsgen, J. Chromatogr., 126 (1976) 465.

8 H. M. Liebich and O. Al-Babbili, J. Chromatogr., 112 (1975) 539.
9 E. Jellum, J. Chromatogr., 143 (1977) 427.
10 A. M. Krstulović, R. A. Hartwick and P. R. Brown, Clin. Chim. Acta, 97 (1979) 159.
11 C. W. Gehrke, K. C. Kuo, G. E. Davis, R. D. Suits, T. P. Waalkes and E. J. Borek, J. Chromatogr., 150 (1978) 455.
12 T. P. Waalkes, C. W. Gehrke, R. W. Zumwalt, S. Y. Chang, D. B. Lakings, D. L. Ahmann and C. G. Moertel, Cancer, 36 (1975) 390.
13 G. E. Davis, R. D. Suits, K. C. Kuo, C. W. Gehrke, T. P. Waalkes and E. Borek, Clin. Chem., 23 (1977) 1427.
14 A. Zlatkis, K. Y. Lee, C. F. Poole and G. Holzer, J. Chromatogr., 163 (1979) 125.
15 A. B. Robinson and L. Pauling, Clin. Chem., 20 (1974) 961.
16 R. A. Hartwick and P. R. Brown, CRC Crit. Rev. Anal. Chem., 10 (1981) 297.
17 D. L. Duewer, J. R. Koskinen and B. R. Kowalski, ARTHUR, available from B. R. Kowalski, Laboratory for Chemometrics, Department of Chemistry, University of Washington, Seattle, WA 98195.
18 A. M. Harper, D. L. Duewer, B. R. Kowalski and J. L. Fasching, in B. R. Kowalski (Ed.), Chemometrics: Theory and Applications, ACS Symp. Ser. No. 52, 1977, pp. 14–52.
19 O. L. Davies and P. L. Goldsmith, Statistical Methods for Research and Production, Hafner, New York, 1972, p. 234.
20 P. R. Bevington, Data Reduction and Error Analysis for the Physical Sciences, McGraw-Hill, New York, 1969, p. 122.
21 J. W. Sammon, D. H. Foley and A. Proctor, Proc. IEEE Symp. Adaptive Processes, University of Texas at Austin, 1970, p. IX.2.1.
22 N. J. Nilsson, Learning Machines, McGraw-Hill, New York, 1965.
23 C. P. Weisel and J. L. Fasching, Anal. Chem., 49 (1977) 2114.
24 T. L. Isenhour and P. C. Jurs, Anal. Chem., 43 (1971) 20A.
25 P. C. Jurs and T. L. Isenhour, Chemical Applications of Pattern Recognition, Wiley, New York, 1975, p. 44.

Analytica Chimica Acta, 150 (1983) 183—198
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

TEACHING THE FUNDAMENTALS OF EXPERIMENTAL DESIGN

STANLEY N. DEMING* and STEPHEN L. MORGAN[a]
Department of Chemistry, University of Houston, Houston, TX 77004 (U.S.A.)
(Received 17th September 1982)

SUMMARY

The success of chemometric techniques that operate on data obtained from experiments is often considerably improved if the data have been acquired by a systematic design. The intimate relationships among good chemometric results and good models, good measurement processes, and good experimental designs are shown. A bibliography of the applications of sequential simplex optimization in chemometrics is presented.

Many chemometric applications make use of analytical data that have been acquired under a variety of circumstances. In many cases, the data fall into the category of "observed results"; i.e., the data have been acquired from a system that has not been controlled or perturbed by the experimenter. The classification of archaeological artifacts by pattern recognition techniques is an example of this type of chemometric data [1]. In other cases, it is possible to acquire data that fall into the category of "experimental results"; i.e., the data have been acquired from a system that has been controlled or perturbed by the experimenter. The optimization and understanding of an automated kinetic method for the continuous-flow determination of creatinine provide an example of this type of chemometric data [2].

The success of chemometric techniques that operate on the latter type of data is often considerably improved if the data have been acquired by a systematic design. Thus, it is important that the training of chemometricians include information about the fundamentals of experimental design. We have found that teaching the fundamentals of experimental design is most easily accomplished if the students have first been taught the fundamentals of least-squares fitting of linear models to data [3, 4]. Two important concepts from model fitting then lead directly into experimental design. These concepts are: (a) the sums of squares and degrees of freedom tree used in the analysis of variance, and (b) the variance-covariance matrix.

---

[a]Present address: Department of Chemistry, University of South Carolina, Columbia, SC 29208 (U.S.A.).

0003-2670/83/$03.00

ANALYSIS OF VARIANCE TREE

Figure 1 shows the additive sums of squares and degrees of freedom tree used in the analysis of variance for linear models containing an offset (or β_0) term [3, 4]. Abbreviations are: SS_T, total sum of squares; SS_{mean}, sum of squares due to mean; SS_{corr}, sum of squares corrected for mean; SS_{fact}, sum of squares due to factors as they appear in the model (also called the sum of squares due to regression); SS_r, sum of squares of residuals (also called the sum of squares about regression); SS_{lof}, sum of squares caused by lack of fit; and SS_{pe}, sum of squares due to purely experimental uncertainty (also called the sum of squares due to pure error). The symbols n, p, and f in Fig. 1 represent the total number of experiments in a set, the number of parameters in the model, and the number of distinctly different factor combinations (treatments) in the set of experiments, respectively. The values of n, p, and f combine as shown to give the degrees of freedom associated with each sum of squares.

Significance of regression

The statistical test for the significance of regression is based upon the Fisher F ratio

$$F_{(p-1,\,n-p)} = [SS_{fact}/(p-1)]\,/[SS_r/(n-p)] \qquad (1)$$

For given numbers of degrees of freedom associated with F ($p-1$, and $n-p$), the larger the calculated value of F, the more probable it is that at least one of the parameters in the model (other than β_0) is significantly different from zero. From Eqn. (1), it is evident that the calculated F ratio will be large when $SS_{fact}/(p-1)$ is relatively large and/or when $SS_r/(n-p)$ is relatively small. This provides a key to important aspects of experimental design.

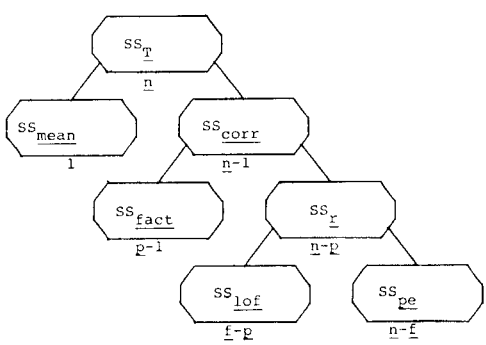


Fig. 1. Sums of squares and degrees of freedom tree for the analysis of variance of linear models. See text for explanation.

Residuals

It can be seen in Fig. 1 that SS_r has two components: SS_{lof} and SS_{pe}. The first of these, SS_{lof}, is caused by a lack of fit of the model to the data. If, for example, a straight-line model is fitted to data that clearly show extensive curvature, then SS_{lof} will be relatively large as a result; if a parabolic model were fitted to the same data, SS_{lof} would be smaller for this model than for the straight line. It should be noted that SS_{lof} has $f - p$ degrees of freedom associated with it.

The other component of SS_r is SS_{pe}, the sum of squares caused by purely experimental uncertainty. The degrees of freedom associated with SS_{pe} is $n - f$ (i.e., the difference between the number of experiments and the number of factor combinations). This, of course, implies replication; i.e., conducting more than one experiment at one or more factor combinations.

Statistical test for lack of fit

The two sums of squares, SS_{lof} and SS_{pe}, can be used to estimate the significance of the lack of fit of the model to the data

$$F_{(f-p,\, n-f)} = [SS_{lof}/(f-p)]\,/\,[SS_{pe}/(n-f)] \tag{2}$$

For given numbers of degrees of freedom associated with F ($f - p$, and $n - f$), the larger the calculated value of F, the more probable it is that any observed lack of fit is real and has not appeared simply by chance. From Eqn. (2) it is evident that the calculated F ratio will be large when $SS_{lof}/(f-p)$ is relatively large and/or when $SS_{pe}/(n-f)$ is relatively small. This also provides a key to important aspects of experimental design.

Fundamentals of experimental design

The following conclusions relate directly to the design of experiments. They are obtained from the structure of the sums of squares and degrees of freedom tree and may be inferred before any experiments are carried out.

(a) If a given model with p parameters is to be fitted to the data, then the number of factor combinations (sets of different experimental conditions) should be greater than p; that is, $f > p$. This will ensure that there are degrees of freedom associated with the sum of squares caused by lack of fit of the model to the data. As a general rule, f should be approximately three larger than p.

(b) If the F-test for lack of fit is to be calculated (see Eqn. 2), then the number of experiments in the set must be greater than the number of factor combinations; that is, $n > f$. This will ensure that there are degrees of freedom associated with the sum of squares arising from purely experimental uncertainty. As a general rule, n should be about three larger than f.

Combination of the recommendations of (a) and (b) above indicates that an adequate experimental design for a given model would require approximately six more experiments than there are parameters in the model.

(c) Precise measurement techniques allow the discovery of small amounts of lack of fit. This is evident from Eqn. (2). It is a useful exercise to ask

students to find examples in the history of science where statistically significant lack of fit has led to important discoveries (e.g., the discovery of the planet Pluto). It is also interesting to observe that (without being taught) some of the students will realize the converse of this principle: i.e., that a bad model can be made to have a lack of fit that is not highly significant simply by being careless about the experimental measurements.

One further point about the precision of the measurement process extends to chemometric methods based on "observed results". The quality of most chemometric methods, those based on "observed results" as well as those based on "experimental results", is influenced by the precision of the measurement process itself. If the analytical data are fuzzy, then the chemometric conclusions will most likely be fuzzy; if the analytical data are crisp, then the chemometric conclusions will also usually be crisp.

VARIANCE–COVARIANCE MATRIX

For a given linear model of the general form

$$y_{1i} = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_{(p-1)} x_{(p-1)i} + r_{1i} \tag{3}$$

where the subscript i refers to a particular experiment and r is a residual, it is possible to write a matrix of parameter coefficients $\mathbf{X}$ of the general form

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} \ldots x_{(p-1)1} \\ 1 & x_{12} \ldots x_{(p-1)2} \\ & \vdots \\ 1 & x_{1n} \ldots x_{(p-1)n} \end{bmatrix} \tag{4}$$

In which each row corresponds to a given experiment, each column corresponds to a given parameter ($\beta_0, \beta_1, \ldots, \beta_{(p-1)}$), and each x corresponds to the coefficient of a given parameter for a given experiment.

The least-squares solution for the matrix of best parameter estimates $\hat{\mathbf{B}}$ is given by $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})$, where $\mathbf{Y}$ is a corresponding matrix of measured responses. The symbol $\mathbf{X}'$ represents the transpose of the $\mathbf{X}$ matrix, and the superscript -1 on $(\mathbf{X}'\mathbf{X})^{-1}$ represents the matrix inverse operation.

In addition to its use in obtaining $\hat{\mathbf{B}}$, the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix is also used to calculate the variance–covariance matrix $\mathbf{V}$:

$$\mathbf{V} = SS_r(\mathbf{X}'\mathbf{X})^{-1}/(n-p) \tag{5}$$

The diagonal elements of this matrix (from upper left to lower right) are the variances associated with the parameter estimates (e.g., $s^2_{b_0}$); the off-diagonal elements are the covariances between parameter estimates.

Confidence intervals for parameter estimates

The variances of the parameter estimates can be used to set confidence intervals that would include the true value of the parameter a certain percentage of the time. The confidence interval for the general parameter β is given by

$$\beta = b \pm (F_{(1,\,n-p)} s_b^2)^{1/2} \tag{6}$$

where β is the (statistically) true value of the parameter, b is its estimated value, s_b^2 is the corresponding variance from the diagonal of the **V** matrix, and $F_{(1,\,n-p)}$ is the tabulated value of F at the desired level of confidence. If $n \approx p + 6$, as suggested before, the F value will be adequately small (and not much improved by increasing n). It is evident that if β is to be estimated precisely, s_b^2 should be small.

Confidence intervals for estimated responses

The variance–covariance matrix can also be used to set confidence intervals that would include the true value of a new experiment a certain percentage of the time. The confidence interval for the estimated response $\hat{y}$ is

$$\hat{y} = \mathbf{X}_0\hat{\mathbf{B}} \pm (F_{(1,\,n-p)}\mathbf{X}_0\mathbf{V}\mathbf{X}_0')^{1/2} \tag{7}$$

where $\mathbf{X}_0$ is a $1 \times p$ matrix of the form shown in Eqn. 4 but corresponding to the particular factor combination of interest. Again, it is evident that if $\hat{y}$ is to be estimated precisely, the elements of **V** should be small.

Fundamentals of experimental design

Given the desirability of minimizing the elements of the variance–covariance matrix (and thereby all of the diagonal elements), the following conclusions relate to the design of experiments and may be inferred before any experiments are carried out.

(d) The quantity $SS_r/(n - p)$ should be small. In previous discussions it was shown that SS_r will be small if both SS_{lof} and SS_{pe} are small. This is achieved in the first place by employing a good model, and in the second place by making precise measurements. It should be noted that increasing the difference $(n - p)$ by increasing n will not usually decrease $SS_r/(n - p)$, the so-called mean square residual. As n increases, SS_r also increases because of the increased number of experiments contributing to it, and the average stays about the same.

(e) The elements of $(\mathbf{X}'\mathbf{X})^{-1}$ should be small. Although not readily evident (but easily shown by a few simple calculations [3]), the elements of the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix can be made smaller by increasing the domain of the factor combinations (i.e., by spreading out the experiments) and/or by increasing the number of experiments, usually an expensive alternative.

APPLICATION TO CENTRAL COMPOSITE DESIGNS

As an example of the application of these fundamentals of experimental design, consider the highly popular central composite design shown in Fig. 2. It is called a central composite design because it is the centered juxtaposition of a two-factor star design (top, bottom, left, right, and center points) and a two-level, two-factor factorial design (upper left, upper right, lower right, and lower left points). It is especially useful because it provides sufficient factor combinations for fitting the full second-order polynomial model.

$$y_{1i} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{11} x_{1i}^2 + \beta_{22} x_{2i}^2 + \beta_{12} x_{1i} x_{2i} + r_{1i} \quad (8)$$

This model can be used to approximate almost any smooth surface over a limited domain.

If the center point is replicated a total of four times, then $n = 12$, $f = 9$, and $p = 6$. Thus, there will be $f - p = 9 - 6 = 3$ degrees of freedom for lack of fit; and $n - f = 12 - 9 = 3$ degrees of freedom for purely experimental uncertainty.

If the experiments are conducted over a broad domain, then the elements of $(X'X)^{-1}$ should be sufficiently small that precise parameter estimates (Eqn. 6) and precise response estimates (Eqn. 7) can be obtained. References [3—8] are recommended for further reading.

Sequential simplex optimization is valuable in connection with experimental design. Applications of sequential simplex optimization in chemometrics, based on computer-assisted literature searches in Chemical Abstracts and Science Citation Index, are listed in Table 1.

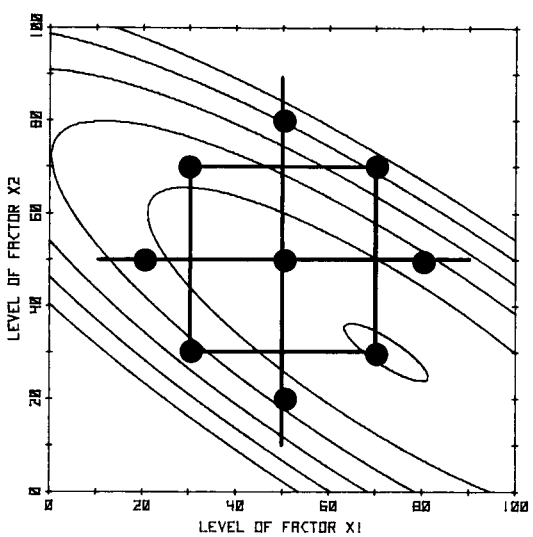


Fig. 2. A two-factor central composite experimental design.

REFERENCES

1 B. R. Kowalski, T. F. Schatzki and F. H. Stross, Anal. Chem., 44 (1972) 2176.
2 A. S. Olansky and S. N. Deming, Clin. Chem., 24 (1978) 2115.
3 S. N. Deming and S. L. Morgan, Clin. Chem., 25 (1979) 840.
4 W. Mendenhall, Introduction to Linear Models and the Design and Analysis of Experiments, Duxbury Press, Belmont, CA, 1968, p. 176.
5 G. E. P. Box, W. G. Hunter and J. S. Hunter, Statistics for Experimenters. An Introduction to Design, Data Analysis and Model Building, Wiley, New York, 1978.
6 M. G. Natrella, Experimental Statistics, Nat. Bur. Stand. Handbook 91, U.S. Govt. Printing Office, Washington, DC, 1963.
7 J. Neter and W. Wasserman, Applied Linear Statistical Models. Regression, Analysis of Variance and Experimental Designs, Irwin, Homewood, IL, 1974.
8 N. R. Draper and H. Smith, Applied Regression Analysis, Wiley, New York, 1966.

TABLE 1

Applications cross-reference for sequential simplex optimization[a]

Category	References
General theory and review articles	1—8, 11, 14, 15, 21—25, 27, 28, 31—34, 36, 39, 42—45, 47—49, 51, 62, 65, 70—72, 75, 76, 93, 113, 119, 120, 130, 135, 138, 154, 160, 161, 169, 174
Wet chemical analysis	11, 16, 29—31, 44, 47, 62, 63, 76, 81, 86, 91, 114, 131, 136, 137, 146, 151, 165, 167, 176, 179
Automated chemical analysis	9, 38, 42—44, 63, 74, 80, 84, 97, 110, 113, 129, 132, 141, 153, 163, 172, 182, 186
Chromatography	64, 69, 79, 85, 112, 125, 127, 133, 148, 149, 152, 162, 173, 186
Spectroscopy, spectrometry (optical and x-ray)	18, 58, 66, 67, 80, 81, 113, 126, 141, 153, 156, 164, 166, 176, 183, 184, 188, 189
Electrochemistry	95, 116, 122
Synthesis	10, 13, 35, 37, 38, 56, 57, 100, 101, 117, 132, 134, 158, 168, 185
Industrial chemistry	2, 3, 5, 7, 8, 13, 15, 17, 19, 26, 35, 37, 45, 101, 106, 132, 154, 155, 161
Pharmaceutical	40, 46, 100, 105, 111, 117, 144, 170, 185
Clinical and biochemical analysis	18, 31, 44, 47, 63, 84, 88, 97, 99, 110, 112, 129, 176
Kinetics and reaction mechanisms	15, 31, 55, 61, 77, 103, 104, 111, 122, 129, 142, 145, 147, 180
Spectral fitting and deconvolution	41, 52, 54, 58, 59, 74, 89, 102, 109, 118, 124, 143, 187
Least-squares and modelling	6, 12, 14, 15, 20, 21, 31, 34, 39, 46, 48, 49, 60, 65, 71, 72, 77, 78, 90, 93, 102, 104, 111, 121, 130, 138, 139, 142, 160, 171, 180, 185, 187
Pattern recognition	68, 73, 83, 115, 123
Nuclear magnetic resonance	9, 54, 59, 73, 74, 157, 175, 182, 187
Theoretical chemistry	53, 89, 94, 96, 98, 128, 140, 145, 157

[a]Prepared with the help of Robert S. Whiton, Department of Chemistry, University of South Carolina.

1 W. Spendley, G. R. Hext and F. R. Himsworth, Sequential application of simplex designs in optimization and evolutionary operation, Technometrics, 4 (1962) 441.
2 C. W. Lowe, Some techniques of evolutionary operation, Trans. Inst. Chem. Eng., 42 (1964) T334.
3 W. D. Baasel, Exploring response surfaces to establish optimum conditions, Chem. Eng. (New York), 72, No. 22 (1965) 147.
4 M. J. Box, A new method of constrained optimization and a comparison with other methods, Computer J., 8 (1965) 42.
5 B. H. Carpenter and H. C. Sweeny, Process improvement with "simplex" self-directing evolutionary operation, Chem. Eng. (New York), 72, No. 14 (1965) 117.
6 J. A. Nelder and R. Mead, A simplex method for function minimization, Computer J., 7 (1965) 308.
7 I. C. Kenworthy, Some examples of simplex evolutionary operation in the paper industry, Appl. Stat., 16 (1967) 211.
8 C. W. Lowe, A report on simplex evolutionary operation for multiple responses, Trans. Inst. Chem. Eng., 45 (1967) T3.
9 R. R. Ernst, Measurement and control of magnetic field homogeneity, Rev. Sci. Instrum., 39 (1968) 988.
10 L. P. Kofman, V. G. Gorskii, B. Z. Brodskii, A. A. Sergo, T. P. Nozdrina, A. I. Osipov and Y. V. Nazarov, Use of optimization methods in developing a process of obtaining prometrine, Zavod. Lab., 34 (1968) 69.
11 D. E. Long, Simplex optimization of the response from chemical systems, Anal. Chim. Acta, 46 (1969) 193.
12 W. Spendley, Nonlinear least squares fitting using a modified simplex minimization method, in R. Fletcher (Ed.), Optimization and its Applications, Symp. Inst. Mathematics, University of Keele, England, 1968, Academic Press, New York, 1969, p. 259.
13 A. Avots, V. Ulaste and G. Enins, Optimization of a heterogeneous catalytic process by the method of simplex planning, Latv. PSR Zinat. Akad. Vestis, Kim. Ser., 6 (1970) 717; Chem. Abstr., 74: 55689q.
14 G. S. G. Beveridge and R. S. Schechter, Optimization: Theory and Practice, McGraw-Hill, New York, 1970, p. 367.
15 D. M. Himmelblau, Process Analysis by Statistical Methods, Wiley, New York, 1970, p. 181.
16 M. J. Houle, D. E. Long and D. Smette, A simplex optimized colorimetric method for formaldehyde, Anal. Lett., 3 (1970) 401.
17 R. Tymczynski, Z. Spych and W. Kupsc, Sequential application of the simplex method for optimization of rubber-blend composition, Polimery, 15 (1970) 530; Chem. Abstr., 74: 127123a.
18 V. G. Belikov, V. E. Godyatskii and A. I. Sichko, Application of simplex planning of experiments to study the optimal conditions for analysis of chloracizin by differential photometry, Farmatsiya (Moscow), 20, No. 3 (1971) 30; Chem. Abstr., 75: 67539z.
19 H. Brusset, D. Depeyre, M. Boeda, R. Melkior and J. L. Staedtsbaeder, Optimization of a multistage reactor for oxidation of sulfur dioxide by three distinct methods (Hooke-Jeeves, Rosenbrock, and simplex), Can. J. Chem. Eng., 49 (1971) 786.
20 R. O'Neill, Function minimization using a simplex procedure, Appl. Stat., 20 (1971) 338.
21 D. M. Himmelblau, Applied Nonlinear Programming, McGraw-Hill, New York, 1972.
22 D. M. Himmelblau, A uniform evaluation of unconstrained optimization techniques, in F. A. Lootsma (Ed.), Numerical Methods for Non-linear Optimization, Academic Press, New York, 1972, Ch. 6.
23 J. M. Parkinson and D. Hutchinson, A consideration of non-gradient algorithms for the unconstrained optimization of functions of high dimensionality, in F. A. Lootsma (Ed.), Numerical Methods for Non-linear Optimization, Academic Press, New York, 1972, Ch. 7.

24 J. M. Parkinson and D. Hutchinson, An investigation into the efficiency of variants on the simplex method, in F. A. Lootsma (Ed.), Numerical Methods for Non-linear Optimization, Academic Press, New York, 1972, Ch. 8.
25 D. A. Phillips, A preliminary investigation of function optimisation by a combination of methods, Computer J., 17 (1972) 75.
26 T. Umeda, A. Shindo and E. Tazaki, Optimal design of chemical process by feasible decomposition method, Ind. Eng. Chem., Proc. Des. Dev., 11 (1972) 1.
27 G. R. Atwood and W. W. Foster, Transformation of bounded variables in simplex optimization techniques, Ind. Eng. Chem., Proc. Des. Dev., 12 (1973) 485.
28 W. E. Biles, An accelerated sequential simplex technique, AIIE Trans., 5 (1973) 127.
29 F. P. Czech, Simplex-optimized J-acid method for the determination of formaldehyde, J. Assoc. Off. Anal. Chem., 56 (1973) 1489.
30 F. P. Czech, Simplex-optimized acetylacetone method for formaldehyde, J. Assoc. Off. Anal. Chem., 56 (1973) 1496.
31 S. N. Deming and S. L. Morgan, Simplex optimization of variables in analytical chemistry, Anal. Chem., 45 (1973) 278A.
32 R. W. Glass and D. F. Bruley, Reflex method for empirical optimization, Ind. Eng. Chem., Proc. Des. Dev., 12 (1973) 6.
33 D. L. Keefer, Simpat: self-bounding direct search method for optimization, Ind. Eng. Chem., Proc. Des. Dev., 12 (1973) 92.
34 J. L. Kuester and J. H. Mize, Optimization Techniques with Fortran, McGraw-Hill, New York, 1973, pp. 298—308.
35 I. I. Pak, I. N. Kim and V. I. Sadovnikova, Use of simplex planning of extreme experiments for optimizing thorough cyanoethylation, Khim. Tekhnol. Tsellyul. Volokna, 6 (1973) 289; Chem. Abstr., 85: 145016z.
36 P. R. Adby and M. A. H. Dempster, Introduction to Optimization Methods, Chapman and Hall, London, 1974, pp. 45—48.
37 A. G. Antonenkov and V. M. Pomerantsev, Use of a simplex method to search for the optimum composition of a catalyst, Zh. Prikl. Khim. (Leningrad), 47 (1974) 899; Chem. Abstr., 81: 30064p.
38 W. D. Basson, P. P. Pille and A. L. DuPreez, Automated in situ preparation of azomethane H and the subsequent determination of boron in aqueous solution, Analyst, 99 (1974) 168.
39 J. M. Chambers and J. R. Ertel, A remark on algorithm AS-47 Function minimization using a simplex procedure, Appl. Stat., 23 (1974) 250.
40 F. Darvas, Application of the sequential simplex method in designing drug analogs, J. Med. Chem., 17 (1974) 799.
41 C. Dauwe, M. Dorikens and L. Dorikens-Vanpraet, Analysis of double decay spectra by the simplex stepping method, Appl. Phys., 5 (1974) 45.
42 S. N. Deming and P. G. King, Computers and experimental optimization, Research/Development, 25, No. 5 (1974) 22.
43 P. G. King and S. N. Deming, Uniplex: single-factor optimization of response in the presence of error, Anal. Chem., 46 (1974) 1476.
44 R. D. Krause and J. A. Lott, Use of simplex methods to optimize analytical conditions in clinical chemistry, Clin. Chem., 20 (1974) 775.
45 J. M. Leaman, A simple way to optimize, Machine Design, 46 (1974) 204.
46 B. W. Madsen and J. S. Robertson, Improved parameter estimates in drug—protein binding studies by nonlinear regression, J. Pharm. Pharmacol., 26 (1974) 807.
47 S. L. Morgan and S. N. Deming, Simplex optimization of analytical chemical methods, Anal. Chem., 46 (1974) 1170.
48 D. M. Olsson, A sequential simplex program for solving minimization problems, J. Qual. Technol., 6 (1974) 53.
49 R. O'Neill, Corrigendum, Function minimization using a simplex procedure, Appl. Stat., 23 (1974) 252.

50 B. N. Shmalei, A. A. Bershitskii, V. K. Rumyantsev and N. N. Khavskii, Use of simplex planning during optimization of the nitric acid decomposition of scheelite, Zavod. Lab., 40 (1974) 581; Chem. Abstr., 81: 173365y.
51 L. A. Yarbro and S. N. Deming, Selection and preprocessing of factors for simplex optimization, Anal. Chim. Acta, 73 (1974) 391.
52 J. W. Akitt, Visual and automatic spectral analysis using a small digital computer, Appl. Spectrosc., 29 (1975) 493.
53 S. Becker, H. J. Kohler and C. Weiss, Geometric optimization of small molecules in the all valence electron—MO formalism using the SIMPLEX and gradient methods, Collect. Czech. Chem. Commun., 40 (1975) 794; Chem. Abstr., 83: 65820g.
54 S. Berger, F. R. Kreissl, D. M. Grant and J. D. Roberts, Determination of anisotropy of molecular motion with ^{13}C spin—lattice relaxation times, J. Am. Chem. Soc., 97 (1975) 1805.
55 J. R. Chipperfield, A. C. Hayter and D. E. Webster, Reactivity of main-group—transition-metal bonds. Part II. The kinetics of reaction between tricarbonyl (η-cyclopentadienyl)(trimethylstannyl)chromium and iodine, J. Chem. Soc., Dalton Trans., 20 (1975) 2048.
56 W. K. Dean, K. J. Heald and S. N. Deming, Simplex optimization of reaction yields, Science, 189 (1975) 805.
57 G. Dumenil, A. Cremieux, R. Phan Tan Luu and J. P. Aune, Bioconversion from DL-homoserine to L-threonine. II. Application of the simplex method of optimization, Eur. J. Appl. Microbiol., 1 (1975) 221.
58 R. H. Geiss and T. C. Huang, Quantitative x-ray energy dispersive analysis with the transmission electron microscope, X-Ray Spectrom., 4 (1975) 196.
59 L. A. Hiscott and P. T. Andrews, Cd 4*d* spin—orbit splittings in alloys of Cd and Mg, J. Phys., F, Metal Phys., 5 (1975) 1077.
60 F. James and M. Roos, MINUIT, a system for function minimization and analysis of the parameter errors and correlations, Comp. Phys. Commun., 10 (1975) 343.
61 G. Just, U. Lindner, W. Pritzkow and M. Roellig, Diels—Alder reactions. V. Kinetic modelling of reactions occurring during the codimerization of cyclopentadiene with 1,3-butadiene, J. Prakt. Chem., 317 (1975) 979; Chem. Abstr., 84: 58166j.
62 P. G. King, S. N. Deming and S. L. Morgan, Difficulties in the application of simplex optimization to analytical chemistry, Anal. Lett., 8 (1975) 369.
63 J. A. Lott and K. Turner, Evaluation of Trinder's glucose oxidase method for measuring glucose in serum and urine, Clin. Chem., 21 (1975) 1754.
64 S. L. Morgan and S. N. Deming, Optimization strategies for the development of gas—liquid chromatographic methods, J. Chromatogr., 112 (1975) 267.
65 D. M. Olsson and L. S. Nelson, The Nelder—Mead simplex procedure for function minimization, Technometrics, 17 (1975) 45.
66 L. R. Parker, S. L. Morgan and S. N. Deming, Simplex optimization of experimental factors in atomic absorption spectrometry, Appl. Spectrosc., 29 (1975) 429.
67 W. E. Rippetoe, E. R. Johnson and T. J. Vickers, Characterization of the plume of a direct current plasma arc for emission spectrometric analysis, Anal. Chem., 47 (1975) 436.
68 G. L. Ritter, S. R. Lowry, C. L. Wilkins and T. L. Isenhour, Simplex pattern recognition, Anal. Chem., 47 (1975) 1951.
69 R. Smits, C. Vanroelen and D. L. Massart, The optimisation of information obtained by multicomponent chromatographic separation using the simplex technique, Fresenius Z. Anal. Chem., 273 (1975) 1.
70 G. R. Walsh, Methods of Optimization, John Wiley, London, 1975, pp. 81—84.
71 M. Auriel, Nonlinear Programming, Prentice-Hall, Englewood Cliffs NJ, 1976, pp. 245—247.
72 P. R. Benyon, Remark: function minimization using a simplex procedure, Appl. Stat., 25 (1976) 97.

73 T. R. Brunner, C. L. Wilkins, T. F. Lam, L. J. Soltzberg and S. L. Kaberline, Simplex pattern recognition applied to carbon-13 nuclear magnetic resonance spectrometry, Anal. Chem., 48 (1976) 1146.
74 D. M. Cantor and J. Jonas, Automated measurement of spin—lattice relaxation times: optimized pulsed nuclear magnetic resonance spectrometry, Anal. Chem., 48 (1976) 1904.
75 S. N. Deming, S. L. Morgan and M. R. Willcott, Sequential simplex optimization, Am. Lab., 8, No. 10 (1976) 13.
76 T. J. Dols and B. H. Armbrecht, Simplex optimization as a step in method development, J. Assoc. Off. Anal. Chem., 59 (1976) 1204.
77 M. J. Faddy, A note on the general time-dependent stochastic compartmental model, Biometrics, 32 (1976) 443.
78 D. J. Finney, Radioligand Assay, Biometrics, 32 (1976) 721.
79 J. Holderith, T. Toth and A. Varadi, Minimizing the time for gas chromatographic analysis. Search for optimal operational parameters by a simplex method, J. Chromatogr., 119 (1976) 215.
80 E. R. Johnson, C. K. Mann and T. J. Vickers, Computer controlled system for study of pulsed hollow cathode lamps, Appl. Spectrosc., 30 (1976) 415.
81 I. Y. Kul and L. E. Kechatova, Use of simplex experiment planning for selecting optimal conditions for the differential spectrophotometric determination of dimecarbine, Farmatsiya (Moscow), 25 (1976) 23; Chem. Abstr., 85: 25446t.
82 G. M. Kuznetsov, M. P. Leonor, S. K. Kuznetsova and V. I. Kovalev, Calculation of the heat of melting of compounds and of simple substances, Zh. Fiz. Khim., 50 (1976) 2517; Chem. Abstr., 86: 22607f.
83 T. F. Lam, C. L. Wilkins, T. R. Brunner, L. J. Soltzberg and S. L. Kaberline, Large-scale mass spectral analysis by simplex pattern recognition, Anal. Chem., 48 (1976) 1768.
84 G. E. Mieling, R. W. Taylor, L. G. Hargis, J. English and H. L. Pardue, Fully automated stopped-flow studies with a hierarchical computer controlled system, Anal. Chem., 48 (1976) 1686.
85 S. L. Morgan and S. N. Deming, Experimental optimization of chromatographic systems. Sep. Pur. Methods, 5 (1976) 333.
86 A. S. Olansky and S. N. Deming, Optimization and interpretation of absorbance response in the determination of formaldehyde with chromotropic acid, Anal. Chim. Acta, 83 (1976) 241.
87 E. Pelletier, P. Roche and B. Vidal, Automatic evaluation of optical constants and thickness of thin films: application to thin dielectric layers, Nouv. Rev. Opt., 7 (1976) 353; Chem. Abstr., 86: 81012f.
88 R. F. Reiss and A. J. Katz, Optimizing recovery of platelets in platelet rich plasma by the simplex strategy, Transfusion, 16 (1976) 370.
89 R. Rousson, G. Tantot and M. Tournarie, Numerical determination of vibrational force constants by means of a simplex minimizing method, J. Mol. Spectrosc., 59 (1976) 1.
90 G. L. Silver, Simplex characterization of equilibrium: application to plutonium, Radiochem. Radioanal. Lett., 27 (1976) 243.
91 C. Vanroelen, R. Smits, P. Van den Winkel and D. L. Massart, Application of factor analysis and simplex techniques to the optimization of a phosphate determination via molybdenum blue, Fresenius Z. Anal. Chem., 280 (1976) 21.
92 R. Wodzki and J. Ceynowa, Simplex design method for planning optimum experiments, Wład. Chem., 30 (1976) 337; Chem. Abstr., 86: 31497x.
93 J. W. Akitt, Function minimization using the Nelder and Mead simplex method with limited arithmetic precision: the self regenerative simplex, Computer J., 20 (1977) 84.
94 W. A. M. Castenmiller and H. M. Buck, A quantum chemical study of the $C_6H_5^+$ potential energy surface. Evidence for a nonclassical pyramidal carbenic species, Recl. Trav. Chim. Pays-Bas, 96 (1977) 207; Chem. Abstr., 87: 183780t.

95 J. Ceynowa and R. Wodzki, Simplex optimization of carbon electrodes for the hydrogen—oxygen membrane fuel cell, J. Power Sources, 1 (1977) 323.
96 J.-L. DeCoen and E. Ralston, Theoretical conformational analysis of Asn^1, Val^5 angiotensin II, Biopolymers, 16 (1977) 1929.
97 S. N. Deming and S. L. Morgan, Advances in the application of optimization methodology in chemistry, in B. R. Kowalski (Ed.), Chemometrics: Theory and Application, ACS Symp. Ser. 52, Am. Chem. Soc., Washington DC, 1977, Ch. 1.
98 P. W. Dillon and G. R. Underwood, Cyclic allenes. 2. The conversion of cyclopropylidenes to allenes. A simplex-INDO study, J. Am. Chem. Soc., 99 (1977) 2435.
99 W. J. Evans, V. L. Frampton and A. D. French, A comparative analysis of the interaction of mannitol with borate by calorimetric and pH techniques, J. Phys. Chem., 81 (1977) 1810.
100 R. D. Gilliom, W. P. Purcell and T. R. Bosin, Sequential simplex optimization applied to drug design in the indole, 1-methylindole, and benzo[b]thiophene series, Eur. J. Med. Chem.-Chim. Ther., 12 (1977) 187.
101 R. Lazaro, P. Bouchet and R. Jacquier, Experimental design. II. Simplex optimization of the synthesis of an industrial pyrazolone, Bull. Soc. Chim. Fr., 11—12, pt. 2 (1977) 1171; Chem. Abstr., 89: 42106g.
102 D. J. Leggett, Numerical analysis of multicomponent spectra, Anal. Chem., 49 (1977) 276.
103 C. L. McMinn and J. H. Ottaway, Studies on the mechanism and kinetics of the 2-oxoglutarate dehydrogenase system from pig heart, Biochem. J., 161 (1977) 569.
104 Y. C. Martin and J. J. Hackbarth, Examples of the application of non-linear regression analysis to chemical data, in B. R. Kowalski (Ed.), Chemometrics: Theory and Application, ACS Symp. Ser. 52, Am. Chem. Soc., Washington DC, 1977, Ch. 8.
105 F. Mayne, Optimization techniques and galenic formulation: example of sequential simplex and compressed tablets, 1st, Expo-Congr. Int. Technol. Pharm., 5 (1977) 65; Chem. Abstr., 90: 43769h.
106 M. D. Medvedev, Simplex planning of experiments in the study of hydrocarbon pyrolysis, Izv. Tomsk. Politekh. Inst., 300 (1977) 126; Chem. Abstr., 88: 52754k.
107 M. D. Medvedev, Study of the pyrolysis of Samotlorsk petroleum fractions, Izv. Tomsk. Politekh. Inst., 300 (1977) 135; Chem. Abstr., 88: 52755m.
108 P. G. Mezey, M. R. Peterson and I. G. Csizmadia, Transition state determination by the x-method, Can. J. Chem., 55 (1977) 2941.
109 B. Moraweck, P. deMontgolfier and A. J. Renouprez, X-ray line—profile analysis. I. A method for unfolding diffraction profiles, J. Appl. Crystallogr., 10 (1977) 184.
110 A. S. Olansky, L. R. Parker, Jr., S. L. Morgan and S. N. Deming, Automated development of analytical methods. The determination of serum calcium by the cresolphthalein complexone method, Anal. Chim. Acta, 95 (1977) 107.
111 P. V. Pedersen, Curve fitting and modeling in pharmacokinetics and some practical experiences with NONLIN and a new program FUNFIT, J. Pharmacokinetics and Biopharmaceutics, 5 (1977) 512.
112 M. L. Rainey and W. C. Purdy, Simplex optimization of the separation of phospholipids by high-pressure liquid chromatography, Anal. Chim. Acta, 93 (1977) 211.
113 M. W. Routh, P. A. Swartz and M. B. Denton, Performance of the super modified simplex, Anal. Chem., 49 (1977) 1422.
114 M. L. H. Turoff and S. N. Deming, Optimization of the extraction of iron(II) from water into cyclohexane with hexafluoroacetylacetone and tri-n-butyl phosphate, Talanta, 24 (1977) 567.
115 C. L. Wilkins, Interactive pattern recognition in the chemical analysis laboratory, J. Chem. Inf. Comput. Sci., 17 (1977) 242.
116 H. Y. Cheng and R. L. McCreery, Simultaneous determination of reversible potential and rate constant for a first-order EC reaction by potential dependent chronoamperometry, Anal. Chem., 50 (1978) 645.

117 F. Darvas, L. Kovacs and A. Eory, Computer optimization by the sequential simplex method in designing drug analogs, Abh. Akad. Wiss. DDR, Abt. Math., Naturwiss., Tech., 1978 (2N, Quant. Struct.-Act. Anal.), 311; Chem. Abstr., 91: 32476e.
118 M. G. Davydov and A. P. Naumov, Optimization of multielement quantitative activation analysis, Radiochem. Radioanal. Lett., 35 (1978) 77.
119 S. N. Deming, Optimization of methods, in R. F. Hirsh (Ed.), Statistics 1977, Eastern Analytical Symposium, The Franklin Inst. Press, Philadelphia, PA, 1978, pp. 31—35.
120 S. N. Deming and L. R. Parker, A review of simplex optimization in analytical chemistry, CRC Crit. Rev. Anal. Chem., 7 (1978) 187.
121 J. W. Evans and A. R. Manson, Optimal experimental designs in two dimensions using minimum bias estimation, J. Am. Stat. Assoc., 73 (1978) 171.
122 M. K. Hanafey, R. L. Scott, T. H. Ridgway and C. N. Reilley, Analysis of electrochemical mechanisms by finite difference simulation and simplex fitting of double potential step current, charge, and absorbance responses, Anal. Chem., 50 (1978) 116.
123 S. L. Kaberline and C. L. Wilkins, Evaluation of the super-modified simplex for use in chemical pattern recognition, Anal. Chim. Acta, 103 (1978) 417.
124 C. H. Lin and S. C. Liu, A new numerical method for automated spectral isolation of component substances in a set of mixtures, J. Chin. Chem. Soc., 25 (1978) 167.
125 V. E. Medyantsev, D. A. Vyakhirev and M. Y. Shtaerman, Improvement in the gas chromatographic separation of resinates using a mathematical method of experiment planning, Gidroliz. Lesokhim. Promst., 5 (1978) 18; Chem. Abstr., 89: 181417x.
126 R. G. Michel, J. Coleman and J. D. Winefordner, A reproducible method for preparation and operation of microwave excited electrodeless discharge lamps: simplex optimization of experimental factors for a cadmium lamp, Spectrochim. Acta, Part B, 33 (1978) 195.
127 S. L. Morgan and C. A. Jacques, Response surface evaluation and optimization in gas chromatography, J. Chromatogr. Sci., 16 (1978) 500.
128 K. Muller and L. D. Brown, Enamines. 1. Vinyl amine, theoretical study of its structure, electrostatic potential, and proton affinity, Helv. Chim. Acta, 61 (1978) 1407.
129 A. S. Olansky and S. N. Deming, Automated development of a kinetic method for the continuous-flow determination of creatinine, Clin. Chem., 24 (1978) 2115.
130 S. S. Rao, Optimization: Theory and Applications, Wiley Eastern Limited, New Delhi, 1978, pp. 284—292.
131 M. Suchanek, L. Sucha and Z. Urner, Optimization of analytical procedures, Chem. Listy, 72 (1978) 1037; Chem. Abstr., 90: 33385r.
132 H. Winicov, J. Schainbaum, J. Buckley, G. Longino, J. Hill and C. E. Berkoff, Chemical process optimization by computer, a self-directed chemical synthesis system, Anal. Chim. Acta, 103 (1978) 469.
133 F. J. Yang, A. C. Brown III and S. P. Cram, Splitless sampling for capillary-column gas chromatography, J. Chromatogr., 158 (1978) 91.
134 D. S. Amenta, C. E. Lamb and J. J. Leary, Simplex optimization of yield of sec-butylbenzene in a Friedel—Crafts alkylation. A special undergraduate project in analytical chemistry, J. Chem. Educ., 56 (1979) 557.
135 G. F. Brissey, R. B. Spencer and C. L. Wilkins, High-speed algorithm for simplex optimization calculations, Anal. Chem., 51 (1979) 2295.
136 D. Brunel, J. Itier, A. Commeyras, R. Phan Tan Luu and D. Mathieu, Les acides perfluorosulfoniques. II. Activation du n-pentane par les systémes superacides du type R_FSO_3H—SbF_5. Recherche des conditions optimales dans le cas des acides $C_4F_9SO_3H$ et CF_3SO_3H, Bull. Soc. Chim. Fr., 5 (1979) 257.
137 K. Doerffel and G. Ehrlich, Process optimization and interpretation of measurements, necessities of modern analytical chemistry, Wiss. Z.-Karl-Marx-Univ., Leipzig, Math.-Naturwiss. Reihe, 28 (1979) 459; Chem. Abstr., 92: 68915g.
138 J. W. Evans, Computer augmentation of experimental designs to maximize [X'X], Technometrics, 21 (1979) 321.

139 C. E. Fiori and R. L. Myklbust, A simplex method for fitting Gaussian profiles to x-ray spectra obtained with an energy-dispersive detector, D. O. E. Symp. Ser. 1978 (1979), 49 (Comput. Act. Anal. Gamma-Ray Spectrosc.) 139; Chem. Abstr., 92: 157089h.
140 E. Gey and W. Kühnel, Berechnung von potentialflachenausschnitten radikalischer anlagerungsreaktionen an olefinische doppelbindungen semiempirische untersuchungene der reaktion $^{\bullet}CH_3 + C_2H_4$, Collect. Czech. Chem. Commun., 44 (1979) 3649.
141 B. B. Jablonski, W. Wegscheider and D. E. Leyden, Evaluation of computer directed optimization for energy dispersive x-ray spectrometry, Anal. Chem., 51 (1979) 2359.
142 M. C. Kohn, L. E. Menten and D. Garfinkel, A convenient computer program for fitting enzymatic rate laws to steady-state data, Comput. Biomed. Res., 12 (1979) 461.
143 C. F. Lam, A. Forst and H. Bank, Simplex: a method for spectral deconvolution applicable to energy dispersion analysis, Appl. Spectrosc., 33 (1979) 273.
144 E. Li-Chan, N. Helbig, E. Holbek, S. Chau and S. Nakai, Covalent attachment of lysine to wheat gluten for nutritional improvement, J. Agric. Food Chem., 27 (1979) 877.
145 K. Mueller and L. D. Brown, Location of saddle points and minimum energy paths by a constrained simplex optimization procedure, Theor. Chim. Acta, 53 (1979) 75.
146 C. L. Shavers, M. L. Parsons and S. N. Deming, Simplex optimization of chemical systems, J. Chem. Educ., 56 (1979) 307.
147 V. J. Shiner, Jr., D. A. Nollen and K. Humski, Multiparameter optimization procedure for the analysis of reaction mechanistic schemes. Solvolyses of cyclopentyl *p*-bromobenzenesulfonate, J. Org. Chem., 44 (1979) 2108.
148 M. Singliar and L. Koudelka, Optimization of the conditions of chromatographic analysis of technical mixtures of glycol ethers, Chem. Prum., 29 (1979) 134; Chem. Abstr., 91: 13182q.
149 X. Tomas, J. Hernandez and L. G. Sabate, The use of the simplex method in the optimization of chromatographic separations, Afinidad, 36 (1979) 485; Chem. Abstr., 93: 32165u.
150 F. R. van de Voort, C. Ma and S. Nakai, Molecular weight distribution of interacting proteins calculated by multiple regression analysis from sedimentation equilibrium data: an interpretation of α_{s1}-κ-casein interaction, Arch. Biochem. Biophys., 195 (1979) 596.
151 F. Vlacil and D. K. Huynh, Determination of low concentrations of dibenzyl sulfoxide in aqueous solutions, Collect. Czech. Chem. Commun., 44 (1979) 1908; Chem. Abstr., 91: 203879v.
152 M. W. Watson and P. W. Carr, Simplex algorithm for the optimization of gradient elution high-performance liquid chromatography, Anal. Chem., 51 (1979) 1835.
153 W. Wegscheider, B. B. Jablonski and D. E. Leyden, Automated determination of optimum excitation conditions for single and multielement analysis with energy dispersive x-ray fluorescence spectrometry, Adv. X-Ray Anal., 22 (1979) 433.
154 W. E. Biles and J. J. Swain, Optimization and Industrial Experimentation, John Wiley, New York, 1980.
155 S. Bolanca and A. Golubovic, The determination of ink composition by means of the simplex method, Hem. Ind., 34 (1980) 168; Chem. Abstr., 93: 96911s.
156 J. Borszeki, K. Doerffel and E. Gegus, Application of experimental planning methods in chemical research. III. Optimization of calcium atomic absorption determination using the simplex method, Magy. Kem. Foly., 86 (1980) 207; Chem. Abstr., 93: 36232m.
157 D. B. Chestnut and F. W. Whitehurst, A simplex optimized Indo calculation of ^{13}C chemical shifts in hydrocarbons, J. Computational Chem., 1 (1980) 36.
158 F. L. Chubb, J. T. Edward and S. C. Wong, Simplex optimization of yields in the Bucherer—Bergs reaction, J. Org. Chem., 45 (1980) 2315.
159 L. Ebdon, M. R. Cave and D. J. Mowthorpe, Simplex optimization of inductively coupled plasmas, Anal. Chim. Acta, 115 (1980) 179.

160 R. Fletcher, Practical Methods of Optimization, Vol. I, Unconstrained Optimization, John Wiley, New York, 1980, pp. 14—16.
161 C. Hendrix, Through the response surface with test tube and pipe wrench, CHEMTECH, 10 (1980) 488.
162 F. Hsu, J. Anderson and A. Zlatkis, A practical approach to optimization of a selective gas chromatographic detector by a sequential simplex method, J. High Res. Chrom. Chrom. Commun., 3 (1980) 648.
163 K. Hyakuna and G. L. Samuelson, Shimplex-automated control system for Y and C shims, Jpn. Electron Opt. Lab. News, 16A, No. 1 (1980) 17.
164 S. V. Lyubimova, S. V. Mamikonyan, Y. N. Svetailo and K. I. Shchekin, Optimization of the composition of substances for an experiment in calibration of a multielement radiometric x-ray analyzer, Vopr. At. Nauki Tekh., [Ser.]: Radiats. Tekh., 19 (1980) 176; Chem. Abstr., 95: 125405d.
165 R. J. McDevitt and B. J. Barker, Simplex optimization of the synergic extraction of a bis-diketo copper (II) complex, Anal. Chim. Acta, 122 (1980) 223.
166 A. Mangia, A new approach to the problem of kaolinite interference in the determination of chrysotile asbestos by means of x-ray diffraction, Anal. Chim. Acta, 117 (1980) 337.
167 T. Michalowski, A. Rokosz and E. Wojcik, Optimization of the conventional method for determination of zinc as 8-hydroxyquinolinate in alkaline tartrate medium, Chem. Anal. (Warsaw), 25 (1980) 563; Chem. Abstr., 94: 131609u.
168 G. Minkova and V. Baeva, Optimization of the isomerization of *p*-(isoamyloxy) anilinium thiocyanate to [*p*-(isoamyloxy)phenyl] thiourea. I. Experimental optimization by the simplex method, Khim. Ind. (Sofia), 9 (1980) 402; Chem. Abstr., 95: 24485z.
169 P. B. Ryan, R. L. Barr and H. D. Todd, Simplex techniques for nonlinear optimization, Anal. Chem., 52 (1980) 1460.
170 E. Shek, M. Ghani and R. E. Jones, Simplex search in optimization of capsule formulation, J. Pharm. Sci., 69 (1980) 1135.
171 S. K. Silber, R. A. Deans and R. A. Geanangel, A comparison of the simplex and Gauss iterative algorithms for curve fitting in Mossbauer spectra, Computers and Chem., 4 (1980) 123.
172 S. Stieg and T. A. Nieman, Application of a microcomputer controlled chemiluminescence research instrument to the simultaneous determination of cobalt (II) and silver (I) by gallic acid chemiluminescence, Anal. Chem., 52 (1980) 800.
173 V. Svoboda, Search for the optimal eluent composition for isocratic liquid column chromatography, J. Chromatogr., 201 (1980) 241.
174 P. F. A. Van der Wiel, Improvement of the super-modified simplex optimization procedure, Anal. Chim. Acta, 122 (1980) 421.
175 J. T. Wrobleski and D. B. Brown, A study of the variable-temperature magnetic susceptibility of two Ti(III) oxalate complexes, Inorg. Chim. Acta, 38 (1980) 227.
176 C. Zimmermann and W. E. Hoehne, Simplex optimization of a fluorimetric determination of the pyruvate kinase and phosphofructokinase activities from rabbit muscle using fluorescent adenine nucleotides, Z. Med. Laboratoriumsdiagn., 21 (1980) 259; Chem. Abstr., 94: 1422s.
177 M. R. Cave, D. M. Kaminaris, L. Ebdon and D. J. Mowthorpe, Fundamental studies of the application of an inductively coupled plasma to metallurgical analysis, Anal. Proc. (London), 18 (1981) 12.
178 L. Ebdon, The optimization of an inductively coupled plasma for metallurgical analysis, in R. M. Barnes (Ed.), Dev. At. Plasma Spectrochem. Anal., Proc. Int. Winter Conf. 1980, Heyden, London, 1981, p. 94; Chem. Abstr., 96: 173462.
179 R. J. Matthews, S. R. Goode and S. L. Morgan, Characterization of an enzymatic determination of arsenic(V) based on response surface methodology, Anal. Chim. Acta, 133 (1981) 169.

180 M. Otto and G. Werner, Optimization of a kinetic—catalytic method by use of a numerical model and the simplex method, Anal. Chim. Acta, 128 (1981) 177.
181 O. V. Sakhartova, V. Sates, J. Freimanis and A. Avots, Chromatography of prostaglandins, their analogs and precursors. I. Gas chromatographic control of the stages of 2-(6-carboethoxyhexyl)-6-endo-vinylbicyclo[3.1.=0] hexan-1-one synthesis, Latv. PSR Zinat. Akad. Vestis, Kim. Ser., 4 (1981) 414; Chem. Abstr., 95: 180312w.
182 M. M. Siegel, The use of the modified simplex method for automatic phase correction in Fourier-transform nuclear magnetic resonance spectroscopy, Anal. Chim. Acta, 133 (1981) 103.
183 I. Taufer and J. Tauferova, Experiment planning in determining the operational regime of an atomic absorption spectrophotometer by the simplex method, Chem. Prum., 31 (1981) 16; Chem. Abstr., 94: 95145d.
184 S. P. Terblanche, K. Visser and P. B. Zeeman, The modified sequential simplex method of optimization as applied to an inductively coupled plasma source, Spectrochim. Acta, Part B, 36 (1981) 293.
185 T. Yoshida, M. Sueki, H. Taguchi, S. Kulprecha and N. Nilubol, Modelling and optimization of steroid transformation in a mixed culture, Eur. J. Appl. Microbiol. Biotechnol., 11 (1981) 81.
186 J. C. Berridge, Unattended optimisation of reversed-phase high-performance liquid chromatographic separations using the modified simplex algorithm, J. Chromatogr., 244 (1982) 1.
187 H. N. Cheng, Markovian statistics and simplex algorithm for carbon-13 nuclear magnetic resonance spectra of ethylene—propylene copolymers, Anal. Chem., 54 (1982) 1828.
188 J. J. Leary, A. E. Brooks, A. F. Dorrzapf, Jr. and D. W. Golightly, An objective function for optimization techniques in simultaneous multiple-element analysis by inductively coupled plasma spectrometry, Appl. Spectrosc., 36 (1982) 37.
189 G. Wuensch, N. Czech and G. Hegenberg, Determination of tungsten with the capacitively coupled microwave plasma (CMP). Optimization of a CMP using factorial design and simplex method, Fresenius Z. Anal. Chem., 310 (1982) 62; Chem. Abstr., 96: 192515y.

Analytica Chimica Acta, 150 (1983) 199—206
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

TEACHING CHEMOMETRICS

BERNARD G. M. VANDEGINSTE
Department of Analytical Chemistry, University of Nijmegen, Nijmegen (The Netherlands)
(Received 4th October 1982)

SUMMARY

Learning the general concepts of the analytical process, which runs from formulation of a problem to its solution, is an essential part of the academic education of analytical chemists. Started as a combination of data processing and evaluation of results, chemometrics now includes formal studies of the analytical process, i.e., the fundamentals of analytical chemistry. In teaching chemometrics, the analytical process must remain the central point, and clear distinctions between auxiliary disciplines and chemometrics are necessary.

In a recent inquiry among Dutch chemical companies on the future needs for post-academic education, some interesting facts were revealed. In its report [1], the committee of inquiry printed two tables. In the first, about twenty topics were listed for which the committee proposed to initiate post-academic education and which the committee considered would cover most of the interests of future participants. Surprisingly, none of the topics was related to chemometrics. Of course, the interests expressed in the various topics were quite different. In the second table, the committee provided a supplementary list of the "missing" subjects. Participants were asked to indicate which topics they had missed in the first table. Two of the topics were in the field of chemometrics this time. The answers were very positive: the score of both chemometric subjects was so high that both topics indubitably deserved to be ranked in the top ten of the first list. Of course, it would be simplistic to maintain that this is an adequate explanation of why chemometrics should be taught. Yet it indicates clearly the importance of chemometrics and the demand for education in this area.

Chemometrics is mainly a discipline within analytical chemistry. Therefore, education in analytical chemistry may be a good starting point for a discussion of the present and future teaching of chemometrics. In the widest sense, the motive for education in analytical chemistry is to ensure that present work can be taken over, as required, by a new generation of young and enthusiastic scientists. In addition to academic education, much practical experience is still needed to become a professional analytical chemist. The link between daily practice and education is essential to ensure that education keeps in step with dynamic developments. This is especially true of analyt-

0003-2670/83/$03.00

ical chemistry, where practical applications quickly follow the results of research.

CURRENT EDUCATION IN ANALYTICAL CHEMISTRY

Present rapid progress in analytical chemistry makes it desirable to look back to the earlier education system for analytical chemistry. The facts are evident: (1) many of today's analytical chemists did not graduate in analytical chemistry; (2) in the present era of computers and automation, the face of analytical chemistry has changed dramatically over a time span of less than 10 years. Accordingly, there is an inevitable gap between the education of today and the analytical profession of tomorrow. It seems very improbable that progress will slow down in the next decade. New topics today may be out-of-date in a few years.

These considerations indicate that the emphasis of education should be more on general principles, rather than on acquiring very specific knowledge. The framework of training in analytical chemistry, one or more decades ago, is shown in Fig. 1. This curriculum still stands as a model of education at many universities, though there is a general awareness of the serious problems associated with it. First, much time is devoted to teaching supporting disciplines under the heading of analytical chemistry. Some courses are largely concerned with wet chemistry or are courses in statistics with a little dressing of analytical chemistry. Teaching the background of instrumental methods

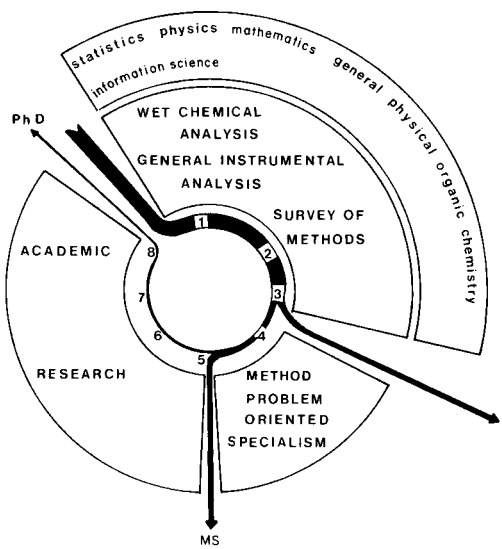


Fig. 1. Classical education in analytical chemistry. The numbers are the time spent in years; the width of the line refers to the number of students. Years 1—3 are the undergraduate level; years 3—8 are the graduate level.

often requires extensive discussion of the principles of physics or physical chemistry. As a result, many students get a wrong impression about the identity of analytical chemistry. Secondly, education in instrumental analysis and applications is not exclusive to analytical chemistry. Biochemists, organic chemists and physical chemists teach and employ their "own" methods: electrophoresis, ultracentrifuge, infrared, nuclear magnetic resonance and mass spectrometry, chromatography, etc. As a result, analytical chemistry as a science is under heavy discussion at many places. So long as analytical chemists are experts in, and teaching, just another subset or another class of instrumental methods, pressure on the credibility of analytical chemistry will continue. What is surely needed, as has been mentioned before [2—5], is to project the peculiar character and the fundamentals of analytical chemistry.

Chemical analysis and analytical chemistry

The analytical process is the key object in this branch of science. Every analytical process requires the application of some general principles of analysis (e.g., calibration), which are independent of the measuring principle. Once these general principles are known, specialization in a particular method takes only a step further by learning the particular physical background and principles. The non-analytical chemist may know a lot about the principles of a particular method, but a bad sampling strategy may render his results worthless. Sampling is one of the most important stages of the analytical process, and the theory of sampling is a typical problem in analytical chemistry. Similar reasoning is valid for every stage of the analytical process

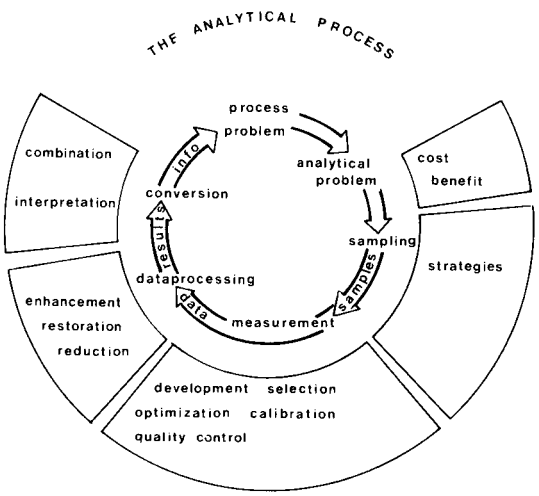


Fig. 2. Stages in the analytical process and associated problems.

(Fig. 2): the selection of an analytical method, the optimization, calibration and quality control of the measurement process, the processing of rough data, the interpretation and combination of analytical results into systematic information, the laboratory organization, etc.

Chemical analysis is done by following well known paths and schemes; that is what non-analytical chemists do. For new problems, new paths and schemes have to be created, using methods developed by the analytical chemist. In a sense, analytical chemistry for the analytical chemist begins where chemical analysis ends for the non-analytical chemist.

THE FORMAL APPROACH TO THE ANALYTICAL PROCESS: CHEMOMETRICS

It is the uneasy task of the analytical chemist to take the right, or optimal, decisions at every stage of the analytical process. Recently, a large part of the decision process was considered as impossible to formulate. Many have put up with the apparent fact that a successful analyst has an inexplicable sense of the right decision. But this reduces analytical chemistry to an art, which it is not.

A study of the analytical process involves the design of robust mathematical models, based on a systems approach to the process. With these models, decisions can be taken on the basis of scientifically sound principles. Chemometrics can play an essential role in that respect. Started as a discipline of data processing and data interpretation, chemometrics has evolved to a discipline that covers the whole analytical process.

Chemometrics can be defined as a chemical discipline that uses mathematical and statistical methods (a) to design or select optimal measurement procedures and experiments; and (b) to provide maximum chemical information by evaluating chemical data. In the field of analytical chemistry, chemometrics is used to obtain information about material systems. Keywords in the definition are "optimal" and "material systems". They express the fact that chemical analysis is related to a problem that should be solved in an optimal way. Optimal implies the definition of optimization criteria and the use of optimization methods. The sample is a carrier of bits of information. The analytical chemist has to combine all the fragmented information into concise system information.

Considerable progress has already been made in the development of valuable mathematical and statistical models. Many decision problems have been rationalized with success, despite the relatively short lifetime of chemometrics. Various examples can be given. Sampling schemes for the description, control and surveillance of processes are completed. The problem of choosing analytical methods can be tackled with "measurability models" based on a time series analysis. Multiparametric optimization methods like simplex optimization and factorial design, are new tools of growing importance for method optimization. The first prototypes of computer-controlled self-optimization systems are already available. The inclusion of artificial intelligence is becoming ever closer.

Pattern recognition has provided capabilities to look into the multi-dimensional space of analytical results. Instead of all analytical results being communicated separately (e.g., clinical tests on a patient), there is an increasing possibility of combining this information into unambiguous systems information (e.g., there is a given probability that the patient has a disease or not). Better tools for calibration and data processing are becoming available, e.g., GSAM, Kalman filtering and curve resolution. Operations research techniques have been invoked to combine analytical methods and procedures in an optimal way. Decisions on laboratory organisation (e.g., on the sample routing) can be made by the application of digital simulation. Finally, methods for the characterization of analytical signals have opened new ways for rational improvement of the performance of analytical methods.

The above examples clearly demonstrate the tremendous complexity of the analytical process, which requires the support of many areas of applied mathematics, e.g., control theory, information theory, operations research, multivariate statistics, pattern recognition (Fig. 3).

EDUCATION IN CHEMOMETRICS

Chemometrics is so strongly allied with basic questions of the analytical process that it is essential in any curriculum for analytical chemists. This probably explains the present big demand for post-academic education in chemometrics.

At the University of Nijmegen, courses in chemometrics are offered according to the global model shown in Fig. 4A. One similarity with classical education in analytical chemistry is very clear; both need the support of auxiliary disciplines. Danger lies in that similarity, however; analytical chem-

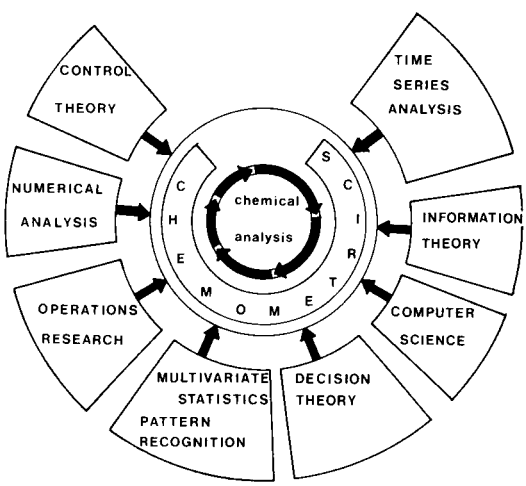


Fig. 3. Mathematical techniques used in chemometrics.

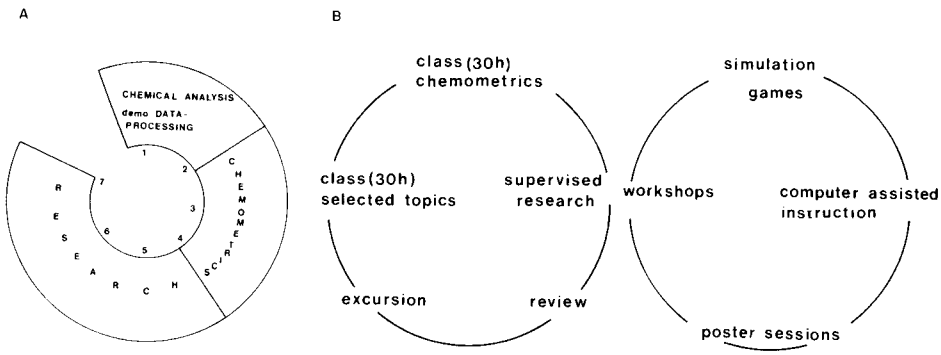


Fig. 4. Curriculum for education in chemometrics at the University of Nijmegen. For explanation, see text.

istry risked its identity by considering the auxiliary disciplines as analytical chemistry and the same mistake should not be made in the field of chemometrics. Therefore, it is not advisable to concentrate education in chemometrics on only a few topics. With this principle in mind, the Nijmegen course in chemometrics was designed around the analytical process. Undergraduate students start to learn the principles of analytical measurements. Sticking to our opinion that chemical analysis is not a synonym for analytical chemistry, these classes are not taught under the heading of analytical chemistry, nor are they necessarily taught by analytical chemists alone. Because all students (in Nijmegen) also follow a class in information science, plans are underway to introduce the students to the field of chemometrics by demonstrating the capabilities of data processing by computer. This will be accomplished essentially during practical work (e.g., a gas chromatograph linked to a microcomputer).

After completing the uniform undergraduate program, students enrol for different curricula. At this stage, they can decide to take a 1/2 or 1 year course in chemometrics. This part of the education is under full responsibility of the department of Analytical Chemistry and comprises the items shown in Fig. 4B. The details are as follows: a 30-h course in chemometrics is accompanied by a 30-h course in selected problem-orientated topics of analytical chemistry, and supervised research in one particular area of chemometrics. A critical review (20—50 pages) on a selected topic from chemometrics has to be written. There are compulsory weekly poster sessions where results of research are presented; this encourages a confrontation with a wide variety of chemometric methods, and each student must present a poster paper every 6 weeks. There are also excursions to an industrial laboratory to obtain a feeling for real analytical chemistry. After graduation, selected students can undertake studies for a doctorate, which requires 3—4 years of research in the field of chemometrics.

The central theme of these courses in chemometrics is a discussion of the various mathematical models available for solving the problems encountered

at each stage of the analytical process (Fig. 3), from sampling up to managerial problems in the laboratory. It is our experience that with the support of workshops, demonstrations and simulation games, it is possible to treat the mathematics to the extent necessary for a clear understanding of the different topics (Fig. 5). Workshops are exercises mostly based on data obtained in completely simulated environment, e.g., evaluation of a detection limit. A package of demonstration programs has been designed [6] for a HP-9845B microcomputer equipped with powerful graphics. Students can run the programs individually or in a group. The package includes demonstrations of most of the topics discussed during the course. Students can vary several parameters and immediately evaluate the effects obtained, which are then compared with the theory [e.g., simplex optimization, curve fitting, digital simulation, sampling of internally correlated lots, threshold control, sequential analysis, filtering (in the Fourier domain), time series (autocorrelation)]. By means of a competition and management computer simulation game, a realistic environment is created where students take decisions for optimal control of, e.g., a fertilizer plant. Students are confronted with decisions about sampling schemes, the selection of measurement procedures (5 options), the design of measurement schemes in order to collect the necessary information for the quantitation of the various parameters in the models used (e.g., time constant of the process, precision/accuracy of the measurement). The game can be played in 2 modes: (i) a competitive game, where all students control processes with a similar characteristic; (ii) a management game, where five plants have to be controlled, having a laboratory available equipped with a limited number of instruments and personnel. Students have to set up a laboratory organization that can manage the control of all the plants; during the course of action, real-time decisions must be taken. Afterwards, results are evaluated.

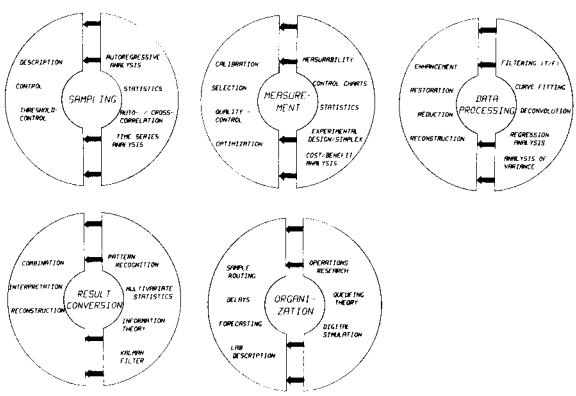


Fig. 5. Topics discussed in a 30-h course in chemometrics at the University of Nijmegen. Arrows indicate the auxiliary mathematical methods necessary to treat the topics shown to the left.

In the recent review, Frank and Kowalski [7] indicated that the computer must become the educational tool for chemometricians. By integrating the computer in a system of courses, workshops, demonstrations and simulation games, an optimal learning environment might be created. The system currently in use in Nijmegen is the product of a joint effort of several people over many years. The exchange of computerized educational tools covering a wide variety of chemometric topics would help greatly in improving the quality of the education in chemometrics.

The opinions presented in this paper are the result of many discussions with several members of the Department of Analytical Chemistry at the University of Nijmegen, and especially with Professor G. Kateman.

REFERENCES

1 Jaarverslag 1981, Centraal Overlegorgaan voor Post-Academisch Onderwijs.
2 D. L. Massart, Fresenius Z. Anal. Chem., 305 (1981) 113.
3 D. Betteridge, Fresenius Z. Anal. Chem., 297 (1979) 265.
4 G. Kateman and A. Dijkstra, Fresenius Z. Anal. Chem., 297 (1979) 249.
5 C. A. Lucchesi, International Laboratory, Nov./Dec., (1980) 67.
6 P. F. A. van der Wiel, G. Kateman, B. G. M. Vandeginste and T. A. H. M. Janse, presented at the CAC-II Conference, Petten, The Netherlands, 1982.
7 I. E. Frank and B. R. Kowalski, Anal. Chem., 54 (1982) 232R.

Analytica Chimica Acta, 150 (1983) 207—217
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

MODELLING COMPONENT COMBINATIONS BY MEANS OF ATTENTION FUNCTION SCORES

HENK M. J. GOLDSCHMIDT and JAN F. LEIJTEN

Department of Clinical Chemistry and Haematology, St. Elisabeth Hospital, Hilvarenbeekseweg 60, 5022 GC Tilburg (The Netherlands)

MARC N. M. SCHOLTEN

National Council for Health Services, J. C. van Markenlaan 5, 2280 AE Rijswijk (The Netherlands)

(Received 7th December 1982)

SUMMARY

The determination of a large number of components in a small sample is common practice in clinical chemistry. The optimal combination of components to be assayed in connection with a particular problem can be difficult to establish. This paper describes an attempt to achieve an optimal combination, not by means of the correlation matrix of component concentrations but by quantification of the attention that is given to the result for each component by the person requesting the analysis. Here, attention is defined as an intention to take the action. It is suggested that little attention is paid when the result is within or near certain expected limits, while a growing deviation from these limits will result in increasing attention. More attention will be paid to a larger deviation, resulting in ultimate saturation. This reasoning results in a sigmoidal curve on each side of the distribution. Missing data are no longer a problem, as they get zero attention. The actual curves, called attention functions, were established with the aid of the persons requesting the analyses, and with a derivation from the cumulative frequency distributions of the component concentrations. A set of 29 attention function scores was collected for each of 298 samples. A hierarchal cluster analysis was applied to these attention function scores to discover component similarities with regard to the attention eventually given to them. The combinations of components found were readily understandable. The advantage of this approach is that the criteria for combining components is directly linked to the daily practice of interpretation of the component concentrations by the people submitting the samples.

Clinical chemistry, in its present state with the aid of automated miniaturized analytical procedures, enables a large number of components to be determined rapidly in a small sample. Test profiles in medical practice are mainly used for screening purposes and confirmation of deviating results. In laboratory practice these profiles are standardized. Physicians can apply for a kidney profile, a heart profile, a liver profile, etc. These standard profiles have been compiled mainly on the basis of intuitive and economic reasonings [1]. A more straightforward methodological approach for the composition of these profiles was sought here.

0003-2670/83/$03.00

A primary consideration was that physicians are mainly interested in deviating results, so that profiles have to contain those components that deviate at the same time. This approach emphasizes the use of abnormal results in a screening situation and is only directed towards actions to be taken. The classical approach to solving a problem like this is to cluster deviations from the mean, measured in terms related to standard deviations (s.d.'s) of a reference population, sometimes preceded by mathematical transformations [2, 3]. These methods were rejected for the following reasons. First, they apply equal mathematical treatment of the results whether the results are lower or higher than their mean. In practice, the physician who requested the analysis does not intuitively assign such equal significance; e.g., a high creatinine result has medical consequences, a low level less so. Therefore it is wrong to compile profiles that give as much weight to a high positive deviation of 4 s.d.'s as to a negative deviation of 4 s.d.'s. Secondly, the classical methods do not take into account that the same results can be interpreted differently by different people, e.g., a cardiologist will interpret a high creatine kinase result quite differently from a physician for internal diseases. Thirdly, comparisons of deviations between tests, and of deviations at each side of a distribution expressed in s.d.-related units, are meaningful only when both the underlying and total a priori probability densities are symmetrical. However, most distributions of clinical chemical test results are skewed. It follows from this that the deviations, when measured in s.d.-related units, are no longer comparable. Normalization transformations are only partially applicable and differ from test to test.

The method of Shannon and Weaver [4] meets the objection of working with s.d.-related units. They calculated the information content, which is applicable to every distribution. However, this is unsuitable for the present purposes because these curves are not "adjustable" on each side of a distribution, i.e., they are independent of the person using the analytical results and the shape does not meet our terms. In the late sixties, Jungner and Jungner [5] introduced the medical weighted standard deviation units. Although based on a similar philosophy, their solution was a linear one and was not directed towards the search of profiles.

The concept proposed here for standardizing different test results is to quantify the attention given by the person requesting the analysis to a certain test result. It was considered that a low deviation from the mean gets little attention whereas a high deviation gets much attention. The feeling was that attention can be expressed as a sigmoidal curve as shown in Fig. 1. Attention is expressed in arbitrary units ranging from 0 (minimal) to 1 (maximal). Compared with the frequency distribution shown, two points of the proposed curve are now set: the mean of this distribution gets zero attention while the most deviating results at both sides get maximal attention. A third point, the point of inflexion, can also be set. This is done by means of a factor A ranging from 0 (minimal) to 1 (maximal) at an appointed attention of 0.5. The factor A is used to determine the shape of the attention curve. The procedure is explained in detail under Statistical Methods.

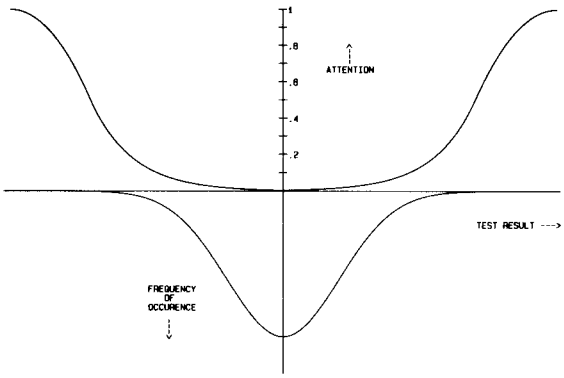


Fig. 1. The assumed path of attention compared with a corresponding relative frequency distribution.

The inflexion point can be interpreted as the action level for that test, as introduced by Elion-Gerritzen [6]. Often a physician still orders a number of laboratory tests even when he finds no abnormalities during a physical examination of a patient with ill-defined complaints. For each test, there are lower and upper limits which, if exceeded, prompt him to an action of the first or second order. These limits are called action levels. An action of the first order will include repeated or additional tests, electrocardiogram, x-ray, initiation or change of diet or medication. An action of the second order is defined as hospital admission or immediate treatment. In the present evaluation, it is assumed that attention grows most rapidly when the test result equals the first-order action level because in our in-patient hospital setting, second-order actions are poorly defined.

The individual differences between physicians, tests and sign of deviation can be built in by choosing different sigmoidal curves for each side of the distribution of each test by each physician. The final result is that each physician owns his own attention curve for each test. These sigmoidal curves are called attention functions. The attention given by a physician to a certain test result is likely to be inversely proportional to the frequency of his contact with very similar results. His experience has ranked the results in a relative cumulative frequency distribution. Based on this, the cumulative value of a result was used as the input variable of the attention functions.

Two transformations are therefore needed. First, the original result is converted to its relative cumulative frequency value; the patient population itself governs this transformation. Secondly, the relative cumulative frequency value is converted to an attention function score; this process depends on the person requesting the analysis. The attention function scores obtained were then clustered to yield profiles.

DATA SET AND METHODS

Data set

This study was restricted to a screening profile consisting of 29 biochemical and haematological tests of the Department of Internal Diseases, St. Elisabeth Hospital, Tilburg. The tests considered were chosen either because they formed part of a subprofile or because there was doubt whether or not they should be included in one. The results for 298 admitted patients were gathered at random. The only restriction for inclusion in the calculation was that at least one test of the 29 had to be requested by a physician. A board of physicians working in the department was then asked for their action levels. Because these physicians had to deal with the same kind of patients, no distinction was made between individual attention behaviour. When they differed in opinion about an action level, discussion was continued until an agreed decision was reached on a certain level. Because certain test results are sex-dependent, there are mixed populations at both sides (e.g., haemoglobin, haematocrit, uric acid). The most outlying population was chosen as the one governing the action limit. To determine the cumulative distributions of low and high deviations, 200 abnormal results of each test were collected on each side of the mean, abnormal being defined as outside the reference region. The distributions together with the action levels obtained gave the factor A values at both sides of the distribution. The A value is, as mentioned above, the relative cumulative frequency value of a test result that equals the action level. This A value is used to describe the attention function. Table 1 shows the selected tests with their action levels, A values and percentage missing data, respectively. Sometimes a deviating result at one side of the distribution does not make any physiological sense (e.g., low CPK values). Therefore the physicians did not apply any action at that side and the cumulative curve was linearized in that part of the distribution curve. Apart from that, linearization was done within all reference values, which for the present purpose introduces negligible errors.

Statistical methods

The method proposed to model component combinations consists of three steps: (a) choice of the particular attention function; (b) the score for a particular patient on the chosen function by means of the cumulative distribution; (c) hierarchal clustering of the attention function scores obtained.

Choice of attention function. Although there is obviously only one cumulative frequency distribution of test results, it was split into two fractions. These two fractions are situated at each side of the mean of the reference region. A set of attention functions for test results was defined at just one side of this mean. The functions depend only on A, the relative cumulative frequency value:

$$y \to 0.5 \exp[2(1-A)^{-1}(y-A)] \qquad \text{for} -2 \leqslant y \leqslant A$$

TABLE 1

The selected tests with the estimated first-order action levels and their A values for both sides of the distribution are shown. The side of the distribution with no action level and therefore no A value, is asterisked. For each test, the percentage of missing results is also listed

No.	Test name	Units	Action levels		A values		Missing results (%)
			low	high	low	high	
1	Sodium	mmol l^{-1}	130.0	147.0	0.915	0.669	20.5
2	Potassium	mmol l^{-1}	3.2	5.5	0.881	0.991	20.5
3	Urea	mmol l^{-1}	1.5	12.0	0.982	0.777	18.5
4	Creatinine	μmol l^{-1}	*	150.0	*	0.602	19.8
5	St. bicarbonate	mmol l^{-1}	20.0	30.0	0.436	0.896	37.9
6	Calcium	mmol l^{-1}	2.05	2.75	0.426	0.807	47.3
7	Total protein	g l^{-1}	55.0	90.0	0.884	0.915	46.0
8	Albumin	g l^{-1}	30.0	60.0	0.688	0.600	48.0
9	Uric acid	mmol l^{-1}	*	0.70	*	0.938	48.7
10	Inorg. phosphate	mmol l^{-1}	0.60	1.70	0.922	0.463	47.3
11	Chloride	mmol l^{-1}	80.0	110.0	0.983	0.891	20.5
12	CPK	U l^{-1}	*	250.0	*	0.788	70.5
13	Triglycerides	mmol l^{-1}	*	2.5	*	0.452	91.9
14	Glucose	mmol l^{-1}	2.0	8.0	0.882	0.655	47.7
15	Cholesterol	mmol l^{-1}	2.0	8.5	0.999	0.494	47.3
16	Haematocrit	$l\,l^{-1}$	0.27	0.51	0.804	0.688	32.2
17	Erythrocytes	10^{12} l^{-1}	3.38	6.05	0.581	0.933	32.2
18	Haemoglobin	mmol l^{-1}	6.9	11.8	0.473	0.917	32.2
19	Leucocytes	10^{9} l^{-1}	2.9	16.5	0.804	0.735	33.2
20	Platelets	10^{9} l^{-1}	50.0	500.0	0.864	0.735	33.6
21	MCHC	mmol l^{-1}	1640.0	2390.0	0.741	0.875	32.9
22	MCV	fl	76.0	115.0	0.846	0.897	32.9
23	MCH	amol	18.0	25.0	0.796	0.925	32.9
24	Alk. phosphatase	U l^{-1}	20.0	150.0	0.999	0.679	36.6
25	Gamma-GT	U l^{-1}	*	100.0	*	0.788	55.4
26	LDH	U l^{-1}	*	275.0	*	0.592	34.6
27	Bilirubin	μmol l^{-1}	*	18.0	*	0.540	35.2
28	SALT (PT)	U l^{-1}	*	65.0	*	0.765	49.7
29	SAST (OT)	U l^{-1}	*	70.0	*	0.654	34.6

$$y \to 0.5 - [2(1-A)^2]^{-1}(y-A)^2 + (y-A)/(1-A)^{-1} \quad \text{for } A < y \leqslant 1$$

The y values between -2 and 0 are the linearized values of the test results within the reference region, where -2 corresponds to the mean of the reference population and 0 to the reference value:

$$y = 2(x-m)/(RV-m) - 2$$

where x are test results within reference region, m is the mean reference population, and RV is the reference value. The x values greater than 0 are relative frequency values of test results. The cumulative distribution is governed only by those test results that are outside the reference region at

that single side. The resulting set of attention functions is shown in Fig. 2 for A running from 0.1 to 0.9.

This set of functions meets the purposes well. The shape of the curve is sigmoidal. Three assigned points govern the path of the curve: $x = A$ is the inflexion point with attention 0.5; the attention has minimal value 0, reached at a test result that equals the mean of the reference region of that component; the attention has maximal value 1, reached at a test result with value 1 in the transformed cumulative distribution.

Attention function score. Knowledge of the action level at a certain con-

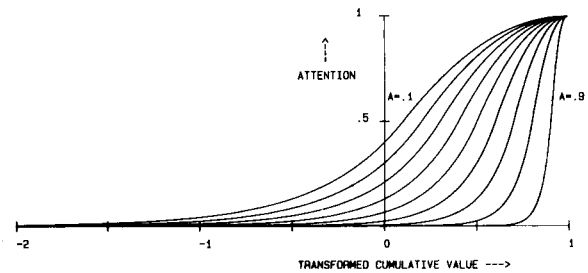


Fig. 2. A set of attention functions, for A values running from 0.1 to 0.9.

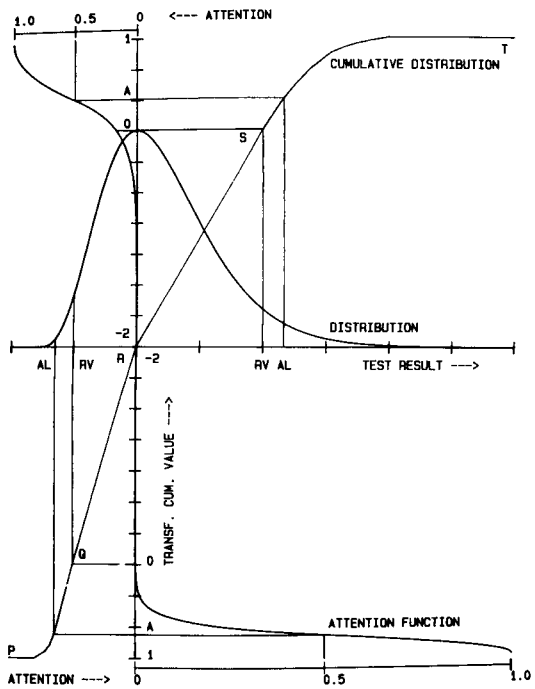


Fig. 3. An arbitrary frequency distribution with its corresponding transformed cumulative distribution and attention function. The reference values (RV) and action levels (AL) are also depicted.

centration of a test result, and knowledge of the transformed cumulative frequency distribution, made it possible to obtain the A value and the attention function of each test. So far, the handling of test results at one side of the mean of the reference region has been discussed. An analogous process can be done, if applicable, for both sides of the frequency distribution. When the attention functions for both sides are joined, the complete range of test results is covered.

Once the exact path of the attention function curves and the cumulative frequency distributions of the outlying results on each side of the population mean are known, the original test results for each patient can be converted to attention function scores. This process is illustrated in Fig. 3. An arbitrarily skewed distribution of test results is shown, together with two reference values (RV) and action levels (AL) relevant to this test. The cumulative distribution, adapted as mentioned above, is shown too. The original test results, shown on the middle horizontal axis, can then be converted via the adapted cumulative distribution to input values of the attention function, shown on the vertical axis, and further to attention scores on the upper or lower horizontal axis.

An example with full details is given in the Appendix. The simplification of linearizing the cumulative function in the reference region is acceptable for two reasons: first, the cumulative function is almost linear within the reference region; secondly, the attention increases very slowly in this part of the curve.

It will be noted that the problem of missing test results, which is inevitable in the classical method, is easy to handle in this approach. The missing data are created through tests that are not requested. In the proposed method, this means that they get zero attention from the physicians. The procedure outlined was done by means of a simple Algol-68 program on the ICL-2966 computer of Tilburg University.

Hierarchal clustering procedure

Finally, an hierarchal clustering procedure was applied to the attention function scores, to reveal any combinations of components with correlated attention function scores. In order to compare the classical approach with the new concept, the correlation matrix was calculated from the raw data as well as from the attention function scores. The hierarchal clustering can be reported as a branching tree diagram, giving the structure of the calculated matrix.

With the raw data, the absolute value of the correlations was used. In such cases, a strong negative correlation is as meaningful as a strong positive correlation, but this is not true in the case of the attention function scores. For example, if there is a perfect negative correlation in attention, then those tests would not fit into the same profile, because if one test receives a high attention score and is therefore important, the other one will not, and vice versa. Thus a negative correlation is a justification for mutual

exclusion of components from a profile. It is clear that very low correlations are not meaningful.

The software package CLUSTAN was used with the option advised by Tyron and Bailey [7]. This option uses the furthest neighbour technique with the correlation coefficients as similarity coefficients. This clustering technique tends towards dense clusters because minimum variance clusters are calculated.

RESULTS AND DISCUSSION

The percentage of missing test results is shown in Table 1; the overall mean is 39.3%. After the attention function scoring procedure, 12.7% of all results were converted to scores of at least 0.10. Figure 4 shows the branching tree diagram derived from the raw data. For example, the MCH and MCV results clustered at a high level (correlation coefficient R = 0.92) and are therefore very similar. In contrast, the triglycerides results showed virtually no correlation with the other test results (R = 0.02). If two variables, like MCH and MCV, are branched together, then another variable that has a relatively high correlation with these two is sought, in this case MCHC. The furthest neighbour criterion means that the correlations between MCHC and MCH—MCV are computed as the minimum of the correlation of MCHC—MCH and MCHC—MCV. Thus the branch is drawn at the minimum correlation available for those three variables, and so on. Figure 5 shows the branching tree diagram derived from the attention function scores.

In the process of evaluation of these diagrams, the profiles show up much more clearly in Fig. 5 than in Fig. 4. The latter figure shows some understandable correlations (e.g., MCH—MCV, haematocrit—erythrocytes, SAST—SALT, alkaline phosphatase—Gamma GT—LDH) but there are no

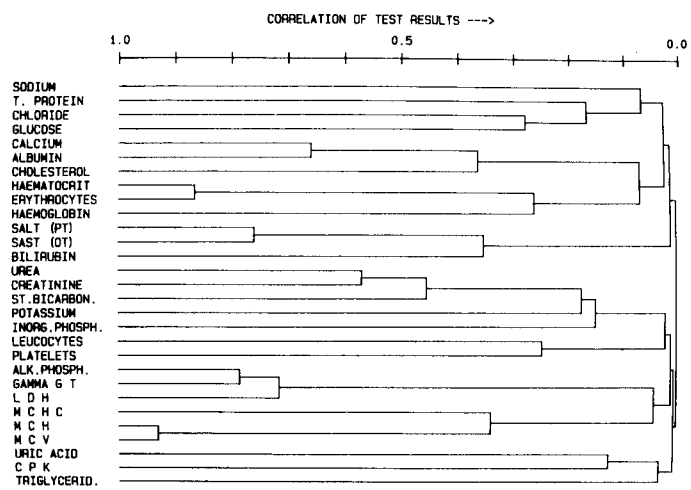


Fig. 4. Branching tree diagram from the raw test results of 298 patients.

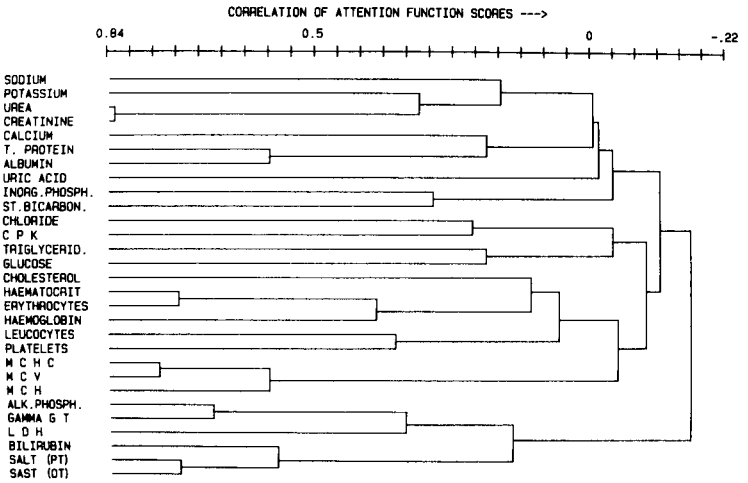


Fig. 5. Branching tree diagram from the attention function scores of test results of 298 patients.

striking profiles. Interpretation of the attention function scores diagram is more profitable with regard to comprehensive profiles. The following features are notable.

(a) A liver function profile (alkaline phosphatase—Gamma GT—LDH—SAST—SALT) is within a strong cluster, whereas the profile used until then did not contain the Gamma-GT test.

(b) Sodium, potassium, urea and creatinine can be regarded as kidney function tests, so this cluster can be described as a kidney function profile.

(c) The haematology tests are amalgamated into one cluster. It was surprising that the erythropoid cells cluster with cholesterol before clustering with the non-erythropoid cells and the blood indices.

(d) The cluster of albumin, total protein and calcium can be understood as a functional protein profile.

(e) The group of chloride, CPK, triglycerides and glucose can be explained as a group of tests that physicians do not pay much attention to in this context. This is understandable for CPK, triglycerides and glucose; CPK, mainly used in cardio-diagnostics, is of minor interest in the department of internal diseases. Glucose, as a pancreas function test, and triglycerides depend very much on the nutritional state of the patient. The interpretation of the chloride, as an outlier, is a problem. What had been expected was a cluster representing the ionic equilibrium, with sodium, potassium and standard bicarbonate which was not calculated.

(f) The group of uric acid, inorganic phosphate and standard bicarbonate is difficult to interpret in relation to each other besides their negative charge they have no physiological mechanism in common, to our knowledge. This says nothing about these tests for general screening purposes, but simply

that, in the patient population reviewed, no attention is given simultaneously to one of these tests and one of the others.

It is interesting to see the general approach of the medical practitioner from Table 1. On the whole, he uses two action limits if the supposed underlying physiological mechanism is that of a balance, but one upper action limit when some kind of cell disruption is assumed. Although they are of the same order, the present action levels tend to be more extreme than those found by Elion-Gerritzen [6].

Conclusions

The proposed method is very promising: the results are readily interpreted and the method is flexible. Subtle differences between physicians, patient populations and tests can either be built into the model or ignored, depending on the problem under investigation. It is clear that this approach, within the medical sciences, can be applied to similar situations in other fields where large data sets are created on a routine basis and their appreciation is very test-dependent.

An advantage of this method is its solution of the problem of missing test results. Besides the attention paid by the physician, it will be easy to introduce other scaling factors such as financial aspects or inconveniences to the patient. For too long, the appreciation of results by the person requesting the tests has been neglected as a valuable piece of information.

APPENDIX

The procedure for attention function scores is illustrated by means of the inorganic phosphate test. The reference region for this test is 0.81—1.54 mmol l^{-1} with a mean of 1.18 mmol l^{-1}. The transformed cumulative distribution is set by the following steps.

(1) Within the reference region, the transformed cumulative distribution is linearized. For test result x, when $0.81 \leqslant x \leqslant 1.18$ mmol l^{-1}, $y = 2(1.18 - x)/(1.18 - 0.81) - 2$. For test result x, when $1.18 \leqslant x \leqslant 1.54$ mmol l^{-1}, $y = 2(x - 1.18)/(1.54 - 1.18) - 2$. In Fig. 3 these are the lines Q—R, and R—S, respectively.

(2) By sampling 200 inorganic phosphate test results which were lower than 0.81 mmol l^{-1}, it was possible to estimate curve P—Q in Fig. 3. This curve represents the cumulative distribution of low test results which are out of the reference region.

(3) An analogous procedure leads to curve S—T in Fig. 3.

The curve constructed, PQRST, the transformed cumulative distribution, makes it possible to convert the test results (middle-x-axis) to transformed cumulative values (y-axis). For example:

Result (mmol l^{-1})	0.65	1.00	1.50	1.89
Trans. cum. value	0.844	—1.041	—0.219	0.727

The action levels, for this test, were set at 0.60 mmol l^{-1} and 1.70 mmol l^{-1}. The corresponding transformed cumulative values are 0.922 and 0.463, respectively. These figures are the A values that determine the specific attention function. Then the transformed cumulative values (y-axis) can be converted to action function scores (upper and lower x-axis):

Trans. cum. value	0.844	−1.041	−0.219	0.727
Att. func. score	0.068	0.000	0.039	0.871

We thank Prof. Dr. R. W. Lent (Albert Einstein College of Medicine, New York) for reviewing the manuscript, and Prof. Dr. J. B. J. Soons (St. Anthonius Hospital, Utrecht) and Prof. Dr. D. L. Massart (Free University of Brussels, Brussels) for encouraging discussions.

REFERENCES

1 Advances in Automated Analysis, Technicon International Congress, 1969, Vol. 3, Biochemical Profiling, Mediad, White Plains, New York, 1970.
2 R. L. Reece and R. K. Hobbie, Am. J. Clin. Pathol., 57 (1972) 644.
3 R. Dybkaer, in R. Gräsbeck and T. Alström (Eds.), Reference Values in Laboratory Medicine, Wiley, New York, 1981, p. 279.
4 C. E. Shannon and W. Weaver, The Mathematical Theory of Communication, University of Illinois Press, Urbana, 1949.
5 G. Jungner and I. Jungner, in E. S. Benson and P. E. Strandjord (Eds.), Multiple Laboratory Screening, Academic Press, New York, 1969, p. 257.
6 W. E. Elion-Gerritzen, Thesis, University of Utrecht, Drukkerij J. H. Pasmans, 's-Gravenhage, 1978, p. 37.
7 D. Wishart, Clustan User Manual, Edinburgh, 3rd edn., in Inter-University Res. Counc. Ser., No. 47, procedure 'DISTIN', 1978.

Analytica Chimica Acta, 150 (1983) 219—231
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

ENHANCEMENT OF PERFORMANCE OF ANALYTICAL LABORATORIES
A Theoretical Approach to Analytical Planning

T. A. H. M. JANSE* and G. KATEMAN

Department of Analytical Chemistry, Faculty of Sciences, University of Nijmegen, Toernooiveld, 6525 ED Nijmegen (The Netherlands)

(Received 6th December 1982)

SUMMARY

The information delivered by an analytical laboratory depends on three factors: the information needed (the aim), the information present in the objects, and the limits on the information obtainable, which are set by the analytical procedures and the means available. In order to reach the specified goals as far as possible, organization is needed, involving many planning decisions. For simplified theoretical models, a theory is developed in which the effect of such decisions on the information yield is computed. In particular, the effect of sample input (related to sampling schemes) and the method of sample handling (allocation of capacity and priorities) is considered in relation to ends such as process control and threshold monitoring.

Most chemical analyses are done on a routine basis, in clinical, environmental or industrial laboratories. The aim of the analysis is to obtain relevant information on a certain object. In this context there are three main items of importance [1] in optimizing the performance of these systems: the purpose of the analysis, the object and the analytical procedure.

First, the need to analyze objects arises from lack of information. This information requirement is often formulated in non-analytical terms by other workers (physicians, politicians, managers or technicians). One of the most difficult tasks for the analytical chemist is to translate the information request into an analytical program; this includes the meaning of the term "relevant" and finding out where information is lacking. Secondly, the object must be considered. The more prior information is available on the object of interest, the better a plan can be developed to gather the missing information. Often an investment of effort in acquiring general information on the object is advantageous and/or necessary. Thirdly, in selecting the analytical method, sample specifications, required accuracy and speed are considered first. Here, limits in obtaining the desired information are set; most analytical research is done in order to reduce these limits. However, often the analytical chemist is confronted with limits of an economical nature, such as equipment and available personnel.

0003-2670/83/$03.00

The mutual relationship of all these features must be considered in planning a system where an analysis provides the maximum amount of relevant information with minimal (or available) means or costs. This asks for knowledge of all three factors, and a plan to interrelate these in order to optimize for the desired information. In this paper, some aspects are developed in order to provide working tools, particularly for chemists concerned with organization and management. However, any chemist using routine analyses or routine laboratories may gain insight in fixing priorities by giving some attention to the planning systems outlined in this paper.

ASPECTS OF INFORMATION THEORY

The three key areas mentioned above can be looked at in terms of fundamental information theory. The average analytical information yield can be quantified with the aid of Shannon's entropy concept [2]: $I = H(\text{before}) - H(\text{after})$, where $H(\text{before})$ is the entropy before analysis, $H(\text{after})$ is the entropy after analysis, and I is the information yield. In this equation $H = -\int f(x)\ \text{ld}\ f(x)\ dx$, where $f(x)$ is the probability distribution of the analytical parameter, and ld denotes dual logarithm. For an analytical parameter having a Gaussian distribution before (uncertainty before) as well as after (uncertainty after) the analysis, it has been shown [3] that

$$I = -\text{ld}\ [Var\ x(\text{before})/Var\ x(\text{after})] \tag{1}$$

where $Var\ x$(after) and $Var\ x$(before) are the variances of x after and before the analysis, respectively.

Consideration of the purpose of the analysis

Three main purposes can be distinguished.

Real-time process reconstruction or control. Dynamic information is needed, and a process reconstruction with a certain minimal reconstruction error is the goal. Here the theory of measurability is applicable [4]. The measurability is defined as

$$m^2 = (Var\ x - Var\ e)/Var\ x \tag{2}$$

where $Var\ x$ and $Var\ e$ are the variances of the uncontrolled process and the reconstruction error, respectively.

With a stationary stochastic process, approximated by a first-order autoregressive A.R. model, the optimal forecast function from the last result into time to the next result is equal to the autocorrelation function:

$$\hat{X}(t + \tau) = X(t)\ \exp(-\tau/Tx) \tag{3}$$

The resulting average reconstruction error, originating from a delay τ after sampling by this forecast function, is

$$Var\ e(\tau) = Var\ x\ [1 - \exp(-2\tau/Tx)] \tag{4}$$

Thus the reconstruction error originating from a sampling interval *Ts*, a dead time *Td* and an analysis error variance *Var a*, expressed in terms of measurability, becomes approximately [4]:

$$m^2 = \exp(-2Td/Tx) \cdot \exp(-Ts/Tx) \cdot [1 - (Var\,a \cdot Ts/Var\,x \cdot Tx)^{1/2}]^2 \quad (5)$$

The measurability is directly related to information yield, as the uncertainty of the process values before analysis is related to *Var x*, and after analysis (or controlling) to *Var e*.

However, it is difficult to compute an information yield per analysis, for the obtained information holds over one forecast period together with the pre-information implied by the forecast function (Eqn. 3). This means that the present information on the object becomes a function of τ: $H(\tau)$ (Eqn. 4). In order to avoid a complex description of this dynamic information behaviour (or entropy flow) the following definition is used: $I = H$(without analysis) $- H$(with analysis).

The information yield of the analysis is expressed as the uncertainty on the object with and without these analyses. The object is not a sample but the process in time itself, and the uncertainty is related to $Var(x)$ without analysis, and to *Var e* after analysis. In this way, again a static picture on a dynamic object is obtained with the resulting analytical information yield:

$$I(\text{dyn}) = -\text{ld}\,(1 - m^2) \quad (6)$$

This dynamic information measure (actually information on a dynamic object) describes not only the effectiveness of the analytical method, but also the way that a sampling scheme functions and the laboratory efficiency, as will be shown. In fact, all factors between the sampling and the actual control action are included.

Threshold monitoring. Again, information on a dynamic process is required. However, a total reconstruction of the process values in time is not of interest. The only relevant information is given by comparison of the process values with a given threshold value. If a first-order A.R. process is analyzed without error and the result is immediately available, then at that moment, there is no uncertainty about the process value. After a while, the process value has changed, and the uncertainty has increased to

$$H = -P(\tau)\,\text{ld}\,P(\tau) - (1 - P(\tau))\,\text{ld}\,(1 - P(\tau)) \quad (7)$$

where $P(\tau)$ is the probability that the process value exceeds the threshold value. When this uncertainty becomes too large, the probability of exceeding the threshold without noticing also becomes intolerable, and a new analysis should be done. Depending on the required reliability, a limit should be set to permissible uncertainties. However, this means that every time this limit is reached, an analysis must be made with an information yield exactly equal to this "uncertainty limit". It has been shown by Muskens [5], that one should sample at a time *Ts* after an analytical result *Xt* according to the equation

$$Ts = -Tx \cdot \ln \left\{ \frac{Tr \cdot Xt + Z\,[Xt^2 - (Tr^2 - Z^2)]^{1/2}}{Xt^2 + Z^2} \right\} \tag{8}$$

where Tr is the threshold value, Xt is the process value at time t (mean zero and unit standard deviation), and Z is the reliability factor.

In this way, optimal use is made of the information given by each analysis. Of course, the resulting distribution of the sampling times is then completely determined by the process values, the threshold unit and the required reliability. Again, an overall information yield can be formulated; it is determined by the probabilities for undetected threshold-crossing when no analyses are done, or with the sampling and measurement scheme.

Static information theory. If some part of the process (e.g., annual means) must be characterized, the desired information can be expressed in acceptable uncertainties concerning the means. Here the theory of sampling-correlated lots is applied, and can be used to formulate the purpose in terms of bits [6].

The object

In all situations mentioned above, some prior information concerning the object is necessary in order to develop a plan. To express this prior information in models is particularly useful and is in fact often done. Modelling is possible over a wide range, extending from a simple Gaussian model to describe a probability distribution, to more advanced modelling as is used in time-series analysis and in very complex models by relations in time, space and components. It should be noted that every relation (usually computed by some correlation analysis) reduces the uncertainty before analysis, and often reduces the analytical information need and yield.

In the above theories concerning the descriptions of the aim, an object model is assumed to be a simple time-series model, the first-order autoregressive model. The entropy before analysis is determined by three parameters: the mean, the variance and a time constant expressing the correlation between two successive process values. In this way, the time-dependent character of "information" can be studied, which is essential in studying laboratory capacity and limits.

The analysis and the laboratory

There are various limitations on the information delivered by an analysis. There are the traditional limits: the detection limit and sensitivity may make it impossible to supply information in the desired region; lack of accuracy and precision may make the uncertainty of the result too large with respect to the uncertainty before analysis; selectivity may be such that unknown matrix effects increase the uncertainty in the result.

However, in dealing with laboratories, limits are more often set by the available equipment and personnel. The main parameters are sampling frequency (how many samples can be done) and measurement time. If the frequency of sampling is too high compared to the measurement time,

queueing of samples occurs. For dynamic information delivery, this means loss of information. The theory of queueing can be applied to study the effects on information yield of such limited facilities. Basically, limited facilities result in a limited number of practicable analytical measurements. If this number is occasionally exceeded, waiting times occur, and a situation may arise in which more samples enter the system than leave it. Besides this quantity, the ways in which a certain number of samples is offered and processed are important. In this respect, organizational aspects can be quantified. In queueing theory, such systems have been extensively studied [7, 8]. Figure 1 shows a representation of one of the simplest queueing models, the M/M/1 model, based on an exponential distribution of the inter-arrival times, as well as an exponential distribution of the analysis times, with one analyst. Such a model could represent a one-analyst laboratory with no organization at all. A perfectly organized system could be modelled by the D/D/1 model (D for deterministic), in which samples arrive at constant intervals, and where every sample takes exactly the same amount of measurement time, with an analyst who is always present. Of course, such models are purely theoretical, but very suitable for studying the effects of limited capacity on the analytical aims. Conclusions related to more realistic situations may be obtained by simulation experiments; such conclusions, pertaining to waiting times, are drawn by Vandeginste [9].

RELATIONSHIPS BETWEEN PURPOSE, OBJECT AND MEASUREMENT

By combination of the theory of measurability with a queueing model, a concrete overall model can be created, in which the effects of management can be studied. In this way, a theory will be developed for planning in laboratories, as an analytical management tool.

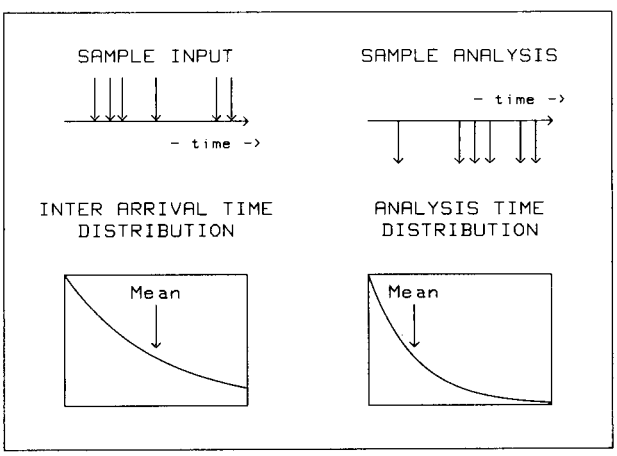


Fig. 1. A representation of a M/M/1 queueing system: stochastic arrival times and stochastic times in finishing jobs.

The basic model is shown in Fig. 2. The way that samples are taken from the process fixes the sample input for the laboratory; the sampling interval time *Ts* equals the inter-arrival time for samples to the laboratory. The mean sampling frequency is λ [λ = 1/mean(*Ts*)]. The laboratory capacity is fixed by one analyst; his average analysis rate is μ(μ = 1/mean(*Ta*)).

As stated above, the absence of a plan can be modelled by a M/M/1 system: the sampling scheme is characterized by a (memory-less) Poisson process, resulting in an exponential distribution for the inter-arrival times. The time required to analyze a sample is also unpredictable and also gives an exponential distribution. As long as the utilization factor (the ratio λ/μ) is below 1, the system functions. However, waiting times occur, and when the utilization factor is increased, the time spent by samples in the system increases exponentially. This "dead time" (waiting + service time) has a negative effect on the process reconstruction, and thus on the information delivered by the chemist.

Investigations of the effects of distribution in the sampling times *Ts* and in the dead times *Td* on the measurability make it possible to relate the information yield directly to the utilization factor. With some approximations (see Appendix), the following general solution is applicable:

$$m^2 = (\lambda\ Tx/2)[(\mathcal{L}(f(Ts)) - 1) \cdot \mathcal{L}(f(Td))]_{s\,=\,2/Tx} \qquad (9)$$

where $\mathcal{L}$ is the Laplace transform, *s* the Laplace operator, *Tx* the time constant of the process, *Ts* the sample interval time, *Td* the time spent in the system (dead time), and λ the mean sample frequency.

The possibility of using Laplace transforms in this equation turns out to be very favourable with respect to systems based on queueing theory; equations for the distribution of the times spent in the system are often obtained in the Laplace domain. In fact, back-transformation to the time domain is often very complicated, if possible at all. For the M/M/1 system, the result is

$$m^2 = [Tx/(Tx + 2\,\bar{T}s)]\ \{(1-\rho)\,Tx/[(1-\rho)Tx + \bar{T}a]\} \qquad (10A)$$

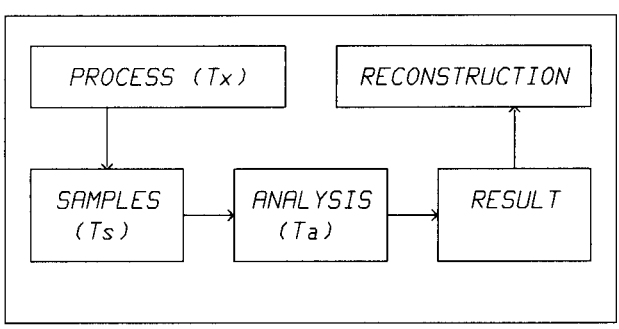


Fig. 2. A model of a sampled and analysed process for the purpose of an on-line process reconstruction.

For the D/M/1 system (fixed arrival rates), the result is

$$m^2 = \exp(-2\,\overline{Ts}/Tx)\{(1-\epsilon)\,Tx/[(1-\epsilon)Tx + 2\,\overline{Ta}]\} \quad (10B)$$

For the M/D/1 system (fixed analysis times), the equation is

$$m^2 = [Tx/(Tx + 2\,\overline{Ts})]\,\{2(1-\rho)\,\overline{Ts}/[(2\,\overline{Ts} - Tx)\exp(2\,\overline{Ta}/Tx) + Tx]\} \quad (10C)$$

In these equations, $\overline{Ta}$ is the mean analysis time, $\overline{Ts}$ the mean sample interval time, and ρ the utilization factor.

In Fig. 3 the information yield (related to m by Eqn. 6) is shown as a function of the utilization factor ρ for the M/M/1 system. As can be seen, there is an optimal mean sampling frequency: however, at the utilization factor 0.5, it is surprisingly low. Only half of the analyst's capacity is used in obtaining the maximal information yield, or in having the best process control by such a system. At lower utilization factors, the information yield decreases because of the lower sampling frequency. At higher utilization factors, the information decreases because of the increasing waiting time in the system. The information-decreasing effect of the waiting times is larger than the information-increasing effect of the higher sampling frequency. This last effect is shown in the same figure by a D/D/1 system (Eqn. 5), in which there are no waiting times. Here, it is obvious that a utilization factor of 1 gives maximal information.

Simulations

In order to verify the above theories and to develop extensions, a simulation program was written in FORTRAN IV, for an IBM 370/158 computer system. This program includes a subroutine package for queueing systems (GASP [10]), and routines for generation and reconstruction of stochastic processes. Eight replicate runs were done over 10 000 standard time units. No advanced techniques for variance reduction were used; some preliminary studies showed no improvements with techniques like antithetic runs, or

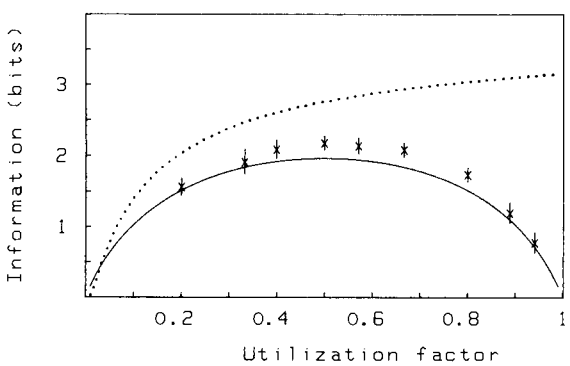


Fig. 3. Theoretical curves for information yield as a function of utilization factor: (——) M/M/1 system; (------) D/D/1 system; (×) simulation results with 95% probability intervals.

with control variates. Common random numbers were used in the various simulation experiments.

Some simulation results are presented in Figs. 3 and 4. All these results show slightly higher values compared with the theoretical curves, mainly because of the discrete character of the simulation (rounding-off favours a proper process reconstruction). Another factor involved is the autocorrelation arising in the waiting times. This factor was neglected in the theoretical treatment. Despite these differences, the figures show a reasonable correspondence between theory and simulation results.

PLANNING

With the theory outlined above, it is possible to develop a theory for analytical planning. First, it is important to distinguish between the influence of a stochastic input (M/D/1 system) and a stochastic processing time (D/M/1 system). The latter is clearly more favourable (compare Figs. 4A and 4B). This is explained as follows: a stochastic input has a direct effect on process reconstruction (stochastic sampling) and an indirect effect through the resulting stochastic dead times. A stochastic processing time gives only the latter. In general, the G/M/1 systems (G stands for general, and so is irrespective of the distribution of the sampling interval times) show an optimal utilization factor of 0.5.

Various plans to improve the performance of the laboratory or to increase the information yield are possible. One conclusion drawn from the above is that smoothing the input is advantageous (from M/M/1 to D/M/1). Of course, this would be no problem if the laboratory itself controls sampling. Otherwise the laboratory could try to establish some other kind of input control in order to assure a steady input, an effort already mostly encountered in practice.

Another possibility for increasing the information yield is to introduce some kind of feedback to the input. Two different feedback schemes are considered (Fig. 5). The first involves feedback from the present capacity and

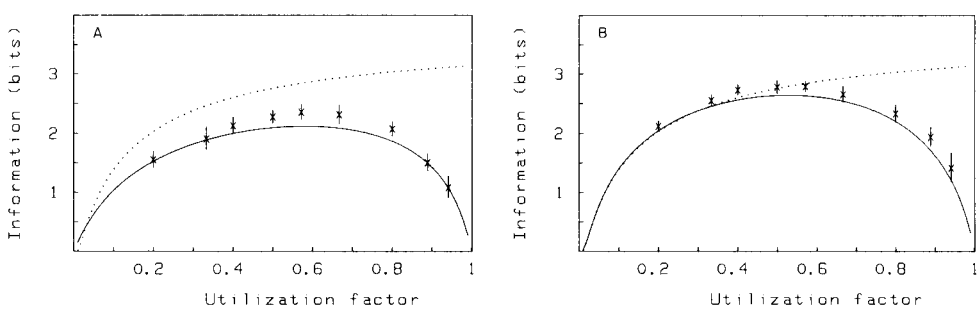


Fig. 4. Theoretical curves for information yield as a function of utilization factor: A, M/D/1 system; B, D/M/1 system; (---) D/D/1 system; (×) simulation results with 95% probability intervals.

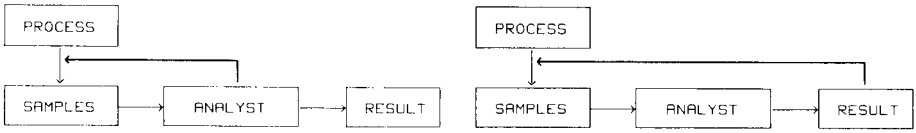


Fig. 5. Schemes influencing sampling by feedback mechanisms. Left: the next sampling is done when the analyst is free. Right: the next sampling is done as computed from the previous result (threshold monitoring).

workload; in the simplest model, this means that every time (and only if) the analyst has finished a job, the next sample is taken. In this way, a complete correlation is created between the inter-arrival times and the analysis times. No queueing occurs and a utilization factor of one can be achieved. The measurability for exponential analysis times is then

$$m^2 = [Tx/(Tx + 2\,\overline{Ta})]^2$$

This means an improvement on the M/M/1 system, but only a slight improvement on the D/M/1 system at a utilization factor of 0.5. The extra effort in trying to obtain as much information as possible is not rewarded; the uncertainty in the analysis times is transmitted to the sampling interval times. With regard to the spare time in the D/M/1 system (for activities with lower priority), a fixed sampling scheme is advantageous, and has the advantage of simplicity.

The second plan involves feedback to sampling by considering the last result, i.e., threshold monitoring. The plan is to analyze as little as possible to obtain a fixed amount of information, and this can be done by the threshold-monitoring scheme. This feedback mechanism is established by considering the difference between the last obtained result and the threshold value. The greater the difference, the more time may pass before the next sample is taken. No queueing occurs when the next time of sampling is computed in this way. The influence of the analysis times is great, however. Table 1 shows an example of the effect on the information yield and the

TABLE 1

The influence of exponentially-distributed analysis times on the dynamic information yield aimed at threshold monitoring. Results are for simulations with a time constant $Tx = 100$, a threshold value of 2, and a reliability factor of 4

Mean Ta	Resulting mean Ts	Utilization factor	Dynamic I (bits)
1	6.2	0.16	0.118
2	8.0	0.25	0.098
3	9.4	0.30	0.070
4	10.6	0.36	0.053
5	11.7	0.40	0.039
7	13.5	0.52	0.025
10	16.5	0.61	0.013

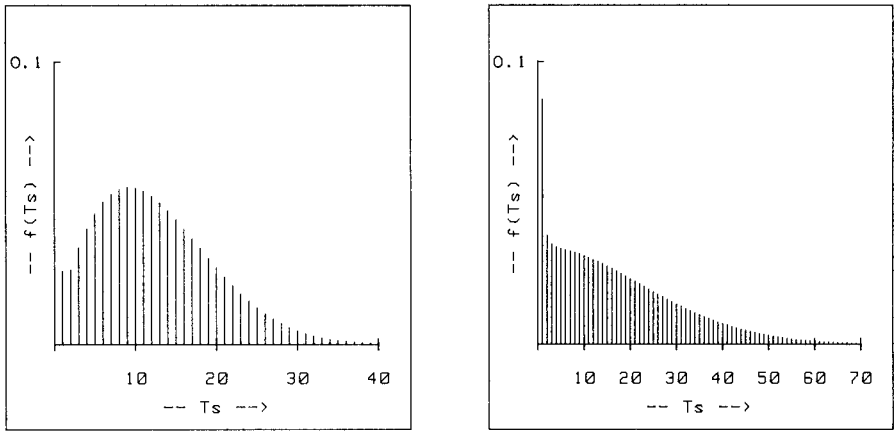


Fig. 6. Resulting inter-arrival time distribution when the sampling scheme for threshold monitoring is applied. The examples are for $Tr = 2$, $Z = 4$, and $Tr = 1$, $Z = 3$.

utilization factor. Both are influenced negatively by larger mean analysis times. This system can work only for relatively low analysis times. Another effect is that such sampling schemes give reliable results only if these samples are given high priority in the laboratory. Other activities must be kept waiting. Because of the unpredictability of the times that these samples enter the laboratory, such a scheme could distort many other activities although these have lower priority. The inter-arrival time distribution generated by threshold monitoring, generally approaches some sort of Erlang distribution (of which the exponential is one), depending on the threshold value, reliability factor and time constant (Eqn. 8.; Fig. 6).

A totally different kind of plan is to fix priorities for the various samples entering the laboratory. Two possibilities are considered. In one, the priorities influence the sequence in which the samples are processed. With respect to process control, it can be shown by simulation that a Last-In-First-Out

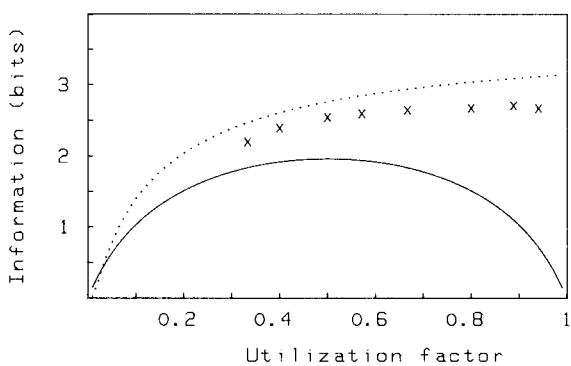


Fig. 7. M/M/1 simulation results: (X) with LIFO priority; (——) the theoretical FIFO curve; (------) the D/D/1 curve.

(LIFO) scheme gives a higher information yield than the standard First-In-First-Out (FIFO) scheme (Fig. 7). This is hardly surprising, as for a first-order A.R. process only the most recent sample gives relevant information. In fact, all the samples that queue could be discarded.

In the second plan, priorities between several processes have to be decided. If the samples of one process are given priority above samples of a second process, this influences both measurabilities. For the first process, Eqn. (10A) is still valid, and the measurability can be computed as if no samples were present from the second process. If priority "with interrupt" (i.e., the analyst should stop his immediate activity and start analyzing the samples) is not given, but priority "without interrupt" is given, then there is a slight decrease in the measurability. As an illustration, some simulation results are presented in Table 2. The total information yield for both process controls is given as a function of the sampling frequency. The results show how the first process is favoured by the priority discipline, and that a LIFO plan for both processes (and no priority for any) gives far better results.

Conclusions

It is shown that computations of information yield with respect to the purpose of an analysis, the objects and their analysis in a laboratory, are possible. Purely mathematical models give insight into the parameters affecting this information yield. Apart from the influence of the utilization factor, it is shown that input description, analysis-time description and priorities can be very significant. Plans can be made to maximize the information gain.

However, when real laboratory organizations are compared with the theoretical models, it is obvious that there is still a gap between theory and practice. This should be filled by further theoretical development or simulation models, with possible interpolations towards practice. Further investigation is particularly useful, even essential, in view of computerization in laboratories, which facilitates the implementation of the schemes developed here.

TABLE 2

Information yield ($I1$ and $I2$) by analyzing aiming for the control of two processes as a function of the utilization-factor and different priority schemes. Results of M/M/1 simulations with a time constant $Tx = 100$ and mean analysis time $\overline{Ta} = 4$

Utilization factor	Dynamic I(bits)		
	FIFO: $I1 + I2$	PRIO: $I1 + I2$ ($I1$)	LIFO: $I1 + I2$
0.94	1.32	2.04 (1.40)	3.33
0.89	2.09	2.59 (1.67)	3.39
0.80	2.54	2.83 (1.73)	3.42
0.67	2.86	2.85 (1.77)	3.31
0.57	3.06	3.10 (1.72)	3.31
0.50	2.65	2.67 (1.38)	2.79
0.40	2.79	2.81 (1.25)	2.84
0.33	2.17	2.17 (1.25)	2.19

APPENDIX

The effect of stochastic sampling times and dead times on the reconstruction error of the process

The mean reconstruction error at a time τ after sampling is given by Eqn. (4). A reconstruction with the analytical result X_i ranges from the time $\tau = Td_i$ to the time $\tau = Ts_i + Td_{i+1}$ when the result of the next sample becomes available. The average reconstruction error over this period is then given by

$$Var\ e\ (Td_i,\ Ts_i + Td_{i+1}) = (1/Ts_i) \int_{Td_i}^{Ts_i + Td_{i+1}} Var\ x \cdot \{1 - \exp(-2\tau/Tx)\}\ d\tau$$

The reconstruction error over a total time T, with N analyses, is

$$Var\ e(T) = (1/T) \sum_{i=1}^{N-1} \int_{Td_i}^{Ts_i + Td_{i+1}} Var\ x \cdot \{1 - \exp(-2\tau/Tx)\}\, d\tau$$

$$= [(N-1)/T] \left\{ \sum_{i=1}^{N-1} \int_{Td_i}^{Ts_i + Td_{i+1}} Var\ x\ [1 - \exp(-2\tau/Tx)]\ d\tau/(N-1) \right\}$$

For $N \to \infty$, this equation can be written in terms of mathematical expectations, resulting in

$$Var\ e = \lambda \cdot \iiint_{-\infty}^{+\infty} f(Td,\ Ts,\ Td') \cdot \int_{Td}^{Ts + Td'} Var\ x\ [1 - \exp(-2\tau/Tx)]\ d\tau \cdot dTd \cdot dTs \cdot dTd'$$

The assumption of independence between Td, Tx and Td' is not entirely correct, because of some autocorrelation that can exist between Td and Td'. However, over a range for ρ, this correlation is low and the following simplification can be made:

$$f(Td,\ Ts,\ Td') \approx f(Td) \cdot f(Ts) \cdot f(Td')$$

Evaluation of the last two equations leads to

$$Var\ e = Var\ x \left\{ (1 + \lambda \cdot Tx/2) \left[\int_{-\infty}^{+\infty} f(Ts) \cdot \exp(-2\ Ts/Tx) dTs \cdot \int_{-\infty}^{+\infty} f(Td') \exp(-2\ Td'/Tx) d\ Td' \right. \right.$$

$$\left. \left. - \int_{-\infty}^{+\infty} f(Td) \cdot \exp(-2\ Td/Tx)\, dTd \right] \right\}$$

With the definition for the Laplace transform in mind, this can be written as

$$Var\ e = Var\ x\ \{(1 + \lambda Tx/2)[\{\mathcal{L}(f(Ts)) - 1\} \cdot \mathcal{L}(f(Td))]_{s = 2/Tx}\}$$

Combined with the definition of the measurability (Eqn. 2) the resulting expression for the measurability is Eqn. (9).

By consideration of the Laplace transforms of the distributions for the inter-arrival times, the analysis times and the resulting waiting times, the solutions for different queueing systems are easily found [7, 8].

The M/M/1 system

Inter-arrival time distribution: $\mathcal{L}[f(Ts)] = \lambda/(\lambda + s)$

Analysis time distribution: $\mathcal{L}[f(Ta)] = \mu/(\mu + s)$

Dead time distribution: $\mathcal{L}[f(Td) = \mu(1 - \rho)/[\mu(1 - \rho) + s]$

Combination with Eqn. (9) gives the following measurability:

$m^2 = [Tx/(Tx + 2\ \overline{Ts})]\{(1-\rho)\ Tx/[(1-\rho)\ Tx + 2\ \overline{Ta}]\}$

The D/M/1 system

Inter-arrival time distribution: $\mathcal{L}[f(Ts)] = \exp(-s/\lambda)$

Analysis time distribution: $\mathcal{L}[f(Ta)] = \mu/(\mu + s)$

Dead time distribution: $\mathcal{L}[f(Td)] = \mu(1-\epsilon)/[\mu(1-\epsilon) + s]$

in which ϵ is computed numerically from $\epsilon = \{\mathcal{L}[f(Ts)]\}_{s=\mu(1-\epsilon)}$ or $\epsilon = \exp[-\mu(1-\epsilon)/\lambda] = \exp[(\epsilon - 1)/\rho]$. This results in

$m^2 = \exp(-2\ \overline{Ts}/Tx)\{(1-\epsilon)\ Tx/[(1-\epsilon)\ Tx + 2\ \overline{Ta}]\}$

The M/D/1 system

Inter-arrival time distribution: $\mathcal{L}[f(Ts)] = \lambda/(\lambda + s)$

Analysis time distribution: $\mathcal{L}[f(Ta)] = \exp(-s/\mu)$

Dead time distribution: $\mathcal{L}[f(Td)] = s(1-\rho)/[(s-\lambda)\exp(s/\mu) + \lambda]$

Combination with Eqn. (9) gives

$m^2 = [Tx/(Tx + 2\ \overline{Ts})]\{2\ (1-\rho)\ \overline{Ts}/[(2\ \overline{Ts} - Tx)\exp(2\ \overline{Ta}/Tx) + Tx]\}$

The D/D/1 system

The inter-arrival time "distribution": $\mathcal{L}[f(Ts)] = \exp(-s/\lambda)$

The analysis time "distribution": $\mathcal{L}[f(Ta)] = \exp(-s/\mu)$

The dead time distribution equals the analysis time distribution. By a Taylor series expansion of $\exp(-2\ \overline{Ta}/Tx)$ the result is the approximate Eqn. (5).

REFERENCES

1 G. Kateman and A. Dijkstra, Fresenius Z. Anal. Chem., 297 (1979) 249.
2 I. Frank, G. Veress and E. Pungor, Hung. Sci. Instrum., 1982 (54) 1.
3 K. Eckschlager, V. Stepanek, Information Theory Applied to Chemical Analysis, Wiley, New York, 1979.
4 F. A. Leemans, Anal. Chem., 43 (1971) 36A.
5 P. J. W. M. Muskens, Anal. Chim. Acta, 103 (1978) 445.
6 G. Kateman, P. J. W. M. Muskens, Anal. Chim. Acta, 103 (1978) 11.
7 L. Kleinrock, Queueing Systems Vol. I: Theory, Wiley, New York, 1976.
8 D. Gross and C. M. Harris, Fundamentals of Queueing Theory, Wiley, New York, 1974.
9 B. G. M. Vandeginste, Anal. Chim. Acta, 112 (1979) 253.
10 J. A. G. M. Kerbosch and R. W. Sierenberg, Discrete Simulatie, Samsom, Alfen aan de Rijn, 1973.

Analytica Chimica Acta, 150 (1983) 233—244
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

ELEMENTAL MAPPING OF TISSUE SECTIONS BY MEANS OF MICRO PARTICLE-INDUCED X-RAY EMISSION SPECTROSCOPY AND COMPUTER GRAPHICS

ULF LINDH

Department of Physical Biology, Gustaf Werner Institute, University of Uppsala, Uppsala (Sweden)

(Received 13th October 1982)

SUMMARY

The elemental distribution within microregions of tissue sections can be revealed by using the nuclear microprobe. The freeze-dried tissue sections are scanned under a 5-μm proton beam and induced x-rays from inner electron shells are continuously recorded along with the actual plane coordinates. The contents of elements heavier than sodium in tissues can be traced in relative amounts as low as 0.5 $\mu g\ g^{-1}$. When energy-dispersive detection of induced x-rays is used, multi-element capacity is achieved. The recorded spectra are stored on tape and processed off-line to produce elemental maps of the tissue section at better than cellular resolution. Although qualitative mapping is economic, quantitative maps can be produced by grey scales or colour graphics.

Sensitive and accurate multi-element methods of analysis have been used by several groups to demonstrate that certain diseases [1] and environmental pollution [2] are accompanied by alterations of trace element levels in the blood. Other tissues have similarly been shown to respond to disease and environmental or occupational pollution [3, 4]. Among the methods which have proved most attractive for research in these areas are atomic absorption spectrometry, neutron activation analysis and various x-ray methods.

Living cells exchange many elements with their environment. Of these, many have been proven to be essential for life. Criteria for essentiality have been established [5]. The list of essential elements is, however, expanding. For an essential trace element, the range of adequate element concentration in the organism is rather narrow. Smaller concentrations result in different anomalies induced by deficiencies which are accompanied by pertinent specific biochemical changes. Higher concentrations become toxic. A typical ratio of toxic level: requirement for domestic animals [6] is only about 50 (80—5000 $\mu g\ g^{-1}$ Fe, 50—2000 $\mu g\ g^{-1}$ Zn, 6—250 $\mu g\ g^{-1}$ Cu, 0.1—5.0 $\mu g\ g^{-1}$ Se). In view of this narrow span between beneficial and toxic concentrations in animals, the establishment of human requirements is especially important. Toxic quantities of essential trace elements for man are known [7] with some certainty only for a few elements such as cobalt (490 mg), fluorine (40 mg)

0003-2670/83/$03.00

and iodine (100 mg). Environmental pollution should thus be considered very seriously.

The bulk of living matter consists of eleven elements of low atomic weight (H, C, N, O, Na, Mg, P, S, Cl, K and Ca). These elements are easy to determine and have thus been known to be essential for life for a long time. The essential trace elements are much more difficult to determine; so far, F, Si, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Se, Mo, Sn and I have been recognised as essential for warm-blooded animals. Most of the essential microelements serve as key components of enzyme systems or of proteins with vital functions. If the metal atom is removed, the protein usually loses its capacity to function as an enzyme. Many pathological disorders arise in animals as a consequence of trace element deficiencies and excesses for which there is, at present, no acceptable explanation in biochemical or enzymatic terms. This suggests either that many element-dependent enzymes of great metabolic significance have not yet been discovered or that these elements participate in the activity of other compounds and in complex interactions in tissues.

Fluorine and silicon are the only essential trace elements with atomic number below 20 discovered so far. Most of the essential trace elements are transition elements; the *d*-orbitals are thus unfilled. Iodine (Z = 53) is the heaviest element known to have any physiological significance. The distribution of trace elements within the periodic table of the elements might have some significance with respect to the development of life in prebiotic times. The need for accurate and sensitive methods to assess elemental abundances in human tissues is thus well established.

Particle-induced x-ray emission (p.i.x.e.) spectroscopy has been widely used for the determination of trace elements since Johansson et al. [8] achieved mass detection limits of the order of picograms. A major problem in quantitative work, however, is associated with sample preparation and handling. Several methods have been devised to prepare thin and thick samples for irradiation with beams of nuclear particles. The techniques and problems associated with the preparation of biological specimens have recently been evaluated [9—11]. Recent years have seen the development of several instruments and techniques for elemental microanalysis; the design of the proton microprobe in 1970 by Cookson et al. [12] was one of the most promising developments. Some thirty nuclear microprobes are now in operation. This development has been reviewed by Cookson [13] and Legge [14].

The objectives of the work reported here were to assess trace element levels, especially for lead resulting from occupational exposure, in tissue sections mainly from compact bone, and to present the results in a visually accessible form. This was accomplished by the use of the nuclear microprobe operated in the p.i.x.e. mode.

EXPERIMENTAL

Tissue material

Autopsy specimens of femur were collected from workers who had been employed more than 10 years in the factory of Rönnskärsverken, Skellefte-hamn, Sweden. The samples (0.5—2 g, wet weight) were cut with a surgical metal saw. To avoid contamination from the cutting procedure, part of the surface was removed with a quartz knife. Very small samples, a few cubic millimetres, of these bone slices ranging from 2 to 5 mm thick were prepared with quartz instruments and freeze-dried at 210 K and 2.7 Pa for 24 h.

As the surface quality is crucial in p.i.x.e. spectrometry, the bone samples had to be prepared prior to examination with the microprobe. The freeze-dried autopsy samples were embedded in a conventional epoxy resin under low pressure (2.7 Pa) and cut in the microtome with glass knives to sections 0.5—5 μm thick (Fig. 1). These sections were transferred either to grids of carbon-coated nylon or kapton foils and stored under vacuum until required.

The nuclear microprobe

The elemental distributions within the samples were examined in the nuclear microprobe at the Studsvik Science Research Laboratory. The microprobe has been in operation since 1975. The principles of construction and performance characteristics were first reported in 1977 [15]. Since then, major improvements have been achieved, permitting the use of a 5-μm probe. The essential idea of the nuclear microprobe is similar to that of the electron microprobe but an ion beam is used. The major advantage of the ion-beam excitation is better sensitivity; improvements by a factor of 100—1000 compared to the electron microprobe are possible. However, the price paid for this improvement is the poorer spatial resolution of the nuclear microprobe.

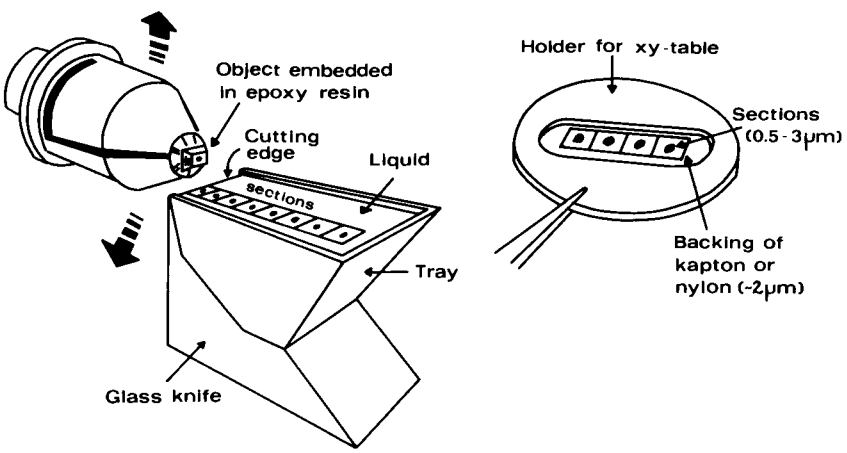


Fig. 1. Illustration of the sample preparation scheme. The sections are cut with glass knives in an ordinary microtome. The support material for the sections is carbon-coated nylon grids or Kapton foils.

The probe consists of four quadrupole magnets, which are required because ion beams are more rigid than electron beams. The attainable beam current for the beam spot used was 0.5 nA. Although this current seems low it can cause serious problems by local heating. The obvious measure to take to avoid such heating would be to decrease the beam current or to scan the beam or the specimen; however, a minimum number of particles is required to provide statistically meaningful measurements in a selected area.

The x-ray emission spectra

Characteristic features of p.i.x.e. spectra (see Fig. 2) include (A) a broad and smoothly varying background (caused primarily by bremsstrahlung radiation of electrons ejected from target atoms by the incident proton beam as well as a small contribution at higher energies from the protons themselves), multiple characteristic x-ray lines identifying elements present in the sample, and interferences between x-ray lines of different elements (B). The vertical scale of the spectrum shown in Fig. 2 represents the square root of the x-ray counts and therefore visually enhances (compared to a linear scale) the regions of low counts as well as the size of the background radiation contribution relative to the peak amplitudes. This effect is by no means as important as when a logarithmic scale is used.

The combined detector efficiency (including the effects of absorbing filters, such as detector window and absorbers used to suppress low-energy parts of the spectrum) and x-ray production cross-sections result in high sensitivity to x-rays with energies between about 1.7 keV and 25 keV. Included in this energy range are the *K*-line radiations (accompanying the filling of vacancies

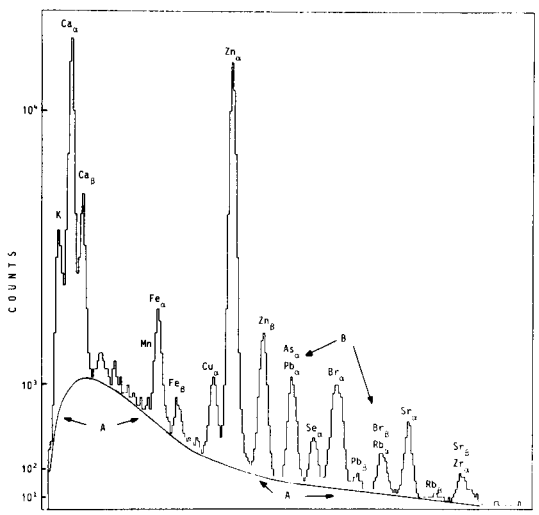


Fig. 2. An x-ray spectrum obtained by irradiation of a crayfish muscle specimen. The horizontal scale (channel number) is proportional to the x-ray energy. A indicates the background and B shows the interferences.

created in the K shell) for elements between silicon ($Z = 13$) and tin ($Z = 50$) and L-line radiations for elements above tin.

Samples of interest in biomedical analysis frequently generate x-ray line spectra of fifteen or more elements simultaneously, giving rise to complex spectra containing more than thirty x-ray lines and complicated by multiple interferences. Estimating the toxicological effects of environmental or occupational exposure may thus present spectra of great complexity. Interelement interferences can occur either as an interference between the K_β signal from element Z and the K_α signal from element $(Z + 1)$ or $(Z + 2)$, or as an interference between a particular K line of a low or medium Z element with an L line of a high Z element. Commonly encountered interferences include the Pb L_α—As K_α interference and the Br K_β—Rb K_α interference (Fig. 2). Fortunately, only a few of the combinations of elements which could be potential interferences are found naturally in sufficient concentrations to be of concern. Assessment of heavy metals from occupational exposure may be an exception. Accurate unfolding of these interelement interferences is a required feature in any spectrum evaluation programme and demands that the K-line and L-line intensity ratios be determined properly.

Data reduction scheme

The evaluation of the x-ray spectra comprises four steps and is an adaptation of a scheme proposed by Nass et al. [16]. The data are smoothed first, to eliminate as much statistical noise as possible. This aids in peak location, as the slope of the data no longer varies widely over small regions. Secondly, the background distribution is estimated and removed from the spectrum. Thirdly, a combination of the correlation technique and an examination of the variation in the slope of the data over a significant range is used to locate the peaks. Finally, identification of the various elements and estimation of the relative proportions of the elements in the sample are done by examining peaks and intensities.

Smoothing. The convolution method basically involves taking a function and sliding it along the spectrum, multiplying at each point the function amplitude and the channel counts and summing [17]. The smoothed function at channel x is

$$s(x) = \sum_{-N}^{+N} G(i)L(x + i)$$

where L is the original datum, G is the function used for smoothing, and $L(j) = 0$ for $j < 1$. For a $2N + 1$ point smoothing function, $G(i) = 0$ for $|i| > N$.

A Gaussian function G was chosen; this convolution will not shift the location of the peaks in the spectrum but will, unfortunately, broaden them. The width at half maximum of the broadened peak will be the sum of the value for the original peak (assumed to be Gaussian in shape, although this is an approximation as the peaks produced by semiconductor detectors tend to

be skew) and that of the Gaussian used for convolution added in quadrature. If a Gaussian of unit area is used for convolution, the height of the data peaks will remain approximately the same, and the background is smoothed without falsely increasing its level. The width chosen for the convoluting Gaussian is the same as that expected for the data peaks, i.e., the resolution. The resolution is of course not a constant parameter throughout the spectrum but is proportional to the square root of the energy.

Background subtraction. After the initial smoothing, the background fit is made by locating several spectrum minima and fitting a line through them (Fig. 2). The minima are located by differentiating the smoothed data. Not all local minima may be used, as some may be due to partially merged peaks. The criterion used to select the appropriate ones was that the minima must increase monotonically up to a certain energy, where the bremsstrahlung shows its maximum, and then decrease monotonically through the rest of the spectrum. Once selected, the fit is made by going back to the unsmoothed data and averaging five points centered about each minimum. At this point, the operator can require the computer to allow for manual selection of the points for the background fit. This is advantageous as the program occasionally produces what appears to be a poor background selection.

Above 4 keV, an excellent fit was possible by using a least-squares method with a fitting function of

$$C = aE^{-2} + bE^{-1} + d$$

where C is the number of counts, E is the x-ray energy, and a, b and d are fitting parameters. For lower energies, the problem is more difficult and the results tend to be less accurate. A third-degree spline fit of the points was used. At about 6 keV, above where the background reaches a maximum, the two background fits are joined and the first derivative is made continuous. The background is then subtracted from the raw spectrum.

Peak location. In the high-energy part, the peaks are usually few in number and well separated. The original spectrum without the background is again smoothed by convoluting as described above. These smoothed data are then differentiated to locate peak positions. Convolution of the raw spectrum is then done once again to establish if a detected peak is single, or a combination of two or more peaks. This convolution is done with a truncated Gaussian which is offset by a specified constant thus forcing its total area to be zero. Hence

$$G(x) = \exp(-x^2/2\sigma^2) - c, \text{ for } |x| < 2\sigma; \; G(x) = 0, \text{ for } |x| > 2\sigma.$$

The constant c must be approximately 0.595 if the area of this function is to be zero. The purpose of this procedure is twofold: first, any slowly changing background still remaining from the subtraction is reduced to insignificant levels; secondly, the action of convoluting is approximately equivalent to taking a second difference of the data. For each peak, there are two zero crossings which correspond to the location of the half-maxima of each peak.

After subtracting the squared width of the convoluting Gaussian, the distance between these zeroes gives an excellent starting value of the widths of the peaks to be found in the fitting routine. This value provides a convenient means to decide whether or not a peak is isolated or in fact the sum of two or more. If this value is significantly larger than that expected for the detector, it is assumed that the peak is in reality the sum of two or more peaks. The f.w.h.m., centroid, and intensity of the isolated peaks are sent to a Gaussian fitting least-squares routine, which is based on that described by Bevington [18]. The fit encompasses all points within one standard deviation of the centroid of the peak. The final fitted peak is then subtracted from the spectrum. Further evaluation of the high-energy end is continued until no single peaks are detected.

Merged peaks are then examined in the high-energy region and also all peaks in the low-energy part. The process described above is repeated twice, but, in this case, the convolution is done with widths of one-half and one-tenth the expected widths of the data to separate merged peaks. Once all peaks have been located, each is examined to establish if it should be fitted separately, or if it is so close to others that it must be fitted as part of a group of peaks. The criterion used was that peak centroids be no closer than three standard deviations, although this condition may be a bit harsh. The widths of the convoluting Gaussians are alternated as $W/2$ and $W/10$, thus enabling the program to pick up closely-packed and single peaks.

Peak identification. The final centroid locations, peak intensities and half-widths are sent to an element identification routine. In order to generate a precise energy calibration, elements known beforehand to be present in the particular sample are identified in the original input data. The x-ray energies of the principal peaks of these elements are found from stored tables and the program centers on the large peaks closest to those in the actual data. A least-squares fit with a second-degree polynomial is done to achieve an absolute energy per channel calibration. Each line in the spectrum is then compared to the stored tables. Peak element identification is verified by cross-checking associated x-rays. Rather than summing over the data, peak areas were measured as described by Nass et al. [16], by using the values already obtained for the intensity and standard deviation. The area is given by $A = I\sigma(2\Pi)^{1/2}$, where A is the area, σ is the standard deviation and I is the intensity (amplitude) of the peak.

Elemental mapping

The specimens were scanned in a pattern like the build-up of a TV picture and the x-ray quanta generated were continuously detected in a semiconductor detector (EDAX, 170-eV resolution at 5.9 keV). Together with the x-ray spectrum were recorded the actual plane coordinates on the specimen. To ensure adequate statistical precision, the same spot on the section had to be passed several times because of the low particle intensity imposed. Acquisition of data and the subsequent processing of elemental maps of the tissue sections were done as described by Legge and Hammond [19].

The x-ray counts registered in a specific channel can be converted to an absolute amount or concentration of an element in several ways. The use of formulae for the x-ray yield involving expressions for stopping power, cross-sections and attenuation coefficients requires not only detailed knowledge of the number of bombarding particles, detection efficiency and detector geometry but also numerical integration of the yield functions. The use of elemental standards allows some of these difficulties to be overcome; monitoring of irradiation conditions and detector geometry is not necessary if standards and samples are irradiated under equal conditions. With a proper choice of standards, even corrections for stopping power and attenuation effects may be neglected [20].

A problem which is particularly pronounced in microprobe work with small structures is the determination of the mass under the probe. Whichever method is selected for yield conversion, the mass must be accounted for. The amount of matrix material being irradiated is proportional to the background radiation [21]; this can be monitored by the choice of a background window in a part of the spectrum where no element is expected. In each window, comprising several channels of the spectrum, where an element can be identified, there are counts from the element and the background under the element. If $C(B)$ denotes the counts in the background window, and $C(E)$ the counts in the element window (i.e., without the background under the element) and $C(b)$ denotes the background counts under the element, then the ratio $k(E) = C(b)/C(B)$ can be used to calculate the fraction of background counts in the element window. The difference $C(E) - C(B)k(E)$ is proportional to the absolute amount of the element and thus the ratio $(C(E) - C(B)k(E))/(k(E)C(B))$ is proportional to the element concentration.

Computer graphics

The procedure described above allows calculation of the concentration of an element present in the sample above the detection limit at arbitrary positions of the sample. The representation of such data in the traditional histogram form is not suitable for more than a few points. In the situation encountered here, other means were thus desirable and computer graphics was employed.

The intensity data from the mapping are transferred to tape, fed into a microcomputer and stored on floppy disks for faster access. The microcomputer (Luxor ABC-800 based on the 8-bit Z80A processor) is then used to produce either coloured or shaded grey maps. A Basic program produces the colour pictures on the high-resolution (240 × 240 pixels) screen. This is a very clear and informative representation but it has the major complication that colour pictures are seldom accepted for publication; and if accepted, large printing costs are involved. A Basic program was therefore designed to translate the intensity or concentration data to shades of grey suitable for reproduction. Starting from the two extremes, black and white, representing 100% and 0% respectively, patterns were designed symmetrically about 50%. To avoid too

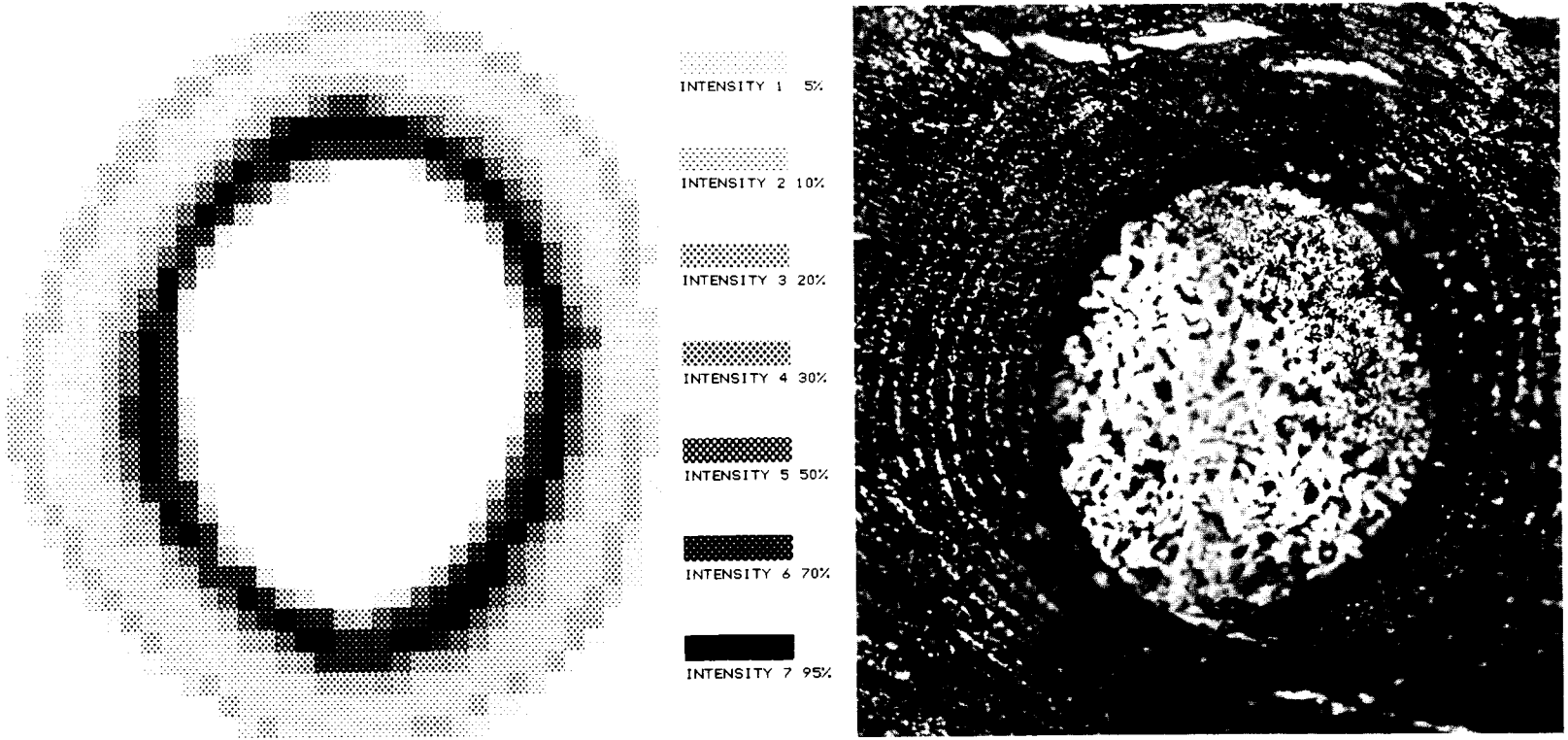


Fig. 3. The mapping of the lead concentration distribution within one osteon together with a micrograph of the section containing the mapped osteon. The intensity scale is also presented.

many levels and for the sake of clarity seven levels were chosen (Fig. 3). When the translation to grey scale is finished, a bit-image addressable printer produces the elemental map of the tissue section.

RESULTS AND DISCUSSION

Because of the kind of production at Rönnskärsverken, emphasis was laid on the determination of concentration profiles of lead within individual osteons of human femur, collected as described above. The main features of the organization of human compact bone tissue are outlined in Fig. 4. One of the osteons selected in a section from the bone sample was thus completely mapped in the microprobe. As the size of this particular osteon was approximately $170 \times 180\ \mu m^2$, this meant the scanning of roughly 1200 points.

The mapping of lead in a section of one osteon is shown in Fig. 3 together with a micrograph of that particular section containing the osteon actually mapped. To translate the intensity to relative amounts of lead, Table 1 can be used. The osteon shown here originated from a worker who had been excessively exposed to lead and had twice suffered from clinical lead poisoning.

Several routines are available for spectrum handling aimed at extracting elemental abundances. The one used here does not differ in philosophy but rather in details. One of the main features is that the program is interactive. Sometimes, features such as background subtraction are better dealt with semi-manually under guidance from an experienced analyst.

The precision and accuracy of the nuclear analytical technique can be assessed on the macro scale by using well-established standards with certified contents of elements. To present a generalized measure of these parameters is not possible because they are likely to vary from element to element and are very sensitive to the kind of matrix in which the elements determined are embedded. Precision and accuracy may be in the range 5%—15% depending on the factors mentioned. Assessment of the accuracy and precision in the use of micro-p.i.x.e. is much more complicated. Unfortunately, there are no good reliable standards from which to make estimates. There is, however, no reason to believe that the micro scale would introduce dramatic changes of the analytical capability.

The use of an alternative scale for presenting x-ray spectra is proposed. A linear scale of the x-ray counts is seldom feasible because of the large span of counts, and the logarithmic scale is normally used. The latter scale,

TABLE 1

Translation of intensity (%) of grey shade to lead concentration ($\mu g\ g^{-1}$, dry weight)

Intensity	5	10	20	30	50	70	95
Pb	5—30	31—45	46—70	71—95	96—130	131—165	166—200

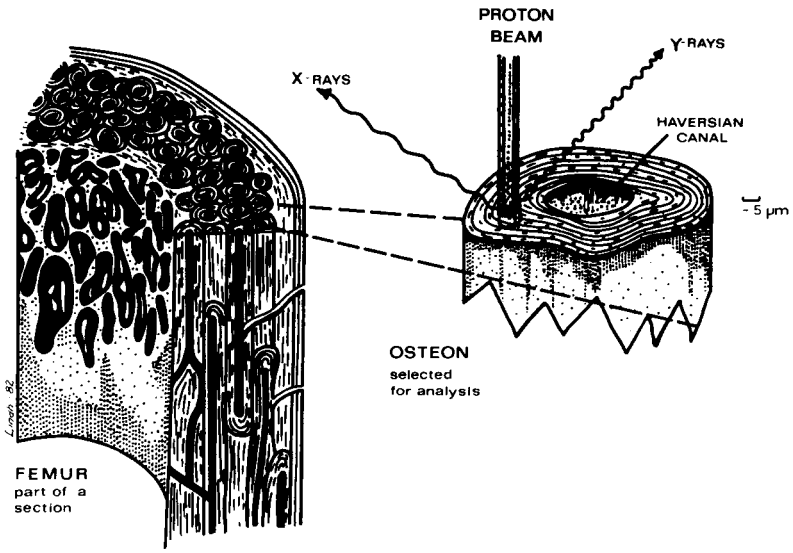


Fig. 4. Schematic drawing showing the essential features of compact bone organization.

however, enhances the background too much and even produces very small peaks which confuse the inexperienced. The square root scale, as used in Fig. 1, appears to be a reasonable compromise between the linear and logarithmic approaches.

This mapping technique is not limited to the kind of tissue dealt with here but can be applied to any soft tissue sample that can be prepared properly for measurements. Soft tissues, however, require specialized treatment differing from that of hard tissues [22]. The careful application of the nuclear microprobe to problems of biomedical origin thus opens a field of great potential.

The efforts of the staff at the Studsvik accelerator laboratory in providing the necessary prerequisites for microprobe work are very much appreciated, as is the skilful preparation of the bone sections by Mrs. L. Näslund. The work was financially supported by the Swedish Work Health Fund.

REFERENCES

1 R. D. Vis, P. M. A. van der Kam and H. Verheul, Nucl. Instrum. Methods, 142 (1977) 159.
2 G. Weber, G. Robaye, J. M. Delbrouck, I. Roelandts, O. Dideberg, P. Bartsch and M. C. de Pauw, Nucl. Instrum. Methods, 168 (1980) 551.
3 N. A. Dyson and A. E. Simpson, Phys. Med. Biol., 21 (1976) 853.
4 D. Brune, G. Nordberg and P. O. Wester, Sci. Total Environ., 16 (1980) 13.
5 E. J. Underwood, Trace Elements in Human and Animal Nutrition, 3rd edn., Academic Press, New York, 1971.
6 W. G. Hoekstra, Ann. N.Y. Acad. Sci., 199 (1972) 182.

7 H. A. Schroeder, The Trace Elements and Man, The Devin-Adair Company, Old Greenwich, CT, 1973.
8 T. B. Johansson, R. Akselsson and S. A. E. Johansson, Nucl. Instrum. Methods, 84 (1970) 141.
9 J. L. Campbell, Nucl. Instrum. Methods, 142 (1977) 263.
10 N. A. Dyson, A. E. Simpson and J. T. Dabek, J. Radioanal. Chem., 46 (1978) 309.
11 B. Meinel, J. Ch. Bode, W. Koenig and F. W. Richter, J. Clin. Chem. Clin. Biochem., 17 (1979) 15.
12 J. A. Cookson, A. T. G. Ferguson and F. D. Pilling, J. Radioanal. Chem., 12 (1972) 39.
13 J. A. Cookson, Nucl. Instrum. Methods, 165 (1979) 477; 181 (1981) 115.
14 G. J. F. Legge, Nucl. Instrum. Methods, 197 (1982) 243.
15 D. Brune, U. Lindh and J. Lorenzen, Nucl. Instrum. Methods, 142 (1977) 51.
16 M. J. Nass, A. Lurio and J. F. Ziegler, Nucl. Instrum. Methods, 154 (1978) 567.
17 H. P. Yule, Anal. Chem., 38 (1966) 103.
18 P. R. Bevington, Data Reduction and Error Analysis for the Physical Sciences, McGraw-Hill, New York, 1969.
19 G. J. F. Legge and I. Hammond, J. Microsc. (Oxford), 117 (1979) 201.
20 U. Lindh, Int. J. Appl. Radiat. Isot., 31 (1980) 737.
21 Y. J. Uemura, Y. Kuno, H. Koyama, T. Yamazaki and P. Kienle, Nucl. Instr. Methods, 153 (1978) 573.
22 U. Lindh, Nucl. Instr. Methods, 197 (1982) 185.

Analytica Chimica Acta, 150 (1983) 245—252
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

OPTIMIZATION OF TEMPERATURE PROGRAMS IN GAS CHROMATOGRAPHY

V. BÁRTŮ* and S. WIČAR

Institute of Analytical Chemistry, Czechoslovak Academy of Sciences, 611 42 Brno (Czechoslovakia)

(Received 17th September 1982)

SUMMARY

Conditions are described for calculating the optimum temperature program for the gas chromatography of any mixture. Calculation of the position of the band maximum and the band width for each component in the temperature-programmed run is based on data obtained by several isothermal runs; the concepts of characteristic temperature and characteristic velocity of a pair of neighbouring peaks in isothermal chromatography are introduced. Calculation of the temperature program is divided into partial trajectories, the number of which is given by the number of difficult-to-separate components in the mixture. An algorithm for the calculation of the optimum temperature program is designed and the individual program blocks of the algorithm are discussed.

The problems of temperature programming in gas chromatography (g.c.) have been dealt with, from different viewpoints, by numerous authors [1]. Relations have been quoted for calculating retention times during a linear temperature rise, and the elution temperatures of components have been discussed. Giddings [2—4] introduced the idea of significant temperature and investigated the possibilities of predicting the optimum parameters in programmed-temperature g.c. from the isothermal characteristics. Knowledge of the properties of the column packing and of the components of the mixture analyzed was needed. Relations have recently been derived in discrete forms for calculating retention times and band widths of the components of the mixture analyzed with any temperature program [5].

Powerful microcomputers are now components of many analytical instruments, serving to monitor and process the data measured. However, these microcomputers do not solve the problem of intelligent automatic setting of operating conditions in replicate analyses of a given mixture, i.e., the problem of optimizing the performance of the instrument according to some preset criterion. In this paper, an algorithm for calculating the optimum temperature program for a gas chromatographic determination is presented; the solution is of general validity, i.e., it can be applied to any mixture that is separable on a given column.

0003-2670/83/$03.00

Isothermal characteristics

An approximate dependence of the retention time, $t_{A,n}$, and peak width at half-height, $s_{A,n}$, on temperature T is used to calculate the optimum temperature program for a mixture. For the nth component, the approximate functions have the form

$$t_{A,n}(T) = A_{tn} \exp(B_{tn}/T) + C_{tn}; \qquad s_{A,n}(T) = A_{sn} \exp(B_{sn}/T) + C_{sn}. \tag{1}$$

The constants in these expressions are obtained from measured values of $t_{R,n}$ and $s_{R,n}$ from at least three isothermal chromatograms of the given mixture [5]. For optimization, mixtures to be processed can be classified into two independent groups. For the first group, $t_{A,z}(T) > \ldots > t_{A,n}(T) > \ldots > t_{A,1}(T)$ is valid within the entire range of working temperatures; for the second group, $t_{A,v}(T) > \ldots \gtrless t_{A,n}(T) \gtrless \ldots > t_{A,1}(T)$ is valid. Thus there may be mixtures of components that can have different mutual positions in the isothermal chromatograms at different temperatures; i.e., there are temperatures T_1 and T_2 at which $t_{A,n}(T_1) > t_{A,n+1}(T_1)$ and $t_{A,n}(T_2) < t_{A,n+1}(T_2)$ (cf. Fig. 1).

Temperature in the column during a temperature program

For temperature control, the column together with the column oven can be considered as a system with first-order delay. The time constant H of the system is calculated from the retention times of the components of the mixture for a temperature program with a single rise [5]. The time course of the

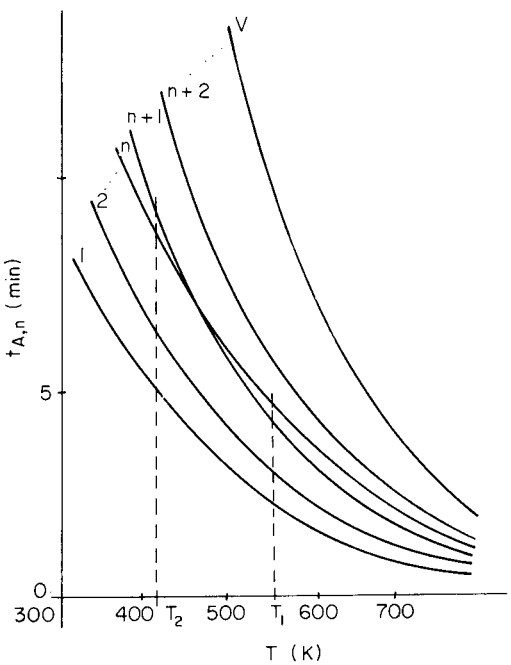


Fig. 1. The dependence of retention time on temperature for the second mixture.

temperature $T(t)$ differs from that set by the temperature program, $T_p(t)$ and is given by the relation

$$T(t) = T_N + (T_N - T_s)[1 - \exp(-t/H)] + D\{t - H[1 - \exp(-t/H)]\} \quad (2)$$

where T_s is the true temperature inside the column at the end of the preceding temperature interval, T_N is the temperature set at the column oven at the beginning of the interval, and D is the rate of temperature rise and/or temperature fall set at the beginning of the interval. The temperature changes inside the column within the given interval depend on the magnitude and sign of $T_N - T_s$ and D.

Calculation of the peak width and retention time

In accordance with expression (1), the band center of the nth component migrates inside the column at temperature $T(t)$, at a velocity

$$v_n(t) = L/[A_{tn} \exp(B_{tn}/T(t)) + C_{tn}] = L/t_{A,n}(T(t)) \quad (3)$$

where L is the column length. At the end of the column, the expression $L = \int_0^{tAP,n} v_n(T(t))\,dt$ is valid for the nth component; here, $t_{Ap,n}$ is the retention time for a temperature program $T_P(t)$. The band width at the column outlet is calculated from

$$S_{AP,n} = s_{AP,n}(T(t))v_n(T(t)) \quad (4)$$

The band width at a distance l from the column inlet is given approximately by

$$S_{AP,n}(l_1) = S_{AP,n} l_1^{1/2} L^{-1/2} = s_{A,n}(T(t))v_n(T(t))l^{1/2}L^{-1/2} \quad (5)$$

If the distance l of the component from the column inlet is expressed as a function of time, $l_t = \int_0^t v_n(T(x))\,dx$, it becomes possible to write for the increment of band width

$$dS_{AP,n}(t) = (s_{A,n}(T(t))v_n(T(t))l^{1/2}(t)L^{-1/2})\,dt \quad (6)$$

If $y_n(t) = l(t)$ is introduced, where $y = 1/t_{A,n}(T(t))$, then $\int_0^{tAP,n} y_n(t)\,dt = 1$, and

$$\begin{aligned} S_{AP,n}(T(t)) = {} & L\{(s_{A,n}(T(t)) - C_{sn}) - (B_{sn}/T^2(t))D\,[1 - \exp(-t/H)]\} \\ & \times \left(\int_0^t y(x)\,dx\right)^{1/2}/t_{A,n}(T(t)) - [s_{A,n}(T(t))\left(\int_0^t y(x)\,dx\right)^{1/2}/t_{A,n}^2(T(t))] \\ & \times (t_{A,n}(T(t)) - C_{tn}) - \{(B_{tn}/T^2(t))D\,[1 - \exp(-t/H)] \\ & + [s_{A,n}(T(t))/2\left(\int_0^t y(x)\,dx\right)^{1/2} t_{A,n}^2(T(t))]\} \end{aligned} \quad (7)$$

Equation (7) and the expression for y constitute a set of two linear differential equations with the initial conditions $l(0)=0, s_{AP,n}(0) = S_{0,n}$ and the boundary condition $l(t_{AP,n}) = \bar{L}$.

Characteristic temperature and velocity

An isothermal chromatogram such as Fig. 2 is considered. For each pair of neighbouring peaks, it is possible to find from expressions (1), a characteristic temperature, T_I such that

$$R_{n,n+1} \approx (t_{A,n+1}(T_{In}) - t_{A,n}(T_{In}))/[1.7(s_{A,n+1}(T_{In}) + s_{A,n}(T_{In}))] = 1 \qquad (8)$$

For an N-component mixture, $(N-1)$ characteristic temperatures T_I can be found. It is possible to write for the velocity at which the centers of the bands of the nth and the $(n+1)$th components migrate at the characteristic temperature in the column:

$$v_{I,n}(T_{In}) \approx v_n \approx v_{n+1} \approx L/t_{A,n}(T_{In}) \approx L/t_{A,n+1}(T_{In}) \qquad (9)$$

The velocity $v_{I,n}(T_{In})$ will be called the characteristic velocity. If the mixture belongs to the second group (cf. [3]), there are two different characteristic temperatures T_I and velocities v_{In1}, v_{In2} for at least one pair of solutes.

OPTIMIZATION OF A TEMPERATURE PROGRAM

A temperature program $T_p(t)$ is called the optimum if the resolution of a pair of successive peaks can be written as

$$R_{1,2} = \ldots = R_{n,n+1} = \ldots R_{N-1,N} = 1 \qquad (10)$$

and, at the same time, $t_{AP,N}(T(t))$ is a minimum. Condition (10) is not often fulfilled, and one has to be satisfied with conditions

$$R_{1,2} \geqslant 1, \ldots, R_{n,n+1} \geqslant 1, \ldots, R_{N-1,N} \geqslant 1 \qquad (11)$$

if all the components of the mixture are of interest. If data on some components (e.g., n, $n+1$, and $n+2$) are not of interest, condition (10) can be modified to

$$R_{1,2} \geqslant 1, \ldots, R_{n,n+1} \gtrless 1, R_{n+1,n+2} \gtrless 1, \ldots R_{N-1,N} \geqslant 1 \qquad (12)$$

i.e., the values of $R_{n,n+1}$ and $R_{n+1,n+2}$ can have any magnitude. Further qualifying conditions of the problem are

$$T_{\min} < T_p(t) < T_{\max} \text{ and } T_{c,\min} < T(t) < T_{c,\max} \qquad (13)$$

where $\langle T_{\max}, T_{\min}\rangle$ is the range of temperature attainable in the column oven and $\langle T_{c,\min}, T_{c,\max}\rangle$ is the range of permissible temperatures for a given column and a given mixture.

Independent variables are T_N and D, for which $T_{\min} < T_N < T_{\max}$ and $D_{\min} < D < D_{\max}$; here, $\langle D_{\min}, D_{\max}\rangle$ is the range of settable temperature increases or decreases in the column oven. Considering an isothermal chromatogram (Fig. 2), the individual pairs have different $R_{n,n+1}$, and it is always possible to find a pair of most-difficult-to-separate components k and $k+1$. For this pair, an increasing temperature program T_t is found such that the mean velocity of components k and $k+1$ equals the characteristic velocity, i.e., for which the following expression is valid:

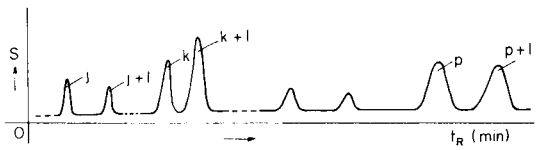


Fig. 2. An isothermal chromatogram.

$$\int_0^{t_{AP,k}} v_k(T(t))\,\mathrm{d}t = \int_0^{t_{AP,k}} v_k(T(In))\mathrm{d}t. \tag{14}$$

Introduction into the temperature program T_t of intervals of decreasing temperature, while the validity of the expression $T_{\min} < T_N < T_{\max}$ is observed, makes it possible to influence the position of components j and $j + 1$, which have retention times shorter than those of components k and $k + 1$. After components k and $k + 1$ have left the column, the next part of the program, for the subsequent pair of the most-difficult-to-separate compounds, p and $p + 1$, is calculated. At the beginning of this part of the program, the bands of the components are at different distances from the column inlet. After components p and $p + 1$ have left the column, the same procedure is applied to the remaining components, etc. The number of partial trajectories is given by the number of difficult-to-separate pairs of components of the mixture, in the direction of increasing retention times. For an N-component mixture, at most $2N - 2$ linear temperature intervals are necessary.

Minimization of the functional values of retention time

Mid-range computers are now equipped with ample libraries of programs for minimization of functions. There are many efficient methods for solving problems such as linear programming, convex programming, or quadratic programming, but efficient algorithms for solving the general problem of minimizing nonlinear functions with nonlinear constraints are few, and the algorithms available seek only local minima. To find global minima, the method can be used several times with different initial estimates of the solution and with introduction of new constraints that would eliminate the local minima already found.

The problem here is to minimize a nonlinear function, $F(x)$, with constraints on the variables (cf. $T_{\min} < T_N < T_{\max}$ and $D_{\min} < D < D_{\max}$)

$l_i < x_i < u_i \qquad (i = 1,\ 2 \ldots n)$

inequality constraints (cf. expressions 11 and 12)

$c_i(x) \geqslant 0 \qquad (i = 1,\ 2 \ldots \mathrm{MIEQ})$

and range constraints (cf. expression 13)

$$v_i \leqslant c_i(x) \leqslant w_i \qquad (i = 1,\ 2 \ldots \text{MRNG})$$

$$\text{M} = \text{MRNG} + \text{MIEQ} \tag{15}$$

Inequality and range constraints are converted within the algorithm to equality constraints by adding new variables. For instance, the constraint $c_i(x) \geqslant 0$ is converted to the equality constraint and single bound,

$$c_i(x) - x_{n+i} = 0; \qquad x_{n+i} > 0$$

and similarly with the constraints (15) cf. [6]. The problem modified in this way can further be converted to a series of bounded and constrained sub-problems. The conversion is based on an augmented Lagrange function defined as

$$L(x, \mu, p) = \mathbf{F}(x) - \sum_{i=1}^{M} \mu_i c_i(x) + p\hat{\mathbf{c}}(x)^T\hat{\mathbf{c}}(x) \tag{16}$$

where the vector μ is an approximation of the Lagrange multiplier, p is a positive number (unknown initially), and $\hat{\mathbf{c}}(x)$ is the vector constraint. The series of sub-problems that has to be solved is

$$k = 1, 2, \ldots \qquad \min[\mathrm{L}(x, \mu^{(k)}, p^{(k)})]$$

with respect to $l_i \leqslant x_i \leqslant u_i$ $(i = 1,\ 2 \ldots n)$. The appropriate routine automatically generates the vectors $\mu^{(k)}$ and $\mathbf{x}$ and also the numbers $\mathbf{p}^{(k)}$ from the initial value input by the user. The routine creates a series that complies with

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha^{(k)}\mathbf{p}^{(k)}$$

where the vector $\mathbf{p}^{(k)}$ is called the direction of search and $\alpha^{(k)}$ is the step length, which is chosen so that $\mathbf{F}(\mathbf{x}^{(k+1)}) < \mathbf{F}(\mathbf{x}^{(k)})$. The step length can be calculated by using some of the methods for one-dimensional optimization. The direction of search can be fixed by using methods for unconstrained minimization. Possibilities are: (1) modified Newton methods, which use $\mathbf{G}^{(k)}$, the Hessian matrix of $\mathbf{F}(x)$, the elements of which are $\partial^2\mathbf{F}(x)/\partial x_i \partial x_j$) at the point $x^{(k)}$ [6]; (2) quasi-Newton methods, which use only the gradient vector $\mathbf{g}^{(x)}$ of $\mathbf{F}(x)$ and creates an approximation to $\mathbf{G}^{(k)}$ by using a recurrent relation; (3) conjugate-gradient methods, in which the direction of search is given by a recurrent relation that uses merely the vector of the first partial differentiation of $\mathbf{F}(x) - \mathbf{g}(x)$.

Structure of the optimization algorithm

A flow chart of the optimization algorithm for temperature programming is shown in Fig. 3. The algorithm consists of five self-contained program blocks. The input data are those obtained from the individual chromatograms for a given mixture (i.e., retention times, peak widths at half-height, peak areas, temperature and/or the course of the temperature program $T_{p,t}$). On the basis of the data obtained in the first isothermal chromatogram at a chosen temperature, program TASAT establishes the number of further

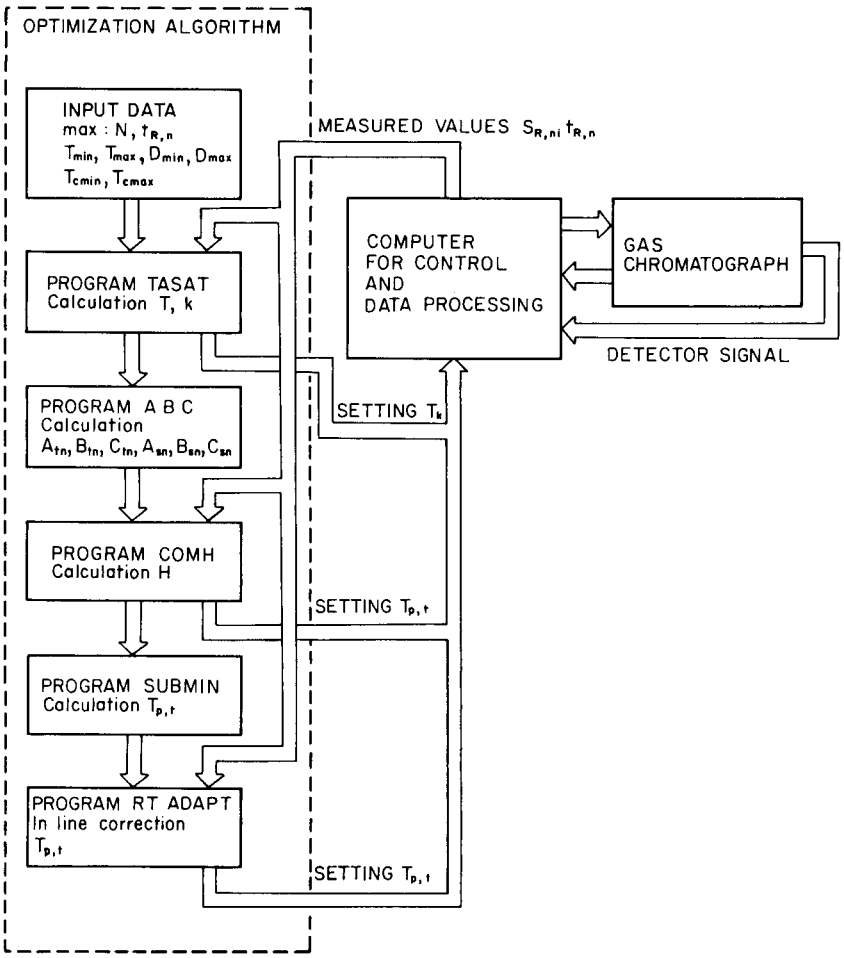


Fig. 3. A block diagram illustrating the information flow in the computer-automated chromatographic system.

isothermal runs and their temperatures and tabulates the data obtained (i.e., the retention times and peak widths for the individual components). The minimum number of isothermal runs is three. Program ABC calculates the approximated functions (1). Program CONH calculates the time constant H of a given column in a given chromatograph on the basis of a run with a single temperature rise; the program first calculates, from the data of programs TASAT and ABC, the initial temperature T and the temperature rise D for this chromatographic run. Program SUBMIN does the minimization proper of the partial temperature program according to the algorithm of the preceding section. Program RT ADAPT requires a test chromatogram at the temperature program $T_{p,t}$ calculated in routine SUBMIN. According to the position of the eluted peaks in the chromatogram, this program corrects

the subsequent parts of the temperature program during the chromatography.

Programs TASAT, ABC, CONH, and SUBMIN are written in Fortran. Program SUBMIN, used for the minimization, was tested on a mid-range mainframe computer ICL 2900, with the use of the NAG Fortran Library Manual (Mk. 8, 1981; ICL). With the above type of problem, the general minimization program is ineffective, because the calculation of the temperature program needs >1 h CPU time. It is therefore better for the calculation of minimization to use a program which will not only calculate the functional value but also choose variables randomly in the direction of the most probable decrease of the minimized function. This program and the RT ADAPT program are under development. The programs are processed off-line except for RT ADAPT, which runs in real time. The amount of calculation required increases roughly proportionately to the number of the components in the mixture.

REFERENCES

1 W. E. Harris and H. W. Habgood, Programmed Temperature Gas Chromatography, Wiley, New York, 1966.
2 J. C. Giddings, J. Chromatogr., 4 (1960) 11.
3 J. C. Giddings, J. Chem. Educ., 39 (1962) 569.
4 J. C. Giddings, 3rd Int. Symp. on Gas Chromatogr., Michigan State University, 1961.
5 V. Bártů, J. Chromatogr., 2 (1983) 255.
6 E. P. Gill and W. Murray, Numerical Methods for Constrained Optimization, Academic Press, London, 1974.

Analytica Chimica Acta, 150 (1983) 253—258
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

Short Communication

AN IMPROVED INJECTION DEVICE FOR QUANTITATIVE CROSS-CORRELATION HIGH-PERFORMANCE LIQUID CHROMATOGRAPHY AT ULTRA-TRACE LEVELS

J. M. LAEVEN, H. C. SMIT* and J. C. KRAAK

Laboratory for Analytical Chemistry, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam (The Netherlands)

(Received 20th October 1982)

Summary. An improved injection device for cross-correlation high-performance liquid chromatography is described. The system is easy to handle and allows rapid sample and eluent changes without significant memory effects. The performance of the injection device in quantitative ultra-trace chromatography is demonstrated by the expansion of the calibration graph for phenol, obtained under common chromatographic conditions, by over two orders of magnitude down to 10 ng l^{-1}. The detection limit for a correlation time of 80 min is 3 ng l^{-1}.

In correlation chromatography [1—5], the single sample injection is replaced by multiple random injections. After cross-correlation of the detector output with the used input function (a pseudo-random binary sequence, PRBS), the resulting correlogram will be equivalent to the chromatogram obtained by single sample injection. The difference is that the signal-to-noise ratio will be considerably improved, so that the detection limit will be much more favourable (Fig. 1). The applicability of the correlation technique has been demonstrated in gas [6, 7] and liquid chromatography [8, 9]. Especially in liquid chromatography, the technique still suffers from experimental imperfections, which have so far hampered its full exploitation in routine analysis. These experimental imperfections are: (1) memory effects caused by adsorption on the inner surface of the sample reservoir; (2) sealing problems; (3) laborious and time-consuming handling to clean and refill the reservoirs. In the present communication, an injection device is described that reduces these drawbacks to a great extent.

Design of the new injection device

It has been shown [4] that distortion of the injection function as well as dynamic phenomena caused by flow changes may increase significantly the baseline noise and peak broadening. Consequently, and also for some practical reasons, the injection device in cross-correlation high-performance liquid chromatography (c.h.p.l.c.) must fulfil the following requirements: First, the injection profile should be reproducible, and the system should be

0003-2670/83/$03.00

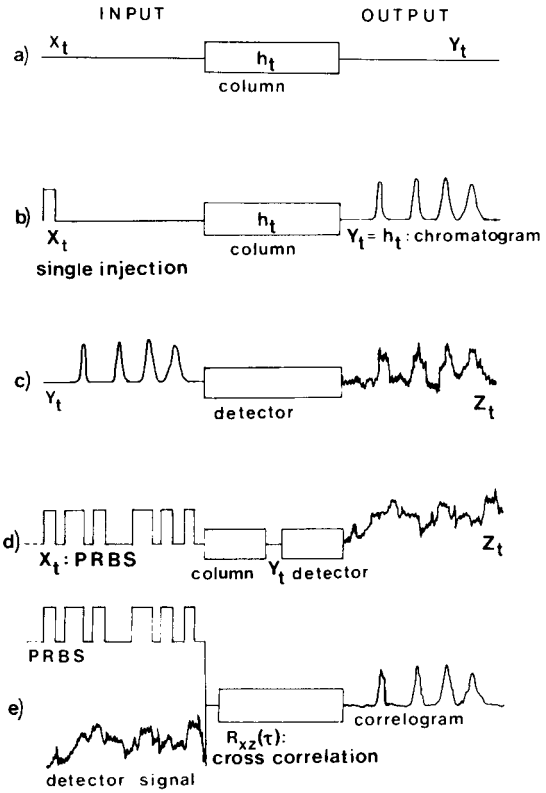


Fig. 1. Correlation chromatography. (a, b) Systems approach to conventional chromatography; an impulse-shaped excitation is applied at the input and the output to be measured is the impulse response (the chromatogram) of the more or less separated compounds. (c) Detection: the column output is measured by the detector, the chromatogram being contaminated by noise added during this step. (d) Correlation chromatography: eluent and sample are injected into the column according to a pseudo-random binary sequence (PRBS); the detector response is the summation of time-shifted chromatograms. (e) Cross-correlation of the detector response with the applied input function (PRBS) yields the chromatogram: the signal-to-noise ratio is considerably improved compared to conventional chromatography.

thermostatted to avoid flow fluctuations caused by temperature changes. Secondly, the materials used should be as inert as possible to minimize adsorption of sample (eluent) constituents, and the device should be resistant to the high pressures needed in h.p.l.c. Thirdly, the sample (eluent) volume capacity must be adequate; this requirement depends on the detection limit demanded by the user, and a large volume was chosen here. Finally, easy and fast reloading of the sample (eluent) reservoir is necessary.

Some of these requirements were fulfilled in the previous injection device [9] and were adopted for the present design. The main difference between the old and new designs is the modification of the sample and eluent reser-

voir. In the original design these were made of stainless steel tubes, which gave rise to serious sealing problems and adsorption phenomena (e.g., memory effects). In the present design, these reservoirs are made of precision glass tubes stored in a pressure-resistant stainless steel container. As the driving liquid pressurizes about equally the outside as well as the inside wall of the glass tube, the system can be applied at the pressures commonly used in h.p.l.c. Because of the smoothness, inertness and precision bore of the glass tubes, no sealing or serious adsorption phenomena occurred.

Experimental

Design of the injection device. The design is shown in Fig. 2. The multiple PTFE ring-sealed movable plunger in the glass tube (Fig. 3) showed no leakage at low or high pressures. This was checked by adding iron(II) to the driving liquid and 2,2'-bipyridine solution to the reservoirs; leakage would be indicated by formation of an intensely red complex. The construction of the plunger allows easy replacement of the PTFE rings. The plunger is fitted with a permanent magnet, which enables the position of the plunger in the glass tube to be determined from outside the steel container by means of a compass or a Hall element. The top of the glass tube is closed by a stainless steel terminator, connected with the outlet capillary. The seal is made by an inert O-ring (Kal-Rez). At the top of the terminator another O-ring is attached to seal the water compartment.

Sample (eluent) exchange is fast with this device: the glass tubes can easily be taken out of the stainless steel containers, and can be replaced in a few minutes by a glass tube containing a new sample (eluent). Thus all the necessary pretreatments of the sample can be done outside the chromatographic system, and the system can be run almost continuously.

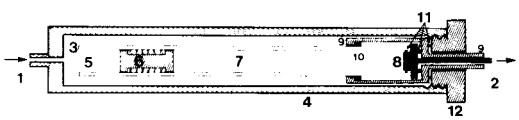


Fig. 2. The sample (eluent) reservoir (simplified). The reservoir has two main parts: the water compartment (5) within the stainless steel outer mantle (4) (63 mm o.d., 40 mm i.d., 85 cm long) up to the plunger (6) (see Fig. 3), and the sample (eluent) compartment (7), a precision glass tube (3) (35 mm o.d., 30 mm i.d., maximum length 65 cm). The water inlet (1) is on the left and the analyte outlet (2) on the right. Both the analyte and the water compartments are sealed by the stainless steel terminator (8) and O-rings (11). The glass tube (3) is attached to a stainless steel holder (9) by a screw-cap (not depicted), which is turned tight in the holder until it stops against a metal ring (10) luted to the glass tube. Exchange of sample (eluent) is done by unscrewing the stainless steel screw (12); the sample (eluent) holder (3 and 9) is pulled out and the glass holder (9) is removed. Finally, the terminator (8) is removed. The glass tube can then be cleaned and refilled. The reverse procedure is then followed and the refilled sample (eluent) holder is replaced. Air bubbles introduced during exchange can escape along the loose O-rings by pumping water into the system (1). After the system has been coupled to the chromatograph, further pumping yields sealing of the O-rings and re-pressurizing of the system.

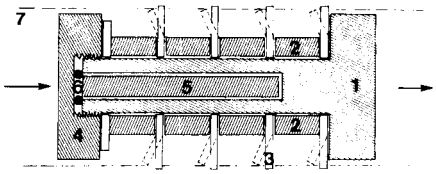


Fig. 3. The plunger: (1) stainless steel shaft; (2) stainless steel rings; (3) PTFE rings with a slightly larger diameter than the glass tube diameter, so that when pressure is applied from the left, the rings seal against the glass tube; (4) stainless steel screw-cap used to seal the replaceable PTFE rings between the steel rings; (5) permanent magnet to trace the position of the plunger; (6) O-ring; (7) glass tube (dashed line).

Both reservoirs are positioned vertically with the water inlet at the bottom, thus permitting gas bubbles, introduced during exchange of the analyte, to escape easily out of the reservoir. The reservoirs are placed in PVC tubes, containing thermostatted water.

The complete c.h.p.l.c. set-up is depicted in Fig. 4.

Performance tests. The performance of the c.h.p.l.c. set-up was tested by measuring a calibration curve of phenol over 5 decades of concentrations: 0.01—100 $\mu g\ l^{-1}$. The injection device was used in combination with conventional h.p.l.c. equipment with fluorimetric detection. The three higher concentrations (1—100 $\mu g\ l^{-1}$) were determined by conventional reversed-phase chromatography and the three lower concentrations (0.01—1 $\mu g\ l^{-1}$) by c.c.h.p.l.c. The chromatographic conditions were as follows: flow rate 1.2 ml min^{-1}; temperature 26°C; detector λ_{ex} = 282 nm, λ_{em} = 296 nm; methanol/water (1:1 v/v) eluent. The retention time for phenol was 223 s. The width of the injections were 2.4 s (48 μl) for c.h.p.l.c. and 1.0 s (20 μl) for h.p.l.c.

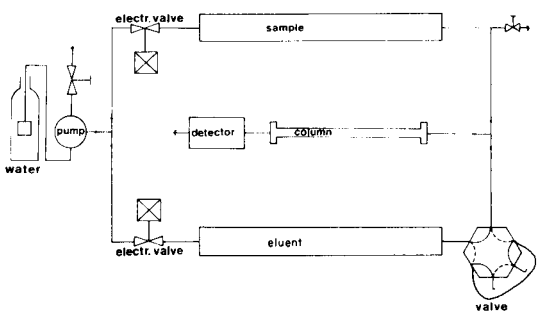
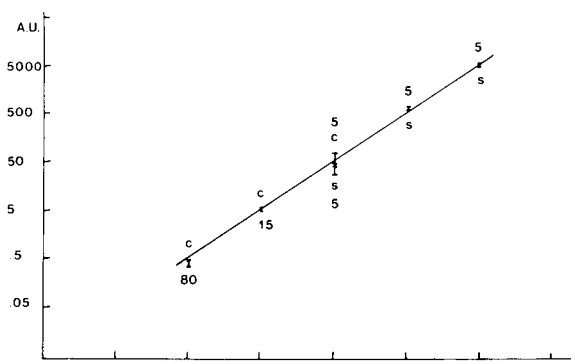


Fig. 4. Complete set-up. The water flow, generated by a constant flow pump, is directed to either the sample or eluent reservoir by two microcomputer-controlled electromagnetic valves, driving the corresponding plunger forward. At the low dead-volume T-joint in front of the column, eluent and sample unit quantities form a pattern according to the input function (PRBS) and enter the column. The detector output is cross-correlated [10] with the PRBS, resulting in the correlogram. At the outlet of the eluent reservoir, a 6-way rotary valve is placed to allow single injection experiments.

Results

The resulting calibration graph is shown in Fig. 5; the bars indicated on the graph represent the peak area $\pm 3\sigma_I$ (arbitrary units), where σ_I is the standard deviation of the integrated noise [11]. Measurements at the 1 μg l^{-1} level were done by conventional chromatography and by c.h.p.l.c. The inner bar represents the cross-correlation results and the outer bar the single-injection results. Under the selected conditions, the detection limit is improved by a factor of about 7. The linear equation for the calibration graph is: $\log y = 1.74 + 1.03 \log x$, where y is the peak area (arbitrary units) and x the concentration (ppb); the $S_a \approx S_b = 0.02$ and the correlation coefficient is 0.9995.

The detection limit for the single-injection experiments, defined as $3\sigma_I$, was about 0.5 μg l^{-1}, whereas from the correlation experiments on 10 ng l^{-1} (80-min correlation time) the detection limit was estimated to be 3 ng l^{-1} (Fig. 6), an improvement by a factor of 170. However, it must be noted that, when the detection limits are compared objectively, the injection volume of a single-injection experiment should be equal to the sample volume injected during one clock-period of the PRBS (the minimal time an injection of sample can last). In this study, the ratio of the injection volumes is 2.4. As the injection volumes in c.h.p.l.c. and h.p.l.c. can be made equal in trace determinations, the detection limits can be compared as described.



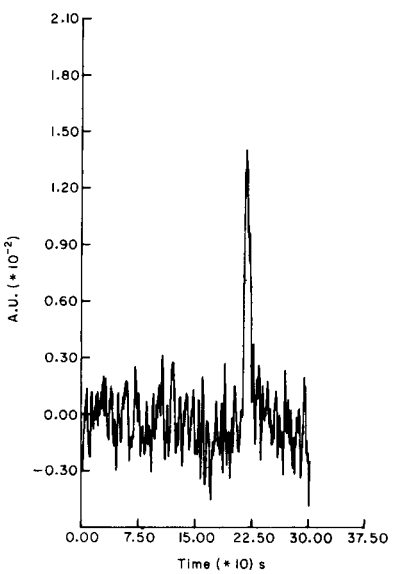


Fig. 5. Calibration graph for phenol with fluorimetric detection for five concentrations of phenol (10^{-4}—10^{-8} g l^{-1}); *s* indicates single injection and *c* indicates correlation. The numbers below the data points indicate the correlation time (min). (ppt = ng l^{-1}.)

Fig. 6. Correlogram of a 10^{-8} g l^{-1} phenol sample. Detection limit for 80-min correlation time is approximately 3 ppt (3×10^{-9} g l^{-1}).

Conclusions

The improved injection device greatly reduces the drawbacks of the earlier devices. The device is easily manageable, sample changing is fast, and memory effects are minimized. It is relatively cheap and can be coupled with conventional chromatographic equipment without problems. Furthermore, the quantitative reliability of c.h.p.l.c. was demonstrated by the expansion of a calibration graph for phenol with by over two decades of concentration compared to conventional chromatography. Peak broadening caused by multiple injections was not observed.

The correlation method is applicable to all forms of liquid chromatography (normal phase, reversed phase, ion exchange, size exclusion). The method offers excellent prospects for ultra trace analysis in cases where preconcentration of the solute fails or, as described by Smit et al. [9], in a differential mode for the determination of trace compounds in the presence of large concentrations of known main compounds with almost similar retention behaviour as the compounds of interest; in this mode the main compounds are added to the eluent in the same concentrations as present in the sample and are thus eliminated from the final chromatogram. Further, the correlation method can be useful for the determination of compounds that cannot be detected with adequate sensitivity under normal h.p.l.c. conditions. Further work is in progress on the application of this technique to the determination of traces of phenol and inorganic ions in river and surface water.

The authors thank Mr. J. K. Clewits, Mr. F. van Eden, Mr. H. Luyten and Mr. K. van de Werff for their valuable contributions.

REFERENCES

1 K. Izawa, K. Furuta, T. Fujiwara and N. Suyama, Ind. Chim. Belg., 32 (1967) 223.
2 H. C. Smit, Chromatographia, 3 (1970) 515.
3 R. Annino and E. Grushka, J. Chromatogr. Sci., 14 (1976) 265.
4 T. T. Lub and H. C. Smit, Anal. Chim. Acta, 112 (1979) 341.
5 M. Kaljurand and E. Küllik, Chromatographia, 11 (1978) 328.
6 M. Kaljurand and E. Küllik, J. Chromatogr., 171 (1979) 243.
7 R. Annino and L. E. Bullock, Anal. Chem., 45 (1973) 1221.
8 T. T. Lub, H. C. Smit and H. Poppe, J. Chromatogr., 149 (1978) 721.
9 H. C. Smit, T. T. Lub and W. J. Vloon, Anal. Chim. Acta, 122 (1980) 267.
10 H. C. Smit, R. P. J. Duursma and H. Steigstra, Anal. Chim. Acta, 133 (1981) 283.
11 R. P. J. Duursma and H. C. Smit, Anal. Chim. Acta, 133 (1981) 67.

Analytica Chimica Acta, 150 (1983) 259—265
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

Short Communication

THE AUTOMATIC OPTIMIZATION OF SEPARATIONS IN HIGH-PERFORMANCE LIQUID CHROMATOGRAPHY

H. J. G. DEBETS*, J. W. WEYLAND and D. A. DOORNBOS

Department of Pharmaceutical and Analytical Chemistry, State University, Ant. Deusinglaan 2, 9713 AW Groningen (The Netherlands)

(Received 24th December 1982)

Summary. Various quality criteria for chromatographic separations have been suggested, but none of them seems to be useful in the automatic optimization of separations by high-performance liquid chromatography. A different optimization strategy is considered: mathematical relations between parameters from the chromatogram and the solvent composition are established. By using one or more of these relations, the optimal composition of the liquid phase can be predicted for particular separations.

During recent years, the optimization of chromatographic separations has been studied thoroughly and researchers have become aware of the fact that a search for a good quality criterion for chromatographic separations is one of the first steps to be made in the optimization of these separations. The essential first step is to understand the basic principles of chromatography. The chromatographic system can be represented as shown in Fig. 1. The input signal $\bar{X}$ represents the sample to be injected into the chromatograph; U represents controllable input parameters such as solvent composition, oven temperature and flow rate. The output signal $\bar{Y}$ represents the chromatogram with chromatographic parameters such as retention times and peak areas. The uncontrollable input parameters Z represent influences on the output signal $\bar{Y}$ (e.g., noise and drift) [1]. If it were possible to describe the chromatographic system in an exact mathematical way, no experimental optimization would be necessary. Unfortunately, the chromatographic process is so complex that attempts to design this mathematical model have not yet been successful [2, 3].

The only things that the researcher knows are empirical rules about the effects of changing the controllable input parameters on the chromatogram. This knowledge may be sufficient for an experienced operator to establish a good separation in a reasonable period of time. However, if the optimization is to be automated, a computer has to judge the quality of the separation and decide which changes have to be made in the input parameters of the chromatographic system. Such computerized judgement is based on quality criteria. A selection of published quality criteria is listed in Table 1.

0003-2670/83/$03.00

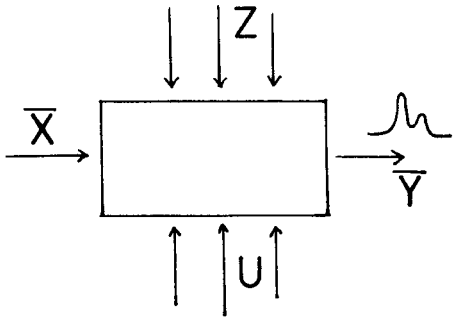


Fig. 1. Schematic representation of the chromatographic system.

Experimental

All the quality criteria listed in Table 1 were tested in simulated hypothetical chromatographic situations as described earlier [12]. The computer used was the Cyber 170/760-computer (CDC) from the Universitair Rekencentrum, State University, Groningen, equipped with several 1200-baud CRT terminals (Beehive Electronics) and an electrostatic plotter (Versatec, V80).

TABLE 1

A selection of quality criteria for chromatographic multicomponent separations

Criterion	Reference
1. Total overlap $\phi = \sum_I \text{Exp}\,(-2R_I)$	Giddings, 1960 [4]
2. Chromatographic response function $CRF = \sum_I \text{Ln}\,(P_I)$	Morgan and Deming, 1975 [5]
3. Chromatographic optimization function $COF = \sum_I \text{Ln}\,(R_I/R_D)$	Glajch et al., 1980 [6]
4. Informing power $P_{INF} = \sum_I {}^2\text{Log}\, S_I$	Massart and Smits, 1974 [7]
5. Separation number $SN = \sum_N {}^2\text{Log}\, P_N$	Spencer and Rogers, 1980 [8]
6. Separation factor $S = \sum_{I<J} (V_J - V_I)/(V_J + V_I)$	Jones and Wellington, 1981 [9]
7. Product R_S Prod. $R_S = \pi_I R_I$	Drouen et al., 1983 [10]
8. Area overlap $A_o = \sum_I (A_{I+1}/A_I)\, Q_{I+1/I}$	Knoll and Midgett, 1982 [11]

All necessary software was written in standard Pascal [13] and standard Fortran IV [14]. The compiler used was the Pascal compiler (ETH Zurich/ University of Minnesota) and the Fortran IV compiler version-4 (CDC, Sunnyvale, CA).

Several scientific subroutines were available from standard libraries. For the second part of the work described here, ordinary least-squares subroutines and non-linear-programming software from the NAG package were used. The Kalman-filter estimates were calculated by using a Kalman-filter algorithm as described by Poulisse [15] especially adjusted for this work.

Results and discussion

As has been argued in previous work [12], most of the quality criteria mentioned are not useful in the automatic optimization of high-performance liquid chromatographic (h.p.l.c.) separations. The most important shortcomings are twofold. First, the number of peaks expected in a chromatogram has to be known before values of the criterion can be calculated correctly. This is best illustrated in Fig. 2, where two of the listed quality criteria are plotted with and without prior knowledge of the number of peaks to be expected. Secondly, the criteria mentioned are not unequivocally related to the quality of separation in a chromatogram, which means that the response surfaces of quality criteria will always show local maxima and/or minima (Fig. 3). This implies that sequential experimental optimization techniques, such as the simplex algorithm, will encounter great difficulties in finding the global optimum (see Table 2). Thus a different approach has to be used in automating the optimization.

Schoenmakers [16] derived the following relation between k' values (capacity factors) and the composition of a binary solvent (θ):

$$\ln k' = A\theta^2 + B\theta + C$$

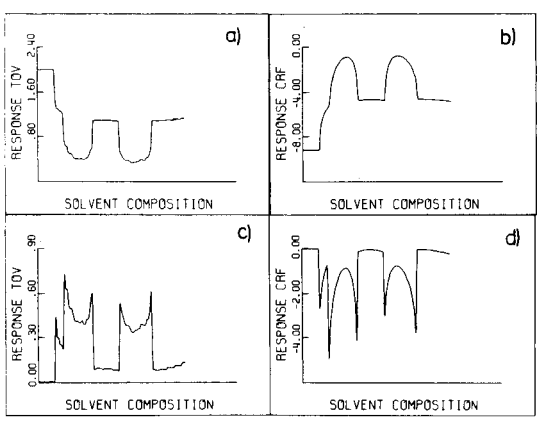


Fig. 2. The response of two quality criteria calculated with (a, b) and without (c, d) prior knowledge of the number of peaks.

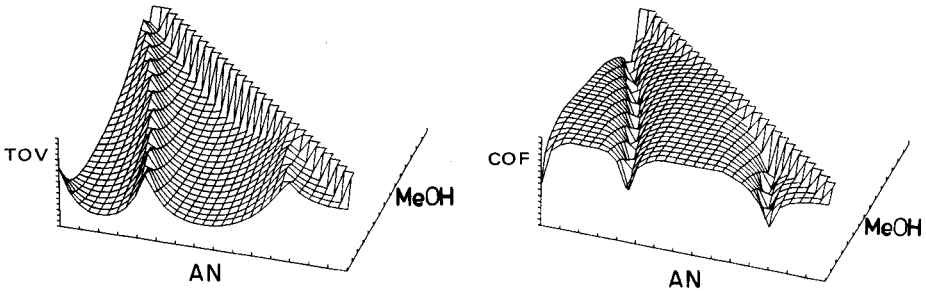


Fig. 3. Response surfaces of the total overlap function and the *COF* function for the separation of five sulfa drugs with mixtures of water, methanol and acetonitrile (AN).

This relation holds for many compounds that are separated by reversed-phase h.p.l.c. and it can easily be transformed to an expression that holds for ternary or even quaternary solvents [16, 17]. According to the expected deviation from linearity of the response surface, a quadratic or special cubic model can be chosen. It should be noted, however, that the number of parameters to be estimated increases rapidly with the complexity of the model. Three ways of estimating the parameters have been considered.

In a simplex lattice design [18], some of the quadratic or higher-order terms are eliminated by using the constraint that the sum of all fractions in a solvent must equal unity (or 100%). By distributing the measurements in a regular way over the factor space, it is possible to calculate the parameters very easily and quickly. Additional measurements can be done in a selected region of the factor space to verify or improve the validity of the chosen model.

When an ordinary least-squares model is used, it is not necessary to distribute the measurements regularly over the factor space, as long as the points at which the measurements are made are chosen so as to avoid singularities in the calculations. The ordinary least-squares method and the simplex lattice design have the advantage of needing only a few measurements to achieve a satisfactory fit.

TABLE 2

Optima found in a response surface during simplex (modified) optimization of a separation of 5 sulfa drugs by using the *COF* function as the criterion[a]

Starting point	End point	*COF*[a]
0.90:0.00:0.10	0.84:0.00:0.16	—0.293
0.90:0.10:0.00	0.95:0.00:0.05	—0.253
0.80:0.10:0.10	0.81:0.04:0.15	—0.312
0.60:0.30:0.10	0.83:0.01:0.16	—0.290

[a]The optimum found by scanning the response surface using a grid of a hundred points was 0.84:0.00:0.16, *COF*: —0.293.

The Kalman-filter estimate needs more measurements but has some important advantages. In this particular problem, the advantage is that the parameters can be estimated for all components at a time. Furthermore, all the known advantages of Kalman-filtering (less computer memory, less computing time, no matrix inversions, etc.) are still valid.

Once the relations between k' values and solvent composition are known, it is possible to predict the solvent composition offering the best minimal α value in a chromatogram; this was described by Laub and Purnell [19]. Scanning the whole factor space by using a grid of a hundred points or more, may yield the global optimum (Table 2), but a somewhat different approach was also tested.

Weyland et al. [20, 21] showed that, just as relations between k' values and solvent composition can be established, it is also possible to fit relations between chromatographic resolution and solvent composition, or between retention times and solvent composition, with satisfactory correlation coefficients. Once these relations are known, a mathematical non-linear programming technique can be used to find the optimum solvent composition. This is done in the following way. The retention time of the last peak is minimized subject to the constraints: (1) $x_1, x_2, x_3 \geqslant 0.0$; (2) $x_1 + x_2 + x_3 = 1$ (or 100%); (3) the chromatographic resolution of two adjacent peaks is larger than a desired value (e.g., 1.25 or 1.50).

In Fig. 4 this procedure is visualized for a selected region of the factor

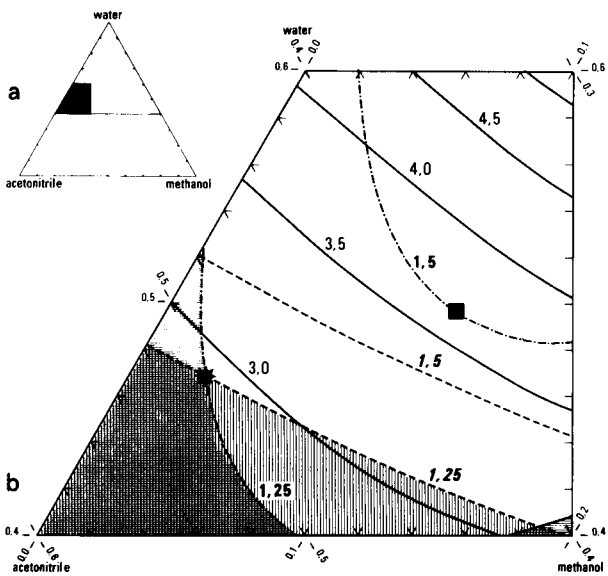


Fig. 4. (a) Accessible part of the factor space (shaded region) and part of the factor space shown in (3b) (black region). (b) Contour plots over the region of the factor space indicated in Fig. (3a): Analysis time (——); resolution between saccharine and caffeine (·····); resolution between caffeine and benzoic acid (-------). The shaded areas are outside the constraints for resolution 1.25. Optimum compositions for minimum resolution 1.25 denoted by (★), and for 1.5 by (■). (from [21]).

TABLE 3

Comparison of experimental results with predictions from the model (non-linear programming) at calculated optimal solvent compositions

	Time (min)	$R_{1,2}$	$R_{2,3}$
Optimal composition calculated 0.470 H_2O, 0.027 CH_3OH, 0.503 CH_3CN			
Measured	2.94	1.18	1.29
Calc./constr.	2.93	1.25	1.25
Optimal composition calculated 0.496 H_2O, 0.108 CH_3OH, 0.396 CH_3CN			
Measured	3.23	1.42	1.80
Calc./constr.	3.63	1.50	1.50

space (between 60% water—40% acetonitrile, 40% water—60% acetonitrile, 40% water—20% methanol—40% acetonitrile, 60% water—30% acetonitrile—10% methanol) for the optimization of a separation of caffeine, saccharin and benzoic acid [21]. The optimal solvent compositions for desired resolutions of 1.25 and 1.50 are listed in Table 3. The results are more or less accurate, depending on the lack of fit of the models used. A severe disadvantage of these methods of fitting relationships between parameters from the chromatogram and the solvent composition is that it is necessary to identify all peaks in the chromatogram. This might be possible by using multichannel detection or by running standards after every change in solvent composition, which means that many chromatograms would have to be run.

Another possibility for identifying peaks in a chromatogram is the use of artificial intelligence. In this approach, all available information from the chromatogram, chromatographic theory, etc., is used to make the computer identify the peaks in a chromatogram with the aid of a logical decision sequence. Research is continuing on that subject in this laboratory.

Conclusions

Neither the quality criteria mentioned nor the methods of determining mathematical relations between parameters from the chromatogram and the controllable input parameters of the chromatograph, are useful unconditionally in the automatic optimization of h.p.l.c. separations. If all peaks can be indentified in a chromatogram, including strongly overlapping peaks, the approach used by Weyland et al. [21] may be successful. Much more attention has to be paid to the possibilities of on-line identification of peaks in a chromatogram.

The authors thank Drs. C. B. M. Didden (D.S.M.) for his work on the adjustments of the Kalman-filter algorithm used.

REFERENCES

1 D. L. Massart, A. Dijkstra and L. Kaufman, Evaluation and Optimization of Laboratory Methods and Analytical Procedures, Elsevier, Amsterdam, 1978, p. 563.
2 J. C. Smit, H. C. Smit and E. M. de Jager, Anal. Chim. Acta, 122 (1980) 1.
3 J. C. Smit, H. C. Smit and E. M. de Jager, Anal. Chim. Acta, 122 (1980) 151.
4 J. C. Giddings, Anal. Chem., 32 (1960) 1707.
5 S. L. Morgan and S. N. Deming, J. Chromatogr., 112 (1975) 267.
6 J. L. Glajch, J. J. Kirkland, K. M. Squire and J. M. Minor, J. Chromatogr., 199 (1980) 57.
7 D. L. Massart and R. Smits, Anal. Chem., 46 (1974) 283.
8 W. A. Spencer and L. B. Rogers, Anal. Chem., 52 (1980) 950.
9 P. Jones and C. A. Wellington, J. Chromatogr., 213 (1981) 357.
10 A. C. J. H. Drouen, et al., Chromatographia, accepted for publication.
11 J. E. Knoll and M. R. Midgett, J. Chromatogr. Sci., 20 (1982) 221.
12 H. J. G. Debets, B. L. Bajema and D. A. Doornbos, Anal. Chim. Acta, 151 (1983) in press.
13 K. Jensen and N. Wirth, Pascal User Manual and Report, Springer, New York, 1978.
14 E. I. Organick and L. P. Meissner, Fortran IV, 2nd edn., Addison-Wesley, Reading, MA, 1974.
15 H. N. J. Poulisse, Anal. Chim. Acta, 112 (1979) 361.
16 P. J. Schoenmakers, Thesis, Technical University, Delft, 1981.
17 G. E. P. Box and K. B. Wilson, J. R. Stat. Soc., Ser. B, 13 (1951) 1.
18 J. W. Gorman and J. E. Hinman, Technometrics, 4 (1962) 463.
19 R. J. Laub and J. H. Purnell, Anal. Chem., 48 (1976) 1720.
20 J. W. Weyland, C. H. P. Bruins, H. J. G. Debets, B. L. Bajema and D. A. Doornbos, Anal. Chim. Acta, submitted.
21 J. W. Weyland, H. Rolink and D. A. Doornbos, J. Chromatogr., 247 (1982) 221.

ANALYTICA CHIMICA ACTA, VOL. 150 (1983)

AUTHOR INDEX

ELECTROPHORESIS

A Survey of Techniques and Applications

edited by Z. DEYL, *Czechoslovak Academy of Sciences, Prague*

JOURNAL OF CHROMATOGRAPHY LIBRARY, 18

PART A: TECHNIQUES

Z. DEYL *(editor)*
F.M. EVERAERTS, Z. PRUSÍK *and*
P.J. SVENDSEN *(co-editors)*

"... provides a sound, state-of-the-art survey of its subject".

– Chemistry in Britain

"... the editors have set out to bring everything together into a coherent whole... they have succeeded remarkably well... the book is bound to be well liked and appreciated by readers".

– Journal of Chromatography

This first part deals with the principles, theory and instrumentation of modern electromigration methods. Both standard procedures and newer developments are discussed and hints are included to help the reader overcome difficulties frequently arising from the lack of suitable equipment. Adequate theoretical background of the individual techniques is given and a theoretical approach to the deteriorative processes is presented to facilitate further development of a particular technique and its application to a special problem. In each chapter practical realisations of different techniques are described and examples are presented to demonstrate the limits of each method.

CONTENTS:
Introduction. Chapters: 1. Theory of electromigration processes *(J. Vacík)*. 2. Classification of electromigration methods *(J. Vacík)*. 3. Evaluation of the results of electrophoretic separations *(J. Vacík)*. 4. Molecular size and shape in electrophoresis *(Z. Deyl)*. 5. Zone electrophoresis (except gel-type techniques and immunoelectrophoresis) *(W. Ostrowski)*. 6. Gel-type techniques *(Z. Hrkal)*. 7. Quantitative immunoelectrophoresis *(P.J. Svendsen)*. 8. Moving boundary electrophoresis in narrow-bore tubes *(F.M. Everaerts and J.L. Beckers)*. 9. Isoelectric focusing *(N. Catsimpoolas)*. 10. Analytical isotachophoresis *(J. Vacík and F.M. Everaerts)*. 11. Continuous flow-through electrophoresis *(Z. Prusík)*. 12. Continuous flow deviation electrophoresis *(A. Kolin)*. 13. Preparative electrophoresis in gel media *(Z. Hrkal)*. 14. Preparative electrophoresis in columns *(P.J. Svendsen)*. 15. Preparative isoelectric focusing *(P. Blanický)*. 16. Preparative isotachophoresis *(P.J. Svendsen)*. 17. Preparative isotachophoresis on the micro scale *(L. Arlinger)*. List of frequently occurring symbols. Subject Index.

1979 xvi + 390 pp. US $83.00/Dfl. 195.00
ISBN 0-444-41721-4

PART B: APPLICATIONS

Z. DEYL *(editor)*
A. CHRAMBACH, F.M. EVERAERTS *and*
Z. PRUSÍK *(co-editors)*

Part B is an exhaustive survey of the present status of the application of electrophoretic techniques to many diverse compounds. Those categories of compounds most suited to these separations, such as proteins and peptides, are dealt with in detail, while the perspectives of the applications of these techniques to other categories of compounds less commonly electrophoresed are given. Special attention is paid to naturally occurring mixtures of compounds and their treatment. This is the first attempt to cover the field on such a broad scale and the book will be valuable to separation chemists, pharmacologists, organic chemists and those involved in biomedical research.

CONTENTS: 1. Alcohols and phenolic compounds *(Z. Deyl)*. 2. Aldehydes and ketones *(Z. Deyl)*. 3. Carbohydrates *(Z. Deyl)*. 4. Carboxylic acids *(F.M. Everaerts)*. 5. Steroids and steroid conjugates *(Z. Deyl)*. 6. Amines *(Z. Deyl)*. 7. Amino acids and their derivatives *(Z. Deyl)*. 8. Peptides and structural analysis of proteins *(Z. Prusik)*. 9. Gel electrophoresis and electrofocusing of proteins *(edited by A. Chrambach)*. Usefulness of second-generation gel electrophoretic tools in protein fractionation *(A. Chrambach)*. Membrane proteins, native *(L.M. Hjelmeland)*. Membrane proteins, denatured *(H. Baumann, D. Doyle)*. Protein membrane receptors *(U. Lang)*. Steroid receptors *(S. Ben-Or)*. Cell surface antigens *(R.A. Reisfeld, M.A. Pellegrino)*. Lysosomal glycosidases and sulphatases *(A.L. Fluharty)*. Heamocyanins *(M. Brenowitz et al.)*. Human haemoglobins *(A.B. Schneider, A.N. Schechter)*. Isoelectric focusing of immunoglobulins *(M.H. Freedman)*. Contractile and cytoskeletal proteins *(P. Rubenstein)*. Proteins of connective tissue *(Z. Deyl, M. Horáková)*. Microtubular proteins *(K.F. Sullivan, L. Wilson)*. Protein hormones *(A.D. Rogol)*. Electrophoresis of plasma proteins: a contemporary clinical approach *(M. Engliš)*. Allergens *(H. Baer, M.C. Anderson)*. 10. Glycoproteins and glycopeptides (affinity electrophoresis) *(T.C. Bøg-Hansen, J. Hau,)*. 11. Lipoproteins *(H. Peeters)*. 12. Lipopolysaccharides *(P.F. Coleman, O. Gabriel)*. 13. Electrophoretic examination of enzymes *(W. Ostrowski)*. 14. Nucleotides, nucleosides, nitrogenous constituents of nucleic acids *(S. Zadražil)*. 15. Nucleic acids *(S. Zadražil)*. 16. Alkaloids *(Z. Deyl)*. 17. Vitamins *(Z. Deyl)*. 18. Antibiotics *(V. Betina)*. 19. Dyes and pigments *(Z. Deyl)*. 20. Inorganic compounds *(F.M. Everaerts, Th. P.E.M. Verheggen)*. Contents of "Electrophoresis, Part A: Techniques". Subject Index. Index of compounds separated.

1982 xiii + 462 pp. US $ 95.75/Dfl. 225.00
ISBN 0-444-42114-9

ELSEVIER
P.O. Box 211, Amsterdam
The Netherlands
52 Vanderbilt Avenue
New York, NY 10017, U.S.A.

7255

(Continued from outside back cover)

0003-2670/83/$03.00

Printed in The Netherlands.

CONTENTS

(Abstracted, Indexed in: Anal. Abstr.; Biol. Abstr.; Chem. Abstr.; Curr. Contents Phys. Chem. Earth Sci.; Life Sci.; Index Med.; Mass Spectrom. Bull.; Sci Citation Index; Excerpta Med.)

International Conference on Chemometrics in Analytical Chemistry, September 15–17, 1982

(Continued on inside back cover)