Vol. 122 No. 4, December 1, 1980
(Computer Techniques and Optimization, Vol. 4 No. 4)
ISSN 0378-4304

ANALYTICA CHIMICA ACTA

International journal devoted to all branches of analytical chemistry

COMPUTER TECHNIQUES AND OPTIMIZATION

ELSEVIER SCIENTIFIC PUBLISHING COMPANY

ANALYTICA CHIMICA ACTA

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

PUBLICATION SCHEDULE FOR 1980 (incorporating the section on Computer Techniques and Optimization).

	J	F	M	A	M	J	J	A	S	O	N	D
Analytica Chimica Acta	113/1 113/2	114	115	116/1	116/2	117	118/1	118/2	119/1	119/2	120	121
Section on Computer Techniques and Optimization			122/1			122/2			122/3			122/4

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

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

Papers from the Americas to: Professor Harry L. Pardue, Department of Chemistry, Purdue University, West Lafayette, IN 47090, U.S.A.

Papers from all other countries to: Dr. A. M. G. Macdonald, Department of Chemistry, The University, P.O. Box 363, Birmingham B15 2TT, England.

For the section on *Computer Techniques and Optimization:* Dr. J. T. Clerc, Universität Bern, Pharmazeutisches Institut, Sahlistrasse 10, CH-3012 Bern, Switzerland.

American authors are recommended to send manuscripts and proofs by INTERNATIONAL AIRMAIL.

Information for Authors. Papers in English, French and German are published. There are no page charges. Manuscripts should conform in layout and style to the papers published in this Volume. Authors should consult Vol. 121, p. 353 for detailed information. Reprints of this information are available from the Editors or from: Elsevier Editorial Services Ltd., Mayfield House, 256 Banbury Road, Oxford OX2 7DE (Great Britain).

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

Advertisements. Advertisement rates are available from the publisher.

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

Publication. *Analytica Chimica Acta* (including the section on *Computer Techniques and Optimization*) appears in 10 volumes in 1980. The subscription for 1980 (Vols. 113–122) is Dfl. 1390.00 plus Dfl. 160.00 (postage) (total approx. U.S. $756.00). The subscription for the *Computer Techniques and Optimization* section only (Vol. 122) is Dfl. 139.00 plus Dfl. 16.00 (postage) (total approx. U.S. $75.50). Journals are sent automatically by airmail to the U.S.A. and Canada at no extra cost and to Japan, Australia and New Zealand for a small additional postal charge. All earlier volumes (Vols. 1–112) except Vols. 23 and 28 are available at Dfl. 153.00 (U.S. $78.50), plus Dfl. 11.00 (U.S. $5.50) postage and handling, per volume.

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

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

Analytica Chimica Acta, 122 (1980) 337—346
Computer Techniques and Optimization

A NEW APPROACH TO BINARY TREE-BASED HEURISTICS

J. ZUPAN
Chemical Institute Boris Kidrič, Ljubljana (Yugoslavia)
(Received 1st February 1980)

SUMMARY

The goals and limitations of some clustering methods are briefly reviewed. In order to avoid the usual methods which are very demanding on computer time and space, a new scheme is proposed for updating (generation) of, and retrieval from, large data bases organized as binary trees. The method is based on calculation of three distances at any given vertex l on the level n, $A(l,n)$ instead of two. The respective distances $d_1(X,A_l)$, $d_2(X,A_r)$, $d_3(A_l,A_r)$ are calculated between the input vector X and the left and right descendant of the vertex $A(l,n)$, A_l and A_r, respectively. The advantage of the method compared to the standard clustering methods or formation of hierarchal trees is that the required memory or computational time is reduced from N^2 to approximately $N \log N$ (N is the number of clustering objects).

The problem of correlation between different properties of objects has been considered in virtually all fields of science. The correlation has been evaluated by different methods ranging from pure statistical correlation analysis [1] to pattern recognition [2, 3]. One of the methods used for this purpose is cluster analysis [4, 5], which became very popular some 15 years ago. The general concepts, and the advantages and limitations of clustering are discussed briefly below in order to explain the new proposed method and its advantages for hierarchal clustering of very large sets of objects.

HIERARCHAL CLUSTERING

The aim of cluster analysis is to group a set of objects into a number of classes (clusters) in such a way that the similarity of the objects within each cluster is greater than their similarity to the objects in other clusters. Each cluster thus represents one or more features of the entire set. A single level of clusters is usually insufficient to describe the entire scheme of features, and so larger clusters on higher levels must be formed from the smaller ones. At the end of such a procedure, only one cluster remains containing all the individual objects.

In order to define precisely the entire procedure of hierarchal clustering two very important definitions must be considered. Firstly, because 'similarity' between objects may be rather hard to define, the inverse quantity, 'distance', between the objects is often preferred. For objects represented as

p-dimensional vectors $X_i(x_{1i}, x_{2i}, \ldots x_{pi})$, the distance between any pair of them in p-dimensional space can easily be calculated and the distance matrix D_N for the entire set of N objects can be written immediately

$$D_N = \begin{vmatrix} 0 & d_{12} & d_{13} \ldots & d_{1N} \\ & 0 & d_{23} \ldots & d_{2N} \\ & & 0 & \ldots\ldots \\ & & & 0 \end{vmatrix} \qquad \begin{aligned} & d_{ij} \geqslant 0, \text{ for all } i,j \\ & d_{ij} = 0, \text{ if and only if } i = j \\ & d_{ij} = d_{ji} \\ & d_{ij} \leqslant d_{ik} + d_{kj} \end{aligned}$$

where d_{ij} is the distance between the ith and jth objects. Any real non-negative function $d(X_i, X_j)$ that satisfies the above four conditions for all objects X from the initial set can be called a distance. Three typical distances are given in Table 1. Secondly, a so-called objective function, according to which the objects (or clusters) can be joined together into larger formations must also be defined. It is evident that the objective function is closely related to the distance function but should not be confused with it: the same distance function can be used to construct many different criteria or object functions. The objective function can be regarded as an algorithm for calculation of the distances between the clusters and then according to some optimal criterion the two most appropriate clusters are fused together. Three different objective functions are listed in Table 2.

After the distance and objective functions have been chosen the procedure for hierarchal clustering is straightforward:
(1) formation of $N \times N$ dimensional matrix $D = \|d_{ij}\|$ for N objects;
(2) search for the minimal element d_{pq} in matrix D;
(3) formation of a new distance matrix D′ from the old D by omitting the qth row and column and replacing the pth row and column by the distances between the new cluster (p,q) and all other objects or clusters;
(4) return to step (2), unless the cluster (p,q) contains all objects from the initial set.

Advantages and drawbacks of clustering

A hierarchal tree for a set of objects is shown schematically in Fig. 1; the objects to be clustered are indicated at the end of each branch as triangles, while the higher points (vertices or knots [10] marked as circles) represent

TABLE 1

Different distance functions. W^{-1} is inverse of the scatter matrix [6]

Name of the distance	Definition
Euclidian	$d(X_i,X_j) = \left[\sum_{k=1}^{p} (x_{ki} - x_{kj})^2 \right]^{1/2}$
Manhattan	$d(X_i,X_j) = \sum_{k=1}^{p} \lvert x_{ki} - x_{kj} \rvert$
Mahalanobis	$d(X_i,X_j) = (X_i - X_j)^T W^{-1} (X_i - X_j)$

TABLE 2

Some common objective functions

Name of the function	Distance between the clusters I and J
Nearest neighbour [7]	$D(I,J) = \min_{i \in I, j \in J} (d_{ij})$
Furthest neighbour [8]	$D(I,J) = \max_{i \in I, j \in J} (d_{ij})$
Group average UPGMA [9]	$D(I,J) = \frac{1}{n_I n_J} \sum_{\substack{i \in I \\ j \in J}} d_{ij}$

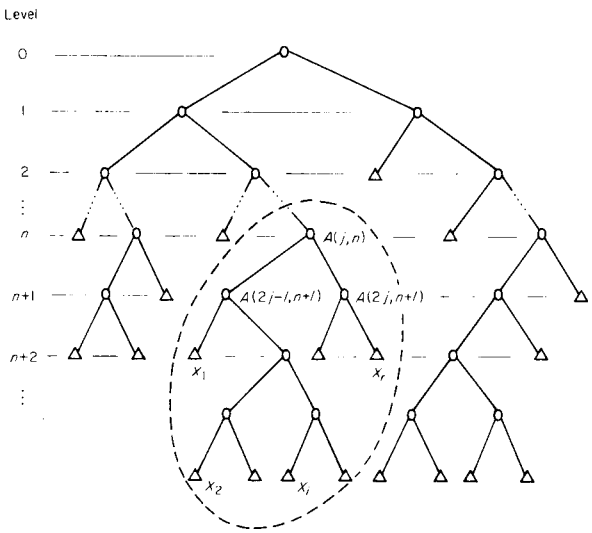


Fig. 1. Binary hierarchal tree with the naming of the vertex representations used throughout the paper. (^) Objects; (○) cluster representatives.

clusters containing at least two objects. Thus at any given level each vertex represents a subtree of objects with one or a few specific features better expressed than in the other clusters. If the level at which the clusters have been joined together is known (from calculation of the object function during the clustering process), it is possible quantitatively to establish the distances not only between single objects but even between any two clusters in the tree.

Quantitative and qualitative comparisons between the subsets of objects have been studied very extensively in many fields including analytical chemistry. There are, for example, classifications of mass spectra [11], hierarchal ordering of infrared spectra [12–14] and correlation of pharmacological activities on the basis of structures [15, 16]. Many strategies for cluster analysis can easily be implemented on most types of computer by using existing high-level programs [17–19].

The most outstanding property of N data items organized for retrieval and/or update as a hierarchal binary tree is that access to any of the desired data is possible in roughly log N steps [20]. As discussed above, the formation of such a tree requires a procedure that needs the space to locate the $N \times N$ distance matrix or the space for storing the vectors representing the data during the entire procedure, in a random-access file which increases the computation time roughly proportionally to N^2. It is evident that even for small data sets of, say, 1000 vectors, the problems cannot be solved economically and some other approximate solution has to be sought.

THE PROPOSED METHOD

A new method for generating hierarchal trees from large data bases, that considerably reduces the computer space and time required, will now be explained and discussed. The method can be applied for clustering objects that can be represented by parameters or some average property.

The final goal of the method is to build a binary tree (Fig. 1) of N objects, each of which is represented by a p-dimensional vector $\mathrm{X}_i(x_{1i}, x_{2i}, \ldots, x_{pi})$. A cluster (the subtree) of r objects $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_r$ is represented in this scheme by a vertex A. At least in principle, components of the vertex A must be obtained by a method that takes into account all components of the vectors $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_r$. For the sake of clarity, each vertex $\mathrm{A}(j,n)$ (jth vertex on the level n) will be taken as a simple unweighted average of the underlying vectors that can be calculated as follows:

$$\mathrm{A}(j,n) = [1/r]\left(\sum_{i=1}^{r} x_{1i}, \sum_{i=1}^{r} x_{2i}, \ldots \sum_{i=1}^{r} x_{pi}\right)$$

where r is the number of vectors $\mathrm{X}_i(x_{1i}, x_{2i}, \ldots, x_{pi})$ in the corresponding cluster.

On its way through the hierarchal tree, a new pattern X must undergo at any vertex $\mathrm{A}(j,n)$ the decision to join either $\mathrm{A}(2j-1, n+1)$ or $\mathrm{A}(2j, n+1)$, its left or right descendant on the level $n+1$ (Figs. 1 and 2). Usually, this is done by comparing two distances

$$d_1 = d[\mathrm{X}, \mathrm{A}(2j-1, n+1)] \tag{1}$$

$$d_2 = d[\mathrm{X}, \mathrm{A}(2j, n+1)] \tag{2}$$

According to the shortest distance $d_{\min} = \min(d_1, d_2)$, the pattern X is moved one level lower to the appropriate point (vertex), and then the procedure is repeated until the bottom of the tree is reached (Fig. 2b), where a new entry is made for a new pattern.

In the scheme proposed here, an additional distance

$$d_3 = d[\mathrm{A}(2j-1, n+1), \mathrm{A}(2j, n+1)] \tag{3}$$

between the two descendants is calculated, and if this happens to be the shortest of the three distances, i.e.

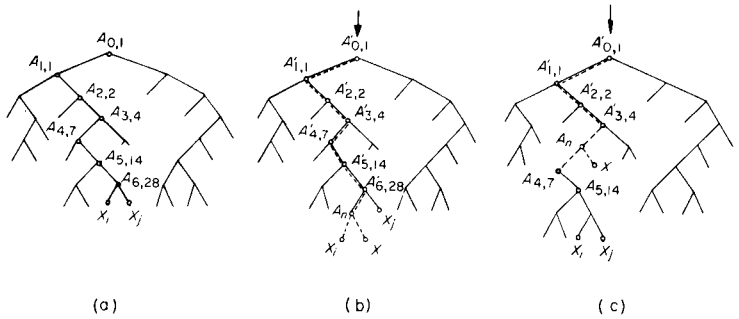


Fig. 2. Comparison of the updated trees after addition of a new object to the old tree (a), by means of the two distance criteria (b), and the proposed tree distance comparison (c).

$$d_3 = \min(d_1, d_2, d_3) \tag{4}$$

the algorithm stops at that vertex. Because it is not logical to join either or both descendants of this vertex, a completely new branch in the tree is created above this vertex. The most significant aspect of this decision is that the new branch is not added at the end of the tree but somewhere in the middle.

For retrieval, such a case means that no exact match for a given query is found in the tree, and that the content of the vertex $A(j,n)$ represents the closest similarity to X that can be found in the tree by this procedure.

For the generation or update of a binary tree, the proposed method means that the new pattern X represents, at level n, a completely different pattern compared to any member of the subtree of the vertex $A(j,n)$. Thus, at this point, the original tree must be disconnected and a new vertex must be inserted in the manner shown in Fig. 2(c). It can be seen that the tree grows as data are added. However, the growth is not bound to the peripheral points but can occur at any point in the tree, similarly to the growth of fractals [21].

THE ALGORITHM

The algorithm for the generation of a binary tree by the method proposed is very simple indeed and is shown in Table 3. To complete the procedure described, two functions must be defined in advance: $d(\mathbf{X}_i, \mathbf{X}_j)$, the distance between the two vectors $\mathbf{X}_i$ and $\mathbf{X}_j$, and $f(A_I, n_I, A_J, n_J)$ the representation of a new cluster $\mathbf{A}_{IJ}$ consisting of two clusters established before and represented by $\mathbf{A}_I$ and $\mathbf{A}_J$, each of which contains n_I and n_J members, respectively.

From the computational point of view, there must be enough space to provide storage for the two 2-dimensional arrays VEC(i,j) and ADD(i,j).For larger trees, however, two random-access files have to be formed instead of the two arrays. In the first array VEC(i = 1,p; j = 1, m), the p-dimensional

TABLE 3

Algorithm A for the clustering of objects represented by vectors X_i. The cluster of objects is represented by VEC(k)

A01. [Initialize] VEC(i,j), ADD(i,j) $\leftarrow$ 0
Set the number of considered vectors $L \leftarrow 1$
Set the number of established vertices $M \leftarrow 1$
A02. [Input first vector] $X_1 \leftarrow$ input, VEC(1) $\leftarrow X_1$
Set the number of vectors in the first cluster:
ADD (3,1) $\leftarrow$ 1
A03. [Input L-th vector] $L \leftarrow L + 1$, $X_L \leftarrow$ input
Stop if there are no more vectors.
A04. [Initialize cluster to be inspected first] $K \leftarrow 1$
A05. [Calculate new cluster] VC $\leftarrow f$(VEC(K), ADD(3,K), X_L,1)
A06. [Find the addresses of both descendants of K-th vertex]
ADLEFT $\leftarrow$ ADD(1,K)
ADRIGHT $\leftarrow$ ADD(2,K)
A07. [Looking for the end of the tree]
if ADLEFT = ADRIGHT = 0, go to step A13.
A08. [Calculate three distances]
$D1 \leftarrow d(X_L$, VEC(ADLEFT))
$D2 \leftarrow d(X_L$, VEC(ADRIGHT))
$D3 \leftarrow d$(VEC(ADLEFT), VEC(ADRIGHT))
A09. [Test of DMIN = min ($D1$, $D2$, $D3$)]
If DMIN = $D1$, set $KN \leftarrow$ ADLEFT, go to step A10.
If DMIN = $D2$, set $KN \leftarrow$ ADRIGHT, go to step A10.
If DMIN = $D3$, go to step A21.
A10. [Store vertex VC on the K-th position] VEC(K) $\leftarrow$ VC
A11. [Increase counter] ADD(3,K) $\leftarrow$ ADD(3,K) + 1
A12. [Next vertex to be inspected] $K \leftarrow KN$, go to step A05.
A13. [Update of the tree: steps A13 to A20]
ADD(3,M + 1) $\leftarrow$ ADD(3,M + 1) + 1
A14. [Store old vertex] VEC(M + 1) $\leftarrow$ VEC(K)
A15. [Store the new vector] VEC(M + 2) $\leftarrow X_L$
A16. [Store new vertex] VEC(K) $\leftarrow$ VC
A17. [Store new addresses]
ADD(1,K) $\leftarrow M + 1$
ADD(2,K) $\leftarrow M + 2$
A18. [Increase counts of cluster members]
ADD(3,K) $\leftarrow$ ADD(3,K) + 1
ADD(3,M + 2) $\leftarrow$ ADD(3,M + 2) + 1
A19. [Increase total count of vertices] $M \leftarrow M + 2$
A20. [Transfer to new vector] Go to step A03.
A21. [Disconnect the tree] Set old attributes of the K-th vertex into new locations:
ADD(1,M + 1) $\leftarrow$ ADD(1,K)
ADD(2,M + 1) $\leftarrow$ ADD(2,K)
ADD(3,M + 1) $\leftarrow$ ADD(3,K)
A22. [Transfer to normal update of the tree] Go to step A14.

vector representations of all vertices A_m ($a_{1m}, a_{2m}, \ldots, a_{pm}$) must be stored. There are $m = (2 * N) - 1$ vertices in the binary tree of N vectors X, and therefore $2N$ records of p words must either be stored or be randomly accessible during the procedure. In the remainder of this paper, the array VEC(i,j) will be indicated only by its second index, as VEC(j), in order to distinguish between the different vertices A_j.

The second array, ADD($i = 1,3; j = 1,m$), serves for storing additional information about each vertex A_j. In ADD(1,j) and ADD(2,j) are stored the addresses of both descendants of vertex A_j, while ADD(3,j) contains the number of vectors represented by A_j. Of course, it is by no means necessary to define the second array ADD: the same items can be stored in an extended array VEC($i = 1, p + 3; j = 1,m$); but for the sake of clarity two arrays are retained in the present description.

DISCUSSION

This discussion is not intended to point out or to analyse the advantages of the proposed method. These advantages are obvious in that the method permits hierarchal clustering of thousands of objects, which is not possible even with 10-times smaller sets by standard methods. Instead, the weak points and limitations of the method are stressed in order to avoid its misuse or a misguided choice of object sets.

First of all, the method is most suitable for large sets of objects (hundreds or even thousands). As shown below, it can be fruitful for small sets as well, but the standard methods are clearly more reliable, considering that the entire distance matrix can be handled in the direct memory. Secondly, the method is much more efficient if the cluster representative A_{IJ} of two combined clusters A_I and A_J can be derived (formed or calculated) directly from both representative vectors A_I and A_J instead of from all n_I and n_J objects in both clusters. An example of a very suitable set is a large infrared spectra collection where the clusters can be represented by the 'average' spectra [12, 22]. Thirdly, it can be seen from the algorithm description that each object is considered only once during the entire procedure; accordingly, the resulting tree is strongly dependent on the sequence in which the objects enter the procedure. The final property, which is not necessarily a disadvantage, is that it is essential to devise a strategy leading to formation of the sequence of given objects that will yield the best possible agreement with the tree obtained by some complete clustering method [5, 7].

In order to clarify these statements, a short example of clustering ten points in 2-dimensional space is shown in detail in Fig. 3. For comparison, the cluster resulting from complete clustering by the unweighted pair–group method using arithmetic averages (UPGMA) [5, 7] of the same set is shown in Fig. 3(a). The case when relation (4) is valid is shown in the fourth tree on Fig. 4 (case d). This situation is illustrated further in Fig. 4(j). The vector X (point 4) that has to be added to the tree (or whose match is sought in the

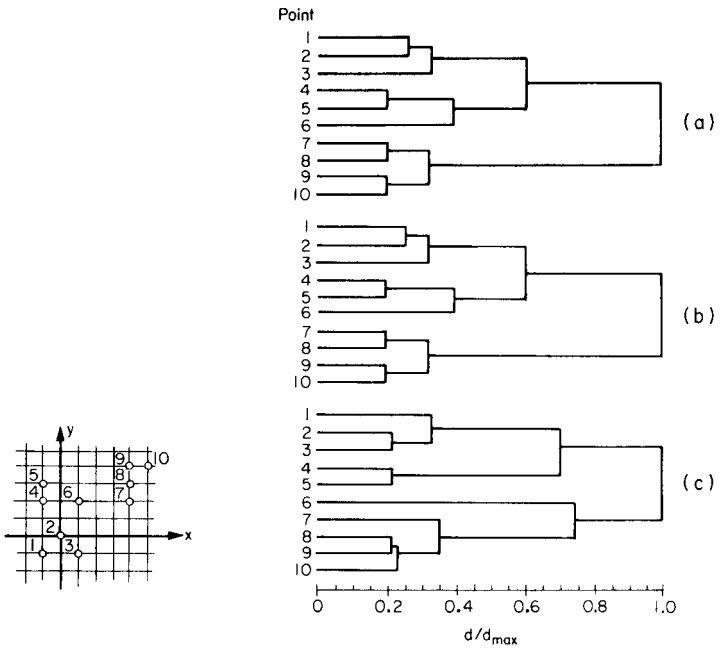


Fig. 3. Clusters of ten points in 2-dimensional space obtained by the complete clustering method UPGMA (a) and by the described procedure using two different sequences of updating the tree. A completely identical tree (b) was obtained by the sequence (1, 2, 7, 8, 10, 4, 5, 3, 6, 9), while the sequence (6, 10, 8, 9, 7, 1, 3, 2, 4, 5) yielded a slightly different tree (c).

case of retrieval) is clearly seen to be outside the range where clustering with the already existing cluster (points 1 and 2) is allowed: the vector **X** at this point is an outlier, so the algorithm inserts a new branch in the tree, and the next point is entered into the process (Fig. 4e).

The most difficult question remaining is how to establish the strategy of queuing the objects needed to obtain perfect clustering. Some general hints, that are indicated by the example shown in Fig. 3, can be outlined as follows. First, the procedure should be started at the edge of the entire cluster (with outliers) rather than somewhere in the middle. Secondly, a small group with few objects should be formed around the first choice and then the next object should be picked up at some middle distance from the first group. Thirdly, if a very distant outlier has to be joined to the tree, an immediate attempt should be made to group around it an appropriate number of similar objects.

These are, of course, only general hints and must be used *cum grano salis*. A study of the reliability of the described method for a large set of infrared spectra will be published in the near future.

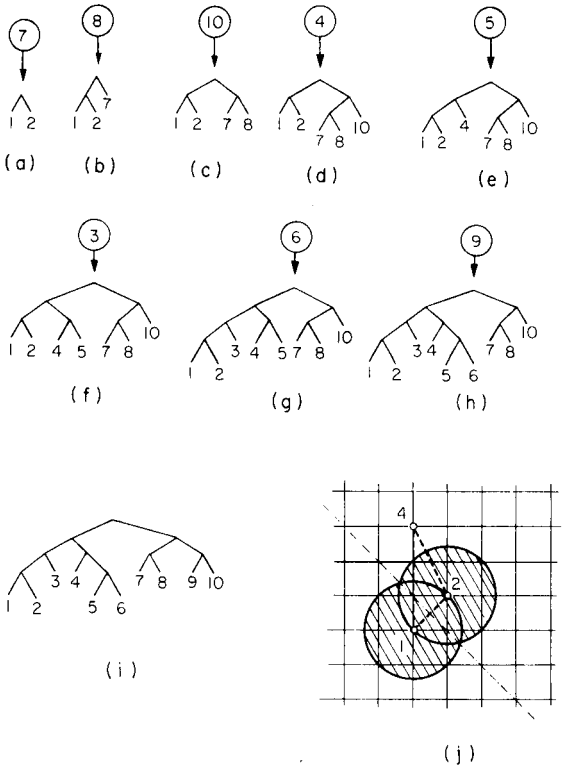


Fig. 4. The sequence of growth steps for the entire updating process as points are added after the first sequence in Fig. 3. The case where a new branch is added in the tree (d) is shown in detail in part (j). The shaded area is the distance zone where an object to be accepted as a new cluster member should lie.

CONCLUSIONS

In contrast to the quadratic growth in space and time with increasing number of objects that is inevitable with standard clustering techniques, the proposed algorithm is comparable with binary tree search algorithms so that growth is proportional to $\log N$ [20].

Despite the enormous saving of computer space and time for clustering of large groups of objects, it should not be forgotten that this proposed method is still only an approximation to total clustering methods where all distances between any pair of objects are considered. Thus the method is heavily dependent on the sequence of objects entering the growing tree. Before complete acceptance of the method can be expected, two points must be resolved: (1) it must be proved that there always exists a sequence of N objects giving a tree identical to that obtained by complete clustering methods, and (2) if this sequence really exists, how can it best be established?

The author is greatly indebted to Professor D. Hadži for many fruitful discussions. Financial support from the Research Community of Slovenia is gratefully acknowledged.

REFERENCES

1 M. Hamburg, Statistical Analysis for Decision Making, Harcourt—Brace, New York, 1970.
2 N. Y. Nilsson, Learning Machines, McGraw-Hill, New York, 1965.
3 P. C. Jurs and T. L. Isenhour, Chemical Applications of Pattern Recognition, J. Wiley, New York, 1975.
4 B. S. Duran and P. L. Odell, in M. Beckmann and H. P. Kuenzi (Eds.), Cluster Analysis, Lecture Notes in Economics and Mathematical Systems, Vol. 100, Springer Verlag, New York, 1970.
5 P. H. A. Sneath and R. R. Sokal, Numerical Taxonomy, W. H. Freeman, San Francisco, 1973.
6 F. J. Rohlf, Syst. Zool., 19 (1970) 58.
7 W. T. Williams and G. N. Lance, Bull. Int. Statist. Inst., 42 (1969) 345.
8 P. MacNaughton-Smith, W. T. Williams, M. B. Dale and L. G. Mocket, Nature, 201 (1964) 426.
9 G. N. Lance and W. T. Williams, Nature, 212 (1966) 218.
10 N. Deo, Graph Theory with Applications to Engineering and Computer Science, Prentice-Hall, Englewood Cliffs, NJ, 1974.
11 W. S. Meisel, M. Jolley, S. R. Heller and G. W. A. Milne, Anal. Chim. Acta, 112 (1979) 407.
12 M. Penca, J. Zupan and D. Hadzi, Anal. Chim. Acta, 95 (1977) 3.
13 F. H. Heite, P. F. Dupois, H. A. van't Klooster and A. Dijkstra, Anal. Chim. Acta, 103 (1978) 313.
14 F. W. Pijpers, H. L. M. van Gaal and J. G. M. van der Linden, Anal. Chim. Acta, 112 (1979) 199.
15 K. C. Chu, Anal. Chem., 46 (1974) 1181.
16 Y. Takahashi, Y. Miyashita, Y. Yotusi, H. Abe, M. Sano and S. Sasaki, Anal. Chim. Acta, 122 (1980) 241.
17 D. L. Duewer, A. M. Harper, J. R. Koskinen, J. L. Fasching and B. R. Kowalski, ARTHUR, Version 3-7-77.
18 D. J. McRae, MIKCA Program, Behavioral Science, 16 (1971) 423.
19 J. Zupan, SOKAL Program, KIBK, Ljubljana, 1975.
20 D. E. Knuth, The Art of Computer Programming, Sorting and Searching, Vol. 3, Addison—Wesley, Reading, MA, 1975, p. 499.
21 B. B. Mendelbrot, Fractals, Form, Chance, and Dimension, W. H. Freeman, San Francisco, 1977, p. 66.
22 M. Penca, Doctoral Thesis, Ljubljana University (1980).

Analytica Chimica Acta, 122 (1980) 347—355
Computer Techniques and Optimization

AN OPERATIONAL RESEARCH MODEL FOR PATTERN RECOGNITION

D. L. MASSART*, L. KAUFMAN and D. COOMANS

Farmaceutisch Instituut, Vrije Universiteit Brussel, Laarbeeklaan 103, 1090 Brussels (Belgium)

(Received 14th January 1980)

SUMMARY

The operational research model is based on a facility location problem and is solved by using heuristic and branch-and-bound methods. The method is particularly useful for clustering because it contains an algorithm that permits a conclusion about the significance of a cluster, without imposing a priori conditions. The method is also applied to supervised learning, for which it is not expected to be better than existing methods. However, it could be an interesting aid to those methods because it allows reduction of large data sets. The application of the method is illustrated with a few examples.

Pattern recognition is one of the more active fields in analytical chemistry, and supervised learning techniques have received much more attention than unsupervised techniques. In recent years, several operational research models that could be useful in this context have been studied [1]. One of these, a facility location model which was introduced in analytical chemistry for other purposes [2], seemed sufficiently interesting to be investigated in more detail for clustering purposes.

THE LOCATION MODEL

This model was applied with success to gas chromatography (g.c.) problems [2]. In g.c., stationary phases are characterized by indices, such as the Rohrschneider index [3], which are calculated by using the retention index of a certain number of probe compounds, i.e. substances with a retention behaviour characteristic of a larger number of substances that can be chromatographed on g.c. stationary phases. The set of probes selected is then considered to be as representative as possible of all solutes which could be chromatographed under analogous conditions. For the Rohrschneider index, the probe set is composed of ethanol, methyl ethyl ketone, nitromethane, pyridine and benzene. In general, this selection is carried out on data sets containing the retention index of n compounds on d g.c. phases, i.e. a data set containing n d-dimensional pattern vectors. The selection of optimum probe sets has been a regularly recurring topic of interest [4—6]. The problem can be restated as follows: how it is possible to choose the best number of probes representative of a given set? This problem can be solved by using a facility

location model. The operational research problem can be stated as follows: for a finite number of users, whose demands for a given service are known and must be fulfilled, and a finite set of possible locations where a given number p of service centres may be located, select the locations of the service centres in order to minimize the sum of transportation costs of the users.

Consider Fig. 1, which is a map of 15 villages, in which three facilities must be located in such a way that they satisfy the conditions given above. If the villages are the users then, clearly, one should select B, G and M. By noting which villages are nearest to each of these three centres, one can make three groups of villages (A—D, E—J and K—O). The solution of this problem is obtained in two steps. First, one determines an approximately optimum solution by using an heuristic method. This solution is then used as the starting solution for a branch-and-bound method, which yields the optimum solution. In many instances, the heuristic method alone gives sufficiently good results. The model and a short outline of the branch-and-bound method are given in the appendix. More details can be found elsewhere [2, 7, 8].

APPLICATION TO UNSUPERVISED PATTERN RECOGNITION

The model extracts p representative patterns from the n patterns present. These p patterns can be visualized as "centres" for a number of other nearby patterns. Suppose that A—O in Fig. 1 are samples for which the values of two variables, X and Y, have been determined; then the selection of B, G and M permits isolation of three groups around these centres. In operational research language, this is called a three-median, while in pattern recognition language these groups are called clusters and the technique is therefore an unsupervised pattern recognition technique.

To illustrate the results obtained by this procedure, it was applied first to a simple two-dimensional case, namely a classification of archaeological artefacts, described by Kowalski et al. [9]. The original data considered of the concentrations of 10 trace elements in 45 samples, the origins of which were known. Thesamples came from four origins and Kowalski et al. reduced the 10-dimensional data to two dimensions (see Fig. 2) using a non-linear mapping method. The distances (in mm) from both axes in Fig. 2 were used as data for the present purpose. Clearly four clusters containing respectively samples 1—9, 10—16, 17—36 + 45, 37—44 should be obtained. The results for p = 1,10 are given in Table 1, for p = 4, the correct result is obtained.

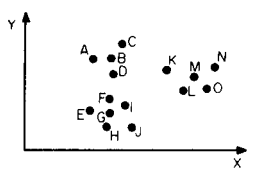


Fig. 1. Map of 15 villages in which three facilities must be located.

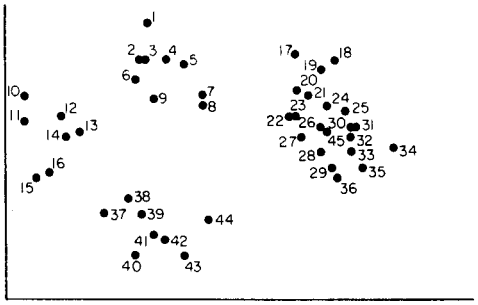


Fig. 2. Two-dimensional representation of 45 samples to be clustered (adapted from Kowalski et al. [9]).

The method was also applied to the 10-dimensional data. These data are available as test data for Kowalski's program package ARTHUR [10]. The data set consists of the original 45 samples with additional ones, making a total of 63. These data are present in a scaled form. The expected classification was obtained. In both cases, Euclidean distances were used to calculate the distances in the operational research algorithm. The clusters selected in this way are more or less round clusters. Wold [11] has shown that, in chemical situations, similar phenomena (or samples) are often situated on lines in the original data space. The operational research algorithm permits, at least in some instances, the detection of such "line clusters".

In the g.c. problem [2], the retention indices of alcohols and ketones, for instance, fall on different lines in the d-dimensional space determined by the retention indices of d stationary phases. By using $1 - \rho$ instead of the Euclidean distances, the correct clusters are found (here ρ is the linear correlation coefficient between the substances and is closer to 1 when it is calculated, for example, between two alcohols than between an alcohol and a ketone).

TABLE 1

Clustering results for p = 1,10

p	Clusters (numbers, see Fig. 2)
2	1—6, 9—16, 37—43/7, 8, 17—36, 44, 45
3	1—14/15, 16, 37—44/17—36, 45
4	1—9/10—16/17—36, 45/37—44
5	1—9/10—16/17—24, 27/25, 26, 28—36, 45/37—44
6	1—9/10—16/17—21/22—28, 45/29—36/37—44
7	1—9/10—16/17—19/20—24, 26, 27/28, 29, 35, 36/25, 30—34, 45/37—44
8	1—4, 6, 9/5, 7, 8/10—16/17—19/10—24, 26, 27/28, 29, 35, 36/25, 30—34, 45/37—44
9	1—4, 6, 9/5, 7, 8/10, 11/12—16/17—19/20—24, 26, 27/28, 29, 35, 36/25, 30—34, 45/37—44
10	1—4, 6, 9/5, 7, 8/10, 11/12—14/15, 16/17—19/20—24, 26, 27/28, 29, 35, 36/25, 30—34, 45/37—44

EXTENSION OF THE MODEL: SIGNIFICANCE OF CLUSTERS

In the example of Fig. 2, there are clearly four clusters. However, the model is not able to perceive this directly. If five clusters are asked for, five will be obtained and some of the clusters obtained in this way are not significant. For example, for $p = 5$ the cluster 17—36 + 45 is split in two (17—24 + 27 and 25, 26, 28—36, 45) while, in fact, there is no reason to split the cluster in this way. This does not give an adequate view of reality, and therefore one must ask whether or not certain clusters should be fused together, and if the clusters obtained are significant.

There may be several ways of answering these questions. What seems to be the best answer is illustrated by Fig. 3. In the $p = 2$ solution the two groups obtained are A—E and F—I. Clearly, the solution is not significant as far as clustering is concerned, although it is correct from the operational research point of view. For $p = 3$, the groups are A—C, D—F and G—I, which is correct. For $p = 2$ cluster D—F is cut in two. In the $p = 2$ solution, D is separated from E—F, but D, E and F come together for $p = 3$. In going from $p = 2$ to $p = 3$, it would be expected that some samples grouped together in the $p = 2$ solution should be separated for $p = 3$, but not the inverse, i.e. that samples separated for $p = 2$ are joined for $p = 3$. When this does occur, it indicates that in the $p = 2$ solution a cluster was cut in two, when it should not have been. The $p = 2$ solution can then be declared insignificant. This leads to the following general rule: let the cluster obtained for a certain value of p be called a p-cluster; a p-cluster is then considered significant when at higher values of p no clusters are formed containing an element or elements of the p-cluster considered together with elements from another p-cluster. In other words, the cluster is significant when its elements are separated definitively from all other elements.

This general rule can now be applied to the results for the classification of the samples shown in Fig. 2. The $p = 2$ to $p = 44$ results were obtained and $p = 2$ to $p = 10$ are shown in Table 1. The 2-clusters are not significant since samples 7 and 8 from the second 2-cluster form a 3-cluster with 1—6 and 9—14 from the first 2-cluster. In the same way, the two first 3-clusters are not significant since 10—14 and 15—16 form a 4-cluster. The third 3-cluster (17—36 + 45), however, is significant, since its elements are definitively separated from all other elements. At the $p = 4$ level, however, all the clusters are significant. As four is the smallest p value for which this is the case, it can be concluded (correctly) that four main clusters can be distinguished in this data set. The same correct conclusion is reached when the 10-dimensional data set described is used.

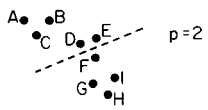


Fig. 3. Example of insignificant clustering.

After the isolation of the main clusters, the structure of each cluster can be investigated. This will not be done here for each cluster, but the kind of results obtained will be demonstrated by considering the simplest cluster (10—16) and then the most complex one (17—36 + 45). The first is separated at the $p = 9$ (see Table 1) level into 10—11 and 12—16. These subclusters are found to be significant. At the $p = 10$ level, one obtains 10—11, 12—14 and 15—16 which is also found to be significant. From Fig. 2, this appears to be correct. The separations concerning 17—36 + 45 are given in Table 2. The table shows that the first significant subcluster to emerge is 17—19 ($p = 7$), then 20—23 ($p = 12$), which is later split up into 20—21, 22—23 and 34 ($p = 14$), which forms a cluster by itself and can therefore be considered as an outlier in this cluster. There is not much sense in going farther because the clusters detached from the main cluster are too small. The picture which emerges is that of a complex cluster in which a few subclusters such as 17—19, 20—21, 22—23 and an outlier (34) are distinguished. Again, this seems to agree with Fig. 2.

The same reasoning was applied to the g.c. problem. Here, too, it was possible to arrive at a conclusion concerning the number of significant clusters. Since these results have been published [2], they are not discussed here.

APPLICATION TO SUPERVISED PATTERN RECOGNITION

There are several ways in which supervised learning can be done with the method proposed. When there are not too many samples to be classified, the preferred method is to apply the location algorithm for all the samples (those from the learning set and the unknowns) together. Since it is now known that there are m groups, the p-median ($p \geqslant m$) is determined. The unknowns are thereby classified as members of one of the p clusters. These are identified each with one of the learning categories and so are the unknowns. This too was applied first to the 2-dimensional data obtained from the paper of Kowalski et al. [9] in the way described above. The 45 samples of Fig. 2 constitute the learning set and 29 other samples have to be classified. The 4-median (i.e. $p = m$) was determined and it was found that all the samples

TABLE 2

Clustering results relating to samples 17—36 and 45

p	Clusters (numbers, see Fig. 2)
5	17—24, 27/25, 26, 28—36, 45
6	17—21/22—28, 45/29—36
7	17—19/20—24, 26, 27/28, 29, 35, 36/25, 30—34, 45
12	17—19/20—23/24, 26, 27, 45/28, 29, 35, 36/25, 30—34
14	17—19/20—23/24, 26, 27, 45/28, 29, 35, 36/25, 30—33/34
16	17—19/20—21/22—23/24, 26, 27, 45/22, 29, 35, 36/25, 30—33/34
19	17—19/20—21/22—23/24, 26, 27, 45/28, 29, 36/33, 35/34/25, 30—32

are correctly classified. Outliers or samples which may be suspected as coming from other unidentified sources are detected because they form clusters consisting of one member when higher medians are also computed. For example, sample B (Fig. 4) is isolated at $p = 15$ and may therefore be considered to be an outlier.

The same classification success is obtained with the variance-weighted 10-dimensional data. Only sample A is not classified with group 1 as proposed by Kowalski et al. [9] but with group 3. However, sample A is an outlier and is isolated as such by the location method at $p = 15$, and from Fig. 4 it can be seen that it is indeed not surprising that A should be located with group 3.

The same procedure was applied to the classification of a data set consisting of two groups of patients in a different functional state of the thyroid (normal, i.e. euthyroid, 150 cases and hyperthyroid, 35 cases). For each of these patients, five chemical variables were measured. The classification of these thyroid patients by linear discriminant analysis, nearest-neighbour techniques and potential functions has been described [12]. The classification results with the present method were as good as with the techniques cited above.

Another problem of supervised learning that can be solved by the present method is the following. When the nearest-neighbour procedure for the classification of unknown individuals is used, it is necessary to store the entire training set in the central memory of the computer, so that for large data sets a very large capacity is required. It would be of interest if it were possible to reduce the training set to a few cluster centroids which are as representative as possible of the whole set. The nearest-neighbour technique can then be carried out by using only the cluster centroids. This was done with the thyroid data set and the results are summarized in Table 3. Admittedly, the data set used is rather well behaved. Nevertheless, Table 3 shows that at least in a number of cases the proposed procedure permits good classification with very few centroids.

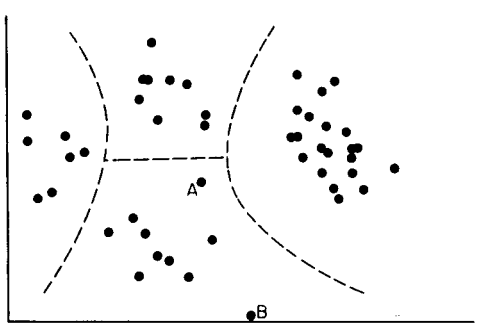


Fig. 4. The samples of Fig. 2, with two outliers (adapted from Kowalski et al. [9]).

TABLE 3

Comparison of the prediction success with nearest-neighbour technique using all the data or a number of centroids

Number of centroids	184	13	10	7	2
% Correct prediction	98.4	96.8	97.3	96.8	91.4

[a] Original data set.

CONCLUSIONS

It is impossible from the study of three data sets to state that the proposed technique is better or worse than existing pattern recognition techniques. This will be decided when it is applied to other data sets and by other researchers.

The most commonly used unsupervised methods are the hierarchic clustering methods and more particularly the agglomerative hierarchic clustering methods [8]. These methods suffer from the drawback that a sequence of decisions is made and that early mistakes in these decisions cannot be corrected. This leads to a sensitivity to outliers and phenomena such as chaining.

Non-hierarchic clustering methods do not suffer from this drawback. In most of these methods one selects centroids around which clusters are built, starting with a set of more or less arbitrarily chosen seed points. Most non-hierarchic clustering methods belong to the so-called nearest-centroid sorting methods (see, e.g. [13]). There are two types among these methods: methods for fixed numbers of clusters (e.g. Forgy [14] or Jancey [15]); and methods for variable numbers of clusters (e.g., Ball and Hall [16]). The extended method described here is a method for variable numbers of clusters. Ball and Hall's method, which is the most perfect non-hierarchic method so far, is an iterative method in which clusters are lumped together or split up according to criteria such as the distance between centroids of neighbouring clusters. These distances must be given by the analyst, which is a drawback of the method. One of the difficulties of this method is also that there is no rule that allows a decision on when to stop the iterations. These difficulties do not exist in the algorithm presented here: no criteria must be preselected by the user and there is no difficulty in deciding when the iteration must be stopped, since the lumping of clusters is not done iteratively.

A distinct advantage of the method presented here is that it distinguishes levels of clusters. For example, in Fig. 2 the four clusters 1—9, 10—16, 17—36 + 45, 37—44 form the first or highest level of clusters. Each of these main clusters can consist itself of smaller clusters (clusters of the second level). Cluster 10—16 consists of the smaller clusters 10—11, 12—14, 15—16. Anderberg states in his authoritative book [13] that this is an attractive

alternative but that there do not seem to be any published instances in which nearest-centroid sorting methods have been used in this way. Therefore, it can be concluded, that, as an unsupervised method, the operational research method proposed here seems to present distinct advantages over other clustering procedures. Its real value compared with other supervised techniques is more uncertain. Clearly, it allows good results for the data sets used here and it has some advantages over techniques such as the nearest-neighbour, because it not only permits classification into known categories but also permits decisions that some of the samples are, in fact, outliers. In Wold's terminology [17], this technique is a level 2 technique while the learning machine, for instance, is a level 1 technique. However, it seems improbable that, in its present state, it is better than SIMCA. One of the possibilities which ought to be investigated, is that this method combined with the SIMCA method may yield a still more powerful and versatile pattern recognition technique than SIMCA. For instance, the initial cluster search might be done with the method proposed here and then the modelling of these clusters by principal component analysis as in SIMCA. In the same way, the reduction of the data sets to a few centroid clusters before the application of the nearest-neighbour method could be valuable when large data sets are studied.

APPENDIX

The following model was used to select p optimum probes out of a set of n probes.

Minimize

$$\sum_i \sum_j d_{ij} x_{ij} \tag{1}$$

subject to

$$\sum_i x_{ij} = 1 \tag{2}$$

$$x_{ij} \leqslant y_i \tag{3}$$

$$\sum_i y_i = p \tag{4}$$

$$y_i, x_{ij} \in \{0, 1\} \tag{5}$$

where i and j are probes ($i, j = 1, 2, 3, \ldots n$); d_{ij} is the distance between i and j and represents the similarity of behaviour for the substances of the data set; y_i is a variable that represents whether a probe is selected or not ($y_i = 1$ if probe i is selected and $y_i = 0$ if not); x_{ij} is a variable that determines which probe is representative of probe j ($x_{ij} = 1$ if j is representative of probe i, which is the case if j is closest to i, and $x_{ij} = 0$ if this is not the case).

Equation (1) represents the total distance of the probes to the representative probes; eqn. (2) indicates that each probe is represented by one of the

selected probes; eqn. (3) that a probe can only be represented by one of the selected probes; and eqn. (4) that exactly p probes are to be selected.

To find the optimum solution of this model all combinations of values that can be taken by the variables can be considered. Such a method, which is called a total combination, is only possible for problems involving up to approximately 30 probes and five selected probes; larger problems contain too many solutions. For such problems a branch-and-bound approach is necessary.

In the branch-and-bound method, the huge set of all possible solutions is divided into subsets (branch) which are examined one at a time. Then for each subset a lower bound is computed: this is a value smaller than or equal to the best solution of the subset. If the lower bound is larger than the best solution found so far, one can say that the subset does not contain the optimum solution; it is then excluded from further consideration and the next subset is considered. Subsequently, one of the remaining subsets is selected and partitioned into smaller subsets. Lower bounds are computed for the smaller subsets and the process is repeated until either a subset is eliminated or it consists of only one element. If in this latter case its value is better (smaller) than the best one found previously, it replaces this solution. If not, it can also be excluded. When all subsets have been excluded, the best solution found so far is the optimum solution.

REFERENCES

1 D. L. Massart and L. Kaufman, Anal. Chem., 47 (1975) 1244A.
2 H. De Clercq, M. Despontin, L. Kaufman and D. L. Massart, J. Chromatogr., 122 (1976) 535.
3 L. Rohrschneider, J. Chromatogr., 22 (1966) 6.
4 W. O. McReynolds, J. Chromatogr. Sci., 8 (1970) 685.
5 A. Hartkopf, S. Grunfeld and R. Delumeya, J. Chromatogr. Sci., 12 (1974) 119.
6 S. R. Lowry, S. Tsuge, J. J. Leary and T. L. Isenhour, J. Chromatogr. Sci., 12 (1974) 124.
7 L. Kaufman, The Location of Economic Activities by 0—1 Programming, Ph. D. Dissertation (1975).
8 D. L. Massart, A. Dijkstra and L. Kaufman, Evaluation and Optimization of Laboratory Methods and Analytical Procedures, Elsevier, Amsterdam, 1978.
9 B. R. Kowalski, T. F. Schatzki and F. H. Stross, Anal. Chem., 44 (1972) 2176.
10 D. L. Duewer, J. R. Koskinen and B. R. Kowalski, ARTHUR; available from B. R. Kowalski, Laboratory for Chemometrics, Dept. of Chemistry BG-10, University of Washington, Seattle, Washington 98195.
11 S. Wold, University of Wisconsin Technical Report No. 357 (1974).
12 D. Coomans, I. Broeckaert, M. Jonckheer, P. Blockx and D. L. Massart, Anal. Chim. Acta, 103 (1978) 409.
13 M. R. Anderberg, Cluster Analysis for Applications, Academic Press, New York, 1973.
14 E. W. Forgy, Biometrics, 21 (1965) 768.
15 R. C. Jancey, Aust. J. Bot., 14 (1966) 127.
16 G. H. Ball and D. S. Hall, Isodata, a novel method of data analysis and pattern classification, Stanford Res. Inst., Menlo Park, CA (1965).
17 C. Albano, W. Dunn, U. Edlund, E. Johansson, B. Nordén, M. Sjöström and S. Wold, Anal. Chim. Acta, 103 (1978) 429.

Analytica Chimica Acta, 122 (1980) 357—361
Computer Techniques and Optimization

SYSTEMATIC COMPUTER-AIDED INTERPRETATION OF VIBRATIONAL SPECTRA

T. VISSER and J. H. van der MAAS*

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

(Received 7th November 1979)

SUMMARY

The interpretation process for vibrational spectra is considered in detail. The concepts of spectral data, structural elements and basic file are described. Two parameters proved to be useful for expressing the degree of correlation of wavenumber regions and structural elements. Three types of intervals are distinguished. A computer program has been developed to obtain these intervals. Factors influencing the results of an interpretation system are discussed and possible criteria are reported.

The interpretation of vibrational spectra requires ample knowledge about correlations between structural elements and spectral data. Frequency or wavenumber, intensity (weak, medium, strong), half bandwidth (broad, sharp, α/β value) and band shape (wedge in carboxylic acids) are used in conjunction with structural elements that may vary from well-defined functional groups like OH, C=C, C=O to functionalities such as cyclohexyl, *para* substitution, etc.

A huge amount of correlations, especially for infrared spectroscopy, is available in the literature. Most of it has been included in Bellamy's books [1, 2] on infrared spectroscopy and the book of Dollish et al. [3] on Raman spectroscopy. Many Colthup tables are also available.

Probably because of the physical impossibility of having that amount of information ready at hand in one mind, interpretation is still an art that has to be learned from experience and that cannot be transferred simply from one person to another. Systematic interpretation would not only exclude accidental and irrelevant factors but would also afford maximum information yield [4, 5]. With the introduction of digitized infrared and Raman spectrophotometers and the decreasing cost of minicomputers, automatic fast interpretation is within reach [6] provided that the software is available. Experience with the development of interpretation systems, including programming, has revealed an unmistakable need for clear-cut rules, which in turn require greater insight into the different relevant factors. A detailed study of the interpretation process seemed to be a logical step.

GENERAL CONCEPTS

The essential feature of systematic interpretation of spectra is to make efficient use of correlations. As the existence of correlations depends on the different types of structural elements and the nature of the spectral data, a study of the development of an interpretation system must be preceded by clear arrangements in these respects.

Spectral data

In practice, the wavenumber of an absorption band is the most conspicuous piece of spectral data and is easy to determine. The intensity of a peak is also useful, but conditions concerning the minimum detection limit, the use of absolute or relative intensity, etc., are required. Other spectral parameters are more difficult to obtain and for that reason the data used are generally limited to wavenumber (regions) and transmittance (thresholds).

In order to be able to use spectral data optimally, recording has to be done under identical scanning conditions (e.g., a wavenumber accuracy of ± 2 cm^{-1}). Although a large amount of spectral data is available in the literature (e.g., Sadtler [7] and DMS [8]), it cannot be used as the scanning conditions are not fulfilled. Therefore it is necessary to compose a basic file of spectra.

Basic file

As the aim of an interpretation system is the detection of structural elements (and not retrieval), a basic file can be composed of a limited number of spectra. For a deliberate selection of compounds, chemical and spectroscopic knowledge, and thus experience, is essential. In order to be able to correlate a structural element with a spectral region, two spectra representing the extreme frequencies of that element will do. Thus the number of structural elements defines the number of spectra in the file. At first glance, further reduction of the amount of data seems possible by combining several elements within one compound but this is not really viable for reasons set out in a later section. The file may, of course, be large, the only objection being that more computer time and space is required. Compounds containing other elements than those to be investigated may be present. In fact, the spectral assembly has to reflect the compounds for which the system will be used.

Structural elements

A structural element is an entity of at least two bonded atoms. For carbon, hydrogen and oxygen, for example, the minimum number of different two-atomic entities is six: CC, CH, CO, HH, OH and OO. However, several types of bonding can be distinguished and so the combinations CC, CO and OO have to be increased to C—C, C=C, C≡C, C—O, C=O, O—O and O=O. Since the combination O=O represents the oxygen molecule and HH the hydrogen molecule these two can be omitted. In this way, the total

number of so-called "main" elements reduces to eight. Other possible combinations will consist of more than two atoms and can be regarded therefore as derivatives or sub-elements of those eight. For example, C≡C—H is a sub-element of C≡C. In this way all elements can be classified systematically.

Correlations

A structural element S_a is said to be correlated if its presence in a compound is generally associated with the appearance of at least one peak within a designated region of the spectrum of that sample. If a file is composed of n compounds, of which m have the element S_a, then if all compounds with S_a show a peak in a particular region, the correlation is at a maximum and the score percentage (*SP*) is $m \times 100/m = 100\%$. As each compound will have at least one peak in a complete spectrum, a 100% score can always be reached if the interval is broad enough. However, with the score percentage, the chance of having interfering peaks of elements different from S_a increases. With l interfering compounds, the interfering percentage (*IP*) is $(l/n-m) \times 100\%$. With *IP* and *SP*, it is possible to distinguish three kinds of regions: (a) $SP = 100\%$ and $IP = 0\%$, which is specific; (b) $SP = 100\%$ and $IP > 0\%$, which is selective; and (c) $SP < 100\%$ and $IP = 0\%$, which is pseudo-specific.

As the intensity threshold determines the presence or absence of a peak, the finding of a region, including *SP* and *IP* and thus its type, is related to that threshold. To find correlations for a structural element in relation to a certain intensity threshold, a program CRISE (Correlation of Raman and Infrared data with Structural Elements) has been developed. The program supplies for any element S_a, at a chosen intensity threshold, all possible (pseudo-)specific and selective regions, provided that the presence/absence of that particular S_a in each compound of the basic file is known. For the pseudo-specific and selective regions, the *SP* and *IP* values, respectively, are calculated.

RESULTS AND DISCUSSION

From preliminary results with CRISE, it appeared that more regions may be found than would be expected theoretically and also that a specific region is very rare; even an *SP* of 70% proved to be rather high. There are various reasons for this. First, regions, which could have been specific, are subdivided into some pseudo-specific regions because of the presence of a few interfering peaks ($\neq S_a$). Secondly, the simultaneous presence of an element S_b ($\neq S_a$) next to S_a is inevitable in a number of compounds and therefore regions may be found which have nothing to do with S_a but all with S_b; these intervals must be excluded. Thirdly, when no maximum is set for the interval width, a large number of very broad selective regions is found. These intervals are chemically and spectroscopically meaningless as functionalities rarely exceed a bandwidth of 300 cm^{-1}. Fourthly, regions may be found that are strongly correlated with vibrations of sub-elements of S_a. These will also be found

when the sub-element itself is investigated and may then even appear to be specific. From this last consideration, it follows that all regions which are (pseudo-)specific or selective for a certain sub-element also form part of the (pseudo-)specific and selective intervals for the corresponding main element. So correlations which are found for S_a may indicate the presence of useful regions for its sub-elements.

If no specific region is found for S_a, one can try to reach a 100% score by combining two or more pseudo-specific regions. With CRISE, it is possible to combine up to five regions (if present). They can be selected interactively. The score *SP* of each of the combinations is calculated. Thus the system has to be constructed from specific, selective and (a combination of) pseudo-specific regions.

The system

The program CRISE (which is available from the authors on request) supplies all data necessary for composing an interpretation system. In the system, regions are used in the form of questions, hence they will be referred to as *Q*'s in the following paragraphs.

A fundamental requirement of a system must be that there are no wrong answers. Another condition may be that for any structural element a certain score percentage (*SP*) has to be reached, which can be used as a criterion. It may be reached either with one or with a combination of pseudo-specific *Q*'s. The maximum number of *Q*'s one is prepared to combine (mQ) can be used as another criterion. Similarly to *SP* for pseudo-specific cases *IP* can be used for selective *Q*'s. These three criteria (*SP*, *IP* and mQ) and the variable intensity threshold determine the system. Accordingly, the following types of *Q*'s can be distinguished in a system: (a) specific: supply complete certainty about the presence or absence of S_a (Fig. 1a); (b) pseudo-specific: may supply certainty about the presence of S_a (Fig. 1b); (c) combined: as (b) (Fig. 1c); (d) selective: may only supply certainty about the absence of S_a (Fig. 1d).

All these questions can be put at any place in the system and they are therefore called unconditioned questions. Thus, results from CRISE that fulfil the mentioned criteria can be simply composed into a system.

Conclusions

Comparison of the system based on the results of CRISE and the earlier developed systems [4, 5] showed that the latter also contain combined answers such as "S_a and/or S_b present". Apart from the usefulness of this kind of information, the data necessary to find such types of question are present implicitly in the results of CRISE; to expose them a modification of the program is required. For the choice of variables and the criteria, and thus for the composition of the system, there is no difference in principle.

In summary, a viable system must be defined by several variables, viz. the basic file, scanning conditions, intensity threshold and the criteria for *SP*, *IP*

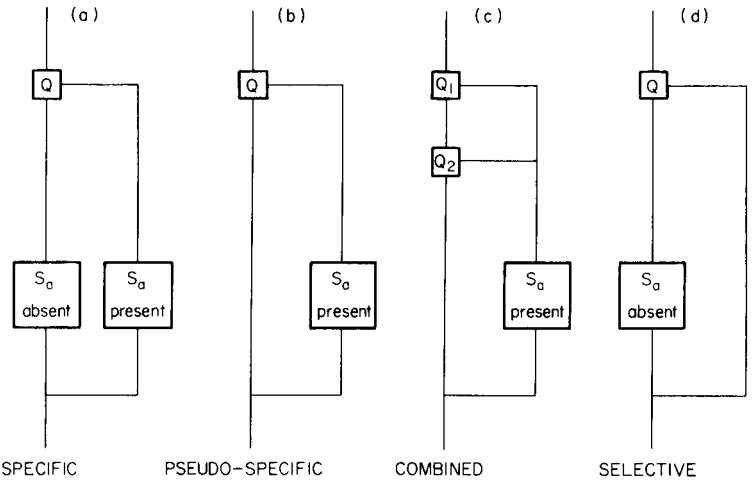


Fig. 1. The four types of questions in an interpretation system. The small squares are question elements (Q's) (positive to the right, negative downwards); the large squares are answers. S_a = structural element.

and mQ. It must be emphasized once more that without a solid knowledge of chemistry and vibrational spectroscopy a deliberate choice of the variables is impossible. An objective comparison of interpretation systems is possible on the basis of the criteria given in this paper.

REFERENCES

1 L. J. Bellamy, The Infrared Spectra of Complex Molecules, Chapman and Hall, London, 1975.
2 L. J. Bellamy, Advances in Infrared Group Frequencies, Methuen, London, 1968.
3 F. R. Dollish, W. G. Fateley and F. F. Bently, Characteristic Raman Frequencies of Organic Compounds, Wiley, New York, 1974.
4 T. Visser and J. H. van der Maas, J. Raman Spectrosc., 7 (1978) 278.
5 C. G. A. van Eijk and J. H. van der Maas, Fresenius Z. Anal. Chem., 291 (1978) 308.
6 J. P. Coates and S. Geary, Anal. Chim. Acta, 103 (1978) 303.
7 The Sadtler Standard Spectra, Sadtler Research Laboratories Inc., Philadelphia.
8 DMS Documentation of Molecular Spectroscopy, Verlag Chemie, Weinheim and Butterworth, London, 1975.

Analytica Chimica Acta, 122 (1980) 363—372
Computer Techniques and Optimization

SYSTEMATIC COMPUTER-AIDED INTERPRETATION OF INFRARED AND RAMAN VIBRATIONAL SPECTRA BASED ON THE CRISE PROGRAM

T. VISSER and J. H. van der MAAS*

Laboratory for Analytical Chemistry, University of Utrecht, Croesestraat 77a, 3522 AD Utrecht (The Netherlands)

(Received 7th November 1979)

SUMMARY

The CRISE computer program is used to correlate wavenumber regions and 6 structural elements containing carbon, hydrogen and oxygen on the basis of 2 standard files with 549 infrared and Raman spectra. The degree of correlation, including score percentages and interfering percentages, is established for different types of intervals in relation to various intensity thresholds. Specific regions (score 100%, interference 0%) proved to be rare, whereas pseudo-specific regions (score < 100%, interference 0%) are normally present. The usefulness of selective regions (score 100%, interference > 0%) is doubtful. The infrared and Raman results for a structural element can differ appreciably, yet neither technique is clearly superior for interpretative purposes.

Interpretation systems for some vibrational spectra can be compared objectively; the necessary parameters can be obtained from a file of coded spectra with the CRISE computer program in limited cases [1]. Files of both the infrared (i.r.) and Raman spectra of over 500 organic compounds are available here, and so it has been possible not only to study the interpretative value of each technique separately, but also to compare their particular utility. It should be noted that the comparison is not perfect as the scanning conditions for the i.r. and Raman spectra were slightly different.

EXPERIMENTAL

Two files, i.r. and Raman, were prepared; both contained the spectra of the same 549 liquid organic compounds. The samples contained the atoms CH (161), CHO (333), CHN (51) and CHNO (4). They were either commercial products or were obtained from the Laboratory of Organic Chemistry of this University. The purity of all compounds was $\geqslant$ 98% (checked by g.c.).

The i.r. spectra were recorded on a Perkin-Elmer 180 spectrometer; the accuracy was 2 cm^{-1} in the region 4000—2000 cm^{-1} and 1 cm^{-1} in the region 2000—600 cm^{-1}. The baseline was adjusted between 100 and 95% transmittance, the transmittance of the most intense band being 3—7% transmittance.

The Raman spectra were recorded on a Spectra-Physics 700 spectrometer with a Spectra-Physics 165 argon ion laser as light source. Spectra were recorded in the ranges 4000—2000 cm^{-1} and 2000—200 cm^{-1} with a precision of 2 cm^{-1}. The sensitivity was adjusted in such a way that the most intense band in the spectral region gave a reading between 85 and 95 scale divisions. For both techniques the peak height (intensity) is defined as the distance between the top of that peak and the baseline on both sides of the peak. A peak is considered to be present if its height is $\geqslant 3$ scale divisions.

The program CRISE is written in FORTRAN EXTENDED IV. Data calculation performed on a CYBER 73-28 computer required a memory of 52K words and, depending on the structural element investigated, 15—30 CPU seconds per structural element. The different wavenumber precisions have been incorporated in the results of CRISE.

RESULTS

The basic files contain the atoms C, H, O and N. Only structural elements of C, H and O have been studied, as the number of compounds with nitrogen was considered to be too small for reliable results. The number of main elements is therefore eight, viz. C—H, C—C, C=C, C≡C, O—H, C—O, C=O and O—O. Of these, O—O, C—H and C—C have not been investigated: for obvious reasons, O—O is not present in the basic files; C—H and C—C would not produce useful intervals in i.r. or in Raman spectroscopy, because C—H is present in every compound and the C—C vibration is usually strongly coupled. The element C—H can be split into three semi-elements: —C—H, =C—H and ≡C—H, which form part of nine sub-elements (Table 1). Four of these are already included in the corresponding main elements, and as neither C—C nor the nitrogen-containing elements are included (see above), this leaves only one extra element, viz. —C—H (alkyl).

As reported previously [1], the CRISE program produces pseudo-specific and selective intervals in addition to specific intervals. Here, the number of non-specific intervals is kept within reasonable limits by setting a score percentage (*SP*) of at least 4% and a maximum interval width of 300 cm^{-1}. The number of combined regions or combined questions (*mQ*'s) has been restricted to five.

TABLE 1

The different types of structural elements

Semi-element	Sub-element	Main element	Semi-element	Sub-element	Main element
H—C—	H—C—C	C—C	H—C=	H—C=C	C=C
	H—C—H			H—C=O	C=O
	H—C—O	C—O		H—C=N	C=N
	H—C—N	C—N	H—C≡	H—C≡C	C≡C
				H—C≡N	C≡N

By means of CRISE, the presence of regions is investigated in relation to the intensity threshold. The results for each of the six primary elements are discussed below and a selection of the regions with the highest *SP*'s or the lowest *IP*'s (interfering percentages) are summarized in separate tables.

The main element —C—H (see Table 2)

Specific regions are not present. Although —C—H vibrations are always active, in Raman spectroscopy some vibrations have insufficient intensity. In i.r. spectroscopy, potentially specific regions are split up by several interfering peaks.

There are several pseudo-specific regions. The highest *SP*'s appear, in both i.r. as well as in Raman spectroscopy, in the —C—H stretching area around 2900 cm^{-1}. This is due both to the generally high intensity of these bands and to the greater chance of interfering bands in other regions. As appears from Fig. 1A, the slope of *SP* as a function of intensity threshold is about the same for both techniques, and there is also no great difference in the maxima.

Selective regions are few, and appear in i.r. spectroscopy only. All regions show IP's over 50%. It should be noted that only 10 compounds of the investigated 549 do not contain an alkyl group.

There are several combined regions. As the starting *SP* is already 90%, the profit gained from these regions is very small.

It can be concluded that alkyl groups can be detected by Raman as satisfactorily as by i.r. spectroscopy.

The main element C=C (see Table 3)

Specific regions are not present. For obvious reasons, a specific interval would have to be found in the C=C stretching area around 1600 cm^{-1}. As C=C=C compounds do not show a peak in this region but at 2000 cm^{-1} instead, a specific region is not found in the Raman or in i.r. spectra. Moreover,

TABLE 2

Selection of the (combined) intervals obtained for the main element —C—H (*IT*, intensity threshold; *SP*, score percentage; *IP*, interfering percentage; *mQ*, number of combined intervals.)

Infrared								Raman						
IT	(Pseudo-) specific	*SP*	Selective	*IP*	*IT*	Combination	*SP*	*IT*	(Pseudo-) specific	*SP*	*IT*	Combination	*SP*
03	2949—2926	57	3007—2887	50	10	$mQ = 2$	95	03	2878—2847	71	20	$mQ = 2$	97
10	2965—2923	88	1473—1383	70	10	$mQ = 3$	96	10	2937—2884	93	10	$mQ = 3$	98
20	2965—2923	87			10	$mQ = 4$	98	20	2997—2884	96	10	$mQ = 4$	99
30	2967—2633	91			10	$mQ = 5$	98	30	2997—2840	91	10	$mQ = 5$	99
40	3007—2675	90						40	2997—2835	89			
50	2988—2707	83						50	2997—2827	84			

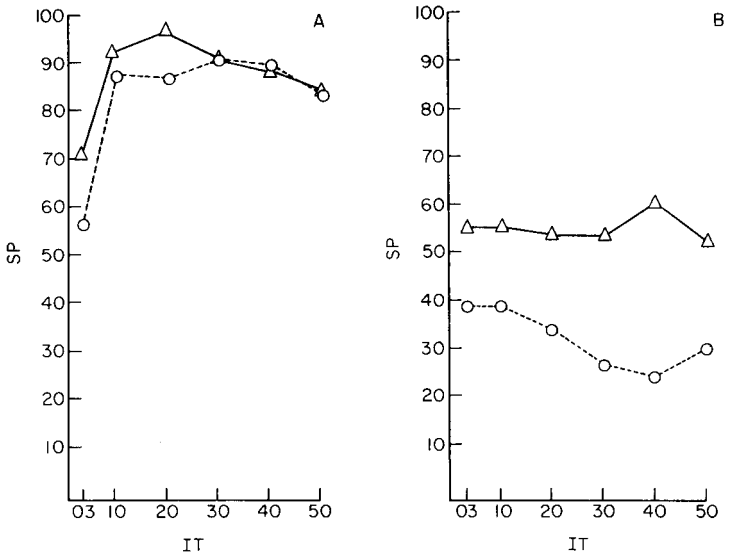


Fig. 1. The score percentage, *SP*, as a function of the intensity threshold, *IT*, for the main elements (A) —C—H, and (B) C=C. (---) Infrared; (—) Raman.

TABLE 3

Selection of the (combined) intervals obtained for the main element C=C (Abbreviations as in Table 2.)

Infrared						Raman					
IT	(Pseudo-) specific	*SP*	*IT*	Combin-ation	*SP*	*IT*	(Pseudo-) specific	*SP*	*IT*	Combin-ation	*SP*
03	3042—3023	39	03	$mQ = 2$	54	03	1612—1569	54	40	$mQ = 2$	87
10	3042—3023	39	10	$mQ = 3$	65	10	1622—1562	54	40	$mQ = 3$	88
20	3042—3023	34	30	$mQ = 4$	69	20	1649—1500	53	40	$mQ = 4$	88
30	3042—3023	27	30	$mQ = 5$	73	30	1655—1500	53	40	$mQ = 5$	88
40	1508—1490	24				40	1669—1500	60			
50	1519—1490	30				50	1669—1500	52			

some double bonds are (pseudo-)symmetric and these vibrations do not absorb in the i.r. Also, in both i.r. and in Raman spectroscopy, there is interference from other structural elements (C=O and C=N) in the C=C region.

There are several pseudo-specific regions in both types of spectra. As can be seen from Fig. 1B, *SP* is only slightly influenced by the increase in the intensity threshold, particularly in Raman spectroscopy; although a raised threshold prevents weak peaks from scoring, it also involves less interference, so that the initial loss of *SP* is compensated by a broadening of the pseudo-specific region (see Table 3). It appears that the regions with the highest *SP*'s are due

to the C=C stretching vibration in Raman spectroscopy and to either the C=C (at an intensity threshold exceeding 30%) or the =C—H stretching around 3030 cm^{-1} in the i.r. spectra. This supports the rough-and-ready rule that the C=C band is more characteristic in Raman than in i.r. spectroscopy.

Selective regions are not present, for the above-mentioned reasons. There are several combined regions. Figure 2 shows that for Raman spectroscopy an *SP* of 87% is already attained with a combination of two *Q*'s, whereas for i.r. a 75% *SP* is not reached even with five. The gradual increase of *SP* in i.r. spectroscopy indicates that the regions are only slightly correlated; they may be related to some of the sub-elements of C=C.

It can be concluded that Raman is more powerful for the detection of C=C than i.r. spectroscopy.

The main element C≡C (see Table 4)

There is one specific region, for Raman only. The region is clearly correlated to the C≡C stretching vibration which is always Raman-active. No other bands are found in this region. The reason for the absence of an interval for i.r. spectroscopy is similar to that for C=C: a (pseudo-)symmetric C≡C is infrared-inactive.

Pseudo-specific regions are few for i.r., and there is only one for Raman spectroscopy. In the i.r. spectra, the regions are related to either the ≡C—H or the C≡C stretching vibration, the latter showing the highest *SP* (see Fig. 3A). For Raman spectroscopy, the specific region becomes pseudo-specific at intensity thresholds exceeding 30%.

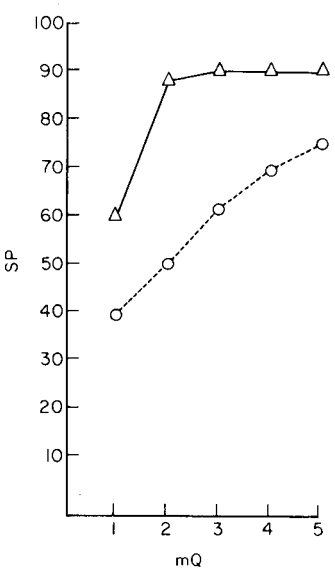


Fig. 2. The score percentage, *SP*, as a function of the number of combined intervals, *mQ*, for the main element C=C. (---) Infrared; (—) Raman.

TABLE 4

Selection of the (combined) intervals obtained for the main element C≡C (Abbreviations as in Table 2.)

Infrared						Raman			
IT	(Pseudo-) specific	*SP*	*IT*	Combination	*SP*	(Pseudo-) specific	*SP*	Selective	*IP*
03	2112—2110	11	10	$mQ = 2$	58	2276—2098	100	2276—2098	0
10	2273—2118	36	10	$mQ = 3$	58	2320—2065	100	2320—2065	0
20	2273—2092	44				2320—2065	100	2320—2065	0
30	2273—2092	30				2303—2065	98		
40	2273—2092	23				2298—2098	98		
50	2273—2092	21				2276—2098	96		

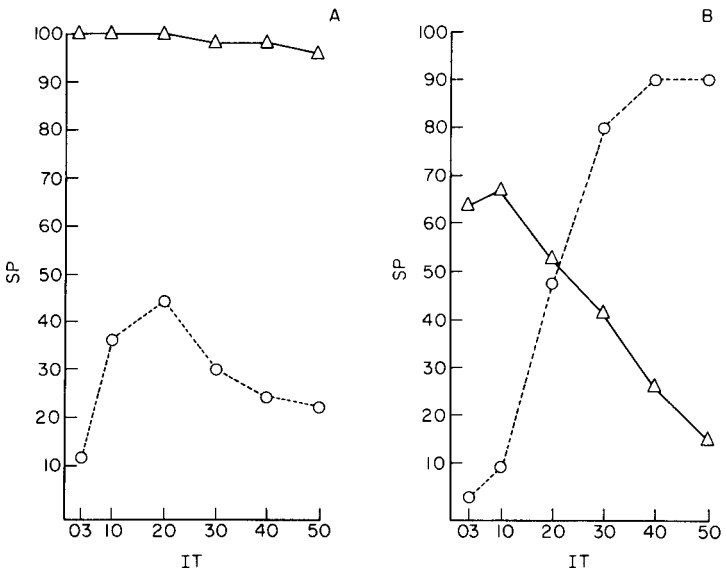


Fig. 3. The score percentage, *SP*, as a function of the intensity threshold, *IT*, for the main elements (A) C≡C, and (B) C=O. (---) Infrared; (—) Raman.

Selective regions are not present in i.r. spectra for the reasons set out above. There are a few combined regions. The maximum *SP* reached is 58% by combining two regions. Addition of an interval caused by the ≡C—H linkage does not increase the *SP*; the interval appears to be completely correlated with the former combination.

It can be concluded that Raman spectroscopy is eminently suited for the detection of this element.

The main element O—H (see Table 5)

Specific regions are not present. Although the O—H stretching vibration is strongly infrared-active, the presence of a specific region is prohibited by the

TABLE 5

Selection of the (combined) intervals obtained for the main element O—H (Abbreviations as in Table 2.)

Infrared					
IT	(Pseudo-) specific	*SP*	*IT*	Combin-ation	*SP*
03	3625—3602	4	50	$mQ = 2$	54
10	3350—3340	22	50	$mQ = 3$	57
20	3350—3323	35	50	$mQ = 4$	57
30	3350—3323	35			
40	3350—3323	35			
50	3350—3323	35			

carboxylic acids which do not show a band around 3300 cm^{-1}. Moreover, the ≡C—H and N—H stretching vibrations give interfering bands. For Raman spectroscopy, the lack of a specific region was expected, as the O—H is only slightly active.

There are several pseudo-specific regions in the i.r. spectra; a maximum *SP* of 35% is found. For Raman spectroscopy, no region reaches the minimum *SP* of 4%. Selective regions are absent for the reasons mentioned above. There are several combined regions for i.r. spectroscopy; a maximum *SP* is reached with five *Q*'s at an intensity threshold of 20%.

In conclusion, the presence of this element cannot be established by Raman and the score is rather small by i.r. spectroscopy.

The main element C—O (see Table 6)

Specific regions are absent, although the C—O stretching vibration is active in both i.r. and Raman spectroscopy. The absence of specific regions can be attributed to (i) the very low intensity (< 3%) of some C—O bands, which in

TABLE 6

Selection of the (combined) intervals obtained for the main element C—O (Abbreviations as in Table 2.)

Infrared							
IT	(Pseudo-) specific	*SP*	Selective	*IP*	*IT*	Combin-ation	*SP*
03			1224—1026	97	50	$mQ = 2$	11
10			1228—1051	83	50	$mQ = 3$	13
20			1285—985	77	50	$mQ = 4$	15
30					50	$mQ = 5$	17
40	1189—1186	4					
50	1022—1018	7					

combination with the scanning conditions leads to the absence of a C—O band; and (ii) the strong coupling of C—O vibrations with skeletal modes; the bands are thus found over a wide region and interferences occur because many skeletal C—C vibrations give rise to peaks in the same region.

There are some pseudo-specific regions in the i.r. spectra, but the highest *SP* is only 9%. There are two selective regions in the i.r. spectra (*IP* > 80%) and several combined regions; for the latter regions, five *Q*'s yield a maximum *SP* of 17%. There are no analogous regions in Raman spectra.

It must be concluded that neither i.r. nor Raman spectroscopy is the obvious technique for the detection of the main element C—O.

The main element C=O (see Table 7)

Specific regions are absent. In the Raman spectra some C=O bands do not reach the minimum intensity of 3%; in both types of spectra, C=C bands interfere in the C=O stretching region.

There are two pseudo-specific regions in each type of spectra. As can be seen from Fig. 3B, the slope of *SP* as a function of intensity threshold for i.r. spectra is opposite to that for Raman spectroscopy. For i.r. spectra, this illustrates that for strongly active vibrations, and thus very intense bands, the influence of interfering bands decreases with an increase in the intensity threshold. For Raman spectra, *SP* decreases, as the intensity of the C=O bands is significantly smaller. Also, it can be seen that there are more interfering peaks at intensity thresholds below 30% in the i.r. spectra, as the *SP* in Raman is higher.

There is one selective region, for the i.r. spectra only. In contrast to all other selective regions, the *IP* found is small (8%). As the C=O band is always very intense in i.r. spectra, a high intensity threshold does not influence *SP* but it does decrease *IP* appreciably. There is only one combined region for each technique. The combination yields a maximum *SP* of 97% for i.r. and 69% for Raman spectroscopy.

It can be concluded that the presence as well as the absence of C=O can be established quite well, especially with the i.r. method.

TABLE 7

Selection of the (combined) intervals obtained for the main element C=O (Abbreviations as in Table 2.)

Infrared								Raman				
IT	(Pseudo-) specific	*SP*	Selective	*IP*	*IT*	Combination	*SP*	(Pseudo-) specific	*SP*	*IT*	Combination	*SP*
03	3425—3423	7	1777—1631	53	50	$mQ = 2$	98	1793—1711	63	10	$mQ = 2$	69
10	1737—1728	10	1777—1631	28				1793—1708	66			
20	1737—1710	49	1777—1631	18				1765—1708	52			
30	1745—1684	81	1777—1631	15				1765—1708	41			
40	1795—1684	91	1777—1631	11				1763—1708	26			
50	1795—1684	91	1777—1631	8				1763—1708	15			

DISCUSSION

Some general remarks can be added to the detailed conclusions outlined above for the different structural elements.

Correlation

In agreement with earlier conclusions [2, 3], the number of specific regions is very small. As the main reason is the presence of interfering peaks, the number will be even smaller for larger basic files.

Pseudo-specific regions were always present in the i.r. spectra whereas in Raman those for C—O and O—H were absent. The *SP* values differ clearly for the structural elements investigated. Selective regions, except for C=O, show interfering percentages over 50%; as this type of region can be used only to establish the absence of an S_a, its usefulness seems very limited.

The combination of intervals (*Q*'s) sometimes produces a considerable increase of *SP*, but when the separate regions involved are mutually correlated, the gain is very small. Combinations of different techniques and of regions with different intensity thresholds have not been studied so far, because of difficulties in programming. These types of combinations may yield higher *SP*'s. A comparison of the *SP* maxima for i.r. and Raman spectroscopy (Table 8) shows that neither of the two techniques is significantly better than the other.

Intensity threshold

A uniform intensity threshold for all structural elements reduces the intensity data to a binary value: a peak is either present or absent. A high threshold reduces the number of peaks significantly, which is advantageous with respect to programming and to storage and processing of the spectral data. However, as can be seen from Table 8, the *SP* maxima of the elements investigated appear to be reached at different intensity thresholds, and it might be beneficial, therefore, to use various thresholds. For each S_a investigated, the maximum *SP* for a threshold of 10% is greater than or equal to the

TABLE 8

The maxima score percentage, *SP*, reached with one interval (*IT*, intensity threshold)

	Infrared		Raman	
	IT	*SP*	*IT*	*SP*
—C—H	30	91	20	96
C=C	10	39	40	60
C≡C	20	44	20	100
O—H	50	35	—	—
C—O	50	7	—	—
C=O	50	91	10	66

SP for a threshold of 3%, which implies that peaks with intensities under 10% can be neglected. This supports a conclusion of van Eijk and van der Maas [4] who reported that loss of information starts at intensity thresholds exceeding 15%. It should be emphasized, however, that this phenomenon is valid for this particular file and the scanning conditions set. As Dupuis et al. [5] reported, there is a relatively small variation in information content of binary-coded spectra for intensity thresholds between 3 and 10%, except for files of saturated carbon—hydrogen compounds. The maximum intensity threshold was limited to 50%, because experiments indicated that the *SP* decreased for higher thresholds.

The study described above is a preliminary one, hence the number of structural elements investigated was limited to the six main ones. The procedure of establishing correlations between molecular structures and spectral data in relation to several variables is independent of the choice of the element and the basic file. The overall results of this study indicate that this approach to the systematic interpretation of vibrational spectra is promising. The CRISE program appears to offer the parameters necessary for the development of an interpretation system and for establishing the validity of the system.

REFERENCES

1 T. Visser and J. H. van der Maas, Anal. Chim. Acta, 122 (1980) 357.
2 T. Visser and J. H. van der Maas, J. Raman Spectrosc., 7 (1978) 125.
3 T. Visser and J. H. van der Maas, J. Raman Spectrosc., 7 (1978) 278.
4 C. G. A. van Eijk and J. H. van der Maas, Fresenius Z. Anal. Chem., 291 (1978) 308.
5 P. F. Dupuis, A. Dijkstra and J. H. van der Maas, Fresenius Z. Anal. Chem., 291 (1978) 27.

Analytica Chimica Acta, 122 (1980) 373—386
Computer Techniques and Optimization

NEW POSSIBILITIES FOR LEAST-SQUARES FITTING OF MÖSSBAUER SPECTRA

H. NULLENS, G. DE ROY, P. VAN ESPEN, F. ADAMS* and E. F. VANSANT

University of Antwerp (U.I.A.), Department of Chemistry, Universiteitsplein 1, B-2610 Wilrijk (Belgium)

(Received 15th January 1980)

SUMMARY

A non-linear least-squares program for the analysis of Mössbauer spectra is presented. The program is capable of resolving very complex spectra and can be used on a mini-computer system, with regard to both calculation time and memory requirements. The commonly used χ^2-minimization algorithm was slightly adapted, so as to broaden its working range. All possibilities for fixing or changing parameters are provided. Two new parameter limitation techniques, which greatly reduce the need for manual intervention during the fitting process, are discussed extensively.

The measurement of the Mössbauer effect leads to digital spectra with a more or less horizontal background and a number of Lorentzian-shaped absorption peaks. These spectra are often very complex, with many overlapping peaks, and can only be resolved by using the sophisticated non-linear least-squares fitting technique. Several computer codes [1—5] have therefore been presented to deal with the problems of Mössbauer analysis. Because of the complexity of the spectra processed, most of these programs have high memory and calculation time requirements, and cannot therefore be used on micro- or mini-computer systems, although many commercial Mössbauer spectrometers are nowadays equipped with a computer-based data acquisition system.

In contrast, the large amount of preliminary physical information concerning spectrum parameters, which is usually available, cannot readily be used by the least-squares algorithm. This has led to the widespread practice of manual program control, i.e., repeatedly fixing and releasing the value of certain parameters during some stages of the χ^2-minimization process. This is, of course, a time-consuming practice which should be avoided as much as possible. A FORTRAN IV program, called ANMOS1, which can take into account virtually all available a priori information concerning the spectrum parameters, and is capable of resolving the most complex spectra when used on a mini-computer system, has therefore been developed.

SPECTRUM DESCRIPTION AND FITTING FUNCTION

The first task to be performed in non-linear least-squares fitting is the construction of a theoretical model to describe the experimental data obtained. For Mössbauer spectroscopy, the problem consists in formulating a fitting function, $f(x_i)$, which will describe the spectra obtained as accurately as possible. The different components of this fitting function are discussed below.

Channel number—velocity relation

Two types of velocity control are usually provided on commercial Mössbauer systems, a triangular and a sinusoidal mode, which are represented in Fig. 1. In the program presented here, which can analyze both types of spectra, the relations between velocity v_i and channel number x_i are given by:

$$v_i = D_v (x_i - x_0) \qquad \text{for data with } x_i \leqslant x_M \tag{1}$$

$$v_i = D_v (2x_M - x_i - x_0) \qquad \text{for data with } x_i > x_M \tag{2}$$

for the triangular mode and

$$v_i = D_v (x_M - x_0) \sin \left[\frac{\pi}{2} (x_i - x_0)/(x_M - x_0) \right] \tag{3}$$

for the sine mode, where the three calibration parameters are: D_v the mean change of velocity per channel, x_0 the channel number of zero velocity, and x_M the channel number of minimum or maximum velocity (folding point). When a triangular mode spectrum, which is not of the usual mirror-image type (reversed scan type), is analyzed, relation (2) is not used.

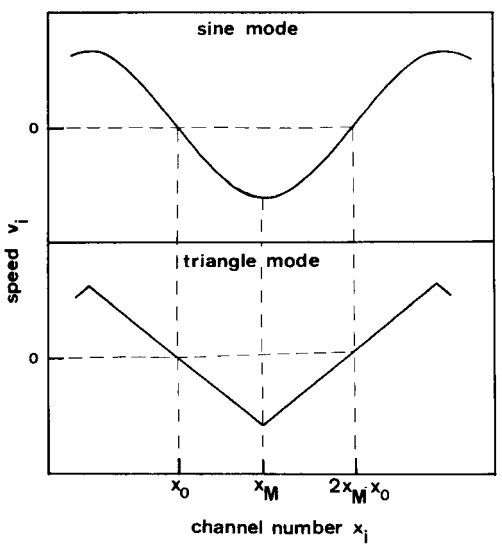


Fig. 1. Relations between channel number x_i and velocity v_i.

Background

As a first approximation, the background of a Mössbauer spectrum can be represented by a horizontal line. Since the background level is inversely proportional to the square of the distance between source and detector, source movement may lead to significant curvature if small distances or long counting times are used. Integration of the velocity relations (1)—(3) and introduction of some simplifications results in a description for the background $B(x_i)$ which contains two independent parameters, b_1 and b_2:

$$B(x_i) = b_1/[1 + b_2 T(x_i)]^2 \tag{4}$$

$$\text{with } T(x_i) = (x_i^2/2) - x_0 x_i - (x_M^2/2) + x_0 x_M \qquad \text{for } x_i \leqslant x_M \tag{5}$$

$$\text{or } T(x_i) = -(x_i^2/2) + (2x_M - x_0)\, x_i - (3x_M^2/2) + x_0 x_M \qquad \text{for } x_i > x_M \tag{6}$$

for the triangular mode, and

$$T(x_i) = \cos\left[\frac{\pi}{2}(x_i - x_0)/(x_M - x_0)\right] \tag{7}$$

for the sinusoidal mode operation. A background fit based on this description is shown in Fig. 2 for a somewhat extreme example.

Peaks

Throughout the program, the simple Lorentzian formula is used to describe the peaks:

$$L_j(x_i) = (2A_j/\pi\Gamma_j)/\{1 + [2(v_i - \mu_j)/\Gamma_j]^2\} \tag{8}$$

or, alternatively

$$L_j(x_i) = a_j/\{1 + [2(v_i - \mu_j)/\Gamma_j]^2\} \tag{9}$$

Here $L_j(x_i)$ is the value of the jth Lorentzian at channel x_i, a_j the amplitude, relative to background, A_j the area, μ_j the centroid velocity, and Γ_j the width in units of velocity. Since the simple Lorentzian should be adequate to describe the peaks for most of the work done in the Mössbauer field, no deviations of the peak shape arising from sample thickness or other effects are taken into account.

Fitting function

The complete fitting function, $f(x_i)$, is obtained by combining eqns. (1)—(9):

$$f(x_i) = B(x_i)\ \{1 - \sum_{j=1}^{n_L} [L_j(x_i)]\} \tag{10}$$

where n_L represents the total number of Lorentzians present.

SPECTRUM PARAMETERS AND FITTING PARAMETERS

The fitting function in eqn. (10) is characterized by a number n_S of physical quantities or "spectrum parameters" S_j, i.e. the calibration par-

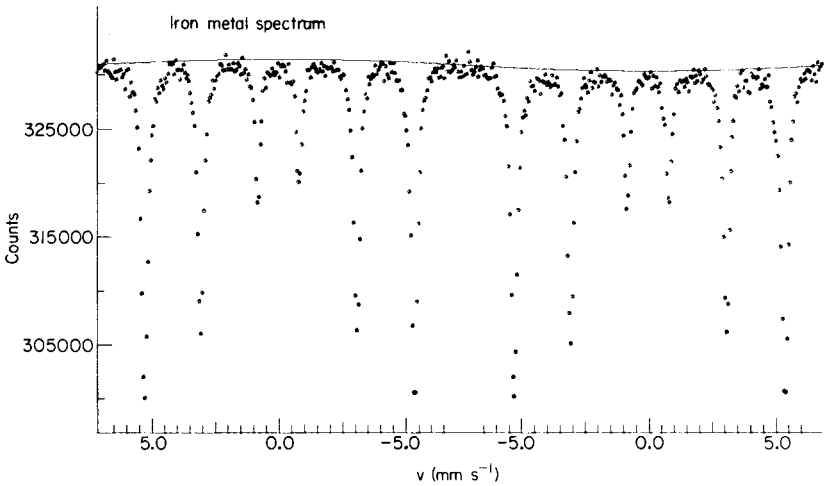


Fig. 2. Fit to iron metal spectrum showing background curvature.

ameters D_v, x_0 and x_M, the background parameters b_1 and b_2 and all the peak parameters A_k or a_k, μ_k and Γ_k, with

$$n_S = 5 + 3n_L \tag{11}$$

Most programs are conceived to optimize these spectrum parameters directly, i.e. they make no distinction between the physical quantities S_j and the values operated on by the least-squares algorithm, which will be called the "fitting parameters", P_j. This leads to highly undesirable situations, however, since a somewhat complex spectrum will then need a large number of fitting parameters P_j for its description, resulting in high memory and computing time requirements and in excessive rounding-off errors in the matrix calculations involved. Furthermore, such a large number of independent parameters is usually not required, since mathematical relations are known to exist between the parameters of several individual peaks in the spectrum; for example, it may be known that the two peaks in a quadrupole doublet have equal areas and/or widths.

Some programs solve the problem by not using the simple Lorentzian formula (8) and (9) to describe the peaks, but a relation describing a whole group of related peaks simultaneously. As an example, the following formula can be used for a quadrupole doublet:

$$D_j(x_i) = (A_j/\pi\Gamma_j)/\{1 + [2(v_i - I_j + \tfrac{1}{2}Q_j)/\Gamma_j]^2\} + (A_j/\pi\Gamma_j)/\{1 + [2(v_i - I_j - \tfrac{1}{2}Q_j)/\Gamma_j]^2\} \tag{12}$$

where A_j is the total intensity of doublet D_j, I_j the isomer shift, Q_j the quadrupole splitting, and Γ_j the width of both peaks. These are the four independent parameters needed to describe the two peaks. Such an approach, however, results either in an extremely complex fitting function which can

accommodate all possible relations between peak parameters, or in a limited field of application, e.g. the program will only be able to resolve spectra consisting exclusively of quadrupole doublets with equal widths and areas for both peaks.

In the program described here, a different strategy was therefore followed. Every spectrum parameter S_j is regarded as being defined by a function g_k of n_k component terms C_m:

$$S_j = g_k(C_m, C_{m+1}, C_{m+2}, \ldots) \tag{13}$$

Every component term C_m can be set equal either to a constant or to a variable fitting parameter P_n. The same function can, of course, define several spectrum parameters, by using the same or different component terms. At present, six such parameter-defining functions are included in program ANMOS1:

$$g_1 = C_m \;;\; g_2 = C_m C_{m+1} \;;\; g_3 = C_m + C_{m+1}C_{m+2} \;;\; g_4 = C_m + C_{m+1}C_{m+2} + C_{m+3}C_{m+4} \;;\; g_5 = C_m C_{m+1}(1 + C_{m+2}) \;;\; g_6 = C_m C_{m+1}(1 - C_{m+2}) \tag{14}$$

Functions g_1, g_2 and g_3, which of course are only special cases of function g_4, were included to save memory space and/or computing time. These six functions are sufficient to describe all spectra measured so far in this laboratory. For example, function g_3 can be used to describe the positions of two quadrupole doublet peaks:

$$\mu_i = I_j \pm \tfrac{1}{2}Q_j \tag{15-1}$$

with isomer shift I_j and quadrupole splitting Q_j as fitting parameters or constants, for example:

$$\mu_2 = C_m + C_{m+1}C_{m+2} \tag{15-2}$$

with $C_m = P_n$, $C_{m+1} = -\frac{1}{2}$, and $C_{m+2} = P_{n+1}$.

Function g_4 can be used to define the position of a peak in a magnetically split sextet, for example:

$$\mu_j = s_C + (\tfrac{3}{2}g_e - \tfrac{1}{2}g_g)\beta B + e^2Qq/4 \tag{16}$$

where g_e and g_g are the nuclear g-factors, and where the centre shift s_C, the magnetic flux density times the Bohr magneton βB, and the quadrupole splitting parameter e^2Qq can be selected as variable parameters. Functions g_5 and g_6 can be used to describe the well-known relations

$$A_1 = A_6 = \frac{3A}{16}(1 + \cos^2\theta)$$

$$A_2 = A_5 = \frac{4A}{16}\sin^2\theta \tag{17}$$

$$A_3 = A_4 = \frac{A}{16}(1 + \cos^2\theta)$$

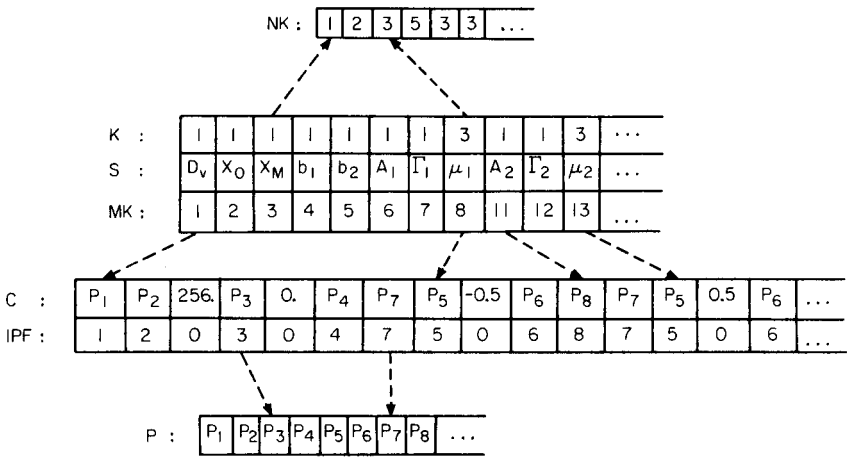


Fig. 3. Part of program memory structure. (S) Array of spectrum parameters; (K) pointers to function g_k used to define this spectrum parameter; (MK) pointers to first component term C_m used by g_k; (C) array of component terms C_m; (IPF) fitting parameter identification flags; (P) array of fitting parameters; (NK) array containing number of component terms n_k used by each function g_k.

As this idea of spectrum parameter-defining functions may be somewhat unusual, a schematic representation of part of the memory structure required by the program is shown in Fig. 3. It refers to a two-peak example where the positions are determined by function g_3 and all the other spectrum parameters by function g_1 as follows:

$$D_v = P_1 \; ; \; x_0 = P_2 \; ; \; x_M = 256. \quad b_1 = P_3 \; ; \; b_2 = 0. \quad A_1 = P_4 \; ; \; A_2 = P_8$$

$$\mu_1 = P_5 - \tfrac{1}{2}P_6 \; ; \; \mu_2 = P_5 + \tfrac{1}{2}P_6. \quad \Gamma_1 = \Gamma_2 = P_7$$

It should be noted that there is absolutely no requirement for linearity of the functions g_k; all the component terms in, for example, function g_5 may be variable parameters. Hence, there are also no objections to the use of relations such as:

$$\mu_j = s_C \pm \frac{e^2 Qq}{4}(1 + \eta^3/3)^{1/2} \tag{18}$$

with parameters s_C, e^2Qq and/or η, as spectrum parameter-defining functions. In fact, the program has been conceived in such a way that the insertion of new functions g_k requires only a minimal change; a single statement should be added to define the function g_k, and n_k statements defining the derivatives of g_k with respect to each of its component terms C_m.

PARAMETER LIMITATIONS AND FLEXIBLE CONSTRAINTS

Once the theoretical model has been properly defined, the problem consists in optimizing the parameters, so as to obtain the best possible agreement

between model and experimental data. Currently, this is done by minimizing chi-square

$$\chi^2 = \sum_{i=1}^{n_D} \{(1/y_i)[y_i - f(x_i)]^2\} \tag{19}$$

where the y_i represents the n_D spectrum data points, assuming Poisson statistics with $\sigma_{y_i}^2 = y_i$. However, the fit with the lowest χ^2 value will frequently be unacceptable, because some of the resulting parameter values are physically impossible. Moreover, the unrealistic values will quite often lead to many unpleasant conditions, such as floating overflows or floating zero divides. Therefore, additional information has to be passed to the program, i.e. limitations must be imposed on the parameters to keep them within a reasonable range.

An often-used type of constraint consists of requiring that some or all of the parameters obey a number of linear relations:

$$\sum_{k=1}^{n_P} (K_{j,k} P_k) = K_{j,0} \quad \text{for } j = 1, 2, \ldots \tag{20}$$

where the $K_{j,k}$ constants are selected by the user. In this case the fitting parameters P_k actually refer to physical spectrum parameters. Generally, these relations are only used to act on parameters of the same kind, e.g. to set the areas of two peaks equal to each other, or to keep certain parameters constant. In the program presented here, all constraints of this type can, of course, be applied completely by means of the spectrum parameter-defining functions described earlier, even without the annoying restriction of linearity.

Frequently, this type of limitation still does not yield satisfactory results; it may be undesirable to fix a certain parameter, whilst allowing it to vary freely or in fixed relationship to other spectrum parameters, may result in an unrealistic value. Some programs try to circumvent the problem by allowing the user repeatedly to fix and free the value of selected spectrum parameters, so as to lead the search manually in the desired direction. Although the possibility of doing this is included in program ANMOS1, the procedure is time-consuming and requires much manual intervention. It is therefore highly desirable to provide a limitation method which can act continuously on each of the fitting parameters separately. Some programs allow the user to select an interval or "cage" within which the value of the fitting parameter must always be contained; this method may be effective, but presents some serious disadvantages. Firstly, the values lying within the interval are all equivalent, which means that when the limitation becomes effective, a rather unlikely parameter value equal to one of the arbitrarily chosen limits will be selected, instead of one of the more likely values lying near the centre of the interval. Secondly, the limitations will sometimes prevent the minimization algorithm from finding the point of lowest χ^2, simply because the path on the χ^2 hypersurface leading towards it passes through a forbidden region. Finally, the program will not be able to calculate reliable error estimates for the limited and related parameters. In view of all these difficulties,

a novel and elegant limitation technique, described below, has been included in the program.

At the beginning of a least-squares search, estimated values E_j for the fitting parameters P_j must always be specified. These values, whether they were determined in an earlier measurement or found in the literature, are always the results of one or more experiments. They can therefore be considered as experimental data with an associated uncertainty σ_{Ej}, in exactly the same way as are the spectrum data points $y_i \pm \sigma_{yi}$. A new definition for χ^2 can therefore be given:

$$\chi^2 = \sum_{i=1}^{n_D} \{(1/y_i)[y_i - f(x_i)]^2\} + \sum_{j=1}^{n_P} [(1/\sigma_{Ej}^2)(E_j - P_j)^2] \qquad (21)$$

This definition implies that different "fitting functions" are used for different data points, e.g. the function $h(P_j) = P_j$ is used to describe data point $E_j \pm \sigma_{Ej}$, and that all these functions have to be optimized simultaneously. This is not so unusual; indeed, for a triangular reversed scan spectrum, two different functions are already used to describe the spectrum data points: eqn. (10) combined with eqns. (1) and (5), for data with $x_i \leqslant x_M$, and eqn. (10) combined with eqns. (2) and (6), for data with $x_i > x_M$.

Consideration of the effect of this new definition of χ^2 on the χ^2 hypersurface may be helpful in understanding the functioning of this type of parameter limitation. As an example, consider the two peaks shown in Fig. 4, where the solid lines represent the "best fit" on the basis of eqn. (19). The first peak is small and poorly defined, whereas the other one is large and well defined. The solid lines in Fig. 5 represent a section through the χ^2 hypersurface, as defined by eqn. (19), for the two peaks, at a constant background, amplitude and width. It is apparent that for the small peak the surface is very flat, making an accurate determination of the peak centroid velocity impossible, and with a minimum at approximately 5.3 mm s^{-1}. If it is known from other experiments that this parameter must be equal to 5.10 ± 0.01 mm s^{-1}, eqn. (21) can be used to compute the modified χ^2 surface, which results in the broken line in Fig. 5(a); clearly, the surface curvature is strongly

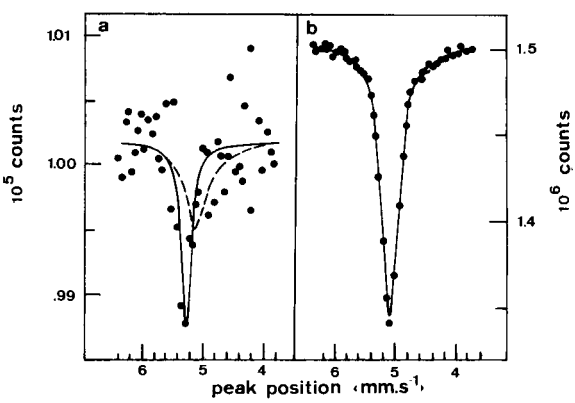


Fig. 4. Fit to small peak (a) and large peak (b).

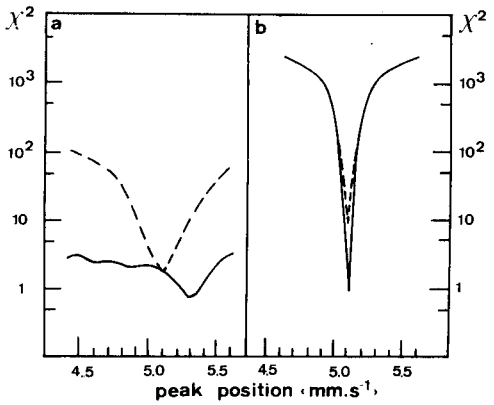


Fig. 5. Section through χ^2 hypersurface for the peaks in Fig. 4.

increased, and the minimum is shifted in the direction of the expected value.

The χ^2 surface (eqn. 19) for the large peak is strongly curved, with a pronounced minimum at 5.10 mm s^{-1}. On using eqn. (21), even with an erroneous centroid estimate of 5.30 ± 0.01 mm s^{-1}, the broken line in Fig. 5(b) will result, showing that this time the minimum is not shifted towards the expected value. Only the erroneous parameter estimate is reflected in a higher χ^2 value. The "best fits" on the basis of eqn. (21) are represented as broken lines in Fig. 4, illustrating the shift towards the expected position for the small peak, and no visible change for the large peak.

From these two examples, it is apparent that the limitations implied by the new χ^2 definition are flexible constraints which become effective only when they are needed, i.e. for statistically poorly defined parameters, as would be the case for very small or strongly overlapping peaks. For parameters P_j that can be accurately determined from the spectrum data, the term $(1/\sigma_{E_j}^2)(E_j - P_j)^2$ will be small compared to the others, and the χ^2 function will behave as though no limitation were present.

The only obvious objection that could arise to the use of such a limitation method is that the standard deviation σ_{E_j} will sometimes not be known. Indeed, for a rigorous mathematical treatment, σ_{E_j} should be the estimate for the width of the gaussian distribution of estimated values E_j around the real value. However, the use of a not too crude estimate for σ_{E_j} in this case will be a valid approximation and will always result in a more reliable optimization of the parameters than would either one of the two alternatives, i.e. keeping the parameter fixed or allowing it to vary freely in a selected interval [6].

Throughout program ANMOS1, formula (21), which overcomes all the disadvantages connected with working with cages, is therefore used as the definition of χ^2, and consequently all fitting parameters must be initialized with an error estimate. It is still possible, however, to leave a parameter completely free by specifying a very large value for σ_{E_j}; in fact, initial values for the parameters determining the peak area are usually chosen as 0 ± 10^9.

MINIMIZATION OF χ^2

The minimization algorithm used by the program is based mainly on the widely used linearization technique [7]. Since the modified χ^2 definition results in some minor modifications to the calculations involved, a short description of the method will be given.

If the initial estimates for the parameters are fairly close to the real values, the χ^2 function (or, alternatively, the fitting function) can be expanded as a first-order Taylor series:

$$\chi^2 = \chi_0^2 + \sum_{k=1}^{n_P} \left(\frac{\partial \chi_0^2}{\partial P_k} \Delta P_k \right) \tag{22}$$

where χ_0^2 represents the χ^2 value at the starting point and ΔP_k the increments for the parameters P_k. Setting the derivatives of χ^2 with respect to the parameters P_j equal to 0 will now yield a set of linear equations for the increments ΔP_k:

$$\sum_{k=1}^{n_P} (\alpha_{j,k} \Delta P_k) = \beta_j \quad \text{for } j = 1, 2, \ldots, n_P \tag{23}$$

with

$$\alpha_{j,k} = \frac{1}{2} \frac{\partial^2 \chi_0^2}{\partial P_j \partial P_k} = \sum_{i=1}^{n_D} \left[\frac{1}{y_i} \frac{\partial f_0(x_i)}{\partial P_j} \frac{\partial f_0(x_i)}{\partial P_k} \right] \quad \text{for } j \neq k$$

$$\alpha_{j,j} = \frac{1}{2} \frac{\partial^2 \chi_0^2}{\partial P_j^2} = \sum_{i=1}^{n_D} \left\{ \frac{1}{y_i} \left[\frac{\partial f_0(x_i)}{\partial P_j} \right]^2 \right\} + \frac{1}{\sigma_{Ej}^2} \tag{24}$$

and

$$\beta_j = -\frac{1}{2} \frac{\partial \chi_0^2}{\partial P_j} = \sum_{i=1}^{n_D} \left\{ \frac{1}{y_i} [y_i - f_0(x_i)] \frac{\partial f_0(x_i)}{\partial P_j} \right\} + \frac{1}{\sigma_{Ej}^2} (E_j - P_j)$$

neglecting the second derivatives of $f_0(x_i)$, representing the fitting function at the starting point. The solution of these equations yields a new set of parameters P_k':

$$P_k' = P_k + \sum_{j=1}^{n_P} (\alpha_{j,k}^{-1} \beta_j) \tag{25}$$

where α^{-1} represents the inverse of the matrix α, and the whole process is repeated until consistent values are obtained; at that time, error estimates for the parameters and related quantities can be calculated from

$$\sigma_{P_k}^2 = \alpha_{k,k}^{-1} \quad \text{and} \quad \sigma_{P_k P_j}^2 = \alpha_{k,j}^{-1} \tag{26}$$

As pointed out above, the linearization algorithm is only valid if the initial guesses are not too far away from the real values. Since this condition is not always met, the program performs a gradient search [7] whenever the linearization method fails, thereby extending the valid range of the mini-

mization algorithm. Experimentally, it has been found that this switching between two techniques is more efficient, regarding memory requirements and/or computing time, than the often-used gradient-expansion algorithm [8]. A flow chart of the complete minimization routine, which performs one iteration step, is shown in Fig. 6.

The derivatives in eqns. (24) are computed analytically as

$$\frac{\partial f_0(x_i)}{\partial P_j} = \sum_{k=1}^{n_S} \left[\frac{\partial f_0(x_i)}{\partial S_k} \sum_{m=M_k}^{M'_k} \left(\frac{\partial S_k}{\partial C_m} \frac{\partial C_m}{\partial P_j} \right) \right] \tag{27}$$

where M_k and M'_k represent the number of the first and last component term C_m used to define S_k, and where $\partial C_m/\partial P_j$ is equal to 1 or 0, depending on whether C_m is set equal to P_j or not.

PROGRAM STRUCTURE

At present, the program can accommodate up to 1024 spectrum data points, 60 independent fitting parameters and 155 spectrum parameters (50 peaks) defined by 300 component terms. When a suitable overlay structure is used, it has a memory requirement of 20K 16-bit words, approximately half of which is used for variable storage.

Program operation starts with reading the spectrum data, selecting the spectrum regions to be used in the fit, and initializing the calibration and background parameters. Next, peak data are entered. A general mode peak

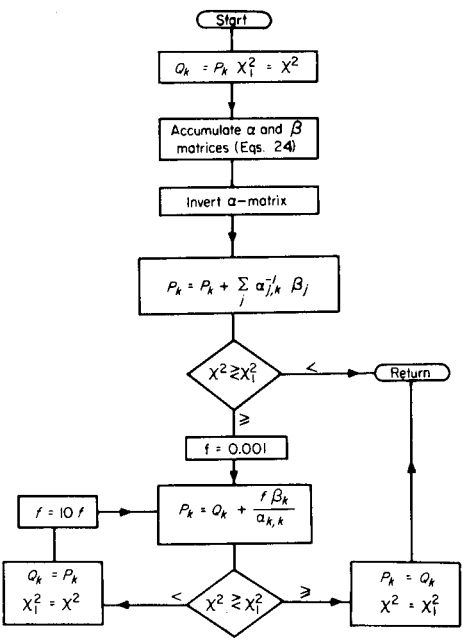


Fig. 6. Flow chart of χ^2-minimization algorithm.

input routine is available, where the user can select the parameter-defining function g_k and all the component terms C_m for each peak-defining parameter, i.e. area or amplitude, width, and position; for operator convenience, however, three small routines are included which handle the input for single peaks, quadrupole doublets and magnetically split sextets with quadrupole interaction, respectively, using default functions g_k and constants C_m. In addition, most of the initial parameter values and/or constants can be taken as defaults if not specified by the user; as examples, the initial value of the background parameter b_1 can be taken as the contents of the first data channel $\pm 10^9$, and peak area parameters are usually initialized as 0 ± 10^9. The latter initialization will cause the first call to the χ^2 minimization routine to act as a linear fit for the areas, all other derivatives (i.e. those with respect to position and width-determining parameters) in eqns. (24) being equal to zero, thus yielding accurate starting values for the next iterations. Input typing can thus be kept to the minimum.

After reading the input, the program will wait for user commands. These can instruct the program to perform a number of χ^2-minimization iterations, to type information on the current parameter values, or to change or fix spectrum and/or fitting parameters. Commands are also available to delete certain peaks or parameters or to insert new ones in the spectrum description, to use a different calibration or background definition, to use different or additional spectrum regions, or even to read an entirely new spectrum. At any time during the fitting process, output tables and/or plots of spectrum and fit can be printed.

Provision is made to use only part of the data points for the calculations, so as to speed up program execution. At the beginning of a fitting procedure, only every second or third data point is used, and derivatives with respect to peak parameters are only calculated in the immediate vicinity of the peaks; when convergence is almost reached, all data points are used, and derivatives will be calculated in the entire spectrum regions selected, so as to obtain optimum accuracy. This speed control feature can work either automatically, based upon the current χ^2 value, or under user control. It allows the deconvolution of even complex spectra in an acceptable calculation time. The analysis of moderately complex spectra can usually be performed in 2—3 min. For the extremely complex example shown in Fig. 7, containing 213 data points and 24 peaks described by 53 independent parameters, the time required to come from the bad initial estimate shown in part (a) of the figure to the final result in part (b) was only 14 min, when a PDP 11/04 minicomputer which has no floating point hardware was used.

DISCUSSION AND CONCLUSIONS

The program has been successfully used for the data reduction of a number of ^{57}Fe spectra obtained with a commercial Mössbauer system. The concept of spectrum parameter-defining functions, together with the χ^2 definition

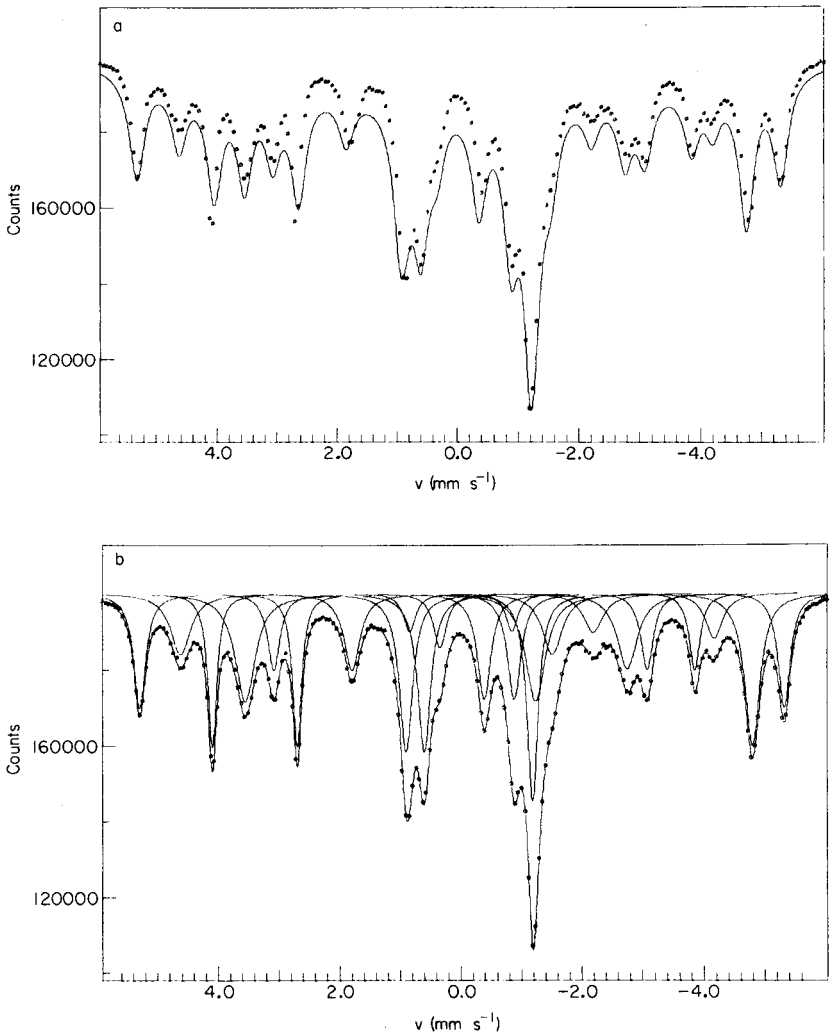


Fig. 7. Fit to a complex spectrum: (a) initial estimate, (b) final result.

used, make it possible for the program to take advantage of virtually all a priori physical information about the spectrum in a very straightforward way; the widely used practice of repeatedly fixing and freeing certain parameter values can therefore largely be avoided. Even vague affirmations such as "the width of the two peaks will differ by no more than 10%" can be entered into the least-squares algorithm by selecting $\Gamma_1 = P_1$ and $\Gamma_2 = P_1P_2$ with $P_2 = 1 \pm 0.05$, using functions g_1 and g_2 (eqns. 14).

Another distinct advantage is that all quantities which are not known with infinite precision can be entered as variable parameters. As an example, both the calibration parameters x_0 and D_v, and the peak positions μ_j can be

selected as fitting parameters. In this way, the uncertainties on the calibration parameter values will be reflected in the error estimates for the peak parameters, calculated by means of eqns. (26); if no χ^2-based limitations were used, the results would almost certainly diverge to something like $x_0 = -\infty$ and $\mu_j = +\infty$, whilst using a constant calibration or imposing cages on the parameters would make it impossible to obtain reliable error estimates for the calculated peak areas, widths and positions directly.

The modest memory and calculation time requirements, which do not inhibit the resolving power for complex spectra, make the program well suited for use on a mini-computer system. It should therefore become a valuable aid for Mössbauer spectroscopy. Some of the features described above could also be of use for the development or improvement of other least-squares programs.

REFERENCES

1 W. Kündig, Nucl. Instrum. Methods, 48 (1967) 219.
2 H. Bokemeyer, R. Meyer, K. Wohlfahrt and W. Wurtinger, Laborbericht 49, Institut für Kernphysik der Technischen Hochschule Darmstadt, 1971.
3 R. Grimm and W. Müller, Laborbericht 1/76, Institut für Anorganische Chemie und Analytische Chemie der Johannes-Gutenberg-Universität, Mainz, 1976.
4 F. James and M. Roos, Minuit-computer code, CERN, Geneva, programme D-506.
5 A. J. Stone, Program MOSSBR, Quantum Chemistry Program Exchange, QCPE Program 276.
6 H. Nullens, P. Van Espen and F. Adams, X-Ray Spectrom., 8 (1979) 104.
7 P. R. Bevington, Data Reduction and Error Analysis for the Physical Sciences, McGraw-Hill, New York, 1969.
8 D. W. Marquardt, J. Soc. Ind. Appl. Math., 11 (1963) 431.

Analytica Chimica Acta, 122 (1980) 387—394
Computer Techniques and Optimization

THE COMPUTERIZED DETERMINATION OF DOUBLE-LAYER CAPACITANCE WITH THE USE OF KALOUSEK-TYPE WAVEFORMS AND ITS APPLICATION IN TITRIMETRY

M. BOS

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

(Received 10th January 1980)

SUMMARY

A method for the rapid determination of double-layer capacitance—potential curves of electrodes is described. An on-line computer is used to apply Kalousek-type waveforms to the electrochemical cell and to measure the accompanying current response. The capacitances are determined from the slope of the plots of log current against time. For 0.1 M KCl, the computerized method agrees well with the bridge method, except for the potential range of 0 to —0.15 V. The method is very useful for automating titrations with tensammetric detection of the end-point. The method is applied to the titration of barium with a macrocyclic compound (kryptofix 222) and the titration of cetyl-trimethyl-ammonium bromide with bromocresol purple. The accuracy of the titrations is ±2%.

Differential capacitance—potential curves of electrodes play an important role in studies of the electrical double layer. Tensammetry is the electroanalytical technique based on the measurement of these curves [1]. Its use has not been widespread, mainly because of the complex relationship between capacitance and concentration of the compounds of interest and also because of the complex equipment required for the measurements. If the technique is used to determine titration end-points, quantitative results can be obtained from the stoichiometry of the titration reaction [2] and the first drawback can thus be circumvented. The computerization of the method described here is meant to simplify the measurements to such an extent that routine application will be possible.

Various methods have been described for the determination of the capacitance of the electrical double layer [3—6]. The impedance bridge method is still believed to be the most accurate one, but it is very time-consuming. All other methods described so far rely on measurements on an oscilloscope screen, which are also rather tedious. Phase-selective tensammetry [7] is an exception in that aspect as there are direct-recording instruments for this technique. However, its use for end-point detection in titrations still requires many manual operations.

This paper describes a computer technique for the determination of the

capacitance of the electrical double layer and its use in computer-controlled titrations.

THEORY

Capacitance measurements

In the absence of faradaic currents, the electrical equivalent of an electrochemical cell consists of a capacitor C and a resistor R connected in series.

If a voltage step of magnitude ΔV is applied to the cell, its current response, from charging of the capacitor, will be

$$i_c = (\Delta V/R) \exp(-t/RC) \quad (1)$$

For a periodic square-wave signal with a period much greater than RC, the current response is given in Fig. 1. At the end of both half-periods, the current has dropped to zero and the voltage across the capacitor has reached the applied voltage, either V_t or V_b. Equation (1) can be plotted logarithmically to give a straight line:

$$\ln i_c = \ln(\Delta V/R) - (t/RC) \quad (2)$$

The slope of this plot is $-(RC)^{-1}$. To be able to determine the double-layer capacitance at a known predetermined value of the d.c. potential V_t of the dropping mercury electrode (DME) the following conditions should be fulfilled: $t \ll RC$ and $T/2 \gg RC$. This means that the current can be measured only during a short interval after the reversals, whereas the frequency of the square wave should be limited to low values. Actual values for these limits will depend on the series resistance and the capacitance of the circuit. For accurate determination of the double-layer capacitance, the current measure-

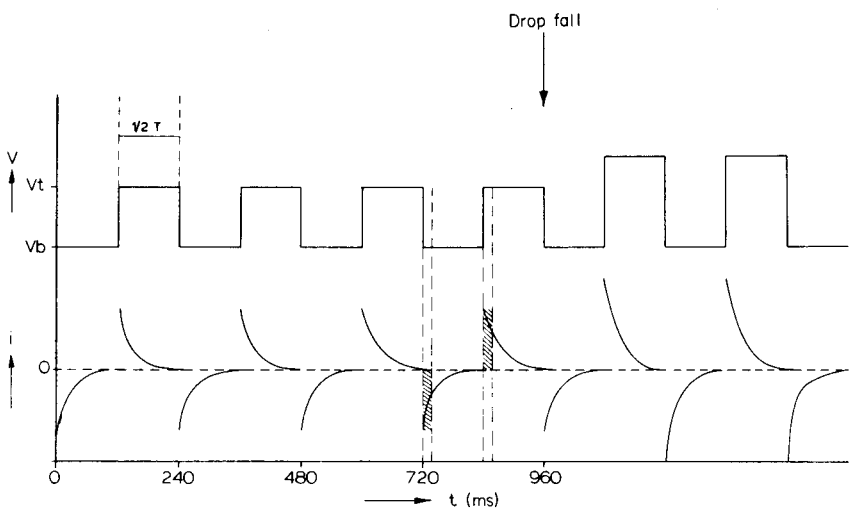


Fig. 1. Applied voltage and current—time patterns in the absence of faradaic currents.

meants should be performed after a reversal near the end of the drop life where the relative change of its surface can be neglected.

In the absence of faradaic currents, one pulse and current measurements directly after the flank of the pulse will suffice to determine the double-layer capacitance. The application of Kalousek-type waveforms offers no special facilities in this case. However, in the presence of electroactive species, capacitance changes at V_b (Fig. 1) should be observed if tensammetrically active substances are being produced during the V_t half-period. For reversible electrode reactions, pseudo-capacitance peaks will be observed in the $C(V_b)$—V_t curves. These can be distinguished from the tensammetric processes by observation of the corresponding Kalousek polarograms: type II polarograms show a wave in the same region.

Titrations with tensammetric end-point detection

Tensammetry has been used for end-point detection in various types of titrations [2, 8]. This type of end-point detection requires that the titration reaction produces or liberates a tensammetric compound or that it consumes such a compound. A distinction should be made between the two different ways of assessing the concentration of a compound tensammetrically, i.e. (a) by measuring the height of the adsorption/desorption peaks and (b) by measuring the lowering of the double-layer capacitance in the potential region of maximum adsorption. Normally method (a) is more sensitive and more selective, but sometimes a compound that is tensammetrically active does not show adsorption/desorption peaks. Then measurement of the lowering of the capacitance of the double layer becomes necessary.

Generally, the adsorption isotherms show a straight line at low concentrations [9]. This gives titration plots from which the end-point can be found as the intersection of two linear parts for titrations of a tensammetrically active substance and for titrations with a tensammetrically active titrant.

EXPERIMENTAL

Chemicals

Potassium chloride (Merck, zur Analyse), mercury (Drijfhout, polarographic grade), alizarin complexone dihydrate (Aldrich), cetyltrimethylammonium bromide (Merck, zur Analyse) and bromocresol purple (Merck, indicator grade) were used as received. Throughout the experiments, deionized water which was also filtered through Millipore Q2 filters was used.

Equipment

A diagram of the equipment is shown in Fig. 2. A Radiometer polarographic stand (type E64) equipped with a drop-life timer (type DLT1) was used. The dropping mercury electrode had the following characteristics: h = 52 cm, m = 2.58 mg s^{-1}, t = 3.25 s (1 M KCl, open circuit). Computer control of the Mettler motor burette (type DV11) was established by connecting its trigger terminals to the D/A converter of the computer.

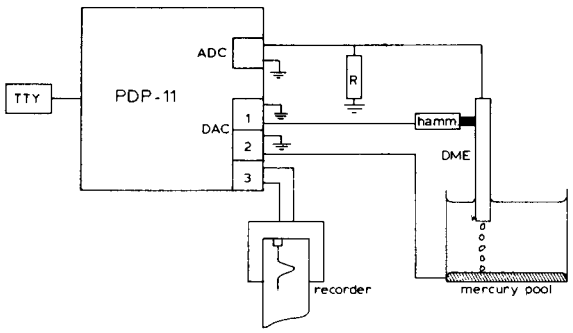


Fig. 2. Equipment for computerized titrations with tensammetric end-point detection.

Procedures

All solutions were freed from oxygen by bubbling with nitrogen for 10 min. During the titrations, the titrant stream and the contents of the cell were mixed by means of the nitrogen stream, except for the titrations of cetyltrimethylammonium bromide where magnetic stirring was used. The time between two successive titrant additions was at least 10 s, to allow for mixing of the titrant. The end-points of the titrations were found by calculating the intersection point of two least-square lines in the titration plots.

COMPUTER PROGRAMS

The program for acquiring the current—time data was the same as used in the KALOUSEK program [10], with the number of pulses per drop-life time restricted to 4. For automatic recording of double-layer capacitance as a function of potential, the data stored by this program on disc were used by a second program that calculated the capacitance values by a least-squares fit of the data to eqn. (2). To fulfil the conditions mentioned for t and $T/2$, only the values of the current up to 4 ms after the reversals of the last period were used in the calculations (shaded area in Fig. 1). The current was sampled at a rate of about 3 kC, so that 12 points were available. The program outputs the results either in tabular form or on a strip-chart recorder. The double-layer capacitance at the pulse base potential and the value at the pulse top potential were plotted.

The operator has control over the following experimental parameters: drop time, pulse base potential and the values that determine the scan of the pulse top potential, i.e. starting value, scan rate and end value.

The program for the automatic titrations with tensammetric detection of the end-point used the same routines for data acquisition and calculations, but was extended by routines for controlling the burette and for averaging the capacitance values obtained for several successive mercury drops. It operated with a fixed pulse base and pulse top potential. The experimental parameters that can be adjusted by the operator are pulse base potential,

TABLE 1

Calibration of equipment with standard capacitors

Capacitance (μF)	Computer result			Mean
0.05	14.56	14.44	14.55	14.51
0.10	29.41	29.64	29.39	29.48
0.15	45.56	45.54	45.56	45.55
0.25	76.31	75.15	74.33	75.39
0.50	143.7	142.9	143.3	143.3
0.75	212.4	213.5	213.9	213.9
1.00	287.2	291.1	281.6	286.6

pulse top potential, drop time, number of mercury drops for which the capacitance values are to be averaged per titration step, titrant volume increment per titration step, and the total volume of titrant to be added during the titration.

RESULTS AND DISCUSSION

Calibration

To calibrate the method, the polarographic cell was replaced by a standard capacitor decade. For a number of known capacitance values the computer result for the slope of the plot of log i against t was obtained with a measuring resistor of 1 MΩ, a pulse base potential of 0 V and a pulse top potential of +1.0 V. The data are given in Table 1. The result can be expressed as capacitance (μF) = (computer result -1.824)/284.2. The value of the intercept (1.824) can be explained by the capacitance of the cable used for the computer connections.

Double-layer capacitance measurements for 0.1 M KCl

For potassium chloride, accurate values are available for the double-layer capacitance at the mercury electrode obtained by the bridge method [11]. These were used to check the computerized method. The comparison is given in Fig. 3. In the computerized measurements, a drop time of 960 ms, a measuring resistor of 1 MΩ and a pulse base potential of -1.6V were used. As can be seen, there is fair agreement except for the potential range of 0 to -0.15 V. No explanation can so far be offered for the discrepancies in this region.

Titration of barium with cryptand 2.2.2

Cryptand 2.2.2 reacts with barium ions to form a 1:1 BaL complex. The formation constant, $K = [LBa^{2+}]/[L][Ba^{2+}]$, has the value of $10^{9.5}$ as determined by Lehn and Sauvage [12]. Therefore, this reaction can be used for the titration of barium ions if a suitable method of end-point indication can

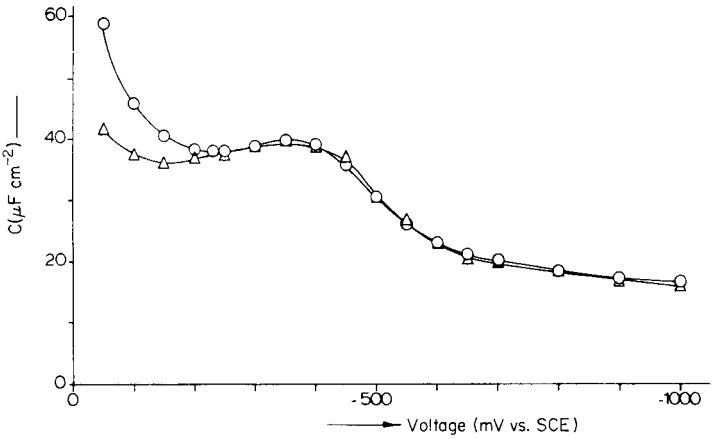


Fig. 3. Differential capacitance data for 0.1 M KCl (○) Grahame [11]; (△) this work.

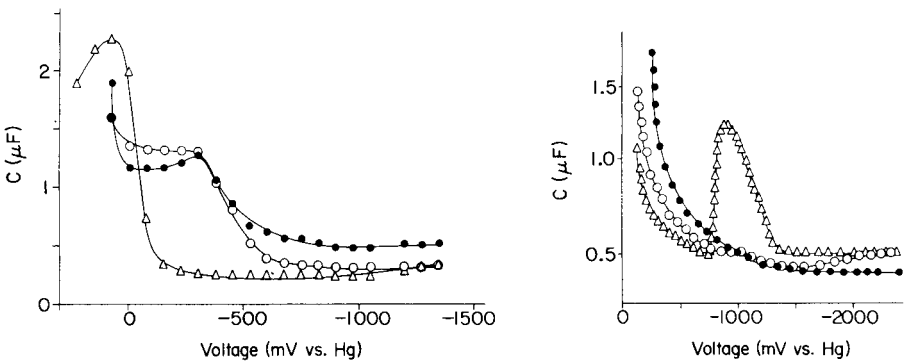


Fig. 4. Differential capacitances in 0.1 M LiCl for (•) 2×10^{-4} M Ba^{2+}, (△) 1.4×10^{-3} M cryptand 2.2.2, and (○) 2.4×10^{-4} M Ba—cryptand 2.2.2 complex.

Fig. 5. Differential capacitances in acetate buffer pH 4.62 for (△) bromocresol purple, (•) CTAB, and (○) bromocresol purple—CTAB complex.

be found. Tensammetric activity has been found [13] for some cryptand complexes. Therefore, it was expected that tensammetry could provide a means of end-point detection of this titration. Figure 4 shows the tensammograms of uncomplexed barium ions, the free cryptand 2.2.2 and a mixture of 1:1 barium and cryptand 2.2.2. The measurements were carried out in 0.1 M LiCl with a pulse base potential of —1.5 V at concentrations of 2.4×10^{-4}, 1.4×10^{-3} and 2.4×10^{-4} M. It can be seen that the presence of uncomplexed cryptand 2.2.2 can be detected from the capacitance values between —0.1 and —0.4 V. With a measuring potential of —0.4 V and a pulse base potential of —1.5 V, the double-layer capacitances were determined during titrations of various amounts of barium ions with 0.0100 M cryptand 2.2.2. Table 2 shows the results of the titrations.

TABLE 2

Tensammetric titration of barium with cryptand 2.2.2 in a 0.1 M LiCl medium

Barium taken (mg)	Barium found (mg)	Error (%)	Barium taken (mg)	Barium found (mg)	Error (%)
0.711	0.738	+3.8	1.401	1.408	+0.5
1.356	1.373	+1.3	1.684	1.717	+2.0
1.378	1.408	+2.1	2.037	2.094	+2.8
1.360	1.394	+2.5	2.836	2.884	+1.7

TABLE 3

Titration of cetyltrimethylammonium bromide with 0.00100 M bromocresol purple in acetate buffer pH 4.62

CTMAB taken ($\times 10^{-3}$ mmol)	CTMAB found ($\times 10^{-3}$ mmol)	Error (%)	CTMAB taken ($\times 10^{-3}$ mmol)	CTMAB found ($\times 10^{-3}$ mmol)	Error (%)
0.482	0.499	+3.5	1.00	0.97	−3.0
0.480	0.499	+4.0	2.00	1.97	−1.5
1.00	1.02	+2.0	2.00	2.00	0.0
1.01	1.04	+3.0	2.50	2.53	+1.2

Titration of cetyltrimethylammonium bromide with bromocresol purple

Cationic detergents form complexes with anionic dyes. These reactions are used in the titrations of detergents [14]. Normally the end-points are determined with the use of an organic solvent immiscible with water, e.g. chloroform. The ion-association complex is insoluble in water and colours the chloroform layer. Both the sample and the titrant of these two-phase titrations generally show tensammetric activity [9]. The feasibility of a tensammetric determination of the end-point will depend on the properties of the complex formed during the titration. From Fig. 5, it can be seen that the adsorption/desorption peak of bromocresol purple at −0.9 V is specific for the free bromocresol purple and can be used to monitor the titration of cetyltrimethylammonium bromide with this compound. Table 3 shows the results of the titration of various amounts of cetyltrimethylammonium bromide with 0.00100 M bromocresol purple. The titrations were carried out in 20 ml of 0.2 M acetate buffer, pH 4.62, at a pulse top potential of −0.9 V and a pulse base potential of 0.0 V at a drop time of 960 ms.

Conclusions

The use of an on-line computer in double-layer capacitance measurements allows routine application of these measurements as a means for the detection of titration end-points. Combination of these measurements with computer control of titrant addition offers an automated technique for the use of titration reactions in which reagents or products are adsorbed at the mercury/

electrolyte interface. The applications given show that this indication method is very useful for monitoring titrations with organic complexing reactions.

The author thanks Mrs. B. Verbeeten-van Hettema for preparing the manuscript and Mr. R. H. Arends for drawing the diagrams.

REFERENCES

1 B. Breyer and H. H. Bauer, Alternating Current Polarography and Tensammetry, Interscience, New York, 1963.
2 G. Nakagawa and T. Nomura, Anal. Lett., 5 (1972) 723.
3 J. W. Loveland and P. J. Elving, J. Phys. Chem., 56 (1952) 250.
4 J. J. McMullen and N. Hackerman, J. Electrochem. Soc., 106 (1959) 341.
5 C. C. Krischer and R. A. Osteryoung, J. Electrochem. Soc., 112 (1965) 735.
6 M. Brzostowska, M. Milkowska, A. Kalinowski and S. Minc, J. Electroanal. Chem., 89 (1978) 389.
7 A. K. Shallal and H. H. Bauer, Anal. Lett., 4 (1971) 205.
8 S. L. Gupta, S. K. Sharma and J. N. Jaitly, Ind. J. Chem., 4 (1966) 166.
9 H. Jehring, J. Electroanal. Chem., 21 (1969) 77.
10 M. Bos, Anal. Chim. Acta, 103 (1978) 367.
11 D. C. Grahame, J. Am. Chem. Soc., 71 (1949) 2975.
12 J. M. Lehn and J. P. Sauvage, J. Am. Chem. Soc., 97 (1975) 6700.
13 D. Britz and D. Knittel, Electrochim. Acta, 20 (1975) 891.
14 G. F. Longman, The Analysis of Detergents and Detergent Products, J. Wiley, New York, 1976, p. 258.

Analytica Chimica Acta, 122 (1980) 395—402
Computer Techniques and Optimization

POLAROGRAPHIC DETERMINATION OF STABILITY CONSTANTS AND THERMODYNAMIC PARAMETERS OF LEAD(II) PROPANOATE AND 2-HYDROXYPROPANOATE COMPLEXES WITH A COMPUTER-CONTROLLED SYSTEM

M. TKALČEC*, B. S. GRABARIĆ, I. FILIPOVIĆ and I. PILJAC

Laboratory of General and Inorganic Chemistry, Faculty of Technology, University of Zagreb, P.O.B. 179, 41000 Zagreb (Yugoslavia)

(Received 14th March 1980)

SUMMARY

A computer-controlled electrometric system is described. It is used for d.c. polarographic determinations of the stability constants of lead(II) propanoate and 2-hydroxypropanoate complexes at four temperatures. From the values of the monoligand complex stability constants obtained at different temperatures, standard thermodynamic functions (ΔH_j and ΔS_j) for the first and second steps of complex formation were obtained. Closed-loop interaction between the minicomputer and electrometric instrument was achieved through computer control of the potentiostat, drop-life timer, burette and valve for nitrogen purging. Computer programs are outlined for numerical and statistical evaluation of the experimental data giving $E_{1/2}$, i_d and slope of logarithmic presentation of polarograms, F_0 functions and cumulative stability constants, β_j, as well as for calculation of the standard thermodynamic functions.

Laboratory minicomputers and microprocessors have led to many important innovations in electroanalytical chemistry [1—3] as well as in many other instrumental techniques. Simply the fact that data can be stored in a computer memory for subsequent operations (numerical computation, smoothing, statistical fitting, numerical transformation, etc.) would fully justify their use and application, especially bearing in mind their relatively low cost. The more powerful features of computerized instrumentation such as input signal generation, real-time data acquisition, control and calculation which permit closed-loop feedback, render such systems even more valuable, as ongoing experiments can be optimized almost as soon as results are obtained.

In the present paper, an on-line computer-controlled electrochemical instrument with closed-loop interaction, based around a PDP-8/E minicomputer (OS-8/10 system) is described. Its use in the d.c. polarographic determination of the stability constants of metal ion complexes at different temperatures is discussed. Measurements of stability constants at different temperatures enabled the standard thermodynamic parameters ΔH_j and ΔS_j to be calculated.

Polarographic determinations of stability constants at constant temperature are very tedious when done by analogue methods. The recording of the polarograms at different ligand concentrations, the evaluation of $E_{1/2}$ and i_d, and the calculation of the experimental Nernst slope are time-consuming when done manually. Accurate final calculations of stability constants, β_j, by statistical least-squares fitting of $F_0 = 1 + \sum_{j=1}^{N} \beta_j [L]^j$ polynomials [4] cannot be achieved without advanced calculators or computers. An on-line computer-controlled electrometric system with a closed-loop interaction is ideally suited to fulfil all the above requirements, especially when the whole measuring and calculation procedure should be repeated at several temperatures.

EXPERIMENTAL

The major part of the system was the same as that used in the electrochemical study of lithium(I) interactions with some radical anions of organic electroactive species [5].

Apparatus

A block diagram of the computerized electrochemical system is shown in Fig. 1; a three-electrode polarographic cell is used. For precise control of the working electrode potential, a potentiostat with operational amplifiers was built, similar to earlier designs [6, 7]. The drop time of the mercury working electrode is set as an initial parameter for the computer program; it is controlled by a programmable real-time clock connected to an electromagnetic drop timer (DLT 1, Radiometer).

The circuitry for current measurements, which converts current to voltage, is an integral part of the potentiostat. The voltage is fed through an active low-pass filter (LPF) into a sample-and-hold amplifier (S/H 1). The corresponding potential from the voltage follower is also fed through another active LPF into another sample-and-hold amplifier. Potential and current are sampled at the end of the mercury drop life. In this way, sampling of potential and current is simultaneous and constant during analogue-to-digital conversion. The outputs of the amplifiers are connected via the scanner (SCAN) to the input of a digital voltmeter (DVM; Hewlett-Packard HP 3480A) which is used as an A/D converter with a resolution of ±0.1 mV for the full-scale range of 1.5 V which was used throughout the experiments. The DVM can be connected to a digital recorder (DR; Hewlett-Packard HP 5055) as a hard copy unit for measured data if required.

The system also contains a microburette (B) and an electromagnetic valve (V), both computer-controlled. The precision of direct reagent addition into the polarographic cell with the microburette was 0.001 cm^3. The valve is programmed to pass purified nitrogen through the solution to remove oxygen and mix the solution after ligand addition, prior to polarographic

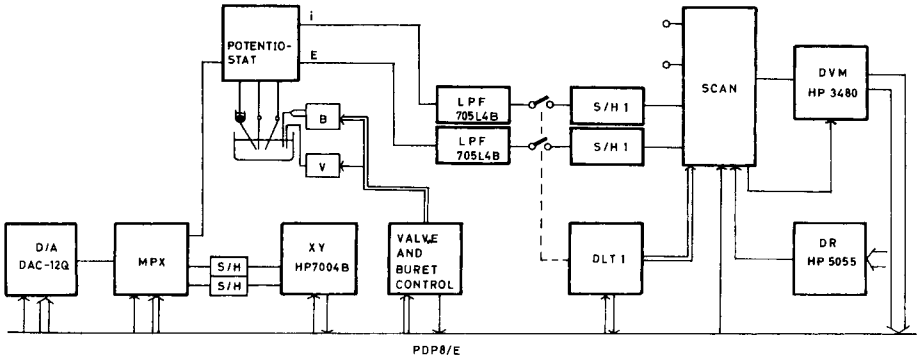


Fig. 1. Block diagram of computer-controlled electrochemical system.

measurements. These features eliminate any human intervention after the initial parameters have been set, and provide a complete closed-loop feedback computer control of the experiment.

The PDP-8/E (Digital Equipment Corp.) OS-8/10 computer system consists of: 8K word memory, RX8 dual floppy disc system, TD8 dual DECtape and LA30 DECwriter as an input/output device. The non-standard features of this system are a programmable real-time clock (DK8/EP) and general-purpose input/output interface boards. The standard precision of calculation of the system is seven significant digits within the range 10^{-38}—10^{38}. The algorithm for least-squares fitting of polynomials requires greater precision, and was evaluated with a precision of 10.5 significant digits.

The measured d.c. polarograms and their logarithmic transforms, as well as the final $\Delta E_{1/2}$, F_0 and/or F_1 as functions of ligand concentration [L], are displayed on an X-Y recorder (XY; Hewlett-Packard HP 7004B) completed with a null detector and point plotter. The recorder is connected to the computer through a 12-bit D/A converter (D/A), multiplexer (MPX) and sample-and-hold amplifiers (S/H).

Computer programs

Subroutines for control of all devices connected to the computer were written in SABR assembler language. The main program, which binds all the assembler subroutines and does the numerical and statistical evaluation of polarograms and weighted least-squares fitting of F_0 polynomials, was written in an extended version of FORTRAN II. A block diagram of the computer program is given in Fig. 2.

The main features of the computer program are as follows. Firstly, the potential range over which the polarograms are recorded changes automatically for the polarogram with the next ligand concentration depending on the shift in $E_{1/2}$. Secondly, the volume of the next addition of ligand solution is calculated via the feedback loop from the shift in $E_{1/2}$ (greater $E_{1/2}$ shifts cause addition of smaller volumes of ligand and vice versa), or the increment of ligand concentration is preselected according to a logarithmic function.

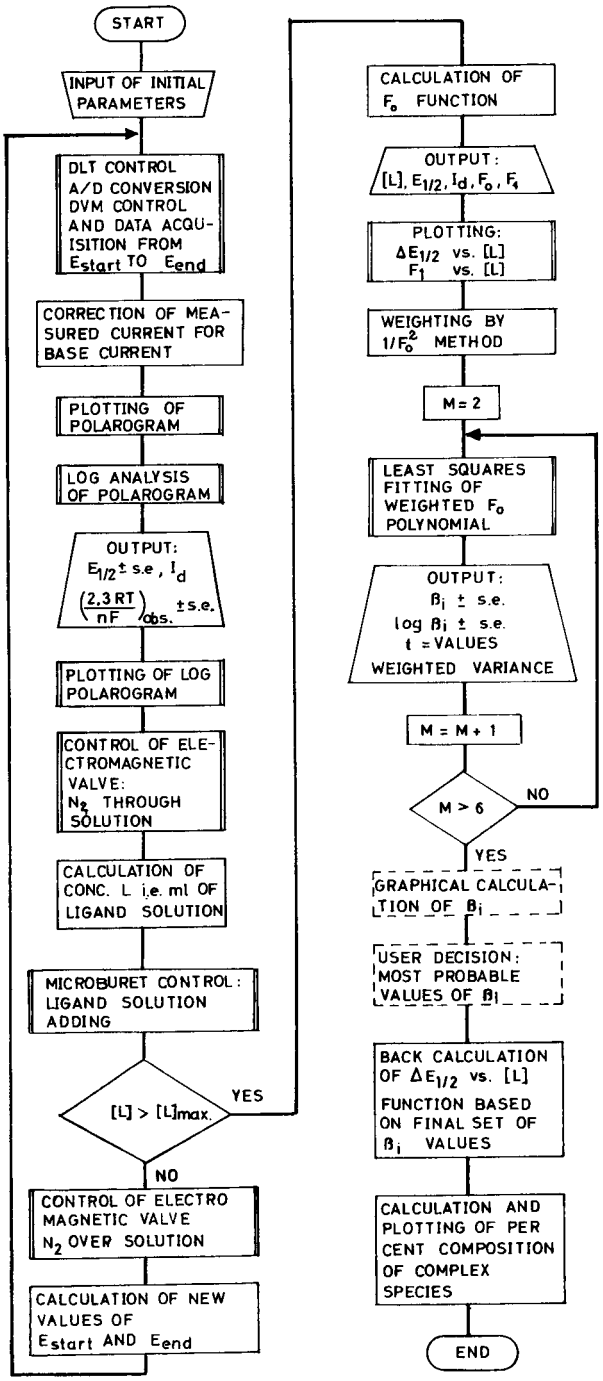


Fig. 2. Block diagram of computer program. Broken lines indicate optional user intervention; s.e. means standard error.

Thirdly, if an obviously erroneous point is sampled (erroneous mercury drop dislodgement), the program either eliminates this point or repeats the recording of the whole polarogram in the same solution. These three features provide a closed-loop interaction that optimizes the experiment on the basis of data recorded. The logarithmic analysis of polarograms and weighted least-squares fitting of F_0 polynomials were the same as described previously [4, 7]. Stability constants could also be obtained graphically by using the subroutine for calculation and plotting of Leden's F_j functions [8] with successively changing extrapolations within preset intervals in order to obtain, on the basis of F_j function shapes, more reliable approximate values of β_j. A further feature is that with the final set of stability constants, it is possible to calculate and plot the percentage of each successive complex species present within the experimental range of ligand concentration. With the final set of stability constants, it is possible to back-calculate $\Delta E_{1/2}$ and/or F_0 functions against ligand concentration together with corresponding confidence intervals. The final feature is the calculation of standard thermodynamic functions by a linear regression subroutine on ΔG_j as a function of temperature.

The program-chaining capability of the OS-8/10 system allowed all these features to be included into several programs chaining automatically.

Experimental conditions

All measurements were in buffered solutions with constant concentration (10^{-2} mol dm^{-3}) of propanoic or 2-hydroxypropanoic acid to avoid hydroxy complex formation. The ionic strength was kept constant (2 mol dm^{-3}) by adding sodium perchlorate. The lead(II) concentration in the test solutions was constant (0.4×10^{-3} mol dm^{-3}). All solutions were thermostated within 0.1 K. The reproducibility of the $E_{1/2}$ values was ±0.1 mV and that of the current measurements was ±0.2%. All the polarograms were reversible on a d.c. polarographic time scale; the experimental value of RT/nF was constant to within ±0.3 mV of the theoretical value at the corresponding temperature.

RESULTS AND DISCUSSION

Plots of $\Delta E_{1/2}$ against free ligand concentration [L], at four temperatures for lead(II) propanoate and 2-hydroxypropanoate complexes are given in Fig. 3. Logarithmic values of the stability constants, $\beta_j = [ML_j]/([M][L]^j)$ and $K_j = [ML_j]/([ML_{j-1}][L])$ with their standard errors are given in Table 1.

The standard state for any reacting species being defined as a hypothetical ideal solution of unit concentration (in aqueous 2 mol dm^{-3} $NaClO_4$ as solvent), standard ΔG_j values were calculated for the first and second steps of complex formation from the relation $\Delta G_j = -RT \ln K_j$. Linear plots were obtained of ΔG_j against T, for the first and second steps of complex formation ($\Delta G_j = \Delta H_j - \Delta S_j T$), and ΔH_j and ΔS_j were obtained for the above-defined standard state by linear regression (Table 1). Log K_3 values and consequently ΔG_3 values for both lead(II) propanoate and 2-hydroxypropanoate

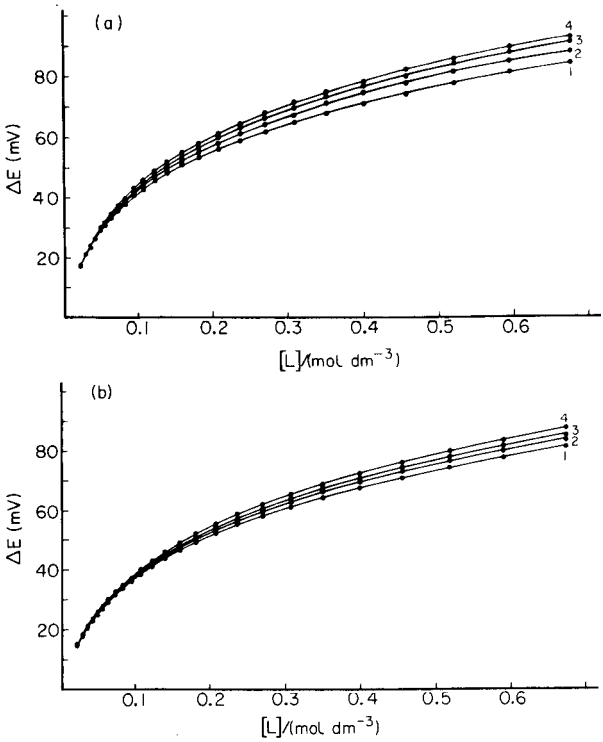


Fig. 3. $\Delta E_{1/2}$ against $[L]$ plots for (a) lead(II) propanoate complexes and (b) lead(II) 2-hydroxypropanoate complexes at various temperatures. Downwards, the four curves correspond to (a) 279.2, 288.7, 298.2 and 303.2 K, and (b) 276.7, 288.7, 298.2 and 307.7, respectively.

TABLE 1

Stability constants and standard thermodynamic parameters of lead(II) propanoate and 2-hydroxypropanoate complexes

T(K)	log K_1	log β_2	log K_2	log β_3	log K_3	$-\Delta G_1$ (kJ mol^{-1})	$-\Delta G_2$ (kJ mol^{-1})
Lead(II) propanoate complexes							
279.2	2.179 ± 0.008	3.24 ± 0.01	1.06 ± 0.02	3.23 ± 0.03	0.01 ± 0.02	11.64 ± 0.04	5.66 ± 0.1
288.7	2.133 ± 0.007	3.23 ± 0.01	1.10 ± 0.01	3.36 ± 0.02	0.13 ± 0.03	11.78 ± 0.04	6.08 ± 0.0
298.2	2.116 ± 0.008	3.22 ± 0.01	1.10 ± 0.02	3.45 ± 0.02	0.23 ± 0.03	12.07 ± 0.05	6.28 ± 0.1
303.2	2.107 ± 0.009	3.27 ± 0.01	1.16 ± 0.02	3.34 ± 0.02	0.07 ± 0.10	12.22 ± 0.05	6.73 ± 0.1
ΔH_1 = −4.7 ± 0.6 kJ mol^{-1}			ΔS_1 = 25 ± 2 J K^{-1} mol^{-1}				
ΔH_2 = 5.7 ± 1.4 kJ mol^{-1}			ΔS_2 = 41 ± 5 J K^{-1} mol^{-1}				
Lead(II) 2-hydroxypropanoate complexes							
276.7	2.115 ± 0.008	3.06 ± 0.02	0.94 ± 0.03	3.34 ± 0.02	0.28 ± 0.04	11.20 ± 0.04	5.00 ± 0.1
288.7	2.086 ± 0.009	3.05 ± 0.01	0.96 ± 0.02	3.34 ± 0.01	0.29 ± 0.02	11.52 ± 0.05	5.30 ± 0.1
298.2	2.052 ± 0.007	3.04 ± 0.01	1.00 ± 0.01	3.26 ± 0.02	0.22 ± 0.03	11.71 ± 0.04	5.71 ± 0.0
307.7	2.037 ± 0.008	3.06 ± 0.01	1.02 ± 0.02	3.24 ± 0.02	0.18 ± 0.03	11.99 ± 0.05	6.00 ± 0.1
ΔH_1 = −4.2 ± 0.2 kJ mol^{-1}			ΔS_1 = 25.1 ± 0.9 J K^{-1} mol^{-1}				
ΔH_2 = 4.2 ± 0.5 kJ mol^{-1}			ΔS_2 = 33 ± 2 J K^{-1} mol^{-1}				

TABLE 2

Stability constants of lead(II) propanoate complexes and lead(II) 2-hydroxypropanoate complexes in aqueous 2 mol dm^{-3} $NaClO_4$ solution at 298.2 K

log K_1	log β_2	log β_3	log β_4	Reference
Lead(II) propanoate complexes				
2.116 ± 0.008	3.22 ± 0.01	3.45 ± 0.02		This work
2.08 ± 0.06	3.35 ± 0.03			[15]
2.34	3.76	3.90	4.18	[14]
2.23	3.34	3.76		[16]
2.08	3.34			[17]
Lead(II) 2-hydroxypropanoate complexes				
2.052 ± 0.007	3.04 ± 0.01	3.26 ± 0.02		This work
2.16 ± 0.02	3.23 ± 0.02	3.67 ± 0.02		[9]
2.15	3.15	4.26	2.95	[14]
1.88	2.82	3.13	3.16	[13]
1.98	2.98			[11]
2.26	3.30	3.33		[12]
2.15	2.88(3.15)	4.30		[17]

complexes at different temperatures are small and some of them are statistically unreliable, so these values were not used as data for ΔH_3 and ΔS_3 calculations.

The stability constants of lead(II) propanoate and 2-hydroxypropanoate complexes obtained in this work are compared in Table 2 with values for the same systems obtained earlier in this laboratory and with literature values for the same medium, ionic strength and temperature. In general, the agreement is very good considering that different experimental methods were applied in different laboratories with different methods of evaluation (graphical and statistical).

The reproducibility of the stability constants obtained by the computerized polarographic method was similar to that obtained by a potentiometric method [9, 15] generally considered to be more precise, although the latter method was not experimentally computerized when the results were reported.

The stability constants of the monoligand complexes confirm the earlier statement that lead(II) 2-hydroxypropanoate complexes have much the same stability as unsubstituted propanoate complexes [9]. Our average values obtained by different methods at 298.2 K are log K_1 = 2.06 and 2.19 for the lead(II) 2-hydroxypropanoate and propanoate complexes, respectively. However, having lower basicity (estimated from the pK_a values of the corresponding monocarboxylic acids which were found potentiometrically to be 4.89 and 3.79 for propanoic and 2-hydroxypropanoic acids, respectively), lead(II) 2-hydroxypropanoate complexes may be slightly more stable owing to the hydroxyl group coordination to the metal ion; this effect is more pronounced for the same ligands bound to smaller metal ions such as Co(II), Ni(II), Cu(II) and Zn(II) [15].

The values of the standard thermodynamic functions (Table 1) and those obtained by calorimetric investigation of lead(II) 2-hydroxypropanoate complexes [10] agree well for the monoligand formation but badly for the biligand formation. Both results, however, suggest a small exothermic enthalpy change for monoligand complex and a small endothermic enthalpy change for the biligand complex, with almost the same magnitude of ΔH_j. ΔH_3 and ΔS_3 values can be obtained only unreliably by calorimetry and even more unreliably from polarographic data. The thermodynamic parameters obtained by the calorimetric method require three separate experiments (i.e. calorimetric, pH and stability constant) while thermodynamic parameters obtained by measuring stability constants at different temperatures require only one experiment. The values obtained earlier [10] for ΔH_j and ΔS_j were evaluated from potentiometrically determined stability constants [9].

The conclusion is that both lead(II) propanoate and lead(II) 2-hydroxypropanoate are weak complexes. The temperature dependence of their stability constants is consequently very small. Therefore, to obtain reliable thermodynamic data from polarographic (or any other) measurements of stability constants, instrumentation giving data of high quality is essential. The computerized electrometric system with closed-loop interaction described above has proved its reliability and suitability for such investigations. Its complete automation, after setting of the initial parameters, for measurements at constant temperature is advantageous for such tedious and time-consuming experimentation. The quality of the results obtained is comparable to, or better than, that obtained by other methods.

REFERENCES

1 J. S. Matheson, H. B. Mark, Jr. and H. C. MacDonald (Eds.), Computers in Chemistry and Instrumentation, Vols. I and II, M. Dekker, New York, 1972.
2 S. P. Perone and D. O. Jones, Digital Computers in Scientific Instrumentation (Application to Chemistry), McGraw-Hill, New York, 1973.
3 S. P. Perone, J. Chem. Educ., 47 (1970) 105.
4 I. Piljac, B. Grabarić and I. Filipović, J. Electroanal. Chem., 42 (1973) 433.
5 M. Tkalčec, I. Filipović and I. Piljac, Anal. Chem., 11 (1975) 1773.
6 E. R. Brown, T. G. McCord, D. E. Smith and D. D. DeFord, Anal. Chem., 38 (1966) 1119.
7 B. Grabarić, M. Tkalčec, I. Piljac and I. Filipović, Anal. Chim. Acta, 74 (1975) 147.
8 I. Leden, Z. Phys. Chem., Leipzig, Abt. A, 188 (1940) 160.
9 I. Kruhak, B. Grabarić, I. Filipović and I. Piljac, Croat. Chem. Acta, 48 (1976) 119.
10 I. Filipović, B. Bach-Dragutinović, N. Ivičić and Vl. Simeon, Thermochim. Acta, 27 (1978) 151.
11 H. Thun, W. Guns and F. Verbeek, Anal. Chim. Acta, 37 (1967) 332.
12 Z. Warnke and A. Basinski, Rocz. Chem., 39 (1965) 1776.
13 J. S. Savić and I. Filipović, Croat. Chem. Acta, 37 (1965) 91.
14 I. Filipović, I. Piljac, A. Medved, S. Savić, A. Bujak, B. Bach-Dragutinović and B. Mayer, Croat. Chem. Acta, 40 (1968) 131.
15 I. Filipović, T. Matusinović, B. Mayer, I. Piljac, B. Bach-Dragutinović and A. Bujak, Croat. Chem. Acta, 42 (1970) 541.
16 I. Filipović, A. Bujak, M. Marač, R. Novak and V. Vukičević, Croat. Chem. Acta, 32 (1960) 219.
17 E. Martell and R. M. Smith, Critical Stability Constants, Vol. 3, Plenum Press, New York, 1977.

Analytica Chimica Acta, 122 (1980) 403—419
Computer Techniques and Optimization

A SENSITIVITY ANALYSIS OF THE SMITH PREDICTOR CONTROLLER

CHARLES J. HERGET, CHARLES L. POMERNACKI and JACK W. FRAZER*

Lawrence Livermore Laboratory, University of California, Livermore, California 94550 (U.S.A.)

(Received 3rd December 1979)

SUMMARY

The sensitivity of the Smith predictor controller to variations in system parameters is analyzed. The parameters considered are system gain, time constant, and delay or dead time. Both simulated and experimental results are discussed. First- and second-order models are studied. For the first-order model, a proportional plus integral control is used with the Smith predictor. A proportional plus integral plus derivative control is used with the Smith predictor for the second-order system. The controller is implemented digitally. An application of the Smith predictor with a model of the system obtained by a nonlinear least-squares parameter identification method is demonstrated on an experimental system.

Efficient methods for controlling large-scale chemical processing plants, from waste treatment to petrochemical plants, are needed to reduce environmental pollution, improve product yield and quality, and conserve energy [1—3]. New processes which produce multiple products, such as coal gasification, would benefit especially from control techniques which would allow selection of product ratios.

The development of integrated circuit technology has resulted in inexpensive micro- and mini-digital computers powerful enough for sophisticated automation and control applications. In addition, many chemical instruments capable of on-line measurements and analysis are available and others are being developed. A profile of a process obtained with these instruments can be used directly by computers to control the process. In spite of this increased availability of hardware, the computer automation and control of chemical processes have proceeded slowly considering the potential benefits of successful implementation. Progress has been slow because of the complexity of chemical systems and the lack of communication between associated disciplines. Typical complexities are the large number of process variables, strong interaction among process variables, process delays, need for good control models for nonlinear distributed processes, need to incorporate control bounds in the controller design, and system disturbances and noise that are difficult to characterize.

In order to provide a systematic procedure to approach these difficult problems, a laboratory bench-scale apparatus with a problem environment similar to that found in a commercial pilot plant has been built, and an interdisciplinary team of chemists, engineers, and mathematicians has been organized at the Lawrence Livermore Laboratory. The goals are to be able rapidly to characterize complex chemical reactions, determine optimum operating conditions, and develop control algorithms to achieve these optimum conditions under adverse conditions.

This paper deals with the problem of time delays in process control. Preceding papers from this research project have treated the problem of characterizing chemical processes [4—6], and future papers will address a number of the other stated problems.

Time delays are known to have detrimental effects in feedback control systems. Smith [7] introduced a method for designing controllers to overcome the difficulties introduced by time delays. This method, which has come to be called the Smith predictor, requires that a model of the process and time delay be built into the controller. If in fact the actual process were to behave exactly as predicted by the model, the detrimental effects of the time delay could be eliminated totally; however, in practice, there is always a mismatch between the model and the actual process.

Several studies on the performance of the Smith predictor have been conducted. Donoghue [8] compared the performance of two controllers for a first-order system with a time delay. One controller is designed using the classical Smith predictor; the other is designed using the optimal linear regulator approach. An optimal controller is recognized to be optimal only in the sense of the performance criterion which it optimizes. Indeed, Donaghue showed that if the optimal linear regulator is subjected to an output disturbance (an occurrence not taken into account in the optimal formulation), then the Smith predictor performs better than the optimal linear regulator. Meyer et al. [9] have compared the Smith predictor with conventional proportional plus integral (PI) control for a system containing time delays in both the forward and feedback paths. They restricted their analysis to a first-order system. Based on simulated results, they concluded that the Smith predictor always gives improved performance over the PI control for setpoint changes, and this improvement becomes significant in systems in which the time delay in the system is large compared to the time constant associated with the system dynamics. They do point out that the Smith predictor may not provide any advantage over PI control against output disturbances.

The Smith predictor was originally formulated for linear, time-invariant systems consisting of a single input (control) and a single output. Several papers have shown that Smith's technique can be applied as well to digital control systems [10, 12] and multivariable systems [11—13]. Since the dynamics of chemical processes are seldom known precisely, a thorough study of the sensitivity of the Smith predictor to parameter variations should be made in the course of considering its implementation. Eisenberg [14]

has performed an analysis of the effects on system stability of various analog approximations for the time delay required in the system model. The effects of mismatch of a variety of system parameters in a digital control implementation will be studied in this paper.

In subsequent sections, a brief description of the system models which will be considered, a discussion of the Smith predictor and its implementation will be given, and results obtained from simulations will be presented. Typical time responses will be shown, and the effect of parameter variations on classical performance criteria will be presented followed by the results of experimental runs on the enzyme reactor. The effects of mismatch between the system and its model will be demonstrated and compared with those predicted by the simulation studies.

SYSTEM DESCRIPTION

There are a number of ways of explaining how the Smith predictor works. A brief explanation of the Smith predictor appropriate for a linear, time-invariant, single-input, single-output sampled data control system will be given here. The concepts can be generalized to a class of control problems containing delays, i.e., multiple-input, multiple-output systems, time-varying, and even nonlinear systems can be treated by the same approach.

The following notation will be adopted. Functions of time, e.g., $u(t)$ and $x(t)$, will be denoted by lower-case letters with t denoting time. The corresponding Laplace transforms of the time functions will be denoted by upper-case letters, e.g., $U(s)$ and $X(s)$, with s denoting the complex variable argument of the Laplace transform. The corresponding Z transforms will be denoted by upper-case letters using the complex variable z as the argument, i.e., $X(z)$ and $U(z)$.

Figure 1 illustrates the implementation of the Smith predictor in a linear, time-invariant, single-input, single-output, sampled data control system. The

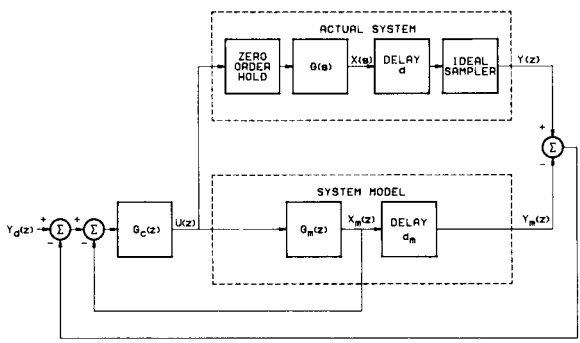


Fig. 1. Implementation of the Smith predictor. $Y_d(z)$ Set point; $U(z)$ control; $X(s)$ system output before delay; d system dead time; $Y(z)$ system output after delay and sample; $G_c(z)$ controller compensation network; $G(s)$ system transfer function; $G_m(z)$ model transfer function; $X_m(z)$ model output before delay; d_m model delay; $Y_m(z)$ model output after delay; T sampling period.

actual system to be controlled is shown within the upper set of dashed lines. It consists of an input, $U(z)$, a zero-order hold, the system transfer function, $G(s)$, the system output before delay, $X(s)$, a dead time of duration d, an ideal sampler with sampling period T, and the measured output, $Y(z)$. The input to the overall system is the setpoint, $Y_d(z)$. To implement the Smith predictor, a model of the system is simulated. The model is shown within the lower set of dashed lines. The input to the model is the same function, $U(z)$, which is the input to the true system. The transfer function $G_m(z)$ is the model approximation to the Z transform, $Z\{[(1 - e^{-sT})/s]G(s)\}$. The model output before delay, $X_m(z)$, is the model's predicted value of $Z\{X(s)\}$. Since $X_m(z)$ appears in the simulation, it is available to the controller whereas $X(s)$ is not. The model contains a time delay, d_m, with output $Y_m(z)$, the model's predicted value of $Y(z)$. The transfer function $G_c(z)$ is the compensation network, usually a digital implementation of proportional plus integral (PI) or proportional plus integral plus derivative (PID) control.

It is apparent that if the model transfer function, $G_m(z)$, is exactly equal to $Z\{[(1 - e^{-sT})/s]G(s)\}$, if the model delay, d_m, is exactly the same as the true system time delay, d, and if the initial conditions on the system model are the same as the initial conditions on the actual system, then $Y(z)$ and $Y_m(z)$ will be identical. If this is so, then the simplified form shown in Fig. 2 will be equivalent to the complete implementation shown in Fig. 1.

The advantage of the simplified model is that design of $G_c(z)$ can be done by classical control system methods. There is no time delay in the feedback loop. Classical design procedures can be applied, or standard methods for tuning PI and PID networks can be used [15, 16].

SIMULATION STUDIES

The system shown in Fig. 1 is studied by the use of simulation. Two forms of $G(s)$ will be considered; one is a first-order transfer function of the form $G(s) = K/(\tau s + 1)$; the second is a second-order transfer function of the form $G(s) = K/(\tau s + 1)^2$, where K is the system gain and τ is the time constant.

The effects of variations of the parameters K, τ, and d on the performance of the Smith predictor will be studied and the results presented in a series of figures.

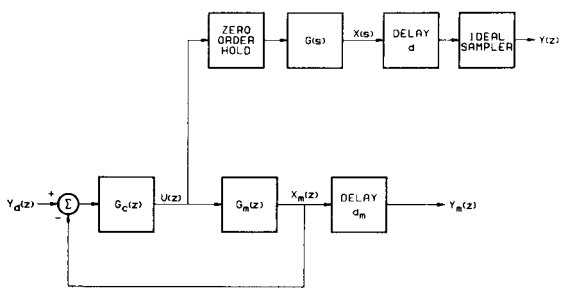


Fig. 2. Simplified form of the Smith predictor.

Although the second-order transfer function is not the most general possible, this form was selected for two reasons. The first reason is that the experimental system which will be used in the next section is of this form. The second reason is that in the more general case in which two time constants are allowed, say τ_1 and τ_2, the number of parameters which are allowed to vary gets quickly out of hand.

The compensation, $G_c(z)$, was designed according to the Dahlin algorithm as described by Chiu et al. [16]. A PI controller was used with the first-order model, and a PID controller was used with the second-order model. The details of the controller design are contained in the Appendix.

The nominal parameter values for the simulated actual system are shown in Table 1. These values were found to describe adequately the experimental system based on a parameter identification method using a nonlinear least-squares technique [6].

The system shown in Fig. 1 was simulated on a digital computer. The parameter values in Table 1 were used to design $G_c(z)$. In each simulation the system responded to a step change in Y_d from 10 to 20. A sampling period of 1 s was used. The parameter values K, τ, and d were varied in the simulated actual system in order to determine the sensitivity of the system to variations in these parameters. The parameters K and τ were varied by $\pm 10\%$ from their nominal values while d was varied by ± 1 s which corresponds to one sampling period.

Figure 3(a) shows the step response for the case in which there is no mismatch between the parameters in the actual system and the system model for both the first-order and second-order systems. The ideal responses are nearly identical except for the difference in the delays exhibited in the output owing to the difference in the delays shown in Table 1. These delays were different for the first- and second-order models simply because the nonlinear parameter estimation algorithm found the best fit with these values.

Typical step responses for variations in the parameters K, τ, and d are shown in Fig. 3(b—d). Figure 3(b) shows the step response of the first-order model in which the gain of the actual system is 5% lower than the nominal value, and the time constant and delay are equal to the nominal values; also shown is the step response of the second-order model with the gain of the

TABLE 1

Nominal parameter values for simulated actual system

Parameters	First-order model	Second-order model
K	1.42	1.41
τ	26.7 s	16.0 s
d	66.8 s	59.6 s
T	1.0 s	1.0 s

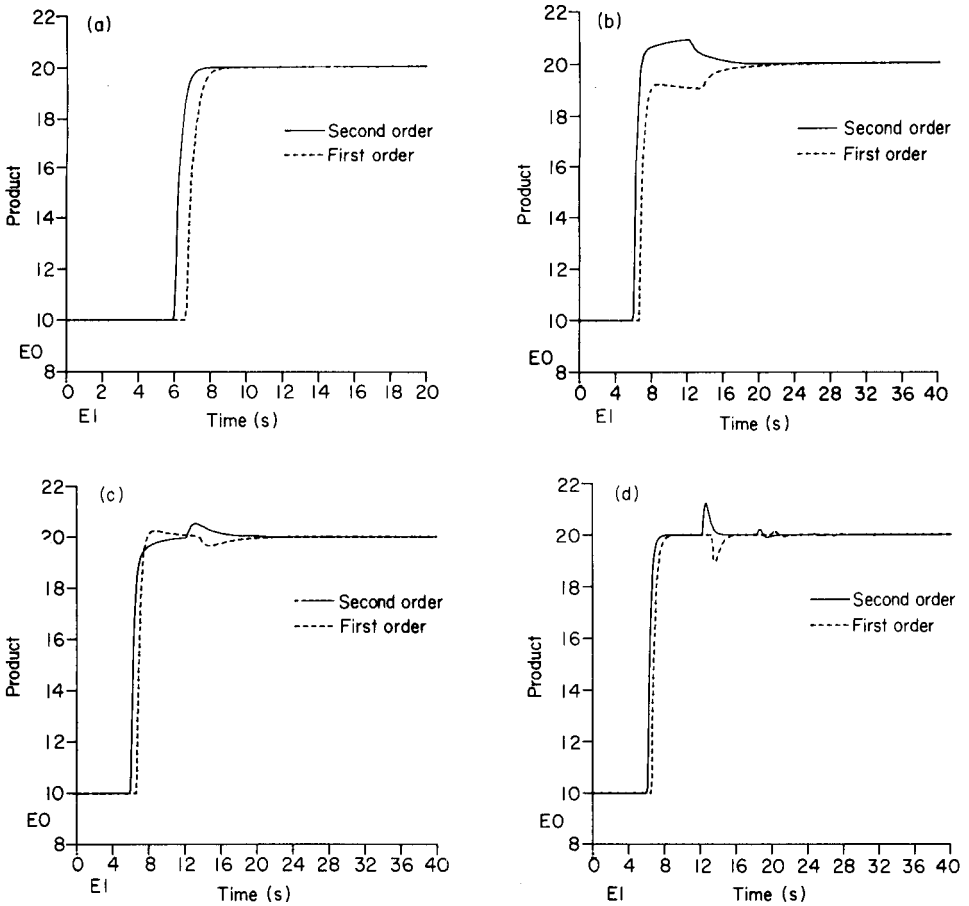


Fig. 3. (a) Step response of the closed loop system with no mismatch in parameters. (b) Effect of gain mismatch; for the first-order system, the actual gain is 5% lower than nominal and for the second-order system, 5% higher than nominal. (c) Effect of time constant mismatch; for the first-order system, the actual time constant is 5% lower than nominal and for the second-order system, 5% higher than nominal. (d) Effect of time delay mismatch; for the first-order system, the actual time delay is 1 s less than nominal and for the second-order system, 1 s higher than nominal.

actual system 5% higher than the nominal value while the remaining parameters are equal to their nominal values. Figure 3(c) illustrates the effect of a 5% error in time constant with no error in gain and delay. For the first-order system, the time constant of the actual system is 5% lower than the nominal value, and the time constant of the second-order system is 5% higher than the nominal value. Figure 3(d) shows the effect of an error in the delay. The delay of the actual system is 1 s (one sampling period) less than the nominal value while the gain and time constant are equal to their nominal values for the first-order system, and the delay of the second-order model is 1 s larger than the nominal value.

A general effect of the modeling errors can be observed in Fig. 3. After the initial delay, a typical step response is initiated. Because of mismatch in parameters, the output of the system model is not the same as the output of the actual system, and an additional error signal is generated and put into the PI or PID controller. The effect of this compensating error is not observed at the output until a full system time delay interval later. Eventually, the effect of these parameter errors is eliminated provided the system remains stable. In the simulated runs the error in time delay had the greatest effect on system stability. Variations in gain and time constant of the order of ±10% have a significant effect on the transient performance; however, these variations did not adversely affect system stability. An error in the delay of magnitude equal to one sampling period (in this example an error of approximately 1.5%) can lead to considerable deterioration of performance, and the simulations showed that an error of approximately three sampling periods led to instability. The system time delay used in digital simulations is usually approximated by the nearest integer multiple of the sampling period because of the ease with which this method can be implemented. However, because of the sensitivity of the response to an error in time delay, a method which allows for delays to be a fraction of the sampling period was implemented in this study. The method is described in the Appendix and adds very little complexity to the overall implementation of the Smith predictor.

The effects of the parameter variations on the classical performance criteria overshoot, settling time and rise time were also determined from simulated runs. Although the classical definitions were applied, a greatly mismatched system and model can lead to results which are difficult to interpret in terms of the classical criteria. The step response of a second-order system in which gain, time constant and delay were all in error is shown in Fig. 4. Normalized rise time is defined as $(t_{90} - t_{10})/\tau$ and normalized settling time is defined as $(t_s - d)/\tau$ where the time constant, τ, and delay, d, of the actual system are used. Overshoot is defined as $(y(t_p) - y_0)/(y_d - y_0)$ if $y(t)$ is greater than y_0 for some t and zero otherwise.

Figure 5 shows three-dimensional plots illustrating the effects of variations in K, τ, and d on settling time, rise time, and overshoot. The figures show results for the second-order system only. The first-order model exhibited similar qualitative behavior. In Fig. 5(b—d), K and τ are varied by ±10% of the nominal values specified in Table 1 while d is held fixed at its nominal value; in Fig. 5(e—g), K is varied by ±10% of its nominal value, d is varied by ±1 s, and τ is held constant at its nominal value.

The purpose of the three-dimensional plots is to present general trends in a visual manner and not to serve as a method of presenting tabulated data. There are some abrupt changes in the surfaces which appear as cliffs. For example, Fig. 5(f) shows a valley enclosed by very steep cliffs and would indicate that the time for a second-order system to settle within 5% of its final value is much greater for a 6% error in gain than for a 4% error in gain (with no error in d). Although this is true, it should not be concluded that

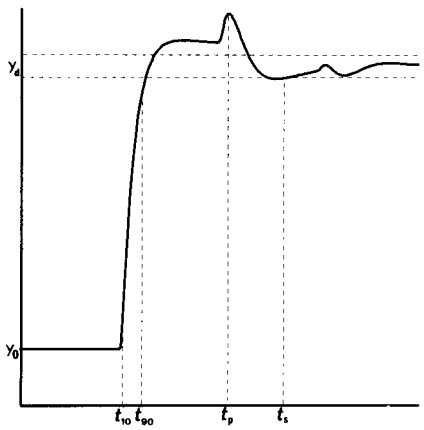


Fig. 4. Definition of classical control system performance criteria parameters. y_0 Initial value of output; y_d new set point; t_p time at which output reaches its maximum value; t_{10} time at which output reaches $y_0 + 0.1(y_d - y_0)$; t_{90} time at which output reaches $y_0 + 0.9(y_d - y_0)$; t_s settling time, the time which it takes for the output to come within a specified percentage (2% or 5%) of y_d and remain there.

there is a marked difference in the time response with a 4% error in gain as opposed to a 6% gain error. Instead, the discrepancy lies in the fact that classical performance criteria are being applied to responses which do not behave in the classical sense. For example, Fig. 4 illustrates the fact that the system response is quite different from the classical response for mismatched parameters. Thus small variations which frequently occur after one or two delay times can drastically affect the settling time criteria. The sharp discontinuities are not observed in the surface showing overshoot. This behavior more accurately describes the continuity of the response with respect to parameter variations.

Some general observations can be made. The Smith predictor is capable of providing satisfactory operation in the presence of reasonable variations in the parameters gain, time constant, and delay. Mismatched parameters result in variations occurring at integral multiples of the system delay time. These variations diminish with increasing time provided the system is stable. Errors of ±10% in gain and time constant can be handled reasonably well if enough time is allowed to achieve the settling criterion, i.e., it takes much longer to attain the 2% settling time than the 5% settling time. The system is much more sensitive to errors in time delay. In this example, an error of 5% led to unstable behavior. However, it is reasonable to assume that this value is known much more accurately than the gain or time constant. Furthermore, it is much less likely to vary under operation than the gain and time constant. Accuracy to one sampling period (about 1.5%) was found to give acceptable performance, and higher accuracy in predicting this value can be expected.

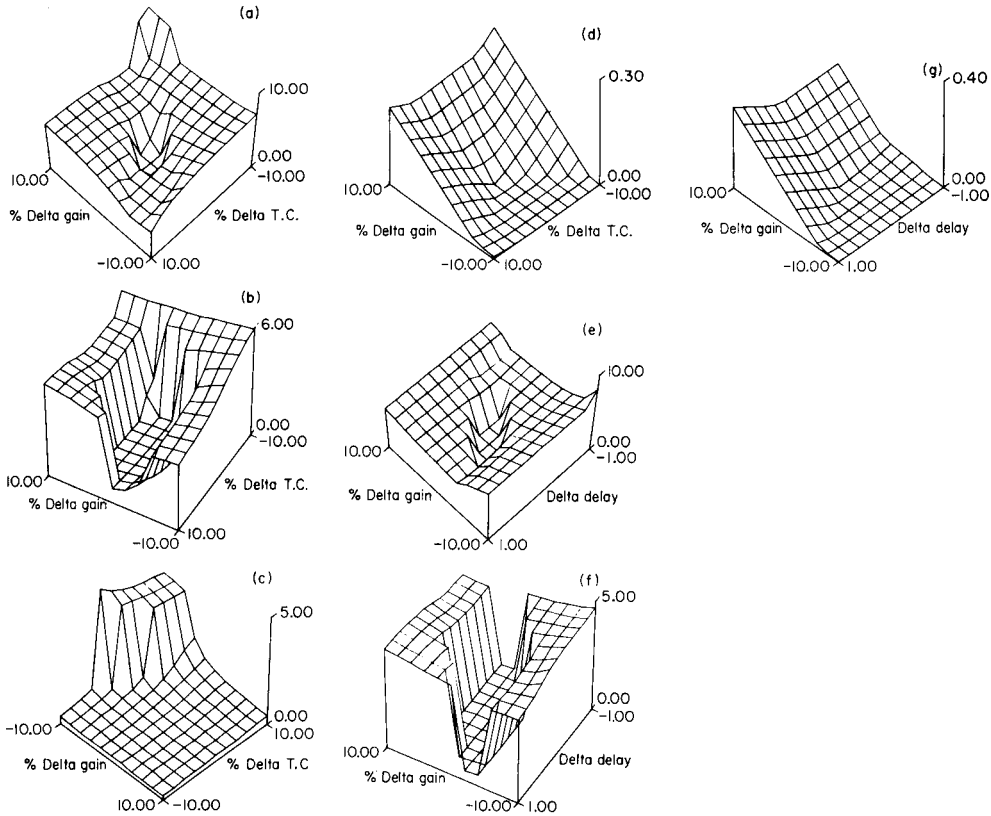


Fig. 5. (a) Normalized settling time to 2% vs. error in gain and time constant. (b) Normalized settling time to 5% vs. error in gain and time constant. (c) Normalized rise time vs. error in gain and time constant. (d) Overshoot vs. error in gain and time constant. (e) Normalized settling time to 2% vs. error in gain and time delay. (f) Normalized settling time to 5% vs. error in gain and time delay. (g) Overshoot vs. error in gain and time delay.

RESULTS

The Smith predictor was implemented in the control of an experimental enzyme reactor [4]. Figure 6 shows a functional block diagram of the computer-controlled apparatus. Input reagents are pumped into the system via stepping motors whose pumping rates are controlled by the digital computer. In this experiment, enzyme (alkaline phosphatase), substrate (*p*-nitrophenylphosphate), buffer (2-amino-2-methyl-1-propanol), and diluent (water) were pumped into the mixer. The substrate was pumped at a constant rate, the enzyme served as the controlled or manipulated variable, the buffer was pumped at a rate to maintain constant pH, and the diluent was pumped at a rate to maintain a constant flow rate through the system. A series of four delay loops and valves permit sixteen possible path lengths which together

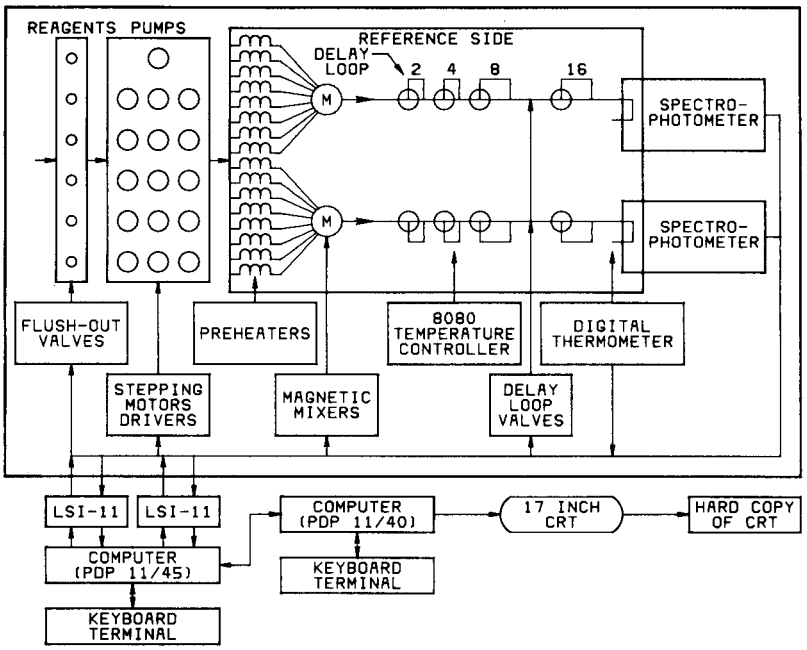


Fig. 6. Functional block diagram of the computerized apparatus showing the flow system and major components.

with pumping rate determine the time allowed for the reaction to occur. The product formed in this enzyme-catalyzed reaction is *p*-nitrophenolate. The absorption of the solution is measured by a spectrophotometer located at the end of the delay lines. The output of the spectrophotometer is sampled every sampling period, converted to digital form, and input to the digital computer. The concentration of the product is determined from the measured absorption. The measured concentration is fed back to the controller which determines the rate at which to pump enzyme, thus completing the feedback loop. A block diagram which illustrates each of these functions is shown in Fig. 7.

The nonlinear least-squares parameter identification method described by Pomernacki et al. [6] was used to obtain the system transfer function. In this case the transfer function is from enzyme concentration to product concentration at the spectrophotometer. The system was operated open loop, and a step change in enzyme concentration was applied. The resulting response in product concentration was measured, and the best fit of a second-order plus delay transfer function was found. The resulting step responses are shown in Fig. 8(a). The values of K, τ, and d of the second-order model found by the parameter identification algorithm are listed in Table 2. These three parameters were used to tune the PID algorithm as described in the Appendix. The closed loop system with a Smith predictor was then

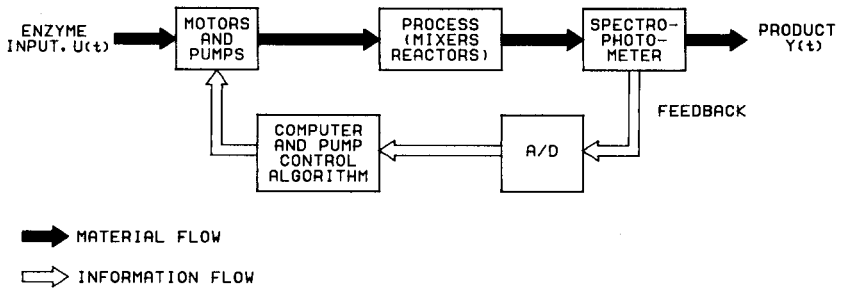


Fig. 7. Block diagram of the enzyme control system.

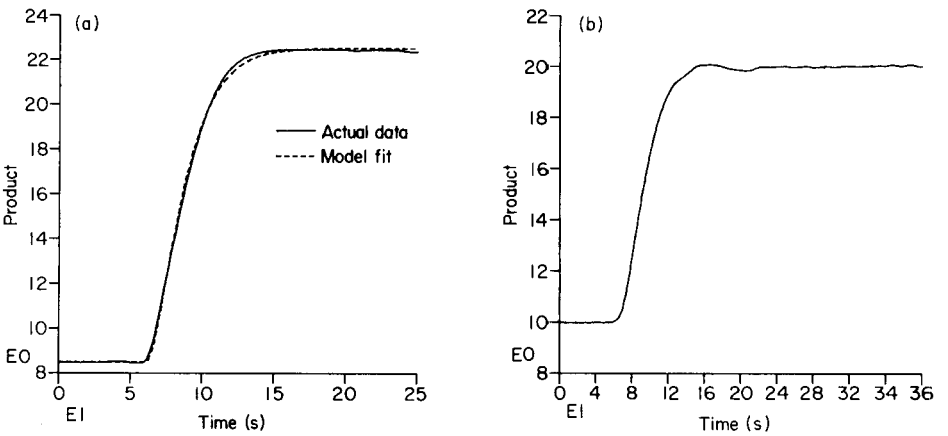


Fig. 8. (a) Open loop step response of actual system and model with identified parameters: pH 10.0; substrate concentration 3 mmol l^{-1}; enzyme concentration step 6.0 to 16.0 U l^{-1}; total flow rate 7 ml min^{-1}; delay path volume 10 ml; temperature 30°C; sampling period 1 s. (b) Step response of the closed loop system with no mismatch introduced.

TABLE 2

Identified parameters K, τ and d found by the identification algorithm for normal and upset conditions

	Normal conditions[a]	Upset conditions			
		pH 9.25	Error	[S] = 4.0	Error
K (μmol/U)	1.41	1.30	−7.8%	1.57	+11.3%
τ (s)	14.22	14.28	+0.4%	14.35	+0.9%
d (s)	60.88	60.49	−0.39	60.70	−0.18

[a]As outlined in the legend to Fig. 8: pH 10, [S] = 3.0.

implemented and tested with a step change in the set point for product concentration. The conditions remained the same as listed in the legend to Fig. 8 except that enzyme concentration was now the control variable. The resulting step response is shown in Fig. 8(b). The response differs some from the ideal step response shown in Fig. 3(a). One of the primary contributors to the discrepancy is the fact that the pumping motors have a limited range of pumping rates while no such limits were placed on the simulation. Overall the agreement between the experimental and theoretical results is quite good in view of the complexity of the system.

To test the system with upsets in operating conditions, pH and substrate concentration were varied from the values used for Fig. 8, but the PID controller was still tuned to the above values for K, τ and d. Figure 9(a) shows the step response of the system at pH 9.25 with all other parameters as before. Figure 9(b) shows the step response with substrate concentration of 4.0 and all other parameters as before. The nonlinear least-squares parameter identification was also used to determine the parameters of the system under the mismatched conditions so that a comparison can be made with the values given in Table 2 for normal conditions. These results are also listed in Table 2.

The step responses of the experimental system with mismatched parameters are not in complete agreement with the simulated system, for example as shown in Fig. 4. The experimental results appear to be much smoother than predicted by the theoretical model. Again, the primary cause of the discrepancy is the presence of limits on the pumping rates. The three-dimensional sensitivity figures (Fig. 5) provide reasonable estimates of the performance of the actual system. For example, the low pH run shown in Fig. 9(a) exhibits no overshoot as predicted by Fig. 5(g). The overshoot for the high

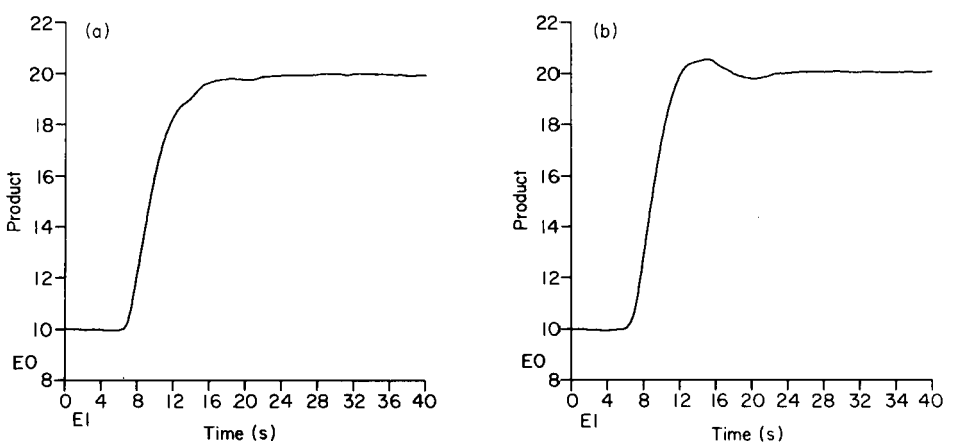


Fig. 9. (a) Step response of the closed loop system with pH low. (b) Step response of the closed loop system with substrate concentration high.

concentration substrate run shown in Fig. 9(b) is approximately 0.10. This value is slightly less than the value predicted by extrapolation of Fig. 5(d). The rise time in the experimental runs are all larger than predicted; however, this can be attributed directly to the limited pumping rates.

Figure 9 illustrates the effect of a mismatch when there is a step change in the set point. In order to illustrate the effect of an upset in the system while attempting to maintain a constant set point, the pH of the system was changed from 10.0 to 9.25 while the system was operating at a fixed set point of 20.0 μmol l^{-1}. The PID controller tuned to the parameters listed as normal conditions in Table 2 was used. Note that these parameters were obtained at a pH of 10.0. The results of the experiment are shown in Fig. 10. Although this upset was quite severe and caused a significant deviation in the output, the Smith predictor with a PID controller did recover and bring the product output back to the desired set point. When the system is operating at constant pH, the response is very nearly linear with respect to changes in enzyme concentration. However, the response with respect to changes in pH is quite nonlinear. The nonlinearity is illustrated in Fig. 11 where the product formed versus pH is shown for constant enzyme and substrate concentrations. More extensive figures illustrating the effects of varying pH, enzyme, and substrate can be found in Frazer et al. [5].

To a lesser extent the response with respect to changes in substrate (PNPP) concentration is also nonlinear as shown in Fig. 12. It is desirable for such nonlinear systems to have control strategies based on the characterization of the chemical reactions to be controlled. The reasons for having control as a function of the chemical characterization are fairly obvious and can be more easily understood by referring to Figs. 10 and 12. Although the pH upset was initiated as a step change in input pH, because of fluid mixing, diffusion and flow dynamics, the plant output tracked the upset in a manner similar to a slower variation in pH. That is, product yield went up as pH in the reaction

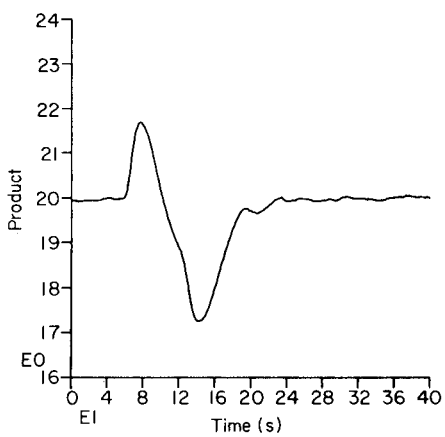
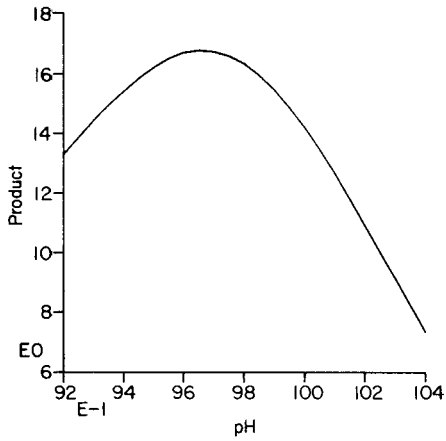


Fig. 10. Run with pH upset while attempting to maintain a set point of 20.



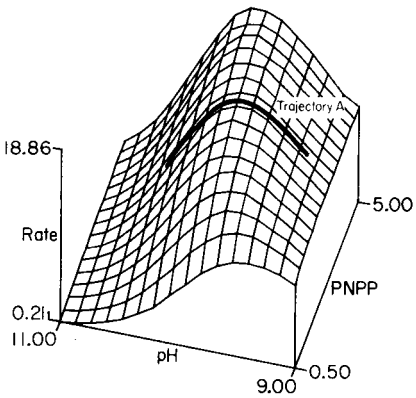


Fig. 11. Product vs. pH at steady state.

Fig. 12. Response surface showing the effect of varying pH and substrate concentration on the reaction rate of alkaline phosphatase.

zone progressed from 9.25 to approximately 9.7 then began to fall as the pH moved towards 10.0, as shown by the trajectory A in Fig. 12. Thus, following the pH upset, the yield increased sharply then fell way below the set point (Fig. 10) before the controller could provide adequate compensation by varying enzyme concentration. Given a good characterization of the chemical system and pH monitoring at the inlet, pH upset could be immediately offset by changes in the control variables such as enzyme and substrate concentrations. Forthcoming work will be directed towards the development of control strategies based on the chemical characteristics of the system to be controlled.

Conclusions

The Smith predictor with PI and PID controllers has been analyzed to determine sensitivity with respect to variations in gain, time constant, and delay. The model under study has a relatively long time delay (approximately 3.5 times the system time constant). A series of figures makes it possible to predict performance based on classical criteria (overshoot, rise time, and settling time) as a function of parameter variations. Experiments on an enzyme reactor were also performed, and the performance of an actual system was compared to that predicted by the simulated results.

It was found that reasonable variation in gain and time constant (±10%) cause significant degradation of the transient response; however, after a sufficiently long settling time, the desired set point can be maintained. The Smith predictor is most sensitive to errors in modeling the time delay. In the example considered, variations of about 1.5% led to significant degradation of performance, and a variation of 5% led to instability.

A nonlinear parameter identification method was used to determine a

model for the experimental system. It was found that the Smith predictor with a PID controller tuned to the identified model parameters perform nearly as predicted by the simulated results. The primary reason for the discrepancy between the simulated and experimental results is that the pumping motors in the experimental system had a limited range of pumping rates. Experiments with a minimum time control algorithm which makes optimum use of these pumping rate limits are now in progress.

In conclusion, the Smith predictor is an effective controller for systems with large time delays provided the controller is well tuned to the actual system. It has been demonstrated on an experimental enzyme reactor that the Smith predictor controller tuned to the parameters identified by a nonlinear least-squares parameter identification algorithm gives nearly ideal performance.

APPENDIX: *Design of the PI and PID controllers*

The design is based on the Guillemin–Truxal method [17]. The development here follows that of Chiu et al. [16] and is called Dahlin's algorithm.

Design of $G_c(z)$. The design of $G_c(z)$ is based on the closed loop shown in Fig. 2 without delay. According to the Dahlin algorithm, $G_c(z)$ is designed such that the closed loop transfer function, $W(z) = X_m(z)/Y_d(z)$, is given by

$$W(z) = (1 - e^{-\lambda T})/(z - e^{-\lambda T}) \tag{A1}$$

where T is the sampling period and λ is a tuning parameter. It can be seen that this transfer function implies that the closed loop system is supposed to behave like the system shown in Fig. A1. Since $W(z) = G_c(z)G_m(z)/(1 + G_c(z)G_m(z))$, the compensation, $G_c(z)$, is

$$G_c(z) = [1/G_m(z)]\,[W(z)/(1 - W(z))] \tag{A2}$$

Both first-order and second-order systems will be considered.

A. First-order system. For the first-order system, it is assumed that $G(s)$ is of the form $G(s) = K/(\tau s + 1)$. Thus

$$G_m(z) = K(1 - e^{-T/\tau})/(z - e^{-T/\tau})$$

Solving for $G_c(z)$ in eqn. (A2);

$$G_c(z) = K_c[1 + (T/T_i)/(1 - z^{-1})], \tag{A3}$$

where $K_c = (1 - e^{-\lambda T})/[K(e^{T/\tau} - 1)]$ and $T_i = T/(e^{T/\tau} - 1)$. The controller,

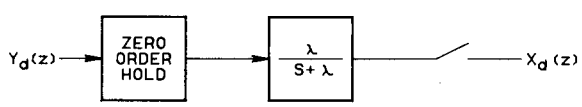


Fig. A1. Desired closed loop behavior.

$G_c(z)$, has been given in the discrete time form of the standard PI controller in eqn. (A3).

B. Second-order system. For the second-order system, it is assumed that $G(s)$ is of the form $G(s) = K/[(\tau_1 s + 1)(\tau_2 s + 1)]$, thus

$$G_m(z) = K(C_3 z + C_4 z)/(z - e^{-T/\tau_1})(z - e^{-T/\tau_2}) \tag{A4}$$

where $C_3 = 1 + (\tau_2 e^{-T/\tau_2} - \tau_1 e^{-T/\tau_1})/(\tau_1 - \tau_2)$

and $C_4 = e^{-(T/\tau_1 + T/\tau_2)} + (\tau_2 e^{-T/\tau_1} - \tau_1 e^{-T/\tau_2})/(\tau_1 - \tau_2)$

For typical values of T, τ_1, and τ_2, C_3 is very nearly equal to C_4, resulting in a "ringing" pole near $z = -1$ when eqn. (A4) is used in eqn. (A2). Elimination of the "ringing" pole, leads to

$$G_c(z) = K_c \left[1 + \frac{T/T_i}{1 - z^{-1}} + \frac{T_d}{T}(1 - z^{-1})\right] \tag{A5}$$

where $K_c = (e^{T/\tau_1} + e^{T/\tau_2} - 2)(1 - e^{-\lambda T})/[K(e^{T/\tau_1} - 1)(e^{T/\tau_2} - 1)]$

$T_i = T(e^{T/\tau_1} + e^{T/\tau_2} - 2)/[(e^{T/\tau_1} - 1)(e^{T/\tau_2} - 1)]$

and

$T_d = T/(e^{T/\tau_1} + e^{T/\tau_2} - 2)$

In eqn. (A5), $G_c(z)$ is written in the discrete time form for the standard PID controller.

The value $\lambda = 0.287682$ (corresponding to $e^{-\lambda T} = 0.75$) was used for all results shown in this paper.

Implementation of the time delay

Both the first-order and second-order systems can be included in a general discussion if state variable notation [17] is used. For the first-order system, let $\mathbf{x}(t) = x(t)$, and for the second-order system, let

$$\mathbf{x}(t) = \begin{bmatrix} x(t) \\ \dot{x}(t) \end{bmatrix}$$

Then the differential equations for the system with time delay can be written in the form $\mathbf{x}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}u(t)$ and $y(t) = \mathbf{C}^T\mathbf{x}(t - d)$.

Let $t_k = kT$, $k = 0, 1, 2, \ldots$, be the sampling times, let N be the largest integer such that NT is less than or equal to d, and let $t_F = T - (d - NT)$. The relative times are illustrated in Fig. A2(a). The value of $y(t_{k+1+N}) = x(t_k + t_F)$. The value $x(t_k + t_F)$ is obtained by

$$x(t_k + t_F) = \mathbf{C}^T\mathbf{x}(t_k + t_F) \tag{A6}$$

where $\mathbf{x}(t_k + t_F) = \Phi(t_F)\mathbf{x}(t_k) + \Theta(t_F)u(t_k)$ and $\Phi(t) = e^{\mathbf{A}t}$ and $\Theta(t) = \int_0^t \Phi(\tau)\,\mathbf{b}\,d\tau$.

The values of $\mathbf{x}(t_k)$ are generated by

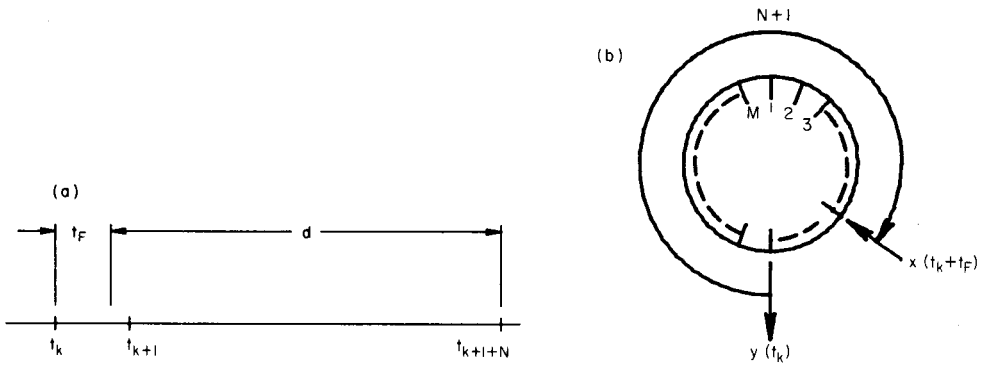


Fig. A2. (a) Illustration of time delay and sampling times. (b) Illustration of circular buffer for time delay.

$$\mathbf{x}(t_{k+1}) = \Phi(T)\mathbf{x}(t_k) + \Theta(T)u(t_k) \tag{A7}$$

Since this equation (or something equivalent) must be implemented in the Smith predictor anyway, eqn. (A6) represents very little additional computational burden. The values of $y(t_{k+1+N})$ are saved in an array of dimension $M \geqslant N$ as illustrated in the circular buffer shown in Fig. A2(b).

Work performed under the auspices of the Department of Energy by Lawrence Livermore Laboratory under contract W-7405-ENG-48, and partly funded by Engineering Sciences, OBES.

REFERENCES

1 J. E. Troyan, Chem. Eng. J., 70 (1963) 120.
2 A. S. Foss, AIChE J., 19 (1973) 209.
3 W. Lee and V. W. Weekman, Jr., AIChE J., 22 (1976) 27.
4 J. W. Frazer, L. P. Rigdon, H. R. Brand and C. L. Pomernacki, Anal. Chem., 51 (1979) 1739.
5 J. W. Frazer, L. P. Rigdon, H. R. Brand, C. L. Pomernacki and T. A. Brubaker, Anal. Chem., 51 (1979) 1747.
6 C. L. Pomernacki, H. R. Brand, T. A. Brubaker and J. W. Frazer, Anal. Chim. Acta, 112 (1980) 287.
7 O. J. M. Smith, Instrum. Soc. Am. J., 6 (1959) 28.
8 J. F. Donoghue, IEEE Transactions on Industrial Electronics and Control Instrumentation, IECI-24 (1977) 109.
9 C. Meyer, D. E. Seborg and R. K. Wood, Chem. Eng. Sci., 31 (1976) 775.
10 J. E. Marshall, Int. J. Control, 19 (1974) 933.
11 G. Alevisakis and D. E. Seborg, Chem. Eng. Sci., 29 (1974) 373.
12 J. O. Gray and P. W. B. Hunt, Electron. Lett., 7 (1971) 131.
13 B. Garland and J. E. Marshall, in M. J. Gregson (Ed.), Recent Theoretical Developments in Control, Academic Press, New York, 1978.
14 L. Eisenberg, Instrum. Soc. Am. J., 6 (1967) 329.
15 E. B. Dahlin, Instrum. Control Sys., 41 (June) (1968) 77.
16 K. C. Chiu, A. B. Corripio and C. L. Smith, Instrum. Control Sys., 46 (Dec.) (1973) 41.
17 D. G. Schultz and J. L. Melsa, State Functions and Linear Control Systems, McGraw-Hill, New York, 1967.

Analytica Chimica Acta, 122 (1980) 421—433
Computer Techniques and Optimization

IMPROVEMENT OF THE SUPER-MODIFIED SIMPLEX OPTIMIZATION PROCEDURE

P. F. A. VAN DER WIEL

Department of Analytical Chemistry, Faculty of Science, University of Nijmegen, Toernooiveld, Nijmegen (The Netherlands)

(Received 3rd December 1979)

SUMMARY

The performance of the super-modified simplex optimization procedure can be improved substantially. Three modifications of this procedure, involving application of Gaussian fit, weighted reflection point, and estimation of response at the reflection point, are described and tested by means of computer simulation. Combined application of the first two modifications leads to an improvement of about 30% in the number of required experiments, and of 56—73% in the variance of this number. Improvements are achieved in 98% of the cases studied.

The performance of an analytical instrument depends strongly on the correct settings of the instrumental parameters. For a flameless atomic absorption spectrometer, for instance, these parameters are wavelength, lamp current, temperature and time of drying, ashing, atomization and burning, inert gas flow, etc.; each type of sample has its own optimal settings. Thus in optimizing the performance of an instrument, a large number of parameters is often involved. A simultaneous technique (e.g. random design) will generally require too many experiments. Among the sequential techniques suitable for attacking this problem, the simplex technique, which also accounts for interactions between the parameters, has found widespread use [1—7]. The simplex technique uses a number of experiments to find better parameter settings, which are in turn used to restart the procedure. This technique normally requires considerably fewer experiments than a simultaneous technique [8], particularly when the number of parameters increases. The disadvantage of all sequential techniques (including the simplex) is that the optimum found can actually be a sub-optimum. Restarting the procedure with different initial parameter settings can usually solve this problem.

The simplex technique was first proposed by Spendley et al. [9] and later modified by Nelder and Mead [10] and Routh et al. [6]. Simplex techniques have been surveyed by Deming and Parker [11]. The methods described in the papers mentioned are referred to as simplex, modified simplex (MS) and super-modified simplex (SMS). The simplex method, described by a set of rules [9, 12], is elegant and involves trivial calculations which can be done

even by hand. One of its disadvantages is that it is not easy to recognize whether an optimum or sub-optimum has been reached, especially with three or more parameters. This problem does not occur in the MS, which simply homes in on an optimum. A disadvantage here is the complexity of the rules [4, 10], although the calculations are simple and fewer experiments are needed than with the original simplex. The SMS incorporates a second-order polynomial fit to estimate better parameter settings and is superior to the MS [6, 7]. The simplicity and ease of calculations decreases from the simplex to the SMS. However, because of the widespread use of programmable calculators in analytical laboratories, the more complex calculations of the SMS are no longer a disadvantage in most practical situations. When computing time is ignored, further improvement of the SMS, in terms of a decrease in the required number of experiments, will have a considerable effect on the time needed to optimize the parameter settings of an analytical instrument.

In this paper, three possible improvements concerning the estimation of better parameter settings by the SMS are investigated by means of computer simulation. The goal is the development of a simplex algorithm for use in optimization of the parameter settings of an analytical instrument (e.g., an atomic absorption spectrometer).

GENERAL CONSIDERATIONS

A simplex consists of an $(n + 1)$-point (usually called a vertex) geometrical figure in an n-dimensional parameter space, n being the number of parameters involved [6, 9, 10]. The response at a vertex is defined as the result of one or more experiments done with the parameter settings indicated by the coordinates of the vertex. The quantity to be optimized depends on the goal of the optimization (mean response, sensitivity, signal/noise ratio, etc.). The simplex moves by replacing the vertex which yields the worst response (vertex $V_{w,j}$, response R_w in Fig. 1) by an estimated better vertex. The SMS incorporates a second-order polynomial fit to obtain the "new" vertex in Fig. 1.

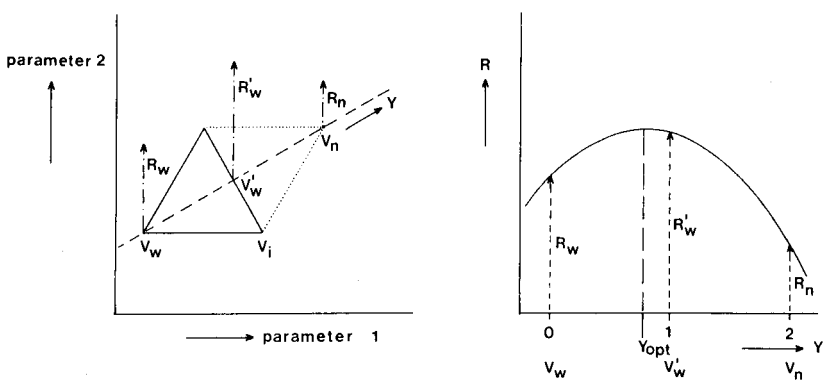


Fig. 1. Second-order fit in the SMS.

The estimated better parameter settings are calculated by using the equation

$$V_{\text{better},j} = Y_{\text{opt}} * V'_{w,j} + (1 - Y_{\text{opt}})\, V_{w,j} \quad (j = 1 \ldots n)$$

$$\text{with } V'_{w,j} = \left(\sum_{i=1}^{n+1} V_{i,j} - V_{w,j}\right)/n \quad (j = 1 \ldots n) \tag{1}$$

$$\text{and } Y_{\text{opt}} = [(R_w - R'_w)/(R_w - 2R'_w + R_n)] + \tfrac{1}{2} \tag{2}$$

where n is the number of parameters involved; $V_{i,j}$ is the jth parameter/coordinate of the ith vertex; $V_{w,j}$ is the jth parameter/coordinate of the "worst" vertex; $V'_{w,j}$ is the jth parameter/coordinate of the so-called reflection point; $V_{n,j}$ is calculated using $Y_{\text{opt}} = 2$ in eqn. (1); Y_{opt} is the expansion factor; and R_i is the response at vertex i.

To find a maximum with eqn. (2), it is obvious that the condition $R_w - 2R'_w + R_n < 0$ should hold. The fitting equation is asymptotic, and so a limit on the expansion factor Y_{opt} must be set. Without such a limit, a minor change in one of the responses could have an unduly large effect on the simplex movement. Although it may seem more logical to set a limit on $\partial Y_{\text{opt}}/\partial R$, this is not practical because it increases the computing time; its value is not yet proven. The new calculated vertex should not equal the rejected vertex ($V_{w,j}$) or the reflection point ($V'_{w,j}$) because then no progress has been made or (part of) the mobility of the simplex is lost.

Some additional rules can be formulated. The "next-to-worst" rule is essential in the normal simplex procedure to prevent the simplex from oscillating between the present and the previous simplex. This rule states that if the new vertex yields the worst response in the simplex, the procedure should be repeated with the vertex which yields the next-to-worst response. This rule forces the simplex eventually to circle around an optimum, and the need for it depends on the procedure. Application of this rule to the MS and SMS is not strictly necessary. An additional rule on noise can be applied to all simplex procedures. This rule states that if a vertex persists in (n + 1) simplexes (n being the number of parameters), the related experiment(s) should be repeated to stop the simplex being pinned down by a false high response. In experimental optimization, the parameters do have a specific possible range, so boundaries must be set. The rule that deals with boundary violations is flexible; of the various possible techniques, none is particularly recommended. The problem of when to stop or alter a simplex procedure can also be solved in various ways, usually depending on the optimization problem itself. A sub-set of rules concerning the SMS originate from the restrictions which have to be set to the SMS. When a simplex procedure is applied, the choice of the aforementioned additional rules should be considered and explicitly stated.

In this investigation on improvements of the SMS the rules described by

Routh et al. [6] were used. The next-to-worst rule and the rule on noise were not applied and no parameter boundaries were set. Three possible improvements of the SMS were tested.

Application of a Gaussian fit

Application of a Gaussian fit instead of a second-order polynomial fit yields the following equations:

$$\text{let } R = f(Y) = A * \exp\left[-(B-Y)^2/2C^2\right]$$

$$\text{then } R_w = f(0) = A * \exp\left[-B^2/2C^2\right]$$

$$R'_w = f(1) = A * \exp\left[-(B-1)^2/2C^2\right]$$

$$R_n = f(2) = A * \exp\left[-(B-2)^2/2C^2\right]$$

Combining these equations (see Fig. 2) yields after some mathematical manipulations:

$$Y_{opt} = \left[(\ln R_w - \ln R'_w)/(\ln R_w - 2\ln R'_w + \ln R_n)\right] + \tfrac{1}{2} \qquad (3)$$

Like the second-order equation, this equation is asymptotic and limits on the expansion factor Y_{opt} should be set; restriction of $\partial Y_{opt}/\partial R$ is practically impossible. Here the condition $\ln R_w - 2\ln R'_w + \ln R_n < 0$ should hold. Comparison of this condition with the previous condition for the second-order fit shows that the Gaussian fit has a greater workable range. This means that in a situation where the second-order fit does not yield a maximum, the Gaussian fit might be capable of finding one. This is demonstrated in Figs. 3 and 4. In Fig. 3 the expansion factor Y_{opt} is given as a function of one of the responses (R'_w) while the other two responses are held constant. Figure 4 gives an example of the arrangement of the responses in the case of a limit. As can be seen in eqns. (2) and (3) the second-order fit maximum is independent of amplification ($R = R * K$) and displacement ($R = R \pm K$) of the responses, while the Gaussian fit maximum is independent of amplifi-

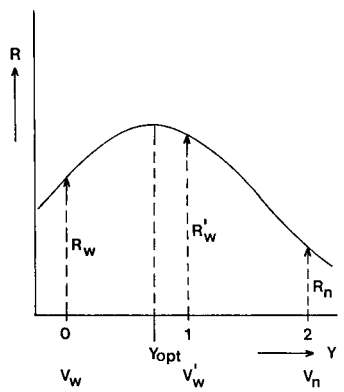


Fig. 2. Gaussian fit in an SMS.

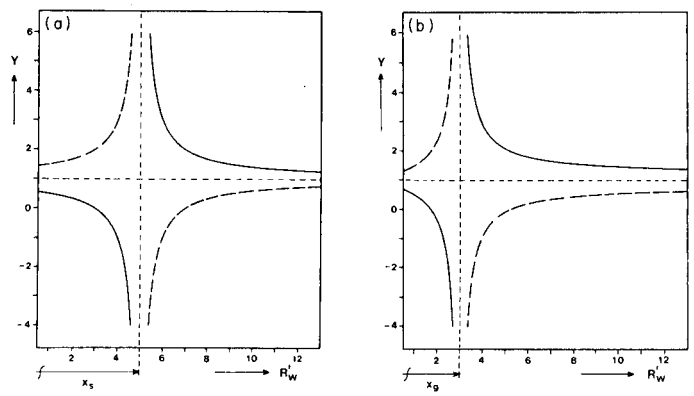


Fig. 3. Y_{opt} as a function of R'_w (eqns. 2 and 3). Comparison of (a) a second-order fit and (b) a Gaussian fit: (——) $R_w = 1$, $R_n = 9$; (---) $R_w = 9$, $R_n = 1$. Note that x_s (the unworkable range of R'_w) is greater than x_g.

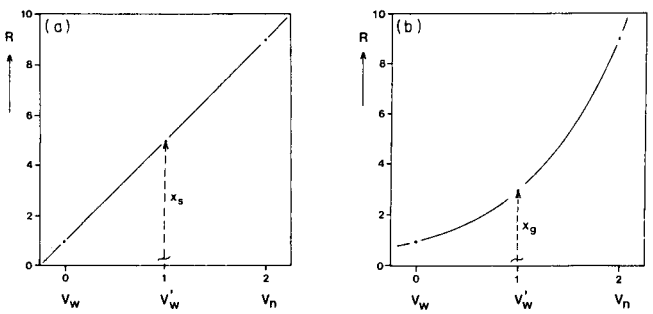


Fig. 4. Arrangement of responses for which a maximum cannot be estimated ($x_s > x_g$): (a) second-order fit; (b) Gaussian fit.

cation but not of displacement. A displacement occurs when the response at a point far remote from an optimum is more or less fixed at a certain base level (see Fig. 5). The effect of a displacement equal to ten times the net maximum response (of a Gaussian response surface) on the Gaussian fit maximum is shown in Fig. 6.

It was expected that careful use of the Gaussian fit could give better results than the second-order fit because of the greater workable range.

Application of a weighted reflection point

For calculation of the parameters/coordinates of the reflection point ($V'_{w,j}$) eqn. (1) is replaced by

$$V'_{w,j} = [\sum_{i=1}^{n+1} (R_i * V_{i,j}) - R_w * V_{w,j}] / [\sum_{i=1}^{n+1} R_i - R_w] \quad (j = 1 \ldots n)$$

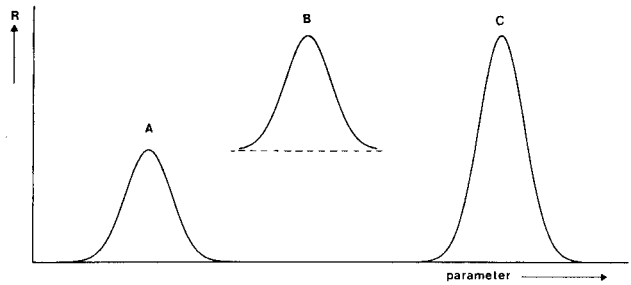


Fig. 5. Simplified response-surfaces: (A) normal; (B) displacement of 1 × height; (C) amplification, factor 2.

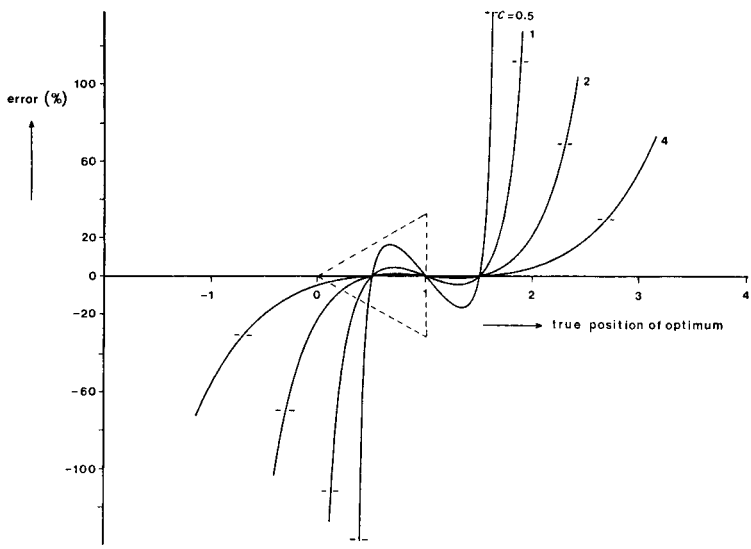


Fig. 6. The effect of a displacement ten times the net maximum response, on the Gaussian fit maximum for a Gaussian response-surface ($R = A * \exp[-(B-Y)^2/2C^2]$, unit of C is the unit of Y) error = (estimate—true position)/true position * 100.

This reflection point fulfils the condition $\sum_{\substack{i=1 \\ i \neq w}}^{n+1} \bar{d} * R_i = 0$, where d is the distance between a vertex and the reflection point and R denotes the response at the vertex. Application of this reflection point in calculating the new parameter settings could give better results than those obtained from a normal reflection point because of better adaptation to the response-surface.

Estimation of the response at the reflection point instead of an experimental check

The advantage is obvious: one experiment is saved in the construction of a new simplex:

$$R'_w = [\sum_{i=1}^{n+1} R_i - R_w]/n$$

The response at a weighted reflection point is estimated from

$$R'_w = [\sum_{i=1}^{n+1} (R_i)^2 - R_w^2]/[\sum_{i=1}^{n+1} R_i - R_w]$$

These modifications were all tested, including all combinations. The methods will be referred to as:

S, SMS (second-order fit); G, SMS with Gaussian fit; E, SMS with estimated reflection point; W, SMS with weighted reflection point.

Valid combinations are S, SW, SE, SWE, G, GW, GE, GWE. An example of one of the algorithms is given in Table 1.

COMPUTATIONAL CONSIDERATIONS

The simplex procedures described were tested by repeated application to a known two-dimensional response-surface. For each procedure the number of experiments needed to reach the optimum within 1% is recorded. The initial parameter settings, defining the start simplex, were chosen randomly for a given combination of simplex size and start area by the following procedure.

The centre point (C) of the simplex was constructed by using a fixed point (P) in parameter space, angle α and distance d (see Fig. 7). The simplex itself was constructed around this centre point by using angle β and sides l, in a manner which defines the start area as enclosed by a circle with centre P and radius ra. Angle β is defined as the angle between line m passing through one simplex point, preferably the "worst" vertex, and centre point C, and the line passing through point C and point P.

Unless otherwise stated,

α: aselect uniform random	$0° < \alpha < 360°$
β: aselect uniform random	$-60° < \beta < 60°$
d: aselect uniform random	$0 < d < d_{max}$, depending on l and ra
l: preset (input)	$0 < l < ra\sqrt{3}$

ra: preset (input) set to 50.0
P: coordinates 100, 100; usually the coordinates of maximum response.

This procedure was followed because it ensures that the whole start area can be covered by the start simplex, which is not the case when a rectangular start area is used.

The simplex methods were tested on five response-surfaces. These response-surfaces are defined by the following equations (no noise is added).

(a) Symmetrical Gaussian: response = $10 * \exp\{-[(100-x_1)^2 + (100-x_2)^2]/1500\}$

TABLE 1

Simplex algorithm GE (simplified)[a]

```
    (Construct start simplex)
    FOR I: = 1 STEP 1 UNTIL D + 1 DO
     R[I]: = measurement(I);
    W: = D + 2; WA: = D + 2; N: = D + 3; NN: = D + 4;
up: C: = W; W: = sort(D); RSUM: = 0.0;
    FOR I: = 1 STEP UNTIL D DO
     BEGIN
        SUM[I]: = 0.0;
        FOR J: = 1 STEP 1 UNTIL D + 1 DO
         SUM[I]: = SUM[I] + V[I,J];
        V[WA,I]: = (SUM[I] − V[W,1])/D;
        V[N,I]: = 2 * V[WA,I] − V[W,I];
     END;
    boundary(N); R[N]: = measurement(N);
    FOR I: = 1 STEP 1 UNTIL D + 1 DO
     RSUM: = RSUM + R[I];
    IF C ≠ W THEN
     R[WA]: = (RSUM−R[W])/D;
    IF R[WA] < sqrt(R[N] * R[W]) THEN
     BEGIN
        IF R[W] > R[N] THEN
                        Y: = −1.0
                     ELSE
                        Y: = 3.0;
        GOTO pass;
     END;
    Y: = (ln R[W] − ln R[WA])/(ln R[W] − 2 * ln R[WA] + ln R[N]) + 0.5;
    restrict(Y);
pass: FOR I: = 1 STEP 1 UNTIL D DO
     V[NN,I]: = Y * V[WA,I] + (1−Y) * V[W,I]; boundary(NN);
    FOR I: = 1 STEP 1 UNTIL D DO
     V[W,I]: = V[NN,I];
    IF test(D) = FALSE THEN
     GOTO up;
```

[a]*boundary* is a routine to detect and correct boundary violations. D is the number of parameters (*D*-dimensional). *measurement* (I) means measure as indicated by vertex I and evaluate the result. *restrict* (Y) is a routine to check the value of Y: $Y > 3 \rightarrow Y := 3$; $Y < -1 \rightarrow Y := -1$; $1.3 > Y \geq 1 \rightarrow Y := 1.3$; $1.0 > Y > 0.7 \rightarrow Y := 0.7$; $0.3 > Y \geq 0 \rightarrow Y := 0.3$; $0 > Y > -0.3 \rightarrow Y := -0.3$. R[I] is the response at the Ith vertex. *sort* (D) is the sort routine which returns the number (W) of the vertex which yields the worst response. *test* (D) is a routine to end the procedure. V[I,J] is the *J*th parameter/coordinate of the *I*th vertex.

(b) B1 [8]: response $= 10 * (0.5 + 0.5Bx_1)^4 * Bx_2^4 * \exp[2-(0.5 + 0.5Bx_1)^4 - Bx_2^4]$

(c) B2 [8]: response $= 10 * (0.3 + 0.4Bx_1 + 0.3Bx_2)^4 * (0.8-0.6Bx_1 + 0.8Bx_2)^4 * \exp[2-(0.3 + 0.4Bx_1 + 0.3Bx_2)^4 - (0.8-0.6Bx_1 + 0.8Bx_2)^4]$

(d) B3 [8]: response $= 10 * Bx_1^2 * \exp[1-Bx_1^2 - 20.25 * (Bx_1-Bx_2)^2]$

(e) B4 [8]: response $= 10 * (0.3Bx_1^2 + 0.7Bx_2^2)^3 * \exp[1-0.6(Bx_1-Bx_2)^2 - (0.3Bx_1^2 + 0.7Bx_2^2)^3]$

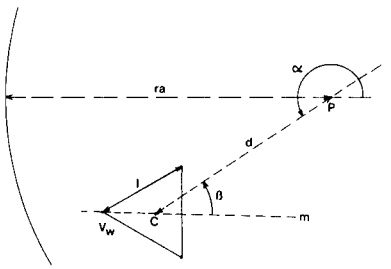


Fig. 7. Construction of the initial simplex.

With (b), (c), (d) and (e): $Bx_1 = [(x_1—100)/50] + 1$; $Bx_2 = [(x_2—100)/50] + 1$. As can be seen above, the maximum response on each response-surface is set to ten. In Fig. 8, some lines giving equal response (response 1, 4, 7 and 10) are shown. The simplex procedures are started as described and are terminated when the mean of the responses in a simplex differs by less than 0.1(1%) from the maximum response. The maximum allowable number of "experiments" (in fact, evaluations of the response-function) is set to 100; every run which requires more experiments is considered as a failure.

The simulation program, written in ALGOL60, is capable of handling four simplex procedures at a time on one response-surface; each procedure was run 800 times. The output of the program consisted of the mean number of experiments required to reach the optimum ($\bar{x}$), the variance in this number (s^2), and the number (N in %) of runs required by each procedure to excel or equal the performance of another procedure (see Table 2). The program was run on an IBM 370/158 computer.

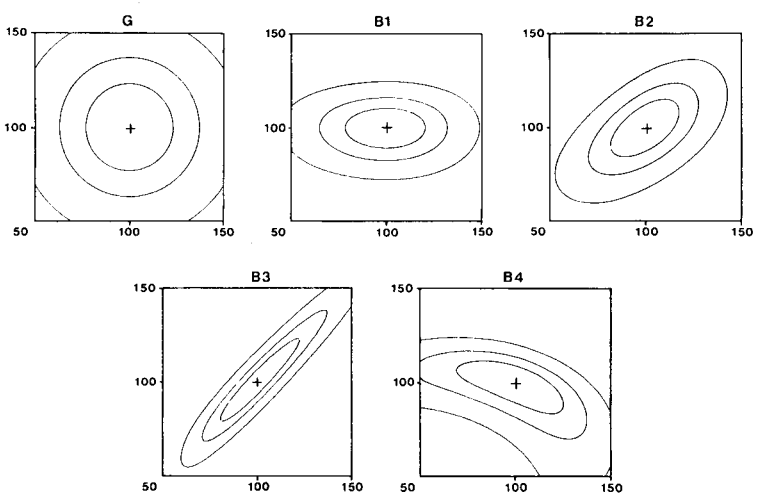


Fig. 8. Equiresponse lines of the response-surfaces involved in this investigation. Response = 1, 4, 7 and 10.

TABLE 2

Effect of the investigated modifications at the minima (see text)

($\bar{x}$, mean required number of experiments; s^2, variance on this number; $\bar{x}_s$, mean required number of simplexes; s_s^2, variance on this number; $N<$, % runs which showed improvement (in x); $N =$, % runs which showed equal x; $N>$, % runs where the modification was not successful; a positive number means improvement.)

Mean of S–G, SE–GE, SW–GW, SWE–GWE		Compared to the SMS	
G.	$\bar{x}$ + 4%	$\bar{x}$ + 5%	
	$\bar{x}_s$ + 4%	$\bar{x}_s$ + 5%	
	s^2 + 23%	s^2 +32%	
	s_s^2 + 23%	s_s^2 + 23%	
	$N<$ 21%	$N<$ 23%	
	$N=$ 69%	$N=$ 67%	
	N_{tot}: 16000	N_{tot}: 4000	
	$N>$ 10%	$N>$ 10%	
Mean of S–SE, SW–SWE, G–GE, GW–GWE			
E.	$\bar{x}$ + 28%	$\bar{x}$ + 27%	
	$\bar{x}_s$ − 0.5%	$\bar{x}_s$ − 1%	
	s^2 + 62%	s^2 + 66%	
	s + 16%	s_s^2 + 23%	
	$N<$ 97%	$N<$ 97%	
	$N=$ 1%	$N=$ 1%	
	N_{tot}: 16000	N_{tot}: 4000	
	$N>$ 2%	$N>$ 2%	
Mean of S–SW, SE–SWE, G–GW, GE–GWE			
W.	$\bar{x}$ − 4%	$\bar{x}$ − 4%	
	$\bar{x}_s$ − 4%	$\bar{x}_s$ − 4%	
	s^2 −64%	s^2 −56%	
	s_s^2 − 66%	s_s^2 − 58%	
	$N<$ 30%	$N<$ 32%	
	$N=$ 42%	$N=$ 39%	
	N_{tot}: 16000	N_{tot}: 4000	
	$N>$ 28%	$N>$ 29%	
GE.		$\bar{x}$ + 30%	$\bar{x}_s$ + 3%
		s^2 + 73%	s_s^2 + 40%
		$N<$ 98%	
		$N=$ 0%	N_{tot}: 4000
		$N>$ 2%	

RESULTS AND DISCUSSION

For all simplex procedures on all response-surfaces involved in this investigation, the mean required number of experiments ($\bar{x}$ of 800 runs) was determined as function of the length (l) of the sides of the starting simplex. The results are given in Fig. 9. Almost all the tested combinations of simplex procedures and response-surfaces yielded a distinct minimum in the mean required number of experiments when a favourable size of starting simplex was used. (Figs. 9 and 10). These minima occur at l = 25, 35, 30, 35 and 30, on the symmetrical Gaussian response-surface B1, B2, B3 and B4, respectively. The existence of these minima means that it is not enough simply to choose a very large starting simplex. For the Gaussian response-surface, it

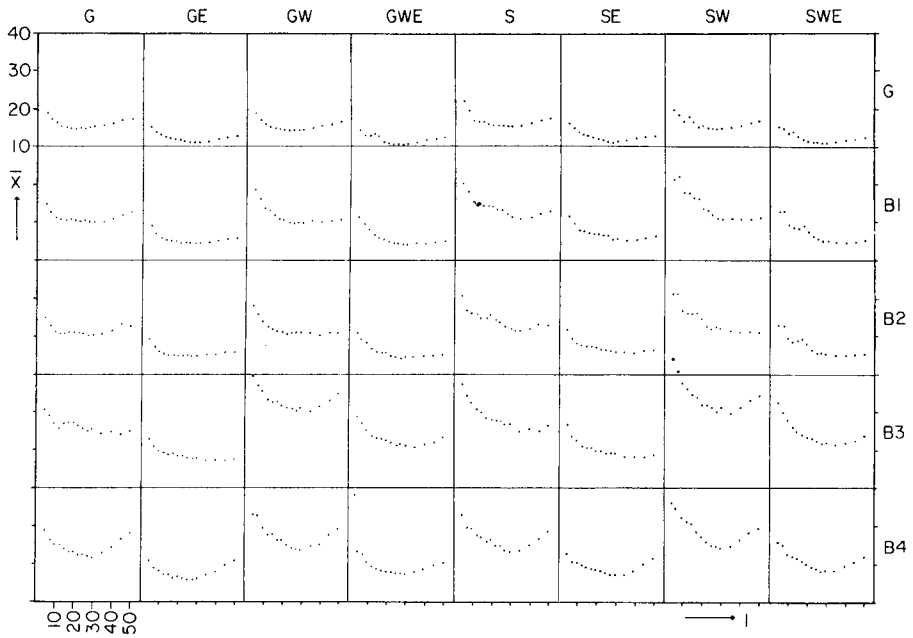


Fig. 9. The mean required number of experiments as a function of the size of the start simplex for eight simplex procedures on five response-surfaces.

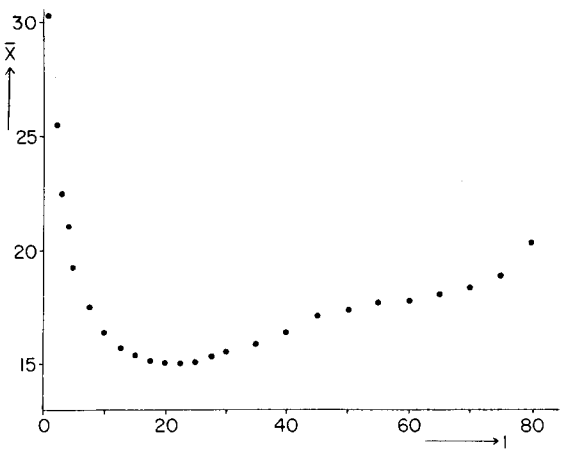


Fig. 10. The mean required number of experiments as a function of the size of the start simplex for one combination of simplex procedure and response-surface. The minimum is clearly visible for simplex G on Gaussian response.

was found that the location of the minimum depends largely on the size of the starting area and that the exact shape of the Gaussian response surface has less influence. Because this study was intended to find improvements in the SMS, the nature of the minima was not further investigated.

Tables 2 and 3 summarize the results of the comparison of the simplex

TABLE 3

Effect of the investigated modifications with random l (5—70)

Mean of S—G, SE—GE, SW—GW, SWE—GWE	Compared to the SMS
G. $\bar{x}$ + 5% $\bar{x}_s$ + 5%	$\bar{x}$ + 7% $\bar{x}_s$ + 7%
s^2 + 28% s_s^2 + 29%	s^2 + 43% s_s^2 + 43%
$N<$ 24%	$N<$ 26%
$N=$ 63% N_{tot}: 16000	$N=$ 64% N_{tot}: 4000
$N>$ 13%	$N>$ 10%
Mean of S—SE, SW—SWE, G—GE, GW—GWE	
E. $\bar{x}$ + 27% $\bar{x}_s$ — 2%	$\bar{x}$ + 27% $\bar{x}_s$ — 2%
s^2 + 36% s_s^2 — 25%	s^2 + 42% s_s^2 — 6%
$N<$ 96%	$N<$ 96%
$N=$ 1% N_{tot}: 16000	$N=$ 1% N_{tot}: 4000
$N>$ 3%	$N>$ 3%
Mean of S—SW, SE—SWE, G—GW, GE—GWE	
W. $\bar{x}$ — 4% $\bar{x}_s$ — 4%	$\bar{x}$ — 4% $\bar{x}_s$ — 4%
s^2 — 31% s_s^2 — 29%	s^2 — 16% s_s^2 — 13%
$N<$ 31%	$N<$ 33%
$N=$ 38% N_{tot}: 16000	$N=$ 35% N_{tot}: 4000
$N>$ 31%	$N>$ 32%
GE.	$\bar{x}$ + 30% $\bar{x}_s$ + 3%
	s^2 + 56% s_s^2 + 19%
	$N<$ 97%
	$N=$ 1% N_{tot}: 4000
	$N>$ 2%

modifications. The application of a Gaussian fit improved the simplex procedures used in terms of a decrease in $\bar{x}$ and a decrease in s^2 (the variance in the number of experiments). The effect of the E modification is a considerable decrease in $\bar{x}$ and s^2 while the mean required number of simplexes is hardly increased. This means that estimation of the response at the reflection point has no or little effect on the performance of the fitting procedure. The application of a weighted reflection point was unsuccessful; $\bar{x}$ and s^2 both increased, except on the symmetrical Gaussian response-surface. The W modification was successful only in the first 4 or 5 simplexes, whereafter the diminished mobility of the simplex prohibited effective advance of the simplex.

A considerable improvement of the SMS (about 30% in $\bar{x}$) was achieved when both the G and the E modifications were used. The value of $\bar{x}$ was improved in 98% of the simulation runs. Table 4 gives the % failure of the simplex procedures investigated on the response-surfaces described.

Because the best procedure (GE) failed in only 0.5% of all simulation runs, no effort was made to reduce this number.

TABLE 4

Percentage failure rates

	At the minimum						Overall (random *l*)					
	G	B1	B2	B3	B4	Mean	G	B1	B2	B3	B4	Mean
S	—	—	—	0.1	0.1	0.05	—	0.9	1.1	0.4	0.6	0.6
SE	—	—	—	—	—	—	—	1.1	2.0	—	0.1	0.6
SW	—	0.1	0.2	0.9	0.1	0.3	—	0.7	0.5	2.2	1.6	1.0
SWE	—	—	—	—	—	—	—	0.4	1.2	0.1	0.6	0.5
G	—	—	—	—	0.2	0.05	—	0.5	0.5	0.4	0.9	0.4
GE	—	—	—	—	—	—	—	1.0	1.4	—	0.4	0.5
GW	—	0.1	—	1.0	0.5	0.3	—	0.4	0.2	2.6	0.4	0.7
GWE	—	—	—	—	—	—	—	0.1	0.2	—	—	0.1

REFERENCES

1 R. R. Ernst, Rev. Sci. Instrum., 39 (1968) 998.
2 M. J. Houle, D. E. Long and D. Smette, Anal. Lett., 3 (1970) 401.
3 F. D. Czech, J. Ass. Offic. Anal. Chem., 56 (1973) 1489.
4 S. L. Morgan and S. N. Deming, Anal. Chem., 46 (1974) 1170.
5 R. Smits, C. Vanzoelen and D. L. Massart, Fresenius Z. Anal. Chem., 273 (1975) 1.
6 M. W. Routh, P. A. Swartz and M. B. Denton, Anal. Chem., 49 (1977) 1422.
7 S. L. Kaberline and C. L. Wilkins, Anal. Chim. Acta, 103 (1978) 417.
8 S. H. Brooks, Oper. Res., 7 (1959) 430.
9 W. Spendley, G. R. Hext and F. R. Himsworth, Technometrics, 4 (1962) 441.
10 J. A. Nelder and R. Mead, Comput. J., 7 (1965) 308.
11 S. N. Deming and L. R. Parker, Crit. Rev. Anal. Chem., 7 (1978) 187.
12 S. N. Deming and S. L. Morgan, Anal. Chem., 45 (1973) 278A.

Analytica Chimica Acta, 122 (1980) 435—454
Computer Techniques and Optimization

DIGITAL SIMULATION OF THE EFFECT OF DISPATCHING RULES ON THE PERFORMANCE OF A ROUTINE LABORATORY FOR STRUCTURAL ANALYSIS

B. G. M. VANDEGINSTE

Department of Analytical Chemistry, University of Nijmegen, Toernooiveld, Nijmegen (The Netherlands)

(Received 3rd December 1979)

SUMMARY

Several output characteristics of a laboratory for structural analysis are shown to be identical with the output of a model of that laboratory, e.g. the histograms of the input and output density (samples/day), the histograms of the number of samples present in the laboratory, the histograms of the delays and several cross-correlations. The effect of various strategies concerning priorities between various groups of samples is forecast, e.g. samples with a different expected analysis time, samples from various sources, samples with a different history in the laboratory. The effects of the introduction of an adaptable routing procedure, several technician assignment decisions and strategies on the termination of the analysis are simulated.

Recently [1], the observed delays of samples in a spectroscopic laboratory (i.r., p.m.r., ^{13}C-n.m.r., and mass spectroscopy) were compared with the delays calculated from a model of that laboratory. The study established the effect of several important parameters, i.e. (i) the utilization factor (ρ), which is the ratio between the mean analysis time and mean inter-arrival time, (ii) the mean analysis time, (iii) the interruptions of the analytical process, and (iv) the number of facilities.

The delay of various types of samples was distinguished by attributing priorities. For example, "easy" samples and "difficult" samples can be distinguished by their short and long analysis times, respectively. Previous study with a simple queueing model forecast that the average waiting time can be improved by giving priority to the "easy" samples which have previously been recognized as such. A common mode of action in a laboratory is to transfer a difficult spectrum after a certain interpretation time to a pile of unfinished spectra and to start the measurement of the next sample or interpretation of the next spectrum. Simulations indicated that this method gave no important discrimination in waiting time in favour of short analyses. Furthermore, it was demonstrated that the number of samples in the various sections of the laboratory could be fitted by an autoregressive model of the first order [1]. Several properties of the laboratory under investigation agreed with the general properties of a network of queues [2]. For example,

no correlation coefficients between the number of samples in the various sections of the laboratory, significantly differing from zero, were found. A typical property of the spectroscopic laboratory considered was that some samples are sequentially analyzed in several sections. Consequently, priority could be assigned depending on the number of analyses done on the sample. A minimal variance of the overall delay was found by giving absolute priority to the samples on which the most methods were tried.

Some further comparisons between the actual laboratory and the previously described model are presented in this paper. The effects of several dispatching rules are simulated. Likewise, the effect of the introduction of a variable routing procedure is simulated, where both the probability that the various spectroscopic methods can solve the analytical problem, and the state (queue lengths) of the laboratory are considered in selecting the analytical method. The conclusion from the calculations on simple queueing systems [1], that shortening the mean analysis time improves the mean delay time more than a reduction of the variations in the analysis time, is validated for the complex laboratory system by simulations. Likewise, the forecast discrimination of the delays of different groups of samples (easy/difficult and various users) is validated for the laboratory by simulation experiments with the model. Other aspects of interest are the flow sensitivity of the delays of sample groups with different priority, and the effect of the priority difference between two sample groups on their delay difference. Both effects are estimated from calculations on a simple queueing system and validated for the considered system by simulation. Very often, the influence of some variables is not independent of the level of the other variables. As a result an interaction is observed between the variables. To minimize the number of requested experiments with the model, the experiments were conducted by means of experimental designs.

VALIDATION OF THE MODEL

The input and output density

The actual laboratory. The laboratory under investigation consists of four sections which receive samples from two different origins (sample flow ratio 1:2.8). The total flows to and from each section are presented in Table 1. These observations indicate that the total flow in the laboratory exceeds the number of received samples (±12 samples/day), as on the average more than one (1.28) analysis is done on a sample. From the flow through the laboratory the probabilities could be calculated that a method will be selected in the first instance or after one or more other methods have failed (Table 2). The analysis of a sample consists of two steps: the measurement of the sample and the interpretation of the spectrum.

The means and variances of input rate, output rate, delays and number of samples were considered earlier [1]. Here the probability density functions of these variables are compared with some theoretical distributions. It is

TABLE 1

Mean and variance of the input and output flow (samples/day)

Section	Input		Output	
	Mean	Variance	Mean	Variance
I.r.	2.8	13.2	2.8	7.5
P.m.r.	7.7	20.3	7.7	50.0
M.s.	2.1	4.2	2.1	5.9
^{13}C-n.m.r.	2.5	4.2	2.5	11.1
Total	15.1		15.1	
Lab.	11.8	44.1	11.8	77.6

TABLE 2

The probabilities that methods are selected and yield the required information (% good)

	I.r.	P.m.r.	M.s.	^{13}C-n.m.r.	Completed samples (%)
First selection	0.17	0.57	0.10	0.16	79.6
% Good	0.65	0.84	0.72	0.84	
Second selection	0.27	0.29	0.29	0.15	94.5
% Good	0.76	0.76	0.64	0.78	
Third selection	0.21	0.29	0.28	0.22	98.3
% Good	0.58	0.72	0.78	0.69	
Fourth selection	0.25	0.27	0.25	0.23	99.8
% Good	0.92	0.85	0.83	0.91	

clear that the histogram of the input rate (α) (Fig. 1) which is close to a Poisson process, differs completely from that of the output rate (γ) (Fig. 2). A goodness of fit test by means of a χ^2 test indicated that, although the differences between the histogram of the actual input rate (Fig. 1) and a Poisson distribution function were small, the difference was significant.

The model. As a dynamic or versatile routing procedure will be used, based on the properties of the sample and on the state of the laboratory, a routing algorithm had to be developed that based its decisions on the observed traffic flow (Table 1) and the probabilities that an analytical problem can be solved by the various methods (Table 2). Two independent sample flows were generated to the laboratory, representing the two main sample sources. Both sample flows were generated by taking random numbers from an exponential density function [3] determining the inter-arrival times of the samples. Kleinrock [4] demonstrated that when the probability of the occurrence of a certain number of events during some given time span is Poisson-distributed, the times between the events are exponentially distributed. As a result, the probability density functions of the number of arrivals in the individual sections of the model are defined by three factors: (i) the probability functions of the inter-arrival times of the samples originating from both sources; (ii) the distribution process of these samples over

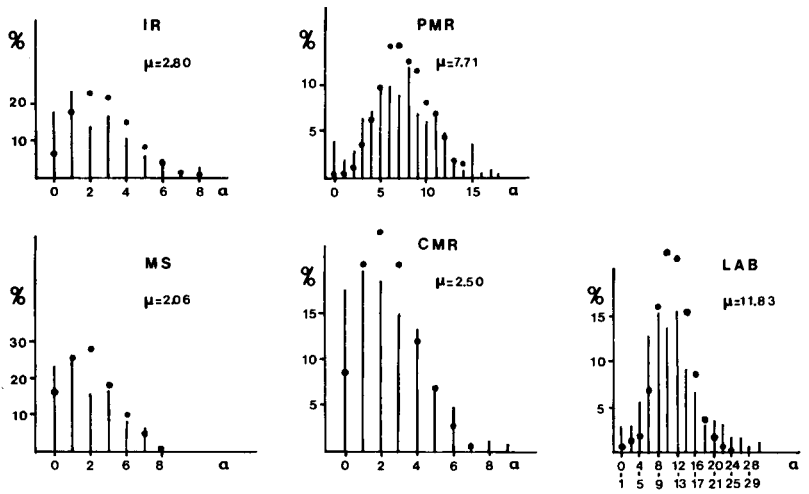


Fig. 1. Histograms of the input density α (samples/day) to the various sections of the laboratory. (•) Fitted Poisson distribution.

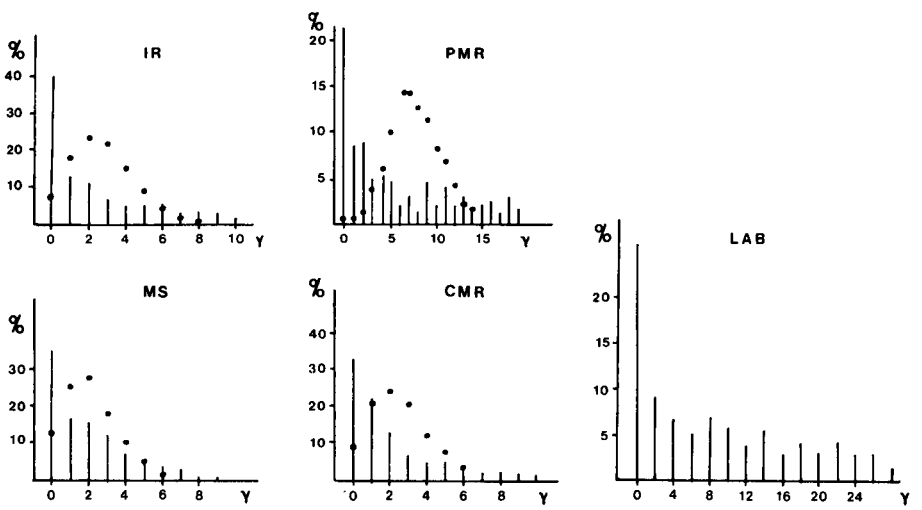


Fig. 2. Histograms of the output density γ (samples/day) from the various sections of the laboratory. (•) Poisson distribution.

the four sections; (iii) the departure process of unsuccessfully analyzed samples.

The histograms of the input and output density are shown in Figs. 3 and 4. The means of these densities did not differ significantly from the actual observations [1]. Here, no significant differences were found between the cumulative density functions of both. This means that the mentioned difference between input and output rate density functions is observed in

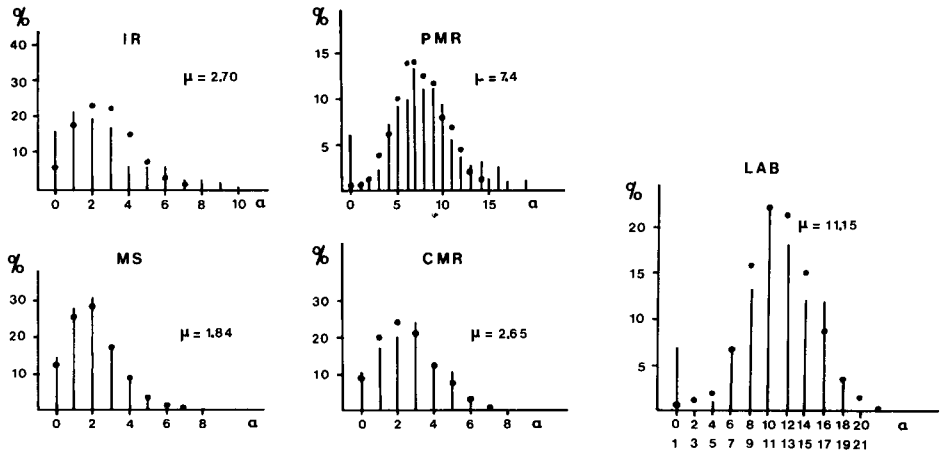


Fig. 3. Histograms of the input density α (samples/day) to the various sections of the simulation model. (•) Fitted Poisson distribution.

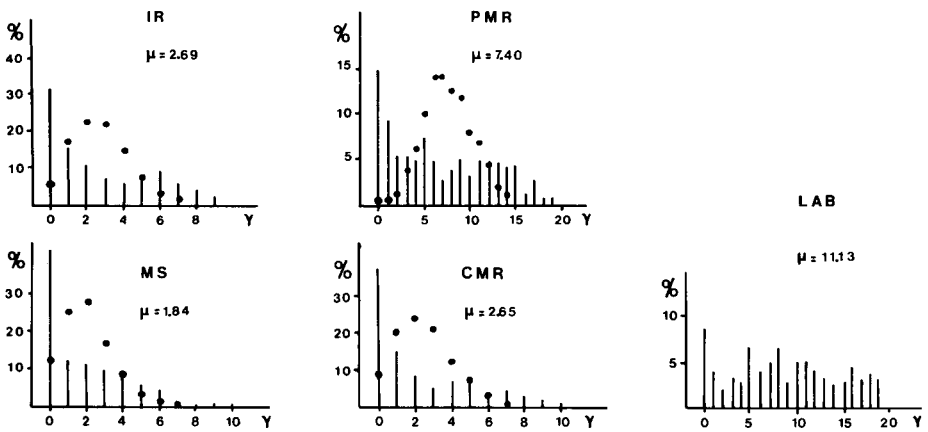


Fig. 4. Histograms of the output density γ (samples/day) from the various sections of the simulation model. (•) Poisson distribution.

the model also. Moreover, although the input densities to the actual sections are significantly different from a Poisson distribution, they can be simulated by the generation of Poisson processes in the model.

Probability density functions of the number of samples

The number of samples in the sections can be adequately described by a first-order autoregressive model [1]. Subsequently, executed Kolmogorov—Smirnov (K—S) tests on the cumulative density function indicated that these observations were Gaussian-distributed. From queueing theory, however, an exponential shape is expected [4].

A preliminary investigation by means of the simulation model pointed out that this unexpected shape is probably observed because the analysts start

measurements only when a given minimal batch of samples is available. As a result, the analysis is not started immediately after the arrival of a sample, as is assumed in the theoretical models of queueing theory.

Probability density function of the delays

Kolmogorov—Smirnov tests executed on the histograms of the delay times (Fig. 5) indicated that the two-stage Erlangian distribution fits these histograms very well. This discrete probability density function is represented by $f(x) = 2\mu(2\mu x)\exp(-2\mu x)$. The fit with a discrete function was permitted as the delays in the laboratory were calculated in terms of whole days. An exponential shape for the cumulative waiting time is a general characteristic for waiting line systems [4].

In the model, in accordance with the actual data, the histograms (Fig. 6) of the delays in the sections could be successfully fitted to a two-stage Erlangian distribution. This proves that the introduced analysis stages (i.e. collection of a batch; measurement and interpretation; collection and com-

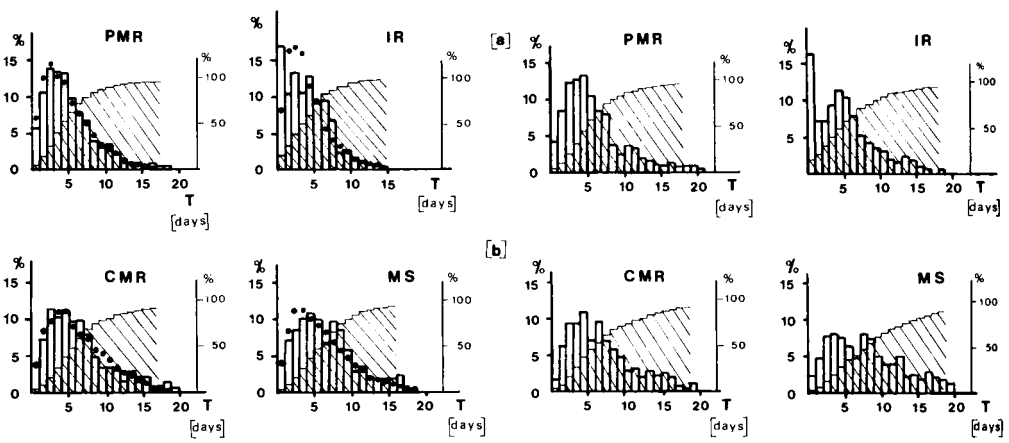


Fig. 5. (a) Histograms of the delays in the sections of the laboratory. (•) Fitted two-stage Erlangian distribution. Shaded figures are the cumulative density functions. (b) Histograms of the samples finished by the same method.

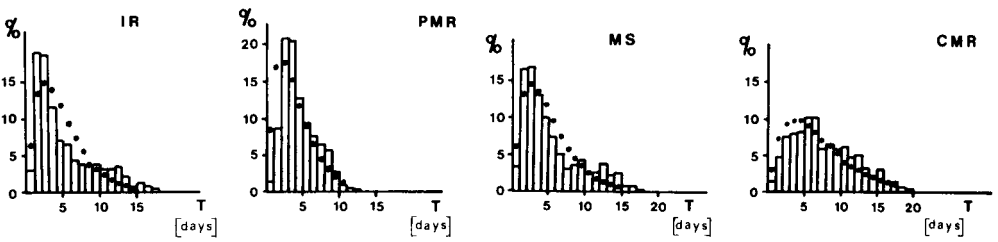


Fig. 6. Histograms of the delays in the sections of the simulation model. (•) Fitted two-stage Erlangian distribution.

munication of the results) with their probability density functions, are sufficiently accurately described in the model to generate identical waiting time distributions, compared with the actual situation.

Cross-correlations

In the laboratory, cross-correlation values between the input flow to the four sections and their number of samples (Table 3) demonstrated that the arrival processes do not depend on the state of the system. The number of samples directed to a section does not depend on the saturation grade of that section, i.e. samples are not preferably moved to that section with the lowest degree of saturation. Likewise, the correlations between the number of samples in a section and the delays are too small, in order to conclude that the delay of a sample can be reasonably forecast from the number of samples present in the section at its arrival (Table 4). Figure 7 illustrates that conclusion, as no dependence is found between the delay whereafter the analytical result will be ready within a 95% probability and the number of samples in the system at the arrival of a sample.

In the model system, simulations carried out with a First-in-First-out (FiFo) analyzing sequence for samples with an equal priority revealed a larger correlation between the number of samples in the section and the delay of a sample, compared with the actual situation (Table 5). Likewise a stronger dependence was found between the delay ($T_{0,95}$) whereafter the

TABLE 3

Correlation (ϕ) between the input flow and the number of samples in the section

Section	I.r.	P.m.r.	M.s.	^{13}C-n.m.r.
ϕ	0.20	0.40	0.26	0.30
95% CI[a]	±0.16	±0.16	±0.16	±0.16
RV[a]	0.91	0.84	0.93	0.91

[a]Confidence interval and residual variance.

TABLE 4

Maximal correlation between the number of samples (x) in the departments and the delay (y) of the samples arriving at the laboratory (data from actual laboratory)

	I.r.	P.m.r.	M.s.	^{13}C-n.m.r.	Lab.
ϕ_{xy}	0.247 (−5)[a]	0.19 (−3)	0.19 (−3)	0.28 (−6)	0.22 (11)
99% CI[b]	±0.18	±0.18	±0.17	±0.20	±0.21
RV[b]	0.94	0.96	0.96	0.92	0.94

[a]Time lag (τ) for maximal correlation. [b]Confidence interval and residual variance.

analytical result will be available (95% probability) and the number of samples in the system at its arrival (Fig. 8). However, by the application of a random analyzing sequence for samples with an equal priority, instead of the FiFo sequence, a decrease in the correlation between delay and number of waiting samples was observed (Table 6). In the p.m.r. section, which receives 60% of all samples, the correlation was not significantly different from zero. Likewise, the dependence previously found between $T_{0.95}$ and the number of waiting samples was removed by applying a random sequence (Fig. 9).

If the maximal delay ($T_{0.95}$) of a sample can be forecast reasonably well from the number of samples present in the system at its arrival, then the delay will remain within certain limits with a given probability, by applying a threshold control of the number of samples in the system. A threshold control system applied to the laboratory under investigation will be described in a further paper.

SIMULATION EXPERIMENTS

The simulation of strategies concerning priorities

Both in the laboratory under investigation and in the model, various priority rules can be applied, depending on the considered property of the sample, i.e. the expected analysis time of the sample, or history and origin of the sample. A priority difference can be obtained between samples by attributing urgency numbers (u_p) to the samples [1]. There are two main

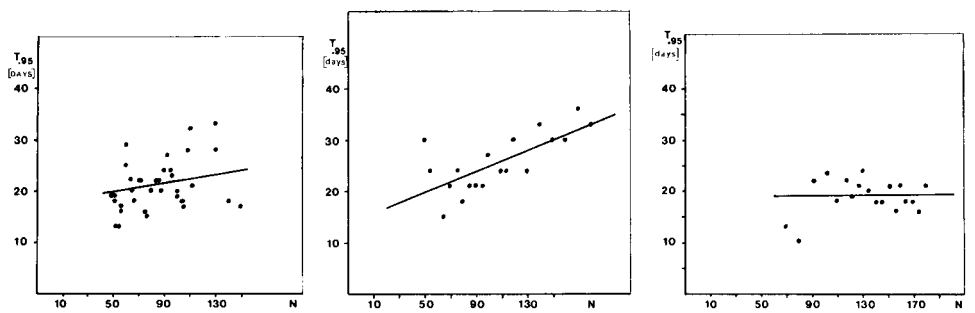


Fig. 7. The delay whereafter the analytical result will be ready ($T_{0.95}$) within a 95% certainty as a function of the number of samples in the system at the arrival of the sample. Data from the actual laboratory.

Fig. 8. The delay whereafter the analytical result will be ready ($T_{0.95}$) within a 95% certainty as a function of the number of samples in the system at the arrival of the sample. Simulated data with FiFo rule.

Fig. 9. The delay whereafter the analytical result will be ready ($T_{0.95}$) within a 95% certainty as a function of the number of samples in the system at the arrival of the sample. Simulated data with random sequence.

TABLE 5

Maximal correlation between the number of samples (x) in the departments and the delay (y) of the samples, arriving at laboratory (model: FiFo sequence)

	I.r.	P.m.r.	M.s.	^{13}C-n.m.r.
ϕ_{xy}	0.619 (−5)[a]	0.703 (−4)	0.635 (−5)	0.431 (−5)
99% CI[b]	±0.30	±0.32	±0.28	±0.21
RV[b]	0.62	0.50	0.60	0.81

[a,b]Footnotes as in Table 4.

TABLE 6

Maximal correlation between the number of samples (x) in the department and the delay (y) of the samples, arriving at the laboratory (model: RANDOM sequence)

	I.r.	P.m.r.	M.s.	^{13}C-n.m.r.
ϕ_{xy}	0.43 (−1)[a]	0	0.49 (−6)	0.48 (−30)
99% CI[b]	±0.37	—	±0.25	±0.26

[a,b]Footnotes as in Table 4.

possibilities. First, the priority u_p may be a function of the time that a sample waits in the system, W_p, and the urgency number b_p, which is a characteristic for the priority group p: $u_p = W_p^r b_p$. The samples are then analyzed in the sequence of decreasing urgency numbers. For large values of r, the sequence becomes independent of b_p and the sequence becomes FiFo. In contrast, for small r, an absolute priority discipline is obtained. Alternatively, the sample with the smallest sum of urgency number (b_p) and arrival date is analyzed next [5]: min(b_p + arrival date). For small values of $\Delta p = b_{p+1} - b_p$ a FiFo sequence is obtained. In contrast, for great values of Δp, an absolute priority discipline is obtained [5]. Unfortunately, no exact analytical results are available for that priority rule. However, the advantage of the rule is that the sample is immediately scheduled to a definite position in the queue. Therefore, that priority rule was used in the model: b_p may be a function of various properties of the sample (e.g. the number of unsuccessfully applied methods). The question arises of how these various priority rules affect the performance characteristics of the laboratory system (e.g. the mean delay and variation coefficient of the delay).

Priority based on the expected analysis time. Priorities based on the analysis time can be attributed in two ways. The samples are subdivided into two (or more) groups (e.g. groups with short and long analysis times), which receive different priorities; within each group, the samples can be analyzed in a FiFo or random sequence. Alternatively, the sample is positioned in the queue, depending on its (expected) analysis time. The effect of the application of the shortest (expected) analysis time (S.(E).A.T.) first priority

rule, along with the introduction of estimated analysis times was simulated with the earlier model [1]. The graph (Fig. 10) of the simulated delay ($\overline{T}$) as a function of the accuracy of the estimated interpretation time (S_{IT}) confirms the expectation that the mean delay ($\overline{T}$) decreases when the samples are positioned in a sequence of increasing analysis time: $\overline{T}_{\text{S.A.T}} < \overline{T}_{\text{S.E.A.T.}} < \overline{T}_{\text{Random sequence}}$. The relative superiority of the S.A.T. first operation rule is consistent with previous research from Conway et al. [6]. The variation coefficient of the delay (C_T^2) was found to be insensitive for that priority rule. When samples with exponentially distributed analysis times ($b(t)$), and a mean $\overline{AT}$, are subdivided into two groups, one with analysis times smaller than x ("easy"), and one with analysis times greater than x ("difficult") (Fig. 11), a mean analysis time ($\overline{AT}$) and sample flow (α) can be calculated for both groups (Fig. 12), where the ratio (R) between the mean analysis time of the easy samples ($\overline{AT_{<x}}$) and the overall mean analysis time ($\overline{AT}$) is plotted against the upper limit of the analysis time of the easy samples ($x/\overline{AT}$). Likewise the fractions (R) of easy $[(\alpha_{<x})/\alpha]$ and of difficult samples $[(\alpha_{>x})/\alpha]$ are plotted.

The delay of each group can be calculated from the equations presented by Kleinrock [4] for a M/M/1 system with head-of-line (HOL) or absolute priorities. The graphical representation (Fig. 13) of the calculated overall system time ($\overline{T}_t/\overline{T}_{\text{FiFo}}$) as a function of the fraction of easy samples $[(\alpha_{<x})/\alpha]$, where HOL priority is assigned to the easy samples, indicated that the improvement of the delay increases with higher utilization factors (ρ). Surprisingly, the overall delay is minimal when a small number of

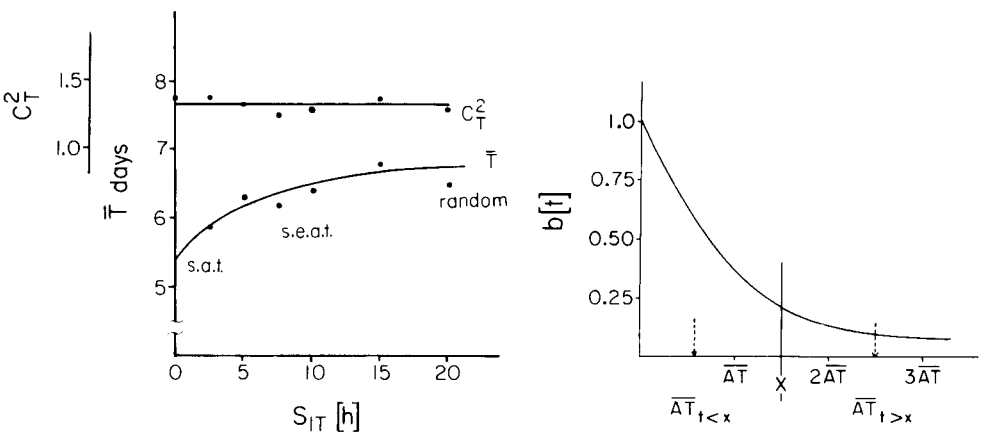


Fig. 10. The effect of the application of the shortest expected analysis time first priority rule on the mean delay ($\overline{T}$) and variation coefficient of the delay (C_T^2) as a function of the accuracy of the estimation of the interpretation time (S_{IT}).

Fig. 11. Distribution of the samples over two groups: probability density function $b(t)$ of the analysis time of: (1) 'easy' samples for which AT $< x$, with a mean $\overline{AT_{<x}}$; (2) 'difficult' samples for which AT $> x$, with a mean $\overline{AT_{>x}}$.

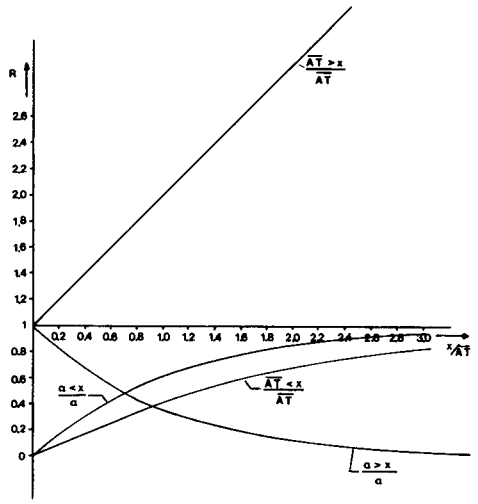


Fig. 12. Distribution of samples with exponentially distributed analysis times over two groups: (1) the mean analysis time ($R = \overline{AT}_{<x}/\overline{AT}$) and fraction ($R = \alpha_{<x}/\alpha$) of easy samples as a function of $x/\overline{AT}$; (2) the mean analysis time ($R = \overline{AT}_{>x}/\overline{AT}$) and fraction ($R = \alpha_{>x}/\alpha$) of difficult samples as a function of $x/\overline{AT}$.

samples (about 10%) with long analysis times should give absolute priority. As expected, the mean system times of both groups ($\overline{T}_{l.p}/\overline{T}_t$ and $\overline{T}_{h.p}/\overline{T}_t$) differ considerably. For these calculations, it was assumed that each sample could be classified in the correct group. As a M/M/1 system is too simple as a representation of the laboratory, simulations were carried out in order to forecast the effect of the mentioned subdivision of the samples in the laboratory under investigation. In the model, easy samples should meet the following conditions: (i) the interpretation time should be smaller than some preset value; (ii) they are analyzed by one method only; (iii) they are measured without waiting until some preset minimal batch size has been reached; and (iv) the analytical results are immediately communicated to the applicant. Comparison of Figs. 14 and 13 demonstrates that the effect of discriminating the samples into two groups in the simulated laboratory is identical to the forecast effect by the theoretical calculations on a M/M/1 system. A consequently executed 4 × 4 × 2 factorial design (Table 7) demonstrates that the accuracy of the forecast of the interpretation time does not significantly influence the delay of both classes of samples. However, it should be stressed that the overall delay is relatively insensitive for the discrimination of easy and difficult samples. Therefore, the subdivision of the samples into two categories with different priorities is only valuable if a shorter delay is the aim for a given group of samples.

Priority differences between samples from various sources. In the laboratory under investigation, samples originate from two different sources (here denoted as F-1 and F-2). A priority difference between both groups results

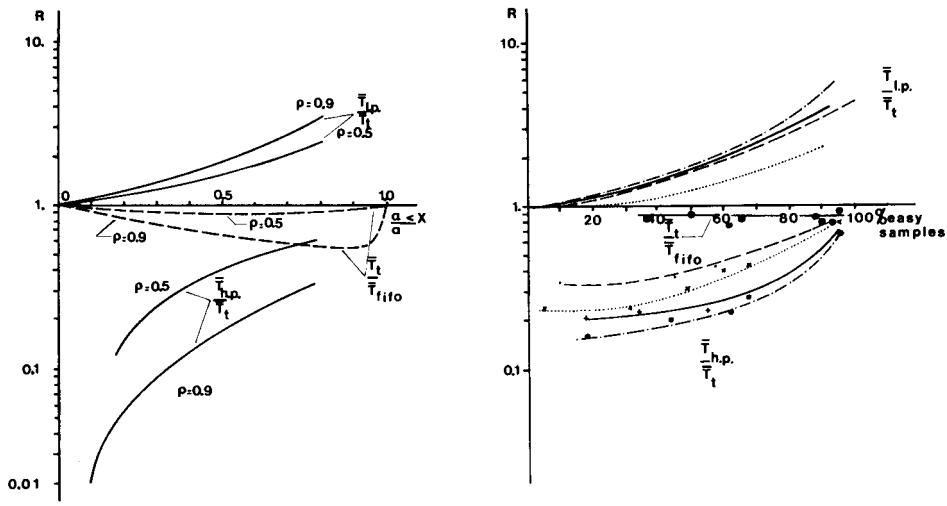


Fig. 13. Ratio (R) of the mean delay of the "difficult" ($\overline{T}_{l.p}/\overline{T}_t$) samples, the "easy" ($\overline{T}_{h.p}/\overline{T}_t$) samples and of the overall delay ($\overline{T}_t/\overline{T}_{FiFo}$), as a function of the fraction of "easy" samples ($\alpha_{<x}/\alpha$) in a M/M/1 system, where the "easy" samples have absolute priority. Broken line, reduction of the mean overall system time ($\overline{T}_t/\overline{T}_{FiFo}$) by discriminating between "easy" and "difficult" samples.

Fig. 14. Simulated mean system time of the "difficult" ($\overline{T}_{l.p}/\overline{T}_t$) and "easy" samples ($\overline{T}_{h.p}/\overline{T}_t$) as a function of the fraction of "easy" samples, where the "easy" samples have absolute priority. Final methods: (—) p.m.r.; (·····) ^{13}C-n.m.r.; (---) i.r.; (· · ·) m.s. (○) Reduction of the overall mean delay ($\overline{T}_t/\overline{T}_{FiFo}$).

TABLE 7

Effect of the accuracy of the estimated interpretation time on the delay of "easy" and "difficult" samples, when "easy" samples have absolute priority (Analysis of variances of a $4^1 \times 4^1 \times 2^1$ design; factor levels: (A) departments; (B) standard error (% of interpretation time) 0, 10, 20, 40; (C) % of easy samples: 35%, 56%.)

Source of variation	"Easy" samples				"Difficult" samples			
	Sum of squares	Degrees of freedom	Mean square	Var. ratio	Sum of squares	Degrees of freedom	Mean square	Var. ratio
(A)	7.84	3	2.61	3.07[a]	326.0	3	108.7	83.8[a]
(B)	0.058	3	0.019	2.18	4.42	3	1.47	1.1
(C)	0.014	1	0.014	1.61	13.9	1	13.9	10.7
residue	0.208	24	0.0087		31.1	24		
total	8.12	31			375.4			

[a]Highly significant (99%).

in a different system time. It is interesting to investigate which group is the more sensitive for the assigned priority and for changes of the sample flow. With regard to sensitivity for the assigned priority, Fig. 15 shows the relative mean delays $\bar{T}_1/\bar{T}$ and $\bar{T}_2/\bar{T}$ of two sample groups plotted against the ratio of their input flows (α_1/α_2) to a M/M/1 system for the situation that the sample group F-1 has absolute priority over group F-2 (full line) and vice versa (dotted line). Figure 15 demonstrates clearly that the delay of the group with the higher input flow is the less sensitive for the attributed priority, e.g. when the input flow $\alpha_1 = 10\,\alpha_2$ and the absolute priority first assigned to group 1 is now attributed to group 2, then the delay $(\bar{T}_2)$ of the latter group is reduced by a factor of 10, while the delay of the former group $(\bar{T}_1)$ is only doubled. As mentioned above, the input flow ratio (α_1/α_2) of both sample sources to the laboratory under investigation is 2.6. For that ratio, the effect of the priority difference $(b_{f_2} - b_{f_1})$ between F-1 and F-2 samples on their mean delay was simulated. The results (Fig. 16) clearly demonstrate that the delay $(\bar{T}_{f_2}/\bar{T})$ of the smaller group of samples is more sensitive for the assigned priority than the delay $(T_{f_1}/\bar{T})$ of the other samples. This completely agrees with the conclusions formulated from the results presented in Fig. 15. Typically, although the various sections have a different number of facilities, different utilization factors, different variation coefficients of analysis time, and different amounts of non-analyzing activities, they exhibit similar behaviour.

Another aspect of the laboratory system is the effect of a variation of the sample flow to the laboratory $(\Delta\alpha/\alpha)$ on the mean delay of sample groups with different priority. The effect of the flow on the delay $(\Delta T/\Delta\alpha/\alpha)$ of the samples originating from both sources was simulated with the earlier

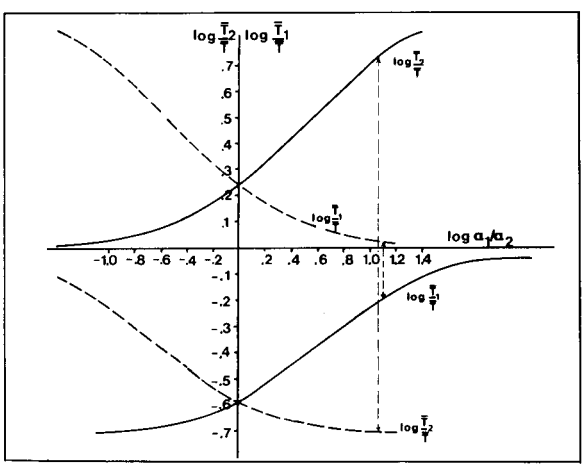


Fig. 15. The system time of two groups of samples with equal analysis time as a function of the ratio of their input density (α_1/α_2) in a system with a total utilization factor $\rho = 0.9$: (—) group 1 has absolute priority; (---) group 2 has absolute priority.

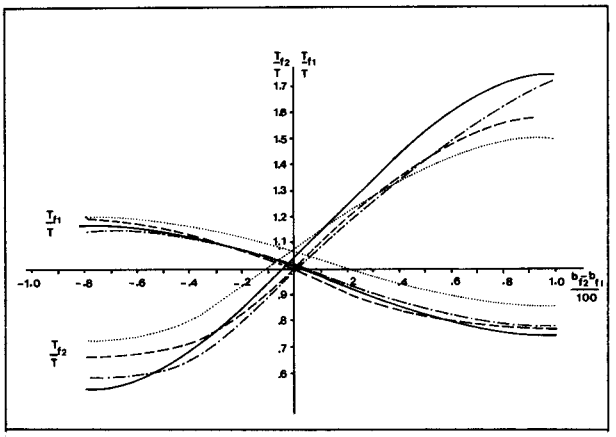


Fig. 16. Simulated mean delay ($T_{f\text{-}1}/T$ and $T_{f\text{-}2}/T$) of a system with two groups of samples (F-1 and F-2) analyzed by the same final method, as a function of the priority difference ($b_{f_2} - b_{f_1}$) between both groups of samples. Final methods: (—) p.m.r.; (·····) ^{13}C-n.m.r.; (---) i.r.; (· · ·) m.s.

model [1]. This sensitivity was simulated for the two extreme situations where one of the groups receives absolute priority over the other. The ratios between the sensitivities of low-priority samples and high-priority samples for the two extreme situations are presented in Table 8 (R_1 and R_2). The values of R_1 and R_2 all exceed unity. This indicates that the high-priority samples are less sensitive to a variation of the flow than the low-priority samples, irrespective of the magnitude of this sample group. Moreover, a comparison of the R_1 and R_2 values (Table 8) indicates that for each section $R_1 > R_2$, i.e. the ratio between the sensitivities of low-priority samples and high-priority samples is the greatest when the largest sample group (F-1) has the absolute priority. As indicated above, the mean delay of this greater group of samples (F-1) in the laboratory scarcely depends on the priority difference with the other group. Now, from the comparison of the

TABLE 8

Sensitivity (s) of the delay for a variation of the input flow ($\Delta\alpha/\alpha = 0.20$) of high- and low-priority samples ($s = \Delta T/\Delta\alpha/\alpha$); flow ratio $\alpha_1/\alpha_2 = 2.6$

Group with absolute priority	f-1 ($\alpha_1 = 8.6$)			f-2 ($\alpha_2 = 3.4$)		
Final method	$s_{f\text{-}1}$	$s_{f\text{-}2}$	$R_2 = s_{f\text{-}2}/s_{f\text{-}1}$	$s_{f\text{-}1}$	$s_{f\text{-}2}$	$R_1 = s_{f\text{-}1}/s_{f\text{-}2}$
I.r.	0.2	0.40	2.0	0.28	0.06	4.7
P.m.r.	0.11	0.24	2.2	0.16	0.05	3.2
M.s.	0.13	0.16	1.2	0.14	0.05	2.8
^{13}C-n.m.r.	0.17	0.21	1.2	0.24	0.08	3.0

effect of the flow on the delay of the greater group of samples, selected as to whether or not they have absolute priority (columns 1 and 4, Table 8), it appears that this effect hardly depends on the attributed priority.

Priority assignment based on the history of the sample. Simulations have demonstrated [1] that the relationship between the delay and the number of analyses executed on the sample is highly dependent on the priority given to the samples, which is correlated with the history of the sample in the laboratory. When priority decreases with the number of instruments involved, there is a strong dependence between delay and number of analyses (Fig. 17). Inversion of this rule leads to an appreciable loss of the dependence. An investigation of the sensitivity of the mean delay of these groups of samples to an increase of sample flow to the laboratory resulted in the observation that this sensitivity is largely dependent on the assigned priority. If samples arriving from outside the laboratory have absolute priority, the delay of the other samples is appreciably sensitive to the total flow (Fig. 18a). In the opposite situation (Fig. 18b), when the priority of the samples increases with the number of visited departments, the dependence of the delay on the flow is similar for all groups. It can also be noted that the delay of the smaller group of samples (Table 9) is more dependent on the attributed priority.

DISPATCHING DECISIONS

The effect of the introduction of a versatile routing procedure in the model was investigated. The routing algorithm takes into account the probability $[j(i)]$ that a section (i) will give the requested information, and the queue length (N_i) in each section, normalized on the total work load of the laboratory:

$$R(i) = f\,(1 - j(i)) + (1 - f)\,N_i/\Sigma N_i \quad (\text{where } 0 < f < 1;\ i = 1,2 \ldots 4)$$

With this algorithm, for each arriving sample, the values $R(1), \ldots R(4)$ are calculated, using some preset value of the weighting factor (f). For $f = 1$, the selection of the section is based only on mentioned probabilities $j(i)$. In contrast, for $f = 0$, the selection is based on the state of the laboratory only. The sample is routed to that section for which the lowest R value is computed. For each sample arriving at the laboratory, the probabilities of a successful analysis were estimated in the model on three levels: $j = 0$: $j = 0.5$ and $j = 1$ (i.e. the analytical procedure is estimated to be incapable of elucidating the requested structure or giving a structure with a probability of 0.5 and 1). The fractions of samples having $j = 0$, 0.5 and 1 were determined for each section (i) from the observed sample flow in the laboratory. The flow to the sections could be reproduced assuming that all samples, for which the probability was estimated that a given method will furnish all requested information, were indeed successfully analyzed. Another possibility is that only a given fraction of these samples are successfully analyzed (i.e. the probability of elucidation of the structure by a given method can be

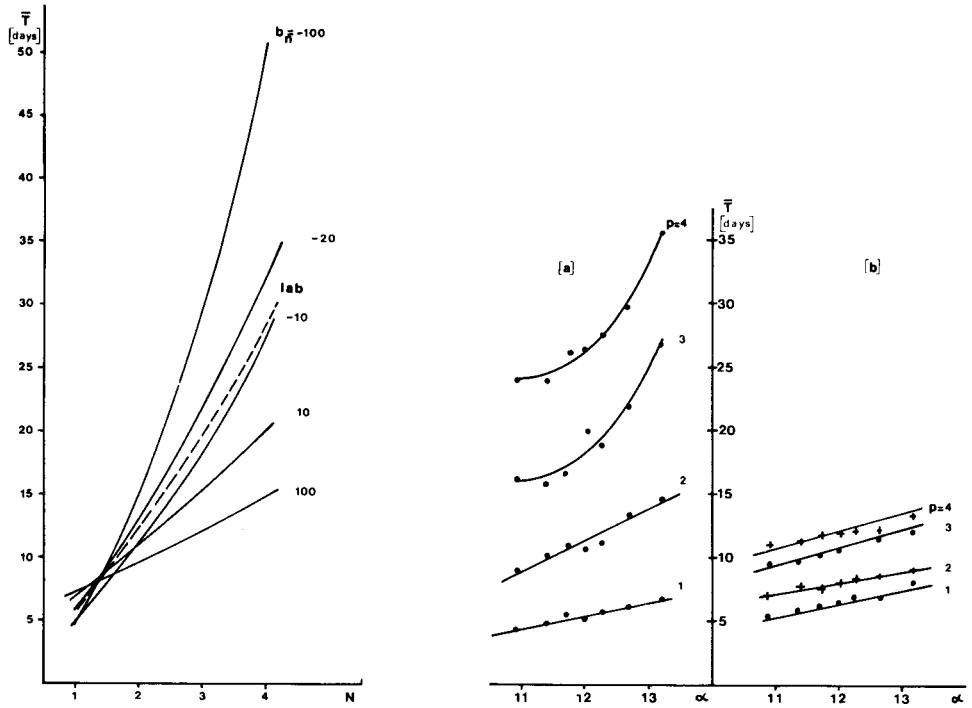


Fig. 17. Simulated mean system time ($\overline{T}$) as a function of the number of visited sections (N) and the priority difference (b_n) between samples that visited n and $n + 1$ sections. Broken line, actual laboratory.

Fig. 18. Flow (α) (samples/day) dependence of the mean delay ($\overline{T}$) of samples as a function of the number of visited sections (p): (a) samples which visited fewer sections have priority; (b) reversed situation.

TABLE 9

Effect of the means and variation coefficients of the measurement (MT) and the interpretation time (IT) on the delay in the i.r. section (Analysis of variance of a 2^4 design. Factor levels: (A) $\overline{MT}$: 0.8 $\overline{MT}$; 1,2 $\overline{MT}$; (B) $\overline{IT}$: 0.8 $\overline{IT}$; 1,2 $\overline{IT}$; (C) C^2_{MT}: 0.5; 2; (D) C^2_{IT}: 0.5; 2)

Source of variation	Mean square	Variance ratio	Effect	Two factor interactions	Mean square	Variance ratio
A	19	6.7[a]	+2.2	AB	6	2.1
B	239	85.3[b]	+7.7	BC	9	3.3
C	16	5.7	−2.0	CD	5	1.7
D	0	—	—	AC	0	—
				AD	0	—
				BD	0	—
				residue	4	2.8

[a]Significant, $1\% < P < 5\%$. [b]Highly significant, $P < 1\%$.

estimated less accurately by the analysts). The flow in the laboratory could be reproduced for up to 16% of unsuccessfully analyzed samples in a method primarily estimated to give the requested information with a probability of 100%. Because balancing the state of the laboratory against the probability of successful analysis will become more worthwhile when these probabilities can be estimated less accurately, the effect of the routing algorithm was simulated for the two extreme situations for which the flow through the laboratory could be reproduced reasonably well.

The model demonstrates that balancing the probability of obtaining the requested information against the number of samples in the sections decreases the delay. However, the effect on the delay is relatively small (ca. 12%) (Fig. 19). In terms of variation coefficients of the overall delay, the model is insensitive to that strategy. Attributing excessive importance to the number of waiting samples ($f < 0.2$), the mean number of visited departments increases from 1.26 to 1.40, resulting in an increase of the delay which is very sensitive to both that number and the flow (Fig. 20). The more

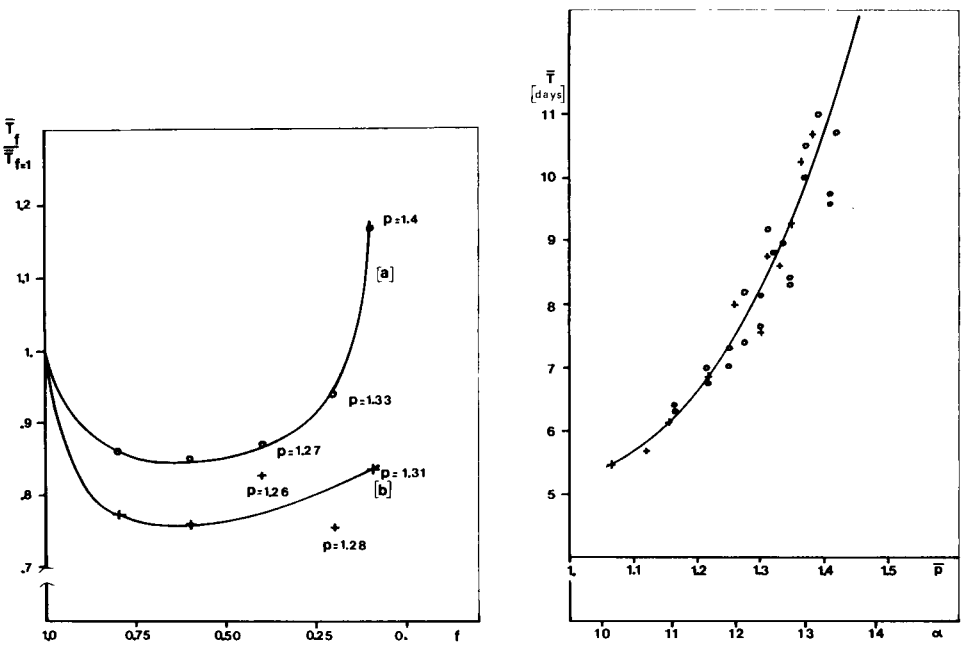


Fig. 19. Simulated mean system time ($\overline{T}_f/\overline{T}_{f=1}$) of all samples as a function of the weighting factor (f), balancing the queue sizes in the sections against the probability for a section to give the requested information. $f = 0$: sample routing based on the queue sizes exclusively. $f = 1$: sample routing based on the probabilities exclusively. (p: mean number of visited sections).

Fig. 20. (○) Simulated effect of the input flow (α) to the laboratory (samples/day) on the overall system time ($\overline{T}$). (+) Simulated effect of the mean number ($\bar{p}$) of visited sections on the mean system time ($\overline{T}$).

inaccurate the estimated probability is that some section will give the requested information (line b in Fig. 19), the more useful it will be to balance this probability against the number of samples in each section. Even when almost only the state of the laboratory is considered in the routing algorithm, no increase of the mean number of visited departments is observed, and the mean delay diminishes.

TECHNICIAN ASSIGNMENT DECISIONS

A completely decentralized organization was assumed for the simulation of the actual laboratory. The technician is always assigned to the same section, irrespective of the state in the other sections. In contrast, in a centralized organization, the experienced technician balances his experience with the different analyses against the state (queue lengths) in the sections. In the model, a relationship is assumed between the experience of a technician (j) and the mean time he needs to do method (i) $\overline{AT}_j = \overline{AT}/\exp(j,i)$. When the technician always selects the method with which he has most experience, regardless of the availability of a more experienced technician, there is a disastrous effect of allowing inexperienced technicians to analyze samples ($\exp(i,j) < 0.8$) (Fig. 21). A reduction of the overall delay is achieved only when all technicians are fully qualified for all methods. When a technician is authorized to execute an analysis for which he is not fully qualified, provided

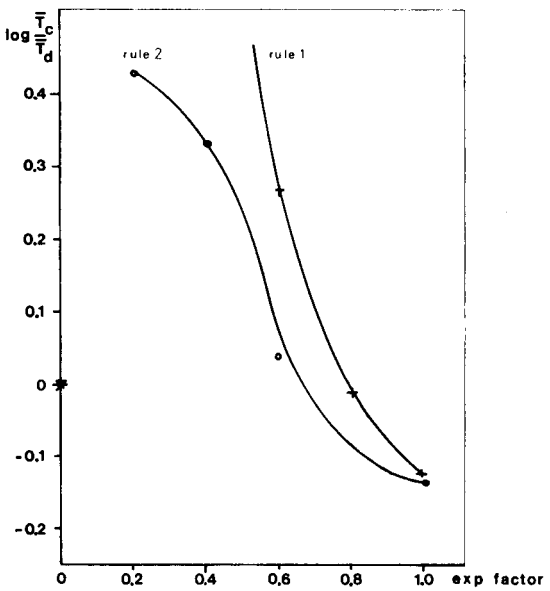


Fig. 21. Simulated effect of the technician assignment decisions on the mean system time. Exp. factor = 0: completely decentralized organization (mean system time: $\bar{T}_d$). Exp. factor = 1: completely centralized organization (mean system time: $\bar{T}_c$). For rules 1 and 2 see text.

that a fully qualified worker is not idle, a somewhat smaller effect on the delay is observed. However, the conclusion remains valid that under the condition of the laboratory the introduction of a decentralized organization is sensible only if all technicians are almost fully qualified in the other methods. Clearly, the inclusion of the lengths of the various queues in the decision as to which sections will be selected next, will not influence the effect of centralizing the organization, when all technicians are fully qualified for all departments.

STRATEGIES ON THE TERMINATION OF AN ANALYSIS

In the model, a maximal allowed analysis time can be selected [1] from the equation $AT_{(\max)} = \overline{AT}(1 + kC^2_{AT})f$, where the value of f depends on the simulated strategy and C^2_{AT} is the variation coefficient of the analysis time. There are then four possibilities: (a) the maximal analysis time is independent of the properties of the sample or state of the laboratory ($f = 1$); (b) the maximal analysis time increases with the number (N) of visited sections ($f = 1 + N$); (c) the maximal analysis time decreases with the queue length in section (i) ($f = \Sigma N_i/N_i$); (d) the maximal analysis time decreases with decreasing probability that the section may solve the analytical problem ($f = 1 + j(i)/50$).

In contrast to the expectation based on the study of a single M/M/1 system [1], which is not embedded in a network, no improvement of the delay could be found. Strategies (a) and (b) cause an increase in the delay, even with 50%. This is due to an increase of the mean number of visited departments from 1.27 to 1.47 and 1.52, respectively. Apparently, the effect of an augmentation of the mean number of visited departments (20% for strategy a) surpasses the effect of a smaller mean analysis time (13.5% for strategy (a)). It causes an increase in the utilization factors, and consequently the delay. By the application of strategies (c) and (d), the overall delay was not significantly influenced with respect to the situation where no maximal analysis times were imposed.

COMPARISON OF THE EFFECTS OF THE MEANS AND VARIATION COEFFICIENTS OF THE MEASUREMENT AND INTERPRETATION TIME ON THE DELAY

Calculations on the M/M/1 system presented previously [1] forecast a greater effect for the means compared with the variation coefficients of the measurement and interpretation time. An analysis of variance on a four-factorial two-level design applied on the i.r. section of the model, confirmed that conclusion for more complex systems (Table 9). A variation of the measurement time from 1.2 to 0.8 times the original mean value reduces the total analysis time by about 10%. As the mean interpretation time is about three times as long as the measurement time, the same variation of the interpretation time reduces the total analysis time by 30%. The variance

ratios (Table 9) demonstrate that a considerable variation of the variance coefficient (i.e. a reduction to 25% of its original value) affects the mean delay to the same extent as a reduction of the analysis time by 10% only. This demonstrates the greater sensitivity of the delay for the mean analysis time in comparison with that for the variation coefficient.

Conclusions

It can be generally concluded that the simulated effects with the laboratory model are less pronounced than predicted for single queueing systems. Because the work load of the laboratory, which is the product of the mean number of arrivals per day and the mean number of visited sections, affects the delay considerably, all decision rules that increase that product affect the delay negatively.

An example is that the effect of reducing the mean analysis time by introducing a maximal analysis time is completely surpassed by the consequent increased number of visited sections. The effect of an increase in the work load is different for the various groups of samples with different flow and priority. The total delay is reduced when absolute priority is attributed to the easy samples. This reduction is relatively insensitive for the limiting analysis time of easy samples and for the estimation error of the analysis time of the samples. The performance of the laboratory model is enhanced if the probabilities that the various departments will give the requested information are considered along with the state of the laboratory, in order to route the sample. The effect is more pronounced when these probabilities can be estimated less accurately. The transition from a centralized to a decentralized organization is only advantageous when all technicians are fully qualified in all methods.

The ultimate delay for an analytical result to be available bears little correlation to the number of samples waiting in the laboratory. This is probably due to a random sequence dispatch of the samples with equal priority. By the change of the random sequence to a FiFo sequence, however, a reasonable correlation is obtained. Combining the $AR(1)$ model of the number of samples in the system with the conditional probability density function of the delay, a threshold control system will be created.

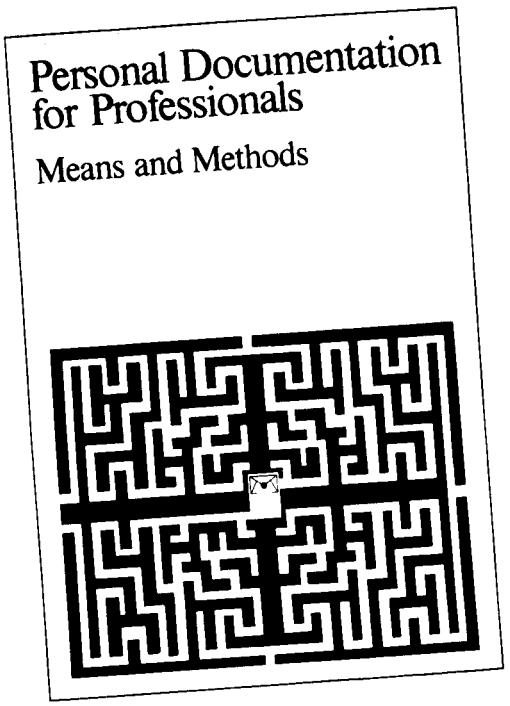
The author thanks F. J. Sprint and all members of the Analytical Research Group of Philips Duphar B. V. Weesp, The Netherlands, for many helpful discussions and the collected data.

REFERENCES

1 B. G. M. Vandeginste, Anal. Chim. Acta, 112 (1979) 253.
2 A. J. Lemoine, Manage. Sci., 24 (4) (1977) 464.
3 T. H. Naylor, J. L. Balintfy, D. S. Durdick and Kong Chu, Comput. Simul. Tech., Wiley, New York (1966).
4 L. Kleinrock, Queueing Systems, Vol. 2, Wiley, New York (1976).
5 J. M. Holtzman, Oper. Res., 18 (1970) 461—468.
6 R. W. Conway, W. L. Maxwell and L. W. Miller, Theory of Scheduling, Addison—Wesley, Reading, MA, 1976.

ANALYTICA CHIMICA ACTA, VOL. 122 (1980)
(Computer Techniques and Optimization, Vol. 4, No. 4)

AUTHOR INDEX



Personal Documentation for Professionals

Means and Methods

by **V. STIBIC**, *Philips, Ltd. Information Systems and Automation, Eindhoven, The Netherlands*

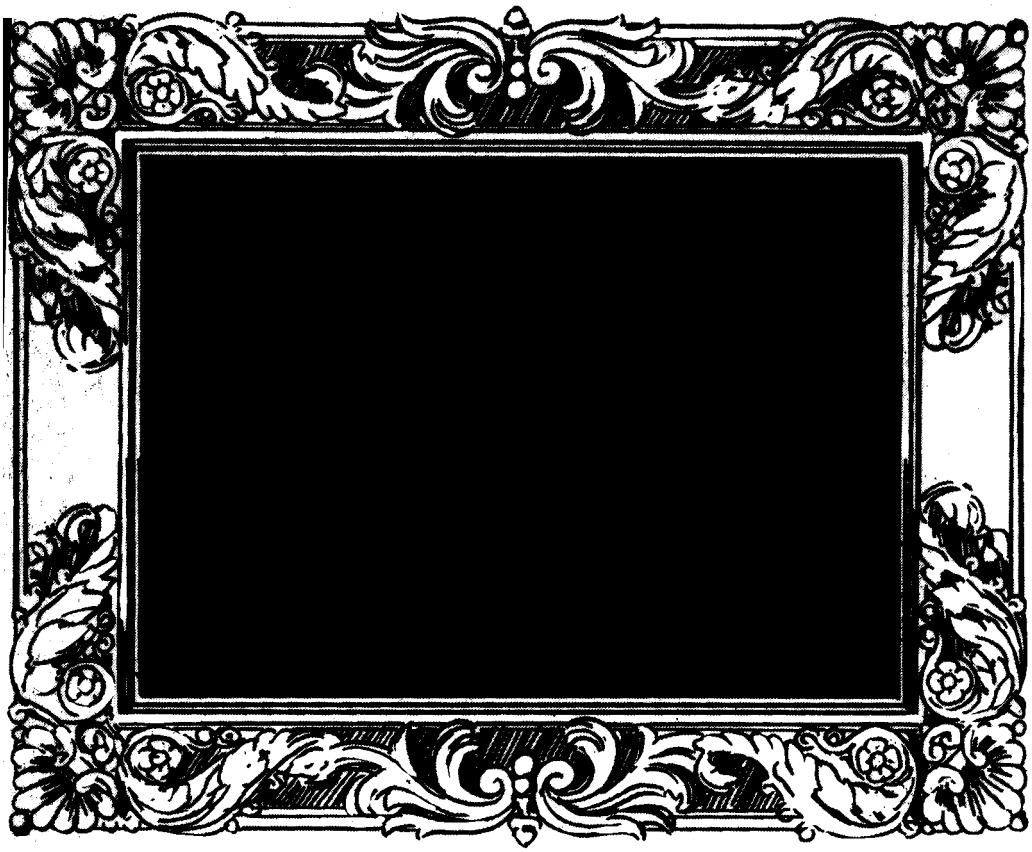
1980 224 pages, 119 fig., 232 ref.
Price: US $29.25/Dfl. 60.00
ISBN 0-444-85480-0

Every professional person - scientist, researcher, technician, manager - owns his/her personal collection of documents (books, reports, journals, cuttings, photocopies, slides, microfiches, drawings, etc.) that must be well organized and accessible at any time. Professionals need their own personal documentation system.

Personal documentation is an important tool that has hitherto received little attention. Fortunately, there are technical and methodological developments now in progress that can help a professional. This book describes the growing need for better personal documentation and the means and methods that can be used to provide it. It is one of the first books aimed at this important subject.

The work describes the organization of personal files, their structure, content, classification and indexing, as well as the technical means that can be employed for filing the original documents and for retrieval. The retrieval instruments, ranging from the cheap and simple card systems and peek-a-boo cards, to the computerized indexes, use of computer terminals, and personal microcomputers, are characterized, compared, and evaluated. The advantages and weak points of these techniques are explained and demonstrated via many examples. Four case histories describe the practice of personal documentation of a professional. Special attention is devoted to the methods of document classification and indexing: use of hierarchical decimal classification (Universal Decimal Classification in simplified form for personal use), free indexing and subject headings in a personal collection, use of existing thesauri and design of one's own personal thesaurus, and the principles of automatic indexing by means of the computer are explained and demonstrated.

This useful and absorbing book contains 119 figures and 232 references, and is a **must** for every professional.



What Industry would be without computers.

Computers?
You as an engineer don't need them as toys. Playing is for kids and Einsteins. But you need them badly as tools.
You want to forget about them as soon as you can. But not sooner. You and your company thrive on them. Or worry about them.
Read Computers in Industry, from cover to cover, 4 times a year and subscribe to it. You need it more than you know. And so does your Industry.

COUPON FOR A FREE COPY

For a free copy of the first issue of **Computers in Industry,** please write or complete the coupon and return it directly to the publisher:

North-Holland Publishing Company
Attn: Mr. J. Dirkmaat
P.O. Box 211,
1000 AE Amsterdam, The Netherlands

Name: _____

_____

Address: _____

_____

_____

STATISTICAL TREATMENT OF EXPERIMENTAL DATA

By J.R. GREEN, *Lecturer in Computational and Statistical Science, University of Liverpool, U.K. and* D. MARGERISON, *Senior Lecturer in Inorganic, Physical and Industrial Chemistry, University of Liverpool, U.K.*

PHYSICAL SCIENCES DATA 2

This book first appeared in 1977. In 1978 a revised reprint was published and in response to demand, further reprints appeared in 1979 and 1980. Intended for researchers wishing to analyse experimental data, this work will also be useful to students of statistics. Statistical methods and concepts are explained and the ideas and reasoning behind statistical methodology clarified. Noteworthy features of the text are numerical worked examples to illustrate formal results, and the treatment of many practical topics which are often omitted from standard texts, for example testing for outliers, stabilization of variances and polynomial regression.

What the reviewers had to say:

"The index is detailed; the format is good; the presentation is clear; and no mathematics beyond calculus is assumed".
–CHOICE

"A lot of thought has gone into this book and I like it very much. It deserves a place on every laboratory bookshelf".
–CHEMISTRY IN BRITAIN

1977. Reprinted 1978, 1979, 1980.
xiv + 382 pages
US $39.25/Dfl.90.00
ISBN: 0-444-41725-7

ELSEVIER
P.O. Box 211, 1000 AE Amsterdam, The Netherlands.
52 Vanderbilt Ave., New York, NY 10017.

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

7/84

VOL. 122, No. 4 ANALYTICA CHIMICA ACTA December 1, 1980

(Computer Techniques and Optimization, Vol. 4, No. 4)

CONTENTS

Printed in The Netherlands.