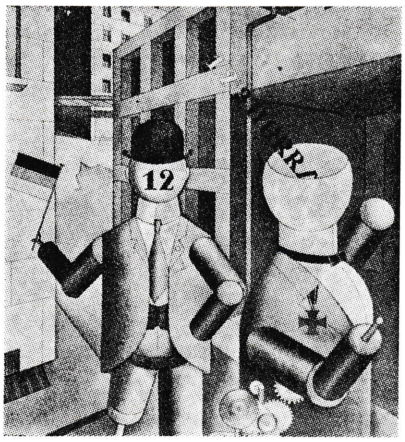
自動分類法のソフトウェア

大隅　昇

『数理科学』No. 230, AUGUST 1982 別刷



G. Grosz：共和国のオートマトン (1920)

自動分類法のソフトウェア

大隅　昇

1. はしがき

ある書によると，“分ける”ということは“分かる”ことに通ずる，あるいは“分かる”ためには“分ける”ことが必要である，という．近来，この分けること（分類）の総称としてクラスター分析という用語をあてることが多い．クラスター分析は，複雑な様相を示す測定データを適当に分類しその分類情報の中からデータに潜在する特徴を把え解釈を容易にするための方法論であり，まさに分けて事を知るための道具である．

分類はあらゆる研究分野において共通にみられる手続きであるから類似の概念が無数にあっても不思議ではない．たとえば，“数値分類”，“自動分類”その他多彩な類似の用語があふれている．Blashfield の報告にもあるように広汎な研究分野で類似の概念にまったく異なる用語があてられ，これに伴う混乱が起きている事も事実である[6]．

これらの事情も含めて，クラスター分析の実状や問題点については本誌を借りてすでに3回にわたり述べたが現在もこれらとさしたる違いはないと考えている[27),28),29)]．本稿はこれらの報告を枕に，分類に関連した計算プログラム・ソフトウェアの紹介に焦点をあてて述べるとしよう．

ところでこの種のソフトウェアについて触れる場合計算機科学そのものの歴史的変遷を無視して語るわけにはいかない．分類手法の多くは計算処理の膨大さ，繁雑さを前提とした，大型電子計算機の利用を不可欠のものとして登場したということがある．これは統計ソフトウェア全般に共通のことであるが，1960年代後半における欧米諸国の大型計算機とそのソフトウェア，とくに体系化されたオペレーティング・システム (OS) の出現により（たとえば IBM 360 の OS），様相が一変しこれを機に加速的に数多くの手法，算法，プログラム・ソフトウェアが登場するのである．

2. 自動分類法とは

すでに本誌の 190 号（1979）で触れたことであるが，分類手続きの繁雑な処理過程を算法として具体化しそのプログラムは人が提供するが，分類処理自体はすべて電子計算機に委ね自動的にデータを分類する，という意味では“自動分類 (Automatic Classification)”という用語がもっとも適切であろう．この意を汲んで本稿では自動分類法とそのソフトウェアということで話を進める．

ところで，電子計算機で利用できる自動分類のソフトウェアといっても様々の段階がある．また実際に多くの分野で多種多様なプログラムが見られるのであるが，大別すると次のようになろう．

(1)特定の手法の解析のために作成された独立プログラム．

(2)複数の手法のサブルーチンあるいは個別プログラムの集まり．

(3)手法や処理課題を統一的に集約した専用システム．

(4)大規模な汎用統計ソフトウェアの一部として収納されている場合．

(1)は，たとえば階層的手法の特定のものだけが利用できるような場合である．著作物，テクニカルレポートや雑誌（*Applied Statistics*, *The*

Computer Journal など）に紹介される算法やそのプログラムリストなどもこれに相当する．（2）は複数の課題（分類手法とその他のデータ加工処理ルーチンなど）をサブルーチンあるいは個別プログラムとして登録してあり，必要に応じて連結利用するもので，この場合には利用者はOSやプログラム作法について若干の知識が要求される．（3）は自動分類に関連した手法を多数収納した上で，データ加工・変換・表示（たとえば類似度・非類似度行列の算出，行列演算機能，結果の図的表示など）の機能を専用の命令語を用いて各課題を自由に連結し自動分類に関わる計算分析処理を統一的に行うことができるものを指している．また，汎用的な統計ソフトウェア，たとえばBMDP, SAS, OSIRIS, GENSTAT などには数が少ないが自動分類手法が取り入れられている．これが（4）である．さらに最近は端末利用が盛んであるから端末モードでの利用の可否も十分考えておかねばならない．

（1）や（2）に属するプログラム類は無数にあると思われるので，ここでは筆者が目を通しあるいは試用したものに限定して最後に文献リストとして挙げておいた．本稿では，（3）または（4）に属するもので実用の見地から注目に値する2，3のソフトウェアを取り上げその特徴を紹介する．

ところで国内は別として，欧米諸国には無数の統計ソフトウェアがある．最近出版されたFrancisの著書で取り上げたものだけでも120種近くあるが，これでもすべてを網羅しているとは思えない[11]．一方，自動分類法のソフトウェアに目を向けるといささか事情が異なるようである．まず第1に，圧倒的に個別のプログラムが多く，システム化の度合が高いソフトウェアは少ないということである．多くは限られた研究分野の専用のデータ解析手法であることから規模が小さいことは予想されることだが，システム化という点でいまだ他の統計ソフトウェアに匹敵するものが見られない．理由はいろいろあろうが，1つには汎用統計ソフトウェアのように組織的な作業グループの支援を得て開発を進めるというよりは，個人的な研究の副産物として誕生した小規模のプログラム集合をシステムとしてまとめるという傾向が見られるということがある．別の理由として，研究分野としての歴史が浅くこの20年位の間に急速に進展したために，他の統計的データ解析のように早くから電子計算機との接触を得ることが少なかったという事情もあろう．またシステム化を阻害する別の理由に，分類手法が関連する研究分野が広範囲にまたがるためそこで用いる概念や知識をある程度まで集約して理解することを強いられる上に，ソフトウェアとして具体化するプログラム作法の技量がかなり要求されるということがある（多くの手法やその算法は，組み合わせ的要素やアドホックで発見的な部分が多いのでプログラム作成の手間がかかる）．

3. 自動分類法のソフトウェア

現存する自動分類法のプログラムを利用する最も手近な方法は雑誌や著作物に掲載のプログラムリストをそのままコピーして用いることである（Anderberg, Hartigan, Späth など[1),13),22)]）．しかし利用者の要求も次第に複雑になり，これだけでは十分とはいえないことも事実である．複数の手法を用いた分類結果を互いに対比分析したい，分類結果を他の副次的要因と対比したい，他の解析ソフトウェアと処理結果を相互利用したい，さらに分類結果が妥当か否か体系的に評価分析したい，……などの要請がそれである．こうした要求を満たした上で，さらに利用者が簡単に利用できるようなシステム化の進んだソフトウェアは極めて数が少ないがそれでも注目すべきものがいくつかある．本稿では，CLUSTAN, NT-SYS, MINTS, CLASS の4種についてその機能を簡単に紹介する．ここでCLASS を除く3種は，全体の機能を統合する主プログラムが1つあってこれにモジュール化された複数の副プログラム群が連結するという構造を持つが，これらに共通した特徴として次の事がある．

（1）専用の命令語を持つので利用者はこれを用いて解析を淀みなく進められる．

（2）複数のデータセットに対して複数の課題処理が可能である．

（3）入力データの型と形式が多様である（質的・量的データ，類似度・非類似度行列など）．

(4)作業ファイルを介して他の統計解析ソフトウェアとの接触面を持ち分析結果のやりとりができる.

(5)データの入力媒体として，カード，磁気テープ，磁気ディスクのいずれも利用できる.

一方，CLASS は複数の課題別の主プログラムの集まりで，これらを個別に呼び出して利用する．課題間のデータのやりとりは作業ファイル，カードパンチ出力によるカードの再入力などによる.

こうしたソフトウェアは，その設計思想，開発意図，利用分野などがその性格，特色に大いに反映すると思われるがこれについての論述はここではさし控え，むしろ例を通して各ソフトウェアの特徴を個別に述べる程度に留める.

3.1 CLUSTAN

この種のソフトウェアとして最も良く知られたもので開発者は D. Wishart である．1969年に Kansas 大学のモノグラフとして CLUSTAN-I が公表され以後 1A, 1B と改良を続け1978年の CLUSTAN-1C の改訂版の登場まで約10年近くの日時を費やしている．実際に開発を思い立ったのは1960年頃であるというから，これを数えると実に20年以上もたつ[36),45)]．1978年版によると CLUSTAN ジョブは原則としてバッチモードで利用する．命令語としてキーワードとパラメータがあり，前者は課題名の指定，後者はその指定課題の処理内容の明細(処理条件，オプション指定)を指示する．ここで簡単な例をみるとしよう(図1).

用いたデータは Slagle からの引用で国連の投票行動に関するもので19ケ国の代表による14項目の審議事項についての賛否を調べたものである[25)]．いまこの19ケ国の分類を考える．まず①は入力データの定義部である．キーワード FILE でデータ入力の発生を指示し続いてタイトルカード，パラメータカード（サンプル数，変数の数，その他のデータ属性を指定），データ入力書式指定カード，データ・デックと置く．以上で作業ファイル上にデータが格納されるので次に課題処理に入る．②でキーワード LABEL によりデータ行列の行（国名）と列（項目名）に名称を付与する．③で CORREL により距離の計算を指示し同時に次のパラメータカードに処理条件を与える(1でユークリッド距離，4で 4-linkage リスト，つまり各個体から第4近傍までの情報の算出とファイル格納)．③で求めた距離行列を用いて④の階層的分類法 (HIERARCHY) を実行する．その条件はパラメータで群平均法(3)，最小クラスター数(2)から最大クラスター数(5)までのクラスター情報の算出を指定する．次に⑤で④の結果をデンドログラムとして表示する．図の型式をパラ

図 1
CLUSTAN の入力例

```
FILE
CLUSTAN TEST SCHEDULE
  19  14   0 C
    (15X,14F2.0)

AUSTRALIA        1 3 3 3 1 3 1 3 3 1 1 3 3 1
BRAZIL           1 2 2 2 1 3 3 1 3 1 1 3 1 1
BULGARIA         1 1 1 1 3 1 1 3 1 3 2 2 1 3
DAHOMEY          1 3 3 3 1 3 1 3 5 1 3 1 2 2
FRANCE           1 3 3 3 3 1 2 3 3 1 1 3 2 2
KENYA            1 3 3 3 3 1 1 3 2 5 3 1 1 3
MEXICO           1 2 2 2 1 3 3 1 3 1 1 1 2 1
NEW ZEALAND      1 3 3 3 1 3 1 1 3 1 1 3 3 1
NORWAY           1 3 3 3 3 1 3 2 3 1 1 3 3 1
SENEGAL          1 3 3 3 1 2 2 2 2 1 3 1 1 2
SWEDEN           1 3 3 3 3 1 2 3 3 1 1 3 3 1
SYRIA            1 2 2 2 3 1 1 3 1 2 3 1 1 3
TANZANIA         1 2 2 2 3 1 1 3 2 5 3 1 1 3
ARAB REPUBLIC    1 3 3 3 3 1 1 3 2 2 3 1 1 3
UNITED KINGDOM   1 3 3 3 1 1 1 3 2 3 1 1 3 3 1
U.S.A.           1 3 3 3 1 3 3 1 3 1 1 1 3 3 1
U.S.S.R.         1 1 1 1 3 1 2 3 1 3 2 2 1 3
VENEZUELA        1 2 2 2 1 3 3 1 3 1 2 1 1 1
YUGOSLAVIA       1 3 3 3 3 1 1 3 1 2 3 1 1 2

LABELS
X X
VAR1    VAR2    VAR3    VAR4    VAR5    VAR6    VAR7    VAR8    VAR9    VAR10
VAR11   VAR12   VAR13   VAR14
AUSTRL  BRAZIL  BULGAR  DAHOMY  FRANCE  KENYA   MEXICO  NWZLND  NORWAY  SENEGL
SWEDEN  SYRIA   TANZNA  ARAB    UK      USA     USSR    VENZEL  YUGO

CORREL
 1 4
HIERARCHY
3  2  5
TREE
2
STOP
```

図 2　CLUSTAN の出力例

```
PROCEDURE TREE
==============

TREE INPUT PARAMETERS

NUMBER OF INDIVIDUALS=  19
SUBTREE PARAMETER K=    0 CLUSTERS
TRANSFORMATION SELECTED= 0
TYPE OF TREE REQUIRED = 2
DENDROGRAM

    ITEMS GROUPED                1       5       7      17      12
                                   8      11      18       6      14
CYCLE   I     J     COEFF          15       9       4      13      19
                                      16       2       3      10
                                 *                                   *
    1     3    17       0.072     * | | | | | | | | | | |_| | | | | | *
    2     5    11       0.143     * | | | | |_| | | | | | | | | | | | *
    3     7    18       0.143     * | | | | | | | | |_| | | | | | | | *
    4    14    19       0.143     * | | | | | | | | | | | | | | | |_| *
    5     5     9       0.215     * | | | | |  |__| | | | | | | | |  | *
    6     6    13       0.215     * | | | |  |    | | | | | | |_| |  | *
    7     1     5       0.266     * |_| | |  |    | | | | |  |   |  | *
    8    12    14       0.266     *  |  | |  |    | |_| | |  |   |  |__| *
    9     2     7       0.358     *  |  | |  |    |  |__| |  |   |  |  | *
   10    15    16       0.358     *  |  |_|  |    |     | |  |   |  |  | *
   11     1    15       0.536     *  |___|  |    |     | |  |   |  |  | *
   12    10    12       0.739     *    |    |    |     | |  |   |  |___| *
   13     1     5       0.768     *    |____|    |     | |  |   |      | *
   14     3     6       1.072     *         |    |     | |  |___|      | *
   15     3    10       1.125     *         |    |     | |      |______| *
   16     1     2       1.180     *         |____|     | |      |      | *
   17     1     4       1.506     *              |_____| |      |      | *
   18     1     3       2.245     *                    |______|      | *
                                 *                                   *
                                 * * * * * * * * * * * * * * * * * * * *

PROCEDURE COMPLETE
```

メータで2（垂直型）と指定する．⑥でCLUSTAN ジョブの終了を指定する．①で格納したデータについてさらに別の課題処理を指定する場合には必要課題のキーワードとパラメータを順次置けばよい．別のデータセットについての解析も希望する場合は FILE 以下を上と同じように反復する（ただし前に格納のデータセットは消滅する）．こうして命令語の簡単な指定で連続的に解析を進めることができる．パラメータの指定が固定書式であるというわずらわしさがあるが利用者が誤まった指定をするとエラーメッセージが出力されてジョブは中断することなく進む．こうしたエラー判定，回避に対する手当てが行き届いているのがシステム化されたソフトウェアの特徴でもある．

CLUSTAN の課題名と機能を表1に一覧にしたがさすがに種類が豊富でしかも処理条件もそろっているが，命令語が固定書式であることや，手法によって扱えるデータ数の制約を受けるといった欠点もある．新版の案内をみると，端末モードで利用できるようにプリプロセッサーが用意されているのでこうした欠点も改良されていると思われる．なお，図2に上の例の出力の一部を挙げておいた．

3.2 NT-SYS

NT-SYS (Numerical Taxonomy System of Multivariate Statistical Programs) は J. Rohlf らによって開発されたソフトウェアで表2の一覧にみるように分類手法以外の多次元データ解析手法も含んでいる．最大の特徴として命令語 (NTSYS コマンド) の指定がほとんど自由書式であるということがある．命令語の規則を簡単に説明しよう．

(1)原則として命令語はカードで与える．

(2)カードの第1欄に命令語を表わす制御記号(*)を1文字与え，これに続いて命令語の機能を指定する課題識別子を与える．

(3)続いて指定課題の明細をキーワードあるいはパラメータにより指示する．これらの指定順は任意でカードの72欄のどこに位置してもよい．

(3)*で始まる各命令語に対し心憶えの注釈文，表題文を120字以内で与えることができる．このとき ¥($) 記号でその文章の前後をくくる必要がある．

(4)課題識別子，キーワード，パラメータは任意の長さに省略縮小できる（ただし他の命令語と識別できる範囲で省略）．

ここで例をみよう(図3)．まず *FILE で作業ファイルの割当てを行う．次に *INPUT で入力データに 'DEMO' と名称を付与し，さらに行列の大きさを (19, 14) と与える．入力データの読み取り条件を指示し (READ, STORE)，データに

表 1　CLUSTAN の命令語とその機能の概略（1978年版）

	課題名 キーワード	概要
分類手法	HIERARCHY	凝集型階層的分類法（いわゆる組合わせ的手法で表わせる8手法がある．Single linkage, Complete linkage, 群平均法，セントロイド法，メジアン法，ウォード法，可変法，Mc-Quitty の方法）
	CENTROID	HIERARCHY のセントロイド法に同じ．
	RELOCATE	k-means 法（初期分割として，直前の手法の結果の利用，ランダム配置，系統的配置などがある）
	MODE	モード法
	DENSITY	モード法の変形
	EUCLID	Gordon-Henderson の方法（非線形計画を用いたクラスター内平方ユークリッド距離の和の最小化）
	DIVIDE	Association Analysis
	KDEND	Jardine-Sibson の B_k 法
	DNDRITE	Calinski-Harabasz の方法（MSTの張り木の長いほうから順次切り落としながらクラスター化を行う）
データ入力と加工	FILE	データ行列の入力，主成分分析の計算
	DISTIN	類似度・距離行列データの入力
	SPSS	SPSS SAVE FILE からの入力
	RESTART	格納ずみデータ・ファイルの再利用指示
	CORREL	各種の類似度，距離の算出（平方ユークリッド距離，相関係数，一致係数，Jaccard 係数等あわせて，40通り近くある）
出力機能	RESULT	分類結果の各種情報の出力（データリスト，平均，分散，相関係数，主成分分析の結果等）
	DUMP	CLUSTAN データ・セット　ファイル CLUSDECK の内容出力
	LABELS	変数，個体（ケース），主成分などへの名称付与
	PLINK	デンドログラム等の出力（XY プロッター）
	TREE	デンドログラムの出力（ラインプリンター）
	SCATTER	各種クラスター化情報の XY プロッター出力（クラスターサークル，MST，散布図，クラスターダイアグラム等）
その他	COMMENT	注釈文の指示
	SIZE	データ格納領域の大きさ変更
	TRACE	ファイル入出力操作に関するシステムテスト用
	STOP	CLUSTAN ジョブの終了指示

図 3　NT-SYS の入力例

```
*FILES SCRSTK=11,SCRATCH=300(12,13,14)
*INPUT NAME='DEMO'(19,14),READ BY ROWS,STORE BY ROWS,
*ETC NUMBER=8,OPTION=LABELS,
*ETC ¥ THIS IS A DEMO-SCHEDULE FOR NTSYS ¥
*FORMAT(15X,14F2.0)
①
AUSTRL  BRAZIL  BULGAR  DAHOMY  FRANCE  KENYA   MEXICO  NWZLND
NORWAY  SENEGL  SWEDEN  SYRIA   TANZNA  ARAB    UK      USA
USSR    VENZEL  YUGO
ITEM1   ITEM2   ITEM3   ITEM4   ITEM5   ITEM6   ITEM7   ITEM8
ITEM9   ITEM10  ITEM11  ITEM12  ITEM13  ITEM14

      データセット(図1と同じもの)

END OF DATA
②*OUTPUT OPER='DEMO'
③ *ALGEBRA EQUATION="'TEST'='DEMO'T"
④ *SIMINT OPER='TEST',EUCLIDEAN='A'
⑤ *OUTPUT OPER='A'
⑥ *MSTSNGL OPER='A',OPTION=LOW,MST='B',TREE='C'
⑦ *OUTPUT OPER='B'
⑧ *PHENOGRAM TREE='C'
⑨ *FINISH
⑩ *FINALE
```

標識の必要があること（OPTION＝LABELS）とそのカード1枚当りの項目数（NUMBER＝8）を指示する．命令語が1枚のカードに入らないので ＊ETC で継続指定を行う．最後の ＊ ETCカードに表題を与える．次にデータの入力書式を指定（FORMAT）し続けてデータブロックを置く．行側の国名，列側の項目名，次に分類したいデータ行列，最後にデータブロック終了を示す END OF DATA カードを置く．続いて，データリストを出力し(②)，それを転置した行列に 'TEST' と名称付与

表 2　NT-SYS の命令語とその機能の概略（1977年版）

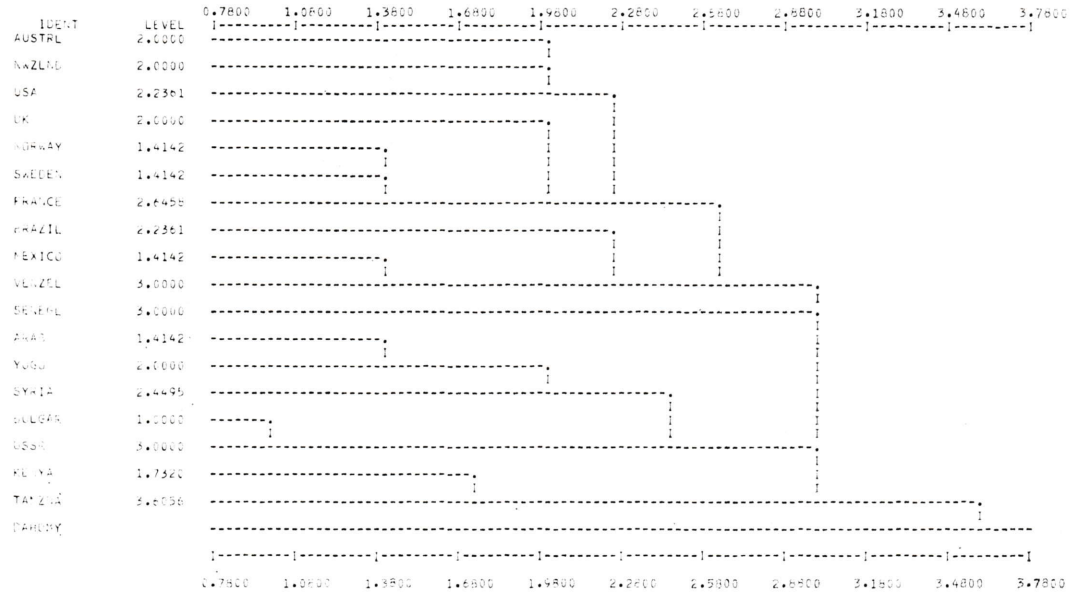
	課題識別子	概　　要
分類手法・その他の解析課題	AHCS	Rohlf の方法 (Adaptive hierarchical method)
	FACTOR	主成分分析，因子分析（手法の選択，固有値解法の選択などが可能）
	GOWER	Gower の主座標分析
	KGRAPH	Jardine-Sibson の B_k 法 (Rohlf の改良アルゴリズム)
	KGROUP	Functionpoint cluster analysis（一種の分割型再配置法）
	MDSCALE	Kruskal の MDSCAL
	MSTSNGL	MST (Minimal Spanning Tree), Single linkage 法
	SUBSETS	Sale の方法 (Complete linkage 法に類似の方法)
	TAXON	凝集型の階層的分類法 (UPGMA, WPGMA, Complete linkage, Single linkage, WPGMS, UPGMC, WPGMC, 可変法)
データ入力・加工	FILES	NT-SYS ファイルの定義
	INPUT	データ行列の入力
	ALGEBRA	行列演算機能（行列の加減，積，転置，逆行列，行列式）
	SIZOUT	平均距離行列から不要の特性の削除
	STAND	行列の行または列の標準化
	SIMINTV	区間尺度データに対する各種類似度・非類似度の算出（相関係数，平均距離，ユークリッド距離，マンハッタン距離等）
	SIMQUAL	質的データに対する各種関連係数の算出（一致係数，Jaccard 係数，ファイ係数など16通り）
	TRANSFORM	関数変換（絶対値，対数，Fisher 変換など），乱数生成（正規乱数，一様乱数）
	WILKSLAMBDA	Wilks の Λ 統計量の算出（およびそれを用いた検定）
	COPHENETIC	Cophenetic value の算出（入力した類似度または非類似度行列と分類結果得られる関連行列との間の要素間の相関係数）
出力表示機能	MXCOMP	行列間の比較 (Cophenetic correlation の算出，2つの行列の要素についての散布図)
	MXPLOT	行列の行（または列）同士のすべての組合わせについて要素間の対のすべてを散布図としてプロットする．
	NETWORK	BKGRAPH, MSTSNGL で求めたグラフ情報の出力
	PHENOGRAM	フェノグラム（デンドログラム）の出力
	STEREO	3次元散布図（立体図）の出力
	OUTPUT	出力制御用の命令語（プリント，カードパンチその他情報の出力をすべてこの命令語で指定する）
	PROJECTION	主成分得点のプロット
その他	ROTATE	行列の各種回転法（バリマックス回転，斜交回転）
	ROTFIT	行列の整合（プロクラスタス法，Gower の方法による）

し(③)，この行列を用いて国名間のユークリッド距離を算出して結果の距離行列を‘A’と名づける(④)．‘A’の内容を出力して(⑤)，‘A’について Single linkage 法を適用し MST (Minimal Spanning Tree) の情報を‘B’，デンドログラム作成用の連結情報を‘C’に格納する(⑥)．‘B’の内容を印刷し(⑦)，‘C’をデンドログラムに表示する(⑧)．第1のデータブロックに対する課題処理の終了を指示し(⑨)，最後に NTSYS ジョブ終了の指示を行う(⑩)．図4はこの例の出力の一部である．

この例にみるように NTSYS コマンドは一種のシンタックス機能を持っておりデータの受け渡しがすべて付与した名称で行われるという特色がある．反面，命令語の解読処理，データのファイル管理などのプログラムルーチンの内部が繁雑となり移植作業に手間どるということがある．

開発者の Rohlf は数値的系統分類学の祖ともいえる Sokal の流れを汲む研究者であるためプログラムの内容もこれを反映したものとなっている．表2にみるように階層的分類手法が主力であり，図的表示として彼らが好んで用いる方法が MXCOMP, MXPLOT, NETWORK, STEREO などの課題名で登録されている．また，多次元尺度解

図 4　NT-SYS の出力例



析（MDSCAL, 主座標分析），主成分分析，因子分析等の関連手法が多い点が CLUSTAN などと異なる．その他行列演算，変数変換機能など簡単なデータ加工ルーチンのあることも特徴である．図4はこの例の出力の一部である．

3.3 MINTS-80

MINTS (MIni-Numerical Taxonomy System) は筆者が独自に開発したソフトウェアで，1980年に初めて公開されたが，それより数年前に開発に着手している[39]．課題数，データ加工・変容機能ともに CLUSTAN, NT-SYS に及ぶものではないが命令語の指定がほとんど自由書式でありバッチモードに限らず端末モードでも利用できるという特色がある．MINTS では分析処理内容の指示は MINTS 命令文によりすべて行う．命令文は処理内容を指示する命令語とその内容明細を与える明細指示語が必要である．命令文には大別して，1) MINTS ジョブ制御用，2)分析課題名（手法名）指定用，3)その他の3種類がある．1)は MINTS ジョブの実行に必要なデータの入力指示と作業ファイル上へのデータの格納，検証，課題ごとのパラメータ指定，オプション指定，ジョブ終了指示などを行うための命令文である．2)は，各手法の課題名を指定する命令文である．3) は，1), 2) に含まれない命令文で，タイトルカードと入力書式指定カードの2つがある．

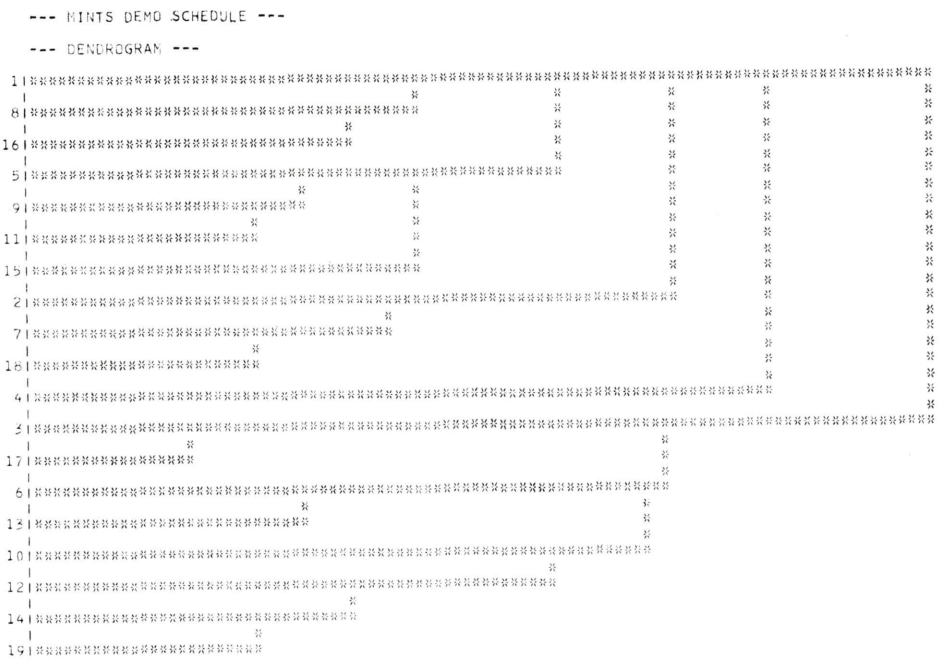
前と同じデータセットに対して類似の分析を，MINTS を用いて行う場合の例が図5である．

```
①  READ DATA CARD
    データセット（図1と同じもの）
②  HIERARCHICAL
    PARAMETER N=19,M=14,MAXCLUS=3,METHOD=3
    MINTS DEAD SCHEDULE
    (15X,14F2.0)
    OPTION OUT=(1,2),FURTHER METHOD=(4,8)
③  STOP
```

図 5　MINTS の入力例

まず READ 文で次にカードデータが続くことを指示する．次にデータカードを置く．第1課題として階層的手法 (HIER) を指定しこのとき用いるデータの条件を PARAMETER 文で与える．ここでは，データ数19，変数の数14，最大クラスター数が3，用いる手法がコード3（WPG 法）である．次に表題と，この課題で用いるデータ入力書式を与える．さらに OPTION 文で出力条件とさらに他の手法も用いることを指示する（(1, 2) および (4, 8)）．最後にジョブ終了を STOP 文で与える．①で定義したデータについて別の課題処理を行う場合は，②に続いて課題名，PARAME-

図 6　MINTS の出力例



TER 文，表題，入力書式，必要があれば OPTION 文と続ければよい．別のデータセットの分析を行う場合は再び READ 文（あるいは INPUT 文）を置きデータを入力の上，前と同じ手順で命令文を反復すればよい．この実行例の出力の一部を図 6 にあげた．

MINTS は命令文として指定した文字や数値のうち，必要なものだけを拾い出して辞書と照合するという解読ルーチンを持つので自由に文章を与えてよいという利点がある．仮に命令文の指定に誤まりがあれば再入力の指示メッセージを出力する．端末モードの場合には正しい命令文の入力があるまで再入力指示をディスプレイに表示する．

表 3 は機能の要約であるがこれにみるように各命令文の下線部だけがキーとして意味を持つので必要に応じて命令文を省略してもよい．たとえば図 5 の例は次の①のように表わしても同じ機能を持つ．もちろん，逆に②のように間に自由に文章を挿入してもかまわない．

＜省　略　例＞

```
READ
(data set)
HIER
PARA     19  14     3  3
*** MINTS DEMO ***
(15X, 14F2.0)
        OPTION 4     8  1     2
NEXT PROCEDURE IS KMEANS
  METHOD
PARAMETER CASES=19 AND
  ITEMS=6, MINCLUS=2
*** MINTS DEMO BY KMEANS ***
(15X, 2F2.0, 4X, 4F2.0)
MINTS JOB STOP HERE
```

①

②

CLUSTAN, NT-SYS が手法，結果の出力，ファイル作成，……と機能を分化させる方針をとっているのに対して，MINTS では手法課題集約型となっており命令文の書式は統一されている．デンドログラムの出力，類似度算出，クラスター情報の出力指定などは各課題のオプションとして組み入れてある．これは，本来個別的に作成したプログラムを解読部を含む主プログラムの下に MINTS として再集約したという設計上の理由に依拠している．プログラムとしては冗長的なきらいがあるが，別の手法を追加したり再編集を行う場合には具合がよいという利点もある．MINTS の

表 3　MINTS 命令文と課題の一覧表（1982年版）

種類		指定方法	
ジョブ制御用	データの入力	READ	カードデータの入力
		INPUT␣TAPE〔*nn*〕	磁気テープデータの入力
		INPUT␣FILE〔*nn*〕	磁気ディスクデータの入力
		INPUT␣SCORE	主成分得点データの入力
	分類結果の出力	OUTPUT	メンバーシップファイルの出力
		OUTPUT␣GENERAL	
	検証用	CHECK␣FORM＝{RAW, SCORE, SIMILARITY, DISTANCE, RELATION}　TYPE＝{CATEGORY, BINARY, QUANTITATIVE, MIXED}	データ表の型と形式の検証
	処理条件指定	PARAMETER　$N=n$, $M=m$, k, l, ……,〔キーワード〕 （＊）　ここで，n＝サンプル数，m＝次元数　である．k, l 以下はその他のパラメータ．	データの属性指定
		〔OPTION　$n_1=m_1$, $n_2=m_2$, ……,〔キーワード〕〕 （＊）　ここで，n_1, n_2……は機能キー，m_1, m_2……は条件キーである．	処理条件の指定
		STOP	ジョブの終了指示
分析課題用		ASSOCIATION ANALYSIS	Association Analysis
		CLUSTERING FOR CATEGORICAL DATA	質的データの階層分類
		FUZZY CLUSTERING	ファジイ・クラスタリング
		HIERARCHICAL ANALYSIS	量的データの階層分類
		ISODATA	ISODATA 法
		KMEANS METHOD	*k*-means 法
		MAID (MULTIVARIATE-AUTOMATIC INTERACTION DETECTOR)	多変量型 AID 法
		PCA (PRINCIPAL COMPONENTS ANALYSIS)	主成分分析法
その他		タイトル・カード	80文字以内で指定
		入力データの書式指定	80欄内で指定

設計方針としてなるべく利用者にとって使い易いことを重視したが，同時にクラスター化後の種々の情報を集計表示する機能を取り入れたことも特徴である．こうした MINTS の特色を次に挙げておこう．

（1）クラスターの諸統計量のきめ細かい出力
（2）種々のクラスター評価基準の算出
（3）乱数を利用したシミュレーション機能
（4）初期条件を変えた分割の多数反復機能
（5）データのランダム抽出，ランダム分割機能
（6）分類データと原データとのクロス表作成
（7）分類パターンの集計機能
（8）入力データと分類でえたメンバーシップデータの合併ファイルの作成と他の解析プログラムへの受け渡し
（9）分類結果の課題間相互利用（HIER の結果を初期値として KMEANS に引き渡すなど）
(10) データの検証機能（データの型・形式と手法との適合性を診断する）

こうした諸機能が必要とされる理由は実は分類手法が抱えるさまざまの問題に対する手当てのためである，といえよう（これについては，28), 29)を参照）．またこれらの機能の大部分は利用者の側からの要請に応じてつけ加えたものであることを強調しておこう．この点では CLUSTAN, NTSYS に引けをとらぬ実用優先のソフトウェアであろうかと自負している．

表 4　CLASS の概略（1978年版）

	プログラム名称	機　能　の　概　要
解析課題	ANCOMP	主成分分析法
	ANCORR	関連分析法（Analyse Factorielle des Correspondances, いわゆる AFC）
	POUBEL STEAK	関連分析法（追補データ処理機能あり）
	CAHPRE	Hubert の方法（階層的分類法の一つ）
	CAH2CO	ユークリッド距離にもとづく，階層的分類法 （数量化得点または主成分得点による分類，分割表の χ^2 量による分類，平方ユークリッド距離による分類）
	CAH2IN	分割表データのエントロピーによる階層的分類法
	SKELET	階層的分類法の一種（MST とそのデンドログラム表示）
データ加工・変容	BURTJJ	Burt 表（多重二元分割表）の計算
	DEDOUB	重複化処理（重複化コードの生成）
	DISJON	データ表の二値化変換（complete-disjunctive form, いわゆるアイテム・カテゴリー化の処理）
	DISLOG	二値データ行列から各種の類似度・非類似度の算出 （Jaccard, Sokal-Sneath, Russel-Rao, Benzécri の χ^2 などあわせて12種の指標）
	HISTOG	ヒストグラムの出力
	RECODE RECOD2	データのコード変換ルーチン
	REDUIT	データの変換（標準化，成分ベクトルの固有値の平方根による標準化など）
	REORDR	行列の行と列のならべかえ（行和，列和の大きさ順にそれぞれならべかえる）
分類結果の各種評価プログラム	CNCOMP	階層分類で求めたクラスターの評価指標（相対寄与，絶対寄与）
	CNCORR	データ表のプロフィル（行和または列和で基準化した行または列ベクトルの分布，数量化得点座標系内でのクラスター間距離など）
	CNINFO	エントロピー基準によるクラスターの評価
	COMPQU	追補データ処理機能を持つ階層的分類法
	COMMUT	AFC で求めた成分軸と階層クラスターとの比較分析
	HISCAH	クラスター別のヒストグラム，諸統計量（平均，分散，相関行列など）の算出
	IMPCAH	分類結果の表示（デンドログラム他）
	PCOORD	数量化得点を用いた行，列のソート（Guttman 効果を見易くするための行，列についてのならべかえ）
	PEPCAH	AFC の結果の数量化得点の散布図にクラスターを楕円表示する，など
	SIMCA1	階層分類の結合水準と AFC の固有値の関係を用いてクラスター化の程度を調べる．シミュレーション機能も含む．

3.4 CLASS

フランスの Jambu が開発した自動分類法のプログラム集合に CLASS がある．これは多数の手法を個別プログラムとして集約したものである（したがってシステムとしての名称が無いので仮に CLASS と名づけた）．各プログラムは完全な固定書式指定をとっており手法ごとに決められたパラメータを指定通りに与える，いわば普通のプログラムである．むしろ前述の3種のソフトウェアに比して際立った差異はその内容にある．それを一言でいえば，これこそフランス流のデータ解析（Analyse des Données）そのものである，ということである．フランスではデータ解析の分野で独自の発展をみせていること，とくに J. P. Benzécri を中心とするグループの活動には特筆すべきものがあるということはすでに本誌 204 号（1980）で紹介したのであるが，CLASS はまさしく Benzécri 流のデータ解析向きのプログラム集合といってよい．表4にみるように，すべての手法が Benzécri の関連分析法（Analyse Factorielle des Correspondances—AFC—，数量化Ⅲ類と同じ方法）を中心に把えられている．この点ですでに他のソフトウェアと異質といわねばならない．たとえば，入力データ表の基本を分割表におく，ということがある．また，プロフィル，追補データ処理，Burt 表，モダリティー（カテゴリーのこ

表 5　プログラムの構成

名称	CLUSTAN	NT-SYS	MINTS	CLASS
版	1978年	1977年	1982年	1978年
使用言語	FORTRAN	FORTRAN	FORTRAN	FORTRAN
総ステップ数（含注釈文）	約 13440 (13437)	約 28400 (28387)	約 10000 (10100)	約 14800 (14823)
プログラムの構成	主プログラム　(1) 副プログラム　(130) ブロックデータ等(62)	主プログラム　(1) 副プログラム　(459) ブロックデータ等(70)	主プログラム　(1) 副プログラム　(124) ブロックデータ等(10)	個別プログラムの集合 プログラム数　(27)
その他	オーバーレイ構造可 XYプロッター使用可 端末モード用新版あり	オーバーレイ構造可	オーバーレイ構造可 端末モード利用可	——

と），雲（数量得点の分布のこと），重複化処理，二値化処理 (complete disjunctive form)，……といった用語が盛んに現われるが，これは彼等の特有の符牒であり，これらを十分理解して初めてプログラムの中味が分るというしくみである[3),15),33)]．

さらに，クラスター化の結果の評価分析を行う各種プログラムが豊富なことも注目してよい．もちろんその大部分が，関連分析法と密接に関連していることはいうまでもないが，数量化法を自動分類法とはっきり関連づけて解析的に体系化している点には大いに見習うものがある．とにかく実際のデータ解析に利用してその効用の程を試したいという食指を動かされるソフトウェアである．

4.　今後の動向

現存する自動分類法のソフトウェアの中からごく少数を眺めてきた．このうち CLUSTAN, CLASS については日本科学技術研修所が開発者の了解のもとに計算サービスを行っている[37)]．また NT-SYS は，農業技術研究所の遺伝資源情報システム (GRIMS) に含まれる情報検索システム EXIR の中のデータ解析用のルーチンとして利用されている[31),38)]．MINTS も，大学の計算センターやマーケティングデータの分析用としてすでに稼働している．ここで参考までに 4 種のソフトウェアの構成を表 5 としてまとめておいた．

こうしたシステムはたえず改編を繰り返しているので必らずしもそれらの現状や内容を正確に伝えることができたとは言い難い．またどれが良いと決めることも難かしい．処理すべきデータに対してどれが適切かを選ぶのは利用者であり，しかもそれは扱うデータに依拠するところが大であるからである．

また別の問題としてシステム化の度合の目標をどの程度に設定するか，ということがある．これはどんな機能までそのソフトウェアに負担をかけるかということであり，汎用化にも関連することである．しかし，自動分類法の場合，その関連分野の多様性から汎用的の意味を把え難いということがあり，OS が進んだ現状にあっては，むしろ個別的に独創的なプログラムを多数用意して利用することが得策ということも考えられる．また実際のデータ解析で経験するところであるが，自動分類法だけを用いてもその効果が十分に期待できないことが多い．分類結果は他の要因との対比分析を通じて初めてその効用が認識されるといってよい．このためには，他の汎用統計ソフトウェアとの連結利用を切り離して考えるわけにはいかない．

さらに大きな問題として計算結果の妥当性あるいはソフトウェアの頑健性ということがある．すでに指摘のように現状では用語や概念にかなりの混乱があるが，これに加えて算法や計算手順の若干の相違から同じ手法を用いても利用したソフトウェアによって結果が異なるという現象が起こる．一般の利用者にとってはプログラムの中味はブラックボックスであるからこうした現象は疑問や誤解を招く．また一知半解のまま無節操にデータ解析を進めるのも誤用の原因となる．これらを防ぐ最小限の手当てとしてソフトウェアに関連した情報の文書の充実化が不可欠である．この点，

CLUSTAN, NT-SYS は利用の手引，システムの解説書，例題など比較的整っているといってよいが全体的には今一歩の感がある．多くの汎用統計ソフトウェアが端末利用に移行する風潮がある中で自動分類法のソフトウェアが今後どういう道程をとるかきわめて興味がある．

文　　献

(A) **Books**

1) Anderberg, M. R.: Cluster Analysis for Applications, Academic Press. 1973.
2) Arthanari, T. S. and Dodge, Y. (1981): Mathematical Programming in Statistics Chapter 7-Cluster Analysis, John Wiley.
3) Benzécri, J. (1973): L'Analyse des Données, Tome I (La Taxinomie), Dunod.
4) Bezdeck, J. C. (1981): Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press.
5) Blackith, R. E. and Reyment, R. A. (1971): Multivariate Morphometrics. Academic Press.
6) Blashfield, R. K. (1976): A Consumer Report on the Versatility and User Manuals of Cluster Analysis Software, Proc. of the Statistical Computing Section. ASA, 31-37.
7) Diday, E. and others (1979): Optimisation en Classification Automatique, Tome 1, Tome 2, INRIA.
8) Dunn, G. and Everitt, B. S. (1982): An Introduction to Mathematical Taxonomy, Cambridge studies in mathematical biology : 5, Cambridge University Press.
9) Eckes, T. and Rossbach, H. (1980): Clusteranalysen, Verlag W. Kohlhammer.
10) Enslein, K., Ralston, A. and Wilf, H. S. (eds.) (1977): Statistical Methods for Digital Computers, Vol. III of Mathematical Methods for Digital Computer, John Wiley.
11) Francis, I. (1981): Statistical software-A Comparative Review-, North Holland.
12) Gordon, A. D. (1981): Classification Methods for the Exploratory analysis of Multivariate Data, Chapman and Hall.
13) Hartigan, J. A. (1975): Clustering Algorithm, John Wiley.
14) Illig, A. (1980): Konkurrenzanalyse mit Hilfe Multivariater Klassifikation, (Reihe Wirtsceaftwissenschaften Bd. 164), Verlag Harri Deutsch.
15) Jambu, M. (1978): Classification Automatique pour L'Analyse des Données. Tome 1-Méthodes et algorithmes, Tome 2-Logiciels, Dunod.
16) Jardine, J. and Sibson, R. (1971): Mathematical Taxonomy, John Wiley.
17) Lerman, I. C. (1970): Les Bases de la Classification Automatique, Gauthier-Villas.
18) Mardia and others (1979): Multivariate Analysis, Academic Press.
19) Mather, P. M. (1976): Computational Methods of Multivariate Analysis in Physical Geography. John Wiley.
20) Opitz, O. (1978): Numerishe Taxonomie in Marktforshung, Verlag F. Vahlen.
21) Späth, H. (1980): Cluster Analysis Algorithms for Data Reduction and Classification of Objects, Ellis Horwood Limited. (次の 22) の翻訳版)
22) Späth, H. (1975):Cluster-Analyse-Algorithmen, R. Oldenbourg Verlag.
23) Steinhausen, D. and Langer, K. (1977): Clusteranalyse, Walter de Gruyter.

(B) **Articles**

24) Pape, U. (1980): Algorithm 562 Shortest path lengths [H], *ACM Transactions on Mathematical Software*, **6**, 3, 450-455.
25) Slagle, J. R. and others (1975): A clustering and data-reorganizing algorithm, *IEEE Transacitions on SMC*. January 1975, 125-128.
26) Sibson, R. (1973): SLINK: An optimally efficient algorithm for the single-link cluster method, *The Computer Journal*, **16**, 1, 30-34.
27) 大隅　昇 (1973)：クラスター分析—SL 法をめぐって—，数理科学，No. 181.
28) 大隅　昇 (1979)：クラスター分析はどう使われるか，数理科学，No. 190.
29) 大隅　昇 (1979)：ファジイ・クラスタリング，数理科学，No. 191.
30) 大隅　昇 (1980)：フランスにおけるデータ解析の動向，数理科学，No. 204.
31) 熊谷申子夫，鈴木茂他 (1977)：遺伝情報システム—EXIR の利用—IBM サイエンティフィック・センター資料 (N: GE 18-1857-0).
32) 矢島敬二，大隅昇：統計，bit 増刊，アプリケーション・プログラム，7月，1977.

(C) **Program manuals**

33) Lebart, L. and Morineau, A. (1982): SPAD-System Portable pour L'Analyse des Données, CESIA.
34) MacRae, D. J. (1970): MIKCA: A FORTRAN IV Iterative k-means Cluster Analysis Program, CTB/McGraw Hill, Del Monte Research Park, Monterey, California.
35) Rohlf, F. J., Kishpaugh, J. and Kirk, D. (1977): NTS-YS-User Manual.
36) Wishart, D. (1978): CLUSTAN User Manual (3rd ed.), Program Library Unit, Edinburgh University.
37) 日本科学技術研修所；CLUSTAN-1C プログラム利用の手引き，1978.
38) 農業技術研究所；遺伝資源情報管理プログラム—利用者マニュアル— (GRIMS: Gene Resources Information System)
39) 大隅昇：クラスター分析プログラム・パッケージ—MINTS 80—(利用の手引)，統計数理研究所研究リポート 51.

(D) **Technical reports**

40) Ball, G. H. and Hall, D. J. (1965): ISODATA, A Novel Method of Data Analysis and Pattern Classification, Stanford Research Institute, Technical Report.
41) Ling, R. (1971): Cluster Analysis, Yale Univ. Tech. Rep. No. 18.
42) Roach, C. D. (1971): An Optimization Approach to a Clustering Algorithm, Ph. D. Thesis, the Faculty of the Institute of Technology, Sourthern Methodist University.
43) Wolfe, J. H. (1967): NORMIX 360 Computer Program, Research Memo., SRM 72-4. Wolfe, J. H.: NORMIX: Computational Methods for Estimating the Parameters of Multivariate Normal Mixtures of Distributions. Research Memo., SRM 68-2.

(E) **Computer Contribution Series, State Geological Survey, The University of Kansas, Lawrence.**

44) Bartcher R. L. (1966): No. 6 FORTRAN IV Program or Estimation of Cladistic Relationships Using the IBM-

7040.
45) Wishart, D. (1969): No. 38 FORTRAN II Programs for 8 Methods of Cluster Analysis (CLUSTAN I).
46) Parks J. M. (1970): No. 46 FORTRAN IV Program for Q-mode Cluster Analysis on Distance Function with Printed Dendrogram.
47) McCammon, R. B. and Wenninger, G. (1970): No. 48 The dendrograph.

(F) **Statistical Algorithms in Applied Statistics**

48) Ross, G. J. S. (1969): AS-13 Minimum Spanning Tree, **18**, 103-104.
49) Ross, G. J. S. (1969): AS-14 Printing the Minimum Spanning Tree, **18**, 105.
50) Ross, G. J. S. (1969): AS-18 Single Linkage Cluster Analysis, **18**, 106.
51) Roger, J. H. (1971): AS-40 Updating a Minimum Spannimg Tree, **20**, 204-206.
52) Sparks, D. N. (1973): AS-58 Euclidean Cluster Analysis, **22**, 126-130.
53) Banfield, C. F. (1976): AS-102 Ultrametric Distances for a Single Linkage Dendrogram, **25**, 313-315.
54) Banfield, C. F. and Bassill, L. C. (1977): AS-113 A Transfer Algorithm for Non-hierarchical Classification, **26**, 206-210.
55) Hartigan, J. A. (1979): AS-136 A k-means Clustering Algorithm, **28**, 100-108.
56) Oehlert, G. W. (1979): AS-140 Clustering the Nodes of a Directed Graph, **28**, 206-214.

(G) **Algorithms Supplement in the Computer Journal**

57) van Rijsbergen, C. J. (1970): Algorithm 47- A Clustering Algorithm, **13**, 113-115.
58) van Rijsbergen, C. J. (1970): Algorithm 52- A Fast Hierarchic Clustering Algorithm, **13**, 324-326.
59) Sale, A. H. J. (1971): Algorithm 65- An Improved Clustering Algorithm, **14**, 104-106.
60) Rohlf, F. J. (1974): Algorithm 81- Dendrogram Plot, **17**, 89-91.

(H) **Communications of the ACM**

61) Whitney, V. K. M. (1972): Algorithm 422- Minimal Spanning Tree, **15**, 4, 273-274.
62) Page, R. L. (1974): Algorithm 479- A Minimal Spanning Tree Clustering Method, **17**, 6, 321-323.

（おおすみ・のぼる，統計数理研究所）