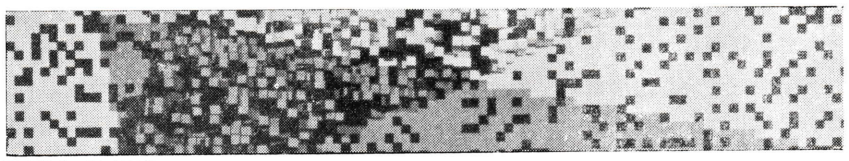
データ解析における統計ソフトウェアの役割

大隅　昇
垂水共之

『数理科学』No. 245, NOVEMBER 1983 別刷

データ解析における統計ソフトウェアの役割

大隅　昇／垂水　共之



R. Winiarski (1968)

データ解析と統計ソフトウェアの接触面

近年，統計的データ解析と計算機科学の接触面でみられる諸問題を論ずることが盛んになってきた．一昔前にくらべると出版される著作物や会議録などが豊富になり諸外国の動向を知ることが容易になっている．たとえば比較的規模の大きい研究集会だけでも次のものがある．

- COMPSTAT—Computational Statistics—（1974年から2年ごとに開催）
- Annual Symposium on the Interface—Computer Science and Statistics—（1967年から年1回開催）
- Statistical Computing Section—American Statistical Association—（米国統計協会の年次会の一部会）
- International Symposium on Data Analysis and Informatics（仏国の INRIA が主催，1977年，1979年そして本年10月に第3回大会開催）

COMPSTAT は ISI (International Statistical Institute) の活動部門として組織されたもので最近は国内からも多くの参加者がある．また上記のいずれの会議も部厚い会議録を出していて大いに役に立つ．こうした定期的に開催される研究集会の他に，不定期のものや計算機科学により近い分野（たとえば国際電気通信学会，パターン認識学会など）で計算機統計学に関連した多数の会議が開かれている．

また今年になって North-Holland 社から，“Computational Statistics and Data Analysis” (CSDA) が創刊された．S. P. Azen（南カリフォルニア大学）を編集主幹とし，各国のデータ解析に関心の高い研究者を編集委員とした定期刊行誌である．日本からは，小柳義夫（筑波大学），丹後俊郎（東京都臨床医学総合研究所）の両氏と筆者が編集委員として参画している．時代の流行に迎合したものという見方もあろうが，やはり時の流れのなせるところであろう．

Azen の編集方針によると，CSDA の内容は三つの部分から構成されている．

（1） 方法論 (methodology)——数値計算，アルゴリズム，データ探査，パターン認識，分類問題，シミュレーション，欠測値・異常値処理の問題など．

（2） 応用，比較・対比分析 (Applications and Comparative Studies)——統計パッケージやアルゴリズムの比較，データ解析に役立つ方略の提案，それらの多分野への応用例など．

（3） ノート，アナウンス，レビュー (Notes, Announcements and Reviews)——関連研究集会の案内，ソフトウェアの紹介など．

こうした分野の研究は急速に進展するのでこれらの話題を早く広く提供し，多様な学際分野への応用普及を図ることが急務である，としている．(1)，(3)はとにかく，(2)のデータ解析の方略，比較分析の提案といった部分は，従来とかく研究の対象として軽視されてきた部分であり，これに積極的に取り組もうという姿勢に注目したい．

フランスでもデータ解析に対する研究者の関心が高く上記の INRIA 主催の研究集会は欧州圏の研究者の参加を得てなかなか盛況である．フランスにおけるデータ解析の動向については本誌 204 号で紹介したのであるが，中でも J. P. Benzécri

を中心とする一派は精力的に広い分野で活動を展開している．彼等の成果は“Les Cahiers des l' Analyse des Données”(Benzécri を主幹とし 1976 年から Dunod 社より発刊) などにみられる．

こうした欧米諸国のやり方に共通にみられることは，現象の体験や実験を通してデータと統計解析手法の接触面で要求されるデータ処理過程の手続き（方略）を重視していることである．手にしたデータが語るものは何か，そこに何がみえるか，これに対して可能な範囲でモデル（理論）を想定し何が適切であるのかを探る方略を考えること，つまり Tukey のいう EDA (Exploratory Data Analysis) に通じるものである[13),18)]．しかしデータに内在する情報を最大限に要約し効率的に読みとるという点では一致していても，分析者の置かれた立場や目的により分析の筋道は一つではあり得ない．

ところでデータ処理には大きく二つの側面がある．一つは扱う“データの性格”である．他はそこで要請される“解析手法の密度”である．前者はデータの規模（量），種類（量的か質的か），形式（データ表の種類）などをいう．後者は，解析手法あるいはモデル化が単純か複雑か，緻密かあるいは粗な分析で済むか，といったことである．こうした多様な処理過程を要約すると次のようになろう．

（1） 実験，調査等の計画（問題の設定）
（2） 実験，調査等の実施
（3） データの測定，収集，整理，点検など
（4） データの編集・加工
（5） 集計，解析，モデル化
（6） 分析結果の表示
（7） 結果の解釈，評価および検証

これらの諸過程が反復，分岐を繰り返しながら即応的に進められることはいうまでもない．（1）～（3）は人的要因や調査・実験の環境要因が大きく影響する部分であり，（3）～（6）は計算機やソフトウェアの支援が期待される部分である．（5）～（7）ではとくに統計的知識やモデル化の技術が強く必要とされる．いわゆる EDA の重点は上の（5）～（7）にあると思うが，現実に統計データ処理に費やす労力の大半はそれ以前の過程にあるということも忘れてはならない．最近はデータの測定，収集等の調査環境が悪化しあらゆる分野で問題となっている．とくにデータ測定法やサンプリング理論など根底から見直す時期にあるとさえいわれている．制約された環境下にあっていかに効率よく情報を集めるかという方法論の検討もデータ解析の重要な側面であるが，この種の実務面での問題に対して，実務家の強い関心とは逆に統計研究者がいま一歩積極的とはいえず，多くの問題を残したままである．

経験科学的な見地から，これらの諸過程をどう結びつけて運用するか，という“方略の設計と選択”を図ることがデータ解析の本質といえるが，実施にあたって求められる，多様にしてしかも高度で複雑な統計データ処理を効果的に進めるには，電子計算機の利用を切り離して考えることは出来ない．しかも特定な計算プログラムがあれば済む，あるいはその都度必要なプログラムを作成する，ということが次第に困難になってきたことも事実である．統計ソフトウェアが強い期待を担って登場する理由がここにあり，SPSS, SAS, BMDP などの現在の隆盛にこれがみられるわけである．

周知のように SPSS は各地の大学大型計算機センターに提供され人文社会科学の分野の分析に適した内容といわれながら，その簡便さと便利さから広い分野の研究者の支持を得てきた．最近は大規模統計システムの代表格である SAS (Statistical Analysis System) の国内市場参加と一部の研究者，実務家の熱心な支持により，SAS か SPSS か，さらに BMDP かといった統計ソフトウェア比較論が盛んである．また国産計算機メーカーの自社機種への移植改編作業や，ユーザー会，統計パッケージ研究会などの活動が盛んになっている．三宅一郎氏（同志社大学）を中心とする少人数のグループが SPSS の移植改編作業に心労を費していた十年前のことを考えると大きな時代の変化を感じないわけにはいかない．

さて欧米諸国における統計ソフトウェアの目ざましい発展にくらべ国内の事情はどうであろうか．統計ソフトウェアの利用者は確実に増えてはいるが，こと開発，製作のように研究の支援的要

素の強い仕事に対して高い評価が与えられないという，欧米との意識風土の差異はいまだに埋められてはいない．もちろん一部の研究者や実務家の努力には注目すべきものがありその成果として，SPMS, SALS, ETPS, NISAN, TIMSAC, EASY, MINTS, MINERVA, MAP, CDA などのソフトウェアが生まれている．しかし利用言語の差異，機種依存度，プログラムの完成度，マニュアルの充実度など，多くの面で問題があり，必らずしも広い支持を得るまでに至っていない．しかも依然として輸入品志向が強いことを感じないわけにはいかない．

ところでデータ解析の道具としての統計ソフトウェアの特徴を語る場合，利用面の問題とアーキテクチュアのあり方の二つの側面から考えることができよう．これらについて，とくに前者について，水野，矢島と筆者は，統計ソフトウェアの現況と問題点として既に意見を述べたのであるが，現在でもこの事情にさしたる変化はない[25),26)]．そこで指摘した統計ソフトウェアのもたらした効用と弊害を改めて列挙してみると次のようになる．

- 何よりも便利であって省力化の有力な道具であること．
- プログラミング能力や統計手法の知識の少ない初心者にも統計データ解析を身近なものにしたこと，そして利用人口の底辺拡大に寄与したこと．
- いわゆる高度で複雑な統計処理を大衆化したこと．また変化に富む多数の統計手法の提供が受けられることを可能にしたこと．
- 既に分析能力や豊富な経験を持つ実務家にまでデータ解析の効用を知らしめ，とくに大量データの多種多様な分析に伴う労力の軽減にはたす役割を強く印象づけたこと．
- 計算機やソフトウェアのメーカーにシステム化の効用を知らしめ商品になると認識させたこと．
- しかし依然として20年近くも欧米にくらべて立ち遅れていること．
- 開発に労力とコストがかかる，従ってかなりの犠牲を強いられるが，研究としても商品としても高い評価が得られにくいこと．
- システムの巨大化に伴い中味がブラックボックスと化し，得られた結果の処理の過程が正しく理解されない，あるいは知りたくても分からない，というおそれがあること．
- 簡単に結果が得られることから，不必要に解析手法を乱用し分析の目標を見失うおそれがあること．
- ソフトウェアを用いることがデータ解析であると錯覚し，過大な期待を抱くおそれがある，あるいは利用の限界を見通せない危険があること．

このようにデータ解析と統計ソフトウェアは良くも悪くも切り離せぬ関係にある．したがって，統計ソフトウェアの比較評価論もいきおい盛んになり，多数の論評や比較分析が試みられている[19),10),16),21)]．とくにアメリカ統計協会（ASA）の Statistical Computing の分科活動として1974年に始まった Francis, Heiberger らの活動と調査に代表されるであろう．この調査の成果の一つが Francis により成書としてまとめられた[9)]．調査は代表的な200のソフトウェアの開発者と利用者にアンケート調査を実施し有効回答のあった117のソフトウェアについて評価や分類を試みたものである．Francis の報告はそれなりに一つの視点を与えるものであるが，ここで改めて，データ解析を望む者にとってその要請に見合った適切なソフトウェアを選ぶための手掛りを得るという観点から彼のデータを再検討しようというわけである．

統計ソフトウェアの分類

Francis が調査にあたって用意した質問項目を表1に挙げてあるが，ソフトウェアの具備すべき技術的，機能的要件はほぼ満たしているといえよう．さて Francis は117のソフトウェアを20の質問項目を用いて11のグループに分類しているが，ここではその中から代表的な33のシステムを選びこれを数量化（パターン分類）により検討した（表2）．33に限定した理由は，まず一次分析で117のパターン分類の結果を要約しこの結果を見て主に調査分析用，統計一般解析用，代表的サブルーチン・ライブラリーのうち比較的知名度が高く筆者の利用経験がある，あるいはマニュアルが

表 1　統計ソフトウェア調査の調査項目

区　分	調　査　項　目
Generality	Subroutines or Package, Generality of Purpose
Filing	**Filing Summary, Data Set Size, Flexible Data Input, *Complex Structures***, Filing Missing Data, **Storage/Retrieval, *Filing Manipulation*, Flexible Output**
Editing	**Editing Summary, Editing Language, *Consistency Checks*, *Probability Checks***, Error Handling
Missing Data	**Missing Data Summary**, Imputation
Display	**Display Summary, *Compute Tables*, *Print Tables*, Graphics**
Extensibility	Language Math. Power, Code Modifiability
Statistical Analysis	Stat. Analysis Summary, ***Multiple Regression*, *Anova/Linear Models*, *Linear Multivariate*, *Multi-Way Tables***, Other Multivariate, ***Time Series*, *Non-Parametric*, *Exploratory*, *Robust*, *Non-linear*, Bayesian, *Econometric***
Sample Survey	***Sample Survey Summary***, Compute Estimates, ***Compute Variances***, Select Sample
Other Mathematics	Simulation, ***Math. Functions*, *Operations Research***
Portability	**Availability, Installations, Computer Makes, Mini Version, Core Requirements, Batch/Interactive**
Ease of learning and using	Stat. Training, Computer Training, **Language Simplicity Documentation**, User Convenience
Reliability and Maintenance	**Maintenance, Tosted for Accuracy**

表 2　分析に用いた統計ソフトウェア

区　分	ソフトウェアの名称
Data Management	SIR
Editing	CONCOR
Tabulation only	COCENTS
Tabulation and Arithmetic	TPL, RGSP
Survey Variance Estimation	CLUSTERS
Survey Analysis	BTFS, EXPRESS, EASYTRIEVE, FOCUS, SURVEYOR/SURV, DATAPLOT, PACKAGE-X, SPSS, SOUPAC, DATATEXT, P-STAT 78, SPMS, OSIRIS IV
General Statistical Analysis	SAS, OMNITAB 80, HP-STATPACKS, MINITAB II, BMDP, NISAN, GENSTAT 4.02, SPEAKEASY III, TROLL, IDA
Speciphic Interactive Statistical Analysis	GLIM
Mathematical Subroutine Libraries	IMSL-LIBRARY, NAG-LIBRARY, EISPACK

手元にあるものに的を絞った．質問項目も表1から37項目を選んで用いる(表1の太字のもの)．

さて図1,2が33のシステムと37項目の布置である．図2の項目の布置をみると1軸の正の側にファイル編集・加工，表示・作表などデータ集計・調査分析に関わる項目がある．一方負の側左上に各統計手法の集まりがみられ，図の左下には，機種の互換性，移植の容易性，保守性などに関連した事項が位置している．

図1のシステムの布置をみると図2に対応して，1軸右側に作表，集計，編集，データ検証機能に優れたソフトウェアが集まっている．SIR, CONCORはデータ編集，検索，検証・自動修正機能で知られたシステムであり，OSIRIS, P-STAT, DATATEXT, SPMSなどはファイル処理，編集・加工機能を重視したソフトウェアである．図の右下のRGSP, TPL, COCENTSなどは集計・作表システムとして知られたものである．

図の中心上部には，SAS, OMNITAB, BMDP, SOUPAC, GENSTATなどいわゆる統計システムが集まっている．SPSS, SOUPACがやや集計システムの集まりに近い点が興味深い(SPSSの編集加工機能が強調されている)．

また図の左に数学・統計計算を含むサブルーチン集合として知られるIMSL, NAG, EISPACK, GLIMが位置する．(3軸まで考慮すると) CLUSTERS, NISAN, BTFSなどはやや異なる位置にある．NISANは国産のパッケージであるが，いわゆる総花的に各種の機能を収録する予定とされており，こうした位置にくる．CLUSTERSは層別サンプリングの標本誤差の計算という特殊な用途のシステムである．

分析の結果から質問項目の構造が項目数の多いわりに単純で，ファイル編集加工機能，統計解析手法，互換性・保守性など機種依存性に関する項目の三つに大別される．統計データ解析を進める

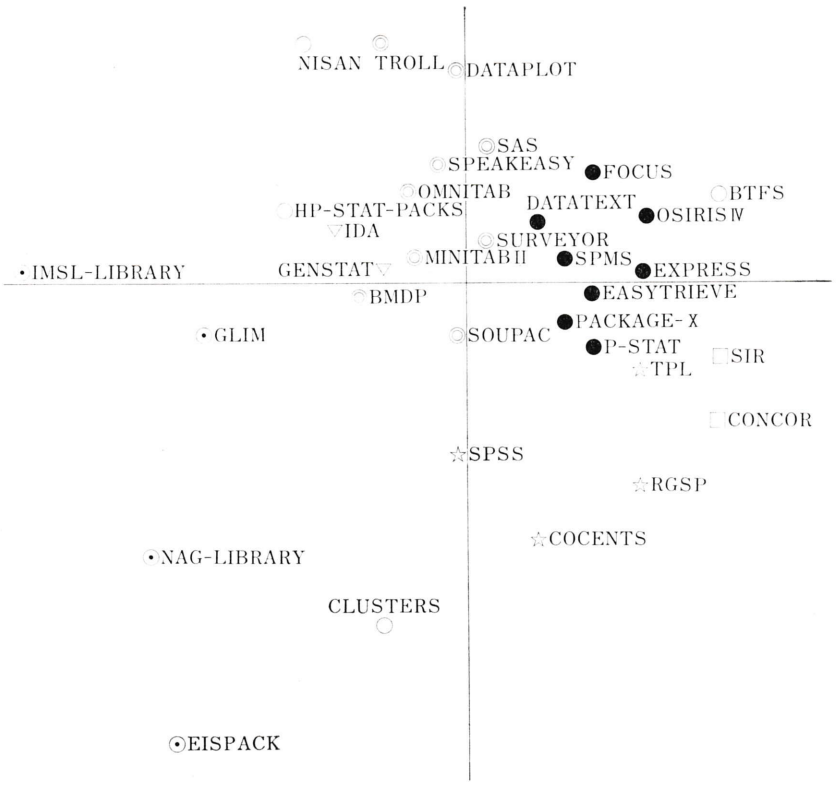


図 1
代表的な統計ソフトウェアの分類

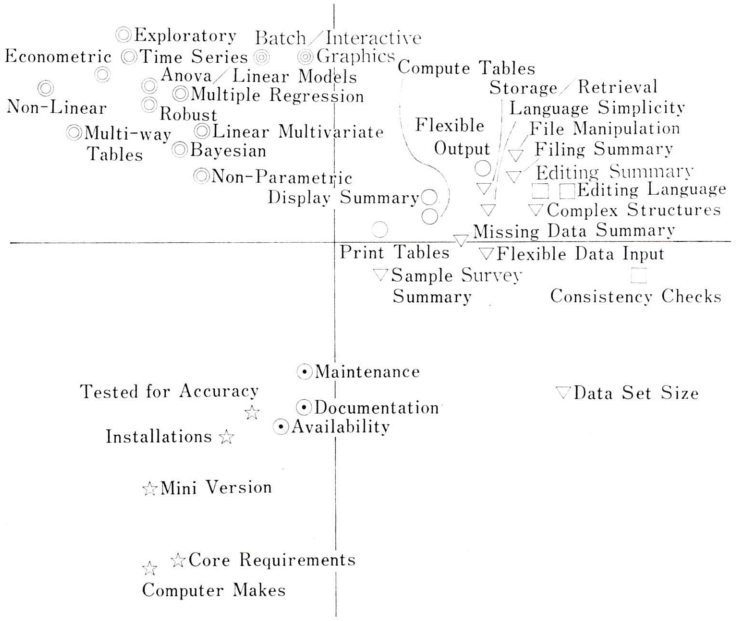


図 2
調査項目の分類

という立場から強い関心があるグループはやはり調査集計・編集分析用と一般統計解析用のシステムであろう．またそれらのシステムの開発者が解析手法に対してどのような志向を示しているかを知りたい．そこで，上の 33 の中から，集計，データ検証など特殊用途のもの (SIR, CONCOR, COCENTS, RGSP, CLUSTERS) を省いた 28 のソフトウェアについて統計解析手法の項目だけを

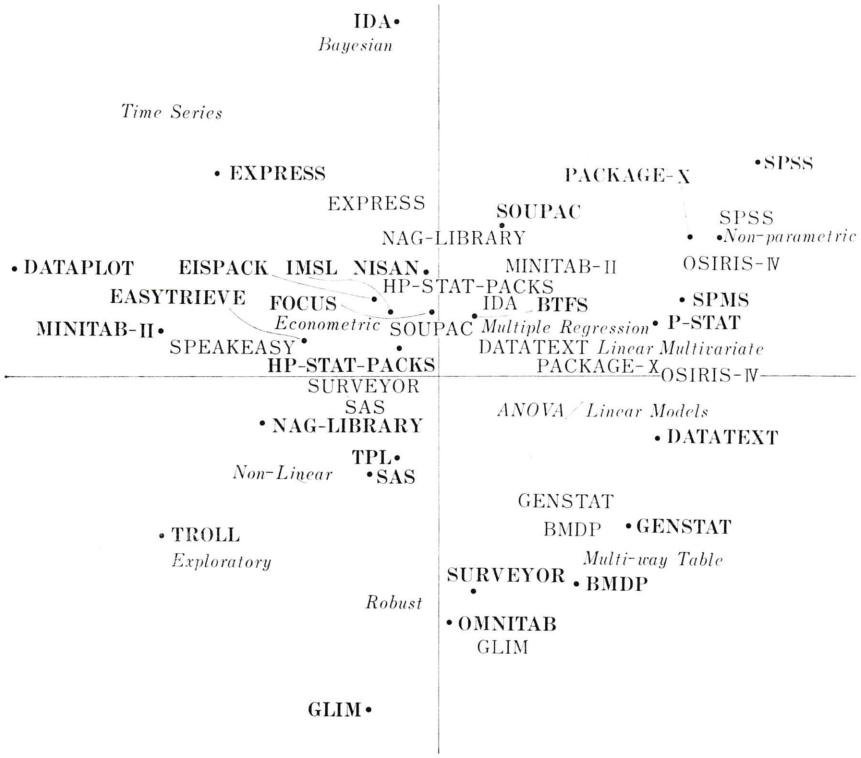


図 3
開発者と利用者の評価の比較分析／統計手法の位置づけ

用いて分析を試みる.

Francis の調査では各ソフトウェアの利用者に対しても開発者とほぼ同じ内容の調査を行っているので(標本数がやや少ないが), ここでは, 開発者の意図する統計解析手法への取り組み方と利用者の考えている評価とがどの程度一致しているかをあわせて検討することにした.

再び数量化法により求めた結果が図 3 である. 黒丸のソフトウェアが開発者の評点を用いた場合の数量化得点の布置である. 一方, 無印のものが利用者の評価点を用いたときの布置である. またイタリックで 11 の解析手法を表わしてある.

まず, 主として社会科学など調査データの分析に関わる利用者の支持を集めてきた, データの編集・加工, ファイル処理などの機能を持つとともにそれらの支援を受けて標準的な統計解析手法も同時に利用可能であるようなソフトウェアの集まりがある. SPSS, OSIRIS, P-STAT, SOUPAC, DATATEXT, PACKAGE-X などに代表されるグループがそれである. 重回帰分析, 因子分析, 分散分析などの他にカテゴリカルデータに関連した手法 (AID, MCA, 多重クロス表分析) や各種連関係数の算出, MDS, クラスター分析, ノンパラメトリック手法などを含むことが特徴である.

次にどちらかというと統計的モデル解析に重点があり新しい手法の導入にも積極的なソフトウェアの集まりがある. たとえば, BMDP, GENSTAT, SAS などの統計システムや GLIM, TPL などのように一般線形モデル解析などを対象としたパッケージなどである. これらは比較的最新のしかもかなり高度の統計知識を必要とし必らずしも初心者向きとは限らない.

最近の統計解析の傾向として, 離散データの解析法, 寿命・生存データの解析法, 時系列解析, クラスター分析, それに EDA に関する各種の表示機能, 不完全データ・欠測値処理法, グラフィクス統計などが挙げられる.

離散データの解析法としては多重クロス表の対数線形モデル解析やロジットモデル, 線形ロジスティック反応モデルなどを収録することが多くなってきた. BMDP のように, モデル選択の方法として特徴のある Brown の方法をいち早く取り

入れるなど積極的である．

寿命・生存データの解析法も BMDP, SAS を始め収録されているが，Cox 流の回帰型モデルの導入が流行のようである．また回帰分析の方法としては逐次選択法はもちろん，非線形回帰，多変量回帰，主成分回帰など多様化する傾向にあり，これにモデル選択の規準（Mallow の C_p 統計量など）を含むことが通例となってきている．さらに BMDP のように当初は取り挙げなかった時系列解析の手法を取り入れたり SAS のように SAS/ETS としてライブラリー化しているものもある．Box-Jenkins 流の方法論を収納するのが共通した傾向であろう（SAS, SPSS-X, BMDP, MINITAB-82).

EDA やグラフィクス統計手法の導入も盛んで，会話型システムにこの傾向が強い．図3の左に位置する TROLL, MINITAB, DATAPLOT, SPEAKEASY, IDA などがそうである．SAS も SAS/GRAPH として一つのライブラリーをもうけている．MINITAB は小規模ではあるが小回りのきく会話型システムであり，DATAPLOT は各種のグラフィクス解析を端末モードで処理できるという利点を持っている．

図3の左上の EXPRESS, IDA などは特殊な手法（経済分析，時系列解析など）まで含むシステムである．さらに数学モデルや非線形モデル，その他標準的な統計手法をサブルーチンとして網羅的に収録してある NAG, IMSL などのライブラリーが図の中心にあり，いわば個性の乏しいパッケージになっている．

次に利用者は各ソフトウェアをどう評価しているのかをみよう．まず，世評の高いシステムは開発者と利用者の要求がほぼ一致しているといってよい．たとえば，BMDP, GENSTAT, SAS, SPSS, GLIM などである．これに対し，システムとしては魅力があるがやや使いにくいとされている OSIRIS, DATATEXT, P-STAT, SOUPAC, PACKAGE-X などがやはり互いに離れて位置している．また，SURVEYOR, NAG も開発者，利用者の一致がみられない．MINITAB も同様の傾向があるが，これは利用者が解析手法としての充実度よりも会話型，教育用としての価値を高く買っているためであろう（実際，大規模データの処理には不向きである）．

最近の SPSS-X, MINITAB-82, SAS (82 年版), GENSTAT など新版や改編版の公開が続いている．とくに SPSS-X はグラフィクス，レポートジェネレーター，時系列解析，対数線形モデルなどを取り入れ，やはり多目的システムへ移行する現象がみられる．

とにかく Francis の調査は現況を反映したものではないのであるが，それでもこの分析でみる限り利用者は自分の専門分野の守備範囲に見合ったシステムを賢明に選んでいるといってよさそうである（といっても調査対象が主として米国であるという事情があるが）．

統計ソフトウェアの移植の問題

どんなに優れた機能をうたったソフトウェアであっても計算機への移植がすみやかに行なわれなければデータ解析の道具としては価値がない．実は大多数の利用者は既に移植ずみのシステムを用いているためにその前提にある移植の適不適の検討の重要性まで知らぬことが多い．SPSS や SAS が国内で定着しつつあるというのも，開発機種である IBM や IBM 互換機への改編・移植作業がその背景にあったからである．

Francis の調査項目にもこれに関連した事項があるが，これと文献 4) にある資料とを合わせて，22 のソフトウェアについてそれがどの計算機種で利用可能かというデータ表を作りこれを分析してみた．結果は図4のようになり次の特徴がみられる．

（1） 機種の能力(大型機か中型，ミニコンか；ユアーサイズ；OS (オペレーティングシステム) の違いなど)，

（2） 使用言語の問題

（3） 開発者の設計志向（多機種を目標とするか，特定機種に限定するか)．

図の右上にある OSIRIS-Ⅳ, CONCOR, DATATEXT, SAS などは利用言語や機種に特徴がある．CONCOR は COBOL 言語，SAS は PL/I が主要言語，また SAS, OSIRIS, DATATEXT にはアセンブラ言語が一部含まれている．しかも IBM

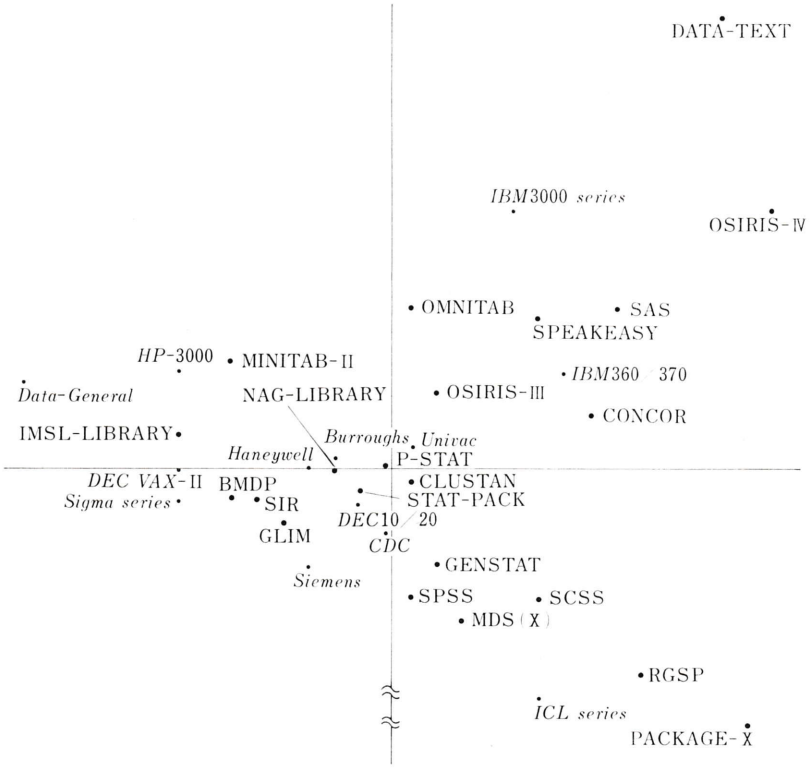


図 4
ソフトウェアと計算機種の関係

機を主計算機として開発されている．一方，右下にある PACKAGE-X, RGSP, MDS (X), GENSTAT, SPSS (ICL 版) などは，それぞれ Dataskil 社，エジンバラ大学の Program Library Unit (PLU), NAG, ロザムステッド農事試験場など英国を中心とする開発グループが英国製計算機 ICL を対象に開発したソフトウェアである．GENSTAT, MDS (X) などは標準的なフォートラン言語で書かれているが，ICL が主機種で IBM, IBM 互換機でも利用できるというわけである．さらに，SIR, BMDP, SPSS, OSIRIS-Ⅲ, CLUSTAN, MINITAB などは広範囲の機種を対象にしかも標準的なフォートラン言語を用いていることもあって多様な機種への移植が可能である．また Data-General (Eclipse, Nova), DEC (VAX-11, PDP-11, DEC-10/20) など中型機やミニコンピュータに適した版の開発にもそれぞれ熱心である．たとえば MINITAB, P-STAT などのように種々の機種に対応できるようにプリプロセサーが考えられていたり BMDP のようにデスクトップコンピュータ用の版 (Stat Cat) まで作られるようになってきている．

ところで最近，利用言語や OS の相違からおこるプログラムの改編という拘束から逃れるための方策の一つとして，プリプロセサー（あるいはシステム・ジェネレータ）の開発や研究が盛んである．

多様な機種や OS をすべて満足するような標準言語は考えにくいし，またハードウェアや OS の進歩が急速で現状が永続するという保証もないから，むしろ，システムの設計・開発・保守，移植作業などを一貫処理できるような“プリプロセサー”を考えようというものである．こうした方向も，統計ソフトウェアを身近なものとし，データ解析の負担の軽減やシステムの普及に結びつくことであり，今後の重要な課題である．

プリプロセサーのはたす役割として，

（1） システムの使用言語の種類に拘束されずに，移植可能な形式に拡張，変形できること

（2） 機種や利用環境依存度を意識させない形に変換できること

（3） ソースシステムの保守，更新，制御はも

```
UNNUMBERED
TITLE BANK EMPLOYMENT STUDY

FILE HANDLE  BANKDATA  NAME='BANKDATA DATA M'
FILE HANDLE  BANK      NAME='BANK SPSSXSF A'
DATA LIST  FILE=BANKDATA/
   ID 1-4  SALBEG 6-10  SEX 12  TIME 14-15  AGE 17-20 (2)
   SALNOW 22-26  EDLEVEL 28-29  WORK 31-34 (2)
   JOBCAT, MINORITY 36-37
VARIABLE LABELS
   ID          'EMPLOYEE CODE'
   SALBEG      'BEGINNING SALARY'
   SEX         'SEX OF EMPLOYEE'
   TIME        'JOB SENIORITY'
   AGE         'AGE OF EMPLOYEE'
   SALNOW      'CURRENT SALARY'
   EDLEVEL     'EDUCATIONAL LEVEL'
   WORK        'WORK EXPERIENCE'
   JOBCAT      'EMPLOYMENT CATEGORY'
   MINORITY    'MINORITY CLASSIFICATION'
VALUE LABELS
   SEX         0 'MALES'  1 'FEMALES'/
   JOBCAT      1 'CLERICAL'  2 'OFFICE TRAINEE'  3 'SECURITY OFFICER'
               4 'COLLEGE TRAINEE'  5 'EXEMPT EMPLOYEE'  6 'MBA TRAINEE'
               7 'TECHNICAL'/
   MINORITY    0 'WHITE'  1 'NONWHITE'
MISSING VALUES  SALBEG, TIME TO EDLEVEL, JOBCAT (0)/
                SEX, MINORITY (9)
PRINT FORMATS  SALBEG SALNOW (DOLLAR 11.2)

COMPUTE  AVGRAISE=(SALNOW-SALBEG)/TIME
VARIABLE LABELS  AVGRAISE 'AVERAGE MONTHLY RAISE'
PRINT FORMATS  AVGRAISE (DOLLAR 9.2)

DO IF AVGRAISE GT 300 OR AVGRAISE LT 30
PRINT  / ID 'BEG: '  SALBEG  'NOW: '  SALNOW  'MONTHS: '  TIME
       'RAISE: '  AVGRAISE  'JOBCAT: '  JOBCAT
END IF

FREQUENCIES  VARIABLES=SALBEG SALNOW AVGRAISE/
             FORMAT=NOTABLE/STATISTICS/HISTOGRAM/
BREAKDOWN TABLES=SALBEG SALNOW AVGRAISE BY MINORITY BY SEX
SAVE OUTFILE=BANK
```

図 5　SPSS-X の記法例（マニュアルから引用）

```
 1 'REFERENCE'   GENIUS
 2 'UNIT'        $ 144
 3 'INTEGER'     DEGREE=2, 3
 4 'NAME'        CARNAME=LAW, POLITICS, ARMS, LETTERS, SCIENCE, POETRY, ART,
                         CHURCH
 5 'NAME'        SEX=MALE, FEMALE
 6 'FACTOR'      CAREER $ CARNAME
 7 'FACTOR'      RELATION $ DEGREE
 8 'FACTOR'      LINE $ SEX
 9 'TABLE'       KINSMEN $ LINE, RELATION, CAREER
10 'READ'        CAREER, RELATION, LINE, NUMBER
11 'TABULATE'    VARIATES=NUMBER ; TOTALS=KINSMEN
12 'PRINT'       KINSMEN $ 9
13 'RUN'
```

図 6　GENSTAT の記法例（文献 2）から引用）

ちろん，新プログラム・モジュールの追加や不要モジュールの削除などが可能なこと

（4） プロセサー自身の言語は簡明平易な自己記述的構造を持つこと

（5） 生成したシステムのテスト機能を持つこと

（6） システムにくらべてプロセサー自体の負荷（大きさ）が大きくないこと

などが挙げられよう[1),11)]．今後，統計ソフトウェアの規模の拡大や分化に伴い，プリプロセサーに対する関心と期待が高まることは間違いない．

代表的な統計ソフトウェアの特徴

統計ソフトウェアの概要を類型化という操作を通してみたわけであるが，改めて国内に目を転じてみると，それほど多くのシステムが普及しているわけではない．おそらく SPSS, SAS, BMDP, OSIRIS, GENSTAT, MINITAB などが主なものであろう．いずれも世界各国に広く普及しており SAS などは最近その配布数が 5000 セットをこえたという．また他のシステムもかなりの普及率であろうと予想される．

いずれのシステムもそれぞれ個性があって，どれが良いと決められない．いわば利用者の好みの問題である．しかし最近 SPSS-X, MINITAB-82 が公開され，SAS や BMDP も精力的に改編を続けている．こうした動きにいくつかの特徴がみえるので，これについて若干触れておこう．個々のシステムの機能，使い方，収録課題などについては既に知られていることが多いと思うので，ここでは SPSS-X, SAS, GENSTAT, MINITAB について，それらにみられる主な傾向についてだけ述べることにしよう．

第一はシステムを利用するための命令語の問題である．いうまでもなく，上に挙げたソフトウェアはいずれもシステム化の度合が高い，いわゆる統計システムであり，それぞれ独自の命令語体系を持っている．図5，図6は各システムにおける記法例である．それぞれどのような処理を意味しているか，大よその見当がつくであろう．これは習得が容易であることをも意味する．統計システムが普及する最大の理由はここにあると思われるのであるが，システム化が進むほど，また命令語体系が自然言語に近い表現をとるほど言語化現象が進み，簡便性や習得の容易性が犠牲になる．またデータ処理の連続性，反復性，変換といった機能や演算処理(行列変換，算術式)，各種ユーティリティ，マクロ機能など多彩な内容を盛り込むあまり（それがシステムでありパッケージというものであるが)，言語がより構造化されいわゆる“メタ言語”へと移行する．こうなるともはや一つのプログラム言語に近く，SPSS-X, SAS などは明らかにこうした方向をたどっているわけである．いずれにしても利用者はシステム独自の“文法”を習得せねばならない．

別の特徴として，新しい統計手法を次々に収録しようという傾向がみられる．前述のように，離散データ解析法，時系列分析，各種非線形モデル，クラスター分析，グラフィク機能，EDA 手法など競って収録する傾向にある．また従来の収録手法のオプションの充実にも熱心である．こうした多目的志向は商品として多くの利用者の要求を満たすためにはやむを得ぬことであるが，必らずしも健全な方向とはいえないのではなかろうか．

今後の方向

統計ソフトウェアが商品としての性格を強め，利用者の要求を広く受け入れようとすればするほど，システムは巨大化への道をたどることになる．しかし総花的にデータ処理機能や解析手法を取り入れることが必らずしも利用者の要求を満たすことにはならない．いわゆる“汎用的”ということはありえず，データ解析とは本来きわめて個別的であり多様な側面を持つものであろう．

しかも巨大化現象はシステムの個性を失わせることにもなる．たとえば，SPSS-X は従来の SPSS にくらべて“社会科学向き”という本来の特色が薄れたようにみられる．また，OSIRIS-Ⅲ にみられた社会調査データ向きの種々の特徴的なデータ処理機能や統計解析手法の良さが OSIRIS-Ⅳ では薄れて，きわめて没個性的なシステムになっている．SAS も巨大化の道をたどっており，個々の解析手法のオプションや作表・グラフ機能など必らずしも実務家の要求を満たしているとはいえな

い.

データ解析の初心者にとっては，統計ソフトウェアは大変重宝であるが，安易に用いるあまり内容の理解がおろそかになったり，良い結果が得られたという錯覚を与えるおそれすらある．またベテランの分析者にとっては，システムの規模が大きいわりには使い方や内容に柔軟性がない，という不満が残る．

こうした個々の不満をすべて吸収した理想的なソフトウェアは考えられない．また計算機の進歩や統計理論の発展がこれを許さない．いずれにしても今後統計ソフトウェアが進む方向としていくつかの道が考えられる．

その一つは巨大化の道であり，他の一つは分化のそれである．前者は，既に SAS や SPSS-X にみられる現象である．巨大化の利点としては，ファイル処理機能，マクロ機能，自己管理機能（保守や各種ユーティリティ），他のシステムとのデータの連結利用機能などが豊富なことである．一方，欠点としては，システム化度が高まるので専用のコマンドが一種の言語に近い形，いわゆる“メタ言語”になり初心者が短時間で習得したりシステムの全容を十分に理解することが困難になることである．しかしこうしたメタ言語化により，行列・数学関数・演算式処理，反復，分岐処理などを自由に用いることができ，処理手順をある程度“構造化”できるという，分析のベテランにとってはきわめて魅力的な一面もある．

分化の方向としては，次の二つが考えられる．

（1） 解析内容の多様化に伴う分化

データ解析といっても専門領域により扱うデータの性格も手法も大きく異なる．この多様化に対応して小回りのきくきめの細かい分析を行うには特定な分野の手法を“専用ソフトウェア”として用意することである．またその分野あるいはある手法については一通りのデータ処理や解析が出来るようになっていることが望ましい．この種のソフトウェアは GLIM, SALS（最小二乗法標準プログラム），LTSM のようにサブルーチン・ライブラリーやプログラム・ライブラリーに多いが，データ編集・作表に適した RGSP, COCENTS, クラスター分析専用システム CLUSTAN, NT-SYS, MINTS のようにシステム化されたものも多い．また SPAD のように数量化法（Ⅲ類）を中心にそれと関連する諸機能（得点の同時布置，追補処理，データ変容機能，自動分類機能など）をコンパクトにまとめたシステムもある．また多重クロス表解析のパッケージとして ECTA, LOLITA, GENCAT, MULTIQUAL 等がある．

専用システムあるいはプログラムに共通した特徴としてデータ編集，ファイル処理などの機能が必らずしも十分でないことがあるが，SAS, SPSS, BMDP, などの大規模システムとのインターフェースを持ちデータの授受が出来るというものが多い．

（2） 利用形態の分化

分化のもう一つの方向として計算機の利用形態の問題がある．端末機の普及やローカルエリアネットワークの進歩に伴い処理形態がバッチモードに限らず会話型に移行する現象もみられる．SCSS, MIDAS, IDA, MINITAB, EASY 等数多く登場しているが，教育的効果という点で高く評価できても大量データを扱う社会調査分析などにはやや不向きであろう．むしろ専用コマンドを会話型で編集・カタログ化し実行はバッチ型で，というハイブリッドな方式が考えられるが，実際多くの統計システムはこれに近い形式をとっている．

ところで，すでに指摘したようにデータ解析の道筋は一つではない．ましてやある方法を用いるとうまい結果が得られる，というものでもない．統計ソフトウェアの便利さに惑わされて安易な紋切り型の分析に終わらせぬためには，分析者の考える独自の分析手順や解析手法を統計システムの中に組み入れることを可能にすることである．分析者はうまい分析手順を工夫しその“分析手順を組み立てる”という操作を計算機に委ねるわけである．

実はこうした試みが全くみられぬわけではない．たとえば，CONCOR というシステムは調査データ（主にセンサスや住宅調査など）の個票管理（ファイル更新，検索，編集，文書化），ファイル処理，データ検証，決めつけ処理（自動修正）など，いわゆる自動編集システムである．利用者がデータ処理内容を CONCOR 専用言語で記

述し入力するとそれに必要な課題を選択し連結してその処理分析に用いる実行モジュールを生成するという，いわゆる“プログラム・ジェネレータ方式”を採用している．総理府統計局が独自に開発した TLOPS (Tree Logic Programming System) がこれに類似の機能を備えており，PL/I に似た簡易言語でデータ処理や作表の形式等の手順をわりつけると，必要な処理が簡単に実行できる．いわばプログラムレス・プログラムを目指したものである．

これとは別に，一つの統計システム機能だけでは満足できないので，複数のシステムを結びつけてそれぞれの持ち味を生かして利用したいというぜいたくな要求も出てくる．とくにデータ解析の経験があってその妙味を体験したベテランほどこう感ずるはずである．また自作のプログラムやサブルーチン・ライブラリーを統計システムに組み入れて併用したいということもあろう．

GENSTAT にみられるように，これを RGSP, GENKEY と連結して利用できる．また CONCOR は生成ファイルを集計プログラム COCENTS に連結可能である．SAS では，サプリメントとして別にプログラムをライブラリー化しておき，これを呼び出して利用できる．このように限られた親類関係にあるシステム間での連結利用は多くの例がある．もちろんシステム間でのデータのやりとりについてはほとんどのシステムが可能であるといってよい．

つまり，分析者のレベルが高まり密度の濃い複雑な分析を柔軟に進めるためには，各システムが他のそれとのインターフェースを用意することである．SAS, SPSS-X, GENSTAT などは，個々のシステム内ではこれに近い機能を備えているが，外部に用意したプログラム類を必要に応じて望む個所に連結して用いることまでは無理のようである．

別の方法として OS の機能を生かすことが考えられる．いわゆるカタログ・プロセデュアやコマンド・プロセデュアのように OS 下で機能する簡易言語を用いて，利用したいプログラムやシステム類を解析手順にあわせて登録しこれを呼び出して処理を行うわけである．これはあくまで“便

法”ではあるが，場合によっては利用者はこうした裏の処置を意識せずに種々のシステムを任意に連結させて有機的に処理分析を進めることができる．

いずれにせよ，技術的な諸問題の解析もさることながら，今後の統計ソフトウェアがとるべき道がどのようなものか，あるいはどうあるべきか真剣に取り組む時期にあるといえよう．それにはまず，多くの統計ソフトウェアの利用体験をまた互いの分析経験を，いろいろの観点から議論しあうことであろう．そして何よりも，統計ソフトウェアを“自ら作る”という体験を通して，それが抱える問題に応えることであると考える．

参 考 文 献

1) Anderson, A. J. B., Aberdeen, G. B. (1980), The use of preprocessors in statistical software, COMPSTAT-80.
2) Alvey, N. and others (1982), An Introduction to GENSTAT, Academic Press.
3) Anscombe, F. (1981), Computing in Statistical Science through APL, Springer Verlag.
4) Arbeitsgemeinschaft Sozialwissenschaftlisher Institute (1981, 1983), Social Science Application Software, European Political Data Newsletter, No. 40, No. 47, Norwegian Social Science Data Services.
5) Benzécri, J-P. and others (1980), Pratique de l'Analyse des Données, Tome 1, 2, 3, Dunod.
6) Chambers, J. M. (1977), Computational Methods for Data Analysis, John Wiley.
7) COMPSTAT—Proceedings in Computational Statistics, 1976, 1978, 1980, 1982, Phisica-Verlag.
8) Dixon, W. J. and others (eds.) (1974), Exploring Data Analysis——The Computer Revolution in Statistics, Univ. of California Press.
9) Francis, I. (1981), Statistical Software, A Comparative Review, North-Holland.
10) Francis, I. (1983), A Survey of Statistical Software, Computational Statistics & Data Analysis, Vol. 1, p. 17-27.
11) Gadzik, W. F., Karpel, L. C., and others (1980), SIRTRAN——a macro preprocessor for software portability and maintenance, COMPSTAT-80.
12) Girish Punj and David W. Stewart (1983), Cluster Analysis in Marketing Research: Review and Suggestions for Application, Journal of Marketing Research, Vol. XX (May 1983), 134-48.
13) Hartwig, F., Dearing, B. E. (1979), Exploratory Data Analysis, Sage Publications. (探索的データ解析の方法，柳井晴夫他訳，朝倉書店).
14) Jambu, M., Lebeaux, M-O., (1983), Cluster Analysis and Data Analysis, North-Holland.
15) Jambu, M., Lebeaux, M. (1982), LTSM-Logiciels de traitements statistiques multidimensionels, A. D. D. A. D.
16) Lauro, N., Serio, G. (1982), Criteria for Evaluating and Comparing Statistical Software: A Multidimensional Data Analysis Approach, Statistical Software Newsletter, Vol. 8, No. 3.
17) Lebart, L., Morineau, A. (1982), SPAD-Systeme portable pour l'analyse des données, Cesia.
18) Tukey, J. W. (1977), Exploratory Data Analysis, Addison Wesley.
19) Velleman, P. E., Hoaglin, D. C. (1981), Applications, Basis, Computing of Exploratory Data Analysis, Duxbury Press.
20) Volle, M. (1981), Analyse des Données (2eme edition), Economica.
21) 浅野長一郎，田中 潔 (1982)，統計プログラムパッケージと多変量解析，数理科学，No. 230, p. 67-75.
22) 大隅 昇 (1979)，データ解析と管理技法，朝倉書店.
23) 大隅 昇 (1980)，フランスにおけるデータ解析の動向—Benzécri の数量化法を中心に—，数理科学，No. 204, p 56-64.
24) 大隅 昇 (1982)，自動分類法のソフトウェア，数理科学，No. 230, p. 22-34.
25) 水野欽司，大隅 昇，桂 康一 (1978, 1979)，統計パッケージ，①〜⑥，(完)，**bit**，Vol. 10, No. 8, 9, 11, 12, Vol. 11, No. 1.
26) 矢島敬二，大隅 昇 (1977)，統計，「アプリケーション・プログラム」，**bit**，Vol. 9, No. 9.

マニュアル類

1) SPSS-X User's Guide (1983), SPSS Inc., McGraw-Hill.
2) SPSS-X Statistical Algorithms (1983), SPSS Inc.
3) SAS User's Guide 1982 edition: Statistics, Basics, SAS Institute Inc.
4) SAS Programmer's Guide 1981 edition, SAS Institute Inc.
5) SAS Views 1980 edition (日本語版).
6) BMDP Statistical Software, 1983 Revised Printing, University of California Press.
7) 医療の為の統計パッケージ，SPMS，利用の手引，(1980年6月)，東京都臨床医学総合研究所 医療情報学グループ
8) ETPS 調査・統計データの編集・作表システム，日本電子計算株式会社
9) クラスター分析プログラム・パッケージ，MINTS-80 (Mini-Numerical Taxonomy System)—利用の手引—，1980年5月，統計数理研究所，研究リポート 51
10) 社会調査データの質の統計的評価のための数理的処理および管理体系の開発，統計パッケージ MINERVA—利用の手引—(第1版)，1980年3月，統計数理研究所，研究リポート 48
11) 会話型統計プログラム・パッケージ EASY 説明書，日本電気株式会社
12) 最小二乗法標準プログラム，SALS (第2版)，利用の手引き (第1部基礎篇)，1979年5月 (改訂2版)，東京大学大型計算機センター
13) 最小二乗法標準プログラム，SALS (第2版)，利用の手引き，(第2部 制御・解法篇) 1979年5月 (改訂2版)，東京大学大型計算機センター
14) TIMSAC-78, Computer Science Monographs, No. 11, 1979年2月，統計数理研究所

(おおすみ・のぼる，統計数理研究所
たるみ・ともゆき，岡山大学・教養部)