医療情報学 6－6
（1980. 6. 28）

ＳＰＭＳ（Statistical Package for Medical Science ）の設計思想とその運用

刈谷丈治　丹後俊郎　東京都臨床医学総合研究所

1、はじめに

ＳＰＭＳは延べ６人によって開発されたプログラムシステムである。各人それぞれの思想と視点があるので、ここでは刈谷の独断と偏見を述べる。

ＳＰＭＳについて種々の学会、論文で宣伝された。しかし、その要素たる手法技法概念等を取り出して、他のシステムと比較すると、特にめずらしいものは数少ない。但し全体としては満足して使用している。これは使用目的に合わせて作製したのだから当然の事であろう。もっとも、作製者すら使用しないシステムの話を耳にする事もあるので、それに比べれば大成功のシステムといえよう。研究会を通して、この満足さにどの程度の賛同が得られるかが判れば幸いである。

2、発端

「医学データを解析して何らかの知見を得たい」という要求仕様から1975年に始まった。この時点での設備は主記憶９２ＫＢ、ディスク３０ＭＢとカードリーダラインプリンタであった。（現在はこの他にディスク３０ＭＢ，ＭＴ２台、ＰＴＲ１台が増設され、PDP11/70上のＢＭＤＰが使用可能）当然にバッチ処理による数値的解析のみが可能であり、統計解析を行うこととなる。当事筆者は、統計パッケージに無知であり、メーカ提供のパッケージに不満足というだけで自力製作を決意した。この時の不満足の理由は次のようなものである。

i)　３次元（Time Oriented）データが扱えない　-----データ構造
ii)　インコア処理だけであり、大量データが扱えない　-----データ量
iii)　かなり整理されたデータが対象であり、生データが扱えない---前処理管理
iv)　手法間の連絡がなく、高次処理が難かしい　-----高次処理
v)　利用者の作った処理を処理過程に組み込むことが難しい　-----特別処理

3、経過

ＳＰＭＳは最初から現在の形をしていた訳でない。これは、必要なデータ表現能力の見切りの誤りと情報処理技術の不足によるものであり、現実のデータを取扱い、勉強を行い、他システムを知る事により現在の姿へ変わって来た。経過を簡単に述べると、

Version 1　単一 flat table で case／record の逐次編成ファイル
Version 2　全体、個人別、時系列の３つの case／record 逐次ファイル上で３次元の処理を行う。（計画だけ）
Version 3　paging した直接編成ファイル上に複数（固定）の flat table を項目毎に記憶した。
Version 4　Version 3 の色々な制限を取払い、コマンドが導入され、ほぼ現在の形となった。
Version 5　マトリクスの導入と、全プログラムの整理、機能の拡張、手法の充実を行った。

この開発以前及び開発中に

i) MULTICSのレベルストアと仮想記憶の管理と効率
ii) DBのデータ中心の考え方と、access method, navigationについて
iii) TOD開発者の保存、管理と解析でのデータ検作の差異に対する意見
iv) relational databaseの思想……SYSTEM R, INGRES, LSL
v) 大須賀先生の論理、集合、DBの並行性、同一性に対する意見
vi) 特定研究「情報システムの形成過程と学術情報の組織化」のM委員会でのPRF(private researcher's file)の議論

等の影響を受け、システム内部はより一般的に単純になっていった。一方、データアクセスを容易にかつ強化されたアプリケーションプログラムは機能を増強し、コマンドにより使用の容易さを加え、外部から見た時の多種多様な機能の、統一的利用が可能となった。

4、要請

臨床研という利用者側での開発なので以下の事が要請された。

i) 作製者は科学計算ができる程度の計算機知識で良い。
ii) 利用者は計算機に関する知識がなくて良い。
iii) 効率の優先順位は、利用、作製、計算の順に低くなる。
iv) 利用者が機能を補完することを、全てのレベルで許す。即ち、利用者と開発者の作製した機能は、汎用性、親切さ等で異るだけである。
v) 2年目以降設備増強はないので、全て92KB中に納める。

以上の要請に基づいて、種々の選択が行われた。

5、階層

データ管理、アプリケーションプログラム、コマンド処理の三層に大別し、それぞれのインタフェースを極力僅少単純にし、各レベルが各々抽象機械を提供する(とみなす)ようにした。現実には理想的に行かないし、機能の追加を許すために、各レベルの仕様も変動し、広域的制約となる機能の作製は困難であった。この為に、重要な機能のうち、利用者の管理に委ねられているものもある。

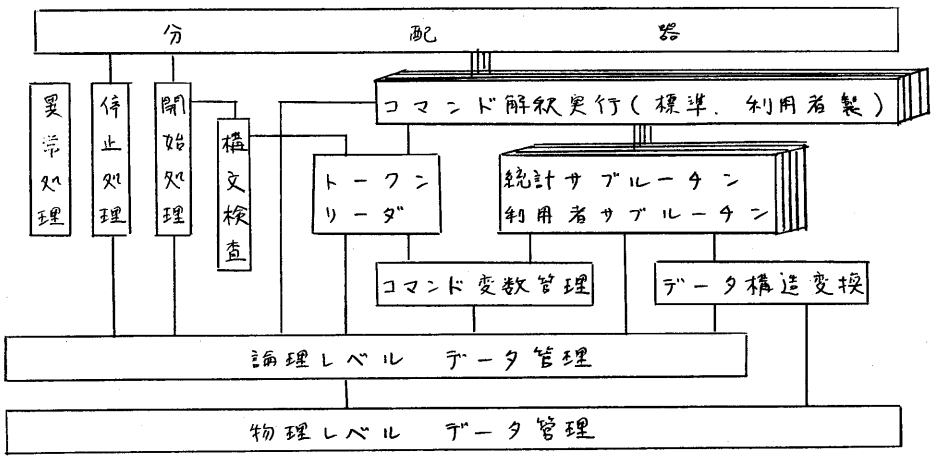

図1 SPMSの構造 3)より修正引用

6．言語

階層化され、かつ開放されたシステムでは、それぞれの成分が何時どんな使われ方をするか不明である。しかもシステム化されているから、実行時に管理情報が主記憶に常駐している。従って異常時に直ちに停止するとデータを失う危険があるので、異常時処理が可能な言語でなければならない。大量の計算をするので、コンパイラ言語が良い。インタプリタならAPLの如く配列演算能力を要する。そこで使用可能言語のうちPL/Iを使うこととした。但し、小さいPL/Iであり、可変長文字列、構造体の配列、言語間通信、アロケート文、可変配列限界等、システム化に必要な機能がなく、製作上の難点となった。

7．情報

データに統計その他の処理を行い何らかの知見を得るという過程での情報の流れを考えると次のことがわかる。対象はある抽象的な実体（entity）であり、観測可能な属性を持っている。この属性の実現値（occurrence）がデータである。これから観測によらない性質や、観測値の予測をする為の性質を発見したい訳である。これら実体の本態的性質は観測データに、仮定、仮説、別の知見を導入して、対象の属性の総合化（主成分分析、関数etc）、対象の層別、分類による統合的抽象化（クラスタ分析、分割表検定etc）を行って表出される。より普遍的抽象的属性の導入により、対象は構造化され、従ってその実現値による表現であるデータは構造変換を受ける。

以上の過程での重要な機能を列挙すると

i) ROI（region of interest）への迅速なアクセス

ii) 新属性の容易な生成と構造表現能力

iii) 動的な構造変換（dynamic restructuring）

である。更に人間がこれを行う時を考えると、ページをめくってノートやメモに書き足し、欄を増やしたり別の表を作ったりする訳で、管理的業務としては、所定の所へノート類をしまうだけで殆んど直接作業であろう。従ってノートをDiskに、鉛筆をシステムに置換して異和感のないようにすることが大事で、これはPRFの思想と一致する。

8．データ管理

datasetはrelationの集合である。relationnal modelのrelationからdependencyやschemaを除いたもの（つまり単なるflat table）と考えて良い。異なる点は、欠損値を許す、tupleの順序が有意である、データ要素に配列を許す（スペクトルや集合データの為）等である。このレベルでのインタフェースはrelationやdomain（itemという）の生成削除、任意のデータ要素への直接アクセス等である。

datasetに対する命令は全てsymbolic/logicalに行い、open等の特別な処理を必要とせず、かつ結果はDisk上に反映される。こうしてデータに関しては論理的なイレベルストアを実現している。この使用法はMUMPS経言語に組み込まれていないが本質的差はない。

ハード的には、Diskをpaging deviceして用い、relationをitem毎に分割して記憶し、対象itemだけの入出力を可能とし、高速化を計った。アクセスにも管理にもhashingを用いて高速化し、overflowを許してrelation、item数の制限を

なくし、garbage collectorの自動起動により、flagmentationに備えた。管理情報はdeffered writeをせず、異常時にデータの一部が失われても他に被害が及ばないようにした。一方データ値は4ノ面のFIFO page bufferにより、多項目の効率的処理と、thrashingを許せば任意の処理ができる。

page I/Oを別タスクとし、アクセスパタンを調べ、逐次処理ならprefetchするようにすれば50%高速化できる。主記憶不足の為現在は出来ない。

9、データ構造の変換と表現

構造変換には、relationのsubsetを取出す(PROJECT:RDBのselection とrestriction)、論理的併合(MERGE)、物理的併合(CONCATENATE)とより基本的なSORTがある。joinやdivisionの要請は多く、作製されていない。

構造表現の機能は基本表現として存在しないため、アプリケーションレベルの抽象データタイプとして導入される。正当性はこのレベルの管理に委ねられ、現在の所利用者の責任となっている。現在の唯一のデータ構造はtree構造であり、sorted keyを持つ2relation間に、relation index(tuple番号)により、1対多の関係をつけたものである。この構造により3Dの表現を行う。この構造だけでも一般的なhierarchyが表現可能であり、操作対象が広くなり、また内部的には多レベルindexによるinverted file等の構築が可能となる。

10、コマンド

利用者にとって使い易いためには、利用者の扱う概念レベルでの表現と操作が可能でなければならない。この為のインタフェースがコマンドである。現在は2D及び3Dデータの表現しかないが、この表現上での操作が数多く作られている。

コマンドはその発生から言って、BMDPのように、単一のコマンドで一定の完結した処理を表わすものが多い。一方複雑な処理を可能とする為に、複数のコマンドで表現するSPSSのようなものもある。SPMSでは、必要なデータしか入出力をしないので、逐次的処理であれば別コマンドとしても効率が余り落ちないという点を生かし、前処理のような事を全て別コマンドとし、コマンドが単一機能の表現になるようにした。層別処理を容易にする為、多くのコマンドは動的な層指定が可能となっている。簡単なコマンドでは、他のパッケージ同様、複数の項目指定や複数の層指定が可能である。

利用者にとって憶え易く、理解し易いという事も重要である。この為に、完全自由書式のやや冗長ではあるが英語に近い表現をとっている。この辺りは会話型に移行する時に最も影響を受ける所であろう。

11、コマンド間関係

コマンド実行結果の殆んどがリスト出力かrelation/itemの生成となるので、コマンド間はほぼ独立である。コマンド間に渡って使用されるものがあると、プログラム言語同様scopeルール等が問題となる。現在この種のものは、対象relationの選択と、コマンド変数、繰返し制御である。同一relationに対し複数の処理を行う事が多いので、対象の指定を分離してコマンドとし、広域的に有効とした。コマンドにより単一の値が得られる時、これをitemとして見るより、そのまま変数値と見た方が理解しやすいので、コマンドレベルでのみ存在する、

コマンド変数というプログラム変数に対応するものを作った。この変数とitemの要素値との変換を許すことにより、コマンド間での単純な値の交換、繰返し制御のパラメータ指定等の能力が強化された。

複数コマンドの繰返しを、対象やパラメータを変えて実行する為の繰返し制御コマンドが作られ、多量連続計算に用いられている。

12、コマンド解析

機能が順次追加され、かつ理解し易い事というので、計算機言語の文法にとらわれない勝手なシンタクスでコマンドが定義されている。従ってコマンド解析の為のツールは開発されておらず、各コマンド毎にインタプリタを作成している。利用者が利用者用コマンドを作り、実行時にパラメータを授受するにはコマンド解析を行わねばならない。この為にシンタクス要素をわかり易くしている。これは逆にプログラム言語から見れば奇異な表現である。

シンタクス要素はキーワード、特殊記号、relation名等の名標、コマンド変数、定数である数値や文字列とコメントである。これらは先頭の一文字で識別され、文脈によらず確定する。

数値やrelation名をパラメータとして受取る事は多くても、出力する事は少い。そこでコマンド解析時にパラメータの値を受取る所では、コマンド上で変数であっても値が入力されるようにした。この機能により、プログラム上で定数変数の区別をしなくても良い。これは普通のlexical analyzerと異る工夫である。

13、運用

利用と開発が同時進行する本システムの場合、(1)実験者－実験を行い目的を持つ人　(2)解析者－手法の選択適用をする人　(3)手法開発者－数値計算を中心とするプログラムを作成する人　(4)基礎開発者－使用しやすい高度環境を整備する人の4層に分かれる。各層の人は少なくとも前後の層の内容をある程度知る事が必要で、そうしてはじめて目的に沿った製作運用ができる。臨床研に於て、理想的とは言い難いが一応各層の意見交換があった事を強調したい。

最も閉鎖的な所を打破する為、小人数ながら医学生等の啓蒙も行っている。

14、結言

本システムの他パッケージと異なる点は、仮想メモリ、1レベルストアやDB等から思想を借り、その上に手法、機能群を乗せたことにある。パッケージとしての有効性は、この手法群が一定のレベルに達し、有機的結合が可能となり、しこうして現実のデータに適当されてはじめて証明される。現在使用可能な機能を引用しておくので、詳細に関心のある人はマニュアル等の資料を請求されたい。portabilityがないため、当研究所においてし分使用できないが希望者はご相談下さい。

15. 機能

標準コマンドと重要な句の機能を文献(4)から引用する。現在更に一部の機能が増強されているが詳細はマニュアルにゆずる。

Table 1. The list of principal keywords of standard commands and their corresponding functions.

Keywords	Functions
IN RELATION	Specify a relation.
TITLE IS	Title setting.
REPEAT	Repeat the procedures. 'REPEAT variables' have 4 types of attributes, viz., 'name' variable --- !N ∿ !N99, 'value' variable --- !V ∿ !V99, 'class' variable --- !C ∿ !C99, and 'string' variable --- !S ∿ !S99, all of which are also called 'command variables'. (see ASSIGN command.)
END REPEAT	End of repetition.
PRESERVE	Preserve the SPMS data file onto tapes.
RELOAD	Reload the SPMS data file preserved on tapes onto disks.
CONVERT	Conversion of the data files between SPMS and BMDP.
PROJECT	Project a relation onto a new relation.
MERGE	Merge two relations into one. (Horizontal connection.)
CONCATENATE	Concatenate several relations.(Vertical connection.)
LINK	Link two realtions by making the item which indicates the correspondence between the cases. This function plays a very important role in representing clinical time series data structures with three dimensions.
SORT	Sort a relation in the ascending order of specified key items.
CREATE	Create new relations, items, and matrices. In the case of entering an initial data set into the SPMS data file, three types of input mediums can be used, viz., Card, Tape and Disk.
MAKE (LET)	Make a new item by re-coding, stratification, standardization, mathematical operations between items and so forth. Several indicators representing, e.g., random sampling or a certain state of clustering process, are also made. Transformations of information between the two relations, BASIC relation and VARIABLE relation, are as follows: 1) VARIABLE → BASIC --- elementary statistics, pattern classification of time series variation, conditional case selections and so on. 2) BASIC → VARIABLE --- copy, time scale change, differentiation and integration of individual time series data, et cetera.
UPDATE	Update an element of item, a relation name, an item name, a matrix name, or relation range (the number of cases in a relation).
DELETE	Delete relations, items or matrices.
SHOW	Show a list of relation names, item names,data of specified items, elements of specified items or individual clinical records.

ASSIGN	Specify mathematical operations or assignment among command variables and data elements. Especially the function 'DATA(relation, case, item)' which has access to an element specified by the three parameters, i.e., relation name, case number and item name, gives flexibility in analyses.
GENERATE	Generate random numbers such as uniform random numbers or normal random numbers. In total, 16 types of random numbers are generated.
COMPUTE	Compute elementary statistics, Kendall or Spearman rank correlation coefficients.
TABULATE	Tabulate cross tables, pairwise or partial correlation coefficient matrices or various kinds of tables.
PLOT	Plot histograms, normal probability curve, scatter diagram, personal variation curve, moving average curve or several types of graphs.
APPLY	Apply advanced statistical methods: 1) Parametric tests --- Homogeneity test of two means, two variances, two correlations or two populations. Normality test and so on. 2) Non-parametric tests --- Kolmogorov-Smirnov two sample, Kruskal-Wallis one-way ANOVA, Cochran Q, Mann-Whitney U, Friedman two-way ANOVA, Sign test. 3) Contingency test --- Yate's correction, Fisher exact probability, Cramer V and Chi-squares. 4) Multivariate analysis --- multiple regression, discriminant function, principal component, factor analysis (varimax rotation), polynomial regression, cluster analysis of cases, canonical analysis, Hayashi quantification type 1-4, non-linear least squares fitting. 5) Other methods --- Clinical normal range setting (IRM), analysis of variance or covariance (fixed-effects and random-effects model).
EXECUTE	Execute user-own coded programs in the standard process of analysis.
SELECTED BY	Case selection and stratification are performed.
SAVE	Save computed results in the form of relation, item or matrix.
WHERE	Specify the parameters or conditions to be used in the specified command.

(C.F.) As to the statistical test, there is an option whether a significance level is to be specified or not, if not, the results are automatically shown at three significance levels, viz,, $p < 0.05$, $p < 0.01$, $p < 0.001$.

16. 文献

(1) Tango, T., Kariya, J. et al. (1978) A Statistical Package SPMS at a computing center for public health care. Proc. of Inter. Conf. on Cybernetics and Soc., 899-903

(2) Tango, T., Kariya, J. et al. (1978) A Statistical Package for Medical Science (SPMS) Proc. of Inter. Symp. on Medical Inf. System 521-524

(3) Kariya, J. and Tango, T. (1979) The design and implementation of SPMS, Proc. 12th Annual Symp. on the interface, computer science and statistics. 476-480

(4) Tango, T. and Kariya, J. (inprinting) on the development of SPMS as an effective tool for medical data analysis in "Recent Development in Statistical Inference and Data Analysis" edited by K. Matsushita, North Holland, (1980)

(5) 丹後俊郎、刈谷文治、倉科国介、神沼二真 (accepted) 医療データの為の統計パッケージSPMSの開発 医用電子と生体工学