1E−3　統計解析パッケージとの連携利用を実現するためのデータベース構築法

桜井尚子*　古川哲也*　川崎晃一**　上園慶子**
*九州大学大型計算機センター　**九州大学健康科学センター

1．まえがき

データベースが実社会の随所で使われるようになってから久しい．マシンの大容量化や性能の飛躍に伴い，様々な分野でデータベースが利用され始めている．一方，応用プログラムパッケージのデータ管理機能についてはあまり開発が進んでおらず，機能強化の必要性が高い．管理機能はパッケージのプログラムに特化しているため，データを一般のプログラムから利用することが困難である．また統計解析では解析するデータの選択など提供されているデータ管理機能では不十分な面が多い．本稿では統計解析パッケージのデータ管理機能の問題点を指摘し，データベース管理システム（ＤＢＭＳ）と連携したシステムの構成を提案する．また解析の対象となるデータの性質を考慮したデータベースの構築方法の一例も併せて提案する．実際の解析例には，九州大学健康科学センターのデータの一部を使用した．

2．統計解析パッケージにおけるデータ管理

一般に用意されている統計解析パッケージは，処理したいデータを持参するだけである一定水準の解析結果が得られ，非常に利用度の高いアプリケーションである．しかし，毎回同じ処理を繰り返すのではなく，多様な処理を試みたい場合などでは以下のようなデータ管理上の欠点が指摘される．

(1)いろいろな側面からの要求に応えられず，処理の度にデータベースの再作成が必要となる．
(2)作成されたデータベースはパッケージ内で閉じておりデータベースの構成が不鮮明なため，他のプログラムから簡単にアクセスすることができない．

統計解析パッケージにおけるデータ管理の関係は図１のとおりである．

①収集データをパッケージ内に取り込む．この処理は統計解析パッケージ内のツールを使って行う．
②統計解析パッケージ独自のデータベースから必要なレコードを取出し，解析処理を行う．
③出力を見てデータベースの中身を変更しながら別の解析処理を行う．

解析の内容にしたがってどのデータを用いるのかを選択するのだが，パッケージ内のデータベースではその選択が巧くできず，図１−①の部分で再びプログラムを実行する必要が生ずる．また，パッケージの外からこのデータベースにアクセスすることは困難である．

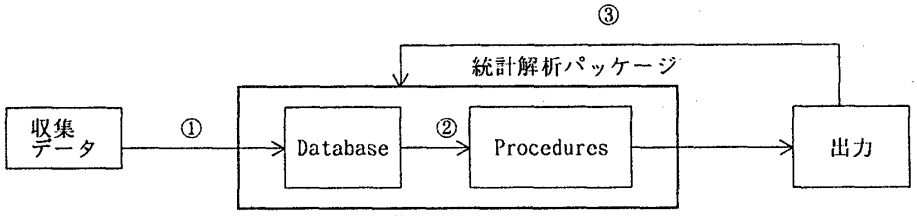

図１．統計解析パッケージでの処理

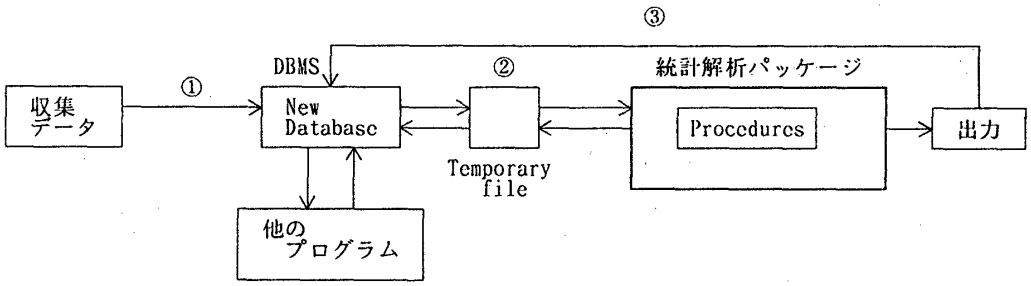

図２．ＤＢＭＳを用いて管理されたデータを使用しての統計解析処理

Constructing Databases to Realize Cooperated Utilization with Statistical Analysis Packages
Naoko SAKURAI, Tetsuya FURUKAWA, Terukazu KAWASAKI, Keiko UEZONO
Kyushu University

3．ＤＢＭＳを用いたデータ管理

２節で示した欠点を改善するため，統計解析パッケージに左右されないようＤＢＭＳを用いて統計解析用のデータを管理する．このような構成では次のようなデータ管理上の長所が生ずる．

(1)解析結果を経て，ＤＢＭＳ配下で作成されたデータベースから統計解析パッケージの処理にわたすデータの再作成が容易になる．

(2)ＤＢＭＳ配下で作成されたデータベースは，統計解析パッケージに関係ない他のプログラムから利用することが可能になる．

以上の処理の流れを図２に示す．

①収集データから解析処理に必要なデータ項目を引き出し，ＤＢＭＳを用いてそれらをデータベースとして構築する．

②①で作られたデータベースを用いて，統計解析パッケージで処理するための中間ファイルを作成する．

③解析結果から，別の解析要求に必要なデータをデータベース操作で拾い出す．

毎回プログラムを作成するのでなく，データベース操作だけで必要なデータを選択しようとする試みが本稿の主旨である．ＤＢＭＳ配下のデータベース操作を用いて図２－②のファイルを出力する行程がこれに相当する．

4．解析用データの実例

健康科学データではひとつの対象についてデータが複数存在する．健康科学データの例を図３に示す．

図３は約９万件に及ぶ健康科学データの一部を抜粋したものである．血圧（ＢＰ）については一回目で基準値内であったものは二回目以降のデータはない．このような性質のファイルの場合，このまま１つの関係とすると空き領域が多く資源節約の観点から好ましくない上，解析操作がしづらい．例えば，ＢＰの平均をとる時に各人について

・最後の値（基準値内の値）を用いる
・測定の平均値を用いる
・測定の中間値を用いる

等が考えられるが，このファイルのままでは上記のデータをデータベース操作で得ることはできない．そこでオリジナルファイルをベースとなるファイルとその関連ファイルに分割し，その融合方法を関係データベースの観点から考えてみる．図４のファイル構成を考える．

処理要求にしたがって必要な項目をデータベース操作で取出し易い形に格納する．但し，どんな処理要求があるのかについては，あらかじめ充分調べておく必要がある．このデータベース構造を持てば，ＩＤが 101 の人についてはこの人のＢＰ（血圧）についてのあらゆる角度からの情報を簡単に引き出すことができる．このデータベース管理機能を用いれば従来の統計解析パッケージに対し，以下の長所が得られる．

(1)統計解析パッケージの一方通行的データベース構造をとおらずに目的の解析を時系列に進めることができる．

(2)専門家の知識をレコード内に反映させ得るので，異なる専攻分野間でもデータを共有でき，それぞれのデータベースを構築できる．

その他の問題としては，測定の事情等によりある年のデータに障害があった場合のデータベースへの対処方法等が挙げられよう．

5．むすび

日頃疑問に感じていた統計解析パッケージ内のデータベースについて統計解析パッケージ外での改善を試みた．本稿で示したデータベース管理機構を用いることにより，眠っているデータの掘り起こしと学際領域の強化に努めたい．

ID	BP一次	BP二次	BP三次	BP四次	BP五次	BP六次	BP七次	EXTRA DATA
101	155	146	122					
102	129							
103	167	158	150	161	152	162	124	
104	128							
・								
・								

図３．健康科学データの実例

ID	項目	LAST?	AVRG	MED	BP
101	1	N	N	N	155
101	2	N	N	Y	146
101	3	Y	N	N	122
101	AVRG	N	Y	N	141
101	NED	N	N	Y	146
102	1	Y	Y	Y	129
103	1	N	N	N	167
103	2	N	N	N	158
103	3	N	N	N	150
・					
・					

ID	EXTRA DATA
101	
102	
103	
104	
105	
・	
・	
・	

図４．新しいＤＢＭＳを登載したデータベースモデル