NOTE

主成分分析における結果解釈過程の構造

——農業部門結合の地域型検出に関連して——

児島俊弘

はじめに

この報告は初めは極めて慎ましい目的をもっていた。それは佐賀県において農業部門別構成比率の市町村データによって、部門結合の地域型を検出する場合に主成分分析を適用するとどういう問題があるか、事実に則して検討する、という目的をもっていた。

しかし、データを検討している間にもう少し一般的な問題、すなわち、主成分分析結果の意味解釈はどのようにして妥当性を獲得できるのであろうか、という問題領域へ入りこまないわけにはいかなかった。それには、主成分分析結果解釈の論理的過程をモデル化して、このモデルに照らして実際のデータ解釈をしてみるというところまで拡大することになってしまった。これは主成分分析そのものを勉強してから一年にしかならないわたくしにとって甚だ大それた試みであって、本来ならば部外の初心者のやるべきことではないように思える。しかし案外初心者だからエキスパートが問題にしていないところに疑問をもつ、ということがあるかもしれない。そう考えてあえて報告にまとめた。

報告は、いまのべたような当初目的をもっていたので、地域農業の部門結合型を求める他の方法との比較から始まっている。一の(2)がそれであって、報告のテーマとは直接関係がないが、これが入らないと形が整わないのでそのままいれてある。

一、なぜ主成分分析を適用するか

(1) 問題の限定

はしがきでのべたように、この報告の当初目的は、農業部門の結合タイプによって県内市町村の型を分類する場合に、主成分分析が適用できるかどうかを検討することにあった。

「適用できるかどうか」という判断の基準はいろいろあるだろう。

なにかの分析目的があって、その目的に対応した市町村のク

ラス分けをまず前作業として行なうのであれば、主成分分析の結果得た市町村のスコアによるクラス分けが、実際に分析目的にとって有用であるかどうかを吟味すればよい。

しかし、ここでは統計データによって農業部門構成の差のタイプ分けを行なう手法を独立にとりあげる。タイプ分けの目的との関連はふれない。その理由は農業部門結合型を見いだして市町村のタイプ分けをした結果は、いろいろな分析目的と結びつくことが可能であって、利用上汎用性をもち独立してとりあげても意味があると考えたからである。

とすれば、「適用できるかどうか」の判断基準は主成分分析の結果得た情報が、もとの原データ群よりも、地域の農業部門結合型の状態をうまく要約して表現していて、しかも原データの情報と矛盾しない、という条件をあげればよいであろう。

この報告は、そのような観点から主成分分析の適用を検討しようとするものである。

しかし、小地域の農業部門結合型を求める方法には、経済地理学の標準的なテキストにのべられている[1]J・C・ウィバーの“Crop Combination”による手法がある。だからなぜウィバーの比較的簡単な方法をとらないで、主成分分析の適用を問題にするか、という点を先にのべておかねばならない。

(2) ウィバーの方法

ウィバーの“Crop Combination”は、あらかじめ部門別分布パターンのモデルをきめておいて、特定地域がこのパターンのどれに属するかを判断する、という手法をとっている。

ウィバーは作物の収穫面積作目別構成比率データによって作物結合型をあつかっている。コポックは、これを拡張して家畜部門を一定の荷重によって作物と結びつけ総合的な部門結合型を考えた[2]。この報告であつかったデータは農業部門別粗生産額構成比率であって、対象は佐賀県市町村である。

ウィバーは極めて単純な前提から出発している。彼はまず「モノカルチュア」のモデルを、ある一作物だけが栽培されている場合、つまり一作物の収穫面積が総収穫面積の一〇〇%を占めている場合と考えた。そしてこの一作物モデルからn作物の場合へと類推する。

「n作物結合」とはその地域の全収穫面積に対する各作物部門の構成比率がどれも、〔1/n×100%〕であるような状態をいう。

その地域で成立している作物の収穫面積は、作物間で均等に分布している状態を理論分布と考えたのである。

特定地域の実際の各作物収穫面積分布が、この均等分布から

どれだけへだたっているかを計算し、そのへだたりの最も小さい作物結合型をその特定地域の「作物結合型」とし「n部門結合型」と特徴づけた。さらにその型の内容として作物名をn個あげるのである。

均等分布と実際との差は、両者の（各作物ごとの）偏差の二乗和平均、つまり分散と同じものによって測定される。

このウィバーの方法には次のような問題がある。

ウィバーが特定市町村の理論分布について前提としていることは、「作物間の均等分布」ということだけであって、具体的な作物の結合型についてのモデルはない。

その市町村に実際栽培されている作物が個別にとりあげられ、したがって、作物の結合は市町村ごとに異なっている。ウィバーの方法で判別できるのは、特定市町村の結合が「n個作物結合」ということだけであって、「その作物が何か」ということについての判別モデルはないといってよいであろう。だが、実際の適用では結合される作物種類のタイプ分類までやっているし、実際にただ結合個数の型がわかっただけではあまり有用ではない。

ウィバーの方法による計算を形式化してみると次の二つのステップがある。（農業部門別粗生産額データを例にとる。）

(a) 前準備。その市町村の部門別粗生産額構成比率を、大きいものから順にならべて順序集合Pをつくる。五部門あるとすると、

$P=(P_1, P_2, P_3, P_4, P_5 | P_1 \vee P_2 \vee P_3 \vee P_4 \vee P_5)$ (P_i は i 部門の構成比率)。

(b) このPから次の五つの部分集合をつくる。

$P_1=(P_1)$, $P_2=(P_1, P_2)$, $P_3=(P_1, P_2, P_3)$, $P_4=(P_1, P_2, P_3, P_4)$, $P_5=(P_1, P_2, P_3, P_4, P_5)$

第一表の理論分布を五つの部分集合に対応させて、偏差二乗和平均が最小のものをその市町村の部門結合型とするのである。

ウィバーの方法によってきまるのは、右の(b)のステップであ

第1表　5部門結合の場合の理論分布

部門結合型	理論分布
1部門結合	{100.0}
2部門結合	{50.0, 50.0}
3部門結合	{33.3, 33.3, 33.3}
4部門結合	{25.0, 25.0, 25.0, 25 0}
5部門結合	{20.0, 20.0, 20.0, 20.0, 20.0}

第2表　ウィバーの方法による市町村農業部門結合型
（佐賀県・昭和35年，40年）

A　昭和40年

	市町村数	割合
		%
1部門結合	9	18
2部門結合	5	10
3部門結合	1	2
4部門結合	—	—
5部門結合	34	69
計	49	100

B　昭和35年から40年への移動

		40年					
		1部門結合	2部門結合	3部門結合	4部門結合	5部門結合	計
35年	1部門結合	5					5
	2部門結合		2				2
	3部門結合						—
	4部門結合			1		1	2
	5部門結合	4	3			33	40
	計	9	5	1	—	34	47

って、(a)のステップは、ウィバーの方法に関係なく事前にきまる。たとえば、ある市町村が「三部門結合」ときまったとしても、実際の結合型が「米・麦・野菜」であるか「果樹・米・いも」であるかの判別については何のモデルもない。部門別構成比率の大きさの順序集合をつくるという操作によって、事前にその市町村固有の状態できまってしまう。ウィバーの操作によりモデルとの関係できまるのは、その順序集合の要素の何番目までを採用するか、ということだけである。

したがって、地域特化の著しいモノカルチュア的な傾向が強い地域では、主要な部門がウィバーの操作によって指定されるからこの方法は有用であろう。その場合には「モノカルチュアの理論分布が一作物一〇〇%」というモデルも生きてくる。しかし、日本のようにどの市町村をとっても多数の作物あるいは多数の農業部門が混在している場合には、多部門結合型が多くなって、地域の部門結合型を抽象化してあらわすという作用はあまり期待できない。

実際に、西日本としては比較的モノカルチュア的市町村を多く含む佐賀県でも、第二表のように七割

が最多部門結合（この場合五部門までをとった）となってしまう。しかも第一位部門のわずかな%差によって「五部門結合」から「一部門結合」に変わってしまう。第二表のBにみられるように三五年に五市町村であった「一部門結合」が四〇年には九市町村にふえているが、これは米部門のわずかな%の変化によって生じているのである。この変動は分類の不安定を示していると思う。

注（一） Maurice H. Yeats, *An Introduction to Quantative analysis in Economic Geography*, 1968, pp. 33～40.

（二） *ibid.*, pp. 33～40.

(3) 部門間相互依存関係の情報を利用する考え方

同じデータ（農業部門別粗生産額構成比率）を使って別な接近をするとすれば、どんな方法が考えられるだろうか。

一つの考え方は、各部門をそれぞれ一つの変量と考え、その変量間に体系的な内部依存関係があると仮定して、この依存関係の情報を利用することである。内部依存因子をとり出すことによって変量の数よりもすくない合成変量に変換し、地域農業部門結合の状態に関する新しい情報があたえられる。

特定地域の部門結合状態を原データよりも抽象度の高い情報によって記述するのには二つの方法がある。

第一は、外部基準をもうけこれを尺度として状態の位置づけ、あるいは状態が分類されるクラスの判別をする方法である。

第二は、市町村が属する県の内部だけで市町村相互間の相対的な位置づけを考える方法がある。各市町村についてあたえられた変量間相互依存関係の情報を利用して、内部基準として使用できる合成変量をつくり、状態の位置づけをするのである。この第二の方法は主成分分析によって行なわれる。この報告で検討しようとするのもそれである。(1)

主成分分析が部門結合の状態を得るのに適当であるのは次の理由によると考えられる。

(1) 変量間の相互依存関係は変量間の単純相関係数であらわせる。この相関係数行列Rによってあたえられる情報を利用して、Rを変換作用素とする$R[x]=\lambda[x]$を考える。その固有ベクトルを求めると、いまのべた結合状態を得るための係数ベクトルがえられる。係数ベクトルを使って各主成分の各市町村に関するスコアが得られる。（各主成分に対応する合成変量のスコアがえられる。）

主成分はもとのデータの情報を抽象度の高い情報に要約できるのがふつうである。

(2) 各主成分は互いに直交していて、その分散が大きい順に

《ノート》 主成分分析における結果解釈過程の構造

えられるから、部門別構成比率の市町村間変動を説明するには、もとのデータを使うよりも自然な座標軸を形成できる。

(3) また固有値の大きい方からとればもとのデータの変量内分散の一定割合（たとえば七〇%）を説明する少数の主成分をえられることができ、よりすくない合成変量によって状態の型を説明できる。

(4) こうして選ばれた主成分に「意味づけ」をすれば、各市町村は一定の意味空間に位置づけられるのである。

ただしデータとして構成比率を使うことには問題がないわけではない。[2] その理由は、データが観察単位ごとに閉鎖的な数体系であたえられている（合計が必ず一〇〇になる）ために、変量間の本来の体系的な相互依存関係以外に形式的な逆相関関係が発生してそれが主成分に反映する可能性がある。構成比率をとったのは市町村間の規模による変動を除去するためであるが、別な問題が変量の共変動に入ってくるのである。今回はこの問題について充分な検討はしていない。ここでは、一一部門の構成比率の中から一〇部門をとり、一部門をはずしてみた。しかしこの操作によって解決できたわけではない。

注 (一) Leslie J. King, *Statistical Analysis in Geography* pp. 174～182 には、やはりウィバーが試みた主成分分析の適用例が紹介されている。

(二) *ibid.*, p. 179.

二、主成分の解釈

(1) 主成分の解釈過程モデル

主成分分析結果利用上の一つの問題は、各主成分の意味づけあるいは解釈が分析者の主観に依存する部分が大きい、ということである。

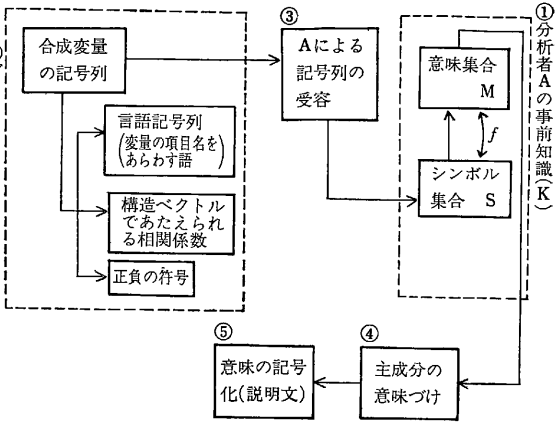
心理学の能力テストや臨床医学における徴候群の場合には、対象領域が限定されている。だから変量が示す性質と状態の型について分析者の間で合意に達しやすい共通の事前知識があることが多いであろう。そこでは主成分分析の結果の意味づけについて人によって大きな差は生じないかもしれない。この分析手法が心理学を中心に発達した理由はそこにあると思う。

社会、経済現象を対象とする場合、対象領域を広くとればとるほど主成分の意味づけは多様な解釈をゆるすことになる。

意味づけはもっぱら分析者が対象についてもっている事前知識にもとづいて行なわれるから、分析者が事実について偏った知識をもっていれば解釈も偏ってくる。

この問題を考えるために、分析者が主成分の意味づけを行なう過程をモデル化して考えてみよう（第一図）。

第1図　主成分の解釈過程モデル



注．①の f は事前にある対応．

① 分析者Aの内部には、対象に関する事前知識（K）があらかじめ存在する。

この事前知識はAの内部で、シンボル集合Sの要素（シンボルSi）と意味集合Mの要素（mi）との特定の対応（f）という形で存在する。

たとえば、分析者Aは養豚の地域分布の性質について、「養豚は都市近郊型農業である」という知識を重視し、「養豚はまた遠隔地でも主産地を形成することがある」という知識は重視していなかったとしよう。

この場合に、分析者Aの事前知識は「養豚」というシンボルと「近郊型農業」という意味内容との対応に限られてくる。

② このような事前知識をもった分析者Aが、都市近郊から遠隔地までを含む広い範囲の市町村の統計データについて主成分分析を行ない、その結果得た第 i 成分の構造ベクトルが養豚に関する変量について正の高い相関をもっていたとしよう。

Aが主成分分析によって得る情報の直接の媒体となるのは、次の三つの記号列である。

（i）第 i 成分の構造ベクトル[1]。その中の大きい絶対値をもつ相関係数。

（ii）相関係数の正負符号。

（iii）その相関係数に対応する変量の項目をあらわす語。

右の（i）と（ii）は数値情報として、（iii）は言語情報としてあたえられる。（係数ベクトルがあたえられている場合でも固有値を使って構造ベクトルに変換できるので、ここでは意味解釈をしやすい相関係数の形で考えている。）

③ この三つのデータはシンボルの形でAに伝達され受容される。第一図の③→①の過程がこれである。

④ Aは受容したシンボルによってAの内部の事前知識（あらかじめAがもっているS_iとm_jとの対応fの知識）を仲立ちとして、意味解釈を行なう。

右の例であれば、Aは1成分が養豚に関する変量と正の高い相関をもつというシンボル群を受けとったので、1成分は「近郊型養豚地帯の性格を示す」状態像をあたえるもの、と解釈するであろう。

⑤ そして、i成分について高い正のスコアを得た市町村に「近郊型農業地域」という意味をあたえるかもしれない。

しかし、この意味づけは一部の市町村については正当であっても、他の一部についてはおそらく誤りとなるだろう。なぜならば、養豚には近郊型のほかに遠隔地型があるからである（佐賀県でも、養豚の特化係数はこの二つの型の両方に高く出てい

佐賀県市町村別農業部門構成比率・昭和40年）

P_5	P_6	P_7	P_8	P_9	P_{10}
0.6957	0.5456	0.4958	0.4055	0.2238	0.0053
7.0	5.5	5.0	4.1	2.2	0.1
83.2	88.7	93.7	97.7	99.9	100.0
0.0267	0.0819	−0.1362	−0.0687	−0 1788	0.7097
0.2927	−0.3671	0.3912	0.3790	0.3586	0.1000
−0 0048	−0 2398	−0.0325	0.2713	−0.7811	0.0858
−0.4803	−0 3293	0.0608	0.0499	0 2724	0.1463
0 1202	−0.1668	0.0672	0.0900	0.0817	0.6071
0.1670	−0.2745	−0.4188	−0.6215	0.2089	0.1305
−0.3089	0.2225	0.0667	0.2595	0.2431	0.1512
−0.3045	0.4596	0.5156	−0.3726	−0 0855	0 1924
0.6716	0.3016	0.2649	−0.0586	0.0359	0.0436
0 0368	0.4876	−0.5507	0.4140	0 1928	0.1001

る)。

注(1)　ここでの主成分分析に関する用語は主に次の著書のどちらかによっている。

(1)　芝祐順『行動科学における相関分析法』。

(2)　伊藤孝一『多変量解析の理論』。

(2)　主成分解釈過程の論理構造分析

このような主成分解釈過程で生じる偏りは、(1)でのべた解釈モデルに則していえば（第一図）①の分析者Aの事前知識Kが、分析対象について偏りをもっていたからである。これが分析者の「主観的意味づけ」といわれるものの実質的内容といってよいであろう。

主成分分析は、その導出過程で(a)理論的には行列の固有値問題という数学が関係し、(b)計算についてはコンピューターで行なわれるので、この方法になじみの薄い人にはデータ処理過程で自動的に意味づけが行なわれる、という誤解をあたえやすいように思える。そのために分析者の意味づけがそのまま受容される傾向がある。もし分析者の意味づけが偏っていれば、分類や意味づけにもとづく推論も偏ってくる。わたくしが読んだ範囲の論文にも、主成分分析結果の意味づけについて、分析者のあたえた意味に直ちに同意してよいかどうか疑問のものがある

第3表　主成分の固有値と固有ベクトル（データ・

		P_1	P_2	P_3	P_4
	固有値	3.5061	2.0510	1.3140	0.7574
	固有値の構成比（%）	35.1	20.5	13.1	7.6
	同累積値（%）	35.1	55.6	68.7	76 3
固有ベクトル	米 X_1	−0.3744	−0.4191	0.2468	0 2386
	麦類 X_2	0 1837	−0.5201	−0.0296	−0.1979
	雑穀・豆・いも X_3	0.4670	0.0119	0.1495	−0.0989
	野菜 X_4	0.3371	0.1272	0.0847	0.6497
	果樹 X_5	−0.0280	0.5734	−0.4378	−0 2204
	その他耕種農 X_6	0.4054	−0.0778	0.1837	−0.2648
	酪 X_7	0.0299	0.2402	0.6775	−0.4287
	にわとり X_8	0.3594	−0.2102	−0.2361	−0.1338
	牛肉 X_9	0.2397	0.2750	0.3163	0.3848
	豚 X_{10}	0.3764	−0.1587	−0.2616	0.0603

《ノート》　主成分分析における結果解釈過程の構造

ように思える。

分析者の主成分に対する意味づけが主観的になることを防ぐ方法はなんだろうか。

主成分の意味づけは第一図の②でのべた三つのデータだけに依存して行なわれるが、この三つのデータによって得られる情報はかなり抽象度の高い、しかもあいまいさを含んだ情報である。ということに注意を向ける必要があると思う。このような性質の情報にもとづいて分析者は意味づけをするが、その「意味づけ」は実は分析者が提出した「仮説」にすぎない。

ふつうの推論であれば、提起した仮説は検証によって推論が妥当であるかどうかを確かめる。しかし、主成分分析で行なわれた意味づけは、しばしば「事実」あるいは「検証ずみの仮説」であるかのようにとりあつかわれる。その意味づけにもとづいて推論が行なわれ、推論もまた検証ずみであるかのようにあつかわれることがすくなくない。

意味づけが主観的になるのを防ぐ方法は、分析者が「意味づけ」の過程で、あるいは意味づけのあとで、検証を行ないその過程を明示することではないだろうか。

問題を明確にするために、少しトリビアルになるが実際の解釈過程を前述のモデルにもとづいて分析してみよう。

第三表は一の(3)でのべたデータによって計算した主成分の固

第4表　第4主成分までの構成ベクトル

		主成分			
		P_1	P_2	P_3	P_4
米	X_1	**−0.701**	−0.600	0.283	0.208
麦類	X_2	0.344	−0.745	−0.034	−0.172
雑穀・豆・いも	X_3	**0.874**	0.017	0.171	−0.086
野菜	X_4	**0.631**	0 182	0.097	**0 565**
果樹	X_5	−0.053	**0.821**	**−0.502**	−0.192
その他新種穀	X_6	**0.759**	−0 111	0.211	−0.230
酪農	X_7	0.056	0.344	**0.777**	−0.373
にわとり	X_8	**0.673**	−0.301	−0.271	−0.116
肉牛	X_9	**0.449**	0 394	0.363	0.335
豚	X_{10}	**0.705**	−0.227	−0.300	0.053

注　各主成分と各変数との相関係数ベクトルである．
太字は 0.400 以上のもの

有値と固有ベクトルである。

また第四主成分（全分散の七六％まで説明する）までを構造ベクトルの形に直したものが第四表である。

第一主成分の解釈をとりあげてみよう。（○の中の番号は第一図の番号に対応するため、②から始まる。）

③　まず、あたえられた情報を三つの要素に分解して考える。（第二図および第一図参照）

これを第一図のモデルと対応させて書くと、次のようになる。

第二図の③の分析		第一図のモデルの②
（ｉ）	相関係数の絶対値の大きい	構造ベクトルの相関係
	七個の変量に注目する。	数
（ii）	X_1は負、他は正であること	相関係数の正負
	を判別する。	
（iii）	七個の変量の項目名によっ	言語記号列
	て部門を判別する。	

④の1　この三つの情報によって、まず意味づけのための基本操作を行なう。それはわたくし（分析者A）がもっている事前知識によって、それぞれの変量の部門名からその部門の〝より一般的な意味〟を抽出することである。この場合に一般的な意味は、いくつかの変量に共通な性質を抽出して次のステップに橋わたしできるようなものでなければいけない。（第二図で、①～⑭↑⑧とあるのは、第一図のモデルで示した分析者が自分の事前知識と、受容したシンボルとを対応させる操作を示す。）

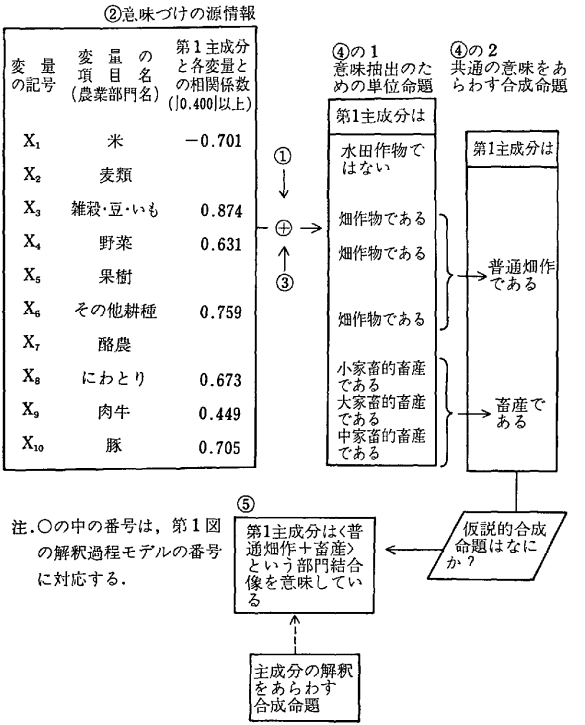
このようにして抽出した個々の変量の一般的な意味は、わたくし（分析者）の判断であって、記号列によって表現され単位命題の形をとる（モデルの⑤）。単位命題は相関係数の符号によって肯定文または否定文（X_1の場合）の形をとる。

④の2　単位命題から、その共通要素をまとめていくつかの合成命題をつくる。この段階で明示的に残るのはふつう肯定命題（相関係数正のもの）であって、否定命題（相関係数負のもの）は参考としてとりあげられることが多いようであるが、人によっては積極的に、明示的に意味づけに使うこともある。否定命題が主成分の意味づけにどのような役割をもつべきかということについて、わたくしの読んだ文献の範囲では何もふれていない。しかし、あとで実際例でみるように否定命題を明示的に意味づけにとり入れることは、誤りをおかす場合もあるように思える。

例では第一主成分は次の二つの合成命題に要約される。合成は論理積をとることによって行なわれる。

(ｲ)　「第一主成分は普通畑作という部門結合状態の性格を意

第2図　第1主成分の解釈過程



注．○の中の番号は，第1図の解釈過程モデルの番号に対応する．

味内容の一つとしてもっている」‥「普通畑作である」と表現する。

(ロ)「第一主成分は畜産という部門結合状態の性格を意味内容の一つとしてもっている。」……「畜産である」と表現する。

⑤ このように正式化した合成命題をさらに合成して、第一主成分の意味について一つの仮説を提示する。

それは「第一主成分は『普通畑作㊉畜産』という部門結合型を意味している」という合成命題であって、これが第一主成分についてわたしくの解釈を表現したものである。

以上の解釈は、とりあげた変量がすくなく、しかも同じ次元に属する属性区分の各カテゴリーを変量にとりあげたので、あまり異論がなく自明のものと考えられるかもしれない。しかし果たして充分自明なものであろうか。よく考えてみるとあいまいな点がある。

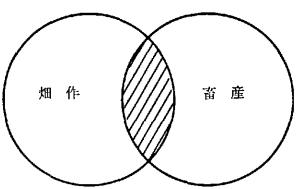
第一に、部門結合型は「普通畑作㊉畜産」を意味するというが、この演算記号㊉は何を意味するのか。第一主成分で高い正のスコアを得た市町村は、(a)「必ず普通畑作と畜産がともに優位を占める」という意味なのか。それとも、その市町村では(b)「普通畑作または畜産のどちらか、あるいは両者がともに優位を占める」という意味なのか。

(a)と考えればそれは二つの命題(「普通畑作である」と「畜産である」)の論理積であって第三図の(a)の斜線の部分にあたる。(b)と考えれば論理和であって第三図の(b)の全領域にあたる。

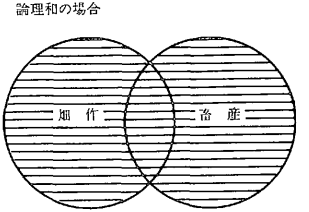
わたくしが㊉と表示したのは何らかの命題合成が行なわれる論理演算の存在を示しただけであって、加法的か乗法的かの判

第3図　合成命題の2つの場合

(a)　論理積の場合



(b)　論理和の場合



断はしていない。

第二に、最初の段階（④の2）で「普通畑作」「畜産」という二つの命題を合成するときには論理積をとる演算を行なった。これは各変量の性質の共通性を抽出する必要があったからである。そのため、そこでとりあげられなかった変量、たとえば「畜産」の場合の酪農（X_7）は「畜産」の中に入っていない。つまりここでいう「畜産」は本来の畜産という概念を局延してはいないのである。畜産の部分概念にすぎない。だから人によってはこれを不正確と考えるかもしれない。しかし、正確に表現するとすれば「肉牛・中小家畜」ということになり列挙的になってしまって情報の要約化から遠くなる。どの程度の論理的正確さを必要とするか、という問題が出てくる。

第三のあいまいさは、否定命題のとりあつかいの部分に残る。「普通畑作㊥畜産」というとき、否定命題としての「水田作ではない」はどう考えたらよいのであろうか。第一主成分のスコアの大小は水田作についてどういう解釈ができるだろうか。

これらの問題は判断の論理形式だけから決定はできない。もとのデータあるいは仮説としてのべた命題と関連のありそうな情報がえられる別な系列のデータにもどって経験的に吟味しなければならないことである。

クラス別特化係数——昭和40年——

野　菜	果　樹	その他の耕種部門	乳用牛	にわとり	肉　牛	豚
%	%	%	%	%	%	%
0.9	0.0	2.1	0.6	2.3	0.2	1.7
2.8	9.5	1.9	2 6	4.0	0.3	1.5
5 5	10.0	2.0	3.1	5.5	1.7	1.5
8.4	10.4	2.9	1 6	4.9	2.1	2.7
6.8	16.9	5.7	1 9	8.5	2.8	3.3
10.6	9.2	6.3	4.2	3.4	3.7	3.5
11.8	6.1	9.7	6.1	17.9	4.1	5.3
5.3	10.0	2.7	2 7	5.2	1.5	2.0
0.17	0.00	0.78	0.22	0.44	0.13	0.85
0.53	0.95	0.70	0.96	0 77	0.20	0.75
1.04	1.00	0.74	1.15	1.06	1.13	0.75
1 58	1.05	1.07	0.59	0.94	1.40	1.35
1.28	1.69	2.11	0.70	1.63	1.87	1.65
2.00	0.92	2.33	1.56	0.65	2.47	1.75
2.23	0.61	3.59	2.26	3.44	2.73	2.65

主成分の解釈についてわたしくが提出した解釈は、一度具体的なデータ群にフィード・バックしてその意味づけを吟味しない限り、主成分解釈のあいまいさは除去できない。いまのべた仮説は命題そのものとしては明確な論理操作によって合成されたようにみえるが、命題と経験的事象との対応をあきらかにしなければあいまいさをのぞけないのである。

(3) 経験的データへのフィード・バック

(α) 吟味のフレーム

主成分についてその性質をのべた仮説の意味内容をMとしよう。ここでMが経験的事実に適合するかどうかはまだ検証されていない。なぜならばMは、分析者としてのわたくしがもっている事前知識から推論したものであって、わたくしの事前知識が検証されない限り主成分の意味づけもまた検証されないからである。

さらに事前知識が正しかったとしても、主成分によってあたえられたシンボル集合と事前知識との対応が正しいかどうか、という問題もあろう。

わたくしの考えでは、この検証の問題は次の二つの明確に定義できる問題に解消できると思う。

第一はこの報告の最初に提起した「判断基準」を適用するこ

第5表　第1合成変量階級別・農業粗生産額構成比率と

	第1合成変量（標準化値）	市町村数	米	麦類	雑穀・豆・いも
			%	%	%
構成比率	～−1.000	2	83.5	6.8	1.0
	−0.999～−0.500	13	67.3	6.7	1.0
	−0.499～ 0.000	18	61.0	6.2	2.0
	0.001～ 0.500	8	56.1	7.0	2.8
	0.501～ 1.000	3	43.6	6.0	3.5
	1.000～ 2.000	3	42.0	7.1	9.1
	2.000～	2	19.7	9.7	9.4
	合　計		60.2	6.5	2.3
クラス別特化係数	～−1.000		1.39	1.05	0.43
	−0.999～−0.500		1.12	1.03	0.43
	−0.499～ 0.000		1.01	0.95	0.87
	0.001～ 0.500		0.93	1.08	1.22
	0.501～ 1.000		0.72	0.92	1.52
	1.000～ 2.000		0.70	1.09	3.96
	2.000～		0.33	1.49	4.09

と、第二はこの判断基準を別なデータによって追加的に吟味すること、の二つである。

第一の判断基準とは、主成分分析の結果えた情報は、

(a) もとのデータ群よりも地域の農業部門の状態をうまく要約して表現しているか、

(b) 原データの情報と矛盾しないか、

という二つの基準によって吟味することを指している。

第二の吟味は、原データとは別な、しかも主成分のあたえる(と仮定した)情報の意味に関連したデータによって、いまのべた判断基準が経験的に成り立つことを追加的に検討することである。

それには、各市町村の第一主成分のスコアによる相対的な位置が、その主成分のMという意味内容に関して正しく表現されているかどうかを吟味すればよい。分析に使った原データと、Mと関連のありそうな別なデータの両面から吟味をするのである。

(β) 第一主成分の吟味

第一合成変量(第一主成分の市町村別スコア)を標準化してクラス分けをし、各クラスに入る市町村の農業部門別構成比率が、

(a) 「普通畑作⊕畜産」という意味づけからみて妥当かどうか、

(b) 佐賀で中心的な部門である米がどういう位置づけになっているか、をみた。

(a)は肯定単位命題の合成命題にもとづく仮説が原データと矛盾しないかどうか、(b)は否定単位命題は主成分の意味づけにどういう役割をもつか、を経験的に検討するための設問である。

変量として採用した一〇部門のうち「普通畑作」としてとりあげられるものは「雑穀・豆・いも」「野菜」「その他の耕種」である。「麦類」は佐賀では水田裏作と本来の畑作と両方あって、前者のウェイトが高いためその性質は畑作であるが除いて考えた。

果樹は「普通畑作」とは異なるカテゴリーに属するものと考えた。

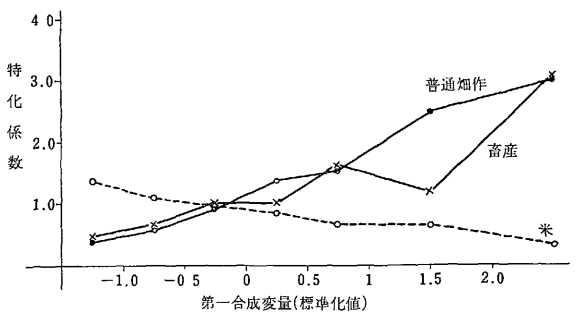
したがって「普通畑作」についてはいまあげた三部門だけを考えればよい。この三部門の構成比率が、第一主成分のスコア階級の低いものほど小さく、スコア階級が高くなる(第五表では下へいく)ほど大きくなっていればよいわけである。なお、判別の便のために第五表の下に県レベルの構成比率を一・〇とする特化係数を出しておいた。

第五表にみられるように、第一主成分のスコア階層はプラス

第6表　第5表を米作・畑作・畜産にまとめたもの

	第1合成変量	米	雑穀・豆・いも・野菜・その他の耕種部門	畜産
		%	%	%
構成比率	～−1.000	83.5	4.0	4.2
	−0.999～−0.500	67.3	5.7	5.8
	−0.499～ 0.000	61.0	9.5	8.7
	0.001～ 0.500	56.1	14 1	9 7
	0.501～ 1.000	43.6	16.0	14.6
	1.000～ 2.000	42 0	26 0	10 6
	2.000～	19.7	30.9	27.3
	計	60.2	10.3	8.7
特化係数	～−1.000	1.39	0.39	0.48
	−0.999～−0.500	1.12	0.55	0.67
	−0.499～ 0 000	1.01	0.92	1.00
	0.001～ 0.500	0 93	1.37	1.06
	0.501～ 1 000	0.72	1.55	1.68
	1.000～ 2.000	0 70	2.52	1.22
	2.000～	0.33	3.00	3.14

第4図　第1合成変量の各階級の特化係数



の値の大きいものほど畑作三部門の特化係数が大きい。したがって「普通畑作」という意味づけはクラス別データに関する限り原データの情報と矛盾せず、また原データの情報をより高い抽象度で要約している。

「畜産」は畜産各部門（酪農は除いてある）の個別よりも、第六表のように一本にまとめてみると、普通畑作とほぼ似た傾向を示している。これは第四図にもみられる。しかし同時に「普通畑作」と全く同一のパターンで動いてはいない。このことはさきに二・(2)で提起した問題、すなわち「普通畑作㊉畜産」の㊉は論理和か論理積かという問題に一つの解答を示唆すると思う。すなわち、それは論理和と考えた方が無難である、ということである。

実際に市町村の第一主成分スコア順位によって、個別に検討してみると、「普通畑作」の構成比率の大いさと「畜産」の構成比率の大いさとは傾向として正の相関関係にあるが、必ず結合しているというものではない。したがって今後、第一主成分の意味は、論理和と考え「普通畑作＋畜産」というように表現しよう。

以上によって第一主成分の意味を、「普通畑作＋畜産」とすることは、クラス分けしたデータに関する限り原データと矛盾しない。

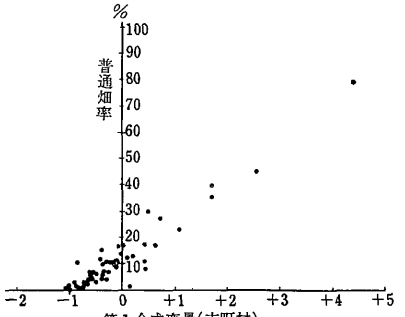
第7表　第1合成変量階級区分別土地利用

	第1合成変量（標準化値）	市町村数	経営耕地の種類別構成比			
			田	普通畑	果樹園	その他の樹園地
			%	%	%	%
構成比率	～−1.000	2	99.1	0.9	—	0.0
	−0.999～−0.500	13	87.6	3.5	8.7	0.2
	−0.499～ 0.000	18	76.3	10.1	13.1	0.5
	0.001～ 0.500	8	74.2	13.0	11.8	1.0
	0.501～ 1.000	3	58.2	20.7	16.0	5.1
	1.000～ 2.000	3	53.5	34.0	12.1	0.4
	2.000～	2	33.5	52.1	14.3	0.1
	計	49				
特化係数	～−1.000		1.31	0.07	0 00	
	−0.999～−0.500		1.14	0.31	0.76	
	−0.499～ 0.000		1.00	0.90	1 15	
	0.001～ 0.500		0.98	1.16	1.03	
	0.501～ 1.000		0.76	1.84	1.39	
	1.000～ 2.000		0.70	3.04	1.08	
	2.000～		0.44	4.60	1.25	
	計		1.00	1.00	1.00	

これを原データとは別の系列に属する情報によってみるとどうなるであろうか。

第七表は、第一主成分の各クラスの土地利用構成をみたものである。また第五図は各市町村の第一合成変量スコアと畑地率（普通畑／耕地）との散布図である。

第5図　各市町村の第一合成変量と普通畑率



第一主成分の一方の意味を「普通畑作」とすることは、このデータによって全く妥当であることが明らかである。

家畜の方は、実物では比較しにくいので次の方法をとって比較してみた。まず、各クラスの家畜頭数を家畜単位によって「家畜単位換算頭数」を出し、そのクラス間分布を求めた。これを各クラスの耕地面積のクラス間分布と比較して家畜のクラス間特化係数を出してみたのが第八表である。これをみると（特化係数の計算の仕方は全く異なるが）第6表の畜産部門の状態ときわめて良く似た分布をしている。

このように第一主成分の意味を「普通畑作＋畜産」とすることは、これを論理和と考える限り、原データにも矛盾しないし、原データ以外の関連データとも矛盾しない。

では、否定単位命題としてあたえられる「水田作ではない」はどうか。

第六表、第四図、第七表をみると、クラス分けをしたデータに関する限り、普通畑作や畜産と逆の関係にあって、この否定命題は経験的データと矛盾しないようにみえる。しかし市町村ごとに当たってみると、一部にこの否定命題をそのまま真としてうけ入れるわけにはいかないことが分かるのである。

いま、第一合成変量のクラスの中から、−0.9～−0.5に入る一三市町村をとり出して、各市町村ごとに「米」と「果樹」の

第8表　家畜頭数の第1合成変量階級別分布

	第1合成変量（標準化値）	家畜単位換算頭数	構成比率	耕地面積のクラス間分布	耕地分布と比較した家畜頭数の特化係数
			%	%	%
構成比率	〜−1.000	593	1.1	2.9	0.38
	−0.999〜−0.500	9,730	17.9	29.8	0.60
	−0.499〜　0.000	23,074	42.5	38.4	1.11
	0.001〜　0.500	7,195	13.2	12.9	1.02
	0.500〜　1.000	7,309	13.4	8.7	1.54
	1.000〜　2.000	3,929	7.2	5.3	1.36
	2.000〜	2,459	4.6	2.0	2.30
	計	54,329	100.0	100.0	

構成比率をみると第九表のようになる。すなわちはじめの一〇市町村とあとの三町村とは全く異質である。

このクラス（−0.9〜−0.5）は、第一主成分の意味づけに関してネガティブな意味内容をかなり強くもつわけである。したがって、もし「水田作ではない」という否定命題を第一主成分の意味の中に入れるとすれば、このクラスでは「否定の否定」すなわち「水田作である」という肯定命題の意味が強くなる。

第九表の上位一〇市町村は「米」の構成比率が七〇％以上であるから、この意味づけは妥当である。下位三町村には、しかしこの意味づけは全く妥当しない。

他方、第一主成分の肯定的な意味、すなわち「普通畑作＋畜産」はこのクラスでは否定的な意味をもつことになるが、それは第九表にみられるようにどの市町村についても妥当である。

「米」部門にみられる異常な要素の混入は、主としてこの三町村で「果樹」の構成比率が著しく高いことに原因する。

なお、第一主成分の固有ベクトルから変量X_1（米）とX_5（果樹）の係数をとり出して、第九表の一三町村のうちの九町村（計算の簡略化のため上位一〇市町村については、うち六町村のみをとった）について「米」「果樹」の第一合成変量推定値を計算すると第一〇表(a)欄、(c)欄のようになる。これを原データの標準化値（(b)欄・(d)欄）と比較すると、下位三町村では推

第9表　第1合成変量の−0.9〜−0.5クラスの市町村の原データ

市町村番号	米	果樹	普通畑作	畜産
	%	%	%	%
1	73.3	1.0	5.9	9.3
6	73.6	—	4.2	3 4
7	77.5	0.0	4.6	4.9
8	72.8	0.0	2.3	6.7
13	75.6	0.0	5.9	3.3
17	75.6	1.3	9.0	3.8
21	70.8	0.0	8.1	3 3
24	73.2	3.9	5.1	6.8
30	73.4	4.7	5.8	7.7
33	73.2	2.1	10.1	4.1
40	**21.2**	**68.9**	3.7	4.3
41	**28.1**	**55.7**	5.7	4.7
42	**29.9**	**55.2**	7.3	2.9

第10表　2つのグループ市町村の変量推定値と実際値

市町村番号	第1合成変量（標準化値）	第1主成分の係数による推定値と原標準化データとの差					
		第1主成分の係数によって推定した，米の推定標準値 (a)	米のもとの標準化した値 (b)	(a)−(b)	第1主成分の果樹の係数によって推定した，果樹の推定標準値 (c)	果樹のもとの標準化した値 (d)	(c)−(d)
13	−0.863	0.59861	1.05338	−0.45477	0.04484	−0.69334	0.73818
17	−0.892	0 61896	1.05765	−0.43869	0.04637	−0.60931	0.65568
21	−0.565	0.39210	0.77320	−0 38110	0.02937	−0.69334	0 72271
24	−0 633	0.43920	0.91295	−0.47375	0.03290	−0.43463	0.46753
30	−0.588	0.40797	0.92245	−0.51448	0.03056	−0.38805	0.41861
33	−0.680	0.47173	0.91599	−0.44426	0.03534	−0.55349	0.58883
40	−0.690	0.47867	−2.14323	**2.62192**	0.03586	3.80109	**−3.76523**
41	−0 520	0.36063	−1.73756	**2.09819**	0.02702	2.95461	**−2.92759**
42	−0.600	0.41653	−1.63039	**2.04692**	0.03120	2.92213	**−2.89093**

農業粗生産額構成比率

野菜	果樹	その他の耕種部門	乳用牛	にわとり	肉牛	豚
%	%	%	%	%	%	%
1.5	0.0	2.9	1.6	4.5	0.2	1.1
3 6	2.2	2.1	3.1	6.2	0 9	2.1
7.0	6.8	2.4	1.8	4.4	0 8	2.1
7.0	9.6	3.8	3.6	3.8	2.4	2.3
8.0	18.2	4.3	2.8	7.6	3.1	3.0
7.4	21.6	1.3	3.7	3.6	3.0	0.5
3.2	63.2	1.0	1 6	2.8	1.1	0.1
	0.00					
	0.22					
	0.68					
	0.96					
	1.82					
	2.16					
	6.32					

定値と原データの間に大きな差があることが分かる。

以上のように個別市町村についてみると、第一合成変量によってあたえられる情報は否定命題つまり高い負の相関係数をもつ変量項目を基礎に推論した意味づけは、それだけでは不完全であって別な情報で補う必要がある。（ただしこのデータに限っているのであって、どのような条件の場合に一般化できるかについては未検討である。）

（ア）第二主成分の吟味

そこで第四表にもどって第二主成分をみると、最も高い正の相関をもつものは「果樹」である。しかも正の相関係数〇・四以上のものは果樹しかない。そこで第二主成分を「果樹」と意味づけることが正しければ、第二合成変量を第一合成変量の補足情報として使えば、いま(ア)でのべたような欠点は補正できるのではないか、と考えられる。

まず、第二主成分の意味づけの吟味のために原データへもどってみると（第一一表）、「果樹」という意味づけの妥当性が確認できる。また原データ以外の系列に属する土地利用データについても（第一二表）同じである。

さらに、第一〇表でとりあげた九市町村について第二主成分の変量X_2（果樹）に対応する係数を使って果樹の推定値を出し実際値と比較してみると（第一三表）、果樹グループ三町村で

第11表　第2合成変量階級別

	第2合成変量（標準化値）	市町村数	米	麦類	雑穀・豆・いも
			%	%	%
構成比率	～−1.000	7	74.8	8.7	1.1
	−0.999～−0.500	13	69.1	7.5	1.6
	−0.499～ 0.000	8	63.5	6.5	1.9
	0.001～ 0.500	5	56.9	5.7	4.0
	0.501～ 1.000	9	42.3	5.8	4.3
	1.000～ 2.000	4	50.7	4.6	2 1
	2.000～	3	24.4	1.6	0.6
三部門の特化係数	～−1.000		1.24	1.34	
	−0.999～−0.500		1.15	1.15	
	−0.499～ 0.000		1.05	1.00	
	0.001～ 0.500		0.95	0.88	
	0.501～ 1.000		0.70	0.89	
	1.000～ 2.000		0.84	0.71	
	2.000～		0.41	0.25	

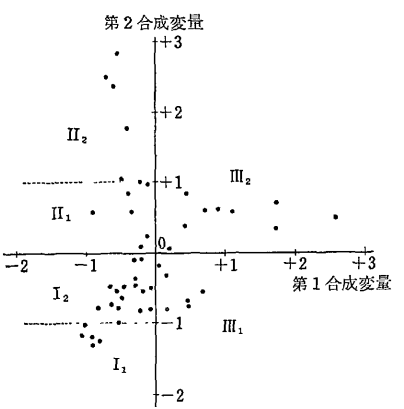
第12表　第2合成変量階級別土地利用

	第2合成変量（標準化値）	田	普通畑	果樹園	茶園・桑園・その他耕地	耕地合計
		%	%	%	%	%
構成比率	～−1.000	96.0	3.8	0.1	0.0	100.0
	−0.999～−0 500	86.5	7.1	4.7	1 6	100.0
	−0.499～ 0.000	84.3	8.1	7.1	0 4	100.0
	0.001～ 0 500	67.8	19 0	12.5	0.7	100.0
	0.501～ 1.000	59.0	21 6	18.8	0 6	100.0
	1.000～ 2.000	64.3	10 2	24.4	1.1	100.0
	2 000～	43.9	5.0	50.4	0.7	100.0

第13表　第10表の市町村の，第2主成分によるX_3推定値と実際値

市町村番号	第2合成変量（標準化値）	第2主成分の果樹の係数によって推定した果樹の推定標準値(a)	果樹のもとの標準化値 (b)	(a)－(b)
13	−0.797	−0.647	−0.693	0.046
17	0.590	0.480	−0.609	1.089
21	−1 001	−0.820	−0.693	−0.127
24	−0.739	−0.601	−0.435	−0.166
30	−0.780	−0.634	−0.388	−0.246
33	−0.479	−0.390	−0.553	0 164
40	2.629	2.137	3.801	−1.664
41	2.847	2.314	2.954	−0.640
42	2.403	1.953	2.922	−0.969

第6図　第1，第2合成変量による市町村の位置づけ

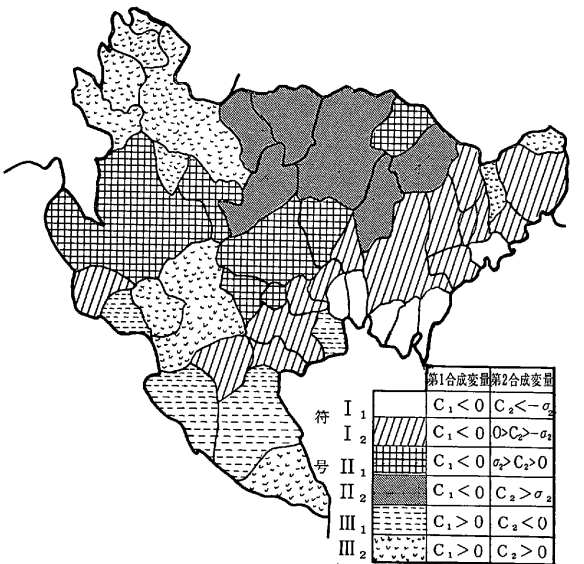


は第一主成分の場合よりも推定誤差がかなり小さい。

(4)　二つの主成分の組み合わせ情報の利用による部門結合型

二つの直交する第一，第二主成分からえられる第一，第二合成変量を直交座標軸に目盛り，各市町村を位置づけると第六図

第7図　第1，第2合成変量による標準化したスコア（昭和40年，佐賀県）



第14表　6つのクラスの農業粗生産額構成比率

	米	麦類	雑穀・豆・いも	野菜	果樹	その他の耕種部門	乳用牛	にわとり	肉牛	豚	その他
I_1	76.91	8.64	0.76	1 13	0.00	2.61	1.60	3.80	0.13	0 71	3.71
I_2	69.47	6.98	1.50	4.32	2 64	1 81	3.06	5.41	0.78	2.01	2.01
II_1	55.74	4.99	2.35	5 38	17.21	2.16	3.17	4.52	2 01	1.55	0.82
II_2	37.15	2.47	1.26	5.62	43.77	1 18	2.63	2.79	1 71	0.21	1.21
III_1	54.94	7.84	2.94	7 57	9.22	4.05	0.85	6.80	1.26	3.06	1.47
III_2	43.95	6.35	5.15	8.93	13.49	5.02	3.26	6 58	3 44	3.20	0.60

のようになる。各象限に入る市町村数によって、これを六つのグループ（I_1～III_2）に分け、各タイプを佐賀県の地図にあらわしたものが第七図である。その地域分布をみると、佐賀平坦沿海部の稲作生産力の高い水田の卓越した市町村を中心に、次第に周辺部へむかって農業部門結合構造を変えていく環状構坦配置がかなりよくあらわれている。各地域タイプの内容は第一四表にみられるが妥当と考えられる。

すなわち、

I_1は水田中核地帯、I_2はそれに次ぐ水田地帯であるが、福富町だけは水田率が極めて高いにもかかわらず野菜のウェイトが高いためIII_1になっている。

II_2は果樹（みかん）の中核地帯で、II_1は水田地帯から果樹地帯への移行地帯である。

III_2は県内の限界地帯で畑作、畜産のウェイトが高い。III_1は亜限界地帯ともいうべき性格のもので水田のウェイトがIII_2よりも高い。

(5) 若干の残された問題

二つの主成分を使うことによってもとのデータよりもすくない変量に要約された農業部門結合状態のタイプをつくる、という目的はある程度達成できた。

しかし、いままでみてきたように個々の市町村のスコアをみると、各主成分の意味づけと必ずしも斉合的でない部分がある。なぜこのようなことが生じるのであろうか。

第一に考えられることは、この報告の始めにのべたように原データに構成比率という閉じた数体系を使ったことである。

第二に考えられることは、原データの各変量における分布型が正規分布から著しく遠いものがすくなくない、ということである。

第一の点について、その影響がどういう形であらわれるのかわたくしには分からない。

第二の点については、主成分分析に使われるデータの備えなければならない性質について、学者の間で必ずしも意見が一致してはいないように思える。

伊藤[1]は観測値を要素とする標本ベクトルの組は「必らずしもあるP変量母集団から抽出された無作為標本とは考えず、単にP次元の統計資料とみなす」として、特別な場合のみ「標本ベクトルの組をP変量正規母集団から抽出された無作為標本と仮定して」母集団の主成分とその分散の推定問題をあつかうとしている。

L・J・キング[2]も「多変量正規分布であるという仮定は推定問題をあつかうのでない限り、主成分分析解にとって最重要な

ものではない（not essential）」といっている。

しかしD・N・ローリー、A・E・マックスウェル[3]は「主成分分析を用いる場合は変量についてなんらの仮説をも設ける必要がない。それは確率変数である必要さえないが、実際的にはそれらの観測値は通常、ある種の母集団からの標本とみられる」と多少ニュアンスの異なるいい方をしている。

いずれにしても、原データを正規型に近いものに変換した場合にどうなるかを再計算すべきであるが、時間の関係で問題の提起にとどめた。（もっとも主成分分析の原データについて変数変換をやっている例はあまり見うけない。）

注（1）伊藤孝一『多変量解析の理論』一六二頁。

（2）L. J. King. *ibid*, p. 167.

（3）D・N・ローリー&A・E・マックスウェル、丘本監修『因子分析法』三頁。

〔付記〕この報告の執筆にあたって三枝研究員、唯是研究員に多くの点で御教示を頂いた。

また、主成分分析の計算は総研のFACOM二三〇―一〇によって行なったが、その際、唯是研究員の開発したアプリケーションを利用させて頂いた。