マッチング、リグレツション、アグリゲーション

前　田　敬四郎

序

社会科学の発展、それに伴う数量化によって、体系的で科学的な単純化に対する要請が生じて来た。大きなデータ量やデータから導き出される特性を吸収，理解，処理しようとの能力には限界がある。電子計算機は，人間によってプログラムされねばならず，それらが作り出す産出物は，人間によって読まれ，理解されねばならぬ。理論家は，理論的分析や説明水準の段階で，先づ，取り扱うとするか，取り扱わないとするかの部分に，分割することが要求される。次には，注意深く選択され，定義され，限定された概念によって，1つの部分を取り扱うことになる。

計量経済学者は，純粋理論家よりも，より厳しい単純化問題に直面される[1]。

例えば，理論家は，「消費者はn財の間から選択する」と仮定し大きなnを持った分析を行うことは不都合でないかも知れないが，計量経済学者はnを合理的に小さくし度い。最初に入手した沢山の数から，如何にしたら少数の財や結合財を選択するかの方法が明きらかでないかも知れない。「データが余りにも多い」，「データが余りにも細かい」ので処理出来ないことを見出し，ある方法で細かいデータをグループにしたり，結合する必要を感ずる時に，我々は1つの単純化，アグリゲーションの問題に遭遇する。

本稿では，最初にアグリゲーションの理論的検討を行い[2]，その後で，Nancy and Richard Raggles のマッチングに関する研究[3]を，回帰とアグリゲーションの観点から取り上げて見た。

(注)

(1) Walter D, Fisher, Clustering and Aggregation in Economics. The Johns Hopkinss Press, 1969.

(2) H. A. John Green, Aggregation in Econnmic Analysis, Princeton University Press, 1964.

J. Vandaal and A. H. Q. M. Merkies, Aggregation in Economic Research, D. Reidel Publishing Company, 1984.

(3) Nancy and Richard Ruggles, A Strategy for Merging and Matching Microdata Sets, Annals of Economic and Social Measurment, 3/2, 1974.

Nancy Ruggles, Richard Ruggles, Edward Wolff, Merging Microdata: Rationale, Practice and Testing, Annals of Economic and Social Measurement, 6/4, 1977.

II　アグリゲーション[1]

アグリゲーションというのは，広義には他の寸法の和としての集計，或は，包括的寸法の計算からなるとして定義された。斯る操作の1例としては，与えられた社会における個々の需要と全体としての集合需要の測度がある。経済学者は，この二種類の寸法を区別するために，ミクロ経済，マクロ経済というjargonを使用した。然しながら，アグリゲーションという言葉の意味は，マクロ経済の寸法の定義や計算の域を越えて急速に拡大した。

一方で，集計の考えは，多かれ少かれ放棄され，データの集りを少数の寸法で要約しようとする操作と呼ばれた。そして，単なる加算の域を脱して，非常に複雑で色々な形の用途を持つ物価，生産指数などの計算が，アグリゲーションの代表的なものと考えられた。

他方，理論の面では，寸法の研究から，寸法間の関係の研究に，その分野を一般化した。斯くして，個人需要関数から出発した集団需要関数の定義が，個々と需要集計に代替された。そして，前者と後者は交換数量と価格の関係を移し変え得る。アグリゲーション問題は，経済局面だけでなく統計的局面も持っていた。従って，両者に考慮が払われねばならぬ。

タイル[2]は「アグリゲーション問題は，展々，不毛の結果のみを生ずる」というが，簡単に諦め切れない魔力を持っている。此処では，一致アグリゲーションは，非常に制約された条件の下でのみ可能というNataf の給論をめぐつた議論を展開する。

II．1　一致アグリゲーション

X_{jm}の要素を持つJ×Mの位数を持つマトリックスXを考え，Y＝H(X)を$U \subset \prod_{m=1}^{M} R^{J}$からRまでのXの関数としよう。

例えば，yはX_{jm}の和とする。すなわち，

(i)　$y=\sum_{j}\sum_{m} X_{jm}$

この場合に，その関数は二つの異なった方法で分解し得る。

(ii a)　$H(X)=\sum_{j} X_{j}$

(ii b)　$X_{j.}=\sum_{m} X_{jm}$　（行jの和）

(iii a)　$H(X)=\sum_{m} X_{.m}$

ここで

(iii b)　$X_{.m}=\sum_{j} X_{jm}$　（列mの和）

そこには，同じ性質を持つ他の関数Rがあるかどうか。すなわち，yがX，からのアグリゲート（和）関数として2つの異なった方法で書かれるという問題が生ずる。そのことは可能で，すべて斯様な関数は形態（i）で書かれる。例えば，

$$\text{(iv)} \qquad y = \Phi\left(\sum_m \sum_j \Phi_{jm}(X_{jm})\right)$$

は，Φが逆転可能であるならば，

$$W = \Phi^{-1}(y)$$

そして

$$Z_{jm} = \Phi_{jm}(X_{jm})$$

の変換を通して

$$W = \sum_m \sum_j Z_{jm}$$

の形で書かれる。

(iv)は，先に述べた二重の分解を許すH(X)の唯一の関数形であることが示される。

この論述は，A. Nataf に源を発する。ここで新しい証明を与え，アグリゲーション理論の枠内に於けるその適切性を示す。

経済関係をアグリゲートする問題のなかで，三つの要素を区別する。

（イ）　ミクロ変数間のミクロ関係の1集合

（ロ）　ミクロ変数がマクロ変数のなかに集計されねばならぬことに応じた規則

（ハ）　これらのアグリゲート間における1つのマクロ関係

アグリゲーションにおける一致の問題は，これらの集合の相互依存から生ずる。一般に，3つの集合の1つが与えられるならば，他の2つの集合は恣意には選択されない。この論述をより正確に再形成するためには，先づ，ある概念を導入して，一致アグリゲーションの概念を定義する。

諸概念

1個人はj文字で示され，個人の全体数はJである。1個人に関する概念は，1つの広い意味に理解されねばならぬ。すなわち，jは1企業と同様，個人又は家計となり得る。X_{jm}という記号は，独立ミクロ変数を示す。ここで，m(=1, ……, M)は(所得，年令，原料などの）変数の形態を示す。変数X_{jm}の各マトリックスXは，ある与えられた領域$U \subset \prod_{m=1}^{m} R^J$に属すると仮定される。

従属ミクロ変数は，y_j(j=1, ……, J) という記号で示される。jに関するアグリゲートは，その添字なき記号で示される。此処でドットをおとし，次のように書く。X_mは各々のmに対しX_{jm}の(jに関する）アグリゲートである。そしてyは$y_{j.}$のアグリゲートである。

ミクロ関係式は，j=1, ……, Jに対して，

$$y_j = f_j(X_{j1}, \cdots\cdots, X_{jM}) \qquad (1)$$

すべての個人jに対して，独立変数の数が，(jと独立した)Mであるという仮定は，そう見えるほど制約的ではない。その理由は，あるmに対して変数x_{jm}の値が，定義によって，jという個人に対してゼロとなり得るからである。更に，アグリゲーションの公式は，m=1, ……, Mに対して，

$$y = G(y_1, \cdots\cdots, y_j) \qquad (2)$$

並びに

$$X_m = g_m(X_{1m}, \cdots\cdots, X_{Jm}) \qquad (3)$$

で表わされる。

対応するマクロ関係式は，

$$y = F(X_1, \cdots\cdots, X_M) \qquad (4)$$

である。

例えば，関係式(1)は，ミクロ生産関数となり得る。(fについている添字jは，これらのミクロ関係式の起り得る不備な規定を暗示するとして考慮される。)その時，アグリゲートな産出物はyで示され，(2)で計算されるが，関

係式(3)は投入物に対するアグリゲーション処理を述べている。そして，(4)は所謂「マクロ生産関数」である。

一致アグリゲーション

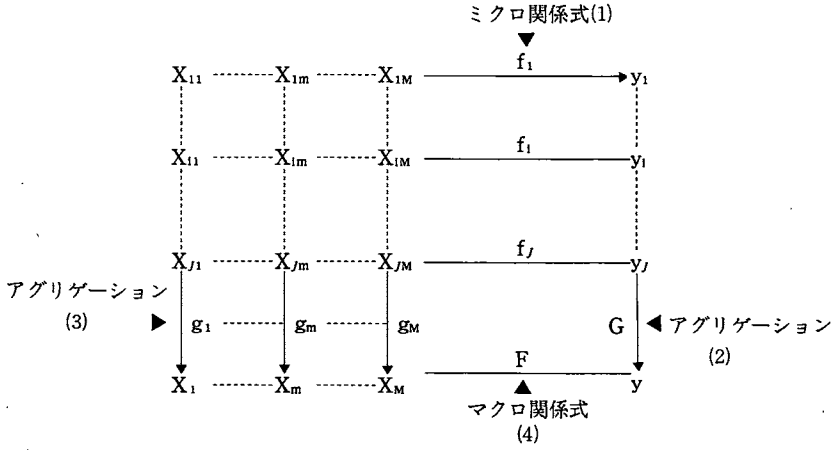
この枠内の一致性は，(3)で示されるM関数g_mを通してX_{jm}から生ずる各ベクトル$(X_1, \cdots\cdots, X_M)$は，関数Gによってy_jが行うように，関数Fを通してyの同一値を生ずることを意味する。これは第1図で明きらかにされている。

公式のなかで，一致アグリゲーションは，それぞれのX∈Uに対して，

$$\begin{aligned} y &= G\{f_1(X_{11}, \cdots\cdots\cdots, X_{1M}), \cdots\cdots, f_J(X_{J1}, \cdots\cdots, X_{JM})\} \\ &= F\{g_1(X_{11}, \cdots\cdots\cdots, X_{J1}), \cdots\cdots, g_M(X_{1M}, \cdots\cdots, X_{JM})\} \\ &= H(X_{11}, \cdots\cdots\cdots, X_{JM}) \qquad (5) \end{aligned}$$

が，成立しなければならぬ。上述の数学的問題の関連に注目すれば，関係式(1)は(ii b)と比較され，(2)は(ii a)，(3)は(iii b)，(4)は(iii a)と比較され得る。

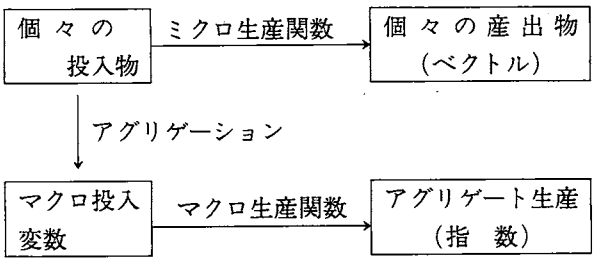
第1図　一致アグリゲーションのシエーマ



それ故，関数Hが前と同じ制約に出合っても不思議ではない。この論旨のなかで，Hは「原子論的マクロ関数」と呼ばれる。

一致アグリゲーションは第2図に示され，便宜上，生産関数を論じていると仮定する。

第2図　生　産　関　数



此処で，一致アグリゲーションに関するNataf（1948）の定理[3]を紹介しておこう。

非ゼロの第一次微分を持つ第1図のシエーマ関数が与えられとせよ。関係式(5)の意味の一致アグリゲーションに対して，すべてのシエーマ関数が加法的に分離可能で，原子論的マクロ関数Hが含まれることが必要で充分な条件である。此処で

$$y=H(X_{11},\cdots\cdots,X_{JM})=\phi\left\{\sum_{m=1}^{M}\sum_{j=1}^{J}\phi_{jm}(X_{jm})\right\} \qquad (6)$$

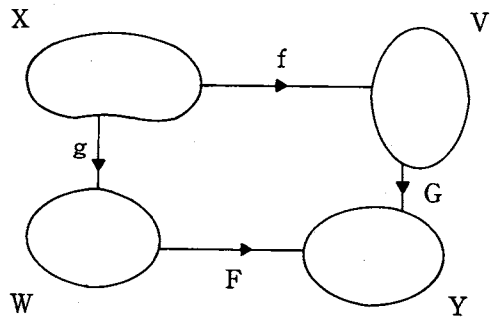
II．2　一致アグリゲーションの一般化

第1，第2図の拡張となる第3図を用いて，一般化を試みる。Xを説明変数の集合，Vを従属変数の集合としよう。猶，YをVのアグリゲートの集合，WをXに含まれる変数のアグリゲートの集合としよう。これらの集合が与えられる時，すべての$x\in X$に対して

$$y=F(g(x))=G(f(x)) \qquad (7)$$

となるように，1つのy∈Yが存在するような，f：X→V，F：W→Y，g：X→W，G：V→Yとなる関数が存在するかどうかを求める。第3図を見れば，f，g，FとGが集合値関数となり得ることを知り得る。すなわち，1つのXはv∈Vのより多くの値を生ずるが，第1図では，f_jという関数はy_jの1つの値のみを提供した。

第3図　集合値関数



一致性の定義(7)は，(5)式に対応している。加えるに，マクロからミクロ体系に戻ることが要求される場合には，それは対応する何れの定義よりも強くない。X，Y，V，Wが与えられる時，f，g，FとGは，(5)の成立するために応ぜねばならぬ最も一般的な充分で必要な条件は，あらゆるx∈Xに対して，Gfという関数の逆像が，gの逆像を含むことである。すなわち，

$$\forall x \in X: \quad g^{-1}(g(x)) \subset (Gf)^{-1}(Gf(x)) \qquad (8)$$

「一般的」なという用語については，f，g，FとGの関数に課される諸条件が存在しないことを意味する。条件(8)は，Ijiri[4]によって説明されるように以下に述べられる。第4図と第5図を見れば解るように，関数GとfのそれぞれぞれがGfという構成のみが果す），第3図のシエーマをX，Y，Z間の「三角関係」に単純化した。以下で明きらかになるように，X，W，Yの選択は，ユニークではない。Vを通してXからYに進む他の順番を考え得る。

先づ，我々は充分性を証明しよう。Gfとgが与えられるとき，(8)がそれぞれのxに対して成立するならば，Gf(x)の逆像は，逆像$g^{-1}(g(x))$の部分集

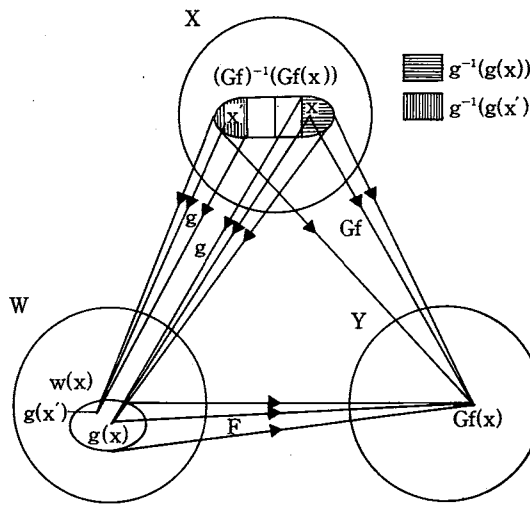
合のなかに分割される。

その理由は，gに関するる逆像がGfに関するある逆像と共通点を持つならば，前者が後者の部分集合とならねばならぬことを(8)式が述べているからである。

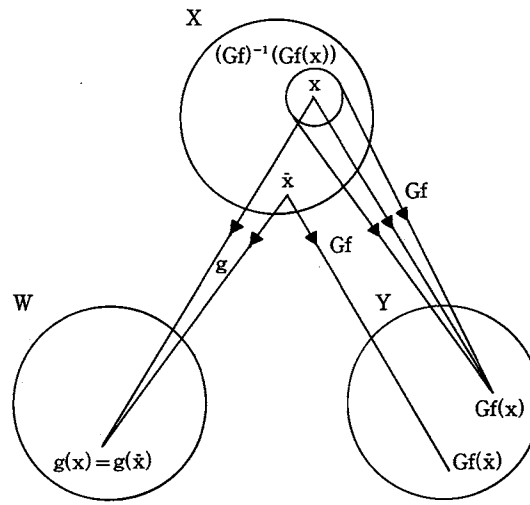
それぞれx∈Xに対して，w(x) の各要素のgに関する逆像が，Gf(x)の逆像の部分集合であるように，そこには1つの集合 w(x)⊂Wがある。あらゆるw∈w(x)に対して，F(w)=Gf(x)が成立するようにFを定義せよ。f，g，Gと一緒に，このFはアグリゲーション処理を完結する。

(8)の必要性を証明するために，(5)式がすべてのx∈Xに成立するような関数f，g，F，Gが存在すると仮定する。その場合に，そこにはg(x̄) = g(x) とx∉(Gf)⁻¹(Gf(x))というような要素xとx̄∈X

第4図　充　分　条　件



第5図　必　要　条　件



が存在し得る。然し，その時，$Gf(\bar{x}) \neq Gf(x)$ を持つ。これは(5)式と矛盾している。

この条件(8)は，経済学における他の一般的結果と共通に，現実の状態では有用でないようなものを持っている。その関係式を規定するなかで，これまでに得られた結果以上に進み得るかどうかは疑問である。Nataf との関係を明きらかにするために $X \subset R^{J \times M}$，$V \subset R^{J}$，$W \subset R^{M}$ と $Y \subset R$ を考えよ。

$$g(x) = (g_1, \cdots\cdots\cdots, g_M) \qquad (9)$$

となるように，$g : X \to W$ としよう。

$x \in X$ が，それぞれ J 次元のM部分ベクトルのなかに分割されるとしよう。そして，$g(x)$ のm番目要素 g_m を部分ベクトル $(X_{1m}, \cdots\cdots, X_{Jm})$ だけの関数としよう。それ故，我々は，

$$g_m = g_m(X_{1m}, \cdots\cdots, X_{Jm}) \qquad (10)$$

と書き得る。

Gf は H として書く単一の価値関数であるならば，$F(g(x)) = H(x)$，或は

$$H(X_{11}, \cdots\cdots\cdots, X_{JM}) = F(g_1(X_{11}, \cdots\cdots\cdots, X_{J1}), \cdots\cdots\cdots, g_M(X_{1M}, \cdots\cdots\cdots, X_{JM})) \qquad (11)$$

が存在するかどうかを求め得る。これは，完全に(5)式の最後の部分となる。

仍，実際には，一致アグリゲーションに関する Nataf 条件は，殆んど充されない。我々は，より多くの矛盾に耐えねばならぬ。次にこの問題を取り扱うことにしよう。

Ⅲ．3　最適アグリゲーション

Nataf の条件が充されない場合，一致アグリゲーションは不可能である。然しながら，科学においては不可能なことをなそうと試みる。

この問題が最適化の文脈のなかに置かれるならば，議論のみはなし得る。このことは実際に行うよりも云うことはより容易である。最適化方法のより

大きな一般性というのは，恐るべき努力の費用に於て達成され得るだけで，その時でも恣意の要素は避けられない。常に，その最も重要な問題は便益と費用の評価であり，これらの便益と費用を互いにどのように考慮するかである。より高い水準のアグリゲーションに於て働く便益は，より大きな明快さと処理し易さ，並びに，より容易な計算であり，費用は，情報の損失，確心や精度の損失である。（アグリゲーションの程度の関数としての）便益や費用は如何に数値的に評価されるであろうか。

我々は斯る一般的問題に巻込まれることに尻込みする。斯様な方法は，未だ余りにも曖昧になっている。それでも，ある特殊な場合には，合理性を持つ解が可能であるかも知れない。そこには我々が未だ躊躇するもう１つの理由がある。すなわち，最適アグリゲーションには，デスアグリゲートされたデータの知識が必要である。これらが利用出来たとしても，猶，アグリゲーションが必要であるかどうかの問題が生ずる。従って，兎に角アグリゲートする決定が行われて来た場合に限定して議論を進めることにした。

非常に限定した解釈においては，最適アグリゲーションというのは，１つの差，又は，差の関数を最小にすることになる。我々の簡単な単一方程式問題のなかで，これは，

$$v = F[g_1(X_{11}, \cdots\cdots\cdots, X_{J1}), \cdots\cdots\cdots, g_M(X_{1M}, \cdots\cdots, X_{JM})] - G[f_1(X_{11}, \cdots\cdots\cdots, X_{1M}), \cdots\cdots\cdots, f_J(X_{1J}, \cdots\cdots, X_{JM})] \quad (12)$$

で定義されるVの絶対値を最小にする形を採用し得る。

この最小化は，これらの剰余が与えられ，現実でなしに想像であろうとそうでなかろうとミクロ・データが与えられるならば，F，G，$f_1, \cdots\cdots, f_J, g_1, \cdots\cdots, g_M$の若干を選択することによって得られるに相違ない。我々のアグリゲーション処理の一般化は，データ・セットが与えられる時，非固定的な集合関数F，G，f_1，……，f_J，g_1，……，g_Mからの適当な選択によって，

$$V = \sum_{t=1}^{T} v_t^2 \quad (13)$$

を最小化することである。不幸にも，どれであれ，１つの解がデータ・セットと独立であるとは期待されない。その問題を解く関数のパラメータが，ア

グリゲート・データにのみ依存する場合に，線型関数に対してのみ諸結果が知られる。然しながら，もっと一般的ケースに於て，これらのパラメータは，ミクロ変数，或は，既知でないアグリゲートに依存したかも知れない。これらの関数（線型，対数，双曲線，など）の性質は，データ・セットの内容と共に変動することが可能でさえある。

線型関数のケースの1つの重要な例を与えよう。関数Gと同様に関数 $f_1, \cdots\cdots, f_J, g_1, \cdots\cdots, g_M$ が，すべての期間1, ……, Tに与えられると仮定しよう。

$$y_{jt} = \sum_{m=1}^{M} \alpha_{jm} X_{jmt} \tag{14}$$

$$X_{mt} = \sum_{j=1}^{J} X_{jmt} \tag{15}$$

$$y_t = \sum_{j=1}^{J} y_{jt} \tag{16}$$

関数Fは線型であると仮定される。

$$y_t = \sum_{m=1}^{M} \beta_m X_{mt} \tag{17}$$

線型で，それ故，加法的に分離可能であるが，α_{jm} が1個人あてに異なるならば，パラメータ $\beta_1, \cdots\cdots, \beta_M$ の何れの選択に対しても(14)～(17)の関係式は同時に成立出来ないような関数である。1つの可能な解は，(15)を異なったアグリゲーション公式 $X_{mt} = \sum_j \alpha_{jm} X_{jmt}$ で置き換えることであるが，方程式(14)～(16)を緩めることが出来ないと仮定するならば，脱出する唯一の方法は，ある最適規準に応じて，パラメータ β_m を選ぶことである。

(14)～(17)を(12)のなかに挿入すれば，

$$v_t = y_t - \sum_{m=1}^{M} \beta_m X_{mt} \tag{18}$$

を生ずる。その結果として，(13)が規準関数ならば，ベクトル $\beta = (\beta_1, \cdots\cdots, \beta_M)'$ の最良の選択は最小二乗回帰ベクトルである。すなわち，

$$X = \begin{bmatrix} X_{11}, \cdots\cdots\cdots, X_{M1} \\ \vdots \\ X_{1T}, \cdots\cdots\cdots, X_{MT} \end{bmatrix} \quad \text{並びに} \quad y = \begin{bmatrix} y_1 \\ \\ y_T \end{bmatrix}$$

を持ち，

$$b=(X'X)^{-1}X'y \qquad (19)$$

これは，特に線型モデルの場合に，アナロージによるアグリゲーションの広く使用される方法を支持する。

II．4　アグリゲーションの処理法

先にも述べたように，一致アグリゲーションが，現実に存在することは極めて難しい。然しながら，現実の経済分析に当たって，アグリゲーションの処理をしなければならないことにしばしば直面する。特に，ミクロ・データの取り扱いにおいては，アグリゲーションを避けて通ることは出来ない。これまでにミクロ・データのアグリゲーションに用いられた方法としては，回帰分析とマッチングの二つがあげられる。

データ・セット間の情報を移動する伝統的な方法の１つは，回帰分析の使用によるものである。標本Ａにおけるそれぞれのケースに対して，標本Ｂに含まれる１変数の推定値を予測する多重回帰モデルを設定することによって，一つのデータ・セットから他のセットえと，情報が帰属される。勿論，この方法が成功するためには，二つの標本が，その回帰方程式のなかで独立変数として役立つ共通の変数を含むことが必要である。例えば，一つの標本が賃金労働者の地位，年令，性，職業，産業，所得などの特性値を示していたならば，同一の年令，性，職業，所得などの特性値を含む，もう１つ別のファイルの各賃金労働者に地位の情報が帰属されたであろう。物論，斯る帰属の有効性は，帰属される変数（地位）が他の共通の変数（特性値）によって如何にうまく説明されるかに依存する。多くの分析目的に対して，個々の観察水準において正確となる必要はない。推定値が既存の変動領域に関する平均に於て，満足に作用することが必要なだけである。回帰のフィットが非常に接近しているならば，現実値の代りに回帰値を代入しても，その後の分析を駄目にすることはない。

回帰による帰属技術は，複雑な情報セットを移動するのには，満足出来な

い。例えば，予算情報が，より豊富な社会学的，地勢学的情報を含んだ一つの標本に帰属されるならば，予算支出が，すべて高度に内部相関するという一つの問題が生ずる。各支出に対する個別推定値は，ある特定の個人に対して矛盾した予算形態を生ずる。更に，予算情報を集めることの主要目的の1つは，予算項目の間の内部相関関係であるが，各予算支出が独立に帰属されたならば，その内部相関関係は失なわれてしまう。各支出項目に対して，元の標本における情報を保つように試みて，已に帰属された要素を考慮するモデルを設計することは可能であったとしても，現実の関係が線型又は対数線型の加法モデルでうまく近似化出来ないならば，斯るモデルは非常に複雑なものとなる。単純でより満足の出来る処理方法は，マッチング処理によって予算情報の完全なセットを，一つの標本観察値から他の標本の観察値えと移動し，両標本において情報セットの完結を留保することである。

マッチング処理を用いることは，重要な方法論上の意味を持っている。回帰による帰属は，通常，平均値を割当てる結果となるが，マッチング技術は，元のデータ・セットの値の分布を再生する。単独の帰属に対しては，平均値は望ましいかも知れないが，繰返しの帰属に対しては，平均値の使用は観察される分散を破壊して仕舞う。マッチング技術の成功は，データ・セットのなかに相似なケースが見出されるように，データが全く密であることに掛っている。マッチングの目的に対して，如何なる特定の関数関係も予め決める必要がないと云うことは特筆すべきことである。その関係が陽表的に非線型であるということを認めることなしに，非線型関係が自動的に線型関係と同様に有効に処理され得る。これは予め正確な関数形の決定を要する回帰技術と著しく対照的である。関数形がよく解っており，データが散乱してマッチングが困難な場合には，回帰分析はより適切な帰属を提供するが，相似したケースが存在する大きなデータ母体では，マッチングによる帰属は元の標本の分布を留め，基本的関係をより正確に反映するという利点を持っている。

(注)

(1) ibid. J. Vandaat and A. H. Q. M. Merkies, Aggregation in Economic Research.

cf. E. Malinvaud, L'Aggregation dans les Modeles Economiques, Cahiers

du Seminaire d'Econometie, 4, Centre National de la Recherche Scientific, Paris, 1956.

前田敬四郎，アグリゲーションの功罪，金沢大学法文学部論集，経済学篇20，1973. pp. 1－49，参照。

(2) H.Theil, Linear Aggregation of Economic Relations, North-Holland Publishing Company, 1954. p. VII.

(3) A. Nataf, Sur la Possibilite de Construction de Certains Macromodeles, Econometrica, 16(1948), pp.232-244.

ibid. H. A. John Green, Aggregation in Economic Analysis, 5 chapter.

(4) Y. Ijiri, Fundamental Queries in Aggregation Theory, Journal of the American Statistical Association, 66(1971) pp.766-782.

Ⅲ．アグリゲーションの１つの応用

National Bureau of Economic Research (ＮＢＥＲ) の人達，特にラグルス夫婦によって展開されたマッチング処理技術の開発は，アグリゲーションの観点から眺めるとき非常に興味ある方法である。先づ，Ａ，Ｂ両ファイルのデータ・セットを移転，統合する時にマッチング技術が使用される。そのマッチングに当って最も重要なのは，両ファイルの共通変数の使用方法である。ＮＢＥＲは，1970年の Public Use Sample（ＰＵＳ）とSocial Security Emplyoer-Employee Dataファイル(ＬＥＥＤ)のマッチでは，共通変数の「賃金」を区間分割するという方法を用いた。その区間分割基準と作製表を検討し，更に，1970 Census15% 1/1000 ＰＵＳと1969 Internal Revenue Service Tax Model （ＩＲＳ)の間で行われたマッチング処理の解説を行う。最後の処では，マッチング技術の信頼性を評価するために，数個の計量経済学的検討を行う。

同一標本の枠を持ち，そして，殆んど同じセットの変数を持ち，同じ母集団から抽出された二つの標本（1970 Census ５％と15％ ＰＵＳ）をマッチすることによって，これを行う。その時，帰属された同時分布が現実の同時分布と統計的に異なるかどうかを決めるために，マッチから得られた帰属値を現実の値と代替する。その結果は，そのケースの95%に於て帰属と現実の

同時分布の間に統計的相違がなかったことを示している。

Ⅲ．１　マッチング処理と区間分割

マッチングの共通変数ｘは，(1)，うまく順序をつけ得る，(2)，順序をつけることが出来ない，(3)，部分的には順序をつけ得る。これら三つのうちのどれかによってｘ変数の分割基準が異なる。賃金所得は(1)に，人種，職種などは(2)で，産業，地域，州は(3)に該当する。

(1)に当るｘ変数が手に負えないほど沢山の生の値を持つことがある。ＰＵＳのなかの「賃金」変数は100ドル刻みで250の区間からなり，ＬＥＥＤファイルは１ドル単位で「賃金」を報告する。従って，生の値を比較する代りに，賃金変数が恣意に少数の区間に分割され，それらが比較された。区間の間に有意差が見出されるならば，これらの区間のそれぞれが二つの区間に分割され，比較される。区間の間に如何なる有意差も見出されないか，或は，生の値に達するまで，この過程が続けられる。

此処では，賃金変数について標本に順序をつけ，それ等を８つの主要線分に分割し，各線分は，同数の観察値を持つ様にすることにした。この方法は，それから生ずる区間が信頼出来る比較を与え，標本のサイズに関しても最適のものであった。

標本が生の値を少ししか持たないｘ変数と沢山の生の値を持つｘ変数を持つ場合の唯一の相違は，前者では，より小さな区間がより大きな区間にアグリゲートされ，後者は，より大きな区間がより小さな区間にデス・アグリゲートされると云うことである。

ｘ変数の生の値の隣り合った区間に対するｙの分布が比較され，カイ二乗と相関測度が計算される。如何なる有意差も見出されないならば，つまり，差の大きさが与えられた水準以下にあるならば，生の値が結合される。それから，新しく結合された区間とそれと隣り合う他の区間の間で一つの比較が行われる。ｘ変数はカイ二乗と相関係数の規定された水準に基づいて，区間集合に分割される。一般に，ｘ変数は，１つ以上のｙ変数によって分析される。それでｙ変数から派生した個別的分割から，どのような方法で一般化された

分割が導出されるかを考慮する必要がある。それには二つの規則が適用される。第一に，個々の分割に代表される最も細かな区間を反映するように一般化分割を構成することが可能である。第二に，y変数に対する百分率の分布をプールし，これらのプールされた分布を基礎に相関係数を計算することが出来る。

これまでに述べた方法に従って，x変数（賃金）を三つの枝分れ集合に，賃金区間を分割した例を第1表に示そう。

生の賃金値は，1～99ドルから25,000ドル乃至それ以上に渡って100ドル刻みで250の賃金層からなった。区間分析をするのに，21個のy変数が使用された。最も細かな階層別水準（水準Ⅲ）において，それぞれのy分布に対して区間の間の差のカイ二乗測度が，0.95以下であった場合に，それらの賃金クラスのみが結合された。この基準は，大さが100ドルから13,200ドルの値域にあって，0.7から13.1パーセントまでの観察値を含んだ21区間を生じた。21番目の区間（11.800～25,000乃至それ以上）に対する広い賃金クラスは，大部分が，この範囲の観察数が相対的に小さいことによるものであることが指摘されねばならぬ。これらのrunがなされた標本は，約20,000の観察を含み，約300の観察が21番目の区間にあったことを意味する。層別標本の使用と一緒に標本の規模における増加は，恐らく，21番目の区間が幾つかの区間に枝分れする結果を生じたであろう。マッチング処理によって斯様に細かい区間にすることは，データの約1.7パーセントのみがマッチされるマッチングを改善するが，最高の賃金クラスの分析が重要である研究に対しては，これらの賃金クラスにおけるマッチングを改善するために，特別の注意が払われるべきである。水準Ⅱに対しては，水準Ⅲに使用された区間を結合するための基準は，カイ二乗が0.95以上で，相関係数が0.9を超えた場合に，加法的に，区間を結合するように緩和された。これは最小賃金クラスの大さが1000ドルで，最小カバレージが観察値の1.5パーセントで，区間の数を8に減少した。Ⅲの階層別水準で特定された21区間のうち4つは，そのまま，Ⅱの階層別水準に持ち込まれた。最後に，相関係数の基準を0.70に緩和することによってⅡの階層別水準の8つの区間は，水準Ⅰに対する2つの区間に崩壊する。この水準において区別された二つの所得クラスは，1～1,799 ドル

第1表　賃金階層の区間分割

賃金階層（ドル）	階層別水準					
	水準 I		水準 II		水準 III	
	区間数	観察のパーセント	区間数	観察のパーセント	区間数	観察のパーセント
1～ 99	1	31.7	1	31.7	1	3.3
100～ 499					2	9.8
400～ 599					3	2.1
600～ 799					4	3.4
800～ 1,799					5	13.1
1,800～ 2,299	2	68.3	2	39.6	6	6.9
2,300～ 2,799					7	6.0
2,800～ 3,499					8	9.1
3,500～ 3,899					9	5.1
3,900～ 4,299					10	6.0
4,300～ 4,499					11	1.7
4,500～ 4,899					12	4.8
4,900～ 5,299			3	14.7	13	6.3
5,300～ 5,499					14	1.4
5,500～ 6,299					15	7.0
6,300～ 7,499			4	5.8	16	5.8
7,500～ 8,499			5	3.0	17	3.0
8,500～ 9,099			6	2.0	18	1.3
9,100～ 9,799					19	0.7
9,800～11,799			7	1.5	20	1.5
11,800～25,000*			8	1.7	21	1.7

＊最高所得層は25,000ドル以上

区間を結合するための仕様書

カイ二乗が0.000と0.94の値域にあるならば，相関係数と関係なしに区間は結合される。

カイ二乗が0.95と1.00の値域にあるならば，相関係数がそれぞれの階層水準に対して下に示される水準以上にあれば結合される。

階層別水準	相関係数
1	0.70
2	0.90
3	1.00

出典　op. cit. Nancy and Richard Ruggles, A Strategy for Merging and Matching Microdata Sets.

と1,800ドル乃至それ以上のものである。最初の区間は32パーセントの観察を含んでいる。物論，望むだけ多くの階層別水準を生ずることは可能である。

Ⅲ．2　マッチング処理の経験的テスト

此処では，1970 Census 15%と1969ＩＲＳの間で行われたマッチの結果を示し，[1] その時行われたマッチング処理の技術的側面を紹介することにしよう。

A．マッチの指令。　１つのファイルBが他のファイルAにマッチされる。このことは，Bファイルの情報がAファイルの各記録に移されることを意味する。例えば，ＰＵＳ－ＩＲＳのマッチにおいて，ＰＵＳファイルにＩＲＳファイルをマッチすることを決定した。その理由は，ＰＵＳファイルはU.S母集団のランダム（代表的）サンプルであるが，ＩＲＳファイルは上層所得グループが過剰に代表した層別標本であるということであった。ＰＵＳファイルにＩＲＳファイルをマッチすることによって，税情報が適切な母集団のウエイトを与えられることを保証し得たであろう。

B．マッチング単位。　ミクロ・データセットは，異なった観察単位を持つから，マッチングの目的には，二つのファイル間に共通した単位を選択することが必要である。このことは，幾つかのファイルの１つのなかに，対応する単位の創設を招く。例えば，ＰＵＳファイルにおける基本単位は家計であるが，家計は家族や個人の観察値のなかに分解される。ＩＲＳファイルにおいて，基本単位は申告で単一または共同である。すべての結婚カップルは，共同申告のなかにファイルすると仮定することによって，ＰＵＳファイルのなかの個人から単独，または，共同申告単位を構成して，税単位に関して二つのファイルをマッチした。

C．マッチを行う変数の選択　マッチング処理のなかで，それぞれのファイルのなかに４種類の変数がある。最初の変数はコホルト(Cohort) 変数である。これらは，正確な値をベースにマッチされる両ミクロ・データセットに共通する変数である。ＰＵＳ－ＩＲＳマッチにおいて，彼等は，税申告の形態，単独申告の場合における性，共同申告のケースにおける家長の年令と人種，単独申告の場合には回答者の年令と人種を選択した。第２表に示さ

れる。コホルト変数は非常に重要なので他のファイルからの近似値ではマッチ出来ないと考えられる。

第２表

1970 Public Use Sample の構造－1969 Internal Revenue Service Tax Model のマッチ。

A．	コホルト変数
1．	税申告の形式
2．	回答者の性（独身の申告）
3．	家長の人種
4．	家長の年令
B．	X 変 数
1．	子供の数
2．	持家－借家
3．	賃金またはサラリー所得
4．	事業所得
5．	農家所得
6．	全 所 得
C．	Y 変 数(a)
1．	教育
2．	出生地
3．	職業
4．	雇用されている産業
5．	労働者の種別
6．	結婚年数（既婚者のみ）
7．	現在地の滞在年数
8．	不動産の金額（持家の人のみ）
9．	家計における自動車の数

(a)はＰＵＳファイルのみ

注．ＰＵＳファイルの収入と所得情報は1969歴年であるからマッチには1969ＩＲＳファイルを使用した。

出典．op. cit. Nancy Ruggles, Richard Ruggles, Edward Wolff, Merging Microdata.

第２の種類はＸ変数である。これら両ファイルに共通した残余の変数であるが，コホルト変数ほど重要ではない。これらの変数は近似値，すなわち，マッチング区間を基礎にマッチされる。（以後を参照）。ＰＵＳ－ＩＲＳマッチにおいて，Ｘ変数は子供の数，家の所有権，賃金や俸給所得，事業所得，農家所得，全所得であった。Ｘ変数はＡとＢファイルの間で概念や分布において相違があるので，それらをX_a及びX_bと指定した。（以後を参照）。

第三番目の種類は，Ｙ変数である。これらは，マッチング区間を構成するのに使用される共通でない変数である。（以後を参照）。原則として，マッチング区間の１集合が，それぞれのファイルに対して発生されるべきで，そして，二つの集合はマッチング区間の１つの統合された集合を作るためにまぜ合される。この場合，各ファイルに対してＹ変数の別々の集合，Y_a，Y_bがある。然しながら，その電算機の処理費用が莫大であるという理由から，Ａファイルのみからマッチング区間を発生した。ＰＵＳ－ＩＲＳマッチのＹ変数は第２表に並べられている。各ファイルにおける残余の変数の集合は，それぞれZ_a，Z_bと指定されている。これは，非共通の変数であるが，表面上は，Ｘ変数と関連がないように見えるので，我々は殆んど関心を持たない。

出生の特定区，経験の地位，仕事えの輸送機関，家で話される言葉などは，ＰＵＳファイルにおけるＺ変数の例である。

Ｄ．マッチング区間の構成。　Ｘ変数に対して，同じ値を持ったＡ，Ｂファイルの記録値を見出すことは殆んどあり得ないので，近似値においてＸ変数をマッチすることが必要である。「充分に近接している」と考えられるＸ変数の値域が「マッチング区間」を形成する。彼等は，X_aに対してX_a ($f(Y_a/X_a)$)に関するY_aの条件分布の感度を分析することによって，これらの値域を構成する。与えられた有意水準に予め設定されたカイ二乗，或は，相関テストを使用して，X_aの何れの値のなかで条件分布 $f(Y_a/X_a)$ が統計的に異なっているかどうかを決定した。条件分布が統計的に異なっているX_aの 値は，異なったマッチング区間のなかにおかれ，条件分布が異なっていない値は，同一マッチング区間のなかに置かれる。猶，カイ二乗や相関水準を変えることによって，異なったマッチング水準にあるそれぞれのマッチング区間の集合を生じ得る。実際に，有意差に対する基準を連続的に緩和することに

よって，マッチング区間の枝分れ集合を生ずることが出来る。ＰＵＳ－ＩＲＳのマッチで使用されたＸ変数の収入に対するマッチング区間の集合とマッチング水準が第3表に示される。例えば，カイ二乗（0.99）水準でＩＲＳ収入の3500ドルは，ＰＵＳ収入の3700ドルに対しては適切なマッチングであるが，ＰＵＳ収入の3900ドルに対してはマッチングしないと考えられる。相関（0.97）水準においては，3401－4100ドルの値域にあるＩＲＳ収入の何れも，その値域にあるＰＵＳ収入に対し適切なマッチと考えられる。第3表から明きらかなようにマッチング区間の値域が広がり，最初と最後のマッチング区間の間でマッチング区間の数は減少する。第4表に示されるように，このことはＰＵＳ－ＩＲＳマッチにおける他のマッチング変数にも成立する。この表から，マッチング区間の数や「崩壊」率がＸ変数の間で本質的に異なることが明きらかにされた。[3]

Ｅ．Ｘ変数の調整。　Ｘ変数がＡ，Ｂファイルの間で概念や標本分布に若干の相違があることが度々起り得る。この二つのファイルがマッチされる前にX_aとX_b変数を調整することが必要である。先づ，二つの変数が概念において異なる場合に，他の情報があらわれるならば，1つの概念を他のものに変換することが，時折，可能になる。例えば，ＰＵＳ－ＩＲＳマッチにおいて，ＩＲＳファイルのなかで調整された粗所得（ＡＧＩ）が，ＰＵＳファイルにおける全体の個人所得にマッチされた。ＩＲＳファイルに現われる他の情報から，個人粗所得を得るために配当留保や他の戻し調整をＡＧＩに加えることが出来た。ＩＲＳファイルの粗所得は，社会保証所得を排除するが，資本収益を含む。処が，ＰＵＳファイルにおける総所得は，社会保証所得を含むが，資本収益を排除するから，二つの概念は同じでなかった。その二つの概念を調整するためには，ＩＲＳファイルの粗所得から資本収益を引き，ＰＵＳファイルの全所得から社会保証所得を引くことが必要であった。標本の枠における相違か，報告される誤差の相違の何れかの理由から，X_aとX_bが標本分布において相違する場合に，二つの分布を調整することが必要になる。彼等はＰＵＳ－ＩＲＳマッチのなかでは，この問題に遭遇しなかった。[4]

Ｆ．選別－統合のマッチと目盛り。　マッチ自身は，次の段階で行われる。先づ，ＡファイルとＢファイルのなかに，それぞれ記録されたＸ変数が，マ

第3表　賃金収入変数に対する枝分れ区間の構造

収入（ドル）	マッチング水準別の区間数					
	カイ二乗	相関 (0.99)	相関 (0.97)	相関 (0.90)	相関 (0.80)	相関 (0.70)	相関 (0.50)
0～ 200	1	1	1	1	1	1
201～ 300	2					
301～ 400	3	2	—	—	—	—
401～ 500	4	3				
601～ 700	5		2			
701～ 800	6	4				
801～ 900	7		—	2	2	2
901～ 1,400	8	5				
1,501～ 1,700	9					
1,701～ 1,800	10	6	3			
1,801～ 2,000	11					
2,001～ 2,200	12		—	—	—	—
2,201～ 2,500	13	7				
2,501～ 2,800	14					
2,801～ 2,900	15	8	4	3		
2,901～ 3,100	16					
3,101～ 3,400	17	9	—	—	3	
3,401～ 3,800	18	10				
3,801～ 4,100	19		5	4		
4,101～ 4,300	20	11				
4,301～ 4,800	21		—	—	—	—
4,801～ 4,900	22					
4,901～ 5,100	23	12	6			3
5,101～ 5,400	24					
5,401～ 5,900	25					
5,901～ 6,400	26		—			
6,401～ 7,100	27	13				
7,101～ 7,500	28	14	7			
7,501～ 8,000	29		—			
8,001～ 8,700	30	15		5	4	
8,701～ 9,700	31	16	8			
9,701～13,600	32	17				
13,601～15,600	33	18	9			
15,600～18,600	34	19				
18,601～25,500	35	20				
25,501～50,000	＋ 36	—	—	—	—	—

出典　op.cit. Nancy Ruggles, Richard Ruggles, Edward Wolff, Merging Microdata.

第4表　PUS－IRSマッチにおけるマッチング水準別のマッチング区間数

	マッチング水準					
	6	5	4	3	2	1
	相関 (0.50)	相関 (0.70)	相関 (0.80)	相関 (0.90)	相関 (0.97)	カイ二乗 (0.99)
X 変数						
1．子供の数	1	1	1	1	4	8
2．家主の地位	1	1	1	2	2	2
3．賃金収入	3	4	5	9	20	36
4．事業収入	1	1	1	1	1	13
5．農家所得	1	1	1	1	1	2
6．全所得	2	2	3	6	16	36

出典．op. cit. Nancy Ruggles, Richard Ruggles, Edward Wolff, Merging Microdata.

ッチング区間のなかに記録される。第2に，各ファイルの記録は，それらのコホルト値を基礎に，コホルト内では，マッチング区間の値を基礎にして選別される。第3に，Aファイルの各記録に対して，最初の（最も細分化された）マッチング水準にあるA記録のそれと同じマッチング区間値を持つB記録の探究が行われる。これが失敗すると，コホルト水準に対して第Ⅲのマッチング水準等々の候補が探し出される。マッチング水準が確立されるや否や，この水準のA記録において，A記録にマッチする全てのB記録から，マッチするB記録がランダムに選択される[5]。選択されたB記録は，それから，A記録と一緒に統合される。第四に，マッチング水準別のマッチの分布は，目盛り調べが行われる。分布が一様でないならば，新しい確率水準の集合と一緒に，新しいマッチング区間が生ぜられる。そして，そのマッチが繰返される[6]。この過程は，それから生ずる目盛りが，相対的に一様となるまで続けられる。PUS－IRSのマッチでは，3度の繰返しが必要であった。最後の目盛りは第5表に示されている。

第5表　PUS－IRSの目盛り調べ

マッチング水準	マッチのパーセンテージ
1. カイ二乗 (0.99)	16.0%
2. 相　関 (0.97)	18.8
3. 相　関 (0.90)	30.6
4. 相　関 (0.80)	14.3
5. 相　関 (0.70)	12.2
6. 相　関 (0.50)	6.2
7. コホルト	3.0
	100.0

出典．op. cit. Nancy Ruggles, Richard Raggles, Edward Wolff, Merging Microdata.

Ⅲ．3　マッチング処理の経験的テスト

Sims[7]が指摘するように，Xを条件とするY_a並びにY_bの同時分布について何も云うことが出来ないことは真実であるが，[8] それでも，Y_a，Y_bの生の同時分布については，何か云うことが出来る。以前に示された如く，[9] Y_aとX，Y_bとXの間の相関が強ければ強いほど，Y_aとY_bの帰属される同時分布は現実の（未知）同時分布に近づく。Y_aとX，Y_bとXが完全に相関する場合には，帰属される同時分布は，正確に，現実の同時分布を模写するであろう。然しながら，相関係数が正，負の1.0からそれるのが多ければ多いほど，帰属されたものと現実の分布間の起り得る誤差は，ますます大きくなる。

ここでマッチング処理の1つの統計的検定を提供する。その検定は1970 Census 1/1000 15%ＰＵＳに対する1970 Census 1/1000 5%ＰＵＳのマッチに関して行われる。（第6表を参照）。

二つのデータ・セットは，約12の変数に対してを除いて，それらの変数リストにおいて全く同一で，[10] 同じ大さのランダム標本である。結果として，殆んどすべてのＹとＺ変数は，二つのデータ・セットのなかで同一となるであろう。[11] 斯くして，これはY_aとY_bの帰属された同時分布と現実の既知同時分布との比較を許す。更に，Y_aとZ_bの帰属された同時分布は，現実のものと比較されるし，Y_a，Y_b，Ｚの帰属された同時分布は，多くの2変数に

第6表　1970 Public Use Sample 5%－15%のマッチ

A．コホルト変数
1．婚姻状態
2．家長の年令
3．家長の性
4．家長の人種
5．持家或は借家

B．X 変 数
1．家計の子供の数
2．財産値，或は，毎月の粗レンタル料
3．家長の賃金収入
4．家長の配偶者の賃金収入（結婚の場合）
5．全家族の収入

C．Y 変 数[a]
1．家長の教育
2．配偶者の教育（結婚の場合）
3．家長の雇用された産業
4．家長の職業
5．家長の生れた場所
6．農業所得
7．専業所得
8．社会保証所得
9．福祉所得

D．Z 変 数[b]
1．家長の週労働時間
2．家長の年労働週
3．家長の労働年数

E．マッチング水準（並びに目盛り調べ）[c]

1．カイ二乗	(.99)	：	18.3%
2．相　関	(.98)	：	19.4%
3．相　関	(.97)	：	25.8%
4．相　関	(.93)	：	17.6%
5．相　関	(.90)	：	13.5%
6．相　関	(.80)	：	3.9%
7．コホルト		：	1.5%

注：a．すべてのY変数は5%と15%標本に共通
b．5%と15%標本に共通したものの部分リスト
c．目盛り調べは，指示された水準でマッチされる家計の全数のパーセント
出典．op. cit. Nancy Ruggles, Richard Ruggles, Edward Wolff, Merging Microdata.

対する現実のものと比較される。

彼等は，二つの同時分布を比較するために「チョウ・テスト」を使用した。[12] 「チョウ・テスト」というのは，１つの標本を使用して推定される１方程式の係数が，異なった標本で推定される同一方程式の係数と比較される場合の１つの回帰技術である。その検定は，係数の全集合，或は，部分集合が二つの推定において統計的に異なっているかどうかを決める。彼等は，原標本に関する回帰の係数推定値を，１つ或はそれ以上の帰属（マッチされた）変数の原変数に代替された場合のものと比較するために「チョウ・テスト」を使用した。係数の二つの集合が統計的に異なっているならば，その時，その指示は，帰属された同時分布が現実の同時分布のよき模写ではないということである。

この基準は回帰技術の特性制限を持つことが記されるべきである。特に，回帰は，回帰変数の同時分布の第一次，第二次積率を捉える総括統計量である。（それは，平均と共分散マトリックスである。）

現実と帰属された同時分布間の関係の包括的様相を得るためにX，Y，Zとコホルト変数の種々の結合に関する「チョウ・テスト」をラシした。総計して，彼等は６つの回帰集合をラシした。各セットにおいて４つの方程式をラシした。先づ，ＰＵＳ15%標本からの変数が使用された。第２に，左辺の変数が15%標本から抽出され，右辺の変数は，マッチされた15%記録からであった。第四に，すべて変数は５%標本から抽出された。彼等は，それぞれの対の方程式に関して，「チョウ・テスト」をラシし，６つの個別的「チョウ・テスト」を生じた。各回帰の観察数は6341であった。

彼等は労働経済学の論文に共通して見出される方程式を使用した。最初の方程式は，学校の修業年数（ｓ）に関する収入の１つの対数回帰(Log E)であった。収入はX変数で修業年数はY変数であった。この「チョウ・テスト」のF統計量は第７表に示される。第７表の一番上で左の記入値は，現実の15%標本からの修業年に関しての収入の回帰と収入が原変数で修業年数が帰属変数である場合の同一方程式の回帰を比較した結果を示す。F統計量は，係数のなかに如何なる有意差をも示さない。第２の方程式のなかでは，収入の代りに所得（もう１つ別の）X変数を代入した。そして対の回帰係数推定値

第7表　PUS標本5%－15%のマッチされた標本に関する回帰ランからのチョウ・テストF統計量

方程式1：　$Log(E)=\beta_0+\beta_1 S+u\ (E>0)$

	(E_{15}, S_5)	(E_5, S_{15})	(E_5, S_5)
(E_{15}, S_{15})	2,140	1,027	2.515
(E_{15}, S_5)	——	0.449	0.081
(E_5, S_{15})	——	——	0.468

方程式2：　$Log(Y)=\beta_0+\beta_1 S+u\ (Y>O)$

	(Y_{15}, S_5)	(Y_5, S_{15})	(Y_5, S_5)
(Y_{15}, S_{15})	0.180	0.923	1.930
(Y_{15}, S_5)	——	1.238	1.545
(Y_5, S_{15})	——	——	0.774

方程式3：　$Log(E)=\beta_0+\beta_1 A+\beta_2 S+u\ (E>O)$

	(E_{15}, S_5)	(E_5, S_5)	(E_5, S_5)
(E_5, S_{15})	1.738	0.823	2.154
(E_{15}, S_5)	——	0.327	0.078
(E_5, S_{15})	——	——	0.388

方程式4：　$Log(Y)=\beta_0+\beta_1 A+\beta_2 S+u\ (Y>O)$

	(Y_{15}, S_5)	(Y_5, S_{15})	(Y_5, S_5)
(Y_{15}, S_{15})	0.947	0.202	1.031
(Y_{15}, S_5)	——	1.612	1.172
(Y_5, S_{15})	——	——	0.634

方程式5：　$Log(E)=\beta_0+\beta_1 A+\beta_2 S+\beta_3 H+\beta_4 W+\beta_5 R+\beta_6 M+u\ (E>O)$

A．すべての係数に関するチョウ・テスト

	$(E_{15}, (S, H, W)_5)$	$(E_5, (S, H, W)_{15})$	$(E_5, (S, H, W)_5)$
$(E_{15}, (S, H, W)_{15})$	0.669	1.419	2.059
$(E_{15}, (S, H, W)_5)$	——	1.933	1.975
$(E_5, (S, H, W)_{15})$	——	——	2.556*

B．(S, H, W) に関するチョウ・テスト

	$(E_{15}, (S, H, W)_5)$	$(E_5, (S, H, W)_{15})$	$(E_5, (S, H, W)_5)$
$(E_{15}, (S, H, W)_{15})$	0.551	1.047	1.238
$(E_{15}, (S, H, W)_5)$	——	1.901	1.484
$(E_5, (S, H, W)_{15})$	——	——	1.192

方程式6：　$W=\beta_0+\beta_1 A+\beta_2 S+\beta_3 M(W>O)$

	(W_{15}, S_5)	(W_5, S_{15})	(W_5, S_5)
(W_{15}, S_{15})	0.977	0.268	1.242

(W_{15}, S_5)	——	0.482	0.882
$(W_5, S_{15})'$	——	——	2.627*

（注）Log：対数，E：収入，Y：所得，A：年令，H：労働時間，W：労働週，R：人種，M：婚姻状態，u：ランダム誤差項

＊　5％有意水準で有意に差あり
＊＊1％有意水準で有意に差あり

出典：op. cit. Nancy Ruggles, Richard Ruggles, Edward Wolff, Merging Microdata.

に有意差を見出さなかった。第3の方程式では，年令（A）のコホルト変数（15%並びに5％の標本記録において同一値を持つ）を最初の方程式に加えたが，如何なる有意差も見出さなかった。第4の方程式では，相似な結果を持つ第二の方程式に年令が附加された。

第5番目の方程式で，収入の対数が修業年数に回帰された。年令，人種(R)並びに結婚状態(M)，それらはすべてコホルト変数である。1週の労働時間(H)，1年の労働週(W)はZ変数である。最初の制約において，収入，修業年数，労働時間，労働週は，15%の標本から導出された。第2の規定では，収入変数は15%の記録から抽出され，他の変数は5％記録から抽出された。第3番目に於て，収入は5％記録から，他のものは15%記録から抽出された。第4番目に於ては，すべての変数が5％記録から抽出された。（残り3つの変数，年令，人種，婚姻状態は，二つのファイルに於て同じ値を持つコホルト変数である）。

すべての係数均等に関する「チョウ・テスト」は，係数が5％有意水準で有意に異なっている場合の唯一の例を暗示した。修業年限，労働時間，労働週の係数均等に関する「チョウ・テスト」は，何れも例外ではなかった。第6番目の方程式で，労働週，Z変数は，二つのコホルト変数とY変数に関して回帰された。そこには，係数が有意に異なっている場合の唯一の例があった。

これらの統計的結果は，マッチング処理から生ずる帰属された同時分布が，現実の同時分布のよき模写であるという強い支持を提供する。彼等が行った42の「チョウ・テスト」のうち，二つのみが原標本変数を含む回帰と，標本

と帰属変数の両者を含む回帰の間で，推定係数に有意差があった。このテストは，選別―統合のマッチング処理が，多種類の統計的適用に対して，信頼すべき総合的データ源を提供出来ることを暗示している。

(注)

(1) op. cit. Nancy Ruggles, Richard Ruggles, Edward Wolff, Merging Microdata.

(2) 税申告に関する現実のソーシャル・セキュリイテイ・ナンバーを使用する社会保障局による特殊なランで，ＩＲＳのファイルに性，人種が附加された。

(3) マッチング区間を構成するのに使用された方法と合理性を持つ完結された結論に対しては，第1表と関連説明を参照。

(4) 1つの起り得る修善方法は，それらのrank orderや百分位の分布を基礎にX_aとX_bの分布を調整することである。実際に，Bファイルのn番目百分位数値がAファイルのn番目百分位数に等しいとして取り扱われる。Bファイルに対するマッチング区間が発生される前にX_b値がX_a変数によって等しい値に変換される。

(5) Bファイルがランダム標本である場合にのみ成立する。Bファイルが層別標本であるならば，標本のウエイトを基礎に，確率基礎の上にB記録が選択される。

(6) この理由は，マッチング水準を再び規定することによって，そのマッチが改善されるということである。例えば，マッチの50%が水準4，相関（0.80）で起ると仮定，然も，水準3は相関（0.90）であると仮定せよ。これは，マッチの大きな割合が0.80と0.90の間にある相関水準，0.85で起ることを暗示する。0.85の相関水準は，0.80の相関水準よりも狭いマッチング区間を生ずるので，AとB記録間のマッチはより近いX値で起り，マッチは改善される。

(7) Sims, Christopher A., Comment to Okner's "Constructing a New Data Base from Existing Microdata Sets", Annals of Economic and Social Measurement, July 1972.

(8) これは確かに情報の消失であるから，条件的同時分布については何も云えない。殆んど全ての帰属処理の場合における如く，その関係は確率的であると仮定する。

(9) Wolff, Edward N., The Goodness of Match, NBER Working Paper No.72, December 1974.

(10) 5%と15%という指定は，それぞれの質問をうけた母集団のパーセントを示す。消費者耐久財保蔵に関する情報を家計のバランス・シートの構成に対して移転するのにこのマッチを行った。

(11) 実際に，すべてのY変数は同じである。

(12) Chow, G., Test of Equality between Sets of Coefficient in Two Linear Regression, Econometrica, 28, pp.591－605.