論　説

統計的照合手法の基礎理論と最近の適用例

美　添　泰　人

1　はじめに

わが国の官庁統計を利用したミクロデータ分析の事例は次第に増加しているが，一橋大学経済研究所附属社会科学統計情報研究センターが，研究者のためのミクロデータ分析拠点として設立されるなど，これからは個体情報をできる限り秘匿した上で，一般的な研究目的でのミクロデータ利用の道が開けてくるものと考えられる。

その際に有効な手法として海外の一部の国で利用されているものが統計的照合の技術である。本研究では，官庁統計の一般公開ミクロデータ (public-use micro data) を念頭におきながら，統計的照合に関わる基本的な問題を検討する。なお，一般に公開することを目的として一部の匿名化した標本を提供する場合，英国では匿名化標本データ (sample of anonymised data, SAR) という呼び方がなされている。

以下，本研究に関連する主要な概念を要約しておく。なお，概念の詳細な整理と関連する話題については美添・荒木 (2000) を参照されたい。

同一の個体 (統計単位) に関する複数のファイルを結合することによって，個々のファイルからは得られない，追加的な情報を得ることができる。そこでは一つのファイルに他のファイルの情報を付加，すなわち**補完** (impute) することになる。元のファイルを**基準ファイル** (base file) と呼び，新たな情報を提供するファイルを**参照ファイル** (reference file) と呼ぶ。参照ファイルは複数のこと

もある。なお，これらを受領ファイル (recipient file)，提供ファイル (donor file) と呼ぶこともある。

この際，ファイルを結合する方法を大別して，完全照合 (exact matching) と統計的照合 (statistical matching) があり，その具体的手順は以下のとおりである。

完全照合　完全照合とは，一連世帯番号などの個体識別符号情報ないし個体を識別できる名称や符号等の情報を用いて複数のファイルの個体を照合し，新たなファイルを構築することを指す。なお，完全照合の場合はレコードリンケージ (record linkage) と呼ぶことも多い。

1. **照合キーの決定**　レコードを結合する際の情報として判別力の高い変数を照合キー (matching key variable)，または共通変数 (common variable) とする。名称，所在地などには小さな誤り等があるため，実際上は照合に際して誤差許容範囲を設ける必要がある。一般に識別力が高く，誤判定率が低い変数により多くの重み付けをする。性別の一致のように重要度の高い場合については，不一致の程度も考慮した重み付けを行うことも考えられる。このような重みを用いて総合得点を求める。

2. **判定基準の決定**　一致，不一致，判定不能の決定を行う。最良の一致とされたレコードが容認できるものかを決定するための判定限界を事前に決めておく必要がある。ある区間を設けて判定不能とし，追加情報を得た後に再検討を行うこともある。

個別情報秘匿の観点からは，識別情報が増加している照合されたファイルにおいては個体が特定化され易くなり，危険性が相対的に高くなるという点に配慮が必要とされる。

企業や事業所の場合，特に規模が大きい場合には複数の独立した調査に回答していることは珍しくない。特に，工業統計，商業統計，事業所・企業統計など，全数調査の場合には，異なる調査および異時点間の調査を組合わせることによって貴重な分析が可能となる。

一方，抽出率が低い世帯調査の場合には複数の統計調査の間では重複するがないため，参照ファイルとして税務情報などの情報を対象としない限り，完全照合が用いられる機会はほとんどない。わが国における例外的な状況として，平成14年に改定される前の「家計調査」と「貯蓄動向調査」では，実際に一部の世帯を共通して調査しているため，調査客体を識別するための情報から完全照合が可能である。

世帯調査の場合には，複数のファイルを組合わせて利用可能な変数を増加させたミクロデータファイルを得るための手法として，完全照合の代わりに統計的照合と呼ばれる手法が利用されることがある。

統計的照合　統計的照合とは，異なったファイルに存在する類似の個体(統計単位)に関するデータを補完することによって，結合された情報を含むファイルを作成することであり，完全照合が不可能な場合の対案として，あるいは個体情報を秘匿するために敢えて完全照合に代わるものとして，利用されている。統計的照合の手順は以下のように整理できる。

1. 母集団および調査単位の概念を調整した後，基準ファイルと参照ファイルで定義が同一であること，その他の変数と高い相関があることなどを考慮に入れて共通変数を選択する。共通変数は一般には多変量であり，その適当な関数として定義されることもある。

2. 基準ファイルの各個体に対して，参照ファイルの個体の情報を補完する。個体の選定は共通変数が最も類似しているという基準で決定される。共通変数の選択，および基準ファイル個体の変数と参照ファイル個体の変数の間の距離関数には，さまざまな変種が考えられる。

さらに，統計的照合は各変数の分布情報をどのように利用するかによって，以下の2つの方法に分類される。

1. **制約付照合** 参照ファイルに含まれていた各変数の周辺分布が不変という制約の下で照合を行うものである。最も簡単な例では，基準ファイルと

参照ファイルのサイズが等しい場合に，参照ファイルのレコードを1回ずつ用いれば実現できる。

2. **無制約照合** 参照ファイルに含まれていた変数の分布に対して制約を設けないもので，多くの単純な照合はこの型となる。複数回補完されるレコードや，全く用いられないレコードが存在することを許容する照合方法である。

統計的照合の利用状況は欧州諸国とアメリカ・カナダで異なっている。官庁統計を対象とする，アメリカやカナダでは統計的照合の主要な役割として個体情報の秘匿があげられている。この記述はU. S. Department of Commerce, Office of Federal Statistical Policy and Standards (1980) による。

統計的照合に関する広汎な評価を行ったRodgers (1984) によれば，アメリカにおける従来の統計的照合は理論的な基礎や経験的な根拠なしに発展してきたが，最近になってこの欠陥を修正するための試みが始められたとされる。また「統計的照合は，2つのデータファイルにある同じ個体を結び付ける完全照合の類似として発展してきたものである」と認識されている。

2　統計的照合の新らしい展開

最初に，美添・荒木 (2000) にも記した筆者の考えを紹介しておく。

> 統計的照合の価値は，形式的な統計分析を可能とするという点にあるのではなく，生成されたミクロファイルがさまざまな目的に利用可能となるという柔軟性にある。たとえば，さまざまな変数の分布の形状や散布図の形状を明らかにすることを通して外れ値処理などの手順を検討することは，ミクロデータがあって初めて可能となる分析である。一方，いくつかのモデルを前提とした仮説の検定などを目的とした回帰分析のために単に分散共分散行列を求めるだけなら，統計的照合は非効率的な方法である。

わが国では，従来アメリカとカナダにおける統計的照合については知られていたが，欧州における事例に関してはほとんど知られていなかった。それは，欧州諸国においては商業的な目的で統計的照合が利用されており，一般には手法が公開されず，その結果として学術誌にも掲載されていないことが，主な原因である。最近になって，Rässler (2002) は，従来の統計的照合手法を展望する中で，メディアと消費者サーベイを中心とする欧州の実情を紹介した。また不十分ではあるがカナダにおける実験にも言及している。

2.1 統計的照合の標本理論

2.1.1 照合の手順

ここで，統計的照合で用いる記号を定義する。基本ファイルと参照ファイルに含まれるかどうかによって，変数は s, t, u の 3 つのグループに分けられる。

- 基本ファイル (A: base file) は大きさ n_A であり，含まれる変数として s と u がある。ここで u は世帯属性などの基本的な変数であり，s は収入，支出などの詳細な変数を想定している。さらに，詳細な調査が必要となる変数 t は，このファイルでは調査されていないために収録されていない。
- 参照ファイル (B: reference file) は大きさ n_B とし，変数 t と共通変数 u が収録されているが，s は観測されていない。

統計的照合で利用される距離関数は i を基本ファイル A の個体, j を参照ファイル B の個体とするとき，$d(i, j)$，または明示的に変数を表示して $d(u_i, u_j)$ と表わされる。なお，共通変数が多変量の場合，変数に依存する適当なウェイト $g(u)$ を用いることもある。

照合の制約に関しては，ファイル B の個体を何回補完に用いるかによって 1 対多 (polygamy, bigamy) および 1 対 1 (monogamy) という場合を区別することがある。また，B ファイルから単一の個体を選んでその値を補完するのではなく，複数の個体の平均値を利用する next three (一般には nearest n) などと呼ばれる手法もある。

重複した個体の利用を制限するためには，たとえば一度採用された個体には距離関数に任意の大きな値を加えるなど，罰金関数 (penalty function) を利用することもある。ただし，その場合には，結果は当該の個体が照合される順序に依存することもある。

2.1.2 照合過程の基本的仮定

母集団における同時分布 $p(s,t,u)$ に関する統計的推論の問題と考えるとき，基本的な仮定として「各ファイルの標本は，それぞれ独立に，同じ母集団から抽出された（無作為抽出）結果とする。」をおく。

基本ファイル A の分布 $p(s,u)$ に参照ファイル B から変数 t を追加して得られた分布を $\tilde{p}(s,t,u)$ と表すと，この分布は照合の方法によって定まる。

u を照合キー変数として t を補完することは，基本ファイル A における観測値 (s,u) に対して，同一の u の値を持つ任意の個体をファイル B から無作為に抽出し，その t の値を代入することを意味する。すなわち

$$\tilde{p}(s,t,u) = p(s,u)p(t \mid u)$$

となる。これは $\tilde{p}(s,t,u) = \tilde{p}(s,t \mid u)p(u) = p(s \mid u)p(t \mid u)p(u)$ または $\tilde{p}(s,t \mid u) = p(s \mid u)p(t \mid u)$ とも表される。

離散形の場合には同じ u をもつ個体が参照ファイルに存在する可能性が高いが，連続形の場合には同一の値を持つ個体が参照ファイルに存在する確率は 0 だから，実際は最も近い個体が選ばれる。このような最近隣個体による照合の影響は，後に数値例によって具体的に紹介する。

周辺分布に関しては $p(s)$, $p(s,u)$ などの分布が保存されることは自明である。具体的には次の関係が成立する。

$$\tilde{p}(s) = p(s), \quad \tilde{p}(t) = p(t), \quad \tilde{p}(s,u) = p(s,u), \quad \tilde{p}(t,u) = p(t,u)$$

しかし，補完された変数を含む同時分布に関しては，

$$\tilde{p}(s,t,u) = p(s,u)p(t \mid u) = p(u)p(s \mid u)p(t \mid u)$$

だから，$\tilde{p}(s,t,u)$ と $p(s,t,u) = p(u)p(s,t \mid u)$ が一致するためには，次の条件が必要かつ十分である。

$$p(s \mid u)p(t \mid u) = p(s,t \mid u)$$

これが**条件付独立性** (Conditional Independence) と呼ばれる条件であり，この仮定 (CIA: Conditional Independence Assumption) が成立する場合に限り，統計的照合が正当化されるものである。

この概念は自明とはいえ，初期の適用例ではその必要性は明確に認識されていなかった。たとえば Sims (1972a, 1972b) の指摘を参照のこと。

2.1.3 補完されたファイルの分散・共分散

分散共分散については以下のような結果が得られる。g を任意の関数とするとき，まず $\tilde{E}g(s) = Eg(s)$, $\tilde{E}g(s,u) = Eg(s,u)$ などは明らかである。ここで $\tilde{E}$ は $\tilde{p}(s,t,u)$ の下での期待値を表す。

他方，統計的照合による共分散は $\widetilde{\mathrm{cov}}(s,t) = \mathrm{cov}(s,t) - E[\mathrm{cov}(s,t \mid u)]$ となる。

これは次のように示される。まず $\mathrm{cov}(s,t \mid u) = E(st \mid u) - E(s \mid u)E(t \mid u)$ の右辺の u についての期待値は，第 1 項が

$$E\{E[st \mid u]\} = \int E[st \mid u]\, p(u)\, du = \int \left[\int \int st\, p(s,t \mid u)\, dsdt \right] p(u)\, du$$
$$= \int \int \int stp(s,t,u) dsdtdu = E(st)$$

および第 2 項が

$$E\{E[s \mid u]E[t \mid u]\} = \int \left[\int s\, p(s \mid u)\, ds \int t\, p(t \mid u)\, dt \right] p(u)\, du$$
$$= \int \int \int st\, p(s \mid u)p(t \mid u)p(u)\, dsdtdu$$
$$= \int \int \int st\, \tilde{p}(s,t,u) dsdtdu = \tilde{E}(st)$$

と評価される。これと

$$\mathrm{cov}(s,t) = E[\mathrm{cov}(s,t \mid u)] + \mathrm{cov}[E(s \mid u), E(t \mid u)]$$

および $\mu_s = E(s) = \widetilde{E}(s) = \widetilde{\mu}_s$，$\mu_t = \widetilde{\mu}_t$ を用いると

$$E\Big[\mathrm{cov}(s,t \mid u)\Big] = E\Big[E[st \mid u] - E[s \mid u]E[t \mid u]\Big] = E(st) - \widetilde{E}(st)$$
$$= \Big(E(st) - \mu_s\mu_t\Big) - \Big(\widetilde{E}(st) - \mu_s\mu_t\Big) = \mathrm{cov}(s,t) - \widetilde{\mathrm{cov}}(s,t)$$

が導かれる。

特に

$$\widetilde{\mathrm{cov}}(s,t) = \mathrm{cov}[E(s \mid u), E(t \mid u)]$$

である。以上の結果から Rässler は共分散が一致するためだけなら条件付独立性より弱い条件，すなわち「平均的な条件付独立性」が満たされればいいと主張している。しかし，これは自然な仮定とは思えないため，実用的とは言えない。結局，条件付独立性に関しては，類似の調査における経験からその妥当性を判断するのが現実的である。

2.1.4 Propensity score による照合

Propensity score とは因果関係の分析において提示された考え方である。Rosenbaum and Rubin (1983, 1985) では，管理された対照を用いた実験が利用できない場合に，処理群と類似の対照群を生成する方法として用いている。おおまかに言えば，$e(x) = \Pr(\text{Treated} \mid x)$ と定義される propensity score とは，処理を受けていない大きな対照群の中から，処理群と類似の共変量 (covariate) x を持つ個体を対応させる方法である。これは処理群のサイズが相対的に小さな場合，特に有効である。この手法を統計的照合に応用することができるが，後の例で見るように，効果的な場合は限られている。なお propensity score の有効性に関しては，別な機会に検討したい。

2.1.5 正規分布の仮定

正規分布の想定の下では，

$$\mathrm{cov}(s,t \mid u) = \Sigma_{st} - \Sigma_{su}\Sigma_{uu}^{-1}\Sigma_{ut}$$

というよく知られた関係が成立する。すでに述べたように

$$\widetilde{\mathrm{cov}}(s,t) = \mathrm{cov}[E(s \mid u), E(t \mid u)]$$

となるが，多変量正規分布の条件付期待値は

$$E(s \mid u) = \mu_s + \Sigma_{su}\Sigma_{uu}^{-1}(u - \mu_u)$$

となることから

$$\mathrm{cov}[E(s \mid u), E(t \mid u)] = (\Sigma_{su}\Sigma_{uu}^{-1})\Sigma_{uu}(\Sigma_{uu}^{-1}\Sigma_{ut}) = \Sigma_{su}\Sigma_{uu}^{-1}\Sigma_{ut}$$

すなわち

$$\widetilde{\mathrm{cov}}(s,t) = \Sigma_{su}\Sigma_{uu}^{-1}\Sigma_{ut}$$

が得られる。

この式の解釈は「不適切な共通変数として s, t のいずれかと無相関であるような u を選ぶと，真の共分散 $\mathrm{cov}(s,t)$ によらず，照合データにおける相関はゼロになる」ということである。この問題については，美添・荒木 (2000) の例を参照のこと。

2.2 統計的照合の問題

2.2.1 統計的照合と無回答の問題

よく知られているように，欠測値に関して Rubin の導入した概念として以下のような整理がある。

1. MCAR (Missing Completely at Random)
2. MAR (Missing at Random)
3. MNAR (Missing Not at Random)

統計的照合が実行される状況を欠測値の問題と考えると，欠測のメカニズムは調査対象の属性によらず，調査設計者の決定が原因である。このことから，欠測メカニズムとしては MCAR が仮定できることになる。したがって，この仮定の下で一般的な欠測値問題の視点から検討することが可能となる。

2.2.2 統計的照合に固有の識別可能性問題

統計的照合の問題では，変数 (s,t) は同時に観測されていないため，本質的な問題として，同時に観測されることのない変数同士の関係を明らかにできないという意味での識別問題がある。

なお，(s,t) の同時分布に関しては，状況によっては若干の情報を得る可能性もある。例として，3変数 (s,t,u) の分散共分散行列として

$$\mathrm{cov}(s,t,u) = \begin{pmatrix} 1 & 0.9 & 0.8 \\ 0.9 & 1 & \mathrm{cov}(s,t) \\ 0.8 & \mathrm{cov}(s,t) & 1 \end{pmatrix}$$

を想定すると，分散共分散行列が非負値定符号であることから，未知の母数である $\mathrm{cov}(s,t)$ に対して，次のような限界が与えられる。

$$0.72 - \sqrt{0.0684} \leq \mathrm{cov}(s,t) \leq 0.72 + \sqrt{0.0684}$$

この不等式は母集団の分散共分散情報を利用したものである。しかし，ファイル A, B のサイズ n_A, n_B が大きくなって，分散共分散が正確に推定できたとしてもこの限界は小さくならないという意味で「識別不能」である。

なお，ここで条件付独立性 $\mathrm{cov}(s,t \mid u) = 0$ が成立すれば，$\mathrm{cov}(s,t) = 0.72$ と「識別可能」になっている。

2.2.3 Rässler の例示

シミュレーションの結果を提示する前に，統計的照合の判断基準として，Rässler は4つの異なった水準で，分布および関連の復元が成功したかどうかを判断することを提案している。

水準 1. 真の値が復元される。$\tilde{t}_i = t_i$ という，完全な一致である。もちろん t が連続変数のときはこの概念は非現実的である。

水準 2. 真の同時分布が復元され，$\tilde{p}(s,t,u) = p(s,t,u)$ が成立する。統計的照合の最も重要な目的は，真の分布 $p(s,t,u)$ に従う単一のファイルを作

成することであり，結合されたファイルを利用して正確な統計分析が実行できることである。すでに確認したように，この性質は条件付独立性が満たされたときにのみ満たされる。

水準 3. 相関構造の保存。すでに見たように $\widetilde{\mathrm{cov}}(s,t) = \mathrm{cov}[E(s \mid u), E(t \mid u)]$ となるが，個別のファイルから得ることができる情報は $E(s \mid u)$ と $E(t \mid u)$ のみである。相関が復元できるためには「平均的な条件付独立性」が満たされなければならない。

水準 4. 周辺分布の保存。$\tilde{p}(s,t,u) = p(s,t,u)$ が最低限の要求である。

まず，Rässler (2002, pp. 22–25) は，条件付独立性が成立しない場合の統計的照合が失敗する例として，garlic pill （健康食品の一種，効果は不明）の購買に関する仮想的な例として，子供と高齢者に関して，消費者調査 (consumer panel) とテレビ視聴者調査 (television panel) を統計的に照合する問題を取り上げている。

この例では高齢者が多く購入する商品である一方，宣伝は子供にも高齢者にも同様に効果的であると想定している。2 つのファイルの照合に年齢，すなわち子供か高齢者か，という共通変数を利用すると，統計的照合の結果では，ほぼ条件付独立が実現されるが，もちろん，その結果は誤りである。この例のように，年齢などの社会経済変数だけでは，共通変数としては一般に不十分であり，購買行動を説明できない。このようなことから，ドイツでは消費者パネルの対象者に対してテレビ視聴に関する追加的な質問がなされていると言う。

続いて，Rässler (2002, pp. 36–42) は条件付独立性が成立していない状況で，最近隣法 (nearest neighbor, **nn**) と傾向スコア法 (propensity score, **ps**) の効果を確認し，また，いくつかの照合基準を比較するシミュレーションを行っているので，それを紹介しよう。ただし記号は本稿に合わせて変更している。

この例では，次の分散共分散行列に対して $u = (u_1, u_2)$ を共通変数として「条件付独立性を想定した統計的照合」を実施する。

$$\Sigma_{uu} = \begin{pmatrix} 1.0 & 0.2 \\ 0.2 & 1.0 \end{pmatrix}, \quad \Sigma_{us} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}, \quad \Sigma_{ut} = \begin{pmatrix} 0.8 \\ 0.6 \end{pmatrix}$$

ここで，真の共分散（相関係数）を $\sigma_{st} = \rho(s,t) = 0.8$ と想定する。さらに $\sigma_{ss} = \sigma_{tt} = 1$ とする。このとき，条件付相関係数は

$$\rho(s,t \mid u) = \frac{\sigma_{st|u}}{\sqrt{\sigma_{ss|u}\sigma_{tt|u}}} = 0.7129$$

となる。また，

$$\tilde{\sigma}_{st} = \Sigma_{su}\Sigma_{uu}^{-1}\Sigma_{ut} = \begin{pmatrix} 0.5 & 0.5 \end{pmatrix} \begin{pmatrix} 1.0 & 0.2 \\ 0.2 & 1.0 \end{pmatrix}^{-1} \begin{pmatrix} 0.8 \\ 0.6 \end{pmatrix} = 0.5833$$

である。採用した方法は以下の4通りである。

(1) Polygamy (nn): 最近隣法，ドナーの繰り返し利用を許す。

(2) Bigamy (nn): 最近隣法，2回までのドナーの利用を許す。

(3) Triple mean: 最も近い3つの個体の平均値を代入する。

(4) Polygamy (ps): 傾向スコア $\hat{e}(u)$ による最近隣法，ドナーの繰り返し利用を許す。

(5) Bigamy (ps): 傾向スコア $\hat{e}(u)$ による最近隣法，2回までのドナーの利用を許す。

具体的には，最初の3つの場合には距離関数は $d(u_i, u_j) = |u_{1i} - u_{1j}| + |u_{2i} - u_{2j}|$ を用い，後の2つでは $d(u_i, u_j) = |\hat{e}(u_i) - \hat{e}(u_j)|$ を用いている。

シミュレーションでは，平均ゼロで上記の分散共分散を持つ正規乱数を $n = 5000$ 個発生し，ランダムに二つのファイルを作成した上で，それぞれから s または t 変数を削除することによって基本ファイルと参照ファイルを作成する。したがって $n_A = 2500, n_B = 2500$ である。これを $K = 50$ 回繰り返して，通常の方法で推定値の平均と分散を評価している。

$$\hat{E}(\hat{\theta}) = \frac{1}{K}\sum \hat{\theta}_k, \qquad s(\hat{\theta})^2 = \frac{1}{K-1}\sum (\hat{\theta}_k - \hat{E}(\hat{\theta}))^2$$

ここでは標準誤差は省略したが，表2.1によればnn法は周辺分布のモーメントをよく再現している。また，条件付独立性は，いずれの照合手法でも達成さ

表 2.1 Simulation with nearest neighbor (nn) and propensity socore (ps)

When $\rho_{st|u} \neq 0$

Technique	$\hat{E}(\hat{\mu}_t)$	$\hat{E}(\hat{\sigma}_t^2)$	$\hat{E}(\hat{\sigma}_{u_1 t})$	$\hat{E}(\hat{\sigma}_{u_2 t})$	$\hat{E}(\hat{\sigma}_{st})$	$\hat{E}(\hat{\rho}_{st})$	$\hat{E}(\hat{\rho}_{st\|u})$
Polygamy (nn)	-0.0009	0.9939	0.7949	0.5953	0.5792	0.5801	-0.0006
Bigamy (nn)	-0.0012	0.9928	0.7944	0.5942	0.5776	0.5788	-0.0015
Triple mean	-0.0008	0.8755	0.7899	0.5909	0.5749	0.6135	-0.0010
Polygamy (ps)	-0.0003	0.9996	0.4149	0.3053	0.3007	0.3021	0.0016
Bigamy (ps)	-0.0023	0.9975	0.4146	0.3045	0.3002	0.3019	0.0016

表 2.2 Simulation with nearest neighbor (nn) and propensity socore (ps)

When $\rho_{st|u} \approx 0$

Technique	$\hat{E}(\hat{\mu}_t)$	$\hat{E}(\hat{\sigma}_t^2)$	$\hat{E}(\hat{\sigma}_{u_1 t})$	$\hat{E}(\hat{\sigma}_{u_2 t})$	$\hat{E}(\hat{\sigma}_{st})$	$\hat{E}(\hat{\rho}_{st})$	$\hat{E}(\hat{\rho}_{st\|u})$
Polygamy (nn)	0.0019	0.9883	0.7921	0.5960	0.5831	0.5853	0.0010
Bigamy (nn)	0.0019	0.9864	0.7912	0.5940	0.5822	0.5849	0.0016
Triple mean	0.0022	0.8745	0.7891	0.5924	0.5794	0.6183	0.0001
Polygamy (ps)	-0.0038	0.9977	0.3459	0.2642	0.2546	0.2552	0.0011
Bigamy (ps)	-0.0006	0.9995	0.3438	0.2634	0.2528	0.2535	0.0004

れている。すなわち $\hat{\rho}_{st|u} \approx 0$ が成立しているし，表には省略したが $\hat{\rho}_{st|u}$ の標準誤差は非常に小さい。

真の相関係数が $\rho(s,t) = 0.8$ であるのに対して，照合結果では 0.58 に近い値が得られているが，これは理論から予想された結果である。

他方 ps 法では条件付独立性は認められるが，(t,u) の同時分布は変化している。この例ではファイル A とファイル B という二つの群で u 変数の分布に差がないため，推定された傾向スコアは $\hat{e}(u) \approx 0.5$ となっている。したがって，$\hat{e}(u_i)$ と $\hat{e}(u_j)$ の近い個体を補完することは，ファイル A の個体 i に対してファイル B の個体 j を無作為に補完する意味になる。(t,u) の相関係数が小さくな

るのは当然である。

もうひとつのシミュレーションとして，$\mathrm{cov}(s,t)=0.6$ のケースも実行され，表2.2にまとめられている。この例では $\rho_{st|u}=0.0550$ となり，近似的に条件付独立性が満たされる。予想のとおり，nn法では (s,t) の相関係数に関して良好な照合結果が得られている。またps法で過小評価されることは前の例と同じ理由である。

なお，Triple meanなど，複数の個体を平均する補完（代入）法を用いた場合には，変数 t の分散は真の分散より小さくなる。一般に，k 個の個体を用いたnn法の性質を考えると，補完された変数 t の分散を理論的に導出することができる。まず，

$$\widetilde{\mathrm{var}}(t)=E\Big[\widehat{\mathrm{var}}(t\mid u)\Big]+\mathrm{var}\Big[\tilde{E}(t\mid u)\Big]$$

という基本的な関係において，$\widehat{\mathrm{var}}(t\mid u)=\mathrm{var}(t\mid u)/k$，および $\tilde{E}(t\mid u)=E(t\mid u)$ となることを用いると，$\widetilde{\mathrm{var}}(t)$ は次のように評価される。

$$\begin{aligned}\widetilde{\mathrm{var}}(t)&=\frac{1}{k}E\Big[\Sigma_{tt}-\Sigma_{tu}\Sigma_{uu}^{-1}\Sigma_{ut}\Big]+\mathrm{var}\Big[\mu_t+\Sigma_{tu}\Sigma_{uu}^{-1}(u-\mu_u)\Big]\\&=\frac{1}{k}\Sigma_{tt}+\left(1-\frac{1}{k}\right)\Sigma_{tu}\Sigma_{uu}^{-1}\Sigma_{ut}\end{aligned}$$

今の例では $\widetilde{\mathrm{var}}(t)=1/3+(2/3)0.8417=0.8944$ となり，シミュレーション結果は，これとほとんど一致している。このように分散が過小評価されることから，相関係数は過大評価となる。なお，Triple meansでは t の分散と共分散に過小推定の傾向が見える。

さらにpp法による照合では t の平均，分散が真の値に近く，条件付共分散がゼロに近いが，その他の推定量は明瞭な偏りを持ち，標本の変動も大きい。また，シミュレーションを理論値と比較するために簡単な t 値に準じた統計量 $\sqrt{k}(\hat{E}(\hat{\theta})-\tilde{\theta})/s(\hat{\theta})$ を試算したところ，nn法を用いたPolygamyとBigamyでは，すべてが絶対値で3より小さかったとしている。

なお，すでに指摘したように，Rässlerは s と t の共分散構造を復元するためには条件付独立性より弱い条件である「平均的条件付独立性」$E[\mathrm{cov}(s,t\mid u)]=0$

が満たされればいいことを強調しているが，このような条件は現実的とはみなせない。他方，条件付独立性なら，当該データの発生メカニズムに関する知識からその妥当性を判断する可能性があり，実用的な基準であろう。

2.3 欧州諸国における統計的照合の適用例

Rässler(2002) によれば，欧州諸国では，アメリカやカナダとは異なる視点から統計的照合の手法が発達してきたとされる。その多くは公開されていないため，Rässler による要約は内部資料による部分と，推測による部分からなっているという。

欧州諸国における統計的照合については，最終的には条件付独立性の議論に依存しているとされる。結論として，伝統的な手法では適切な共通変数を用いることで少なくとも水準 4 の妥当性は確保するように工夫されているらしい。しかしながら，原理的な問題として，用いられている共通変数が，同時に観測されない変数間の関連を正しく表現しているのかどうかは知ることができない。以下，この節の記述はすべて Rässler (2002) による。なお，欧州諸国では Data Fusion という名称が一般的とされる。

2.3.1 欧州諸国における統計的照合 (Data fusion)

ドイツ，フランス，イギリスにおいて，1960 年代において，ほぼ同時期に，相互に独立に，統計的照合の手法が適用されている。各国とも，消費者パネルとテレビパネルが，それぞれ別個に実施されており，これらの調査を組合わせることによってより多くの情報を抽出したいという要望があった。

イギリスでは Beale がテレビとメディアのデータの照合に線形計画法を利用し，フランスでは Boucharenc and Bergonier が分割，距離関数，無作為の順序での割り当てを用いた。ドイツでも Wendt が分割 (segmentation) したファイルを無作為に照合するという手法を用いたが，それは，その後の欧州における統計的照合で利用されているものとほとんど同じである。

歴史的にはドイツにおける多数の調査を実施してきたAG.MA (Arbeitsgemeinschaft Media Analyse, 出版社，テレビ・ラジオ，広告代理店の連合会）が読者を対象とした調査を他の調査と統計的に照合し，広範に利用したことがあるという。

それらの手法は基本的には水準4の基準を満たすように構築されており，ドイツにおけるtopological conceptやフランスにおけるFRF algorithm (Fusion sur Référenciel Factoriel，そこでは主成分分析やcorrespondence analysisが利用されているらしいが，詳細は未確認) が用いられているとのことである。

1980年代後半から，イギリス，ベルギー，フィンランド，スペインを含む欧州諸国では照合の実験プロジェクトを開始したが，その内容はAntoine and Santiniによって公表されている。イギリスの例では，TGI (Target Goup Index, メディア調査からの成果) とBARB (Broadcasters' Audience Research Boardの視聴情報) の統計的照合が行われたが，そこでは提供側と需要側のファイルの照合に11個の変数だけが利用可能であり，追加的にテレビ視聴と宣伝のウェイトに関する2つの共通変数が作成された。効果的な照合を実現するためには，このような共通変数が重要であることが強調されている。

いずれにせよ，これらの手法が広く用いられているのは，それがメディアの計画立案にとって最適であるというより，単一の情報源が存在しないことから唯一の現実的な解法であることによる，と結論されている。

2.3.2 Topological Concept

その内容は，クラスター分析によって導かれる“typology-based-solution”を利用することにあるとされるが，詳細は不明である。

まず，基本ファイルおよび参照ファイルを，性別などの先験的に重要と考えられる変数を用いて，分割 (segmentation) する。次に，参照ファイルBの各分割ごとに，変数t, uを用いた“typological analysis”（クラスター分析を含めた用法と思われる）によって，できるだけ同質の部分群に分割する。つぎに基本ファイルの各個体ともっとも近いクラスターを求め，その「クラスター平均」で補

完する。

2.3.3 特定変数の多重割当て

最近の欧州諸国では統計的照合の妥当性はキー変数の予測力に大きく依存していることが認識されてきた。そのため英国の例ではTV調査であるBARB (British Broadcasters' Audience Research Board) と英国の企業AGBによって実施されている大規模市場調査であるAGB Superpanelをリンクする前に，Superpanelの被調査者に対してTV視聴行動に関する情報を収集しており，これによって社会経済変数だけを利用するのに比較してはるかに高い説明力を与えることができるという。ここでの統計的照合の目的は，参照ファイルであるBARBにおけるコマーシャルを見る確率を基本ファイルであるSuperpanelの各人に補完することである。

消費者パネルの各個人について，追加的な調査によって得られた時間帯と番組に関する視聴行動を，機械的な測定によって得られるTVパネルの視聴行動と照合するが，1分ごとの対応はできないため，時間帯と番組に関しては約100変数を定義し，これに主成分分析を適用することによって，各人に30次元のスコアとして視聴者行動の特性を定義している。この特性を共通変数として，最近隣法によって3人の変数の加重平均を補完している。これがmultiple ascriptionと呼ばれる手順である。

2.4 アメリカ，カナダにおける無条件照合 (Statistical Matching)

アメリカとカナダに関しては，比較的情報の入手が容易であり，これまでにも主要な応用例は知られている。初期にはOkner (1972, 1974), Ruggles and Ruggles (1974) などをはじめとする研究があり，当時の理論的，実際的な状況はGoel and Ramalingam (1989) がモノグラフとして取りまとめている。

民間のデータを対象とする欧州諸国とは大きく異なり，アメリカとカナダでは連邦統計局が，異なる資料の情報を照合しようと試みている。たとえば1960年代のアメリカでは，1967 Survey of Economic Opportunityと1966 Tax fileの情

報を組み合わせている。このような統計的照合の導入は，コンピュータの発展とともに，1960年代から1970年代にかけて，アメリカとカナダにおいて，詳細なミクロデータに対する需要が高まった一方で，整合的かつ包括的な世帯の収入などに関する統計調査が存在していないということから，新たなミクロ分析のためのデータベースを構築する試みが行われたという事情による。

ここで1970年代に，アメリカにおいて「完全照合」の適用に対する法律および規制の厳格化があった。さらに，Privacy Act (1974, 個人情報保護法) およびTax Reform Act (1976, 税制改革法) によって，連邦政府の機関 (Federal agency) の所有するデータが他の政府機関のデータや他の組織のデータと結合されるときの規制が強められた。一方で，ミクロデータ分析の目的で複数の資料を照合する場合には，それぞれが大きなサイズであっても，同一の世帯が含まれていることはほとんどありえない。したがって，統計的照合が適用される必要がある。

これらの手法は，暗黙のうちに条件付独立性の仮定を利用している点で批判されてきた。この指摘は，主としてSims (1972a, 1972b) やRodgers (1984) によるが，Okner (1972a) その他の初期の統計的照合で用いられていた，観測値 (u, t) を無作為に (u, s) に組合せる手法に対する批判として，的を得たものであった。

Rodgers (1984) ではシミュレーションによってさまざまな照合手法および条件付独立性の仮定が，水準2および水準4を満たすかどうかについて検討された。Judkins (1998) は，アメリカにおいて統計的照合の重要性が低下したのは，この結果によるものであろうと指摘している。

カナダ統計局 (Statistics Canada) では1990年代に，統計的照合の技術に関して集中的な研究が実施された。Kovačević などは，「カテゴリカル制約付き照合」という新たな手法を提案し，公開ミクロデータ（1986 Public Use Micro File）とセンサス（1991 Census）を用いて，その他の手法との比較を行っている。そこで用いられた手法は，すべてホットデックである。この手法については，節を改めて紹介する。

ここでも，統計的照合は，税と移転の計画，公衆衛生と福祉，教育の達成程度など，政策の分析を目的としている。このような目的のためには広範なデータ

ベースが必要であり，カナダ統計局は異なる出所の情報を組合わせて構築している。たとえばSocial Policy Simulation Databaseは，統計局による調査結果からCanadian Survey of Consumer Finances（消費者金融資産調査），Canadian Family Expenditure Survey（家計収支調査）と行政記録であるCanadian Personal Income Tax Returns（納税申告書）とCanadian Unemployment Insurance Claim history（失業保険給付記録）を組合わせている。

欧州の統計的照合と比較すると，アメリカとカナダでは，大きいデータセットが少ない共通変数で照合されていることになる。コンピュータの処理能力が向上するとともに，大規模データセットと補助情報の利用が可能となってきているが，Rässlerによれば，欧州におけるような多数の変数を対象とする場合は，カテゴリカル制約つき照合のように補助情報ファイルを利用する手法は，依然として計算量の面からの障害となっているという。

3　カナダ統計局の手法

ここでは，入手可能な資料であるKovačević and Liu (1994)を参考にしながら，カナダ統計局の手法を紹介する。その前に，そこで利用される「伝統的な手法」について概観しておく。

無制約照合の問題点

無制約照合では，基本ファイルの各個体に対して，参照ファイルにおいて何らかの距離で最も近い個体を組み合わせるが，このときの問題は，統計的照合ファイルにおいて補完された変数の分布 $p(t)$，特に平均 $E(t)$ と分散 $\mathrm{var}(t)$ が参照ファイルとは違ってしまうことにある。

基本ファイルと参照ファイルが同一の母集団からの無作為標本であれば，理論的には補完された変数の周辺分布 $\tilde{p}(t)$ は参照ファイルと同じく，正しい母集団を反映する。したがって，問題は「経験分布」に差が生じるということである。

人工的な例として有名なRodgers (1984)およびRubin (1986)の例では，基本

ファイルは $n_A = 8$, 参照ファイルは $n_B = 6$ の個体からなる。共通変数 u としては性 u_1 と年齢 u_2 を用い，特に性を**照合クラス** (matching class) とする。すなわち男性と女性は別々に照合する。年齢 (u_2) に関しては通常の距離 $d_{ij} = |u_{i2}^A - u_{j2}^B|$ を用いる。通常の例では補完された変数 t の平均と分散は参照ファイルの平均，分散とは一致しないが，形式的な「平均の差の検定」を適用すると，有意な差は見出せない。

また u が与えられたときに s と t が条件付独立かどうか，すなわち $\rho_{ts \cdot u} = 0$ の検定は，回帰モデル $t = \beta_0 + \beta_1 u_1 + \beta_2 u_2 + \beta_3 s + u$ における係数に関する $\beta_3 = 0$ という仮説と同等である。Rubin (1986) の例でこの検定を行っても s と t の条件付独立性を疑わせる結果は生じない。

距離の定義を拡張したり，最小の d_{ij} に対応する個体を組み合わせる代わりに，許容量 (tolerance) δ を導入して距離が $d_{ij} + \delta$ の範囲内にあるすべての個体 j の中から無作為に組み合わせるという方法もあるが，基本的には照合方法には条件付独立性が想定されており，この手法の妥当性は，共通変数である u の説明力にかかっている。

3.1 制約付き照合の手順

制約付き照合 (Constrained matching) では「経験分布」を一致させるように t を補完する。

いま，基本ファイルにおけるウェイトを w_i^A，参照ファイルにおけるウェイトを w_j^B とする。たとえば，単純無作為抽出なら $w_i^A = N_A / n_A$，すなわち抽出率の逆数（線形推定の乗率と呼ばれることもある）であり，一般に，母集団推定は $\hat{\mu}_s = \sum_i w_i s_i$ という形式でなされる。

問題の設定は以下のようになる。基本ファイル A の第 i 個体に参照ファイル B の第 j 個体を組み合わせ，その観測値にウェイト w_{ij} を付与するものとする。このウェイトは母集団推計に利用されるものであり，これを適切に定めることが「制約付き照合」の課題である。

経験分布が一致するための条件は次の式で与えられる。

$$\sum_{j=1}^{n_B} w_{ij} = w_i^A, \qquad \sum_{i=1}^{n_A} w_{ij} = w_j^B$$

このとき，照合の良さを測定するためには次の目的関数を最小化することが自然であろう。

$$\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} w_{ij}\, d(i,j) \tag{1}$$

経験分布が一致し，かつ $w_{ij} \geq 0$ という条件の下で (1) 式を最小にする問題は，線形計画法の典型的な問題であり，これを解くことはそれほど難しくはない。特に u が 1 変量のときには，Rubin が提示した簡単な方法が適用できる。すなわち，ウェイトが整数であれば，各個体をウェイトだけ複製すなわち拡大 (explode) したそれぞれのファイルを作成し，共通変数 u について大きさの順に整列する。その上で，二つのファイルにおける個体を順番に組み合わせればいい。この方法では，一般にはすべての (i, j) について $w_{ij} > 0$ とはならず，少数の (i, j) の組合せ以外は $w_{ij} = 0$ となる。

共通変数が多変量で，u の次元が高いときには，1 変数ずつ，この方法を適用することができるが，もちろん，その結果は整列に利用する変数の順番に依存する。

なお，Rubin (1986) は制約付き照合も無制約照合と同様に，補完方法のひとつであり，両者ともに補完にかかわる誤差の評価を与えない点を批判して多重補完 (multiple imputation) を推奨している。

3.2 カテゴリ制約付き照合

ここではカナダ統計局 (StatCan) において試みられている手法である補助データファイル (auxiliary data file) の利用とカテゴリ照合 (categorical matching) の紹介を扱う。詳細な論文とされる Liu, Liu and Kovačević によるカナダ統計局の Working Paper は入手できていないが，要約に相当する Kovačević and Liu (1994)

に基づいて，カナダ統計局で実施されたシミュレーションの内容を紹介する。ただし記号はこれまでの表現に合わせて若干修正している。

以下では，変数 s, u を含む基本ファイル A および t, u を含む参照ファイル B のほかに，第3の補助ファイル C が存在する状況を想定する。補助ファイル C にはすべての変数 (s, t, u) が含まれるか，あるいは少なくとも (s, t) が同時に観測されているものとする。したがって，これまでの議論のように条件付独立性 (CI) の成立が統計的照合の前提条件であるが，補助的な情報を利用することによって，ある程度は CI が成立しない状況にも対応した照合が実現できる。

さらに，異なる調査では各個体のウェイトが違っているため，その調整がひとつの課題である。その上，カナダ統計局においては統計的照合に関して，いくつかの制約条件が課されている。たとえば Social Policy Simulation Data Base (SPSD) では，次の条件があるという。

(i) 参照ファイル B の条件付周辺分布 $p(t \mid u)$ をできるだけ保存し，歪みは最小限にとどめること。

二つのファイルにおけるウェイトが異なる場合には歪みが大きくなる可能性がある。SPSD にとって歪みが問題になる分布としては，$p(t)$, $p(t, u)$, $p(s, t, u)$ の三つが重要である。最初の二つは目標の分布 $p(t \mid u)$ に直接影響する。ファイル B は母集団からの無作為標本だから，照合されたファイルにおける分布 $p(t \mid u)$ は標本変動の範囲に収まるべきである。

(ii) 両方のファイルに含まれるすべての個体を利用すること。この要請は SPSD における独特のものでファイル B の情報を重視することによる。通常の統計的照合ではファイル A を補完すればよく，このような要請は不要である。

(iii) 基準ファイル A の拡大を最小に抑制し，照合されたファイルのサイズを操作可能な大きさに留めること。これはファイルの維持，管理と操作に関する費用の面からの制約である。

(iv) 照合ファイルのウェイトを整数にすること。このファイルを真の母集団からの標本とみなすとき，ウェイトは各レコードが代表する母集団の個体数を表すためとしている。

3.2.1 CIA に基づく照合手順

補助情報を利用する統計的照合の前に，まず，CIA に基づく照合の手順とウェイトの分割方法を紹介する。最初に，共通変数 u を K 個の照合階級 (matching class, データリンケージでは pocket，統計的補完では imputation class と呼ばれる）に分割する。もともとカテゴリカル変数であれば，u の値を用いるが，そうでない場合には適当な階級に分割することでカテゴリー変数 u^* を作成する。それぞれの階級 u^* においては距離関数は元の u とウェイト w によって定められる。照合は K 個の照合階級ごとに実行され，そのの手順は共通だから，以下 K を省略すると，次のように記述できる。

まず u 変数だけを考慮する場合は，距離の許容範囲法 (fixed distance tolerance) または最近隣法 (nearest available) が用いられる。前者は距離の上限を与えた上で，この境界内にある最も近い個体を用いて補完する方法である。しかし，すべての個体を利用するという制約の下では最近隣法が適当である。他方，ウェイトが問題となる場合には，u の相対累積ウェイト (RCW: relative cumulative weight value) を利用して t の値を補完する方法が提案されている。この方法は後に説明する。

u-距離法　一般的には，ある A レコードに対して u-距離 $d(u_i, u_j)$ を最小とする B レコードをひとつ補完する。複数ある場合には無作為に選べばいい。

いくつかの B レコードが用いられないこともあるが，それは要請 (ii) に反することになる。そのようなときには，利用されなかった B レコードに対して最近隣の A レコードを見つける。したがってひとつの A レコードに複数の B レコードが用いられる多重補完となるから，ウェイトの調整が必要である。

i 番目の A レコードに対して J_i 個の異なる B レコードが補完されたときは，

初期のウェイト w_i^A は対応するBレコードのウェイト $w_{i:j}^B$ を用いて次のように配分される。ここで，ファイルAの i レコードに対応しないファイルBのレコードについては $w_{i:j}^B = 0$ である。

$$w_{ij} = w_i^A \cdot w_{i:j}^B \Big/ \sum_{k=1}^{n_B} w_{i:k}^B$$

この方法によると周辺分布 $p(s, u)$ は保存されるが，Bファイルにおける条件付分布 $p(t \mid u)$ はゆがみを持つ。

周辺分布の整合性のためには，次の条件が満たされなければならない。

$$\sum_{i=1}^{n_A} w_{ij} = w_j^B, \quad \sum_{j=1}^{n_B} w_{ij} = w_i^A, \quad \sum_{i=1}^{n_A} w_i^A = \sum_{j=1}^{n_B} w_j^B$$

最適なウェイトは，$\sum_i \sum_j w_{ij} d_{ij}$ を以上の制約条件の下で最小にする線形計画法の解として得られる。

ウェイト分割法 (weight-split method)　ウェイトと u の情報をともに利用する方法としてウェイト分割法が用いられる。u の値が大きさの順に整列されていない場合は「無作為ウェイト分割」，整列されている場合には「順位ウェイト分割」となる。いずれの場合も最も近いRCWを持つファイルBのレコードから t の値が補完される。以下では，両方のファイルが u によって整列されている場合について記述する。

各照合階級において，それぞれのレコードのRCWを評価する。ここでレコード u_i のRCWは

$$F_i = F(u_i) = \sum_{j=1}^{i} w_j, \quad (i = 1, \cdots, n)$$

で定義される。著者たちは相対累積ウェイトと呼んでいるが，ウェイトの合計は母集団の個体数 N に一致することになる。これを用いて，ファイルの各照合階級をRCWの順位に整列する。照合するファイルのRCWを $\{F_i^A\}, \{F_i^B\}$ と表す。

順位による補完は以下の手順で行われる。第1段階（下向ステップ）として，すべてのAレコードに対して

$$F_{j-1}^{B} < F_{i}^{A} \leq F_{j}^{B}$$

となる j 番目のBレコードを補完する。

第2段階（上向ステップ）では，等しいRCW F を持つAレコードが存在しないようなBレコードだけについて，$F_i^A > F_j^B$ となる最初のAレコード i に対して t_j の値を補完する。

このようにして照合されたファイルの大きさは，$n = n_A + n_B - n_0$ となる。ただし，n_0 は $F_i^A = F_j^B$ となるようなレコードの数である。最後に，合成されたレコードのRCWは

$$F_{ij} = \min(F_i^A, F_j^B)$$

と定める。

こうして得られた合成ファイルにおける周辺分布 $p(s), p(t), p(u)$ が保存されることは容易に確かめられるが，実際的にはいくつかの欠点もある。

まず，得られた相対ウェイトが $w_i < 1$ となる場合を問題であるとし，このようなレコードを棄却する。次に合成されたファイルの大きさが問題で，通常は非常に大きなサイズとなる。そのため「逐次的ファイル縮小法 (sequential file reduction procedure) を適用する」と記されている。具体的な方法は示されていないが，すべてのレコードを利用するという制約条件の下で，周辺分布の多少の歪みを許容するものと考えられる。

最終的に得られたファイルのウェイトを決定するときは「距離による照合」と同じ方法を採用する。

3.2.2 補助情報が利用可能な場合の照合手順

補助ファイルCでは (s, t, u) または (s, t) が同時に観測されるものとする。

(s,t,u)-距離法　最初の段階で，ファイルAにファイルCのtの値を補完するために，最近隣レコードを特定する。ここで距離関数は，ファイルCとの間で利用可能な共通変数である(s,u)またはsのみによって定められる。このようにして，ファイルAの各レコードに対してファイルCからtの値を補完した中間ファイルが作成される。この段階ではファイルAのウェイトとサイズは変更されない。

次の段階で，(t,u)を共通変数としてファイルAとファイルBとの距離を評価し，ファイルBからtの値を補完する。その手順は，u-距離法と本質的に同じである。

ウェイト分割法$(s,t,u$順位$)$　補助ファイルCがウェイト情報も持っていれば，RCWが利用できる。この場合の中間的補完では最も近いRCWを持つレコードのtの値を補完する。共通変数が多次元の場合には，まずu_1で整列し，同じu_1の値を持つレコードについてはさらにu_2で整列する。以下同様にして，u,sのすべてについて順番に整列する。この段階で累積ウェイトRCWを定め，これを$F_{us}(u_i)$と表す。ここで添え字は整列の順番を表している。

このようにして作成される中間ファイルにはファイルAと同じサイズとウェイトを与える。つまり，補助ファイルCのウェイトは中間的補完の目的のためだけに利用し，ウェイトを変更するためには利用されない。

次の段階では中間ファイルとファイルBのウェイト分割による照合が適用される。利用される共通変数は(t,u)であるが，最初にtで整列し，次にuで整列する。したがってRCWは$F_{tu}(u_i)$である。この後の手順はCIAの場合と同様である。

3.2.3　カテゴリ制約つき照合手順

基本的な考え方は以下のとおりである。(i) 変数(s,t,u)を適当な最適分割の基準を利用してカテゴリ変数(s^*,t^*,u^*)に変換する，(ii) カテゴリ変数(s^*,t^*,u^*)の分布を推定する，(iii) 推定された(s^*,t^*,u^*)の分布を用いて，$B(t^*,u^*)$から

$A(s^*, u^*)$ への補完を行う。目的はカテゴリ間の関連を保存することにある。

まずファイルBによって条件付分布 $p(t^*|u^*)$ を推定することができる。ここでカテゴリ変数における条件付独立性

$$p(t^*|s^*, u^*) = p(t^*|u^*)$$

が想定されれば，$w_{s^*,t^*,u^*} = w_{s^*,u^*} \cdot w_{t^*|s^*,u^*} = w_{s^*,u^*} \cdot w_{t^*|u^*}$ から，次の比例関係が導かれる。

$$w_{s^*,t^*,u^*} = w^A_{s^*,u^*} \cdot (w^B_{t^*,u^*}/w^B_{u^*}) \tag{2}$$

このようにして作成されたファイルはカテゴリ分布として元のファイルの分布を保存している。ファイルA, Bにウェイトがない場合にはウェイトの代わりに(2)式においてレコード数を用いることができる。

照合の候補はファイルA, Bにおいて同一の u^* カテゴリに属するレコードである。補完はファイルAの各カテゴリ (s^*, u^*) ごとに独立になされ，そこではHOD法を用いるという。[1] ファイルの縮小とウェイトの調整も前と同様に行うことによって，中間ファイル $A'\{s, t, u, w'\}$ が生成される。

このファイル A' におけるカテゴリ (s^*, t^*, u^*) のウェイトを $w'_{(s^*,t^*,u^*)}$ とし，$w_{(s^*,t^*,u^*)}$ を(2)式で調整されたウェイトとする。ここですべてのカテゴリ (s^*, t^*, u^*) においてウェイトの差 $|w_{(s^*,t^*,u^*)} - w'_{(s^*,t^*,u^*)}| < 1$ であれば，中間ファイル A' は最終的な照合ファイルとみなされる。[2]この条件が満たされないときにはminimum move and split procedureを用いて差を縮小させる。[3]

変数 s, t, u を階級値を用いるなどしてカテゴリ化することから，Categorical matchingと呼んでいる。

---

1) HOD法はCIAによる照合の箇所で説明したと書いてあるが，見当たらない。Hot Deckでもないようで，意味不明である。また「同一の u^* カテゴリに属するレコードを候補にする」という説明と，「補完はファイルAの各カテゴリ (s^*, u^*) で独立に行う」という説明では若干矛盾する。この部分は著者に確認が必要。

2) 原文には $w_{(s^*,t^*,u^*)} - w'_{(s^*,t^*,u^*)} < 1$ とあるが，誤りと思われる。

3) このminimum move and split procedureという方法も詳細は不明。t^* が二つのカテゴリの場合には (s^*, t^*_1, u^*) と (s^*, t^*_2, u^*) の間でレコードを移動させる，複製する，ウェイトを分割する，階級区分を修正するなどが，その内容らしい。

最近のワーキングペーパー　カナダ統計局では2001年以降のWorking Paperは掲載されているが，そこには統計的照合関連の文献は見当たらない。また“constrained matching”というキーワードで検索しても1件も関連の記事が見つからない。Rässlerは内部資料を入手したと記述しているが，現時点ではカナダ統計局はこのような照合手法を多用していない可能性もある。

ただし，補助変数を利用した照合方法は自然なものであり，この項に関しては，もう少し検討が必要であろう。また，美添・荒木(2000)に紹介した実験から示されるように，過去の類似の統計によって「条件付独立性」が近似的に成立するようなキー変数群を特定し，それを利用した照合を行うという方法も現実的である。

参 考 文 献

[1] Goel, P. K. and T. Ramalingam (1989) *The Matching Methodology: Some Statistical Properties,* Lecture Notes in Statistics, Springer.

[2] Judkins, D. R. (1998) Not Asked and Not Answered: Multiple Imputations for Multiple Surveys: Comment, *Journal of the American Statistical Association*, vol. 93, 861–864.

[3] Kovačević, M. S. and T.-P. Liu (1994) Statistical Matching of Survey Datafiles: a Simulation Study, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 479–484.

[4] Okner, B. A. (1972a) Constructing a New Data Base from Merging Microdata Sets: The 1966 Merge File, *Annals of Economic and Social Measurement*, vol. 1, 325–341.

[5] Okner, B. A. (1972b) Reply and Comments, *Annals of Economic and Social Measurement*, vol. 1, 359–362.

[6] Okner, B. A. (1974) On Data Matching and Merging: An Overview, *Annals of Economic and Social Measurement*, vol. 3, 347–352.

[7] Rässler, S. *Statistical Matching*, Springer (2002)

[8] Rogers, W. L. (1984) An Evaluation of Statistical Matching, *Journal of Business and Economic Statistics*, vol. 2, 91–102.

[9] Rosenbaum, P. R. and D. B. Rubin (1983) The Central Role of the Propensity Score in Observational Studies for Causal Effects, *Biometrika*, vol. 70, 41–55.

[10] Rosenbaum, P. R. and D. B. Rubin (1985) Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score, *The American Statistician*, vol. 39, 33–38.

[11] Ruggles, N. N. and R. Ruggles (1974) Merging Microdata: Rationale, Practice and Testing, *Annals of Economic and Social Measurement*, vol. 3, 407–428.

[12] Sims, C. A. (1972a) Comments, *Annals of Economic and Social Measurement*, vol. 1, 343–345.

[13] Sims, C. A. (1972b) Rejoinder, *Annals of Economic and Social Measurement*, vol. 1, 355–357.

[14] 美添泰人・荒木万寿夫 (2000)「5.1 完全照合」,「5.2 統計的照合」, いずれも松田芳郎・伴金美・美添泰人編『ミクロ統計の集計解析と技法』(講座ミクロ統計分析第 2 巻) 日本評論社, 第 5 章「ミクロデータのリンケージ」に所収