統計数理(2003)
第 51 巻 第 2 号 321–335

特集「個票開示問題の統計理論」
[総合報告]

PRAM の理論とその実用上の諸問題

藤野 友和[1] ・垂水 共之[2]

(受付 2003 年 2 月 3 日；改訂 2003 年 7 月 25 日)

要　　旨

Post Randomization Method (PRAM)は，Kooiman et al. (1997)により提案された匿名化標本データ(マイクロデータ)に対する統計的開示制御の方法である．これはデータの公開者が，マイクロデータにおける各標本がとる値に対して，事前に定めた確率構造に基づいて攪乱を与えることにより，個体の再識別や情報漏洩の発生する危険性を低下させるものである．本稿ではこの方法について紹介し，実際の適用に関して考慮すべき事項や適用されたデータの分析結果に与える影響について議論する．また，現在開発を行っている PRAM のためのソフトウェア環境に関しての提案も行う．

キーワード：マイクロデータ，マルコフ連鎖，局所再符号化，EM アルゴリズム．

1. はじめに

現在行われている各種統計調査などにより得られた匿名化標本データ(マイクロデータ)は，ほとんどの場合，調査主体が目的とする調査項目に対する集計結果のみを外部に提供するという形をとっている．これらの統計調査は大規模のものになればなるほど，多額な経費と大きな労力をかけて実施されるが，様々な統計分析手法が確立し，個人レベルでの情報処理能力が大幅に向上した現在，調査主体だけでなく，外部の企業や研究者がマイクロデータを分析することを望むのは自然の流れである．しかしながら，マイクロデータを外部に公開すると，第三者により公開されたデータの中のあるレコードが，ある個人のものであると特定される(個体識別)可能性が生じる．個体識別の発生は，第三者が本来知り得なかったその個人の属性データや調査項目に対する回答内容を知ってしまう(情報漏洩)ことにつながる．これらのことは，個人のプライバシーの侵害に関する問題はもちろんのこと，情報漏洩が発生したということが世間一般に知れ渡ることによる統計調査に対する信頼の失墜という問題をも引き起こす．これを解決するために，マイクロデータに対して一定の処理を施し，個体識別が発生する可能性を低減させた上で公開するためのいくつかの方法(統計的開示制御)が研究されてきた．マイクロデータに対する統計的開示制御の方法は 2 種類に大別できる．一つはカテゴリ併合やトップコーディング，ボトムコーディングに代表されるようなデータに攪乱を与えない方法である．もう一つはランダム化応答やデータスワッピング，ノイズの付加などのようなデータに攪乱を与える方法である．本稿で紹介する Post Randomization Method (PRAM)は後者の部類に属する．PRAM は Kooiman et al. (1997)により提案された方法で，データの公開者が，マイクロデータにおける各標本がとる値に対して，事前に定めたマルコフ遷移行列(PRAM 行列)に基づいて攪

[1] 福岡女子大学 人間環境学部: 〒813–8529 福岡市東区香住ヶ丘 1–1–1; fujino@fwu.ac.jp
[2] 岡山大学 環境理工学部: 〒700–8530 岡山市津島中 3–1–1; tarumi@ems.okayama-u.ac.jp

乱を与えることにより，個体識別や情報漏洩の発生する危険性を低下させるものである．

Warner (1965)は，デリケートな質問を対面式で行う場合に，調査員にもその回答内容が明らかにならないための手法としてランダム化応答を考案した．これは，ある質問に対して，事前に定めた確率によりその質問に答えるか，逆の質問に答えるかを決定するものである．確率構造により，回答内容が攪乱されるという意味では PRAM と類似の手法であるため，ランダム化応答に関する文献は PRAM に関する議論を行う際には参考になることが多い．しかしながら，あらかじめ計画した調査に対してしか適用できないランダム化応答に対して，PRAM は既存のカテゴリデータについては全て適用可能であり，データに従って確率構造を決定できるため，適用範囲が広い．

Gouweleeuw et al. (1998)は，PRAM の紹介を行うと共に，Kooiman et al. (1997)で提案された Invariant PRAM を適用するための PRAM 行列をより一般化したものを示した．さらにはオランダ国民の旅行調査(Dutch National Travel Survey)への PRAM の適用例を示し，PRAM により攪乱される期待度数との関連で考察を与えている．また，PRAM を適用する際には，PRAM 行列をどのように決定するかが重要な問題になる．これに関しては，前述の Invariant PRAM の他に，Willenborg (2000)や Willenborg and De Waal (2000)の第 5 章において，行列の正則性や危険なカテゴリの組み合わせに対する安全性を条件として，PRAM を適用することによる情報の損失を最小にするような最適化問題を解くことで，PRAM 行列を決定するという方法についての議論が行われている．Hout (1999)によるレポートでは，PRAM が適用されたデータと PRAM 行列に基づく統計解析の方法について詳細に述べられており，EM アルゴリズムによって PRAM が適用される前のデータに対する分割表のセル確率を最尤推定する方法が示されている．さらには PRAM が適用されたデータに対する対数線形モデルやロジスティック回帰による分析も EM アルゴリズムにより修正されるという示唆を与えている．Fujino and Tarumi (2001)は PRAM が適用されたデータに対して，数量化の各手法を実行した例を示し，考察を与えた．しかし，Hout (1999)では 2×2 分割表の場合のサンプルデータ，Fujino and Tarumi (2001)は簡単な人工データのみを数値例として扱っており，実際のマイクロデータへの応用を考えるとより大きな実データによる数値的検討が必要とされるであろう．Fujino and Tarumi (2002)は実データに対して簡単な数値実験を行った．本稿では，PRAM についての理論的背景を簡単に紹介すると共に，2000 年に実施された岡山行動圏調査のデータを利用して，PRAM が適用されたデータに対する統計解析が EM アルゴリズムによって十分修正されるかどうかをシミュレーションを使って得られる各種統計量の分布として評価する．特に，PRAM の適用方法によるそれらの違いに関する検討も行う．

PRAM の原理は比較的単純であるが，大規模なデータに対して PRAM 行列を定めて PRAM を適用したり，PRAM が適用されたデータの分析を EM アルゴリズムによって修正したりするというような作業にかかる労力は非常に大きい．さらに，統計にあまり詳しくない一般の利用者にとって PRAM を適用されたデータを扱うことは困難であると考えられる．Willenborg and Hout (2000)は，このような問題を解決するためのソフトウェアの必要性を指摘している．後半では，我々が開発している PRAM を運用するためのソフトウェアのプロトタイプを示し，PRAM を利用する環境についての提案といくつかの考察を与える．

2. PRAM の原理

2.1 PRAM 行列

本節では，PRAM を適用する際に必要ないくつかの定義について述べる．まず適用する変量が一つの場合について考える．K カテゴリを持つカテゴリ変量を A，PRAM が適用された

あとのカテゴリ変量を A^* とするとき，A に対する PRAM 行列 P_A は要素 (k,l) に

$$p_{kl} = P(A^* = l | A = k) \tag{2.1}$$

を持つ $K \times K$ のマルコフ遷移行列として定義される．すなわち，N 個の個体を含むマイクロデータにおける変量 A に PRAM を適用するということは，各個体の変量 A がとる値を，対応する P_A の行における確率に従って変更するということを意味する．

2.2 頻度ベクトルの推定

大きさ K のベクトル T_A, T_{A^*} をそれぞれ，A, A^* の頻度を示すベクトルとし，$T_A(k), T_{A^*}(l)$ によりその第 k 要素を表すものとする．T_A が与えられたとき，T_{A^*} の各要素 $T_{A^*}(l)$ は，二項分布 $B(T_A(k), p_{kl}), k = 1, \ldots, K$ の和の分布に従う確率変数の実現値であると考えられる．よって，

$$E(T_{A^*} \,|\, T_A) = P_A^T T_A \tag{2.2}$$

が成り立つので，T_A が未知であると考えると

$$\hat{T}_A = (P_A^T)^{-1} T_{A^*} \tag{2.3}$$

は，T_A の不偏推定量となる．すなわち，PRAM 行列と PRAM が適用されたデータを受け取ったデータの利用者は，これらの情報から PRAM が適用される前のデータにおける頻度の不偏推定量を計算することができる．しかしながら，逆行列をかけるという操作に伴い，頻度の推定量として負の値を生じる可能性がある．これは，度数表に対する統計解析を行う場合などには不都合である．これを避けるため，データの公開者によって PRAM 行列 P_A が

$$P_A^T T_A = T_A \tag{2.4}$$

を満たすように定められれば，T_{A^*} 自身が T_A の不偏推定量となる．PRAM 行列がこの条件を満たすとき，PRAM は特に Invariant PRAM と呼ばれる．Invariant PRAM を適用するための PRAM 行列の定め方としては 2 通りの方法が提案されている．Gouweleeuw et al. (1998) により提案された PRAM 行列は次のようなものである．T_A の要素を大きさの順に並び替えて，1 以上の頻度を持つ要素の個数を K_0 とする．このとき，任意の $0 < \theta < 1$ に対して，

$$p_{kl} = \begin{cases} 1 - \theta T_A(K_0)/T_A(k) & \text{if} \quad l = k \ \text{ and } \ l, k \le K_0 \\ \theta T_A(K_0)/((K_0 - 1)T_A(k)) & \text{if} \quad l \ne k \ \text{ and } \ l, k \le K_0 \\ 1 & \text{if} \quad l = k \ \text{ and } \ l, k > K_0 \\ 0 & \text{otherwize} \end{cases} \tag{2.5}$$

と P_A の要素を定めると，これによる PRAM は Invariant PRAM になる．ここで，データの公開者が事前に定めるパラメータ θ は頻度が 0 の T_A の要素以外で最小の頻度を持つ要素に属する個体が，同じ要素に留まらない確率であると解釈され，大きい値に定めるほどデータの安全性は高まる．このように PRAM 行列の特性を解釈することができ，さらに θ を与えることで一意的にデータに対して PRAM 行列が定まるということがこの方法の利点である．ただし，この方法によって定められた PRAM 行列で PRAM を適用した場合，値が変化する個体数の期待値は $K_0 T_A(K_0)\theta$ であり，実際のマイクロデータにおいては多くの場合，$T_A(K_0) = 1$ となることが予想され，カテゴリ数未満にしか設定することができない．

De Wolf et al. (1997) は Invariant PRAM を実現する方法として two stage PRAM を提案した．これは，任意の PRAM 行列 P_A に対して，$\tilde{P}_A^T (P_A^T T_A) = T_A$ となるように，$\tilde{P}_A$ を定めて，$R = P_A \tilde{P}_A$ により PRAM を適用するというものである．しかしながら，これにより得られた

PRAM 行列の性質は明らかではない．以降では，式(2.5)により設定された PRAM 行列による PRAM のことを Invariant PRAM と呼ぶことにする．

2.3 2 変量以上への適用

これまでは，単一の変量に対して PRAM を適用することを考えてきたが，実用上は複数の変量に対して適用することになる．今，PRAM を適用する p 個の変量 $A_1,\ldots,A_p$ に対して，適用後の変量を $A_1^*,\ldots,A_p^*$ とする．このとき，各個体の取る値の組み合わせに対する推移確率を，

$$p_{(k_1,\ldots,k_p),(l_1,\ldots,l_p)} = P(A_1^* = l_1,\ldots,A_p^* = l_p | A_1 = k_1,\ldots,A_p = k_p) \tag{2.6}$$

のように表現し，全ての値の組み合わせに対して推移確率を与えれば本質的には単一の変量に対する PRAM と同様に考えることができる．これにより得られた PRAM 行列を $P_{A_1\cdots A_p}$ とする．任意の推移確率について，

$$\begin{aligned} &P(A_1^* = l_1,\ldots,A_p^* = l_p | A_1 = k_1,\ldots,A_p = k_p) \\ &\quad = P(A_1^* = l_1 | A_1 = k_1,\ldots,A_p = k_p)\cdots P(A_p^* = l_p | A_1 = k_1,\ldots,A_p = k_p) \end{aligned} \tag{2.7}$$

が成り立つとき，PRAM の適用は各変量について独立であるという．また，任意の変量 A_i について，

$$P(A_i^* = l_i | A_1 = k_1,\ldots,A_i = k_i,\ldots,A_p = k_p) = P(A^* = l_i | A_i = k_i) \tag{2.8}$$

が成り立つとき，PRAM の適用は各変量について non-differential であるという．PRAM の適用が各変量に対して独立でかつ non-differential であるとき，

$$\begin{aligned} &P(A_1^* = l_1,\ldots,A_p^* = l_p | A_1 = k_1,\ldots,A_p = k_p) \\ &\quad = P(A_1^* = l_1 | A_1 = k_1)\cdots P(A_p^* = l_p | A_p = k_p) \end{aligned} \tag{2.9}$$

が成り立つ．以降ではこの場合を単に独立であると呼ぶ．この場合は，各変量に対してそれぞれ PRAM 行列 $P_{A_i}, i = 1,\ldots,p$ を与えて PRAM を適用することと同等になる．これらの PRAM 行列から，各変量の組み合わせを単一の変量とみなした場合の PRAM 行列 $P_{A_1\cdots A_p}$ は，各 PRAM 行列のクロネッカー積

$$P_{A_1\cdots A_p} = P_{A_p}\otimes\cdots\otimes P_{A_1} \tag{2.10}$$

により得られる．

3. EM アルゴリズムによる頻度の推定

式(2.3)による頻度ベクトルの不偏推定を実行すると，負の頻度を生じる場合があり分析を行う場合には都合が悪い．Hout (1999)は EM アルゴリズムによって，PRAM を適用する前の頻度ベクトルのセル確率を最尤推定する方法を提案した．最尤推定されたセル確率に基づく期待頻度をもって元の頻度の推定量とすることにより，負の頻度を生じない推定を実行することができる．ここでは，単一の変量の場合についてのみ述べるが複数の変量の場合についても同様に適用することができる．

データの利用者には，PRAM 行列 P_A と攪乱されたデータ，すなわち攪乱された頻度ベクトル T_{A^*} のみが渡されており，これらから元の頻度ベクトル T_A のセル確率 $\phi = (\phi(1),\ldots,\phi(K))$ を最尤推定することが目的である．N をデータに含まれる個体数とするとき，T_{A^*} はパラメータとして N と表 1 のようなセル確率を持つ多項分布に従う確率変数の実現値であると仮定する．ϕ に関する尤度を直接最大化するのは困難であるので，A と A^* の分割表 T_{AA^*} を表 2 の

表 1. 分析者に渡る観測（不完全）データ.

カテゴリ	セル確率	観測データ
1	$\sum_{j=1}^{K}\phi(j)p_{j1}$	$T_{A^*}(1)$
2	$\sum_{j=1}^{K}\phi(j)p_{j2}$	$T_{A^*}(2)$
⋮	⋮	⋮
i	$\sum_{j=1}^{K}\phi(j)p_{ji}$	$T_{A^*}(i)$
⋮	⋮	⋮
K	$\sum_{j=1}^{K}\phi(j)p_{jK}$	$T_{A^*}(K)$

表 2. 想定する完全データ T_{AA^*}.

カテゴリ	セル確率	完全データ
$(1,1)$	$\phi(1)p_{11}$	$T_{AA^*}(1,1)$
$(1,2)$	$\phi(1)p_{12}$	$T_{AA^*}(1,2)$
⋮	⋮	⋮
(i,j)	$\phi(i)p_{ij}$	$T_{AA^*}(i,j)$
⋮	⋮	⋮
(K,K)	$\phi(K)p_{KK}$	$T_{AA^*}(K,K)$

ようなセル確率をパラメータとして持つ多項分布からの実現値としたものを完全データと考えて EM アルゴリズムを実行する．すなわち，完全データに対する条件付き尤度を繰り返し最大化することによって ϕ の最尤推定量を得る．実際のアルゴリズムは次のようになる．

$$\text{初期値}: \phi^{(0)}(i) = \frac{T_{A^*}(i)}{N} \tag{3.1}$$

$$E \text{ ステップ}: T^{(v)}_{AA^*}(i,j) = \frac{\phi^{(v)}(i)p_{ij}}{\sum_{k=1}^{K}\phi^{(v)}(k)p_{kj}} T_{A^*}(j) \tag{3.2}$$

$$T^{(v)}_{A}(i) = \sum_{j=1}^{K} T^{(v)}_{AA^*}(i,j) \tag{3.3}$$

$$M \text{ ステップ}: \phi^{(v+1)}(i) = \frac{T^{(v)}_{A}(i)}{n} \tag{3.4}$$

4. 数値的検討

はじめに述べたように，Hout (1999)は 2×2 の分割表においては，EM アルゴリズムによって対数線形モデルやロジスティック回帰による分析が修正されることを例示した．その中で，大規模なデータによる検証や多変量解析において分析結果が修正可能かどうかについての調査の必要性を述べている．そこで本節では実際に行われた統計調査のデータを利用して，PRAM が適用されたデータに対する分析がどの程度影響を受けるか，また EM アルゴリズムによってどれくらい修正されるかを検証する．さらに，PRAM の適用が独立でない場合と独立な場合での影響の違いについても調べる．

利用したデータは 2000 年に実施された岡山行動圏調査により得られたものである．この調査は 1979 年から実施されており，岡山県とその周辺地域の住民を対象に，基本情報，通勤・通学圏，医療圏，交際圏，商圏，観光圏などに関するアンケート調査が行われている．2000 年の調査で得られた有効回答数は 7797 であった．ここでは，調査項目のうち，「年齢」「性別」「職業」をキー変数として選び，これらの変数に PRAM を適用する．カテゴリ数は無回答を含み，それぞれ 7,3,8 カテゴリである．なお，これらのキー変数の組み合わせにおいて，標本一意となっているのは 20 個体である．

適用方法としては独立でない場合と独立な場合の 2 通りを考える．独立でない場合には，キー変数の値の組み合わせをカテゴリとして持つような単一の変数に対して PRAM を適用する．

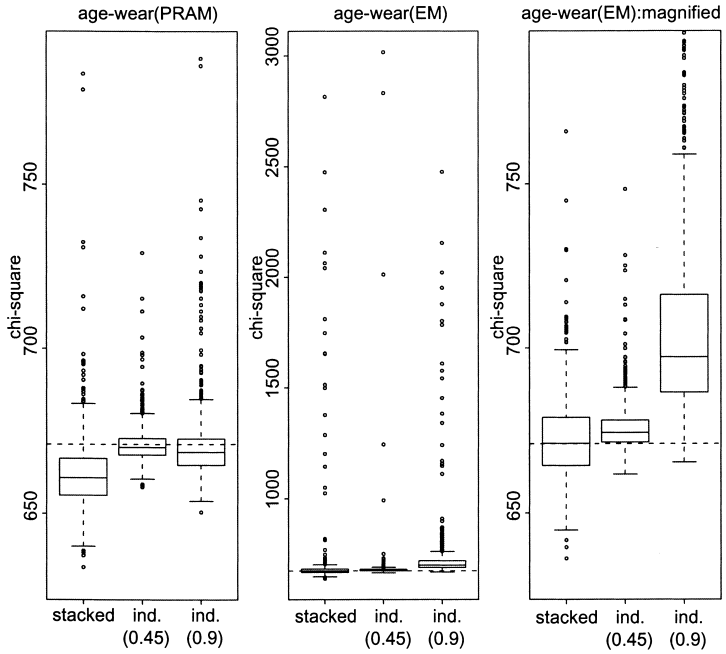


図 1. 年齢-洋服の購入に関するカイ 2 乗統計量の分布.

今回は特に Invariant PRAM を $\theta = 0.9$ として適用した(方法 1). このとき, キー変数の値が変化する期待個数は 88.2 (1.1%)である. 比較を行うために, 独立な場合にはこれと同程度のキー変数の値が変化する期待個数を持つように Invariant PRAM を各変数に対して $\theta = 0.45$ として適用した(方法 2). さらに, 各変数に対して $\theta = 0.9$ として独立に Invariant PRAM を適用した場合(方法 3)についても検討している. それぞれにおいてキー変数の値が変化する期待個数は 85.25 (1.0%)と 170 (2.1%)である. 方法 1 のように PRAM が適用された場合, 適用されたデータのキー変数同士の分割表がそれ自身, 元のデータの分割表の不偏推定量となるため, これらの変数に関する分析には PRAM が適用されたデータをそのまま利用することにより, 平均的には元データによる分析結果と同様の結果が得られることが期待される. よって, PRAM が適用された変数とそうでない変数について同時に分析するとき, PRAM がどの程度分析結果に影響を与えるかが興味の対象となる. そこで, PRAM が適用されていない変数として洋服を購入した店舗形態(7 カテゴリ)を考え, PRAM が適用された変量との分割表に対する分析について考察を行う. また, 分析により得られた統計量を分布として評価するために, 前述のような方法で PRAM を元データに対して 1000 回適用した結果に対して分析を行った.

図 1 は, 各 PRAM の適用方法に対する, 独立モデルを仮定した場合の年齢と洋服の購入店舗形態に関するカイ 2 乗統計量の分布を示している. 左側の図は PRAM を適用したデータから計算した統計量の分布であり, 2 番目の図は EM アルゴリズムにより推定した分割表に基づいて計算したものである. 3 番目の図は, 2 番目の図に対して中央値付近の拡大を行ったものである. 全ての図において, 点線が元データに基づく値 671 を示している. EM アルゴリズムを適用した場合には, 元の値から大きく外れたものがいくつか生じるが, 中央値付近の拡大図を見てみると方法 1 においては平均的には修正されているように見える. しかしながら, 方法

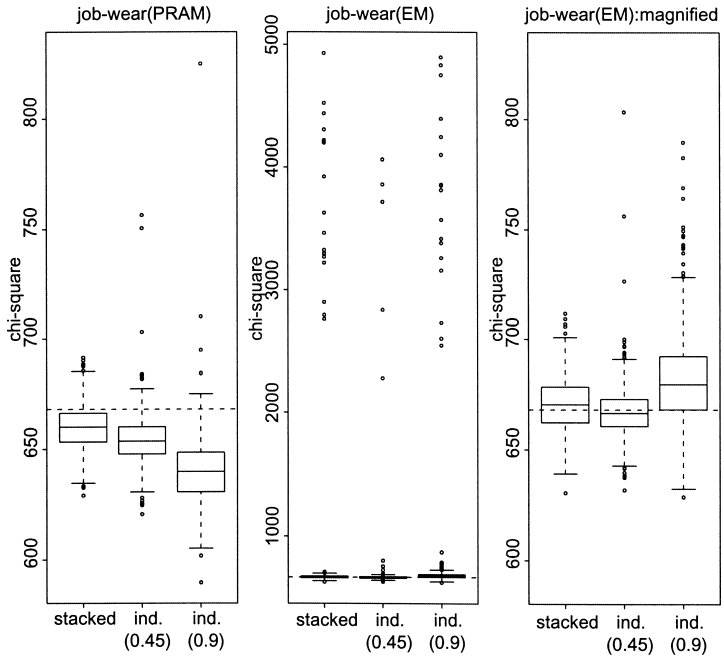


図 2. 職業–洋服の購入に関するカイ 2 乗統計量の分布.

3 においては PRAM を適用した場合，EM アルゴリズムを適用した場合のいずれにおいても元データによる値からは乖離している．図 2 は，職業と洋服の購入店舗形態に関して同様の処理を行った結果を示している．この場合の元データに対する値は 668 である．

次に，PRAM が多変量解析に与える影響を見るため，個体数，変量数の多い問題に対してよく適用される数量化 II 類に関して調べた．ここでは，洋服の購入店舗形態を外的基準，PRAM が適用された 3 つのキー変数を説明変数として扱い，数量化 II 類における固有値問題の最大固有値である相関比の分布を評価した(図 3)．この結果に関してもカイ 2 乗統計量の場合とほぼ同様の傾向を示している．ただし，2 番目の図に関しては，EM アルゴリズムにより推定された分割表を完全データとみなして標本相関比を計算しており，直接母相関比を最尤推定しているわけではない．従って，ここで示される相関比の標準誤差には PRAM による標準誤差だけでなく，標本相関比自体の標準誤差も含まれていることに注意しなければならない．

EM アルゴリズムを適用したデータによる統計量の分布は，方法 1 においては平均的には改善する傾向がみられるものの，大きく外れた値がいくつも出現する．これに対して PRAM が適用されたデータ自体を用いて分析した場合には，分析結果の解釈が大きく変わるほどの影響は出ておらず，EM アルゴリズムによる推定を行うことが常によい結果を与えるとは限らないことが分かる．また，独立な場合とそうでない場合の間には，分析に与える影響としては大きな違いは見られない．すなわち，遷移する個体数が同程度の場合には，変数の組み合わせに対して，少数個体を含むセルに属する標本の遷移確率，すなわち安全性を制御できるという点では方法 1 による適用が望まれる．これに対し，変数の組み合わせやカテゴリ数が多い場合には PRAM 行列が大きくなるため，これを避ける手段として方法 2 による適用が考えられる．

以上の結果を見ると，EM アルゴリズムによる分析の修正はあまり効果的に行われないとい

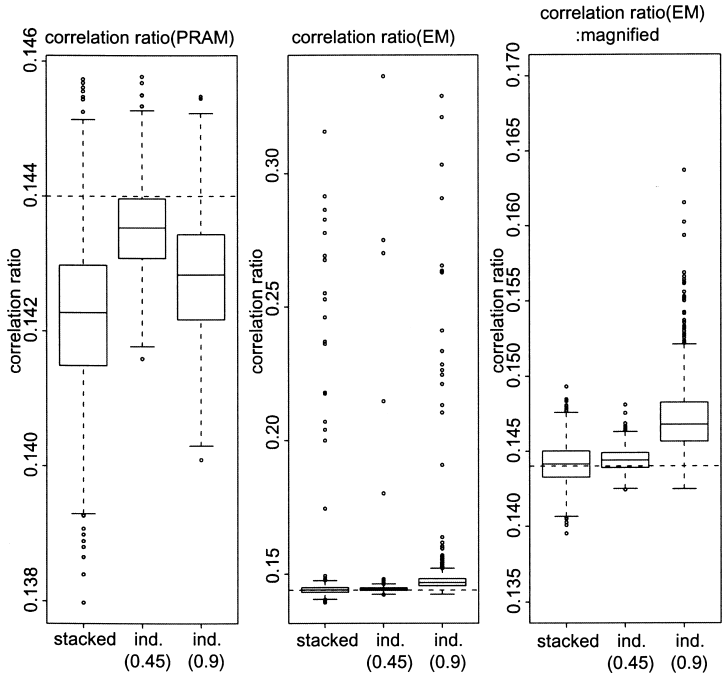


図 3. 数量化 II 類における相関比の分布.

うことになる．しかし，これらの方法においては遷移する個体数が標本総数に対して十分小さい．そこで，参考のために，同様のキー変数の組み合わせに対し，全てのセルに対して個体が同じセルに留まる確率を 0.5，遷移する確率を 0.5（セル毎の遷移確率は等確率）とした PRAM 行列による PRAM を 1000 回適用し，同様の統計量の分布と EM アルゴリズムにより修正されたデータによる統計量の分布を調べた（図 4）．これを見ると明らかにどの統計量においても EM アルゴリズムによる分析の修正が効果的であることが分かる．

5. サポート環境の開発

統計的開示制御の手法や開示リスクに関する研究と共に，マイクロデータを提供または利用する環境を構築することは，マイクロデータの利用促進にとって重要であると考えられる．特に PRAM を利用する場面においては，マイクロデータの提供者，利用者双方にデータに対する一定の処理が要求される．これらの処理にかかる負担を軽減することが一つの課題となるが，本節ではソフトウェアを利用してこの問題を解決する方法を提案する．

われわれは，マイクロデータの提供者，利用者の双方のための PRAM を利用するためのソフトウェアをそれぞれ開発した．提供者向けのソフトウェアを PPE（PRAM provider's environment），利用者向けのソフトウェアを PUE（PRAM user's environment）と呼ぶことにする．これらは Windows 環境で動作するソフトウェアであるが，Linux 環境への移植は容易である．

PPE は，CSV 形式で保存されているマイクロデータファイルを読み込んで，これに対して乱数を発生させることにより PRAM を適用し，PRAM 行列と攪乱されたデータファイルを出力するものである．ここで，PRAM 行列をどのように設定するかが問題となるが，PPE では

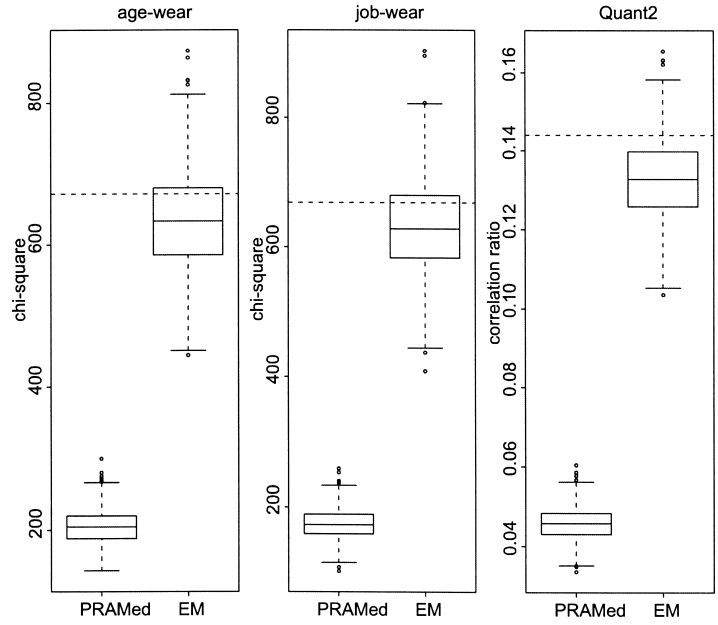


図 4. 遷移個体数が多い場合の各種統計量の分布.

以下の方法をサポートしている.

1. あらかじめ作成した任意の PRAM 行列を与えるファイルを読み込み，これを PRAM 行列とする
2. 式(2.5)による Invariant PRAM 行列をウィザード形式で設定する
3. 1 により読み込まれた PRAM 行列に対して two stage PRAM を実行するための PRAM 行列を生成する

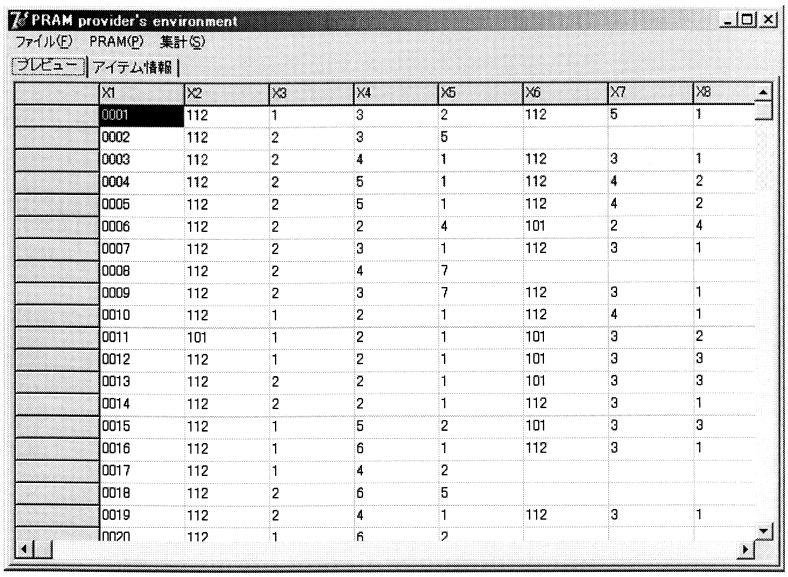
いずれの場合も最初に PRAM を適用する変数と，独立に適用するか否かというオプションの設定が要求される．その後，1, 3 の場合は独立であれば各変数に対してそれぞれ PRAM 行列のファイルを，そうでなければ単一の PRAM 行列のファイルを読み込む作業を行う．2 の場合は独立であれば各変数に対して θ を，そうでなければ単一の θ の値を設定することになる．最後に出力ファイルを指定してから PRAM を実行する．実行結果は攪乱されたデータを保存するファイルと PRAM に関する情報を保存するファイルにそれぞれ記録される．また，PRAM を適用することにより値が変わった個体数とその全体の標本数に対する割合が最後に表示される．

PPE により出力されたファイルが一般向けとして公開されることになるが，これらのファイルを処理するために，ユーザーは PUE を利用することができる．PUE の持つ機能は以下の通りである.

1. 単純集計，クロス集計機能
2. 集計表に対する統計解析機能

PUE の機能の特徴はその集計機能にある．集計を実行する際に，PUE は必要に応じて EM アルゴリズムを適用して PRAM を適用する前の集計表の各セルにおける期待度数を計算する.

すなわち，Invariant PRAM があるキー変数の組に対して適用されたときに，その変数の部分集合に関する PRAM が適用されたデータにおける集計表は元データの集計表に対する不偏推定量となっているため，EM アルゴリズムは適用されない．また，独立に Invariant PRAM が適用されている場合は適用されている変量の単純集計のみ，PRAM が適用されたデータにおけ



	X1	X2	X3	X4	X5	X6	X7	X8
	0001	112	1	3	2	112	5	1
	0002	112	2	3	5			
	0003	112	2	4	1	112	3	1
	0004	112	2	5	1	112	4	2
	0005	112	2	5	1	112	4	2
	0006	112	2	2	4	101	2	4
	0007	112	2	3	1	112	3	1
	0008	112	2	4	7			
	0009	112	2	3	7	112	3	1
	0010	112	1	2	1	112	4	1
	0011	101	1	2	1	101	3	2
	0012	112	1	2	1	101	3	3
	0013	112	2	2	1	101	3	3
	0014	112	2	2	1	112	3	1
	0015	112	1	5	2	101	3	3
	0016	112	1	6	1	112	3	1
	0017	112	1	4	2			
	0018	112	2	6	5			
	0019	112	2	4	1	112	3	1
	0020	112	1	6	2			

図 5. PPE 実行画面～データプレビュー.

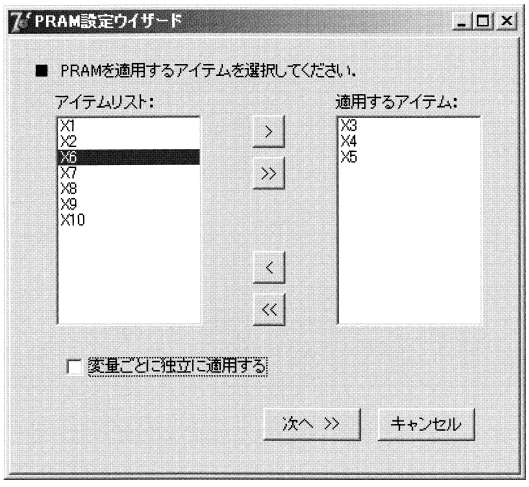


図 6. PPE 実行画面～適用方法と適用変数の選択.

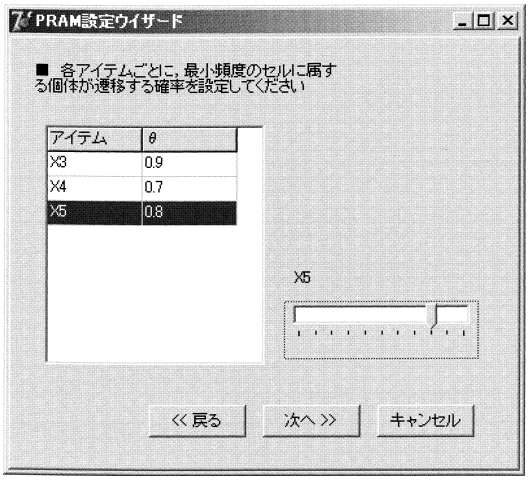


図 7. PPE 実行画面〜Invariant PRAM における θ の設定.

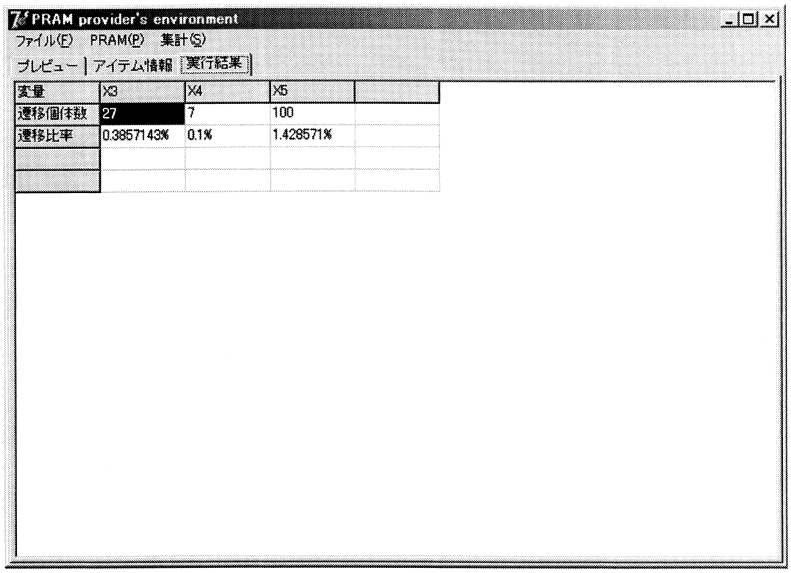
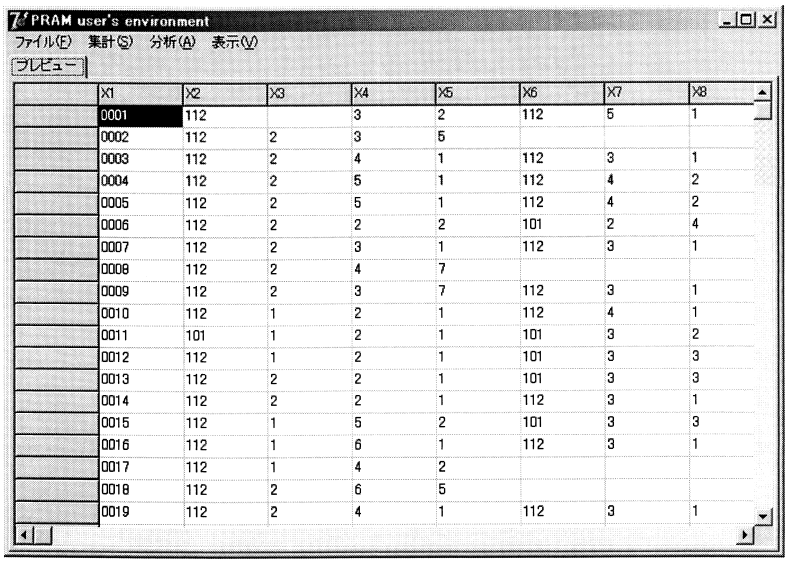


図 8. PPE 実行画面〜適用結果：遷移標本数の表示.

る集計がそのまま出力される．その他の場合はオプションで指定した場合を除き，EM アルゴリズムが適用される．統計解析機能としては，分割表の独立性検定の他，数量化 II 類や対応分析が実行可能となっている．

図 5 から図 8 には，PPE における一連の処理の様子が示されている．PPE を起動後に CSV 形式のデータファイルを読み込むと，図 5 のようなデータのプレビューが表示される(空白のセルは無回答を示す)．同時にアイテム情報のタブでは，各変量に対するカテゴリがリスト表



	X1	X2	X3	X4	X5	X6	X7	X8
	0001	112		3	2	112	5	1
	0002	112	2	3	5			
	0003	112	2	4	1	112	3	1
	0004	112	2	5	1	112	4	2
	0005	112	2	5	1	112	4	2
	0006	112	2	2	2	101	2	4
	0007	112	2	3	1	112	3	1
	0008	112	2	4	7			
	0009	112	2	3	7	112	3	1
	0010	112	1	2	1	112	4	1
	0011	101	1	2	1	101	3	2
	0012	112	1	2	1	101	3	3
	0013	112	2	2	1	101	3	3
	0014	112	2	2	1	112	3	1
	0015	112	1	5	2	101	3	3
	0016	112	1	6	1	112	3	1
	0017	112	1	4	2			
	0018	112	2	6	5			
	0019	112	2	4	1	112	3	1

図 9. PUE 実行画面〜データプレビュー.

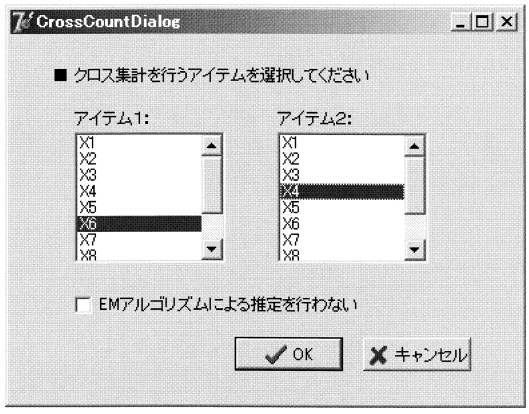
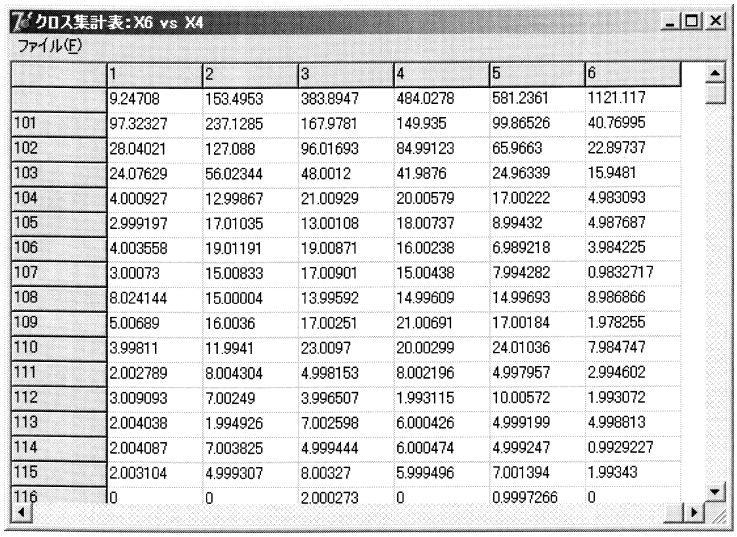


図 10. PUE 実行画面〜クロス集計変量選択.

示される．このデータに対して，Invariant PRAM を適用する場合には PRAM 設定ウィザードを起動し，適用する変量と独立に適用するかどうかを指定する(図 6)．ここで，独立に適用するよう指定すると，次の画面では各変量に対する θ の値を設定することになる(図 7)．最後に，PRAM により攪乱されたデータを保存するファイルと，PRAM 行列などの付随する情報を含むファイルの保存先を指定すると PRAM が実行される．実行後に，PRAM によって実際に遷移した個体数がその割合と共に各変量ごとに表示される(図 8)．

図 9 から図 11 は，PUE における処理の様子を示している．ユーザーは初めに PPE から出力されて公開された 2 つのファイルを PUE に読み込ませ(図 9)集計や統計解析を実行する．


クロス集計表:X6 vs X4

ファイル(F)

	1	2	3	4	5	6
	9.24708	153.4953	383.8947	484.0278	581.2361	1121.117
101	97.32327	237.1285	167.9781	149.935	99.86526	40.76995
102	28.04021	127.088	96.01693	84.99123	65.9663	22.89737
103	24.07629	56.02344	48.0012	41.9876	24.96339	15.9481
104	4.000927	12.99867	21.00929	20.00579	17.00222	4.983093
105	2.999197	17.01035	13.00108	18.00737	8.99432	4.987687
106	4.003558	19.01191	19.00871	16.00238	6.989218	3.984225
107	3.00073	15.00833	17.00901	15.00438	7.994282	0.9832717
108	8.024144	15.00004	13.99592	14.99609	14.99693	8.986866
109	5.00689	16.0036	17.00251	21.00691	17.00184	1.978255
110	3.99811	11.9941	23.0097	20.00299	24.01036	7.984747
111	2.002789	8.004304	4.998153	8.002196	4.997957	2.994602
112	3.009093	7.00249	3.996507	1.993115	10.00572	1.993072
113	2.004038	1.994926	7.002598	6.000426	4.999199	4.998813
114	2.004087	7.003825	4.999444	6.000474	4.999247	0.9929227
115	2.003104	4.999307	8.00327	5.999496	7.001394	1.99343
116	0	0	2.000273	0	0.9997266	0

図 11. PUE 実行画面～クロス集計実行結果.

図 10 はクロス集計を実行するためのダイアログであり，集計を行う変数を指定する．また，EM アルゴリズムを適用した集計表を出力するかどうかをオプションとして指定することが可能である．図 11 は EM アルゴリズムにより推定されたクロス集計の出力画面である．

6. まとめと今後の課題

PRAM はマイクロデータをより安全にして公開するための方法であると同時に，PRAM 行列を使うことによって PRAM により攪乱されたデータに対する統計解析を修正することができるという可能性を持っている．本論文では，PRAM の理論の紹介と，いくつかの数値実験，さらに PRAM を実行する環境についての提案を行った．

PRAM におけるデータの提供者側の問題は，PRAM 行列をどのように与えればよいかということであるが，今のところ第 4 章において述べた方法 1 が，データの安全性や遷移する個体の少なさという面から最も合理的な方法であると考えられる．ただし，θ の値をどのように設定すればよいかということに関しては検討の余地がある．現状では安全性を最優先に考慮して方法 1 においては $\theta = 0.9$ という値を設定したが，θ の値をどの程度まで小さくすることが可能であるかということは課題の一つである．また，今後さらに計算機の処理能力が向上することにより，情報量の損失と安全性に関する最適化問題を解くことによる PRAM 行列の生成方法も現実的になってくるだろう．PRAM に関する情報をどの程度公開するのかという問題も考えられる．第 5 章で提案したシステムは，PRAM 行列と適用した変量，適用方法(独立かそうでないか)に関する情報を分析者に渡すという設定で設計されている．これらの情報を公開すること自体がデータの安全性を低下させる要因となると考えると，PRAM の枠組み自体を考え直す必要があるかもしれないが，システムとしてはこれらの情報を暗号化するなどして直接分析者から見えない形にするなどして対応することが可能である．これらの提供者側の問題が解決すれば，データの分析者側の問題設定がより明確になり，分析の修正方法に対する研究も進むであろう．

このように検討課題はいくつか存在するが，PRAM を単独で用いるのではなく，一定の開示制御がなされたデータに対して，より安全性を高める手段として PRAM を導入するということは効果的であると考えられる．今後は，PRAM を利用した統計調査データの公開実験を何らかの形で行っていきたいと考えている．

参 考 文 献

De Wolf, P. P., Gouweleeuw, J. M., Kooiman, P. and Willenborg, L. C. R. J. (1997). Reflections on PRAM, Report, Department of Statistical Methods, Statistics Netherlands, Voorburg/Heerlen.

Fujino, T. and Tarumi, T. (2001). PRAM and its influence of multivariate analysis, *Bulletin of the International Statistical Institute* (*53rd Session Contributed Papers Book* 1), 153–154.

Fujino, T. and Tarumi, T. (2002). Evaluate the influence of PRAM to statistical analyses using simulation, *Proceedings of the 4th ARS Conference of the IASC*, 89–92.

Gouweleeuw, J. M., Kooiman, P., Willenborg, L. C. R. J. and De Wolf, P. P. (1998). Post randomization for statistical disclosure control: Theory and implementation, *Journal of Official Statistics*, **14**, 463–478.

Hout, A. D. L. van den (1999). *The Analysis of Data Perturbed by PRAM*, Delft University Press, Delft.

Kooiman, P., Willenborg, L. C. R. J. and Gouweleeuw, J. M. (1997). PRAM: A method for disclosure limitation of microdata, Report, Department of Statistical Methods, Statistics Netherlands, Voorburg/Heerlen.

Willenborg, L. C. R. J. (2000). Optimality models for PRAM, *Proceedings of Compstat 2000*, 505–510.

Willenborg, L. C. R. J. and de Waal, Ton (2000). *Elements of Statistical Disclosure Control*, Springer, New York.

Willenborg, L. C. R. J. and Hout, A. D. L. van den (2000). Possibilities of PRAM to protect statistical data, Report, Department of Statistical Methods, Statistics Netherlands, Voorburg/Heerlen.

Warner, S. (1965). Randomized response: A survey technique for eliminating answer bias, *J. Amer. Statist. Assoc.*, **60**, 63–69.

Proceedings of the Institute of Statistical Mathematics Vol. 51, No. 2, 321–335 (2003) 335

Theory of PRAM and Problems in Practical Use

Tomokazu Fujino

(Department of Environmental Science, Fukuoka Women's University)

Tomoyuki Tarumi

(Department of Environmental & Mathematical Sciences, Okayama University)

Post Randomization Method(PRAM) is a statistical disclosure control method for anonymized sample data. It was proposed by Kooiman et al. (1997). With this method, a data provider can decrease the risk of individual identification and information disclosure by perturbing data for each record based on a pre-determined probability structure. This paper discusses some possible problems in practical use and its influence on statistical analysis results. It also proposes a software environment for PRAM.

Key words: Microdata, Markov chain, local recoding, EM algorithm.