集計表セル秘匿問題の拡張による データ効用保持の有効性評価

南 和宏[1,2]　阿部 穂日[2]

概要：クロス集計表のセル秘匿問題は統計開示制御の分野で長年研究されてきた．しかし通常は表内の内部セルを秘匿対象とし，行計，列計に関するセルは公開することを前提としてきた．しかし，集計表の各セルに集計される客体数が非常に少ない場合，集計表の安全性を保証するような秘匿処理が不可能な場合が多く存在する．我々はこの問題を解決するため，通常の秘匿処理アルゴリズムを拡張し，集計セルも秘匿対象とした．実データによる予備実験は，この拡張した秘匿処理アルゴリズムは従来手法よりも情報損失を削減できることが示した．

キーワード：統計開示制御，セル秘匿，組み合わせ最適化

On Suppressing Marginal Cells in the Cell Suppression Problem

Kazuhiro Minami[1,2]　Abe Yutaka[2]

Abstract: The cell suppression problem (CSP) for tabular data has been studied for many years. In CSP, we prefer to suppress regular internal cells to marginal cells because information on marginal sums is more likely to be available elsewhere. We, however, find that there is a class of tables with no cell suppression pattern for regular cells that protects sensitive cell values properly. Such a situation occurs when a marginal sum of multiple cell values is smaller than a given threshold value or when a difference between a cell value and its marginal sum is smaller than that threshold; the marginal sum restricts the upper bound of each cell contributing to that sum narrowing the ranges of their possible values. To study positive aspects of suppressing marginal cells in CSP, we compare the approach of suppressing both marginal cells and regular cells with that of suppressing only regular cells after preprocessing unsafe rows and columns. Our initial experimental results show that suppressing marginal cells is an effective way of reducing information loss in terms of the number of suppressed cells.

Keywords: statistical disclosure control, cell suppression, combinatorial optimization

1. はじめに

クロス集計表のセル秘匿問題は公的統計の統計開示制御の研究領域で長年研究されており，Castro [4] により線形計画法の最適化問題として定式化されている．このセル秘匿問題は NP 困難であり，近似解のための多項式時間アルゴリズム [5], [6], [11] が考案されている．オランダ統計局は，これらのアルゴリズムを実装した秘匿処理ツール τ-Argus [2] を開発し，一般に公開している．これらのアルゴリズムは，表データの各セルに機密性ルール（p%ルール等） [3], [7] を適用し，機密性の高いセルを 1 次秘匿する．さらに表データが内包する線形の行計，列計の関係式から 1 次秘匿したセルの値が復元されないように，追加の 2 次秘匿処理を行う．セル秘匿による情報損失を最小化するため，セル秘匿処理アルゴリズムの主な目的は 2 次秘匿するセル数を最小化する最適な秘匿パターンを決定することである．

[1] 統計数理研究所
Institute of Statistical Mathematics
[2] 独立行政法人 統計センター
National Statistics Center

クロス集計表のセルを秘匿する場合，行計や列計の集計値を示すセルよりも通常の内部セルの秘匿が望ましい．なぜなら，集計値は他の関連するクロス集計表から入手可能な場合が多いからである．悪意のある攻撃者であれば，そのような類似する複数の表データの差分から集計値を推測する可能性がある [10], [12].

しかし我々が開発した秘匿処理ツールを総務省統計局が実施した平成２６年全国消費実態調査 [1] の個票データから作成した集計表に適用したところ，内部セルのみを秘匿処理では安全性の要件を満足するようなセル秘匿パターンが存在しない（解が存在しない）集計表が存在することが判明した．そうした状況は，行計，列計の集計値が，集計に含まれる内部セルの取りうる値の上限を制限することに原因がある．詳しくは，以下の２つのケースに分類できる．一つは，集計値の値が各セル値が満足すべき最小値より小さい場合である．もう一つは，集計値とその集計に含まれる一番上位のセル値の差がセルの取るべき最小値よりも小さい場合である．この２つのいずれかに該当する場合，集計表のすべての内部セルを秘匿しても安全性の要件は満足できない．

我々はこの問題を解決するため，２つの解決策を検討した．一つは，秘匿処理を行う前に前処理として，集計表から集計値による制限で安全性の要件を満足できない行と列を削除する方法である．この前処理の後，残りの表データに対して，内部セルの秘匿処理を行う．この手法は，前処理後の集計表の行計，列計の集計値をすべて保持して公開することを可能にするが，前処理での行，列を削除する処理は情報損失が大きくしがちである．

もう一つは，集計セルも秘匿対象に拡充する方法である．従来のセル秘匿問題の定式化 [4] では，集計セルが秘匿対象に含まれるかどうかは明確に示されていなかった．本論文では，集計セルを秘匿対象に含めるセル秘匿問題を明示的に「拡張セル秘匿問題」として定式化する．この拡張セル秘匿問題の枠組みでの秘匿処理では，前者の手法で削除対象となった安全でない行または列を削除する代わりに集計セルの秘匿を行い，その結果、削除対象の行、列に含まれる機密性が高くないセルの値を保持することを可能とする．

本論文では，我々の検討する２つの解決策を実際に全国消費実態調査 [1] から作成した度数分布表で実証的に評価した．評価実験は，２つの手法を集計表に適用した場合の秘匿セル数を比較し，集計セルを秘匿対象する方法のほうが，前処理で行，列を削除する方法よりも２次秘匿するセル数が削減できることを示した．前者の手法では，少数の集計セルを秘匿することで安全な秘匿パターンの解が存在しないという問題を回避し，情報損失を効果的に削減していることが分かった．

2. 背景

本章では，セル秘匿問題を紹介し，安全でない行，列が原因で生じる安全なセル秘匿パターンの存在性の問題について解説する．

2.1 セル秘匿問題

まず，セル秘匿問題を例を用いて簡潔に紹介する．ただし、説明は，安全なセル秘匿パターンの存在性に関する部分に限定する．この問題の定式化の詳細は [4] を参照することとする．

図１は度数分布表に対するセル秘匿アルゴリズムの概要を示す．このアルゴリズムは元テーブルと安全性要件のパラメータを入力とし，秘匿したテーブルとその秘匿箇所を示す秘匿パターンを出力する．秘匿パターンは 0/1 の２値テーブルであり，セル値が１の場合，対応する元テーブルのセルは秘匿され，セル値が０の場合は，元テーブルのセル値が保持される．セル秘匿アルゴリズムは，秘匿パターンに含まれる ‘1’ の値の数が目的関数であり，その最小化を目指す．

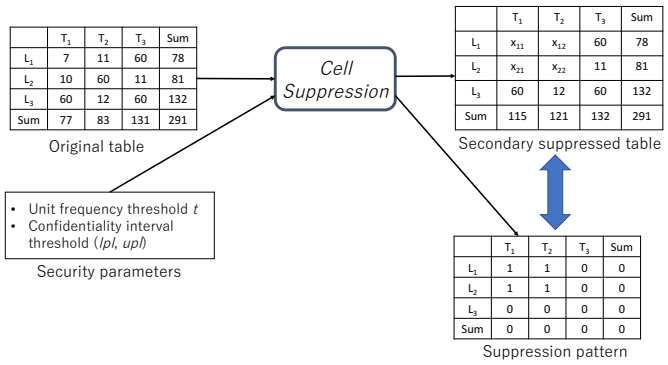

図 1 セル秘匿アルゴリズムの概要．秘匿したセルの値は変数 x_{11}, x_{12}, x_{21},x_{22} で参照する．

入力パラメータは２つある．最小度数しきい値 t は表セルの１次秘匿の判断に用いられる．もしセル i について，値 v_i が $v_i < t$ の場合，セル i は秘匿される．２つめのパラメータは秘匿インターバルのしきい値であり，２つの値 lpl, upl の対から成る．秘匿されたセルに対し、集計表の行計，列計の拘束条件から，秘匿インターバルと呼ぶセル変数が取りうる可能な値の範囲が算出できる．もし秘匿セル p に対応する変数を a_p とすると，その秘匿インターバルは，$(\underline{a_p}, \overline{a_p})$ で表される．ここで，$\underline{a_p}$, $\overline{a_p}$ はそれぞれ，変数 a が取りうる最小値，最大値である．下記の２つの条件 (1),(2) はセル変数 a_p が十分な幅の秘匿インターバルをもつことを保証する．

$$\underline{a_p} < a_p - lpl_p \quad (1)$$

$$a_p + upl_p < \overline{a_p} \quad (2)$$

例えば，図 1 において，セル (L_1, T_1) が 1 次秘匿セル，(L_1, T_2), (L_2, T_1), (L_2, T_2) の 3 つが 2 次秘匿セルとする．このとき，行計，列計に関する下記の 4 つの等式をセル変数が満足する必要がある．

$$x_{11} + x_{12} = 18 \tag{3}$$

$$x_{21} + x_{22} = 70 \tag{4}$$

$$x_{11} + x_{21} = 17 \tag{5}$$

$$x_{12} + x_{22} = 71 \tag{6}$$

もし，$lpl = upl = 5$ であれば，

$$\underline{x_{11}} = 0 < 2 = x_{11} - lpl \tag{7}$$

$$x_{11} + upl = 12 < 17 = \overline{x_{11}} \tag{8}$$

となり，秘匿インターバルの条件は満たされる．

2.2 可能な秘匿パターンの存在性

我々が開発した秘匿処理ツール [9] を実データで評価した際，安全な秘匿パターンが全く存在しない種類の集計表が存在することが判明した．このような状況が生じる 2 つの原因を表 1 の度数分布表を例に説明する．

	a_1	a_2	a_3	a_4	Sum
l_1	0	1	7	8	16
l_2	1	4	6	8	19
l_3	0	1	1	10	12
Sum	1	6	14	26	47

表 1 可能な秘匿パターンが存在しない度数分布表. 最小度数しきい値 $t = 5$ の場合, 列 a_1 と a_2 は安全ではない.

1 つ目は，集計セルの値が秘匿インターバルの最大値の下界 upl よりも小さい場合である．例えば，最小度数しきい値 $t = 5$ の場合，第 1 列の各セル値 a_p は秘匿される．ここで，秘匿インターバルの最大値の下界を $upl = 5$ とすると，秘匿したセルの秘匿インターバルの最大値は 5 より大きくなくてはいけない．しかし，列計 1 が公開される場合，秘匿インターバルは $[0, 1]$（つまり，$\underline{a_p} = 0, \overline{a_p} = 1$）となり，秘匿インターバルに関する式 (2) の条件を侵害する．

2 つ目は，1 次秘匿されたセル値 a_p とその行計または列計 a_{sum} の差 $a_{sum} - a_p$ が upl より小さい場合である．表 1 の第 2 列がこの場合に該当する．第 2 列の各セル値も最小度数しきい値 $t = 5$ より小さいため，1 次秘匿される必要がある．この場合，列計 6 は $upl = 5$ より大きな値で，1 つめのケースには該当しない．しかし，セル (l_2, a_2) の値と列計の値の差は $6 - 4 = 2$ であり，セル (l_2, a_2) の取りうる最大値は実際の値よりせいぜい 2 大きいに過ぎない．その結果，式 (2) の条件を満たすことができない．2 つ目のケースは，同一の行または列に属する多くのセルがゼロ値またはそれに近い小さな値を取るときに生じる．我々は全国消費実態調査の個票データから「年齢」の属性に関する詳細な区分をもつ集計表を作成した際，このような状況が頻繁に生じることが分かった．

3. 提案手法

可能な秘匿パターンが存在しない問題の解決策として，2 つの手法を検討した．一つは，2.2 章で説明した秘匿インターバルの条件を満足しない行および列を前処理段階で削除する方法であり，もう一つは，セル秘匿問題を拡張し，集計セルも秘匿対象とする拡張セル秘匿問題を解く手法である．

3.1 前処理

最初に各行および列に対して，下記の 2 つの条件

(1) 行計または列計の値が最小度数しきい値 t より小さい.

(2) 行計または列計とその行または列に含まれるセル値の差が秘匿インターバルの最大値の下界 upl より小さい.

を検証し，どちらかに該当する場合は，安全でない行または列の集合 D_1 または D_2 に追加する．その後，集合 D_1, D_2 に含まれる行と列を元の集計表から削除する．最後に [4] にある内部セルに対する既存の秘匿処理手法を実施する．この手法の欠点は安全でない行および列を全て削除するために情報損失が大きくなる点である．通常は機密性をもたないゼロ値のセルであっても，安全でない行または列に含まれる場合は，一緒に秘匿されてしまうことになる．

3.2 拡張セル秘匿問題

もう一つの解決策は，セル秘匿の対象を集計セルに拡充する方法である．集計セルを秘匿することによりその集計に含まれる個々のセル値の秘匿インターバルの幅に十分な長さを確保することが可能になる．

我々は Castro のセル秘匿問題の定式化 [4] を拡張し，明示的に拡張セル秘匿問題として定義することとする．唯一の違いは表データが内包する線形の拘束条件の形式であり，それを以下に説明する．行計，列計に関する拘束条件は下記の線形式で表される．

$$Ax = b \tag{9}$$

ここで A は係数行列，x は内部セル変数のベクトルである．ただし，[4] では行列 A をどのように構成すべきか具体的な記述はなく，ベクトル b をどのように解釈すべきか明らかではない．この論文では，b は行計，列計の集計値のベクトルを示すと仮定する．

行列 A の各行は集計値の拘束条件を記述するための係数を格納する．もしセル j が i 番目の関係式に含まれるとすると，$A[i, j] = 1$ であり，そうでなければ $A[i, j] = 0$ で

ある．ここで，ベクトル x に対応する別のベクトル $\bar{x}$ を導入し，その各要素 $\bar{x_j}$ がセル j の真の値 x_j とその秘匿インターバルの中の取りうる値 x' との差分を表すとする．その場合，この差分ベクトル $\bar{x}$ に対して，

$$A\bar{x} = 0. \tag{10}$$

が成り立つ．なぜなら，一つの関係式に含まれるベクトル $\bar{x}$ の値の増減は互いに打ち消し合ってゼロでなければ式 (9) で規定された集計値と同じにならないからである．我々が前回開発した秘匿処理ツール [9] は式 (10) を拘束条件とすることで，集計表の内部セルのみを秘匿対象とする秘匿処理を実現した．

秘匿対象に集計セルを加える場合，行列 A を下記のように修正する．まず，集計セルに対応するベクトル $\hat{x}$ を内部セル変数ベクトル $\bar{x}$ に追加する．集計セルが m 個あるとすると，m 列の係数行列を元の係数行列 A の右側に結合し，行列 B とする．しかし，集計値は拘束条件の各関係式の右辺に現れるため，それらに対する係数は通常の内部セルに対するものとは異なる方法で指定する．集計セル $\hat{x}_k$ が関係式に含まれる場合，内部セル変数の合計は同じ値を取る必要がない．その増減に応じて集計セルの値を変えればよいからである．したがって，集計セル $\hat{x}_k$ が i 番目の関係式に現れる場合，$B[i,k] = -1$ であり，そうでなければ，$B[i,k] = 0$ となる．下記の拘束条件

$$B\tilde{x} = 0. \tag{11}$$

に対して，[9] の秘匿処理アルゴリズムを適用し，集計セルを対象として拡充したセル秘匿処理が実現できる．

4. 評価

我々は，3 章で説明した 2 つの提案手法を実証的に比較した．2014 年全国消費実態調査の個票データから作成した度数分布表を評価対象とし，情報損失の指標として，秘匿セルの個数を用いた．調査データの標本数は 51,768 世帯であり，地域情報は日本を 10 地域に分割したものを用いている．我々は地域区分と各歳の年齢の 2 つの属性をクロスした度数分布法を複数の異なる年齢範囲について数種類作成した．

表 2 に両者の結果を比較する．情報損失の度合いを評価するため，それぞれの手法で秘匿された秘匿セルの数を評価指標とした．最初の列の「前処理」は 3.1 章にある行，列の削除の前処理と内部セルの秘匿処理を組み合わせた手法である．2 番めの列の「拡張セル秘匿」は 3.2 章にある集計セルを秘匿対象に含めた手法である．最初の 3 つの行は秘匿対象とする表データのサイズを示している．4 行目は前処理で秘匿されたセルの数を示す．5 行目は 1 次秘匿されたセルの数を示す．拡張セル秘匿では集計セルも秘匿対象としているため，若干秘匿されたセルが前処理の場合

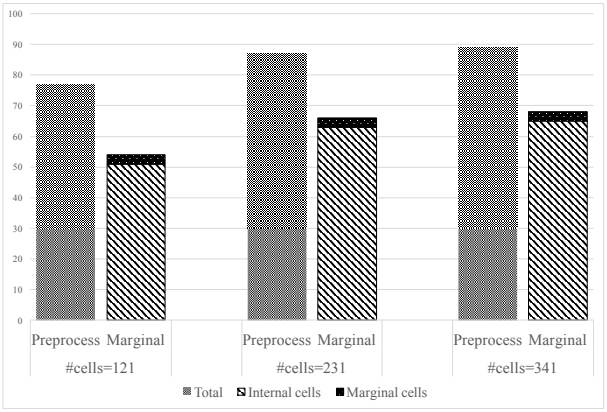

図 2 秘匿セルの合計の比較．左の棒は，前処理の結果，右の棒は拡張セル秘匿の結果を示す．棒グラフの黒色は集計セルの内訳を示す．

よりも多くなっている．6 行目，7 行目は 1 次秘匿されたセル数の内訳を内部セル，集計セルの別に示している．8 行目は 2 次秘匿されたセルの数を示し，9 行，10 行はその内訳を示す．11 行は秘匿セルの合計，12 行目はその全体に対する割合を示す．

図 2 の棒グラフは 2 つの手法による秘匿セルの合計を比較したものである．3 つの表データに対する全ての場合で，拡張セル秘匿の手法の情報損失が少なくなっている．表 2 の 12 行の合計の秘匿セル数を参照すると，前処理の手法に対して，秘匿セル数を約 6%から 20%削減できていることがわかる．

少数の集計セルを秘匿することで，秘匿する内部セルの数を大幅に減らすことができることが分かった．表のサイズが大きくなるに従い両者の差が少なくなるのは，追加する年齢範囲が人口分布の中心に近い部分であったため高い度数をもつセルが大半であり，秘匿処理の必要なセルが少なかったためである．

表 2 2 つの提案手法の情報損失の比較．

	前処理			拡張セル秘匿		
セル数	121	231	341	121	231	341
行数	11	11	11	11	11	11
列数	11	21	31	11	21	31
前処理	30	30	30	0	0	0
1 次秘匿セル数	45	53	54	54	62	63
内部セル	45	53	54	51	59	60
集計セル	0	0	0	3	3	3
2 次秘匿セル数	2	4	5	0	4	5
内部セル	4	4	5	0	4	5
集計セル	0	0	0	0	0	0
合計	77	87	89	54	66	68
秘匿セルの割合	0.63	0.37	0.26	0.44	0.29	0.19

5. 関連研究

τ-Argus [2] は 3.2 の拡張セル秘匿を実装しているが，拡

張セル秘匿問題の定式化が明示的に記述されていない．また集計セルを秘匿対象とする優位性はこれまで実証的に評価されていなかった．

τ-Argus [2] のドキュメントは同一の行または列に２つの度数１のセルが現れるシングルトンの問題を議論している．調査に参加した内部攻撃者を考慮する場合，もしどちらかの度数１のセルの客体が攻撃者であれば，もう一方の度数１の客体の機密情報が推測可能である．本論文では外部の攻撃者のみを考慮したが，集計セルの秘匿処理はシングルトンの課題に対する有望な解決策になりうると考える．

6. おわりに

本論文では，集計表に対する内部セルの秘匿問題において，可能な秘匿パターンが全く存在しない場合が実際に起きうることを示し，その解決策として，集計セルを秘匿対象とする拡張セル秘匿問題を定式化し，秘匿処理ツールに実装した．予備的な評価実験では提案手法が秘匿パターンの存在性の問題を回避し，さらに情報損失の削減にも効果的であることを示した．今後は，集計値を公開する場合の懸念となる差分攻撃のリスクを評価する予定である．

参考文献

[1] National survey of family income and expenditure, http://www.stat.go.jp/english/data/zensho/index.html

[2] τ-Argus homepage, http://neon.vb.cbs.nl/casc/tau.htm

[3] Bring, J., Wang, Q.: Comparison of different sensitivity rules for tabular data and presenting a new rule – the interval rule. In: Proceedings of the 2014 Privacy in Statistical Databases (2014)

[4] Castro, J.: Recent advances in optimization techniques for statistical tabular data protection. European Journal of Operational Research **216**(2), 257 – 269 (2012)

[5] Cox, L.H.: Network models for complementary cell suppression. Journal of the American Statistical Association **90**(432), 1453–1462 (1995)

[6] Giessing, S., Repsilber, D.: Tools and Strategies to Protect Multiple Tables with the GHQUAR Cell Suppression Engine, pp. 181–192. Springer Berlin Heidelberg, Berlin, Heidelberg (2002)

[7] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K., de Wolf, P.P.: Statistical Control Disclosure. Wiley (2012)

[8] Kikuchi, R., Minami, K.: On-site service and safe output checking in japan. In: Proceedings of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality (2017)

[9] Minami, K., Abe, Y.: Statistical disclosure control for tabular data in r. Romanian Statistical Review (4), 67–76 (2017)

[10] Sukasih, A., Jang, D.H., Edson, D.: Using tau-argus and sdctable to conduct secondary cell suppression for linked tables. In: Proceedings of the 2011 Joint Statistical Meetings (2011)

[11] de Wolf, P.P.: HiTaS: A Heuristic Approach to Cell Suppression in Hierarchical Tables, pp. 74–82. Springer Berlin Heidelberg, Berlin, Heidelberg (2002)

[12] de Wolf, P.P., Hundepool, A.: Three ways to deal with a set of linked sbs tables using τ-argus. In: Domingo-Ferrer, J., Magkos, E. (eds.) Privacy in Statistical Databases. pp. 66–73. Springer Berlin Heidelberg, Berlin, Heidelberg (2010)