統計数理(2003)
第 51 巻 第 2 号 337–350

特集「個票開示問題の統計理論」
[総合報告]

集計表におけるセル秘匿問題とその研究動向*

瀧 敦弘†

(受付 2003 年 2 月 7 日；改訂 2003 年 6 月 16 日)

要　　旨

集計表による調査データの公表においても，個票開示の問題と同様に，プライバシー保護の点から秘匿を必要とする箇所が発生する．このような秘匿すべきセルをいかに取り扱うか，すなわち，セル秘匿問題がこの小論で取り扱う問題である．セル秘匿には 2 つの問題が存在する．第 1 は，セルを秘匿する基準の問題である．第 2 は，いかに秘匿するかという方法についての問題である．この小論では，第 2 節で基準の問題を解説し，次に，第 3 節でセル秘匿方法について解説した．第 4 節では，おもに欧米の研究動向を詳しく解説した．欧米では，オランダ統計局を中心とするヨーロッパの研究グループと，米国およびカナダのセンサス局を中心とする研究グループの 2 つの大きなグループにより，近年，非常にさかんに研究されている．さらに，セル秘匿処理を自動化して行うためのソフトウェアの開発も進んでいる．

キーワード： 集計表，セル秘匿問題，SDC プロジェクト，CASC プロジェクト，セル秘匿処理ソフトウェア．

1. はじめに

電子媒体による官庁統計の公表が容易となった現時点においても，集められたデータは一般には印刷された集計表として公表されている．このような集計表による調査データの公表においても，個票開示の問題と同様に，プライバシー保護の点から秘匿を必要とする箇所が発生する．すなわち，集計表において公表してはならない行と列が交わった場所(「セル」と呼ばれる)が発生するのである．このような秘匿すべきセルをいかに取り扱うかがここでの問題である．

具体的には，「商業統計表」「工業統計表」の市区町村表，または「サービス業基本調査」をみると，かなりの数のセルに伏字「x」が付いていることに気づく．これらの「x」について，たとえば「商業統計表」の利用上の注意には，つぎのように説明している．

> 「x」は，その数字に該当する商店数が 1 又は 2 であるため，ここの申告者の秘密保護の観点から数字を秘匿したことを示したものである．なお，この秘匿によっても数値 x が算出される恐れがあるものについては，商店数が 3 以上でも「x」で秘匿した個所があります．

この利用上の注意に指摘されているように，集計表における秘匿には 2 つの問題が存在する．第 1 の問題は，「その数字に該当する商店数が 1 又は 2 であるため」とあるように，そのセルを秘匿すべきかどうかの基準の問題である．秘匿する箇所を増やすと集計表から得られる情報

† 広島大学 経済学部： 〒739–8525 広島県東広島市鏡山 1–2–1

* 本稿は，科学研究費補助金の研究成果報告書に作成した 2 つの拙稿(瀧(1997, 2002))をもとに，大幅に加筆・改稿し，さらに最新の研究動向を付け加えたものである．

の量は減少するので，どれだけ秘匿しなければならないかについての基準を考察しなければならない．この問題については，次節で述べる．

第 2 の問題は，いかに秘匿するかという方法についてである．「商業統計表」などでは，単純に非表示(伏字「x」を付ける)で秘匿している．非表示以外にはどのような方法が考えられるのか．また，非表示する場合には，どのような点を考慮して伏字を付けなければならないのかについて，第 3 節で取り扱う．

付け加えて言えば，秘匿の基準や方法を公表しないということが，それ自体で有力な秘匿の方法でもある．データを利用する立場からすると，秘匿の基準や方法を公表しないことは，利用するデータの信頼性に問題が生じることとなるが，既存の研究において，秘匿の基準や方法の非公開による影響や問題点の議論はみあたらない．

そして，第 4 節で，欧米の研究動向を解説することにする．その際に，実用のためのソフトウェアの開発も重要な話題であり紙幅をさくこととする．そして，わが国の現状についても述べる．最後に，第 5 節で，第 4 節で述べる日本の現状も踏まえて，今後の課題を述べ，この小論をまとめる．

ただし，この小論では，集計表は 2 次元(平面の表のイメージ)のみを考えている．Willenborg and De Waal (1996, 2000), Domingo-Ferrer (2002), Doyle et al. (2001)によると，3 次元や 4 次元などの多次元表についてのセル秘匿問題の研究が，近年進んでいることがわかるが，それについての解説は別の機会としたい．

2. 秘匿すべきセルの基準

あるセルの値を，プライバシー保護の点から公開すべきでないことを，「そのセルはセンシティブ(sensitive)である」と表す(直訳からすれば，「感応的」ということであろうが，「ヤバい」くらいのニュアンスであろう．なお，デニング(1988)の訳者は，「機密扱い」という訳語を与えている)．センシティブかどうか，すなわち，公表すべきではないかどうかの判断基準として n-k 占有ルール(n-k dominance rule)と (p, q)-事前事後ルール(prior-posterior rule)が一般に用いられている．秘匿すべきセルの基準については Loeve (2001)に詳しいが，Cox (1981)による (p, q)-事前事後ルールの提案以後，新しい基準についての議論はなく，これら 2 つのルールについて検討することで十分であろう．

まず，n-k 占有ルールについて述べよう．これは，特定のセルに落ちる個体のうちの上位 n 個体がセルのすべての個体の和の k%を占める場合，すなわち，セルの合計の k%以上を最上位の n 個の個体の和で占められてしまった場合に，そのセルはセンシティブとするのである．前節に示した「商業統計表」の秘匿の基準を，n-k 占有ルールで解釈すると，$n = 2$，$k = 100$ となる．このルールの考え方について，デニング(1988)では，つぎのような例をあげている．

> 例：IBM 社の事業所得とインディアナ州のアーリーミュージック店の事業所得額を合計した統計があったとすれば，これは 1 個体，99%占有ルールに照らして機密扱いの統計であろう．アーリーミュージック店の事業所得額については，これだけでは何の情報を与えるものではないが，IBM 社の事業所得額の情報は十分に得られる．

普通，n は小さい値(一般に 5 まで)，k は大きい値(たとえば，80)を用いる．たとえば，表 1 のような集計表において，$n = 3$，$k = 75$ の場合を考える．

A1-RB のセルについて，値の大きいほうから 3 つの企業について，企業 1 が 16，企業 2 が 10，企業 3 が 6 であるとする．その他の企業の計が 18 であるが，$(16 + 10 + 6) \div 50 = 0.64$ であり，3-75 占有ルールを満たさない．すなわち，企業 2 と 3 が結託したとしても，企業 1 につ

表 1.

	R A	R B	R C	計
A 1	20	50	10	80
A 2	8	19	22	49
A 3	17	32	12	61
計	45	101	44	190

いては，最小では企業 2 の値と同じ 10，最大では 50 から企業 2 と 3 の合計の値を引いた 34 としかわからないからである．

ところが，A2-RC のセルについて，値の大きいほうから 3 つの企業について，企業 1 が 8，企業 2 が 6，企業 3 が 5 であれば，$(8+6+5) \div 22 = 0.86$ となり，このセルは 3-75 占有ルールではセンシティブなセルであることがわかる．実際に，企業 2 と 3 が結託すると企業 1 の値は，企業 2 の値である 6 よりは大きく，最大では 22 から 11 を引いた 11 となり，かなり狭い区間で推定が可能である．

もう一つの判断基準として，(p,q)-事前事後ルールがある．このルールについて，n-k 占有ルールのようには邦語の文献がないので，少し詳しく解説する．

予めある値に固定しておいたパラメータ p と q ($p < q$ とする．p が n-k 占有ルールにおける k のような閾値を示す)について， すべての回答者は他の回答者がどのような値を回答したかを，それぞれの値の q パーセント以内に推定できると仮定する．ここで，結果的に，真の値の p パーセント以内に推定できてしまうならば，センシティブなセルであるとする．

まず，最も大きな値を回答したものが，2 番目に大きな値を回答したものの値を正確に推定できるかどうかを考えよう．ただし，ここでは 2 番目に大きな値を回答したものが，自分が 2 番目に大きい値を回答したという事実を知っているかどうかは重要ではない．この 2 番目に大きい回答者は，つぎのように第 1 番目に大きな値 x_1 に対する上限を計算できる．合計値 $\sum_{i=1}^{N(X)} x_i$ はわかっているので，自分自身の回答した値 x_2 を差し引くことによって，他のものが回答した値は， $i = 3, \ldots, N(X)$ に対しては(ここで，セル X には $N(X)$ の数の回答者がいることを表している)，少なくとも $x_i - (q/100)x_i$ であるので，この結果を利用すれば，第 1 番目の値の上限は，つぎのように推定できる．

$$E(x_1) = x_1 + (q/100) \sum_{i=3}^{N(X)} x_i \tag{2.1}$$

もし $E(x_1) \leq x_1 + (p/100)x_1$ であれば，センシティブなセルであることとなる．同様に考えれば，第 m 番目の回答者は第 n 番目の回答者の値の上限をつぎのように推定できる．

$$E(x_n) = x_n + (q/100) \sum_{\substack{i=1 \\ i \neq n,m}}^{N(X)} x_i \tag{2.2}$$

従って， $E(x_n) \leq x_n + (p/100)x_n$ であれば，センシティブなセルであることとなる．

回答者が降順に番号がふられていれば，

$$(q/100) \sum_{\substack{i=1 \\ i \neq n,m}}^{N(X)} x_i \geq (q/100) \sum_{i=3}^{N(X)} x_i \tag{2.3}$$

となる．したがって,(2.1)式が $x_1+(p/100)x_1$ より大きければ，すなわち，2 番目に大きな値を回答したものが原数値の p パーセント以内で第 1 番目の回答者の値を推定できないならば,(2.2)式は， $x_n+(p/100)x_n$ より大きくなるであろう．そして，任意の回答者の値も p パーセント以内では推定できない．このことから，センシティブなセルかどうかを判定するために,(2.1)式が $x_1+(p/100)x_1$ より大きいかどうかを判定すればよいこととなる．

以上により，事前事後ルールは，つぎの基準による．

$$S_q(X)=-(q/x_1)\sum_{i=3}^{N(X)}x_i \tag{2.4}$$

事前事後ルールにおいて，$S_q(X)>-p$ であればセンシティブなセルであることになる．この基準は，事前の値(x_i の q パーセント推定値)と事後の値の両方を必要とするので，事前事後ルールと呼ばれている．

p と q をもっと簡単にいうと，q は社会一般の知識や情報によって予測できる範囲であり，p は当該企業がこの範囲でわかってしまうと困ると考えている範囲であると考えてよい(したがって，$p<q$ でなければならない)．

例をあげよう．$p=10$, $q=50$ とする．あるセルの回答者が 4 個体で, それぞれの値が 100, 90, 10, 6 であるとする．たとえば, 最も大きな値 100 を回答した個体は, 2 番目に大きな値を回答した個体の値を, $206-100-0.5\times(10+6)=98$ と推定され, これは真の値 90 の幅 10% 以内となってしまう．実際には，式(2.4)で示される指標によっても，$S_q(X)=-(50/100)\times(10+6)=-8$ となり，セルは秘匿すべきとなる．

Robertson and Ethier (2002)では，2 つのルールを幾何学的に視覚で捉えて，そのルールの相違を明らかにしている．また，Cox (2001)では,(p,q)-事前事後ルールの特別な場合である p% ルールについて，n-k 占有ルールと比較している．それによれば，一般的に n-k 占有ルールのほうが p% ルールより厳しいことを明らかにしている．しかし，この 2 つのルールのどちらが優れているかという議論は適当ではない．セルに存在する個体の分布などのデータの特性からどちらが適当かという議論でなければならない．

Cox (2001)によれば，n-k 占有ルールは，1940 年代まで遡ることができ，古くから知られている秘匿の基準であり，多くの国の多くの統計で利用されていると思われる．一方，1982 年以後の米国経済センサスは，p% ルールを用いている．さらに，どのような統計に用いているかなどの詳細は不明であるが，カナダ統計局は(p,q)-事前事後ルールを用いている．なお，後述する集計表の秘匿のためのソフトウェア τ-ARGUS では，両方のルールを利用可能である．

3. 秘匿の方法

秘匿の方法として，つぎの 4 つの方法が考えられる．

1) 表の再設計(table redesign)
2) セルの非表示(cell suppression)
3) 区間開示(feasibility intervals)
4) 丸め法(rounding)

これらの方法でどの方法が優れているかという議論は，簡便さなどの基準も考えられるが，本来は，情報量損失の視点から考察されるべきである．しかしながら，この小論では，これらの方法がどのようなものであるかを，おもに例をもちいて説明するにとどめ，情報量損失の議論については複雑になるので取り扱わない．秘匿方法については，Cox (1980)や Willenborg and De Waal (1996)を参照されたい．また，情報量損失は，たとえば，Willenborg and De Waal

(2000)第 7 章や Massell (2001)を参照されたい.

1) 表の再設計

表の再設計は，単純な方法である．松本(1996)は「合併法」と呼んでいる．表の行または列の分割を見直して，秘匿する必要がないよう，ある行と別のある行(または，ある列と別のある列)を統合して新たな表につくりかえる方法である．たとえば，表 2 が表 1 に対する再設計の例である．この秘匿の方法は簡単であるが，せっかくの詳細な情報がたくさん脱落する．また，もし地域別など意味があって行や列に項目を立てて集計されたものであれば，ある地域と別のある地域を秘匿の問題だけで統合して合計値で表示すること自体が問題である．

2) セルの非表示

秘匿の基準から判断して表示すべきではないセルを隠してしまう方法である．先に述べたように「工業統計表」「商業統計表」にも用いられている．具体的に，表 3 で示す.

この表 3 において，セル A2-RC が秘匿の基準よりセンシティブであり，非表示とすべきとなったので，伏字 x を付けている．これを 1 次秘匿とよぶ．この 1 次秘匿だけでは，行の和，列の和が示されているので，簡単な計算でそのセンシティブなセルの値がわかってしまう．したがって，1 次秘匿したセルの値を秘匿するためには，さらにいくつかのセルを非表示にして秘匿しなければならない．これが 2 次秘匿(secondary suppression または complementary suppression)である．これに関しては，先に示した「商業統計表」の利用上の注意では，「…この秘匿によっても数値 x が算出される恐れがあるものについては，商店数が 3 以上でも『x』で秘匿した個所があります」が，この具体的な記述となっている．非表示法について最も重要な点は，いかに情報量損失を少なくして，2 次秘匿を付けるかである．ただし，数理的な情報の損失のほかにも，考慮しなければならないケースが存在する [1)].

表 4 は，表 3 の 1 次秘匿に 2 次秘匿を施したものである．1 次秘匿しなければならないセルの数が多くなると，2 次秘匿を施す作業は非常に複雑なものとなる．たとえば，セルの値が非

表 2.

	R A	R B	R C	計
A 1	20	50	10	80
A 2 + A 3	25	51	34	110
計	45	101	44	190

表 3.

	R A	R B	R C	計
A 1	20	50	10	80
A 2	8	19	x	49
A 3	17	32	12	61
計	45	101	44	190

表 4.

	R A	R B	R C	計
A 1	20	50	10	80
A 2	x	19	x	49
A 3	x	32	x	61
計	45	101	44	190

表 5.

	C 1	C 2	C 3	計
R 1	x_1	1	x_2	80
R 2	x_3	2	x_4	49
R 3	70	3	2	75
計	192	6	6	204

表 6.

	C 1	C 2	C 3	C 4	計
R 1	x_1	x_2	x_3	2	13
R 2	x_4	2	x_5	7	13
R 3	3	x_6	8	x_7	14
R 4	4	x_8	3	x_9	18
計	11	10	19	18	58

負という制約があると，秘匿されるべきセルの値を非常に短い区間で推定できてしまう場合がある．表 5 のケースは，その具体例を示したものである．

x_1 と x_3 が 1 次秘匿であり，x_2 と x_4 が 2 次秘匿である．

$$x_1 + x_2 = 80 - 1 = 79$$
$$x_3 + x_4 = 49 - 2 = 47$$
$$x_1 + x_3 = 192 - 70 = 122$$
$$x_2 + x_4 = 6 - 2 = 4$$

それぞれが非負であれば，$0 \leq x_2 \leq 4$，$0 \leq x_4 \leq 4$ なので，

$$75 \leq x_1 \leq 79,\ 43 \leq x_3 \leq 47$$

と，セルの値 x_1 と x_3 はかなり狭い区間でわかってしまう．

しかし，セルの値が非負という制約は統計データの集計表の場合，非常に一般的であるので，充分な範囲が確保されるように 2 次秘匿を見いだすことが難しい場合が存在する．

また，表 6 は，各行と各列にたくさんの非表示があって，一見，秘匿として十分であると思われるが，

1 列と 3 列の合計：$x_1 + x_3 + x_4 + x_5 = 12$
1 行と 2 行の合計：$x_1 + x_2 + x_3 + x_4 + x_5 = 15$

となり，これら 2 式より，$x_2 = 3$ とわかってしまう．このように 2 次秘匿をどのように行うかは非常に難しい．

さらに，セルそれぞれの情報に重みを持たせてウエイト付けし，秘匿による情報の損失量を定量化することも考えられる．たとえば，セル (i, j) を w_{ij} でウエイト付けして表示するとして，つぎの 3 つの方法が考えられる(他にも対数変換もウエイト付けのひとつと考えられる)．

i) 秘匿されるセルの数を最小とする方法… $w_{ij} =$ 一定と考える方法
ii) 秘匿されるセルの値の合計を最小とする方法… $w_{ij} = d_{ij}$ (セルの値)
iii) 秘匿されるセルに対応する respondents (回答者)の数を最小とする方法… $w_{ij} = N_{ij}$ (回答者の数)とし，N_{ij} の表を別に用意しておく．

表 7.

	ＳＣ１	ＳＣ２	ＳＣ３	ＳＣ４	ＳＣ５	計
ＥＡ１	80	253	54	-	-	387
ＥＡ２	641	3694	2063	746	-	7143
ＥＡ３	592	88	329	1449	1440	3898
ＥＡ４	57	x	946	x	2027	4281
ＥＡ５	78	-	890	1719	1743	4430
計	1448	4353	4281	4847	5210	20139

表 8.

	ＳＣ１	ＳＣ２	ＳＣ３	ＳＣ４	ＳＣ５	計
ＥＡ１	80	253	54	-	-	387
ＥＡ２	641	x	2063	x	-	7143
ＥＡ３	592	88	329	1449	1440	3898
ＥＡ４	57	x	946	x	2027	4281
ＥＡ５	78	-	890	1719	1743	4430
計	1448	4353	4281	4847	5210	20139

表 9.

	ＳＣ１	ＳＣ２	ＳＣ３	ＳＣ４	ＳＣ５	計
ＥＡ１	x	x	54	-	-	387
ＥＡ２	x	3694	2063	x	-	7143
ＥＡ３	592	88	329	1449	1440	3898
ＥＡ４	x	x	946	x	2027	4281
ＥＡ５	78	-	890	1719	1743	4430
計	1448	4353	4281	4847	5210	20139

表 7 に対する 2 次秘匿を考えよう．x が 1 次秘匿であり，– はもともと回答者がいないことを示している．付け加えて言うと，このような回答者がいないセル(空のセル)や値がゼロを示すセルは，回答者がいないという事実や値がゼロであるという事実が重要な情報であり，そのようなセルを秘匿すべきでないとされている．

まず，表 8 に秘匿するセルの数を最小にした表を示そう．2 次秘匿されたセルは 2 つであるが，結果的には 3694 という大きな値を示しているセルが秘匿されてしまう．

つぎに，秘匿されるセルの値の合計が最小になるように作成したのが，表 9 である．

この表 9 では，2 次秘匿の数が 5 つになってしまう．表 8，表 9 いずれが実用に適しているかには，利用する立場から別に基準が必要となる．

容易にわかるように，これらは整数計画法の問題と考えることができる．単純な集計表であれば，このような整数計画法は簡単に解くことができるが，集計表が複雑になると解くのは非常に難しくなる．秘匿処理のためのソフトウェアの開発のためには，これらを整数計画法によって定式化して解くアルゴリズムの開発が不可欠である．たとえば，De Waal (1994)のアルゴリズムの研究を例にあげることができる．第 4 節で後述するように，欧米では，ソフトウェアの開発と関連して，これらの数学モデルが非常に詳細に研究されてきた．このような整数計画モデルについては，Salazar-Conzalez (2002)に詳細に述べられている [2)]．また，ネットワークの理論を援用してこのような整数計画モデルの解法についての研究も Kelley et al. (1990)によって着手された．その後，Cox (1995)で詳細に検討されたが，この Cox (1995)以後には，このア

表 10.

	ＲＡ	ＲＢ	ＲＣ	計
Ａ１	20	50	10	80
Ａ２	0-25	19	5-30	49
Ａ３	0-25	32	4-29	61
計	45	101	44	190

表 11.

	ＲＡ	ＲＢ	ＲＣ	計
Ａ１	20	28-60	0-32	80
Ａ２	8	9-41	0-32	49
Ａ３	17	32	12	61
計	45	101	44	190

表 12.

1	2	計
2	2	4

→

1	2	計
0	0	5

プローチからの研究は進んでいないようである.

3) 区間開示

非表示(伏字「x」をつける)ではなく，区間で開示したり，他の数値に置き換えたりする方法も考えられる(他の数値に置き換える方法のうち，丸め法は次小節で取り上げる).

セルの値ではなく区間で開示する方法は，簡単であり，非表示よりも情報量損失が少ない. 先に示した表 3 に対して，表 10 や表 11 が作成できる.

表 10 では，範囲は 25 であるが，表 11 では，32 となる. もちろん，秘匿の意味からすると，範囲(レインジ)は広ければ広いほどよいが，非表示の場合の非負制約同様に，範囲の小さいほうは必ず 0 以上であり，かならずしも長い範囲を確保できない場合がある.

4) 丸め法

この方法は，セルの値をあらかじめきめられた値に置き換えてしまうことにより秘匿する. たとえば，丸めのベースを 5 として処理する伝統的丸め法(conventional rounding)の例を示す. 丸めのベースを 5 とする方法とは，0 から 2 までの値はすべて 0，3 から 7 までを 5，8 から 12 までを 10，13 から 17 までを 15，…とすることであり，簡単な例では表 12 のようになる.

問題点としては，必ずしも合計があわないこと(事実，表 12 の右側は合計が合っていない)，また，この表 12 の例では，計が 5 であることより，セルの値の合計が 3 より大きいことがわかる. 0，1，2 から 2 つを選んだ組み合わせで，合計が 3 以上のものは, (2, 1), (1, 2), (2, 2)の 3 種類しかないので，少なくともどちらかのセルの値が 2 であることが簡単にわかってしまう.

この表 12 の例では，丸めのベースを整数値としているが，乱数を加えてノイズを含んだ値として公表する方法も考えられる. これをランダム・ラウンディングとよぶ. 丸め法の一つではなく，独立した方法として，錯乱法(perturbation)とよぶ場合がある. どのような乱数が適当であるのか，また，そのようなノイズを含んだデータを利用する場合，どのような問題がおきるのかなど課題は多いと思われる. 日本においては，小野(1988)の解説があるが，実用には至っていない.

さらに最近提案された方法である PRAM (Post Randomization Method の略．マルコフ連鎖を用いた秘匿方法)の集計表のセル秘匿問題への応用も研究されている(Duncan and Fienberg (1999)を参照されたい)．これも，錯乱法の一つであると考えられる [3]．

4. 研究の動向

欧米では，オランダ統計局を中心とするヨーロッパの研究グループと，米国およびカナダのセンサス局を中心とする研究グループの 2 つの大きなグループにより，近年，非常にさかんに集計表のセル秘匿問題が研究されている．

ヨーロッパでは，すでに 1970 年代後半から，この分野への明確な問題意識が存在した．1980 年代になると，オランダ統計局の内部の研究報告(internal note)が作成され [4]，その後，第 4 次の EU 加盟国のプロジェクトとして，統計的開示を制御する問題(Statistical Disclosure Control，略して SDC)が取り上げられた(Esprit 20462-SDC)．SDC のプロジェクトのホームページ(アドレス http://www.cbs.nl/sdc/)によると，オランダ，イタリア，英国の統計局やオランダの Eindhoven University of Technology，英国の University of Manchester や University of Leeds，イタリアの Consorzio Padova Richerche (CPR)が参加し，1996 年 2 月に開始された [5]．Willenborg and De Waal (1996)は，オランダ統計局による統計データの秘匿に関するマニュアルとして利用でき，このプロジェクトの大きな研究成果のひとつである．さらに，Willenborg and De Waal (2000)は，その後の研究成果を含めて，増強された解説書である．両書を比べれば，その間の研究の進展に驚かされる．

この SDC プロジェクトは，秘匿問題の全体を取り扱っているが，重要な目的は秘匿処理のためのソフトウェアの開発であった．秘匿作業の実務をマニュアル作業から自動化した作業へと効率化を図ることが目的とされていた．後で述べるようにアメリカでは，すでに 70 年代からこのような方向でソフトウェア開発が進められており，ヨーロッパも追随したような形となった．

オランダ統計局がこのプロジェクトにより開発したソフトウェアは ARGUS (Anti-Re-identification General Utility System の略であるが，ギリシャ神話に登場する怪物の名前とかけている)と称し，個票データを秘匿処理して，個票データのまま開示するための μ-ARGUS と，この小論で取り扱う個票データを集計表で公表する際のセル秘匿問題に対処するための τ-ARGUS の 2 つのソフトウェアが開発された．ARGUS の開発については，Willenborg and Hundepool (1999)，μ-および τ-ARGUS のマニュアルを参照されたい．

SDC プロジェクトが 1998 年 7 月に終了した後，CASC プロジェクトに発展的に引き継がれた．CASC プロジェクトは，Computational Aspects of Statistical Confidentiality を意味し，秘匿方法の研究開発よりもコンピュータを利用した実用に重点をおいたプロジェクトである．これも第 5 次の EU 加盟国のプロジェクトのひとつであり，現在も進行中である．ホームページのアドレスは http://neon.vb.cbs.nl/casc/で，プロジェクトの詳細を知ることができる．μ-ARGUS，τ-ARGUS の開発は，両者ともに引き継がれ，先に示したホームページから入ってソフトウェアおよびマニュアルがダウンロードできる．このホームページによれば，ARGUS の利用などの各国の実務者向けの講習会も，EU のひとつの機関である AMRADS (Accompanying Measure to Research and Development in Official Statistics)の講習会として行われる(AMRADS については，http://amrads.jrc.it/を参照されたい)．

また，CASC プロジェクトの研究成果も，このホームページを通して知ることができる．出版物として Domingo-Ferrer (2002)が，CASC プロジェクトの成果である．これは，2001 年 11 月に催されたコンファレンスに集められた論文集であり，米国人研究者も寄稿している [6]．これが Lecture Notes in Computer Science シリーズの一冊であることからわかるように，プロ

ジェクトはいかに実用のために秘匿処理を行うかの方向に変化したことがわかる．

一方，この分野に関するアメリカにおける研究と実務については，米国センサス局が中心となり，1970 年代からすでに重要な研究テーマとなっていた．また一方では，調査とプライバシー保護の本質にかかわる問題も議論されてきたと思われる（Committee on Privacy and Confidentiality (1998)は，最近のものであるが，政府のサポートにより学会がこの問題を取り扱ってきたことがわかる）．

さらに，コンピュータの発達とデータベースの構築が進むとともに，その方面の関心として，セル秘匿問題も考察されるようになった．センサス局などの政府機関やカーネギーメロン大学などで研究は進展中であり，2001 年には，研究成果である Doyle et al. (2001)が出版された [7]．NISS（国家統計科学研究所， National Institute of Statistical Science の略）では，デジタル政府プロジェクト（Digital Government Project），Web を利用した統計情報の公開プロジェクトに取り組んでおり，開示リスクについても，この一環として研究されている．プロジェクトの詳細については，http://www.niss.org/dg/ で知ることができる．

先に述べたように，アメリカでは，秘匿作業を手作業からコンピュータによって自動化する研究も，70 年代から始まり，センサス局は，自動化のための秘匿のためのプログラムの開発と実際のデータへの応用に 30 年以上の歴史があると，Massell (2001)では自負しているほどである．

この開発を簡単に紹介すると，センサス局は，1977 年の経済センサスから，数学的な理論による自動化された秘匿プログラムを開発し，それを利用した．ただし，当時のそれは，単純な 2 次元表のなかで組み合わせ最適化のアルゴリズムに基づいたものであり，そのためのモジュールは INTRA という名称のもので，秘匿による情報量損失は，秘匿されたセルの数としていた．このモジュール INTRA を用いた基本的なプログラムは，1982 年の経済センサスでも踏襲されたが，1987 年の経済センサスでは，センサス局の Cox により提案された数理的ネットワークを用いた方法を利用した．第 3 節で述べたように，これは 2 次元表をネットワークと解釈したもので，この数理的ネットワークに基づく秘匿プログラムが開発された．このプログラムでは関連ある複数の表の間の秘匿の問題も，ある程度考慮されていたというが，しかし，2 次元表では完全な秘匿ができるが，3 次元表では秘匿漏れの可能性があったという．

1992 年の経済センサスでは，ネットワーク理論に基づいたプログラムがさらに改良され応用された．モジュールも，INTRA から，プログラム言語 FORTRAN のサブルーチン MCF（テキサス大学の Klingman によって開発されたものの名称）に置き換わった．

しかし，ネットワークの理論に基づくプログラムでは，3 次元表での秘匿漏れが懸念されるとして，一方では，線型計画法を用いた新たな秘匿プログラムが開発されており，これによって，3 次元表での秘匿まで考慮されることとなった．モジュールも，メリーランド大学の Kelly により開発された XMP に置き換わった．さらに，この線型計画法に基づくプログラムが改良され，XMP は 1999 年にセンサス局統計学研究部門によって新たにライセンスを取得した CPLEX に置き換わり，計算のスピードアップと 3 次元表や 4 次元表への発展が試みられている．

最後に，日本の状況について述べよう．「はじめに」で述べたように「商業統計表」「工業統計表」などの集計表は秘匿処理を施して公表されている．松本(1996)が「サービス業基本調査」における秘匿がどのように処理されたかを書き留めているが，セル秘匿問題についての近年の研究は，具体的には見あたらない．また，秘匿されたセルを含むデータを分析する方法として，稲葉・岩崎(1996,1997)が「商業統計表」からの実例を示している．

秘匿の基準についても，n-k 占有ルールで表現すれば，$n = 2, k = 100$ となるルールがかなり以前から応用されてきたようだが，これが適切かどうかの議論もなされていない．たとえば，先に述べた秘匿ソフトウェア τ-ARGUS では，$k = 100$ は入力できないので（$k = 99$ まで），日

本のようなルールは想定されていないと考えられる．

また，総務省統計局の「事業所・企業統計調査」では，従業員数については，秘匿を施さないことになっているが(たとえば，当該産業(産業小分類)の企業数が 1 の市区町村であっても，その産業の従業者数をそのまま開示するので，結果として，その事業所あるいは企業の従業者数がわかってしまうこととなる [8])，「工業統計表」や「商業統計表」では，従業者数は秘匿の対象である．このように秘匿の対象について同じ指定統計であっても整合性がない．

集計表自体の作成については，自動化作業がおこなわれているが，秘匿作業については，自動化が行われていない状況である．1998 年時点では，「工業統計表」「商業統計表」では，秘匿すべきセルはすべて自動的に(計算機により)位置が判明するので，それを人海戦術で秘匿するということを行っていた(実際は，絶対に危険セルが表示されないように完全に秘匿して，後で開示できるものだけをピックアップして修正するという方法をとっていた)．秘匿処理まで，すべてを自動化することは将来的には導入したいということであったが，プログラム開発などは未着手であった．

5. おわりに

ここまで，集計表におけるセル秘匿問題とはどのようなものか，また，おもに欧米の研究とその実用の進展について述べた．近年の欧米の複数の研究者は，グループとして非常な勢いでこの分野の研究や開発に着手しているのに対し，集計表のセル秘匿問題について，日本は，実務上どのように秘匿処理が行われているかという情報さえ非常に少ない．それは，実務においてずっと以前に(正確な年月の特定ができない)できあがった方法を踏襲しているためである．さらに，集計表を作成する作業は，定期的に繰り返されるので，新しい方法や自動化された作業手順を研究し実用に供してみる余裕がないことにもよるのであろう．しかし，コンピュータによる事務処理の自動化が非常な勢いで進展している昨今では，日本においても，このようなセル秘匿問題が，コンピュータプログラム化されて，秘匿作業が自動化される日も近いと思われる．ただし，その際に，今一度，統計を設計した当時のように，秘匿すべきセルの基準や秘匿方法，その際の情報量損失について，議論を深めることを期待したい．

注.

1) 数理的な情報損失のほかにも，どうしても秘匿してはならないような場合が存在する．筆者のヒアリングでは，日本において「工業統計表」では防衛産業に関するものは秘匿しないというルールがあるとのことであった．
2) もちろん，実用のために，これらの数理計画問題を解くことは簡単ではない．後述するように，秘匿のためのソフトウェア ARGUS においても，数理計画問題を解く最適化のルーチンは，開発チームが独自に開発したものではなく，外部の企業にライセンス料を支払って利用する．
3) 錯乱法については，Willenborg and De Waal (2000)の第 9 章が詳しいが，完成された方法ではないという印象を受ける．
4) たとえば，Bemelmans-Spork, M. (1983). The subroutine SUPPRESSION for the protection of a two-dimensional tables, internal description (in Dutch), Internal Note, Statistics Netherlands をあげることができる．これは，De Waal (1994)の参考文献欄にあるが，内部報告書のために未入手である．
5) 上記のホームページによると，共同プロジェクトとしては 1998 年 7 月まで続くとされている．また，このホームページは，1999 年 8 月までアップデートされ，その時点まで

は，さかんに研究がなされたようである．

6) 雑誌 *Statistical Journal of the United Nation Economic Commission of Europe* の Vol. 18, No. 4 (2001)も特集号である．
7) こちらについても，ヨーロッパの研究者が参加しているので，まったくの米国の研究成果とは言えないが．
8) 筆者が研究会の席上で聴講したことによると，「事業所統計調査」が設計された当初は，石炭産業が従業員数の開示に反対したが，その後，石炭産業の斜陽化とともに反対が弱くなった(しなくなった？)ことによるらしい．

謝　辞

この小論を丁寧に読んで，ご助言をくださった匿名査読者に感謝いたします．また，小論は統計数理研究所共同研究プログラム(14-共研-2024)「個票データの開示におけるリスクの評価と官庁統計データの公開への応用」の援助を受けている．

参 考 文 献

Committee on Privacy and Confidentiality (1998). *Surveys and Privacy*, American Statistical Association, Alexandria, Virginia.

Cox, Lawrence H. (1980). Suppression methodology and statistical disclosure control, *J. Amer. Statist. Assoc.*, **75**, 377–385.

Cox, Lawrence H. (1981). Linear sensitivity measures in statistical disclosure control, *J. Statist. Plann. Inference*, **5**, 153–164.

Cox, L. H. (1995). Network models for complementary cell suppression, *J. Amer. Statist. Assoc.,* **90**, 1453–1462.

Cox, L. H. (2001). Disclosure risk for tabular economic data, network models for complementary cell suppression, *Confidentiality, Disclosure, and Data Access*: *Theory and Practical Applications for Statistical Agencies* (eds. P. Doyle, J. I. Lane, J. J. M. Theeuwes and L.V. Zayatz), 167-183, Elsevier, Amsterdam.

De Waal, A. G. (1994). The hypercube method for suppression in tables, Report 1–11, Department of Statistical Methods, Statistics Netherlands, Voorburg.

デニング，D. E. R. (1988)『暗号とデータセキュリティ』(上園忠弘，小嶋　格，奥島昌子 訳)，培風館，東京．(Denning, D. E. R. (1982).*Cryptography and Data Security*, Addison-Wesley, Reading, Massachusetts.)

Domingo-Ferrer, J. (ed.)(2002). *Inference Control in Statistical Databases*, Lecture Notes in Comput. Sci., **2316**, Springer-Verlag, New York.

Doyle, P., Lane, J. I., Theeuwes, J. J. M. and Zayatz, L. V. (2001). *Confidentiality*, *Disclosure*, *and Data Access*: *Theory and Practical Applications for Statistical Agencies*, Elsevier, Amsterdam.

Duncan, George and Fienberg, Stephan (1999). Obtaining information while preserving privacy: A Markov perturbation method for tabular data, *Statistical Data Protection Proceedings of the Conference*, *Lisbon*, *25 to 27 March 1998*, EUROSTAT, Luxemburg.

稲葉由之，岩崎　学(1996). クロス集計表における秘匿の影響に関する数値的評価，応用統計学，**25**，61–72.

稲葉由之，岩崎　学(1997). 統計表における秘匿の補完法，日本統計学会誌，**27**，263–280.

Kelly, J., Golden, B. L. and Assad, A. A. (1990). Controlled rounding of tabular data, *Oper. Res.*,

38, 760–772.

Loeve, J. A. (2001). Notes on sensitivity measures and protection levels, Division of Technology and Facilities, Methods and informatics department, Statistics Netherlands, Voorburg.

Massell, Paul B. (2001). Cell suppression and audit programs used for economic magnitude data, Statistical Research Report Series, No. RR2001/01, Statistical Research Division, U. S. Bureau of the Census, Washington, D. C.

松本正博(1996). サービス業基本調査の秘匿処理方法について，統計局研究彙報，**54**，57–86.

小野達也(1988). ランダム・ラウンディングの理論と方法，統計局研究彙報，**6**，97–135.

Robertson, D. A. and Ethier, R. (2002). Cell suppression: Experience and theory, *Inference Control in Statistical Databases* (ed. J. Domingo-Ferrer), Lecture Notes in Comput. Sci., **2316**, 8–20, Springer-Verlag, New York.

Salazar-Gonzalez, J. J. (2002). *A Unified Mathematical Programming Framework for different Statistical Disclosure Limitation Methods for Tabular Data*, CASC-Project, Deliverables of the CASC-project No. 4.1-D1, Statistics Netherlands, Voorburg.

瀧　敦弘(1997). 表に整理されたマイクロデータの秘匿と開示，「ミクロ個票データの開示とプライバシー保護に関する理論的考察とその現実データへの適用」，平成 8 年度文部省科学研究費補助金報告書(重点領域，課題番号 08209107，研究代表者　加納　悟)，45–67.

瀧　敦弘(2002). 表に整理されたマイクロデータの秘匿と開示，「統計データの個票開示における局所秘匿方法の確立」，平成 11 年度～平成 13 年度文部省科学研究費補助金報告書(基盤研究(B)(1)，課題番号 11558026，研究代表者　竹村彰通)，173–178.

Willenborg, L. and De Waal, T. (1996). *Statistical Disclosure Control in Practice*, Lecture Notes in Statist., **111**, Springer-Verlag, New York.

Willenborg, L. and De Waal, T. (2000). *Elements of Statistical Disclosure Control*, Lecture Notes in Statist., **155**, Springer-Verlag, New York.

Willenborg, L. and Hundepool, A. (1999). ARGUS for statistical disclosure control, *Statistical Data Protection Proceedings of the Conference, Lisbon, 25 to 27 March 1998*, EUROSTAT, Luxemburg.

350 Proceedings of the Institute of Statistical Mathematics Vol. 51, No. 2, 337–350 (2003)

Statistical Disclosure Control for Tabular Data and Current Trends in Theoretical and Applied Research

Atsuhiro Taki

(Faculty of Economics, Hiroshima University)

We introduce the theory of statistical disclosure control (SDC) for tabular data and survey the current status and perspectives of theoretical research. The second section shows the idea of cell sensitivity and its measures. Both the n-k percent rule and the prior-posterior rule are influential dominance rules. The third section outlines control methods for tabular data. These methods can be roughly classified into two groups: cell suppression, and perturbation. This section mainly treats cell suppression. Recently these methods have been aggressively researched and developed by groups in Europe and America. The fourth section gives a brief sketch of their activities and introduces SDC software.

Key words: Statistical disclosure control, tabular data, SDC-project, CASC-project, SDC software.