差分プライバシー基準に基づく情報秘匿手法の一考察

寺田 雅之[1]　竹内 大二朗[2]　齊藤 克哉[2]　本郷 節之[2]

概要：集計データのプライバシーを差分プライバシー基準に基づいて保護する上で，データの統計的正確性と計算効率に着目した手法を提案する．差分プライバシー基準は，安全性に対する数学的な裏付けが保障されているものの，(1) 非負データが負の値となってしまう場合が生ずる，(2) 広範囲のデータの集計値において真値からの偏差が大きくなる，(3) 疎なデータ分布を密なデータ分布へと変化させることにより計算量の著しい増大を招く，などの実用上の課題を持つ．本報告では，これら三点の課題を解決する手段として、Wavelet 変換と Top-down 精緻化処理と呼ぶ方法を組み合わせた，差分プライバシー基準を満たす新たなプライバシー保護方式を提案し，国勢調査データを用いた提案手法の評価結果を示す．

A Practical Differentially Private Method for Publishing Large and Sparse Tabular Data

Masayuki Terada[1]　Daijiro Takeuchi[2]　Katsuya Saito[2]　Sadayuki Hongo[2]

1. はじめに

人々に関係するデータベースから作成された集計データを公開するにあたっては，プライバシー保護への十分な配慮が必要となる．本稿では，これら集計データのプライバシーを差分プライバシー (differential privacy) 基準に基づいて保護する上で，データの統計的な正確性と計算効率を改善した，新たなプライバシー保護手法を提案する．なお，本稿における集計データとは，元のデータベースに含まれる個々のデータ群 *1 から作成した，「ある条件を満たす」データの個数を数えあげた数値データ (セル) の集まり *2 を意味し，特に広範囲の空間分布を表す集計データ (たとえば人口分布や交通量分布など) などの大規模で広い属性空間を持つ集計データを主な検討の対象とする．

集計データに対するプライバシー保護の必要性は，統計分野において古くから議論されてきた．これらの統計分野におけるプライバシー保護手法は，統計的開示制御 (statistical disclosure control, SDC) と総称される [9], [10], [11], [20]．たとえば，セル秘匿 (cell suppression) 基準や $n-k\%$ 基準などに基づく各種の秘匿方式が集計データに対する SDC の手法として挙げられる [21]．SDC は，国内外における公的統計，すなわち国勢調査や各種の社会・経済統計などを公開するにあたって，統計から個人のプライバシーが侵されることがないよう専門家により注意深く適用されており，その安全性に関しても長年の実績を持つ．

その一方，公的統計分野のみならず，近年の計算機能力の向上や記憶装置の容量拡大を背景として，各種分野でのビッグデータ活用への期待が高まっている．これらの応用では，公的統計と比べてより高次元のデータを扱うことが多いことなどから，従来の SDC の手法では必ずしも十分にデータの有用性を確保できない懸念がある．そこで，近年になり，情報セキュリティ分野やデータベース処理分野などにおいて，プライバシーを保護しつつ有用なデータを公開するのための様々な新しい基準や手法が提案されている．これらの技術は，プライバシー保護データ公開 (privacy-preserving data publishing, PPDP) もしくは出力プライバシー (output privacy) などと呼ばれる [6]．本稿では，以降 PPDP 技術と総称する．PPDP 技術の例としては，k-匿名性 (k-anonymity) 基準 [15] およびその変形に基づく手法や，エントロピー基準に基づく手法，差分プライバシー基準 [3] に基づく手法などが挙げられる．

[1] (株)NTT ドコモ 先進技術研究所
Research Laboratories, NTT DOCOMO, Inc.
[2] 北海道科学大学 工学部
Faculty of Engineering, Hokkaido University of Science

*1 個票あるいは生データとも呼ばれる．
*2 度数表 (frequency table) とも呼ばれる．

しかし，これらの PPDP 技術は，それぞれ攻撃者が持つ目的および能力や背景知識に関する前提が異なり，その安全性について一概に議論することが困難であることから，実際のデータ活用における適用は容易ではない．すなわち，これらの技術を実際に適用する上では，扱うデータの性質や応用ごとに，「どのプライバシー保護基準に基づいて，どの手法によりプライバシーを保護するべきか」を適切に判断することが求められるが，これをすべてのデータ活用の現場に求めることは現実的とは言いがたい．

そこで，本研究では Dwork らにより 2006 年に提案された差分プライバシー基準 (differential privacy)[3] に着目する．これは，「ある人が (加工データを作成する上での元データとなる) データベースに含まれるか否かの，加工データからの判別困難性」を安全性の根拠とするプライバシー保護基準である．差分プライバシーは，他の多くのプライバシー保護基準と異なり，任意の背景知識を持つ攻撃者や未知の攻撃に対して数学的な安全性が与えられているという優れた性質を持つ．

差分プライバシー基準を実現する代表的な手段としては，集計データの各セルに対して，それぞれに平均 0 の Laplace 分布に従う独立した乱数 (Laplace ノイズ) を付与する手法 (Laplace メカニズム) が挙げられる．この手法は，実装が簡単であり，かつ Laplace ノイズの付与は集計データに含まれる個々のセルに対して独立に実施できるため並列処理による高速化が容易である，という実装上の利点を持つ．

しかし，Laplace メカニズムをそのまま実際の集計データのプライバシー保護に適用することは，特に大規模な集計データにおいて実用上の困難を伴う．その理由として，Laplace メカニズムが適用された集計データは，(実際の集計データではありえない) 負数を多く含みうるためその後の利用に困難を伴うこと (**非負制約の逸脱**)，複数セルの部分和を取った際の誤差が大きく有用性が劣化すること (**部分和精度の劣化**)，集計データの密度 (非 0 値の割合) を大きく増大させてしまい，大規模な集計データに適用した際に計算量や出力データ量が現実的ではなくなること (**計算量の増大**)，の三点の課題が挙げられる．

これらの課題に対して，いくつかの部分的な改善方式が提案されている [1], [2], [8], [12], [16], [17]．しかし，いずれの方式においても，前述の三点の課題を同時に解決することはできず，またこれらの方式を単純に組み合わせることも困難である．

そこで，本稿では Wavelet 変換と Top-down 精緻化と呼ぶ手法を組み合わせた，差分プライバシー基準を満たす新たなプライバシー保護手法を提案する．本提案手法は，Wavelet 変換を適用した上で Laplace メカニズムにより差分プライバシーを保証する点で，Xiao らの手法 [16], [17] を高度化したものとみなすこともできる．提案手法では，Wavelet 変換の適用がもたらす効果であるところの，部分和誤差の増大への対処が可能である特長を活かし，さらに Top-down 精緻化という独自の手法を導入することにより，併せて非負制約への対処と計算量の増大に対する解決を可能にする．

2. 従来技術と課題

2.1 集計データ

集計データとは，1 もしくは複数の属性を持つレコードの集合から構成されるデータベースにおいて，ある属性 (もしくは属性の組み合わせ) に該当するレコードの個数を数えあげた値の集合である．集計データは，様々な統計分析における基礎データとして広く使われている．たとえば，国勢調査の結果に基づく各種の地域別人口や，パーソントリップ調査結果に基づき (出発地，到着地) の組ごとに移動人数を集計した OD (origin-destination) 表などの各種の公的統計や，携帯電話の運用データから日本全国の属性別人口を時間帯別に推計したモバイル空間統計 [13], [18] などが相当する．

集計データは以下のように定義できる．l 個のレコードから構成されるデータベース $D = \{x_1, x_2, \ldots, x_l\}$ を考える．ここで，各レコード x_i は d 次の属性空間 $A = A_1 \times A_2 \times \cdots \times A_d$ に属するベクトル値を持つ ($x_i \in A$)．集計データとは，ある与えられた A の部分空間の集合 $C = (C_1, C_2, \ldots, C_n)$ $(C_j \subseteq A)$ に対する，D 内の各部分空間に属するレコードの個数の集合 $V = (v_1, v_2, \ldots, v_n)$ である．ここで，$v_j = Count(D, C_j)$ であり，これは C_j に属する D 中のレコードの個数 $|x|(x \in D, x \in C_j)$ を意味する．

一般的には，各属性の定義域 A_k の互いに素な部分空間集合の直積が C として用いられる．この時の集計データは分割表 (contingency table) と呼ばれる．たとえば，$A_1 =$ { 男性, 女性 }, $A_2 =$ 年齢 としたとき，20 歳を境として A_2 を成人と未成年の 2 つの部分空間に分け，$C =$ { 男性, 女性 } × { 成人, 未成年 } = { 男性・成人, 男性・未成年, 女性・成人, 女性・未成年 } を与えて作成した集計データ V は分割表である．分割表におけるそれぞれの値 v_j を，セルもしくはセル値 (cell / cell value) と呼ぶ．以降，本稿では断りがない限り集計データは分割表の形式をとるものとする．

高次元のデータベースから作成された集計データや，集計データの作成にあたって属性を多数の部分空間に分割する場合など，実際の集計データは 0 値のセルを多数含む疎なデータ (sparse data) になることが多い．すなわち，論理的なセルの総数を $n(= |C|)$，そのうち非 0 値を持つセル数を m とすると，$m \ll n$ となる．たとえば，ある日時における日本全国の属性別人口分布を，500m メッシュを単位として，5 歳区分の年齢層別，男女別，居住市区町村

別に集計したとする．日本の国土にかかる 500m メッシュの数は約 1,500,000[19]，年齢層の数は約 20，市区町村数は約 2,000 であるため，論理的なセルの総数 $n = |C|$ はおよそ $1.5 \cdot 10^6 \times 20 \times 2 \times 2 \cdot 10^3 = 1.2 \cdot 10^{11}$ となる．これは日本の総人口，約 $1.2 \cdot 10^8$ を 1,000 倍ほども上回る数字であり，そのまま計算機上で扱うには極めて効率が悪い．

そのため，実際の集計データは，実装上は非 0 値を持つ m 個のセルのみについての，(j, v_j) の組のリスト (長さ m) として表現されることが多い．これは COO 形式 (coordinate format) とも呼ばれる [14]．

2.2 差分プライバシー

差分プライバシー [3], [5] は，識別不能性に基づくプライバシー基準の一種である．直感的には，「ある個人のデータを含むデータベースに対する問い合わせ結果が，その個人のデータを含まないデータベースへの問い合わせ結果と区別できないなら，その問い合わせは安全である (個人に関するプライバシーを開示しない)」という考え方によりプライバシーを規定する．

たとえば，個人に関するデータの集合から構成されるデータベース D と，データベース問い合わせ $f(\cdot)$ を考える．なお，$f(\cdot)$ は (出力に摂動を加えるなど) 確率的な出力を持つ関数 (randomized function) である．このとき，データベース D への問い合わせ $f(D)$ と，ある個人 i に関するデータ $x_i (\in D)$ を D から取り除いたデータベース $D' (= D \setminus x_i)$ への問い合わせ結果 $f(D')$ が区別できないなら，$f(D)$ から x_i に関して意味がある情報を抽出することはできない．すなわち，個人 i のプライバシーは保護される．

より厳密には，差分プライバシーはパラメータ ϵ を用いて以下のように定義される．

定義 1 任意の隣接した *(互いにたかだか 1 要素しか異ならない)* データセット D_1 および D_2 *($D_1, D_2 \in \mathcal{D}$)* に対し，ランダム化関数 *(randomized function)* $\mathcal{K} : \mathcal{D} \to \mathcal{R}$ が下式を満たすとき，$\mathcal{K}$ は ϵ-差分プライバシーを満たす．ただし，ここで S は $\mathcal{K}$ の出力空間 $\mathcal{R}$ の任意の部分空間である *($S \subseteq \mathcal{R}$)*.

$$\frac{\Pr[\mathcal{K}(D_1) \in S]}{\Pr[\mathcal{K}(D_2) \in S]} \leq e^{\epsilon}. \tag{1}$$

なお，上記のランダム化関数 $\mathcal{K}$ は「メカニズム (mechanism)」とも呼ばれる．

差分プライバシーの特徴として，その安全性定義がデータの性質や攻撃者の能力 (攻撃手段や攻撃者の背景知識) に依存しないことが挙げられる．すなわち，データベースに異常値が混入していても安全性が損なわれることがなく，また任意の背景知識を持つ攻撃者や未知の攻撃に対して安全である．これは，差分プライバシー基準を正しく満たしたデータは，データ作成時には未知であった新たな攻撃手法が発見されたり，もしくは想定外の背景知識を持つ攻撃者が現われたとしても，その安全性が損なわれないということを意味する．Dwork は，差分プライバシーが持つこの性質について，差分プライバシーは (“ad hoc” ではなく) “ad omnia” なプライバシー保証を与える，としている [4]．

2.3 Laplace メカニズム

差分プライバシーを実現するためには，定義 1 を満たすメカニズム $\mathcal{K}$ が必要となる．差分プライバシーを実現する代表的なメカニズムとしては Laplace メカニズム [*3] が挙げられる．

Laplace メカニズムは「問い合わせ結果に対して Laplace ノイズを加える」だけという簡単な手段によって実現することができる．ここで Laplace ノイズとは，0 を平均とした Laplace 分布から独立に抽出された乱数である．Laplace 分布の確率密度分布は平均 μ とスケール λ を用いて下式で与えられる．

$$f(x; \mu, \lambda) = \frac{1}{2\lambda} e^{(-|x-\mu|/\lambda)}. \tag{2}$$

以降，平均 0，スケール λ の Laplace 分布に従って発生させた Laplace ノイズを $\mathrm{Lap}(\lambda)$ とし，k 個の独立した $\mathrm{Lap}(\lambda)$ からなるベクトル列を $\mathrm{Lap}(\lambda)^k$ と記載する．

Laplace メカニズムで用いられる Laplace ノイズのスケール λ は，定義 1 におけるパラメータ ϵ と，問い合わせの種類ごとに定まる「(大域的) 感度 (global sensitivity, GS)」によって与えられる．具体的には，GS_f を問い合わせ $f : \mathcal{D} \to R$ の感度としたとき，f に対応するメカニズム $\mathcal{K}_f$ は下式で定義される．

$$\mathcal{K}_f(X) = f(X) + \mathrm{Lap}(GS_f/\epsilon), \tag{3}$$

$$GS_f = \max_{D_1, D_2} ||f(D_1) - f(D_2)||_1. \tag{4}$$

ここで，D_1 および D_2 は，任意の隣接 [*4] したデータセットのペアである．

2.4 集計データへの Laplace メカニズムの適用

差分プライバシーが保証された集計データは，理論的には Laplace メカニズムを用いることにより簡単に作成することができる．前述の通り，集計データは計数問い合わせ $v_j = Count(D, C_j)$ の集合である．計数問い合わせに対する感度 GS_{count} は 1 であることが知られている．そのため，各セル v_i にスケール $1/\epsilon$ の Laplace ノイズを加

[*3] 集計データの構成要素である計数問い合わせ (count query) に対しては，幾何分布に従う乱数を用いた幾何メカニズム (geometric mechanism) もしばしば用いられる ([2], [7] など)．幾何メカニズムは Laplace メカニズムの特殊形とみなすことができるため，本稿では代表して Laplace メカニズムのみを扱う．

[*4] 定義 1 参照.

えた値，

$$v_i^* = v_i + \mathrm{Lap}(1/\epsilon) \tag{5}$$

は ϵ-差分プライバシーを満たす．

ここで，C に含まれる各部分集合が互いに素であるとする．すなわち，$\forall i, j (i \neq j), C_i \cap C_j = \phi$ が成立するとする．V が分割表の場合はこの条件を満たす．このとき，差分プライバシーの並列合成則により，上記の Laplace メカニズムを適用した v_i^* の集合 $V^* = \{v_1^*, v_2^*, \ldots, v_n^*\}$ もまた ϵ-差分プライバシーを満たす．言い換えると，集計データ V の全セルにそれぞれ Laplace ノイズ $\mathrm{Lap}(1/\epsilon)$ を加えることにより，ϵ-差分プライバシーを満たした集計データ V^* を得ることができる．

しかし，実際の集計データにこの方法を適用することは，しばしば現実的ではない．その理由として，以下の 3 点が挙げられる．

第 1 の問題は，非負制約の逸脱である．集計データは計数値の集合であるため，その定義から各セルの値は非負でなくてはならない．しかし，Laplace ノイズは正負いずれの値も取りうるため，Laplace メカニズムを適用したデータは (実際の集計データではありえない) 負数を多く含みうる．特に，0 値のセルは 1/2 の確率で負値をとるため，疎な集計データではほぼ半数のセルが負値をとることになる．これは，データの利用者にとって不自然に感じられるだけでなく，分析プログラムの予期せぬ異常動作を引き起こす可能性をもたらすなど，データの利用に著しい困難を生じさせる．なお，ナイーブな対策として，Laplace メカニズムの適用後に負数のセルを 0 値のセルに校正することにより，見かけ上は非負制約に従う集計データを生成することもできる．しかし，この方法はセル値の平均や部分和に大きな過大バイアスがかかる (平均や部分和の期待値が，元の集計データにおける値に対して大きく「上ぶれ」する) ことになり，実用に耐えがたい．

第 2 の問題は，部分和の劣化である．集計データを利用する際には，個々のセルの値だけではなく，複数のセルの値を合算した部分和が用いられることも多い．しかし，Laplace メカニズムを適用した集計データでは，その精度が大きく劣化する．たとえば，500m メッシュを単位とした人口分布から，2x2 個のメッシュ人口を合算して 1km メッシュ人口として利用することなどは一般的である．しかし，Laplace メカニズムを適用した集計データにおける部分和には，和をとる対象のセル数と等しい数の Laplace ノイズが重畳して加算されることになる．たとえば，上記の例では，算出された 1km メッシュには 4 個分のノイズが重畳される．2km メッシュであれば 16 個分となる．すなわち，部分和の対象範囲が広ければ広いほどノイズによる真値からの偏差が大きくなり，その有用性も大きく劣化する．

第 3 の問題は，データ密度の増大である．前述の通り，たとえば日本全国の属性別人口分布など，大規模な集計データは疎なデータであることが多い．すなわち，論理的なセルの総数を n とし，そのうち 0 以外の値をとるセルの個数を m とすると，$m \ll n$ となる．ここで，Laplace メカニズムによるノイズの付与は，(0 値のセルを含めた) n 個のセルに対して行なう必要がある [*5]．すなわち，COO 形式などにより $O(m)$ のデータ量で表現された集計データに対し，$O(n)$ の計算量による Laplace ノイズの付与により $O(n)$ のデータ量を持つ集計データを出力することになる．これは $m \ll n$ の場合に非効率であるだけでなく，そもそも前述の日本全国の属性別人口の例のように n が非常に大きくなる場合には現実的ではない．

2.5 関連研究

これらの課題に対し，これまでにいくつかの部分的改善手法が提案されている．

Barak ら [1] は，集計データに対する離散 Fourier 変換の導入と，周波数領域における Laplace メカニズムの適用により部分和を改善し，さらに線形計画法に基づいて非負制約の逸脱を解消する方法を提案している．この手法では，元の集計データに離散 Fourier 変換を適用した上で，各周波数成分 (部分和をとる範囲に相当する) に対応する Fourier 係数にそれぞれ Laplace メカニズムを適用することにより，部分和の精度を改善する．また，Laplace メカニズムの適用後に，(原データを参照することなく) 適用後データのみを参照して線形計画法を適用することにより，差分プライバシーを保ちつつ非負制約の逸脱を解消する．しかし，データ密度の増大への対処はなされておらず，また線形計画法の計算負荷が大きいことから，大規模な集計データへの実用的な適用は困難である．

Xiao ら [16], [17] は，部分和の改善に離散 Wavelet 変換とその概念的な拡張である Nominal Wavelet 変換と呼ぶ方式を用いる方式を提案している．この提案手法を Xiao らは “Privelet” と名付けている．Privelet では，Nominal Wavelet 変換の導入により Barak らの方式では扱えなかった階層的な名義尺度を持つ属性 (地方-都道府県-市区町村など) への適用を可能にしている．

Xiao らの手法では，Barak らが Fourier 変換を導入したのに対し，Haar 基底に基づく離散 Wavelet 変換を導入し，Wavelet 係数に対して Laplace メカニズムを適用することにより部分和精度を改善する．Haar Wavelet 変換/逆変換は概念や実装が単純であり，Fourier 変換に基づく手法では自明ではない階層的な名義尺度への適用を，比較的簡単な拡張で可能としている．しかし，その一方で非負制約の逸

[*5] もし 0 値を無視してノイズ付与を実施した場合，ある条件を満たす人が存在したか/しなかったかが開示されることになる．すなわち個人のプライバシーの開示になりうる．

脱を解決する手段は与えられていない．また，Barak らの手法と同様にデータ密度の増大についても解決されない．

なお，Barak らの手法や Xiao らの手法は，いずれも線形変換の一種である Fourier 変換や Wavelet 変換の適用後に Laplace メカニズムを適用し，その後に逆変換をかけることにより差分プライバシーを満たす集計データを得る (Barak らの手法では，さらにその後に線形計画法による精緻化を適用する)．すなわち，これらの手法は，線形変換 → Laplace メカニズムの適用 → 逆線形変換，と一般化できる．このアプローチにより差分プライバシーを満たす手法は，Li ら [12] により Matrix メカニズムと名付けられている．

Cormode ら [2] は，「ある閾値」を越える値を持つセルの値だけがあれば良いという応用を前提とした上で，計算量の増大を回避する方式を提案している．この方式では，Laplace ノイズの付与により 0 値のセルが「閾値」を超える (= 出力の対象となる) 確率をあらかじめ計算しておき，その確率に従った個数のセルをランダムに抽出する．そして，以降の処理は，ここで抽出されたセルと非 0 値を持つセルのみを対象とする．これにより，閾値が十分に大きければ計算量・データ量ともに大きく削減されるため，計算量の問題は回避される．また，閾値は正の値をとることから，非負制約の問題も生じない．

その一方，Cormode らの手法は「閾値」を下回る値を持つセルは無視されてしまうことから，集計データに含まれるロングテイル部分を分析することが全く不可能になるという新たな問題が発生する．また，部分和にノイズだけでなく (過少方向の) バイアスを生じさせることにもなるため，単に Laplace メカニズムを適用したデータ以上に部分和の利用は困難なものとなる．

このように，いずれの手法も前述の 3 つの問題の全てを同時に解決しない．また，これらの手法を単純に組み合わせて問題を解決することも困難である．

3. 提案方式

前節で上げた 3 点の課題を解決する手段として，Wavelet 変換と Top-down 精緻化処理と呼ぶ方法を組み合わせた，差分プライバシー基準を満たす新たなプライバシー保護方式を提案する．本提案方式は，集計データに Wavelet 変換を適用した上で，Wavelet 係数に対して乱数ノイズの付与を行っている点で，Xiao らの手法 (Privelet 法) を高度化したものと見做すこともできる．

提案方式では、Wavelet 変換による処理がもたらす効果であるところの，部分和精度の改善が可能である特長を活かし，さらに Top-down 精緻化という手法を導入することで，非負制約からの逸脱への対処と計算量の著しい増大に対する解決を可能にする．また，Top-down 精緻化により非負制約からの逸脱を排除することにより，提案手法による部分和の精度は，特に狭範囲の部分和において Privelet 法と比較してさらに改善される．

以下，提案方式の考え方と振る舞いを説明する．ここでは，簡単のため一次元のデータ系列を対象にして述べるが，Wavelet 変換の標準分解 (standard decomposition) の適用により，多次元のデータ系列に拡張することができる．

まず，提案方式と等価な出力を得るアルゴリズムを説明する．これをアルゴリズム 1 と呼ぶ．アルゴリズム 1 は，前節で示した 3 つの問題のうち，非負制約の逸脱と部分和の劣化の 2 つの問題を解決するが，データ密度の増大を解決しない．本アルゴリズムを改善し，データ密度の増大を解決する方法については後述する．

処理対象の集計データを $V = (v_1, v_2, \ldots, v_n)$ とする．これを原データ系列と呼ぶ．簡単のため，$n = 2^k (k \in \mathbb{N})$ であるとする．以下，アルゴリズム 1 の概略手順を示す．アルゴリズム 1 は，V を入力として差分プライバシーを満たした $V^+ = (v_1^+, v_2^+, \ldots, v_n^+)$ を出力する．

(1) 原データ系列 V に Haar Wavelet 変換を適用する．出力として得られた Wavelet 係数系列を W とする．
(2) W に Laplace メカニズムを適用する．得られたノイズつき Wavelet 係数系列を W^* とする．
(3) W^* に Top-down 精緻化を適用する．得られた精緻化ずみ Wavelet 係数系列を W^+ とする．
(4) W^+ に逆 Haar Wavelet 変換を適用し，出力 V^+ を得る．

なお，アルゴリズム 1 は，Privelet 法に Top-down 精緻化の手順を加えたものとみることができる．言い替えると，アルゴリズム 1 において Top-down 精緻化を省略し，$W^+ = W^*$ とした場合，その出力は Privelet 法の出力と等価になる．

以下，アルゴリズム 1 で用いられる各手順について詳述する．

3.1 Haar Wavelet 変換

Haar Wavelet 変換 $\mathcal{H} : \mathbb{R}^n \to \mathbb{R}^n$ は，階段関数の一種である Haar 関数を母 Wavelet とした離散 Wavelet 変換の一種であり，長さ $n = 2^k (k \in \mathbb{N})$ のベクトル列 $V = (v_1, v_2, \ldots, v_n)$ を，同じ長さを持つベクトル列 $W = (w_1, w_2, \ldots, w_n)$ に変換する．$\mathcal{H}$ は逆変換関数 $\mathcal{H}^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ を持ち，任意の $V \in \mathbb{R}^n$ について $V = \mathcal{H}^{-1}(\mathcal{H}(V))$ が成立する．

$\mathcal{H}$ は Haar 分解 $\mathcal{H}_1 : \mathbb{R}^n \to \mathbb{R}^{n/2} \times \mathbb{R}^{n/2}$ を再帰的に k 回適用することにより構成できる．Haar 分解 $\mathcal{H}_1$ は，長さ 2^l のベクトル列 $Y = (y_1, y_2, \ldots, y_{2^l})$ を，長さ 2^{l-1} のベクトル列 cA, cD に分解する．

$$cA|cD = \mathcal{H}_1(Y), \tag{6}$$

$$cA = (\frac{y_1+y_2}{2}, \frac{y_3+y_4}{2}, \ldots, \frac{y_{2^l-1}+y_{2^l}}{2}), \tag{7}$$

$$cD = (\frac{y_1-y_2}{2}, \frac{y_3-y_4}{2}, \ldots, \frac{y_{2^l-1}-y_{2^l}}{2}). \tag{8}$$

cA と cD はそれぞれ，Y において隣り合う 2 つの値の平均のベクトルと差分のベクトルとなる [*6]．cA を近似係数ベクトル，cD を詳細係数ベクトルと呼ぶ．

生成された近似係数ベクトル cA を入力として，再び Haar 分解をほどこすと，長さ 2^{l-2} の近似係数ベクトルと詳細係数ベクトルの組が得られる．すなわち，V を初期入力として，この分解を再帰的に k 回繰り返すと，最終的には k 個の詳細係数ベクトルと近似係数ベクトルが得られる．

$$cA_1 \mid cD_1 = \mathcal{H}_1(V), \tag{9}$$

$$cA_i \mid cD_i = \mathcal{H}_1(cA_{i-1}). \quad (i \in \{2..k\}) \tag{10}$$

ここで，cA_i, cD_i は i 回目の Haar 分解により得られた出力であり，これをレベル i の係数ベクトルと呼ぶ．このとき，以下の連接により W を構成できる．

$$W = (cA_k \mid cD_k \mid cD_{k-1} \mid \ldots \mid cD_1) \tag{11}$$

なお，$|cA_k| = 1$ であり，$|cD_i| = 2^{k-i}$ であることから，W の長さ $|W|$ は，$|W| = 1 + \Sigma_{i=0}^{k-1} 2^i = 2^k = |V|$，すなわち V の長さと等しくなる．

3.2 Laplace メカニズムの適用

この手順では，Wavelet 変換により得られた係数系列 W の各要素に Laplace メカニズムを適用して差分プライバシーを満たす係数系列 W^* を得る．

ここで，(Laplace メカニズムにより) 付加する Laplace ノイズのスケールは，Haar Wavelet 変換におけるレベルにより異なる．具体的には，$\lambda = 2(1+k)/\epsilon$ として，W に含まれるレベル i の係数にそれぞれ $\mathrm{Lap}(\lambda/i)$ を加える．これを W^* とする．

これにより e-差分プライバシーを満たす W^* が得られることは，直感的には以下の考え方により理解できる [*7]．

(1) Haar Wavelet 変換の定義により，v_j が 1 変化すると，レベル i の係数は $1/i$ 変化する．

(2) すなわち，W に含まれる各係数の GS は $1/i$ であるため，各係数単体だけを見た場合，$\mathrm{Lap}(\lambda/i)$ の付加により $lambda^{-1}$-差分プライバシーが満たされる．

(3) ただし，データベース中の 1 つのデータの変化は，V における 2 つの値 v_{j1}, v_{j2} にそれぞれ変化をもたらしうる (片方の値が 1 増加し，もう片方の値が 1 減少しうる)．

(4) v_{j1}, v_{j2} の変化は，W においてそれぞれが k 個の詳細係数と 1 個の近似係数の値に影響する．すなわち最大で $2(1+k)$ 個の係数に影響しうる．

(5) したがって，差分プライバシーの直列合成則により，W^* 全体では，$1/\lambda \times 2(1+k) = 2(1+k)/\lambda = \epsilon$-差分プライバシーが満たされることになる．

3.3 Top-down 精緻化

Top-down 精緻化は，$W^* = (cA_k^* \mid cD_k^* \mid cD_{k-1}^* \mid \ldots \mid cD_1^*)$ に含まれるそれぞれの Wavelet 係数を検証し，非負制約を逸脱させるような係数が存在した場合にそれを補正する．この処理をレベル k の係数から Top-down に始め，最終的にレベル 1 までの係数全てについて検証・補正が終了したとき，非負制約を逸脱しないことが保証された $W^+ = (cA_k^+ \mid cD_k^+ \mid cD_{k-1}^+ \mid \ldots \mid cD_1^+)$ を得る．

具体的には，Top-down 精緻化は，cA_i^+ の全要素が負値をとることがないように cD_{i+1}^* の値を精緻化する．なお，以降では説明の便宜上 i は 0 から (k まで) の値を取るとする．このとき，$V^+ = cA_0^+$ となる．

まず，$i = k$ において，cA_k^+ が非負制約を満たすようにする．

$$cA_{k,1}^+ = \max(cA_{k,1}^*, 0). \tag{12}$$

次に，$i < k$ について議論する．このとき，cA_i^+ の各要素 $cA_{i,x}^+$ は，1 レベル上の近似係数 $cA_{i+1,\lceil x/2 \rceil}^+$ と詳細係数 $cD_{i+1,\lceil x/2 \rceil}^+$ とを用いて再帰的に定義される．

$$cA_{i,x}^+ = cA_{i+1,\lceil x/2 \rceil}^+ + g(x) \cdot cD_{i+1,\lceil x/2 \rceil}^+. \tag{13}$$

ここで $g(x)$ は符号関数であり，

$$g(x) = \begin{cases} +1 & (x = 1 \pmod 2), \\ -1 & (x = 0 \pmod 2). \end{cases} \tag{14}$$

の値をとる．

すなわち，式 (13) によれば，

$$|cD_{i+1,\lceil x/2 \rceil}^+| \leq cA_{i+1,\lceil x/2 \rceil}^+ \tag{15}$$

を満たすことができるならば，

$$cA_{i,x}^+ \geq 0 \tag{16}$$

が成立する．$V^+ = cA_0^+$ であるため，これは $cA_{1,k} > 0$ であり，かつ W^+ の全要素について式 (15) が成立するなら，V^+ は非負制約を逸脱しないことを意味する．

すなわち，W^* を入力として，下記のアルゴリズムを実行することにより，Wavelet 逆変換 $\mathcal{H}^{-1}$ の出力 $V^+ = \mathcal{H}^{-1}(W^+)$ が非負制約を逸脱しない W^+ を得ることができる．

[*6] 一般的には，Haar 分解の定義として，Wavelet 係数の信号エネルギー保存のために cA, cD の各要素を $\sqrt{2}$ 倍する定義が用いられることが多いが，本稿では Xiao ら [16], [17] にならってこれを乗じない定義を用いる．

[*7] 詳細な証明は [17] に示されている．

(1) $cA^+_{k,1} = \max\{cA^*_{k,1}, 0\}$
(2) $i = \{k-1, ..., 1\}$ について，降順に下記を実行する．
　(a) $x = \{1, ..., 2^{k-i}\}$ について下記を実行する．
　　(i) 式 (13) により $cA^+_{i,x}$ を算出する．
　　(ii) $cD^+_{i,x} = \begin{cases} -cA^+_{i,x}, & (cD^*_{i,x} < -cA^+_{i,x}) \\ cA^+_{i,x}, & (cD^*_{i,x} > cA^+_{i,x}) \\ cD^*_{i,x}. & (\text{otherwise}) \end{cases}$
(3) $W^+ = (cA^+_k \mid cD^+_k \mid cD^+_{k-1} \mid \ldots \mid cD^+_1)$.

3.4 Wavelet 逆変換

Haar Wavelet 逆変換 $\mathcal{H}-1$ は，精緻化ずみ Wavelet 係数系列 $W+$ から，差分プライバシーを満たし，かつ非負制約を満たす V^+ を得る．

これは，W^+ を入力として，式 (13) を再帰的に cA^+_0 を得るまで適用することにより，$V^+ = cA^+_0$ として得ることができる．

3.5 アルゴリズム 1 の性質

前節までの処理により得られた $V^+ = (v^+_1, v^+_2, \ldots, v^+_n), n = 2^k$ は，以下の性質を満たす．

(1) 適用する Laplace メカニズムにおけるスケールパラメータを λ としたとき，V^+ は $\epsilon = 1/\lambda \times 2(1+\log_2 n) = 2(1+\log_2 n)/\lambda$ の ϵ-差分プライバシーを満たす．
(2) V^+ は非負制約 $\forall i v^+_i \geq 0$ を満たす．
(3) V の各セルの総和が λ に対して極端に小さくない (確率 $\Pr[Lap(\lambda/n) > \Sigma_{v_i \in V}]$ が無視できる) ならば，V^+ の任意のセルもしくはその部分和の平均 $\bar{v^+}_C = \Sigma_{v^+_i \in C} v^+_i / |C| (C \subseteq A)$ の期待値は，原データ系列における平均と等しい．すなわち，原データ系列に対して過大/過小いずれのバイアスも発生することはない．
(4) V^+ を 2^l ごとのブロックに分割したとき，その部分和に含まれるノイズ (原データ系列との差) の分散は，$2\lambda^2 \cdot (1/2^k + \Sigma^k_{i=k-l+1} 1/2^i) = 2\lambda^2/2^{k-l}$ と同程度か，より小さい．すなわち，上記ブロックの部分和のノイズは，ブロック長が長いほど小さくなる．
(5) 計算量は $O(n)$ に従う．すなわち，アルゴリズム 1 では**計算量の問題は解決されない**．

以下，上記の性質のそれぞれが V^+ について成立することを説明する．

性質 1 は，W^* が ϵ-差分プライバシーを満たすこと，および差分プライバシーの事後処理則から導出できる．すなわち，Laplace メカニズムの適用により ϵ-差分プライバシーを満たす W^* を生成した後に，以後の Top-down 精緻化の過程でも Wavelet 逆変換の過程でも (W^* そのものを除き) V に関する知識を用いていない (事後処理則の適用条件を満たす) ため，V^+ も ϵ-差分プライバシーが維持される．

性質 2 は，すでに議論したように，Top-down 精緻化による出力 W^+ が式 (12), (15) を満たすことにより，V^+ は非負制約を逸脱しないことが保証される．

性質 3 について，Wavelet 変換の性質から Wavelet 変換/逆変換の過程でこの性質が保存されることは明らかである．そこで，Laplace メカニズムと Top-down 精緻化について議論する．

Laplace メカニズムの適用において，cA^*_k と cD^*_k にそれぞれ Laplace ノイズが付加されるが，その平均は 0 であるため，まず下記が成立する．

$$E(cA^*_{k,1}) - cA_{k,1} = 0 \tag{17}$$

また，$i \leq k-1$ において，$E(cD^*_{i-1,x}) - cD_{i-1,x} = 0$ が成立するため，式 (13) より，

$$E(cA^*_{i,2x-1}) - cA_{i,2x-1} = E(cA^*_{i,2x}) - cA_{i,2x} = 0 \tag{18}$$

も成立する．したがって，Laplace メカニズムの適用においても平均は保存される．

次に，Top-down 精緻化の過程においては，式 (12) により $cA^+_{k,1}$ に正のバイアスがかかる可能性がある，しかし，その確率は $Pr[\Sigma_{v_i \in V} v_i < Lap(\lambda/n)]$ であり，これは一般的な (人為的に作成されたものではない) 集計データではほぼ無視できると考えられる．また，cD^+_i の精緻化は cA^+_{i-1} の期待値に影響を与えないため，上記の $cA^+_{k,1}$ に正のバイアスがかかる可能性が無視できれば，$cA^+_0 = V^+$ の任意の要素について，期待値は V の対応する要素に等しい．

性質 4 について，Haar Wavelet 変換の性質より，V^+ を 2^l ごとのブロックに分割したときの部分和は，x を何個目のブロックであるかを示すインデックスとしたとき，$2^l \cdot cA^+_{l,x}$ で与えられる．Laplace メカニズムにより付与される Laplace ノイズは互いに独立であるため，$cA^+_{l,x}$ が Top-down 精緻化の影響を受けなかったとき，そのノイズの分散は，$cA^+_k, cD^+_k, cD^+_{k-1}, \ldots, cD^+_{l+1}$ にそれぞれ与えられるノイズの分散の総和になる．

$cA^+_{l,x}$ が Top-down 精緻化の影響を受けるときには，その分散は集計データの分布に依存するため解析的に示すことは難しいが，人口分布などのいわゆる「自然な」集計データ，すなわちロングテイル性を持ち，0 値や小さい値を持つ v_i が連続するようなデータにおいて，Top-down 精緻化は，それらの値が大きく上振れ/下振れすることを防ぐ効果を持つため，条件によってはノイズがより小さくなる傾向を持つと予想される．これについては第 4 章で定量的に評価する．

性質 5 について，アルゴリズム 1 における各処理の計算量はいずれも $O(n)$ であるため，全体の計算量も $O(n)$ となる．以下において，その改善可能性について議論する．

3.6 計算量の改善

アルゴリズム 1 は，非負制約の逸脱，部分和精度の劣化の問題を解決するが，計算量は $O(n)$ になるものであった．これは，Balak らの手法 [1] (n の多項式時間) より優れているが，前述の通り $m \ll n$ となる大規模な集計データでは実用的とは言えない．そこで，アルゴリズム 1 を改良し，等価な出力を得つつ計算量を削減するアルゴリズム 2 を構成する．

アルゴリズム 1 の各処理において，Wavelet 変換/逆 Wavelet 変換については，ナイーブに実装すると $O(n)$ の計算量となるが，これは COO 形式で V および W を表現する工夫により，Wavelet 変換全体の計算量は比較的簡単に $O(km) = O(m \log n)$ に削減することができる．

具体的には，V を COO 形式 (j, v_i) の集合の形式で入力し，$v_i \neq 0$ の (m 個の) セル値についてのみ，

$$cA_{1,\lceil i/2 \rceil} := cA_{1,\lceil i/2 \rceil} + v_i, \tag{19}$$

$$cD_{1,\lceil i/2 \rceil} := cA_{1,\lceil i/2 \rceil} + g(i) \cdot v_i. \tag{20}$$

を計算し (cA, cD はそれぞれ COO 形式などの疎データ形式で保持するものとし，その初期値はいずれも $cA = cD = \{0\}^{n/2}$ とする)，それを再帰的に cA_k, cD_k まで算出すれば，それぞれの非 0 値を持つセル v_i あたり高々 $k = log_2 n$ 回の上記計算で W を得ることができる．非 0 値を持つセルは m 個のため，全体の計算量は $O(m \log n)$ となる．

逆 Wavelet 変換についても，出力である V^+ に含まれる非 0 値のセル数を m^+ とすれば，同様の計算方法で $O(m^+ \cdot \log n)$ の計算量で W^+ から V^+ を得ることができる．

しかし，Laplace メカニズムの適用においては，0 値を持つ w_i に対してもノイズを付与する必要があるため，このような単純なアプローチにより計算量を削減することはできない．

そこで，Top-down 精緻化によりほとんどのノイズは精緻化過程で「捨てられて」しまうことに着目する．Top-down 精緻化の過程で $cD^*_{i,x}$ に精緻化が適用される ($cD^+_{i,x} \neq cD^*_{i,x}$ となる) とき，$cA^+_{i+1,2x-1}$ か $cA^+_{i+1,2x}$ のいずれかは必ず 0 値をとる．そのため，0 値をとったほうの値の部分木に含まれる，2^{i-1} 個の Laplace ノイズが出力値に影響する可能性はなくなる．したがって，Laplace メカニズムの適用と精緻化を，非 0 値をとる $cA^+_{i,x}$ の部分木についてのみ再帰降下で順に実施していくことにより，そのような無駄なノイズを発生させることなしに，差分プライバシーを満たすことができる．

上記の点に鑑み，アルゴリズム 2 では Laplace ノイズの付与と Top-down 精緻化を「同時に」再帰降下で実施する．その手順を以下に示す．

(1) Laplace メカニズムにより，$cA^*_{k,1}, cD^*_{k,1}$ をそれぞれ計算する．

(2) $cA^+_{k,1} = \max\{cA^*_{k,1}, 0\}$

(3) $cD^+_{k,1} = \begin{cases} -cA^+_{k,1}, & (cD^*_{k,1} < -cA^+_{k,1}) \\ cA^+_{k,1}, & (cD^*_{k,1} > cA^+_{k,1}) \\ cD^*_{k,1}. & (\text{otherwise}) \end{cases}$

(4) $i = \{k, ..., 2\}$ について，降順に下記を実行する．

(a) $\forall(x \mid cA^+_{i,x} \neq 0)\}$ について下記を実行する．

(i) $cA^+_{i+1,2x-1} = cA^+_{i,x} + cD^+_{i,x}$.

(ii) $cA^+_{i+1,2x} = cA^+_{i,x} - cD^+_{i,x}$.

(iii)Laplace メカニズムにより，$cD^*_{i+1,2x-1}$ と $cD^*_{i+1,2x}$ とをそれぞれ計算する．

(iv)$cD^+_{i+1,2x-1} =$

$$\begin{cases} -cA^+_{i+1,2x-1}, & (cD^*_{i+1,2x-1} < -cA^+_{i+1,2x-1}) \\ cA^+_{i+1,2x-1}, & (cD^*_{i+1,2x-1} > cA^+_{i+1,2x-1}) \\ cD^*_{i+1,2x-1}. & (\text{otherwise}) \end{cases}$$

(v) $cD^+_{i+1,2x}$ も同様に計算する．

(5) $W^+ = (cA^+_k \mid cD^+_k \mid cD^+_{k-1} \mid \ldots \mid cD^+_1)$.

なお，ここで，精緻化ずみの Wavelet 係数 W^+ 自体に興味がない場合は，最後に W^+ を出力するかわりに，$i = 2$ の計算途中で現れる cA^+_1 と cD^+_1 とを用い，$\forall(x \mid cA^+_{1,x} \neq 0)\}$ について下記を実行することにより，(あらためて Wavelet 逆変換を実施することなしに) $V^+ = cA_0$ を得ることもできる．

$$cA^+_{0,2x-1} = cA^+_{1,x} + cD^+_{1,x}, \tag{21}$$

$$cA^+_{0,2x} = cA^+_{1,x} - cD^+_{1,x}. \tag{22}$$

ここで，この cA^+_1 からの cA^+_0 の導出に要する計算量は $O(m^+)$ である．また，$i \in \{1, \ldots, k-1\}$ において，cA^+_{i+1} から cA^+_i を導出するのに要する計算量がそれを上回ることはない．したがって，本処理全体の計算量は，たかだか $O(km^+) = O(m^+ \log n)$ となる．

4. 部分和精度の評価

提案方式について，部分和精度は Xiao らの Privelet 法と同等か，条件によってはそれ以上に改善されることが期待されるが，その程度を解析的に示すことは難しい．そこで，H22 国勢調査から得られた地域メッシュ人口の一部をサンプルデータとして用い，Privelet 法と提案方式をそれぞれ適用した．

4.1 データセットと評価方法

本評価では，H22 国勢調査に基づく地域メッシュ統計 [19] の 500m メッシュ人口から $n = 2^{19} = 524,288$ 個のデータをランダムに抽出したものを V として用いた．なお，外れ値等による評価の偏りを防ぐため，このランダム抽出を 10 回行ない，10 個のデータセットにより評価を実施した．ただし，いずれの方式でもデータセットの違いによって結

表 1　2km メッシュ (16 セル) の部分和の分散

データセット	Privelet 法 (σ_P^2)	提案方式 (σ_T^2)
データセット 1	6,628.1	2,065.0
データセット 2	6,564.6	2,075.4
データセット 3	6,660.5	2,027.5
(平均)	6,617.8	2,055.8

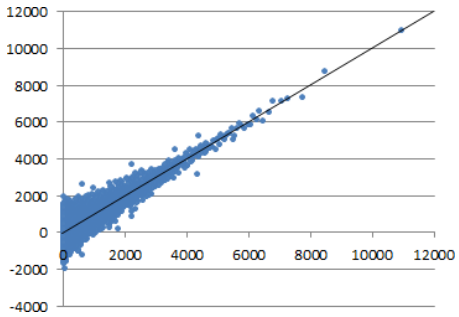

図 1　Privelet 法による 2km メッシュ (4 セル) の部分和の散布図 (データセット 1)

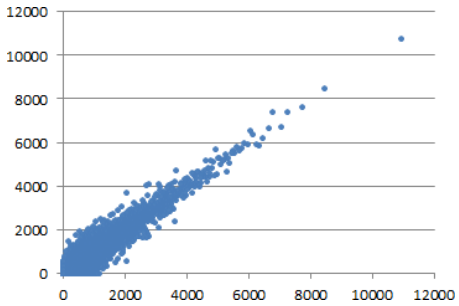

図 2　提案方式による 2km メッシュ (4 セル) の部分和の散布図 (データセット 1)

果に大きな差が見られなかったことから，本稿ではそれらのうち最初の 3 回の試行結果のみを代表例として示す．

これらのデータに，Privelet 法および提案方式をそれぞれ適用した．ここで，$\epsilon = 0.1$ と設定した．すなわち，$\lambda = 2(1 + \log n)/\epsilon = 2 \cdot 20/0.1 = 400$ となる．

両方式の部分和精度の比較を行なうため，まず比較的狭い範囲の部分和として (1) 2km メッシュ相当となる $2^4 = 16$ 個のメッシュ人口の部分和，そして広い範囲の部分和として (2) 16km メッシュ相当となる $2^{10} = 1,024$ 個のメッシュ人口の部分和を，それぞれの方式で算出した．

また，参考として単純な Laplace メカニズムを適用した場合の理論値についても併せて示した．単純な Laplace メカニズムを適用した場合，0.1-差分プライバシーを満たすためには各セルの値に $\mathrm{Lap}(1/0.1) = \mathrm{Lap}(10)$ を付与することになる．この分散は $2 \cdot 10^2 = 200$ となるため，k セルの部分和の分散の期待値は $200k$ となる．

4.2　2km メッシュ (16 セル) の部分和

2km メッシュ相当となる 16 セルのメッシュ人口の部分和について，Privelet 法，提案方式を適用した結果について，その分散を**表 1** に示す．

表 2　16km メッシュ (1,024 セル) の部分和の分散

データセット	Privelet 法 (σ_P^2)	提案方式 (σ_T^2)
データセット 1	102.8	106.7
データセット 2	95.8	112.0
データセット 3	92.3	113.9
(平均)	97.0	110.8

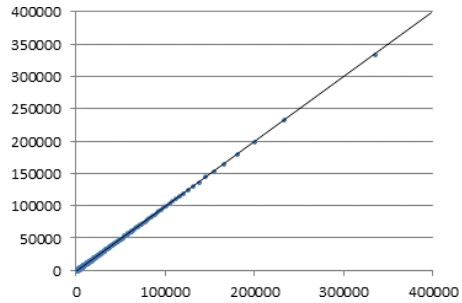

図 3　Privelet 法による 16km メッシュ (1,024 セル) の部分和の散布図 (データセット 1)

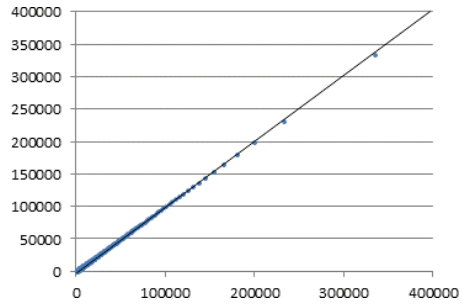

図 4　提案方式による 16km メッシュ (1,024 セル) の部分和の散布図 (データセット 1)

Privelet 法の分散を σ_P^2，提案方式の分散を σ_T^2 としたとき，それぞれの平均は $\bar{\sigma_P^2} = 6,617.8$ および $\bar{\sigma_T^2} = 2,055.8$ となった．なお，単純な Laplace メカニズムを適用した場合の分散の期待値は，$E(\sigma_L^2) = 200 \cdot 16 = 3,200$ となる．

提案方式は，この条件では Privelet 法，単純な Laplace メカニズムのいずれに対しても精度に優れていることが見てとれる．また，Privelet 法などの Matrix メカニズムに属する方式は，少数セルの部分和の精度が良くない傾向があるが，Privelet 法は 16 セルでも単純な Laplace メカニズムより精度に劣る．

また，データセット 1 における散布図を**図 1**，**図 2** に示す (他のデータセットについては図を省略するが，ほぼ同じ傾向を示す)．これらの散布図について，図の横軸はいずれも原データ系列の値であり，縦軸は適用後の値である．Privelet 法が多数の負の人口値を持つセルを出力している (非負制約を逸脱している) のに対し，提案方式は負の値を出力せず，非負制約を充足していることがわかる．

4.3　16km メッシュ (1,024 セル) の部分和

次に，16km メッシュ相当となる 1,024 セルのメッシュ人口の部分和の分散について，同様に**表 2** に結果を示す．

本条件では，$\hat{\sigma_P^2} = 97.0$ および $\hat{\sigma_T^2} = 110.8$ となった．2km メッシュの結果と比べ，両方式とも分散が大幅に小さくなっているが，Privelet 法のほうが若干ではあるが高い精度を持つ．なお，単純な Laplace メカニズムを適用した場合は，$E(\sigma_L^2) = 200 \cdot 1,024 = 204,800$ となり，両方式と比較して大きく劣るようになる．

データセット 1 の結果における散布図についてそれぞれ**図 3**，**図 4** に示す．図の縦軸と横軸は図 1，図 2 と同じである．いずれの方式でもほぼ $y = x$ にデータが沿っており，高い精度を持つことがわかる．

5. まとめ

本稿では，集計データのプライバシーを差分プライバシー基準に基づいて保護する上で，データの統計的正確性と計算効率に着目した手法を提案した．差分プライバシーは，数学的な安全性保証が与えられているという優れた特徴を持つ一方で，大規模な集計データの公開におけるプライバシー保護に適用するためには，(1) 非負制約の逸脱，(2) 部分和精度の劣化，(3) 計算量の増大，という 3 つの課題を解決する方法があることを示した．

上記問題を解決するために，Wavelet 変換を用いて部分和精度の問題を解決する手法である Privelet 法を応用し，Top-down 精緻化と呼ぶ手順を導入する方式を提案した．まず，アルゴリズム 1 として部分和精度の問題に加えて非負制約の逸脱を解決する手法を示し，次にアルゴリズム 2 として，アルゴリズム 1 と等価な出力をしつつ，さらに計算量の増大の問題を併せて解決する方式を示した．単純な Laplace メカニズムや Privelet 法の計算量が $O(n)$ であるのに対し，提案方式の計算量は $O(m^+ \log n)$ である．ここで，n はセル空間全体の大きさ，m^+ は非 0 値を持つ出力セルの数である．したがって，特に $m^+ \ll n$ となるような疎な集計データにおいて，提案方式は計算効率に優れる．

また，国勢調査によるメッシュ人口に基づくサンプルデータセットへの適用を通じた比較評価により，提案方式を適用した集計データは Privelet 法とほぼ同等以上の部分和精度を持ち，特に少数セルの部分和について提案方式は Privelet 法と比べて精度に優れることが明らかとなった．

参考文献

[1] Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., Talwar, K. and Berkeley, U. C.: Privacy, accuracy, and consistency too: a holistic solution to contingency table release, *Proc. 26th ACM SIGMOD-SIGACT-SIGART symp. Principles of database systems - PODS '07*, ACM Press, pp. 273–282 (2007).

[2] Cormode, G., Procopiuc, M., Srivastava, D. and Tran, T.: Differentially Private Publication of Sparse Data, *Proc. intl. conf. Database Theory (ICDT2012)* (2012).

[3] Dwork, C.: Differential Privacy, *Proc. 33rd intl. conf. Automata, Languages and Programming - Volume Part II* (Bugliesi, M., Preneel, B., Sassone, V. and Wegener, I., eds.), Lecture Notes in Computer Science, Vol. 4052, Springer, pp. 1–12 (2006).

[4] Dwork, C.: An ad omnia approach to defining and achieving private data analysis, *Proc. 1st ACM SIGKDD intl. conf. Privacy, security, and trust in KDD*, Springer-Verlag, pp. 1–13 (2007).

[5] Dwork, C.: Differential privacy: a survey of results, *Proc. 5th intl. conf. Theory and applications of models of computation*, Springer-Verlag, pp. 1–19 (2008).

[6] Fung, B. C. M., Wang, K., Chen, R. and Yu, P. S.: Privacy-preserving data publishing, *ACM Computing Surveys*, Vol. 42, No. 4, pp. 1–53 (2010).

[7] Ghosh, A., Roughgarden, T. and Sundararajan, M.: Universally Utility-maximizing Privacy Mechanisms, *SIAM J. Computing*, Vol. 41, No. 6, pp. 1673–1693 (2012).

[8] Hay, M., Rastogi, V., Miklau, G. and Suciu, D.: Boosting the accuracy of differentially private histograms through consistency, *Proc. VLDB Endowment*, Vol. 3, No. 1-2, VLDB Endowment, pp. 1021–1032 (2010).

[9] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Nordholt, E. S., Seri, G. and de Wolf, P.-P.: *Handbook on statistical disclosure control*, Statistics Netherlands (2010).

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

[11] Hundepool, A. and Wolf, P.-p. D.: *Statistics Netherlands Methods Series : Statistical disclosure control*, No. July (2011).

[12] Li, C., Hay, M., Rastogi, V., Miklau, G. and McGregor, A.: Optimizing linear counting queries under differential privacy, *Proc. 29th ACM SIGMOD-SIGACT-SIGART symp. Principles of database systems of data (PODS '10)*, New York, New York, USA, ACM Press, pp. 123–134 (2010).

[13] Makita, N., Kimura, M., Terada, M., Kobayashi, M. and Oyabu, Y.: Can mobile phone network data be used to estimate small area population ? A comparison from Japan, *Statistical J. IAOS*, Vol. 29, No. 3, pp. 223–232 (2013).

[14] Stanimirovic, I. P. and Tasic, M. B.: Performance comparison of storage formats for sparse matrices, *Ser. Mathematics and Informatics*, Vol. 24, No. 1, pp. 39–51 (2009).

[15] Sweeney, L.: k-anonimity: a model for protecting privacy, *Intl. J. Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10, No. 05, pp. 557–570 (2002).

[16] Xiao, X., Wang, G. and Gehrke, J.: Differential privacy via wavelet transforms, *Proc. 26th intl. conf. Data Engineering (ICDE 2010)*, IEEE, pp. 225–236 (2010).

[17] Xiao, X., Wang, G., Gehrke, J. and Jefferson, T.: Differential Privacy via Wavelet Transforms, *IEEE Trans. Knowledge and Data Engineering*, Vol. 23, No. 8, pp. 1200–1214 (2011).

[18] 寺田雅之：モバイル空間統計の試み： 携帯電話ネットワークによる人口変動の推計とその応用，統計， Vol. 63, No. 9, pp. 29–36 (2012).

[19] 総務省 統計局：地域メッシュ統計の特質・沿革.

[20] 統計センター：統計データ開示抑制に関する用語集 (改訂版), 製表関連国際用語集 No.2 (2005).

[21] 瀧　敦弘：集計表におけるセル秘匿問題とその研究動向，統計数理， Vol. 51, No. 2, pp. 337–350 (2003).