情報処理学会論文誌 Vol.56 No.9 1801–1816 (Sep. 2015)

推薦論文

大規模集計データへの差分プライバシの適用

寺田 雅之[1,a)] 鈴木 亮平[1] 山口 高康[1] 本郷 節之[2]

受付日 2014年12月7日, 採録日 2015年6月5日

概要：データの有効な活用による社会・産業の発展への期待が高まる中，プライバシを保護したうえでデータを利用するための技術が注目を集めている．そのなかで，Dwork らによる差分プライバシは，その高い安全性から大きな期待が寄せられているが，特に大規模データへの適用においてデータの有用性や処理効率などの観点から実用上の課題を持つ．本稿では，地理空間データなどの大規模な集計データに差分プライバシを適用するうえでの課題を示すとともに，これを解決する手法について安全性証明と実データに基づく評価を与える．本手法は，集計データの非負制約に着目し，その逸脱を Wavelet 空間において補正する過程を導入することにより有用性と処理効率の向上を実現するとともに，局所性保存写像（locality preserving mapping）の一種である Morton 順序写像を用いることにより，地理空間データなどの多次元集計データへの適用時の精度劣化を抑制することを特徴とする．

キーワード：プライバシ保護，差分プライバシ，ウェーブレット変換，非負制約

On Publishing Large Tabular Data with Differential Privacy

Masayuki Terada[1,a)] Ryohei Suzuki[1] Takayasu Yamaguchi[1] Sadayuki Hongo[2]

Received: December 7, 2014, Accepted: June 5, 2015

Abstract: Big data become widely expected to enhance the quality and efficiency of our daily life, and methods to prevent privacy information included in the data from being disclosed by data utilization become attracting wide attention therewith. Differential privacy is a promising paradigms to achieve proven privacy, but previous methods to assure the differential privacy have several drawbacks on data utility and scalability in practice, in particular when applied to publishing large and sparse tabular data such as geospatial data. This paper proposes a novel differentially private method that simultaneously solves these problems, and demonstrates its evaluation results. The proposed method introduces a process to correct for the non-negative restriction of the output data by modifying the wavelet coefficients of the perturbed data, and this correction process enables the proposed method to efficiently process large sparse data in terms of scalability and accuracy. In addition, the proposed method effectively suppresses the amount of noise required to process multi-dimensional data by reducing its dimensionality using a locality-preserving mapping method called Morton order mapping.

Keywords: privacy-preserving data utilization, differential privacy, wavelet transform, non-negativity constraints

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を差分プライバシ基準に基づいて保護するうえで，データの統計的な正確性と計算効率を改善した，新たなプライバシ保護手法を提案する．なお，本稿において集計データと

[1] 株式会社 NTT ドコモ先進技術研究所
Research Laboratories, NTT DOCOMO, Inc., Yokosuka, Kanagawa 239–8536, Japan
[2] 北海道科学大学工学部
Faculty of Engineering, Hokkaido University of Science, Sapporo, Hokkaido 006–8585, Japan
[a)] teradam@nttdocomo.com

本稿の内容は 2014 年 7 月のマルチメディア，分散，協調とモバイル（DICOMO2014）シンポジウムにて報告され，コンピュータセキュリティ研究会主査により情報処理学会論文誌ジャーナルへの掲載が推薦された論文である．

は，元のデータベースに含まれる個々のデータ群（個票，あるいは生データ（raw data）とも呼ばれる）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）の集まりである．本稿では特に，人口分布や交通量分布などの，大規模な地理空間に関する集計データ（地理空間データ）を主な適用対象として議論する．

集計データの代表例として，国勢調査をはじめとした公的統計があげられる．これらの公的統計は，主に国民や事業者に対する調査票を用いたアンケート結果を集計したものであり，公共政策の立案や企業の経営計画の策定，およびそれらの実施効果の検証など，公共分野・産業分野における社会活動に広く活用されてきた．さらに，近年における計算機の処理能力や記憶容量の向上にともない，いわゆるビッグデータに基づく大規模かつ高次元な集計データの作成が現実になりつつある．その普及と活用は，「証拠に基づいた政策・方針策定（Evidence-based Policy Making）[2]」，すなわち客観的な事実に基づく社会活動の最適化を実現する鍵として期待される [15]．

ただし，これらの集計データは個々の人々の活動を集積したものであることに留意が必要である．すなわち，集計データを公開することにより個人のプライバシが侵されるようなことはあってはならない．

集計データに対するプライバシ保護の必要性やその方法は，統計分野において古くから議論されてきた．これらは統計的開示制御（statistical disclosure control, SDC）と総称される [10], [11], [23]．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，具体的な SDC の手法としてあげられる [24]．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・経済統計を公開するにあたり，統計から個人のプライバシ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長年の実績を持つ．

その一方で，近年に注目を集めているビッグデータに基づく大規模集計データは，従来の SDC の手法では十分にデータの安全性や有用性を確保できない懸念がある．たとえば，集計データの規模（セル数）が大きくなると，データがロングテイル性を持つ（大多数のセルが小さいセル値をとる）ようになる．しかし，そのようなデータに対してセル秘匿基準などを機械的に適用すると集計データのロングテイル部分の情報が完全に失われ，データの有用性が大きく損なわれることになる．

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，プライバシを保護しつつ有用なデータを公開するのための様々な基準や手法が提案されている．これらの技術は，プライバシ保護データ公開（privacy-preserving data publishing, PPDP）技術などと呼ばれる [7]．PPDP 技術の例としては，k-匿名性（k-anonymity）基準 [17] や，その変形に基づく手法 [13], [18] などが代表的である．

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うえでは，扱うデータの性質や応用ごとに，「どのプライバシ保護基準に基づいて，どの手法によりプライバシを保護するべきか」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いいにくい．

そこで，本研究では，集計データのプライバシを保護するための基準として，Dwork により 2006 年に提案された差分プライバシ（differential privacy）[4] に着目する．これは，データベースから集計データを作成した際に，「ある個人がデータベースに含まれていたか否かを，集計データから判別することは困難である」ことを安全性の根拠とするプライバシ保護基準である．差分プライバシは，k-匿名性基準などの他のプライバシ保護基準と異なり，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，平均 0 の Laplace 分布に従う独立した乱数（Laplace ノイズ）を付与する手法があげられる．この手法は Laplace メカニズムと呼ばれる．

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ともなう．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ともなうこと（**非負制約の逸脱**），複数セルの部分和をと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 3 点の課題があげられる．

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], [3], [9], [12], [19], [20] が，いずれの方式においても，前述の 3 点の課題を同時に解決することはできず[*1]，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．

本稿では，これらの問題を解決するため，(1) Wavelet 変換と Top-down 精緻化と呼ぶ手法を組み合わせた，差分プライバシ基準を満たすプライバシ保護方式を提案するとともに，(2) 局所性保存写像（locality preserving mapping）の 1 種である Morton 順序写像を用いることにより，地理空間データなどの多次元データへの適用時における精度を

[*1] 3.3 節において詳細に検討する．

向上させる適用方式を与える．

2. 準備

本章では，議論の準備として，本稿で対象とする集計データの定義を与えるとともに，差分プライバシ [4], [6] の定義と，差分プライバシを実現するための代表的な手段として知られている Laplace メカニズムについて説明する．

2.1 集計データ

集計データとは，1 以上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（もしくは属性の組合せ）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 表[*2]などの各種の公的統計，および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 [14], [21] などが相当する．以下に集計データの定義を与える．

定義 1. l 個のレコードから構成されるデータベース $D = \{x_1, x_2, \ldots, x_l\}$ において，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 $(x_i \in A)$ とする．また，D において，A の部分空間 $C_j (\subseteq A)$ に属するレコードの個数を $Count(D, C_j) = |\{x \in D \mid x \in C_j\}|$ とする．これを計数問合せ（count query）結果と呼ぶ．このとき，任意の C_j からなる列である集計条件 $C = (C_1, C_2, \ldots, C_n)$ に対して，集計データ $V(D, C)$ は対応する計数問合せ結果 $Count(D, C_j)$ の列として与えられる．

$$V(D,C) = (v_1, v_2, \ldots, v_n), \quad v_j = Count(D, C_j). \quad (1)$$

集計データ $V(D, C)$ の作成にあたり，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集計条件 C として用いられる．このときの集計データは分割表（contingency table）と呼ばれる．たとえば，$A = A_1 \times A_2$ $(d = 2)$ において，$A_1 = \{\text{男性}, \text{女性}\}, A_2 = \{\text{年齢} \mid \text{年齢} \in \mathbb{N}\}$ とする．20 歳を境として A_2 を 2 つの部分空間，未成年 $= \{\text{年齢} \in A_2 \mid \text{年齢} < 20\}$ と 成人 $= \{\text{年齢} \in A_2 \mid \text{年齢} \geq 20\}$ に分け，$C = (\{\text{男性}\} \times \{\text{未成年}\}, \{\text{男性}\} \times \{\text{成人}\}, \{\text{女性}\} \times \{\text{未成年}\}, \{\text{女性}\} \times \{\text{成人}\})$ を与えて作成した集計データ $V(D, C)$ 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（cell）もしくはセル値と呼ぶ．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し，簡単のため $V(D, C)$ を単に V と表記する．

実世界に基づく大規模な集計データは，0 値のセルを多数含む，疎（sparse）なデータになることが多い．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，500 m メッシュ単位に，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 500 m メッシュの数は約 150 万 [22]，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 $1.5 \cdot 10^6 \times 20 \times 2 \times 2 \cdot 10^3 = 1.2 \cdot 10^{11}$（約 1,200 億）となる．これは日本の総人口である約 $1.2 \cdot 10^8$ 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きわめて効率が悪い．

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（もしくは j のリストと v_j のリストのペア）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 [16]．集計データを COO 形式で表現したとき，そのデータ量は $O(m)$ となる．

2.2 差分プライバシ

差分プライバシ [4], [6] は，識別不能性に基づくプライバシ基準の一種である．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合せ結果が，その個人のデータを含まないデータベースへの問合せ結果と区別できないなら，その問合せは安全である（個人に関するプライバシを開示しない）」という考え方によりプライバシを規定する．

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，データベース問合せ $\mathcal{K}(\cdot)$ を考える．なお，$\mathcal{K}(\cdot)$ は（出力に乱数ノイズを加えるなど）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 への問合せ $\mathcal{K}(D)$ と，ある個人 i に関するレコード $x_i (\in D)$ を D からとり除いたデータベース $D' (= D \setminus x_i)$ への問合せ結果 $\mathcal{K}(D')$ が区別できないなら，$\mathcal{K}(D)$ 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 のプライバシは保護される．

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

定義 2. 任意の隣接した（互いにたかだか 1 レコードしか異ならない）データベース 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 $(S \subseteq \mathcal{R})$．

$$\Pr[\mathcal{K}(D_1) \in S] \leq e^{\epsilon} \cdot \Pr[\mathcal{K}(D_2) \in S]. \quad (2)$$

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

差分プライバシの特徴として，その安全性定義がデータ

[*2] 「Origin-Destination 表（matrix）」の略．起終点表とも呼ばれる．

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あげられる．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，差分プライバシが持つこの性質について，差分プライバシは（“ad hoc” ではなく）“ad omnia” なプライバシ保証を与える，としている [5]．

2.3 Laplace メカニズム

差分プライバシを実現するためには，定義 2 を満たすメカニズム $\mathcal{K}$ が必要となる．その代表的なものとして Laplace メカニズム[*3]があげられる．

Laplace メカニズムは，0 を平均とした Laplace 分布に従う乱数である Laplace ノイズを問合せ結果に加算することにより実現できる．Laplace 分布の確率密度 $\ell(x)$ は，平均 μ とスケール λ を用いて下式で与えられる．

$$\ell(x;\mu,\lambda) = \frac{1}{2\lambda} e^{(-|x-\mu|/\lambda)}. \tag{3}$$

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 個の互いに独立した $\mathrm{Lap}(\lambda)$ からなるベクトル列を $\mathrm{Lap}(\lambda)^k$ と記載する．

Laplace メカニズムにおける Laplace ノイズのスケール λ は，定義 2 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 を問合せ $f : \mathcal{D} \to \mathbb{R}^d$ 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．

$$\mathcal{K}_f(X) = f(X) + \mathrm{Lap}(GS_f/\epsilon)^d, \tag{4}$$

$$GS_f = \max_{D_1,D_2} ||f(D_1) - f(D_2)||_1. \tag{5}$$

ここで，$D_1(\in \mathcal{D})$ および $D_2(\in \mathcal{D})$ は任意の隣接したデータベース（定義 2 参照）のペアである．

3. 従来技術と課題

3.1 集計データへの Laplace メカニズムの適用

理論的には，Laplace メカニズムを用いることにより差分プライバシが保証された集計データを簡単に作成することができる．

前述のとおり，集計データ V は計数問合せ結果 $v_j = Count(D, C_j)$ の集合である．ここで，V は分割表，すなわち C を構成する各部分集合は互いに素であるとする $(\forall i, j \mid i \neq j, C_i \cap C_j = \emptyset)$．

計数問合せの感度 GS_{count} は 1 であることが知られている [4], [6] ため，集計データのセル v_i にスケール $1/\epsilon$ の Laplace ノイズを加えた値，

$$v_i^* = v_i + \mathrm{Lap}(1/\epsilon) \tag{6}$$

は ϵ-差分プライバシを満たす．

また，v_i は互いに素な集合から生成されていることから，差分プライバシの並列合成則により，v_i^* の集合 $V^* = \{v_1^*, v_2^*, \ldots, v_n^*\}$ も ϵ-差分プライバシを満たす．すなわち，

$$V^* = V + \mathrm{Lap}(1/\epsilon)^n \tag{7}$$

により ϵ-差分プライバシが保証された集計データ V^* を得ることができる[*4]．

3.2 Laplace ノイズの適用における実用上の課題

しかし，実世界から得られた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 点があげられる．

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 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．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．この問題に対する直接的な対処として，Laplace メカニズムの適用後に負値のセルを 0 値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（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，部分和精度の劣化である．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．たとえば，500 m メッシュを単位とした人口分布から，2×2 個のメッシュ人口を合算して 1 km 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．たとえば，上記の例では，1 km メッシュには 4 個分のノイズが，2 km メッシュであれば 16 個分のノイズが重畳することになる．すなわち，部分和の対象範囲が広ければ

[*3] 計数問合せに対しては，幾何分布に従う乱数を用いた幾何メカニズム（geometric mechanism）のほうが適しているとされる（文献 [3], [8]）．しかし，ϵ が小さいときには両者の出力はほぼ変わらないことから，本稿では代表して Laplace メカニズムを扱う．

[*4] $\mathrm{Lap}(1/\epsilon)^k$ は，k 個の独立した $\mathrm{Lap}(1/\epsilon)$ からなるベクトル列を表す（2.3 節参照）．

広いほどノイズによる真値からの偏差が大きくなり，その有用性が大きく劣化する．

第3の問題は，計算量の増大である．前述のとおり，実世界から得られた大規模な集計データは疎であることが多い．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．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う必要がある[*5]．すなわち，$O(m)$ 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．

3.3 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，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精度を改善し，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．具体的には，集計データに離散 Fourier 変換を適用したうえで，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．その後に，（元の集計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，差分プライバシを保ちつつ非負制約の逸脱を解消する．しかし，計算量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 [19], [20] は，部分和精度の改善に離散 Wavelet 変換を用いる手法を提案している．この手法は “Privelet” と名付けられている．Barak らが Fourier 変換を導入したのに対し，Privelet では Haar 基底に基づく離散 Wavelet 変換（HWT）を導入し，その Wavelet 係数に対して Laplace メカニズムを適用する．HWT 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脱を解決する手段は与えられていない．また，その計算量は $O(n)$ であり，Barak らの手法に対して改善はあるものの，単純な Laplace メカニズムと同等であり，計算量の増大の問題を解決しない．

部分和の精度を改善する他のアプローチとして，Hay ら [9] による階層的な部分和に基づく方式があげられる．この方式は，入力データを 2^i（$0 \le i \le k$）個のセルから成るブロックに分割した際の部分和をすべて計算し（2^{k+1} 個の部分和が得られる），それらに対して Laplace ノイズを付与する．ノイズ付与により，部分和の間で値の不整合が生じうる[*6]ため，部分和間での整合性の確保を制約条件として，ノイズ付与後の（不整合が生じている）部分和との L_2-ノルムが最小となる系列を計算し，これを最終的な部分和とする．しかし，Privelet 法と同様に，この方式では非負制約の逸脱が解決されず（今後の課題としている），また計算量の問題も解決されない．

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 変換の適用後の値に Laplace ノイズを加算し，その結果に逆変換をかけることにより差分プライバシ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 ノイズの加算 → 逆線形変換，と一般化できる．上記の線形変換に対応する行列を $A = \{a_{ij}\}$ とすると，差分プライバシを満たす集計データ V^* は以下の式で与えられる．

$$\begin{aligned} V^* &= A^{-1}\left(AV + \mathrm{Lap}(\Delta A/\epsilon)^n\right) \\ &= V + A^{-1}\mathrm{Lap}(\Delta A/\epsilon)^n. \end{aligned} \tag{8}$$

ここで，ΔA は A のクエリ行列感度（query matrix sensitivity）と呼ばれ，$\Delta A = \max_j \sum_{i=1}^{n} |a_{ij}|$，すなわち A の各列の L_1-ノルムの最大値である．

このアプローチにより差分プライバシを満たす手法は，Li ら [12] により Matrix メカニズムと名付けられている[*7]．Matrix メカニズムは，Barak らや Xiao らの方式を包含する汎用的なものであるが，その柔軟性の代償として計算量のさらなる増大を招く．たとえば，高速 Fourier 変換の計算量は $O(n \log n)$，Wavelet 変換の計算量は $O(n)$ であるが，Matrix メカニズムにおける行列計算の計算量は一般的に $O(n^2)$ となる（$|V| = |V^*| = n$ の場合）．

Cormode ら [3] は，「ある閾値」を超える値を持つセルの値だけがあればよいという応用を前提としたうえで，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．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

[*5] 厳密には，構造的ゼロ（structual zero），すなわち「0以外の値をとることが論理的にありえない」セルに対してはノイズの付与は不要であるが，それ以外の0値をとるセルに対してはノイズを付与しなくてはならない．

[*6] たとえば，4個ずつのブロックの部分和は，隣接する2個ずつのブロックの部分和を足したものになるはずであるが，ノイズ付与によりその関係が崩れる．

[*7] より正確には，文献 [12] における Matrix メカニズムの定義は，式 (8) に対し，さらに最終出力を抽出するための行列（workload matrix）B を乗じた形式（$BV + BA^{-1}\mathrm{Lap}(\Delta A/\epsilon)^n$）として与えられている．

表 1 従来方式による課題の解決状況
Table 1 Previous works.

方式	非負制約	部分和精度	計算量
Barak ら [1]	×	○	$O(n \log n) \sim$
Xiao ら [19], [20]	×	○	$O(n)$
Hay ら [9]	×	○	$O(n) \sim$
Li ら [12]	×	○	$O(n^2)$
Cormode ら [3]	○	×	(閾値による)

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 らの手法では「閾値」を下回る値を持つセルを無視してしまうことから，部分和にノイズだけでなく閾値に応じたバイアスを生じさせる．Zipf の法則などが示唆するように，実世界から得られた集計データは一般にロングテイル性を持つ（大多数のセルが小さいセル値をとる）ことが多いため，特に広範囲の部分和においてその影響は無視できなくなる．そのため，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が困難となる．

表 1 に，本節で示した従来方式が，前節にあげた 3 つの課題（非負制約の逸脱，部分和精度の劣化，計算量の増大）のいずれを解決しているかについてまとめる．

4. 提案方式

本章では，前章であげた 3 点の課題を解決する新たなプライバシ保護方式を提案する．なお，本章では一次元の集計データを対象として議論するが，6 章で示す Morton 順序写像を用いた次元変換により，地理空間データなどの多次元集計データに適用することも可能である．

提案方式は，集計データに Haar Wavelet 変換（HWT）を適用した上で，その 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なすこともできる．Privelet 法では，W へのノイズ付与後に，単に逆 HWT $\mathcal{H}^{-1}$ を適用することにより出力となる集計データを得る．しかし，この方法で得られた集計データは，ノイズの影響により非負制約を逸脱する．さらに，計算量および出力データ量のいずれも $O(n)$ となるため計算量の問題も解決しない．

これらの問題を解決するために，提案方式は，Top-down 精緻化と呼ぶ処理過程を導入する．Top-down 精緻化 $\mathcal{T}$ は，HWT により得られた Wavelet 係数に対し，Laplace ノイズの付与と非負精緻化，および逆 Haar Wavelet 変換を適用することにより，差分プライバシを満たした出力 V^+ を出力する．

すなわち，提案方式は，元の集計データ V に対し，HWT $\mathcal{H}$，Top-down 精緻化 $\mathcal{T}$ を以下の二段階の手順で適用することにより，ϵ-差分プライバシを満たす集計データ V^+ を得る．

$$W = \mathcal{H}(V), \tag{9}$$

$$V^+ = \mathcal{T}(W). \tag{10}$$

提案方式は，前章であげた 3 つの課題をいずれも解決したうえで差分プライバシを満たす．これらの安全性と有用性に関する提案方式の性質については 4 章で詳しく議論する．

以下，HWT $\mathcal{H}$ について簡単に説明したのち，Top-down 精緻化 $\mathcal{T}$ の構成法として $\mathcal{T}_1$（直列構成）と $\mathcal{T}_2$（並列構成）の 2 種類を与える．$\mathcal{T}_1$ はノイズ付与，非負精緻化，逆 HWT を順に適用する構成である．簡潔な構成であるが，計算量の増大を部分的にしか解決しない．その一方，$\mathcal{T}_2$ は各 Wavelet 係数に対して，上位の係数から再帰的に上記の各処理を「まとめて」適用する．構成が若干複雑であるが，（最終的な出力 V^+ に寄与しない）無駄な処理を「枝刈り」することにより，計算量の増大を解決することができる．

これらの 2 種類の構成法は計算手順と計算量が異なるが，出力に関して等価である．すなわち以下の補題が成立する．この証明は付録 A.2 で与える．

補題 1. $n(=2^k)$ 要素からなる任意の実数ベクトル $V(\in \mathbb{R}^n)$ を $\mathcal{T}_1$ と $\mathcal{T}_2$ に対してそれぞれ与えたとき，その出力分布は等しい．すなわち，$\mathcal{T}_1$，$\mathcal{T}_2$ の出力空間 $\mathbb{R}^n$ における任意の部分空間 $S(\subseteq \mathbb{R}^n)$ について下記が成立する．

$$\forall V \forall S, \Pr[\mathcal{T}_1(\mathcal{H}(V)) \in S] = \Pr[\mathcal{T}_2(\mathcal{H}(V)) \in S]. \tag{11}$$

4.1 Haar Wavelet 変換

Haar Wavelet 変換（HWT）$\mathcal{H} : \mathbb{R}^n \to \mathbb{R}^n$ 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 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．

$$\mathcal{H}_1(Y) = cA \mid cD, \tag{12}$$

$$cA = \left(\frac{y_1 + y_2}{2}, \frac{y_3 + y_4}{2}, \ldots, \frac{y_{2^l-1} + y_{2^l}}{2}\right), \tag{13}$$

$$cD = \left(\frac{y_1 - y_2}{2}, \frac{y_3 - y_4}{2}, \ldots, \frac{y_{2^l-1} - y_{2^l}}{2}\right). \tag{14}$$

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である．cA を近似係数ベクトル，cD を詳細係数ベクトルと呼ぶ．

Haar 分解により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．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 1 個の近似係数ベクトルが得られる．これらの連接が HWT の出力 W となる．すなわち，$W = \mathcal{H}(V)$ は Haar 分解 $\mathcal{H}_1$ を用いて以下の式により定義される．

$$cA_0 = V, \tag{15}$$

$$cA_i \mid cD_i = \mathcal{H}_1(cA_{i-1}) \quad (i \in \{1..k\}), \tag{16}$$

$$W = (cA_k \mid cD_k \mid cD_{k-1} \mid \ldots \mid cD_1). \tag{17}$$

式 (16) の手順は，そのまま実装すると $O(n)$ の計算量となる．しかし，非 0 値を持つセルのみに着目するアルゴリズムを用いることにより，これを $O(km) = O(m \log n)$ に削減することができる．ここで，m は V に含まれる非 0 値の個数である．その具体的な構成法を付録 A.1 に示す．

$\mathcal{H}$ 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．$W = (cA_k \mid cD_k \mid cD_{k-1} \mid \ldots \mid cD_1)$ としたとき，以下の式を再帰的に適用することにより $V = \mathcal{H}^{-1}(W)$ を得ることができる．

$$cA_{i-1} = \mathcal{H}_1^{-1}(cA_i, cD_i), \quad (i \in \{k..1\}).$$

ここで，$\mathcal{H}_1^{-1}$ は $\mathcal{H}_1$ の逆変換であり，$|X| = |Y| = 2^l$ として以下の式により与えられる．

$$\begin{aligned}&\mathcal{H}_1^{-1}(X, Y)\\ &\quad = (x_1 + y_1, x_1 - y_1, x_2 + y_2, x_2 - y_2, \ldots,\\ &\qquad x_{2^l} + y_{2^l}, x_{2^l} - y_{2^l}).\end{aligned} \tag{18}$$

4.2 Top-down 精緻化

HWT により得られた Wavelet 係数系列 $W = \mathcal{H}(V)$ から，ϵ-差分プライバシを満たし，かつ非負制約を充足する集計データ V^+ を得る．説明のため，まず $O(n)$ の計算量を持つ構成法 $\mathcal{T}_1$（直列構成）を示し，その後に計算量を $O(m^+ \log n)$ に効率化した，$\mathcal{T}_1$ と出力等価な構成法 $\mathcal{T}_2$（並列構成）を示す．ここで，m^+ は出力 V^+ における非 0 値の個数である．

4.2.1 $\mathcal{T}_1$ の構成法

HWT において，近似係数ベクトル cA_i $(0 \le i \le k)$ の各要素は，それぞれ V における 2^i 個のセル値の集合の平均であることに着目する．ある集合の平均が負値をとるとき，少なくとも 1 つ以上の要素の値は必ず負値をとることから，cA_i の x 番目（$1 \le x \le 2^{k-i}$）の要素を $cA_{i,x}$ としたとき，$cA_{i,x} < 0$ となることは 1 つ以上の V の要素が負値をとることを意味する．すなわち，任意の $cA_{i,x}$ について $cA_{i,x} \ge 0$ が成立することは，V が非負制約を満たすための必要条件である．また，$cA_0 = V$ より，これは明らかに十分条件でもある．すなわち，

$$\forall v_j \in V, v_j \ge 0 \iff \forall i \forall x, cA_{i,x} \ge 0. \tag{19}$$

次に，$cA_{i,x} \ge 0$ が成立する条件を考える．HWT の定義より，$1 \le i \le k$ に対して，

$$cA_{i,x} + cD_{i,x} = cA_{i-1,2x-1}, \tag{20}$$

$$cA_{i,x} - cD_{i,x} = cA_{i-1,2x} \tag{21}$$

であることから，$cA_{i,x} \ge 0$ であるためには，

$$cA_{i+1,\lceil x/2 \rceil} \ge |cD_{i+1,\lceil x/2 \rceil}| \tag{22}$$

が成立すればよいことが分かる．

V の全要素が非負であるとき，$W = \mathcal{H}(V)$ は上式を満たす．しかし，W の各要素に Laplace ノイズを付加した系列 W^* では明らかにこの性質が維持されることは保証されない．これが，W^* にそのまま逆 HWT $\mathcal{H}^{-1}$ を適用する Privelet 法の出力が非負制約を逸脱する理由となる．逆にいえば，式 (22) を満たすよう W^* を精緻化することができれば，その逆 HWT 後の出力は非負制約を満たすことになる．

この性質を利用し，アルゴリズム $\mathcal{T}_1$ は W へのノイズ付加後に式 (22) を満たすよう各 Wavelet 係数を修正し，最後に逆 HWT を適用することにより出力 V^+ を得る．具体的には，W を入力として，以下の三段階の手順を順に適用する．なお，下記手順の説明において，$g(\cdot)$ は整数値を入力として以下の値をとる符号関数である．

$$g(x) = \begin{cases} +1 & (x = 1 \pmod 2) \\ -1 & (x = 0 \pmod 2). \end{cases} \tag{23}$$

(1) W の各要素に Laplace ノイズを付加することにより，ϵ-差分プライバシを満たす係数系列 W^* を得る．これは，$\lambda = (1 + \log_2 n)/\epsilon$ として，下式で導出される．

$$cA_k^* = cA_k + \mathrm{Lap}(\lambda/2^k), \tag{24}$$

$$cD_i^* = cD_i + \mathrm{Lap}(\lambda/2^i)^{2^{k-i}} \quad (i \in \{1..k\}), \tag{25}$$

$$W^* = (cA_k^* \mid cD_k^* \mid cD_{k-1}^* \mid \ldots \mid cD_1^*). \tag{26}$$

(2) W^* において，各 Wavelet 係数に対応する部分和が非負制約を逸脱しないよう係数値を補正（非負精緻化）し，精緻化済み Wavelet 係数系列 W^+ を得る．

$$cA_{i,x}^+ = \begin{cases} \max(cA_{i,x}^*, 0) & (i = k) \\ cA_{i+1,\lceil x/2 \rceil}^+ + g(x) \cdot cD_{i+1,\lceil x/2 \rceil}^+ & (\text{otherwise}), \end{cases} \tag{27}$$

$$cD_{i,x}^+ = \begin{cases} -cA_{i,x}^+ & (cD_{i,x}^* < -cA_{i,x}^+) \\ cA_{i,x}^+ & (cD_{i,x}^* > cA_{i,x}^+) \\ cD_{i,x}^* & (\text{otherwise}), \end{cases} \tag{28}$$

$$W^+ = (cA_k^+ \mid cD_k^+ \mid cD_{k-1}^+ \mid \ldots \mid cD_1^+). \quad (29)$$

(3) W^+ に対して逆 HWT $\mathcal{H}^{-1}$ を適用し，出力 V^+ を得る．

$$cA_{i,x}^+ = cA_{i+1,\lceil x/2 \rceil}^+ + g(x) \cdot cD_{i+1,\lceil x/2 \rceil}^+, \quad (30)$$

$$V^+ = cA_0. \quad (31)$$

4.2.2 $\mathcal{T}_2$ の構成法

$\mathcal{T}_1$ の計算量は $O(n)$ となる．これは，$\mathcal{T}_1$ を構成する 3 つの手順の計算量がいずれも $O(n)$ である[*8]ことによる．この計算量は Barak らの手法 [1]（n の多項式時間）より優れており，Xiao らの Privelet と同等であるが，前述のとおり $m \ll n$ となる大規模な集計データでは実用的とはいえない．そこで，$\mathcal{T}_1$ と等価な出力を得つつ計算量を削減するアルゴリズム $\mathcal{T}_2$ を構成する．

計算量の削減にあたり，V が疎であるときには，ほとんどのノイズは実際には手順 2（非負精緻化）の過程で「捨てられる」ことに着目する．すなわち，$cD_{i,x}^*$ 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 となる．このとき，0 値をとるほうの部分木に含まれる 2^{i-1} 個の Laplace ノイズが出力値 V^+ に影響する可能性はなく，安全性にも寄与しない．したがって，安全性を損なうことなく，ノイズの付与を省略することができる．

そこで，非 0 値をとる $cA_{i,x}^+$ のみを対象として，Laplace メカニズムの適用と非負精緻化を再帰降下により同時に実施する．これにより，出力に寄与しない無駄なノイズを発生させることなく，出力 V^+ の差分プライバシを満たす．

(1) まず，最上位の Wavelet 係数である cA_k, cD_k に関し，それぞれ対応する cA_k^+, cD_k^+ を計算する．

$$cA_k^* = cA_k + \mathrm{Lap}(\lambda/2^k), \quad (32)$$

$$cD_k^* = cD_k + \mathrm{Lap}(\lambda/2^k), \quad (33)$$

$$cA_{k,1}^+ = \max\{cA_{k,1}^*, 0\}, \quad (34)$$

$$cD_{k,1}^+ = \begin{cases} -cA_{k,1}^+ & (cD_{k,1}^* < -cA_{k,1}^+) \\ cA_{k,1}^+ & (cD_{k,1}^* > cA_{k,1}^+) \\ cD_{k,1}^* & (\text{otherwise}). \end{cases} \quad (35)$$

(2) $i = \{k, ..., 2\}$ について，$\forall(x \mid cA_{i,x}^+ \neq 0)$ に対して下記を実行することにより，再帰的に cA_{i-1}^+ と cD_{i-1}^+ が得られ，最終的には cA_1^+ と cD_1^+ を得る．

$$cA_{i-1,2x-1}^+ = cA_{i,x}^+ + cD_{i,x}^+, \quad (36)$$

$$cA_{i-1,2x}^+ = cA_{i,x}^+ - cD_{i,x}^+, \quad (37)$$

$$p^* = cD_{i-1,2x-1} + \mathrm{Lap}(\lambda/2^i), \quad (38)$$

$$q^* = cD_{i-1,2x} + \mathrm{Lap}(\lambda/2^i), \quad (39)$$

[*8] 正確には，手順 3 の $\mathcal{H}^{-1}$ の計算は，付録 A.1 と同様の工夫により容易に $O(m^+ \log n)$ とすることができる．

$$cD_{i-1,2x-1}^+ = \begin{cases} -cA_{i-1,2x-1}^+ & (p^* < -cA_{i-1,2x-1}^+) \\ cA_{i-1,2x-1}^+ & (p^* > cA_{i-1,2x-1}^+) \\ p^* & (\text{otherwise}), \end{cases} \quad (40)$$

$$cD_{i-1,2x}^+ = \begin{cases} -cA_{i-1,2x}^+ & (q^* < -cA_{i-1,2x}^+) \\ cA_{i-1,2x}^+ & (q^* > cA_{i-1,2x}^+) \\ q^* & (\text{otherwise}). \end{cases} \quad (41)$$

(3) $\forall(x \mid cA_{1,x}^+ \neq 0)$ に対して下記を実行することにより，$V^+ = (v_0^+, v_1^+, ..., v_n^+)$ を得る．

$$v_{2x-1}^+ = cA_{1,x}^+ + cD_{1,x}^+, \quad (42)$$

$$v_{2x}^+ = cA_{1,x}^+ - cD_{1,x}^+. \quad (43)$$

5. 提案方式の安全性と有用性

3 章で提案した方式 $\mathcal{T}_1$，$\mathcal{T}_2$ により得られた出力 $V^+ = (v_1^+, v_2^+, \ldots, v_n^+)$ は，安全性と有用性に関する以下の性質を満たす．

- （安全性）V^+ は ϵ-差分プライバシを満たす．
- （非負制約の充足）V^+ は非負制約 $\forall v_i^+ \geq 0$ を満たす．
- （部分和劣化の抑制）V^+ を 2^l ごとの $q = 2^{k-l}$ 個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の分散は，$\frac{2}{3}\lambda^2(1 + \frac{2}{q^2})$ より小さい．これは，$l = k$ の（総和を表す）とき最大値 $2\lambda^2$ をとり，l が小さくなるに従って急速に $\frac{2}{3}\lambda^2$ に収束する．

さらに，構成法 $\mathcal{T}_2$ を用いた場合，提案方式は以下の性質を満たす．

- （計算量の抑制）m^+ を出力 V^+ における非 0 値の個数としたとき，$\mathcal{T}_2$ の計算量は $O(m^+ \log n)$ である．

以下，上記のそれぞれの性質について証明する．なお，定理 1〜3 については，$\mathcal{T}_1$ の出力に対する証明のみを示すが，補題 1 により $\mathcal{T}_2$ の出力に対してもそれぞれ成立する．

5.1 安全性

定理 1. V^+ は ϵ-差分プライバシを満たす．

証明. 提案方式 $\mathcal{T}_1$ の Matrix メカニズム [12] $\mathcal{M}_A(B, x) = Bx + BA^{-1}\mathrm{Lap}(\Delta A/\epsilon)$ への帰着と，差分プライバシの事後処理則を用いて，$\mathcal{T}_1$ が ϵ-差分プライバシを満たすことを証明する．

$W = \mathcal{H}(V)$ は V の一次変換であるため，HWT を表す行列 H を用いた行列積により $W = HV$ と表すことができる．ここで，W，V はそれぞれ列ベクトルとして表現されているとする．たとえば $V = (v_1, v_2, v_3, v_4)^T$ のとき $(n = 4)$，

$$W=\frac{1}{4}\begin{pmatrix}1&1&1&1\\1&1&-1&-1\\2&-2&0&0\\0&0&2&-2\end{pmatrix}\begin{pmatrix}v_1\\v_2\\v_3\\v_4\end{pmatrix}=\begin{pmatrix}\frac{v_1+v_2+v_3+v_4}{4}\\\frac{v_1+v_2-v_3-v_4}{4}\\\frac{v_1-v_2}{2}\\\frac{v_3-v_4}{2}\end{pmatrix}.$$

W^* は，n 次の対角行列 $I_H=\{h_{ij}\}$ を用いて Laplace ノイズをスケールさせることにより，

$$W^*=HV+I_H\mathrm{Lap}(\lambda)^n \tag{44}$$

と表すことができる．ここで，I_H の各要素 h_{ij} は以下の値をとる．

$$h_{ij}=\begin{cases}0 & (i\neq j)\\ 2^{-k} & (i=j=1)\\ 2^{-(k-\lceil\log_2 i\rceil)+1} & (\text{otherwise}).\end{cases} \tag{45}$$

たとえば $n=4$ のとき，$I_H=(1/4,1/4,1/2,1/2)I$ となる．

式 (44) において，両辺に H^{-1} を左から乗ずると下式を得る．

$$\begin{aligned}H^{-1}W^*&=V+H^{-1}I_H\mathrm{Lap}(\lambda)^n\\&=V+(I_H^{-1}H)^{-1}\mathrm{Lap}(\lambda).\end{aligned} \tag{46}$$

すなわち，$H^{-1}W^*$ は Matrix メカニズム $\mathcal{M}_{I_H^{-1}H}(I,V)$ に帰着されるため，$H^{-1}W^*$ は $\Delta(I_H^{-1}H)/\lambda$-差分プライバシを満たす．ここで，$\Delta(I_H^{-1}H)=1+\log_2 n$ である [12] ことから，

$$\begin{aligned}\Delta(I_H^{-1}H)/\lambda&=(1+\log_2 n)/\{(1+\log_2 n)/\epsilon\}\\&=\epsilon.\end{aligned} \tag{47}$$

したがって，$H^{-1}W^*$ は ϵ-差分プライバシを満たす．これに H を左辺から乗じた $HH^{-1}W^*=W^*$ も，H は V の値に関する情報を持たないことから，事後処理則により ϵ-差分プライバシを満たす．

また，$\mathcal{T}_1$ において，W^* を生成した後の手順，すなわち W^* から W^+ および V^+ を導出する過程において，V の値に関する知識は用いられていない[*9]ため，事後処理則の適用条件を満たす．したがって，$\mathcal{T}_1$ の出力である V^+ も ϵ-差分プライバシが保証される． □

5.2 非負制約の充足

定理 2. V^+ は非負制約 $\forall v_i^+\geq 0$ を満たす．

証明. $\mathcal{T}_1$ において，$i\in\{0..k-1\}$ のとき，

$$cA_{i,x}^+=cA_{i+1,\lceil x/2\rceil}^+ + g(x)\cdot cD_{i+1,\lceil x/2\rceil}^+ \tag{48}$$

であることから，

$$|cD_{i+1,\lceil x/2\rceil}^+|\leq cA_{i+1,\lceil x/2\rceil}^+ \tag{49}$$

が満たされるならば，$cA_{i,x}^+\geq 0$ が成立する．$V^+=cA_0^+$ であるため，これは $\forall v_i^+\geq 0$ の十分条件である．

すなわち，$cA_{k,1}^+\geq 0$ であり，$i\in\{1..k\}$ において，任意の $cA_{i,x}^+$ と $cD_{i,x}^+$ の組に対し，$|cD_{i,x}^+|\leq cA_{i,x}^+$ が成立すれば良い．これらは $\mathcal{T}_1$ の構成法における手順 (2) により明らかに満たされる． □

5.3 部分和劣化の抑制

定理 3. V^+ を 2^l ごとの q $(=2^{k-l})$ 個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の分散は，$\frac{2}{3}\lambda^2(1+\frac{2}{q^2})$ より小さい．

証明. V^+ を 2^l ごとの q $(=2^{k-l})$ 個のブロックに分割したときの部分和を $p_1^l,p_2^l,\ldots,p_q^l$ とする．HWT の性質により，$p_j^l=2^l\cdot cA_{l,j}^+$ となる．$\mathcal{T}_1$ において付与されるノイズは互いに独立であるため，$cA_{l,j}^+$ のノイズの分散は，$cA_k^+,cD_k^+,cD_{k-1}^+,\ldots,cD_{l+1}^+$ に与えられるノイズの分散の和になる．

ここで，$cD_{i,x}^+$ に与えられるノイズの分布は，$\mathrm{Lap}(\lambda/2^i)$ の分布の両端（具体的には $\pm cA_{i,x}^+$ の外側）を非負精緻化により「カット」した分布となる．したがって，その分散は $\mathrm{Var}(\mathrm{Lap}(\lambda/2^i))=2(\lambda/2^i)^2=2\lambda^2/4^i$ よりも小さい．

すなわち，cA_l^+ に含まれる各要素のノイズ分散 $\sigma_{k,l}^2$ について，以下が成立する．

$$\begin{aligned}\sigma_{k,l}^2&<\mathrm{Var}(\mathrm{Lap}(\lambda/2^k))+\sum_{i=l+1}^{k}\mathrm{Var}(\mathrm{Lap}(\lambda/2^i))\\&=2\left(\frac{\lambda}{2^k}\right)^2+\sum_{i=l+1}^{k}2\left(\frac{\lambda}{2^i}\right)^2\\&=2\lambda^2\left(\frac{1}{2^{2k}}+\sum_{i=1}^{k}\frac{1}{2^{2i}}-\sum_{i=1}^{l}\frac{1}{2^{2i}}\right)\\&=2\lambda^2\left\{\frac{1}{2^{2k}}+\frac{1}{3}\left(\frac{1}{2^{2l}}-\frac{1}{2^{2k}}\right)\right\}\\&=\frac{2}{3}\lambda^2\left(\frac{1}{2^{2k-1}}+\frac{1}{2^{2l}}\right).\end{aligned} \tag{50}$$

$p_j^l=2^l\cdot cA_{l,j}^+$ であるため，その分散 $\mathrm{Var}(p_j^l)$ は，

$$\begin{aligned}\mathrm{Var}(p_j^l)&=2^l\sigma_{k,l}^2\\&<2^l\cdot\frac{2}{3}\lambda^2\left(\frac{1}{2^{2k-1}}+\frac{1}{2^{2l}}\right)\\&=\frac{2}{3}\lambda^2\left(1+\frac{1}{2^{2(k-l)-1}}\right)\\&=\frac{2}{3}\lambda^2\left(1+\frac{2}{q^2}\right).\end{aligned} \tag{51}$$

□

以下，式 (51) で示される部分和劣化の抑制効果について説明を加える．

まず，Laplace メカニズムにおける部分和精度の劣化につ

[*9] V^+ の導出に用いられる非負制約は V の値域を定義するドメイン知識であり，V の具体的な値には関係しない．

いて定量的に議論する．単純な Laplace メカニズムは，個々のセルに対してそれぞれ独立した Laplace ノイズ $\mathrm{Lap}(\epsilon^{-1})$ を付与することにより ϵ-差分プライバシを達成する．互いに独立した確率変数の和の分散はそれぞれの分散の和になることから，Laplace メカニズムにおける部分和 p_j^l のノイズ分散 $\mathrm{Var}_{Lap}(p_j^l)$ は，$\mathrm{Var}_{Lap}(p_j^l) = 2(2^l)\epsilon^{-2}$ となる．すなわち部分和 p_j^l のノイズ分散は，その領域サイズ 2^l に比例して大きくなる（部分和劣化が発生する）．特に $l = k$ のとき，すなわち全体の総和におけるノイズ分散は n に比例することになるため，$m \ll n$ であるような疎なデータへの適用は現実的とはいえなくなる．

その一方，提案方式のノイズ分散は，式 (51) が示すとおり $\mathrm{Var}(p_j^l) < \frac{2}{3}\lambda^2(1+\frac{2}{q^2})$ となる．（非負精緻化の効果により）左辺と右辺は不等号で結ばれていることに留意する．$q = 2^{k-l}$ であることから，右辺の値は l に対して単調増加する．具体的には，$l = 0$（個々の v_i^+ に相当）のとき最小値 $\frac{2}{3}(1+\frac{2}{n^2})\lambda^2$ をとり，$l = k$（V^+ の全領域の総和に相当）のときに最大値 $2\lambda^2$ をとる．すなわち部分和 p_j^l のノイズ分散の上限は，その領域サイズ 2^l の拡大にともなって単調に増加するが，たかだか定数倍（具体的には 3 倍未満）にとどまり，部分和の劣化が抑制されている．

しかし，その代償として，提案方式は Laplace メカニズムと比較して小領域の部分和の精度が劣る可能性がある．たとえば $l = 0$ のとき，Laplace メカニズムにおけるノイズ分散 $\mathrm{Var}_{Lap}(p_j^0)$ は（n に依存せず）$\mathrm{Var}_{Lap}(p_j^0) = 2\epsilon^{-2}$ となる．その一方，提案方式においては，$\lambda = (1+\log_2 n)/\epsilon$ より，

$$\mathrm{Var}(p_j^0) < \frac{2}{3}\left(1+\frac{2}{n^2}\right)\lambda^2 \simeq \frac{2}{3}(1+\log_2 n)^2\epsilon^{-2}. \quad (52)$$

すなわち $\mathrm{Var}(p_j^0)$ の上限はおよそ $\log^2 n$ に比例する．

ただし，この上限値は，非負精緻化によるノイズカットがまったく発生しない場合の値に相当することに留意する．実際のデータ，特に $m \ll n$ となるような疎な集計データに提案方式を適用した場合，非負精緻化によるノイズ削減効果により，実際にはこれを大きく下回るノイズ分散が得られることが期待される．

したがって，小領域の部分和の精度について，単純な Laplace メカニズムと提案方式とでどちらが優れるかは一概に結論づけることはできず，データ依存となる．ただし，定性的には，対象データが疎であるほど（非負精緻化によるノイズカットが多く発生するため）提案方式に有利で，逆に密なデータでは Laplace メカニズムに有利であることが予想される．

なお，最後に Privelet 法との比較について簡単に述べる．Privelet 法のノイズ分散は式 (51) の不等号を等号に置き換えたものになる（提案方式でまったく非負精緻化が起こらない場合に相当する）ため，理論的に提案方式の精度は Privelet 法よりつねに優れる．また，Privelet 法は，部分和の領域拡大にともなう精度劣化に関しては提案方式と同様の抑制効果を持つが，小領域の部分和精度に関しては，（提案方式と異なり）非負精緻化によるノイズ削減効果を得ることができない．したがって，Privelet 法における小領域の部分和の有用性は，Laplace メカニズムと比較して劣ることになる．

5.4 計算量の抑制

定理 4. $\mathcal{T}_2$ の計算量は $O(m^+ \log n)$ である．

証明. まず $\mathcal{T}_2$ における手順 (3) に着目する．$cA_{1,x}^+ \neq 0$ のとき，v_{2x-1}^+, v_{2x}^+ の少なくともいずれか片方は非 0 値をとる．そのため，cA_1^+ に含まれる非 0 値の個数はたかだか m^+ 個である．したがって，手順 (3) の計算量は $O(m^+)$ となる．

次に，手順 (2) に着目する．同様に，$cA_{i,x}^+ \neq 0$ のときに $cA_{i+1,2x-1}^+, cA_{i+1,2x}^+$ の少なくともいずれかは非 0 値をとることから，cA_i^+ に含まれる非 0 値の個数は，たかだか cA_{i+1}^+ に含まれる非 0 値の個数となり，m^+ を超えることはない．したがって，手順 (2) は $O(m^+)$ の処理を $k-1$ 回実行することになるため，その計算量は $O(m^+k) = O(m^+ \log n)$ となる．

また，手順 (1) の計算量は明らかに $O(1)$ のため，$\mathcal{T}_2$ の計算量は，$O(m^+) + O(m^+ \log n) + O(1) = O(m^+ \log n)$ となる． □

6. 地理空間データへの適用

提案方式の実問題への適用例として，地理空間データ（geospatial data）への適用を考える．地理空間データとは，人口分布や降雨量分布，交通量分布など，地理的な場所に応じて変化する「量」を表すデータである．

地理空間データは一般的に二次元以上のデータであるため，提案方式への適用にあたっては，提案方式を多次元の入力に対応できるよう拡張する必要がある．提案方式で用いている HWT は標準分解（standard decomposition）などの構成法を用いて簡単に多次元化することができるため，提案方式も比較的容易に多次元データに対応させることができる．しかし，多次元 HWT に基づいて差分プライバシを満たすために必要なノイズの強度は，一次元 HWT に基づく際に必要なノイズの強度よりも大きい[*10]ため，出力精度の大幅な劣化を招く．

そこで，提案方式の地理空間データへの適用に際し，まず多次元空間から一次元空間への全単射を用いてデータの次元を圧縮し，それにより得られた一次元写像に対して提案方式を適用するアプローチをとる．これにより，多次元

[*10] より具体的には，必要ノイズ強度のデータ量 n に対するオーダーが，一次元 HWT に基づく場合には $O(\log n)$ であるのに対し，d-次元 HWT に基づく場合には $O(\log^d n)$ となる [20].

HWT を用いることなく多次元の地理空間データを提案方式で扱うことができる．以下，簡単のため対象データを二次元の地理空間データとして議論を進める．

二次元データを一次元に次元圧縮するための単純な方法としては，二次元データを縦方向（緯度方向）もしくは横方向（経度方向）に順に走査していくことにより一次元に再配列し，これを提案方式の入力 V として用いることが考えられる．この方法では，V における連続した領域が，二次元空間上の「縦一列（もしくは横一列）」の領域に対応することになる．提案方式では，連続領域の部分和の精度が高くなるため，この方法で次元圧縮をした場合は「縦一列」もしくは「横一列」に並んだ領域について，高い部分和精度を持つことになる．

しかし，実際の地理空間データの応用においては，一般にこのような部分和が用いられることはほとんどない．たとえば 500 m メッシュ人口における 2×2 の 4 セルの部分和を 1 km メッシュ人口として用いるなど，正方領域（もしくは正方領域に近い形）の部分和が用いられることが多い．

そこで，次元圧縮にあたっては，地理空間データの分析の際に頻繁に用いられる正方領域に関する部分和の精度に着目する．正方領域の部分和精度を向上させるためには，二次元空間上の正方領域を，一次元空間上の連続した領域に射影できることが望ましい．

このような変換として，局所性保存写像の一種である Morton 順序写像を用いる方式を提案する．Morton 順序写像は，$2^k \times 2^k$ の大きさの二次元領域を等分割する $2^l \times 2^l$ の正方領域を，長さ 2^{2k} の一次元ベクトルにおける連続した 2^{2l} 個の領域に射影する性質を持つ．また，次節で示すように，Morton 順序写像による射影の計算は bit interleaving により簡単かつ高速に実行することができる．

6.1 Morton 写像

Morton 順序写像（Morton order mapping）とは，多次元空間から一次元空間への全単射であり，元の空間上で距離の遠近が投影先の距離の遠近に反映される性質を持つ，局所性保存写像の一種である．二次元空間を再帰的に “Z” 字状に走査した順序で一次元に投影することから，Z-曲線順序写像（Z-curve order mapping）とも呼ばれる．

二次元 Morton 順序写像 $\mathcal{F}$ の具体的な構成方法について簡単に説明する．Morton 順序写像の適用 $V = \mathcal{F}(M)$ において，$\mathcal{F}$ は 2^n 行 2^n 列の行列からなる入力 M の要素 m_{ij} を，長さ 2^{2n} の出力 V の要素 v_k に射影する（簡単のため，i, j, k はいずれも 0 から始まるとする）．このとき，k は以下の二進数表現により得られる．

$$i = (i_n\ i_{n-1}\ \cdots\ i_1)_2, \tag{53}$$

$$j = (j_n\ j_{n-1}\ \cdots\ j_1)_2, \tag{54}$$

$$k = (i_n\ j_n\ i_{n-1}\ j_{n-1}\ \cdots\ i_1\ j_1)_2. \tag{55}$$

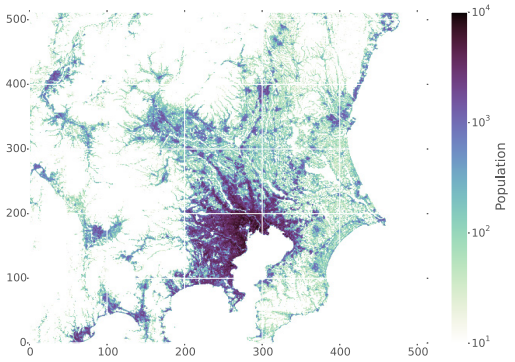

図 1 評価対象データの分布

Fig. 1 Geographic distribution of the data used in the evaluations.

すなわち，i, j をそれぞれビット分解した上で，それぞれから得られた各ビットを「互い違いに」合成する，いわゆる bit interleaving 操作のみによって k を得られることが Morton 順序写像の特長である．

なお，$\mathcal{F}$ の逆変換 $\mathcal{F}^{-1}$ についても，式 (53) から容易に導くことができる．

6.2 地理空間データに基づく評価

Morton 順序写像により一次元化された地理空間データを用い，提案方式から得られる部分和の精度について評価した結果を示す．評価対象データとして，実世界に基づく地理空間データの 1 つである H22 国勢調査による地域メッシュ人口を用い，Laplace メカニズム [4] および Privelet 法 [20] との比較を通じて提案方式の改善効果を定量的に評価する．

本評価では，地理空間データの例として，H22 国勢調査に基づく地域メッシュ統計 [22] の 500 m メッシュ人口（1/2 地域メッシュ人口）から，首都圏周辺の 256 km 四方（$n = 512 \times 512 = 2^{18}$）を抽出したものを評価対象データとして用いる（以下，対象データと呼ぶ）．なお，対象データのうち，非 0 値を持つセル数 m は 95,317 セルであった．データ密度 m/n は約 36.3%であり，人口データとしてはかなり密度が高い．これは，対象データは（性別，年代別などの）属性ごとに分計していないこと，世界有数の人口集中地域である首都圏を対象としていることによる．**図 1** に対象データの分布を示す．

対象データに再配列関数 $\mathcal{F}$ を適用して得られたデータに対し，単純な Laplace メカニズム $V + \mathrm{Lap}(1/\epsilon)^n$，Privelet 法，提案方式をそれぞれ適用し，出力におけるノイズの大きさ（対象データに対する誤差）を評価する．Morton 順序写像の導入による効果を確認するため，Privelet 法については，（Morton 順序写像を用いずに）[20] の多次元 Privelet 法をそのまま適用した結果と，提案方式と同様に Morton 順序写像を用いて一次元化した値に対して一次元 Privelet を適用した結果の両方を検証する．以降，多次元 Privelet

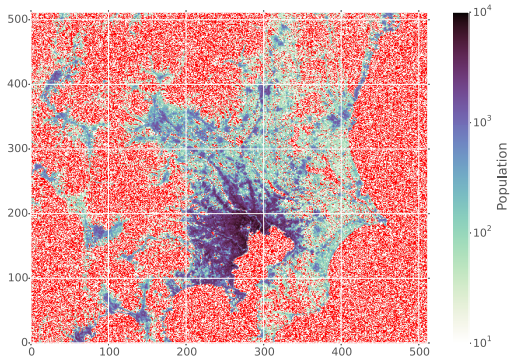

図 2 Laplace メカニズムの適用結果
Fig. 2 Geographic distribution of the output of Laplace mechanism.

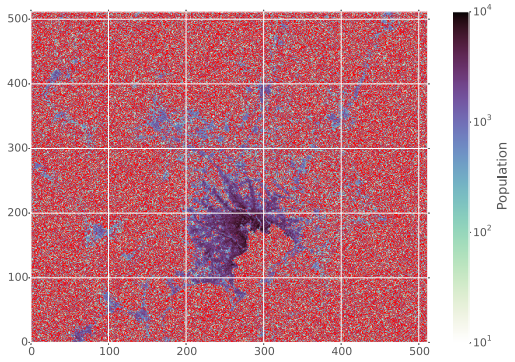

図 3 Privelet 法の適用結果（二次元 Privelet）
Fig. 3 Geographic distribution of the output of 2D-Privelet.

法を Privelet（2D），Morton 順序写像を適用した Privelet 法を Privelet（Morton）と表記する．

評価にあたり，部分和の範囲の大きさを変動させ，誤差との関係を見る．誤差指標としては MAE（mean absolute error）および RMSE（rooted mean squared error）を用いた．なお，それぞれにおいて，$\epsilon = 0.1$ と設定した[*11]．

6.3 評価結果

図 2，**図 3**，**図 4**，**図 5** に，対象データに対してそれぞれ Laplace メカニズム，Privelet（2D），Privelet（Morton），提案方式を適用した結果の地理分布を示す．

各図中の赤色の領域は，非負制約を逸脱する（負値をとる）セルを示す．対象データと提案方式の出力において負値をとるセルはなく，非負制約が充足されていることが分かる．

それに対し，Laplace メカニズムによる出力と 2 種類の Privelet 法による出力は，いずれも多数のセルで非負制約を

[*11] 文献 [19], [20] では，同文献中で提案された generalized sensitivity という概念を用い，Wavelet 係数に与えるノイズスケールの母数 λ を $\lambda = 2(1 + \log_2 n)$ としているが，Matrix メカニズムに基づくと提案方式と同様に $\lambda = (1 + \log_2 n)$ とできる（ノイズの強度が半分になる）．提案方式との対比をより明確にするため，本実験では後者の条件（提案方式と同じパラメータ）で評価した．前者の条件（文献 [19], [20] で与えられたパラメータ）で評価する場合，Privelet 法の MAE と RMSE は，本稿による評価結果の約 2 倍の値となることに注意が必要である．

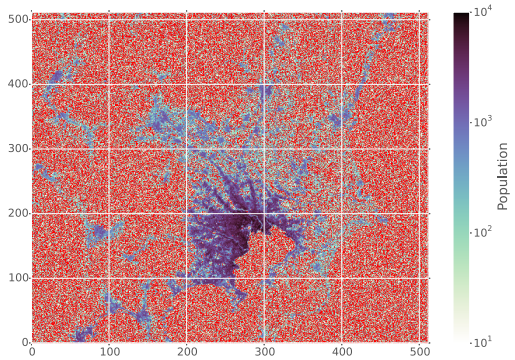

図 4 Privelet 法の適用結果（Morton 順序写像を適用）
Fig. 4 Geographic distribution of the output of Privelet (Mortonized).

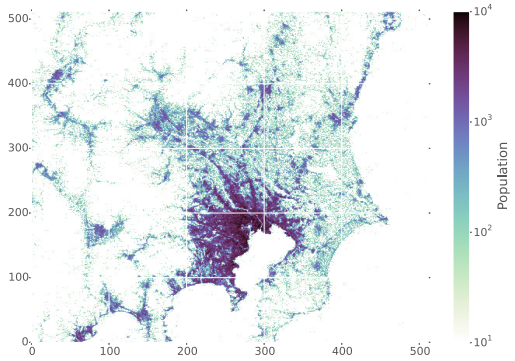

図 5 提案方式の適用結果
Fig. 5 Geographic distribution of the output of the proposed method.

逸脱している．具体的には，Laplace メカニズムで 88,399 セル（全体の 33.7%），Privelet（2D）で 114,533 セル（同 43.7%），Privelet（Morton）で 105,275 セル（同 40.2%）が負値をとった．

表 2 に部分和の領域サイズ（セル数）と，それぞれの部分和における誤差（MAE，RMSE）の関係を示す．試行は 100 回行い，その平均値を評価結果とした．

図 6 に表 2 を図示する．同図において x 軸（Area size）は部分和の領域サイズを 2 の対数で示し，y 軸（Error）はそのときの誤差の大きさ（MAE，RMSE）を示している．たとえば，グラフ上の $x = 6$ 上の点は，$2^6 = 64$ セルの部分和，すなわち 8×8 メッシュ（4 km メッシュ相当）の正方領域に含まれる人口の合算値の誤差を表す．

これらの結果から，Laplace メカニズムは領域サイズが小さいときには誤差が低く抑えこまれているものの，領域サイズのほぼ二乗根に比例して RMSE が，それよりやや大きい割合で MAE が増大しており，部分和精度の劣化が発生していることが確認できる．それに対し，提案方式と Privelet 法では部分和の領域サイズが大きくなっても誤差の増大が抑えこまれている．

提案方式と Privelet 法との比較のため，**図 7** に図 6 のうち $y \leq 800$ の部分を拡大したものを示す．さらに，$x \leq 10$，

表 2 部分和の領域サイズと誤差の関係
Table 2 Relationship between area size and error.

手法	誤差指標	部分和の領域サイズ（Area size）									
		2^0	2^2	2^4	2^6	2^8	2^{10}	2^{12}	2^{14}	2^{16}	2^{18}
Laplace [4]	MAE	10.00	21.87	44.77	90.03	180.69	363.21	722.03	1,402.73	2,806.58	5,553.95
	RMSE	14.14	28.28	56.55	113.06	226.48	455.65	906.12	1,789.27	3,474.87	6,907.69
Privelet（2D）[20]	MAE	361.51	361.27	361.67	361.93	361.25	366.09	378.30	414.57	593.86	1,178.56
	RMSE	471.37	471.01	471.66	472.52	471.26	479.17	494.20	531.12	756.30	1,631.86
Privelet（Morton）	MAE	115.74	115.73	115.73	115.56	115.57	116.55	116.69	118.01	121.70	209.08
	RMSE	155.16	155.16	155.09	154.90	155.00	155.88	155.69	155.93	161.45	317.79
Proposed	MAE	28.73	44.49	60.14	74.91	89.41	101.02	111.14	119.87	124.95	139.32
	RMSE	66.20	87.54	106.18	121.09	135.07	145.33	152.92	158.95	168.02	195.30

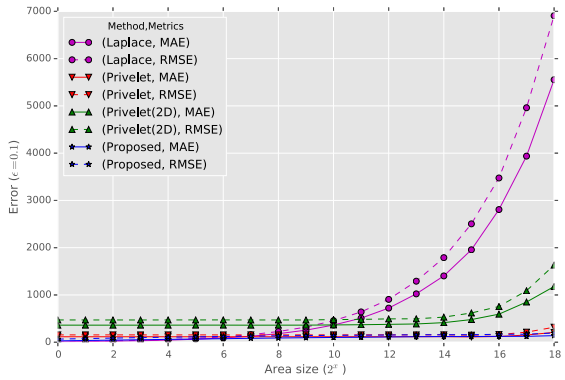

図 6 部分和の領域サイズと誤差の関係
Fig. 6 Relationship between area size and error.

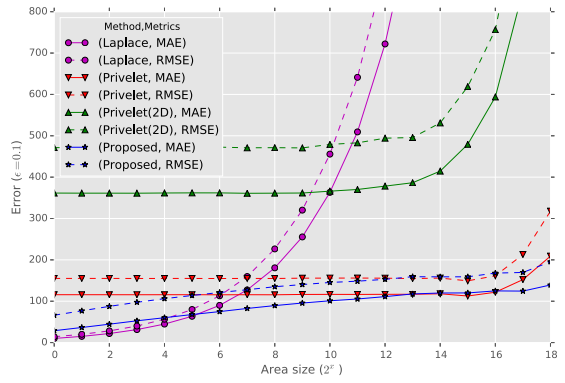

図 7 部分和の領域サイズと誤差の関係（$y \leq 800$ 部分を拡大）
Fig. 7 Relationship between area size and error ($y \leq 800$).

すなわち $2^{10} = 32 \times 32$ メッシュ（16 km メッシュ相当）以下の大きさの部分和において，提案方式は Morton 順序写像を適用した Privelet に対しても精度が改善しており，特に部分和サイズが小さいほど改善効果が大きいことが確認できる．

6.4 考察

前節の評価結果について，3.2 節で示した課題に照らしあわせて議論する．

まず非負制約の逸脱について，Laplace メカニズム，Privelet（2D），Privelet（Morton）のいずれも，多量の負値を含む結果となり，非負制約を逸脱している．その一方で，提案方式の出力は負値を含まず，非負制約が満たされていることが確認された．これらはいずれも理論通りの結果である．

部分和精度の劣化に関しては，それぞれの方式で異なる傾向を示した．まず，Laplace メカニズムは，小領域の部分和において最も精度が高い（MAE では $\sim 2^4$（2 km メッシュ）の領域，RMSE では $\sim 2^6$（4 km メッシュ）の領域）が，部分和の領域サイズが広くなるにつれて大きく誤差が増大する．5.3 節で議論したとおり，Laplace メカニズムにおいて，部分和のノイズ分散は部分和の領域サイズに比例するので，これをそのまま裏付ける結果となった．

その一方，Privelet（2D）と Privelet（Morton）は，いずれもおよそ $\sim 2^{12}$（32 km メッシュ）の領域まで MAE，RMSE ともにほぼ一定であり，そこから領域サイズが広がるにつれて単調に増加するものの，2 倍から 3 倍程度の増加までにとどまる．すなわち，部分和の精度が維持されていることが確認された．

また，Privelet（Morton）の RMSE は，5.3 節で示した提案方式のノイズ分散の上界とほぼ等しくなることが確認できた．実際に $n = 2^{18}, \epsilon = 0.1$ の条件下で式 (52) を計算すると，$\frac{2}{3}(1 + \log_2 n)^2 \epsilon^{-2} = \frac{2}{3} \cdot 19^2 \cdot 10^2 \simeq 24{,}067$ となる．その平方根をとると約 153.13 となり，これは Privelet（Morton）における 2^0 の RMSE とほぼ等しい．

ただし，Privelet 法の小領域の部分和の精度は Laplace メカニズムと比較して大きく劣る．具体的には，最も差が大きい 2^0 の部分和（= 個々のセル値）において，Privelet（Morton）で 10 倍程度，Privelet（2D）で 30 倍程度の開きがある．これも 5.3 節の議論のとおりである．

なお，これら 2 種類の Privelet 法で比較すると，ほぼ一貫して Privelet（2D）の MAE，RMSE は Privelet（Morton）と比較して 3 倍程度大きい，すなわち，本稿で示した Morton 順序写像を用いる手法が，Privelet 法においても

多次元データの精度改善に有効であることが確認された．

その一方，提案方式の MAE，RMSE は，領域サイズの拡大につれてなだらかに単調増加を続ける傾向が見られた．これは，領域サイズが小さいほど非負精緻化によるノイズカットの効果が大きいためである．たとえば，領域サイズ 2^0 において，前述のとおり RMSE の理論的な上界は約 153.13 であるが，本評価では約 66.23 となっており，半分以下（約 43%）の誤差にとどまる．ただし，その改善効果は領域サイズの拡大にともない徐々に弱まり，領域サイズ 2^{12} でほぼ見られなくなる．

この効果により，Privelet 法の弱点であった小領域での部分和の精度悪化が緩和されていることが確認できる．領域サイズ 2^0 において，前述のとおり Privelet（2D）の誤差は MAE と RMSE のいずれも Laplace メカニズムの 30 倍程度，Privelet（Morton）は 10 倍程度であったが，提案方式では MAE で約 2.9 倍，RMSE で約 4.7 倍に抑えられている．この改善は非負精緻化が発生することにより得られるものであるため，その効果の大きさは対象データの分布に依存する．非負精緻化の発生条件から，具体的にはデータが疎であるほど高い効果が得られると考えられる．今後の課題として，対象データの分布と部分和精度の改善効果の関係の定量化ないし定式化があげられる．

次に，計算量の増大について考察する．提案方式以外は，いずれも出力に 0 値を含まず，データ密度 m/n は元となる対象データの 36.3%から 100%に増大した．それに対し，提案方式による出力のデータ密度は約 27.5%であり，ほぼ元のデータ密度を維持している（若干の減少となった）．データ密度に変化がある理由は，Laplace ノイズの付与により 0 値から正値に変化するセルがある一方で，非負精緻化により逆方向に変化するセルが存在するためである．この増減がどの程度発生するかはデータの分布により定まり，データのロングテイル性が高いほどデータ密度は減少する傾向を持つことが予想される．その検証と定量化は今後の課題である．

最後に，本評価における計算時間について簡単に触れる．今回の対象データの評価は，Python 2.7.9 と数値演算ライブラリ NumPy 1.9 を用いた自作プログラムにより，intel Core i5-2520M CPU（2.5 GHz）を備えたノート PC 上で実施した．本評価の対象データはセル数 $n = 2^{18}$ という（実データとしては）それほど大規模ではないセル空間を持つデータであること，対象データのデータ密度が比較的高めであったことから，いずれの方式でも 20〜30 ms 程度で 1 回の試行を完了した．提案方式の計算量改善による計算性能の向上効果を定量的に検証するためには，より大規模かつ高次元のデータセットを用いた評価が必要となる．

7. まとめ

本稿では，地理空間データなどの大規模な集計データのプライバシを差分プライバシ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の向上に着目した手法を提案した．差分プライバシは，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必要があることを示した．

上記問題を解決するために，Wavelet 変換と Top-down 精緻化と呼ぶ手法を導入する方式を提案した．提案方式は，差分プライバシを満たしつつ，上記 3 点の課題を同時に解決することを証明した．具体的には，部分和精度に関し，提案方式は集計データを 2^l 個のブロックに分割した部分和において，l の大きさにかかわらず（部分和の範囲の大きさにかかわらず）その精度を一定以上に保つことが保証され，さらに集計データが疎であるほどその精度は向上する．また，計算量に関し，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，$m^+ \ll n$ 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．

さらに，提案方式を多次元の地理空間データに適用するための方式として，局所性保存写像の一種である Morton 順序写像を導入する方式を示した．これにより，精度劣化の要因となる多次元 HWT を用いることなく多次元データへの対応を可能とした．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，上記であげた非負制約の逸脱，部分和精度の劣化，計算量の増大について，いずれも改善が確認された．提案方式を適用した集計データは，Privelet 法と比較して同等以上に部分和の精度を向上させ，特に小範囲の部分和について大きく精度に優れることが明らかになった．

ただし，提案方式による部分和精度や計算量の改善効果は，適用対象となるデータの密度や分布に依存して変動すると考えられる．これらの定量的な検証のために，より大規模かつ高次元なデータセットへの適用および評価を進めることが今後の課題としてあげられる．

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*, pp.273–282, ACM Press (2007).

[2] Black, N.: Evidence based policy: Proceed with care, *British Medical Journal*, Vol.323, No.7307, pp.275–279 (2001).

[3] Cormode, G., Procopiuc, M., Srivastava, D. and Tran,

T.: Differentially Private Publication of Sparse Data, *Proc. Intl. Conf. Database Theory* (*ICDT2012*) (2012).
[4] 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, pp.1–12, Springer (2006).
[5] 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*, pp.1–13, Springer-Verlag (2007).
[6] Dwork, C.: Differential privacy: A survey of results, *Proc. 5th Intl. Conf. Theory and Applications of Models of Computation*, pp.1–19, Springer-Verlag (2008).
[7] 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).
[8] Ghosh, A., Roughgarden, T. and Sundararajan, M.: Universally Utility-maximizing Privacy Mechanisms, *SIAM J. Computing*, Vol.41, No.6, pp.1673–1693 (2012).
[9] 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, pp.1021–1032, VLDB Endowment (2010).
[10] 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).
[11] 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).
[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*), pp.123–134, ACM Press (2010).
[13] Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M.: *l*-diversity: Privacy Beyond *k*-anonymity, *ACM Trans. Knowledge Discovery from Data* (*TKDD*), Vol.1, No.1 (2007).
[14] 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).
[15] Misuraca, G., Mureddu, F. and Osimo, D.: Policy-Making 2.0: Unleashing the Power of Big Data for Public Governance, *Open Government*, Gascó-Hernández, M. (Ed.), Public Administration and Information Technology, Vol.4, pp.171–188, Springer (2014).
[16] 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).
[17] Sweeney, L.: k-anonimity: A model for protecting privacy, *Intl. J. Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol.10, No.5, pp.557–570 (2002).
[18] Xiao, X. and Tao, Y.: m-Invariance: Towards Privacy Preserving Re-publication of Dynamic Datasets, *Proc. 2007 ACM SIGMOD Intl. Conf. Management of Data*, pp.689–700, ACM (2007).
[19] Xiao, X., Wang, G. and Gehrke, J.: Differential privacy via wavelet transforms, *Proc. 26th Intl. Conf. Data Engineering* (*ICDE 2010*), pp.225–236, IEEE (2010).
[20] 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).
[21] 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).
[22] 総務省統計局：地域メッシュ統計の特質・沿革，入手先 ⟨http://www.stat.go.jp/data/mesh/pdf/gaiyo1.pdf⟩.
[23] 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).
[24] 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).

付　　録

A.1 Haar Wavelet 変換の計算量削減

4.1 節の式 (16) の計算量を $O(m \log n)$ に削減する構成法を示す．

V は COO 形式などの疎データ形式で入力されるとする．$\forall v_i (v_i \neq 0)$ に対し，順番に，

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

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

を計算することにより Haar 分解 $\mathcal{H}_1$ を実現する．ここで，$g(\cdot)$ は式 (23) で与えられる符号関数である．なお，cA_1, cD_1 もそれぞれ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 $cA_1 = cD_1 = \{0\}^{n/2}$ とする．

この手順による Haar 分解の計算量は明らかに $O(m)$ であり，cA_1 に含まれる非 0 値の個数を m_1 とすると，$m_1 \leq m$ となる．これを再帰的に cA_k, cD_k まで繰り返せば，$k = log_2 n$ 回の Haar 分解により $W = \mathcal{H}(V)$ を得ることができる．その計算量は，cA_i に含まれる非 0 値の個数を m_i としたとき，$\forall i, m_i \leq m$ であることから，

$$\begin{aligned} O(\mathcal{H}(V)) &= O(m) + O(m_1) + \cdots + O(m_{k-1}) \\ &= O(km) = O(m \log n). \end{aligned} \tag{A.3}$$

すなわち，この構成法に基づく HWT $\mathcal{H}$ の計算量は $O(m \log n)$ となる．

A.2 補題 1 の証明

$\mathcal{T}_1$，$\mathcal{T}_2$ において，cA_1^+ および cD_1^+ の分布がそれぞれ等しいなら，V^+ の分布も等しくなり，補題 1 が成立する．そこで，$\forall i \in \{1..k\}$ において，cA_i^+ および cD_i^+ が $\mathcal{T}_1, \mathcal{T}_2$ のいずれにおいても等しい分布を持つことを数学的帰納法で示す．

証明．まず，$i = k$ のとき，すなわち cA_k^+ と cD_k^+ については，それぞれの定義から明らかに $\mathcal{T}_1$，$\mathcal{T}_2$ で同一である．

次に，$1 \leq i < k$ のときを考える．$\mathcal{T}_1$ の手順 2 における $cA_{i,x}^+$ の導出式において，$j = i+1$，$y = \lceil x/2 \rceil$ と置換すると，これは $\mathcal{T}_2$ の手順 2 における導出式と等価となる．す

なわち，cA^+_{i+1} および cD^+_{i+1} が $\mathcal{T}_1$，$\mathcal{T}_2$ でそれぞれ等しい分布をとるならば，cA^+_i の分布も等しい．

また，cD_i について，$cA^+_{i,x} = 0$ のときは，$\mathcal{T}_1$，$\mathcal{T}_2$ のいずれにおいても $cD^+_{i,x} = 0$ となる．$cA^+_{i,x} \neq 0$ のとき，$\mathcal{T}_2$ の手順 2 における $cD^+_{i-1,2x-1}$ の導出において，$j = i-1$，$y = 2x-1$ と置換すると，$\mathcal{T}_1$ の手順 2 における $cD^+_{i,x}$ の導出と等価な式となる．これは $cD^+_{i-1,2x}$ の導出においても同様である．したがって，cA^+_i の分布が $\mathcal{T}_1$，$\mathcal{T}_2$ で等しいならば，cD^+_i がとる分布も等しい．

すなわち，$\mathcal{T}_1$，$\mathcal{T}_2$ において，(1) $i = k$ のとき，cA^+_i および cD^+_i の分布はそれぞれ等価である．(2) $1 \leq i < k$ のとき，cA^+_{i+1} および cD^+_{i+1} の分布が等価であるならば，cA^+_i の分布も等価であり，(3) cA^+_i の分布が等価であるならば，cD^+_i の分布も等価である．したがって，cA^+_1 および cD^+_1 は $\mathcal{T}_1$，$\mathcal{T}_2$ で等しい分布を持つ． □

推薦文

本稿では，集計データのプライバシを差分プライバシ基準に基づいて保護するうえで，データの統計的正確性と計算効率を効果的に向上させる手法を提案している．Wavelet 変換と Top-down 精緻化処理と呼ぶ方法を組み合わせた，差分プライバシ基準を満たす新たなプライバシ保護方式を提案し，国勢調査データを用いて提案手法の評価した．理論的な新規性はもとより，ビッグデータサイエンスにおける実応用面の貢献も大きいことが期待される，良質な研究成果の報告となっている．

（コンピュータセキュリティ研究会主査　西垣正勝）

寺田 雅之 （正会員）

1995 年神戸大学大学院工学研究科修士課程修了，同年日本電信電話（株）入社．同社情報通信研究所，情報流通プラットフォーム研究所を経て，2003 年（株）NTT ドコモへ転籍．2008 年電気通信大学大学院電気通信研究科博士後期課程修了．博士（工学）．2009 年より現職．情報セキュリティ技術，プライバシ保護技術，大規模データ処理技術の研究開発に従事．電子情報通信学会会員．

鈴木 亮平

2010 年東京大学大学院情報理工学系研究科電子情報学専攻博士課程修了．博士（情報理工学）．同年（株）NTT ドコモ入社．ユビキタスコンピューティング，分散処理基盤の研究開発に従事．電子情報通信学会会員．

山口 高康 （正会員）

2001 年電気通信大学大学院電気通信学研究科博士前期課程修了．同年 (株) NTT ドコモ入社．以後，携帯端末での撮影対象判別技術，権利価値流通技術，コンテンツ検索技術，統計情報作成技術，プライバシ保護技術の研究開発に従事．

本郷 節之 （正会員）

1984 年岩手大学大学院工学研究科修士課程修了．同年日本電信電話公社入社．1987 年 ATR 視聴覚機構研究所へ出向．1991 年 NTT ヒューマンインタフェース研究所へ復帰．この間，視覚情報処理モデルの研究に従事．著書『脳・神経システムの数理モデル』（共著）ほか．工学博士．1999 年 NTT ドコモマルチメディア研究所へ転籍．2001 年セキュリティ方式研究室長．2010 年北海道工業大学（現北海道科学大学）教授に着任，現在に至る．モバイルセキュリティならびにプライバシ保護技術の研究開発に従事．電子情報通信学会，IEEE 各会員．