非負集計データのための部分和精度に優れた差分プライバシー適用手法二次元化の試み

本郷 節之[1]　杉尾 信行[1]　寺田 雅之[2]

概要：本研究では，元のデータベースに含まれる個々のデータの集合体（個票）から，何らかの条件を満たすデータの個数を数えた数値データの集合体であり，さらに，全体的に疎な分布をとるような集計データを対象とする．Dwork らが提案した差分プライバシー基準は，データベースへの問い合わせを行った際に，「ある特定のデータがデータベースに含まれているか否かを問い合わせ結果から判別することが困難である」ことを安全性の根拠とするプライバシー保護基準である．差分プライバシー基準を満たす代表的な手法に Laplace メカニズムがあるが，大規模集計データに適用する場合には，「非負制約の逸脱」「部分和精度の劣化」「疎データの密度急増」といった問題への対処が必要となる．我々は以前，これら 3 点の課題を同時に解消・改善する手法として，「非負精緻化を伴う Privelet 法」を提案した．しかしこの手法は一次元データ系列を対象としており，例えば地理空間上に配置されたデータへ適用する際には，二次元のデータを一次元に変換する前処理と，その逆の後処理を行う必要がある．そこで本研究では，本来一次元データを対象とするこの手法を二次元化する方法を検討している．本稿では，今回構築した実現方法を解説するとともに，一次元方式との比較を通じて，二次元方式の特性評価を試みる．

1. はじめに

あらゆる日常シーンが情報通信ネットワークと融合していると言っても過言ではない現代社会においては，そこで得られた多種多様なデジタルデータを蓄積する取り組みが，随所で盛んに行われている．蓄積された，いわゆるビッグデータから抽出・加工された大規模集計データの有効活用を図ることは，新たな産業分野の創出に向けての強力な足がかりとなる可能性が期待され，近年，大きな注目を集めている．例えば，2022 年 4 月から全面施行された「個人情報の保護に関する法律等の一部を改正する法律」[1] なども，ビッグデータを有効活用することの重要性，そしてその社会的潮流に呼応したものと言えるであろう．この見直しは，データ主体の権利保護強化，法令違反に対するペナルティの強化といった，有効活用に伴うリスクを低減するための保護強化という観点に加え，データの利活用促進（仮名加工情報に対する事業者の義務の緩和）という積極的な視点をも含んでおり，ビッグデータ有効活用という大きなうねりを視野に入れたものとも言えよう．しかしこれらのデータは，もともと人々の消費活動や，日常的な生活行動に基づくものであるであることから，大規模集計データの有効活用においては，将来発生するであろう脅威までをも見据えた高い安全性を有するプライバシー保護技術を適用することが，きわめて重要となってくる．

一方，さまざまな産業活動で計量・蓄積されている集計データを広く見渡すと，商品の数や生物の個体数，事象の発生件数や人口など，自然数を含む非負値から構成される集計データであることも少なくない．また，データが，人口密集地や商業地などのような固有の特性を有するエリアのみに集中するような，全体的には疎（スパース）な分布をとる傾向もしばしば見られ，そしてこの傾向は，特に，集計データの規模が大きくなるほど生ずる可能性が高まる．そこで本稿では，元のデータベースに含まれる個々のデータの集合体（個票）から，何らかの条件を満たすデータの個数を数えた数値データの集合体であり，さらに，全体的に疎な分布をとるような集計データを対象とする．

1.1 差分プライバシー

集計データに対するプライバシー保護に関しては，古くから検討が行われて来ている．これらは統計的開示制御（statistical disclosure control）[2], [3] と呼ばれ，そこではセル秘匿基準や $n-k\%$ 基準などに基づく各種の秘匿方式が専門家によって注意深く適用されており，長年にわたっ

[1] 北海道科学大学
Hokkaido University of Science
[2] 株式会社 NTT ドコモ
NTT DOCOMO Inc.

て安全性が確保されてきた [4], [5]．しかし，ビッグデータに基づく大規模集計データでは，値の小さな大量のセル値に対しても，切り捨てるのではなく，活用することが望まれる．そこで近年，プライバシーを保護しつつも有用なデータを有効に活用するための，新たな基準や手法に対する様々な研究が盛んに進められている．こうした技術はプライバシー保護データ公開（PPDP）と呼ばれ [6]，$k-$ 匿名性 [7] や $l-$ 多様性 [8]，$m-$ 不変性 [9] など，さまざまな基準や手法が提案されている．

しかし，これらの PPDP 技術で前提とするところの，攻撃者の目的や能力，保有知識はそれぞれ異なっており，安全性を統一的に議論することは難しい．そうした中，2006 年に Dwork らが提案した差分プライバシー基準 [10], [11] が，高い安全性を実現するための基準として注目を集めている．差分プライバシー基準は，ϵ-差分プライバシー基準とも呼ばれ，データベースへの問い合わせ (クエリ) を行った際に，「ある特定のデータがデータベースに含まれているか否かを問い合わせ結果から判別することが困難である」ことを安全性の根拠とするプライバシー保護基準である．

いま，ϵ をある小さな正の定数とし，データベース D に対して確率的問い合わせ M を適用したときに，問い合わせ結果として t が得られる確率を $Pr[M(D) = t]$ とする．ここで，任意の隣接する (たかだか 1 レコードしか異ならない) データベース D_1, D_2 に対して次式の関係が成立するとき，この問い合わせ M は ϵ-差分プライバシー基準を満たす．

$$\frac{Pr[M(D_1) = t]}{Pr[M(D_2) = t]} \leq e^{\epsilon}$$

この基準を満たす処理手法は，従来手法と異なり，背景知識や攻撃手法に依存しない数学的な安全性を備えることが保証されていることから，その実現方法に関する様々な研究が行われている．

例えば最近では，差分プライバシー（DP）に基づいて集計データに秘匿処理を施す従来手法よりも脅威にさらされる領域が限定的な，局所差分プライバシー（LDP）に関する研究が盛んに行われている [12], [13], [14], [15]．これは，集計前の個々のデータに対して数学的根拠に基づいた確率的ノイズを付加するものであり，データの集計機関に生データを渡す必要が無いことから，集計機関が信頼できない場合にも適用が可能な手法である．また，局所差分プライバシーに基づくことで，個々のデータが秘匿された状態で集計機関に提供されることとなり，悪意の第三者が何らかの手段を用いて，集計機関との通信，あるいは，集計機関が保有する個々の提供データを盗聴することができたとしても，プライバシーの保護を保証できるという点でも優れた手法と言える．

一方で，駅の自動改札機の位置や有料道路のゲートの位置，携帯電話端末の位置に関する情報のように，現行システムにおいて集計機関（事業者）側がデータを既に保有しているようなケースも存在する．これらの情報は，決済や通信といった本来のサービスを実現する上で事業者側がもともと取得する必要がある情報であり，ノイズを加えたデータに基づいてサービスを提供することは，そもそもできないものである．このようなサービスモデルにおいては，局所差分プライバシーを適用するのではなく，事業者側で集計を行ったデータに対して差分プライバシーを適用する方法を適用することとなる．そして例えば携帯電話端末の位置情報などは，新型コロナウイルス感染症（COVID-19）拡大時期以降，繁華街や観光地などにおける人出を観測する手段として広く活用されており，社会的役割や存在意義が広く認知されている．すなわち，集計データに対する差分プライバシーの適用を検討することには，いまだ十分な意義があると考えられる．

この差分プライバシー基準を満たす代表的な手法に Laplace メカニズムがある．この手法は，データベースへの問い合わせ結果に対して，平均値が 0 の Laplace ノイズ（Laplace 分布に従う独立な乱数）を付加するものである．適用対象が集計データの場合には，単に集計データに含まれる各セル値に対してそれぞれ Laplace ノイズを付加すれば良い．Laplace 分布の確率密度 $\ell(x)$ は，平均値 μ とスケール λ を用いて次式で与えられる．

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

ここで，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 $Lap(\lambda)$ とし，k 個の互いに独立な $Lap(\lambda)$ から成るスカラベクトルを $Lap(\lambda)^k$ と記すこととする．Laplace メカニズムにおける Laplace ノイズのスケール λ は，パラメータ ϵ と，問い合わせの種類ごとに決まる大域的感度（global sensitivity, GS）によって与えられる．具体的には，GS_f を問合せ f の感度としたとき，f に対応するランダム化関数は次式で表される．

$$f(X) + Lap(GS_f/\epsilon)^d$$

$$GS_f = \max_{D_1, D_2} \parallel f(D_1) - f(D_2) \parallel_1$$

ここで D_1 および D_2 は任意の隣接したデータベースのペアであり，d はスカラベクトルデータ X の要素数を表すものとする．V が分割表，すなわち，構成する部分集合が互いに素であるとき，計数間問い合わせの大域的感度は 1 であることが知られている [16], [17]．したがって，集計データの各セルにスケール $\lambda = 1/\epsilon$ の Laplace ノイズを加えることで，ϵ-差分プライバシを満たすことができる．

1.2 非負精緻化を伴う Privelet 法

差分プライバシーのさまざまな適用手法が提案される中，

Xiao らにより Privelet 法 [18] と呼ばれる，新たな差分プライバシーの適用方式が提案された．Xiao らが提案した二分木に基づく Privelet 法は，長さ $n = 2^H$ のスカラベクトルデータ $V = (v_0, \ldots, v_{n-1})$ に対して Haar 基底に基づく離散 Wavelet 変換 (HWT)$\mathcal{H}$ を導入し，その Wavelet 係数に対して Laplace メカニズムを適用した上で，逆 HWT $\mathcal{H}^{-1}$ 処理を施すことにより，差分プライバシー基準を満たすスカラベクトルデータ V^* を得るという手法である．この手法は，原理的に，部分和精度に優れており，たとえば，ある高解像度の空間に配置された集計値の部分和としてより大局的な空間配置の集計値を求めるような処理を行っても精度の低下が発生しないことから，実用面での利用価値が高いと考えられる．しかし，この方法で得られた集計データは，他の差分プライバシー適用手法と同様，ノイズ付加の影響により，本来負値が存在しないという，非負制約を逸脱する．さらに，もともとゼロ値であった大量のデータが非ゼロ値となるため，疎データの密度急増も伴うことになる．

一方，この Privelet 法を発展させた，非負精緻化を伴う Privelet 法 [17] の場合には，Privelet 処理中の HWT により得られた Wavelet 係数に対し，Laplace メカニズムの適用，および，非負精緻化を伴う逆 HWT の適用を通じて，差分プライバシを満たすスカラベクトルデータ V^+ を得ている．この手法では，オリジナルの Privelet 法に非負精緻化処理を組み込むことで，非負制約の逸脱を回避するとともに，疎データの密度急増を抑制している．以下に，非負精緻化が組み込まれた逆 HWT について，**図 1** を用いて説明する．

HWT により得られた係数ベクトルの各要素は，図 1 のように 2 分木の各ノードに対応させて配置することができる．これらのうち，HWT の出力 W として保存されているのは，ひとつの近似係数 $cA_{H,0}$ および $n-1$ 個の詳細計数 $cD_{h,x}$ である．ここで，x は階層内におけるノードの位置座標（$0 \leq x \leq n/2^h - 1$）を，h はノードの階層（$1 \leq h \leq H$）を，そして，H はツリー（ノード層）の高さ（階層数）をそれぞれ表している．そしてこの段階では，Laplace メカニズムの適用により $cA_{H,0}$ および $cD_{h,x}$ には Laplace ノイズが付加されており，それぞれ $cA^*_{H,0}$ および $cD^*_{h,x}$ へと値が変化している．近似係数 $cA_{h-1,2x}$ や $cA_{h-1,2x+1}$ の値は，

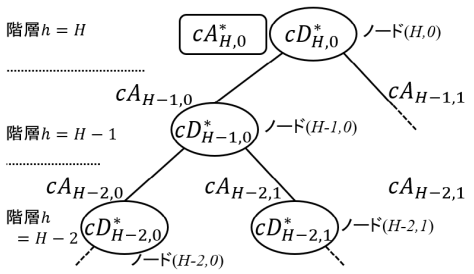

図 1 逆 Haar Wavelet 変換

それぞれ二分木に沿って対応するノード $(h-1, 2x)$ および $(h-1, 2x+1)$ に集約される入力データの平均値であるから，HWT への入力データが非負値であった場合には，本来，必ず非負値となるべきである．しかし，Laplace メカニズムの適用により，$cA^*_{h-1,2x}$ や $cA^*_{h-1,2x+1}$ へと値が変化した結果，これら近似係数の値が負値となってしまう場合も生じ得る．上述した通り，これらの値はともに，二分木に沿って対応するノードに集約される入力データの平均値であるから，これらの値が負値となることで，出力データに多数の負値が発生する事態を招く可能性がある．そこで，$cA^*_{h-1,2x}$ や $cA^*_{h-1,2x+1}$ の値に負値が生じないよう，一階層上の $cD^*_{h,x}$ の値を，符号を維持したまま $|cD^*_{h,x}| = cA^*_{h,x}$ となるように精緻化する．この精緻化処理により，全ての階層において $cA_{h,x}$ の値が非負値となり，出力値である $n = 2^H$ 個のベクトルデータ $V^+ = (v^+_0, \ldots v^+_{n-1})$ も全て非負値となる．なお，本稿では精緻化処理の結果得られた値には，$+$ を付して表す．

Privelet 法はもともと一次元データの処理を前提として提案された技術であるが，実社会の問題に対して実装する場合には，例えば地理空間上に分布した各種の集計データのように，しばしば二次元状に分布していることも少なくない．そこで二次元データに適用可能な形での Privelet 法の拡張が望まれる．

Privelet 法の多次元化については，Xiao らが提案している，標準分解（Standard decomposition）をベースとする手法がある [18]．この手法は，標準分解を行ってそれぞれの次元について一次元 Privelet 法を適用するものである．しかしこの手法の場合，大域的感度（GS）の値が大きくなってしまい，その利用が現実的ではないという問題があった．そこで我々は，非負精緻化を伴う Privelet 法の開発にあたって，単に Xiao 等が提案している二次元 Privelet 法に対して非負精緻化を適用するのではなく，標準分解を行わずに，局所性保存写像の一種である Morton 写像を用いて二次元データを一次元化する手法を採用することにより GS の増大を抑えた [17]．

しかしながらこの手法の場合には，秘匿対象のデータを一旦一次元データ配列に変換した上でプライバシー保護処理を適用し，その上で，処理された一次元データを，改めて二次元データへ戻すという，事前および事後処理が必要となる．やはり，二次元データを，次元変換する手間や時間を必要とすることなく，直接処理できる手法が望まれる．そしてその手法は，標準分解による GS の増大を招かない手法であることが望ましい．

そこで我々は，非負精緻化を伴う Privelet 法を，GS の増大を招く標準分解を行わずに二次元化する試みを行った．本稿では，我々が行った二次元化の方法を紹介するとともに，一次元方式との比較を通じて，二次元方式における非負精緻化の効果や秘匿データの精度，演算処理速度につい

て評価・考察を行う．

2. 非負精緻化を伴う二次元 Privelet 法

2.1 二次元ツリー構造

いま，**図 2** に示す 2×2 基本構造からなる二次元 Haar Wavelet を採用し，**図 3** に示す二次元ツリー構造を構成する．階層 $h = 1$ には，秘匿対象となる二次元集計データ $V = (v_{0,0}, \ldots, v_{x,y}, \ldots, v_{X-1,Y-1})$ が入力され，一方，演算処理後の各階層のノードには演算過程で得られた Wavelet 係数 ($cA_{h,x,y}$ または $cD_{h,x,y}$) が格納される．ここで，h（$1 \leq h \leq H$）は階層位置を，x, y（$0 \leq x \leq \sqrt{n} - 1$ かつ $0 \leq y \leq \sqrt{n} - 1$）は階層内におけるノードの位置座標を，$X = Y$ は X 軸座標・Y 軸座標それぞれに関する二次元データのサイズを，そして，$n = X \cdot Y$ はデータの総数をそれぞれ表している．

図 2 2×2 基本構造

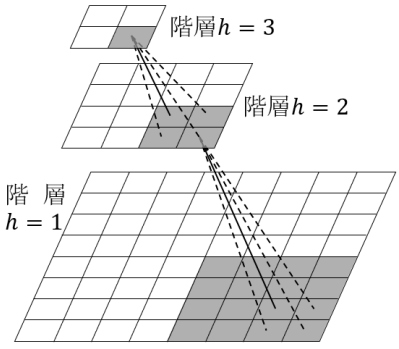

図 3 二次元ツリー構造

2.2 二次元 Wavelet 変換と逆変換

いま，階層が $h = (1, 2, 3, \ldots, H)$ のように表される H 階層のノードからなる二次元ツリー構造を考える．最初の Wavelet 変換 ($h = 1$) を行った際の近似係数 $cA_{h,x,y}$ および詳細係数 $cD_{h,x,y}$ の値は，次式で求められる．ここで，$x = (0, 1, 2, \ldots . X - 1)$ ならびに $y = (0, 1, 2, \ldots . Y - 1)$ は各階層における二次元ノードの位置を表す（但し，各階層のノードサイズは $X_h = Y_h = \sqrt{n}/2^{h-1}$ とする）．また，x, y の値は，図 2 の基本構造を踏まえて，2 ずつインクリメントする．

$$
\begin{aligned}
cA_{1,x,y} &= (v_{1,x,y} + v_{1,x+1,y} + v_{1,x,y+1} \\
&\qquad + v_{1,x+1,y+1})/4 \\
cD_{1,x+1,y} &= (v_{1,x,y} - v_{1,x+1,y} + v_{1,x,y+1} \\
&\qquad - v_{1,x+1,y+1})/4 \\
cD_{1,x,y+1} &= (v_{1,x,y} + v_{1,x+1,y} - v_{1,x,y+1} \\
&\qquad - v_{1,x+1,y+1})/4 \\
cD_{1,x+1,y+1} &= (v_{1,x,y} - v_{1,x+1,y} - v_{1,x,y+1} \\
&\qquad + v_{1,x+1,y+1})/4
\end{aligned}
\tag{1}
$$

続いて次階層 ($h = 2$) 以上では，各ノードに対して，上記式 (1)に準ずる次式の演算を行う．

$$
\begin{aligned}
cA_{h,x,y} &= (cA_{h-1,2x,2y} + cA_{h-1,2x+2,y} \\
&\qquad + cA_{h-1,2x,2y+2} + cA_{h-1,2x+2,2y+2})/4 \\
cD_{h,x+1,y} &= (cA_{h-1,2x,2y} - cA_{h-1,2x+2,y} \\
&\qquad + cA_{h-1,2x,2y+2} - cA_{h-1,2x+2,2y+2})/4 \\
cD_{h,x,y+1} &= (cA_{h-1,2x,2y} + cA_{h-1,2x+2,y} \\
&\qquad - cA_{h-1,2x,2y+2} - cA_{h-1,2x+2,2y+2})/4 \\
cD_{h,x+1,y+1} &= (cA_{h-1,2x,2y} - cA_{h-1,2x+2,y} \\
&\qquad - cA_{h-1,2x,2y+2} + cA_{h-1,2x+2,2y+2})/4
\end{aligned}
\tag{2}
$$

以後，上記式 (2)の演算処理を最上位層まで再帰的に繰り返すことで二次元 Wavelet 変換を実現できる．

一方，逆 Wavelet 変換の処理は，Wavelet 変換処理のプロセスを逆にたどる．

$$
\begin{aligned}
cA'_{h,x,y} &= cA_{h+1,x/2,y/2} + cD_{h,0,1} + cD_{h,1,0} \\
&\qquad + cD_{h,x+1,y+1} \\
cD'_{h,x+1,y} &= cA_{h+1,x/2,y/2} - cD_{h,0,1} + cD_{h,1,0} \\
&\qquad - cD_{h,x+1,y+1} \\
cD'_{h,x,y+1} &= cA_{h+1,x/2,y/2} + cD_{h,0,1} - cD_{h,1,0} \\
&\qquad - cD_{h,x+1,y+1} \\
cD'_{h,x+1,y+1} &= cA_{h+1,x/2,y/2} - cD_{h,0,1} - cD_{h,1,0} \\
&\qquad + cD_{h,x+1,y+1}
\end{aligned}
\tag{3}
$$

2.3 ノイズ付加

Wavelet 変換（順変換）の後，全ての詳細係数 $cD_{h,x,y}$ および最上位層の近似係数 $cA_{H,0,0}$ に対して確率分布 $\ell(x : \lambda'(h)) = (1/2\lambda'(h))e^{-x\lambda'(h)}$ に従う Laplace ノイズを付加する．階層 h における Laplace ノイズのスケールは，マトリクスメカニズム [17] に基づき，

$$
\lambda'(h) = \begin{cases} \frac{3 \cdot \lambda}{4^h} & \text{for } cD_{h,x,y} \\ \frac{\lambda}{4^h} & \text{for } cA_{H,0,0} \end{cases}
\tag{4}
$$

とする．ここで ℓ，x および $\lambda = (H + 1)/\epsilon$ はそれぞれ確率密度，確率変数，および，スケールを表す．なお，本稿では，ノイズ付加の結果得られた値には $*$ を付して表す．

2.4 非負精緻化

非負精緻化処理は，逆 Wavelet 変換処理の過程で負値の発生を排除する処理である．いま，h 層での逆 Wavelet 変換処理の結果得られた 3 つの詳細計数 $cD^*_{h,x+1,y}$, $cD^*_{h,x,y+1}$, $cD^*_{h,x+1,y+1}$ のうちのいずれか（複数もあり得る）に負の値が現れたら，次式に従って，$cD^*_{h,x+1,y}$, $cD^*_{h,x,y+1}$, $cD^*_{h,x+1,y+1}$ に対して非負精緻化処理を施し，精緻化後の詳細計数 $cD^+_{h,x,y}$ を用いて，改めて式 (3)に従って逆 Wavelet 変換の演算処理を行う．本稿では，非負精緻化の結果得られた値には $+$ を付して表す．

$$
\begin{aligned}
cD^+_{h,x+1,y} &= \beta \cdot cD^*_{h,x+1,y} \\
cD^+_{h,x,y+1} &= \beta \cdot cD^*_{h,x,y+1} \\
cD^+_{h,x+1,y+1} &= \beta \cdot cD^*_{h,x+1,y+1} \qquad (5) \\
\beta &= \frac{cA^*_{h,x,y}}{|Min(cD^*_{h,x+1,y}, cD^*_{h,x,y+1}, cD^*_{h,x+1,y+1}) - cA^*_{h,x,y}|}
\end{aligned}
$$

2.5 枝刈り

非負精緻化を伴う Privelet 法では，逆 Wavelet 変換プロセスに非負精緻化処理が付加されることによって，オリジナルの Privelet 法よりも演算効率が低下することが避けられない．この問題を解決するため，非負精緻化を伴う Privelet 法のための演算効率化手法が提案された [17], [20]．非負精緻化を伴う Privelet 法では，逆 Wavelet 変換の際に，あるノードの値が 0 になると，そこに連結した下層のノード値は全て 0 となるため演算を省略できる．この「演算の省略」により効率化を 図る手法が枝刈りである．枝刈りの手法として，水平型実装法 [19] と垂直型実装法 [20] が提案されているが，ここではより効率性の高い垂直型実装法を採用する．これは，逆 Wavelet 変換処理を行う際に，階層ごとに処理を行うのではなく，ひとつの枝に沿って Tree の深さ方向に演算を進めて行き，0 値となるノードが現れた時点で，そこから下層の演算を省略するという手法である．

3. 評価

3.1 評価方法

ここでは，異なる複数のエリアにおける人口分布データに対して二次元 Privelet 法による秘匿処理を適用し，「非負制約の逸脱の回避」，「データ密度急増の抑制」，「部分和精度劣化の抑制」，「演算効率化処理の効果」について評価する．評価には，平成 22 年度国勢調査に基づく地域メッシュ人口（1km メッシュ）のデータを使用した．差分プライバシーを規定するパラメータの値は $\epsilon = 0.1$ とした．

評価の実施にあたっては，日本全国（2^{11} メッシュ $\times$ 2^{11} メッシュ）のデータから切り出した，(1) 関東（$2^8 \times 2^8$），(2) 四国（$2^8 \times 2^8$），ならびに，(3) 北海道 1/4（$2^8 \times 2^8$）を使用した．(3) 北海道 1/4 は，切り出した北海道（$2^9 \times 2^9$）エリアに対し，他の 2 エリアとサイズを合わせる目的で，隣接する縦横 2×2 メッシュの人口をひとまとめにして生成したものである．

比較用となる一次元方式での処理を行うにあたっては，二次元上に配置された地域メッシュ人口データを一次元化する手法として，Morton 変換を使用した．Morton 変換は，多次元空間から一次元空間への全単射を行う写像を構成し，もとの空間上における距離の遠近が写像先の空間における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．

評価には Intel Core i7-9700 CPU (3GHz)，実装メモリ 16GB のデスクトップ PC を使用した．処理時間を計測する際には，同一処理を 10,000 回繰り返した時間を計測して 1/10,000 し，計測時間の精度向上を図った．

3.2 評価結果

3.2.1 非負制約逸脱の回避とデータ密度急増の抑制

まず，二次元 Privelet 法が，一次元方式と同様に，非負制約の逸脱を回避する特性と，データ密度の急増を抑制する特性とを有しているか否かについて評価する．**図 4** に，秘匿処理前後におけるメッシュ人口のヒストグラムを示す．グラフの横軸はメッシュ人口の値を，縦軸は度数を表している，各グラフはそれぞれ，(a) 秘匿処理前のデータ，(b) 非負精緻化なしの場合の秘匿結果，(c) 非負精緻化ありの場合の秘匿結果である．ここには関東のデータのみを示すが，全ての地域で同様の結果が得られている．

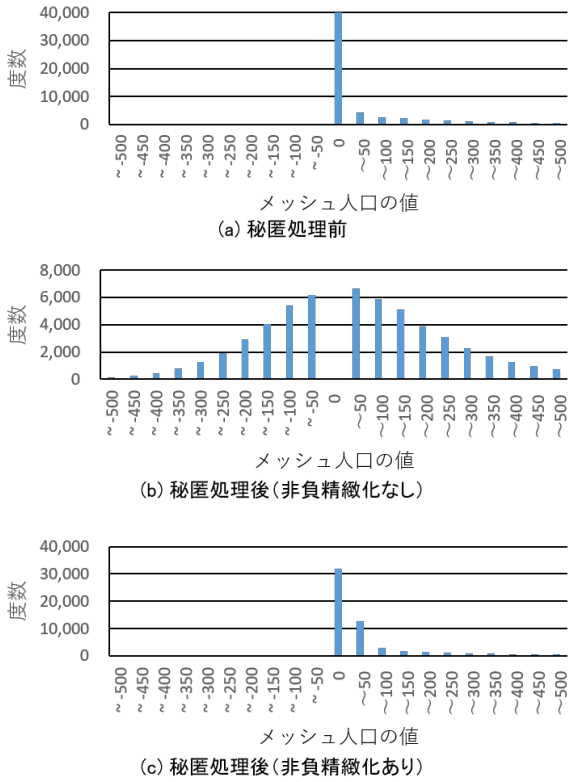

図 4 秘匿処理前後におけるメッシュ人口ヒストグラム（関東）

(a) のグラフを見ると，当然のことながら，もともとのメッシュ人口データには負値が存在しないこと，そして，ゼロ値が非常に多いスパースなデータであることがわかる．しかし，ノイズを加えて秘匿を行うことにより，(b) のグラフのように，大量の負値が発生するとともに，ゼロ値のないデンスなデータとなってしまう．ところが，非負精緻化処理を導入することにより，(c) のグラフのように，負値データの発生が解消されるとともに，ゼロ値データの数に回復傾向が見られ，秘匿によるデータ密度の増大が抑制される．以上の結果から，本研究で採用した二次元 Privelet 法において，非負精緻化処理が，狙い通りに上記 2 つの機能を発揮していることがわかる．

3.2.2 部分和精度劣化の抑制

次に，部分和をとっても精度が低下せずに一定程度維持されるという Privelet の性質が，二次元方式においても維持されているか否か，さらに，非負精緻化の導入により部分和精度の低下を招いてはいないかについての評価を行う．**図 5** に，二次元 Privelet 法を用いた秘匿結果の部分和に対する誤差量の様子を示す．グラフの横軸は，隣接する 2×2 メッシュ位置データの和をとって 4 メッシュ分をひとまとめにする操作を必要回数繰り返して得られる部分和に対応させて，ひとつの位置に含まれる元データのメッシュ数 s を $\log_4 s$ として表している．一方縦軸には，各位置における秘匿前と秘匿後のデータの差の 2 乗の平均値の平方根である RMSE の値をとっている．

図 5(a) は，特に部分和サイズの小さい領域において，Privelet 法がもつ特徴である，部分和をとっても精度が一定程度維持される性質を，二次元方式も有していることを示している．さらに，図 5(b) からは，特に部分和サイズの小さい領域において，非負精緻化の導入が部分和精度の劣化を招くものではないことがわかる．

3.2.3 演算効率化処理の効果

非負精緻化を伴う Privelet 法では，非負精緻化処理の導入によって，非負精緻化を行わない場合よりも演算時間が増大する．そしてこの問題に対処するために，「枝刈り」と呼ばれる演算効率化手法が提案されている [17], [20]．ここではこの枝刈りによる演算効率化の効果について評価を行う．

図 6 に二次元 Privelet 法の演算処理に要する時間を示す．縦軸は対象とするエリアを，横軸は演算時間の値をそれぞれ表している．「非負精緻化なし」は非負精緻化を用いない二次元 Privelet 法の演算時間を，「枝刈りなし」は枝刈り処理を用いずに非負精緻化処理のみを導入しだ場合の演算時間を，また，「枝刈りあり」は非負精緻化に加えて枝刈り処理をも導入して演算の効率化を図った場合の演算時間をそれぞれ表している．グラフから，いずれのエリアにおいても，非負精緻化の導入により演算時間が増加すること，および，枝刈りによる演算効率化によって，非負精緻化導入前とほぼ同程度まで演算時間が抑制されていることがわかる．

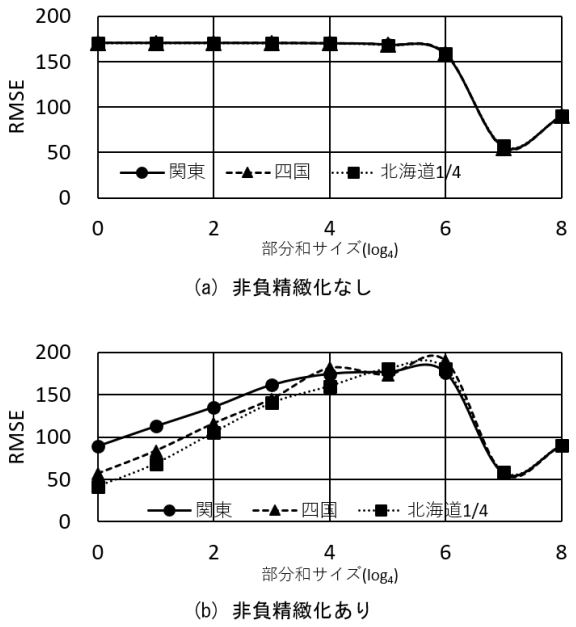

図 5 二次元 Privelet 法による秘匿結果の RMSE 値

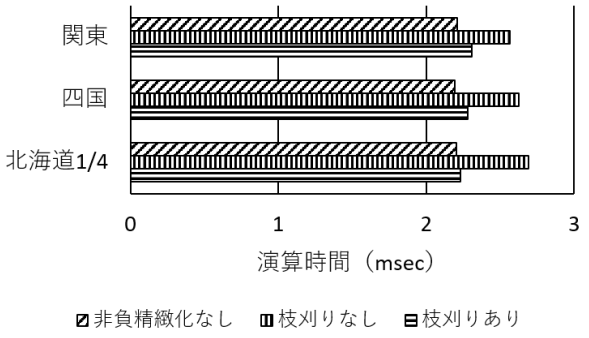

図 6 演算時間

4. 考察

第 4 章では，非負精緻化を伴う二次元 Privelet 法の有用性について，一次元方式との比較を通じて，部分和精度，ならびに，演算処理効率（時間効率）の観点から考察する．

4.1 部分和精度

まず，部分和精度の違いに着目する．**図 7** に一次元方式，および，二次元方式における，部分和に対する誤差量の様子を示す．グラフにおいて，横軸は部分和サイズを，また，縦軸は誤差量（RMSE）の値を表している．

「(a) 非負精緻化なし」のグラフを見ると，部分和サイズが小さい領域においては，一次元方式に比して，二次元方式の誤差量の方が約 30%ほど高く，秘匿処理後のデータの精度が低いことがわかる．一方で，部分和サイズの大

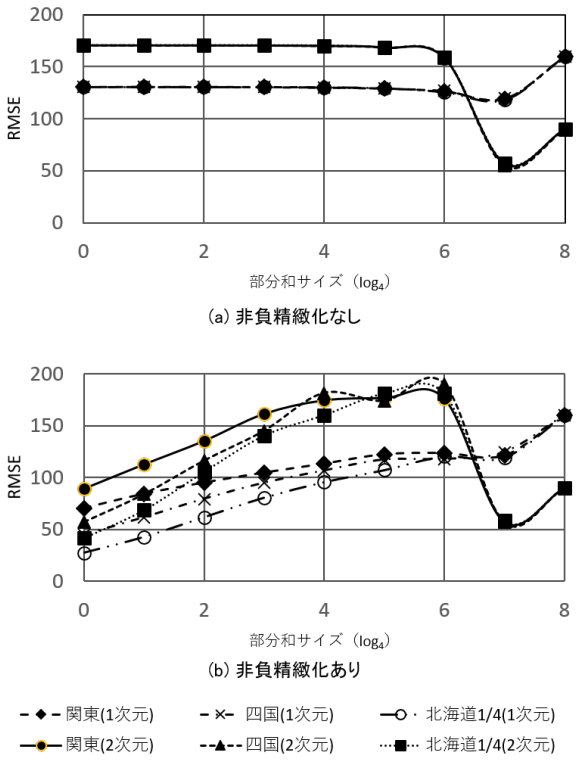

図 7 部分和サイズに対する誤差量

きい領域では，逆転する現象が起きている．こうした特性は，詳細係数 $cD_{h,x,y}$ と近似係数 $cA_{H,0,0}$ とにそれぞれ加えられているノイズの大きさの違いに起因していると考えられる．

Privelet 法の場合，部分和における誤差の大きさは，部分和サイズの小さい領域においては詳細係数 $cD_{h,x,y}$ に加えられているノイズの寄与が支配的であり，部分和サイズの大きな領域に行くほど近似係数 $cA_{H,0,0}$ に加えられているノイズの寄与が支配的となる．二次元方式の場合，一次元方式に比べて，$cD_{h,x,y}$ にはより大きいノイズが (約 1.5 倍程度)，そして逆に $cA_{H,0,0}$ にはより小さいノイズが（約 0.5 倍程度)，それぞれ加えられている．したがって，(a) のグラフに見られる，一次元方式と二次元方式における誤差値の違いは，木構造の違いと，この $cD_{h,x,y}$ と $cA_{H,0,0}$ に加えられているノイズの大きさの違いに起因するものと考えられる．

「(b) 非負精緻化あり」のグラフにおいても，(a) の場合と同様に，二次元方式の誤差量の方が，一次元方式よりも大きいことがわかる．なお，グラフにおける部分和サイズが 0（$= log_4 1$）の誤差値は，部分和をとらない，もともとのメッシュサイズにおける誤差値であることから，二次元方式で秘匿されたデータは，一次元方式で処理した場合よりも部分和精度の点で劣っていることがわかる．

4.2 演算処理時間

次に，一次元方式と二次元方式の処理に要する時間に着目する．**図 8** に一次元方式および二次元方式の処理に要する演算時間を示す．グラフにおいて，縦軸は一次元方式，および，二次元方式それぞれについて，対象とするエリアを表しており，横軸は，演算時間を表している．

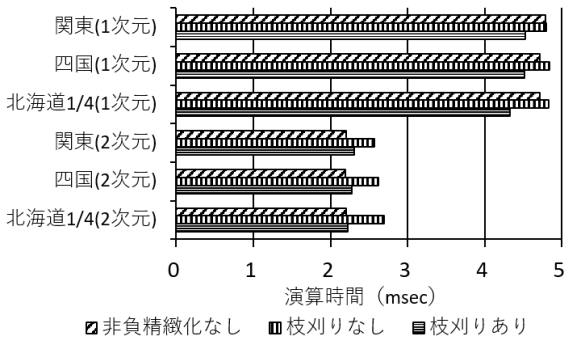

図 8 一次元方式・二次元方式に要する演算時間

グラフから，二次元方式の処理時間が，一次元方式の処理時間よりも少ないことがわかる．非負精緻化なしの場合には，二次元方式の方が，約 55%ほど演算時間が短くなっている．また，演算効率化手法である枝刈りを行わない非負精緻化を伴う Privelet 法においては約 45%ほど，枝刈りを行う非負精緻化を伴う場合には約 50%ほど二次元方式の演算時間が，一次元方式に比して短いことがわかる．

また，一次元方式の場合には，ここに示した処理時間に加えて，二次元から一次元への変換を行うための前処理と，一次元から二次元への変換を行うための後処理とが必要になる．実験に使用した 256×256 メッシュサイズのデータを対象として計測してみたところ，両方の処理で，合計約 5.4[msec] ほどの変換時間を要することが確認された．以上の結果から，非負精緻化を伴う二次元 Privelet 法は，処理速度の点において，一次元方式よりも優れていることがわかる．

5. おわりに

本稿では，本来一次元データを対象とする「非負精緻化を伴う Privelet 法」を二次元化する試みについて述べた．まず，第 1 章において非負精緻化を伴う Privelet 法の概要を説明した上で，第 2 章において二次元化を行う上での基本的な考え方，ならびに，方法について紹介した．特に，本検討の対象とした二次元化方式については，第 2 章において，基本となる二次元ツリー構造，二次元 Wavelet 変換と逆変換，ノイズ付加，および，枝刈りに分けて詳細な説明を行った．

続いて，コンピュータ上に実装した実験用プログラムを用いた評価実験について，第 3 章で紹介した．評価実験では，まず，非負精緻化処理によって，「非負制約逸脱の回避」，ならびに，「データ密度急増の抑制」が実現されていることを示した．また，同じく非負精緻化処理によりもた

らされる効果である，「部分和精度劣化の抑制」が実現されていることも示した．さらに，二次元方式に合わせて実装した枝刈りが，非負精緻化処理による演算時間の増大を抑えることも併せて確認した．

そして第 4 章では，非負精緻化を伴う二次元 Privelet 法の有用性について，一次元方式との比較を通じて考察を行った．一次元方式との比較は，「部分和精度」，ならびに，「演算処理時間」を対象として行った．比較の結果，二次元方式が，秘匿後データの誤差の少なさ（精度）の点では一次元方式よりも劣っていること，ならびに，演算処理速度の点では一次元方式よりも優れていることが明らかとなった．

秘匿データの精度については，詳細計数と近似係数との間におけるノイズレベルの振り分けかた（ϵ の分配問題）に関連する問題であると考えられる．今後，この分配の最適化手法を明らかにすることが必要である

謝辞　本評価の初期段階において協力をしていただいた石井貴己氏に感謝する．本研究は日本学術振興会科学研究費補助金基盤研究 (C)（課題番号：19K11970）による助成を受けて行なわれたものである．

参考文献

[1] 個人情報保護委員会：令和 2 年 改正個人情報保護法について, 個人情報保護委員会 (Web)：入手先 <https://www.ppc.go.jp/personalinfo/legal/kaiseihogohou/>(参照 2022-05-05).

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

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

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

[5] 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).

[6] Fung, B., Wang, K., Chen, R. and Yu, P.: Privacy-preserving Data Publishing, in *ACM Computing Surveys*, Vol. 42, pp. 1–53 (2010).

[7] 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).

[8] Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M.: l-diversity: Privacy Beyond k-anonymity, *ACM Trans. Knowledge Discovery from Data (TKDD)*, Vol. 11, No. 1 (2007).

[9] Xiao, X. and Tao, Y.: m-Invariance: Towards Privacy Preserving Re-publication of Dynamic Datasets, in *Proc. 2007 ACM SIGMOD Intl. Conf. Management of Data*, pp. 689–700, ACM (2007).

[10] Dwork, C.: Differential Privacy, in *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).

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

[12] Kasiviswanathan, S. P. e. a.: What Can We Learn Privacy?, *SIAM Journal on Computing*, Vol. 40, No. 3, pp. 793–826 (2011).

[13] Acharya, J. e. a.: Test without Trust: Optimal Locally Private Distribution Testing, *Priceedings of Machine Learning Research*, Vol. 89, pp. 2067–2076 (2019).

[14] Duchi, J. M. I., J. C. and WainWright, M. J.: Local Privacy and Statistical Minimaxrates, *IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 429–438 (2013).

[15] al., et. J. M.: The Role of Interactivity in Local Differential Privacy, *IEEE 60th Annual Symposium on Foundations of Computer Science*, pp. 94–105 (2019).

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

[17] 寺田雅之, 鈴木亮平, 山口高康, 本郷節之：大規模集計データへの差分プライバシの適用, 情報処理学会論文誌, Vol. 56, No. 9, pp. 1801–1816 (2015).

[18] Xiao, X., Wang, G. and Gehrke, J.: Differential Privacy via Wavelet Transforms, in *Proc. 26th Intl. Conf. Data Engineering (ICDE 2010)*, pp. 225–236 (2010).

[19] 本郷節之, 寺田雅之, 鈴木昭弘, 稲垣潤：非負精緻化をともなう Privelet 法の演算効率化手法, 情報処理学会論文誌, Vol. 61, No. 2, pp. 474–485 (2020).

[20] 本郷節之, 寺田雅之, 鈴木昭弘, 稲垣潤：非負精緻化をともなう Privelet 法における演算効率化手法の性能向上, 情報処理学会論文誌, Vol. 61, No. 9, pp. 1458–1471 (2020).