大規模集計データへのzCDPの適用

石岡 卓将[1,a)]　寺田 雅之[1,2]

概要：企業や公的機関におけるデータ活用や合理的根拠に基づく政策立案（EBPM: Evidence-Based Policy Making）の推進など，データに基づく社会・産業の最適化の重要性はますます高まっている．一方で，攻撃技術の進化により従来のプライバシー保護技術の陳腐化も進んでおり，その対策が急務である．この問題に対し，任意の攻撃に対して数理的な安全性を与えるプライバシー保護フレームワークである差分プライバシーが注目されている．しかしながら，大規模高次元データに対して適切な安全性強度を保証しつつ差分プライバシーを実用的に適用することは容易ではない．そこで本論文では，安全性定義にレニー情報量を取り入れた差分プライバシーの拡張の一種である，Zero-Concentrated Differential Privacy（zCDP）に着目し，大規模高次元データのプライバシー保護に対して zCDP を導入することによる改善効果を，国勢調査に基づく人口統計データを用いて定量的に評価した．評価の結果，単純な Laplace メカニズムと zCDP を満たす最も単純なメカニズムである Gaussian メカニズムの比較では，zCDP の導入による改善効果が見られず，かえって出力データの精度が悪化した．一方で，NN-Wavelet などの複雑でかつ大規模データへの適用に適したメカニズムに対しては，zCDP の導入により顕著な改善効果を得られることが明らかになった．

キーワード：差分プライバシー，レニー情報量，zCDP

On Publishing Large Tabular Data with Zero-Concentrated Differential Privacy

Takumasa Ishioka[1,a)]　Masayuki Terada[1,2]

Abstract: The importance of data-driven optimization in society and industry is increasing, as evidenced by the promotion of data utilization in businesses and public institutions and evidence-based policymaking (EBPM). Conversely, conventional privacy protection techniques are becoming obsolete due to the evolution of attack techniques, and measures to address this issue are urgently needed. To address this issue, differential privacy, a privacy protection framework that provides mathematical security against arbitrary attacks, has gained attention. However, applying differential privacy to large-scale, high-dimensional data while ensuring appropriate security strength is a complex and challenging task requiring innovative solutions. This paper focuses on Zero-Concentrated Differential Privacy (zCDP), an extension of differential privacy that incorporates Renyi Divergence into its security definition. We quantitatively evaluated the improvement effects of introducing zCDP for privacy protection of large-scale, high-dimensional data using population statistics based on census data. The evaluation results showed that when comparing the simple Laplace mechanism with the Gaussian mechanism, which is the most straightforward mechanism satisfying zCDP, introducing zCDP did not improve the output data's accuracy but decreased it. On the other hand, the introduction of zCDP resulted in significant improvements for complex mechanisms suitable for large-scale data applications, such as the non-negative wavelet method.

Keywords: Differential Privacy, Renyi Divergence, zCDP

1. はじめに

近年，企業や公的機関におけるデータ活用や，合理的根拠に基づく政策立案（EBPM: Evidence-Based Policy Making）の推進など，データに基づく社会・産業の最適化の重要性がますます高まっている．大規模かつ詳細なデータの分析は，効率的な資源配分，的確な政策決定，そして新たな価値創造の基盤となっている．特に，国勢調査データのような大規模高次元の集計データは，社会の実態を把握し，将来を予測する上で極めて重要な役割を果たしている．しかしながら，このようなデータの活用には常にプライバシー保護の課題が付随する．近年の攻撃技術の急速な進化により，従来のプライバシー保護技術の陳腐化が進んでおり，より強固な保護手法の開発が急務となっている．

この問題に対し，任意の攻撃に対して数理的な安全性を与えるプライバシー保護フレームワークである差分プライバシー（DP: Differential Privacy）が注目を集めている [1]．差分プライバシーは，データベースに対するクエリの結果に制御された確率的ノイズを加えることで，個々のデータ主体の情報を保護しつつ，全体的な統計的特性を保持する．従来の k-匿名性基準 [2] などのプライバシー保護基準と異なり，差分プライバシーでは未知の攻撃に対して汎用的に安全性を担保することができる．

大規模高次元データに対して適切な安全性強度を保証しつつ差分プライバシーを実用的に適用することは容易ではない．特に，データの有用性とプライバシー保護のバランスを取ることが極めて困難になるという問題が存在する．本研究では，この問題に対する新たなアプローチとして，Zero-Concentrated Differential Privacy（zCDP）に着目した．zCDP は，安全性定義にレニー情報量を取り入れた差分プライバシーの拡張の一種である．zCDP は，従来の ϵ-差分プライバシー（ϵ-DP: ϵ-Differential Privacy）よりも柔軟なプライバシー保証を提供し，特に繰り返し計算や複合的なデータ分析において優れた性能を発揮することが期待されている．

本稿では，大規模高次元の集計データに対して，zCDP を用いた効果的なプライバシー保護メカニズムを開発し，その性能を理論的および実験的に評価する．具体的には，以下の点に焦点を当て評価を行う．

- 非負 Wavelet（NN-Wavelet: Non-Negative Wavelet）を，zCDP の枠組みで再解釈し改良する．
- 実際の大規模高次元データセットを用いて，提案手法の有効性を検証する．
- 改良された手法と既存手法との比較を行う．

[1] 京都橘大学 工学部
Faculty of Enginering, Kyoto Tachibana University
[2] (株) NTT ドコモ モバイルイノベーションテック部
Mobile Innovation Tech Department, NTT DOCOMO, Inc.
[a)] ishioka@tachibana-u.ac.jp

本論文の構成は以下の通りである．まず第 2 節で従来技術について概説し，差分プライバシーの基本概念と大規模集計データに対する既存の手法を紹介する．第 3 節では zCDP によるプライバシー保護の理論的基礎を説明し，Gaussian メカニズムについて述べる．第 4 節で本研究の提案方式である NN-Wavelet への zCDP 導入について詳述し，その実装方法を示す．第 5 節では提案方式の有用性評価として，実験設計と結果を示す．最後に第 6 節で結論と今後の課題について述べる．

2. 関連研究

2.1 差分プライバシーの安全性定義

2.1.1 ϵ-DP

ϵ-DP は，個々のデータの存在や不在が統計的クエリの結果に大きな影響を与えないことを保証する数学的概念である [3]．この概念は，データベース D とその隣接データベース D'（D から 1 つのレコードを追加または削除したもの）に対して適用される．ランダム化アルゴリズム $\mathcal{M}$ が ϵ-DP を満たすとは，任意の出力集合 S に対して以下の条件を満たすことである．

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} \cdot \Pr[\mathcal{M}(D') \in S] \qquad (1)$$

ここで，ϵ はプライバシーパラメータである．ϵ が小さいほど強力なプライバシー保護を意味する．

ϵ-DP は複数の重要な性質を持つ．合成定理により，複数のメカニズムを組み合わせた場合のプライバシー保証が定量化できる．後処理不変性は，保護されたデータの柔軟な利用を可能にする．グループプライバシーにより，小規模なグループに対するプライバシー保護の程度も定量化できる．これらの性質により，ϵ-DP は強力かつ柔軟なフレームワークとして機能し，ϵ パラメータを統計的仮説検定の枠組みに基づいて定量的に解釈することができるため，多様なアプリケーションに適用されている [4]．

一方で，ϵ-DP には重要な課題が存在する．最も顕著な例は，プライバシー保護の強度とデータの有用性のトレードオフである [5]．強力な保護を要求すると，データの有用性が大きく損なわれる可能性がある．このトレードオフを適切に管理するには，データの生成プロセスや相関関係を考慮し，アプリケーションの特性に応じてプライバシー保護メカニズムを慎重に選択する必要がある [6,7]．また，同じデータセットに対する多数のクエリ実行により，プライバシー保証が急速に劣化する．これは，合成定理がプライバシー損失を線形に蓄積することに起因する [8]．複数のクエリを組み合わせる場合，プライバシー予算を適切に配分する必要がある．さらに，高次元データやスパースデータへの適用も課題となっている．データの次元増加に伴う必要なノイズ量が増大や，スパースデータにおいて不必要に大きなノイズが追加される可能性があるため，結果の有

用性が低下する [9].

2.1.2 Zero-Concentrated Differential Privacy

zCDP は、レニー情報量を用いてプライバシー保証を定量化する手法として注目を集めている．zCDP の基礎となるレニー情報量は，二つの確率分布間の差異を測る尺度であり，確率分布 P と Q の間のオーダー $\alpha(\alpha > 0, \alpha \neq 1)$ のレニー情報量 $D_\alpha(P||Q)$ は以下で定義される．

$$D_\alpha(P||Q) = \frac{1}{\alpha - 1} \log \left(\sum_x P(x)^\alpha Q(x)^{1-\alpha} \right) \quad (2)$$

ランダム化メカニズム $\mathcal{M}$ が ρ-zCDP を満たすとは，任意の隣接データセット D, D' と任意の $\alpha \in (1, \infty)$ に対して以下を満たすことである．

$$D_\alpha(\mathcal{M}(D)||\mathcal{M}(D')) \leq \rho\alpha \quad (3)$$

ここで，$D_\alpha(\cdot||\cdot)$ は α のレニー情報量を表し，$\rho > 0$ はプライバシーパラメータである．

zCDP はレニー情報量を用いてプライバシー損失を測定する．この方法により，zCDP は複数のクエリの組み合わせに対してより柔軟な評価を可能にする．zCDP の合成定理では，プライバシー損失の累積は平方根に比例するため，ϵ-DP と比較してより緩やかである．この特性により，zCDP は大規模集計データや複雑な分析タスクにおいて，より効率的なプライバシー保護を実現する．

zCDP の根本的な目的は，ϵ-DP と同様に，個人のデータがクエリ結果に与える影響を制限することである．しかしながら，これらのメカニズムは，プライバシー損失の測定と累積の方法において実践的な違いを示す．ϵ-DP は，各クエリに対するプライバシー損失の上限を明確に定義する．この特性により，単一のクエリに対しては直感的で強力な保証を提供する．しかし，複数のクエリを組み合わせる際，ϵ-DP はプライバシー損失を線形的に累積させるため，特に繰り返しクエリに対して過度に悲観的な見積もりをもたらす可能性がある．zCDP は，プライバシー損失の累積が劣線形的となるため，特に繰り返しクエリを含む複雑なシナリオにおいて，プライバシー保証と有用性のバランスをより適切に取ることができる．より具体的には，zCDP では ϵ-DP よりもプライバシー損失の累積を正確に反映して，プライバシーリスクを増大させることなくデータの有用性を向上させることができる．

zCDP を実現する主要手法のひとつが，Gaussian メカニズムである．Gaussian メカニズムでは，クエリ結果に Gaussian ノイズを加えることでプライバシーを保護する．具体的には，与えられた関数 f に対して，以下で示されるノイズ付与を行う．

$$M(x) = f(x) + N(0, \sigma^2) \quad (4)$$

ここで，$N(0, \sigma^2)$ は平均 0，分散 σ^2 のガウシアン分布からのノイズを表す．Gaussian メカニズムの重要な特性として，その zCDP との自然な適合性が挙げられる．感度 Δf のクエリに対して標準偏差 σ の Gaussian ノイズを加えるメカニズムは，以下の関係を満たす．

$$\rho = \frac{\Delta f^2}{2\sigma^2} \quad (5)$$

この関係は，Gaussian メカニズムが ρ-zCDP を満たすことを示している．したがって，所望の ρ-zCDP を達成するために必要な σ の値は以下の式で決定される．

$$\sigma = \frac{\Delta f}{\sqrt{2\rho}} \quad (6)$$

この関係式を通じて，プライバシー保証レベル ρ と関数 f の感度 Δf に基づいて，適切なノイズ量を設定することが可能となる．

Gaussian メカニズムは従来の Laplace メカニズムと比較して，高次元データや複数のクエリを組み合わせる場合により適していると考えられる．これは，Gaussian ノイズの特性と中心極限定理に基づく統計的性質に起因する．具体的には，独立な確率変数の和の分布は，サンプル数が増加するにつれてガウス分布に近づくため，高次元データや複数クエリの合成では，Gaussian ノイズがより自然な選択となる．感度が高い場合や複数のクエリが合成される場合において，Gaussian メカニズムはより少ないノイズ量で所定のプライバシー保証を達成できる可能性がある．しかしながら，zCDP と Gaussian メカニズムの組み合わせがすべてのシナリオで最適であるわけではない．例えば，感度が低いクエリやデータセットが小さい場合，Laplace メカニズムがより適切な選択となることがある．また，特定の高次元データにおいてはノイズが集中し，データの有用性が損なわれる可能性もある．データの性質やプライバシー要件に応じて適切なメカニズムを選択することが重要である．

2.2 大規模集計データへの差分プライバシー適用

大規模集計データに対する差分プライバシーの適用は，プライバシー保護と統計的有用性の両立を目指す重要な研究領域である．本節では，本研究が対象とする大規模集計データへの差分プライバシー適用における主要なアプローチについて詳述する．

2.2.1 Top-down 方式

Top-down 方式では，最も粗い地理的レベルから始め，徐々により細かい地理的レベルへとデータを処理していく階層的なアプローチを採用している．2020 年の米国国勢調査局（USCB: United States Census Bureau）が行った国勢調査 [10, 11] では，まず国全体のレベルから処理を開始し，州，郡，センサストラクトなど，より小さな地理的単位へと段階的に進める形で Top-down 方式によるプライバ

シー保護を実現した．各レベルでのノイズ追加後，上位レベルとの整合性を維持するための調整が行われ，これにより全体的なデータの一貫性が保たれる．具体的な実装として，国レベルの集計から始め，差分プライバシーノイズを追加する．次に州レベルの集計にノイズを追加し，国レベルの集計と整合するよう調整する．この過程を郡レベル，センサストラクトレベルと，順次下位の地理的単位に対して繰り返し適用する．最終的に，個人レベルのミクロデータを再構築する．

Top-down 方式の利点として，データの階層的な性質を保持しつつ，各レベルでプライバシー保護を適用できる点が挙げられる．上位レベルの集計値の精度を高く保ちながら，下位レベルでより強力なプライバシー保護を適用できるため，高い柔軟性がある．Top-down 方式では結果が非負の整数値となるよう制約を課すことで，現実世界のセンサスデータの性質の保持して，人口データの特性を反映することが可能となる．一方で，Top-down 方式では，下位レベルのデータ精度が上位レベルに比べて低下する可能性がある．さらに，異なる地理的レベル間でのデータの一貫性を維持することが技術的に困難な場合がある．

2.2.2 Haar Wavelet 変換

Haar Wavelet 変換（HWT）は，データの多重解像度解析を可能にする数学的手法である．その基本原理は，データを隣接する要素のペアに分割し，各ペアの平均（近似係数）と差の半分（詳細係数）を計算することにある．例えば，データ列 [4, 6, 3, 5] に対して，HWT は最初のステップで $[(4+6)/2, (3+5)/2, (4-6)/2, (3-5)/2]$ より，データ列 [5, 4, -1, -1] を生成する．HWT は可逆性，局所性，線形性，疎表現という重要な性質を備えている．可逆性により元のデータを完全に再構成でき，局所性によりデータの局所的な変化が変換後の係数に局所的な影響しか与えない．線形性は，入力データの線形結合に対する HWT の出力が，個々の入力の HWT の線形結合に等しいことを意味する．また，多くのデータセットにおいて，HWT は多くの要素が 0 または非常に小さい値となる疎表現を生成する．これらの特性は，プライバシー保護への応用において重要な利点をもたらす．特に，局所性，線形性，疎表現の特性により，少量のノイズ追加で効果的なプライバシー保護が可能となり，データの有用性を維持しつつプライバシーを保護できる．

2.2.3 NN-Wavelet

NN-Wavelet は，HWT の特性を活かした手法である [12]．NN-Wavelet では，入力データへの HWT 適用後，変換後係数への Laplace ノイズ付加によって差分プライバシーを達成する．さらに，非負制約を満たすように再帰降下的な各係数への補正を加えながら逆 HWT を適用することで，非負制約を逸脱しない出力データを生成する．

NN-Wavelet の理論的性能は，プライバシー保証と精度の両面で特徴づけられる．プライバシー保証に関しては，NN-Wavelet は Privelet と同様に ϵ-DP を満たす．これは，ノイズ追加と調整過程が非ゼロ要素のみに限定されているにもかかわらず，数学的に全要素処理と等価であることが証明されているためである．精度面では，NN-Wavelet はスパース性を活かした効率的な処理を行うことで，Privelet よりも優れた性能を示す．具体的には，最上位レベルでの非ゼロ要素に対応する係数へのノイズ付与と非負化，各レベルでの非ゼロ要素に関連する詳細係数のクリッピング，そして逆変換過程での非ゼロ再構成値の調整を通じて，スパース性と非負性を保証しながら元のデータ分布との乖離を抑制している．NN-Wavelet の実装上の特徴は，データの階層構造と Wavelet 変換の性質を利用した効率的な処理にある．ゼロ要素に対する変換結果もゼロとなることを利用し，非ゼロ要素のみを処理することで計算効率を高めている．この手法により，本研究が対象とするセンサスデータをはじめとしたスパース性の高い大規模データセットに対して，精度と計算効率において特に有効な手法となる．

3. 提案方式

本研究では，NN-Wavelet に zCDP を導入する新たな手法を提案する．zCDP を適用することで，より柔軟で効果的なプライバシー保護メカニズムの実現を目指す．本研究では，ρ 値をプライバシー強度の主要なパラメータとして使用する．ρ 値が小さいほど強力なプライバシー保護が実現されるが，同時にデータの有用性が低下する可能性がある．

3.1 zCDP を用いた NN-Wavelet の実装

NN-Wavelet に zCDP を適用するため，アルゴリズムにいくつかの主要な修正を行う．まず，従来の Laplace メカニズムに対して，提案手法では Gaussian メカニズムを採用する．Gaussian メカニズムは，zCDP の枠組みに適したメカニズムであり，高次元データに対してより適切なノイズ付加を可能にする．次に，ρ に基づいて各 Wavelet 係数レベルに適切なノイズを生成する新たなノイズスケール計算関数を導入する．ノイズのスケールは後述の式により決定し，Wavelet 変換の階層構造に応じた効率的なプライバシー保護を実現する．各レベルの感度は，そのレベルの Wavelet 係数の最大変化量に基づいて計算される．さらに，zCDP の合成性に基づいて，Wavelet 変換の各レベルに対してプライバシー予算を配分する方法を導入して，全体的な精度を向上させることができる．

3.2 アルゴリズムの詳細

提案手法のアルゴリズムは，大きく分けて三つの段階で構成される．まず，入力データに対して HWT を適用する．次に，各 Wavelet 係数レベルに対して，zCDP に基づいて

計算されたガウスノイズを付加する．最後に，非負性を保証しながら逆 Wavelet 変換を行う．

具体的なアルゴリズムは以下のように定式化される．入力として，データベース V とプライバシーパラメータ ρ を受け取る．まず，入力データ V に HWT を適用し，Wavelet 係数 W を得る．

$$W = \mathcal{H}(V) \tag{7}$$

ここで，$\mathcal{H}$ は HWT を表す．

次に，データの階層数 k を取得する．ここで，k は Wavelet 変換の最大分解レベルを表し，入力データのサイズ n に対して $k = \log_2 n$ で与えられる．zCDP のプライバシー予算 ρ を各レベルに配分する．各レベル i $(i = 0, 1, ..., k)$ に対して，プライバシー設定に応じて以下のようにプライバシー予算 ρ_i を割り当てる．

$$\rho_i = \frac{\rho}{k+1} \tag{8}$$

各レベル i のノイズスケール σ_i は，プライバシー設定に応じて以下の式で計算される．

$$\sigma_i = \frac{\Delta_i}{\sqrt{2\rho_i}} \tag{9}$$

ここで，Δ_i はレベル i の感度であり，以下のように定義される．

$$\Delta_i = \frac{1}{2^i} \tag{10}$$

次に，各レベル i $(i = k, k-1, ..., 1)$ の係数に対して，スケール σ_i のガウスノイズを加える．最上位レベル $(i = k)$ の近似係数 cA_k と詳細係数 cD_k に対して以下のようにノイズを加える．

$$cA_k^* = cA_k + \mathcal{N}(0, \sigma_k^2)$$
$$cD_k^* = cD_k + \mathcal{N}(0, \sigma_k^2)$$

そして，非負性を保証するために以下の処理を行う．

$$cA_k^+ = \max(cA_k^*, 0)$$
$$cD_k^+ = \begin{cases} -cA_k^+ & \text{if } cD_k^* < -cA_k^+ \\ cA_k^+ & \text{if } cD_k^* > cA_k^+ \\ cD_k^* & \text{otherwise} \end{cases}$$

その他のレベル $(i = k-1, ..., 1)$ に対しては，非負制約を考慮しながら逆 HWT を適用する．cA_i^+ が 0 でない各 x に対して，以下の処理を行う．

$\rho \ \backslash \ \delta$	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}
10^{-1}	1.76	2.02	2.25	2.45	2.64	2.81
10^{-2}	0.54	0.62	0.69	0.75	0.81	0.87
10^{-3}	0.17	0.19	0.22	0.24	0.25	0.27

表 1: 各 (ρ, δ) 値に対応する ϵ 値

Table 1 Corresponding ϵ values for (ρ, δ)

$$cA_{i-1,2x-1}^+ = cA_i^+ + cD_i^+$$
$$cA_{i-1,2x}^+ = cA_i^+ - cD_i^+$$
$$p^* = cD_{i-1,2x-1} + \mathcal{N}(0, \sigma_{i-1}^2)$$
$$q^* = cD_{i-1,2x} + \mathcal{N}(0, \sigma_{i-1}^2)$$
$$cD_{i-1,2x-1}^+ = \begin{cases} -cA_{i-1,2x-1}^+ & \text{if } p^* < -cA_{i-1,2x-1}^+ \\ cA_{i-1,2x-1}^+ & \text{if } p^* > cA_{i-1,2x-1}^+ \\ p^* & \text{otherwise} \end{cases}$$
$$cD_{i-1,2x}^+ = \begin{cases} -cA_{i-1,2x}^+ & \text{if } q^* < -cA_{i-1,2x}^+ \\ cA_{i-1,2x}^+ & \text{if } q^* > cA_{i-1,2x}^+ \\ q^* & \text{otherwise} \end{cases}$$

最後に，cA_1^+ が 0 でない各 x に対して，以下の処理を行い，出力データ $V^+ = (v_0^+, v_1^+, ..., v_n^+)$ を得る．ここで，v_{2x-1}^+ と v_{2x}^+ は，元のデータ空間における隣接する 2 つの要素を表している．

$$v_{2x-1}^+ = cA_1^+ + cD_1^+ \tag{11}$$
$$v_{2x}^+ = cA_1^+ - cD_1^+ \tag{12}$$

この一連の処理により，zCDP に基づくプライバシー保護と非負性の保証を両立したデータ出力が実現される．

3.3 パラメータの選択とその影響

ρ-zCDP を満たすメカニズムは，任意の $\delta > 0$ に対して (ϵ, δ)-DP を満たすことができる．したがって，ρ と δ の組み合わせにより，従来の ϵ-DP と比較可能な ϵ に変換することができる．ここで，ϵ は ρ と δ を用いて以下のように変換される．

$$\epsilon = \rho + 2\sqrt{\rho \log(\frac{1}{\delta})} \tag{13}$$

この関係により，zCDP は従来の ϵ-DP の枠組みでも評価可能であるプライバシー保証を提供する．

表 1 に，式 (13) に関して ρ 値と対応する ϵ 値を示す．提案手法では，主に ρ 値を調整してプライバシー強度を制御する．本稿では，δ 値をプライバシー保護メカニズムが ϵ-DP のプライバシー保証を超えて，プライバシー損失が発生する確率の上限を表す値であるとする．

ρ や ϵ，δ の値の選択は，プライバシー保証に影響を与える．ρ 値が大きい場合，全体的に緩やかなプライバシー保

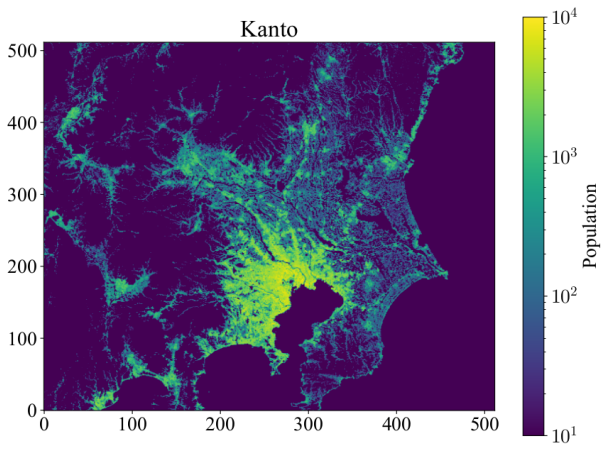


図 1: 首都圏周辺のエリアの人口統計データ

Fig. 1 Demographic data for areas surrounding the Tokyo metropolitan region

護が提供され，ρ が小さい場合には，大多数のケースでは強力なプライバシー保護を提供する．$\delta = 0$ の場合は純粋な ϵ-DP に相当し，強力なプライバシー保証を提供する．一方，$\delta > 0$ の場合は，δ 値に応じた確率で，ϵ-DP の保証を超えるプライバシー損失が許容される．通常 δ 値は極めて小さな値（例えば 10^{-6}）に固定する．これにより，大多数のケースで予測可能なプライバシー保護を提供しつつ，極端なプライバシー損失のリスクを最小限に抑えることができる．δ 値を大きく設定した場合には，大きなプライバシー損失が発生することが許容されることになる．

zCDP の枠組みにおいて，ρ 値は，対象とするデータセットの特性やアプリケーションの要件に応じて慎重に選択する必要がある．例えば，非常に機密性の高いデータを扱う場合は，ρ をより小さく設定する必要がある．一方，データの有用性をより重視する場合は，やや大きめの ρ を許容することで，平均的なプライバシー保護レベルを維持しつつ，データの精度を向上させることが可能である．具体的な選択例として，$\delta = 10^{-6}$ を選択し，目的に応じて ρ を 10^{-2} から 10^{-1} の範囲で調整することが考えられる．このとき，ϵ 値に換算すると表 1 より，約 0.24 から 2.45 の範囲となり，従来の差分プライバシーの基準と比較しても十分な保護強度を確保できる．ただし，ρ 値の最適な選択は，データの性質，利用目的，求められるプライバシー保護の強度など，様々な要因に依存する．そのため，実際の適用においては，これらの要因を慎重に考慮し，適切なパラメータを選択することが重要である．

4. 提案方式の有用性評価

本節では，提案方式の有用性を評価するための実験設計について詳述し，その評価結果を考察する．4.1 節では，実験に用いたデータセットの詳細と，評価指標および比較対象とした手法について説明する．4.2 節では，zCDP の適用が異なるメカニズムにもたらす影響を分析し，提案手法の性能を他の手法と比較しながら考察する．

4.1 実験設計

4.1.1 データセット

本評価では，データセットの例として，H22 国勢調査に基づく地域メッシュ統計 [13] を 500m メッシュに分割し，首都圏周辺のエリアの人口統計データを抽出したものを用いた．**図 1** に，有用性評価において対象した人口統計データを可視化したものを示す．データセットは，256km 四方の領域を対象とし，$512 \times 512 = 2^{18}$ のメッシュで構成されている．図 1 より，評価に用いたデータセットでは人口密度の高い都市部と，人口の疎な地域の対比が明確に見られる．疎な地域は、データの特性としてスパース性を生み出し、プライバシー保護メカニズムの適用において重要な考慮事項となる。

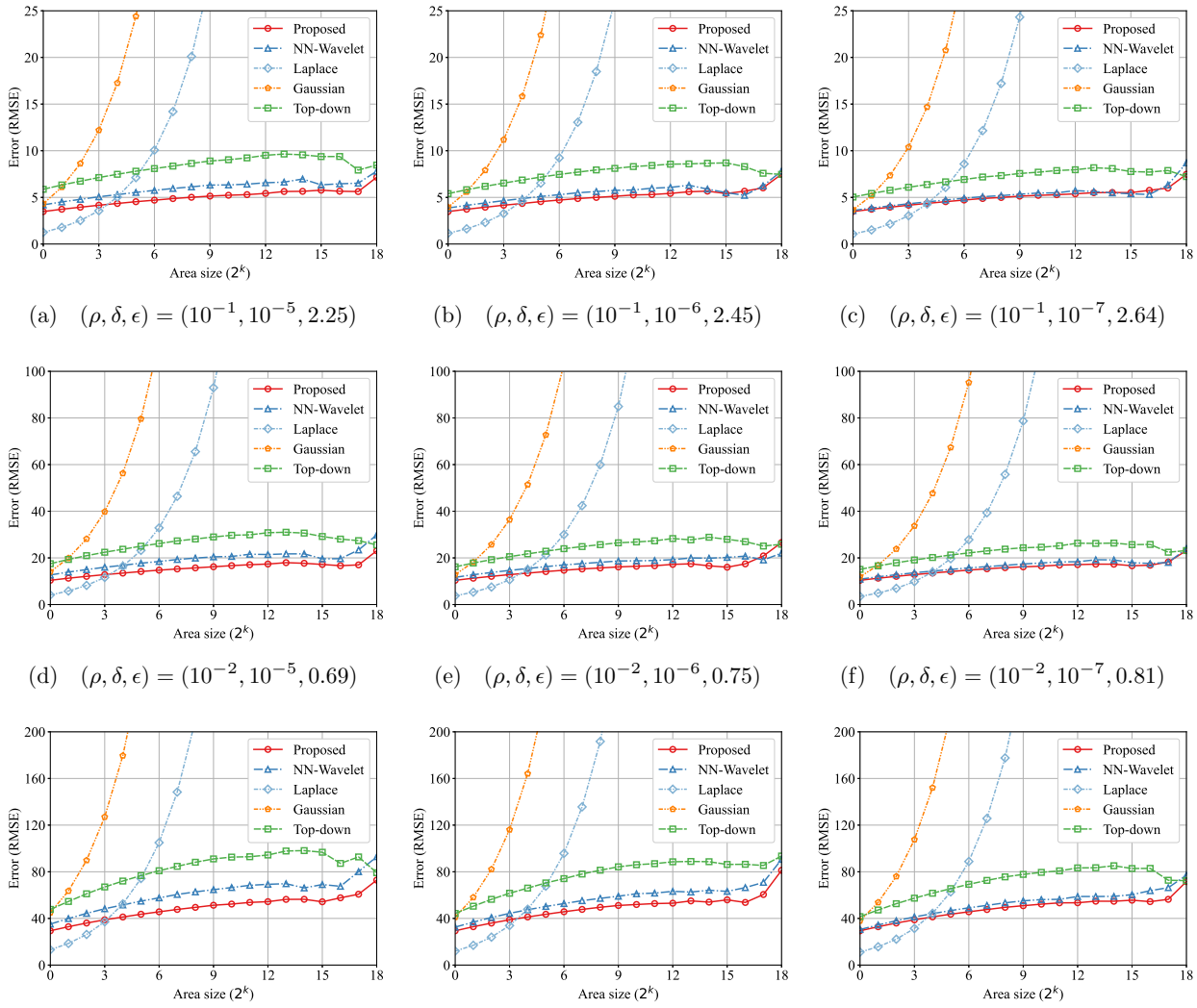
4.1.2 評価指標と比較手法

本評価では，プライバシーパラメータとして ρ 値を使用し，10^{-3}，10^{-2}，10^{-1} の 3 段階で設定した．δ 値は 10^{-5}，10^{-6}，10^{-7} の 3 段階で設定した．zCDP を用いない手法については，式 (13) より，ρ 値を ϵ 値に変換して使用した．評価指標として，二乗平均平方根誤差（RMSE: Root Mean Squared Error）を用いた．比較対象として，基礎的なメカニズムである Laplace メカニズムと Gaussian メカニズムに加えて，既存手法である NN-Wavelet，Top-down を用いた．提案手法は zCDP を用いた NN-Wavelet であり，NN-Wavelet は ϵ-DP を用いた手法である．Laplace メカニズムは ϵ-DP を満たす基礎的な差分プライバシー手法であり，Gaussian メカニズムは zCDP を満たす基礎的な差分プライバシー手法である．Top-down は，評価領域を再帰的に二分割しつつ処理を行う階層的なアプローチにより，ϵ-DP を満たす差分プライバシー手法である．各実験は 100 回繰り返し実行し，結果の平均を取った．

4.2 zCDP の適用による影響

図 2 に評価結果を示す．図 2 より，基礎的な手法である Laplace メカニズムと Gaussian メカニズムの比較では，zCDP を満たす Gaussian メカニズムの精度がより低くなることが確認された．両メカニズムともエリアサイズに対して誤差が線形に増加する傾向を示したが，Gaussian メカニズムの誤差は Laplace メカニズムよりも常に大きくなった．複雑なアプローチをとらない単純なメカニズムにおいては，zCDP を適用することで精度が悪化することが確認された．

一方，zCDP を用いた提案手法と NN-Wavelet との比較では，異なる傾向が確認された．図 2 より，δ が大きいほど，同程度の ϵ-DP に対して精度が向上した．この傾向は，zCDP がプライバシー損失の累積をより正確に捉えること



(a) $(\rho, \delta, \epsilon) = (10^{-1}, 10^{-5}, 2.25)$ (b) $(\rho, \delta, \epsilon) = (10^{-1}, 10^{-6}, 2.45)$ (c) $(\rho, \delta, \epsilon) = (10^{-1}, 10^{-7}, 2.64)$

(d) $(\rho, \delta, \epsilon) = (10^{-2}, 10^{-5}, 0.69)$ (e) $(\rho, \delta, \epsilon) = (10^{-2}, 10^{-6}, 0.75)$ (f) $(\rho, \delta, \epsilon) = (10^{-2}, 10^{-7}, 0.81)$

(g) $(\rho, \delta, \epsilon) = (10^{-3}, 10^{-5}, 0.22)$ (h) $(\rho, \delta, \epsilon) = (10^{-3}, 10^{-6}, 0.24)$ (i) $(\rho, \delta, \epsilon) = (10^{-3}, 10^{-7}, 0.25)$

図 2: 異なる (ρ, δ, ϵ) 値における RMSE の比較

Fig. 2 Comparison of RMSE for different (ρ, δ, ϵ) values

ができることを反映している．δ 値を大きくすることで，極めて低確率でのプライバシー損失をより許容し，その代わりに全体的なノイズレベルを低減させることができるためである．zCDP がプライバシー損失をより正確に捉えることができ，同じプライバシー保証レベルをより少ないノイズで実現できることが確認された．

さらに，図 2 より，ρ が小さいほど，同程度の ϵ-DP に対して精度が向上しやすい傾向が見られた．これは，zCDP が低いプライバシー予算においてもノイズの分配を効率的に行えることを示している．zCDP は，プライバシー損失の累積を平方根のスケールで評価するため，特に低いプライバシー予算下での性能が優れている．zCDP が低いプライバシー予算においても効率的にノイズを分配できることが示された．

また，δ 値が小さくなるにつれて，zCDP の精度は同程度の ϵ-DP と同等となった．これは，δ 値が極めて小さい場合，zCDP と ϵ-DP の理論的な違いが縮小することを反映している．本稿において，より小さい δ 値は，プライバシー保証の失敗確率をより低く抑えることを意味する．したがって，δ を小さくするにつれて，zCDP の柔軟性による利点が減少したと考えられる．提案手法が各パラメータ設定の意義を正しく反映したプライバシー保証を実現していることが確認された．一方で，例えば図 2 (e) の最大エリアサイズにおいては，提案手法の精度が悪化した．この現象は，ノイズ付加の確率的性質に起因するものであり，特に大きなエリアサイズでは，累積されたノイズの影響がより顕著になることを示している．

zCDP の導入の本質的な価値は，より適切なプライバシー保証と有用性のバランスを実現することにある．このバランスは，特に複数のクエリや複雑なデータ分析タスクにおいて重要となる．zCDP は，プライバシー損失の累積をより正確に追跡し，必要以上にプライバシー予算を消費

せずに済むため，長期的または複雑なデータ利用シナリオにおいて優位性を発揮する．評価の結果，複数のパラメータ設定において，提案手法は従来手法と比較して同等以上の精度を実現することが確認された．この結果は，zCDP が単にノイズを減らすだけでなく，プライバシー保証と有用性のトレードオフをより効果的に管理できることを示している．特に，データの複雑な利用パターンや長期的なプライバシー保護が要求される状況下で，zCDP の利点がより顕著になると考えられる．zCDP が従来の ϵ-DP よりも様々なプライバシー要件に対する堅牢性と汎用性を持ち，柔軟なプライバシー保証を実現可能であることが示された．

5. 結論

本研究では，大規模集計データに対する zCDP の適用効果を，理論的および実験的に詳細に評価した．具体的には，既存手法に zCDP を導入して，その性能を従来の ϵ-DP の枠組みで比較した．評価の結果，NN-Wavelet などの Wavelet 変換を用いた複雑な手法に対して zCDP を導入することで，精度の向上が確認された．また，提案手法はエリアサイズの増加に伴う誤差の増加を効果的に抑制し，大規模データに対する優れた適用性を示した．zCDP が大規模高次元データのプライバシー保護において有効であることが確認された．

さらに，提案手法は異なる ρ 値，δ 値において，一貫して優れた精度を持つことが確認された．提案手法が様々なプライバシー要件に対して堅牢性と汎用性を持つことが示された．zCDP は，企業や公的機関におけるデータ活用や EBPM，疫学研究などの大規模なデータセットを扱う分野において，従来困難とされてきた個人情報の保護と統計的有用性の両立を実現する新たな解決策となる可能性がある．

今後の課題として，従来の ρ-ϵ 変換式の妥当性を実験的に検証し，必要に応じて改良することが挙げられる．この検証により、zCDP と ϵ-DP の間のより正確な比較が可能となる．また，zCDP の実用性評価として，特定の有意水準における検出力の向上度合いを定量的に評価する必要がある．具体的には，従来の ϵ-DP と比較して，zCDP がどの程度効果的に偽陽性を制御しつつ，偽陰性を減少させられるかを分析する．zCDP がプライバシー保護の強度を維持しつつ，より高い統計的検出力を達成できる可能性を探る．

また，δ 値の選択に関するより詳細な考察も必要である．δ は，プライバシー保護メカニズムが完全に失敗する確率の上限を表すパラメータであり，その適切な設定はプライバシー保護の実効性に大きな影響を与える．本研究では δ を固定値として扱ったが，データセットの特性やアプリケーションの要件に応じて，データセットのサイズや感度，想定される攻撃モデルなどを考慮に入れた δ の動的な決定方法を開発することで，より柔軟で効果的なプライバシー保護フレームワークの構築が可能になると考えられる．

謝辞　本研究は JSPS 科研費 JP24K23872 の支援の下で行った．

参考文献

[1] Dwork, C.: Differential privacy, *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II*, Berlin, Heidelberg, Springer-Verlag, pp. 1–12 (2006).

[2] Sweeney, L.: K-anonymity: A Model for Protecting Privacy, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10, No. 5, pp. 557–570 (2002).

[3] Dwork, C. and Roth, A.: The Algorithmic Foundations of Differential Privacy, *Foundations and Trends® in Theoretical Computer Science*, Vol. 9, No. 3–4, pp. 211–407 (2014).

[4] Kairouz, P., Oh, S. and Viswanath, P.: The Composition Theorem for Differential Privacy, *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, JMLR.org, pp. 1376–1385 (2015).

[5] Sarathy, R. and Muralidhar, K.: Evaluating Laplace Noise Addition to Satisfy Differential Privacy for Numeric Data, *Transactions on Data Privacy*, Vol. 4, No. 1, pp. 1–17 (2011).

[6] Kifer, D. and Machanavajjhala, A.: No free lunch in data privacy, *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, New York, NY, USA, Association for Computing Machinery, pp. 193–204 (2011).

[7] Kifer, D. and Lin, B.-R.: Towards an axiomatization of statistical privacy and utility, *Proceedings of the Twenty-Ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, New York, NY, USA, Association for Computing Machinery, pp. 147–158 (2010).

[8] Dwork, C., Naor, M., Pitassi, T. and Rothblum, G. N.: Differential privacy under continual observation, *Proceedings of the Forty-Second ACM Symposium on Theory of Computing*, New York, NY, USA, Association for Computing Machinery, pp. 715–724 (2010).

[9] Li, N., Qardaji, W. and Su, D.: On Sampling, Anonymization, and Differential Privacy Or, K-anonymization Meets Differential Privacy, *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security*, New York, NY, USA, Association for Computing Machinery, pp. 32–33 (2012).

[10] Abowd, J. M.: The U.S. Census Bureau Adopts Differential Privacy, *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, New York, NY, USA, Association for Computing Machinery, p. 2867 (2018).

[11] Abowd, J. M., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P., Machanavajjhala, A., Moran, B., Sexton, W., Spence, M. and Zhuravlev, P.: The 2020 Census Disclosure Avoidance System TopDown Algorithm, *Harvard Data Science Review*, No. Special Issue 2 (2022).

[12] 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).

[13] 総務省 統計局：地域メッシュ統計の特質・遠隔.