Vol.2016-CSEC-73 No.26
Vol.2016-IOT-33 No.26
2016/5/27

# 匿名化個票への差分プライバシー基準の適用に関する一考察

寺田 雅之[1]　山口 高康[1]　本郷 節之[2]

**概要**：個人に関わるデータの公開・提供にあたっては，開示されたデータから個人のプライバシーが漏洩することを防ぐ必要がある．本稿では，それらのデータを匿名化された個票データ (microdata) である匿名化個票として開示する際において，強い数学的な安全性が示されている差分プライバシー基準を充足しつつ，データの有用性を高く保つ方式を提案する．提案方式は，集計データの一種である完全分割表に対する Laplace メカニズムの適用と，ベクトル空間における最近傍探索に基づく非負制約，整数制約，総数制約の充足により匿名化個票を得る．また，提案方式の有用性を評価するために，売上履歴を模したロングテイル性を持つ擬似的な個票データを用い，$L_2$ ノルム，KS-距離を評価指標として従来方式と定量的に比較評価し，従来方式と比べて元データの性質をより強く保持する匿名化個票が得られることを示す．

# On Releasing Anonymized Microdata with Differential Privacy

Masayuki Terada[1]　Takayasu Yamaguchi[1]　Sadayuki Hongo[2]

## 1. はじめに

個人に関わるデータの利活用にあたっては，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，それらのデータを匿名化された個票データ (匿名化個票) の形態で開示する際において，開示データの有用性をなるべく高く保ちながら，強い数学的な安全性が示されている差分プライバシー基準を充足する方式について考察する．

あるシステムで収集・作成されたデータベースを，第三者に公開・提供することを考える．これをデータの開示という．データ開示の形態は，大きく「個票データ (microdata)」の開示と「集計データ (aggregated data)」の開示に大別される．個票データとは，個人を単位とした情報 (レコードと呼ばれる) の集合として開示されるデータである．集計データとは，それらのデータをなんらかの条件で集約し，その個数を数えた数値データ [*1] (セルと呼ばれる) の集合などとして開示されるデータである [1], [20].

これらのデータの開示は，開示されるデータに含まれる個々人のプライバシーを保護しつつ行なわれなければならない．しかし，個票データの開示にあたっては，その有用性を保ちながらプライバシーを十分に保護することは容易ではない．これは，個人を単位として表現されるデータから個人のプライバシーが暴露されることを防ぐという，そもそも原理的な困難が含まれるためである．

データ開示にあたってプライバシーを保護するための指標としては，$k$-匿名性 ($k$-anonymity) 基準 [14] や差分プライバシー (differential privacy) 基準 [2] などが挙げられる．$k$-匿名性は，特に日本で言及されることが多い指標であるが，単に $k$-匿名性を満たすだけではプライバシーを十分に保護できないことが知られており，数々の改善手法が提案されている [13], [16]．しかし，それらの「改善」を施すことにより，データの有用性が劣化することから，広く使われるには至っていない．

その一方，差分プライバシーは，その安全性が数学的に保証可能であるという特長を持ち，米国などにおけるプライバシー保護研究の分野で大きく注目を集めている．しかし，差分プライバシーは集計データのプライバシー保護に有用である一方で，個票データへの効果的な適用法は明らかではない．たとえば，差分プライバシーを満たすため

[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] 一般には，データの平均や総計など，他の統計量の場合もあるが，本稿では個数であるとする (詳細は第 2 章で定義する)．

に用いられる最も一般的な手法である Laplace メカニズム [4], [5] は，集計データの各セル値に対して Laplace ノイズを加算するのみという簡単な処理によって，加算したノイズ強度に従った強度の差分プライバシーを保証できるが，これをそのまま個票データのプライバシー保護に適用することはできない．

個票データの開示において差分プライバシーを保証するための従来手法としては，サンプリングと大域的再符号化 (global recoding) に基づく手法 [9] や，個票データの分布からの再サンプリングに基づく手法 [15]，統計分野における開示制御手法 [7] の一つである PRAM (post randomizaion)[6] に基づく手法 [5], [8], [11] などが挙げられる．しかし，これらの手法で得られるプライバシー保護とデータの有用性のトレードオフの関係はあまり芳しいものではない．すなわち，十分な強度でプライバシーを保護しようとすると，著しいデータの有用性の低下を招き，実用的なプライバシー保護の手段とは言えなくなる．

本稿では，上記の問題を解決し，差分プライバシーによる安全性を保証しつつ，かつ実用的な有用性を備えた個票データの開示をするための手法について検討し，属性を完全分割した高次元集計データを経由して個票データの差分プライバシーを保証する方式を提案する．

提案方式では，個票データの各属性の値域の直積をセル集合とした集計データ (完全分割表) は，元の個票データと可換であることに着目し，個票データに対してそのままプライバシー保護のための処理を施すのではなく，いったん上記の完全分割表を作成した上で，Laplace メカニズムにより差分プライバシーを満たす匿名化集計データを作成する．しかし，そのように作成された匿名化集計データは，そのままでは個票データに対応づけられない (個票データに戻すことができない)．そこで，ある集計データを個票データに「戻す」ためには，集計データが非負のセル値のみから構成されること (非負制約)，各セル値が整数を値としてとること (整数制約)，セル値の合計が個票データのレコード数と等しいこと (総数制約)，の 3 つの制約を充足すればよいことを示し，完全分割表に含まれるセルの値域から構成されるベクトル空間において，この制約を充足する上記の匿名化集計データからの最近傍点を探索することにより，差分プライバシーを満たす匿名化個票を生成する．

## 2. 準備

本章では，議論の準備として，個票データおよび集計データの定義を与えるとともに，差分プライバシー [2] の定義と，差分プライバシー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 個票データ

個票データとは，それぞれが個人に対応づけられた 1 つ以上のレコードから構成されたデータベースであり，各レコードは 1 つ以上の属性値を持つ．これは，各レコードを元とした多重集合 (multiset)[*2] として定義される．

ある個人 $i$ に対応づけられたレコードを $x_i$ とする．$x_i$ は，その個人に関する何らかの情報を表す，$d$ 個の属性値 $x_{ij}(1 \le j \le d)$ の組 (順序対) から構成される．任意のレコードにおける，$j$ 番目の属性値は集合 $A_j(1 \le j \le d)$ に属する $(\forall i, x_{ij} \in A_j)$．ここで，$A_j$ を属性と呼び，すべての属性の直積 $A = A_1 \times A_2 \times \cdots \times A_d$ を属性空間と呼ぶ．

このとき，$n$ 個のレコードから構成される個票データ $D$ は，属性空間 $A$ を台集合 (underlying set)[*3] とする，以下の $n$ 元の多重集合として表される．

$$\begin{aligned} D &= \{\{x_1, x_2, \ldots, x_n\}\}, \\ x_i &= (x_{i1}, x_{i2}, \ldots, x_{id}) \quad (\in A), \\ A &= A_1 \times A_2 \times \cdots \times A_d. \end{aligned} \tag{1}$$

ここで，$D$ を (集合や順序対ではなく) 多重集合として定義する理由は，個票データには同一の属性値の組み合わせを持つレコードが複数存在しうること，レコードの並び順のみが異なる個票データは本質的に等価であることによる．

$x_i \in A$ より，各レコードがとりうる属性値の組み合わせの数は，属性空間 $A$ の濃度 (cardinality) $|A|$ に等しい．これは，各属性 $A_j$ の濃度 $|A_j|$ の総積である．すなわち，

$$|A| = \prod_{j=1}^{d} |A_j|. \tag{2}$$

一般には，属性 $A_j$ は有限集合 (カテゴリ属性や上限/下限を持った自然数をとる数値属性など) もしくは無限集合 (実数を値としてとる数値属性など) のいずれもとりうる．本稿では，以降の議論において $A_j$ は有限集合であるとする．すなわち，(実数値をとる) 数値属性は，階級化などの処理により，有限なカテゴリとして表現されているものとする．このとき，$|A_j|$ は (有限な) 自然数となることから，$|A|$ もまた自然数となる．

### 2.2 集計データ

集計データは，個票データ $D$ において，ある定められた条件を満たす属性値 (もしくは属性値の組み合わせ) を持つレコードの個数を数えあげた値の集合である．

$A$ を $D$ の属性空間とするとき，$A$ の部分空間 $C_k(\subseteq A)$ に属するレコードの個数を $Count(D, C_k) = |\{x \in D \mid x \in C_k\}|$ とする．これを計数問い合わせ (count query) と呼ぶ．このとき，任意の $C_k$ からなる順序対である集計条件 $C = (C_1, C_2, \ldots, C_p)$ に対して，集計データ $V$ は，$C$ の各元に対応する計数問い合わせ $Count(D, C_k)$ からなる順序対として与えられる．すなわち，

[*2] 同値を持つ元が重複して存在することを許す (順序なし) 集合．
[*3] 多重集合の元が属する集合．

$$V = (v_1, v_2, \ldots, v_p), \quad v_k = Count(D, C_k). \qquad (3)$$

集計データ $V$ の作成において，一般的には各属性の値域 $A_j$ の互いに素な部分集合の直積が集計条件 $C$ として用いられる．このとき，集計データは分割表 (contingency table) と呼ばれる．分割表の各要素 $v_k$ を，セル (cell) もしくはセル値と呼ぶ．

### 2.3 完全分割表

$A = \{a_1, a_2, \ldots, a_p\}$ を属性空間とする集計データ $V$ において，$C_k = \{a_k\}$ $(1 \leq k \leq p)$，すなわち集計条件の各要素 $C_k$ は，属性空間 $A$ のいずれかの元のみを含む集合であり，重複なくすべての $A$ の元に対応づけられる (全単射を持つ) とする [*4]．この集計データは分割表であり，これ以上に集計条件を細かくした分割表は作れないことから，本稿ではこれを完全分割表と呼ぶ．完全分割表 $V$ のセル数は，$A$ の濃度 $|A|$ と等しい ($|V| = |A| = \prod_{j=1}^{d} |A_j|$)．

このとき，完全分割表 $V$ における各要素 $v_k$ は，(多重集合である) 個票データ $D$ における，(その台集合である) 属性空間 $A$ の元 $a_k$ の多重度 (multiplicity) $m_D(a_k)$ に他ならない．任意の多重集合は，台集合と，その各元の多重度により一意に定義されるため，$(V, A)$ の組が与えられれば $D$ は一意に定まる．すなわち，以下の定理が成立する．

**定理 1.** 属性空間 $A$ を持つ個票データ $D$ から完全分割表 $V$ を生成する写像を $f_A$ とする ($V = f_A(D)$)．このとき，$f_A$ は $D = f_A^{-1}(V)$ となる逆写像 $f_A^{-1}$ を持つ．

### 2.4 差分プライバシー

差分プライバシー [2], [3] は，識別不能性に基づくプライバシー基準の一種であり，パラメータ $\epsilon$ 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 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 $\epsilon$-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 $(S \subseteq \mathcal{R})$．

$$\Pr[\mathcal{K}(D_1) \in S] \leq e^{\epsilon} \cdot \Pr[\mathcal{K}(D_2) \in S]. \qquad (4)$$

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．

### 2.5 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．代表的なメカニズムとしては Laplace メカニズムが挙げられる．

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 ノイズを問い合わせ結果に加算することにより実現される．Laplace 分布の確率密度 $\ell(x)$ は，平均 $\mu$ とスケール $\lambda$ を用いて下式で与えられる．

$$\ell(x; \mu, \lambda) = \frac{1}{2\lambda} e^{(-|x-\mu|/\lambda)}. \qquad (5)$$

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

Laplace メカニズムにおけるノイズのスケール $\lambda$ は，定義 1 における安全性パラメータ $\epsilon$ と，問い合わせの種類ごとに定まる感度 (sensitivity) によって与えられる．具体的には，$S_f$ を問い合わせ $f : \mathcal{D} \to \mathbb{R}^q$ 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．

$$\mathcal{K}_f(X) = f(X) + \mathrm{Lap}(S_f/\epsilon)^q, \qquad (6)$$

$$S_f = \max_{D_1, D_2} |f(D_1) - f(D_2)|_1. \qquad (7)$$

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 1 参照) のペアである．

Laplace メカニズムを用いることにより，差分プライバシーを満たす集計データを簡単に作成することができる．特に $V(|V| = p)$ が分割表であるとき，計数問い合わせの感度 $S_{count}$ は 1 であることから，差分プライバシーの並列合成則により，

$$V^* = V + \mathrm{Lap}(1/\epsilon)^p \qquad (8)$$

により $\epsilon$-差分プライバシーを満たす集計データ $V^*$ を得る．

## 3. 従来技術と課題

前述の通り，集計データに対しては，Laplace メカニズムを用いることにより差分プライバシーを容易に保証することができる [*5]．しかし，Laplace メカニズムの適用には「問い合わせ」と，それにより定まる「感度」が定義できることが必要となるが，個票データにはそれらは存在せず，直接 Laplace メカニズムを適用することはできない．

[*4] このときの $a_k$ の並び順は，なんらかの規則 (辞書順など) によりあらかじめ定められているとする

[*5] 十分な有用性が得られるかどうかは別の話である．単に Laplace メカニズムを適用するだけでは十分な有用性を得られない場合における改善方式は近年盛んに議論されている．[10], [17], [18], [19]

そのため，差分プライバシーを保証した個票データを開示する従来技術は，主に再符号化やサンプリング，PRAM などの他のプライバシー保護手法を適用し，それが差分プライバシーを保証することを証明するアプローチを採っている．

Li ら [9] は，個票データからのランダムサンプリング結果に対して大域的再符号化と ($k$ を閾値とした) セル秘匿 [7] によって $k$-匿名性を実現することにより，差分プライバシーを満たす匿名化個票を作成する手法を示している．ここで，$k$-匿名性の実現手段として (データの内容に依存しない) 大域的再符号化を用いる必要があることに留意が必要である．一般的な $k$-匿名化手法では，有用性の劣化を抑制するために，データの内容に応じて適応的に $k$ 個のレコードを集約する (clustering and local recoding, CLR) 方法が採られるが，CLR は差分プライバシーを満たすことができないとされる．そのため，この手法では一般的な $k$-匿名化手法と比較して有用性が低くなる．また，大域的再符号化を伴うことにより，出力される匿名化個票の属性空間 $A'$ は，元の属性空間 $A$ に対し，より「丸められた」ものとなる ($|A'| < |A|$)．

Wasserman ら [15] は，元の個票データから得られる頻度分布に「似た」分布を作成し，その分布からのサンプリングにより差分プライバシーを満たす匿名化個票を得る方法を示している．ここで，元の頻度分布に「似た」分布を作成する手法として，スムージングに基づく手法，攪乱に基づく手法，指数メカニズム (exponential mechanism) に基づく手法の 3 種類を挙げている．ここで，攪乱に基づく手法は，基本的に Laplace メカニズム適用後の負値の切り捨てに相当すること，指数メカニズムに基づく手法は攪乱に基づく手法に劣ることが示されていることから，以下ではスムージングに基づく手法について説明する．

この手法では, 個票データ $D$ の頻度分布 $f(i) = v_i/n$ を定数 $\delta$ でスムージングしたヒストグラム $f^*(i) = (1-\delta)p(i)+\delta$ を作成し，そのヒストグラムを正規化した確率分布からの独立試行により得られた標本を匿名化個票とする．ここで，$\delta$ は 安全性強度を制御するパラメータとして機能する．試行の回数を $k$ とするとき，この匿名化個票は，

$$\epsilon = k \ln \left( \frac{(1-\delta)p}{n\delta} + 1 \right) \tag{9}$$

の $\epsilon$-差分プライバシーを満たす [*6]．

Lin ら [11] と Ikarashi ら [8] は，統計的開示制御手法の一つである PRAM が差分プライバシーを満たすことを示している [*7]．PRAM とは，個票データの各レコードに含まれる属性値を，ある定められた確率 (遷移確率) に従って他の (嘘の) 属性値に書き換えることによってプライバシーを保護する手法である．特に，この遷移確率が属性ごとに一様であるもの，すなわち，それぞれの属性値 $x_{ij}$ について確率 $\rho_j$ で属性値を維持し，確率 $1-\rho_j$ で属性値を $x'_{ij} \in A_j$ へと一様ランダムに置換するものを維持置換攪乱 (retention-replacement perturbation) と呼び，$\rho_j$ を維持確率 (retention probability) と呼ぶ．$d$ 次の属性を持ち $n$ レコードからなる個票データに対し，属性 $A_j$ に対応する維持確率を $\rho_j$ とする維持置換攪乱は，以下の $\epsilon$-差分プライバシーを与える．

$$\epsilon = \sum_{j=1}^{d} \ln \frac{1 + (|A_j| - 1)\rho_j}{1 - \rho_j}. \tag{10}$$

しかし，これらの手法が与える差分プライバシーの安全性強度はあまり高くない．言い換えると，十分な安全性を与えるようにこれらの手法を適用すると，有用な匿名化個票を出力できるとは言えなくなる．

以下，具体例を用いて定量的に議論する．たとえば，$A = A_1 \times A_2$, $|A_1| = |A_2| = 10$ という二次の属性空間を持つ，レコード数 100 の (簡単な) 個票データを考える．この個票データに対し，集計データに対する差分プライバシーの議論で (十分な安全性を持つとして) しばしば用いられる $\epsilon = 0.1$ の差分プライバシーを満たす，同数 (100 個) のレコードを持つ匿名化個票を得るために必要となるパラメータを具体的に計算する．

[15] のスムージングに基づく手法では，式 (9) の変形により，安全性パラメータ $\delta$ は以下のように導出される．

$$\delta = \frac{p}{(e^{\epsilon/k} - 1)n + p} = \frac{1}{(e^{0.001} - 1) + 1} \simeq 0.999. \tag{11}$$

これを $f^*(i)$ の定義に代入すると，匿名化個票を得るためのサンプリング元となる分布は，その 99.9% が定数部分となることがわかる．すなわち，ほとんど元の個票データの情報は残らず，一様分布からサンプリングされたようなデータが匿名化個票として出力されることになる．

次に，PRAM に基づく手法について議論する．$\epsilon = \sum_{j=1}^{d} \epsilon_j$ とするとき，式 (10) の変形により維持確率 $\rho_j$ について以下の式を得る．

$$\rho_j = \frac{e^{\epsilon_j} - 1}{|A_j| + e^{\epsilon_j} - 1}. \tag{12}$$

上記と同様に，$A = A_1 \times A_2$, $|A_1| = |A_2| = 10$ の個票データに対し，$\epsilon = 0.1$ の条件で維持置換攪乱を施すとする．このとき，$\epsilon_1 = \epsilon_2 = 0.05$ とするならば，式 (12) より維持確率 $\rho_1, \rho_2$ は以下の値をとる．

$$\rho_1 = \rho_2 = \frac{e^{0.05} - 1}{10 + e^{0.05} - 1} \simeq 0.005. \tag{13}$$

[*6] なお，頻度分布のスムージングとそこからのサンプリングが差分プライバシーを満たすことは，[15] 以前に Machanavajjhara ら [12] によっても示されている．

[*7] さらに [8] は，PRAM が ($k$-匿名性を一般化したプライバシー指標である) $Pk$-匿名性を満たすことを併せて証明しており，PRAM を媒介として差分プライバシーと $k$-匿名性との関係を定量的に示している．

すなわち，上記の個票データに対して維持置換攪乱により 0.1-差分プライバシーを満たそうとすると，約 0.5% の確率でしか元の個票データの属性値は維持されず，約 99.5% の確率で属性値がランダムに置き替わることを意味する．

## 4. 提案方式

前章で示したように，従来手法では個票データを差分プライバシーを満たすように加工すると，その有用性が大きく劣化する．そこで本稿では，個票データを直接処理するのではなく，完全分割表を経由して個票データに対する差分プライバシーを実現する方式を提案する．

提案方式は，個票データと可換な集計データである完全分割表に対して差分プライバシーを満たすためのノイズ付与を行なう．個票データに対して直接差分プライバシーを与える方式に比べ，(Laplace メカニズムを代表とする) 集計データに差分プライバシーを与える方式は，データの有用性を効率的に保つ特長を持つ．そのため，提案方式は従来方式に比べ，より高い有用性を備える出力が得られることが期待される．ただし，ノイズが付与された完全分割表は，個票データの形にはそのまま変換できない．そこで，個票データの形式に変換可能な完全分割表の条件と，その導出手段を与える．

### 4.1 処理の流れ

本方式は，以下に示す流れにより，個票データ $D$ から差分プライバシーを満たす匿名化個票 $D^+$ を作成する．

(1) 個票データ $D$ に対応する完全分割表 $V$ を作成する．
(2) $V$ に対して (Laplace メカニズムなど) 集計データに対して差分プライバシーを与えるメカニズムを適用し，完全分割表 $V^*$ を作成する．
(3) $V^*$ から個票データに対応づけられる完全分割表 $V^+$ を導出する．
(4) 最後に $V^+$ を対応する匿名化個票 $D^+$ に変換する．

提案方式において差分プライバシーを与えるためのメカニズムは Laplace メカニズムに限らない (たとえば，幾何メカニズム (geometric mechanism) を用いてもよい) が，以降では Laplace メカニズムの適用を例として議論を進める．

上記手順において，手順 (1) および (2) は，それぞれ式 (3) および式 (8) により容易に実現することができる．しかし，式 (8) により得られた $V^*$ は，そのままでは個票データに対応づけられない．これは，Laplace メカニズムによりノイズが加算されたセル値は，負の値や小数を取りうるため，$V^*$ に対応する個票データが存在するとは限らないためである．言い換えると，$V^*$ が $f_A^{-1}$ の定義域に属するとは限らず，このとき $f_A^{-1}(V^*)$ は値を取りえない．

そこで，手順 (3) として，上記の $V^*$ を入力として，そこから最も「近い」$f_A^{-1}$ の定義域に属する (個票データと可換な) 完全分割表 $V^+$ を導出した後に，手順 (4) で $D^+ = f_A^{-1}(V^+)$ により差分プライバシーが保証された匿名化個票 $D^+$ を生成する．

### 4.2 個票データと可換な完全分割表の条件

可換な個票データを持つ完全分割表 $V^+$ の導出にあたり，まず最初に $V^+$ が満たす必要がある条件を議論する．

第 2 章で議論したように，完全分割表 $V$ のセル値 $v_i(1 \le i \le p)$ は，個票データ $D$ の台集合 $A$ における各元の多重度 $m_D(a_i)$ $(a_i \in A)$ である．ここで，任意の多重集合において多重度の値域は非負整数であり，かつ，(台集合と) 多重度を与えることにより対応する多重集合が一意に定まることから，$V^+$ の要素 $v_i^+$ について，

$$\forall v_i^+ \in V^+, v_i^+ \in \mathbb{Z}, v_i^+ \ge 0 \qquad (14)$$

が成立すれば，$V^+$ は対応する多重集合を持つ ($A$ を併せて与えることにより個票データの形に戻すことができる)．すなわち，$V^+$ のすべての要素は非負の値を持ち (非負制約)，かつ整数でなくてはならない (整数制約)[*8]．以降，$V^+$ により与えられる個票データを $D^+$ と表記する．

また，個票データ $D$ のレコード数 $n$ は，$D$ の多重度の総和である $(n = \sum m_D(a_i) = \sum v_i)$．したがって，$D^+$ のレコード数を保存する ($D$ と同じレコード数にする) ためには，

$$\sum v_i^+ = n \qquad (15)$$

という束縛条件を併せて満たす必要がある (総数制約)．

これら 3 つの制約条件を合わせると，以下の定理を得る．

**定理 2.** $A$ を属性空間とする完全分割表 $V^+$ は，その要素 $v_i^+$ について非負制約，整数制約，総数制約をいずれも満たすとき，対応するレコード数 $n$ の個票データ $D^+$ を持つ．すなわち，

$$\exists D^+ \mid \{D^+ = f_A^{-1}(V^+),\ |D^+| = n\},\quad \forall v_i^+ \in V^+, v_i^+ \in \mathbb{Z}, v_i^+ \ge 0 \ \text{ s.t. } \sum v_i^+ = n. \qquad (16)$$

### 4.3 最近傍点の探索

定理 2 により，手順 (3) は，$V^*$ から最近傍にある (式 (16) の条件を満たす) $V^+$ を探索する問題に帰着される．ここで何が「最近傍」であるかは $(V^*, V^+)$ 間の距離の定義により異なるが，以下では $p$ 次元の実数ベクトル空間 $\mathbb{R}^p$ におけるユークリッド距離 (二次ノルム) を用いるとする．すなわち，以下の問題を解くことにより $V^+$ を得る．

$$V^+ = \underset{\Phi \in \mathbb{N}_0^p}{\arg\min} |V^* - \Phi|_2 \ \text{ s.t. } |\Phi|_1 = n. \qquad (17)$$

[*8] 二つの制約条件を合わせて「各要素は (0 を含む) 自然数である $(\forall v^+ i, v_i^+ \in \mathbb{N}_0)$」と一言で表すこともできるが，以降の議論のため非負制約と整数制約を分けて表現する．

幾何学的には，これは $p$ 次元の実数ベクトル空間 $\mathbb{R}^p$ において，$p-1$ 次元の超平面 $Z : \sum_{\varphi_i \in \Phi} \varphi_i = n$ 上に配置された $\mathbb{N}_0^p$ 上の格子点を候補点とし，$V^*$ から最近傍に存在する候補点を抽出する問題に相当する．

$n$ や $p$ が十分に小さいときは，これはすべての候補点に対する距離計算や，$k$d 木などを用いた探索など，一般的な最近傍探索手法により求めることができる．しかし，$n$ もしくは $p$ が大きくなると候補点の数は階乗オーダーで増加するため，計算量の観点からあまり実用的ではない．

式 (17) における候補点の数は，$D$ のレコード数 $n$ と $V^*$ のセル数 $p(=|A|)$ によって定まる．具体的には，$p$ 元からなる集合から (重複を許した) $n$ 元を選ぶ組み合わせ (重複組み合わせ) の数 ${}_pH_n$ となるため，

$$ {}_pH_n = {}_{p+n-1}C_n = \frac{(p+n-1)!}{n!(p-1)!}. \qquad (18) $$

すなわち，$O((p+n)!)$ の計算量オーダーとなる．$k$d-木を用いれば，時間計算量はその対数となるが，それでも現実的とは言えない．

ただし，候補点が超平面上の非負格子点として規則的に並んでいることを利用すれば，下記の手順により，より効率的に $V^+$ を探索することができる．

(1) $V^*$ から最近傍にある，超平面 $Z$ 上の点を求める．これは，$V^*$ から $Z$ への垂線の足に相当するため，$Z$ の法線ベクトル $\mathbf{z} = (1, 1, \ldots, 1)$ とパラメータ $s$ を用い，直線 $V^* + s\mathbf{z}$ と $Z$ との交点となる．これを $s$ について解くことにより下式を得る．

$$ \Phi_1 = V^* - \frac{(|V^*|_1 - n)}{p}\,\mathbf{z}. \qquad (19) $$

ここで $\Phi_1$ は $Z$ 上にあるため，総数制約を満たす．

(2) $\Phi_1$ から最近傍にある，超平面 $Z$ 上にある非負領域 $(\forall \varphi_i \in Z, \varphi_i \geq 0)$ 上の点を求める．これは，$\Phi_1$ の最小値 $\varphi_i = \min(\Phi_1)$ が負であったとき，$Z$ と $\varphi_i = 0$ とが交わる ($|Z|-1$ 次元の) 超平面上にある $\Phi_1$ との最近傍点 $\Phi_1'$ を求め，得られた $\Phi_1'$ について同様の処理を (負値がなくなるまで) 再帰的に行うことにより得ることができる．

これをナイーブに実装すると $O(p^2)$ の計算量となるが，たとえば下記の通り複数の負値をまとめて処理することにより $O(p + n\log n)$ で求めることができる．

(a) $\Phi_1$ の要素を，値の正負により $\Phi_1^+ = \{\varphi_i\} \mid \varphi_i > 0$ と $\Phi_1^- = \{\varphi_i\} \mid \varphi_i < 0$ に分割する．

(b) $\Phi_1^+$ の各要素に対し，$\Phi_1^-$ の総和を $\Phi_1^+$ の要素数で除した値 $-|\Phi_1^-|_1/|\Phi_1^+|$ を加算する．

(c) $\Phi_1^-$ の要素を全て 0 にする．

(d) もし $\Phi_1^+$ が負値を含まないなら，処理を終了する．負値を含むのであれば，$\Phi_1 := \Phi_1^+$ として本手順を再実行する．

上記の手順により書き換えられた $\Phi_1$ を $\Phi_2$ とする．$\Phi_2$ は，$Z$ 上の非負領域にあるため，総数制約と非負制約をいずれも満たす．

(3) 最後に，$\Phi_2$ の最近傍にある $Z$ 上の整数格子点を探索する．これは，$\Phi_2$ の各要素について，小数部分が大きいものは小数点以下を切り上げ，小さいものは同じく切り下げたものになる．具体的には，$\Phi_2$ の各要素の小数点以下を切り下げたものを $\lceil \Phi_2 \rceil$ とするとき，$|\Phi_2 - \lceil \Phi_2 \rceil|_1$ 個の要素を (小数部分が大きい順に) 切り上げ，それ以外の要素を切り下げる．これを $V^+$ とする．

以上の手順により，総数制約，非負制約，整数制約のいずれも満たす，$V^*$ の最近傍点 $V^+$ を得る．これにより，定理 1 が示すように，$f_A^{-1}$ を用いて容易に匿名化個票 $D^+ = f_A^{-1}(V^+)$ を得ることができる．

## 5. 評価

提案方式の有用性を検証するため，提案方式により得られる匿名化個票 $D^+$ がどのくらい元の個票データ $D$ と「近い」かを評価する．個票データ間の距離の定義としては経験分布による Kolmogorov Smirnov (KS)-距離に基づくもの，各レコードの出現頻度の $L_1$ もしくは $L_2$ ノルムに基づくもの，属性間の相関に基づくものなど様々な指標が考えられるが，ここでは [15] にならい，代表的な距離定義として $L_2$ ノルムと KS- 距離を指標として評価する．なお，KS-距離とは，ある 2 つの標本群があったときに，それらが同じ確率分布に従うか否かを検査する検定 (KS-検定) で用いられる統計量であり，それぞれの累積分布 (経験分布) が最大でどのくらい異なるかを距離としたものである．

### 5.1 評価データ

評価のための個票データ (評価データ) としては，商品の売上履歴を擬似的に模した，下記の 3 属性からなるロングテイル性を持つデータを用いる．ある商品が，どのような顧客に売れたかを記録した売上履歴を考える．ここで，商品種別は $r$ 種類あり，それぞれの商品を購入した顧客について性別 (男性/女性) と年齢層 (20 代から 60 代まで 10 歳刻み) が記録されているとする．すなわち，評価データ $D$ は，以下の属性空間 $A$ を台集合とする，$n$ 個の要素から構成される多重集合となる．

$$ \begin{aligned} A &= A_1 \times A_2 \times A_3, \\ A_1 &= \text{商品種別} = \{h_1, \ldots, h_r\}, \\ A_2 &= \text{性別} = \{\ \text{男性}, \text{女性}\ \}, \\ A_3 &= \text{年齢層} = \{20\ \text{代}, \ldots, 60\ \text{代}\ \}. \end{aligned} \qquad (20) $$

パレートの法則 (Pareto principle) が示唆するように，商品売上など，自然現象や社会現象による事象から得られる高次元データの頻度分布はロングテイル性を持つ (べ

き乗則もしくは近い性質が成り立つ) ことが多いとされる．そこで，評価データにおいて商品 $h_k$ が売れる確率 $\Pr[x_{i1} = h_k]$ は，Zipf の法則に従うものとした．すなわち，

$$\forall i, \Pr[x_{i1} = h_k] = \frac{k^{-1}}{\Sigma_{j=1}^{r} j^{-1}}. \qquad (21)$$

また，男女による嗜好差を表すものとして，$k$ が奇数の商品 $h_k$ は男性が女性の二倍多く買う (2/3 の確率で男性が購入，1/3 で女性が購入) とし，$k$ が偶数の場合はその逆とした．年齢層による差は特に導入していない．購入した商品種別や性別に関わらず，顧客の年齢層は 20 代から 60 代まで一様に分布しているものとする．

本章の評価は，この評価データを $r = 100, n = 10,000$ の条件で生成したものを用いた．すなわち，100 種類の商品に関する，10,000 件から成る売上履歴に相当する．

### 5.2 評価結果

前述の評価データを，$\epsilon = 0.1$ の条件下で差分プライバシーを満たしたデータを比較し，その有用性を調べる．比較の対象は，提案方式，維持置換攪乱，Wasserman ら [15] における攪乱サンプリングに基づく手法の 3 種類とし．維持置換攪乱については，すべての属性において維持確率は等しいとした ($\rho_1 = \rho_2 = \rho_3 = \rho$)．なお，一回の試行ごとに評価データを上記の分布に従って作成し，その評価データに対して上記の 3 種類の手法をそれぞれ適用した．

まず，直感的な理解のために，ある試行における，評価データ (“original”)，提案方式の出力 (“proposed”), 維持置換攪乱による出力 (“RRP”)，Wasserman らの攪乱サンプリングによる出力 (“Wasserman”) について，商品種別を $x$ 軸とした度数分布を**図 1** に，累積度数分布を**図 2** に示す．なお，本評価で評価指標の一つとして用いる KS-距離は，ほぼ図 2 におけるグラフ間の最大距離に比例する (これを標本数で除して正規化した値で近似できる)．

次に，定量的な評価のため，この試行により得られた匿名化個票に関し，評価指標として元の評価データとの $L_2$ ノルム，および KS-距離を計算した．この試行を 1,000 回繰り返した平均値と標準誤差 (括弧内) を**表 1** に示す．いずれの指標も，値が小さいほうが元の評価データの分布に近い (より良い) ことを示す．

## 6. 考察

第 5 章の評価結果に基づき，提案手法の有用性に関して従来手法と比較して議論する．

図 1，図 2 のいずれも，維持置換攪乱の出力は元の評価データの性質をほとんど残しておらず，ほぼ一様分布と等価な出力となっていることを示している．これは本評価における条件下において $\rho$ を具体的に計算すると，約 $1.68 \times 10^{-5}$ となり，確率的には 10 万回に 1 ～ 2 回しか元の属性値が維持されない (それ以外は一様ランダムに属性値が選択される) ことを反映している．なお，本来は維持置換攪乱は再構築と呼ばれる操作と組み合わせて用いるべきであるが，基本的に再構築は (維持置換攪乱によってスムージングされた) 分布の山と谷を再拡大する操作であるため，この結果からは再構築を施したとしても精度を向上させることは難しいと考えられる．

それに比べると，提案方式と攪乱サンプリングはいずれも元データの性質を残していると言える．度数分布では差異は一見して明らかではないが，累積度数分布で見ると提案方式のほうがより良好に元データの性質を反映している．これは，攪乱サンプリングはノイズの付与後に負値を単純に切り捨ててしまうことにより，分布が「ファットテイル化 [*9]」する影響を持つことを反映していると考えられる．

上記の考察は，表 1 における定量評価の結果によっても裏付けられる．いずれの評価指標においても，維持置換攪乱の結果は他の 2 手法と比べて明らかに良くない．提案手法と攪乱サンプリングは，$L_2$ ノルムの比較では提案手法のほうが有意に優れているように見えるが，その差はあまり大きくない (10%程度) ように見える．しかし，KS-距離で見ると提案方式は攪乱サンプリングの半分強の距離にまで改善されており，最大でも約 8.5% しか評価データと異ならないことが示されている．すなわち，提案方式は元の評価データの性質をより強く残している．

## 7. まとめ

本稿では，差分プライバシーを満たす匿名化個票の作成方法について検討し，集計データの一種である完全分割表に対する Laplace メカニズムの適用と，ベクトル空間における最近傍探索に基づく非負制約，整数制約，総数制約の充足によりこれが得られることを示した．

提案方式の有用性を評価するために，売上履歴を模したロングテイル性を持つ擬似的な個票データを用いて評価を行ない，従来方式として PRAM の一種である維持置換攪乱方式 [8], [11]，および Wasserman らによる攪乱サンプリングに基づく方式 [15] と比較した．その結果，提案方式は $L_2$ ノルム，KS-距離のいずれの評価指標からも，従来方式と比べて元となる個票データの性質をより強く保持することを示した．

### 参考文献

[1] Departement of Economic and Social Affairs, United Nations: *Multilingual Demographic Dictionary*.
[2] Dwork, C.: Differential Privacy, *Proc. 33rd intl. conf. Automata, Languages and Programming - Volume Part II*, LNCS, Vol. 4052, Springer, pp. 1–12 (2006).
[3] Dwork, C.: Differential privacy: a survey of results,

[*9] ロングテイル性を持つ分布において，テイル部分が全体的に持ち上がり，その分だけ相対的にヘッド部分が頭打ちになる．

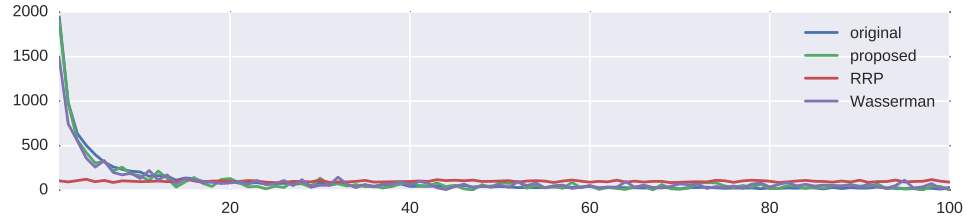


**図 1** 度数分布の比較
**Fig. 1** Comparison of frequency distributions.

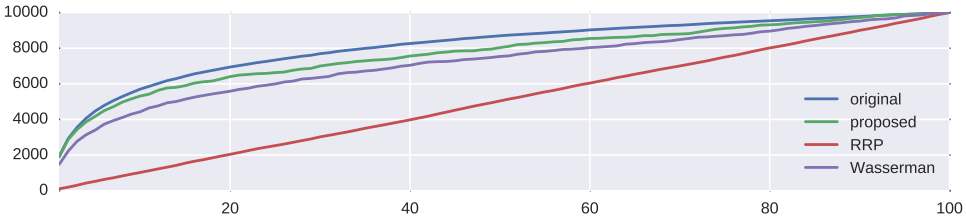


**図 2** 累積度数分布の比較
**Fig. 2** Comparison of the cumulative frequency distributions.

**表 1** 評価データとの $L_2$ 距離と KS-距離の比較
**Table 1** Comparison on $L_2$ distance and KS-distance.

| | 提案方式 | 維持置換攪乱 | 攪乱サンプリング |
|---|---|---|---|
| $L_2$ ノルム | 297.5 | 771.3 | 335.2 |
| | (±13.00) | (±12.60) | (±14.76) |
| KS-距離 | 0.085 | 0.496 | 0.154 |
| | (±0.013) | (±0.006) | (±0.014) |

*Proc. 5th intl. conf. Theory and applications of models of computation*, Springer-Verlag, pp. 1–19 (2008).
[4] Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I. and Naor, M.: Our Data, Ourselves: Privacy via Distributed Noise Generation, *Proc. 25th Annual intl. Cryptology conf. (EUROCRYPT 2006)*, pp. 486–503 (2006).
[5] Dwork, C. and Roth, A.: The Algorithmic Foundation of Diferential Privacy, *Foundations and Trends in Theoretical Computer Science*, Vol. 9, No. 3-4, pp. 211–407 (2014).
[6] Gouweleeuw, J., Kooiman, P., Willenborg, L. and de Wolf, P.-P.: The Post Randomisation Method for Protecting Microdata, *Quaderns d'Estadística i Investigació Operativa (QÜESTIIÓ)*, Vol. 22, No. 1, pp. 145–156 (1998).
[7] 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).
[8] Ikarashi, D., Kikuchi, R., Chida, K. and Takahashi, K.: $k$-anonymous Microdata Release via Post Randomisation Method, *eprint arXiv*, Vol. 1504.05353, pp. 1–22 (2015).
[9] Li, N., Qardaji, W. and Su, D.: On Sampling, Anonymization, and Differential Privacy: Or, $k$-Anonymization Meets Differential Privacy, *Proc. 7th ACM symp. Information, Computer and Communications Security (ASIACCS '12)*, ACM, pp. 32–33 (2012).
[10] Li, Y. D., Zhang, Z., Winslett, M. and Yang, Y.: Compressive Mechanism: Utilizing Sparse Representation in Differential Privacy, *Proc. 10th annual ACM workshop on Privacy in the electronic society (WPES '11)*, New York, New York, USA, ACM Press, p. 177 (2011).
[11] Lin, B.-R., Wang, Y. and Rane, S.: A Framework for Privacy Preserving Statistical Analysis on Distributed Databases, *IEEE intl. Workshop on Information Forensics and Security (WIFS)*, IEEE, pp. 61–66 (2012).
[12] Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J. and Vilhuber, L.: Privacy: Theory meets Practice on the Map, *24th intl conf. Data Engineering*, IEEE, pp. 277–286 (2008).
[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] 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).
[15] Wasserman, L. and Zhou, S.: A Statistical Framework for Differential Privacy, *J. American Statistical Association*, Vol. 105, No. 489, pp. 375–389 (2010).
[16] Xiao, X. and Tao, Y.: $m$-Invariance: Towards Privacy Preserving Re-publication of Dynamic Datasets, *Proc. 2007 ACM SIGMOD intl. conf. Management of Data*, ACM, pp. 689–700 (2007).
[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] Yuan, G., Zhang, Z., Winslett, M., Xiao, X., Yang, Y. and Hao, Z.: Low-rank mechanism: optimizing batch queries under differential privacy, *Proc. VLDB Endowment*, Vol. 5, No. 11, pp. 1352–1363 (2012).
[19] 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).
[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).