Computer Security Symposium 2017
23 - 25 October 2017

統計値を用いたプライバシ保護擬似データ生成手法

岡田 莉奈[1]　正木 彰伍[1]　長谷川 聡[1]　田中 哲士[1]

概要：近年，ビッグデータの一つであるパーソナルデータの分析が注目されるようになってきた．パーソナルデータの分析では，不意・故意に関わらずデータ内の個人のプライバシが侵害されることがある．これを防ぐ方法として，個人の情報ではない集約情報である統計値を分析者へ公開することが考えられる．しかしながら，統計値の公開は分析者にとって有効ではない．なぜならば，公開された統計値の組み合わせから様々な分析結果を得ることは容易でないからである．本研究では，プライバシ保護のため統計値から，分析に必要な生のデータの複数の性質を保持するミクロデータ形式のデータ (擬似データ) を生成する手法を提案する．ミクロデータ形式のデータであれば，分析者は統計解析ソフトウェアによって様々な分析を容易に行うことができる．我々の提案する擬似データは，生のデータにおける各属性のヒストグラムと等しくなるランダムなミクロデータを生成し，それを線形変換することによって，各属性の平均，分散共分散を正確に保持，各属性のヒストグラムと各二属性間のクロス集計表の近似を保持する．このため，ビッグデータ分析でよく用いられる線形回帰モデルを誤差なく計算できる．また，本手法は数値属性だけでなくカテゴリ属性を含むデータに対しても適用可能である．ヒストグラムとクロス集計表の近似度合いは，実データセットを用いて実験的に評価した．

キーワード：擬似データ，プライバシ保護データ分析

Privacy preserved synthetic datasets generation by using statistics

Rina Okada[1]　Shogo Masaki[1]　Satoshi Hasegawa[1]　Satoshi Tanaka[1]

Abstract: Recently, it has been paid attention to analyses of personal data which is one of big data. Individuals who is in the personal data could be violated them privacy by an analyst, no matter whether intentionally or accidentally. To prevent this, we can consider a way that is to disclose statistcs which is aggregated information, not the individual's information. However, the disclosure is unkindness for analysts because it is not easy for analysts to obtain several information which are calculated from combinations of given statistcs. In this work, we propose a method to generate data (we call them synthetic datasets) whose form is same to micro data and which holds some properties of the raw data. It is easy for analysts to obtain some information from the data which is micro data format. Paticuraly, our proposing synthetic datasets holds the mean in each attribute and the variance-covariance exactly, the histgram in each attribute and the cross-tabulation table in each two attributes approximately. Therefore, we can calculate exactly linear regression models which are often used in big data analysis from the synthetic datasets.Our method is able to apply not only numerical attributes but also categorical attributes. In this paper, we experimentaly evaluated utility of histgrams and cross-tabulation table by using a real datasets.

Keywords: Synthetic dataset, Privacy preserved data mining

1. はじめに

近年，ビッグデータ分析が盛んになり，パーソナルデータも分析対象になりつつある．パーソナルデータには個人の情報が含まれるため，分析者が不意・故意に関わらずデータ内の個人のプライバシを侵害してしまうことがある．このようなプライバシ侵害を防ぐ方法として，統計値

[1] NTT セキュアプラットフォーム研究所
NTT Secure Platform Laboratories

表 1　ミクロデータ形式の例

氏名	年齢	性別	身長	体重	年収	学歴	職業
佐藤優子	25	女性	165	46	380	大卒	教師
鈴木悠太	52	男性	169	59	1200	博士卒	医者
高橋博	38	男性	175	61	600	高卒	花屋
田中茂	40	男性	178	63	740	修士卒	会社員

を分析者へ公開することが考えられる．2017 年 8 月現在の国内の法制度では，集約情報である統計値は個人情報でないとされているため，法律上は生のパーソナルデータでなく統計値公開することはプライバシ保護されているとみなすことができる．しかしながら，統計値の公開は分析者にとって有効ではない．なぜならば，公開された統計値の組み合わせから様々な分析結果を得ることは容易でないからである．

本研究では，パーソナルデータ内に含まれる個人のプライバシを保護のため，個人情報でないとされている統計値から分析に必要な生のデータの複数の性質を保持するミクロデータ形式の擬似データと呼ぶデータを生成する手法を提案する．ミクロデータ形式のデータに注目した理由は，ミクロデータ形式のデータであれば分析者は既製の統計解析ソフトウェアを用いることで機械学習などの様々な分析を容易に行うことができるからである．ミクロデータ形式のデータの例を表 1 に示す．以降では，ミクロデータ形式のデータの列を属性，行をレコードと呼ぶ．

生データにおける多くの分析を行うためには，生データのあらゆる情報が必要である．しかし，プライバシ保護の観点から，分析者が生データの統計値しか手に入れることが出来ない場合，より多くの統計値を保持する擬似データがあると有効である．しかしながら，複数の統計値を保持するデータを生成することは難しいく，我々の知る限りでは，三つ以上の統計値を保持するデータを生成する方法は無い．このことから，あらゆる分析に有用な複数の統計値を保持するデータを生成することは難しいため，我々は様々な分析の中でも分析者が主に用いる三つの分析が達成できる擬似データを生成することに焦点を当てる．一つ目は，機械学習でよく用いられる線形回帰 [9] である．これが出来ると，様々な事象の需要予測や疾患リスク予測などを行うことができる．線形回帰の計算は，各属性の平均と分散共分散から計算可能である．二つ目は，マーケティング分析でよく用いられる相関分析である．相関分析では，各二属性間の関連性を調べることができ，例えば，購買データから「20 代の人はアイスを購入しやすい」といったような関連性を知ることができる．相関分析では，二属性クロス集計表を用いる．三つ目は，各二属性間の順位相関分析である．これが出来ると，例えば，二つの属性において片方の属性の順位が高いともう片方の属性の順位は低いということが分かる．順位相関分析の計算は，各属性のヒストグラムと二属性クロス集計表から計算可能である．本稿では，これら三つの分析が達成できる擬似データを提案する．

1.1　関連研究

本節では，統計値と乱数を用いて擬似データを生成する研究について記す．パーソナルデータには，数値属性 (年齢，身長，体重など) とカテゴリ属性 (性別，学歴，職業など) がある．統計値と乱数の数値を利用することによって，数値属性の擬似データを作る研究は多く存在する．しかしながら，カテゴリ属性は数値で扱うことが困難であることから，その研究は少ない．

まず，数値属性のみの擬似データを生成する研究について記す．数値属性のみの擬似データ生成方法には，[6,2,11,7,3] がある．特に，[7] は，属性間の依存関係と周辺分布の情報を含むガウスコピュラと呼ばれる関係式を用いて擬似データを生成する方法であり，本稿で焦点を当てている順位分析と相関分析に有用である．また，[3] は，生データの各属性の平均と分散共分散に完全一致するよう乱数を整形し，擬似データを生成する方法であり，線形回帰に有用である．しかしながら，数値属性だけでも本稿で焦点を当てている線形回帰，順位分析，相関分析を同時に達成できる擬似データ生成手法はこれまでに無い．

次に，カテゴリ属性を含む擬似データを生成する研究について記す．我々の知る限り，これまでにカテゴリ属性を含んだ属性間の関係性を保持し，統計値から擬似データを生成する方法として，擬似ミクロデータ生成方法 [12]，一般ミクロデータ生成方法 [13] がある．[12,13] では，カテゴリ属性によってグループ化されたレコードごとに生データの数値属性の相関係数行列を算出し，そのグループ内にて算出した相関係数を保持する数値属性のデータをランダムに生成する．しかしながら，カテゴリ属性が多い場合や各カテゴリ属性の取りうる値の種類が多い場合，そのグループ数は組み合わせ爆発を起こし，一般的な生データの擬似データを作り出すことは困難である．

1.2　貢献

本研究では，プライバシ保護されたミクロデータ形式の擬似データと呼ぶデータを生成する方法を提案する．この擬似データは，数値属性の各属性の平均，分散共分散が生データのものと完全一致し，数値属性だけでなくカテゴリ属性を含む各属性のヒストグラム，各二属性クロス集計表が生データのものに近似する．

本稿では，上記の擬似データがヒストグラムの近似を保持しつつ，各属性の平均と分散共分散が生のデータのものと完全一致することを理論的に示す．また，ヒストグラムと分散共分散が保持されるならば各二属性のクロス集計表と順位相関も再現できるという仮説を立て，実験では生データと擬似データのクロス集計表と順位相関の誤差が十

表 2　記号一覧

記号	説明
D	生のデータ
D'	擬似データ
N	D のレコード数
N'	D' のレコード数
A_n	D や D' の数値属性の集合
A_c	D や D' のカテゴリ属性の集合
$\boldsymbol{\mu}$	平均ベクトル (各属性の平均の集合)
Σ	分散共分散行列
f	ヒストグラム
F	ヒストグラム集合 (各属性のヒストグラムの集合)

分に小さいことを確認した．これらのことから，提案手法による擬似データは，本稿で焦点を当てている線形回帰，順位分析，相関分析を同時に達成することができる．

本稿の構成は次のようになっている．2 節にて，上記特徴を保持する擬似データを生成するための提案手法について述べる．3 節では，上記で述べた各種近似値の精度を確かめるための実験について記す．最後に 4 節にて，まとめと今後の課題を述べる．

2. 提案手法

2.1 準備

一般的に，擬似データとは生データに似せたデータのことを指す．よって，匿名化データも擬似データに含まれる．また，有用性を全く考慮しない生データと同じ形式のデータも擬似データと呼ばれ，システムのデバッグを行う際に用いられることもある．このように，擬似データは非常に幅の広い用語ではあるが，本稿では擬似データの定義を定義 2.1 に示すものとする．

定義 2.1 (擬似データ)．統計値集合 S を入力にとり，その S と乱数のみを利用し，ミクロデータ形式であり，可能な限り S を保持するように生成されたデータことを擬似データと呼ぶ．

本稿で扱う記号の一覧を表 2 にまとめる．

2.2 提案アルゴリズム

本稿にて提案するアルゴリズムを Algorithm 1 に示す．擬似データのレコード数 N' は，データ提供者または加工者が任意の指定できるパラメータとする．

Algorithm 1 は 4 つの関数を含んでおり，そのアルゴリズムをそれぞれ．Algorithm 2〜Algorithm 5 に示す．関数 FixHistogram は，擬似データの各属性のヒストグラムを生データのものと等しくする．関数 Coding は，各カテゴリ属性の値を符号化 (数値ベクトル化) する．関数 FixMeanCov は，生データの平均ベクトル $\boldsymbol{\mu}_D$ と分散共分散行列 Σ_D と

Algorithm 1 擬似データ生成アルゴリズム

Input: 生のデータの平均ベクトル $\boldsymbol{\mu}_D$，生のデータの分散共分散行列 Σ_D，生のデータのヒストグラム集合 F_D，擬似データのレコード数 N'

Output: $\boldsymbol{\mu}_D$ と Σ_D を保持，F_D の近似を保持する擬似データ D'

1: $X \leftarrow \mathrm{FixHistogram}(\mathrm{F_D}, \mathrm{N'})$
2: $Y, T \leftarrow \mathrm{Coding}(\mathrm{X})$
3: $Z \leftarrow \mathrm{FixMeanCov}(\mathrm{Y}, \boldsymbol{\mu}_\mathrm{D}, \Sigma_\mathrm{D})$
4: $D' \leftarrow \mathrm{DeCoding}(\mathrm{Z}, \mathrm{T})$
5: **return** D'

乱数群を入力にとり，その乱数群の平均ベクトル，分散共分散行列が $\boldsymbol{\mu}_D$，Σ_D と完全一致するよう整形する．関数 DeCoding は，符号化されたカテゴリ属性の値をカテゴリ値に復号する．

まず，Algorithm 1 の設計理由を述べる．数値属性のヒストグラム，平均，分散共分散を保持する方法として二つ考えられる．一つ目は，ヒストグラムを一致させた後に平均と分散共分散を一致させる方法である．二つ目は，平均と分散共分散を一致させた後にヒストグラムを一致させる方法である．Algorithm 1 は，前者の方法を採用した．なぜならば，本稿で焦点を当てている三つの分析をする際に，ヒストグラムは多少誤差があっても順位や相関が保持されていれば良いが，平均と分散共分散から計算される線形回帰は少しでも誤差が出ると結果に大きく影響を及ぼすためである．

次に，Algorithm 2〜Algorithm 5 の詳細を述べる．Algorithm 2 の 4 行目の関数 UniRand$(0,1)$ は，実数範囲 $[0,1]$ から一様ランダムに実数値をサンプリングする関数である．関数 FixMeanCov は，擬似データの各数値属性の平均と分散共分散を生データのものと等しくする．Algorithm 3 内の 3 行目の関数 1-of-K は，機械学習でよく用いられる符号化手法であり，あるカテゴリ属性の取りうる値のバリエーションが $\{c_1, c_2, c_3, \cdots, c_v\}$ であるとき，$c_1 = (0, 0, \cdots, 0)$，$c_2 = (1, 0, \cdots, 0)$，$c_3 = (0, 1, \cdots, 0)$，$\cdots$，$c_v = (0, 0, \cdots, 1)$ に変換 (数値ベクトル化) する．例えば，あるカテゴリ属性の取りうる値のバリエーションが { 公務員，会社員，無職 } であるとき，バリーション数 $v = 3$ であるため，各カテゴリ値はサイズ $\ell = v - 1 = 2$ の数値ベクトルへ変換される．この場合，公務員 $= (0, 0)$，会社員 $= (1, 0)$，無職 $= (0, 1)$ となる．この変換は一対一対応であり，Algorithm 3 ではこの変換を行うたびに，変換の記録を符号化テーブル T へ記録する．この記録した符号化テーブルの反対の変換を T^{-1} とし，Algorithm 5 の 12 行目にて用いる．Algorithm 1，Algorithm 4 の入力の $\boldsymbol{\mu}_D$ と Σ_D は，関数 Coding を用いて D 内のカテゴリ属性を符号化した後に算出する値とする．Algorithm 3 と同様に Algorithm 5 内では i 番目のレコードにおける符号化されたカテゴリ値を $c^{(i)} = (c_1^{(i)}, \cdots, c_\ell^{(i)})$ としている．Algorithm 2〜Algorithm 5 内の $\boldsymbol{x}$，$\boldsymbol{y}$，$\boldsymbol{z}$，$\boldsymbol{d'}$ はそれぞれ

Algorithm 2 関数 FixHistogram

Input: 生のデータのヒストグラム集合 F_D，擬似データのレコード数 N'

Output: F_D に近似するヒストグラムを保持する擬似データ X

1: 属性 $a \in A_n \cup A_c$ のヒストグラム h_a の逆関数 f_a^{-1} を数値的に導出する
2: **for** $a \in A_n \cup A_c$ **do**
3: **for** i to N' **do**
4: $w_{(i,a)} \leftarrow \mathrm{UniRand}(0,1)$
5: **end for**
6: $\boldsymbol{x}_a \leftarrow f_a^{-1}(\boldsymbol{w}_a)$
7: **end for**
8: **return** X

Algorithm 3 関数 Coding

Input: 関数 GenRand によって出力される X

Output: カテゴリ属性が符号化された擬似データ Y，符号化テーブル T

1: **for** $a \in A_c$ **do**
2: **for** i to N' **do**
3: $y_{(i,a)}, T_{(i,a)} \leftarrow \text{1-of-K}(x_{(i,a)})$
4: **end for**
5: **end for**
6: **return** Y

Algorithm 4 関数 FixMeanCov

Input: 生のデータの平均ベクトル $\boldsymbol{\mu}_D$，生のデータの分散共分散行列 Σ_D，関数 Coding によって出力される Y

Output: $\boldsymbol{\mu}_D$ と Σ_D を保持，F_D の近似を保持する擬似データ Z

1: Y の平均ベクトル $\boldsymbol{\mu}_Y$ と分散共分散行列 Σ_Y を算出する
2: $\Sigma_Y = U_Y \Lambda_Y U_Y^T$ となる U_Y，Λ_Y を得る
3: $Q_Y = U_Y \Lambda_Y^{1/2}$ を計算する
4: **for** i to N' **do**
5: $\boldsymbol{z}_i \leftarrow Q_Y^{-1}(\boldsymbol{z}_i - \boldsymbol{\mu}_Y)$
6: **end for**
7: $\Sigma_D = U_D \Lambda_D U_D^T$ となる U_D，Λ_D を得る
8: $Q_D = U_D \Lambda_D^{1/2}$ を計算する
9: $Z = YQ_D^T + I\mathrm{diag}(\boldsymbol{\mu}_D)$ を計算する
10: **return** Z

X，Y，Z，D' 内のレコードとする．

目的の (生データの) 平均や分散共分散行列を保持するデータにするためには，平均ベクトルを $\mathbf{0}$，分散共分散行列を単位行列 I に正規化する必要がある．Algorithm 4 の 4〜6 行目では，その正規化の操作を行っている．文献 [3] より，関数 FixMeanCov の入力の Y を正規分布に従うデータとする場合は正規分布の確率密度関数を利用して正規化が可能であることを示している．しかし，関数 FixMeanCov の入力の Y は必ずしも正規分布に従っているとは限らないため，任意の確率密度関数に従うデータである場合，正規化が可能か確かめる必要がある．これを確かめた結果を定理 2.1 に示す．

定理 2.1. $M \in \mathbb{N}^+$ とし，確率変数を $\boldsymbol{r} \in \mathbb{R}^M$ とする．任意の確率密度関数 $g(\boldsymbol{r})$ に従う $\boldsymbol{r}$ の平均値ベクトルを $\boldsymbol{\mu} \in \mathbb{R}^M$，分散共分散行列を $\Sigma \in \mathbb{R}^{M \times M}$ とする．このとき，$\boldsymbol{z} = Q^{-1}(\boldsymbol{r} - \boldsymbol{\mu})$ であるならば，z を確率変数とする確率密度関数の平均ベクトルは $\boldsymbol{\mu}_Z = \mathbf{0}$，分散共分散行列は $\Sigma_Z = I$ となる．ただし，Q^{-1} は $\Sigma = QQ^T$ となる Q の逆行列であり，I は単位行列である．

Algorithm 5 関数 DeCoding

Input: 関数 FixMeanCov によって出力される Z，符号化テーブル T

Output: カテゴリ属性が復号化された擬似データ D'

1: **for** $a \in A_c$ **do**
2: 属性 a の取りうるカテゴリ値のバリエーション数を v とし，$\ell = v - 1$ とする
3: $\bar{c} = \sum_{i=1}^{N'} \{c_1^{(i)} + \cdots + c_\ell^{(i)}\}/(n \cdot \ell)$ を計算する
4: **for** i to N' **do**
5: $\mathrm{maxIndex} = \mathrm{argmax}_{\mathrm{j}=1}^{\ell}(\mathrm{c}_{\mathrm{j}}^{(\mathrm{i})})$
6: **if** $c_{\mathrm{maxIndex}}^{(i)} \geq \bar{c}$ **then**
7: $c_{\mathrm{maxIndex}}^{(i)} \leftarrow 1$
8: maxIndex 番目以外の $\boldsymbol{c}^{(i)}$ の要素は 0 にする
9: **else**
10: $\boldsymbol{c}^{(i)}$ の全ての要素を 0 にする
11: $d'_{(i,a)} \leftarrow T^{-1}(\boldsymbol{c}^{(i)})$
12: **end if**
13: **end for**
14: **end for**
15: **return** D'

定理 *2.1* の証明． まず，$\boldsymbol{z} = Q^{-1}(\boldsymbol{r} - \boldsymbol{\mu})$ の平均値ベクトルが $\mathbf{0}$ であることを示す．$\boldsymbol{r}' = \boldsymbol{r} - \boldsymbol{\mu}$ であるならば，明らかに，$\boldsymbol{r}'$ をサンプルに持つ確率密度関数の平均値ベクトルは $\mathbf{0}$ である．平均値ベクトルが $\mathbf{0}$ である確率分布に従うサンプルにどのような行列をかけても平均ベクトルは $\mathbf{0}$ のままなので，$\boldsymbol{z} = Q^{-1}\boldsymbol{r}'$ の平均ベクトル $\boldsymbol{\mu}_Z$ は $\mathbf{0}$ である．

次に，$\boldsymbol{z} = Q^{-1}(\boldsymbol{r} - \boldsymbol{\mu})$ の分散共分散行列 Σ_Z が単位行列 I であることを示す．分散共分散行列は，実対称行列であるため特異値分解可能であり，$\Sigma = U\Lambda U^T$ とすることができる．ここから，$Q = U\Lambda^{1/2}$ とする [*1]．ただし，この $\Lambda^{1/2}$ は Λ の各要素に対して平方根を取った行列とする．また，Q は正則であるため，逆行列が存在する．このとき，$\boldsymbol{z} = Q^{-1}(\boldsymbol{r} - \boldsymbol{\mu})$ とすると，$E[\boldsymbol{z}] = \boldsymbol{\mu}_Z = \mathbf{0}$ なので，Σ_Z は次のようになる．

$$
\begin{aligned}
\Sigma_Z &= E[(\boldsymbol{z} - E[\boldsymbol{z}])(\boldsymbol{z} - E[\boldsymbol{z}])^T] \qquad (1)\\
&= E[\{Q^{-1}(\boldsymbol{r} - \boldsymbol{\mu})\}\{Q^{-1}(\boldsymbol{r} - \boldsymbol{\mu})\}^T]\\
&= Q^{-1}E[\boldsymbol{r}\boldsymbol{r}^T](Q^{-1})^T - Q^{-1}\boldsymbol{\mu}\boldsymbol{\mu}^T(Q^{-1})^T\\
&= Q^{-1}(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^T)(Q^{-1})^T - Q^{-1}\boldsymbol{\mu}\boldsymbol{\mu}^T(Q^{-1})^T\\
&= Q^{-1}\Sigma(Q^{-1})^T\\
&= Q^{-1}QQ^T(Q^{-1})^T\\
&= I(Q^{-1}Q)^T = I
\end{aligned}
$$

[*1] 理論上，Σ は正定値行列であるため，コレスキー分解から $\Sigma = QQ^T$ となる Q を求めることも可能であるが，数値計算では $\boldsymbol{z}$ が少ないときは Q の値は不安定になることが多い．そのため，ここでは特異値分解を経由する．

ゆえに，任意の確率密度関数に従うサンプル $\boldsymbol{r}$ を $\boldsymbol{z} = Q^{-1}(\boldsymbol{r} - \boldsymbol{\mu})$ のように線形変換することによって変換された $\boldsymbol{z}$ の平均ベクトルは $\boldsymbol{\mu}_Z = \boldsymbol{0}$，分散共分散行列は $\Sigma_Z = I$ となる．□

平均ベクトルが $\boldsymbol{0}$，分散共分散行列が I であるデータを得ることができれば，そのデータに目的の (生データ D の) 分散共分散行列 Σ_D から得られる回転行列 U_D と拡大縮小行列 $\Lambda_D^{1/2}$ を掛け，平均ベクトルを $\boldsymbol{\mu}_D$ 足せば，目的の平均ベクトルと分散共分散を保持するデータへ変換することができる．このことと定理 2.1 より，関数 FixMeaCov では関数 FixHistogram，関数 Coding から出力されたデータ入力にしたとき，正しく目的の平均ベクトルと分散共分散を保持する．

以上より，Algorithm 1 によって得られる擬似データは，生データのヒストグラムの近似，数値属性の平均・分散共分散の一致を保持する．しかし，提案手法による擬似データが本稿で焦点を当てている相関分析や順位分析にて有用であるかは分からないため，それらについては 3 節にて評価する．

3. 実験

本節では，2 節にて提案した Algorithm 1 によって生成する擬似データがどの程度，生のデータの性質を保持するのか数値実験にて評価する．

3.1 設定

本実験では，UCI Machine LearningRepository [8] にて公開されている国勢調査の結果である Adult データセット[*2]と Census-Income(KDD)[*3] データセットの二種類を用いた．以降では，Adult データセットを Adult，Census-Income(KDD) データセットを Census と呼ぶ．どちらのデータセットも前処理として，欠損値を含む行 (レコード) を削除した．最終的に用いたデータの行数 (レコード数) は，Adult が $30,162$，Census が $95,130$ である．属性数は，Adult が 15(うち，6 属性が数値属性，9 属性がカテゴリ属性)，Census が 41(うち，12 属性が数値属性，29 属性がカテゴリ属性) である．また，生成する擬似データのレコード数は，どちらの生データに対しても $300,000$ とした．

本節で示す実験は次の三種類である．

- 実験 1：線形回帰で有用な擬似データであるか確かめるために，擬似データの各属性の平均と分散共分散の誤差を評価する．
- 実験 2：相関分析で有用な擬似データであるか確かめるために，各二属性間のクロス集計表の誤差を評価する．
- 実験 3：順位分析で有用な擬似データであるか確かめるために，各二属性間の順位相関の誤差を評価する．

比較手法は，数値属性にもカテゴリ属性にも適用可能な，Algorithm 2 のみから擬似データを得る方法とカテゴリ属性は符号化せずに数値属性のみ Algorithm 4 を適用する方法の二種類とする．以降では，前者の方法を比較手法 A，後者の方法を比較手法 B として記述する．

実験 1 にて用いる各属性の平均の誤差と分散共分散の誤差を式 2，式 3 に示す．

$$\mathrm{AveMean} = \frac{\sum_{a in A_n} |\mu_D^{(a)} - \mu_{D'}^{(a)}|}{|A_n|} \tag{2}$$

$$\mathrm{AveCov} = \frac{\sum_{a_1 in A_n} \sum_{a_2 \in A_n} |\sigma_D^{(a_1,a_2)} - \sigma_{D'}^{(a_1,a_2)}|}{|A_n|^2} \tag{3}$$

ただし，生データ，擬似データから算出される平均ベクトルの属性 $a \in A_n$ に関する要素をそれぞれ，$\mu_D^{(i)}$，$\mu_{D'}^{(i)}$ とし，生データ，擬似データから算出される分散共分散行列の属性 $a_1 \in A_n$ と属性 $a_2 \in A_n$ に関する要素をそれぞれ，$\sigma_D^{(a_1,a_2)}$，$\sigma_{D'}^{(a_1,a_2)}$ とする．

実験 2 にて用いるクロス集計の誤差を式 4 に示す．生データと擬似データは，行数 (レコード数) や取りうる値の範囲が異なるため，対応するセル同士の比較を簡単に行うことができない．そのため，ヒストグラムの各セル内の値は，度数の割合とし，その誤差を計測する．数値属性における各セルの範囲は，生データと擬似データの最大値のうち大きい方を max，最小値のうち小さい方を min とし，範囲 $[min, max]$ を 10 分割したものとした．カテゴリ属性における各属性の範囲は，生データにおける各属性の取りうる値を一つのセルの範囲とする．このとき，生データ，擬似データの属性 $a_1 \in A$，属性 $a_2 \in A$ の i 番目のセルの割合をそれぞれ，$c_D^{(a_1,a_2,i)}$，$c_{D'}^{(a_1,a_2,i)}$ とする．

$$\begin{aligned}&\mathrm{AveCross} \\ &= \sum_{a_1 \in A} \sum_{a_2 \in A} \frac{\sum_{i=1}^{|C^{(a_1,a_2)}|} |c_D^{(a_1,a_2,i)} - c_{D'}^{(a_1,a_2,i)}|}{|A|^2 \cdot |C^{(a_1,a_2)}|}\end{aligned} \tag{4}$$

ただし，属性 a_1，属性 a_2 のクロス集計表のセルの個数を $|C^{(a1,a2)}|$ とする．

実験 3 にて用いる順位相関の誤差を式 5 に示す．順位相関は，スピアマンの相関係数を用いて計測する．スピアマンの相関係数は，生データ D，擬似データ D' 内の各値を各属性内の度数の順位に変換した後に計算する．以降では，$A = A_n \cup A_c$ とする．

$$\begin{aligned}&\mathrm{AveSpearCorr} \\ &= \frac{1}{|A|^2} \cdot \sum_{a_1 \in A} \sum_{a_2 \in A} |\mathrm{SpearCorr}(\boldsymbol{a}_1(D), \boldsymbol{a}_2(D)) \\ &\quad -\mathrm{SpearCorr}(\boldsymbol{a}_1(D'), \boldsymbol{a}_2(D'))|\end{aligned} \tag{5}$$

[*2] https://archive.ics.uci.edu/ml/datasets/adult
[*3] https://archive.ics.uci.edu/ml/datasets/Census-Income+(KDD)

表 3　生データと擬似データの平均ベクトルと分散共分散行列の誤差

	比較手法 A	比較手法 B	提案手法
AveMean (Adult)	$6.58E2$	0	0
AveMean (Census)	29.25	0	0
AveCov (Adult)	$1.92E6$	0	0
AveCov (Census)	$2.53E5$	0	0

表 4　生データと擬似データの二属性クロス集計表の誤差

	比較手法 A	比較手法 B	提案手法
AveCross(%)(Adult)	0.1191	0.2518	0.2331
AveCross(%)(Census)	0.5684	0.2385	0.7447

表 5　生データと擬似データの順位相関係数の誤差

	比較手法 A	比較手法 B	提案手法
AveSpearCorr(Adult)	0.0858	0.0858	0.0666
AveSpearCorr(Census)	0.1192	0.1242	0.1232

ただし，D 内の任意の二属性 $a_1 \in A$ と $a_2 \in A$ における列ベクトルを $\boldsymbol{a}_1(D)$，$\boldsymbol{a}_2(D)$ とし，D と D' のそれぞれにおけるスピアマンの相関係数を $\mathrm{SpearCorr}(\boldsymbol{a}_1(D), \boldsymbol{a}_2(D))$，$\mathrm{SpearCorr}(\boldsymbol{a}_1(D'), \boldsymbol{a}_2(D'))$ とする．

3.2　結果と考察

実験 1：各手法により生成した擬似データの各属性の平均と分散共分散の誤差を評価した結果を表 3 に示す．比較手法 A は，平均ベクトルと分散共分散行列を生データのものに一致させる操作をしていないため，誤差が生じた．一方，比較手法 B と提案手法は，平均ベクトルと分散共分散行列を生データのものに一致させる操作をするため，実験でも誤差が無いことが確かめられた．このことから，提案手法による擬似データの数値属性の線形回帰モデルは，生データから得られる線形回帰モデルと一致する．

実験 2，3：各手法により生成した擬似データの二属性クロス集計表の誤差を評価した結果を表 4 に，順位相関係数の誤差を評価した結果を表 5 に示す．我々は，比較手法 A による擬似データは，各属性ごとに独立にデータが生成されるため，クロス集計表や順位相関は保持されないという仮説を立てていた．しかしながら，比較手法 A によって生成した擬似データはそれらの誤差が比較手法 B や提案手法より小さいことが分かった．このため，線形回帰はせずに相関分析と順位分析のみを対象とする場合は，比較手法 A を用いることが良いと言える．線形回帰，相関分析，順位分析の全ての分析を対象とする場合は，比較手法 B と提案手法のどちらを使うべきかは生データの性質次第である．

生データの性質次第である理由を次に述べる．Adult のクロス集計表の誤差が最も大きい場合を図 1 に，誤差が最も大きい場合を図 2 に示す．Census は，前処理の欠損値を含む行 (レコード) の削除によって図 2 のように取りう

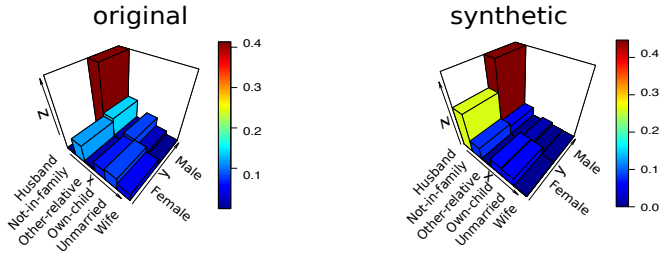

図 1　Adult データセットの二属性クロス集計表の中で最も誤差の少ない場合 (属性名：Relationship，Sex)．左側の original は生データの結果であり，右側の synthetic は擬似データの結果．x 軸は Relationship，y 軸は Sex，z 軸は度数の割合．x 軸，y 軸は，左から右に向かってアルファベット順に並べて表示している．

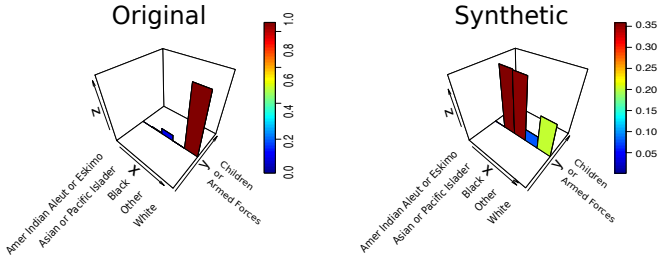

図 2　Census データセットの二属性クロス集計表の中で最も誤差の少ない場合 (属性名：Race，Full- or part-time employment status)．左側は生データの結果であり，右側は擬似データの結果．x 軸は Race，y 軸は Full- or part-time employment status，z 軸は度数の割合．

る値の種類が一種類である属性が含まれている．これにより，二属性間の相関を評価することができないにも関わらず，Algorithm 4 内の分散共分散を一致させようとする操作を加えるため，誤差が発生させていることが分かった．逆に，取りうる値の種類が二種類以上の属性から構成されている Adult では，最も誤差が大きい場合でも図 1 のようになり，生データに近い分布を形成することができることが分かった．

実験 1 から 3 より，生データに取りうる値が一種類である属性が含まれなければ，線形回帰，順位分析，相関分析の全ての分析を行う際には提案手法を用いると良いということが分かった．

4.　まとめと今後の課題

本稿では，生データの平均，分散共分散，ヒストグラムから，それらと二属性クロス集計表と順位相関係数の近似を保持する擬似データ生成手法を提案した．特に，実験を通して，生データに取りうる値が一種類である属性が含まれなければ，数値属性に関しては正確に線形回帰を行うことができ，かつ，順位分析，相関分析の誤差の少ない擬似

データを生成できることを示した．

今後の課題は，次の四点である．一点目は，あり得ない属性の組み合わせの発生を制限した擬似データを生成することとである．あり得ない属性の組み合わせとは，{ 夫,5歳 } のような社会通念上あり得ない属性の組み合わせのことである．本稿での提案手法では，このようなあり得ない属性の組み合わせの制限を設けずに，関数 FixMeanCov 内でデータの回転操作を施したため，そのようなセルの度数が増加してしまった．今後は，このようなセルへ制限を設けた擬似データ生成手法を検討する．

二点目は，数値属性において最大値，最小値を保持する擬似データを生成することである．これは，上記課題にも関係する．本稿での提案手法では，擬似データ内のカテゴリ属性は生データに含まれない属性値は存在しないよう調整したが，数値属性に関しては，最大値や最小値の保持はしていない．それゆえ，平均ベクトルや分散共分散行列は完全一致するものの，生データに含まれない極端に大きな値や小さな値が擬似データの中には存在してしまった．さらに，学歴年数のような非負の値しか存在しないはずの属性についても負の値を生成してしまっている．今後は，平均ベクトル，分散共分散行列の誤差を最小にしつつ，生データ内の最大値，最小値内に存在する擬似データを生成することを目指す．

三点目は，時系列の性質も保持する擬似データを生成することである．本稿での提案手法では，“ある時点” の生データに対する擬似データであるため，複数の時点のデータから得られる傾向を保有できるものなのか分からない．複数の時点のデータから得られる傾向を測るものとして，自己回帰モデル [1] や移動平均モデル [5] がある．今後は，複数の時点の生データに対するこれらのモデルと等しいモデルを保持する擬似データを生成することを目指す．

四点目は，複数の統計値を保持する擬似データの匿名性を調査することである．本稿での提案手法や今後の検討手法では，プライバシ保護のために個人情報ではない統計値と乱数を用いて擬似データを生成することを考えている．しかしながら，そのような方法によって生成される擬似データから生データのいくつかの行 (レコード) が特定できないことを保証する必要がある．統計値から擬似データを生成する場合，生データに関する情報は統計値にしかないため，疑似データ生成に用いる統計値の開示における匿名性を考えることと等しい．統計値の開示に関するプライバシ保護方法として，差分プライバシ [4] や統計的開示制御 [10] がある．しかし，差分プライバシでは個人を特定しようとする攻撃者の仮定が厳しいため，ほとんどの場合で正確な統計値は公開されず，場合によっては誤差の大きい統計値を出力することがある．実用化においては，差分プライバシほど厳しい攻撃者の仮定をするのではなく，現実的な攻撃者の仮定をおくことによって，正確な統計値の公開をしたとしても個人の特定には繋がらないという場合も検討すべきであると考えている．また，これまでの統計的開示制御では単一の統計値からの個人特定リスクを考慮しているが，複数の統計値の公開からの匿名性は評価されていない．これらのことから，今後は複数の正確な統計値の公開における匿名性について検討する．

参考文献

[1] Hirotugu Akaike. Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21(1):243–247, 1969.

[2] Rui Chen, Qian Xiao, Yu Zhang, and Jianliang Xu. Differentially private high-dimensional data publication via sampling-based inference. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 129–138. ACM, 2015.

[3] Chuong B Do. The multivariate gaussian distribution. *Section Notes, Lecture on Machine Learning, CS*, 229, 2008.

[4] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9:211–407, August 2014.

[5] Walter Enders. Stationary time-series models. Applied Econometric Time Series (Second ed.):48–107, 2004.

[6] Zhengli Huang, Wenliang Du, and Biao Chen. Deriving private information from randomized data. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 37–48. ACM, 2005.

[7] Haoran Li, Li Xiong, and Xiaoqian Jiang. Differentially private synthesization of multi-dimensional data using copula functions. In *Advances in database technology: proceedings. International Conference on Extending Database Technology*, volume 2014, page 475. NIH Public Access, 2014.

[8] M. Lichman. UCI machine learning repository, 2013.

[9] George AF Seber and Alan J Lee. *Linear regression analysis*, volume 936. John Wiley & Sons, 2012.

[10] Leon Willenborg and Ton de Waal. *Statistical Disclosure Control in Practice*. Lecture Notes in Statistics. Springer-Verlag New York, 1996.

[11] Jun Zhang, Graham Cormode, Cecilia M Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes: Private data release via bayesian networks. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 1423–1434. ACM, 2014.

[12] 独立行政法人 統計センター. **教育用擬似ミクロデータの開発とその利用 ～平成 *16* 年全国消費実態調査を例として～**. 2012.

[13] 滝澤 有美 and 平澤 鋼一郎. 一般用ミクロデータ（仮称）の作成及び利活用について. **統計関連学会連合大会**, 2015.