Scientific and Educational Reports of the Faculty
of Science and Technology, Kochi University
Vol. 1 (2018), No. 5

個人情報のないビッグデータ再生成法とその検証

A Privacy Free Big-Data Reproduction Method and Verification

筒井 真璃菜, 三戸 理誠, 江口 大介, 豊永 昌彦

Marina Tsutsui, Masataka Mito, Daisuke Eguchi, Masahiko Toyonaga

〒780-8520 高知市曙町 2-5-1

高知大学大学院理学専攻 情報科学

Information Science, Graduate School of Science, Kochi University

Kochi 780-8520 Japan

あらまし

近年, ビッグデータは, 業務効率や売上げ改善のため多様な産業で利用されている. しかしながら, ビッグデータの中には, 個人情報保護問題を含むものがあり, 何らかの保護方法が求められる. 従来の個人情報保護技術として, 氏名の記号化, 個人を特定する数値の乱数化や除去, データ全体の暗号化などが挙げられる. 一方, 様々に保護されたデータから名寄せなどにより個人特定される危険性も指摘されている.

本研究は, ビッグデータの様々な統計情報のみからデータ再生成する方法を提案するものであり, 個人情報保護問題を根本的に解消するビッグデータ利用環境の提案である. 提案手法は, ヒストグラムと相関係数を入力として, モンテカルロ法でヒストグラムの再生, およびデータレコード順の交換で相関係数の再現をする.

生成法によるデータがビッグデータとして利用可能かどうかを検証するため, 医療データを生成し, これと実際の医療データの SOM 分析を比較する. 実験の結果, 生成データは, ほぼ同等の SOM 分析ができることが確認できた.

Keyword:**ビッグデータ, ヒストグラム, 相関係数, SOM**

1 はじめに

近年, 様々な分野で, 業務効率改善や売上げ向上のために, ビッグデータが利用されている[1]. ビッグデータのデータ属性の統計のみでは発見できない知見が, データマイニング技術により発見され, 役立てられている. 一方, これらビジネスで利用されるビッグデータの多くは, 顧客情報などの個人情報の保護が社会的な関心事となっている. そのため, 今後のビッグデータ分析技術の発展のためには, 何らかの個人情報保護技術の確立が不可欠である.

従来行なわれている個人情報保護技術には, 氏名の記号化, 個人を特定する数値の乱数化や除去, データ全体の暗号化などが挙げられる[2][3][4].

ここで, 表 1.1 にビッグデータの例を示す. 第 1 カラムに氏名(Name), 第 2 カラムに年齢(Age)や性別(Sex), さらに最後のカラムには年収(Income)などの個人に関する多種の情報(属性データ)で構成されている. これを「氏名の記号化」をしたものを表 1.2(a), (b)に示す.

それぞれ単独では, 氏名が C と番号, D と番号などと氏名が記号化されており, かつ分割されているため, それぞれの表で個人を特定することは難しい.

しかし, このような情報保護策を加えても, 個人情報が復元される危険が排除できない. 例えば表 1.2, 表 1.3 において 24 歳が 1 人であることがわかれば, C12 と D4 が同一人物であることが特定されてしまい, 氏名まで特定される危険性が残る.

ビッグデータ分析の研究を進展させるためには, 新たな個人情報保護技術が望まれる.

そこで, 本研究は個人情報を含まない統計データのみからビッグデータ分析を可能にするデータ生成法を提案し, 個人情報保護問題の解決を目指す.

なお本論文において「ビッグデータ分析を可能にするデータ」とは, 1)データ属性の分布特徴がビッグデータと同様であること, および 2)データ属性の相関性の特徴がビッグデータと同様であることと仮定する. これは, データ分析で扱う統計値であるデータ属性の平均値, 分散, 最大値, 最小値が属性データの分布により算出できること, および自己組織化マップ(SOM)[6]などの分析ツールでは, データ属性の相関性でデータマイニングしているためである[7.8].

提案するデータ生成手法は, 1)の分布関数としてヒスト

Received: March 5, 2018
Reviewed by anonymous referee, and accepted: March 15, 2018
Published: March 31, 2018

グラム H, 2)の相関性として相関係数表 R を入力として任意レコード数 N のビッグデータを再現するデータを生成する. 入力ヒストグラム H からヒストグラムの階級 a_i (値範囲 a_{i-1} から a_i) のデータ数(頻度)に基づくモンテカルロ法によりヒストグラム H を再現し, 次に属性データのレコード交換により相関係数を再現する. 属性データのレコード番号の変更は, 分布を変化させないため, 相関係数調整後もヒストグラムを再現した分布は, 損なわれず, 上記の2段階を独立して実施できる.

提案するデータ生成法がビッグデータ分析で有用なデータを生成したかどうかを検証するため, SOM[6]により評価する. 入力データとしては, 筒井等が医療データの SOM 分析[5]での個人情報を含まない統計データ, すなわちデータ属性のヒストグラムとデータ属性間の相関係数表を用いる. 同ヒストグラムとデータ属性間の相関係数を入力として, N=1,000 のヒストグラムと相関係数表の比較, 生成データに対する SOM 分析結果を比較した. その結果, ヒストグラム, 最大値・最小値・平均値・標準偏差の再現, およびデータ属性間の相関係数の再現, さらに, ビッグデータと同等の SOM 分析による知見が得られることが確認された.

以下の構成は, 第 2 節で提案するデータ生成法を説明し, 第 3 節で実証のための実験と考察を行い, 第 4 節で全体をまとめる.

表 1.1 ビッグデータの例

Name	Age	Sex	- -	Income($)
Tom	19	male	- -	1,000
Tony	22	male	- -	6,000
Linda	33	female	- -	40,000
Mark	38	male	- -	30,000
- -	- -	- -	- -	- -
Lerey	48	female	- -	35,000
Tuka	30	female	- -	25,000
Lisa	40	female	- -	40,000
Solo	24	male	- -	60,000

表 1.2 個人情報隠蔽(a)

Name	Age	Sex
C1	19	male
C2	22	male
C3	33	female
C4	38	male
- -	- -	- -
C9	48	female
C10	30	female
C11	40	female
C12	24	male

表 1.3 個人情報隠蔽(b)

Name	Income($)	Age
D1	1,000	19
D2	40,000	33
D3	25,000	30
D4	60,000	24
- -	- -	- -
D9	30,000	38
D10	35,000	48
D11	6,000	22
D12	40,000	40

2 データ生成法

本論文で提案するデータ生成法は, ビッグデータの 1)分布と 2)相関係数表を入力として, これらを再現するデータの生成をおこなう. 以下, 1)および 2)を再現する方法について説明する.

2.1 分布を再現するデータ生成法

分布を再現するデータ生成法では, ヒストグラム H を入力して, ヒストグラムを再現するレコード数 N のデータの生成を目指す.

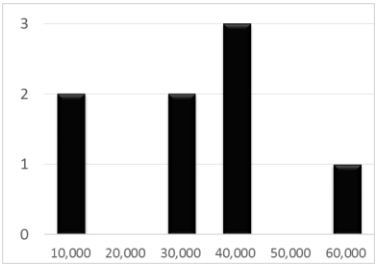


図 2.1 表 1.1 の Incom のヒストグラム H

ヒストグラムとは, 多数の数値データについて数値幅 W の範囲毎を階級と称して, これら階級内の値をもつデータの頻度を表したものである. 先ほどの表 1.1 におけるデータ属性 Income を, 数値データとしてそのヒストグラム H(Income)を図 2.1 に示す. 図 2.1 に示すように, Income のヒストグラム H は, 数値幅 W=10,000 で x 軸が区切られている. 各範囲である階級ラベル CV は, $CV-W$ から CV 以下の数値をもつデータ数(度数)を縦軸 y にプロットしたものである. 例えば, 階級 CV=40,000 において, 30,000 から 40,000 以下の数値をもつデータが 3 個あることを示す.

いま, レコード数 N_B のデータ D_B のヒストグラム H_B が与えられたとし, H_B が幅 w で C_0 から $Cmax$ まで階級で区分されているとし, i 番目の階級 C_i には $w\times(i\text{-}1)$ から $w\times i$ 以下の数値のデータが $P(C_i)$個あるとする. 各階級の度数 $P(C_i)$の総和は式(2,1)となる.

$$\sum_{i=0}^{max} P(C_i) = N_B \tag{2.1}$$

従って, 全データにおける階級 C_i の範囲の値をもつデータ値の相対度数 $p(C_i)$は式(2.2)となる,

$$p(C_i) = \frac{P(C_i)}{N_B} \tag{2.2}$$

0 以上 1 未満の擬似乱数 $Rand$ が式(2.4)を満たすとき階級 C_k のデータを生成し, その値 v を式(2.3)に従って決定すれば, ヒストグラム H_B を再現する任意数のデータの生

成ができる.

$$v = \frac{1}{2}w\big(k + (k+1)\big) \tag{2.3}$$

where $$\sum_{i=0}^{k-1} p(C_i) < Rand < \sum_{i=0}^{k} p(C_i) \tag{2.4}$$

以上の実装に際して, 疑似乱数から階級 k を容易に得るため, 式(2.2)の$p(C_i)$に代わり, これを λ 倍した式(2.2b)の$ps(C_i)$, およびそれらを積算したルックアップテーブル L[], 疑似乱数 *Rand* を λ 倍した *Rand* ・λ を導入する.

```
Input:
  Number of record N_B
  Histogram
       H_B={freq[i] | i=0, 1, 2, ..., C_max}
        with Class width w
  Required Number of data as N
Output:
  Data D_H[i]  (i=1 to N)

Histogram Data Generator(H_B,N_B,w,N,D_H) {
 // calculate probability for each class
   for (i = 0 to C_max){
      p [i] =freq[i]/_HB ;
   }
 // set probability line L[]
   k = 0;
   for (i = 0 to C_max){
      for (j = 1 to p[i]*Λ){
         k++; L[k] = i;
      }
   }
   for (i = 1 to N){
       r = rand()*Λ
       k = L[r]
       D_H[i] = w*(k+(k+1))/2;
   }
   return ;
}
```

図 2.2 ヒストグラムを再現するデータ生成法

ここで, $ps(C_i)$を導入する理由は, 0 から 1 の疑似乱数 *Rand* を 0 からλまでの整数化し, 配列 L[]の最大インデックスを限定するためである.

ルックアップテーブル L[]は, 式(2.4b)により生成する.

$$ps(C_i) = p(C_i)\lambda \tag{2.2b}$$

$$L[i] = k \quad \text{where} \sum_{i=0}^{k-1} ps(C_i) < i < \sum_{i=0}^{k} ps(C_i) \tag{2.4b}$$

配列 L[]と *Rand*・λ により階級 k を(2.4c)から計算し, データ値 v を式(2.3)から決定する.

$$k = L[Rand \cdot \lambda] \tag{2.4c}$$

以上を指定された生成データ数 N 回繰り返せば, ヒストグラム H_B と同等の分布をもつデータ D_H が得られる.

提案アルゴリズムを図 2.2 に示す. 多次元のビッグデータを再現する場合は, それぞれのデータ属性のヒストグラムについて図 2.2 のアルゴリズムを用いて生成する.

2.2 相関係数の再現法

ここでは, 前節 2.1 における分布を再現するデータ生成法によりデータ属性 $a1$,, 属性 $a2$, 属性 $a3$ について各属性データ間の相関係数 R_B を再現する方法について説明する.

相関係数を再現するためには, 同一データ属性内の2つのレコードの数値を交換することで相関係数 R_B に繰り返し近づける方法をおこなう. なお, ヒストグラムがある階層(値の範囲)に含まれるデータ数であることから, それらデータのレコード順に依らないため, これらの処理で元のヒストグラムは変わらない.

相関係数は, データ属性 x, y 間における依存性(線形関係性)を表す数値であり, 2 つのデータ属性 x, y について, 相関係数 R_{xy} は, それぞれの属性の i 番目のレコード値 x_i, y_i により式(2.5)として定義されている.

$$R_{xy} = \frac{\frac{1}{N}\sum_i (x_i - \langle x \rangle)(y_i - \langle y \rangle)}{\sqrt{\frac{1}{N}\sum_i (x_i - \langle x \rangle)^2}\sqrt{\frac{1}{N}\sum_i (x_i - \langle x \rangle)^2}} \tag{2.5}$$

式(2.5)における総和は, データ数 N についてであり, <x>, <y>はそれぞれの属性の平均値を示す. 分母は, 属性 x, y の標準偏差の積であり, 分子は同一レコード番号の属性 x, y のそれぞれの平均値から差の積を総和した値(共分散)である. 同一属性の相関係数は, y=x と置くと R_{xx}=1 となる.

いま, 式(2.6)と式(2.7)にヒストグラムを再現した M 次元の属性をもつ生成データ D_H, およびビッグデータの相関係数 R_B を入力として, 相関係数 R_B を再現するデータ D_{HR} の改善法を説明する.

$$D_H = \left\{ d[m][i] \mid m = 1,2,\dots,M, \quad i = 1,2,\dots N \right\} \tag{2.6}$$

$$R_B = \left\{ r[m_1][m_2] \mid m_1 = 1,2,\dots,M, \quad m_2 = 1,2,\dots M \right\} \tag{2.7}$$

```
Input:
  M dimension data D_H
     D_H = {data[m][i] | m=1 to M, i=1 to N}
  Correlation Table R
     R_B = {r[m1][m2] | m1=1 to M, m2=1 to M}
Output:
  Correlated Data with dimension M
     D_HR = {data[m][i] | m=1 to M, i=1 to N}

Correlated Data Improvement(D_H,R_B, D_HR ) {
 // calculate correlation R_D of D_H
     R_D = cal_corr(D_H, M, N);

 // calculate probability for each class
   for (k = 1 to Iter){
     for(m=1 to M){
       i = rand()*N;  j = rand()*N;
       diff1 = m_corr(D_H, R_B, R_D, m);
       exchange(D, m, i, j);
       diff2 = m_corr(D_H, R_B, R_D, m);
       if(diff1 < diff2)  exchange(D_H,m, i, j) //reject
     }
   }
   D_HR =D_H;
   return ;
}
```

図 2.3 相関係数 R を再現するデータ改善法

まず, 相関係数の再現アルゴリズムを図 2.3 に示す. まず, 入力データ D_H から式(2.5)により相関係数表 R_D を作成する関数 *cal_corr*()をおこなう. 次に, データ属性 m のレコード変更で相関係数 R_D を R_B に近づける改善処理を m＝1 から M まで繰り返し, これを任意数 Iter 回繰り返す.

具体的には, 属性 m について m と m'(m'≠m)の相関係数を式(2.5)から計算し, R_B の属性 m と m'の全相関係数の差 diff1 を関数 *m_corr*(D_H, R_B, R_D, m)により計算する. 次に, 属性 m 内のランダムに選択したレコード i, j を交換し, diff2 として関数 *m_corr*(D_H, R_B, R_D, m)を再計算する. もし, diff2 が diff1 より小さければ, R_D が R_B に近づいているので, 交換をそのまま受け入れ, そうでなければレコード i, j を再交換して元に戻す.

以上を全属性についていて行い, さらに十分な回数 Iter だけ繰り返すことで, D_H の全属性データのレコードの改善を行う.

このレコード交換で各データ属性のヒストグラムは変化しない. 従って, 式(2.5)における分母の標準偏差は変わらないため再計算は不要となる. また, 全レコードの平均値からの差をデータ属性ごとに記録保存しておけば, 再計算は, 交換するレコードの値2つに関してのみで, 変更後の相関係数が求められる. すなわち, 1 回の交換にともなう手間は, 2 レコード×M 程度の計算量で実行可能である.

3 SOM による検証実験

前章で提案したヒストグラムと相関係数を再現するデータ生成法によるデータを用いてビッグデータ相当のデータマイニングができるかどうかを評価するため, 筒井ら[5]が SOM（自己組織化マップ）で分析用いた医療データのヒストグラムおよび相関係数を用いてデータ生成し, SOM 分析結果を比較する.

3.1 評価データと実験条件

評価データは, 過去に高知大学医学部と筒井らが共同研究をおこなった際に使った高知大学医学部附属病院の患者検査データのうち論文[5]で公開している分析結果と, ヒストグラム, 相関係数表のみを入力としている.

同ビッグデータのヒストグラムによれば, データ属性数 M=302, データ数 N_B＝45,289 である.

評価実験では, 生成データ数 N=1,000, M=302, λ=100, および繰り返し回数 Iter＝10,000（全交換回数は, N×10 程度としている）としてデータを生成した.

3.2 生成データの再現性の概要

生成結果の一部を表 3.1 に示す. 1列目は患者 Customer の ID, 2 列目以降は検査などのデータ属性値（検査値や症状の有無など）である. それぞれの属性の名称は, eGFR, UTP や病歴などである.

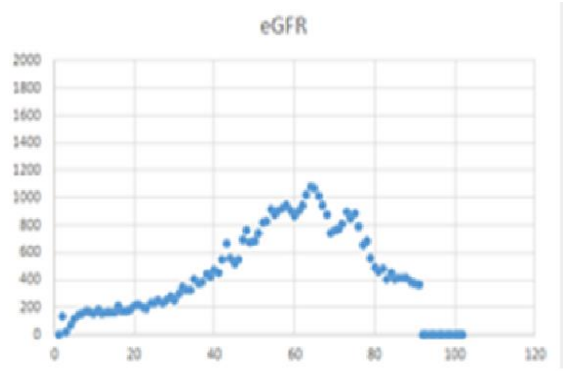
ビッグデータのヒストグラムと, N=1,000 としたときの生成データのヒストグラムを図 3.1（a）,（b）にそれぞれ示す. ビッグデータのデータ数 N_B に比べて生成データ数 N が少ないため, 図 3.1（b）にばらつきが見られるが, ほぼ同等の分布であることが確認できる.

次に, ビッグデータと生成データの相関係数の一部を表 3.2(a), (b)に示す. それぞれ 1 行目および1列目が属性を示し, 各行と各列が交差する部分がそれぞれの属性間の相関係数である. 同一属性間の相関係数は 1.0 となることが自明であるため, 記載していない.

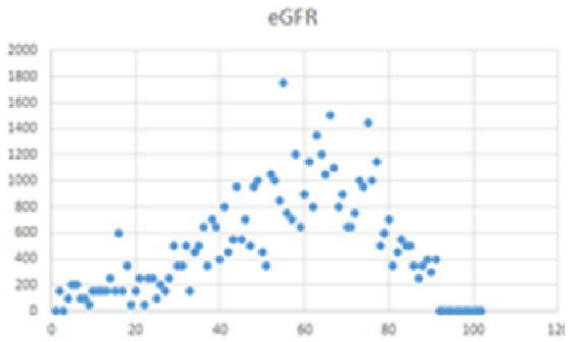
例えば，ビッグデータの2つの属性 eGFR と UTP の相関係数は －0.3068 であり，生成データは －0.29341 である．ともに負の相関であること，数値誤差は 0.01339 であることがわかる．また，胃潰瘍と UTP の相関係数では，ビッグデータが 0.075092，再現データが 0.087371 である．ともに正の相関であること，誤差が 0.012279 である．このようにしてビッグデータと生成データのすべての属性間の相関係数を比較すると，平均誤差が 0.00597 であった．ゆえに，提案手法により生成したデータが，ビッグデータの相関係数をほぼ再現していることが確認できる．

表 3.1 ヒストグラム再現データ D_H の一部

Customer	eGFR	UTP	胃潰瘍	高血圧症	腰痛症	便秘症	糖尿病
C1	63.25386	50.93868	1.000002	1.000002	1.000002	1.000002	0.004717
C2	45.84844	74.54717	0.004717	0.004717	1.000002	0.004717	0.004717
C3	50.94271	74.54717	0.004717	0.004717	1.000002	0.004717	1.000002
C4	24.1978	50.93868	1.000002	0.004717	1.000002	0.004717	0.004717
C5	57.31055	253.9716	0.004717	1.000002	0.004717	0.004717	0.004717
C6	25.04684	8.443397	1.000002	0.004717	0.004717	0.004717	0.004717
C7	60.2822	74.54717	0.004717	0.004717	0.004717	0.004717	0.004717
C8	45.42392	13.16509	1.000002	0.004717	0.004717	0.004717	0.004717
C9	40.32965	3.721698	0.004717	0.004717	0.004717	0.004717	0.004717
C10	80.23476	13.16509	0.004717	0.004717	0.004717	0.004717	0.004717
C11	61.55577	253.9716	1.000002	0.004717	0.004717	0.004717	0.004717
C12	66.65005	3.721698	1.000002	0.004717	0.004717	1.000002	0.004717
C13	70.89527	22.60849	0.004717	0.004717	0.004717	0.004717	1.000002
C14	58.15959	22.60849	0.004717	0.004717	0.004717	0.004717	0.004717
C15	49.66914	13.16509	0.004717	0.004717	0.004717	0.004717	0.004717



(a) 入力データのヒストグラム



(b) 生成データのヒストグラム

図 3.1 ヒストグラムの再現の比較

表 3.2(a) ビッグデータの属性間の相関係数 R_B

属性	eGFR	UTP	胃潰瘍	高血圧症	腰痛症	便秘症	糖尿病
eGFR		−0.3068	−0.1116	−0.17844	−0.09503	−0.05728	−0.10591
UTP			0.075092	0.10691	0.06536	0.061758	0.10591
胃潰瘍				0.125708	0.173091	0.185885	0.110058
高血圧症					0.098686	0.124336	0.180193
腰痛症						0.187169	0.086543
便秘症							0.080726
糖尿病							

表 3.2(b) 生成データの属性間の相関係数 R_D

属性	eGFR	UTP	胃潰瘍	高血圧症	腰痛症	便秘症	糖尿病
eGFR		−0.29341	−0.13274	−0.19779	−0.09944	−0.07273	−0.10323
UTP			0.087371	0.106561	0.071291	0.09896	0.108063
胃潰瘍				0.143143	0.217401	0.175514	0.153528
高血圧症					0.13495	0.15429	0.18084
腰痛症						0.181265	0.094285
便秘症							0.109723
糖尿病							

ビッグデータと生成データのその他の統計値，すなわち最大値・最小値・平均値・標準偏差の一部を表 3.3 (a), (b) に示す．例えば，属性 eGFR の最大値，平均値，標準偏差では，ビッグデータに比べて生成データが誤差 1%以内で再現している．また，最小値の誤差が目立つのは式(2.2)で生成する最小値が $w/2$ となるためと思われる．

表 3.3(a) ビッグデータの属性毎の統計値

	eGFR	UTP	胃潰瘍	高血圧症
最大値	89.99885942	1000	1	1
最小値	0.0000121	−1	0	0
平均値	55.02199458	45.89931	0.214622	0.20367
標準偏差	19.87480914	110.8691	0.41056	0.402726

表 3.3(b) 生成データの属性毎の統計値

	eGFR	UTP	胃潰瘍	高血圧症
最大値	89.99878	999.9986	1.000002	1.000002
最小値	0.849058	3.721698	0.004717	0.004717
平均値	54.85214	43.96945	0.234628	0.191831
標準偏差	19.62445	97.69982	0.419485	0.38887

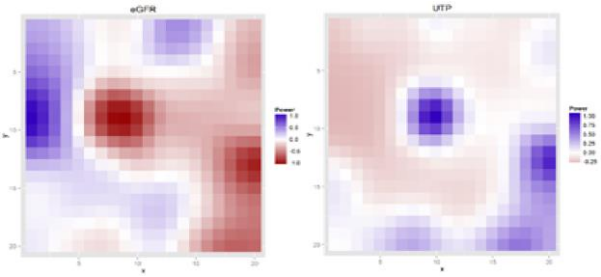
3.3 SOM による評価

続いて，ビッグデータと生成データの自己組織化マップ SOM（Self-Organizing Map）による分析結果を比較する．SOM は，多次元のデータを指定した次元の地図として可視化する技術であり，コホネン(T. Kohonen)により提案された分析法である[6]．多次元の属性を持つデータ群の関係を 2 次元のマップの位置関係で表すことで数値のみから見いだせない関係を明らかにすることができる．

本実験では，SOM 分析に言語 R[9]の SOM パッケージを用いている．マップを表示するグリッド数は 20×20 である．図 3.2(a), (b)にビッグデータ，および図 3.3(a), (b)に生成データについて属性 eGFR と UTP の SOM を示す．

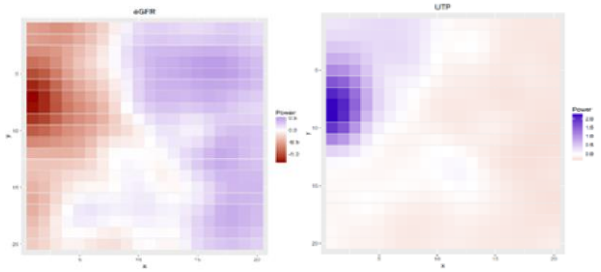
筒井ら[5]によれば，図 3.2(a)にみられる eGFR の青い部分(検査値の高い患者集団)が UTP における赤い部分(検査値の低い患者集団)と反転している様子がみられる．この分析結果は，すでに医療関係者に知られていたことであるが，SOM 分析により初めて視覚化されたものである．

生成データの eGFR と UTP の SOM を図 3.3(a), (b)にそれぞれ示す．図 3.2 と形は異なるが，eGFR と UTP において赤と青の部分における反転がみられ，実際のビッグデータから得られる知見とほぼ一致する分析が可能であることがわかる．



(a) eGFR の SOM　　(b)UTP の SOM

図 3.2　ビッグデータの SOM 分析[5]



(a) eGFR の SOM　　(b)UTP の SOM

図 3.3　生成データの SOM 分析

4　まとめ

本論文において，個人情報を含まない統計情報のみからビッグデータを再生成する方法を提案した．提案手法は，ヒストグラムと相関係数を入力として，モンテカルロ法でヒストグラムを再現し，データレコード順の交換により相関係数を再現するものである．生成されたデータを検証するため，医療データの SOM 分析と比較したところ，生成データからほぼ同等の SOM 分析ができることを確認した．本手法で生成した個人情報保護問題を生じないデータは，今後のビッグデータ分析研究に役立つと期待される．

参考文献

[1] McAfee, Andrew, et al. "Big data: the management revolution." *Harvard business review* 90.10 (2012): pp. 60-68.

[2] Wu, Xindong, et al. "Data mining with big data." *IEEE transactions on knowledge and data engineering* 26.1 (2014): pp.97-107.

[3] Abowd, J., Alvisi, L., Dwork, C., Kannan, S., Machanavajjhala, A., & Reiter, J. “Data Analysis for the Federal Statistical Agencies.” *arXiv preprint arXiv:1701.00752* (2017).

[4] Bello-Orgaz, Gema, Jason J. Jung, and David Camacho. "Social big data: Recent achievements and new challenges." *Information Fusion* 28 (2016):pp. 45-59.

[5]筒井真璃菜，村岡道明，“自己組織化マップによる解析結果を用いた病態の可視化および類似検索システムの構築，”平成 27 年度　高知大学　卒業論文

[6] Kohonen, Teuvo. "Essentials of the self-organizing map." *Neural networks* 37 (2013): pp.52–65

[7] Obayashi, Shigeru, and Daisuke Sasaki. "Visualization and data mining of Pareto solutions using self-organizing map." *International Conference on Evolutionary Multi-Criterion Optimization*. Springer, Berlin, Heidelberg (2003): pp. 796-809.

[8] Siddiqui, Muazzam, Morgan C. Wang, and Joohan Lee. "A survey of data mining techniques for malware detection using file features." *Proceedings of the 46th annual southeast regional conference on XX*, ACM (2008):pp. 509-510.

[9] The R Project for Statistical Computing
https://www.r-project.org/