データ分析支援のための匿名化データ候補自動作成支援機能の検討

石井 陽介† 水野 和彦† 田中 剛†

†株式会社日立製作所 〒185-8601 東京都国分寺市東恋ヶ窪 1 丁目 280

E-mail: †{yohsuke.ishii.bz, kazuhiko.mizuno.pq, tsuyoshi.tanaka.vz@hitachi.com

あらまし データ分析により，多種多様なデータから新たな価値を生み出す取り組みが広がっている．対象データの中には個人情報など匿名化が必要なものがあり，匿名度の確保と匿名化に伴う丸め込みによる分析精度の低下防止を両立することが課題となる．著者らは，分析対象データの匿名化データ候補を機械的に作成し，有用性指標に基づいて選択可能な匿名化データ候補自動作成支援機能を検討した．本稿では，有用性指標として，k 匿名性の k 値，データ欠損率および匿名化データと元データの相関値を利用したケースの評価結果を報告する．

キーワード データ管理，匿名化データ，データ分析，データ欠損

1. はじめに

近年，様々なデータを分析し，その分析結果を利用したデータの利活用への取り組みがなされるようになってきている．データ分析結果を利用することで，現行業務における課題解決ならびに新サービスや新事業創生につなげることが期待されている．

このデータ利活用を進めるためには，当該データをそのままの形式で利用するだけでなく，データの一部を加工することで，その用途を広げる取り組みもなされてきている．例えば，そのままの形式だとプライバシー保護の問題や法規制対応などの問題がクリアできないデータに対して，匿名化などによるデータ加工を実施し，加工データという形式でデータを利活用する取り組みがある．データを匿名化する手法としては，同じ属性を持つデータが一定数以上存在するようにデータを加工することで個人が特定される確率を低減する k 匿名化[1][2][3][4]等がある．しかし，匿名化の際に匿名化のレベルは一通りではない．また，データ利用者による利用目的ならびにデータ提供者の提供条件に基づいて，匿名度の確保と匿名化に伴う丸め込みによる分析精度の低下防止を両立させるために，その加工内容について試行錯誤が必要になる場合もある．このためデータ利活用とデータ匿名化を両立可能なデータ管理の仕組みの構築が必要になってきていると言える．

著者らは，データ利用者がデータ利活用を目的に匿名化データをデータ提供者から提供を受ける現状のフローの問題点に基づき，その改良フローとその実現に必要な匿名化データ候補自動作成支援機能の検討を進めている．匿名化データ候補自動作成支援機能とは，分析対象データの匿名化データ候補を機械的に作成し，有用性指標に基づいて選択可能にするものである．機械的に複数の匿名化データ候補を作成することになるので，その中から有益な候補を容易に選択できるような有用性指標の利用が課題となる．本稿では，本機能実現に向けた検討内容と，選定した有用性指標を利用したケースの評価結果を報告する．

2. 匿名化データ提供の流れ

2.1. 匿名化データの作成

データ分析を実施する際，分析対象データに特定個人に関するプライバシー情報が含まれていると，対象者が意図しない形で当該情報が流布されてしまう可能性があり，その取り扱いには細心の注意を払う必要がある．そこで，対象データにて個人を特定できないように加工し，その加工データを利用してデータ分析を行う形態が広く利用されている．この加工のことを匿名化と呼び，対象データ群の中に，同じ属性を持つデータが k 件以上存在するようにデータを加工する技術のことを k 匿名化と呼ぶ．

表形式の構造化データを対象とした場合の匿名化処理の例を表 1 に示すデータを対象に説明する．表形式のデータには，複数の属性情報で構成されるレコードを基本単位とし，複数のレコードが存在する．属性情報は，識別子，準識別子およびその他情報に分類することができる．識別子とは，その情報だけで特定の個人や対象を特定可能な情報であり，ID や名前などが該当する．準識別子とは，それ単独では個人や対象を識別できないが，組み合わせることで個人や対象を特定可能な情報である．例えば，性別や年齢などが該当することが多い．その他情報には，それら以外の属性が該当し，この中に対象者に関するセンシティブな情報が含まれることがある．

データを加工し，匿名化する手法として，抑制と一般化という手法がある[2]．抑制とは，ある属性の値を画一的に置換し値を隠すことである．例えば，性別欄の属性値を「＊」に置き換えることに相当する．一般化とは，ある属性の値を広義の意味に置換することである．例えば，年齢欄の属性値を「23」から「20-29」に置き換えることに相当する．本検討では，匿名化デ

ータを対象としたデータ分析を行うことが目的であるので，匿名化データを作成する場合は，元データの情報量がなるべく欠損しない形式で加工することが望ましい．そこで，本検討では，一般化手法をメインに匿名化データを作成するケースを想定することにした．

表 1　対象データ例

識別子		準識別子			その他情報			
ID	名前	性別	年齢	・・・	年収	車	住居	・・・
001	A	男	39		1200万円	あり	分譲マンション	
002	B	男	34		540万円	あり	賃貸アパート	
003	C	女	25		670万円	なし	賃貸アパート	
004	D	女	23		440万円	あり	一軒家	

一般化によってデータ加工を行う際，図 1 に示す一般化階層定義を利用することが多い．一般化階層定義とは，対象とする属性の値を一般化の度合いに応じてグループ化し，そのグループ同士の一般化による関係を階層的に示すものである．図 1 はその一例であり，木構造の場合は一般化階層木とも呼ばれる．匿名化処理は，対象属性毎に指定された一般化階層レベルに基づいて実施される．

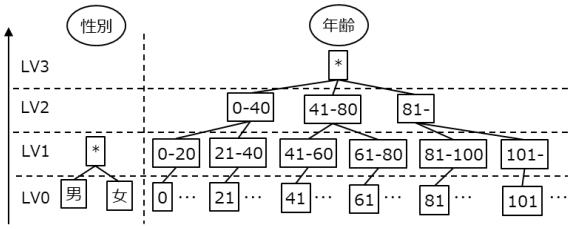

図 1　一般化階層定義の例

2.2. 従来の匿名化データ提供の流れ

データ利用者がデータ分析をする場合，一般的には図 2 に示すようなフローに沿って，対象データを管理する人(データ提供者)の許可を得て，対象データを提供してもらう必要がある．対象データが匿名化を要するデータである場合，対象データを加工して作成した匿名化データは当該データ提供者によって定められた匿名性を担保したものである必要がある．例えば，k 匿名性の k 値が 10 以上といった条件が課されることになる．その一方で，データ利用者にとっては，データ分析の目的を達成するために，匿名化のための加工処理によって，対象データの情報量の欠損がなるべく少なくなるようにしなければならない．具体的には，データ利用者が提示するデータ要件として，データ利用者が分析向けに着目したい属性については一般化階層レベルの値を低くし，それ以外の属性については一般化階層レベルの値を高く設定することになる．当該データ要件を提示されたデータ提供者は，指定の条件で加工して k 匿名性の k 値を算出し，その値に基づいて当該匿名化データの提供可否を判断し，その結果をデータ利用者に提示する．データ利用者は，当該匿名化データが自身の分析に使えるか否かを検証し，使えないもしくは十分な結果が得られないと判断した場合は，再度データ要件を見直したうえで本フローを繰り返す必要がある．この一連のフローは，試行錯誤を伴うことが多く，効率よく実行することが難しい．

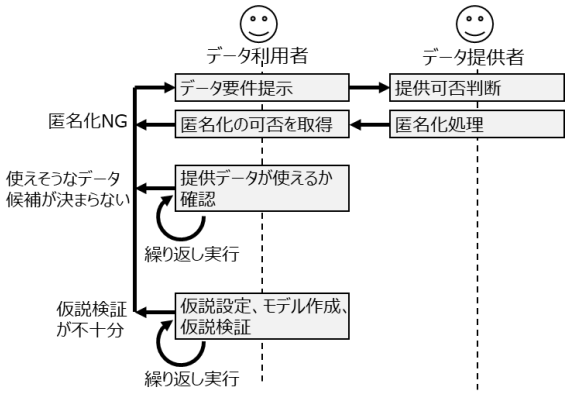

図 2　従来の匿名化データ提供の流れ

3. 匿名化データ候補自動作成支援機能の検討

3.1. 匿名化データ候補自動作成支援機能

前節で説明した従来のフローを改良するために，著者らは匿名化データ候補自動作成支援機能を検討した．

匿名化データ候補自動作成支援機能は，データ利用者が提示したデータ要件に基づいて，匿名化データの候補を機械的に作成する．具体的には，データ要件で示される各属性の一般化階層レベルの値をもとに，対象データの各属性で加工可能な一般化階層レベルの値を全て列挙し，それらを組み合わせることで対象データ候補の加工条件を導出する．

次に，導出した加工条件に基づいて，すべての組み合わせを対象に匿名化処理を実施し，作成した匿名化データそれぞれに対して k 匿名性の k 値を算出する．算出した k 値が所定の値を満たさない場合は，k 値が所定の値になるよう，k-1 以下の匿名性を持つレコード群を削除する．レコード削除によって対象データの情報量は欠損するものの，分析対象データ候補の数を可能な範囲で減らさないようにできる．

次に，作成した匿名化データの一覧をデータ提供者に提示し，作成した匿名化データの中でデータ利用者に提供してもよい匿名化データ候補を選択してもらう．ここでの選択は，k 値で候補を絞り込むことをベースに，必要に応じて個別に除外すべきデータ候補を選択してもらう．

最後に，データ提供者の提供許可を得た匿名化データ候補の一覧をデータ利用者に提示する．以上により，

データ利用者のデータ要件とデータ提供者の提供条件との両方に折り合える匿名化データ候補の一覧を準備することが可能になる.

3.2. 有用性指標による選択

匿名化データ候補自動作成支援機能は，提示されたデータ要件に基づいて匿名化データ候補を機械的に複数作成することになる．匿名化加工対象となる属性数や，各属性で指定可能な一般化階層レベルの種類が多い場合，匿名化データ候補も数百，数千といった数になる可能性がある.たくさんのデータ候補群の中から，データ利用者が分析対象データを選択する際，その選択肢の数が多いとどのデータ候補を選んでよいか判断が難しくなる.

そこで，著者らは，各々の匿名化データ候補に有用性指標を算出し，データ利用者が匿名化データ候補の中から利用対象データ候補を選択する際に参照できるようにした．匿名化データの有用性指標に関する例は多数存在する[5][6]．本検討では，簡便な演算で算出可能な次の 3 つを有用性指標として利用することにした.

・k 匿名性の k 値
・データ欠損率
・匿名化データと元データの相関値

k 匿名性の k 値は，匿名化データ候補自動作成支援機能によって作成された匿名化データ候補から算出した値を利用する．k 値が所定の値に近いほど，有用性が高いと考えることができる.

データ欠損率は，作成された匿名化データ候補が所定の k 値を満たさない際に，匿名化データ候補自動作成支援機能によって対象レコードを削除する場合があり，その時の削除レコード数と元のレコード数との比率から算出する．データ欠損率が低いほど，有用性が高いと考えることができる.

匿名化データと元データの相関値は，はじめに，元データを構成する属性間の相関係数行列ならびに匿名化データを構成する属性間の相関係数行列を算出する.次に，２つの行列の差分の絶対値からなる差分絶対値行列を算出する．次に，差分絶対値行列の上三角部分の行列要素の値の平均，分散を算出する．この平均の値を相関係数差分平均，分散の値を相関係数差分分散と呼び，有用性指標として利用する．相関係数差分平均の値がゼロに近い，または相関係数差分分散の値が小さいほど，元データの相関係数と匿名化データの相関係数の一致度合いが高く，匿名化処理による相関係数分布の変化が小さいと考えられるため，当該匿名化データ候補は元データの特徴を残せている可能性があり，有用性が高いと考えることができる.

3.3. 匿名化データ提供の流れ改良型

匿名化データ候補自動作成支援機能を利用し，前節で説明した従来のフローを改良したフローを図 3 に示す．改良型のフローでは，データ利用者は，提示された匿名化データ候補の一覧の中から，有用性指標を参照して任意のデータ候補を選択して取得し，データ分析を行うことができる．もし，当該データ候補が分析用途に使えない，あるいは分析結果が思わしくなかった場合は，別のデータ候補を選択し，同様にデータ分析を実施する．このようなフローにすることで，データ利用者とデータ提供者との間で発生していたやり取りを減らし，作業を効率化することが期待できる.

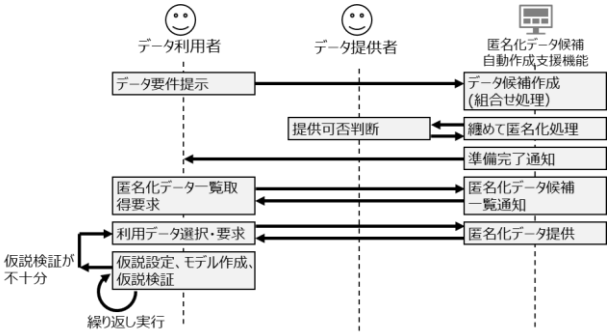

図 3　匿名化データ提供の流れ改良型

4. 匿名化データ候補自動作成支援機能の検証

4.1. 検証方法

匿名化データ候補自動作成支援機能を試作し，その実現可能性と匿名化データ候補に付与する有用性指標の価値を検証した.

匿名化データ候補を作成する処理については，データ要件として提示される一般化階層レベルに基づいて対象データの属性値の加工することで実現可能である.加工処理プログラムを試作し，機能検証することができた．前述した有用性指標の算出についても，試作によって実現できることを確認できた.

有用性指標の価値を検証するために，元データとそれを加工した匿名化データ群に対して同じ手法でデータ分析を実施し，元データを対象に分析した結果から得られる知見と，当該有用性指標を用いて選択した匿名化データを対象に分析した結果から得られる知見が同等になるか否かを検証することにした．もし，有用性指標を用いて選択した匿名化データ候補から同等の知見を得ることができれば，対象データと分析手法が限定的ではあるものの，当該有用性指標に一定の価値があると言えると考える.

今回，対象データとして，機械学習のベンチマークで広く活用されているサンプルデータ[7]を利用した.対象データの属性群の中から，今回の検証のため便宜的に準識別子相当の属性群を 3 つ選択し，それぞれ一

般化階層レベルを 4 つ定義した．定義した一般化階層レベルを図 4 に示す．この一般化階層レベルに基づいて，元データから 4*4*4=64 個の匿名化データ候補を作成した．今回，k 値が 10 以上の条件で匿名化を実施した．

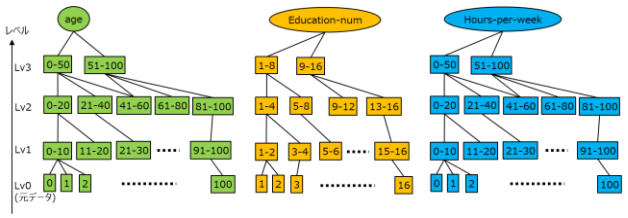

図 4　検証用の一般化階層レベル定義

想定分析シナリオとして，はじめに，対象データのレコード各属性値を表 2 に示すようにカテゴリ変数化し，ダミー変数化したものを説明変数とする．当該レコードの年収属性値が所定の値以上か否かを示すフラグを目的変数とし，ロジスティック回帰を用いてモデル化する．

表 2　属性のカテゴリ変数

属性	元データを対象としたカテゴリ変数	匿名化データを対象としたカテゴリ変数
age	17-30(young), 31-55(middle-aged), 56-100(old)	指定された一般化階層レベルと同じグループでカテゴリ化
workclass	Private, Self, gov, others	同左
education-num	1-6(low), 7-11(medium), 12-16(High)	指定された一般化階層レベルと同じグループでカテゴリ化
marital-status	married, single	同左
occupation	highskill, lowskill	同左
race	white, other	同左
hours-per-week	0-35(low), 36-40(medium), 41-60(high), 61-100(veryhigh)	指定された一般化階層レベルと同じグループでカテゴリ化
その他	カラムの各値をそのまま利用	同左

次に，モデル化で導出した回帰式において，係数が非ゼロの項に着目し，その項に対応する説明変数を特定する．これらの説明変数は，当該モデルによって，年収が所定値以上か否かを推論する際に寄与する説明変数であると言える．そこで，元データを対象に分析した場合に特定した説明変数群と，匿名化データ候補を対象に分析した場合に特定した説明変数群を比較する．これらが同じであれば当該匿名化データ候補からは元データと同等の知見を得ることができたと言える．

検証では，匿名化データ候補の中から，有用性指標を用いて絞り込みならびに整列したリストの中から，任意の件数の匿名化データ候補を選択し，それらを対象に分析し得られる知見として，作成した回帰モデル式で係数が非ゼロとなる説明変数群を調べた．

4.2. 元データを対象とした分析

前述した元データを対象に回帰モデル式を作成し，係数が非ゼロになる説明変数群を抽出した．今回のデータでは，以下の 4 項目を抽出できた．この結果により，対象データから立案した回帰モデル式から得られる知見として，対象者の年収が所定値以上か否かを推論する際に，以下の 4 項目の影響が大きいという知見を得たことになる．

- Age
- Education Num
- Marital Status
- Occupation

4.3. 匿名化データ候補を対象とした分析

匿名化データ候補自動作成支援機能によって作成した 64 個の匿名化データ候補を対象に回帰モデル式を作成し，係数が非ゼロになる説明変数群を抽出した．その結果を表 3 に示す．64 個の匿名化データ候補の中で，元データと同じ知見，すなわち抽出した説明変数群が AEMO(Age, Education Num, Marital Status and Occupation)だったものを得られた匿名化データ候補の件数は 20 件であった．

表 3　匿名化データ候補の分析結果

抽出した説明変数群	匿名化データ候補の件数
AEMO	20
AMO	8
EMO	25
HMO	1
EM	2
MO	7
M	1

A:Age, E:Education Num, M:Marital Status, O:Occupation, H:Hours per week

次に，有用性指標を利用して匿名化データ候補を絞り込みならびに整列した場合における結果を検証する．絞り込み/整列条件とその条件に合致する匿名化データ候補の上位 5 件について，抽出できた説明変数群を表 4 に示す．

表 4　有用性指標を利用して選択した匿名化データ候補の分析結果

#	有用性指標による絞り込み/整列条件	上位 5 件の中で、元データと同じ知見を得られた件数	抽出した説明変数群の内訳
1	K値が10以上 and データ欠損率で昇順整列	0	EMO:2, MO:3
2	K値が10以上 and 相関係数差分分散で昇順整列	1	AEMO:1, AMO:3, EMO:1
3	K値が10以上 and データ欠損率が5%以下 and 相関係数差分分散で昇順整列	1	AEMO:1, AMO:3, EMO:1

条件#1 では，データ欠損率が低いものを抽出条件とし，結果は上位 5 件に所望の知見を得られた匿名化データ候補は含まれなかった．この場合，一般化階層レベルの値が高い組合せからなる匿名化データ候補が該当したため，匿名化加工によるデータの丸め込みの影

響が大きくなったと考えることができる．

条件#2 では，相関係数差分分散が小さいものを抽出条件とし，結果は上位 5 件に所望の知見を得られた匿名化データ候補は 1 件であった．上位 5 件に同等の知見を得られる匿名化データ候補が含まれていることで，匿名化データ候補自動作成支援機能が作成したすべての匿名化データ候補をしらみつぶしに探すのではなく，有用性指標に基づいてあたり付けをして効率的にデータ分析を進められる可能性があると言える．ただ，同じ知見を得た匿名化データ候補のデータ欠損率が約 11%であり，欠損データによる副作用も考えられる．このため，データ欠損率も絞り込み条件に追加し検証することにした．

条件#3 では，条件#2 にデータ欠損率が 5%以下という条件を追加した．結果は，結果は上位 5 件に所望の知見を得られた匿名化データ候補は 1 件であった．同じ知見を得た匿名化データ候補のデータ欠損率が約 2%であり，欠損データによる副作用は比較的小さいと考えられる．

以上の検証結果により，有用性指標による絞り込みならびに整列を利用することで分析対象とする匿名化データ候補数を絞り込み，それらを優先的に利用してデータ分析を行うことで，元データを対象に分析した際に得られる同等の知見を効率よく取得できる可能性があることがわかった．匿名化データ候補自動作成支援機能を利用するにあたり，有用性指標に基づいて対象データを選択可能にすることに価値があると言える．

5. 関連研究

データ分析などデータ利活用目的で匿名化データを提供する際に生じる課題に対して様々な取り組みがなされている．文献[8]では，データベース管理システムに格納されたデータを一般化階層定義における任意の要求レベルに対応した加工をする際に，非順序型実行原理を活用して処理を高速化している．これにより，大規模データを加工に要する時間を短縮することが可能になり，匿名化データ提供に要する期間短縮に貢献し，データ利用者とデータ提供者が対話的に提供すべき匿名化データの仕様の落とし込みとその匿名性の確認を効率よく行うことができる．また，文献[9]では，対象データの分布を正規分布と仮定した累乗近似式に基づく k 匿名性予測式を提供している．本予測式を使うことで，匿名化データ加工のために指定する一般化階層定義における任意の要求レベルに対して，当該レベルで加工したデータの k 匿名性が達成可能か否かを簡便に判定できる．データ利用者は，データ提供者に対して，要求レベルを指定して匿名化加工したデータ提供を依頼する前に，当該要求レベルでデータ提供者が所望する K 匿名性が達成できるかどうかを事前に予測することで，無駄な匿名化加工データ提供依頼回数を削減することができる．

著者らは，データ利用者が提示する一般化階層定義における要求レベルに基づいて，匿名化データ候補を機械的に抽出し，匿名化処理前後の情報損失や分布の変化などの有用性指標とともに匿名化データ候補群を提示する方式を提案した．本方式により，匿名化データの加工条件がデータ利用者のデータ分析目的に合致するか否か，ならびに当該加工条件に基づいて加工された匿名化データがデータ提供者の提供条件に合致する k 匿名性を満たすか否かのすり合わせを容易にできる．また，有用性指標を活用することで，匿名化データ候補の絞り込みを容易にできる．

6. まとめ

本稿では，データ利用者とデータ提供者との間における匿名化データのやり取りの問題点を改善するための匿名化データ候補自動作成支援機能について説明した．本機能の実現にあたり，有用性指標による匿名化データ候補の選択の実現可能性とその提供価値について，サンプルデータとサンプル分析シナリオに基づいた検証を実施した．その結果，有用性指標として，k 匿名性の k 値，データ欠損率および匿名化データと元データの相関値を利用することで，匿名化データ候補の絞り込みを容易にできる可能性があることがわかった．

今後の課題として，より汎用的に適用できるような有用性指標の追加やブラッシュアップについて取り組む必要があると考える．また，様々なデータや分析シナリオを対象に，元データと同等の知見を得られる匿名化データ候補抽出の実現可能性についても検証する必要があると考える．

参　考　文　献

[1] Grigorios Loukides and Jianhua Shaom, “Data utility and privacy protection trade-off in k-anonymisation,” In Proceedings of the 2008 international workshop on Privacy and anonymity in information society (PAIS '08). Association for Computing Machinery, New York, NY, USA, pp.36–45 (2008).

[2] Latanya Sweeney, “Achieving k-anonymity privacy protection using generalization and suppression,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 10, No.5, pp.571–588, (2002).

[3] Latanya Sweeney, “A Model for Protecting Privacy,” International Journal on Uncertainty, Fuzziness and Knowledgebased Systems, Vol.10, No.5, pp.557-570, (2002).

[4] 原田邦彦，佐藤嘉則，“一般化階層木の自動生成と

情報エントロピーによる歪度評価を伴う k-匿名化手法," 情報処理学会研究報告, Vol.2010-CSEC-50 No.47, (2010).

[5] Ayala-Rivera, Vanessa & Mcdonagh, Patrick & Cerqueus, Thomas & Murphy, Liam & Thorpe, Christina, "Enhancing the Utility of Anonymized Data by Improving the Quality of Generalization Hierarchies," Transactions on Data Privacy. Vol.10, No.1, pp.27-59, (2017).

[6] 村本俊祐,上土井陽子, 若林真一, "データを極小歪曲し k-匿名性を保持したデータに変換するプライバシー保護アルゴリズム," 日本データベース学会 Letters (DBSJ Letters) 6 巻 1 号, pp.97-100, (2007).

[7] US Adult Income: Salary Prediction, https://www.kaggle.com/marksman/us-adult-income-salary-prediction (2020/12/11 アクセス確認).

[8] 西川記史, 礒田有哉, 茂木和彦, 清水晃, 早水悠登, 合田和生, 喜連川優, "非順序型データベースエンジンを用いた大規模データの対話的非特定化手法の性能評価," 第 11 回データ工学と情報マネジメントに関するフォーラム DEIM Forum 2019 J3-2, (2019).

[9] 小栗秀暢,曽根原登,松井くにお,モハマド ラスール サラフィ アグダム, "累乗近似式を用いた k-匿名化処理の効率化," 情報処理学会論文誌 Vol.57, No.9, pp.2034-2044, (2016).