

モバイル環境センシングにおける Perturbation 後の 復元精度推定手法の提案と評価

鈴木 孝男^{1,a)} 伊藤 昌毅^{1,b)} 瀬崎 薫^{1,2,c)}

概要：スマートフォンに代表される、様々なセンサを搭載したモバイルデバイスを持つ一般ユーザーがセンシングに参加するモバイルセンシングにおいて、センサデータからユーザーのプライバシー情報が流出してしまう危険性がある。Perturbation を用いるプライバシー保護手法は、ユーザーがセンシングによって取得した真のデータに対してノイズ付加などの偽装処理を施して生成した偽装データを提供することで、ユーザーのプライバシーを保護しつつ、データ回収者はこれら偽装データを大量に集め、特定の処理を加える事で、真のデータ分布を復元することができる。この復元精度は回収するデータ数に依存するという性質を持つため、データ回収者は所望の復元精度を満たすためにユーザーの参加を促すインセンティブ付与をする必要があると考えられるが、その際現在の復元精度を知ることができず、適切なインセンティブ付与ができないという問題が存在する。本稿では、環境情報を測定対象としているモバイル環境センシングにおいて、環境情報が時空間的に相関を持つことに着目し、Perturbation 後の復元精度を推定する手法を提案する。オープンデータの解析により、環境情報の一例である一酸化炭素の値が対数正規分布に高い確率で従うことを確認し、それを利用した復元精度の推定手法の提案及び評価を行った。

キーワード：モバイルセンシング、プライバシー保護、Perturbation、復元精度、環境情報

1. はじめに

スマートフォンに代表される、様々なセンサを搭載したモバイルデバイスが普及したことにより、一般ユーザーがセンシングに参加するモバイルセンシングに注目が集まっている [1]。モバイルデバイスを持ち歩きながら取得したセンサデータを多くのユーザーに提供してもらうことで、大規模にデータを取得でき、それらデータを解析することによって統計データや新たな知見を得ることが期待されている。モバイルセンシングはユーザーの動きをセンシングするものと、周囲の環境情報をセンシングするものの2つに大きく分けることができ、本稿では後者のセンシング手法をモバイル環境センシングと呼ぶ [1]。モバイル環境センシングについては多くのアプリケーションが提案されており、都市空間の大気汚染情報や、騒音情報、気候情報などが測定対象の例として挙げられる [2], [3], [4], [5]。

このようなモバイル環境センシングにおいて問題となるのは、センサデータからユーザーのプライバシー情報が流出するリスクがあることである。スマートフォンなどのモバイルデバイスは基本的にユーザーと同じ空間を共有しているため、それらデバイスから送られてくるセンサデータがユーザーのプライバシー情報に結びついている可能性は高い [6]。また、近年センサデータを用いたユーザーに関するコンテキスト推定技術が盛んに研究されており、このような背景からもモバイル環境センシングにおいてプライバシーを保護することがアプリケーションの普及には不可欠であると言える。

モバイルセンシングにおいて多くのプライバシー保護手法が提案されているが、その中でも Perturbation を利用したプライバシー保護手法は、第三者機関に頼らずユーザーの手元でプライバシー保護処理が可能であり、また、個々のデータを知ることなく必要な統計情報を抽出できる点で他の手法よりも優れている。この手法は、真のセンサデータに対してノイズ付加などの偽装処理を施すことでセンサデータからユーザーのプライバシー情報が流出することを防ぐことができる。アプリケーション提供者は多くのユーザーからこれら偽装データを回収し、特定の処理を施すことで、真のセンサデータの分布を復元できる [7], [8], [9], [10]。

¹ 東京大学 生産技術研究所
Institute of Industrial Science, The University of Tokyo

² 東京大学 空間情報科学研究センター
Center for Spatial Information Science,
The University of Tokyo

a) t-suzuki@mcl.iis.u-tokyo.ac.jp

b) mito@iis.u-tokyo.ac.jp

c) sezaki@iis.u-tokyo.ac.jp

Perturbation によって偽装したデータを回収した後に復元処理において、その復元精度は回収したデータ数に依存するという性質を持つ。そのため、データ数が少ないと正確に真のデータ分布を復元することができず、間違った情報を取得することに繋がるため、現在の復元精度に応じてユーザーに対してお金などの報酬をインセンティブとして付与し、追加のデータを提供してもらうことで所望の復元精度を満たすシステムが必要であると考えられる。

このようなシステムを実現する上で問題となるのが、Perturbation を用いた場合、元々の真のデータ分布を知ることができないため、復元処理を行った後に得られる分布の復元精度が分からないことである。現在の復元精度が分からないと、追加のデータが必要なのかどうか、どの程度必要なのかといったことが判断できず、適切なインセンティブ付与ができない。

本稿ではこの問題の解決手法として、モバイル環境センシングで測定対象となる環境情報が従うモデルを利用して復元精度を推定する手法を提案する。環境情報は時間的、空間的相関を持つあるモデルに従うことも多いと考えられる。このような測定対象が従う時空間的なモデルがあらかじめ既知であれば、そのモデルを利用することで真のデータ分布を知らなくても復元精度を推定できると考えられる。本稿では大気汚染をモニタリングするモバイル環境センシングを想定し、大気汚染の原因の1つである一酸化炭素のオープンデータの解析を行った。解析の結果、対数正規分布に従う確率が高いことを示し、このことを利用した復元精度推定手法の提案及び評価を行った。Perturbation を利用したプライバシー保護手法はいくつか存在しているが、本稿では Negative Surveys を用いて評価を行った [10]。

本稿の構成は以下の通りである。2章で関連研究について述べ、3章で Perturbation を用いたプライバシー保護手法の問題について述べる。4章で復元精度推定のための提案手法の説明と、オープンデータの解析について述べる。4章でのデータ解析の結果を元に詳細な復元精度推定手法を5章で提案した後、手法の評価を6章で行う。考察を7章で行い、8章でまとめを行う。

2. 関連研究

2.1 モバイル環境センシング

モバイル環境センシングとは、スマートフォンなどのモバイルデバイスを用いて、周囲の大気汚染度や騒音などの環境情報を取得するセンシング手法である。専用のモバイルデバイスを用いて都市空間の大気汚染度を測定する HazeWatch[2] や Common Sense[3] といったシステムや、NoiseTube[4] や Ear-Phone[5] など、スマートフォンに搭載されているマイクを用いて周囲の騒音状況をモニタリングするシステムなどが提案されている。このような環境センシングにおいては、収集データから求められるデータ分

布や平均などの統計情報が重要となる。

2.2 モバイル環境センシングにおけるプライバシー問題

スマートフォンなどのモバイルデバイスはユーザーと同じ空間を共有しているため、それらデバイスから送られてくるセンサデータがユーザーのプライバシー情報に結びついている可能性は高く [6]、また、センサから得られる環境情報から個人のコンテキストを推定する研究が盛んになされている。[11] ではユーザーが生活のなかで取得した二酸化炭素濃度の推移を見ることによって、ユーザーが何をしているかといったコンテキスト情報が取得できる可能性を指摘している。また、単一のセンサ以外に、他のセンサの値や公開情報を組み合わせることで得られる情報量が増大し、より高度にユーザーのコンテキストを推定できる。一例として [12] では気温や湿度など、複数の環境情報を利用することでユーザーの位置を特定する手法を提案している。センサの組み合わせ以外にも、センサ値の特異値を利用したり、大気汚染度と健康状態のような相関情報を利用することでユーザーのプライバシー情報が流出する危険性も指摘されている [13]。

このような背景からユーザーが安心してデータを提供できる土壌を整えるために、モバイル環境センシングにおいてプライバシー保護を行うことは重要であるといえる。

2.3 モバイルセンシングにおけるプライバシー保護手法

データを匿名化することによって、データとユーザーの結びつけをなくし、ユーザーのプライバシーを保護する手法が数多く提案されている。 k 匿名化手法 [14] はその一例であり、あるデータが誰の情報なのか k 人未満に絞ることをできなくする手法で、ユーザーから得られたデータを元に作成したデータベースに対して処理を行うものである。また、MIX Network[15] などの送信者の秘匿化技術を用いたモバイルセンシングにおける匿名化手法も提案されている [16]。しかし、これら匿名化手法ではアプリケーション提供者や第三者機関などに自分のデータが送られた後にプライバシー保護処理が行われるので、それら機関に悪意がある場合にユーザーのプライバシーが侵害される危険性がある。ユーザーが安心してセンシングに参加するにはユーザーの手元でプライバシー保護処理を行うことが望ましい。

ユーザーの手元でプライバシー保護処理を行う手法に Perturbation を用いる手法がある。Perturbation を利用したプライバシー保護手法は、偽装データをアプリケーション提供者に送信することでセンサデータからユーザーのプライバシー情報が流出することを防ぎつつ、アプリケーション提供者は多くのユーザーからこれら偽装データを回収し、復元処理を施すことで、真のセンサデータの分布を復元することができる。個々のデータを知ることなく、必

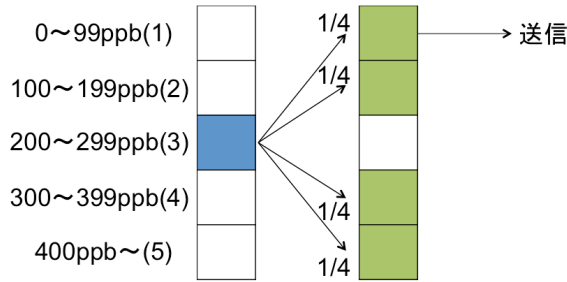


図 1 Negative Surveys の偽装処理例

要な統計情報のみを抽出できる点でも、モバイル環境センシングに適したプライバシー保護手法である。[7], [8], [9]では、センサデータにノイズを付加してデータの偽装を行う Perturbation 手法を提案している。

本稿では、提案手法の評価のために Perturbation 手法の1つである Negative Surveys を用いている。次節で Negative Surveys について説明する。

2.4 Negative Surveys

Negative Surveys は Perturbation 手法の1つであり、アンケート調査などで用いられるプライバシー保護手法である。複数選択肢（カテゴリ）がある質問に対して回答者は常に自分の真の答えとは違う偽のカテゴリを当確率で選んで回答することでプライバシーを保護しつつ、集計者は得られた回答を復元することで全体的なデータ分布を得る手法のことである [10]。以下では説明のために、 X を真のデータのカテゴリ分布、 X_i を真のデータのカテゴリ i の集計、 Y を偽装処理後のカテゴリ分布、 Y_i を偽装処理後のカテゴリ i の集計、 A を復元後のカテゴリ分布、 A_i を復元後のカテゴリ i の集計、 N をデータ数、 α をカテゴリ数として表記する。

本来連続値を取るセンサデータをカテゴリ化して扱うことで Negative Surveys をプライバシー保護手法として適応できる。図 1 に Negative Surveys の偽装処理例を示す。この図では、測定対象に一酸化炭素を想定しており、一酸化炭素の値に対応するカテゴリ値を括弧内に記述している。例えば一酸化炭素の測定値が 250 ppb である場合、そのカテゴリ値は 3 となる。Negative Surveys では常に真でないカテゴリを等確率で選択することでデータを偽装するので、図 1 の例ではカテゴリ 3 以外の 4 つのカテゴリを等確率にランダムに 1 つ選択し、送信する。

アプリケーション提供者はこれら偽装データを集計し、以下の式 (1) による処理を行うことで、真のカテゴリデータの分布を復元する。

$$A_i = N - (\alpha - 1) Y_i \quad (1)$$

図 1 と同じカテゴリ区分 $\alpha = 5$ を持ち、 $N = 10000$ として生成したデータに Negative Surveys を施し、復元した例

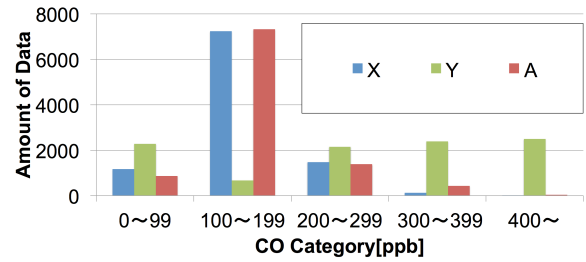


図 2 Negative Surveys の復元処理例

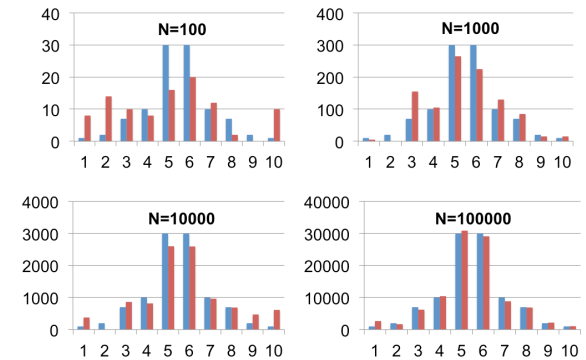


図 3 複数のデータ数での Negative Surveys の復元例。いずれの図も横軸がカテゴリ、縦軸がデータ数を表し、青色の棒グラフが X 、赤色の棒グラフが A である。

を図 2 に示す。 A は真のデータ分布 X を復元できていることが分かる。また、Negative Surveys における復元精度を、式 (2) のように X と A の誤差 $TrueError$ で定義する。 $TrueError$ が 0 に近いほど復元精度が高い。このときデータ回収者は X を知り得ないので、厳密な復元精度を知ることはできない。

$$TrueError = \frac{\sum_{i=1}^{\alpha} |X_i - A_i|}{N} \quad (2)$$

3. Perturbation を用いたプライバシー保護手法の問題

Perturbation によって偽装したデータを回収した後に行う復元処理において、その復元精度はデータ数に依存するという性質を持ち、データ数が多いほど復元精度が良くなる。図 3 に Negative Surveys においてデータ数を変えて復元した例を示す。この図では後述する手法で A の負の値を補正している。データ数が多くなるほど復元精度が良くなっていることが分かる。逆にデータ数が少ないと正確にデータ分布を復元することができず、間違った情報を取得することに繋がるため、現在の復元精度に応じてユーザーに対してお金などの報酬をインセンティブとして付与し、追加のデータを提供してもらうことで復元精度を向上させるシステムが必要であると考えられる。

このようなシステムを実現する上で問題となるのが、Perturbation を用いた場合、元々の真のデータ分布を知ることができないため、復元処理を行った後得られる分布の

表 1 各地点, 各月の一酸化炭素データ数

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	total
AMT	2022	1878	2246	1834	2449	2398	2359	2358	2239	2278	1975	2220	26256
BAO	2479	2232	2234	2230	2179	2393	2480	2480	1995	1983	2154	2476	27315
SCT	2478	1210	2479	2400	2480	1781	2479	2457	2278	2480	2193	2480	27195
SNP	1835	2156	2429	2394	2198	147	2400	1745	1960	1999	2399	2223	23885
WBI	2480	2239	2480	2338	2480	2200	2186	2320	1959	2394	2388	2438	27902
WGC	2121	1848	1801	2053	2174	2129	2201	1993	1238	1190	2022	2047	22817
WKT	2480	2218	2496	2086	2359	2399	2473	2449	2132	2477	2400	2360	28329

復元精度が分からないことである。現在の復元精度が分からないと、追加のデータが必要なのかどうか、どの程度必要なのかといったことが判断できず、適切なインセンティブ付与ができない。

Negative Surveys のようにデータ数から復元精度を推測できる Perturbation 手法も存在するが [17], データ数のみでは測定データの復元精度を推定できない Perturbation 手法も存在する。例えば Negative Surveys を拡張し、＜位置情報、環境情報＞のような多次元データの相関を維持しながらプライバシー保護を実現する Multidimensional Negative Surveys [18] では、特定の位置の環境情報の復元精度をその位置のデータ数のみで推定することはできない。さらに、[7], [8], [9] のように、データにノイズを付加して偽装処理を行う Perturbation は、同じデータ数でも付加するノイズによって復元精度が変わってくる。これらノイズは [8], [9] のように複数のパラメータを用いて生成されることもあり、それら複数パラメータやデータ数と復元精度の対応付けは筆者の知る限りなされておらず、復元精度の推定ができないと考えられる。

そこで本稿では、モバイルセンシングの中でも特にモバイル環境センシングにおいて Perturbation を用いてプライバシー保護を行った場合の復元精度の推定手法を提案する。

4. 復元精度推定手法の提案とデータ解析

Perturbation 後の復元精度推定手法として、測定対象のモデルを利用するという方法を提案する。モバイル環境センシングで測定されている環境情報は時間的、空間的な相関に起因する特定のモデルに従うことも多いと考えられる。例えば大気汚染度や天気などは、ある地点での値がインパルス的に急激に変化することはめったになく、一定の時間窓内では似た値を取る時間的相関を持つ。また、近くの地点で測定されたデータは似た値を取る空間的相関も持つ。あらかじめこのような時空間的相関に起因する測定対象のモデルが既知であれば、そのモデルと復元後のデータを比較することで Perturbation を用いた場合でも復元精度を推定できると考えられる。

そこで実際のモバイル環境センシングで測定対象となる時空間データがどのようなモデルに従うかを実データを解析することによって調べた。測定対象として一酸化炭



図 4 一酸化炭素の測定地点, 測定高度

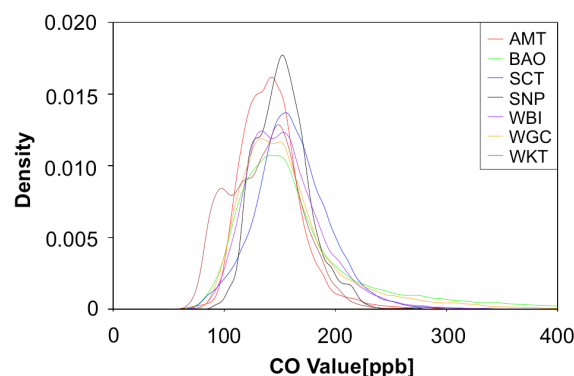


図 5 各地点における年間一酸化炭素データの確率密度曲線

素を取り上げる。一酸化炭素は大気汚染の原因の 1 つであり、人々の健康に影響を与える。このことから都市空間の大気汚染をセンシングする必要があるとして、実際のモバイル環境センシングにおいても測定対象としているアプリケーションが存在している [2], [3]。今回はアメリカ海洋大気局の Earth System Research Laboratory が公開しているアメリカ大陸の各地点でセンシングした一酸化炭素データを解析した [19]。解析対象として選んだデータの測定地点、測定地点名、及び測定高度の詳細を図 4 に示す。解析したデータの測定期間は 2010 年の 1 月から 12 月であり、各月、各場所のデータ数を表 1 に示した。1 日のデータ数の平均は 76 個である。また、データの概要を示すために、各地点の年間のデータの確率密度曲線を図 5 に示した。

モバイル環境センシングを想定する場合に、センシングのサイクルを決めて、そのサイクルごとにデータを集計し分布を知るといった状況が考えられる。このような状況のもと、センシングサイクルを 1 日として 1 日毎の一酸化炭

表 2 各地点の 1 日のサンプルが各確率分布に従う割合

	正規分布 [%]	対数正規分布 [%]	ガンマ分布 [%]
AMT	90.6(308/340)	92.4(314/340)	92.4(314/340)
BAO	56.3(198/352)	69.3(244/352)	65.3(230/352)
SCT	66.3(228/344)	71.8(247/344)	69.2(238/344)
SNP	69.3(212/306)	69.9(214/306)	70.6(216/306)
WBI	70.5(251/356)	75.8(270/356)	73.3(261/356)
WGC	72.4(257/355)	78.3(278/355)	76.9(273/355)
WKT	83.5(298/357)	87.7(313/357)	86.8(310/357)
全体	72.7(1752/2410)	78.0(1880/2410)	76.4(1842/2410)

素データの分布がどのような確率分布に従うかをパラメトリックに推定した。その際、考えられる母集団のパラメトリックな確率分布として、図 5 及び日毎の分布の様子から、正規分布、対数正規分布、ガンマ分布を選択した。まず 1 日のデータのサンプルに対して、上記で述べた 3 つの確率分布に当てはめる場合の最適なパラメータをパラメトリックな推定手法である最尤推定によって求め、母集団がそのパラメータを持つ確率分布に従うとして、母集団の確率分布とサンプルの確率分布が一致するかどうかの検定をコルモゴロフ・スミルノフ検定を用いて行う。これら操作には R を使い、最尤推定は MASS パッケージの `fitdistr` 関数、コルモゴロフ・スミルノフ検定は `ks.test` 関数によって行った。地域毎に 1 年間この処理を行い、指定した母集団の確率分布に従うと検定された日数を集計し、それを 1 年間の日数で除算した結果を表 2 に示す。

この結果より 3 つの分布いずれの場合に当てはめても、各地点とも 1 年のうち 70% 以上の日数が各確率分布に従うと分かる。その中でも特に対数正規分布は平均で 78.0%、最大で 92.4% と、他の 2 つの分布よりも適合していることがわかった。このように 1 日間の一酸化炭素データの分布が対数正規分布に高い確率で従うのは、時間的相関に起因していると考えられる。1 日という短い測定期間において一酸化炭素の値が大きく変化しないため、対数正規分布のようなピークがある確率分布に従う。

以下では実際のモバイル環境センシングを想定して、このようなモデルを利用した具体的な Perturbation 後の復元精度推定手法を提案する。

5. 復元精度推定手法

4 章のオープンデータの解析の結果、環境情報の一例である一酸化炭素が対数正規分布に高い確率で従うことを示した。そこで本章では、

- 一酸化炭素を測定するモバイル環境センシングを想定し、センシングサイクルを 1 日とする
- 1 日間の一酸化炭素の分布は対数正規分布に従う
- プライバシー保護手法に Negative Surveys を用いるという設定のもと Negative Surveys の復元精度である式 (2) で表される $TrueError$ を復元後のカテゴリデータ分

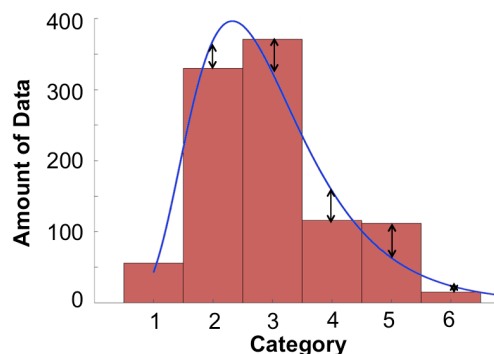


図 6 推定手法 1

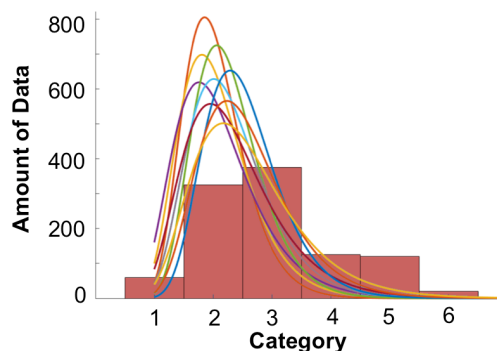


図 7 推定手法 2

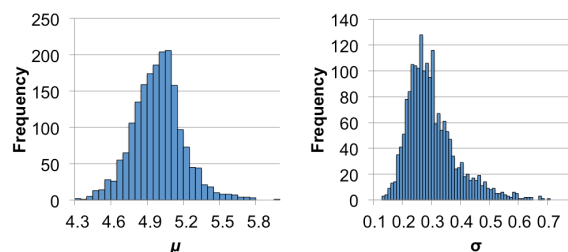


図 8 対数正規分布に従うと判断されたデータの各パラメータの分布

布である A からパラメトリックに推定する具体的な手法を 2 つ提案する。Perturbation 手法の一例として Negative Surveys を用いているが、提案手法は他の Perturbation 手法にも適用できる。以下では $f_{LN}(x|\mu, \sigma)$ を対数平均 μ 、対数標準偏差 σ をパラメータとする対数正規分布の確率密度関数とする。また、 $TrueError$ の推定値を $EstError$ と定義する。

手法 1 点 $(1, A_1), (2, A_2), \dots, (\alpha, A_\alpha)$ から A を最尤推定によって対数正規分布でフィッティングする際のパラメータ $(\mu_{fit}, \sigma_{fit})$ を求め、その対数正規分布と確率密度化した A との誤差を $EstError$ とする (図 6 参照)

$$EstError = \sum_{i=1}^{\alpha} \left| f_{LN}(i|\mu_{fit}, \sigma_{fit}) - \frac{A_i}{N} \right| \quad (3)$$

手法 2 μ, σ の取りうる範囲のパラメータを持つ対数正規分布を複数生成し、それらと確率密度化した A との誤差の平均値を $EstError$ とする (図 7 参照)。式 (4)

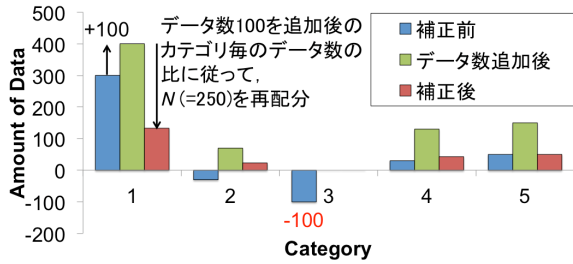


図 9 A の負の値補正例

において $n_\mu + 1$ と $n_\sigma + 1$ はそれぞれ推定に用いる μ と σ の個数である。

$$EstError = \frac{\sum_{i=0}^{n_\mu} \sum_{j=0}^{n_\sigma} \sum_{k=1}^{\alpha} |f_{LN}(k|\mu_i, \sigma_j) - \frac{A_k}{N}|}{(n_\mu + 1) \times (n_\sigma + 1)} \quad (4)$$

手法 2 については、測定対象によって取りうるパラメータの範囲が限られることを利用している。一例として 4 章で行ったデータ解析において対数正規分布に従うと判定されたデータの μ と σ のヒストグラムを図 8 に示した。この図を見ると各パラメータとも偏りがあることが見て取れる。このような特性を利用することで測定対象が従う対数正規分布に当たりをつけて、復元精度を推定できると考えられる。

6. 評価

シミュレーションを行い、提案手法による推定精度の評価を行った。以下では 4 章で行ったデータ解析において対数正規分布に従うと判定されたデータの μ の平均と標準偏差、 σ の平均と標準偏差をそれぞれ、 $\bar{\mu}$, μ_{sd} , $\bar{\sigma}$, σ_{sd} と記述する。

6.1 シミュレーションの流れ

シミュレーションの流れは以下である。

- (i) $f_{LN}(x|\bar{\mu}, \bar{\sigma})$ に従う一酸化炭素データを N 個生成する。
- (ii) 生成されたデータをカテゴリ数 α のカテゴリデータに変換して X を生成する。その際、 $\frac{CO_{max}}{\alpha-1} \times (i-1)$ 以上 $\frac{CO_{max}}{\alpha-1} \times i$ 未満の値のデータをカテゴリ i とし ($i = 1, \dots, \alpha-1$)、 CO_{max} 以上の値のデータをカテゴリ α とする。
- (iii) X の個々のデータに Negative Surveys の偽装処理を施して Y を生成する。
- (iv) 式 (1) に従って復元処理を行い、 A を生成する。この際データ数が少ないと A_i の値が負となることがあり、この負の値を補正する必要がある。補正手法として [18] を参考に、 $\min(A_i) < 0$ ($i = 1, \dots, \alpha$) の場合、全カテゴリのデータ数に $|\min(A_i)|$ を加える事で全てのカテゴリのデータ数を 0 以上にし、そのデータ数の比で N を再配分するという処理を施した。補正の一例を図 9 に示す。この補正処理により、手法 1 での対

表 3 シミュレーションに用いたパラメータ

α	4, 7, 15, 50
N	100, 1000, 10000, 100000, 1000000
$\bar{\mu}$	4.99020284
$\bar{\sigma}$	0.301474648
μ_{sd}	0.21004708
σ_{sd}	0.086562118
n_μ	10
n_σ	10
loop	100
CO_{max} [ppb]	400

数正規分布へのフィッティングが可能となる。

- (v) X と A を用いて復元精度の真値 $TrueError$ を求める。また、 A を用いて式 (3) と式 (4) で表される復元精度の推定値 $EstError$ を求め、それぞれ評価指標

$$Error = |TrueError - EstError| \quad (5)$$

を求める。 $Error$ は 0 に近いほど提案手法による推定精度が良いことになる。

- (vi) (i) から (v) の一連の流れを loop 回繰り返す、 $Error$ の平均値と標準偏差を求める。

シミュレーション (v) で手法 2 を用いて $EstError$ を求める際、対数正規分布のパラメータ μ_i と σ_j を設定する必要がある。今回行ったシミュレーションでは、図 8 を参考に、 μ , σ 共に平均から 2 標準偏差内にパラメータの範囲を絞り、その範囲を各々 n_μ , n_σ 個に等分したときの値を用いた。 μ_i , σ_j はそれぞれ以下の式 (6) (7) で表される。

$$\mu_i = \bar{\mu} - 2\mu_{sd} + i \times \frac{4\mu_{sd}}{n_\mu} \quad (i = 0, 1, \dots, n_\mu) \quad (6)$$

$$\sigma_j = \bar{\sigma} - 2\sigma_{sd} + j \times \frac{4\sigma_{sd}}{n_\sigma} \quad (j = 0, 1, \dots, n_\sigma) \quad (7)$$

6.2 結果

カテゴリ数 α 、データ数 N を変化させてシミュレーションを行い、提案手法の性能評価を行った。用いたパラメータを表 3 に記す。図 5 を参考にして、 CO_{max} を 400 ppb とした。シミュレーションの結果を図 10 に示す。図 10 の凡例において括弧内は用いた推定手法を表しており、手法 1 を用いた場合は実線を、手法 2 を用いた場合は破線を用いて描画し、平均値からの標準偏差をエラーバーで表している。縦軸の $Error$ は式 (5) の評価指標であり、0 に近いほど復元精度の推定性能が良いことを表す。この結果を見ると、 $\alpha = 4$ 以外のカテゴリにおいては、 N が小さい場合は手法 2 を、 N が大きい場合は手法 1 を用いると精度よく推定できることが分かる。特に N が大きい場合は誤差なく復元精度を推定できていることが分かる。また、 $\alpha = 4$ を除く各カテゴリにおいて、2 つの手法が交わる点の $Error$ が 0.2 から 0.3 程度となっていることが読み取れ、2 つの手法を組み合わせることで $Error$ を 0.2 から 0.3 以下

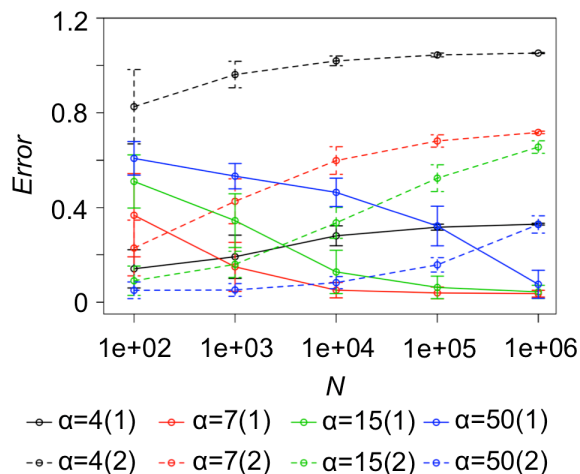


図 10 推定手法の性能評価

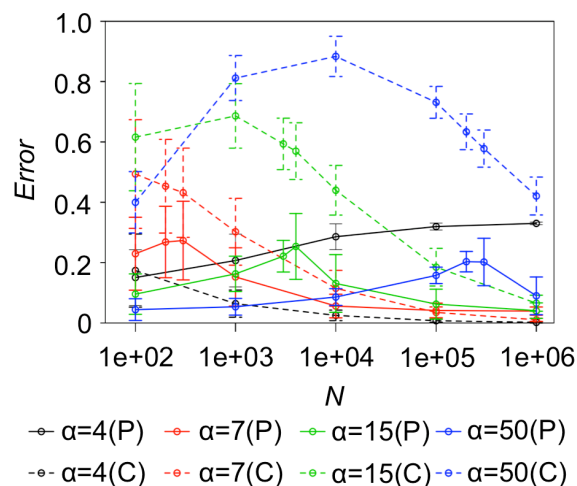


図 11 比較結果

に抑えることができると考えられる。

次に他の復元精度推定手法に対する優位性を確認するために、比較実験を行った。[7] 内で提案されている Perturbation の性能評価手法を比較対象とする。アプリケーション提供者の手元には偽装カテゴリデータ分布 Y と復元後データ分布 A がある。そこで A に再び Negative Surveys の偽装処理を施して得られる擬似的な偽装カテゴリデータ分布 $\hat{Y}$ と Y を比較することで誤差を推定するという手法である。数式としては式 (8) で記述される。

$$EstError = \sum_{i=1}^{\alpha} \frac{|Y_i - \hat{Y}_i|}{N} \quad (8)$$

表 3 のパラメータを用いてシミュレーションを行い、得られた比較結果を図 11 に示す。図 11 の凡例において括弧内が P の場合は提案手法を、C の場合は比較手法を表しており、提案手法を実線、比較手法を破線で描画している。提案手法については、図 10 を参考に、 $\alpha = 4$ のときは常に手法 1 を、 $\alpha = 7$ のときは $N = 200, 300$ を追加して、 $N \leq 200$ では手法 2、 $N \geq 300$ では手法 1 を適応した。 $\alpha = 15, 50$

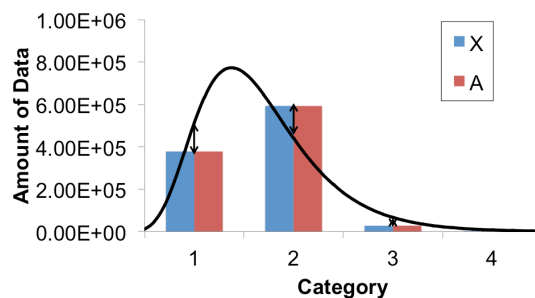


図 12 カテゴリ数が少ない場合の誤差原因

のときはそれぞれ $N = 3000, 4000$, $N = 200000, 300000$ を追加して、 $\alpha = 7$ のときと同様に手法 1 と手法 2 を使い分けた。図 11 を見ると、カテゴリ数が少ない $\alpha = 4$ 以外のときはほとんどの N において提案手法が優れていることがわかり、 N が大きいときに比較手法の方が優れていることもあるが、提案手法との差はごく小さいものであることが見て取れる。

7. 考察

7.1 提案手法の性能について

図 10 を見ると N が増加すると手法 1 の精度が向上する傾向が見て取れるが、 $\alpha = 4$ においてはこの傾向が見て取れない。この理由は、カテゴリ数が少ないことが原因で対数正規分布とフィッティングする際に誤差が生じてしまうからである。図 12 に誤差が生じる例を示した。A は X を復元できているが、対数正規分布へのフィッティングに用いる点の数が少ないため、誤差が生じてしまっているのが分かる。このことが原因で N が増えても Error の値が 0 に近づかないと考えられる。 $\alpha = 3, 4, 5, 6$ でシミュレーションしたところ、 $\alpha = 5$ 以上において、 N の増加と共に推定精度が向上する結果が得られた。このことから Negative Surveys においては $\alpha = 5$ 以上の場合に提案手法を用いるべきであることがわかる。

また、 $\alpha = 4$ 以外のいずれのカテゴリにおいても、手法 1 と手法 2 のグラフが交わる点において最も悪い推定精度であり、その時の Error は 0.2 から 0.3 となっている。提案手法を用いて復元精度を推定した後、インセンティブを付与して復元精度を向上させることを考えると、この誤差がインセンティブ付与にかかるコストに影響を与える。用いるインセンティブ付与手法に依存するが、Error を小さくすることがアプリケーション提供者にとって望ましく、提案した 2 つの手法以外の推定手法が必要と考えられる。

7.2 提案手法の適応範囲について

本稿で述べた復元精度の推定手法は、測定対象の時空間的相関を利用したものである。モバイル環境センシングの測定対象は、大気汚染度や騒音、気温や湿度のような気候情報など多岐に渡り、それぞれの時空間的特性は異なると

考えられる。例えば騒音情報等は大気汚染度と異なりインパルスの信号が発生することがあるため、時間的な相関を持つ時間窓が小さい。また音の伝播距離を考えると、大気汚染度よりも空間的な相関を持つ範囲も狭くなる。一方で気温や湿度などの情報は大気汚染度と似た挙動を取ると考えられる。本提案手法を用いて復元精度を推定する場合は、このような測定対象の特性を加味する必要がある、時間空間的な相関を持つようにセンシングのサイクルなどのパラメータを決める必要があると考えられる。

8. まとめと今後の課題

本稿では、モバイル環境センシングにおいて Perturbation を用いてプライバシー保護を行った場合に、環境情報が時空間的に相関を持つことを利用した Perturbation 後の復元精度推定手法を提案した。環境情報がどのようなモデルに従うかを調べるために実際のモバイル環境センシングで測定対象としてとりあげられる一酸化炭素のオープンデータを解析し、対数正規分布に高い確率で従うことを示した。そこで一酸化炭素を測定するモバイル環境センシングにおいて Perturbation 手法の 1 つである Negative Surveys を用いてプライバシー保護を行うことを想定し、一酸化炭素が対数正規分布に従うことを利用した復元精度の推定手法を 2 つ提案した。シミュレーションによって復元精度の推定精度を検証したところ、データ数に応じて 2 つの手法を組み合わせることで精度良く復元精度を推定できることを示した。さらに、他の復元精度推定手法との比較を行い、結果としてカテゴリ数が少ない場合を除いては、比較手法よりも優れた推定精度であることを示した。

今後は実データを用いた検証を行うとともに、復元精度に応じたインセンティブ付与手法の提案を行う予定である。

参考文献

- [1] Ganti, R. K., Ye, F. and Lei, H.: Mobile crowdsensing: current state and future challenges., *IEEE Communications Magazine*, Vol. 49, No. 11, pp. 32–39 (2011).
- [2] Sivaraman, V.: HazeWatch, The University of New South Wales (online), available from <http://www.hazewatch.unsw.edu.au/> (accessed 2016-04-09).
- [3] Dutta, P., Aoki, P. M., Kumar, N., Mainwaring, A., Myers, C., Willett, W. and Woodruff, A.: Common sense: participatory urban sensing using a network of handheld air quality monitors, *Proceedings of the 7th ACM conference on embedded networked sensor systems*, ACM, pp. 349–350 (2009).
- [4] Maisonneuve, N., Stevens, M., Niessen, M. E. and Steels, L.: NoiseTube: Measuring and mapping noise pollution with mobile phones, *Information Technologies in Environmental Engineering*, Springer, pp. 215–228 (2009).
- [5] Rana, R. K., Chou, C. T., Kanhere, S. S., Bulusu, N. and Hu, W.: Ear-phone: an end-to-end participatory urban noise mapping system, *Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks*, ACM, pp. 105–116 (2010).
- [6] Christin, D., Reinhardt, A., Kanhere, S. S. and Hollick, M.: A survey on privacy in mobile participatory sensing applications, *Journal of Systems and Software*, Vol. 84, No. 11, pp. 1928–1946 (2011).
- [7] Agrawal, R. and Srikant, R.: Privacy-preserving data mining, *ACM Sigmod Record*, Vol. 29, No. 2, ACM, pp. 439–450 (2000).
- [8] Ganti, R. K., Pham, N., Tsai, Y.-E. and Abdelzaher, T. F.: PoolView: stream privacy for grassroots participatory sensing, *Proceedings of the 6th ACM conference on Embedded network sensor systems*, ACM, pp. 281–294 (2008).
- [9] Zhang, F., He, L., He, W. and Liu, X.: Data perturbation with state-dependent noise for participatory sensing, *INFOCOM, 2012 Proceedings IEEE*, IEEE, pp. 2246–2254 (2012).
- [10] Esponda, F.: Negative surveys, *arXiv preprint math/0608176* (2006).
- [11] 青木俊介, 岩井将行, 瀬崎薫: ユーザ参加型環境センシングにおける状態依存型プライバシー保護手法, 電子情報通信学会技術研究報告. ASN, 知的環境とセンサネットワーク, Vol. 113, No. 38, pp. 73–78 (2013).
- [12] Chen, S., Chen, Y. and Trappe, W.: Exploiting environmental properties for wireless localization and location aware applications, *Pervasive Computing and Communications, 2008. PerCom 2008. Sixth Annual IEEE International Conference on*, IEEE, pp. 90–99 (2008).
- [13] di Vimercati, S. D. C., Genovese, A., Livraga, G., Piuri, V. and Scotti, F.: Privacy and Security in Environmental Monitoring Systems: Issues and Solutions, *Computer and Information Security Handbook, 2nd Edition*, J. Vacca (ed.), Morgan Kaufmann, Boston, pp. 835–853 (2013).
- [14] Sweeney, L.: k-anonymity: A model for protecting privacy, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10, No. 05, pp. 557–570 (2002).
- [15] Chaum, D. L.: Untraceable electronic mail, return addresses, and digital pseudonyms, *Communications of the ACM*, Vol. 24, No. 2, pp. 84–90 (1981).
- [16] Shin, M., Cornelius, C., Peebles, D., Kapadia, A., Kotz, D. and Triandopoulos, N.: AnonySense: A system for anonymous opportunistic sensing, *Pervasive and Mobile Computing*, Vol. 7, No. 1, pp. 16–30 (2011).
- [17] Horey, J., Groat, M. M., Forrest, S. and Esponda, F.: Anonymous data collection in sensor networks, *Mobile and Ubiquitous Systems: Networking & Services, 2007. MobiQuitous 2007. Fourth Annual International Conference on*, IEEE, pp. 1–8 (2007).
- [18] Groat, M. M., Edwards, B., Horey, J., He, W. and Forrest, S.: Application and analysis of multidimensional negative surveys in participatory sensing applications, *Pervasive and Mobile Computing*, Vol. 9, No. 3, pp. 372–391 (2013).
- [19] Andrews, A., Kofler, J., Bakwin, P., Zhao, C. and Tans, P.: Carbon Dioxide and Carbon Monoxide Dry Air Mole Fractions from the NOAA ESRL Tall Tower Network, 1992–2009, Version: 2011-08-31 (2009). <ftp://ftp.cmdl.noaa.gov/ccg/towers/>.