明治大学総合数理学部

2020 年度

卒　業　研　究

差分プライバシのプライバシー費用による一般化 k-匿名化の評価手法の提案

学位請求者　先端メディアサイエンス学科

堀込　光

目次

第 1 章

はじめに

2017 年 5 月に施行された改正個人情報保護法では，本人の同意なく個人情報を第三者に提供できる匿名加工情報が導入された．しかし，k-匿名化 [2] における k などの明確な規定はなく，加工をどのレベルまで行うか不確かなことが課題であった．加工による有用性低下は定義しやすいが，その時の安全性は攻撃者の仮定に大きく依存して，精密に評価するのがむずかイいためである．例えば，Terrovitis らは再符号化の方程式を提案している [7] が，安全性評価は不十分である．

そこで本研究では，攻撃者の計算能力向上により差分プライバシに着目する．差分プライバシを満たすデータベースの作成手法，ラプラスメカニズム [1] や Radamized Response[6] では，プライバシ費用 ϵ によって決まるノイズを付加することにより差分プライバシを実現する．

これに対して，本研究では，ランダムサンプリングを行ったデータに一般化 k-匿名化を適用したデータからプライバシ費用 ϵ を計算することにより，差分プライバシの観点から k-匿名化の安全性を評価する手法を提案する．

第 2 章

従来技術

2.1 差分プライバシ

差分プライバシ [1] とは，2006 年に Dwork が提唱したプライバシの定義である．

定義 1. 任意の $S \in Range(A)$ と任意の隣接するデータセット D と D' についてランダムアルゴリズム A がある実数 ϵ について，

$$e^{-\epsilon} \leq \frac{Pr[A(D) = S]}{Pr[A(D') = S]} \leq e^{\epsilon}$$

を満たすとき，ϵ-差分プライバシを満たすという．

この時，プライバシ費用である ϵ が小さいほどデータベース D，D' の区別が難しくなり，安全性は高くなる．差分プライバシでは，任意の背景知識を持つ攻撃者や未知の攻撃者に対して安全性を保障する．

2.2 k-匿名

k-匿名 [2] は，2002 年に Sweeney により提唱されたプライバシ保護技術である．

定義 2. 準識別子の集合を $QI = \{A_1, \ldots, A_d\}$ とする．任意のレコード $t \in D$ に対して，任意の準識別子の値の組 $(a_1, \ldots, a_d)$ を持つレコードが全ての QI について k 以上存在しているとき，データベース D は k-匿名性を満たす．

2.3　一般化

k-匿名化の手法の一つとして一般化がある．一般化とは，属性の要素をより広い意味合いを持つ要素に置き換える操作である．一般木の例を図 1 に示す．属性値が近い要素同士で階層化クラスタリングし，構成していく．一般木の階層の深さは最大で 3 である．Local-gov（地方公務員）や State-gov（州公務員）は Government（公務員）のように一般化される．

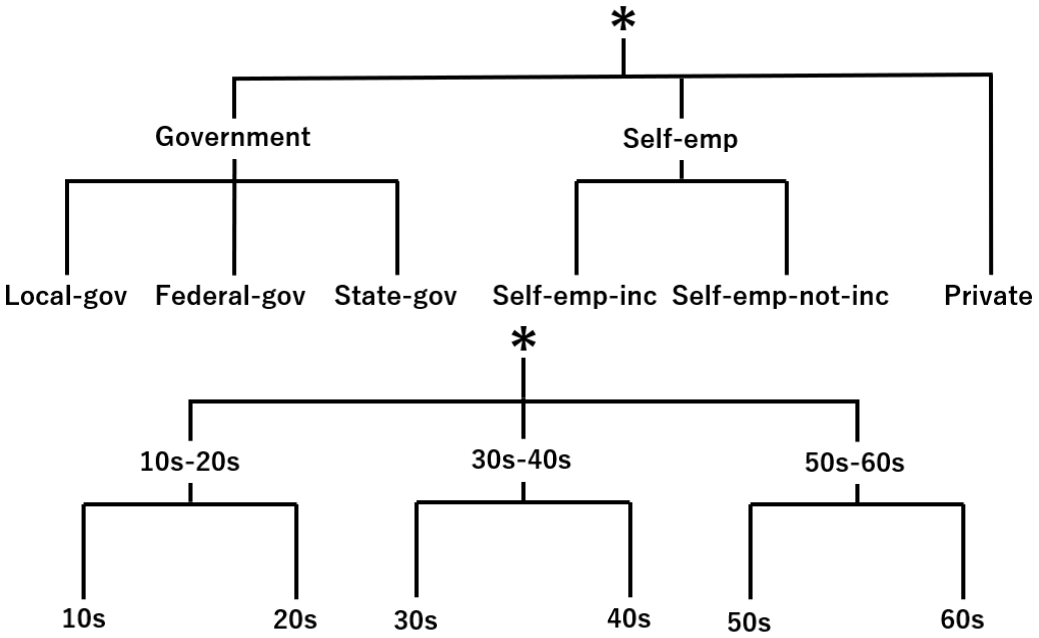

図 2.1　業種 (workclass) と年齢 (age) の一般木

2.4 NCP(Normalized Certainty Penalty)

一般化による k-匿名化アルゴリズム，NCP(Normalized Certainty Penalty) 指標による細分化 [3] である．データベース D 内のレコード t_i の属性 Q_j の要素が $q_{j,u}$ であるとき，レコード t_i の属性 Q_j の NCP は，以下のように求められる．

$$NCP_{Q_j}(t_i) = \sum_{j=1} \frac{|q_{j,u}|}{|Q_j|}$$

この時，$|q|$ は，その要素 q が持つ葉ノードの数である．例えば，$t_i = [10s - 20s, Goverment]$ とすると，

$$NCP(t_i) = \frac{2}{6} + \frac{3}{6} = \frac{5}{6}$$

となる．詳細化可能な属性を詳細化した際に，D 内の全レコードの NCP_{Q_j} の和，

$$NCP_{Q_j} = \sum_{i=1} NCP_{Q_j}(t_i)$$

が最小となる属性を詳細化の対象とし，詳細化後の k-匿名を検討し，満たしている場合，詳細化を行う．また，満たしていない場合は，詳細化を行わず他の NCP が最小となる属性で詳細化を試みる．これを全ての属性が細分化できなくなるまで繰り返す．

Terrovitis らは大域的再符号化 [7] を提案している．この手法では，ある要素全てを同じ値に一般化する手法である．

$k = 5$ と $k = 10$ のときの Age(年齢) 属性の一般木の変化を図 2，図 3 に示す．ともに点線で詳細化の程度を示している．図 2 の k=5 のとき，age(年齢) 属性は，$10s - 20s$，$30s$，$40s$，$50s$，$60s$ に一般化される．$k = 10$ のとき，50 代と 60 代は $50s - 60s$ に一般化されいるのに対し，$k = 5$ では，一般化されず，k が小さいとき詳細なデータとなる．

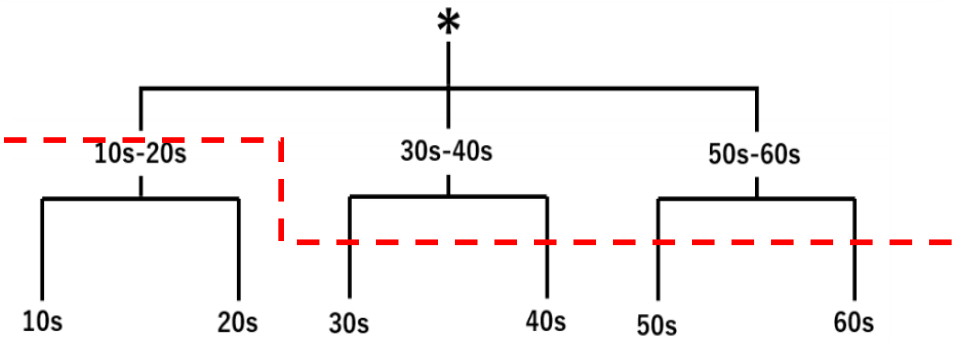

図 2.2　$k = 5$ の際の一般木

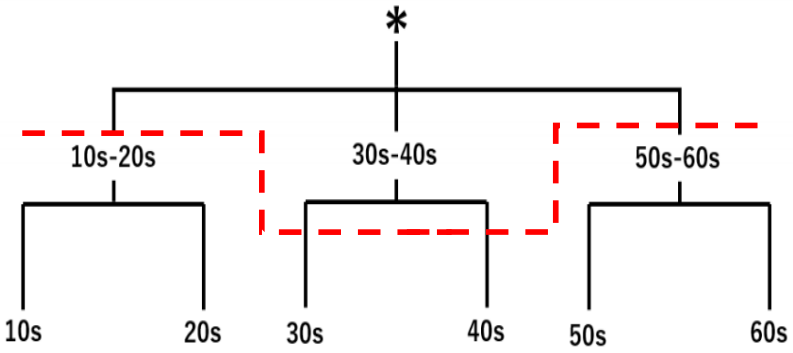

図 2.3　$k = 10$ の際の一般木

第 3 章

提案手法

3.1　サンプリングによる差分プライバシ

Kamalika らは差分プライバシの観点からランダムサンプリングの安全性について示している [4]．これを参考として，サンプリングを行ったデータベースに一般化 k-匿名化を行ったデータの差分プライバシについて命題 1 が証明できる．

Proof. アルゴリズム A をランダムサンプリングを行うアルゴリズム，サンプリング率を β とし，隣接するデータセットをデータベース D と任意のレコード t を抜いたデータベース D_{-t} とする．また，S をランダムサンプリングを行ったデータベースとし，データベース S 内のレコード t の数を s_t とする．今，D 内のレコード t の数を n_t とすると，D_{-t} 内のレコード t の数は $n_t - 1$ となる．

任意のレコード u を考えた場合，n_u 個のレコードから s_u 個サンプリングされる確率は，

$$Pr[n_u \in A(D) = s_u] = {}_{n_u}C_{s_u}\beta^{s_u}(1-\beta)^{n_u - s_u}$$

$Pr[A(D) = S]$ は，すべてのレコードのサンプリング確率の積となるが，隣接するデータベース D と D_{-t} を考えると，レコード t 以外のレコード数は同じであるため，レコード t のみについてサンプリング確率を求めると，

$$\begin{aligned}\frac{Pr[A(D) = S]}{Pr[A(D_{-t}) = S]} &= \frac{{}_{n_t}C_{s_t}\beta^{s_t}(1-\beta)^{n_t - s_t}}{{}_{n_t-1}C_{s_t}\beta^{s_t}(1-\beta)^{n_t - s_t - 1}} \\ &= \frac{n_t(1-\beta)}{n_t - s_t} = \frac{1}{1 - \frac{s_t}{n_t}}(1-\beta) \leq e^{\epsilon}\end{aligned}$$

両辺の対数を取り，命題を得る．　□

ランダムサンプリングでは，ランダムサンプリングにより得られたデータ S と元のデータベース D のレコード数からプライバシ費用 ϵ を求めることができる．サンプリング後に k-匿名化を行う場合も同様に，出力データベース S' とデータベース D からプライバシー費用 ϵ を計算できることを次節で示す．

命題 1. レコード t を n_t 個持つデータ D から，s_t 個を β のサンプリング率でランダムサンプリングしたデータ $A(D)$ は，

$$\epsilon = \log(1-\beta) - \log(1 - \frac{s_t}{n_t})$$

について ϵ-差分プライバシを満たす．

3.2　サンプリング +k-匿名化による差分プライバシ

一般化による k-匿名化では，確率的に処理が行われないため，差分プライバシのプライバシ費用を求めることができない．そこで，確率的なランダムサンプリングを行ったデータベースに k-匿名化を行うことで，差分プライバシのプライバシ費用を求める．

レコード数 n のデータベース D からサンプリング率 β でランダムサンプリングを行い，k-匿名化するアルゴリズムを A とする．また，作成されたデータを $S = A(D)$ とする．この時，データベース D が S となる確率は，S 内のあるレコード s_i の数を n_i，D 内の一般化された際に s_i となり得るレコード数を d_i とすると，

$$Pr[A(D) = S] = \prod_{i=1} {}_{d_i}C_{n_i}\beta^{n_i}(1-\beta)^{d_i-n_i}$$

となる．

しかし，この時 $A(D) = S$ とはならない場合がある．例えば，$[age, workclass]$ の場合を考えてみる．age 属性の要素 $30s - 40s$ について詳細化を考えると，$[30s, Goverment]$, $[40s, Goverment]$，$[30s, Private]$, $[40s, Private]$，$[30s, Self]$，$[40s, Self]$ のすべてが k 以上のとき，age 属性の要素 [30s-40s] は詳細化してしまう．このような条件は，データ S ないのレコードが $2k \leq [30s - 40s, Goverment]$ かつ $2k \leq [30s - 40s, Private]$ かつ $2k \leq [30s - 40s, Self]$ である．同様に，各要素についても全ての組み合わせが k 以上存在したデータとなる可能性を排除する必要が，本実験では，得られたデータ S のすべてのレコードに対して上記の条件とならないデータ S でプライバシ費用 ϵ を算出する．

データベース D から任意のレコード t を削除した隣接するデータベースを D_{-t} とする．この時，削除したレコード t を一般化した際になり得るレコード数 d_i が 1 少なくなる．そのため他のレコード数 d_i は同数である．$Pr[A(D) = S]$ と $Pr[A(D_{-t}) = S]$ の比を以下に示す．

Proof.

$$\begin{aligned}\frac{Pr[A(D) = S]}{Pr[A(D_{-t}) = S]} &= \frac{{}_{d_t}C_{n_t}\beta^{n_t}(1-\beta)^{d_t-n_t}}{{}_{d_t-1}C_{n_t}\beta^{n_t}(1-\beta)^{d_t-n_t-1}} \\ &= \frac{d_t(1-\beta)}{d_t - n_t} \leq e^{\epsilon}\end{aligned}$$

□

第 4 章

実験

4.1　実験目的

k-匿名化による差分プライバシのプライバシ費用を評価する．

4.2　使用データ

本実験では Adult Data Set[5] を使用し，欠損値を含むレコードを削除した 45,223 レコードで実験を行った．使用する属性は，Age（年齢），Workclass（職業），Education（学歴），Income（所得）であり，Age 属性に関しては，一の位を切り捨てた数値を離散値として扱う．使用データのレコード分布を図 4 に示す．

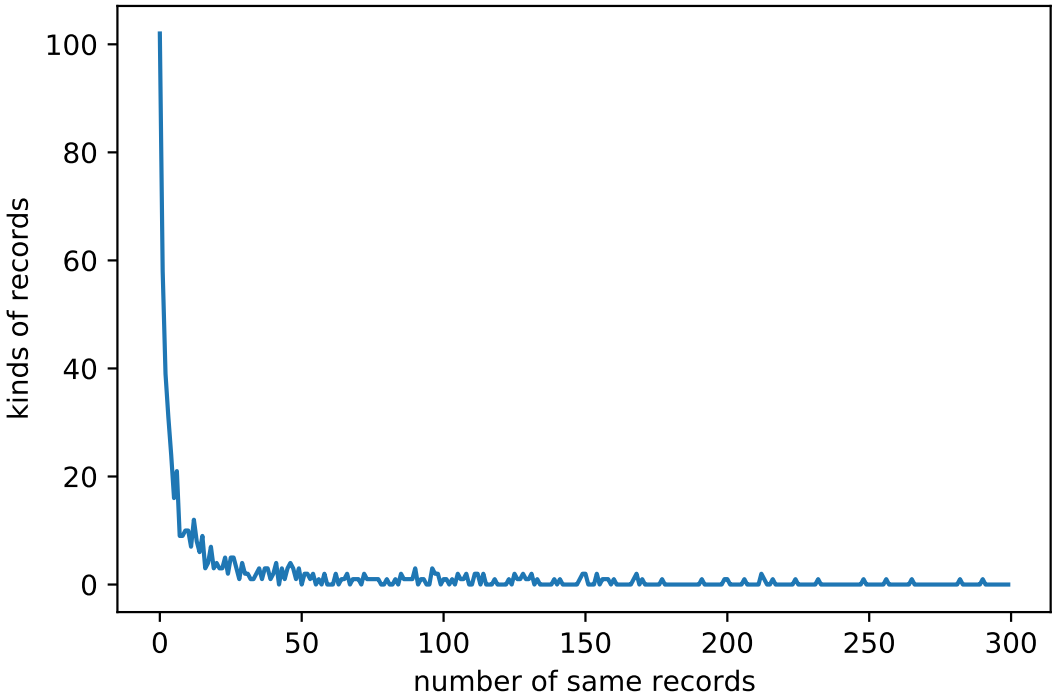

図 4.1　レコード数分布

4.3　実験結果

本実験では，匿名化指標 k とサンプリング率 β を変化させ一般化 k-匿名化のプライバシ費用を算出する．この時，5.2 節で述べた $A(D) = S$ とならない条件を無視する．匿名化指標を $3 \leq k \leq 15$，サンプリング率を $0.1 \leq \beta \leq 0.35$ についてプライバシ費用 ϵ 算出する．k を $k = 10$ に固定し，サンプリング率 β を変化させた際の ϵ の変化を図 5 に示す．図 6 には，サンプリング率 β を $\beta = 0.2$ に固定し，k を変化させた際の ϵ の変化を示す．プロットは 10 回ずつ S を出力した際の平均である．また，平均と標準偏差を表 1 に示す．

4.4　考察

図 5 よりサンプリング率 β が増加するに伴い，ϵ も増加する．定数 $\alpha = 0$ の時，ϵ は，得られたデータ S と S 内の任意のレコードになり得る元のデータベース D 内のレコード数を比較することで算出できる．β を大きくすることで，d_i と n_i の値の差が小さくなり，分母である $d_i - n_i$ が小さくなる．一般化による k-匿名化では，$k = 10$，$\beta = 0.35$ で $\epsilon = 2.0$ 程度である．

また，図 6 より k が大きくなると ϵ が小さくなり，安全性が向上する．$k = 10$ 以上では，ϵ は 1 以下となり安全性は高いと言える．

表 4.1　β と k における ϵ の平均と標準偏差

β	k	平均	標準偏差
0.1	10	0.7571	0.0134
0.15	10	0.8618	0.0203
0.2	10	1.0252	0.0331
0.25	10	1.1838	0.0321
0.3	10	1.4256	0.0362
0.35	10	1.9488	0.0612
0.2	3	1.5973	0.0296
0.2	5	1.0652	0.0262
0.2	15	0.4128	0.0191

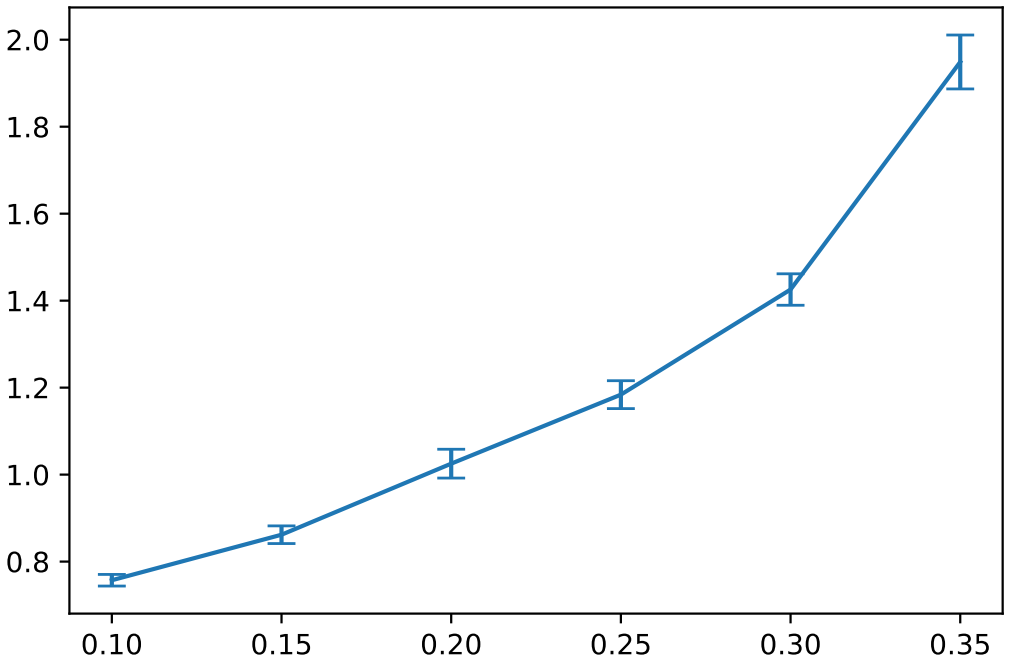

図 4.2　サンプリング率 β による ϵ の変化

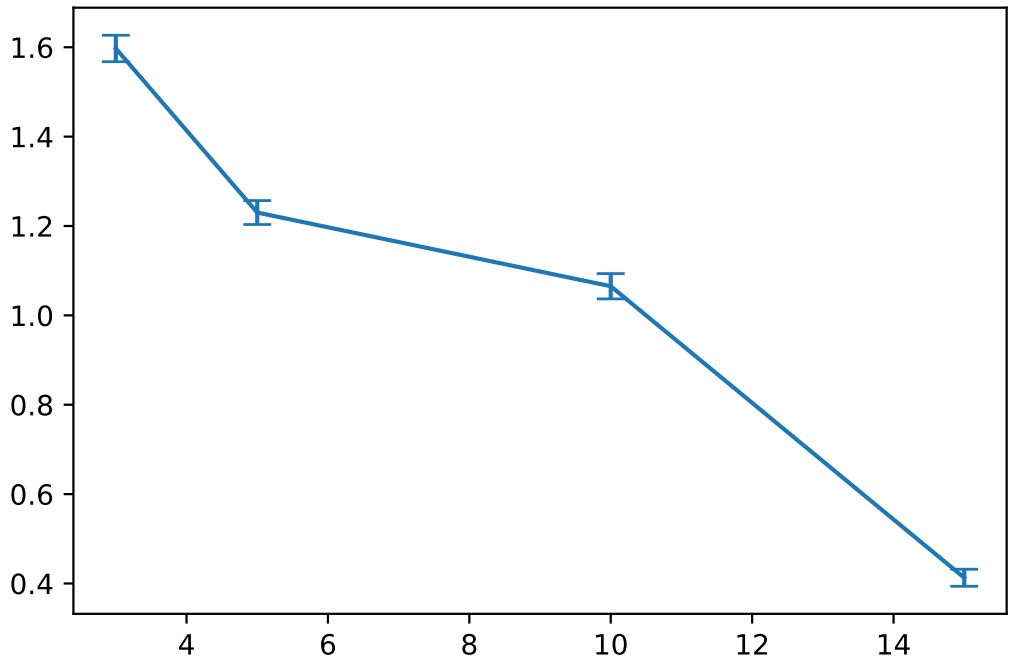

図 4.3　k による ϵ の変化

第 5 章

おわりに

本実験では，Adult Data set を用いて差分プライバシの観点から一般化 k-匿名化を評価した．3.2 節で述べた $A(D) = S$ とならない組み合わせを定式化することが今後の課題である．また，他の差分プライバシ手法の有用性比較を行う必要がある．

謝辞

本研究を行うにあたり，多くの方より御指導いただきました．特に，多大なる御指導を受け賜りました，明治大学総合数理学部先端メディアサイエンス学科，菊池浩明教授に深く感謝申し上げます．実験等に協力してくださった菊池研究室の皆様並びに先端メディアサイエンス学科の方々に深く感謝の意を表するとともに，謝辞とさせていただきます．

参考文献

[1] L.Sweeney, “k-anonymity: a model for protecting privacy”, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5), pp. 557-570, 2002.

[2] J.Xu, W.Wang, J.Pei, X.Wang, B.Shi, A.W.Fu, “Utility-based anonymization using local recoding”, Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 785-790, 2006.

[3] Kamalika Chaudhuri, Nina Mishra, “When Random Sampling Preserves Privacy”, 26th Annual International Cryptology Conference (CRYPTO), pp. 198-213, 2006.

[4] UCI, “Adult Data Set”, (`https://archive.ics.uci.edu/ml/datasets/Adult/`, 2020 年 6 月参照).

[5] C.Dwork, A.Roth, “The Algorithmic Foundation of Differential Privacy”, Foundations and Trends in Theoretical Computer Science, Vol. 9, No. 3-4, pp. 211-407, 2004.

[6] M.Terrovitis, N.Mamoulis, P.Kalnis, “Privacy-preserving Anonymization of Set-valued Data”, Proceedings of the VLDB Endowment, 1(1), pp. 115-125, 2008.

付録 A

Local Differential Priacy によりプライバシーを考慮した位置情報分布推定と人口推測

A.1　はじめに

近年大幅に普及したスマートデバイスを私たちは常に所持している．これにより，サービス企業は人々のあらゆる行動を分析できるようになった．その統計情報は，防災や観光，交通など様々な分野で利用されている．例えば，株式会社 NTT ドコモが提供するモバイル空間統計 [1] では，年齢や性別についての 10 分おきの分布を提供しており，利用者は，任意の居住地を選択し人口分布をみることができる．しかし，事業者には全利用者の正確な位置履歴が管理されており，過失による情報漏洩や不正な内部犯行者によるプライバシ侵害の危険性がある．

個人情報を保護した情報出版技術の一つに Differential Privacy[2] がある．これは，収集した情報を公開する際に確率的なノイズを付加するなどして個人情報を保護する．確率的なノイズが用いられているので，値を曖昧にする匿名加工情報よりも統計的な値を算出するのに適している．しかし，匿名加工情報は，データは安全管理措置が適切である仮定の下で管理されており，プライバシ保護に関しての保証はない．そこで，Local Differential Pricacy[3] が注目されている．スマートデバイスからの情報を収集する際に確率的なノイズを付加してから収集するという技術である．これにより，サービス事業者でさえも真の値はわからない．

Local Differential Privacy の技術に関していくつかの手法が提案されている．Google は，2014 年に Erlingsson らによって提案された RAPPOR[4] という手法を，Apple は，2017 年に Privacy Count Mean Sketch(CMS)[5] という手法を提案し，情報を収集している．RAPPOR は，Randomize Response[6] に基づいており，真の値と偽の値を確率的に入れ替えることによりプライバシ情報を保護する．しかし,[4]，[5] の先行研究では，真の値を推測する際，最尤推定法を用いているが，データの量が少ない場合や値に偏りがあるときには誤差が大きい．

そこで，本研究では，EM アルゴリズム [7] を用いて人口を推定する方式を提案する．株式会社ナイトレイが公開する疑似人流データ [8] から抽出した 6258 名のデータを用い，RAPPOR ベースの LDP により情報を収集し，この収集したデータから最尤推定と EM アルゴリズムにより東京 23 区のそれぞれの人口を推測する．すべての安全指標 ϵ の値で EM アルゴリズムのほうが先行研究の RAPPOR の最尤推定法よりも誤差が小さいことを報告する．

表 A.1　記号表

n	真の人口
n'	LDP における人口
$\hat{n}$	最尤推定法による推定人口
$n^{(*)}$	EM アルゴリズムによる推定人口

A.2　Local Differential Privacy

A.2.1　定義

クライアントは確率メカニズム Q を通してデータ v を変化させて送信することで，サーバにクライアントの持つ真のデータを秘匿する．これによりユーザのプライバシーを保護する．Local Differential Privacy メカニズムは以下のように定義される．

定義 3. *[3]* Q を集合 V の要素 v を受けとって, 集合 Z の要素 z を出力する確率メカニズムとする．$\epsilon \geq 0$ においてハミング重み $\omega_H(v)$，$\omega_H(v')$ が 1 となる任意のペア $v, v' \in V$ と任意の部分集合 $S \subset Z$ に対して,Q が次の条件を満たす時, メカニズム Q は ϵ-*LDP* であるという.

$$Pr[Q(v) \in S] \leq e^{\epsilon} Pr[Q(v') \in S]$$

A.2.2　RAPPOR

RAPPOR は，2014 年に Erligsson らによって提案された手法である [4]．RAPPOR は，クライアント側が持っている情報の集合 V の要素 v_i に対して確率メカニズム Q を適用する．確率 p で出力 $z_i = v_i$，確率 $q = 1 - p$ で出力 $z_i = 1 - v_i$ とする．すなわち，

$$z_i = \begin{cases} v_i & w/p \quad p, \\ 1 - v_i & w/p \quad q \end{cases}$$

例えば，あるユーザ 1 が $\boldsymbol{v} = (0, 1, 0, 0)$ という情報を持っており，別のユーザ 2 が $\boldsymbol{v'} = (0, 0, 1, 0)$ という情報を持っていたとする．ユーザ 1 の情報 $\boldsymbol{v}$ に確率アルゴリズム Q を通すことにより，出力 $\boldsymbol{z} = (0, 1, 0, 1)$ となったとすると，入力 $\boldsymbol{v} = (0, 1, 0, 0)$ という情報が出力 $\boldsymbol{z} = (0, 1, 0, 1)$ となる確率は，

$$Pr[Q(\boldsymbol{v}) = \boldsymbol{z} | \boldsymbol{v}] = p^3 q$$

となる．また，一方でユーザ 2 の入力 $\boldsymbol{v'} = (0, 0, 1, 0)$ という情報が出力 $\boldsymbol{z} = (0, 1, 0, 1)$ となる確率は，

$$Pr[Q(\boldsymbol{v'}) = \boldsymbol{z} | \boldsymbol{v'}] = pq^3$$

となる．Kairouz らの binary mechanism[9] より，

$$\begin{cases} p = \frac{e^{\frac{\epsilon}{2}}}{1 + e^{\frac{\epsilon}{2}}}, \\ q = \frac{1}{1 + e^{\frac{\epsilon}{2}}} \end{cases}$$

とすると，

$$\frac{Pr[Q(\boldsymbol{v}) = \boldsymbol{z}|\boldsymbol{v}]}{Pr[Q(\boldsymbol{v'}) = \boldsymbol{z}|\boldsymbol{v'}]} = \frac{p^3 q}{p q^3} = e^{\epsilon}$$

となり，この確率メカニズム Q は，Local Differential Privacy を満たす．つまり，出力 $\boldsymbol{z} = (0, 1, 0, 1)$ のとき，この出力はユーザ 1 の入力 $\boldsymbol{v} = (0, 1, 0, 0)$ からなのか，ユーザ 2 の入力 $\boldsymbol{v}$’ $= (0, 0, 1, 0)$ からなのか区別することができず，個人のデータを保護することができる．また，ϵ の値が小さいとき，データの変化量が大きくなり，安全性が高くなる．ϵ における確率 q の値を図 1 に示す．

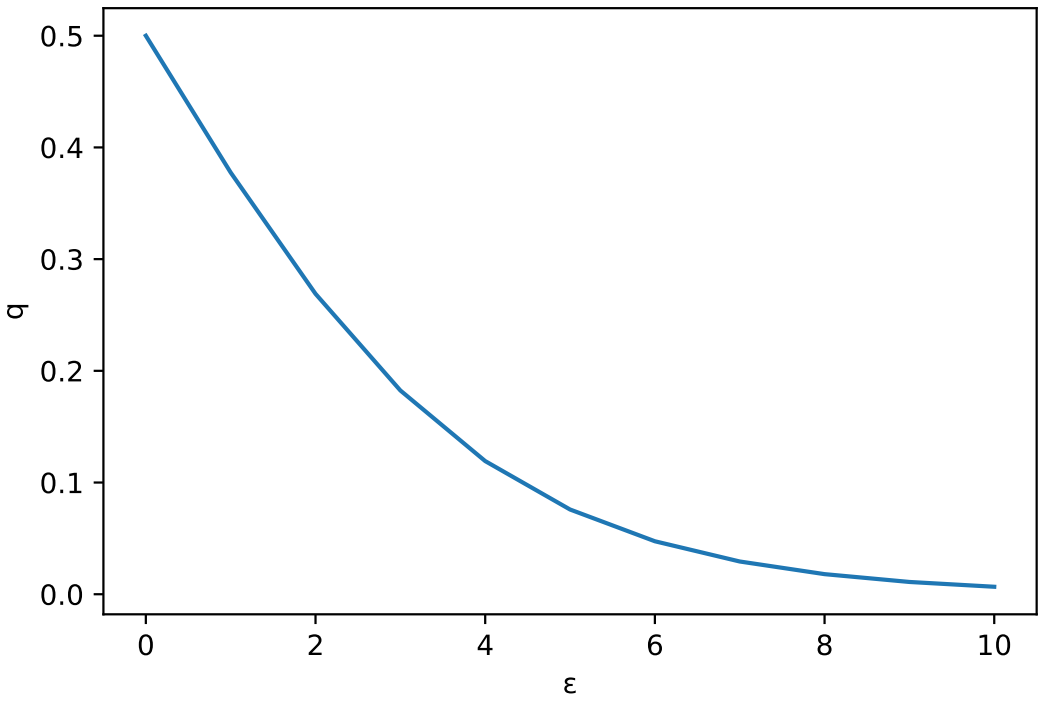

図 A.1　ϵ における確率 q

(証明) 入力 $\boldsymbol{v}$，$\boldsymbol{v'}$ を 23 ビットとする．$\boldsymbol{v}$，$\boldsymbol{v'}$ は 23 区のとこか 1 つについてのみ滞在しているデータであるため，それぞれある 1 ビットが 1，それ以外が 0 である．そのため，$\boldsymbol{v}$ と $\boldsymbol{v'}$ の異なるビット数 Δf は最大で 2 となる．

$$\Delta f = \sum_{k=1}^{23} ||\boldsymbol{v_i} - \boldsymbol{v'_i}||_1 \leq 2$$

また，出力 $\boldsymbol{z}$ と入力 $\boldsymbol{v}$，$\boldsymbol{v'}$ の異なるビット数をそれぞれ r，r' とすると，

$$\frac{Pr[Q(\boldsymbol{v}) = \boldsymbol{z}|\boldsymbol{v}]}{Pr[Q(\boldsymbol{v'}) = \boldsymbol{z}|\boldsymbol{v'}]} = \frac{p^{n-r}q^r}{p^{n-r'}q^{r'}} = (\frac{p}{q})^{r'-r}$$

となる．この時，

$$r' - r = \Delta f$$

であり，

$$\frac{Pr[Q(\boldsymbol{v}) = \boldsymbol{z}|\boldsymbol{v}]}{Pr[Q(\boldsymbol{v'}) = \boldsymbol{z}|\boldsymbol{v'}]} = (\frac{p}{q})^{\Delta f} = e^{\frac{\epsilon \Delta f}{2}} \leq e^{\epsilon}$$

となり，任意の $\boldsymbol{z} \in V$ について言えるので，ϵ-LDP の定義を満たす．

A.3　提案手法

A.3.1　最尤推定

RAPPOR により 区$_i$ の真の人口 n_i を推定していく．最尤推定法による推定方法は以下の通りである．区$_i$ の真の人口を n_i，ユーザ数を ℓ，RAPPOR アルゴリズムにより操作された値のを加算して得られた i 区の人口を n_i' とすると，確率 p，q を用いて，ある区 i のビットが実際に 1 であった n_i 人のうち n_ip 人が 1 を送信する確率が最も高い．また，i 区におらず，そのビットが実際は 0 であった $(\ell - n_i)$ 人のうち $(\ell - n_i)q$ 人が 1 と送信する．従って，

$$n_i' = n_ip + (\ell - n_i)q$$

が成り立ち，n_i について解くと最尤値 $\hat{n}$ は，

$$\hat{n} = \frac{n_i' - \ell q}{p - q}$$

となり，n_i を推測できる．

また，n 人のうち h 人が 1 を送信したとすると，n' を得る条件付き確率分布，すなわち，n_i' のユーザ数 ℓ に対する割合は，

$$Pr[n_i'|n,h] = \left(\begin{array}{c} n \\ h \end{array} \right) p^h q^{n-h} + \left(\begin{array}{c} \ell - n \\ n' - h \end{array} \right) p^{\ell - n - n' + h} q^{n' - h}$$

となる．図 2 に $h = np$ として求めた 14;00 の新宿区の人口 n' の確率分布を示す．

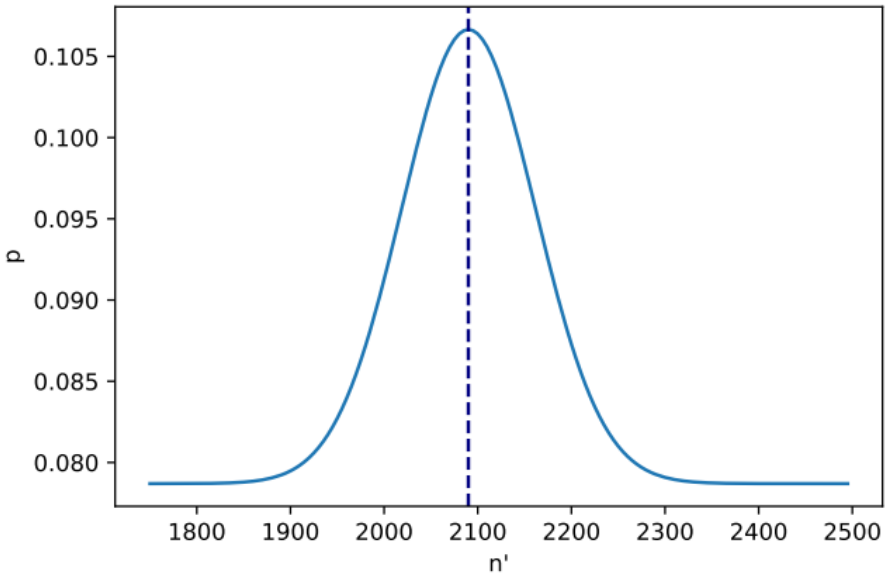

図 A.2　確率分布　$Pr[n'|n,h](\epsilon = 0.5, n = 505, l = 4640)$

A.3.2 EM アルゴリズム

EM アルゴリズムは，ベイズの定理に基づいて，確率パラメータを求める手法である．この手法では，E(Expectation)-step と M(Maximaization)-step を推定値が収束するまで反復することにより，推定する．区$_i$ の全人口の割合を θ_i，各時刻におけるユーザ数を ℓ，RAPPOR アルゴリズムにより算出された 区$_i$ の人口を n'_i とする．E-step と M-step を以下に示す．

E-step

各ユーザ $u \leq \ell$ と 区$_i \leq 23$ について，処理を行う．ユーザ u の 区$_i$ における入力を $\boldsymbol{v_i} = (v_1, \ldots, v_{23}) \in \{0,1\}^{23}$，出力を $\boldsymbol{z_i} = (z_1, \ldots, z_{23}) \in \{0,1\}^{23}$ とする．θ_i の初期値 $\theta^{(0)_1} = \cdots = \theta_{23}^{(0)} = \frac{1}{23}$ とする．

条件付き確率より，入力が $v_i = 1$ だった場合，出力が z_i である確率は，

$$Pr[z_i|v_i = 1] = \frac{Pr[z_i, v_i = 1]}{Pr[v_i = 1]}$$

となる．また，ベイズの定理より，出力が z_i であった場合，入力 $vi = 1$ である確率は，

$$\begin{aligned}\theta_{i,u}^{(k)} = Pr[v_i = 1|z_i] &= \frac{Pr[z_i|v_i = 1]Pr[v_i = 1]}{\sum_{j=1}^{23} Pr[z_i|v_j = 1]Pr[v_j = 1]} \\ &= \frac{Pr[z_i|v_i = 1]\theta_i}{\sum_{j=1}^{23} Pr[v_j = 1|z_j]\theta_j}\end{aligned}$$

となる．

M-step

すべてのユーザで E-step の計算が終わったら，$\theta_i^{(k)}$ をすべてのユーザーの平均とする．

$$\theta_i^{(k)} = \frac{1}{\ell}\sum_{u=1}^{\ell} \theta_{i,u}^{(k-1)}$$

$\theta_i^{(k)}$ が収束して $\theta_i^{(k)} - \theta_i^{(k-1)} \leq \epsilon_2$ となるまで，E-step と M-step を反復する．その収束値 $\theta_i^{(*)}$ を用いて，

$$n_i^{(k)} = \ell\theta_i^{(k)}$$

で推定する．反復回数 k における推定人口 $n_i^{((k))}$ は図 3 のように変化する．ϵ の値が小さいほど収束するまでの反復回数 k は大きくなる．

例

以下に EM アルゴリズムの例を示す．入力 $\boldsymbol{v}$ をハミング重み $\omega_H(v) = 1$ の 4 ビットのベクトルとし，出力 $\boldsymbol{z}$ =(1, 0, 1, 0) であるとする．まず，それぞれの人口の割合，$Pr[v_i = 1] = \theta_i^{(0)} = \frac{1}{4}$ とする．今，$v_i = z$ となる確率はそれぞれ，$Pr[\boldsymbol{z}|\boldsymbol{v_1}] = p^3q$，$Pr[\boldsymbol{z}|\mathrm{v_2}]$=pq^3，$Pr[\boldsymbol{z}|\boldsymbol{v_3}] = p^3q$，$Pr[\boldsymbol{z}|\boldsymbol{v_4}] = pq^3$ となり，$\theta_i^{(1)}$ を計算

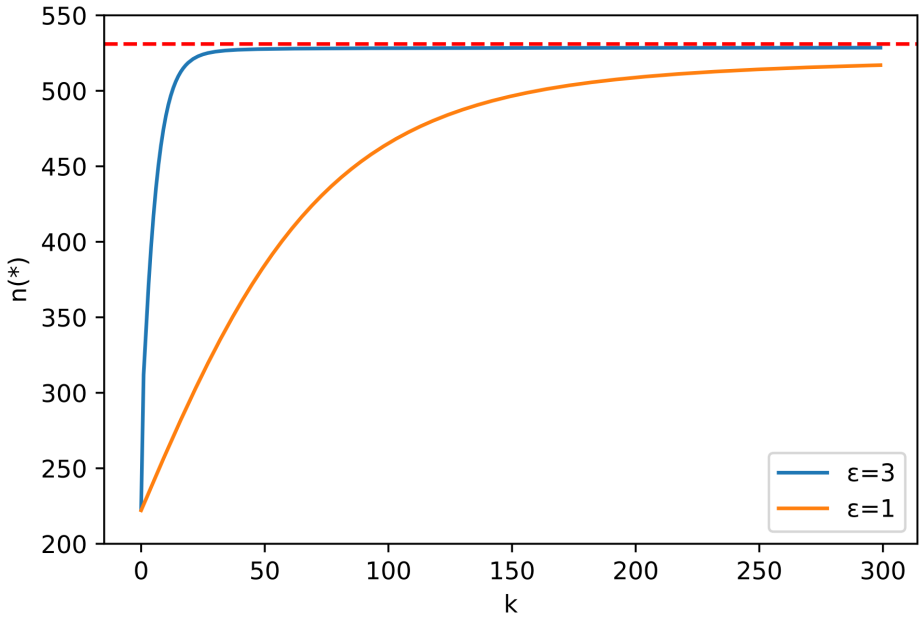

図 A.3　反復回数 k による推定値の変化

する．

$$\theta_1^{(1)} = Pr[\boldsymbol{v_1}|\boldsymbol{z}] = \frac{p^3 q \theta_1^{(0)}}{p^3 q \theta_1^{(0)} + p q^3 \theta_2^{(0)} + p^3 q \theta_3^{(0)} + p q^3 \theta_4^{(0)}} = \frac{\frac{1}{4} p^3 q}{\frac{1}{4} p^3 q + \frac{1}{4} p q^3 + \frac{1}{4} p^3 q + \frac{1}{4} p q^3}$$

同様に $\theta_2^{(1)}$，$\theta_3^{(1)}$，$\theta_4^{(1)}$ を計算する．各ユーザについて同様に $\theta_i^{(1)}$ を求め，その平均を $\theta_i^{(1)}$ とする．これを $\theta_i^{(k)}$ が収束するまで反復させる．

A.4　実験

A.4.1　実験目的

LDP アルゴリズムにおける最尤推定と EM アルゴリズムの精度を明らかにする．

A.4.2　収集データ

本研究では，株式会社ナイトレイより無料公開されている疑似人流データ [8] を使用する．このデータには 6,258 人の一日の位置情報が格納されている．ユーザの位置情報は緯度，経度で示されている．

この緯度，経度から Goole Map API を用いて 8:00 から 3 時間ごとのユーザの位置する市町村区を求めた．そして，各時間における 23 区に属するユーザを収集した．

表 2 は，各時間ごとの東京 23 区に属している全人口である．また，1 日の平均人口が多い 10 区の各時刻における人口を表 3 に示す．表 4 では，各時刻の人口の平均値，最大値，最小値を示している．

そして，図 4 は，中野区，江東区，中央区，文京区における時間ごとの人口推移である．東京 23 区の 8;00 と 14:00 のヒートマップを図 6，図 7 に示す．

表 A.2　時間ごとの 23 区の人口

時間	人数
8:00	2,957
11:00	3,922
14:00	4,640
17:00	4,793
20:00	4,300
23:00	3,283

表 A.3　時間ごとの 23 区の人口

番号	時間	8:00	11:00	14:00	17:00	20:00	23:00
n_1	渋谷区	262	394	533	532	479	351
n_2	新宿区	278	414	505	531	454	304
n_3	港区	267	393	509	479	416	284
n_4	千代田区	186	381	506	496	476	248
n_5	世田谷区	295	331	367	403	368	317
n_6	杉並区	165	209	227	246	187	188
n_7	中央区	121	177	216	188	148	118
n_8	文京区	98	166	181	197	206	143
n_9	品川区	98	147	182	173	147	99
n_{10}	中野区	154	117	116	133	141	142

表 A.4　時間ごとの 23 区の人口

時間	8:00	11:00	14:00	17:00	20:00	23:00
平均	192.4	272.9	334.2	337.8	302.2	219.4
最大値	295	414	533	532	479	351
最小値	98	117	116	133	141	99

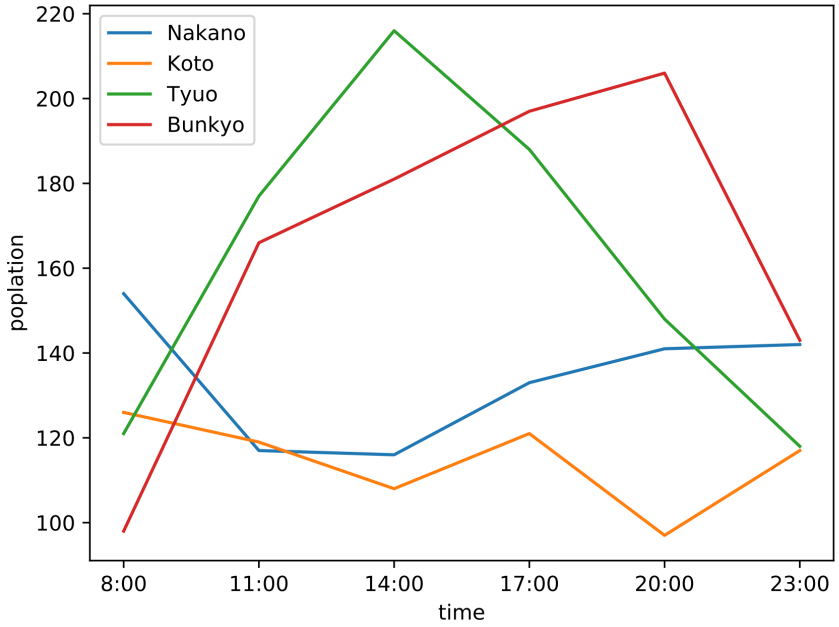

図 A.4　各区の人口推移

各時間ごとにそれぞれのユーザーを位置情報をもとに位置情報の入力 $\boldsymbol{v}$ を作成していく．東京 23 区を列名とし，その時間にユーザが位置していた区を 1 とし，それ以外を 0 となる行列 $\boldsymbol{v}$ (23 行 1 列) を作成する．その行列 $\boldsymbol{v}$ に RAPPOR アルゴリズムを適用し，送信する．収集したデータに対し，最尤推定法と EM アルゴリズムを適用し実際の値を推測し，誤差を求める．例えば，ユーザがある時間に中野区にいるとすると入力 $\boldsymbol{v}$ は，

$$\begin{aligned} v &= (n_{\text{世田谷区}}, n_{\text{中野区}}, n_{\text{渋谷区}}, \ldots, n_{\text{葛飾区}}, n_{\text{江戸川}}) \\ &= (0, 1, 0, \ldots, 0, 0) \end{aligned}$$

となる．このデータを RAPPOR アルゴリズムを適用すると，出力 Z は，

$$\begin{aligned} z &= (n_{\text{世田谷区}}, n_{\text{中野区}}, n_{\text{渋谷区}}, \ldots, n_{\text{葛飾区}}, n_{\text{江戸川}}) \\ &= (1, 0, 0, \ldots, 1, 0) \end{aligned}$$

となったとする．

各ユーザーから $\boldsymbol{z}$ を出力し，出力 $\boldsymbol{z}$ の各区の和 $n'_i = \sum_{i=1}^{23} z_i$ を求める．この n' から最尤推定法と EM アルゴリズムを用いて各区の実際の人口 $\hat{n}$ と $n^{(*)}$ を推定していく．システム構成図を図 5 に示す．

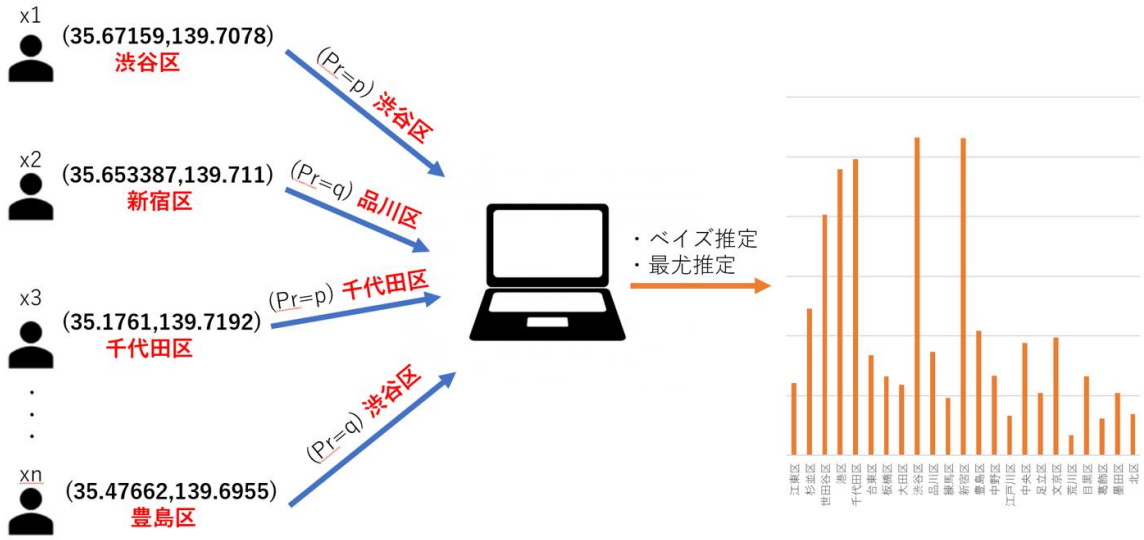

図 A.5　システム構成図

図 A.6　8:00 における人口分布

A.4.3　実験結果

最尤推定法と EM アルゴリズムを用いて RAPPOR アルゴリズムで収集したデータから推定した人口と実際の人口との誤差を ϵ の値を変化させて求めた．誤差を絶対誤差 (AE) の和として，次のようにして求める．

(1) 実際の人口データから各ユーザーが位置する区を 1，それ以外を 0 とする入力 $\boldsymbol{v}$ を作成する．

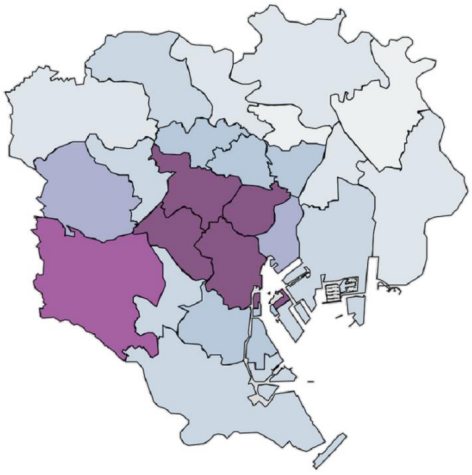

図 A.7　14:00 における人口分布

(2) 入力 v 用いて ϵ の値を 0.5 から 5 まで 0.5 ずつ変化させながら，RAPPOR を適用し出力 $\boldsymbol{z}$ を求める．
(3) 各ユーザの出力 $\boldsymbol{z}$ から各区の人口 n' を求める．
(4) (3) で求めた人口 n' に最尤推定法と EM アルゴリズムを適用し，各区の実際の人口 $\hat{n}$ と $n^{(*)}$ と推定する．
(5) 各区ごとに推定した人口と実際の人口の差の絶対値を求め，その和を S とする．
(6) (2) から (5) を 10 回行い，S の平均 $\hat{S}$ を求める．

S は，東京都の区の数は 23 で真の各区の人口を n，推定した各区の人口を $\hat{n}$，$n^(*)$ とすると，

$$S = \sum_{i=1}^{23} |n_i - \hat{n}_i|$$

として求める．ϵ の値を変化させた各時間の誤差 $\hat{S}$ を実験結果を図 15 から図 20 に示す．

また，17:00 における $\epsilon = 0.5$ の際の最尤推定と EM アルゴリズムにおける推定人口は図 11 のようになる．図 12 には，各区の真の人口とそれぞれの推定人口の散布図を示している．図 11，図 12 ともに実験を 10 回行い，各推定 $\hat{n}, n^{(*)}$ の平均したものである．表 5 には，ϵ ごとの各時刻の誤差 S の平均を示す．

図 A.8　最尤推定と EM アルゴリズムによる各区の推定

表 A.5　ϵ における各時刻の誤差 S の平均

ϵ	EM	MLE
0.5	2971.31	4885.28
1.0	1846.69	2256.53
1.5	1404.14	1614.74
2.0	982.69	1072.24
2.5	780.58	849.41
3.0	628.41	710.34
3.5	524.80	601.56
4.0	407.92	503.13
4.5	349.73	434.63
5.0	287.42	363.16

A.4.4　考察

図 8 と図 15 から図 20 より，最尤推定に比べ EM アルゴリズムでは，すべての ϵ の値において MAE が小さかった．17:00 において $\epsilon = 0.5$ のとき，最尤推定と EM アルゴリズムの MAE の差は 1,871.515 であり，他の ϵ の MAE の差よりも大きい．$\epsilon \leq 0.5$ のときには *epsilon* が小さくなるにつれ両推定の MAE の差は大きくなっていくと考えられる．また，最尤推定による MAE と真の人口との相関係数は 0.255 であったのに対し，EM アルゴリズムによる推定の MAE と真の人口との相関係数は 0.870 であった．つまり，EM アルゴリ

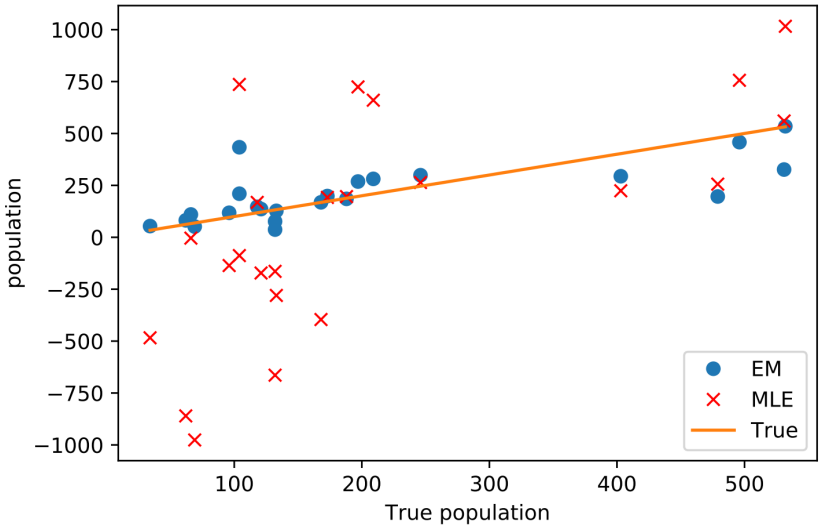

図 A.9　最尤推定と EM アルゴリズムによる推定人口

ズムによる推定では，人口数と誤差には相関があると言え，図 13 に示すように，人口が多い区では誤差が大きくなる．

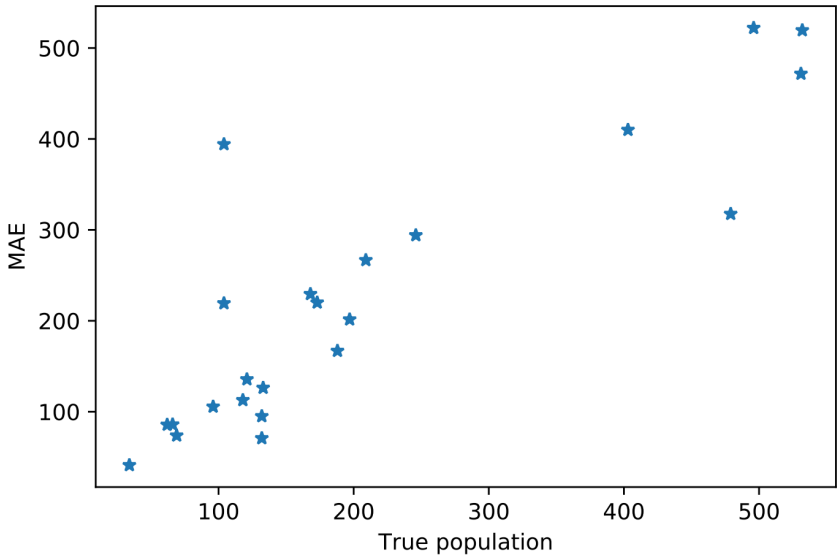

図 A.10　EM アルゴリズムによる真の人口との平均絶対誤差 MAE 分布

A.5　おわりに

本研究では，Local Differential Privacy のアルゴリズムのひとつである RAPPOR を用いて疑似人流データからデータ収集を行い，最尤推定法と EM アルゴリズムを用いて人口推定を行った．EM アルゴリズムを用いて推定を行う際，ϵ の値が小さいと反復回数が大きくなる．データ数が大きくなれば，収束値のわずかな違いで推定値に大きな影響を与える．収束回数の適切な決定が今後の課題である．

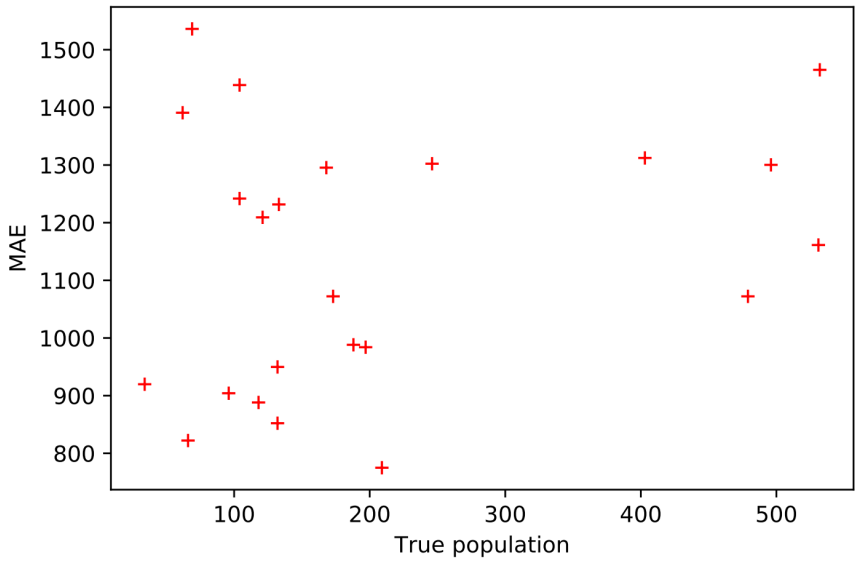

図 A.11　最尤推定による真の人口との平均絶対誤差 MAE 分布

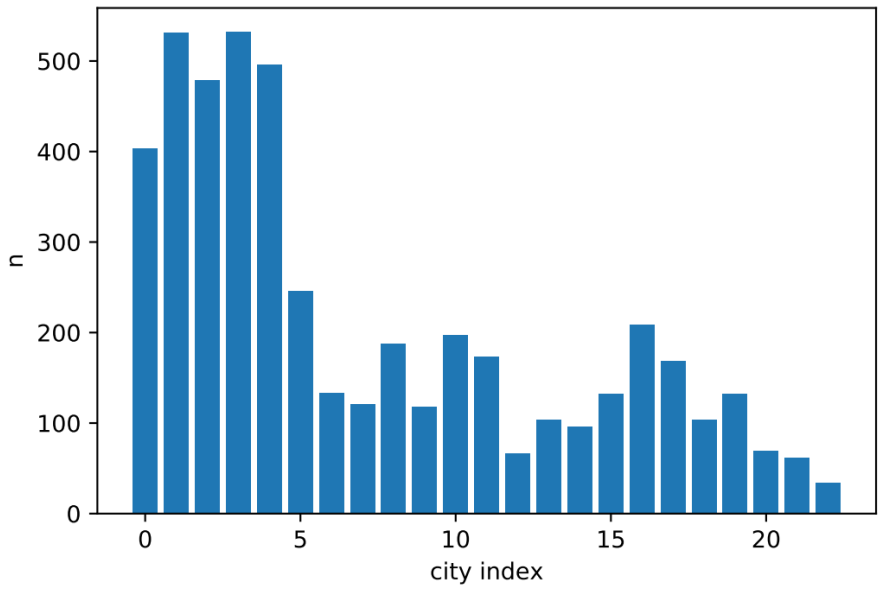

図 A.12　各区の真の人口

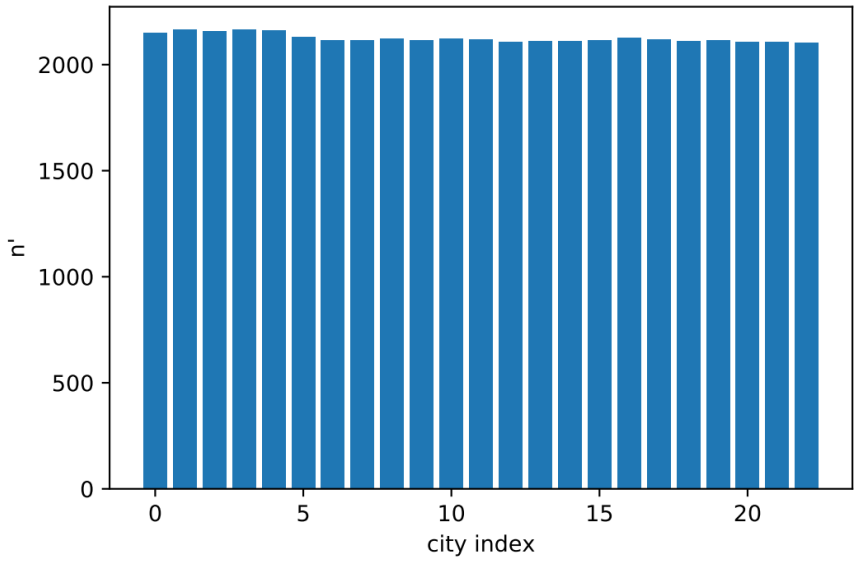

図 A.13　RAPPOR における人口 n’

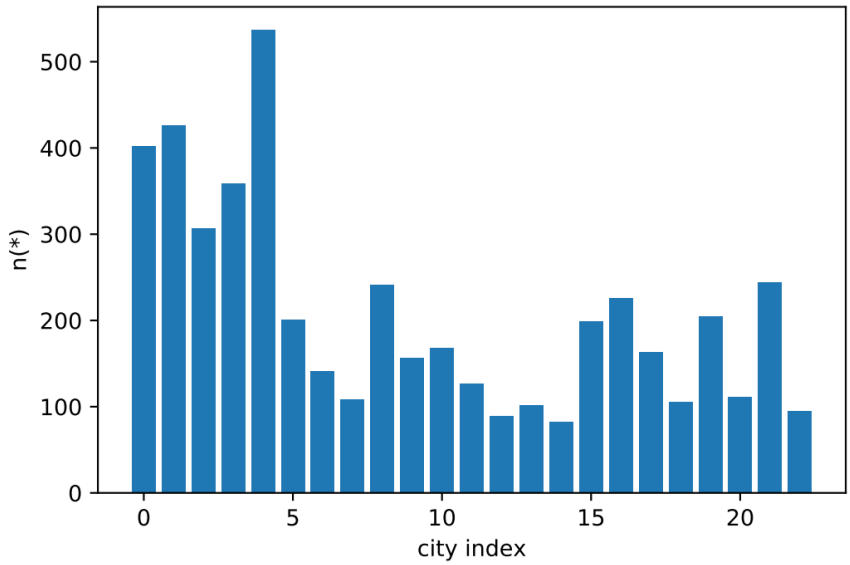

図 A.14　EM アルゴリズムによる推定人口 n(*)

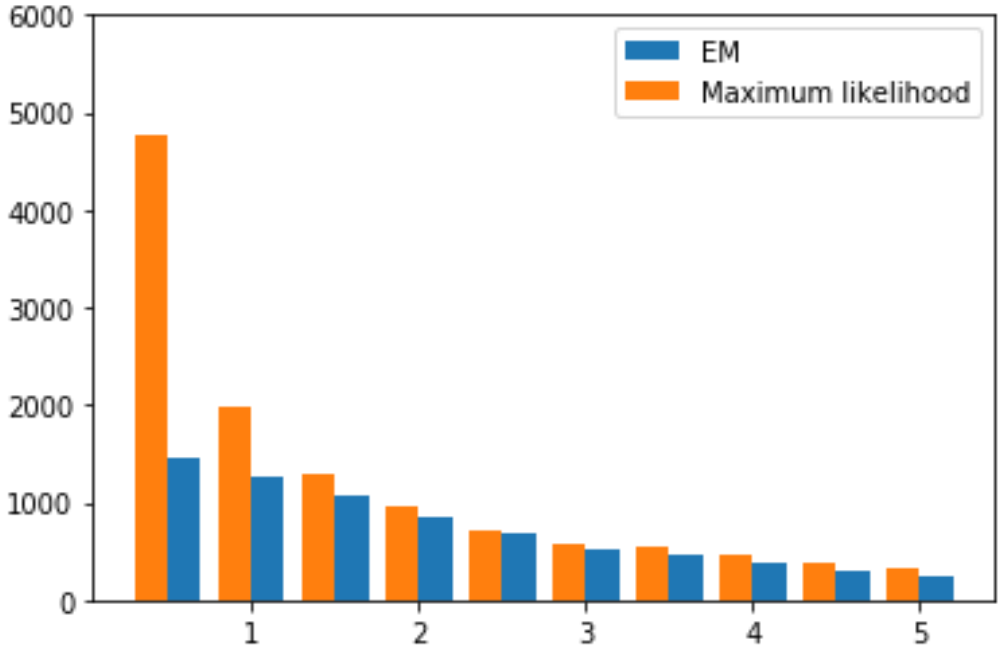

図 A.15　8:00 の誤差 S

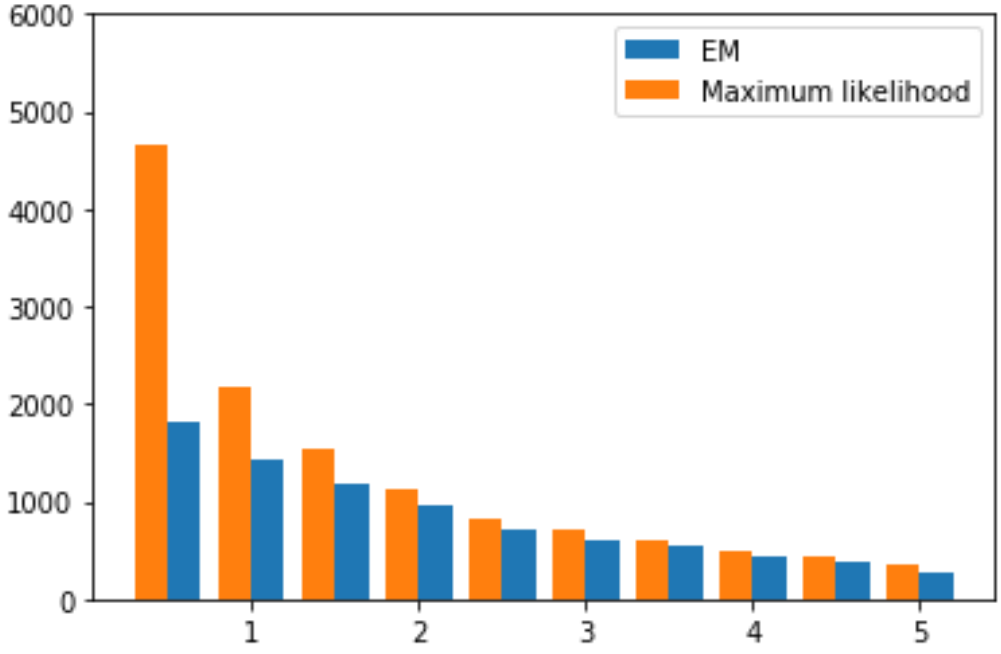

図 A.16　11:00 の誤差 S

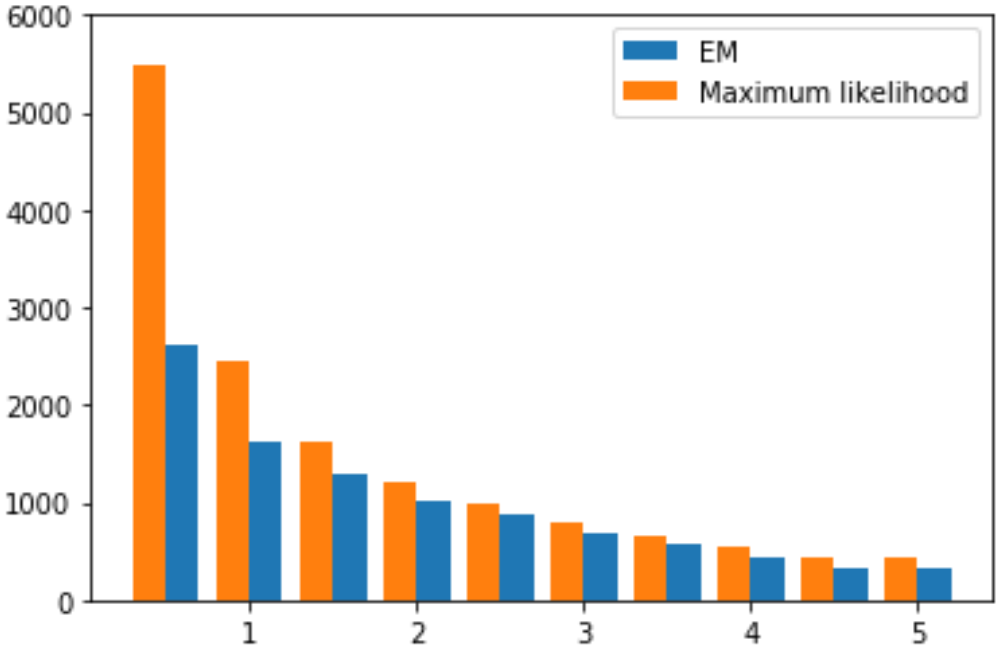

図 A.17　14:00 の誤差 S

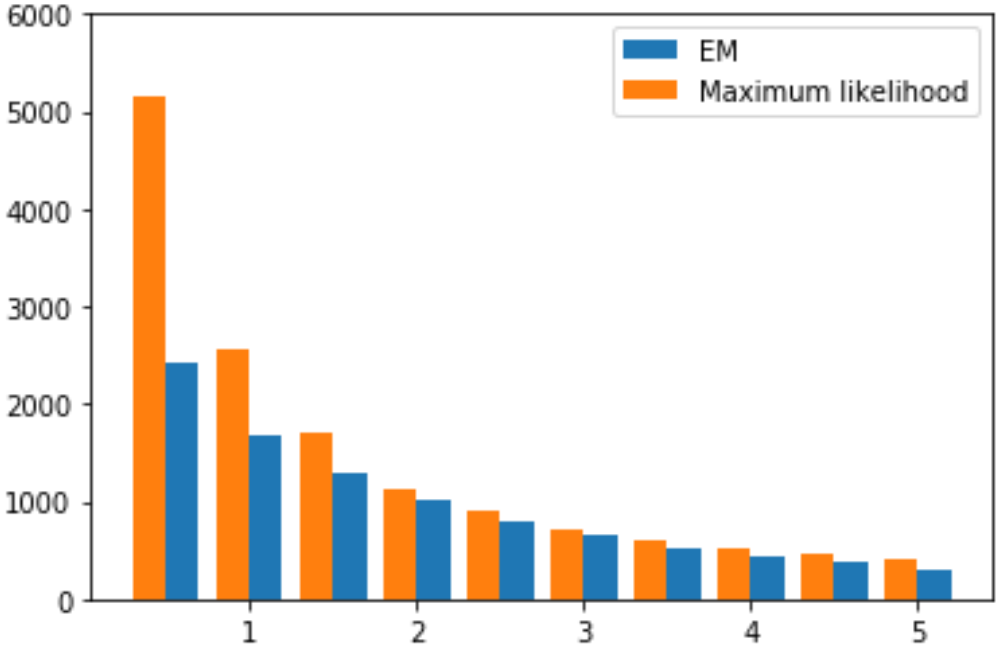

図 A.18　17:00 の誤差 S

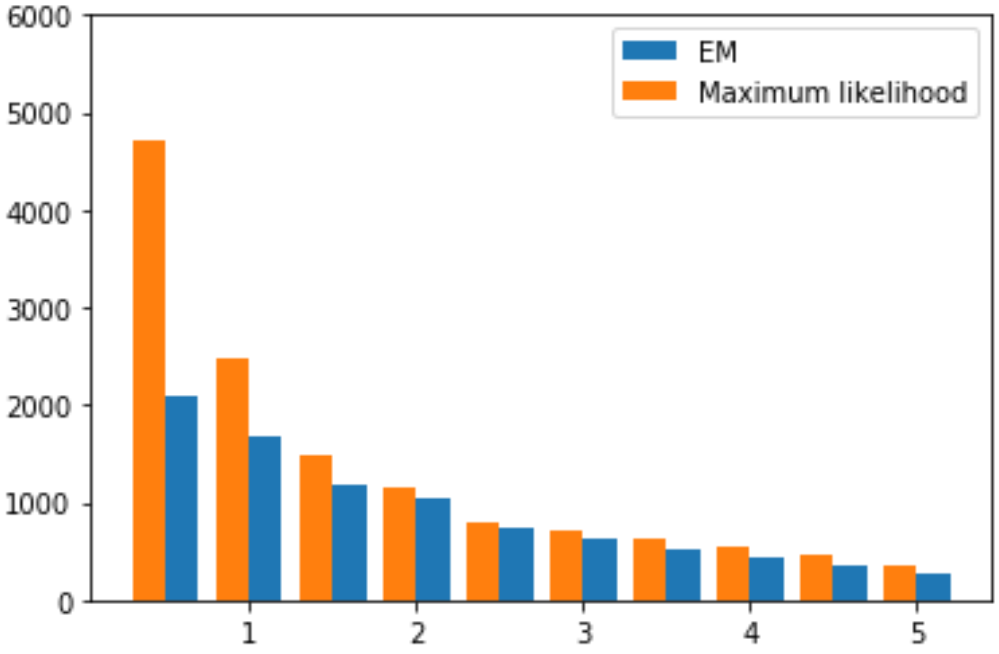

図 A.19　20:00 の誤差 S

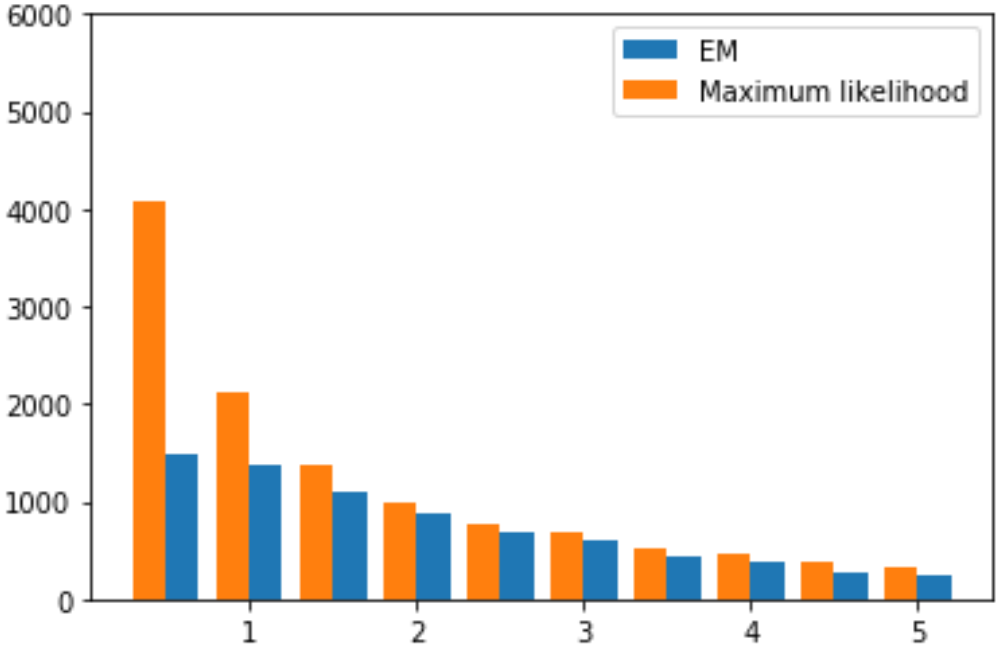

図 A.20　23:00 の誤差 S

参考文献

[1] NTT Docomo, “モバイル空間統計の「国内人口分布統計（リアルタイム版）」の提供開始” (https://www.nttdocomo.co.jp/info/news_release/2019/12/03_00.html/ ,2020 年 4 月参照).

[2] Cynthia Dwork,Frank McSherry,Kobbi Nissim,Adam Smith,“Calibrating noise to sensitivity in private data analysis”, TCC, Vol. 3876, pp. 265–284, 2006.

[3] John C Duchi,Michael I Jordan,Martin J Wainwright, “ Local privacy and statistical minimax rates”, FOCS, pp. 429–438, 2013.

[4] Úlfar Erlingsson, Vasyl Pihur,Aleksandra Korolova,“RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response”,ACM Conference on Computer and Communications Security, pp. 1054-1067, 2014.

[5] Diferential Privacy Team, Apple, “Learning with privacy at scale”, 2017 (https://machinelearning.apple.com/docs/learning-with-privacy-at-scale/appledifferentialprivacysystem.pdf, 2012 年 4 月参照).

[6] Stanley L.Warner,“Randomized Response:A Survey Technique for Eliminating Evasive Answer Bias”,Journal of the American Statistical Association, pp. 63-69, 1965.

[7] 宮川雅巳,“EM アルゴリズムとその周辺”, 応用統計学, Vol 16, No. 1, pp. 1-19, 1987.

[8] Nightley, “疑似人流データ’’ (https://nightley.jp/archives/1954/, 2019 年 10 月参照).

[9] Peter Kairouz, Sewoong Oh, and Pramod Viswanath, “Extremal mechanisms for local differential privacy”, In Advances in neural information processing systems, pp. 2879–2887, 2014.

[10] Anna Mochizuki,Hiroaki Kikuchi, “Priacy-preserving Collaborative Fitering Using Randomized Response”, Journal of information Processing, Vol. 21, No.4, pp. 617-623，2013.

[11] 小野元，福地一斗，佐久間淳，“局所差分プライバシー制約下における遂次 heavy hitters 検知”，DEIM Forum 2018, E1-3, 2018.

[12] Zhan Qin，Yin Yang，Ting Yu, “Heavy Hitter Estimation over Set-Valued Data with Local Differential Privacy”, ACM CCS 2016, pp. , 2016.

[13] Hiroaki Kikuchi，Jin Akiyama，Howard Gobioff，“Stochastic Voting Protocol To Protect Voters Privacy”, WIAPP’99, pp. 103-111, 1999.