人の移動の特徴を考慮した人口統計情報からの移動人数推定

赤木　康紀† 　西村　拓哉† 　倉島　健† 　戸田　浩之†

† 日本電信電話株式会社　〒 239–0847 神奈川県横須賀市光の丘 1-1

E-mail: †{yasunori.akagi.cu, takuya.nishimura.fk, takeshi.kurashima.uf}@hco.ntt.co.jp, hirotoda@acm.org

あらまし　群衆の移動傾向を把握する上で，十分なサンプル量が確保されつつプライバシーに配慮がなされ入手・利用しやすいデータとして，人口統計情報が注目を集めている．人口統計情報とは，各エリア・各時刻の人口を集計し統計化したデータである．本研究では，このような人口統計情報から各タイムステップ間の各エリア間の移動人数を推定する手法を提案する．既存手法では，エリア間の遷移確率パラメータを出発エリアと到着エリアのペアそれぞれについて用意し推定を行っている．しかしこの手法には，モデルの自由度が高くなりすぎるために解が絞りきれず，推定精度が低くなってしまうという問題点が存在する．本論文では，人間の移動の特徴をもとにしたモデリングを行うことによって，高精度な推定を行う手法を提案する．カープローブデータから生成した人口統計情報を用いて評価実験を行い，提案手法は既存手法より高い精度で推定を行うことができることを示した．

キーワード　人口統計情報，時空間データ，Collective Graphical Model

1. はじめに

近年，群衆の位置情報を解析する上で，人口統計情報が注目を集めている．人口統計情報とは，各エリア・各時刻の人口を集計し統計化したデータであり，例えば「午前 7 時にエリア A には 50 人いた」などの情報である．一例として，株式会社ドコモ・インサイトマーケティングが販売している「モバイル空間統計」[14] などが挙げられる．また，海外の事例として，フランスの XData Project では携帯電話の位置情報を各地域の人口の形に集計したデータの公開を検討している [1]．人口統計情報は，位置情報活用において問題となるプライバシー保護の問題に対して，統計量データのみを公開することで対処しており，集団の位置情報を解析する上で現実的に利用可能な数少ないデータとなっている．これらのデータは，タクシーや乗り合いバスのための需要予測 (注1)，マーケティングや防災計画への活用が検討されている．

このように様々な場面での活用が検討されている人口統計情報であるが，明示的には人々の移動に関する情報が含まれていない．ここでいう移動に関する情報とは，各時刻における各エリア間の移動人数のことであり，例えば「午前 7 時から午前 8 時にかけてエリア B からエリア C に 30 人移動した」などの情報である．

そこで本研究では，人口統計情報からエリア間の移動人数を推定するタスクに取り組む（図 1 参照）．エリア間の推定移動人数を利用することで，タクシーや乗り合いバスの需要予測の際に，エリアごとの人口だけではなく人の移動の方向を活用することができ，より高精度な予測やより高度なアプリケーションを実現することができるようになると考えられる．

このタスクを解くためのアプローチの 1 つとして，Collective Graphical Model (以下 CGM) が挙げられる．CGM は集計されたデータ（今回のタスクの場合，人口統計情報）から，その背後に隠れた確率モデルのパラメータ及び集計される前の値（今回のタスクの場合，移動人数）を推定するフレームワークであり，Sheldon らによって提案された [11]．Iwata らはこのフレームワークを人口統計情報から移動人数を推定するタスクに適用し，ナイーブな手法と比較して高い精度での推定が行えることを示した [6]．

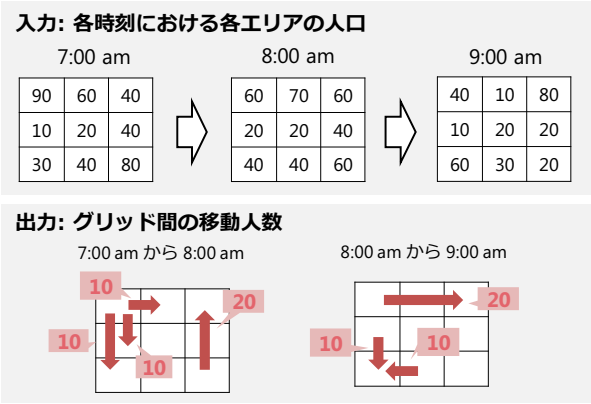

図 1　取り組むタスクの入力と出力．

しかしこれらの既存手法には，人口統計情報の種類や離散化のスケールによっては推定精度が低くなってしまうという問題点が存在する．CGM を使った既存手法では，セル間の遷移確率パラメータを出発セルと到着セルのペアそれぞれについて用意し，それぞれの推定を行っている．しかしこの手法には，モデルの自由度が高くなりすぎるために解が絞りきれず，推定精度が低くなってしまうという問題点が存在する．すなわち，モデルの自由度が高くなることにより，観測された人口統計情報をうまく説明するパラメータ及び移動人数が多く存在することになってしまうため，「真のパラメータ・移動人数とは異なるものの，観測された人口統計情報をうまく説明するパラメータ・

(注1)：https://www.nttdocomo.co.jp/info/news_release/2017/09/20_00.html

移動人数」を推定してしまうのである．このようなモデルの自由度の高さに起因する問題は，推定するパラメータが多くなる場合，すなわち離れたセルへの移動も考慮しなければならない場合などに顕著になる．したがって，セルサイズが小さい場合や離散化の時間間隔が長い場合には推定精度が非常に悪くなってしまう．

そこで，本研究では確率モデルに人間の移動に関する知見を組み込み，少ないパラメータで人間の移動をモデリングすることによって上記の問題を解決する手法を提案する．具体的には，人間のエリア間の移動の確率が (1) エリア間の距離 (2) エリアへの集まりやすさ (3) エリアから出発する確率という 3 つの要素によって近似的に計算できると仮定したモデリングを行い，CGM のフレームワークを適用する．この結果，離れたセルへの移動も考慮しつつ，モデルの自由度を下げることで高精度な推定が可能になる．

また，推定の際の最適化問題に対して，Minorization-Maximization アルゴリズム [5] の枠組みを適用することで効率的なアルゴリズムを導出した．導出されたアルゴリズムでは，閉形式での更新と 1 変数の凸最適化問題を解くことによる更新の繰り返しによってパラメータを求めることができる．

本研究の貢献は以下のように整理される．

（1） モデルの自由度が高くなることによる推定精度の悪化という既存手法の問題点を解決するための，人間の移動の特徴をもとにした遷移確率のモデリング方法の提案．

（2） 上記のモデルにおける，効率的なパラメータ推定アルゴリズムの提案．

（3） カープローブデータから生成した人口統計情報を用いた評価実験による提案手法の有効性の確認．

本論文の構成は以下の通りである．まず 2 章で取り組む問題を数学的に定式化し，記号の導入を行う．3 章で関連研究を紹介し，4 章で既存手法の詳細及び問題点を説明する．5 章で提案手法の詳細について述べ，6 章で提案手法の有効性を確認するために行った実験について説明する．最後に，7 章でまとめを述べる．

2. 問題設定

本章では，本研究において取り組むタスクを数学的に定式化し，必要となる記号を導入する．本論文で使用される記号を表 1 に示す．

本論文においては，地理空間を緯度経度と平行な線で等間隔に区切ることで作られるグリッド空間を考え，それぞれのセルを 1 つのエリアとする．このような形式を想定する理由は，現実の多くの人口統計情報がこの形で提供されるためである．ただし，空間の分割され方がグリッド状ではなくても，適切にエリア間の距離関数を設計することができれば，本論文で提案する手法は適用することが可能である（たとえば市区町村などの行政区画など）．このセル全体の集合を V で表す．また，時間方向には T 個のタイムステップに離散化されているとする．人口情報が取得されているタイムステップはそれぞれ $t = 0, 1, \ldots, T-1$ で表される．

記号	定義
V	セル全体の集合.
T	タイムステップ数.
N_{ti}	タイムステップ t におけるセル i の人口.
M_{tij}	タイムステップ t から $t+1$ にかけてセル i からセル j に移動する人数.
θ_{ij}	セル i からセル j への遷移確率.
Γ_i	セル i から出発した人の移動先の候補となるセルの集合.
$d(i,j)$	セル i とセル j の距離.
π_i	セル i からの出発確率パラメータ．セル i から「セル i 以外のセル」へ移動する確率を表す.
s_i	セル i への人の集まりやすさを表すパラメータ.
β	セル間の距離と移動確率の関係を表現するためのパラメータ.
λ	人数保存制約を破ることへのペナルティをコントロールするためのハイパーパラメータ.

表 1　本論文で用いる記号.

以上の記号のもと，今回取り組む問題の入出力は以下のように書くことができる．

- 入力: 時刻 t でのセル i における人口 N_{ti} $(t = 0, \ldots, T-1, i \in V)$.
- 出力: 時刻 t から 時刻 $t+1$ にかけて，セル i からセル j に移動した人数 $M_{tij} (t = 0, \ldots, T-2, i, j \in V)$.

3. 関連研究

本章では，人口統計情報からの移動傾向の推定に関する既存研究を紹介する．人口統計情報から移動傾向を推定する方法は，近年盛んに研究されている．これらの研究は，想定する入力の形式や手法などにそれぞれ違いがある．

3.1 Collective Graphical Model を用いた研究

Collective Graphical Model (CGM) は，集計化されたデータからその背後にあるモデルを推定するための一般的なフレームワークであり，Sheldon らによって提案された [11]．以降，道路ネットワーク [7] やアミューズメントパーク [3] における人流の解析などに応用されてきた．特に，Iwata ら [6] は本研究と同様にセル毎の人口分布から移動を推定する問題を取り扱っており，時間方向のクラスタリングと変分ベイズ推定を導入することで高精度な推定を行うアルゴリズムを提案している．ここで挙げた応用例では, [7] の入力形式は各地点の入人数・出人数を想定しており, [3] [6] においては各地点の人口を入力とすることを想定している．また，推定の方法については，MCMC サンプリングによる方法 [11]，事後確率最大化による方法 [12]，メッセージパッシングによる方法 [13] など様々な方法が提案されている．これらの研究と本研究の関係については，4 章で詳細に述べる．

3.2 統計化されたデータからの移動の推定に関する研究

Kumar らは，ウェブグラフ上における各ページの閲覧数という統計化された情報から，ページ遷移の傾向を推定する研究を行っている [8]．この研究では各ページの閲覧数はマルコフ連鎖の定常分布によって決まると仮定し，定常分布が観測され

たもとで，もとのマルコフ連鎖の遷移確率行列を推定する問題を解いている．通常，この問題は条件が足りず一意的に解を求めることができないが，Kumar らは Luce の選択モデル [9] と呼ばれる人の選択行動のモデルを仮定することによって，一定の条件のもとで解が一意に決まることを示し，解を求めるアルゴリズムを与えた．さらに，Maystre らはこの研究を拡張し，Luce の選択モデルの仮定のもとで，各地理空間上のある地点（もしくはウェブグラフ上のある頂点）の入人数及び出人数からもとのマルコフ連鎖を推定する手法 Choice Rank を提案し，Wikipedia のクリックストリームデータや NYC のバイクシェアのデータを用いて有効性を検証している [10]．しかし，これらの手法は「各タイムステップごとの統計化されたデータ」という形式のデータを想定しておらず，本研究で考えるデータに対して適用することができない．

Xu らは，人口統計情報（特に携帯電話の基地局ごとの人口情報）の公開に関するプライバシーリスクを評価する研究の中で，各タイムステップごとの人口情報から，個人の移動軌跡を復元する手法を提案している [15]．この手法は，実データの分析で得られた知見をもとに決めたコスト関数を用いて，各時間ごとに割当問題を解くという方法をとっている．コスト付き割当問題を解くという点では CGM に類似しているが，コスト関数を定めるパラメータを推定する機構が含まれていないという点で CGM とは本質的な差がある．パラメータ推定が行えないため，この手法が適用できるデータは非常に限定的である．我々の手法は，ヒューリスティックにコスト関数を定めることなく移動人数の推定を行うことが可能である．

3.3 人口統計情報の予測に関する研究

過去の人口統計情報及びその他の特徴量から，将来の人口統計情報を予測するタスクを扱っている既存研究は多数存在する．例えば，Hoang らは都市のそれぞれの地域への入人数・出人数を予測するタスクを扱い，季節成分・トレンド成分・それ以外の成分に分解してモデリングする手法を提案した [4]．また，Zhang らはディープラーニングベースの手法である ST-ResNet を提案し，セルごとの入人数・出人数を予測するタスクに適用した [16]．これらの研究は，未来の予測を行うという点で本研究とは異なるタスクを取り扱っている．これらの手法と我々の提案する手法と組み合わせることによって，過去の人口統計情報から未来の移動人数を予測するなどの，より発展的なタスクを解くことができるようになる可能性もある．

4. CGM を用いた移動人数推定とその限界

本章では，既存研究 [6] における，CGM を用いた人口統計情報からの移動人数推定の方法を紹介し，その限界について指摘する．

4.1 モデルと推定方法

CGM を人口分布に適用した研究としては，Iwata らの先行研究 [6] がある．この研究では，移動が起こりうるセルの対について遷移確率を表すパラメータ θ_{ij} を用意し，CGM のフレームワークを適用している．

セル i からの移動先候補となるセルの集合を Γ_i とする．Γ_i の決め方については後述する．時刻 t におけるセル i からの移動人数 $\boldsymbol{M}_{ti} = \{M_{tij} \mid j \in V\}$ は，$\boldsymbol{\theta}_i = \{\theta_{ij} \mid j \in \Gamma_i\}$ をパラメータとする多項分布から生成されると仮定する:

$$P(\boldsymbol{M}_{ti} \mid N_{ti}, \boldsymbol{\theta}_i) = \frac{N_{ti}!}{\prod_{j\in\Gamma_i} M_{tij}!} \prod_{j\in\Gamma_i} \theta_{ij}^{M_{tij}}.$$

したがって，$\boldsymbol{N} = \{\boldsymbol{N_{ti}} \mid t = 0, \ldots, T-1, i \in V\}, \boldsymbol{\theta} = \{\boldsymbol{\theta}_i \mid i \in V\}$ が与えられたとき，$\boldsymbol{M} = \{\boldsymbol{M}_{ti} \mid t = 0, \ldots, T-2, i \in V\}$ の尤度関数は

$$P(\boldsymbol{M} \mid \boldsymbol{N}, \boldsymbol{\theta}) = \prod_{t=0}^{T-2} \prod_{i\in V} \left(\frac{N_{ti}!}{\prod_{j\in\Gamma_i} M_{tij}!} \prod_{j\in\Gamma_i} \theta_{ij}^{M_{tij}} \right)$$

となる．対数を取ると

$$\begin{aligned}
&\log P(\boldsymbol{M} \mid \boldsymbol{N}, \boldsymbol{\theta}) \\
&= \sum_{t=0}^{T-2} \sum_{i\in V} \left(\log N_{ti}! - \sum_{j\in\Gamma_i} \log M_{tij}! + \sum_{j\in\Gamma_i} M_{tij} \log \theta_{ij} \right) \\
&\approx \sum_{t=0}^{T-2} \sum_{i\in V} \sum_{j\in\Gamma_i} (\log \theta_{ij} M_{tij} + M_{tij} - M_{tij} \log M_{tij}) \quad (1)
\end{aligned}$$

を得る．ただし，途中の変形でスターリングの近似 $\log n! \approx n \log n - n$ を使った．また，推定したい変数 $\boldsymbol{\theta}, \boldsymbol{M}$ に依存しない部分に関しては定数として省略した．この最右辺を $\mathcal{L}(\boldsymbol{M}, \boldsymbol{\theta})$ と置く．また，人数の保存則を表す制約

$$N_{ti} = \sum_{j\in\Gamma_i} M_{tij} \quad (t = 0, 1, \ldots, T-2), \qquad (2)$$

$$N_{t+1,i} = \sum_{j\in\Gamma_i} M_{tji} \quad (t = 1, 2, \ldots, T-1) \qquad (3)$$

が成立する．

$\boldsymbol{M}, \boldsymbol{\theta}$ を変数とみなし，制約 (2)(3) のもとで (1) を最大化することによって $\boldsymbol{M}, \boldsymbol{\theta}$ を推定する．最適化は，$\boldsymbol{M}$ と $\boldsymbol{\theta}$ に関する交互最適化によって行う．M に関する最適化問題は凸計画問題になっているため，準ニュートン法などの手法によって大域的最適解を求めることができる．また，$\boldsymbol{\theta}$ に関する最適化問題は，ラグランジュの未定乗数法により閉形式で解くことが可能である．この交互最適化のプロセスは，EM アルゴリズムにおいて E ステップを近似計算したものと解釈することが可能である [12].

セル i からの移動先候補 Γ_i の決め方として，[6] ではセル間の L_∞ 距離が 1 以下であるセルを採用している．

4.2 既存手法の限界

既存手法では，セル間の遷移確率を出発セルと到着セルのペアそれぞれについて用意し，それぞれの推定を行っている．しかしこの手法には，モデルの自由度が高くなりすぎるために解が絞りきれず，推定精度が低くなってしまうという問題点が存在する．すなわち，モデルの自由度が高くなることにより，観測 $\boldsymbol{N}$ をうまく説明する $\boldsymbol{\theta}, \boldsymbol{M}$ が非常に多く存在することになってしまうため，「真の $\boldsymbol{\theta}, \boldsymbol{M}$ とは異なるものの観測 $\boldsymbol{N}$ をうまく説明する $\boldsymbol{\theta}, \boldsymbol{M}$」を出力してしまうのである．図 2・図 3 にこの

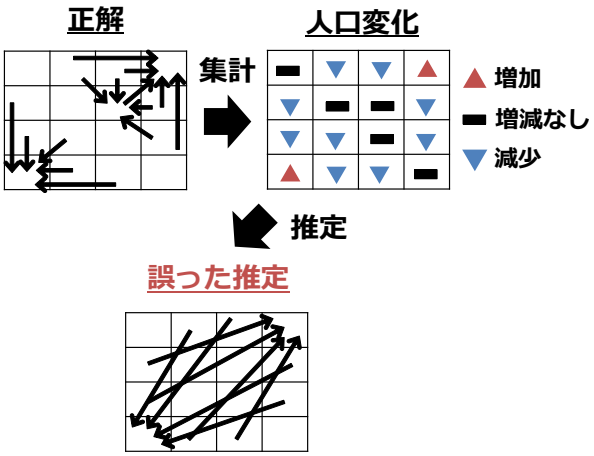

図 2　既存手法がうまくいかない 1 つ目の例．セル間の位置関係を考慮していないため人口の増減を説明する方法が多数存在し，解が絞りきれないため誤った推定が行われてしまっている．

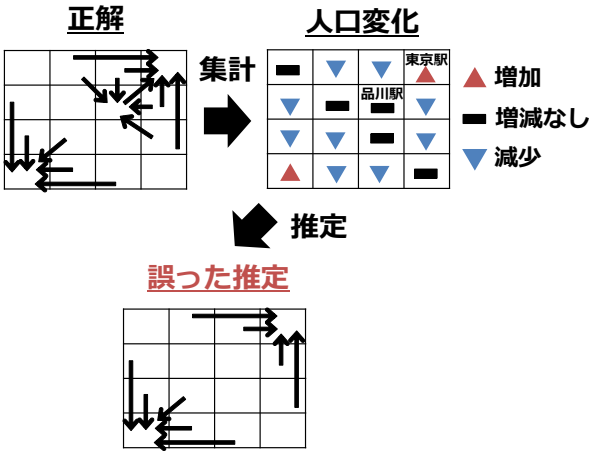

図 3　既存手法がうまくいかない 2 つ目の例．人口が変化していない「品川駅」と書かれているセルについて，「誰も流入しておらず，かつ誰も流出していない」と「流入した人数と流出した人数が等しい」のいずれでも数を合わせることができてしまうため，解が絞りきれず誤った推定を行ってしまっている．

問題が発生している 2 つの例を示している．左上の「正解」におけるそれぞれの矢印が，対象となるタイムステップ間において，人間の移動が存在していることを表している．右上の「人口変化」がそれらを集計した結果である，各セルにおけるタイムステップ間の人口の変化を表しており．▲は増加，▼は減少，-は増減なしを表す．そして，下側の図が推定された人の移動の様子を表している．図 2 の例では，セル間の位置関係を考慮していないため人口の増減を説明する方法が多数存在し，解が絞りきれないため誤った推定が行われてしまっている．図 3 の例では，セルの人口が変化していない「品川駅」と書かれているセルについて，「誰も流入しておらず，かつ誰も流出していない」と「流入した人数と流出した人数が等しい」のいずれでも数を合わせることができてしまうため，解が絞りきれず誤った推定を行ってしまっている． このようなモデルの自由度の高さに起因する問題は，推定するパラメータが多くなる場合，すなわち離れたセルへの移動も考慮しなければならない場合などに顕著になる．したがって，セルサイズが小さい場合や離散化の時間間隔が長い場合，推定精度が非常に悪くなってしまう．

5.　提 案 手 法

本章では，4.2 節で指摘した既存手法の問題点を解決し，精度の高い推定を行う手法を提案する．

5.1　アプローチ

本研究では，モデルに人の移動に関する前提知識を組み込んで推定を行う．これにより，従来手法では絞り込むことのできなかった解候補から，「人間の移動らしさ」を備えた解を選んで出力できるようになることが期待される．

本研究では，群衆の移動に関する以下の特徴を用いたモデル化を行った．

- 2 つのセル間の距離が移動のしやすさに影響する．近いセルには移動しやすく，遠いセルには移動しにくい．例えば朝の通勤時間帯において，横須賀→東京と移動する人の数よりも横須賀→横浜と移動する人の数が多いと考えられる．
- 人を集めやすいセルが存在する．例えば，朝の通勤時間帯においては，オフィス街があるセル（例えば東京駅を含むセルなど）には人が集まりやすいのに対し，ベッドタウンのある郊外のセル（例えば横須賀周辺のセル）には人が集まりにくい．逆に，夕方の帰宅時間帯においては，オフィス街セルに人は集まりにくくベッドタウンのあるセルには人が集まりやすくなると考えられる．
- セルごとに別のセルへの移動しやすさ（そのセルからの出発しやすさ）が異なる．朝の通勤時間帯においては，ベッドタウンのあるセルからは多くの人が流出するが，オフィス街セルからは人があまり流出しない．夕方の帰宅時間帯においては，逆の傾向を見て取ることができるようになると考えられる．

これらの知識を組み込むと，既存手法ではうまくいかないケースにおいても正しい推定を行うことが可能になる．このことを，4.2 節で示したケースを例に説明する，図 2 に示した誤った推定結果に関しては，セル間の位置関係を考慮し，近いセル間の移動は多く遠いセル間の移動は少ないという上に挙げた人間の移動の特徴を考慮した推定を行うことで，このような推定結果が出力されることを防ぐことができる．また，図 3 に示した誤った推定に関しては，「東京駅と書かれたセルには人が集まりやすい性質があるのにも関わらず，品川駅→東京駅という移動が少ない」という点が，上に挙げた人間の移動の特徴に符合していないため，このような推定結果が出力されにくくなる．

5.2　提案モデル

5.1 節で説明した人間の移動の特徴をもとに，セル間の遷移確率のモデリングを行う．セル i からの移動候補先 Γ_i は，セル間の距離関数 $d : V \times V \to \mathbb{R}_{\geqq 0}$ と，閾値 K を用いて $\Gamma_i = \{j \mid j \in V, d(i,j) \leqq K\}$ とする．距離関数としては，L_p-距離 $(p = 1, 2, \infty)$ などを使うことが考えられる．

セル i から セル j への遷移確率を θ_{ij} とする．提案手法では，θ_{ij} が

$$\theta_{ij} = \begin{cases} 1-\pi_i & (i=j) \\ \pi_i \cdot \frac{s_j \cdot \exp(-\beta \cdot d(i,j))}{\sum_{k \in \Gamma_i \setminus \{i\}} s_k \cdot \exp(-\beta \cdot d(i,k))} & (j \neq i, j \in \Gamma_i) \\ 0 & (\text{otherwise}) \end{cases} \quad (4)$$

という形で書くことができると仮定する．

π_i はセル i からの別のセルへの移動しやすさであり，$0 \leqq \pi_i \leqq 1$ を満たす．s_i はセル i への人の集まりやすさを表す値であり，$s_i \geqq 0$ を満たす．$\exp(-\beta \cdot d(i,j))$ は，移動確率と距離の関係を表す項であり，β は距離による移動確率の変化を表現するパラメータである．

この遷移確率のモデリングを直観的に説明する．セル i にいる個人が次のタイムステップにおいてどのセルにいるかは，以下のように決定される．

（1） それぞれの個人が，セル i から移動するかとどまるかが確率 π_i によって決まる．

（2） そのセルから移動することになった場合，移動先候補 $\Gamma_i \setminus \{i\}$ の中からそれぞれの候補に対して決まる「スコア」に比例した確率で移動先が選択され，そのセルに移動する．移動先候補 j の「スコア」は，セル j の人の集まりやすさ s_j 及びセル i とセル j の距離 $d(i,j)$ から計算される値（この場合 $\exp(-\beta \cdot d(i,j))$）の乗算によって計算される．

上記 (2) のように，遷移確率が移動先の「スコア」によって決まるという考え方は，Luce の選択モデルを用いた研究 ([10] などを参照) における考え方に近いが，エリア間の距離を考慮している点が先行研究と異なっている．遷移確率の距離に対する依存性に関しては，指数関数以外の関数（例えばベキ関数など）を使うことも可能であるが，以下説明の簡単のために指数関数の形に限定して説明を行う．

5.3 推定のための最適化問題

対数尤度関数の導出

対数尤度 (1) に (4) を代入すると

$$\begin{aligned} &\log P(\boldsymbol{M} \mid \boldsymbol{N}, \boldsymbol{\pi}, \boldsymbol{s}, \beta) \\ &= \sum_{t=0}^{T-2} \sum_{i \in V} \log(1-\pi_i) M_{tii} \\ &+ \sum_{t=0}^{T-2} \sum_{i \in V} \sum_{j \in \Gamma_i \setminus \{i\}} \Bigg[\log \pi_i + \log s_j - \beta \cdot d(i,j) \\ &- \log \sum_{k \in \Gamma_i \setminus \{i\}} s_k \cdot \exp\left(-\beta \cdot d(i,k)\right) \Bigg] M_{tij} \\ &+ \sum_{t=0}^{T-2} \sum_{i \in V} \sum_{j \in \Gamma_i} \left(M_{tij} - M_{tij} \log M_{tij} \right) \end{aligned}$$

を得る．ただし定数部分に関しては省略した．この右辺を $\mathcal{L}(\boldsymbol{M}, \boldsymbol{\pi}, \boldsymbol{s}, \beta)$ とおく．

$\boldsymbol{M}, \boldsymbol{\pi}, \boldsymbol{s}, \beta$ を変数として，制約 (2)(3) のもとで $\mathcal{L}(\boldsymbol{M}, \boldsymbol{\pi}, \boldsymbol{s}, \beta)$ を最大化することによって，$\boldsymbol{M}, \boldsymbol{\pi}, \boldsymbol{s}, \beta$ の推定を行う．ただし，制約 (2)(3) に関しては，観測にノイズがある場合なども考慮して，ソフトな制約として目的関数に組み込む．すなわち，

Algorithm 1 Estimation Algorithm

1: **Input**: population data: $\boldsymbol{N}$
2: **Output**: movement $\boldsymbol{M}$, parameter $\boldsymbol{\pi}, \boldsymbol{s}, \beta$
3: Initialize parameters $\boldsymbol{M}, \boldsymbol{\pi}, \boldsymbol{s}, \beta$
4: **while** condition f **do**
5: $\quad \boldsymbol{M} \leftarrow \arg\max_{\boldsymbol{M}} \mathcal{L}(\boldsymbol{M}, \boldsymbol{\pi}, \boldsymbol{s}, \beta)$
6: $\quad \boldsymbol{\pi} \leftarrow \arg\max_{\boldsymbol{\pi}} \mathcal{L}(\boldsymbol{M}, \boldsymbol{\pi}, \boldsymbol{s}, \beta)$
7: $\quad \boldsymbol{s}, \beta \leftarrow \arg\max_{\boldsymbol{s}, \beta} \mathcal{L}(\boldsymbol{M}, \boldsymbol{\pi}, \boldsymbol{s}, \beta)$
8: **end while**
9: **return** $\boldsymbol{M}, \boldsymbol{\pi}, \boldsymbol{s}, \beta$

$$\begin{aligned} &\mathcal{L}'(\boldsymbol{M}, \boldsymbol{\pi}, \boldsymbol{s}, \beta) := \mathcal{L}(\boldsymbol{M}, \boldsymbol{\pi}, \boldsymbol{s}, \beta) \\ &- \frac{\lambda}{2} \sum_{t=0}^{T-2} \left| N_{ti} - \sum_{j \in \Gamma_i} M_{tij} \right|^2 - \frac{\lambda}{2} \sum_{t=0}^{T-2} \left| N_{t+1,i} - \sum_{j \in \Gamma_i} M_{tji} \right|^2 \end{aligned}$$

と置いて，

$$\begin{aligned} \text{Max.} \quad & \mathcal{L}'(\boldsymbol{M}, \boldsymbol{\pi}, \boldsymbol{s}, \beta), \\ \text{s.t.} \quad & M_{tij} \geqq 0 \\ & (t = 0, 1, \ldots, T-2, i \in V, j \in \Gamma_i), \\ & 0 \leqq \pi_i \leqq 1 \quad (i \in V), \\ & s_i \geqq 0 \quad (i \in V), \quad \beta \in \mathbb{R} \end{aligned} \quad (5)$$

という最適化問題を解く．ただし，λ は制約をどれだけ強く守らせるかをコントロールするハイパーパラメータである．

5.4 推定アルゴリズム

最適化問題 (5) に基づいて $\boldsymbol{M}, \boldsymbol{\pi}, \boldsymbol{s}, \beta$ を推定する．最適化は，$\boldsymbol{M}, \boldsymbol{\pi}, \boldsymbol{s}, \beta$ に関する交互最適化によって行う．推定アルゴリズムの概略は， Algorithm1 に示されている．以下，それぞれのステップについて詳しく説明していく．

$\boldsymbol{M}$ の最適化

目的関数 $\mathcal{L}'$ は $\boldsymbol{M}$ について凹であり制約も凸集合であるため，$\boldsymbol{M}$ に関する最適化問題は凸計画問題になる． この最適化問題は，L-BFGS-B 法 [2] などの凸最適化法を適用することで大域的最適解を求めることができる．

$\boldsymbol{\pi}$ の最適化

目的関数 $\mathcal{L}'$ を $\boldsymbol{\pi}$ に関して整理すると，

$$\begin{aligned} \mathcal{L}' = \sum_{i \in V} \Bigg[&\left(\sum_{t=0}^{T-2} M_{tii} \right) \cdot \log(1-\pi_i) \\ &+ \left(\sum_{t=0}^{T-2} \sum_{j \in \Gamma_i \setminus \{i\}} M_{tij} \right) \cdot \log \pi_i \Bigg] \end{aligned}$$

となる．ただし，$\boldsymbol{\pi}$ に依存しない部分については省略した．これを最大化する $\boldsymbol{\pi}^*$ は，ラグランジュの未定乗数法より

$$\pi_i^* = \frac{\sum_{t=0}^{T-2} \sum_{j \in \Gamma_i \setminus \{i\}} M_{tij}}{\sum_{t=0}^{T-2} \sum_{j \in \Gamma_i} M_{tij}}$$

と閉形式で記述することができる．

$\boldsymbol{s}, \beta$ の最適化

目的関数 $\mathcal{L}'$ を $\boldsymbol{s}, \beta$ に関して整理すると，

$$\mathcal{L}' = \sum_{i \in V} \left[A_i \log s_i - B_i \log \left(\sum_{k \in \Gamma_i \setminus \{i\}} s_k \exp(-\beta \cdot d(i,k)) \right) \right] - \beta D \tag{6}$$

となる．ただし，

$$A_i := \sum_{t=0}^{T-2} \sum_{j \in \Gamma_i \setminus \{i\}} M_{tji}, \quad B_i := \sum_{t=0}^{T-2} \sum_{j \in \Gamma_i \setminus \{i\}} M_{tij},$$
$$D := \sum_{t=0}^{T-2} \sum_{i \in V} \sum_{j \in \Gamma_i \setminus \{i\}} d(i,j) \cdot M_{tij}$$

とおき，$\boldsymbol{\pi}, \beta$ に依存しない部分に関しては省略した．簡単のため，式 (6) の右辺を $f(\boldsymbol{s}, \beta)$ とおく．

$f(\boldsymbol{s}, \beta)$ の最大化を行うため，Minorization-Maximization アルゴリズム [5](以下 MM アルゴリズム) と呼ばれる枠組みを利用する．MM アルゴリズムは，関数を直接最大化するのが困難な場合に，その下界となる近似関数の最大化問題を逐次的に解くことによって，解の候補点系列を生成する手法である．MM アルゴリズムの枠組みを適用することによって，閉形式による更新と 1 変数の最適化問題を解くことによる更新を繰り返す，効率的なアルゴリズムを導出することができる．

MM アルゴリズムの具体的な適用方法について説明する．$x, y > 0$ について

$$-\log x \geqq 1 - \log y - \frac{x}{y} \tag{7}$$

が成り立つ．ここで，

$$x_i = \sum_{k \in \Gamma_i \setminus \{i\}} s_k \exp(-\beta \cdot d(i,k)),$$
$$y_i = \sum_{k \in \Gamma_i \setminus \{i\}} s_k^{(u)} \exp(-\beta^{(u)} \cdot d(i,k))$$

として (7) を $i \in V$ に適用することで，

$$f^{(u)}(\boldsymbol{s}, \beta) := \sum_{i \in V} \left[A_i \log s_i - C_i^{(u)} \sum_{k \in \Gamma_i \setminus \{i\}} s_k \exp(-\beta \cdot d(i,k)) \right] - \beta D$$

に対して

$$f(\boldsymbol{s}, \beta) \geqq f^{(u)}(\boldsymbol{s}, \beta) \quad (\forall \boldsymbol{s}, \beta) \tag{8}$$

を得る．ただし，

$$C_i^{(u)} := \frac{B_i}{\sum_{k \in \Gamma_i \setminus \{i\}} s_k^{(u)} \exp(-\beta^{(u)} \cdot d(i,k))}$$

とおいた．すなわち，$f^{(u)}(\boldsymbol{s}, \beta)$ は $f(\boldsymbol{s}, \beta)$ の下界となっている．

この下界 $f^{(u)}(\boldsymbol{s}, \beta)$ の最大化を繰り返すことで，$f(\boldsymbol{s}, \beta)$ の最大化を行う．アルゴリズムの概略は Algorithm 2 に示されている．Algorithm 2 中の $\boldsymbol{s}^{(u+1)} \leftarrow \arg\max_{\boldsymbol{s}} f^{(u)}(\boldsymbol{s}, \beta^{(u)})$ については，

Algorithm 2 Maximization of $f(\boldsymbol{s}, \beta)$

1: Initialize $\boldsymbol{s}^{(0)}, \beta^{(0)}$
2: $u \leftarrow 0$
3: **while** improving f **do**
4: $\quad \boldsymbol{s}^{(u+1)} \leftarrow \arg\max_{\boldsymbol{s}} f^{(u)}(\boldsymbol{s}, \beta^{(u)})$
5: $\quad \beta^{(u+1)} \leftarrow \arg\max_{\beta} f^{(u)}(\boldsymbol{s}^{(u+1)}, \beta)$
6: $\quad u \leftarrow u + 1$
7: **end while**
8: **return** $\boldsymbol{s}^{(u)}, \beta^{(u)}$

$$\frac{\partial f^{(u)}(\boldsymbol{s}, \beta^{(u)})}{\partial s_i} = 0 \Leftrightarrow$$
$$s_i = \frac{A_i}{\sum_{k \in \Gamma_i \setminus \{i\}} C_k^{(u)} \exp\left(-\beta^{(u)} \cdot d(k,i)\right)}$$

が成り立つため，$\boldsymbol{s}^{(u+1)}$ は閉形式で得ることができる．また，$\beta^{(u+1)} \leftarrow \arg\max_{\beta} f^{(u)}(\boldsymbol{s}^{(u+1)}, \beta)$ については，閉形式での更新を行うことができない．しかし簡単な計算より，$\forall \beta \in \mathbb{R}$ について $\frac{\partial^2 f^{(u)}(\boldsymbol{s}^{(u+1)}, \beta)}{\partial \beta^2} < 0$ が確かめられる．すなわち，$f^{(u)}$ は β に関して凹関数になっている．よって，$\beta^{(u+1)}$ を求めるためには β に関する 1 変数の凹関数最大化問題を解けばよく，これは黄金分割探索やニュートン法などによって効率的に行うことができる．

式 (7) の等号成立条件が $x = y$ であることから，

$$f(\boldsymbol{s}^{(u)}, \beta^{(u)}) = f^{(u)}(\boldsymbol{s}^{(u)}, \beta^{(u)}) \tag{9}$$

が成立する．式 (8)(9) より，

$$\begin{aligned} f(\boldsymbol{s}^{(u+1)}, \beta^{(u+1)}) &\geqq f^{(u)}(\boldsymbol{s}^{(u+1)}, \beta^{(u+1)}) \quad (\because (8)) \\ &\geqq f^{(u)}(\boldsymbol{s}^{(u+1)}, \beta^{(u)}) \geqq f^{(u)}(\boldsymbol{s}^{(u)}, \beta^{(u)}) \\ &= f(\boldsymbol{s}^{(u)}, \beta^{(u)}) \quad (\because (9)) \end{aligned}$$

を得る．すなわち，目的関数 $f(\boldsymbol{s}, \beta)$ は Algorithm 2 中において単調増加することが保証される．

6. 実　　験

6.1 実験設定

データ

実験においては，NAVITIME 社のカーナビアプリによって集められたカープローブデータ（実際に走行している車両から収集したデータ）を利用した．カープローブデータはルート案内を利用している間の移動軌跡しか残らないため，データがない部分を線形補間し，擬似的な一日の移動軌跡を作成した．作成した移動軌跡を時間間隔 1 時間・セルサイズ 5000m 及び時間間隔 1 時間・セルサイズ 10000m という 2 通りの離散化方法で集計化することによって，人口統計情報データを作成した．こうして作成された人口統計情報データに対して推定アルゴリズムを適用することによって得られた各タイムステップ間の推定移動人数と，もとの移動軌跡から計算した各タイムステップ間の正解移動人数を比較することによって, 推定精度の評価を

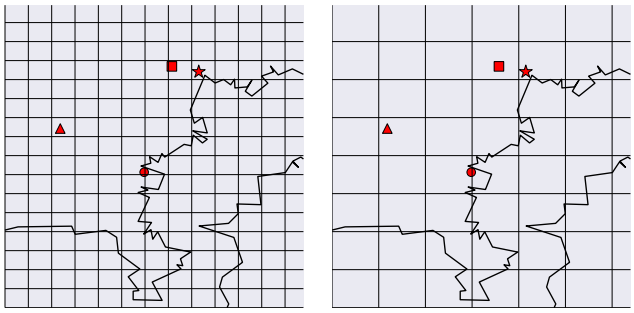

図 4　グリッド分割の様子．左側がセルサイズ 5000m，右側がセルサイズ 10000m．★は東京駅, ■は新宿駅, ●は横浜駅, ▲は町田駅をそれぞれ表している．

行った．推定の際には「各タイムステップにおける各セルの人口」という情報のみを入力としていることに注意されたい．

データに含まれるユーザ数は 8694 人であり，セル数はセルサイズ 5000m の場合 $16 \times 13 = 208$ 個，セルサイズ 10000m の場合 $8 \times 7 = 56$ 個である．グリッド分割の様子を図 4 に示している．

ベースライン手法

評価のために，4.1 章で説明したベースライン手法（他のパラメータによる表現を行わず $\boldsymbol{\theta}$ を直接推定する手法）との比較を行った．移動先候補セルの集合 Γ_i として,「セル i から L_∞ 距離で K 以内にあるセル全体の集合」を用い，K を 1, 5, 10 で動かすことにより 3 種類のベースライン手法を用意した．K が小さければ近くのセルへの移動のみを考慮し，K が大きければ遠くのセルへの移動も考慮することになる．先行研究 [6] では $K = 1$ の場合の Γ_i を利用している．提案手法においては，$K = 10$ とした．

評価指標

評価指標として，既存研究 [6] で用いられた指標である NAE (Normalized Absolute Error) 及び推定誤差を評価する一般的な指標である MAPE (Mean Absolute Percentage Error) を用いた．NAE は，M_{tij} を推定した移動人数，M^*_{tij} を真の移動人数として

$$\frac{\sum_{t=0}^{T-2}\sum_{i\in V}\sum_{j\in V}\left|M_{tij}-M^*_{tij}\right|}{\sum_{t=0}^{T-2}\sum_{i\in V}\sum_{j\in V}M^*_{tij}}$$

で定義される．この指標は 0 から 2 までの間の値をとり，推定が正しければ正しいほど小さな値をとる．また，MAPE は

$$\frac{1}{|V|^2\cdot T}\sum_{t=0}^{T-2}\sum_{i\in V}\sum_{j\in V}\frac{\left|M_{tij}-M^*_{tij}\right|}{M^*_{tij}}$$

と定義される．MAPE も，推定が正しければ 0 に近い値をとる．

データの時間帯分割

人々の移動の傾向（すなわちグリッド間の遷移確率）は時間帯によって変化すると考えられる．実験においては 1 日のデータ（6:00–21:00）を 5 つの時間帯（6:00–9:00，9:00–12:00，12:00–15:00，15:00–18:00，18:00–21:00）に分割し，それぞれの時間帯に対して推定アルゴリズムを適用した．すなわち各時間帯内においては遷移確率は共通であるとして推定が行われている．

6.2　実 験 結 果

図 5 に各データ・手法ごとの，それぞれの時間帯及び 1 日全体（total）の推定結果についての NAE および MAPE を示す．この実験結果から以下が観察される．

(1) **全体の推定精度**

提案手法はいずれのセルサイズ・指標においても，total スコアにおいて全てのベースライン手法よりも精度良く推定を行うことができている．

(2) **時間帯ごとの推定精度**

3 つのベースライン手法はそれぞれ得意とする時間帯が異なる．例えば，セルサイズ 5000m における NAE の結果に注目する．Baseline ($K = 1$) は 9–12, 12–15, 15–18, 18–21 などの時間帯においてはベースライン手法の中では優れた精度を示しているが，6–9 という時間帯においては非常に精度が悪くなっている．それに対し，Baseline ($K = 10$) は 6–9 という時間帯において比較的高い精度を達成している．これは，6–9 という時間帯は通勤・通学などにより長距離の移動が起こりやすく，その結果隣接グリッド以外への移動多くなるが，それ以外の時間帯では長距離の移動が比較的少ないことが理由であると考えられる．そして，ほとんど全ての時間帯において，提案手法はベースライン手法を上回る精度を達成している．

(3) **セルサイズごとの推定精度**

3 つのベースライン手法はそれぞれ得意とするセルサイズが異なる．セルサイズが 10000m の場合には Baseline ($K = 1$) が高い精度を示しているが，セルサイズが 5000m の場合には Baseline ($K = 5$), Baseline ($K = 10$) の方が高い精度を示している．これは，セルサイズが大きい場合はほとんどの移動が隣接セルへの移動になるが，セルサイズが小さくなると隣接セル以外への移動が多くなることが原因であると考えられる．そして，いずれのセルサイズにおいても，提案手法は高い精度を達成している．

6.3　推定結果の例

午前 7 時から午前 8 時の間に，★マークのついたセル（東京駅付近）へ移動した人数の真値（左）と提案手法による推定値（中央）とベースライン手法による推定値（右）を図 6 に示す．ただし，ここでいずれの手法においても $K = 5$ としている．ベースライン手法では，離れたセルからの移動人数が過大評価され，真値から大きく離れた推定結果になっているのに対し，提案手法は真値に近い推定を行うことができていることを読み取ることができる．この実験結果は，ベースライン手法は「観測をうまく説明するモデル」を多数持ち，そのようなモデルのうちの 1 つにトラップされて真値からかけ離れた解で推定が止まっているのに対し，提案手法は人間の移動の特徴を考慮することで解を絞り込み，正しい解を出力していることに起因していると考えられる．

7.　ま　と　め

本論文では，人口統計情報から移動人数を推定する問題を解くためのアプローチとして，Collective Graphical Model (CGM) と呼ばれる確率モデルに着目し，人間の移動の特徴をもとにし

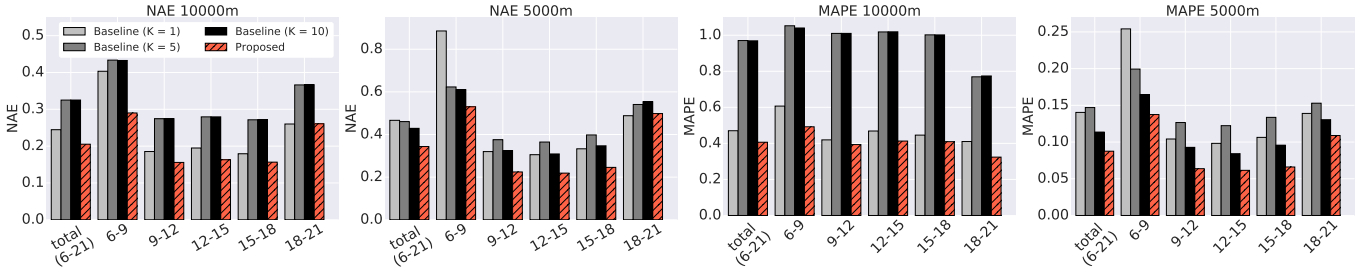

図 5　セルサイズ 5000m・10000m，それぞれの時間帯，提案手法・ベースライン手法の NAE と MAPE.

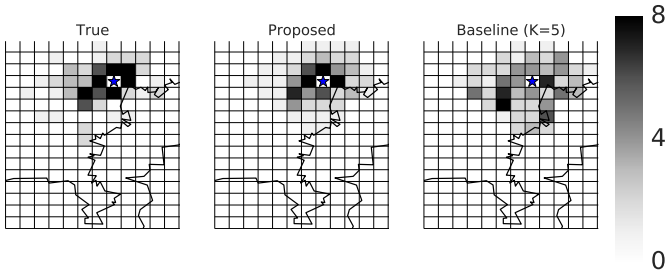

図 6　午前 7 時から午前 8 時の間に，★マークのついたセル（東京駅付近）へ移動した人数の真値（左）と提案手法による推定値（中央）とベースライン手法による推定値（右）.

たモデリングを行うことによって，既存研究では精度が低下してしまうデータでも高精度に推定が行える手法を提案した．手法の評価のために，カープローブデータから生成した人口統計情報で評価実験を行い，提案手法は既存手法よりも小さな誤差で推定を行うことができることを示した．

文　献

[1] G. Acs and C. Castelluccia. A Case Study: Privacy Preserving Release of Spatio-temporal Density in Paris. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*, pp. 1679–1688. ACM Press, 2014.

[2] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, Vol. 16, No. 5, pp. 1190–1208, sep 1995.

[3] J. Du, A. Kumar, and P. Varakantham. On Understanding Diffusion Dynamics of Patrons at a Theme Park. *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pp. 1501–1502, 2014.

[4] M. X. Hoang, Y. Zheng, and A. K. Singh. FCCF: forecasting citywide crowd flows based on big data. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '16*, pp. 1–10, New York, New York, USA, 2016. ACM Press.

[5] D. R. Hunter. MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, Vol. 32, No. 1, pp. 384–406, feb 2003.

[6] T. Iwata, H. Shimizu, F. Naya, and N. Ueda. Estimating People Flow from Spatiotemporal Population Data via Collective Graphical Mixture Models. *ACM Transactions on Spatial Algorithms and Systems*, Vol. 3, No. 1, pp. 1–18, may 2017.

[7] A. Kumar, D. Sheldon, and B. Srivastava. Collective Diffusion Over Networks: Models and Inference. *Proceedings of the 29th Conference on Uncerainty in Artifical Intelligence*, 2013.

[8] R. Kumar, A. Tomkins, S. Vassilvitskii, and E. Vee. Inverting a Steady-State. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pp. 359–368, New York, New York, USA, 2015. ACM Press.

[9] R. D. Luce. *Individual Choice Behavior: A Theorerical Analysis*. Jhon Wiley & Sons, Inc., New York, 1959.

[10] L. Maystre and M. Grossglauser. ChoiceRank: Identifying Preferences from Node Traffic in Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 2354–2362, jul 2017.

[11] D. R. Sheldon and T. G. Dietterich. Collective Graphical Models. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, pp. 1161–1169, 2011.

[12] D. Sheldon, T. Sun, A. Kumar, and T. Dietterich. Approximate Inference in Collective Graphical Models. *Proceedings of the 30th International Conference on Machine Learning*, pp. 1004–1012, feb 2013.

[13] T. Sun, D. Sheldon, A. Kumar, and A. E. Sg. Message Passing for Collective Graphical Models. *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, 2015.

[14] M. Terada, T. Nagata, and M. Kobayashi. Population Estimation Technology for Mobile Spatial Statistics. *NTT DOCOMO Technical Journal*, Vol. 14, No. 3, pp. 10–15, 2013.

[15] F. Xu, Z. Tu, Y. Li, P. Zhang, X. Fu, and D. Jin. Trajectory Recovery From Ash: User Privacy Is NOT Preserved in Aggregated Mobility Data. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 1241–1250, New York, New York, USA, 2017. ACM Press.

[16] J. Zhang, Y. Zheng, and D. Qi. Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction. In *Proceedings of the 31st AAAI Conference*, pp. 1655–1661, 2017.