トピックモデルを利用した地域別人口特性の把握手法の提案

神谷　啓太[1]・布施　孝志[2]

[1]学生会員　東京大学 工学系研究科 社会基盤学専攻・日本学術振興会特別研究員 DC（〒 113-8656 東京都文京区本郷 7-3-1）
E-mail: kamiya@trip.t.u-tokyo.ac.jp
[2]正会員　東京大学教授　工学系研究科 社会基盤学専攻（〒 113-8656 東京都文京区本郷 7-3-1）
E-mail: fuse@civil.t.u-tokyo.ac.jp

近年の情報通信技術と観測技術の発達に伴い、メッシュ人口データなどのデータ蓄積と利用が急速に進んでいる．これらの高分解能なデータを用いることで、より詳細な移動情報をもとにした人口特性の把握が期待できる．そこで本研究ではトピックモデルに基づく地域別人口特性の把握手法を提案する．メッシュを文書，滞在者の居住地を一つの単語とし，人口特性としてのトピックの意味解釈を行う．また，トピック数の事前設定が困難である課題に対し，データに応じてトピック数を自動的に推定することができる HDP-LDA を援用する．500m メッシュのモバイル空間統計データに対する適用実験を通じ，推定されたトピック分布を用いることで，地域別人口特性を解釈するための新たな指標となる可能性を確認した．

Key Words: *topic model, human dynamics, Dirichlet process, population data*

1. はじめに

流動的な人口の把握が多分野から着目されている[1)−5),6)]．例えば都市内の人々の移動や分布等を分析することで都市計画や災害対策，マーケティング等に有益となる．また，GPS 等を利用した測位技術の発達により，時間的・空間的に高分解能かつ低コストで位置情報の入手が容易となっている．例えばモバイル空間統計やメッシュ型流動人口データに代表されるように，匿名性および入手可能性が高いメッシュ人口データの活用がより期待される．中でも，これら高分解能で得られたメッシュ人口データの分析を通じ，より詳細な移動情報をもとにした人口特性の把握に活用することが期待できる．

人口特性を把握するための手法としても様々考えられるが、上述の大規模データに対して適用可能な手法の一つに、トピックモデルや潜在的意味解釈によるアプローチが考えられる．これは，観測された人口分布や OD データに潜むパターンを抽出し，そのパターンの意味解釈を行おうとする統計手法である．特にトピックモデルを基にした分析手法は，そのモデリングの拡張可能性の高さや大規模データへの適用可能性などの理由から，自然言語処理をはじめ，画像や購買データなどの大規模かつ非構造化データに対して盛んに適用されている．

ただし，トピックモデルを用いた分析手法では，トピック数を事前に設定する必要がある．適切な，もしくは分析目的に応じたトピック数の決定が常に大きな課題となるが，残念ながらメッシュ人口データが取りうる適切なトピック数は未知である．しかしながら，データから適切な状態数を推定することができれば，この問題は克服できると考えられる．そのため本研究では，データに応じてトピック数を自動的に推定することができる HDP-LDA（Hierarchical Dirichlet Process Latent Dirichlet Allocation）[7)] を援用する．以上の背景に基づき，本研究では HDP-LDA に基づく分析を実施し，地域別人口特性の把握手法の提案を行う．

本論文の構成は以下の通りである．第 1 章では本研究の背景を述べた．第 2 章で，ノンパラメトリックベイズについて概説した後，HDP-LDA を導入する．次に第 3 章では，本研究において HDP-LDA の枠組みをいかにメッシュ人口データに適用するかを議論する．第４章で手法の適用結果と考察を述べた後，最後に本研究の成果と今後の課題を第 5 章でまとめる．

2. HDP-LDA の導入

本章では階層ディリクレ過程の枠組みをトピックモデルに応用した HDP-LDA[7)] の導入を行う．このモデルはトピック数をデータから自動的に推定することができるため，トピック数を事前に設定することが難しいと考えられるメッシュ人口データに対する適用が有効である．まずディリクレ過程について概説した後，本研究で用いる HDP-LDA を説明する．

(1) ディリクレ過程

ディリクレ過程（Dirichlet process; DP）[8] は「分布に対する分布」と呼ばれており，パラメータ空間 Φ 上の基底測度 H と集中パラメータ $\alpha > 0$ によって一意に定義される．ここでは，DP(α, H) と記す．端的には，基底測度 H を，それに似た無限次元の離散分布によって近似した分布 G_0 を生成する確率過程であると換言でき[9]，$G_0 \sim \mathrm{DP}(\alpha, H)$ と記述される．また，G_0 は具体的には以下のように立式できる．

$$G_0 = \sum_{k=1}^{\infty} \pi_k \delta(\phi_k), \quad \phi_k \sim H \tag{1}$$

ここで，$\delta(\cdot)$ はディラックデルタである．なお，$\delta(\phi_k)$ は ϕ_k 上のアトムと呼ばれ，パラメータ空間 Φ から抽出したパラメータ ϕ_k に対応する離散分布である．

また，重み π_k は以下の棒折り過程（GEM）と呼ばれる確率過程から生成される．

$$\nu_k \sim \mathrm{Beta}(1, \alpha) \tag{2}$$

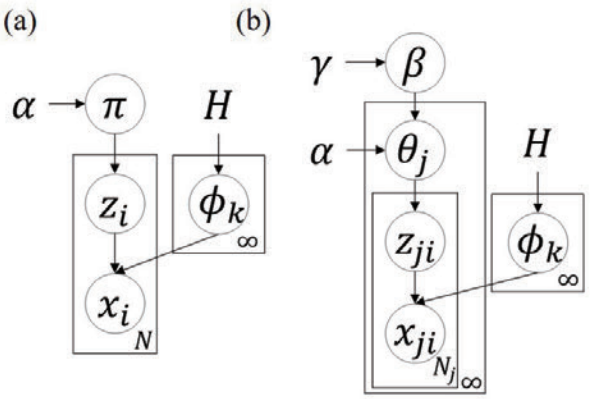
$$\pi_k = \nu_k \prod_{l=1}^{k-1} (1 - \nu_l) \tag{3}$$

ここで，これらの重みをまとめて $\boldsymbol{\pi} = \{\pi_1, \pi_2, ..., \pi_\infty\}^{\mathrm{T}}$ と記す．なお，T は転置である．棒折り過程ではベータ分布 Beta$(1, \alpha)$ から抽出した ν_k 分だけ，確率の総和である長さ 1 の棒を無限に切っていき，無限次元ベクトル $\boldsymbol{\pi}$ を生成する．棒折り過程は無限次元の多項分布を生成する確率過程であり，$\boldsymbol{\pi} \sim \mathrm{GEM}(\alpha)$ と記される．集中パラメータ α が大きいと多項分布 $\boldsymbol{\pi}$ の始めの項に重みが平均的に集中するため，基底測度 H をより少ない離散分布 $\delta(\phi_k)$ で近似することになる．

ディリクレ過程を混合分布モデルの事前分布として用いたものをディリクレ過程混合モデルと呼ぶ．そのグラフィカルモデルは**図–1** (a) に示す通りである．ここで，i 番目の出力を $\boldsymbol{x}_i$，出力分布パラメータを紐付けるインデックスを z_i，データサイズを N とする．このモデルでは，棒折り過程 GEM(α) より構成された多項分布 $\boldsymbol{\pi}$ からインデックス z_i が生成され，そのインデックス $k = z_i$ に対応するパラメータ ϕ_k が割り当てられる．その後，ϕ_k によって定義された出力分布 F によって観測値 $\boldsymbol{x}_i$ が出力される．なお，このモデルではパラメータ空間から抽出されるアトム $\delta(\phi_k)$ が出力毎に異なるため，潜在状態 z_i は統一されていない．

(2) 階層ディリクレ過程

階層ディリクレ過程（Hierarchical Dirichlet Process; HDP）[7] では，ディリクレ過程を別のディリクレ過程の事前分布として設定することでアトムを共有して潜在状態の統一化を図る．まず，1 層目のディリクレ過程 DP(γ, H) によりグローバルな確率測度 G_0 を生成する．



図–1 (a) ディリクレ過程混合モデル, (c) 階層ディリクレ過程.

さらに G_0 が 2 層目のディリクレ過程 DP(α, G_0) の基底測度となり，ローカルな確率測度 G_j が生成される．ここで，グローバルな確率測度 G_0 は，ローカルな確率測度 G_j の平均的な分布と解釈される．ディリクレ過程の特徴により G_0 は離散分布になるため，G_j はインデックス j が異なっていても同様のアトム $\delta(\phi_k)$ を共有することになる．これにより，異なる混合分布間で同一の潜在状態集合および出力分布パラメータ集合を共有することができる．

図–1 (b) に階層ディリクレ過程のグラフィカルモデルを示す．ここでは，棒折り過程 GEM(γ) よりグローバルな無限次元多項分布 $\boldsymbol{\beta}$ が生成され，次にローカルな無限次元多項分布 $\boldsymbol{\theta}_j$ が DP$(\alpha, \boldsymbol{\beta})$ から生成される．そして，潜在状態 z_{ji} と観測値 $\boldsymbol{x}_{ji}$ が DP と同様の機構で出力される．

(3) Latent Dirichlet Allocation

次に，トピックモデルとしての LDA について概説する．なお，LDA についての解説は参考文献[10)11)12)] をもとに記す．

LDA はもともと文書の確率的生成モデルとして提案されたモデルである．ただし，文章の順序は無視し，Bag of Words(BoW) 表現と呼ばれる単語と出現頻度のペアの集合をモデル化する．ここで，BoW 表現は単語が共起している現象を表しており，この共起性を用いて単語や文書をクラスタリングする手法である，

LDA では，文書中の単語に対応する潜在変数（トピック）を導入する．具体的には，文書 d の i 番目の単語を $w_{d,i}$ として，対応する潜在変数を $z_{d,i}$ と定義する．トピックの添字集合を $\{1, 2, ..., K\}$ とする（$z_{d,i} \in \{1, 2, ..., K\}$）．また，各トピックはそれぞれに対応した単語の出現分布 $\phi_k (k = 1, 2, ..., K)$ を有している．

文書数を M，文書 d の文章長（総単語数）を n_d，総トピック数を K とする．LDA では，文章は複数のト

ピックから構成され，その構成比を離散分布としてもつ．$\theta_{d,k}$ を，文章 d でトピック k が出現する確率（文章 d でのトピック k の構成比率）とし，トピック分布を $\boldsymbol{\theta}_d = (\theta_{d,1}, \ldots, \theta_{d,K})$ とする．$\phi_{d,v}$ をトピック k における単語 v の出現確率とし，単語の出現分布を $\boldsymbol{\phi}_k = (\phi_{k,1}, \ldots, \phi_{k,V})$ とする．生成モデルとしては，$\boldsymbol{\theta}_d$ や $fat\phi_k$ は Dirichlet 分布 (Dir と表記) による生成を仮定する．すなわち，

$$\boldsymbol{\theta}_d \sim Dir(\boldsymbol{\alpha}), d = 1, \ldots, M \tag{4}$$

$$\boldsymbol{\phi}_k \sim Dir(\boldsymbol{\beta}), k = 1, \ldots, K \tag{5}$$

ここで、$\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)$ は K 次元ベクトル、$\boldsymbol{\beta} = (\beta_1, ..., \beta_V)$ は V 次元ベクトルで、いずれも Dirichlet 分布のパラメータである。単語 $w_{d,i}$ や潜在トピック $z_{d,i}$ は離散値なので、多項分布（Multi と表記）を生成分布として仮定する。すなわち、各文書 $d(= 1, ..., M)$ において、各単語は以下の生成過程を仮定する。

$$z_{d,i} \sim Multi(\boldsymbol{\theta}_d) \tag{6}$$

$$w_{d,i} \sim Multi(\phi_{z_{d,i}}), i = 1, ..., n_d \tag{7}$$

(4) HDP-LDA

上記の LDA モデルでは，総トピック数 K を事前に設定する必要がある．そこで階層ディリクレ過程を LDA に拡張し，原理的には無限個のトピック数を許容したものが HDP-LDA である．HDP-LDA では，1 層目のディリクレ過程によりグローバルなトピック分布 $\boldsymbol{\beta}$ を生成した後，2 層目のディリクレ過程によってローカルな（各文書の）トピック分布 $\boldsymbol{\theta}_d$ を生成する．一方単語分布では単語空間上の一様事前分布 H からディリクレ過程による多項分布 ϕ_k のサンプリングを無限回行う．その後の，トピック $z_{d,i}$ の決定と，そのトピックに紐づいた単語分布 $\phi_{z_{d,i}}$ からの単語の生成過程は通常の LDA と同様である．

計算上は全てのトピックが用いられるわけではないため，観測データに応じてトピック数を決定することのできるモデルであると解釈可能である．この特性により，トピック数が未知である場合のクラスタリング問題に多く適用されている．

3. メッシュ人口データにおけるトピックモデルの考え方

本研究において HDP-LDA の枠組みをいかにメッシュ人口データに適用するかを議論する．自然言語処理では，以下の仮定に基づいてトピックモデルを適用している．

- 文書は互いに独立である
- 文書は複数のトピックの重み付き和を潜在的に保有している
- 文書は、トピック→単語の順で構成する単語数をサンプリングしたものである

本研究では，文書に対応するものが「メッシュ」であり，そのメッシュに存在する滞在者の「居住地」が単語に対応する概念と考える．それゆえトピックに該当するものが,「ある地域からの滞在者が潜在的に多い」などのパターンであり，これを「人口特性」の一つと考える．ただし，どの地域にも大勢が存在するような「居住地」はトピックモデルの枠組みによって特徴的な単語とはならないため，単なる OD データの大小に基づく分析では得られない潜在的なトピックが得られると考えられる．また，文書ごとに存在するトピック分布から文書特徴を把握するように,「メッシュ」ごとに存在するトピック分布から「地域別人口特性」を把握しようとするのが，本研究におけるアイディアである．

これら前提の上で，メッシュ人口に対してトピックモデルを適用するにあたり、上記に対応するような以下の仮定を置く．

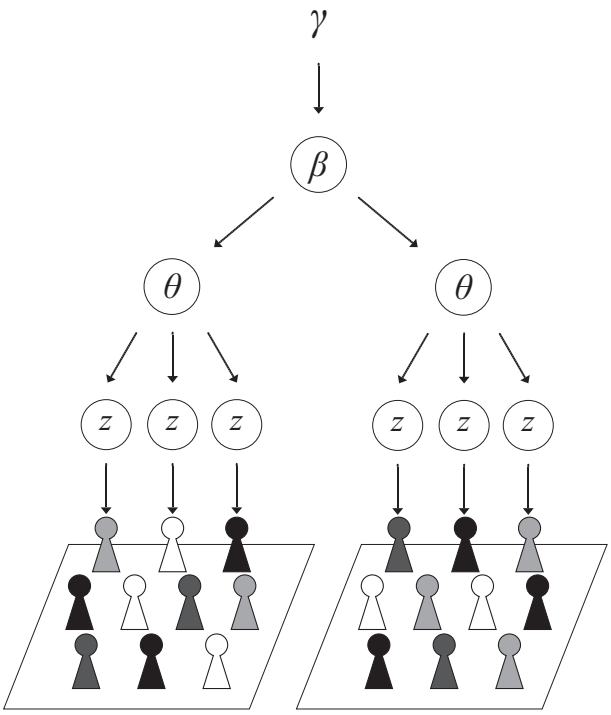
- 本研究では各時刻・各メッシュを独立に取り扱う
- 各メッシュは、複数のトピック（人口特性）の重み付き和で表現される
- 各メッシュに存在する人口そのものの値については分析対象とせず，所与のものとする
- トピックに応じて滞在者の「居住地」がサンプリングされる機構とする

また，図-2 は本研究におけるメッシュ人口データの生成過程を模式的に表したものである．各時刻・各地点で独立したメッシュに対してそれぞれトピック分布 θ が割り当てられている．その θ からトピック z がサンプリングされ，その時刻・そのメッシュに存在する滞在者数に応じて「どこから来たか（居住地）」を割り振っていく，という流れである．図中の人型シルエットの濃淡の違いが居住地の違いを表しており，その濃淡（居住地）の共起性をもとにした人口特性を推定する．

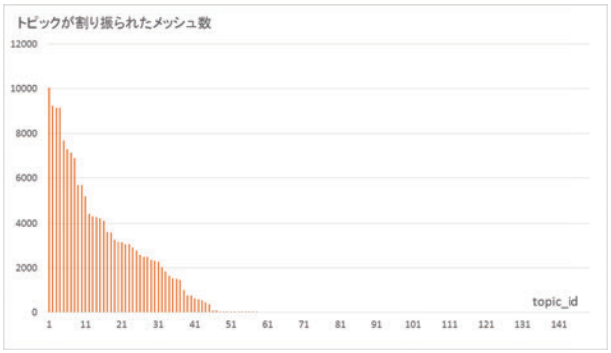
4. 手法の適用とその結果

(1) 適用データと実験条件

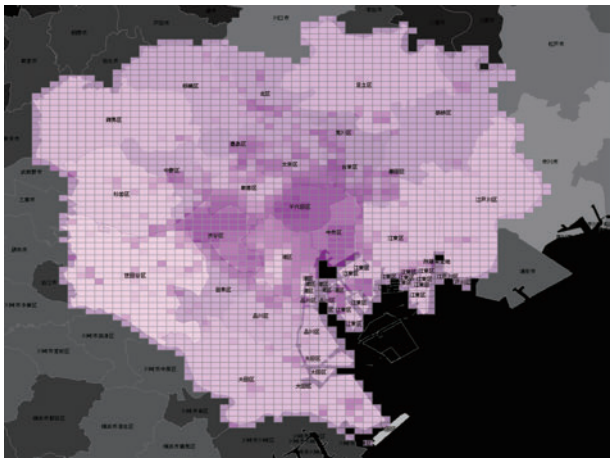
東京 23 区を対象とし，2015 年 6 月平日の平均値を表す，約 500m メッシュのモバイル空間統計データを使用した．対象範囲は 2,607 メッシュで構成され，1 時間ごとの居住地域別人口が得られている．本研究では各時刻・各メッシュを独立した文書として扱うため，文書数としては 2,607 メッシュ× 24 時間＝ 62,568 文書相当である．単語として用いる居住地としては，モバイル空間統計のデータ性質上，都道府県レベル，市区町



図–2 HDP-LDA を用いたメッシュ人口の表現方法.



図–3 各トピックの文書採択数.



図–4 15 時におけるトピック 2 の強度分布.

村レベル，町丁目レベルでの入手が可能である．トピックの意味解釈を行う上ではなるべく高分解能が望まれ，一方で安定した推定のためには単語数が少ない，すなわち低分解能が望まれる．そこで，本研究では市区町村レベルを居住地の単位として利用した．

HDP-LDA の実装には python 用ライブラリ gensim を使用した．また，HDP-LDA のパラメータ事後分布の推定には，オンライン変分推定手法[13] を用いた．これは，理論上は無限次元で行われる第 1 層および第 2 層のディリクレ過程を，十分な大きさのトピック数上限値 T および K を設定し，ディリクレ分布として近似する過程を含む手法である．今回の実験においては，$T = 150$, $K = 15$ を採用した．またそれぞれ，学習係数 $\kappa = 1.0$, ミニバッチサイズ $S = 256$, 学習制御パラメータ $\tau_0 = 64$ とした．これらのパラメータの詳細については，参考文献[13] を参考にされたい．また，ハイパーパラメータについては $\gamma = 1.0$, $\alpha = 1.0$ とした．

(2) トピック数の推定結果

上記の設定の下，HDP-LDA をメッシュ人口データに適用し，モデルの推定を行い，各トピックにおける単語分布 ϕ_k，および各文書（メッシュ）におけるトピック分布 θ_d を推定した．図-3 は，各トピックが，いくつのメッシュから採択されたのかを示している．すなわち，各トピックがいくつのメッシュの構成トピックとなっているかを示している．なお，トピック ID 自体は意味を持たないため，この値を降順にソートし，トピック ID を振りなおしている．

この結果から，上位 45 個のトピックが 100 を超える文書（メッシュ）の構成トピックとなっていることが確認された．103 個のトピックについては数メッシュから数十メッシュの構成トピックとなっていることが確認されたが，これらは主要なトピックではないと考えられる．残る 2 個のトピックについては確率が付与されなかった（どのメッシュの構成トピックともなっていない）．以上より，45 個の主要なトピックが HDP-LDA により自動推定されたことが確認された．

(3) 推定されたトピックの意味解釈

次に，各トピックにおいてどのような単語分布 ϕ_k が得られたのかを確認する．ただし，推定された 45 個の主要トピックすべてに関して提示することができないため，ここではそのうちの上位 10 トピックに限って図-5 に可視化した．

この図において，より赤く塗られた市区町村がそのトピックにおける高い出現確率を有していることを意

味する．例えば，トピック 0 においては足立区や荒川区などを筆頭に，東京都心から見て北東部に位置する市区町村が集まったトピックと解釈できる．また，トピック 1 では板橋区，北区，練馬区から川越市までの北西部に位置する市区町村が集まるトピックと解釈できる．トピック 2 においては，世田谷区，渋谷区などが出現確率が特に高い地域であるとともに，23 区の出現確率が高いという特徴から，23 区の内内移動に関するトピックであると考えられる．以降の主要トピックにおいても，あるまとまった地方が同じトピックに集中する様子が確認された．また，46 番目以降のマイナーなトピックに関しては，関東全域に単語出現分布が広がっているなど，地理的に顕著な傾向は見られなかった．

今回の実験において，各単語間，すなわち市区町村間の距離や近接性などの情報は与えておらず，メッシュ人口における居住地の共起性のみを用いて各トピックの単語分布を推定した．そのため，あるトピック内でともに高い出現確率を有する関係にある市区町村は，その居住者の滞在先や移動先に共起性があるということである．特に本実験では 500m メッシュ人口データを用いているため，市区町村レベルでの OD 情報に基づく分析よりも，より詳細な移動情報をもとにした人口特性把握となることが期待できる．

(4) トピック分布を用いた分析への考察

最後に，各文書（メッシュ）におけるトピック分布 θ_d 側からの分析を行う．図-4 に示すのは，15 時におけるトピック 2 の強度分布，すなわち各メッシュにおけるトピック 2 の重みを可視化したものである．濃い紫色のメッシュほどトピック 2 の重みが大きいことを示しており，背景の市区町村は明るいほど単語分布の出現確率が高いことを示している．渋谷区，千代田区，中央区内のメッシュに特にトピック 2 の重みが大きい地域が存在していることが確認できる．一方で，単語分布にて出現確率が高い世田谷区内においてはトピック 2 の重みが大きなメッシュは少ないことが分かる．

ただし，HDP-LDA によって自動的にトピック数を推定したとはいえ，今回の結果における 45 個の主要トピックおよび 24 時間分すべてについて上記の分析を行うことは困難である．また，本研究においては各時刻で独立してトピックモデルへと適用したが，同一メッシュにおける時系列性を考慮する必要があると考えられる．そのため，Dynamic topic model[16] や Topics over time[15] など、トピックの時間変化を含めたモデルに拡張することにより，検討すべきトピック分布項目を減らすなどが有効であると考えられる．

5. おわりに

本研究ではメッシュ人口データを対象に，トピックモデルに基づく人口特性の把握手法を提案した．メッシュを文書，滞在者の居住地を単語，人口特性をトピックと捉えたうえで，データに応じてトピック数を自動的に推定することができる HDP-LDA によってモデル推定を行う機構である．

500m メッシュのモバイル空間統計に対して本手法を適用したところ，事前情報を与えていないのに関わらず，まとまった地方が同じトピックに集中する様子が確認された．あるトピック内でともに高い出現確率を有する関係にある市区町村は，その居住者の滞在先や移動先に共起性があるということであり，市区町村レベルでの OD 情報に基づく分析よりも，より詳細な移動情報をもとにした人口特性把握となることが期待できる．

今後，同一メッシュにおける時系列性を考慮するため，Dynamic topic model[16] や Topics over time[15] など、トピックの時間変化を含めたモデルに拡張することが課題として挙げられる．また，複数の単語分布を許容できる Multiple-Source Latent-Topic[?] などを参考にモデル拡張を行い，メッシュ人口データ以外の交通・都市データを統合したうえで人口特性の把握手法の深化を行うことが望まれる．

謝辞: 本研究は JSPS 科研費 16J06150 の助成を受けたものです．

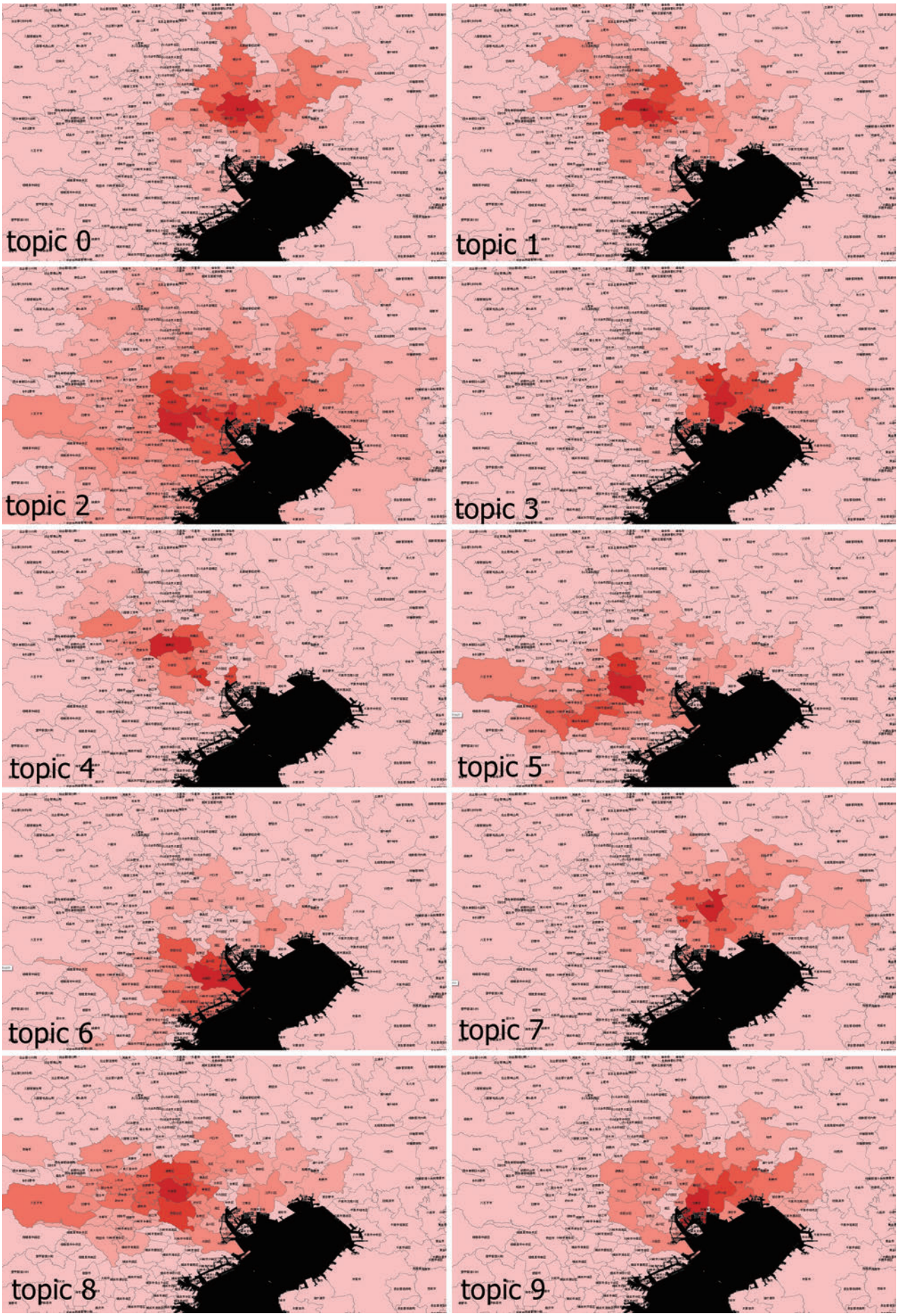
参考文献

1) Barabási, A. L.: The origin of bursts and heavy tails in human dynamics, *Nature*, Vol. 435, pp. 207-211, 2005.
2) Sun, J. B., Yuan, J., Wang, Y., Si, H. B. and Shan, X. M.: Exploring space-time structure of human mobility in urban space, *Physica A: Statistical Mechanics and its Applications*, Vol. 390, No. 5, pp. 929-942, 2011.
3) Hasan, S., Schneider, C., Ukkusuri, S. V. and González, M. C.: Spatiotemporal patterns of urban human mobility, *Journal of Statistical Physics*, Vol. 151, No. 1-2, 2013.
4) Peng, C., Jin, X., Wong, K. C., Shi, M. and Lió, P.: Collective human mobility pattern from taxi trips in urban area, *PLoS ONE*, Vol. 7, No. 4, pp. 1-8, 2012.
5) Reades, J., Calabrese, F., Sevtsuk, A. and Ratti, C.: Cellular census: Explorations in urban data collection, *IEEE Pervasive Computing*, Vol. 6, No. 3, pp. 30-38, 2007.
6) 神谷啓太，布施孝志:メッシュ人口データに対するノンパラメトリックベイズに基づく統計的異常検知手法の適用可能性の検証，土木学会論文集 *D3*（土木計画学）, Vol. 72, No. 5, pp. I759-I769, 2016.
7) Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M.: Hierarchical Dirichlet processes, *Journal of the American Statistical Association*, Vol. 101, No. 476, pp. 1-30, 2006.
8) Ferguson, T. S.: A bayesian analysis of some nonparametric problems, *The Annals of Statistics*, Vol. 1, No. 2, pp. 209-230, 1973.

9) 持橋大地：最近のベイズ理論の進展と応用 [III] ノンパラメトリックベイズ，電子情報通信学会誌，Vol. 93, No. 1, pp. 73-79, 2010.
10) Blei, D. M., Andrew, Y. N. and Michael I. J.: Latent dirichlet allocation, *Journal of machine Learning research 3*, pp. 993-1022, 2003.
11) 岩田具治：トピックモデル，講談社，2015.
12) 奥村学：トピックモデルによる統計的潜在未解釈，コロナ社，2015.
13) Wang, C., Paisley, J. W. and Blei., D. M.: Online Variational Inference for the Hierarchical Dirichlet Process, *AISTATS*, Vol. 2, No. 3, pp. 752-760, 2011.
14) Blei, D. M. and Lafferty., J. D.: Dynamic topic models, *Proceedings of the 23rd international conference on Machine learning*, 2006.
15) Xuerui, W. and McCallum, A.: Topics over time: a non-Markov continuous-time model of topical trends, *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006.
16) Zheng, Y., Zhang, H. Yu, Y.: Detecting collective anomalies from multiple spatio-temporal datasets across different domains, *SIGSPATIAL'15*, No. 2, 2015.

PROPOSAL OF UNDERSTANDING TECHNIQUE OF THE REGIONAL POPULATION CHARACTERISTICS USING TOPIC MODEL

Keita KAMIYA and Takashi FUSE



図–5 トピック 0 からトピック 9 における単語分布の可視化結果.