GPSログデータを用いた訪日外国人旅行者の訪問パターンの分析手法の開発

本研究は，訪日外国人旅行者の訪問パターンの特性把握をGPSログデータを用いて行うことを目的とする．論文では，はじめに訪問地の組み合わせである訪問パターンの抽出にトピックモデルを用いることを説明するとともに，潜在クラス分析との差異を明らかにした．トピックモデルは教師データなしの機械学習（セグメンテーション手法）の1つと位置づけられ，セグメントの確率的導出過程が明示でき，セグメント相互が排他的でない特色を有する．分析では，訪日外国人向けナビゲーションアプリにより取得できたGPSログ情報を用いて，トピックモデルによって訪問場所の組み合わせパターンを幾つかのクラスに分類した．

キーワード　**GPSログデータ，訪日外国人旅行者，訪問パターン，トピックモデル**

古屋秀樹
FURUYA, Hideki
博士（工学）　東洋大学国際観光学科教授

岡本直久
OKAMOTO, Naohisa
博士（工学）　筑波大学システム情報系社会工学域教授

野津直樹
NOZU, Naoki
株式会社ナビタイムジャパン交通コンサルティング事業部事業責任者
和歌山大学国際観光学研究センター客員研究員

1——はじめに

訪日外国人旅行者数は，過去最高の2,403.9万人（2016年）となった[1]．訪日外国人旅行者数の増加は，大きな経済効果や外国への情報発信が期待でき，今後の東京オリンピック（2020年）をはじめとするビッグイベントへの布石として，一層のインバウンド振興の戦略的対応が必要不可欠と考えられる．政府は，あらたに訪日外国人旅行者数を2020年に4,000万人，2030年に6,000万人を目標として設定し，さまざまな施策が検討されている[2]．さらに，この観光ビジョンを推進するため短期的な行動計画として「観光ビジョン実現プログラム2016」（観光ビジョンの実現に向けたアクション・プログラム2016）も決定されている[3]．

この取り組みが効果を生むためには，継続的な取り組みに加え，関係主体が互酬性を持つ組織形態で取り組む必要性が高いと考えられる[4]．このような機能を持つものとして，近年，DMO（Destination Management Organization）が着目されており，その機能の1つに各種データ等の継続的な収集・分析とそれを参考にした明確なコンセプトに基づいた戦略の策定が示されている[5]．そこでは，既存データとともに，近年着目されているビッグデータの利用が考えられ，すでに様々な視点からの研究[6]に加えて，観光庁でもローミングデータ，アプリデータ，SNSを活用した観光振興の検討がなされている[7]．本研究は，この中でもGPSログデータを活用し，ヒートマップのようなある1時点における来訪者数，前後のトリップ目的地の連関などの実態把握と異なり，同一旅行者がどのような地点を訪問しているのか個人単位の訪日旅行に着眼し，訪問地の組み合わせである訪問パターンの分析手法を開発することを目的としている．この分析によって，来日回数や国籍，季節によって旅行者の行動特性にどのような違いがあるのか把握でき，当該地のみならず他の訪問地との関係性を踏まえながら観光振興の検討に資することができる．

一方，旅行者に着目した視点から考えると，訪日旅行者のニーズや来日動機を的確に把握することが重要と考えられるものの，旅行者を対象としたアンケート調査を随時実施することはコストなどの課題があり，ネット調査による意向把握も，その信頼性について十分担保されているとは言い難い．そこで，訪問地の組み合わせから，旅行者の来日動機を類推する方法も考えられ，移動データを的確に取得できた場合，実行動データに基づくために旅行行動の忘却やアンケート回答欄数の制約による影響も小さくなると考えられる．さらに，旅行者の移動データが例えば複数年など継続的に収集が可能であるとすると，季節別，年別にどのような差異が生じているか，今までになかった訪問パターンが顕在化しているのか，といったモニタリングが可能と考えられる．

以上を背景としながら，本分析では，訪日外国人向けナビゲーションアプリにより取得できたGPSログ情報を用いている．ユニークユーザー数（アプリの利用者数単位．以下，UU）が約5万人を数えるため，訪問パターンも多様である．そのため，組み合わせを効率的かつ論理的に集約し，

類似した訪問パターンをセグメントすることが重要と考えられる．そこで，本論文では多数のデータをセグメントでき，論理的整合性や具体的なセグメントの導出過程が明示できるトピックモデルを用いて分析を行う．

2——データ・分析手法ならびに本研究の位置づけ

2.1 利用データ

本研究では，株式会社ナビタイムジャパンが収集している訪日外国人旅行者の位置データ（GPSデータ）を用いる．このデータは，日本全国を対象としたスマートフォン向けアプリ「NAVITIME for Japan Travel」（iPhoneならびにAndroid用）をインストールした旅行者の中で，データ提供を承諾した旅行者の位置情報である．アプリの機能は，トータルナビ，路線図からの乗換検索，GPSを利用した現在地確認，約400万件のスポット検索，観光スポット情報，スポット・ルートの履歴/お気に入り，オフラインでの無料Wi-Fiスポット検索である．データの概要を下記に示す．

データ取得期間　2015年1月1日～12月31日

対象UU　国籍：日本国以外，居住国：日本国以外

データ項目　1）相対日（利用開始時からの日数），2）時刻，緯度経度，3）状態（stay：同一1kmメッシュ内に30分以上の連続した測位データが存在，break：同一1kmメッシュ内に7日以上の連続した測位データが存在，other：その他），4）個人属性（国籍・地域，性別，訪日回数，訪日目的，測位年月）

UU数　55,199（4,051万行，734レコード/UU）

なお，位置情報は2分ごと，かつ200m以上移動した場合に測位データを蓄積するものであり，個人情報秘匿の観点から，該当期間・該当メッシュでUU数が3以上のレコードのみがデータとして提供されている．本研究では，国籍を日本国以外，居住国を日本国以外で，かつデータ取得に協力したユーザのデータのみを分析に用いている．

2.2 既存研究と本論文の位置づけ

インバウンド観光の振興では，国籍・地域や年齢，旅行への嗜好を考慮しながら，適切にセグメントに分割して特性把握をすることがマーケティング活動に有効と考えられる．これらは，セグメンテーション，ターゲティング，ポジショニングの段階で示されるSTPマーケティングの考えに沿うものであり，その第一段階に相当する部分に着眼しているといえる．

本論文では，日本国内における訪問地およびその組み合わせ（訪問パターン）が，日本に対する旅行動機・ニーズ，旅行需要を反映していると仮定して，それに基づくセグメント導出を考える．このような訪問地点の訪問の有無といった多変量を集約して特徴量を抽出する方法として因子分析，主成分分析，数量化Ⅲ類があげられるが，多変量からの次元の圧縮を共分散行列のみによって行うため，外れ値が存在する場合に出現回数の多い特徴的なベクトルを抽出することが困難である[8]．また，区分分けでしばしば用いられるクラスター分析でも，何故異なるのかという因果構造を明示できていないとともに，妥当なクラスター数の決定が分析者の主観に依存すること，各サンプルが属するクラスが排他的であることなどの問題点・特徴がある．また，特徴的なカテゴリーの組み合わせを抽出するマーケットバスケット解析では，特徴的なペアリングを抽出できるものの，生成過程について内生化できていない問題点がある．また，立寄り行動を選択モデルとして表現してモデリングする事例[9),10)]は，各種政策変数による影響を推定できるものの，数多い訪問パターンや多様な行動規範を考慮することは困難と考えられる．

これらの問題点を解決する手法として，潜在クラス分析があげられる．この手法を適用した研究として，地方区分単位で潜在クラス分析を用いた研究[11)]や都道府県単位で分析した研究[12),13)]があり，そこでは外国人旅行者の訪問パターンが複数存在し，旅行者1名はいずれか1つのクラスに属すると仮定している．しかしながら，1旅行の中には，例えば自然資源への訪問と人文資源への訪問という複数の「トピック」の混在が考えられ，（訪問地の同定は）単独のトピックに基づきなされることを仮定する潜在クラスは制約の強いモデルであるといえる．また，モデル推定においてパラメータの事前確率を想定していないことから，データによる過学習（overfitting）の恐れもある．すなわち，極めて少数の他と傾向が大きく異る訪問地の組み合わせデータに過度にフィットしたトピック抽出を過学習と考え，それを避けて未知のデータに対するクラスの同定も正しくできる能力（汎化能力）を有するモデルを構築する必要がある．これらの問題点を改善するために，本研究ではトピックモデルを適用することとした．このトピックモデルは，主にテキストマイニングで用いられているが，訪問パターンに対して適用された事例は著者の知る限りない．

2.3 トピックモデルの概要[14),15),16),17),18)]

それぞれの旅行には，自然観光地の周遊，都市観光の実施，ゴールデンルートの体験などトピックが複数かつ確率的に存在し，そのトピックごとに訪問地への訪問比率が異なると仮定する．この時，各旅行はいずれのトピックに所属するのかあらかじめ与えられていないため，観測できていない潜在的トピックとして抽出できるようにモデル化を行う必要がある．トピックモデルには，各旅行のトピック構成率を最尤推定によって導出する確率的潜在意味解析

(Probabilistic Latent Semantic Analysis (PLSI)) があるが, 本論文では過学習をおさえ, 汎化性能 (generalization ability) の向上が期待できるLDAモデル (Latent Dirichlet Allocation) を用いる. LDAの生成過程は下記のとおりである.

2.3.1 訪問地別訪問比率分布の設定

トピック総数をKとすると, 各々のトピックkごとに訪問地別訪問比率Φ_kが存在すると仮定する (v:訪問地v, データセットに含まれるユニークな目的地の総数 (ユニーク訪問地数) :V. 例えば, ある個人が2つの地点A, Bを1回ずつ訪問した場合は, $\phi_A=0.5$, $\varphi_B=0.5$となる). なお, 訪問比率に各サンプルの訪問個所数を乗じると, 訪問地別訪問率を算出できるものとする.

$$\boldsymbol{\Phi}_k=(\varphi_{k1},\ \cdots,\ \varphi_{kV}) \tag{1}$$

ここで, $\varphi_{kv}=p(v|\boldsymbol{\Phi}_k)$, $\varphi_{kv}\geqq 0$, $\Sigma_v\varphi_{kv}=1$.

上記から, 同一の訪問地でも異なるトピックに含まれることが推察でき, 出現する訪問地の組み合わせによって異なった旅行トピックが存在するとみなせるといえる. そして, Φ_kは, パラメータベクトルβで規定されるDirichlet分布から生成されると仮定する.

$$\boldsymbol{\Phi}_k\sim \text{Dirichlet}(\boldsymbol{\beta}) \tag{2}$$

ここで, $\boldsymbol{\beta}=(\beta_1,\ \cdots,\ \beta_V)$, $\beta>0$.

なお, Φは$K\times V$次元, βはV次元であり, Dirichlet分布は, 多項分布の共役事前分布 (conjugate prior) である. 実際の訪問比率を用いず, このような過程を踏まえて訪問比率を導出する理由は, 過学習を避けるために事前分布としてDirichlet分布を設定すること, データ数に対してパラメータが多い場合や比率が小さいセルが多い場合に偏った結果が導かれる危惧があるため, 汎化性能を高める方法として, 最大事後確率 (maximum a posteori, MAP) 推定を用いてベイズ更新を行うことによる.

2.3.2 旅行別トピック分布の設定

1訪日旅行tにトピック確率分布θ_tが存在すると仮定する. (旅行総数:T)

$$\boldsymbol{\theta}_t=(\theta_{t1},\cdots,\theta_{tK}) \tag{3}$$

ここで, $\theta_{tk}=p(k|\boldsymbol{\theta}_t)$, $\theta_{tk}\geqq 0$, $\Sigma_k\theta_{tk}=1$.

2.3.1と同様に, θ_tは, パラメータベクトルαで規定されるDirichlet分布から生成されると仮定する.

$$\boldsymbol{\theta}_t\sim\text{Dirichlet}(\boldsymbol{\alpha}) \tag{4}$$

ここで, $\boldsymbol{\alpha}=(\alpha_1,\ \cdots,\ \alpha_T)$, $\alpha>0$.

なお, θは$T\times K$次元, αはT次元である.

2.3.3 データの生成過程

各旅行がどの潜在トピックによって生成されたかを示す離散型潜在変数を定義する. 具体的には, 旅行tのi番目訪問地をw_{ti}, 旅行tにおける訪問地総数N_tのもとで, 各訪問地がいずれのトピックに属するかを示す離散型潜在変数z_{ti}を定義する. z_{ti}は, i番目訪問地がトピックkに含まれるとすると, $z_{ti}=k$となるものである ($z_{ti}\in\{1,\ \cdots,\ K\}$). この$z_{ti}$は, Dirichlet分布から導かれるパラメータθ_t ($\theta_t\sim$ Dirichlet (α)) に従う多項分布からランダムで生成されると仮定する.

$$z_{ti}\sim\text{Multi}(\boldsymbol{\theta}_t) \tag{5}$$

次に, 割り当てられたトピックz_{ti}, ならびに訪問地別訪問比率Φ_{zti}に従ってi番目訪問地 (w_{ti}) が生成される.

$$w_{ti}\sim\text{Multi}(\boldsymbol{\Phi}_{zti}),\ \ i=1,\cdots,N_t \tag{6}$$

以上から, 旅行tの生起確率は (7) 式のように示すことができる.

$$p(w_t|\boldsymbol{\theta}_t,\boldsymbol{\Phi})=p(\boldsymbol{\theta}_t|\alpha_t)\prod_{i=1}^{Nt}p(z_{ti}=k|\boldsymbol{\theta}_t)\,p(w_{ti}|\boldsymbol{\Phi}_{z_{ti}}) \tag{7}$$

また, 全旅行データの生起確率を (8) 式に示す.

$$p(\boldsymbol{w}|\boldsymbol{\theta},\boldsymbol{\Phi})=\prod_{t=1}^{T}\int p(\boldsymbol{\theta}_t|\boldsymbol{\alpha})\prod_{i=1}^{Nt}p(z_{ti}=k|\boldsymbol{\theta}_t)\,p(w_{ti}|\boldsymbol{\Phi}_{z_{ti}})\,d\alpha \tag{8}$$

以上より, LDAモデルは, 旅行のトピック分布を表すDirichlet分布パラメータ (α) ならびにθ_tと, 訪問地分布を示すDirichlet分布パラメータ (β) ならびにΦ_kによって規定される.

2.3.4 パラメータ推定について

Dirichlet分布パラメータ (α, β) を推定するためには, (8) 式の尤度最大化 (maximum likelihood estimation) が考えられるが, 先に示したような懸念があるため, 汎化性能を高める方法として, 最大事後確率 (maximum a posteori, MAP) 推定を考える.

MAP推定では, データWを観測したあとのパラメータ (α, β) の事後確率が最大となるパラメータを導出するものである. パラメータ (α, β) の事後確率は, ベイズの定理を用いて下式によって示すことができる (添字b, aは, それぞれ事前, 事後を示す).

$$p(\boldsymbol{\alpha}_a, \boldsymbol{\beta}_a | \boldsymbol{W}, \boldsymbol{\alpha}_b, \boldsymbol{\beta}_b) = \frac{p(\boldsymbol{\alpha}_a, \boldsymbol{\beta}_a | \boldsymbol{\alpha}_b, \boldsymbol{\beta}_b) p(\boldsymbol{W} | \boldsymbol{\alpha}_a, \boldsymbol{\beta}_a)}{p(\boldsymbol{W} | \boldsymbol{\alpha}_b, \boldsymbol{\beta}_b)} \quad (9)$$

ここで, $p(\boldsymbol{\alpha}_a, \boldsymbol{\beta}_a | \boldsymbol{\alpha}_b, \boldsymbol{\beta}_b)$：データを観測する前のパラメータの確率を示す事前確率, $p(\boldsymbol{W} | \boldsymbol{\alpha}_a, \boldsymbol{\beta}_a)$：尤度.

そして, $p(\boldsymbol{W} | \boldsymbol{\alpha}_b, \boldsymbol{\beta}_b)$は, 事後のパラメータに関係しないことから, MAP推定量は下記のように算出できる.

$$\begin{aligned} &\operatorname{argmax} p(\boldsymbol{\alpha}_a, \boldsymbol{\beta}_a | \boldsymbol{W}, \boldsymbol{\alpha}_b, \boldsymbol{\beta}_b) \\ &= \operatorname{argmax} \{\log\left(p(\boldsymbol{\alpha}_a, \boldsymbol{\beta}_a | \boldsymbol{\alpha}_b, \boldsymbol{\beta}_b)\right) + \log\left(p(\boldsymbol{W} | \boldsymbol{\alpha}_a, \boldsymbol{\beta}_a)\right)\} \end{aligned} \quad (10)$$

以上から, 汎化性能を高めるには尤度最大化（(10) 式右辺第2項）に加えて, パラメータの事前分布（(10) 式右辺第1項）が必要となる. 例えば, サイコロをふった場合を想定すると, 試行回数が少ない際に特定の目に偏った事象が生起したとしても, その前提として各々1/6の確率が期待されるため, それを補正するのが (10) 式右辺第1項と考えられる. この観点からも実際に生じた確率（尤度最大化）のみならず, Dirichlet分布を用いてパラメータを推定する利点といえる. なお, 過学習を避け, 汎化性能を向上させるために, (10) 式を用いているが, 事前分布の設定は一様分布以外も考えられるため, 今後の検討が必要と考えられる. また, 様々な確率密度分布が考えられるが, 複数訪問地の立寄り確率分布を多項分布で示せること, この多項分布の共役事前分布であることからDirichlet分布とあわせて採用している.

さて, パラメータはDirichlet分布によって規定されているため, (10) 式の推定にあたっては確率密度を考慮する必要がある. そのために積分計算が必要であり, 解析的に解くことができない. そのための解法として変分ベイズ推定があるが, ここでは計算速度は遅いものの, 誤差が小さいといわれているギブスサンプリング（Gibbs sampling）手法を用いた[19].

さて, 導出されたモデルの妥当性評価であるが, 平均分岐数（perplexity, PPL）によって行われる. PPLは下記のように示すことができる.

$$\begin{aligned} \mathrm{PPL} &= p(\boldsymbol{w} | \boldsymbol{\alpha}_a, \boldsymbol{\beta}_a)^{-\frac{1}{V}} = \exp\left(-\frac{1}{V}\log\left(p(\boldsymbol{w} | \boldsymbol{\alpha}_a, \boldsymbol{\beta}_a)\right)\right) \\ &= \exp\left(-\frac{1}{V}\log\left(p(\boldsymbol{w} | \boldsymbol{\theta}_a, \boldsymbol{\Phi}_a)\right)\right) \end{aligned} \quad (11)$$

=exp（一対数尤度/ユニーク訪問地数）

これは, データの出現確率を最大にするパラメータを推定することが最適とする中で, 尤度自体がユニーク訪問地数Vに依存することから, 相加平均ではなく相乗平均によって同時確率のもとでの確率の逆数（分岐数）を示すものといえる. その表す意味であるが, 例えばある旅行の訪問地1つが隠されていたとする. PPL=1/100の場合, 隠された訪問地の選択肢数を訪問地総数から100まで減少させたことを示し, より小さい指標であるほど絞り込みの性能が高いことを示す.

なお, トピック数を多くするとマーケットセグメンテーションの差異を考慮できるためにPPLは小さくなるが, 一定以上増加するとトピックの増加によって尤度自体が大きくならないため, PPLが増加傾向を示す.

一方, パラメータ数が多いことによる識別問題への指摘, 初期値の設定によって収束先が異なるなどの特徴がある. 前者については明確な指標は提案されていないが, 後者に対しては, 複数回推定を行いPPLが小さいものを検索する必要があると考えられる.

3――訪問パターンの分析結果

3.1 ユニークユーザーの概要

まず, 分析対象UUの概要を示す. 表―1は国・地域別UU数を示したものである. 地域区分はナビタイム独自のものであるが, 1~7までが国・地域で, 8以下は国単位の構成割合が5%未満であるため, 秘匿処理があらかじめ行われた地域コードに集約されたものである.

平成27年の訪日外客数上位3ヶ国（実績値）は, 中国：25%, 韓国：20%, 台湾：19%となっているが, 本調査データではUUが少ないため中国, 韓国は東アジアに含まれている. その理由として中国ではGoogleが使用できないため

■表―1 国・地域別UU数並びに平均訪日回数

	国・地域	UU数	構成割合	構成割合*	拡大係数	平均訪日回数
1	台湾	7,972	14%	19%	5.0	4.1
2	タイ	6,699	12%	4%	1.2	3.1
3	アメリカ合衆国	6,496	12%	5%	1.7	2.3
4	香港	3,880	7%	8%	4.4	4.3
5	オーストラリア	3,320	6%	2%	1.1	1.9
6	シンガポール	3,257	6%	2%	1.0	2.7
7	マレーシア	2,882	5%	2%	1.1	1.9
8	北米（3を除く）	1,593	3%	1%	1.9	2.1
9	ロシア	231	0%	0%	2.9	3.2
10	西ヨーロッパ	4,974	9%	5%	2.0	2.2
11	東南アジア	4,456	8%	3%	1.7	2.0
12	東アジア	4,142	7%	46%	23.9	2.5
13	南アジア	484	1%	1%	2.7	2.1
14	北ヨーロッパ	143	0%	0%	5.6	3.3
15	オセアニア	136	0%	0%	5.2	3.0
16	南アメリカ	127	0%	0%	4.1	2.9
17	中央アメリカ	118	0%	0%	4.1	2.6
18	中東	32	0%	0%	7.1	4.0
19	その他	4,527	8%	2%	1.3	2.2
	合計（平均）	55,469	100%	100%	4.1	2.7

※構成割合*：訪日外客数（実績値：JNTO，平成27年）

Android利用者が少なく，団体旅行が多いなどの理由がある．また，韓国はハングルで使用できる他アプリが存在するため，本アプリの利用者が少ないことが考えられる．実際の訪問パターンの分析では，これらの偏りを補正するために，i国・地域別に全データに対するUU構成割合（$P_{i,U}$）を構成比率実績値（訪日外客数データ（JNTO），$P_{i,A}$）で除した値の逆数に，この値の最も大きいシンガポールの値（$P_{SIN,U}/P_{SIN,A}$）を乗じて拡大係数（E_i）を算出した．

$$Ei = \frac{P_{SIN,U}/P_{SIN,A}}{P_{i,U}/P_{i,A}} \quad (12)$$

この拡大係数をUUに乗じてデータを拡大処理した．

表—2は，滞在日数分布を示したものである（本データでは，アプリ利用開始からの通算日であるため厳密には滞在日数ではないが，便宜上滞在日数と表記する）．サンプリング調査である参照調査結果と比較すると，1〜3日の短期滞在者が約半数を占め，開始時点が来日後と予想されること，スマホの電源状態に依存することが原因として考えられる．訪問パターン分析の目的からは，訪問地点が少なくなったり，アプリ開始までの地点の欠落に留意する必要がある．また，観光やビジネスなど訪日目的による影響（訪日外国人消費動向調査で平均泊数全目的：10.2泊，観光・レジャー目的：5.9泊）が考えられるが，データにこの区分がないため，全データを用いた．

次に，緯度経度データをみると，位置情報は2分ごと，かつ200m以上移動した場合に測位データとして蓄積され，さらに状態別に3区分される（stay，break，other）．表—3より，“stay”が約73%占め，時間帯別平均レコード数では，昼間に移動と思われる“other”が多くみられた（図—1）．なお，訪問パターン分析には，訪問を同定するために“stay”データを用いた．その際，訪問の有無に加え，その順序自体の考慮も考えられるが，本研究では訪問順序は捨象する．すなわち，パラメータの導出にあたっては，訪問順序による影響はなく，訪問地の組み合わせのみが影響をあたえることとなる．これは，テキストマイニングの分野において，単語の順序を無視し，文書を単語の集合として捉えるbag-of-wordsの考え方と同一である．なお，28日以上は構成比率が10%を下回ること，“stay”データ数がそれ未満の約14.4倍になって行動がより広範・長期化すると考えられるため，27日以下のUU（54,055UU）に分析対象を限定した（全UU平均滞在日数：6.0日，27日以下滞在者平均滞在日数：4.5日）．

3.2 分析データの作成

使用する緯度経度データは世界測地系1984（WGS84）で記録されており，訪問地を同定するために基準地域メッシュ（概ね1Km四方）に変換した．これは，データの抽出基準に基準地域メッシュが用いられていたためである．この場合，日本全体の基準地域メッシュ総数は461,497となる．

さらに，データクリーニングを下記に則り行った．

1) GPS位置情報が欠損しているレコード，

2) stay状態以外のレコード，

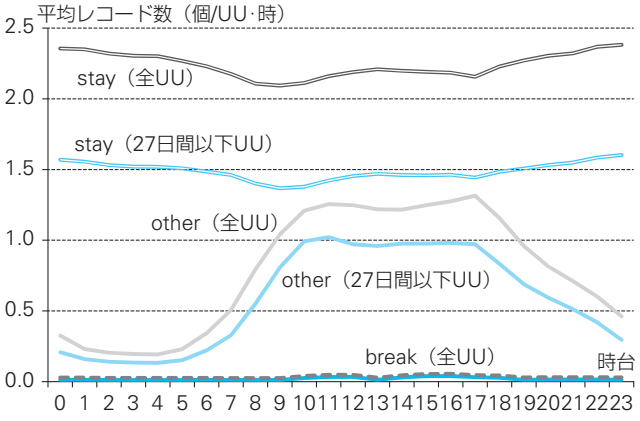
■表—2　滞在日数分布

	分析データ			参照調査結果*		
日数	UU数	構成比率	構成比率	全目的	業務	観光
1日	11,020	20%	51%	12%	15%	11%
2日	9,931	18%				
3日	7,400	13%				
4日	6,227	11%	26%	48%	45%	49%
5日	4,866	9%				
6日	3,410	6%				
7日	2,617	5%	16%	24%	23%	25%
8日	2,022	4%				
9日	1,389	3%				
10日	1,056	2%				
11日	767	1%				
12日	664	1%				
13日	468	1%				
14−20日	1,671	3%	3%	6%	6%	6%
21−27日	547	1%	1%	2%	2%	2%
28−90日	1,177	2%	2%	5%	7%	5%
91日−	237	0%	0%	2%	2%	2%
合計	55,469	100%	100%	100%	100%	100%

※参照調査結果*：訪日外国人消費動向調査（平成27年）

■表—3　状態別レコード数

	Stay	Break	Other	合計
全UU レコード数	29,684,874	466,529	10,355,788	40,507,191
構成比率	73%	1%	26%	100%
平均レコード数	538	8	188	734
滞在日数：〜27日 レコード数	19,226,277	256,953	7,554,163	27,037,393
構成比率	71%	1%	28%	100%
平均レコード数	357	5	140	502



■図—1　時間帯別状態別平均レコード数

3) 成田空港, 羽田空港, 中部国際空港, 関西国際空港など空港の訪問レコード,

4) 少数訪問者数のメッシュ訪問のレコード,

5) 28日以上滞在者のUU.

3) は, 入国・出国場所が大きく影響してしまうための措置である. 特に, アプリ使用開始時からデータ取得がなされるため, 個々人でのばらつきが考えられため, 分析から除外した. 一方, 各トピックでどの空港 (港) を経由しているのかという情報も, 広域的観光ルートを形成する上で貴重な情報となりうるため, これらについては今後の課題としたい. 次に, 4) は, 特異なパターンの抽出を除外するためで, 4) は具体的には54,055UUから推定されるメッシュ訪問率の標準誤差率 (訪問率標準偏差/訪問率) が33.3%を上回らない, という基準を設けた. なお, 訪問率の標準偏差は2項分布に従うと仮定している. これより, 訪問UU数が8以下であるメッシュを除外した. 以上, 3), 4) により10,901メッシュを除外し, 訪問がみとめられたメッシュは1,834メッシュとなった. それに対して, 最終的に分析対象は33,279UUとなった.

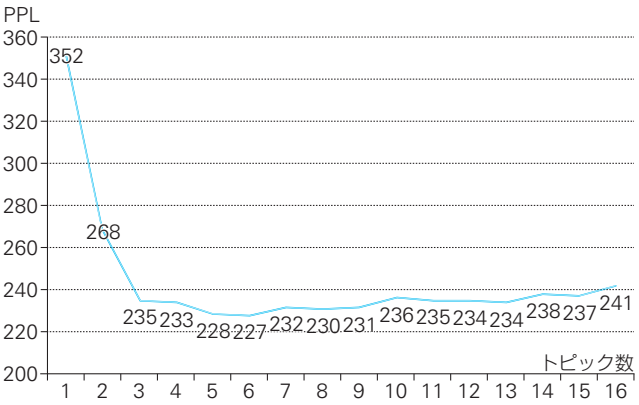
このUUは訪日外国人旅行者の国籍・地域構成割合実績値から偏りがあるため, 表—1に示す拡大係数によって拡大処理を行った. トピックモデルでは, 個体として認識させる必要があるため, 国籍・地域ごとに拡大係数を平均値とした乱数によってUUをクローニングした結果, 124,872サンプルとなった.

以上の結果, 47都道府県毎の訪問率上位5位は, 東京都:68%, 大阪府:24%, 京都府:22%, 千葉県:20%, 神奈川県:19%となり, 訪日外国人消費動向調査とある程度整合していることが確認できた. また, 都道府県単位の訪問パターン数は3,124に上り, 上位10位は, 1) 東京都のみ:22.9%, 2) 東京都, 千葉県:9.2%, 3) 東京都, 神奈川県:7.8%, 4) 北海道のみ:4.8%, 5) 大阪府のみ:4.0%, 6) 大阪府, 京都府:3.6%, 7) 京都府のみ:2.2%, 8) 東京都, 神奈川県, 千葉県:2.1%, 9) 東京都, 京都府:1.7%, 10) 東京都, 大阪府:1.5% (累積構成割合:59.8%) となった.

3.3 トピックモデル分析結果ならびに評価

3.3.1 トピックモデル分析結果

上記データを用いて, トピックモデルによる分析を行った. まず, トピックス数を1から16まで設定して各々で3回推定を行い, PPLの平均値を算出した (図—2). その結果, トピックスの増加にともない減少し, 6トピック時に最小となった (227). これは分析対象メッシュ1,834が存在する中で, 隠された訪問地の選択肢を227まで絞り込めたことを示す. しかしながら, トピック数が7を超えるに従って逓増し, モデルの説明力は低下していると考えられる. 先行研究ではPPLの妥当な目安や個々のパラメータ検定は提案されていないことから, 本分析では6トピックが適当と考え, その中で最もPPLの小さい推定結果について引き続き考察を行う.



■図—2　トピック数別PPL

■表—4　トピック別構成割合・上位訪問地

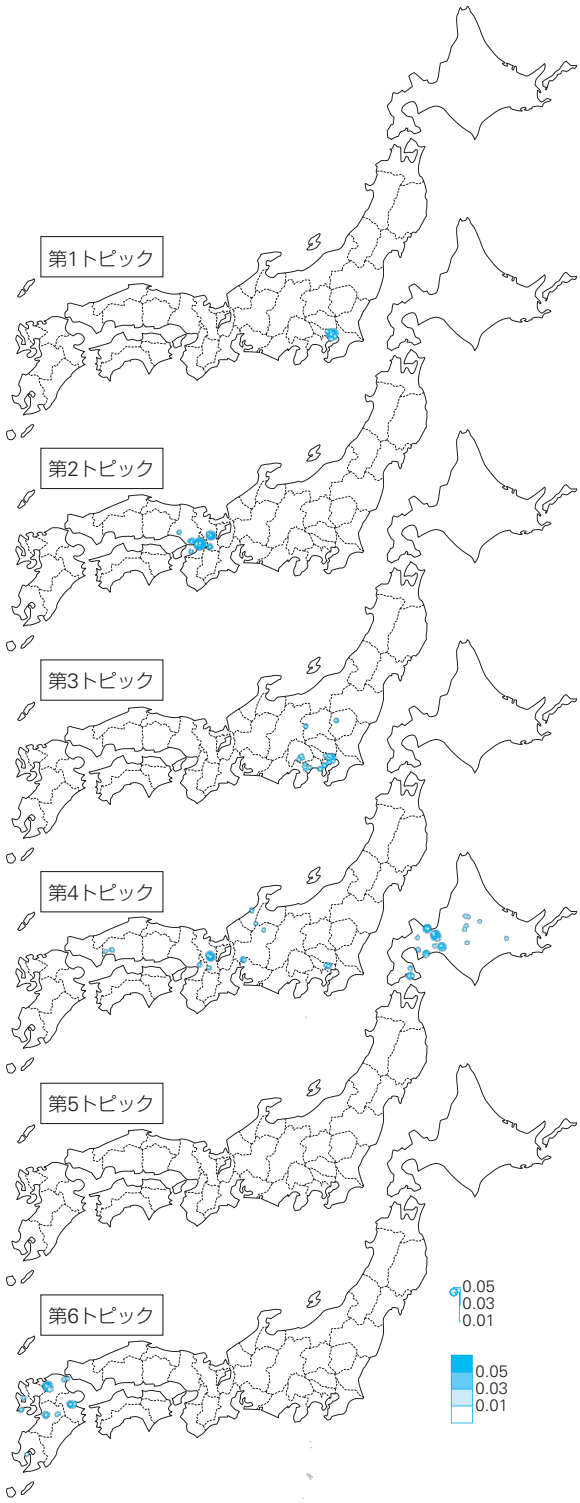
	第1トピック	第2トピック	第3トピック	第4トピック	第5トピック	第6トピック
名称	東京のみ	京阪	東京, 箱根, 富士他	ゴールデンルート, 昇龍道	北海道, 東北	九州, 沖縄
構成比率	47%	18%	14%	11%	6%	4%
第1訪問率	東京・浅草駅	大阪・道頓堀	東京・浅草駅	京都・京都駅	北海道・すすきの	福岡・博多駅
第2訪問率	東京・原宿駅	大阪・梅田駅	*東京・都庁*	京都・河原町	北海道・小樽運河	福岡・中洲南
第3訪問率	東京・新宿駅	大阪・難波駅	*東京・新宿駅*	京都・清水寺	北海道・時計台	福岡・天神
第4訪問率	東京・銀座	京都・清水寺	*東京・歌舞伎町*	京都・金閣寺	北海道・札幌駅	沖縄・国際通り
第5訪問率	東京・歌舞伎町	京都・京都駅	*神奈川・みなとみらい*	*京都・八坂神社*	北海道・登別温泉	福岡・中洲北

表—4は, 各々のトピックの構成割合, 訪問比率上位5箇所を示したものである. 構成割合は, 旅行別トピック確率分布:θ_tを規定するDirichlet分布パラメータ (α) の「サンプル全体の和」に対する「サンプル全体の各トピックの総和」の比率によって算出される. なお, 訪問比率はDirichlet分布パラメータ (β) から導かれるものである. 訪問比率をゴシックで示した地点は5%以上, 斜体は3%未満を示している. 多項分布に基づき尤度が設定されているため, 先行研究でも訪問率 (選択比率もしくは出現率) は10%以下の事例が多く, 表—4の値も必ずしも高くない. なお, 個々の地点は基準地域メッシュの中で主要な駅・施設を例示したものである.

第1トピックは, 浅草, 新宿, 銀座など東京を代表するメッシュを訪問しているのに対して他の道府県の訪問比率が小さいことから「東京のみ」トピックと考えられ, その構成割合は47%と非常に高い. 同様に一地方内でのトピックとしては, 第2トピック (大阪, 京都を重点的に訪問する京阪トピック, 18%), 第3トピック (東京, 箱根, 富士, 軽井沢,

日光などを訪問する関東近隣トピック，14%），第5トピック（北海道の自然，文化を重点的に体験，6%），第6トピック：九州・沖縄トピック（4%）が抽出された．

それに対して，複数地域を横断するものとして，第4トピック（東京，名古屋，昇龍道，京阪奈，広島などを訪問するゴールデンルート（GR）トピック，11%）が抽出できた．なお，東北，四国地方を訪問するトピックを明確に抽出できなかったが，試みに15トピックに設定すると抽出できたため，トピック数設定には検討の余地が残る．



■図―3　トピック別メッシュ訪問率（スペースの制約により一部地域を除外）

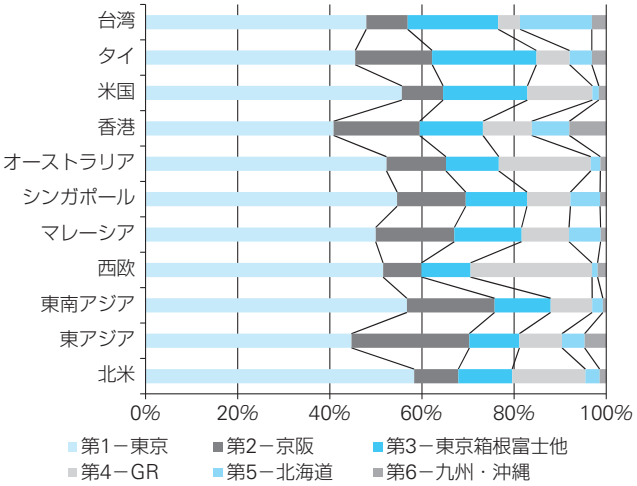
また，図―3はトピック別メッシュ訪問率を示しており，灰色で示された最大円が訪問率5%以上である．なお，個々のトピックの妥当性検証は，外的基準として予め存在していないため，その解釈を通じて判断する必要がある．

3.3.2　トピックモデルの評価

トピックモデルによって，多数の訪問地組み合わせから導出された訪問パターンと個人属性との関連性に着目する．図―4は，1旅行での最大帰属トピックで集計した国籍・地域別トピック割合である．特徴とその理由（括弧内）として，台湾：第5トピック多（雪を目的とした旅行形態等），米国：第4トピック多（GRへの旅行志向等），香港：6トピック多（クルーズ利用が多く，沖縄が寄港地になっている等）などが考えられる．なお，中国が含まれる東アジアで第4トピック（GRに立ち寄る傾向：高）が多くないが，これはアプリの利用ニーズが高い個人旅行では必ずしもGRを選好していないことが考えられる．なお，（　）内に示した理由以外にも多様なものが考えられる．旅行者がどのような理由でその地を訪れたのか，どうしてそのような訪問ルートを辿ったのか，来訪動機との関連性については，さらに検討の余地があると考えられる．

次に，訪日外国人旅行者数の多い台湾，東アジアならびに観光庁で強化するターゲット層として設定している米国を取り上げ，来日回数別トピック構成割合を示す（図―5～7）．いずれも，初回来日のトピック別構成割合が来訪を重ねても大きく変化していないものの，台湾では2―3回，4―9回で第2トピックが増える一方，10回以上では初回来訪とほぼ同一の構成比率になっている．東アジアも比較的台湾に似ているが，米国では10回以上で第2，第3，第4が微増しており，多様な地域への訪問ニーズを反映していることが考えられる．

以上から，国籍・地域別構成比率の差異は来日回数よりも大きく，訪問パターンを規定していると考えられること，



■図―4　国籍・地域別トピック構成割合

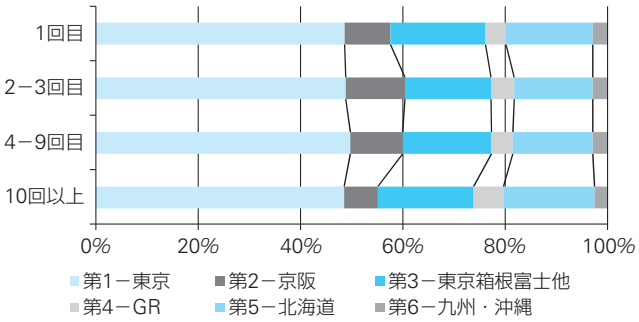
リピーター醸成に向けては，各トピックに対応した多層的，多面的な魅力創出を検討する必要があること等が推察される．

図―8は，月別トピック別サンプル数を示している．これより，第1トピック（東京）は通年で高いものの，第2トピック（京阪）は下半期のほうが構成比率が高いこと，第5トピック（北海道）は1―2月や7―9月で，第4トピック（GR）は3―4月，9月で構成比率が増加していることがわかり，観光資源と強く関連しているといえる．なお，2015年ではアプリ利用者が10月以降で特に多く，その中で第1，第2トピック数自体が増加していた．

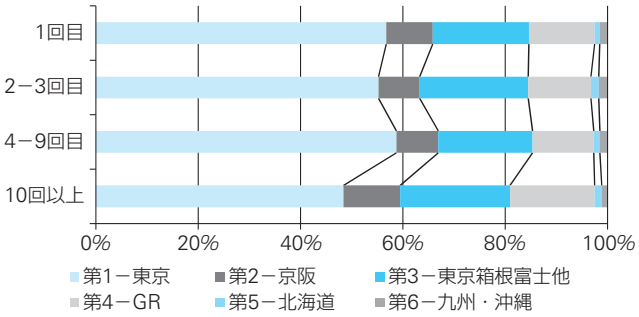
また，トピックモデルでは，1訪日旅行が複数トピックを有すると仮定している．この確率はDirichlet分布パラメータαから算出することができるため，構成割合が50%を超えるものを優位トピックとして設定し，個々人のトピック構成割合を示す．優位トピックの構成比率を折れ線グラフ（右目盛），その他のトピックを積み上げ棒グラフとして示した．なお，横軸は優位なトピック構成割合を降順に並べ替えしている．構成比率が10%を超えるトピックの中から，複合の傾向が大きく異なると考えられる2つを比較すると，図―9（第1トピック）は，当該トピックの占める割合が高く，複合する他トピックの割合は比較的小さいのに対して，図―10（第4トピック）は他トピックが複合している割合が高く，特に第1トピックとの複合傾向が多いと判断できる．

以上の分析は，旅行者に着眼したものであったが，訪問地からの視点として，どのようなトピックをもった訪問者が存在するのか，集計することも可能である．図―11は，東京，京阪の主要訪問地（基準地域メッシュにおける代表的地点）を取り上げ，その各訪問者で最も構成比率の高いトピックを集計したものである．

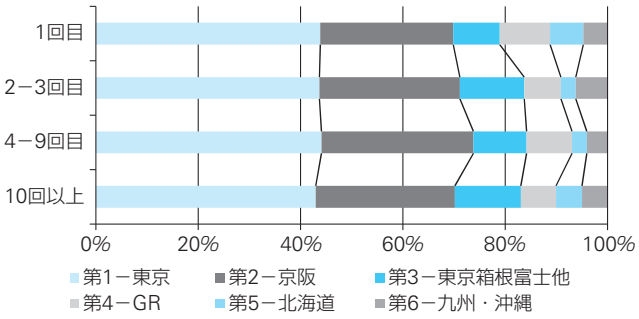
例えば，東京都・浅草駅では，約8割の訪問者が第1トピック（東京のみ）を保有しており，2割弱が第3トピック，5%前後が第4トピックとなっている．原宿駅から都庁まで，



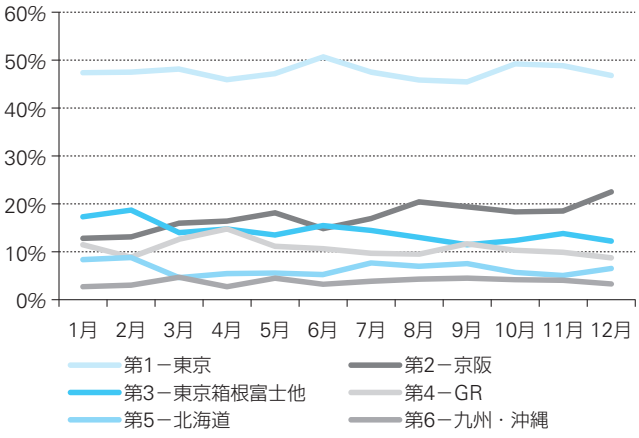
■図―5　来日回数別トピック構成割合（台湾）



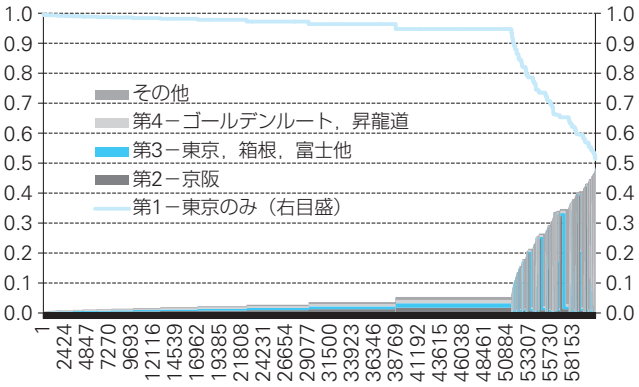
■図―6　来日回数別トピック構成割合（米国）



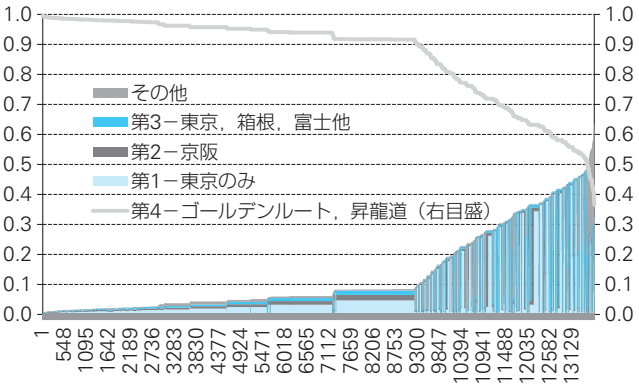
■図―7　来日回数別トピック構成割合（東アジア）



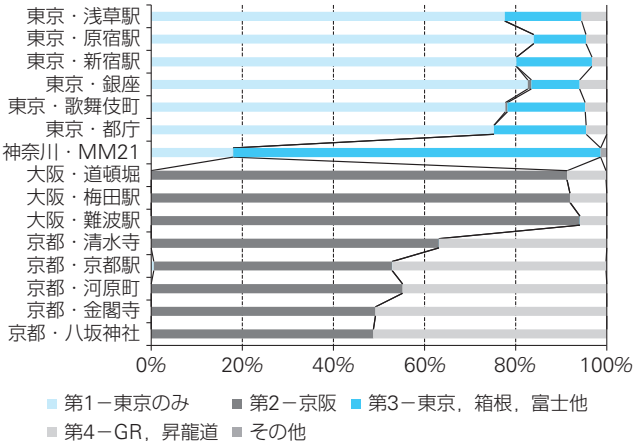
■図―8　月別トピック構成割合（全サンプル）



■図―9　第1トピック優位者のトピック構成割合（60,568サンプル）



■図―10　第4トピック優位者のトピック構成割合（13,669サンプル）



■図―11　メッシュ別トピック構成割合

ほぼ同様の構成比率であることから，これらの地域の観光への取り組みとして，東京の魅力とともに箱根，富士山などとあわせた東日本の主要観光地を一体的にしたイメージ醸成等が重要といえる．それに対して，MM21地区（神奈川）では第3トピックが優勢となることから，MM21自体の魅力向上に加えて，東京，箱根，富士山との連関を考慮したテーマ性，周遊ルートを訴求する必要性が高いと考えられる．

一方，大阪についてみると第2トピック以外が少ないため，大阪自体の魅力の向上が重要であるのに対して，京都では第4トピックが多いため，GR周遊の利便性向上が不可欠であると考えられる．着地における調査では，前後の訪問地の聞き取りケースがあるが，訪日旅行者は多くの地点に立ち寄るため，本論文で行った訪問パターン分析により包括的把握が可能となり，旅行動機・ニーズを踏まえることができると考えられる．

4――まとめ

本研究は，インバウンド観光振興において需要，供給両面の定量的ポジショニング，セグメントをおさえた戦略策定への情報獲得を念頭に置きながら，訪日外国人旅行者の訪問パターンの分析手法を開発することを目的として，トピックモデルを用いた分析を行った．GPSデータによる位置情報から旅行者ならびに目的地をいくつかに区分できることが本分析の特徴であり，トピックモデルには個人が有するトピック割合を訪問の有無（実績）のみで推定するのではなく，Dirichlet分布を前提とした生成モデルで過学習を抑制できること，個人個人が異なる複数トピックを有することを考慮できることが利点としてあげられる．

国籍・地域によるサンプルの偏りならびに少数サンプル訪問地の除去を行った後，トピックモデルで分析したところ，PPLより6トピックに区分することが適当と考えられた．その結果から，旅行者の訪問パターンには，主として大都市を中心に訪問するトピックと比較的広域を周遊するパターンの2つが抽出できた（図―3）．また，国籍・地域（図―4），来日回数（図―5～7），訪問時期（図―8）によってトピック構成率が異なること，個人個人が有する複数トピックの割合推定（図―9～10）など訪日旅行者側（需要側）の特徴を把握でき，これらは行動の把握だけでなくターゲット別のルート設定，プロモーションの検討に際して参考になると考えられる．さらに，図―11に示すように訪問地においてどのようなトピックが多いか把握することもでき，地域の取り組みに対する参考情報の1つとして位置づけできる．　このように類型化が困難であった数多くの組み合わせを適当に区分できたことにより，需要，供給の特徴，整合性，関連性把握が容易になったと考えられる．

今後の課題として，推定方法ではPPL以外の評価指標を検討し，より細かな訪問パターンを導出する方法や，類似したトピックの取り扱い方法，トピック数の設定方法に関する検討があげられる．また，使用データについては，訪日外国人旅行者の大きなシェアを有する中国，韓国が集約されてしまうことから，旅行者全体の特性把握が困難な面があると考えられる．他データの有効活用を含めた検討を行う必要性がある．以上を解決しながら，地域のインバウンド振興に対して実際的な情報提供，提言に直結するアウトプットを行うことが重要といえる．

参考文献

1) 国土交通省［2016］,“大臣会見要旨”,（オンライン）, http://www.mlit.go.jp/report/interview/daijin170110.html, 2017/1/14.
2) 首相官邸［2016］,“第2回明日の日本を支える観光ビジョン構想会議”,（オンライン）, http://www.kantei.go.jp/jp/singi/kanko_vision/dai2/gijisidai.html, 2017/1/14.
3) 観光庁［2016］,“観光ビジョン実現プログラム2016の策定”,（オンライン）, http://www.mlit.go.jp/kankocho/topics01_000208.html, 2017/1/14.
4) 古屋秀樹［2016］,“群馬県館林市における観光振興への取り組み”,「東洋大学地域活性化研究所報」, No.13, pp.33-40.
5) 観光庁［2015］,“日本版DMOとは”,（オンライン）, http://www.mlit.go.jp/kankocho/page04_000048.html, 2017.1.14.
6) ナビタイムジャパン［2015］,“ビッグデータを用いた訪日外国人の観光分析～発見！意外なホットスポット～”,（オンライン）, http://consulting.navitime.biz/pdf/monograph_20150619_01.pdf, 2017/1/14.
7) 観光庁［2015］,“観光ビッグデータを活用した観光振興/GPSを利用した観光行動の調査分析”,（オンライン）, http://www.mlit.go.jp/kankocho/shisaku/kankochi/gps.html, 2017/1/14.
8) 古屋秀樹［2015］,“旅行要因の関連性を考慮した宿泊観光旅行のクラス構築に関する基礎的分析”,「都市計画論文集」, Vol. 50, No. 3, pp. 337-344.
9) 森地茂・兵藤哲朗・岡本直久［1992］,“時間軸を考慮した観光周遊行動に関する研究”,「土木計画学研究論文集」, No.10, pp.63-70.
10) 森川高行・佐々木邦明・東力也［1995］,“観光系道路網整備のための休日周遊行動モデル分析”,「土木計画学研究・論文集」, No.12, pp.539-547.
11) 劉瑜娟・古屋秀樹［2015］,“潜在クラス分析を用いた訪日外客の訪問パターンに関する基礎的分析”,「第52回土木計画学研究発表会講演集」, No.52, CD-ROM.
12) 古屋秀樹・劉瑜娟［2015］,“訪日外客の47都道府県の訪問パターン分析”,「日本観光研究学会第30会全国大会研究発表論文集」, CD-ROM.
13) 古屋秀樹・劉瑜娟［2016］,“潜在クラス分析を用いた訪日外国人旅行者の訪問パターン分析”,（オンライン）,「土木学会論文集D3（土木計画学）」,

Vol.72, No.5, pp.I_571-I_583.
14) 佐藤一誠 [2015],『トピックモデルによる統計的潜在意味解析』, コロナ社.
15) 前掲14
16) 岩田具治 [2015],『トピックモデル』, 講談社.
17) Graham Neubig [2013], "NLP Programming Tutorial 7 -トピックモデル", (オンライン), 奈良先端科学技術大学院大学 HP, http://www.phontron.com/slides/nlp-programming-ja-07-topic.pdf, 2017/1/14.
18) 塚井誠人･椎野創介 [2016], "討議録に対するトピックモデルの適用", (オンライン),「土木学会論文集D3 (土木計画学) 」, Vol.72, No.5, pp.I_341-I_352.
19) 伊庭幸人･種村正美他 [2005],『計算統計II－マルコフ連鎖モンテカルロ法とその周辺－』, 岩波書店.

(原稿受付2017年1月20日)

Development of the Combination Analysis Method of Visited Places of Foreign Visitors to Japan by GPS Log Data

By Hideki FURUYA, Naohisa OKAMOTO and Naoki NOZU

The purpose of this paper was to develop the combination analysis method of visited places by foreign visitors to Japan. For identification of combination patterns, the topic model was applied using GPS tracking data of smartphone applications of foreign tourists. This model was supposed as a method of identifying segments of destinations and visitors simultaneously. The number of unique users collected in 2015 was 55,199. We had data cleaning for bias caused due to sampling ratio and the noise of GPS positioning data. In the estimation, it was found appropriate to set the number of topics as six. The visited rates and visitors' topic distribution of each zone were also estimated by this model.

Key Words : ***GPS Log Data, foreign tourist, combination of visited places, Topic Model***