トピックモデルによる訪日外国人旅行者の訪問パターンの基礎分析

古屋 秀樹[1]

[1]正会員　東洋大学国際地域学部国際観光学科教授（〒112-8606 東京都文京区白山）
E-mail:furuya@toyo.jp

本研究は，訪日外国人旅行者の訪問パターンの抽出を都道府県区分よりも細かい地区レベルで行うことを目的としている．論文では，はじめに訪問地の組合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，観光庁が実施した「訪日外国人消費動向調査（平成26年）」データを用いながら，トピックモデルによって訪問場所の組合せパターンを幾つかのクラスに分類した．

Key Words: *Combination of visiting place, Latent Dirichlet Allocation Model*

１．はじめに

2015年の訪日外国人旅行者数は1973.7万人を数え，過去最高を記録した．2020年の東京五輪開催などをひかえ，インバウンド観光の一層の促進にむけた様々な取り組みがなされている．訪日外国人旅行者数増加は，外国への情報発信や大きな経済効果発現につながると考えられる．今後は，訪日外国人旅行者数といった量の視点からだけではなく，初回訪日旅行者の満足度を高めてリピーターへの変容をはかる取り組みも重要といえる．

そのために，需要側である旅行者の観光行動特性を把握し，訪日回数や国籍別，季節別にどのような違いがあるのか把握するとともに，時系列で比較しながら訪問パターンをモニタリングしていく必要性が考えられる．このように訪日旅行者を幾つかのセグメントに分割を行い，それぞれの特徴把握によって，より詳細に誘客のための効果的施策の検討，実施が可能と考えられる．

そこで，本研究は観光・レジャーを旅行目的とした訪日外国人旅行者の訪問パターン把握を目的とする．この訪問パターンとは，日本における「訪問地の組み合わせ」と設定し，分析では観光庁が実施した「訪日外国人消費動向調査」データを用いている．サンプル数が約2万人を数えるため，訪問パターンも多様な中で，それらを効率的かつ論理的に集約し，セグメントを導出しなければならない．

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ものとする．

２．推定手法と本研究の位置づけ

(1)　訪日外国人消費動向調査[1)]

各種アンケート調査をはじめとして訪日外国人の訪問パターンを調査する方法があるが，より多くのサンプル収集が可能なものとして，訪日外国人消費動向調査（観光庁）ならびに携帯電話の位置情報データの利用が考えられる．後者は，訪日外国人旅行者の増加によってより多くのサンプルを獲得できるとともに，あらためて調査自体を行う必要がないなどのメリットがある一方，移動自体を取り扱っているためにその移動目的や同行者，訪日滞在中の消費額や満足度などを把握できないと考えられる．それに対して，前者は聞き取り調査を行う手間がかかるものの，被験者の振り返りを通じた訪問地点，満足度などを明確に把握することができるため，その信ぴょう性が高いと考えることができる．

そこで，本研究では，観光庁が実施した「訪日外国人消費動向調査（平成26年）」の個票データを用いて分析を行った．本調査は，11空海港の国際線ターミナル搭乗待合ロビーで出国を待つ訪日外国人旅行者に対して，外国語対応のタッチパネル式PCまたは紙調査票を用い，外国語を話せる調査員によって個人属性ならびに訪日旅行に関する訪問地，同行者，旅行支出，旅行情報源，満

足度と再訪意向の聞き取りを実施したものである．

本研究では，その中から，「今回の日本滞在中における訪問地」回答データを利用しているが，これは宿泊，日帰りの区分はされていない点に留意する必要がある．さらに，個人属性として，国籍・地域，性別，年齢，訪日回数，訪日時期を活用した．

⑵ 潜在クラス分析の概要と問題点[2)]

インバウンド観光の振興では，訪日外国人旅行者の行動特性を把握する必要があるとともに，国籍・地域や年齢，旅行への嗜好が異なることから，複数のセグメントに分割することが現状の把握，今後のマーケティング活動の検討に有効であると考えられる．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．

さて，訪日外国人旅行者のセグメンテーションをどのような要因に基づき行うかが重要であるが，本論文では，日本国内における訪問地およびその組合せ（訪問パターン）が，日本に対する観光動機，観光需要を的確に反映していると仮定した．この訪問パターンに着眼しながら，潜在クラス分析によって地方区分単位で分析した文献3や都道府県単位に分析した文献4，5がある．

この潜在クラス分析を簡単に示すと，旅行者の訪問地の組み合わせを考えるとき，訪問パターンを規定するクラスがX種類存在し，クラス tの構成比率をπ_t^X とする．また，ある個人のある旅行nのゾーンkの立ち寄りの有無を$\delta_{k1,n}$，$\delta_{k2,n}$（訪問有り：$\delta_{k1,n}$ =1，$\delta_{k2,n}$=0，訪問無し：$\delta_{k1,n}$=0，$\delta_{k2,n}$=1）で示す．さらに，旅行nがクラスtに属すると仮定した場合に，ゾーンkの訪問率を$\pi_{k,1,t}^X$，非訪問率を$\pi_{k,2,t}^X$とすると，全ゾーンKの訪問の有無の組み合わせを示す同時確率$P_{n,t}$ は下記のように示すことができる．

$$P_{n,t} = \pi_t^X \cdot \prod_{k=1}^{K} \pi_{k,1,t}^{X\ \delta_{k,1,n}} \cdot \pi_{k,2,t}^{X\ \delta_{k,2,n}} \quad \text{…(1)式}$$

ここで，$0 \le \pi_t^X$，$\sum_{t=1}^{X} \pi_t^X = 1$，

$$0 \le \pi_{k,1,t}^X, \pi_{k,2,t}^X,\ \pi_{k,1,t}^X + \pi_{k,2,t}^X = 1$$

なお，π_t^X は多項分布で，$\pi_{k,1,t}^X$と$\pi_{k,2,t}^X$は2項分布に従うとする．さて，(1)式で示すように，所属クラスtのもとでは，クラスtの構成比率にクラスtに固有な訪問率・非訪問率を乗じることによって同時確率$P_{n,t}$が求められるとする．これより，多数の組み合わせがある訪問パターンを，(1)式で定義した尤度に基づき少数パターンに集約するのが潜在クラス分析といえる．

さて，潜在クラス分析では，外国人旅行者1名の旅行が1つのクラスに属することを仮定している．しかしながら，1旅行の中には，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」が混在していることも考えられ，単独のトピックならびにすべての被験者が同一の分布を有することを仮定する潜在クラス分析は制約の強いモデルであると考えられる．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．これらの問題点[6)]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．

⑶ トピックモデルについて[2),7),8),9),10)]

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複数のトピックが存在し，そのトピックごとに訪問地への訪問確率(分布)が異なると仮定する．なお，各旅行にはトピックの情報が先験的に与えられておらず，観測できていない潜在トピックとして抽出できるようにモデル化を行う．トピックモデルには，各旅行のトピック比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．

1) 訪問地別訪問率分布の設定

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率分布を規定するパラメータ $\boldsymbol{\Phi}_k$ が存在する（訪問地総数：V）．

$$\boldsymbol{\Phi}_k = (\varphi_{k1},\ \cdots,\ \varphi_{kv}) \quad \cdots\text{(2)式}$$

ここで，$\varphi_{kv} = p(v|\boldsymbol{\Phi}_k)$

（トピックkにおける訪問地vの訪問比率を規定するパラメータ），

$\varphi_{kv} \ge 0$，$\Sigma_v \varphi_{kv} = 1$.

上記から，同一の訪問地でも異なるトピック複数に出現することを推察でき，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，確率ベクトルである $\boldsymbol{\Phi}_k$ は，確率ベクトル上の確率分布であるディリクレ分布から生成されると仮定する($\boldsymbol{\Phi}_k$~Diriclet($\boldsymbol{\beta}$)，$\boldsymbol{\beta}=(\beta_1,\ \cdots,\ \beta_V)$，$\beta_k>0$).

なお，ディリクレ分布は，その総和が 1.0 となるものであり，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る理由は，過学習を避けるために，共役事前分布に尤度を乗じるベイズ更新を行うためである．

2) 旅行別トピック分布の設定

1つの旅行tにトピック確率分布が存在し，それを規定するパラメータ$\boldsymbol{\theta}_t$が存在すると仮定する．(旅行総数：T)

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

ここで，$\theta_{tk} = p(k \mid \boldsymbol{\theta}_t)$（旅行tにトピックkが割り当てられる確率を規定するパラメータ），

$\theta_{tk} \geqq 0,\ \Sigma_k \theta_{tk} = 1.$

1)と同様に，確率ベクトルである $\boldsymbol{\theta}_t$は，確率ベクトル上の確率分布であるディリクレ分布から生成されると仮定する($\boldsymbol{\theta}_t \sim$Diriclet($\boldsymbol{\alpha}$)，$\boldsymbol{\alpha}=(\alpha_1, \cdots, \boldsymbol{\alpha}_K)$，$\alpha_k>0$).

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潜在変数を導入する．具体的には，旅行 t の訪問地 i を w_{ti}，旅行 t における訪問地総数 N_tのもとで，各旅行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_t を定義する．z_{ti}は，例えば旅行 t 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（$z_{ti} \in \{1, \cdots, K\}$）．この時，旅行 t において，

a) Diriclet 分布のパラメータ（$\boldsymbol{\theta}_t \sim$Diriclet($\boldsymbol{\alpha}$)，$\boldsymbol{\alpha}=(\alpha_1, \cdots, \boldsymbol{\alpha}_K)$）に従って，旅行 t のトピック z_{ti} が生成される（$z_{ti} \sim$Multi($\boldsymbol{\theta}_t$)，i=1,$\cdots$，N_t）．

上に示すようにトピックが割り当てられる確率は，Diriclet 分布が共役事前分布である多項分布と仮定している．

c) 一方，割り当てられたトピック z_{ti}，ならびに訪問地分布パラメータ $\boldsymbol{\Phi}_{zti}$ に従って訪問地 i が生成される（w_{ti}）．

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

d) 旅行 t の生起確率は下記のように示すことができる．

$$\mathrm{P}(w_t \mid \boldsymbol{\theta}_t, \boldsymbol{\Phi}) = \prod_{i=1}^{Nt} \sum_{k=1}^{K} p(z_{ti}=k \mid \boldsymbol{\theta}_t)\, p(w_{ti} \mid \boldsymbol{\Phi}_k) \quad \cdots(4)式$$

また，全旅行データの生起確率は，(4)式から以下のように示すことができる．

$$\mathrm{P}(\boldsymbol{w} \mid \boldsymbol{\theta}, \boldsymbol{\Phi}) = \prod_{t=1}^{T} \prod_{i=1}^{Nt} \sum_{k=1}^{K} p(z_{ti}=k \mid \boldsymbol{\theta}_t)\, p(w_{ti} \mid \boldsymbol{\Phi}_k) \quad \cdots(5)式$$

以上より，LDA モデルのパラメータは，旅行ごとのトピック分布を表す Diriclet 分布パラメータ($\boldsymbol{\alpha}$（T×K）)，ならびにトピックごとの訪問地分布を示す Diriclet 分布パラメータ($\boldsymbol{\beta}$（K×V）)によって規定される．

4)パラメータ推定について

Diriclet 分布パラメータ($\boldsymbol{\alpha}, \boldsymbol{\beta}$)を推定するためには，(5)式で表される尤度最大化（maximum likelihood estimation）が考えられるが，データ数に対してパラメータが多い場合や訪問分布の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．このような過学習を抑制し，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($\boldsymbol{\alpha}, \boldsymbol{\beta}$)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($\boldsymbol{\alpha}, \boldsymbol{\beta}$)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）．

$$\mathrm{p}(\boldsymbol{\alpha}_a, \boldsymbol{\beta}_a \mid \boldsymbol{W}, \boldsymbol{\alpha}_b, \boldsymbol{\beta}_b) = \frac{\mathrm{p}(\boldsymbol{\alpha}_a, \boldsymbol{\beta}_a \mid \boldsymbol{\alpha}_b, \boldsymbol{\beta}_b)\mathrm{p}(\boldsymbol{W} \mid \boldsymbol{\alpha}_a, \boldsymbol{\beta}_a)}{\mathrm{p}(\boldsymbol{W} \mid \boldsymbol{\alpha}_b, \boldsymbol{\beta}_b)} \quad \cdots(6)式$$

ここで，$\mathrm{p}(\boldsymbol{\alpha}_a, \boldsymbol{\beta}_a \mid \boldsymbol{\alpha}_b, \boldsymbol{\beta}_b)$：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

$\mathrm{p}(\boldsymbol{W} \mid \boldsymbol{\alpha}_a, \boldsymbol{\beta}_a)$：尤度

そして，$\mathrm{p}(\boldsymbol{W} \mid \boldsymbol{\alpha}_b, \boldsymbol{\beta}_b)$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．

$$\begin{aligned}&\operatorname{argmax} \mathrm{p}(\boldsymbol{\alpha}_a, \boldsymbol{\beta}_a \mid \boldsymbol{W}, \boldsymbol{\alpha}_b, \boldsymbol{\beta}_b) = \\ &\operatorname{argmax} \{\log\big(\mathrm{p}(\boldsymbol{\alpha}_a, \boldsymbol{\beta}_a \mid \boldsymbol{\alpha}_b, \boldsymbol{\beta}_b)\big) + \log\big(\mathrm{p}(\boldsymbol{W} \mid \boldsymbol{\alpha}_a, \boldsymbol{\beta}_a)\big)\}\end{aligned} \quad \cdots(7)式$$

以上から，汎化性能を高めるには尤度最大化（(7)式右辺第 2 項）に加えて，パラメータの事前分布（(7)式右辺第 1 項）が必要となり，この観点からも Diriclet 分布を用いる意味を確認することができる．なお，様々な確率密度分布を考えられるが，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 Diriclet 分布とセットで採用している．

さて，パラメータは Diriclet 分布によって規定されているため，(7)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，誤差が小さいと言われているギブスサンプリング（Gibbs sampling）手法を用いた[11]．

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．

$$\mathrm{PPL} = \mathrm{p}(\mathrm{w} \mid \alpha, \beta)^{-\frac{1}{V}} = \exp\left(-\frac{1}{V}\log\big(\mathrm{p}(\mathrm{w} \mid \alpha, \beta)\big)\right)$$

=exp(-対数尤度/訪問地総数)

･･･(8)式

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考えながら，尤度自体が訪問地総数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=100の場合，隠された訪問地の選択肢数を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現象がみとめられるため，PPLが再度大きくなる傾向になるといえる．

その他の指標として，最終対数尤度（$LL(\theta)$）と自由度（df）から導かれる赤池情報量規準（AIC，小さいほど好ましいモデルと判断）や初期対数尤度（$LL(0)$）と最終尤度の比から導出できる尤度比などが考えられる．

$$AIC=-2\times LL(\theta)\ +2\times df \qquad \text{…(9)式}$$

$$\rho^2=1-LL(\theta)/LL(0) \qquad \text{…(10)式}$$

なお，パラメータ数が多いため識別問題への指摘，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．後者に対しては，複数回推定を行いパープレキシティが小さいものを探索する必要がある．

3．分析データの概要

今回の分析では業務目的の場合，訪問地に制約が課されていると考え，全データ（27,680 サンプル）から訪日目的を業務ならびにトランジット，その他を除きながら「観光・レジャー」，「親族・知人訪問」にサンプルを限定した（18,012 サンプル）．**表-1** に対象サンプルの国籍・地域別訪日回数別サンプル数を示す．表中の比率*1は，2014 年出入国管理統計データによる構成比率であり，本調査データはそれに対して大きな偏りがないといえ，本調査のメリットの 1 つと考えられる．

さて，「訪日外国人消費動向調査」では日本滞在中における訪問地の設問があり，最大 10 地点を回答できる．回答は，地名を記入するようになっているが，データ入力の過程で訪問地点は，東京都を例とすると，東京，新宿，銀座，浅草，渋谷，秋葉原，上野，原宿，六本木，お台場，御徒町，アメ横，東京ドームなどのスポット単位でコーディングされ，その他は「東京（都）」として集約されている．全国では合計 526 スポットが設定されている．

さて，都道府県単位の訪問率を算出したところ，**表-2** に示すように，上位 10 位は，東京都：51%，大阪府：35%，京都府：28%，千葉県：14%，神奈川県：14%，福岡県：11%，愛知県：10%，北海道：10%，山梨県：7%，兵庫県：7%となった．平均都道府県立寄り数は 2.43 であった．

次に都道府県別の訪問地の組合せを算出したところ 1,964 パターンが存在し，上位 5 位には，東京都のみ：12%，北海道のみ：7%，沖縄県のみ：5%，大阪府のみ：4%，大阪府・京都府：4%となった．これ以下は構成比率が 4%を下回っており，多数の組合せパターンがあることがわかる．

表-1　国籍・地域別サンプル数・訪日回数別構成比率

国･地域	合計	比率	比率*1	1回目	2-3回目	4-5回目	6-9回目	10回目～
韓国	4,155	23%	21%	34%	32%	11%	7%	16%
台湾	3,546	20%	20%	25%	32%	16%	11%	15%
香港	1,204	7%	6%	19%	28%	18%	14%	21%
中国	4,365	24%	18%	67%	21%	4%	3%	5%
タイ	473	3%	5%	45%	27%	13%	7%	8%
ｼﾝｶﾞﾎﾟｰﾙ	294	2%	2%	36%	34%	12%	7%	11%
ﾏﾚｰｼｱ	465	3%	2%	56%	26%	8%	3%	6%
ｲﾝﾄﾞﾈｼｱ	252	1%	1%	55%	23%	8%	4%	10%
ﾌｨﾘﾋﾟﾝ	193	1%	2%	54%	22%	9%	3%	11%
ﾍﾞﾄﾅﾑ	67	0%	1%	67%	24%	1%	1%	6%
インド	49	0%	1%	65%	16%	6%	2%	10%
英国	366	2%	2%	57%	19%	8%	5%	10%
ドイツ	270	1%	1%	51%	23%	8%	8%	11%
フランス	311	2%	1%	57%	24%	9%	4%	6%
ロシア	154	1%	0%	47%	23%	9%	9%	11%
米国	709	4%	7%	57%	22%	7%	5%	10%
カナダ	551	3%	1%	58%	23%	7%	4%	8%
ｵｰｽﾄﾗﾘｱ	462	3%	2%	58%	26%	6%	4%	5%
その他	126	1%	7%	66%	20%	3%	7%	4%
合計	18,012	100%	100%	44%	27%	10%	7%	11%

表-2　都道府県別訪問率

都道府県	訪問率	都道府県	訪問率	都道府県	訪問率
北海道	10%	石川県	1%	岡山県	1%
青森県	1%	福井県	0%	広島県	3%
岩手県	0%	山梨県	7%	山口県	0%
宮城県	1%	長野県	3%	徳島県	0%
秋田県	0%	岐阜県	3%	香川県	1%
山形県	0%	静岡県	5%	愛媛県	0%
福島県	0%	愛知県	10%	高知県	0%
茨城県	1%	三重県	1%	福岡県	11%
栃木県	2%	滋賀県	1%	佐賀県	0%
群馬県	1%	京都府	28%	長崎県	2%
埼玉県	1%	大阪府	35%	熊本県	4%
千葉県	14%	兵庫県	7%	大分県	6%
東京都	51%	奈良県	6%	宮崎県	0%
神奈川県	14%	和歌山県	2%	鹿児島県	1%
新潟県	1%	鳥取県	0%	沖縄県	6%
富山県	1%	島根県	0%		

さて，訪日外国人マーケットを対象とした地域を考え

る場合，都道府県単位では粗いため，先に示した訪問スポット単位の訪問パターン分析を行うこととした（平均訪問スポット数：3.92）．訪問の有無に加え，その順序自体を考慮することも考えられるが，本研究では訪問順序は捨象する．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（BOW）の考え方と同一である．

4．推定結果

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，トピックスを 2~18 を設定して各々で 3 回推定を行い，PPL の平均値を算出した（**表-3**）．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（72）．これは分析対象ゾーン 526 が存在する中で，隠された訪問地の選択肢を 72（72/526≒1/73）まで絞り込めたことを示す．さらに，トピック数が 7 を超えるに従って PPL は逓増するためモデルの説明力は低下する．

また，AIC についてみると，トピックが 1 つ増加すると，ゾーン数（526）と構成比率パラメータ(1)の数だけパラメータ数が増加するため，最終対数尤度に大きな変化がなければ AIC 指標自体は増加し，モデルの評価は悪化することになる．一方，トピック数が増加すると対数初期尤度が大きくなるため，モデルによる説明力の改善度合いを示すと考えられる尤度比をみるとトピック数が増加するほど説明力の向上となっていることがわかる．

これから判断すると，PPL が最小の場合に最も良好なモデルであると統計指標から判断できることになるが，外国人旅行者の多様な訪問パターン抽出を行うことができれば，実務への適用性が広がるとも考えられる．

以上から，①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 PPL が最小となる 6 トピックが適当と考えられること，②先行研究 [3]では 18 クラスが適当との結果が示されていることから 18 トピックも比較対象として取り上げること，以上にもとづき 6 トピックならびに 18 トピックの推定結果について引き続き考察を行う．

表-4，**表-5** は，各々のトピックの構成割合，訪問率上位 8 地点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を規定するディリクレ分布パラメータ（Diriclet(α)，$\alpha=(\alpha_1, \cdots, \alpha_K)$）の「サンプル全体の和」相互の比率によって算出される．また，訪問率はディリクレ分布パラメータ（Diriclet(β)，$\beta=(\beta_1, \cdots, \beta_V)$）から導かれる訪問割合にトピック別平均訪問地点数を乗じた．

表-3　モデル説明力に関する指標

ﾄﾋﾟｯｸ数	Perplexity	最終対数尤度	AIC	初期対数尤度	尤度比
2	95	-2,396	5,845	-12,485	0.81
4	77	-2,285	6,678	-24,970	0.91
5	77	-2,286	7,207	-28,989	0.92
6	72	-2,251	7,664	-32,273	0.93
7	73	-2,253	8,196	-35,050	0.94
8	73	-2,257	8,731	-37,455	0.94
10	76	-2,279	9,829	-41,474	0.95
12	80	-2,302	10,927	-44,758	0.95
14	81	-2,313	12,003	-47,535	0.95
16	82	-2,319	13,069	-49,940	0.95
18	86	-2,345	14,177	-52,061	0.95

表-4　トピック別構成比率･主要訪問地(6 トピック)

ﾄﾋﾟｯｸ	構成比率	名称	訪問地1	訪問地2	訪問地3	訪問地4	訪問地5	訪問地6	訪問地7	訪問地8
1	31%	東京区部	新宿	浅草	渋谷	銀座				
2	31%	東京･京阪	大阪市	名古屋	大阪（府）	京都（府）	東京	京都市		
3	16%	京阪神	大阪市	京都市	神戸					
4	10%	九州	福岡市	由布院	別府	阿蘇	熊本市	博多		
5	8%	北海道	札幌	小樽	函館	登別	洞爺			
6	5%	沖縄	那覇	美ら海水族館	沖縄本島	国際通り	首里城	恩納	名護	万座毛

表-5　トピック別構成比率･主要訪問地(18 トピック)

ﾄﾋﾟｯｸ	構成比率	名称	訪問地1	訪問地2	訪問地3	訪問地4	訪問地5	訪問地6	訪問地7	訪問地8
1	10%	東京･京阪1	東京	名古屋	大阪（府）	大阪市	京都（府）			
2	9%	九州	福岡市	由布院	別府	阿蘇				
3	8%	東京1	新宿	浅草	渋谷					
4	8%	東京･京阪2	大阪市	銀座	成田	秋葉原	箱根	浅草	京都市	名古屋
5	8%	東京2	新宿	浅草	銀座	渋谷	TDR			
6	8%	北海道	札幌	小樽	函館	登別	洞爺			
7	7%	京阪神1	大阪市	京都市	神戸	清水寺	奈良市			
8	6%	東京3	新宿	渋谷	浅草	原宿	銀座	東京（都）	上野	お台場
9	6%	京阪神2	大阪市	京都市	神戸					
10	5%	東京･京阪3	京都（府）	大阪（府）	東京					
11	5%	沖縄	那覇	美ら海水族館	沖縄本島	国際通り	首里城	恩納	名護	万座毛
12	5%	京阪広島	京都市	大阪市	広島市					
13	4%	東京･京都	新宿	渋谷	浅草	銀座	原宿	秋葉原	京都（府）	
14	3%	東京4	浅草	池袋	銀座	新宿				
15	3%	東京･京阪4	大阪市	京都市	成田	新宿	浅草	名古屋	箱根	富士五湖
16	2%	昇龍道	名古屋	高山	金沢					
17	2%	京阪	大阪市	難波	京都市	梅田				
18	1%	四国	高松	香川（県）	大阪市	松山				

※赤：訪問率≧60%，橙≧40%，青≧20%

表-4 は，6 トピックの推定結果で，構成比率の多いトピック順に示しており，単独の地方内で訪問しているものが 5 トピック，複数地方を周遊しているものが第 2 トピックの 1 つ（構成比率：31%）になっている．それに対して，**表-5** は 18 トピックの推定結果であり，複数地域訪問（表中，太線で表示）は 30%と同様の結果となった．さらに，**表-6** は訪問地点の類似から，おおまかに両モデルのトピックの関連性を示したものである．これより，6 トピックで構成比率の高かった第 1~3 トピックが細分化されるとともに，新たに抽出されたものとして，第 12 トピック（京阪広島，構成比率：5%），第 16 トピック（昇龍道，構成比率：2%），第 18 トピック（四国，構成比率：1%）をあげることができる．

表-6　6 トピックと 18 トピックとの比較

6トピック				18トピック						
ﾄﾋﾟｯｸ	構成比率	名称	主要訪問地	ﾄﾋﾟｯｸ	構成比率	構成比率	名称	訪問地1	訪問地2	特徴的な訪問地
1	31%	東京	新宿	3	8%	26%	東京1	新宿	浅草	渋谷
			浅草	5	8%		東京2	新宿	浅草	TDR
			渋谷	8	6%		東京3	新宿	渋谷	お台場
			銀座	14	3%		東京4	浅草	池袋	銀座
2	31%	東京・京阪	大阪市	1	10%	30%	東京・京阪1	東京	名古屋	大阪（府）
			名古屋	4	8%		東京・京阪2	大阪市	銀座	成田
			大阪（府）	10	5%		東京・京阪3	京都（府）	大阪（府）	東京
			東京	15	3%		東京・京阪4	大阪市	京都市	成田
			京都市	13	4%		東京・京都	新宿	渋谷	浅草
3	16%	京阪神	大阪市	7	7%	15%	京阪神1	大阪市	京都市	清水寺
			京都市	9	6%		京阪神2	大阪市	京都市	神戸
			神戸	17	2%		京阪	大阪市	難波	京都市
4	10%	九州	福岡市	2	9%		九州	福岡市	由布院	別府
5	8%	北海道	札幌	6	8%		北海道	札幌	小樽	函館
6	5%	沖縄	那覇	11	5%		沖縄	那覇	美ら海水族館	沖縄本島
				12	5%		京阪広島	京都市	大阪市	広島市
				16	2%		昇龍道	名古屋	高山	金沢
				18	1%		四国	高松	香川（県）	松山

表-7 は国籍・地域別トピック構成比率を示したものである．特徴的なものとして，韓国（九州（第 2 トピック）が多い），台湾・香港（北海道（第 6 トピック）や沖縄（第 11 トピック），九州（第 2 トピック）や昇龍道（第 16 トピック）が多い），中国（ゴールデンルートである東京-京阪 1，2 の割合が高い），タイ・シンガポール（北海道（第 6 トピックの割合が高い））などのほか，欧米で広島を訪問する第 12 トピックの高さも特徴的である．このように旅行者の国籍・地域によっても選好される訪問地の組合せ（トピック）が異なることがわかり，地域におけるターゲティング戦略策定の際に参考情報になると考えられる．

表-7　国籍・地域別トピック構成比率

(上：6 トピック，下：18 トピック)

ﾄﾋﾟｯｸ	構成比率	名称	韓国	台湾	香港	中国	タイ	ｼﾝｶﾞﾎﾟｰﾙ	ﾏﾚｰｼｱ	ｲﾝﾄﾞﾈｼｱ	英国	ドイツ	フランス	米国	カナダ	ｵｰｽﾄﾗﾘｱ	総計
5	8%	北海道	6%	12%	11%	6%	12%	11%	7%	2%	2%	1%	1%	3%	2%	3%	7%
1	31%	東京1	25%	33%	30%	28%	45%	47%	31%	37%	53%	44%	44%	51%	52%	36%	33%
2	31%	東京・京阪	13%	19%	21%	52%	28%	27%	43%	44%	38%	39%	43%	36%	36%	51%	31%
3	16%	京阪神	24%	20%	21%	8%	8%	11%	17%	16%	5%	11%	10%	7%	7%	8%	15%
4	10%	九州	26%	8%	6%	1%	6%	5%	2%	1%	1%	4%	2%	2%	1%	2%	9%
6	5%	沖縄	5%	8%	11%	4%	0%	0%	0%	0%	1%	0%	0%	1%	0%	0%	5%
6	8%	北海道	7%	12%	13%	7%	15%	10%	8%	3%	2%	1%	2%	2%	3%	4%	8%
3	8%	東京1	6%	9%	8%	7%	14%	9%	10%	12%	14%	12%	8%	16%	13%	6%	9%
5	8%	東京2	5%	10%	8%	6%	12%	13%	9%	8%	13%	10%	8%	13%	13%	10%	8%
8	6%	東京3	8%	6%	6%	4%	7%	11%	4%	10%	12%	11%	9%	9%	11%	7%	7%
14	3%	東京4	1%	4%	2%	4%	3%	3%	1%	1%	2%	2%	3%	2%	3%	2%	3%
13	4%	東京・京都	2%	3%	3%	3%	4%	9%	6%	6%	11%	8%	14%	8%	10%	13%	4%
1	10%	東京・京阪1	5%	5%	6%	23%	8%	9%	15%	8%	11%	11%	9%	9%	5%	11%	11%
4	8%	東京・京阪2	3%	3%	4%	19%	6%	6%	9%	15%	5%	6%	6%	6%	6%	7%	8%
10	5%	東京・京阪3	3%	3%	3%	4%	8%	4%	4%	9%	6%	7%	14%	8%	8%	13%	4%
15	3%	東京・京阪4	1%	2%	2%	6%	3%	2%	4%	2%	2%	1%	1%	1%	2%	3%	3%
17	2%	京阪	5%	1%	2%	1%	1%	0%	2%	2%	0%	1%	0%	1%	0%	0%	2%
7	7%	京阪神1	11%	10%	10%	4%	3%	5%	10%	8%	1%	3%	3%	4%	1%	3%	7%
9	6%	京阪神2	6%	8%	9%	4%	3%	5%	6%	4%	2%	6%	5%	3%	4%	4%	5%
12	5%	京阪広島	1%	3%	2%	2%	3%	6%	3%	5%	15%	15%	14%	13%	15%	12%	4%
16	2%	昇龍道	1%	5%	4%	1%	2%	1%	5%	5%	1%	1%	0%	1%	1%	2%	2%
18	1%	四国	1%	1%	1%	0%	0%	0%	0%	1%	1%	1%	0%	1%	1%	0%	1%
2	9%	九州	31%	8%	7%	2%	7%	5%	2%	1%	1%	4%	3%	3%	3%	2%	10%
11	5%	沖縄	6%	9%	11%	4%	1%	1%	0%	0%	1%	0%	0%	2%	1%	1%	5%

先に示したように，トピック数の決定には様々な考えがあるが，本論文では昇龍道や瀬戸内地方における DMO（Destination Management/Marketing Organization）活動なども視野に入れることを意図して，18 トピックの結果に着目する．

トピックモデルでは，1 つの訪日旅行が複数トピックを有すると仮定している．この確率はディリクレ分布パラメータ $\boldsymbol{\alpha}$ から算出することができるため，構成割合が多いものから第 1 位から第 3 位トピックと設定し，第 1 位の構成比率を昇順にしめした（**図-1**）．横軸はサンプル番号を便宜的に示しているが，約 12,000 番目までは第 1 位トピックの構成比率が 0.9 以上と高く，優位となっていることがわかる．

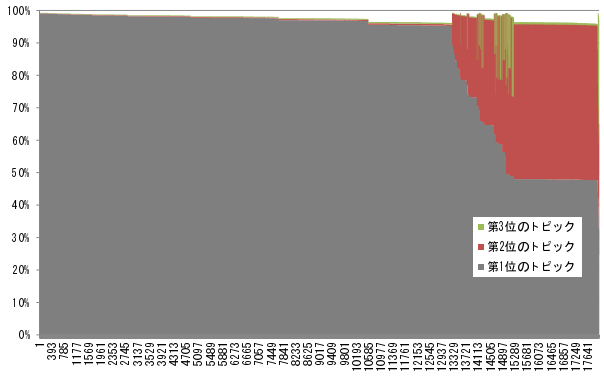

図-1　旅行者別第 1 位~第 3 位までのトピック構成比率

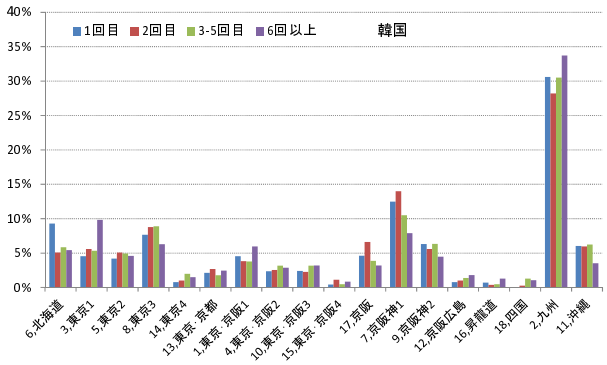

図-2　訪日回数別第 1 位トピックの構成比率（韓国）

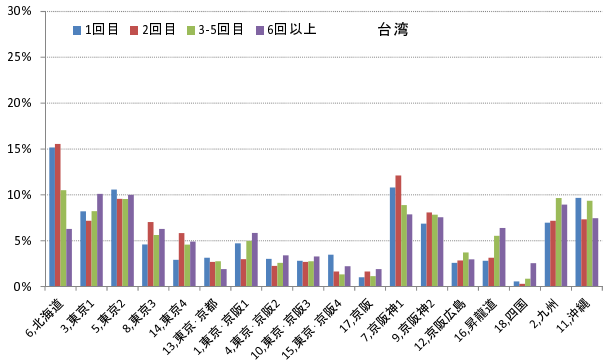

図-3　訪日回数別第 1 位トピックの構成比率（台湾）

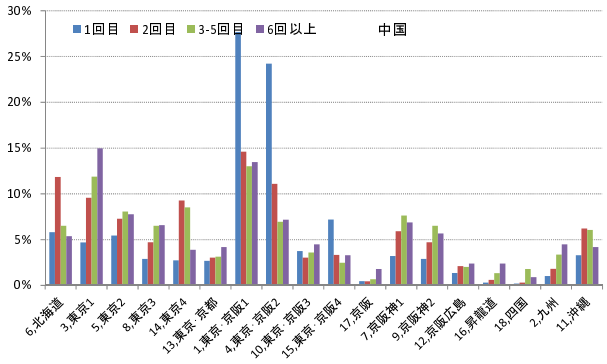

図-4　訪日回数別第 1 位トピックの構成比率（中国）

さて，**表-7** における国籍・地域別トピック構成比率に加えて，訪日回数を考慮した主要国・地域の訪日回数別トピック構成比率を示した（**図-2~6**）．

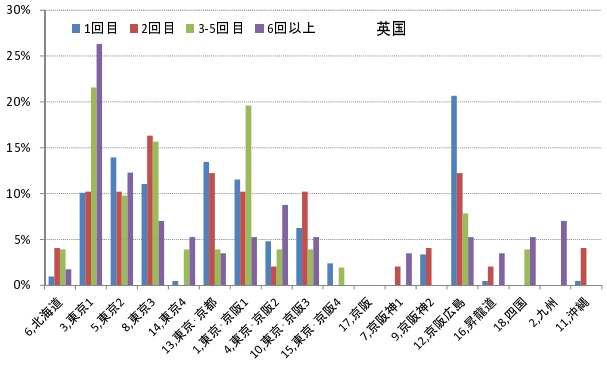

図-5　訪日回数別第 1 位トピックの構成比率（英国）

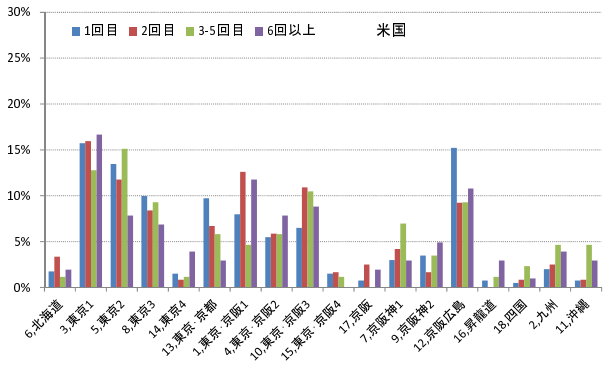

図-6　訪日回数別第 1 位トピックの構成比率（米国）

韓国では九州（第 2 トピック）が高く，訪日回数の増加による変化は顕著に見られないが，京阪神 1 が減少傾向を示していることがわかる．その他の図より，下記のような特徴がみられる．

・台湾　北海道,京阪神 1：減少，東京 1･3 の微増
・中国　東京-京阪 1･2：減少，東京 1,京阪神 1･2：微増
・英国　東京･京都,京阪広島：減少，東京 1･3･4：増加
・米国　東京･京都,東京 3：微減

以上から，初訪日時は比較的広域に訪問を行うものの，訪日回数が増加するに従って特定の都市への訪問に移行すると考えられる．

次に，四半期別のトピック構成比率を算出した（**図-7**）．観光資源の特性，訪日外国人の休暇と関連して，トピック構成比率が変化していると考えられる．

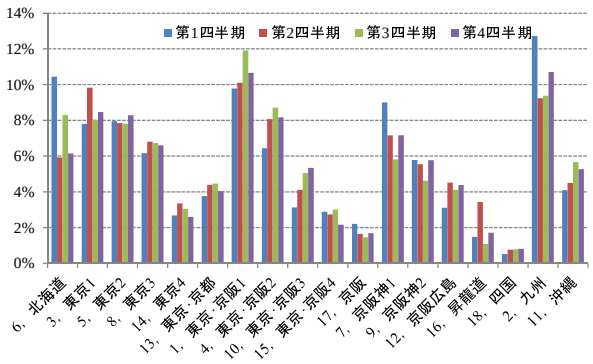

図-7　四半期別構成比率

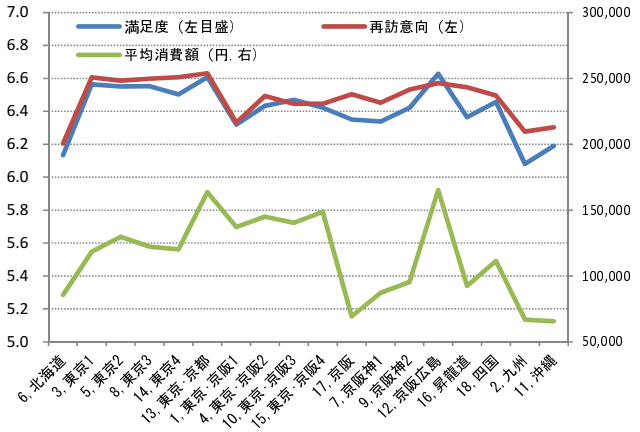

図-8　トピック別平均満足度，再訪意向，平均消費額

図-8 は，トピック別の平均満足度，再訪意向，平均消費額である．満足度，再訪意向は 7 段階評価（大変満足・必ず来たい：7 点，満足・来たい：6 点，やや満足・やや来たい：5 点，普通・何ともいえない：4 点，やや不満，不満，大変不満：1 点）となっているため，いずれのトピックも 6 点以上であることがわかる．しかしながら，東京 1~東京・京都の満足度は 6.5 点を超えるため，概ね 7 点と評価した被験者が多かったと推察できる．それと比較すると，北海道，東京-京阪 1，九州，沖縄の満足度や再訪意向が若干低い．リピーターの確保，良好な口コミ発信のために，さらなる改善が急務と考えられる．また，消費額については，複数地方の訪問など移動距離や滞在日数の増加によって増加する傾向が見られる．

5．まとめ

本研究は，観光・レジャー目的の訪日外国人旅行者の訪問パターンの特性把握を目的として，分析に用いるトピックモデルの概念整理や潜在クラスモデルとの比較を行った．次に，「訪日外国人消費動向調査（平成 26 年）」データを用いてトピックモデルの推定を行った結果，1 地方のみの訪問が約 7 割を占めること，国籍・地域，訪日回数，訪問時期によってトピック構成比率が異なることが確認できた．

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，各訪問パターンと来訪者属性との関連性把握が容易になり，来訪者へのプロモーションへの示唆，ゾーンから見たターゲット設定に対して有益な情報になると考えられる．

今後の課題として，最新データへの適用，分析ゾーンの細分化，より効果的なプロモーションの検討，効果的な観光振興策への落とし込みが考えられる．

【謝辞】

本研究の分析にあたり，データを提供いただきました観光庁ならびに有益なコメントを頂きました西井和夫教授（流通科学大学），岡本直久教授（筑波大学），野瀬元子准教授（静岡英和学院大学）をはじめとする関係諸氏に深謝の意を表します．

【参考文献】

1) 観光庁(2012,2013)（訪日外国人消費動向調査），http://www.mlit.go.jp/kankocho/siryou/toukei/syouhityousa.html（2016.2.6 閲覧）
2) 古屋秀樹：訪日外国人旅行者の地区レベル訪問パターンの基礎分析，第 53 回土木計画学研究・講演集(CD-ROM)，No. 53，2016
3) 劉瑜娟，古屋秀樹：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(CD-ROM)，No.52，2015
4) 古屋秀樹，劉瑜娟：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)，2015
5) 古屋秀樹，劉瑜娟：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，土木学会論文集Ｄ３・特集号（土木計画学研究・論文集），投稿中
6) 佐藤一誠：トピックモデルによる統計的潜在意味解析，コロナ社，2015
7) 前掲 6
8) 岩田具治：トピックモデル，講談社，2015
9) Graham Neubig：奈良先端科学技術大学院大学 HP（NLP Programming Tutorial 7 - トピックモデル），http://www.phontron.com/slides/nlp-programming-ja-07-topic.pdf，2016.2.16 閲覧
10) 伊塚井誠人，椎野創介：討議録に対するトピックモデルの適用，第 52 回土木計画学研究発表会講演集(CD-ROM)，No. 52，2015
11) 伊庭幸人，種村正美他：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，2005 庭幸人，種村正美他：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，2005

(2016. 7. 31 受付)