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

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特性把握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23，24，25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 Alocation 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．さらに，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おり，外国人旅行者の増加は，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，初回来日者の満足度を高めてリピーターへの変容をはかる質的観点からの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現状の観光行動特性を把握し，来日回数や国籍別，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して，それぞれの特徴把握によってより詳細に効果的施策の検討，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5万人を数えるため，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

⑴　訪日外国人消費動向調査[1)]

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23~26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おける訪問地，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．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，旅行形態（団体，個人旅行）を活用した．

⑵　潜在クラス分析の概要と問題点

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2や都道府県単位に分析した文献3，4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$\delta_{k,1,n}$，$\delta_{k,2,n}$（訪問有り：$\delta_{k,1,n}$ =1，$\delta_{k,2,n}$=0，訪問無し：$\delta_{k,1,n}$=0，$\delta_{k,2,n}$=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．そして，(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（類似度）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[5]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[6),7),8),9)]

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（Probabilitsic 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）．

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

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

（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複数のトピック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($\boldsymbol{\beta}$)，$\boldsymbol{\beta}$=(β_1，$\cdots$，β_V)，β_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)

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

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），

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

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($\boldsymbol{\theta}_t$~Diriclet($\boldsymbol{\alpha}$)，$\boldsymbol{\alpha}$=(α_1，$\cdots$，$\boldsymbol{\alpha}_K$)，α_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}∈{1，･･･，K}）．この時，旅行 t において，

a) Diriclet 分布のパラメータ（$\boldsymbol{\theta}_t$ ~Diriclet($\boldsymbol{\alpha}$)，$\boldsymbol{\alpha}$=(α_1，･･･，$\boldsymbol{\alpha}_K$)）に従って，旅行 t のトピック z_{ti} が生成される（z_{ti}~Multi($\boldsymbol{\theta}_t$)，i=1,･･･，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．

b) 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}~Multi($\boldsymbol{\Phi}_{zti}$)，i=1,･･･，N_t）

c) 旅行 t の生起確率は下記のように示すことができる．

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

･･･(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．

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

･･･(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）．

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

･･･(6)式

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

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

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

$$\operatorname{argmax}\mathrm{p}(\boldsymbol{\alpha}_a,\boldsymbol{\beta}_a|\boldsymbol{W},\boldsymbol{\alpha}_b,\boldsymbol{\beta}_b)=\operatorname{argmax}\{\log\big(\mathrm{p}(\boldsymbol{\alpha}_a,\boldsymbol{\beta}_a|\boldsymbol{\alpha}_b,\boldsymbol{\beta}_b)\big)+\log\big(\mathrm{p}(\boldsymbol{W}|\boldsymbol{\alpha}_a,\boldsymbol{\beta}_a)\big)\}$$

･･･(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 分布を採用している．

さて，各パラメータは 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 smapling）手法を用いた[10]．

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．

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

=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=1/100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各々のマーケット（トピック）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逆に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．

一方，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検索する必要がある．

３．訪問パターンの推定結果並びに検証

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，業務ならびにトランジット，その他の訪日目的を除く「観光・レジャー」に限定した．

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 箇所を回答できる．回答は，地名を記入するようになっている．

訪問パターンは，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．

なお，詳細な分析結果は，発表時に譲る．

４．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，まずはじめに分析で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．潜在クラス分析では，訪問率に 2 項分布を仮定しているのに対して，トピックモデルでは多項分布を用いているため，直接的にパラメータを比較できないものの，各旅行で異なったトピック構成比率を設定できること，汎化性能の高いパラメータを導出できることが特徴と考えられる．

分析では，「訪日外国人消費動向調査（平成 24，25，26 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 地方のみの訪問が複数地方にまたがる周遊型パターンよりも多い構成比率を占めること，国籍・地域，来日回数，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・着地からみ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）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) 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
3) 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
4) 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
5) 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
6) 前掲 5
7) 岩田具治：トピックモデル，講談社，2015
8) 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
9) 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
10) 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

(2016.4.16 受付)