観光スポット間遷移データ収集において異常回答率を低減させる Web アンケート回答フローの評価

長谷川 凌真† 渡邉 貴之†

本研究では、Web アンケートによる観光行動調査手法の有効度を、回遊行動モデルから算出される異常スコアを指標として評価する。Web アンケートを用いることで、多様な個人属性と紐付けた行動データを多数のユーザから収集することが期待できるが、ユーザによる異常回答の存在が課題となる。回答フローの変更により異常回答率を低減し、高精度な行動データの収集に有効な Web アンケート手法を検討する。異なる回答フローを用いたデータセットごとに異常スコアを算出することで、回答フローの改善による異常回答率の低減効果を自動的かつ定量的に評価できることを示す。

1. はじめに

観光の利便性やサービスを向上し地域や個人消費を活性化させるために、観光客の行動実態の把握は重要な情報資源となる。また、観光客のデモグラフィック属性や旅行形態、嗜好といった個人属性データと行動データを紐付けて調査することで、より詳細な行動実態把握やシミュレーション、施策の効果測定が期待できる[1]。ここで、GPS データやアンケートの回答データから得られる、2 つの観光スポット間の遷移を表すデータを「遷移データ」、遷移データの集合体を「行動データ」と定義する。

観光地に赴き質問紙調査を行うことで、多数の観光客からどの観光スポットにどの順番で立ち寄ったかの行動データを収集することができるが、対象は現地調査を行う日時に限られ、継続的なデータ収集に基づく分析への活用には課題がある。GPS 調査[2][3]は観光客の行動を詳細に追跡することが可能であるが、GPS データを調査主体に提供することに対して、個人情報保護の観点から抵抗感を持つ観光客が存在し、高精度な分析に足りうるサンプル数の確保には困難が伴う。また、GPS 調査であることを隠してアプリケーションやキャリアデータから位置情報を取得する場合、利用者情報は匿名環境下で得られる限定的なものとなり、個人属性に基づく行動分析は困難となる。

そこで、行動分析を行うための十分なサンプル数を確保しつつ、多様な個人属性データと紐付けされた行動データを得るために、Web アンケートによる観光行動調査の有効性が検討されている[4]。Web アンケートでは、宿泊施設のフロントなどで二次元バーコードを印刷したチラシなどを宿泊客に配布することで、中長期間に渡る継続した行動データの収集が可能となる。Web アンケートでは、観光客の位置情報の粒度は GPS 調査と比較して圧倒的に粗いため、調査協力への抵抗感を抑え、より多くのサンプルデータの収集が期待できる。また、観光客の詳細な個人属性データについても Web アンケートで収集できることから、観光客の多様な属性と行動データを紐付けた分析が可能となる。

しかし、Web アンケートによる行動データは GPS 調査のように自動的に収集されるものではなく、観光客が自ら周遊した観光スポット等を選択もしくは入力する必要がある。そのため、回答の信頼性は観光客の善意と正確な入力に委ねられ、「異常回答」すなわち不正確な回答や架空の回答が少なからず発生する。本研究における「異常回答」とは、アンケートの回答者が実際に訪問しなかった観光スポットを選択すること、また実際に行った移動とは異なる順番で観光スポットを選択して行動データを回答することを指す。異常回答の発生要因には、回答者が回答フローを誤認して誤った選択を行う場合と、故意に乱雑な選択を行う場合が考えられる。異常回答が多いほど高精度な分析は困難となるため、操作の誤認を防ぐために分かりやすく、同時に乱雑な選択を誘発しない回答フローを用いる必要がある。

本研究では、Web アンケートを用いて高精度な行動データの収集を可能とするために、その回答フローの良否を数値指標に基づいて評価できる手法について示す。本手法では、まずアンケートによって収集した行動データから回遊行動モデル[5]のパラメータを推定する。求められた回遊行動モデルから各ユーザの行動データの異常スコアを計算する。異常スコアの統計量から、行動データを収集する際の回答フローの良否が評価できる。本研究では、複数回の観光行動調査を通じて回答フローの改良を行い、回答フローの変更による改善を異常スコアによって評価する。

2. 先行研究

文献[4]では、「静岡・ふじのくに割アンケート」（以降「ふじのくに割アンケート」と表記）において、スマートデバイス向けの Web アンケートに GPS 調査を組み合わせた観光行動調査システムを構築している。また、Web アンケートにより収集される個人属性データと、GPS 調査の行動データを紐付けして収集し分析を行うことを試みて

† 現在，静岡県立大学経営情報学部
Presently with University of Shizuoka.

いる。しかし、Web アンケートの回答数に比較して GPS 調査のサンプル数は少数にとどまることが予想された。そのため、「ふじのくに割アンケート」に訪問した観光スポットとその訪問順を尋ねる設問を設けることで、GPS 調査に頼らずに観光客の行動データを収集することを試みている。また、GPS 調査と Web アンケートによる行動データ収集の双方を行い、その回答率とデータ整合性の検討を行っている。

図 1「ふじのくに割アンケート」遷移回数の分布

文献[4]における Web アンケート回答者のうち GPS データを提供した利用者は 7.7%にとどまった。その一方で、Web アンケートにおける行動データの回答は高い回答率が得られている。また Web アンケートと GPS 調査の間でエリア別訪問者数を比較したところ、Web アンケートと GPS 調査のエリア別訪問者数には相関係数 R=0.83 の高い相関が見られ、Web アンケートによる観光行動調査の結果は GPS 調査との整合性を有すると述べている。

しかし、先述の通り Web アンケートによる行動データの収集の際には、異常回答を誘発しにくい回答フローを用いて乱雑な回答や誤答を防ぐことが求められる。文献[6]で示された「ふじのくに割アンケート」のユーザの遷移回数の分布（図 1）を見ると、一部のユーザが突出して多くの遷移を回答していることが分かり、異常回答を行っていると推測される。また、それ以下の遷移回数を回答したユーザの中でも、距離が大きく離れたスポット間の遷移を複数回行うなど、実際の観光行動では選択し難い異常回答が行われている可能性が指摘されている。

文献[6]において大石らは、Web アンケートにより収集された行動データの中から異常回答を検知・除外するため、Tandon らの文献[7]に基づき異常スコア（Anomaly Score）を算出し、ユーザの行動データが異常であるかを数値化する手法を提案している。本研究においても、同様の手法を用いて、行動データを収集するための複数の異なる回答フローの良否を評価する。以降では、本研究で使用する回遊行動モデルおよび異常スコアについて説明する。

3. 回遊行動モデル

鈴木ら[5]は、Web アンケート回答者の行動データから観光行動をモデル化するため、スポット間の距離とスポットの人気度に基づく条件付き確率によって回遊行動モデルを定義する手法を提案している。選択肢となるスポット集合を S={r,s,q･･･}、あるスポット r から別のスポット s に移動する際の最短距離を d(r,s)とする。このとき、モデルを用いた 2 つのスポット r, s 間の遷移確率 p_1 は次式で定義される。

$$p_1(s \mid r;\theta_1) = \frac{d(r,s)^{-\theta_1}}{\sum_{q\in S} d(r,q)^{-\theta_1}} \tag{1}$$

ここで、θ_1 は距離に対するパラメータであり、Levy Flight [5]のベキ係数 λ に対応する。また、スポットの人気度を加味するために、スポット s の人気度 f(s)に対するパラメータを θ_2 とすると、スポット s の選択確率 p_2 は次式で定義される。

$$p_2(s;\theta_2) = \frac{f(s)^{\theta_2}}{\sum_{q\in S} f(q)^{\theta_2}} \tag{2}$$

なお本研究では、人気度 f(s)をデータセット中の全ユーザにおける各スポットの訪問者数とする。

p_1 と p_2 はそれぞれ独立しており、これらを組み合わせることにより、回遊行動モデルは以下の条件付きスポット遷移確率 p で定義される。

$$p(s \mid r;\theta_1\theta_2) = \frac{p_1(s \mid r;\theta_1)\, p_2(s;\theta_2)}{\sum_{q\in S} p_1(q \mid r;\theta_1)\, p_2(q;\theta_2)} \tag{3}$$

本研究では、パラメータ $\theta = (\theta_1, \theta_2)^T$ の推定に文献[5]と同様の機械学習的アプローチを用いる。まず、ユーザ u が h 番目に訪問したスポット s(u,h)からなる行動データ集合を D={･･･,s(u,h),･･･}、ユーザ u が訪れた総スポット数を H(u)とする。このとき、目的関数である次式の対数尤度式の最大化から θ を推定する。ここで、U はユーザの集合であり u∈U である。

$$L(\theta;D) = \sum_{u\in U}\sum_{h=1}^{H(u)-1} \log p\left(s(u,h+1) \mid s(u,h);\theta\right) \tag{4}$$

$$\hat{\theta} = arg\max_{\theta} L(\theta;D) \tag{5}$$

ここで、$x = (-\log d(r,s), \log f(s))^T$ と定義されたベクトルを導入すれば、上記の対数尤度式は次式で書き表せる。

$$L(\theta;D) = \sum_{u\in U}\sum_{h=1}^{H(u)-1}\left[\theta^T x\left(s(u,h), s(u,h+1)\right) - \log\sum_{q\in S}\exp\left(\theta^T x\left(s(u,h), q\right)\right)\right] \tag{6}$$

よって、目的関数の勾配ベクトルとヘス行列が以下で算出される。

$$\frac{\partial L(\theta;D)}{\partial \theta} = \sum_{u\in U}\sum_{h=1}^{H(u)-1}\left[x(s(u,h),s(u,h+1)) - \sum_{q\in S} p(q\mid s(u,h);\theta)\ x(s(u,h),q)\right] \quad (7)$$

$$\frac{\partial^2 L(\theta;D)}{\partial\theta\partial\theta^T} = \sum_{u\in U}\sum_{h=1}^{H(u)-1}\left[\sum_{q\in S} p(q\mid s(u,h);\theta)\ x(s(u,h),q)\ x(s(u,h),q)^T - \left(\sum_{q\in S} p(q\mid s(u,h);\theta)\ x(s(u,h),q)\right)\left(\sum_{q\in S} p(q\mid s(u,h);\theta)\ x(s(u,h),q)\right)^T\right] \quad (8)$$

反復計算を用いてパラメータ推定を行う。推定は、次式の修正ベクトルによるニュートン法を用いる。

$$\delta = -\frac{\partial L(\theta;D)}{\partial\theta}\left(\frac{\partial^2 L(\theta;D)}{\partial\theta\partial\theta^T}\right)^{-1} \quad (9)$$

A1：パラメータベクトル θ を適当な値で初期化する
A2：修正ベクトル δ を計算する
A3：定数 $\varepsilon = 10^{-8}$ とし、$\|\delta\| < \varepsilon$ ならば反復を終了する
A4：パラメータベクトルを $\theta \leftarrow \theta + \delta$ で更新し、A2 に戻る

4. 異常スコア

機械学習による対数尤度式の最大化から推定されたパラメータ θ を実際に回遊行動モデルに当てはめ、データセット内の異常データの検知を行う。Tandon らは、確率的アプローチに基づいた異常スコアを用いて無線ネットワーク環境を利用し不正アクセスを行うユーザの特定を試みている[7]。具体的には、損失関数（遷移確率の対数の負値）の和を異常スコアとし、これが高いものほど異常なアクセスであると判定している。

また、大石らは損失関数を回遊行動モデルによる条件付きスポット遷移確率とし、行動データの異常検知に文献[7]と同様な異常スコアを用いている[6]。具体的には、式(10)で示される異常スコアがより高い行動データほど、ユーザが実際の回遊行動とは異なる異常回答を行った可能性が高いものと判定している。

$$AnomalyScore = -\sum_{h=1}^{H(u)-1}\log p(s(u,h+1)\mid s(u,h);\theta) \quad (10)$$

式(10)から、遷移確率が毎回同じであるならば、遷移回数が多いユーザほど異常スコアが大きくなることがわかる。また遷移回数が同じユーザ同士であれば、遷移確率の低い遷移を繰り返しているユーザほど異常スコアの値は大きくなる。文献[6]では、異常スコアの低いユーザほど、距離の近い少数のスポットを効率的に遷移する傾向が見られたのに対して、異常スコアの高いユーザは、多数のスポットを大きく距離の離れた遷移を交えながらランダムに遷移する傾向が見られたとしている。

表 1 データセットの比較

	ふじのくに割	意外と熱海	対面調査
実施時期	2015 年 9 月～2016 年 2 月	2017 年 2 月～2019 年 12 月	2019 年 10 月
回答形式	Web アンケート	Web アンケート	対面・質問紙調査
選択方法	一覧	一問一答	一覧
対象地域	静岡県全域	熱海市内	熱海市内
回答者数	1751	840	651
訪問スポット回答者数	1651	591	636
上記回答率	94.3%	70.4%	97.7%
遷移データ回答者数	1505	507	469
上記回答率	86.0%	60.4%	72.0%
有効行動データ件数	1259	430	456
上記回答率	72.0%	51.2%	70.0%
有効遷移総数	4571	1036	1002

5. データセット

本研究では、以下の３回のアンケート調査で得られた行動データについて、回遊行動モデルのパラメータ推定と異常スコアの計算を行う。

- 静岡・ふじのくに割アンケート（ふじのくに割）
- 意外と熱海アンケート（意外と熱海）
- 2019 年度 熱海市観光客実態調査（対面調査）

各データの詳細を表 1 に示す。ここで、「対面調査」は紙ベースの調査であり、調査員と対面した状態で回答していることから異常回答は低いことが予測されるため Web アンケートの異常回答率の評価基準として用いることができる。また、「ふじのくに割」の Web アンケートは回答フローの改良前、「意外と熱海」の Web アンケートは回答フローの改良後のアンケートである。以下、各アンケートの詳細について説明する。

5.1 2019 年 熱海市観光客実態調査（対面質問紙調査）

Web アンケートの異常回答率の良否を評価するため、2019 年 10 月 5 日・6 日に実施された「2019 年度 熱海市観光客実態調査」（以降「対面調査」と表記）で収集した対面による質問紙調査のデータを比較対象とする。アンケート対象者は調査実施日に熱海市内を訪れた日本人観光客とし、主要な観光スポット 7 地点で調査員が声をかけ、協力を得られた観光客から回答を得た。観光客に質問紙を渡し、調査員が必要に応じて質問項目の説明および回答の確認をしながら記入を行ったため、質問紙から得られる行動データは異常回答率の低いものであると考えられる。アンケート内の「訪問地」に関する質問では、24 箇所の主要な観光スポットを選択肢とし、観光客が回答時点で 15 分以上滞在

したスポット、または質問紙の回答後に立ち寄りを予定しているスポットを、各スポットに割り振られた番号で訪問順に記入する。

「対面調査」では、2 日間の合計で 651 件の回答が得られた。ここから、立ち寄りスポット数が 2 箇所に満たない回答は行動データとして取り扱うことができないため除外する等のデータクリーニングを行い、結果として 456 件の行動データをデータセットとして得た。遷移総数は 1036 回となり、データセット上に含まれる全観光客の行動データを地図上に可視化したものを図 2 に示す。

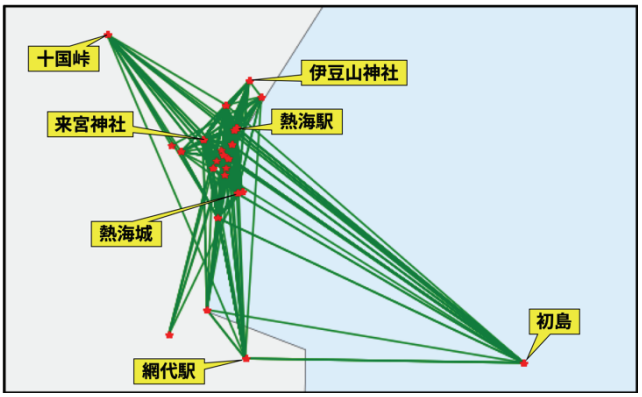

図 2「対面調査」全ユーザーの行動データ

5.2 静岡・ふじのくに割アンケート（Web アンケート回答フロー改善前）

行動データの回答フローを改良する前の Web アンケートとして、2015 年 9 月 11 日から 2016 年 2 月 29 日までの 171 日間に実施された「静岡・ふじのくに割アンケート」で収集したデータを使用する。

アンケート対象者は原則として静岡県内を旅行する静岡県外からの日本人観光客とし、静岡県内の宿泊施設 578 箇所で調査を実施した。対象者が宿泊先にチェックインした際に、Web アンケートサイトの URL を埋め込んだ QR コード付きの名刺を配布し、アンケートへの回答を依頼した。アンケート内の「訪問地」に関する質問では、静岡県内およびその周辺の 1556 箇所の観光スポットの中から、ユーザの宿泊地を中心として半径 40km 圏内のスポットを、スポットの属するカテゴリ別にチェックボックス形式で一覧表示した。実際の Web アンケート上での画面例を図 3 左に示す。ユーザには、宿泊前に訪問したスポットと、宿泊後に訪問を予定しているスポットを全て選択するよう依頼した。その後、図 3 右の画面においてスマートフォンのスワイプ操作でスポットを訪問順に並び替えるという回答フローを用いた。

この回答フローの問題点としては、スポットの選択と並べ替えのフローを切り離したことによって、異常回答を誘発する可能性が生じる点にある。すなわち、ユーザは訪問順を考えずにスポットを乱雑に選択することができ、その後の並べ替えではユーザが訪問順に並べ替え忘れる、もしくは意図的に並べ替えをスキップすることが可能となっている。

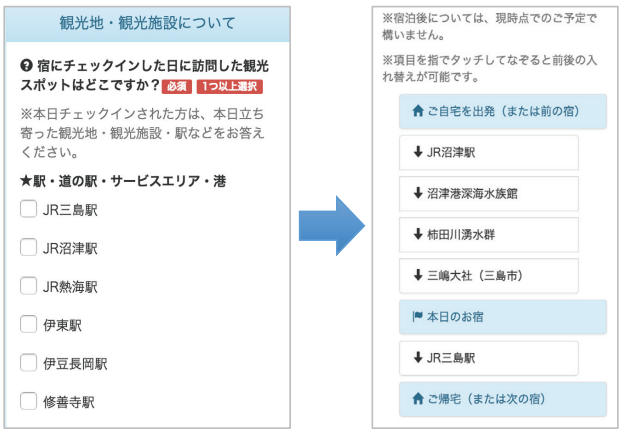

図 3「ふじのくに割アンケート」遷移データ回答画面

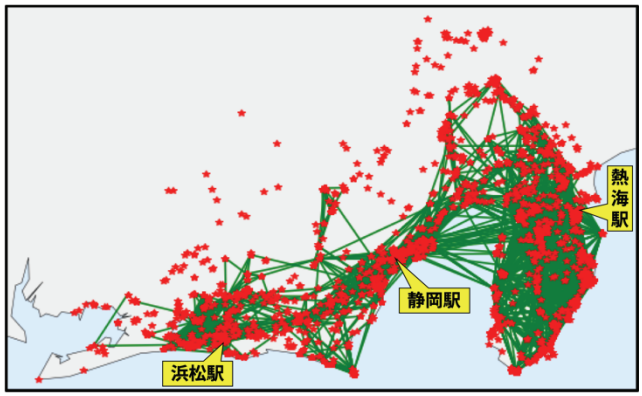

図 4「ふじのくに割アンケート」全ユーザの行動データ

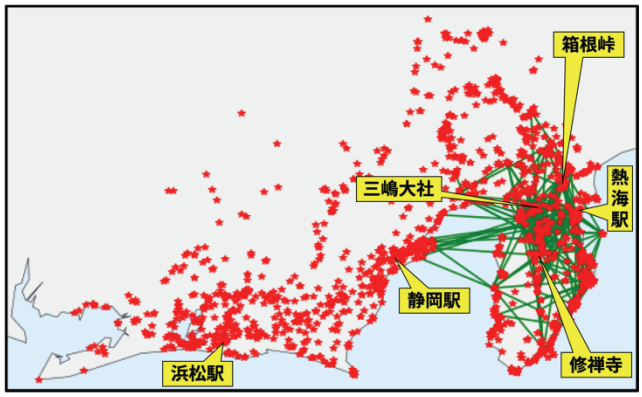

図 5「ふじのくに割アンケート」
熱海市を経由する全ユーザの行動データ

「ふじのくに割アンケート」では、1751 件の回答が得られた。本研究では、宿泊施設を除いた選択スポット数が 2 箇所に満たない回答を除外するなどのデータクリーニングを行い、結果として 1259 件の回答をデータセットとして用いる。遷移総数は 4571 回となり、データセット上に含まれる全ユーザの行動データを地図上に可視化したものを図

4 に示す。

また、対象エリアをより狭域に限定したデータセットとして、1556 箇所の観光スポットのうち、熱海市内にある 52 箇所のいずれかを経由する行動データのみを抽出したものを「熱海市を経由したユーザ」のデータセットとして用いる。熱海市を経由したユーザとして 71 件の回答が該当し、遷移総数は 401 回であった。データセットに含まれる全ユーザの行動データを地図上に可視化したものを図 5 に示す。

5.3 意外と熱海アンケート（Web アンケート回答フロー改善後）

我々は、2017 年 2 月 2 日から 2019 年 12 月 22 日までの 1053 日間にスマートフォンでの回答を前提とした Web アンケートとして「意外と熱海アンケート」を実施した。アンケート対象者は熱海市内を旅行する日本人観光客とし、Web アンケートサイトの URL を埋め込んだ三角 POP を、熱海温泉ホテル旅館協同組合に加盟する宿泊施設および調査協力を得られた飲食店に設置した。

アンケート内の「訪問地」に関する質問では、熱海市内の主要な観光施設や商業施設、駅など 52 箇所のスポットと、「その他（飲食店・施設等）」「その他日帰り温泉施設」「熱海海上花火大会」「ユーザによる自由回答」「無し」の合計 57 個を、カテゴリ別に一覧表示して選択肢とした。

本アンケートでは、「ふじのくに割アンケート」の行動データの回答フローを踏まえて、異常回答率の低減を狙って回答フローの改良を行った。実際の「意外と熱海アンケート」での行動データの回答画面を図 6 に示す。本アンケートでは、熱海に到着後最初に立ち寄ったスポットを選択し「次へ」をタップする（図 6 左上）。続いて立ち寄った時間帯を選択し「次へ」をタップする（図 6 右上）。続いて、次に立ち寄ったスポットを選択して「次へ」をタップする（図 6 左下）。このように、スポットの訪問順に 1 スポット 1 画面という一問一答方式の回答フローを採用した。スポットと訪問時刻を交互に聞くことで、スポットの訪問順とおおよその訪問時刻を合わせて収集することができ、滞在時間の推察も可能となる。回答画面の右下に配置された「旅程」ボタンを押すことで、その時点までに選択したスポットとその訪問時刻を選択順に図示し、入力の確認を可能とした（図 6 右下）。選択肢のうち「無し」を選択することでスポットの選択を打ち切り、それまでに選択したスポットと訪問時刻を行動データとして記録した。スポットの選択により訪問順が確定し、ユーザによる並べ替えを必要としない回答フローとすることにより、「ふじのくに割アンケート」における課題であった並べ替えのスキップによる異常回答の発生を抑止することが可能であると考えた。

「意外と熱海アンケート」では 840 件の回答が得られた。本研究では、選択スポット数が 2 箇所に満たない回答を除外するなどのデータクリーニングを行い、結果として 430 件の回答をデータセットとして用いる。遷移総数は 1036 回であり、データセット上に含まれる全ユーザの行動データを地図上に可視化したものを図 7 に示す。

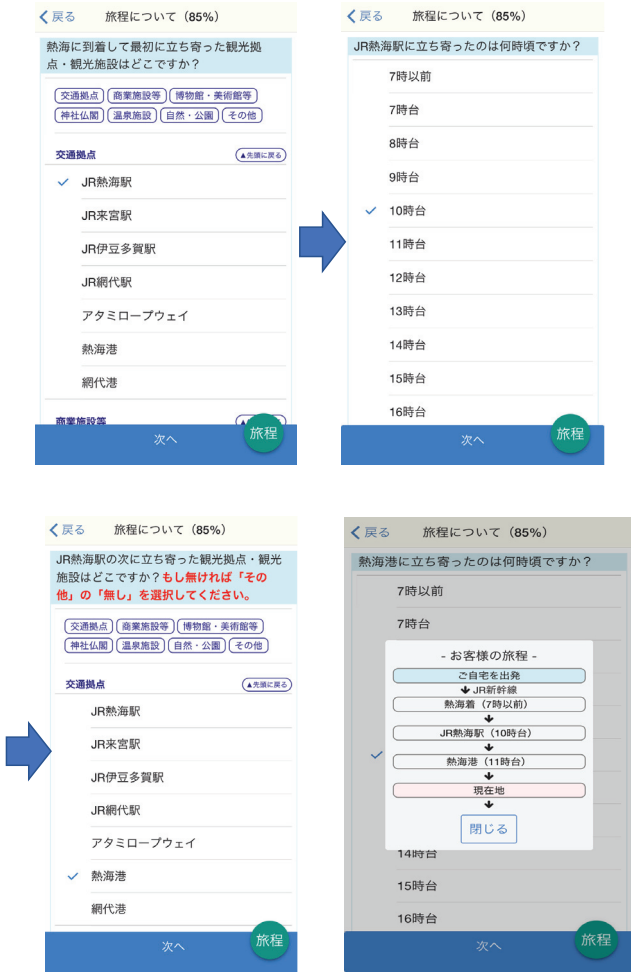

図 6「意外と熱海アンケート」行動データ回答画面

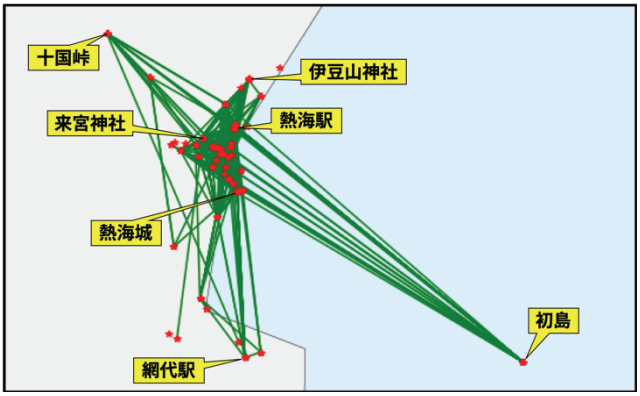

図 7「意外と熱海アンケート」全ユーザの行動データ

道路網データ

第 3 章で述べた回遊行動モデルについて、あるスポット r から別のスポット s に移動する際の最短距離 d(r,s) の計算は、Open Street Map から取得したノードの座標（交差点の緯度・経度）とノード間を結ぶリンク（道路）の長

さを用いて計算する[8]。具体的には、Open Street Map のデータ上に、それぞれのアンケートの観光スポットをノードとして追加し、追加したノードから最も近い交差点のノードまでリンクを結ぶ。以上の処理によって、ユーザが選択したスポットの間の最短距離を、ダイクストラ法を用いて求める。

6. パラメータ推定および異常スコア判定結果

3 種類のデータセットごとに算出された回遊行動モデルのパラメータおよび異常スコアの統計量を表 2 に示す。「ふじのくに割アンケート」については、エリア全域のデータセットと熱海市内を経由したユーザに限定したデータセットに分けて記載している。

各データセットにおけるユーザごとの異常スコアを図 8 から図 11 に示す。「ふじのくに割アンケート」では、異常スコアの平均値が 14.8 点（熱海市内を経由するユーザに限ると 21.1 点）であるのに対し、回答フローを改良した「意外と熱海アンケート」では 6.8 点と相対的に低いスコアとなった。標準偏差についても、「ふじのくに割アンケート」が 16.9（熱海市内を経由するユーザに限ると 23.4）であるのに対し、改良後の「意外と熱海アンケート」では 5.1 となり、異常スコアのばらつきが小さくなり、突出した異常回答が減少したものと考えられる。また、「対面調査」の異常スコアの平均値は 5.9 点、標準偏差は 3.9 となり、改良後の「意外と熱海アンケート」により近い結果が得られている。このことから、「意外と熱海アンケート」では、回答フローの改良によって「対面調査」に近い回答品質が得られていることを、異常スコアという統一的な指標によって定量的に示すことができたと考える。

表 2 パラメータと異常スコアの比較

	ふじのくに割（全域）	ふじのくに割（熱海）	意外と熱海	対面調査
有効行動データ件数	1259	71	430	456
有効遷移総数	4571	401	1002	1036
スポット数	1556	1556	52	24
人気度 $f(s) \geqq 1$ スポット数	347	104	45	24
尤度	-18656	-1517	-2934	-2710
距離パラメータ θ_1	0.999	0.778	0.396	0.306
人気度パラメータ θ_2	0.863	0.766	0.965	0.871
異常スコア平均	14.8	21.1	6.8	5.9
異常スコア標準偏差	16.9	23.4	5.1	3.9
異常スコア最大値	342.6	118.5	54.1	22.1
異常スコアが 20 点以上の行動データ割合(件)	20.5% (258)	29.6% (21)	2.3% (10)	0.7% (3)
目視異常回答率（異常回答件数）	-	16.9% (12)	0.9% (4)	0.0% (0)

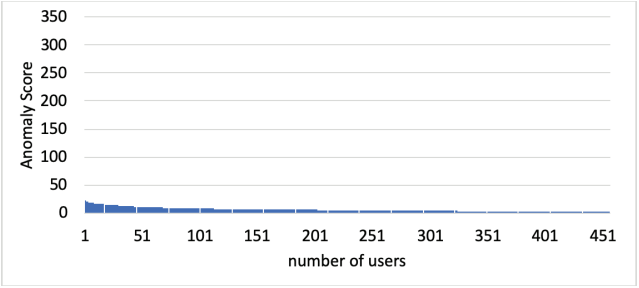

図 8「対面調査」ユーザの異常スコア

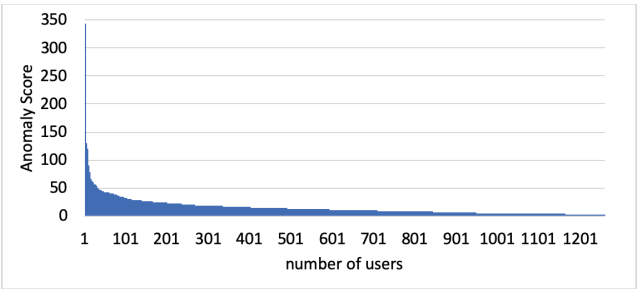

図 9「ふじのくに割アンケート」ユーザの異常スコア

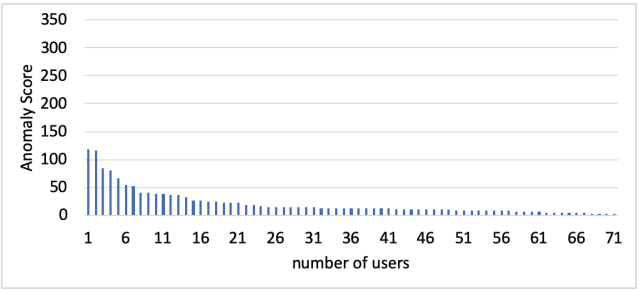

図 10「ふじのくに割アンケート」
熱海市を経由するユーザの異常スコア

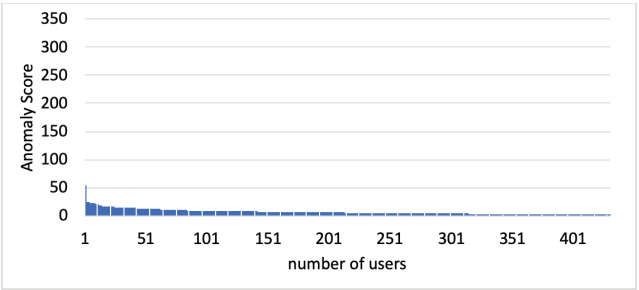

図 11「意外と熱海アンケート」ユーザの異常スコア

続いて、異常スコアの高いユーザについて、行動データを地図上に描画して目視で確認し、どの程度のスコアから異常回答と判定すべきかを考察する。本研究では、行動データを地図上に描画して目視で確認した際に、明らかに異常な回答であると著者が主観的に判断した行動データを目視異常回答とした。データセットに含まれる全ての行動デ

ータの件数を N、目視異常回答の件数を n とすると、目視異常回答率 R は次式で定義される。

$$R = \frac{n}{N} \tag{11}$$

まず、「ふじのくに割アンケート」では、異常スコアの最大値が 342.6 点（熱海市内を経由するユーザに限ると 118.5 点）と高く、20 点を超える回答件数が 258 件と有効行動データ全体の 20.5%（熱海市内を経由するユーザに限ると 21 件、有効行動データ全体の 29.6%）となった。熱海市を経由するユーザの中で最も異常スコアの高い行動データを地図上に描いたものを図 12 に示す。地図上で異常スコアの高い行動データを目視で確認すると、20 点を境に目視異常回答が増え始め、40 点台から異常と見られる行動データがさらに増え、80 点を超えた 10 件の回答については明らかに異常な選択が行われていた。

「ふじのくに割アンケート」は、異常スコアの高い行動データの件数が多いため、目視異常回答率の算出は熱海市内を経由するユーザのデータセットのみで行った。異常スコアが 20 点を超える行動データを確認すると、21 件の回答のうち目視異常回答であると判断したものが 12 件、目視異常回答ではないと判断したものが 9 件となった。よって、表 2 に示した通り、目視異常回答率は 16.9%となった。

続いて「対面調査」では、異常スコアの最大値が 22.1 点と低く、20 点を超える回答が 3 件となり有効行動データ全体の 0.7%となった。地図上で異常スコア上位の行動データを確認すると、いずれの回答も目視異常回答ではないと判断でき、異常回答率は 0.0%となった。

最後に、「意外と熱海アンケート」では、異常スコアの最大値は 54.1 点となり、20 点を超える回答は 10 件と有効行動データ全体の 2.3%となった。最も異常スコアの高い行動データを図 13 に示す。また、異常スコアが 30 点を超える回答は 1 件のみであった。地図上で異常スコア上位の行動データを確認すると、20 点以上の 10 件の回答のうち目視異常回答であると判断したものが 5 件、目視異常回答ではないと判断したものが 5 件となった。よって目視異常回答率は 0.9%となった。

異常スコアの計算結果と目視による地図上の行動データの確認によって、回答フローを改良した「意外と熱海アンケート」では、異常スコアが数値的に低くなっただけでなく、著者の主観的な判断ではあるものの、異常回答が抑制されていることが確認できた。

また、表 2 に示した異常スコアが 20 点以上の行動データの割合と、目視異常回答率から、異常回答と判定すべき異常スコアの目安は 20 点から 30 点の範囲と考えられるが、より詳細な検証は今後の課題としたい。

図 12「ふじのくに割アンケート（熱海市を経由する）」
において異常スコアの最も高いユーザの行動データ

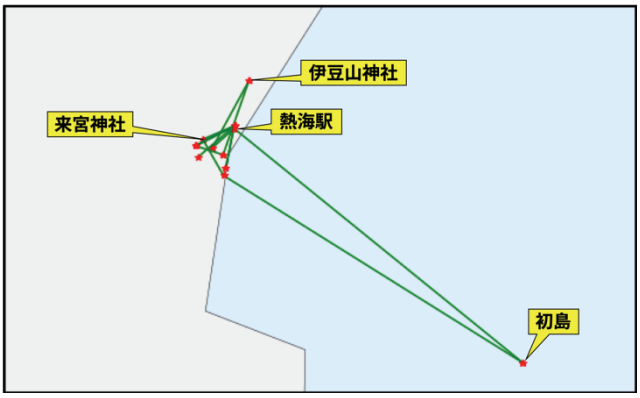

図 13「意外と熱海アンケート」において
異常スコアの最も高いユーザの行動データ

7. まとめ

本研究では、Web アンケートを用いて高精度な行動データの収集を可能とするために、その回答フローの良否を異常スコアによって定量的に評価できる手法について示した。具体的には、「ふじのくに割アンケート」から改良した行動データの回答フローを、「意外と熱海アンケート」に実装し、後者では異常スコアの平均値が確かに減少することを確認した。

本研究は、Web アンケートによって行動データを収集するための回答フロー（ユーザインタフェース等）の改良に活用することができる。異常スコアを数値指標として用いることで、異常回答をさらに抑制できる回答フローが実現できれば、GPS データを用いることなく、ユーザの多様な個人属性と紐付けられた高精度な回遊行動の分析を行うことが可能であると考えられる。実際に、本研究で改良を行った「意外と熱海アンケート」における行動データの回答フローは改善の余地が残されている。具体的には、回答の際に立ち寄ったスポットの数だけスポット選択のページ遷移が繰り返され手間がかかること、これにより回答の途中離脱率が増加する可能性が挙げられる。今後さらなる改良を行い、行動データの回答の容易さと異常回答の抑制の両

立を図った回答フローのデザインを目指したい。

謝辞

本研究の調査には、株式会社 JTB 静岡支店にご支援いただいた。ここに深謝する。

参考文献

1) “平成 27 年度 ICT を活用した訪日外国人観光動態調査報告書”, 国土交通省 観光庁 観光地域振興課, 2016.
2) 矢部, 有馬, 岡村, 角野, “GPS を用いた観光行動調査の課題と分析手法の検討”, 観光科学研究 第 3 号, 2010.
3) 相, “観光研究への位置情報ビッグデータ展開の可能性”, 観光科学研究 第 7 号, 2014.
4) 渡邉, 長島, 大石, 湯瀬, 武藤, 大久保, 木村, “観光行動分析のための Web アンケートシステム”, 第 14 回観光情報学会全国大会, 2017.
5) 鈴木, 斉藤, 風間, “最尤推定に基づく回遊行動モデリング”, ネットワークが創発する知能研究会, 2015
6) 大石, 鈴木, 斉藤, 渡邉, “アンケート調査による観光スポット遷移データからの異常回答検知”, 第 13 回ネットワーク生態学シンポジウム, 2016.
7) Tandon, Chan, “Tracking User Mobility to Detect Suspicious Behavior”, Proceedings of the 2009 SIAM International Conference on Data Mining, 2009.
8) Open Street Map, https://www.openstreetmap.org/