レビューやSNSにおける位置情報付き投稿と地図上の施設や用地の比較分析と可視化

松下　将也† 　楊　 子正† 　荘司　慶行† 　Martin J. Dürst†

† 青山学院大学理工学部情報テクノロジー学科　〒 252–5258 神奈川県相模原市中央区淵野辺 5–10–1
E-mail: †{matsushita,yang}@sw.it.aoyama.ac.jp, ††{shoji,duerst}@it.aoyama.ac.jp

あらまし　本論文では，ソーシャルメディアやレビューサイトから収集した位置情報付き投稿と，地図上の施設情報を，個人の消費行動に紐づけて分析する．Tripadvisor などのレビューや，Twitter などの位置情報付きソーシャルメディア投稿では，その場所で利用者がどんな行動をとったかに関する情報が含まれる．本研究では，これらの行動を「何かを購入した」「サービスを受けた」「何かを食べた」「宿泊した」の 4 種類に分類して，位置情報と紐づけた．これらの消費行動を国土地理院が公開している用地情報や OSM（Open Street Map）の地物情報と比較することで，ソーシャルメディアが現実の地物をどう反映しているかを分析した．

キーワード　情報検索，Twitter，地理情報検索

1　はじめに

近年，位置情報機能付きのモバイル端末や，それを受け入れる様々なオンラインサービスが普及したことにより，Web 上に高精度な位置情報付きの情報が溢れかえってきている．ソーシャルメディアや，オープンソースサイトに，だれもが気軽に，地物に関する情報を投稿できるようになっている．このようなユーザが投稿した地物に関する情報は，旧来の，国家や地図作成会社の作成したデータと，データの性質や信頼性に，大きな差があると考えられる．

例えば，地域の「雰囲気」などは，地図製作会社の作成した正式な地図には反映されない．具体例として，同じ飲食店が立ち並ぶ地域でも，サラリーマンの町である新橋の飲食店街と，若者の多い渋谷の飲食店街，新宿歌舞伎町では，その性質が異なると考えられる．このような情報は,「飲食店が何件ある」という公的な情報とは別に，そこで実際に何が行われているか，だれがそこを使っているかなど，利用者目線での情報が必要になる．こういった情報は，位置情報付きのソーシャルメディア投稿や，地物レビューであれば，含まれている可能性が高い．

一方で，データの信頼性という面を考えると，政府や企業が作成したデータと比較して，ソーシャルメディアやレビューサイトの投稿は，信頼に劣ると予想される．第一に，ソーシャルメディア上の位置情報付き投稿にはノイズが多い．例えば，マイクロブログの位置情報つき投稿は，その投稿が場所に関係するとは限らない．普段から位置情報を付与する設定にしている利用者がいた場合に，友人との会話などの，日常的な内容までもが位置情報に紐づけられてしまう．第二に，だれでも自由に投稿できるサイトの情報は，利用者の悪意や間違いに弱い．具体例として，ある飲食店や施設に対して，攻撃的なレビューやサクラ行為が一般的に行われることが指摘されている．また，GPS の精度が十分でなかったり，投稿者の勘違いで，実際に意図した地物とは別の地物に投稿が紐づけられる場合もありうる．

そこで，本研究では，一般市民の消費行動に注目することで，様々なデータソースにおける，地物情報に関する性質の違いを分析可能にする．具体的には，落語 寿限無で言うように「くうねるところにすむところ」，糸井重里による日産セフィーロのキャッチコピーである「くう，ねる，あそぶ」に代表されるような，人間の根源的な消費行動を，地物に関連させて集計する．

そのために，まず，ソーシャルメディアやオープンソースサイト，公的機関から地物情報に関連するデータを収集した．具体的には，

- Twitter の位置情報付き投稿,
- Trip Advisor のレビュー,
- Open Street Map の店舗情報,
- 国土地理院の用地情報

を，それぞれ東京都と神奈川県に関して収集した．

次に，Twitter の位置情報付き投稿や，Trip Adovisor のレビュー中に含まれるそれぞれの文について,「食べる」,「宿泊する」,「遊ぶ」,「購入する」の 4 種類に分類した．分類のために，BERT の言語モデルを用いた．言語モデルは，このタスクに合わせて，ファインチューニングを施した．

こうして得られた各データについて，地域をグリッドで区分けして，各セルにどの程度含まれるかを集計した．そして，地域ごとに，それぞれに関する投稿や，それぞれの施設が，どの程度含まれるかを比較した．このように，セルごとの各トピックの投稿数や，施設数を比較することで，たとえば「飲食店が多い地域では，食に関する位置情報付き Tweet が多くなる」などの直感的な考えが，実際に正しいものであるかを検証可能になる．

具体的に、集計した結果を可視化したものを、図 1 に示す。この図では、新橋駅周辺のそれぞれの異なるデータソースにおける投稿の出現比率を、シアン、マゼンタ、イエローに割り当てて可視化している。このようなインタフェース上で、様々な

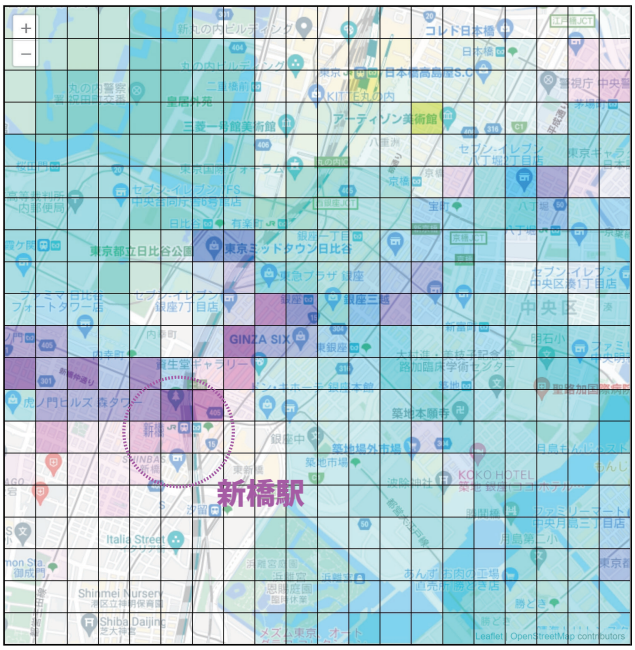

図 1　「食」に関する Tweet、「買」に関する Tripadvisor の新橋駅周辺での比較

データソースに表れた消費に関する情報を、それぞれの色に割り当てて比較することで、データソースの性質や、地域の性質について分析可能にできる。

以下に，本論文の構成について述べる．本章では，研究の背景と目的について説明した．第 2 章では，目的や使用技術の類似した研究分野として，位置情報付きツイートからの情報，地物レビューサイトにおけるレビュー内容の分類についてそれぞれ紹介して論じる．第 3 章では本研究で提案する手法を述べる．第 4 章では実際に収集したデータを使った分析とその結果について説明し，第 5 章では結果をもとに分析手法について考察する．第 6 章では，本研究の結論と，今後の展望について述べる．

2　関 連 研 究

本研究では，位置情報付きのソーシャルメディア情報や，位置情報付きのオープンデータを分析する．これは，広義の位置情報に基づくソーシャルセンシングの一部である．また，本研究では地物に対するレビューを分析する．これらの観点から既存の関連研究を整理し，本研究の位置付けを示す．

2.1　位置情報付きデータの分析

一方で，ソーシャルメディアにおける位置情報付き投稿には，さまざまな偏りがあることも指摘されている．例として Malik ら [1] は，Twitter 投稿に人口のバイアスがかかっていることを指摘している．また，位置情報付きツイートにはノイズが含まれていることが多く，内容的にユーザの位置と関係のないツイートが存在することを Hiruta ら [2] は指摘している．本研究でも，ソーシャルメディア投稿の分野や，オープンソースにおける偏りについて分析することを目的の一つとしている．

本研究著同様に，位置情報に関連するソーシャルメディア上の投稿のデータセットごとの差に注目した研究もいくつか存在する．例として，Lei ら [3] は Tripadvisor のレビュー内容や評点が，英語，オランダ語，イタリア語でどのように異なるかを分析している．本研究では日本語でデータセット間での差異を分析するが，地域や言語差も本来であれば考慮すべきである．また Jurgens [4] らは，Tweet から位置を推定するタスクにおけるデータセットにおいて，その規模によって制度が大きく変わる現象を指摘している．本研究においては，GPS および付与された座標情報を用いるが，地域の消費動向を測るソーシャルセンシングとして考えた際に，地域内でのソーシャルメディア投稿の総数や施設数などが分析精度に影響することが考えられる．

Quan ら [5] は，旅行体験の分析の際に，満足度などの表層的な特徴ではなく，食事，宿泊，交通などのそれぞれの要素に注目する必要性を指摘している．本研究でもこれに倣い，投稿数やレビューの星の数などではなく，消費形態に注目している．

2.2　位置情報に基づくソーシャルセンシング

位置情報付きのソーシャルメディア投稿を利用して現実世界の動向を分析する研究は古くからソーシャルセンシングの一分野として扱われる [6]．本研究では，位置情報付きの情報を，ある地域でどのような消費が行われているかを推定するために用いている．多くの研究で，様々なものを推定し発見するために，位置情報つきの投稿が用いられている．例として，Gurajala ら [7] は，Twitter 上の投稿から，その地域の大気の汚染度を推定している．また，本研究と類似した例として，Nguyen ら [8] は Tweet を健康や幸福，飲食などのトピックに分類することで，地域ごとに人の生活を分析している．

位置情報付きのソーシャルメディア投稿を用いたソーシャルセンシングの代表的な研究例として，地域の人口推定があげられる．例として，Sloan ら [9] は，Twitter における位置情報サービスを有効にしている人とそうでない人との人口統計学的な差異を探っている．また，観光地や飲食店などの，地図上の地点（POI：Point of Interest）を発見するタスクも，ソーシャルセンシングでは一般的である．例として，Gao ら [10] は FourSquare などの LBSN（Location-based Social Network）サービス上で，単にどこにいたという意味のチェックインだけでなく，チェックイン先のメタデータを用いることで，より高精度に POI を発見できることを示している．地物の使われ方に注目した POI の研究として，Debnath ら [11] は利用者と POI をカテゴリ，利用時間，場所の嗜好，POI の人気度という観点に分解して紐づけ，推薦に利用している．本研究では，メタデータやカテゴリなどではなく，独自に推定した消費形態に基づいて地域を分析する．近年では，このようなソーシャルメディアからの POI 発見にもディープラーニング技術が一般的に用いられるようになってきている [12]．本研究でも，レビューや Tweet 本文のテキスト分析に，BERT というニューラルネッ

トワーク技術を用いている．

本研究では，ソーシャルメディアやレビューから，ある地域でどうお金が使われているかと，そのデータセットへの現れ方がデータセットごとでどう異なるかを論じることを目的としている．本研究が正しくデータセットの特性を導き出せれば，これらの既存のソーシャルセンシングに関する研究において，それぞれのセンシングに適したデータセットが定量的に検討できるようになる可能性がある．

2.3 地物レビューの分析

本研究では，地物レビューをもとに地域とその消費傾向を分析しているが，地物レビューの分析も盛んに研究が進んでいる分野である．

地物レビューの重要性を示す例として，Xiang ら [13] は，人々が旅行を計画する際にウェブ上で情報を探した際に，検索結果の多くが公的な情報ではなく投稿レビューサイトであることを指摘している．

3 提 案 手 法

本研究では，ある場所がどのように用いられているかを表す情報が，データセットごとにどう異なって表れているかを分析する方法を提案している．そのために，ソーシャルメディアやレビューサイト上の投稿内容が,「食べる」,「宿泊する」,「遊ぶ」,「購入する」のうちどれかに属しているかを分類し，位置情報と紐づける．そして，ある地域におけるそれぞれに関する投稿の量が，オープンソースや日本政府の提供している地図情報と，どう関係を持つかを分析可能にする．

3.1 データの収集と整形

はじめに分析対象とするデータを収集し、位置情報によってそれらを一元的に扱えるようにした．ソーシャルメディア，レビューサイト、オープンソース情報，そして国土交通省が提供する用地データについて、実際に分析対象として収集した．

ソーシャルメディアには，情報を発信する際に位置情報を付与できる機能が存在している．この時の位置情報は緯度と経度で示されており，また，この経緯度情報は 1 つの点として表される．さらに，ソーシャルメディアから発信される情報をこの経緯度情報と組み合わせることで，位置情報を持つソーシャルメディアを整形することができる．

レビューサイトの多くは，マップ API によってサイト上に地物の位置を表示している．そして API は多くの場合，マップの URL を示しており，その URL に経緯度情報が付与されている．従ってサイト内のレビューと URL を取得し，URL から経緯度情報を抽出し，レビュー内容と組み合わせることで，位置情報付きのデータを取得することが可能となる．

オープンソースマップは，登録されているポイントごとに経緯度情報とタグが付与されている．タグはポイントの特徴で分類されており，タグと経緯度を組み合わせることで，位置情報付きのデータを取得することができる．また，国土交通省の用地データは，メッシュデータという形で配布されており，中には地域名称，地域の用地情報，用地における面積，経緯度情報などが含まれているため，このまま使用することができる．

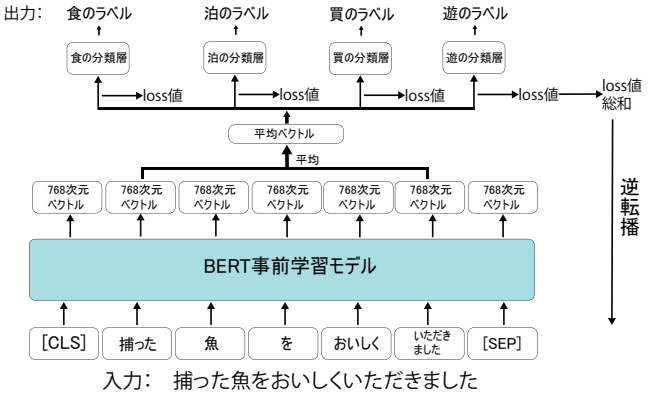

図 2 BERT によるファインチューニング

3.2 位置情報付きデータの分類

ソーシャルメディアとレビューサイトから取得できる情報は，オープンソースマップや国土交通省が提供している用地データから得られる情報と違い，どの消費行動に分類されているのかを示す情報が存在しない．そのため，データをあらかじめ分類しておかなければ，比較を行うことは不可能である．そこで本研究では，BERT によるファインチューニングを行うことで分類器を作成し，データの分類を実現可能にする．

図 2 は，BERT によるファインチューニングの流れを表している．まずは，収集したデータからレビュー内容の一文を取り出し,「食」、「泊」、「買」、「遊」という 4 種類の消費行動に分類するためのラベル付けを行い，学習データの作成を行った．これを BERT の入力とする．BERT の事前学習モデルに収集済みデータを入れる前作業として，形態素解析を用いてレビューの解析を行い，BERT 専用の特殊トークンである [CLS] と [SEP] を付与した．

これをモデルによるベクトル化を行うことで，特殊トークンが付与されたベクトル以外のベクトルの平均値を取得することを可能にした．そして，この平均ベクトルで「食」「泊」「買」「遊」という 4 種類の消費行動が表すラベルの分類層を通して，予測ラベルを出力させた．出力したラベルが入力とは異なっている場合も存在する．その場合はこの 2 つのずれを loss 値とした．そして，4 つの分類の loss 値の総和をモデルに逆伝播し，モデルのバイアスを変更し，BERT のファインチューニングを完成させることで，本研究用に用いる分類器の作成が可能となった．

3.3 情報の比較の前処理と分析手順

分類した情報から，位置ごとに一部抽出を行い，それを地図上のグリッドに表示する．まず始めに，分析を行いたい地域の経緯度情報を探す．その地域の左上と右下の座標を経緯度情報から決定し，分析を行う地域を決定する．その際にグリッドの周りの座標や，グリッド内のセル 1 つ 1 つの大きさも決定する．

次に，整形したデータの中から，緯度経度情報をもっている

データの処理を最初に行う．そのデータに含まれる経緯度情報が分析を行う地域のグリッドの範囲内に含まれていた場合はデータを取り出す．その際にグリッド内のどのセルに入るかの計算もあらかじめ行う．ここで，ソーシャルメディアとレビューサイトのデータ**??**章で作成した分類器を用いて分類を行い，分析用のデータセットを作成する．また，国家機関の情報には緯度経度情報が含まれていないため，そのまま使用する．

データセットを用意した上で，まず始めに緯度経度情報を持つデータセットの比較を行う．同じグリッド内のセルで数個の緯度経度情報が入っており，且つこのグリッドを行列とすると，ピアソンの相関係数

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

を用いて，2 つのデータセット A，B の相関度

$$r_{AB} = \frac{\sum_{i=1}^{n}(A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum_{i=1}^{n}(A_i - \bar{A})^2}\sqrt{\sum_{i=1}^{n}(B_i - \bar{B})^2}} \quad (2)$$

を計算する．

国土地理院の配布している用地情報は，緯度経度からなる点座標ではなく、ポリゴン形式で表される地域メッシュである。そのため，任意のデータセット内の地物と国土地理院のメッシュを比較する際には、あるセル内に含まれるそれぞれの地物が、メッシュ内にどれだけ含まれるかをそれぞれのセルの値として、その大小によって相関を計算した。

4　実際のデータを用いた分析

本節では，実際に収集したデータをもとに，それぞれのデータを比較し，相関のあるデータどうしを発見する．そして，相関のあるデータセット間で，外れ値的な地域を探し，その特徴について人手で分析し、考察する．

4.1　データセット

はじめに，実際に分析するために，東京都内と神奈川県内の各位置情報付きの情報源のデータを収集し，データセットを作成した．データセットの詳細を表 1 に示す．

位置情報付きのソーシャルメディア投稿のデータとして，Twitter の位置情報付き Tweet を収集した．Tweet は 2018 年から 2019 年に投稿された Tweet で，緯度経度情報から東京都内と神奈川県内を被覆する矩形エリア内のもののみを抽出した．注意事項として，Twitter は 2019 年 6 月以降，Tweet に詳細な緯度経度による位置情報付与を廃止しており，現在では大まかな市区町村のみが利用可能である．そのため今回の実験では，2019 年までの Tweet のみを用いた．

地物レビューサイトの情報として，Tripadvisor[1]のレビュー投稿を収集した．実際の収集の際には，地域のカテゴリツリーを利用して，東京都内の全市区と，神奈川県の全市区について，そこに含まれる地物のメタデータとレビューを収集した．

表 1　分析に使用した東京都内と神奈川における各データセットのデータ数

収集元	データ	件数
Trip Advisor	ホテル レビュー	100,510
	観光地 レビュー	414,283
	レストラン レビュー	522,317
Twitter	日本語 都内	11,014,285
Open Street Map	地点数（Point 型要素）	2,344,067
国土地理院	用地情報メッシュ数	11,177

Tripadvisor ではレビューがホテル，レストラン，観光地に分かれているため，それぞれに対してレビューを個別に収集した．今回の分析では，それぞれの地物に対するレビューを文単位に分割して，文ごとに分類機で分類して用いている．つまり,「ホテルに対するレビュー中の文であるので，宿泊に関する記述である」というような判定はしないようにしている．これは，例えば，ホテルであれば結婚式などのイベントに使われることがあり，その際には食事に関する感想も，結婚式に関するサービスへの言及もレビューに含まれるためである．

OSM（Open Street Map）からは，東京都と神奈川県内に含まれるすべての Node（地点）型のエントリを収集した．OSM におけるエントリは，道などの長さを持つ線形式である Way，公園などの面積を持つポリゴン形式である Area，商店などの点形式である Node のデータが存在する．実際の Node データは，その地物のカテゴリと詳細が，メタデータとして登録されている [2]．メタデータは大分類と詳細のペアで表され，例えばハンバーガーショップであれば「amenity=fast_food」のように，amenity（生活施設）という大分類と fast_food（ファストフード店）という詳細を持つ．本分析では，このメタデータの大分類をその施設の種別として，実際の分析に用いる．

4.2　データセット間の相関分析

はじめに，本節では，各データセット間の相関を計算する．具体的には，ある地域内において，地域を複数のセルに区切った際に，2 つのデータセットにおいてそこに含まれる要素の数が相関を持つかどうかを総当たりで計算する．例えば，OSM データでレストランの多いセルでは，同様に食に関するレビューや食に関する Tweet も増えることが予想される．また，観光スポットには飲食店が併設されがちなので,「遊ぶ」と「食べる」の Tweet にも相関が現れるかもしれない．このようなことを，地域ごとに表としてまとめ，相関のあったデータセットペアどうしをランキングとして示した上で．分析を行う．図 3 は，OSM 上の施設の種類，Twitter 上での消費行動ごとの Tweet，Tripadvisor 上の消費行動ごとの文の数の，東京都内での場所ごとの相関の可視化結果を示している．またここでの赤いセルは，同じエリア内では 2 つのデータセットが似た値を取ることを示す。東京都内に該当のないカテゴリは相関を 0 として扱う．

表 2 は，それぞれのデータセット間の相関分析を行った際に，

1：https://www.Tripadvisor.jp/

2：OpenStreetMap Wiki: Map features
https://wiki.openstreetmap.org/wiki/Map_features

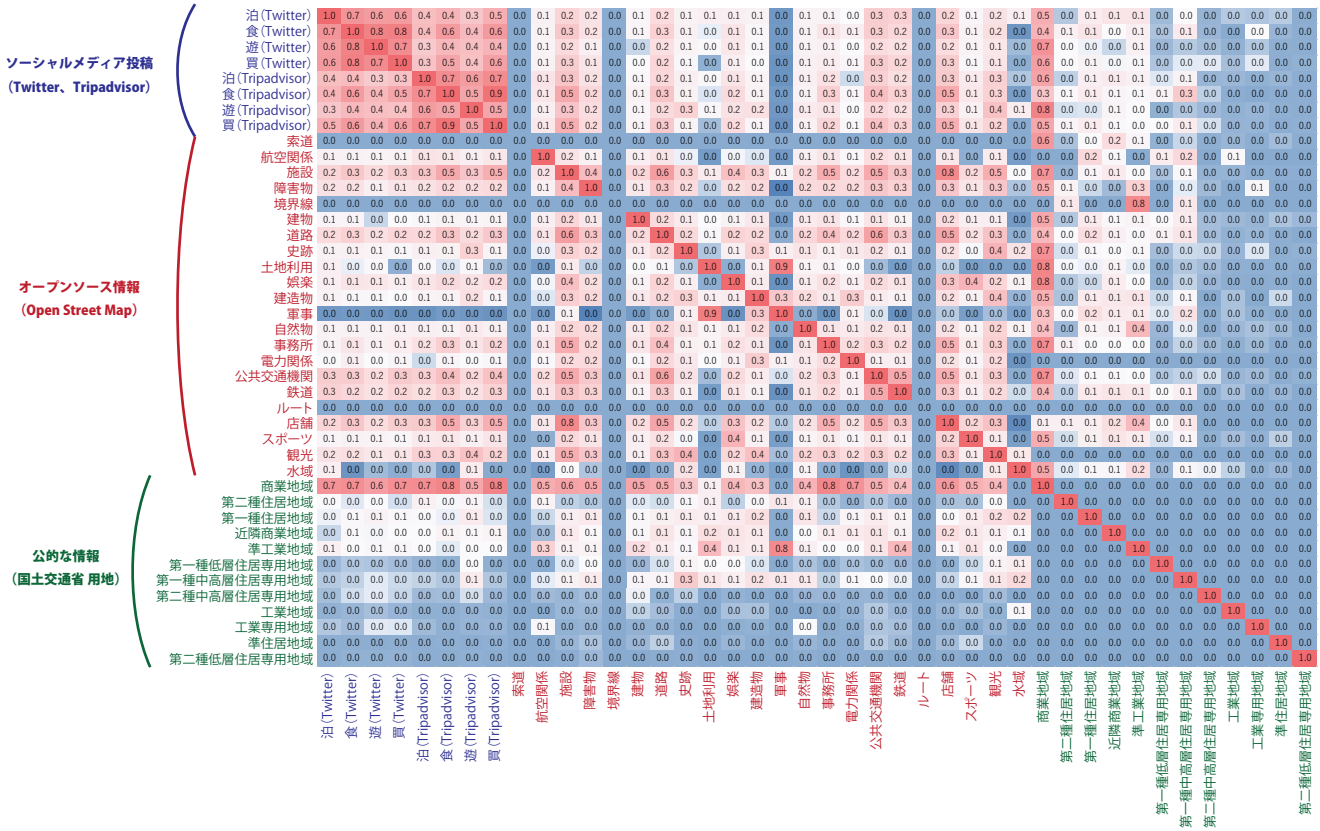

図 3　東京都内での場所ごとの相関の可視化結果

相関が強いデータどうしを順番にランキング付けを行い，表にまとめたものである．この結果から，Tripadvisor に掲載されている地物の中で「食」に関するレビューが多い地物には，「食」と同様に「買」に関するレビューも多く書かれていることが分かる．これは Tripadvisor に掲載されている「食」に関する地物，つまりレストランやカフェのような施設の付近には，食べ物をその場で食べるだけでなく，食べ物を購入して持ち帰ることができたり，その店舗に関するグッズを購入できる店舗が多いことが考えられる．同様に，レストランやカフェのような「食」に関する施設が集まっている場所には，買い物ができる施設や観光名所とされている場所が集まっていることが，Tripadvisor のデータだけでなく Twitter のデータからも分かる．

また，国土交通省が提供している用地データのうち商業地域として定められている場所に対して，Tripadvisor 上の「遊」に関するレビューや，Twitter 上の「遊」,「買」に関する Tweet が多いことから，国が商業地域に定めている場所には観光名所や買い物ができる施設が多く存在するということが分かる．

4.3　データセットを跨いだ消費行動の可視化インタフェース

実際に、複数のデータセットにおける、4 種類の消費動向を、シアン、マゼンタ、イエローの各色に割り当てて、地図上に可視化できるウェブアプリケーションを実装した。黒字で地物名の表示される地図上で混色した際に見やすいように、CMYK 色空間で可視化するように実装した。図 1 に実際のウェブアプリケーションのスクリーンショットを示す。このシステムでは，可視化したい地域と使用するデータセット，そして可視化する際の色の見え方を自身で選択可能である。例えば,「食」に関する Tweet をシアン,「買」に関する Tripadvisor のレビュー投稿をマゼンタにし，場所を新橋駅周辺に設定して実行すると，図 1 のように表示される．すると，新橋駅周辺から有楽町駅，日比谷駅，銀座駅に向かって，特に駅周辺には飲食店が数多く存在することが分かる．また，東京タワー付近ではシアンとマゼンタの混色である青に近い色が強く出ており，この場所では飲食店と同様に買い物ができる店舗が集中していることが分かる．

データセットと色の割り当ては，図 4 のように自由に選択できるようにしている．

4.4　地域ごとの外れ値分析

地域ごとの外れ値に注目することで，実際にこのような分析が何に使えるかについても議論する．

図 5 は，データセットに「遊」に関する Tweet と「遊」に関する Tripadvisor のレビュー投稿，そして OSM 施設情報を用い，色をそれぞれシアン，マゼンタ，イエローにした時のディズニーランド周辺の可視化である．Twitter と Tripadvisor の情報を比較して場所がずれていることから，Tripadvisor 上の位置情報は敷地内の一点で定められているが，Twitter の位置情報は敷地内でも細かく分かれていることが分かる．

また，図 6 は，データセットに OSM の娯楽と「遊」に関する Tweet を用い，色をそれぞれシアン，マゼンタにした時の歌舞伎町周辺の消費行動を可視化したものである．OSM に登録

表 2　分析に使用したデータセット間の相関ランキング

順位	データセット 1	データセット 2
1 位	食 (Tripadvisor)	買 (Tripadvisor)
2 位	土地利用 (OSM)	軍事 (OSM)
3 位	食 (Twitter)	買 (Twitter)
4 位	食 (Twitter)	遊 (Twitter)
5 位	商業地域 (国土交通省)	遊 (Tripadvisor)
6 位	商業地域 (国土交通省)	土地利用 (OSM)
7 位	商業地域 (国土交通省)	娯楽 (OSM)
8 位	店舗 (OSM)	施設 (OSM)
9 位	準工業地域 (国土交通省)	境界線 (OSM)
10 位	買 (Tripadvisor)	泊 (Tripadvisor)
11 位	買 (Twitter)	遊 (Twitter)
12 位	食 (Tripadvisor)	泊 (Tripadvisor)
13 位	商業地域 (国土交通省)	事務所 (OSM)
14 位	商業地域 (国土交通省)	公共交通機関 (OSM)
15 位	商業地域 (国土交通省)	史跡 (OSM)
16 位	商業地域 (国土交通省)	施設 (OSM)
17 位	商業地域 (国土交通省)	遊 (Twitter)
18 位	食 (Twitter)	泊 (Twitter)
19 位	商業地域 (国土交通省)	索道 (OSM)
20 位	買 (Tripadvisor)	食 (Twitter)
21 位	買 (Twitter)	泊 (Twitter)
22 位	食 (Tripadvisor)	食 (Twitter)
23 位	遊 (Tripadvisor)	泊 (Tripadvisor)
24 位	遊 (Twitter)	泊 (Twitter)
25 位	道路 (OSM)	施設 (OSM)
26 位	商業地域 (国土交通省)	買 (Twitter)
27 位	買 (Tripadvisor)	買 (Twitter)
28 位	公共交通機関 (OSM)	道路 (OSM)
29 位	商業地域 (国土交通省)	泊 (Tripadvisor)
30 位	食 (Tripadvisor)	買 (Twitter)

されている娯楽施設が少ないにもかかわらず，遊んだことを示唆する Tweet が多くされていることから，歌舞伎町には，公的ではない娯楽施設が集まっているのではないかということが分かる．

図 7 は，データセットを「泊」に関する Tweet にしたときの青山学院大学相模原キャンパス周辺の消費行動である．図 7 と同じ地域で，データセットを Tripadvisor に掲載されている宿泊施設のレビュー投稿にした時は，色がついているグリッドは表示されなかった．このことから，Tripadvisor のようなレビューサイトには登録されていない隠れた宿泊施設が存在することや，そもそも宿泊施設ではない場所で寝泊まりしている人が存在するのではないかということが考えられる．

図 8 は，OSM の店舗情報をシアン,「食」に関する Tweet をマゼンタ，そして OSM の施設情報をイエローにした時の青山学院大学相模原キャンパス周辺の消費行動である．図 8 を見ると，OSM で飲食店として登録されていないのにもかかわらず「食」に関する Tweet が多いエリアが存在していることが分かり，そのエリアには，OSM 上からは判断できないが飲食ができる場所，もしくはできたばかりの新しい飲食店が存在するということが分かる．

図 4　実際に実装したウェブアプリケーション上でのデータソースと色の選択画面

5　考　　察

本節では，得られた実験結果をもとに，実際にデータセット間の差の特徴と，実際にこの分析手法がどのように用いることができるかについて論じる．

まず，国土交通省の用地データで定められている商業地域や，OSM で娯楽施設として登録されている施設が多く存在するエリアには,「食」に関する Tweet が多かったり Tripadvisor のレストランに関するレビューが多かった。このため、「飲食店が多い地域では，食に関する Tweet が多くなる」という直感的な考えは，実際に正しい可能性が示唆される。しかし,「食」に関する Tripadvisor のレビュー投稿と「食」に関する Tweet を比較すると，相関ランキングを示した際に相関は低くなっていることも分かった．このことから，Tripadvisor に掲載されている「食」に関する施設は店内で食事ができる飲食店が多いことに対して，Twitter ではコンビニで食べ物を買って他の場所で食べた，などのように購入した場所とは別の場所で食事に関わる Tweet をしているケースが多いからだと考えられる．

また，データセットを「遊」に関する Tweet と「遊」に関する Tripadvisor のレビュー投稿にして地域ごとに可視化すると，それぞれの可視化されたエリアにずれが生じているケースも

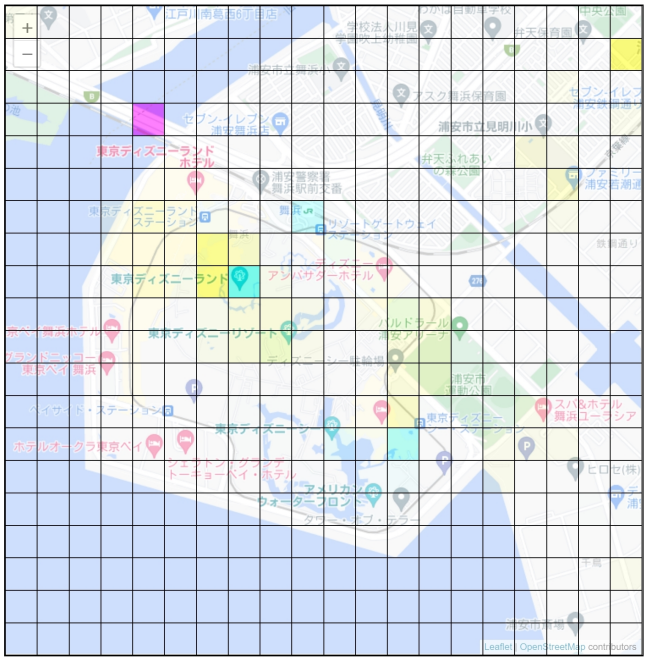

図 5　「遊」に関する Tweet をシアン,「遊」に関する Tripadvisor のレビュー投稿をマゼンタ，OSM 施設をイエローにした時のディズニーランド周辺の可視化

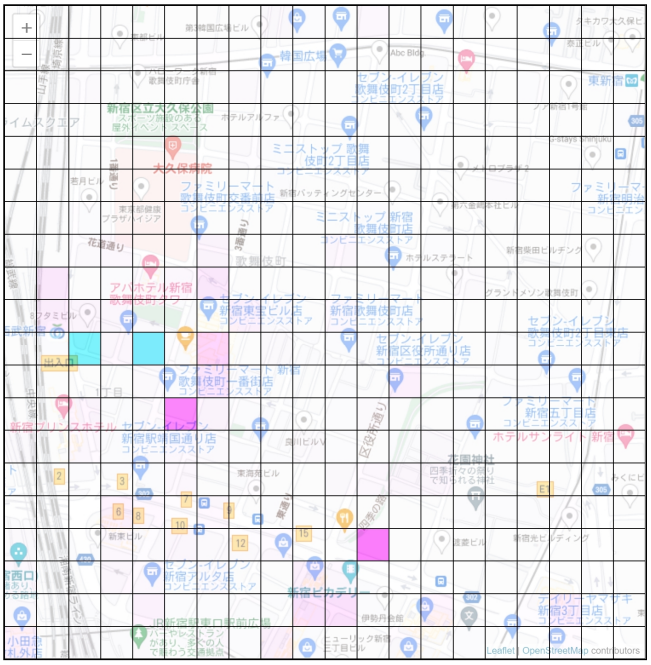

図 6　OSM 娯楽をシアン,「遊」に関する Tweet をマゼンタにした時の歌舞伎町周辺の娯楽情報

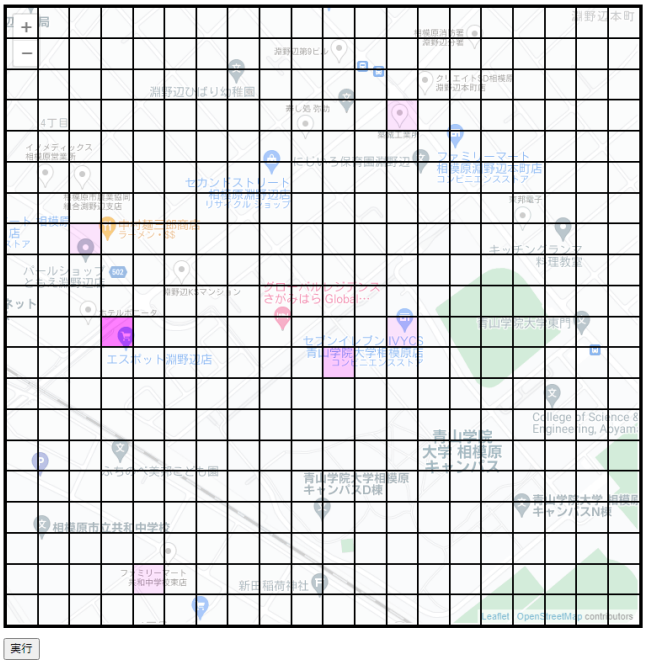

図 7　「泊」に関する Tweet をマゼンタにした時の相模原キャンパス周辺の宿泊情報

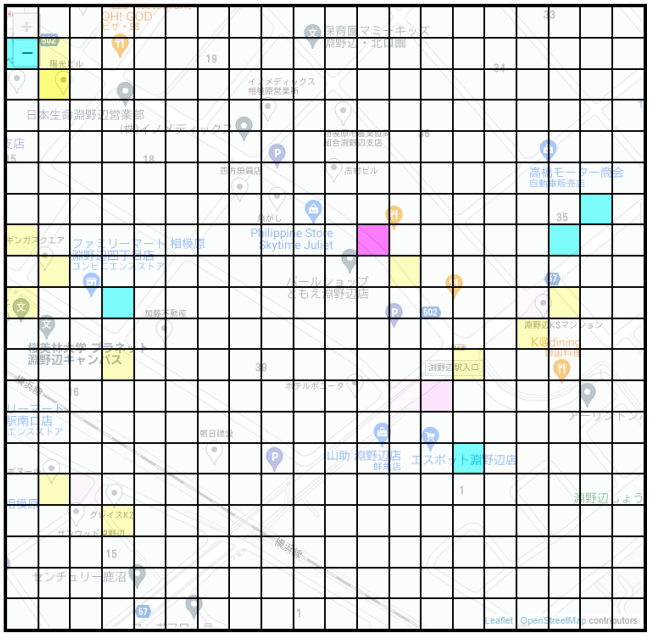

図 8　OSM 店舗をシアン，OSM 施設をイエロー,「食」に関する Tweet をマゼンタにした時の青山学院大学相模原キャンパス周辺の可視化

あった．これは，飲食店のような狭い施設と，ディズニーランドのような広大な娯楽施設における位置情報の決められ方の違いによるものだと考えられる．実際に，Tripadvisor 上では位置情報が施設内の一点に定められていることに対して，Twitter の位置情報は施設内でも細かく分けられているのである．

次に，図 6 より，OSM に登録されているような公的な娯楽施設が存在しないにもかかわらず，遊びに利用されている場所が存在することが Twitter 情報より分かった．このような地域には非公的な娯楽施設が存在したり，猥雑な店舗やサービスが行われていたりするような娯楽施設が集まっているのではないかと考えられる．また，OSM のような公的なデータよりも Twitter のような SNS の情報からの方が，公的であるかどうかに捉われることなく，消費行動を広く得られることが分かる．

そして，図 7 より，Tripadvisor 上では宿泊施設が存在しないものの，Twitter の情報によると宿泊に利用されている場所が存在することが読み取れた．またその地点を詳しく調べてみ

ると，青山学院大学相模原キャンパスの敷地内であることが分かった．このことから，青山学院大学の学生寮に住んでいる学生が自室から Tweet していることや，研究で忙しい教授や学生が自宅に帰ることができず，研究室で寝泊まりしているのではないかということが考えられる．また，図 8 より，青山学院大学相模原キャンパスの周辺では，OSM 上では店舗として登録されていないが,「食」に関する Tweet が多くされているエリアが存在することが分かった．実際にそのエリアを詳しく見てみると，お昼時には並ぶ人ができる程に人気のあるラーメン屋が存在することが分かった．このように，外れ値とされる地域には，知る人ぞ知る隠れた名店が存在することが多いのではないかと考えられる．

6 まとめと今後の課題

本研究では，ユーザ投稿型サイト，オープンソースの地理情報，公的な地理情報をそれぞれ収集し，消費行動に注目してデータセット間の比較と分析を行った．実際に，Tripadvisor と Twitter における地物に関連する投稿，Open Street Map における地物のメタデータ，国土地理院の用地情報を収集し，地図上の座標と紐づけて比較可能にした．分析のために，レビューとソーシャルメディア投稿を，消費行動に注目して,「食べる」「寝る」「遊ぶ」「泊まる」に分類した．分類には分類タスクに特化するようファインチューニングした BERT モデルを用いた．

今回の実験では，あくまでも初歩的なデータセット間の相関分析にとどまった．今後は，より大規模なデータセット間の可視化を行い，実際の有識者評価を通じて，正しく地域の消費行動を抽出できているかを評価したい．また，これらの分析結果が，実際に出店計画や政策などのための地域の分析に用いることができるか，有用性に関する質的評価を行いたい．

謝　　辞

本研究は JSPS 科研費 18K18161（代表:荘司慶行），21H03775（代表：大島裕明）の助成を受けたものです．ここに記して謝意を表します．

文　　献

[1] Momin M Malik, Hemank Lamba, Constantine Nakos, and Jürgen Pfeffer. Population bias in geotagged tweets. In *Ninth international AAAI conference on web and social media*, 2015.

[2] Shinya Hiruta, Takuro Yonezawa, Marko Jurmu, and Hideyuki Tokuda. Detection, classification and visualization of place-triggered geotagged tweets. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, pp. 956–963, 2012.

[3] Irene Cenni and Patrick Goethals. Negative hotel reviews on tripadvisor: A cross-linguistic analysis. *Discourse, Context & Media*, Vol. 16, pp. 22–30, 2017.

[4] David Jurgens, Tyler Finethy, James McCorriston, Yi Tian Xu, and Derek Ruths. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. In *Ninth international AAAI conference on web and social media*, 2015.

[5] Shuai Quan and Ning Wang. Towards a structural model of the tourist experience: An illustration from food experiences in tourism. *Tourism management*, Vol. 25, No. 3, pp. 297–305, 2004.

[6] Charu C. Aggarwal and Tarek Abdelzaher. *Social Sensing*, pp. 237–297. Springer US, Boston, MA, 2013.

[7] Supraja Gurajala and Jeanna N Matthews. Twitter data analysis to understand societal response to air quality. In *Proceedings of the 9th international conference on social media and society*, pp. 82–90, 2018.

[8] Quynh C Nguyen, Matt McCullough, Hsien-wen Meng, Debjyoti Paul, Dapeng Li, Suraj Kath, Geoffrey Loomis, Elaine O Nsoesie, Ming Wen, Ken R Smith, et al. Geotagged us tweets as predictors of county-level health outcomes, 2015–2016. *American journal of public health*, Vol. 107, No. 11, pp. 1776–1782, 2017.

[9] Luke Sloan and Jeffrey Morgan. Who tweets with their location? understanding the relationship between demographic characteristics and the use of geoservices and geotagging on twitter. *PloS one*, Vol. 10, No. 11, p. e0142209, 2015.

[10] Huiji Gao, Jiliang Tang, Xia Hu, and Huan Liu. Content-aware point of interest recommendation on location-based social networks. In *Twenty-ninth AAAI conference on Artificial Intelligence*, 2015.

[11] Madhuri Debnath, Praveen Kumar Tripathi, and Ramez Elmasri. Preference-aware successive poi recommendation with spatial and temporal influence. In *International Conference on Social Informatics*, pp. 347–360. Springer, 2016.

[12] Hanqing Bao, Dongping Ming, Ya Guo, Kui Zhang, Keqi Zhou, and Shigao Du. Dfcnn-based semantic recognition of urban functional zones by integrating remote sensing data and poi data. *Remote Sensing*, Vol. 12, No. 7, p. 1088, 2020.

[13] Zheng Xiang and Ulrike Gretzel. Role of social media in online travel information search. *Tourism management*, Vol. 31, No. 2, pp. 179–188, 2010.