ソーシャルセンサを用いた特定空間における

母集団推定に関する研究

２０２２年　３月

原　大樹

ソーシャルセンサを用いた特定空間における
母集団推定に関する研究

原　大樹

システム情報工学研究科

筑波大学

２０２２年　３月

概要

本論文は，ソーシャルセンサと呼ばれ，ソーシャルメディアの１つであるTwitterデータを特定空間における母集団推定に活用できるかどうかの分析をおこなったものである．

近年，スマートフォンの普及，情報技術の発展により，人々は常時オンラインとなり，様々なデータを発信している．このような状況において，ソーシャルメディアに蓄積されたデータを研究や企業活動に活用することは，不可欠な要素となっている．

そのような中，本論文では，日本国内プロ野球パ・リーグのメインスタジアムで発信されたTwitterデータを収集，分析することで，該当空間に対する観客数とTwitterデータの関係性や特徴を明らかにするとともに，Twitterデータを用いて観客数を推定することの有効性を確認している．

また，Twitterデータの活用，母集団推定精度の向上に資するため，日本国内プロ野球に関連するTwitterデータを収集し,該当空間内での投稿データと該当空間外での投稿データに分離し，それぞれの投稿内容や付加されたデータを比較することで，投稿場所によってツイートに含まれるURL情報やメディア情報の量が大きく異なることを明らかにした．その上で，明らかにした特徴とディープラーニング（BERT）及び機械学習の解釈手法（LIME）を用いることにより，該当空間でツイートされたものであるかの二値分類を高い精度で実現するとともに，分類に寄与した特徴量を明らかにしている．

本研究は日本国内プロ野球という限定された空間に対しての分析であるものの，これらの結果はTwitterデータを活用する上で，Twitterの投稿場所と投稿内容の関係性や特徴を考察する際の重要な情報であり，本研究は実用的な研究成果である．また，Twitterデータを用いて，母集団の推定や予測を検証したものはなく，取得が容易であり，ユーザのメッセージが含まれるTwitterデータを母集団推定に用いることは，母集団推定だけでなく，その空間に存在する集団の思いや意見など，様々な分析へ拡張できる可能性があり，意義があると考え

る．

本論文は，7章で構成される．第1章では，本論文で取り上げる研究の背景と目的を述べている．第2章では，ソーシャルセンサの特徴およびソーシャルセンサを用いた事例，観戦需要に関する先行研究について調査し，本研究の位置付けを明確にしている．第3章では，本研究の全体像および前提を示した上で，母集団推定で使用するデータの収集方法および抽出されたデータを示すとともに，対象空間における観客数および該当空間におけるTwitterデータの特性を示している．また，それぞれのデータの関係性を分析している．第4章では，第3章で収集，加工したソーシャルセンサデータと関連情報を用いた複数の重回帰モデルにより，該当空間における母集団を推定し，考察するとともに，母集団推定におけるソーシャルセンサデータの有用性を評価している．第5章では，第3章で収集，加工したソーシャルセンサデータと関連情報を用いたランダムフォレスト回帰モデルにより，該当空間における母集団を推定し，考察するとともに，母集団推定におけるソーシャルセンサデータの有用性を評価している．第6章では，機械学習アルゴリズムBERT（Bidirectional Encoder Representations from Transformers）を活用した分類器を提案し，ツイート投稿場所が特定空間内外のどちらであるかを分類することに取り組んでいる．さらには，特定空間内外におけるツイートの特徴分析を行い，ツイートに付与されたURL数やメディア数の違いを明らかにするとともに，これら属性情報と投稿内容を組み合わせることで分類精度が向上することを考察している．最後に第7章では，結論として本研究の成果をまとめるとともに，今後の取り組みについて述べている．

目次

図目次

表目次

第1章　緒論

近年，インターネットの普及に加え，スマートフォンやタブレットなどのモバイル端末の性能向上や普及とあいまって，ソーシャルメディアサービス（マイクロブログ，SNS，動画投稿，画像投稿）の利用者が急速に増加している．例えば，世界的に展開する最大のソーシャルネットワーキングサービスを提供しているFacebookの月間アクティブユーザは，全世界では既に27億人を超え［Meta, 2020］，国内での月間アクティブユーザは 2600 万人と言われている ［Social Media Experience, 2021］.

「情報通信白書」の平成 29 年版 ［総務省， 2017］，平成 30 年版 ［総務省，2018a］では，パソコンや携帯電話に比べて，スマートフォンやタブレット端末の利用者の方が，ソーシャルメディアサービスの利用率が高くなる傾向にある．身近にいつでもアクセスできるスマートフォン等がさらに普及すれば，ソーシャルメディアの利用は，さらに広がる可能性があることが示されている．実際，図 1-1，図 1-2 に示すように，スマートフォンの普及率は年々高まり，20 代，30 代では既に 90%を超え，SNS の利用率も増加傾向である．

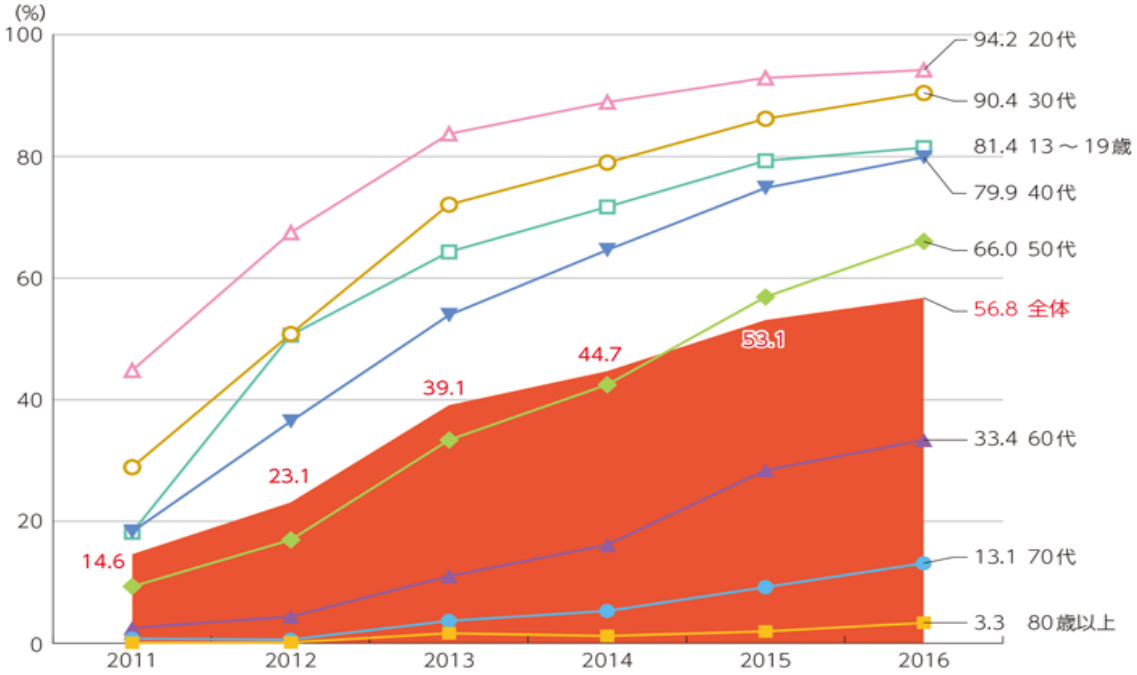


図 1-1　スマートフォン

（出典）総務省 通信利用動向調査

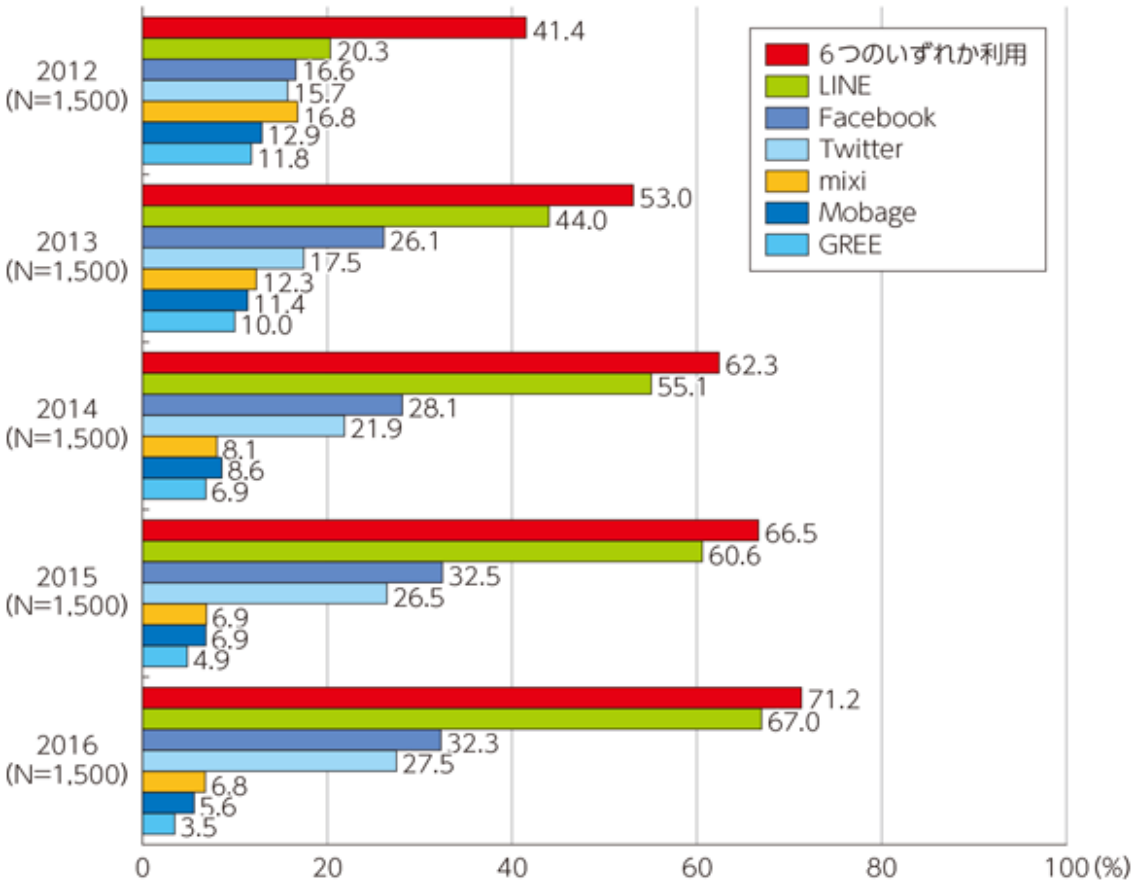


図 1-2　代表的 SNS の利用率の推移（全体）

（出典）総務省「情報通信メディアの利用時間と情報行動に関する調査」

ソーシャルメディアの多くは，SNS（Facebook や Twitter，mixi 等），ブログ（Ameba ブログ，Yahoo!ブログ等），LINE，掲示板，ミニブログなど文字（テキスト情報）によるコミュニケーションを主たる目的としたものである．

「平成 29 年情報通信メディアの利用時間と情報行動に関する調査」[総務省，2018b]によると，コミュニケーション手段としての利用時間（一日平均）は，平日では電子メール（以下，「メール」と略す）が一番長く，30.4 分，次いでソーシャルメディアが 27.0 分，休日ではソーシャルメディアが一番長く，31.2 分，次いで動画投稿・共有サービス，オンラインゲーム，ソーシャルゲームがともに 26.1 分となっており，平日と休日の両方で，テキスト情報を使ったメールやソーシャルメディアが，音声によるコミュニケーション手段である携帯電話や固定電話の利用時間を上回っており，現在の主要なコミュニケーション手段になっていることがわかる．

このようにソーシャルメディアの利用が広がるなか，ソーシャルメディア上でユーザが書き込むプロフィールやコメント等の構造化されていない非定型のテキスト情報やユーザの位置情報などをビジネスに活用する動きが現在既に進んでいる．近年，これを急速に加速させている背景にビッグデータ活用の拡大がある．

ビッグデータ活用とは，データの利用者やそれを支援するサービスの提供者それぞれの観点によって捉え方が様々であるが，ここでは，多種多量なデータを生成・収集・蓄積をリアルタイムで行い，このデータを分析することで未来の予測や異変の察知等を行い，利用者の個々のニーズに即したサービスの提供や業務の効率化，新サービスの創出に活かす取り組みとする．

特に，ソーシャルメディアへ書き込んだプロフィールやコメント，位置情報を収集・分析するソーシャルリスニングが広がりつつある．ソーシャルリスニングとは，人々がソーシャルメディア全体で日常的に語っている会話や自然な行動に関する投稿情報を収集・分析し，マーケティングや業務改善，発生したイベント検知に活かす手法である．前述のように，ビッグデータを収集・分析して社会問題の解決，マーケティング戦略立案や業務改善などのビジネスに活かす取り組みが急速に拡がっており，ソーシャルメディアがその情報源のひとつとして注目されている．

中でも，代表的なソーシャルメディアである Twitter では，ユーザは 140 文字以内のツイートと呼ばれるメッセージを使い，日々の生活体験や思いを投稿できる．投稿された情報は日常的に人から人へ伝わり，多くのユーザによってシェアされる．ツイートには，ユーザの日々の行動，生活体験，ユーザが購入した商品，サービスの選択基準や購入後の感想などが書き込まれる．企業にとっては自社のビジネスに役立つ知見の抽出やマーケティングに活用するために，このようなユーザが自ら発信する投稿情報を収集・分析することの重要性が増している．

さらに Twitter のデータはリアルタイム性が高く，位置情報など付帯する情報もあるため，多数決などの静的な情報分析だけでなく，場所や時系列に沿った動的な情報分析の研究などにも用いられている．

本研究は，ソーシャルメディアの１つであり，取得が容易な Twitter データの

リアルタイム性や位置情報を利用し，特定空間において発信された Tweet 数もしくは Tweet したユーザ数を計測し，分析に用いることで，特定空間における母集団を推定することを目的とする．母集団推定を行う対象は，国内プロ野球のパ・リーグで開催された試合におけるスタジアム観戦者数を取り上げる．

本研究によって期待される成果は 2 点ある．学術面については，その新規性が挙げられる．Twitter データを用いた，実際に発生したイベントの検出やテキストデータからのユーザ評価などの研究例は存在するが，母集団の推定や予測に用いた先行研究は見当たらない．

取得が容易でかつ観客でもある Twitter ユーザ自身が発生させる Twitter データを用いる点，またそれらデータとスタジアム観客数にはどのような特徴や関係性があるのかを明らかにすることができる点に意義があると考える．

2 点目は，実務面において，データの取得が容易な Twitter データからスタジアム等の観客数や来場者数など，特定空間における母集団を推定することができれば，観客数や来場者数の推定結果に基づく企業業績の予測や，ビジネス活動における計画の立案など，様々な分野への応用が期待できることが挙げられる．

以降の各章では，次の流れで議論を展開していく．第 2 章では，ソーシャルセンサの特徴およびソーシャルセンサを用いた事例，観戦需要に関する先行研究について説明する．

第 3 章では，本研究の全体像および前提を示した上で，母集団推定で使用するデータの収集方法および抽出されたデータを示すとともに，対象空間における観客数と Twitter データの特性を示す．また，それぞれのデータの関係性を分析する．

第 4 章では，第 3 章で収集，加工したソーシャルセンサデータと関連情報を用いた，複数の重回帰モデルによる該当空間における母集団推定を評価し，考察する．また母集団推定におけるソーシャルセンサデータの有用性を評価する．

第 5 章では，第 3 章で収集，加工したソーシャルセンサデータと関連情報を用いた，ランダムフォレスト回帰モデルによる該当空間における母集団推定を評価し，考察する．また母集団推定におけるソーシャルセンサデータの有用性を評価する．

第 6 章では，機械学習を用いた分類器を実装し，ツイート投稿場所が特定空間内外のどちらであるかを分類することに取り組む．また，特定空間内外におけるツイートの特徴分析をおこない，ツイートに付与された URL 数やメディア数の違いを明らかにするとともに，これら属性情報と投稿内容を組み合わせることで分類精度が向上することを考察する．なお，本取り組みに際しては，第 3 章のデータは用いず，新たに Twitter データの収集・加工をおこなう．また，機械学習のアルゴリズムには BERT（Bidirectional Encoder Representations from Transformers）を用いる．

第 7 章は，結論であり，本研究を総括する．

第2章　ソーシャルセンサの活用

2.1 緒言

本章では，ソーシャルセンサの特徴およびソーシャルセンサを活用した先行研究を概観するとともに，ソーシャルセンサを用いて母集団推定をする意義について説明する．

2.2 ソーシャルセンサの特徴

近年，スマートフォンの保有率が高まり，ソーシャルメディアが発展したことにより，これまでの新聞，テレビ，Web などからの情報収集に加え，Twitter や Facebook，Instagram などの SNS を利用して情報を収集する人が増えている．

SNS では企業による情報発信だけではなく，SNS の利用者が実際に購入した商品の評価，電車の遅延や事故などに遭遇した際の状況，自らがインフルエンザに感染したことなど,様々な情報がリアルタイムに発信され,蓄積されていく．そうした結果，SNS で情報にアクセスすることで電車の運行状況や流行のファッション,購入を悩んでいる商品の評判などを調べることができるためである．

榊らが提案しているソーシャルセンサという考え方は，その根幹的な研究である［榊 & 松尾，2012］．この研究では，図 2-1 で示されるように Twitter を通じて実世界を観測する場合において，各 Twitter ユーザを一種のセンサと考える．機械的なセンサが温度や振動など，様々な実世界を観測し，出力するように，Twitter ユーザが投稿する情報も人間というセンサが観測し，出力した情報として,実世界の観測に用いようとするものである.このような考え方は，いくつかの研究において「Social Sensor」,「Citizen Sensor」などと呼ばれている（［Sakaki，et al.，2010］，［Zhao，et al.，2006］）．

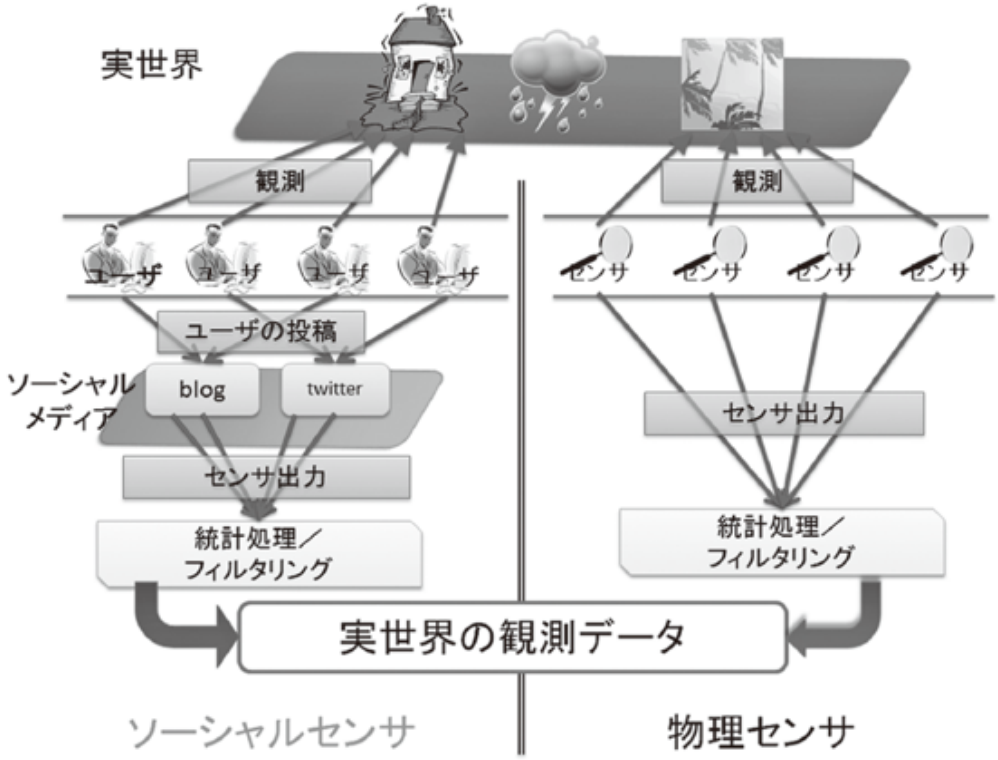


図 2-1 ソーシャルセンサとしてのソーシャルメディアユーザと物理センサの相似性

（出典）榊，松尾（2012）

ソーシャルセンサが観測するデータは，物理センサが観測するデータと異なり，発信するユーザの興味・状態など様々な定性的情報を含んだ実世界のリアルタイムな状態の観測出力である．例えば，2011 年 3 月 11 日に発生した東日本大震災では，Twitter が一般的なセンサでは観測できない人的被害の状況や支援物資の不足を観測し，伝達手段として利用されている［吉次，2011］．

このように既存のメディアと異なる Twitter データの特徴が活用され，様々な研究がされている．ソーシャルセンサとしての Twitter の特徴について［榊 & 松尾，2012］らの内容を踏まえつつ，一部，情報を加筆のうえ，説明する．

時間情報の利用

Blog をはじめとする既存のソーシャルメディアでは，投稿時間と実際のイベント発生時間のタイムラグが大きく［Zhao，et al.，2006］，イベント発生時間を推定するのが困難であった．また，検索エンジンの検索履歴を用いてイン

フルエンザの流行を検出する，といった実世界でのイベントを観測する研究も行われてきた［Ginsberg, et al., 2009]．しかしながら，検索エンジンの検索履歴は特定の事業者しか手に入れられないデータであるため，余り多くの研究は行われていない．

近年のモバイルデバイスの発達に伴い，イベントを観測してからユーザが投稿するまでのタイムラグが非常に小さくなり，投稿時間をイベント発生時間とみなすことができるようになった．これにより，今までは使うことができなかった時間情報をイベント抽出・同定の手掛かりとして用いることができる．

空間情報の利用

実時間性と同様に，モバイルデバイスの物理センサの発展に伴い，ユーザが位置情報を付加した投稿をすることができるようになっている．これにより投稿に付加された空間情報（位置情報）を用いてイベントが発生している場所を推測することができる．

ソーシャルメデイアではGPSにより投稿に付加されたgeotagと呼ばれる位置情報を用いる．geotag 付きのツイートは全体数の 0.42%と述べられている[Cheng, et al., 2010]．また，他にも位置情報が付与されたTweetの割合について，［鳥海，2015]は，0.1%程度，［山田 & 斉藤，2010]は0.6%程度，［橋本 & 岡，2012]は日本語のツイートにおいても0.18%程度という結果を示しており，いずれの結果においても，位置情報が付与されたTweetは非常に少ないことがわかる．

著者属性の利用

ユーザの多くがプロフィールにて居住地や居住地域のタイムゾーン，言語などを公開しており，また，ユーザ間インタラクションの際にユーザの同定が容易なため，著者属性を活用しやすい．

いくつかの研究では，Twitterユーザプロフィールの locationに記載されている地名を位置情報として用いている．しかしながら，自由記述であるため，

ユーザごとに地名の詳細さが異なり，地名でない記述をしているユーザも存在する．location に意味のある地名を記入しているユーザは全体の 66%，そのうち，都市単位まで記述しているユーザは 64%程度，州単位まで記述しているユーザは全体の 20%程度である［Hecht，et al.，2011］．

他にも，ユーザのリンク情報やユーザとリンク関係にあるユーザの位置情報を用いてユーザの位置情報を推定する研究［Backstrom，et al.，2010］や Twitter ユーザプロフィールの timezone の情報を位置情報として用いている研究［Song，et al.，2010］などがある．

このような Twitter データの特徴を用いて，観測するイベントを設定し，関連する Tweet を収集・分析することで，そのイベントの発生や移動などの検出をおこなう研究が存在する．

2.3 ソーシャルセンサの活用事例

Twitter に投稿されたデータを用いることで，現実に起こっているイベントを検出することが可能であり，このようなソーシャルセンサを用いた先行研究について述べる．

ソーシャルメディアによって物理的なイベントを観測する研究としては，山火事などの自然災害［Longueville，et al.，2009］や鳥インフルエンザや流行性感染症・季節性アレルギーであるインフルエンザ，花粉症などの流行（［Chew & Eysenbach，2010］，［Aramaki，et al.，2011］）などがある．

［Sakaki，et al.，2010］らは，地震に関するツイートを収集し，分析することで書き込み情報だけから地震の発生を 96%の精度で検出することに成功しており，地震の震源地を Twitter から検出し，地図上に提示している．

商品・企業に関する研究では，株価の推移予測（［Bollen，et al.，2011a］，［Schumaker，2010］）や一般に販売されている商品やサービスに対する意見・評判の分析［Jansen，et al.，2009］，ヒットする映画の予測［Asur & Huberman，2010］などがある．［Bollen，et al.，2011a］によるダウ平均株価の予測では，Twitter データのうち心的ツイートなどの書き手の感情が入っているツイート

に絞って感情分析をおこない，どのような心理状態がダウ平均株価と強い相関があるかを調査している．

他にも，スポーツの試合のシーンごとに要約を作成する研究［Chakrabarti & Punera, 2011]，コンサートに関する意見の構造化を行う研究［Benson, et al., 2011]，有名人の目撃情報を抽出する研究［榊 & 松尾, 2011]，スポーツや政治討論のテレビ中継に対する意見・感想（［Diakopoulos & Shamma, 2010]，[Shamma, et al., 2010])，大統領選挙や国会議員選挙などにおける人々の意見・評判（［Tumasjan, et al., 2010]，［那須野 & 松尾, 2014]）の分析，世論調査（［Akcora, et al., 2010]，［Bollen, et al., 2011b]）に関するものなど様々な研究がある．

また，イベントを観測する研究以外にも，マーケティングの観点から Twitter ユーザのプロフィールを推定する研究もおこなわれている．

［池田, et al., 2012]はテキストの中から重要なキーワードを検出することでプロフィール推定を行う手法を提案している．［榊 & 松尾, 2014]ははじめて職業を推定する手法を提案した．会社員か否かの二値分類で，適合率 85%，再現率 77%という結果を得ている．［米田 & 前田, 2017]は Twitter ユーザの性別を推定する手法を提案している．他にも Twitter の周辺ユーザの属性補完を利用する提案［上里, et al., 2015]や Twitter のメンション情報を使ってプロフィール推定［奥谷 & 山名, 2014]をおこなった研究がある．

このように Twitter データおよびその特徴は，イベント検出や場所特定，世論調査，商品・企業に対する評判の分析，投稿者のプロフィール推定など，様々な形で利用，研究がされているが，Twitter データを用いた特定空間における母集団推定に関する研究は見当たらない．

2.4 観客数の決定要因及び推定に関する先行研究

本研究では特定空間における母集団推定を扱い，その母集団は観客数とすることから，観戦需要研究に関する先行研究について概説する．

「観戦需要研究」とは，観客数の決定要因を探る研究分野であり，主に欧米において盛んに行われてきた．1970 年代に英国で［Hart, et al., 1975], 米国で ［Noll, 1974]が観戦需要研究を始めたのを契機に多くの論文が出されている（［Bird, 1982], ［Garcia & Rodriguez, 2002], ［Allan, 2004], ［河合, 2008]).

［Borland & Macdonald, 2003］は，観客需要研究には 2 つの大きな特徴があるとしている．1 つ目は，観客需要研究における研究対象スポーツは，英国におけるサッカーと米国における野球に集中していることであり，2 つ目は観客需要を測る明確なモデルは未だに開発されていないことである．

日本においては，［河合，2008]の J リーグの試合（サンプル数 2699 試合）を用いて，観客数を規定する要因に関する研究がある． ［河合， 2008]の研究では，観客数を規定する要因として，経済的要因，試合要因，観戦要因，人気要因，J リーグ要因という 5 つの要素にデータを分類している．

2.5 ソーシャルセンサを母集団推定に用いる意義

母集団推定に用いるデータとして Twitter データ以外にも株式会社 NTT ドコモが提供する「モバイル空間統計」［NTT ドコモ，2018]といった携帯電話事業者が有する携帯電話の接続情報が考えられる．

実際に，モバイル空間統計を用いた人口推計技術に関する研究［寺田，2014]やイベント開催による市街地への動向調査，都市拠点地区の人口特性分析［清家，et al.，2015]，その他，まちづくり分野に関する研究（［清家，et al.，2011], ［清家，et al.，2013]）などに用いられている．

このような携帯電話の接続データは情報の精度が高く，母集団推定の情報として有効と考えられるが，利用のためのコストが高く，また携帯電話契約者のプライバシーやセキュリティといった観点から非識別化処理，集計処理，秘匿処理が加えられている［NTT ドコモ，2021].

非識別化処理

モバイル空間統計では，携帯電話サービスを提供する上で必要となるデータのうち電話番号のような個人を識別できる情報を使用しない．また，生年月日を年齢層に変換するなど，情報の要約をおこなう．

集計処理

性別・年代別などの属性別に携帯電話の台数を数え，さらに，ドコモの携帯電話の普及率を加味することで，ドコモの利用者以外も含む人口を推計する．

秘匿処理

少人数エリアの数値を除去する．統計的に少数であることで個人を推測されやすくなる場合があり，これを防ぐためにおこなう．

個人情報保護等の観点から提供事業者が元データに非識別化処理，集計処理，秘匿処理を加えたうえで，データが提供されることから，提供されるデータに含まれる内容は少なくなり，利用に際しての柔軟性や応用性に欠けることが考えられる．

そこで，本研究では取得が容易であり，また位置情報以外にも投稿者であるユーザ自らが書き込むメッセージや属性情報など，様々な非定型情報を含み，活用することができる Twitter データを用いることとした．

2.6 結言

本章では，まずソーシャルセンサの特徴について説明した．次にソーシャルセンサを活用した先行研究および観戦需要研究について概説した．そのうえで Twitter データおよびその特徴を用いたイベント検出や場所特定に関する研究は数多く存在するが，Twitter データを用いた特定空間における母集団推定に

関する研究は見当たらないことを示した.

また，携帯電話事業者が有する携帯電話の接続情報について概説するとともに，ソーシャルセンサである Twitter データを母集団推定に用いる意義を説明した.

ユーザのメッセージが含まれる Twitter データを母集団推定に用いることは,母集団推定だけでなく,該当空間に存在する集団の思いや意見の抽出など,様々な分析への応用，拡張につながる可能性がある.

第3章　ソーシャルセンサデータと特徴分析

3.1 緒言

本章では，まず母集団推定の全体像および前提を示したうえで，本研究で扱うソーシャルセンサの特徴および第 4 章，第 5 章で使用するデータの収集方法と加工手順について説明する．

3.2 母集団推定の全体像と用いるデータ

本研究では，ソーシャルセンサデータを用いた，母集団を推定することに取り組む．具体的には 図 3-1 に示すイメージの通り，母集団推定対象である空間においてユーザが出力した位置情報付きの Twitter データを収集し，そこから抽出，集計した Tweet 数や Tweet したユーザ数のデータを用いて，対象空間の母集団を推定する．

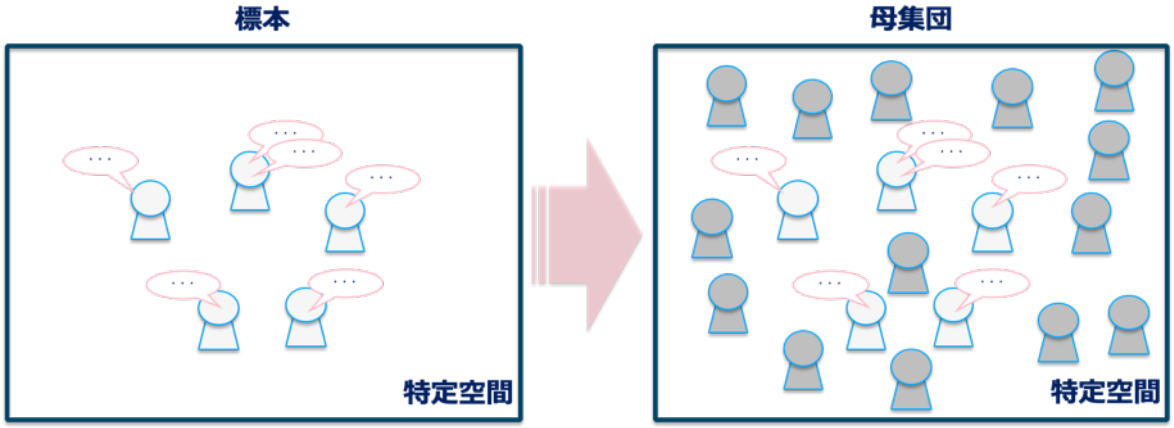


図 3-1　Twitter データを用いた母集団推定イメージ

母集団推定の対象とする特定空間は，日本国内プロ野球のパ・リーグ 6 球団間でおこなわれたメインスタジアムでの試合とし，母集団推定対象は該当スタジアムでの観客数とする．対象空間に関する詳細情報を表 3-1 に示す．

表 3-1　母集団推定対象とするスタジアムとチームの関係

対象スタジアム	ホームチーム	対戦相手チーム
福岡ドーム	ソフトバンク（hawks）	パ・リーグのチームのみ
西武ドーム	西武ライオンズ（lions）	*セ・リーグとの交流戦は含まない
宮城球場	楽天イーグルス（eagles）	
大阪ドーム	オリックス（buffaloes）	
札幌ドーム	日本ハム（fighters）	
千葉マリンスタジアム	千葉ロッテ（marines）	

なお，日本国内プロ野球のパ・リーグにおける試合観客数を母集団推定対象に選んだ理由は，推定対象の母集団である試合観客数が日々公表され，かつ比較的データ数が多いこと．また，パ・リーグは地理的にスタジアムが散らばっていることから選定した．

母集団推定に用いる機械学習モデルは各データ間での影響や特徴量の重要度を考察することができるモデルであることを考慮する．目的変数をスタジアムの観客数，説明変数に該当空間で出力された Twitter データを用いる．また，そのほかに説明変数として観客数を規定しうるホームチーム，対戦相手，試合実施の曜日，祝祭日，気温，降水量，風速などを加える．

3.3 ソーシャルセンサとしての Twitter データ

母集団推定をおこなうために用いる Twitter データの詳細について，［鳥海，2015]がまとめた内容を中心に一部情報を最新化したうえで説明する．

3.3.1 Twitter とは

Twitter とは，半角 280 文字（日本語・中国語・韓国語は全角 140 文字）以内のメッセージや画像，動画，URL などを投稿・閲覧できるサービスであり，投稿される記事は Tweet（つぶやき）と呼ばれる．通常のブログや SNS と異なり，文字数が制限されていることが，Twitter の特徴の一つであり，この文字数制

限が，Tweet を行うための労力を減らしており，気軽に Tweet を行えるようにしている要因であると考えられている．

基本的には Twitter は単に今何をしているかを 140 文字で書くだけのサイトであるが，コミュニケーションサイトとしての機能も有している．一つは，返信機能（Reply）である．返信機能は，友人などが行った Tweet に対して反応をしたいときに利用する．公開されては困るような内容についてやりとりをしたい場合は，Direct Message と呼ばれる機能を利用することになる．これは，相手と自分のみが見られる完全にプライベートなコミュニケーションになる．ただし，Direct Message でスパムが送られてくる事を防止するため，Direct Message は相手が自分を Follow している場合にしか送ることが出来ない．

Reply と Direct Message 以外に Twitter が持つ特徴的なコミュニケーション機能として，Retweet 機能がある．Retweet とは，他人が投稿した Tweet をそのまま他のユーザに転送する機能である．通常，各ユーザの Timeline には Follow しているユーザの Tweet しか表示されない．そのため，Follow 外のユーザがどのような Tweet を行なっているかは意識的に調べない限り見ることが出来ない．Retweet は，Follow していないユーザの Timeline に Tweet を表示させるための機能である．

これ以外にも Follow をしていないが気になるユーザをまとめるためのリスト機能や良いと思った Tweet をチェックするお気に入り機能（Favorite）が存在する．直接的なコミュニケーション機能ではないが，これらの機能によって Tweet の人気度などを測ることができ，Tweet とは異なるタイプの情報を獲得できる利点がある．

3.3.2 Twitter データの基本

Twitter のデータは，ユーザが非公開モードを選択しない限り誰でも参照可能な Tweet としてインターネット上に公開される．Twitter におけるデータ分析を行う場合，これらのデータを収集して分析を行うことになる．Twitter のデータには大きく，以下の 3 つのものがある．

Tweet データ

ユーザが投稿した Tweet 自体のデータであり，Tweet のメッセージ内容や投稿ユーザ，投稿時間，位置情報などのデータが含まれる．

ユーザデータ

Twitter のユーザに関するデータであり,アカウント名である ScreenName や,設定したユーザ名，プロフィールデータ，壁紙の色などのデータが含まれる．

Follow/Follower データ

ユーザ同士の関係性を示したデータであり，誰を Follow しているか，誰に Follow されているかを示すデータである．本データは，一種のソーシャルネットワークを表す．

本研究では母集団を推定するためのデータとして前述した Tweet データを用いるため，Tweet データについて，詳細を説明する．Tweet データは，一つの Tweet が持つ情報をあらわすものであり,そこから得られる主なデータは表 3-2 にまとめたものとなる．

TweetID は各 Tweet にユニークに付与される ID であり，created_at は，Tweet が投稿された日時を GTM で表記したものであり，秒単位まで記録されている．また，当該 Tweet を投稿したユーザに関する情報（User）も取得できる．

Text は Tweet されたメッセージそのものである．Source は Tweet を行ったアプリである．ブラウザから投稿した場合は Web と表示され，それ以外のアプリで投稿を行った場合，アプリ名と関連 URL が表示される．

GeoLocation データは，ユーザがどの位置で Tweet を行ったかをモバイル端末の GPS 情報などから取得したものである．Tweet 位置と Tweet 内容の関係などはマーケティングや観光情報などの分野で有効に利用できるが，多くの場合，ユーザはプライバシー保護の目的で位置情報投稿を停止している場合が多い．

その他，Retweet された数や Favorite された数などが獲得可能である．

表 3-2　Tweet から得られる主なデータ

データ名	説明
TweetID	Tweet ID
created_at	Tweet 作成日．GMT 表記
User	Tweet を行ったユーザの情報であり， UserID や ScreenName などが含まれる
text	Tweet 本文
source	Tweet を投稿したアプリ
GeoLocation	Tweet を行った緯度経度
retweet_count	Retweet された数
favorite_count	Favorite された数

[鳥海，2015]による表を一部修正

3.4 データの収集と加工

本研究において使用するデータの収集方法と抽出したデータおよびそのデータに加えた処理について説明する．

3.4.1 Twitter データの収集

Twitter データは,Twitter 社が無償で公開している Application Programming Interface（以下，API）を通じて収集する．API では全てのデータにアクセスする方法が提供されている． API は大きく二つの種類があり，一つは RestAPI と呼ばれる静的なアクセス方法，もう一つは StreamingAPI と呼ばれるリアルタイムアクセス方法である．

本研究で必要となるデータはリアルタイムな Tweet ではなく，ある特定日時・場所における Tweet であるため，Tweet データの収集には RestAPI を用いる．

Tweet に関するデータであれば，任意のユーザによって投稿された直近の Tweet や，特定の文字列を含んだ Tweet の検索，自分の Timeline 上の友人の Tweet など，様々な方法でアクセスすることが可能である．また，ユーザのプロファイルデータ，Follower 一覧からバックグラウンドの色までユーザに関する多くの情報を獲得可能である．

RestAPI を通じて Tweet データを収集する際の設定条件として，検索キーワードとツイートされたエリア範囲を指定することで，特定空間に関連性の高い Tweet データを収集することが可能である．

本研究では検索キーワード「*（アスタリスクを指定)」，収集するエリアの範囲を「スタジアムの中心（緯度経度で指定）から半径 120 メートル」の条件を設定し，2018 年 3 月 30 日から 2018 年 10 月 11 日の期間において，表 3-1 でまとめた国内プロ野球パ・リーグの 6 つのスタジアムでつぶやかれた Twitter データを収集した．2018 年における 10 月 12 日以降の試合はペナントレースに反映されないという通常と異なる特殊な試合となるため，本研究の対象にはしないこととする．

各スタジアムでつぶやかれた Tweet データを収集するために用いた設定値を表 3-3 に示す．なお，各スタジアムの中心となる緯度経度の測定には Web サイトで公開されているツール「Google マップに同心円を描画する」［マルティスープ，2018]を用いた．

表 3-3　Tweet データ収集に用いた設定値

対象スタジアム	緯度	経度	半径
福岡ドーム	33.595345	130.362214	0.12km
西武ドーム	35.768567	139.420524	0.12km
宮城球場	38.256346	140.902619	0.12km
大阪ドーム	34.669297	35.476103	0.12km
札幌ドーム	43.015081	141.409814	0.12km
千葉マリンスタジアム	35.645240	140.030909	0.12km

無償の API を通じた Twitter データの収集では,収集条件に該当した Tweet すべてを収集することはできず，Twitter 社にてサンプリングされたデータしか取得できないという制約 [Twitter, Inc., 2018]がある．本研究では Twitter 社によるサンプリングはある一定の条件に基づいたものであり，日々の取得可能なデータの割合などは同一であるという前提をおいたうえで，収集した Twitter データを取り扱うこととする．

また，本設定による収集エリアのイメージを図 3-2 に，取得した Tweet データをその位置情報を用いて地図上にプロットした結果を図 3-3 に示す．

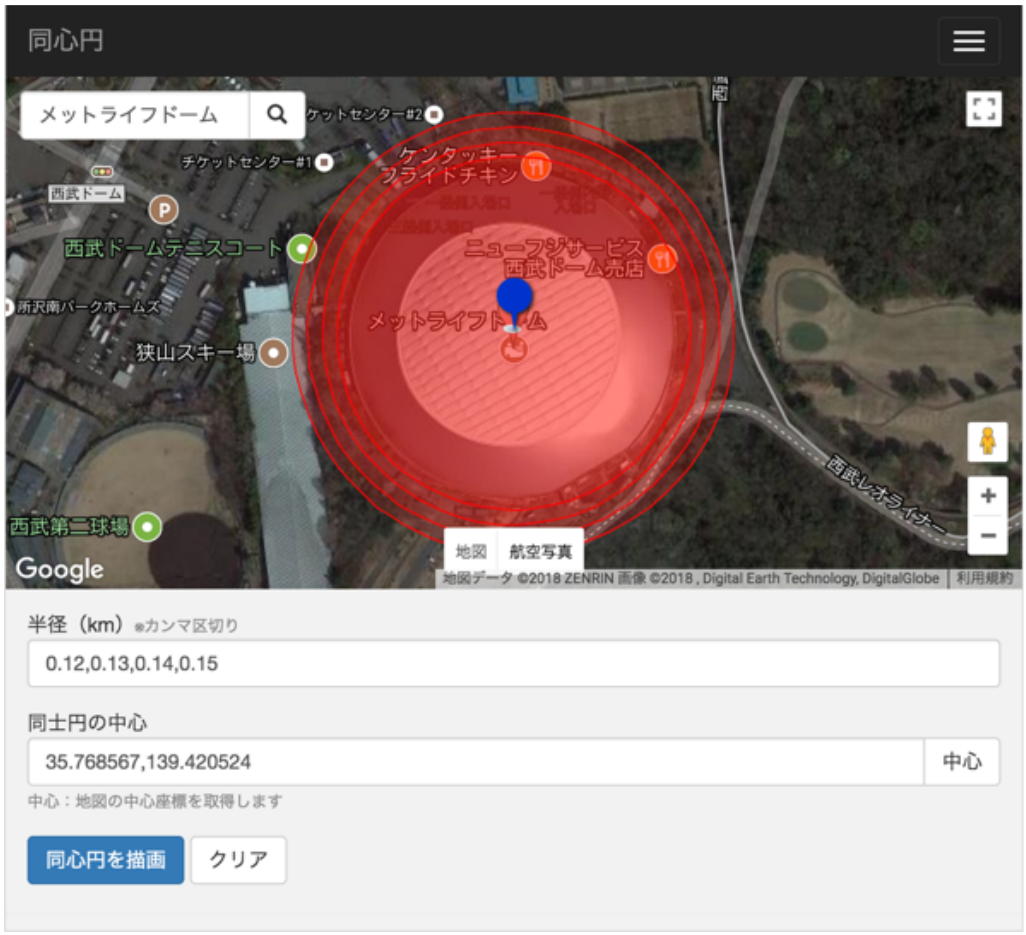


図 3-2　緯度経度の算出と Tweet 収集範囲イメージ

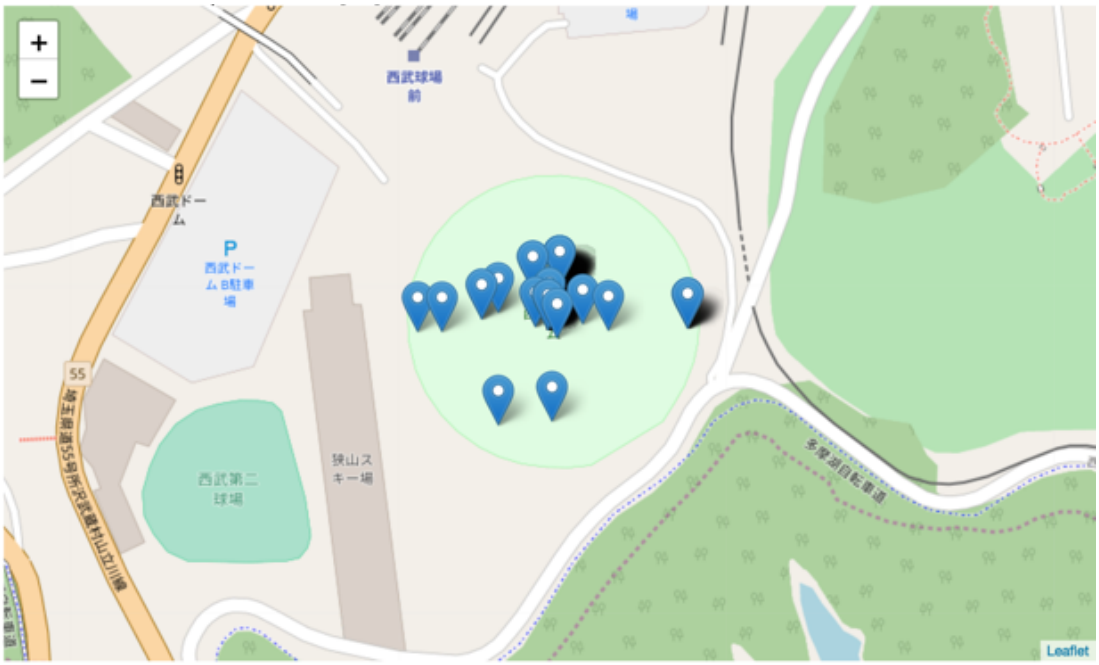


図 3-3　収集した Tweet を地図上にプロットした結果

3.4.2 プロ野球試合データの収集

プロ野球の試合日程，試合結果，観客数，試合時間などの情報はプロ野球の情報まとめサイトであるプロ野球 Freak［ロプロス，2018]から Web スクレイピングにより取得した．データの取得対象とした Web サイトに掲載されている情報項目の一例を表 3-4 および表 3-5 に示す．

表 3-4　プロ野球 Freak の球団別の試合日程・結果ページ

日付	勝敗	スコア	対戦相手	先発投手	責任投手	球場	開始
3月30日(金)	○	2 - 0	オリックス	千賀	○岩嵜	ヤフオクドーム	18:30
3月31日(土)	●	4 - 8	オリックス	武田	●武田	ヤフオクドーム	13:00
4月1日(日)	○	12 - 3	オリックス	中田	○石川	ヤフオクドーム	13:00
4月3日(火)	●	4 - 7	西武	東浜	●東浜	メットライフ	18:00
4月4日(水)	●	1 - 2	西武	バンデンハーク	●バンデンハーク	メットライフ	14:00

表 3-5　プロ野球 Freak の球団別の観客動員数ページ

日付	勝敗	スコア	対戦相手	先発投手	観客数	試合時間	球場
3月30日(金)	○	2 - 0	オリックス	千賀	38,530 人	2:46	ヤフオクドーム
3月31日(土)	●	4 - 8	オリックス	武田	38,530 人	3:29	ヤフオクドーム
4月1日(日)	○	12 - 3	オリックス	中田	38,325 人	3:53	ヤフオクドーム
4月10日(火)	●	1 - 4	日本ハム	東浜	31,774 人	3:29	ヤフオクドーム
4月11日(水)	○	8 - 5	日本ハム	バンデンハーク	32,255 人	3:30	ヤフオクドーム
4月12日(木)	○	3 - 0	日本ハム	石川	30,861 人	2:32	ヤフオクドーム
4月14日(土)	-	中止	ロッテ				熊本
4月15日(日)	○	7 - 6	ロッテ	中田	19,124 人	3:30	鹿児島

3.4.3 天候データの取得

対象スタジアムにおける天候情報は国土交通省気象庁の Web サイト［気象庁，2018]から該当スタジアムと住所が最も近い地域のデータを取得し，用いる．なお，取得したデータ項目は，気温，降水量，風速の 3 つである．スタジアムと取得した天候情報との地点の関係を表 3-6 に，取得した天候データ内容の一例

を表 3-7 に示す.

表 3-6　各スタジアムの天候情報とした地点

対象スタジアム	地点
福岡ドーム	福岡（福岡県）
西武ドーム	所沢（埼玉県）
宮城球場	仙台（宮城県）
大阪ドーム	大阪（大阪府）
札幌ドーム	札幌（北海道）
千葉マリンスタジアム	千葉（千葉県）

表 3-7　天候データ内容

年月日	福岡	福岡	福岡	福岡
	気温(℃)	降水量(mm)	風速(m/s)	風速
				風向
2018年4月6日10時	16.8	5.0	4.5	西
2018年4月6日11時	14.3	3.0	5.7	北
2018年4月6日12時	13.5	2.0	4.8	北北西
2018年4月6日13時	12.9	1.5	6.7	北
2018年4月6日14時	13.0	0.0	5.3	北
2018年4月6日15時	12.9	0.0	6.0	北北西

3.5 分析用データの作成

Twitter データ，プロ野球試合データおよび天候データを以下に示す手順で抽出・結合し，本研究で用いるための分析用データを作成した.

① プロ野球試合データより，各試合の試合開始２時間前と試合終了２時間後の時間を算出し，その間の時間を対象空間における当該試合のデータ収集の対象時間とする.

② 各スタジアムで収集した Twitter データのうち，試合開催日の①で算出した時間内に取得したデータを集計し，1 試合あたりの Tweet 数および 1 回以上ツイートを行った Tweet ユーザ数を集計する．なお，天候等の理由により中止となった試合および Tweet 数が 0 件の試合は外れ値として分析用データから除外する．

③ 天候データについても①で算出した時間内におけるデータを集計し，平均化することで，該当試合における平均気温，平均降水量，平均風速を算出する．

上記手順によるデータ集計・結合の結果，パ・リーグの全試合数のうち，メインスタジアムで開催された試合（セ・リーグとの交流戦を除く）かつ位置情報付きの Tweet が 1 件以上あった試合は 326 試合となった．結合したデータのサンプルを表 3-8 に示す．それぞれの列項目の説明を表 3-9 に示す．

表 3-8　分析用データ

date1	day_of_week	opposite	spectators	Num_of_Tweets	Num_of_Users	temperature	precipitation	wind_velocity	dome	home_team	day
2018-03-30 00:00:00+09:00	Fri	buffaloes	38530	112	62	14.928571	0.000	2.828571	1	hawks	weekday
2018-03-30 00:00:00+09:00	Fri	eagles	30051	0	0	11.533333	0.000	3.266667	0	marines	weekday
2018-03-30 00:00:00+09:00	Fri	lions	38693	74	41	5.442857	0.000	2.314286	1	fighters	weekday
2018-03-31 00:00:00+09:00	Sat	buffaloes	38530	167	55	19.462500	0.000	3.550000	1	hawks	weekend
2018-03-31 00:00:00+09:00	Sat	eagles	28203	2	1	15.728571	0.000	3.771429	0	marines	weekend
2018-03-31 00:00:00+09:00	Sat	lions	41138	50	23	12.228571	0.000	2.457143	1	fighters	weekend
2018-04-01 00:00:00+09:00	Sun	lions	41138	41	19	6.012500	0.375	2.112500	1	fighters	weekend
2018-04-01 00:00:00+09:00	Sun	eagles	29073	28	25	18.055556	0.000	7.600000	0	marines	weekend
2018-04-01 00:00:00+09:00	Sun	buffaloes	38325	149	45	21.825000	0.000	3.137500	1	hawks	weekend
2018-04-03 00:00:00+09:00	Tue	fighters	26622	47	28	17.212500	0.000	2.487500	0	eagles	weekday

表 3-9　分析用データ項目の説明

列名	変数属性	定義
date1	日付データ	試合実施日（タイムゾーンは東京）
day_of_week	カテゴリ変数	試合実施日の曜日
opposite	カテゴリ変数	対戦相手チーム
Num_of_Tweets	量的変数	試合開始２時間前から試合終了２時間後までの期間において，該当スタジアムで発信された位置情報付き Tweet 数
Num_of_Users	量的変数	試合開始２時間前から試合終了２時間後までの期間において，該当スタジアムで位置情報付き Tweet を発信したアカウント数
temperature	量的変数	試合開始２時間前から試合終了２時間後までの期間における平均気温
precipitation	量的変数	試合開始２時間前から試合終了２時間後までの期間における平均降水量
wind_velocity	量的変数	試合開始２時間前から試合終了２時間後までの期間における平均風速
dome	カテゴリ変数	スタジアムのドーム有無
home_team	カテゴリ変数	試合実施スタジアムのチーム ※本研究ではスタジアムとホームチームは同義となる
day	カテゴリ変数	試合実施日が平日か，土日祝日であるかを識別するフラグ

集計したデータから各スタジアムにおける対戦相手との試合数の関係を表 3-10 に示す．表 3-10 の赤字部分は開催された試合において位置情報付きの Tweet データが 0 件だったものを除外したということを表し，実際の総試合数は 333 件だが，本研究に用いるデータ数は 326 件となったことを意味する．

表 3-10　パ・リーグの試合データ数

パ・リーグ		対戦相手別の試合データ数（Tweet数が0件の試合を含まない）						試合データ数の合計
チーム名（略称）@対象スタジアム	ドーム	hawks	lions	eagles	buffaloes	fighters	marines	
福岡ソフトバンクホークス（hawks）@福岡ドーム	○	－	11	12	13	11	8 (-2)	55試合 (-2)
埼玉西武ライオンズ（lions）@西武ドーム	○	12	－	10	12	12	12	58試合
東北楽天ゴールデンイーグルス（eagles）@宮城球場	×	12	13	－	10	13	12	60試合
オリックス・バファローズ（buffaloes）@大阪ドーム	○	10 (-1)	9	11	－	9	9 (-1)	48試合 (-2)
北海道日本ハムファイターズ（fighters）@札幌ドーム	○	7	8	12	13	－	8	48試合
千葉ロッテマリーンズ（marines）@千葉マリンスタジアム	×	12 (-1)	11 (-1)	11 (-1)	11	12	－	57試合 (-3)
パ・リーグ合計試合数								326試合 (-7)

3.6 対象空間における特徴分析

前節で得られた分析用データを用いて，本研究の対象とするパ・リーグのプロ野球スタジアム観客数および Tweet 数，Tweet ユーザ数の特徴を考察する．

3.6.1 観客数に関する特徴分析

分析用データに含まれる連続値同士の散布図を描いた結果を図 3-4 に示す．観客数に対して，Tweet 数，Tweet ユーザ数，平均気温，平均降水量，平均風速のいずれも強い相関関係はないことがわかる．また，Tweet 数，Tweet ユーザ数の Twitter データに対して，平均気温，平均降水量，平均風速の天候データの相関も強いと言えないことがわかる．

観客数をヒストグラム（色分けはホームチーム）にしたものが図 3-5 である．ここから，福岡ドームにおけるホークス（hawks）の試合では観客数が多い傾向があり，反対に大阪ドームにおけるバッファローズ（buffaloes）の試合では観客数が少ない傾向があることがわかる．

観客数を各スタジアムに分離し，かつ試合開催日が平日（weekday）であるか，土日祝日（weekend）であるかで色分けしたヒストグラムが図 3-6 である．図 3-5 と同様に，スタジアムごとに観客数の偏り方が異なるとともに，全体を通

じて平日におこなわれる試合に比べ，土日祝日に行われる試合のほうが観客数は多くなる傾向がわかる．

これらの結果から観客数については，試合が開催されるホームスタジアムおよび試開催日が平日であるか土日祝日であるかが影響すると推測することができる．

図 3-4　分析用データの各項目間における散布図

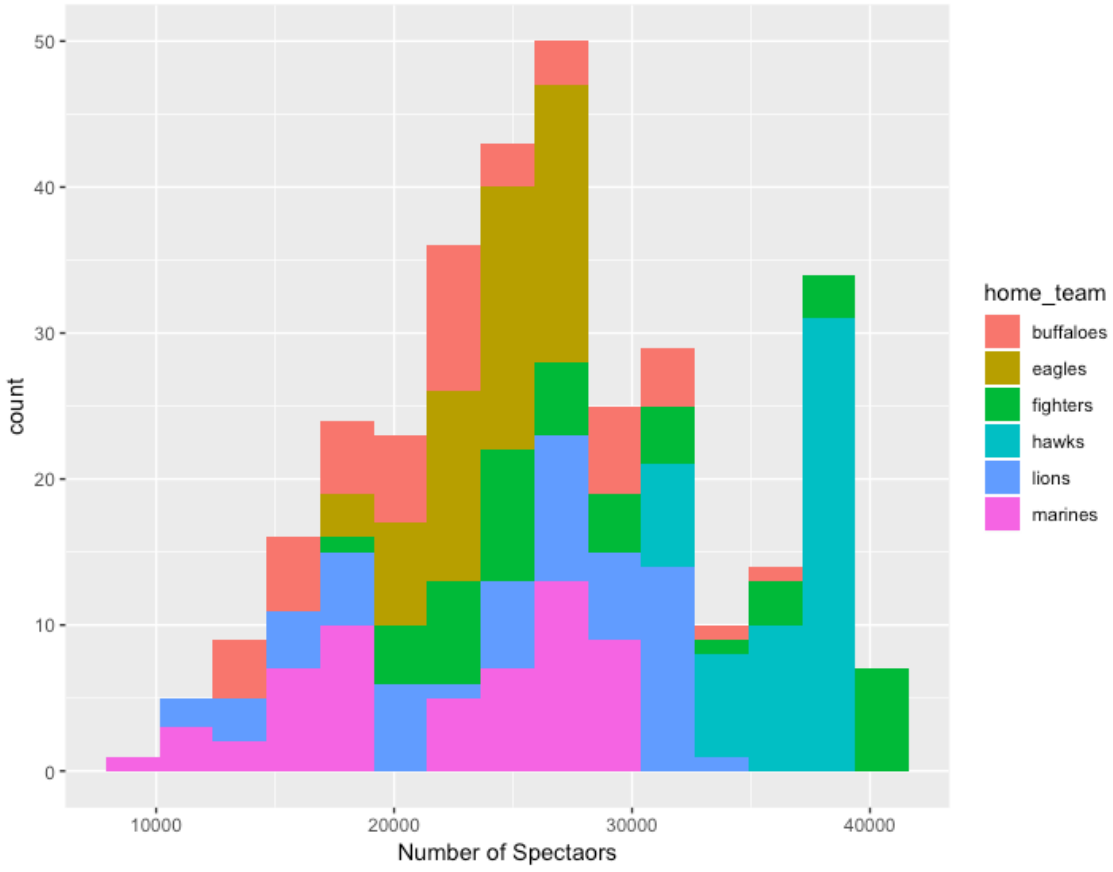


図 3-5　パ・リーグ観客数のヒストグラム

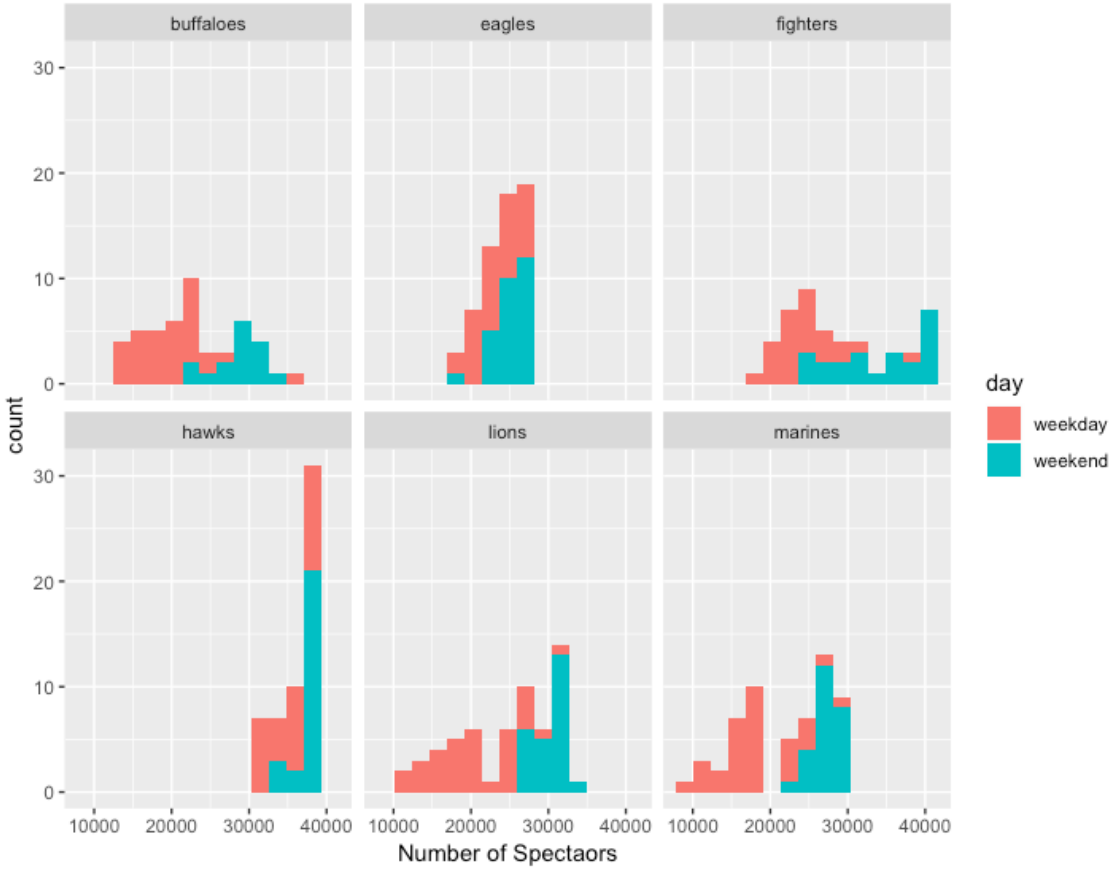


図 3-6　パ・リーグのスタジアム別観客数のヒストグラム

3.6.2 ソーシャルセンサの特徴分析

分析用データに含まれる Twitter データについて特徴を考察する．各試合で Tweet された件数をヒストグラムにしたものが，図 3-7 である．Tweet 件数のレンジとしては 1 試合あたり 100 件弱が多く，千葉マリンスタジアムにおけるマリーンズ（marines）の試合では回数は少ないながらも 200 件を超えるツイートがあることが見て取れる．

Tweet 件数を各スタジアムに分離し，かつ試合開催日が平日（weekday）であるが，土日祝日（weekend）であるかで色分けしたヒストグラムが図 3-8 である．平日とくらべて土日祝日にツイート件数が増える試合が多いことがわかる．

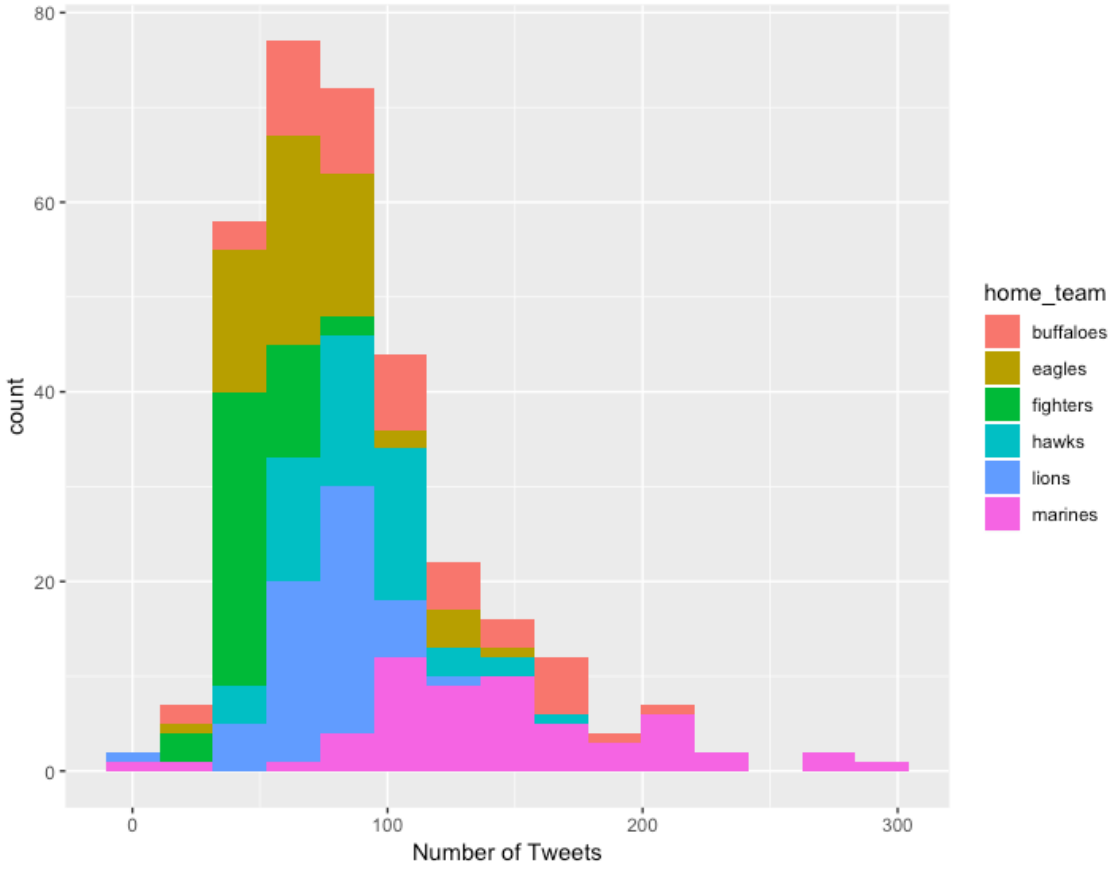


図 3-7　パ・リーグの Tweet 数／試合のヒストグラム

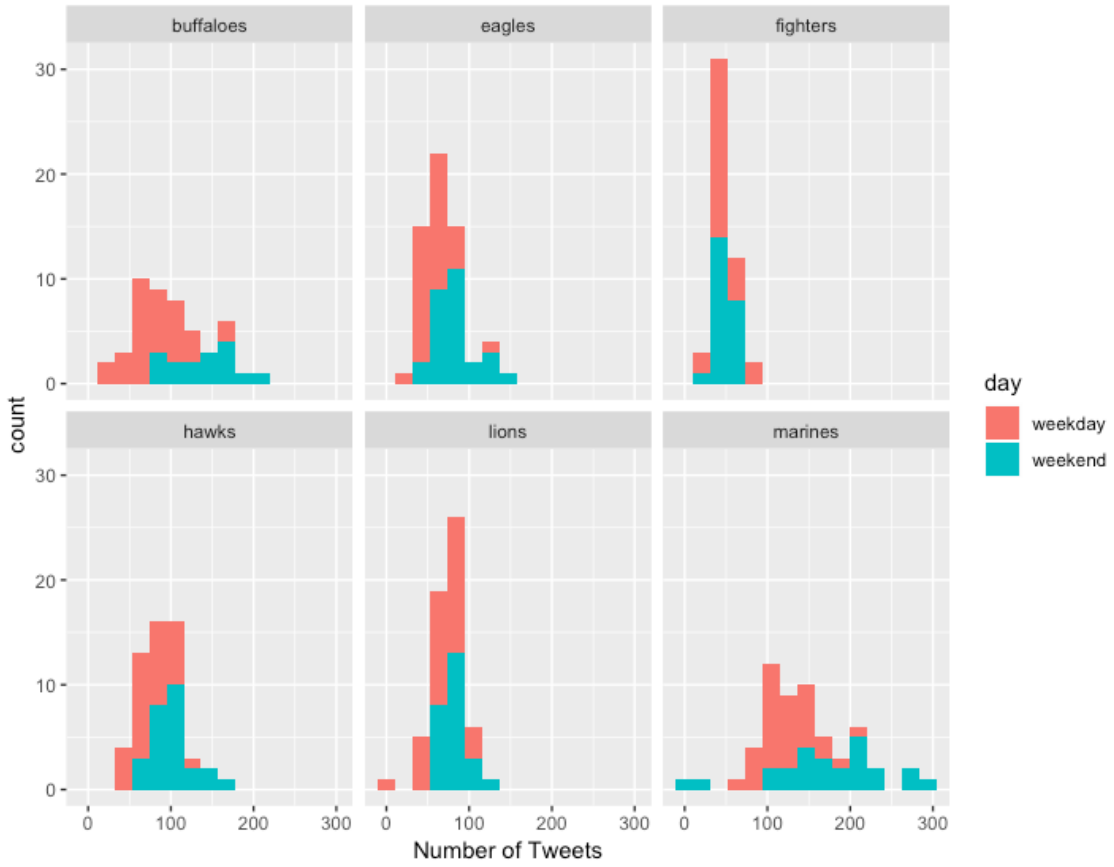


図 3-8　パ・リーグのスタジアム別 Tweet 数／試合のヒストグラム

各試合における Tweet したユーザ数をヒストグラムにしたものが，図 3-9 である．Tweet ユーザ数のレンジは，1 試合あたり 50 人弱が多く，千葉マリンスタジアムにおけるマリーンズ(marines)の試合では，回数は少ないながらも 100 名を超えるユーザがツイートをする場合があることがわかる．

Tweet ユーザ数を各スタジアムに分離し，かつ試合開催日が平日（weekday）であるか，土日祝日（weekend）であるかで色分けしたヒストグラムが図 3-10 である．平日とくらべて土日祝日のほうが Tweet をするユーザが増える試合が多いことがわかる．

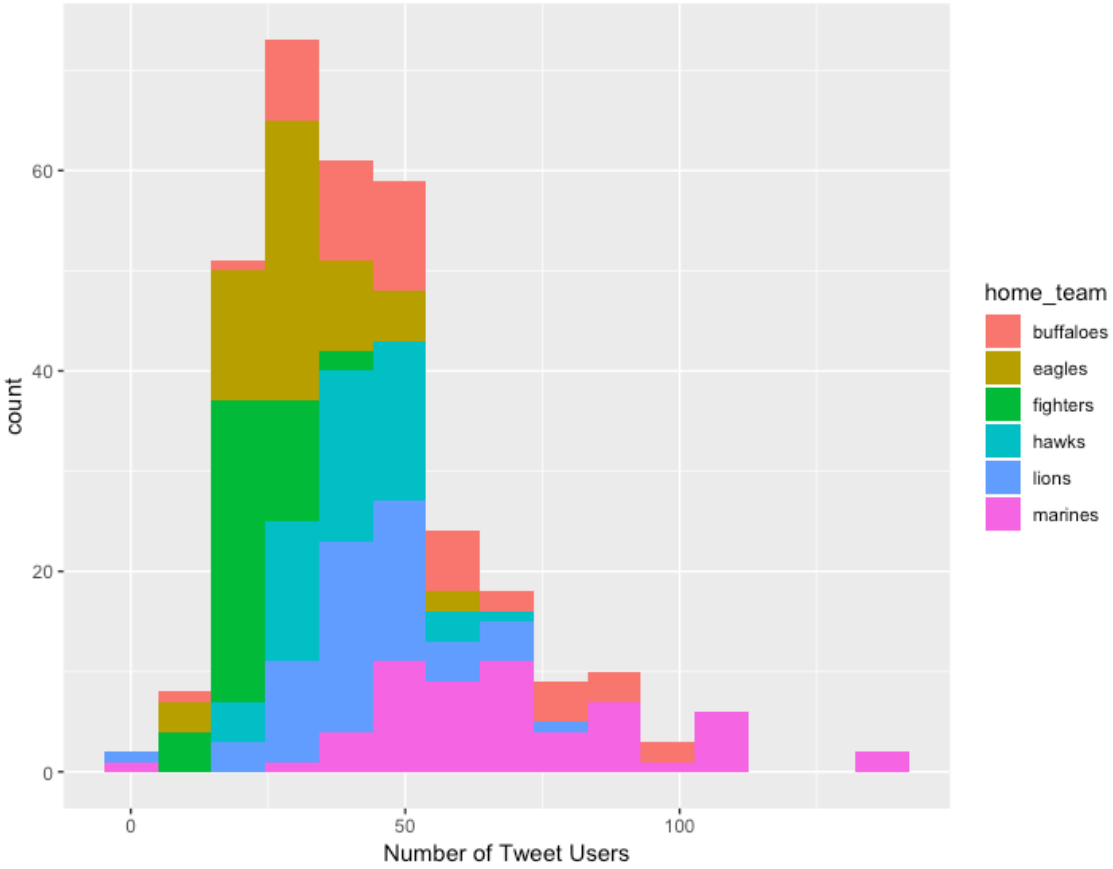


図 3-9　パ・リーグの Tweet ユーザ数／試合のヒストグラム

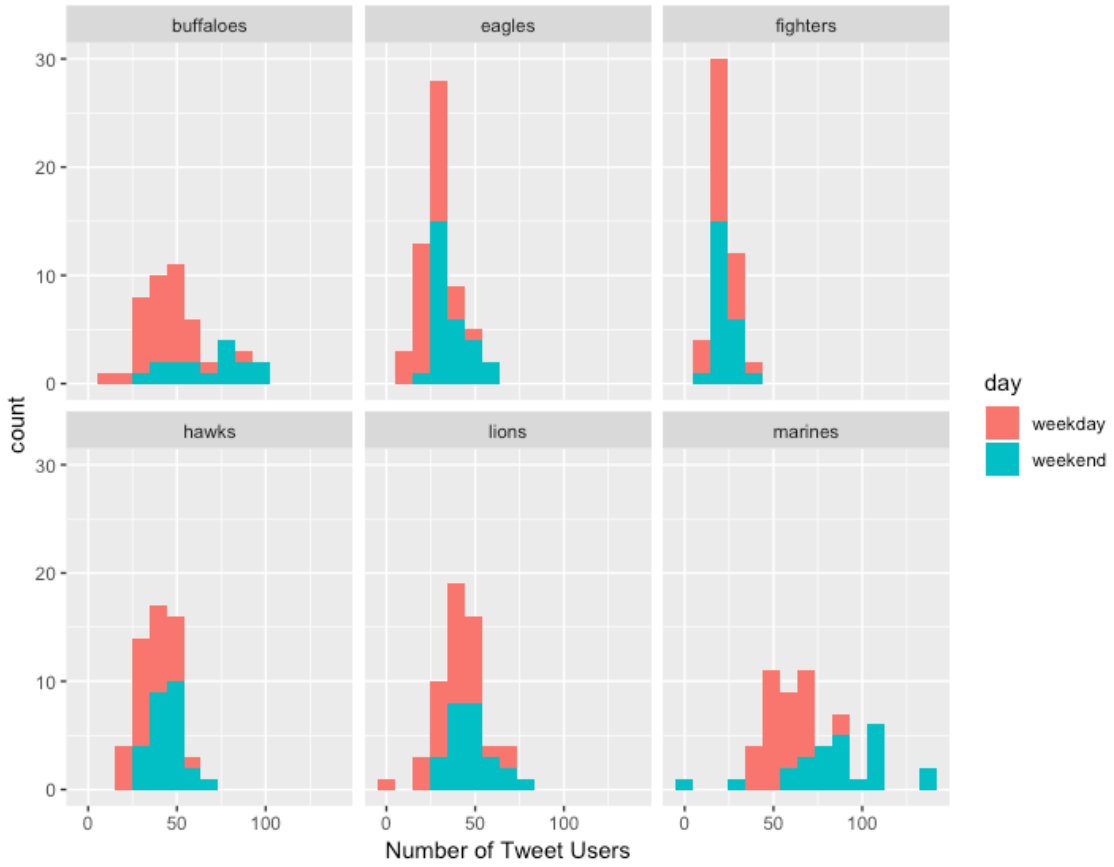


図 3-10　パ・リーグのスタジアム別 Tweet ユーザ数／試合のヒストグラム

3.7 対象空間とソーシャルセンサの関係性分析

3.7.1 観客数と Tweet 数の相関関係

観客数と Tweet 数の散布図を図 3-11 に示す．パ・リーグ全体として観客数と試合あたりの Tweet 数との相関は弱いが，ホームチームごとの色分けで固まりの特徴があることを読み取ることができる．実際に相関係数を算出すると，パ・リーグ全体での観客数と Tweet 数との相関係数は 0.065 である．

各スタジアムに分割した観客数と Tweet 数の散布図が図 3-12 である．スタジアムごとに特徴がありつつも，いずれも正の相関を見て取れる．一例を挙げると，大阪ドームにおけるバッファローズ（buffalos）戦の観客数と Tweet 数との相関係数は 0.684 となり正の相関があることがわかる．各スタジアムに分割した各チームの試合における観客数と Tweet 数の相関係数を算出した結果を図

3-15，表 3-11 にまとめる．

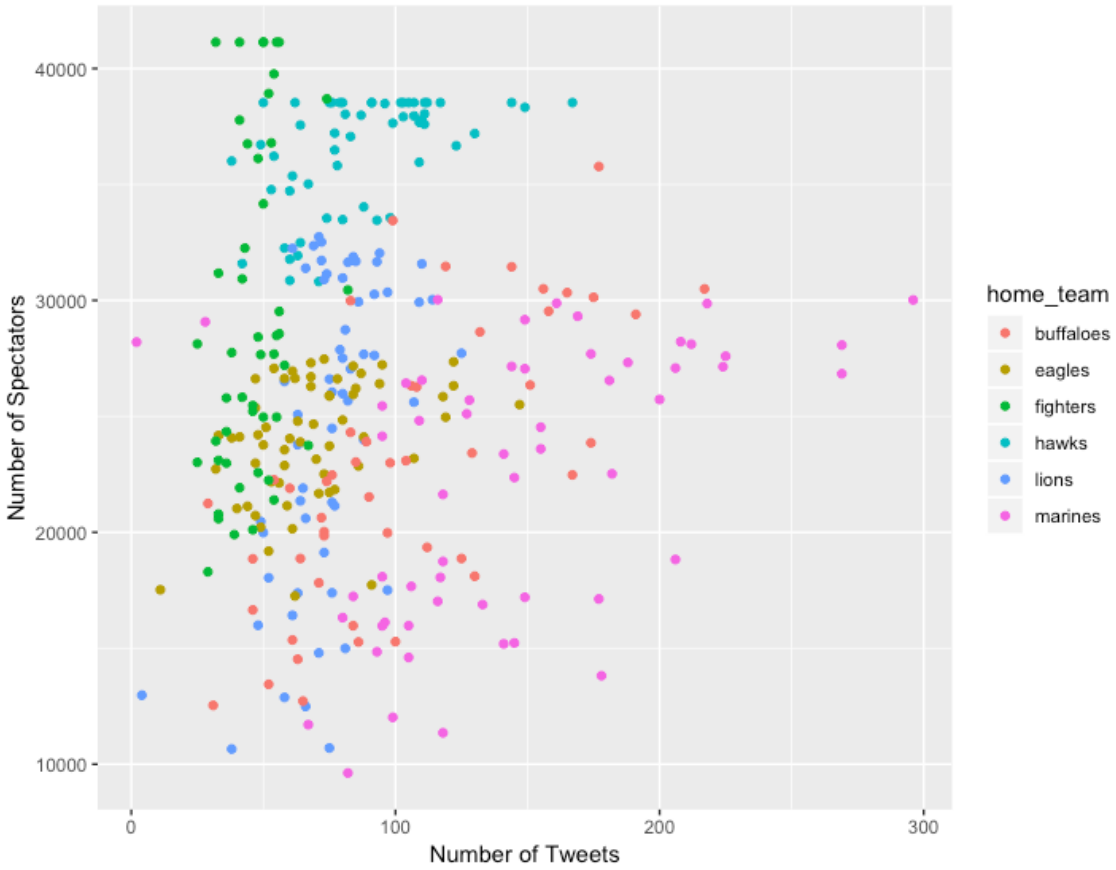


図 3-11　パ・リーグの観客数と Tweet 数の散布図

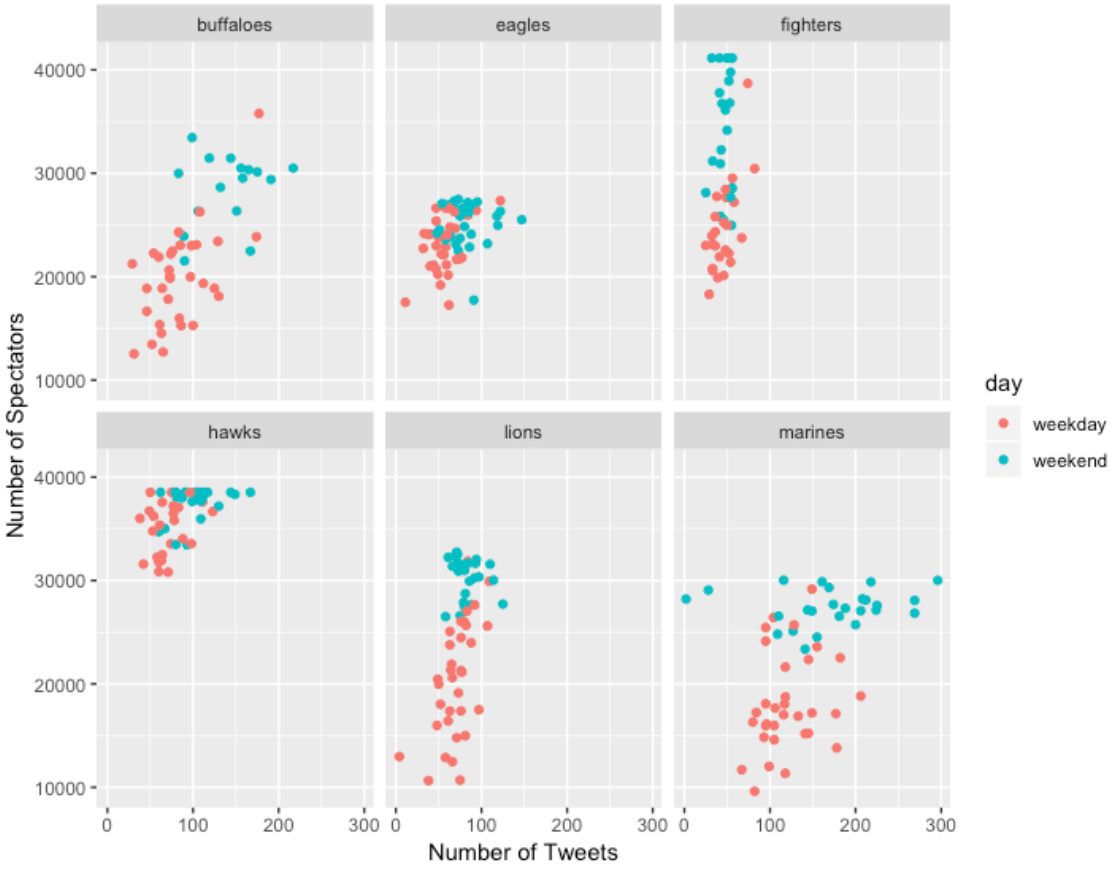


図 3-12　パ・リーグのスタジアム別の観客数と Tweet 数の散布図

3.7.2 観客数と Tweet ユーザ数の相関関係

観客数と Tweet ユーザ数の散布図を図 3-13 に示す．パ・リーグ全体として観客数と試合あたりの Tweet ユーザ数との相関は弱いが，ホームチームごとの色分けで固まりの特徴があることを読み取ることができる．実際に相関係数を算出すると，パ・リーグ全体での観客数と Tweet ユーザ数との相関係数は 0.064 である．

各スタジアムに分割した観客数と Tweet ユーザ数の散布図が図 3-14 である．スタジアムごとに特徴がありつつも，いずれも正の相関を見て取れる．一例を挙げると，大阪ドームにおけるバッファローズ (buffalos) 戦の観客数と Tweet ユーザ数との相関係数は 0.736 となり強い正の相関があることがわかる．各スタジアムに分割した各チームの試合における観客数と Tweet ユーザ数の相関係数を算出した結果を図 3-15，表 3-11 にまとめる．

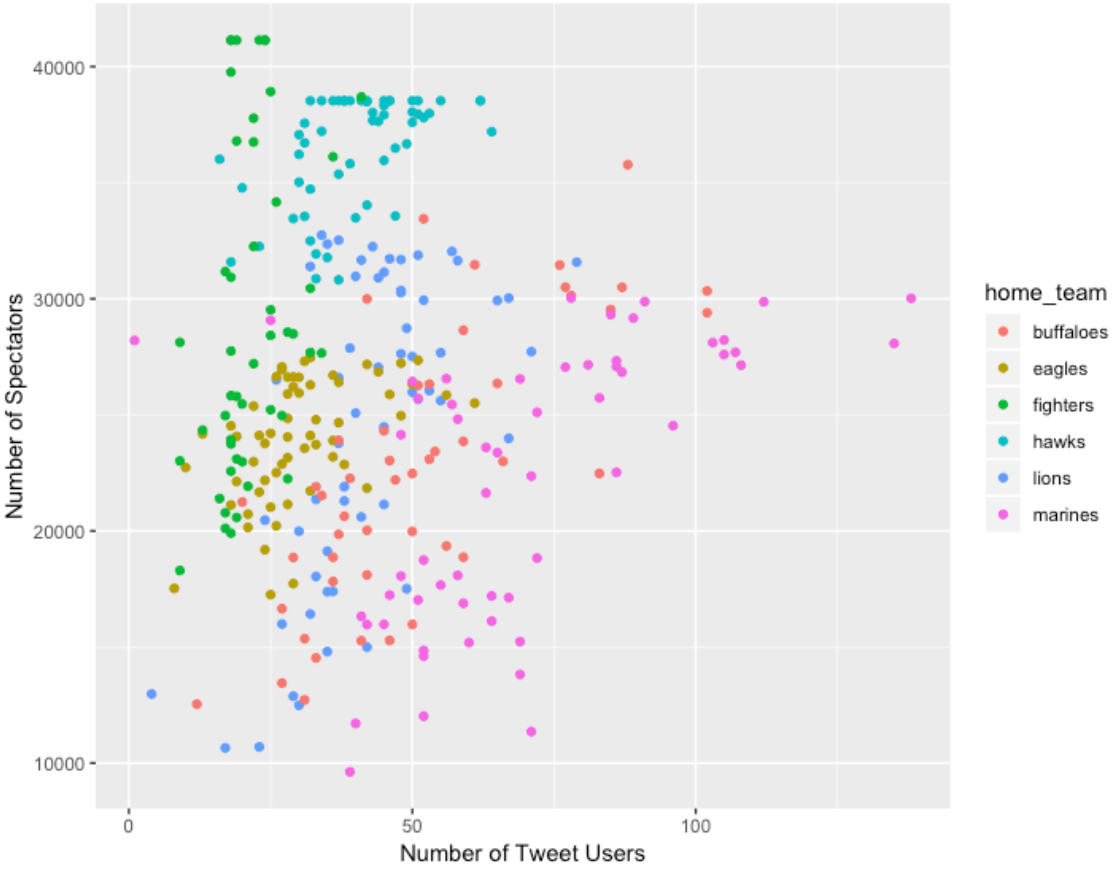


図 3-13　パ・リーグの観客数と Tweet ユーザ数の散布図

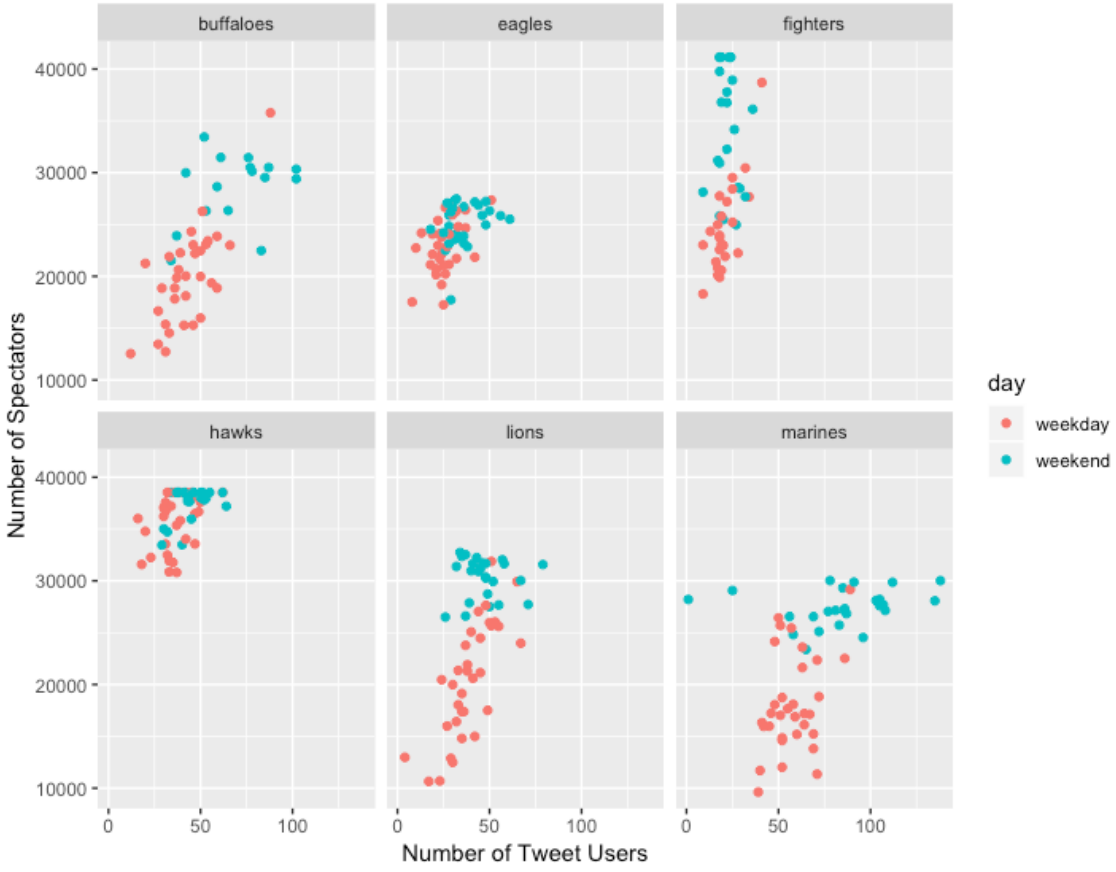


図 3-14　パ・リーグのスタジアム別の観客数と Tweet ユーザ数の散布図

3.7.3 観客数とソーシャルセンサの関係性の考察

パ・リーグ全体，各スタジアムでの観客数と Tweet 数および観客数と Tweet ユーザ数の相関係数を算出し，まとめたものが図 3-15，表 3-11 である．

観客数と Tweet 数の相関係数と観客数と Tweet ユーザ数の相関係数では後者の値が大きくなる傾向が見られる．これは，Tweet 数は試合の盛り上がりなどによって変動する要素が多いが，Tweet したユーザ数は盛り上がりなどの影響を受けにくいためと考える．このような Tweet 数と Tweet ユーザ数の関係については，［三田村，2014］の研究における「述べ数」と「異なり数」との関係に近いと考えることができる．

またパ・リーグ全体とした場合の観客数と Tweet 数および観客数と Tweet ユーザ数の相関係数が小さいことは，それぞれのスタジアムの特徴が異なっており，各々がそれらの特徴を打ち消しあった結果と考えることができる．

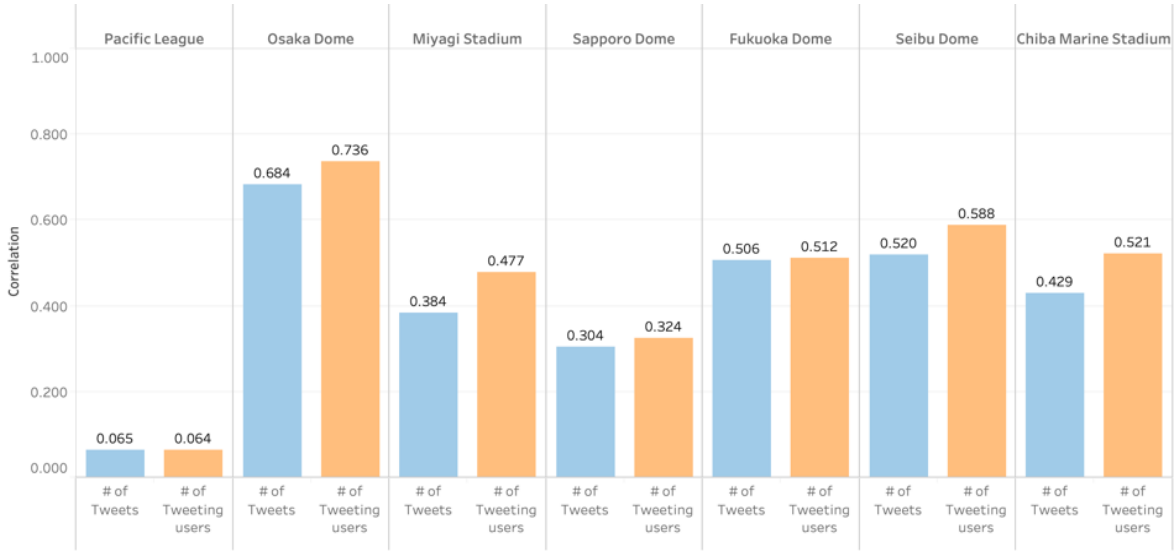


図 3-15　観客数と Tweet 数／Tweet ユーザ数の相関係数

表 3-11　観客数と Tweet 数／Tweet ユーザ数の相関係数

スタジアム名	Tweet 数との相関係数	Tweet ユーザ数との相関係数
パ・リーグ全体	0.065	0.064
福岡ドーム	0.506	0.512
西武ドーム	0.520	0.588
宮城球場	0.384	0.477
大阪ドーム	0.684	0.736
札幌ドーム	0.304	0.324
千葉マリンスタジアム	0.429	0.521

3.8 結言

本章では，まず母集団推定の全体像と Twitter 及びそのデータの基本について述べた．次に第 4 章，第 5 章で用いるデータについて，Twitter データ，プロ野球試合データ及び天候データの収集方法とそれらデータから分析用データを作成する際の加工方法について述べた．Twitter データの収集については緯度経度を指定することで，スタジアム内で出力された Tweet データを正確に取得することができ，また検索キーワードを「*」と設定できることで分析者の主

観に基づくキーワード条件を用いないことに意義はあると考える．しかしながら，位置情報付きの Tweet データが少なくなることは課題である．

また，収集したプロ野球データより，観客数の増減は，ホームスタジアムおよび試合開催日が平日か土日祝日であるかの要因が関わっていると考えることができる．

収集した Twitter データからは，観客数と Tweet 数よりも観客数と Tweet したユーザ数の相関係数を比較し，前者にくらべて後者の値が大きくなる傾向を示した．したがって，Twitter データを用いてプロ野球の試合観客数を推定する場合，Tweet 数よりも Tweet したユーザ数を用いることで推定精度が高くなる可能性がある．

第4章　重回帰モデルを用いた特定空間における母集団推定

4.1 緒言

第 3 章で示したデータおよび重回帰モデルを用いて，日本国内プロ野球のパ・リーグ 6 球団間で，かつ各チームのメインスタジアムでおこなわれた試合における観客数の推定をおこなう．

Twitter データの特徴を踏まえ，Tweet 数および Tweet ユーザ数等，複数の特徴量を組み合わせ,またステップワイズ法による変数選択等をおこなうことで，複数の重回帰モデルを構築し，母集団推定の精度を比較・評価する．

また，Twitter データを母集団推定モデルに加えることで推定精度が向上するかを評価する．

4.2 重回帰分析における変数選択

本章では目的変数である観客数と説明変数に用いる各種データの影響の度合いも含め，比較，評価をおこなうため，母集団推定モデルとしては回帰モデルを用いることとした．本節では回帰モデルを用いる上でどのデータを説明変数に用いるべきかの変数選択をおこなう．

説明変数を増やすことで決定係数は大きくなるが，説明変数の多いモデルが良いモデルであるとは限らない．予測精度が同程度であるならば，説明変数のより少ないシンプルなモデルのほうがデータ収集コストも少なくなるため，優れていると考えることができる．言い換えると，予測に寄与しない不要な説明変数は，モデルから取り除くほうが望ましいと考えることができる．

統計的基準をもとに予測精度を低下させることなく，変数選択を行うためには，主に変数増加法，変数減少法，ステップワイズ法（変数増減法）の 3 つの方法がある［川端，et al.，2018］.

変数増加法

切片のみの（説明変数のない）モデルを基準に，予測に有効な説明変数を 1 つずつ追加していく方法

変数減少法

全ての説明変数を用いて予測を行い，予測に寄与しない説明変数を 1 つずつ削除していく方法

ステップワイズ法（変数増減法）

変数増加法と変数減少法を組み合わせたもので，予測に有効な変数を取り入れ，有効でない変数を削除することを繰り返し，最適な組み合わせを探る方法

本章では，ステップワイズ法（変数増減法）を用いて説明変数の選択をおこなった.なお,今回の変数選択は,統計ソフト R のパッケージ MASS の関数 stepAIC を用いておこない，統計モデルの良さを評価するための指標である AIC（赤池情報量規準） [Akaike, 1973]が最小になる説明変数の組み合わせを探索している.

ステップワイズ法による変数選択の結果，得られた重回帰モデル式（Model0）が式 4-1 である．なお，(*) は説明変数がカテゴリ変数であることを表す.

$$\boldsymbol{spectators} = \boldsymbol{home\_team}(*) + \boldsymbol{day}(*) + \boldsymbol{Num\_of\_Users} + \boldsymbol{opposite}(*) + \boldsymbol{Num\_of\_Tweets} + \boldsymbol{temperature} + \boldsymbol{wind\_Velocity} \quad \text{(式 4-1)}$$

Model0 に対するモデルの各種統計量を表 4-1 に，各説明変数に対する偏回帰係数および各説明変数間の多重共線性を確認するための GVIF（Generalized Variance Inflation Factor：一般化分散拡大要因）の結果を表 4-2，表 4-3 に

まとめる.

ステップワイズ法により求められた Model0 の決定係数は 0.7440 であり，目的変数の分散のうち，説明変数で説明できた割合（分散説明率）として高い数値を示している．また，調整済み決定係数も 0.7317 と決定係数とくらべて極端に小さくなっていないことから目的変数に寄与しない説明変数はないと考えることができる.

表 4-1　ステップワイズ法により得られた重回帰モデルの統計量（Model0）

決定係数	0.7440
調整済み決定係数	0.7317
AIC	5385.753
BIC	5446.343
Residual standard error: 3776 on 310 degrees of freedom	
F-statistic: 60.08 on 15 and 310 DF, p-value: < 2.2e-16	

得られた偏回帰係数については，ホームチーム（home_team）および平日か土日祝日（day）の情報が大きな係数を保持しており，第 3 章での考察と同じ結果が得られた.また,Tweet ユーザ数(Num_of_Users)については偏回帰係数 149.69 が得られた．これは，全ての説明変数を固定した際，ツイートユーザが 1 名増えると観客数が約 150 名増えることを意味する．Tweet 数（Num_of_Tweets），平均気温（temperature），平均風速（wind_velocity）の説明変数もステップワイズ法で求めた Model0 では採用されているが，　有意水準 5%では有意でないという結果になった.

表 4-2　得られた偏回帰係数等の詳細情報（Model0）

変数名	推定値	標準誤差	95%CI [2.5%, 97.5%]	t 値	p 値	
(Intercept)	11538.09	1476.98	[8631.90946, 14444.26562]	7.812	8.80e-14	***
home_teameagles	3572.14	851.43	[1896.82261, 5247.45052]	4.195	3.56e-05	***
home_teamfighters	9924.24	962.67	[8030.04994, 11818.43759]	10.309	< 2e-16	***
home_teamhawks	14740.48	790.12	[13185.80991, 16295.15714]	18.656	< 2e-16	***
home_teamlions	2626.88	785.78	[1080.74928, 4173.00870]	3.343	0.000930	***
home_teammarines	-3120.56	859.81	[-4812.35836, -1428.75517]	-3.629	0.000332	***
dayweekend	5224.51	478.75	[4282.49584, 6166.51624]	10.913	< 2e-16	***
Num_of_Users	149.69	28.89	[92.83335, 206.54060]	5.181	3.99e-07	***
oppositeeagles	2150.76	735.00	[704.53485, 3596.97929]	2.926	0.003685	**
oppositefighters	2311.93	734.35	[866.99676, 3756.86911]	3.148	0.001802	**
oppositehawks	2304.69	749.63	[829.69063, 3779.69290]	3.074	0.002297	**
oppositelions	2264.74	746.62	[795.66511, 3733.82359]	3.033	0.002623	**
oppositemarines	128.50	762.11	[-1371.05515, 1628.06343]	0.169	0.866208	
Num_of_Tweets	-22.72	13.24	[-48.76410, 3.33124]	-1.716	0.087160	.
temperature	67.56	40.69	[-12.49513, 147.61851]	1.661	0.097817	.
wind_velocity	193.64	135.82	[-73.60633, 460.89262]	1.426	0.154957	

Model0 における GVIF の結果を表 4-3 に示す.Tweet ユーザ数(Num_of_Users」と Tweet 数（Num_of_Tweets）の値が高いことがわかる．Tweet 数と Tweet ユーザ数は相関があることから多重共線性が考えられる.

表 4-3　GVIF の算出結果（Model0）

変数名	GVIF	Df	GVIF(1/(2*Df))
home_team	4.534962	5	1.163208
day	1.290346	1	1.135934
Num_of_Users	9.156137	1	3.025911
opposite	1.538866	5	1.044047
Num_of_Tweets	8.373707	1	2.893736
temperature	1.358457	1	1.165529
wind_velocity	1.258216	1	1.121702

4.3 重回帰モデル式の検討と評価

前節での変数選択されたモデル（Model0）およびその詳細を確認した結果，Tweet 数（Num_of_Tweets）と Tweet ユーザ数（Num_of_Users）においては多重共線性が生じている可能性,平均気温(temperature),平均風速(wind_velocity)の説明変数は有効性への疑問が考えらえる．これらを考慮した 4 つのモデルを作り，その結果を分析・比較する.

4.3.1 重回帰モデル（Model1）の評価

Model0 の説明変数から Tweet 数（Num_of_Tweets）と平均気温（temperature）および平均風速（wind_velocity）を除いたモデルを Model1 とし，そのモデル式を式 4-2 に示す．得られたモデルの各種統計量を表 4-4，偏回帰係数の詳細および GVIF の結果を表 4-5，表 4-6 にまとめる.

$$\boldsymbol{spectators} = \boldsymbol{home\_team}(*) + \boldsymbol{day}(*) + \boldsymbol{Num\_of\_Users} + \boldsymbol{opposite}(*) \quad (式\ 4\text{-}2)$$

表 4-4　重回帰モデルの統計量（Model1）

決定係数	0.7368
調整済み決定係数	0.7267
AIC	5388.855
BIC	5438.085
Residual standard error: 3811 on 313 degrees of freedom	
F-statistic: 73.02 on 12 and 313 DF, p-value: < 2.2e-16	

表 4-5　得られた偏回帰係数等の詳細情報（Model1）

変数名	推定値	標準誤差	95%CI[2.5%, 97.5%]	t 値	p 値	
(Intercept)	13476.16	1074.96	[11361.0890, 15591.2294]	12.536	< 2e-16	***
home_teameagles	3271.65	846.90	[1605.3028, 4937.9889]	3.863	0.000136	***
home_teamfighters	9607.17	939.88	[7757.8948, 11456.4407]	10.222	< 2e-16	***
home_teamhawks	14637.33	795.49	[13072.1437, 16202.5082]	18.400	< 2e-16	***
home_teamlions	2897.73	776.03	[1370.8317, 4424.6383]	3.734	0.000224	***
home_teammarines	-3053.85	811.14	[-4649.8246, -1457.8790]	-3.765	0.000199	***
dayweekend	5266.10	475.58	[4330.3730, 6201.8298]	11.073	< 2e-16	***
Num_of_Users	110.33	15.61	[79.6206, 141.0442]	7.069	1.02e-11	***
oppositeeagles	2362.20	733.83	[918.3323, 3806.0673]	3.219	0.001421	**
oppositefighters	2308.86	726.16	[880.0837, 3737.6370]	3.180	0.001623	**
oppositehawks	2226.18	744.06	[762.1884, 3690.1647]	2.992	0.002993	**
oppositelions	2129.96	742.57	[668.8925, 3591.0271]	2.868	0.004407	**
oppositemarines	154.70	763.45	[-1347.4394, 1656.8385]	0.203	0.839554	

表 4-6　GVIF の算出結果（Model1）

変数名	GVIF	Df	GVIF(1/(2*Df))
home_team	2.975537	5	1.115210
day	1.250211	1	1.118128
Num_of_Users	2.623588	1	1.619749
opposite	1.411179	5	1.035043

4.3.2 重回帰モデル（Model2）の評価

Model1 の説明変数からさらに対戦相手（opposite）を除いたモデルを Model2 とし，そのモデル式を式 4-3 に示す．得られたモデルの各種統計量を表 4-7，偏回帰係数の詳細および GVIF の結果を表 4-8，表 4-9 にまとめる．

$$\boldsymbol{spectators} = \boldsymbol{home\_team}(*) + \boldsymbol{day}(*) + \boldsymbol{Num\_of\_Users} \qquad \text{(式 4-3)}$$

表 4-7　重回帰モデルの統計量（Model2）

決定係数	0.7180
調整済み決定係数	0.7118
AIC	5401.38
BIC	5431.67
Residual standard error: 3914 on 318 degrees of freedom	
F-statistic: 115.7 on 7 and 318 DF, p-value: < 2.2e-16	

表 4-8　得られた偏回帰係数等の詳細情報（Model2）

変数名	推定値	標準誤差	95%CI[2.5%, 97.5%]	t 値	p 値	
(Intercept)	15841.08	918.28	[14034.40953, 17647.747]	17.251	< 2e-16	***
home_teameagles	2590.14	834.18	[948.93804, 4231.342]	3.105	0.002074	**
home_teamfighters	8684.99	934.08	[6847.23158, 10522.745]	9.298	< 2e-16	***
home_teamhawks	14048.64	800.22	[12474.24901, 15623.024]	17.556	< 2e-16	***
home_teamlions	2293.98	780.21	[758.94977, 3829.001]	2.940	0.003520	**
home_teammarines	-2922.04	807.86	[-4511.47333, -1332.605]	-3.617	0.000347	***
dayweekend	5447.94	481.55	[4500.50478, 6395.376]	11.313	< 2e-16	***
Num_of_Users	99.45	15.08	[69.77935, 129.128]	6.594	1.78e-10	***

表 4-9　GVIF の算出結果（Model2）

変数名	GVIF	Df	GVIF(1/(2*Df))
home_team	2.180376	5	1.081068
day	2.322628	1	1.524017
Num_of_Users	2.623588	1	1.619749

4.3.3 重回帰モデル（Model3）の評価

Model0 の説明変数から Tweet ユーザ数（Num_of_Users）と平均気温（temperature）および平均風速（wind_velocity）を除いたモデルを Model3 とし，そのモデル式を式 4-4 に示す．得られたモデルの各種統計量を表 4-10，偏

回帰係数の詳細および GVIF の結果を表 4-11，表 4-12 にまとめる．

$$\boldsymbol{spectators} = \boldsymbol{home\_Team}(*) + \boldsymbol{day}(*) + \boldsymbol{Num\_of\_Tweets} + \boldsymbol{opposite}(*) \text{(式 4-4)}$$

表 4-10　重回帰モデルの統計量（Model3）

決定係数	0.7156
調整済み決定係数	0.7047
AIC	5414.069
BIC	5463.299
Residual standard error: 3961 on 313 degrees of freedom	
F-statistic: 65.64 on 12 and 313 DF, p-value: < 2.2e-16	

表 4-11　得られた偏回帰係数等の詳細情報（Model3）

変数名	推定値	標準誤差	95%CI[2.5%, 97.5%]	t 値	p 値	
(Intercept)	15065.75	1105.37	[12890.85572, 17240.64008]	13.630	< 2e-16	***
home_teameagles	2069.92	845.75	[405.84390, 3733.99590]	2.447	0.014937	*
home_teamfighters	8336.29	951.71	[6463.73045, 10208.84285]	8.759	< 2e-16	***
home_teamhawks	13988.26	815.86	[12382.99872, 15593.52884]	17.145	< 2e-16	***
home_teamlions	2875.89	822.30	[1257.95081, 4493.82470]	3.497	0.000538	***
home_teammarines	-2440.76	840.99	[-4095.46834, -786.05497]	-2.902	0.003968	**
dayweekend	5726.39	492.09	[4758.16318, 6694.62195]	11.637	< 2e-16	***
Num_of_Tweets	35.45	7.40	[20.89057, 50.01168]	4.791	2.57e-06	***
oppositeeagles	2303.77	769.76	[789.20998, 3818.32258]	2.993	0.002984	**
oppositefighters	2495.63	759.98	[1000.31182, 3990.95353]	3.284	0.001140	**
oppositehawks	2655.12	770.28	[1139.52728, 4170.70654]	3.447	0.000645	***
oppositelions	2437.61	770.21	[922.17332, 3953.04897]	3.165	0.001704	**
oppositemarines	727.85	785.38	[-817.44558, 2273.14939]	0.927	0.354772	

表 4-12　GVIF の算出結果（Model3）

変数名	GVIF	Df	GVIF(1/(2*Df))
home_team	2.684902	5	1.103806
day	1.238940	1	1.113077
Num_of_Tweets	2.378124	1	1.542117
opposite	1.393610	5	1.033747

4.3.4 重回帰モデル（Model4）の評価

Model3 の説明変数からさらに対戦相手（opposite）を除いたモデルを Model4 とし，そのモデル式を式 4-5 に示す．得られたモデルの各種統計量を表 4-13，偏回帰係数の詳細，および GVIF の結果を表 4-14，表 4-15 にまとめる．

$$\boldsymbol{spectators} = \boldsymbol{home\_team}(*) + \boldsymbol{day}(*) + \boldsymbol{Num\_of\_Tweets} \qquad (式\ 4\text{-}5)$$

表 4-13　重回帰モデルの統計量（Model4）

決定係数	0.6968
調整済み決定係数	0.6902
AIC	5424.949
BIC	5455.244
Residual standard error: 4058 on 318 degrees of freedom	
F-statistic: 104.4 on 7 and 318 DF, p-value: < 2.2e-16	

表 4-14　得られた偏回帰係数等の詳細情報（Model4）

変数名	推定値	標準誤差	95%CI[2.5%, 97.5%]	t 値	p 値	
(Intercept)	17645.168	904.107	[15866.38027, 19423.95554]	19.517	< 2e-16	***
home_teameagles	1487.608	833.780	[-152.81469, 3128.03038]	1.784	0.07535	.
home_teamfighters	7389.106	938.880	[5541.90491, 9236.30659]	7.870	5.61e-14	***
home_teamhawks	13343.498	816.062	[11737.93546, 14949.06060]	16.351	< 2e-16	***
home_teamlions	2200.212	821.837	[583.28682, 3817.13752]	2.677	0.00781	**
home_teammarines	-2409.784	839.768	[-4061.98689, -757.58161]	-2.870	0.00439	**
dayweekend	5891.654	497.882	[4912.09408, 6871.21358]	11.833	< 2e-16	***
Num_of_Tweets	30.675	7.178	[16.55338, 44.79633]	4.274	2.54e-05	***

表 4-15　GVIF の算出結果（Model4）

変数名	GVIF	Df	GVIF(1/(2*Df))
home_team	1.984944	5	1.070964
day	1.208582	1	1.099355
Num_of_Tweets	2.131863	1	1.460090

4.3.5 重回帰モデル式の比較と評価

ここまでに検討した Model0 から Model4 までの 5 つの重回帰モデルを評価する．評価は表 4-16 に示す手法および評価指標を用いて，総合的におこなう．

表 4-16　モデルの比較に用いた手法と評価指標

手法	説明
クロスバリデーション - MSE - Adjusted.MSE	データを重複しない k 個に分割．k-1 個のデータにモデルを適合させ，k 個目のデータの予測に利用する．k 個に分割されたデータそれぞれをテスト事例として k 回検証を行う．k 回の結果から平均平方誤差と調整済み平均平方誤差の値を求める． MSE：Mean Squared Error ／平均二乗誤差を意味する $$MSE = \frac{1}{k}\sum_{i=1}^{k}\left(y^{(i)} - \hat{y}^{(i)}\right)^2$$
ANOVA（分散分析）	残差二乗和（RSS）による評価．変数を追加するごとに改善してしまうことに注意が必要
AIC（赤池情報量規準）	モデルの複雑性に対してペナルティを与える指標． 値が小さいモデルのほうが，適合が良好と判断する $$AIC = -2\ln(\mathcal{L}) + 2p$$ $\ln(\mathcal{L})$は最大の対数尤度，pはモデル中の係数の数
決定係数（R squared） 自由度調整済み決定係数	目的変数の分散のうち，説明変数で説明できた割合（分散説明率）を表す．値が 1 に近いほど，説明変数で目的変数を良く説明できていることになる．説明変数が多くなる分，決定係数を下方修正した値を自由度調整済み決定係数と呼び，今回は本値を用いる．
MAPE（平均絶対誤差率）	実測値から予測値を引いた値の絶対値を実測値で割った絶対誤差率を平均したものであり，定義式は以下となる． $$MAPE = \frac{100\%}{n}\sum_{t}^{n}\left\lvert\frac{y_t - \hat{y_t}}{y_t}\right\rvert$$ 実測値と予測値の算出には LOOCV を用いた．
LOOCV （ Leave-One-Out Cross-Validation）	データセットから 1 つの個体を除いて学習を行い，学習データに用いていない 1 つの個体で判別モデルの評価を行う．

Model0 から Model4 を評価した結果を図 4-1 に示す．ステップワイズ法で導かれた Model0 において偏回帰係数として有意ではなく，Tweet ユーザ数と多重共線性の疑いの強い Tweet 数（Num_of_Tweets）および偏回帰係数として有意でない平均気温（temperature）と平均風速（wind_velocity）を Model0 の説明変数から除いたモデルである Model1 が他のモデルと比べて，良い結果が得られている．ここでは各モデルにおける説明変数の有効性とデータ処理コストを考慮した結果，Model1 を最良のモデルとした．

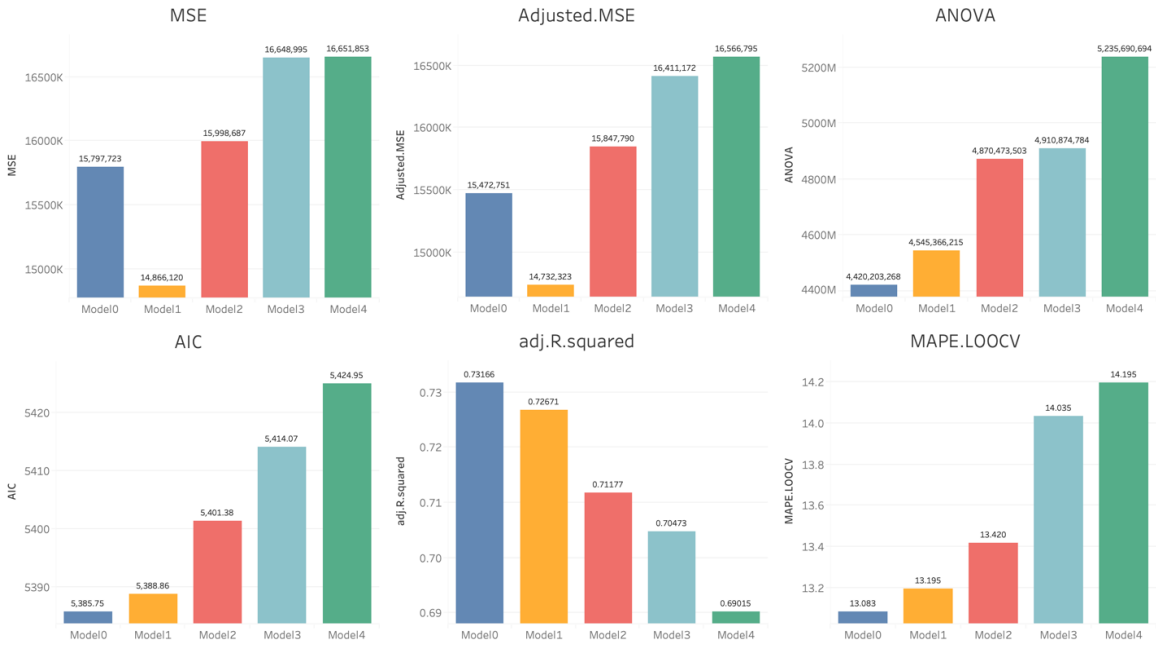


図 4-1　モデルの比較結果①（Model0~4）

4.4 Twitter データを用いた回帰分析の有効性

前節で得られた Model1 から Tweet ユーザ数（Num_of_Users）を除いたモデルを Model5 とし，そのモデル式を式 4-6 に示す.

Model1 と Model5 を比較することで，観客数を推定するための説明変数として Tweet ユーザ数（Num_of_Users）が有効であるかどうかを評価する．式 4-6 から得られたモデルの各種統計量を表 4-17，偏回帰係数の詳細，および GVIF の結果を表 4-18，表 4-19 にまとめる.

$$\boldsymbol{spectators} = \boldsymbol{home\_team}(*) + \boldsymbol{day}(*) + \boldsymbol{opposite}(*) \qquad \text{(式 4-6)}$$

表 4-17　重回帰モデルの統計量（Model5）

決定係数	0.6948
調整済み決定係数	0.6841
AIC	5435.136
BIC	5480.579
Residual standard error: 4097 on 314 degrees of freedom	
F-statistic: 64.98 on 11 and 314 DF, p-value: < 2.2e-16	

表 4-18　得られた偏回帰係数等の詳細情報（Model5）

変数名	推定値	標準誤差	95%CI[2.5%, 97.5%]	t 値	p 値	
(Intercept)	18562.5	858.6	[16873.109376, 20251.8294]	21.619	< 2e-16	***
home_teameagles	557.9	811.6	[-1038.987504, 2154.7723]	0.687	0.492344	
home_teamfighters	6113.7	859.5	[4422.570713, 7804.8386]	7.113	7.72e-12	***
home_teamhawks	13315.9	831.3	[11680.261715, 14951.5599]	16.018	< 2e-16	***
home_teamlions	1775.6	816.7	[168.726614, 3382.5418]	2.174	0.030441	*
home_teammarines	-1080.8	818.8	[-2691.915543. 530.2541]	-1.320	0.187806	
dayweekend	6740.1	459.5	[5835.990020, 7644.2995]	14.667	< 2e-16	***
oppositeeagles	1542.7	779.1	[9.887835, 3075.5945]	1.980	0.048549	*
oppositefighters	2026.7	779.5	[492.942623, 3560.5438]	2.600	0.009766	**
oppositehawks	2702.2	796.7	[1134.698260, 4269.7361]	3.392	0.000783	***
oppositelions	2477.1	796.6	[909.718401, 4044.5179]	3.110	0.002046	**
oppositemarines	1058.1	809.2	[-534.085142, 2650.3244]	1.308	0.191980	

表 4-19　GVIF の算出結果（Model5）

変数名	GVIF	Df	GVIF(1/(2*Df))
home_team	1.259561	5	1.023345
day	1.009825	1	1.004901
opposite	1.249298	5	1.022508

Model1 および Model5 に対する，各評価指標（表 4-16）の適用結果を図 4-2 にまとめる．調整済み決定係数では Model1 は 0.7267，Model5 は 0.6841 と Model1 が良い結果を示すとともに，one-leave-out クロスバリデーションにより母集団推定精度を比較した結果である平均絶対誤差率においても Model1 では 13.19502%，Model5 では 14.77547%であり，1.58045%の差異が見られた．

また，モデルの複雑性にペナルティを与える指標である AIC や BIC の値においても，Model5 に対して Tweet ユーザ数（Num_of_Users）の説明変数が増えている Model1 のほうが良い結果を示すことから，Tweet ユーザ数を説明変数に加えることの有効性が示された．

図 4-2　モデルの比較結果②（Model1，Model5）

4.5 結言

本章では，第 3 章で示したデータおよび重回帰モデルを用いて，日本国内プロ野球のパ・リーグ 6 球団間で，かつ各チームのメインスタジアムでおこなわれた試合における観客数の推定をおこなった．

まず，観客数を推定するための説明変数として，Twitter データ，プロ野球試合データ，天候データを用い，ステップワイズ法による変数選択などをおこうことで複数の重回帰モデルを構築した．それらのモデルを比較・評価し，最良としたモデル（Model1）では平均絶対誤差率 13.19502%という観客数の推定精度を得ることを示した．

また，Twitter データである Tweet ユーザ数（Num_of_Users）を特徴量として重回帰モデルに加えた場合（Model1）と加えなかった場合（Model5）でのモデル評価をおこない，Tweet ユーザ数（Num_of_Users）を特徴量に加えることでスタジアムの観客数である母集団推定の平均絶対誤差率が 1.58045%向上することを示した．

第5章　ランダムフォレストモデルを用いた特定空間における母集団推定

5.1 緒言

第 3 章で示したデータおよびランダムフォレスト回帰モデルを用いて，日本国内プロ野球のパ・リーグ 6 球団間で，かつ各チームのメインスタジアムでおこなわれた試合における観客数の推定をおこなう．

ランダムフォレストは決定木を弱学習器とするアンサンブル学習の一種である． [Breiman, 2001]によって提案されたものであり，パターン識別をはじめとして，回帰，クラスタリングに利用でき，特徴量のスケーリングに配慮する必要のない，機械学習アルゴリズムである．また広く用いられ，良好な結果が得られるとされている［波部，2012]．

ランダムフォレスト回帰による母集団の推定精度を評価するとともに，説明変数に用いた各特徴量の重要度を測定することで，Twitter データを加える効果を評価する．ランダムフォレストの基礎と動向については，［波部，2016]を参照されたい．

5.2 グリッドサーチによるハイパーパラメータ探索

ランダムフォレストにはいくつかのユーザが選択・決定しなければならないハイパーパラメータと言われる設定箇所が存在する．具体的には,「決定木の個数 T」,「決定木の最大深さ D」,「ランダム性を制御するパラメータ ρ」,「ノードでの分割関数」,「学習時の目的関数」,「適用先に応じた特徴選択」，などがある．

一般的には初期値の設定であっても良い精度がでると言われているが，本研究では，グリッドサーチによるハイパラーパラメータ探索をおこなうことで，設定値を決定した．グリッドサーチでのパラメータ探索項目と範囲および探索

された結果を表 5-1 に示す.

表 5-1　ハイパーパラメータの探索項目と範囲

ハイパーパラメータ	探索範囲	探索結果
n_estimators	10, 50, 100, 300, 500	50
max_depth	2, 3, 5, None	None
max_features	1, 2, 3, 4	4
min_samples_split	2, 5, 10, 15, 20	2
min_samples_leaf	1, 3, 5, 10, 15, 20	3
bootstrap	True, False	True
criterion	mse	mse

5.3 推定精度の評価

前節でのグリッドサーチ により求めたハイパーパラメータ（表 5-1）を設定したランダムフォレスト回帰モデルを用いて，表 5-3 に示した LOOCV （Leave-One-Out-Cross-Validation)を用い，MAPE（平均絶対誤差率)を求めた結果を表 5-2 に示す.

LOOCV とはデータセットから 1 つの個体を除いて学習を行い，学習データに用いていない 1 つの個体で判別モデルの評価をおこなう手法である．推定精度を評価する際の方法として，MAPE（平均絶対誤差率）を用いた.

ランダムフォレスト回帰モデルによる観客数の推定における Mean Squared Error (MSE),決定係数および分類精度は,学習データでは MSE は 6882106.108, 決定係数は 0.869,平均絶対誤差率 8.437,評価データでは MSE は 9792775.607, 決定係数は 0.820，平均絶対誤差率 10.453％を示した.

第 4 章で取り組んだ重回帰モデルを用いた観客数の推定結果と比べると，推定精度は 2.74%ほど向上することを示した.

表 5-2　ランダムフォレスト回帰による推定精度

評価項目	学習データ	評価データ
Mean Squared Error (MSE)	6882106.108	9792775.607
Coefficient of determination (R^2)	0.869	0.820
Mean Absolute Percentage Error (MAPE)	8.437%	10.453%

表 5-3　モデル比較に用いる手法と評価指標

手法	説明
クロスバリデーション - MSE - Adjusted.MSE	データを重複しない k 個に分割．k-1 個のデータにモデルを適合させ，k 個目のデータの予測に利用する．k 個に分割されたデータそれぞれをテスト事例として k 回検証を行う．k 回の結果から平均平方誤差と調整済み平均平方誤差の値を求める． MSE：Mean Squared Error ／平均二乗誤差を意味する $$MSE = \frac{1}{k}\sum_{i=1}^{k}\left(y^{(i)} - \hat{y}^{(i)}\right)^2$$
ANOVA（分散分析）	残差二乗和（RSS）による評価．変数を追加するごとに改善してしまうことに注意が必要
AIC（赤池情報量規準）	モデルの複雑性に対してペナルティを与える指標． 値が小さいモデルのほうが，適合が良好と判断する $$AIC = -2\ln(\mathcal{L}) + 2p$$ $\ln(\mathcal{L})$は最大の対数尤度，pはモデル中の係数の数
決定係数（R squared） 自由度調整済み決定係数	目的変数の分散のうち，説明変数で説明できた割合（分散説明率）を表す．値が 1 に近いほど，説明変数で目的変数を良く説明できていることになる．説明変数が多くなる分，決定係数を下方修正した値を自由度調整済み決定係数と呼び，今回は本値を用いる．
MAPE（平均絶対誤差率）	実測値から予測値を引いた値の絶対値を実測値で割った絶対誤差率を平均したものであり，定義式は以下となる． $$MAPE = \frac{100\%}{n}\sum_{t}^{n}\left\lvert\frac{y_t - \hat{y_t}}{y_t}\right\rvert$$ 実測値と予測値の算出には LOOCV を用いた．
LOOCV （ Leave-One-Out Cross-Validation）	データセットから 1 つの個体を除いて学習を行い，学習データに用いていない 1 つの個体で判別モデルの評価を行う．

5.4 Twitter データに関する特徴量の重要度

ランダムフォレストでは，データが線形分離可能かどうかについても，フォレスト内のすべての決定木から計算された不純度の平均的な減少量として特徴量の重要度を測定できる．特徴量の重要度の測定は scikit-learn の RandomForestRegressor を適合させた後に，featuer_importances_属性を使って取得した．

取得した特徴量の重要度を図 5-1 に示す．特徴量の重要度は合計して 1.0 になるように正規化されている．

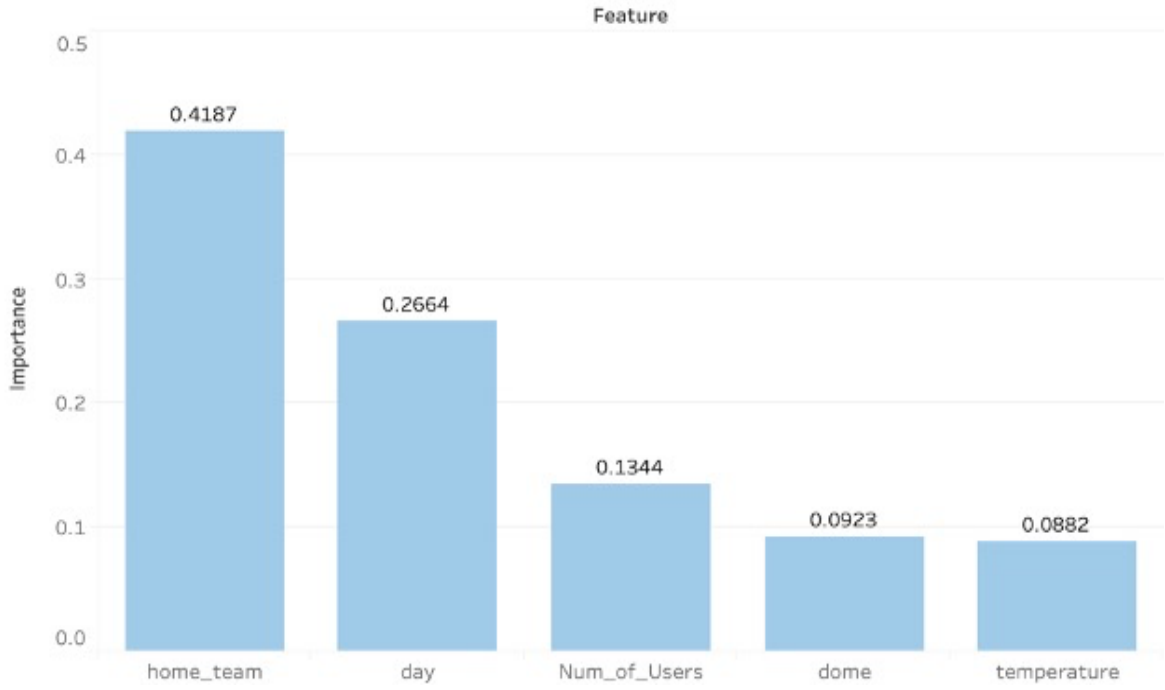


図 5-1　特徴量の重要度（featuer_importances_属性）

特徴量の重要度として，ホームチームの値が 0.4187，試合開催日の祝祭日有無の値が 0.2664 という高い結果を示した．この結果は，3.6 節で示した観客数の増減は，試合が開催されるホームスタジアムおよび試開催日が平日であるか土日祝日であるかが影響することとも一致する．

また，Tweet ユーザ数も 0.1344 という相対的に高い値を示しており，観客数を推定するための特徴量として有効であると考えることができる．

5.5 結言

本章では，第 3 章で示したデータおよびランダムフォレスト回帰モデルを用いて，日本国内プロ野球のパ・リーグ 6 球団間で，かつ各チームのメインスタジアムでおこなわれた試合における観客数の推定をおこなった.

具体的にはグリッドサーチによるハイパーパラメータの選択をおこない，そのパラメータを適用したランダムフォレスト回帰により，観客数を推定した.

ランダムフォレスト回帰モデルを用いた結果，推定精度である平均絶対誤差率は 10.453%で観客数の推定ができることを示した.

また，ランダムフォレストの featuer_importances_属性から Tweet ユーザ数が観客数を推定するうえで特徴量として有効であることを示した.

第6章　特定空間における Twitter 投稿場所の分類

6.1 緒言

Twitter をデータとして活用する際，位置情報が付加されたデータが少ないことが問題になる場合がある．第 2 章で述べたが，ソーシャルメデイアでは GPS により投稿に付加された geotag と呼ばれる位置情報を用いるが，geotag 付きのツイートは全体数の 0.1%から 0.6%程度であると言われており，全体の 1%に満たないほど少ない（［Cheng, et al., 2010]，［山田 & 斉藤, 2010]，［鳥海, 2015]，［橋本 & 岡, 2012]）．

そのため，特定のエリアを指定して Twitter データを収集しても，データ量が少ないため，活用することが難しい場合がある．このため，位置情報が付加されていない多くのツイートの中から,特定空間で発信されたツイートを分類し，抽出することができれば，この問題を克服する一助になるとともに，場所に関連した情報推薦や該当空間に対する利用者の集合的な分析や評価，ツイート数を用いた該当空間における母集団推定への活用など，様々な分野への応用が期待できる．

本章では，　Twitter データのテキスト情報や付随する属性情報を利用し，発信された各ツイートが特定空間から発信されたものであるかを分類することに取り組んだ．対象とする特定空間は，日本国内プロ野球のスタジアムとした．但し，日本国内プロ野球で使用する全 12 球団のメインスタジアムをまとめて，野球スタジアム（Baseball Stadium）というひとつの空間として扱う．

分類手法には自然言語処理モデルの一つである BERT（Bidirectional Encoder Representations from Transformers）［Devlin, et al., 2019]を用いて，分類結果の評価をおこなった．

また，近年，多くのシステムで機械学習が活用され，機械学習モデルの説明可能性が重要となっている．そのため，機械学習モデルの内部の数値データ等を可視化するだけでなく，入力とその予測結果を人間が解釈できる説明方法が求められている［Hind, et al., 2019]．文書分類など自然言語を対象とした機

械学習モデルでは，入力テキストの単語や文およびそれらの関係を用いてモデルの分類結果を説明する必要がある［Lei, et al., 2016].

特定空間内外における投稿内容の違いや傾向を明らかにするため,LIME(Local Interpretable Model-agnostic Explanations) ［Ribeiro, et al., 2016]を用い，分類予測に寄与した単語を抽出・集計することで，投稿内容における特徴分析をおこなった.

ツイートされる内容は，発信された空間に関する情報やツイートの目的に応じた文章構成など，様々な影響が含まれていると考えられる．例えば，特定空間で発信されたツイートはその空間における投稿者の視点や体験を含んだ投稿となり，特定空間外における，特定空間に関連するツイートは，一般的な情報や周知の事実を含んだ投稿となる可能性が考えられる.

6.2 自然言語の情報抽出手法

本節では自然言語処理によるテキストデータからの情報抽出の方法について，その概要を説明する．まず言語の解析の方法について述べたのち，ニューラルネットワークを用いた情報抽出の手法について述べる.

6.2.1 言語の解析方法

形態素解析

文章は，複数の語の系列として表されており，その構造を解析するためには，文を語の単位へ分解する必要がある．この最小の単位の語へ分解するための処理が形態素解析処理である．ここで形態素とは意味を持つ最小の言語単位であり，形態素解析とは与えられた文を形態素単位に区切り，各形態素に品詞などの情報を付与する処理である．日本語における形態素解析では，辞書に記載されている見出し語を形態素と見なして文を形態素に分割する．さらに，見出し語にひもづく品詞や標準形などを形態素に付与する処理となる.

英語のように空白文字によって明確に単語が分かち書きされるような言語に

おいては，その単語の品詞の同定が重要な問題となるが，日本語の解析では，形態素の情報を記述した辞書を参照して形態素解析をするのが普通となる．日本語の難しいところは，分かち書きをしない言語であり，かつ一部の品詞が活用を行うため，最小単位の単語を同定しつつ，その品詞の原形を求める処理となり，構文解析や構造解析といったより高度な自然言語処理の基礎となる非常に重要な処理である．

日本語の代表的な形態素解析の手法として 2 つ挙げられる．1 つは規則による形態素解析である．これは事前に用意しておいた規則や辞書と形態素解析をしたい入力文とを照らし合わせ，どの規則や品詞にマッチするのかを延々と試していき分解していくものである．しかしながら，このやり方では，規則や辞書に全通り記載する必要があり，また総当たりとなる可能性もあるため時間がかかる．そこで近年では，確率的言語モデルによる形態素解析が主流となっている．この手法では，単語の表記，品詞，活用などの情報を記述した単語辞書と，どのような単語または，品詞および活用が日本語文中で連続して出現しうるかを記述した連接可能辞書を用いる．

具体的な例として，図 6-1 に「ねたら元気になった」という日本語分の形態素解析の様子を示す．まず，単語辞書を参照して，入力文の各位置からはじまる部分文字列で辞書にマッチするものをすべて取り出し，その各候補に対応するノードを作成する．さらに入力文の各位置において，その位置までの語の候補と，その位置からの語の候補が連携するかどうかを，連接可能辞書を用いて調べる．連接可能であれば，リンクする．このような処理によって，1 文の形態素解析の可能性は図 6-1 に示すようなラティス構造で表現される．

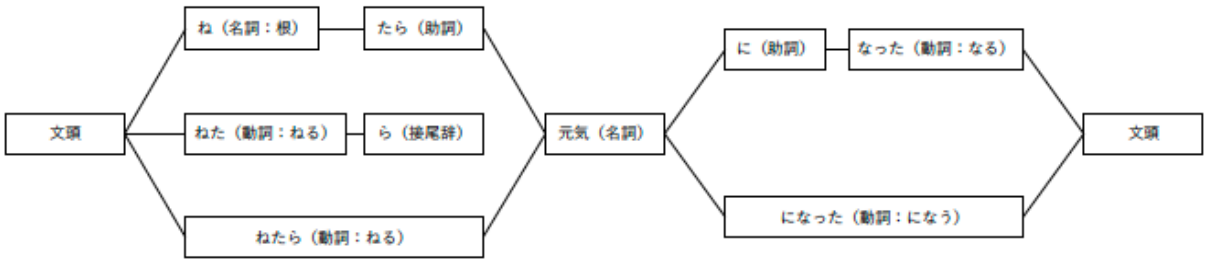


図 6-1　ラティス構造の例

次にこのすべてのパスのうち，どれが尤もらしいのかを統計的に決定するのである．適切なパスの選択方法の１つとしてとしてビタビアルゴリズムが用いられる．このアルゴリズムは，語やその連接にコストを与えて，パスのコストの和が最小になるようにパスを選択するというものである．このコストを与える方法として，コーパスに基づいて学習する方法などがある．単純な分であればパスは限られているため，すべて計算すればよいが複雑な文になるとその計算通りは爆発的に増えることになり，総当たりでは解けない．そこで，ビタビアルゴリズムはダイナミックプログラミングの考え方に基づいて効率的に問題を解決している．

その他にも，隠れマルコフモデルや条件付確率場という手法も用いられる．さらに近年ではニューラルネットワークなどの技術が用いられる．

日本語の形態素解析を行うソフトウェアとしては MeCab や JUMAN といったものがよく利用される．

構文解析

文は 1 次元の語の並びであるが，その中には構文と呼ばれる語の結びつきの構造が存在する．構文解析は，形態素に分けられた単語の系列を用いて，文全体の構造を明らかにするものである．構文は一般的に木構造によって表現することができる．この関係を係り受け構造とよび，係り受け構造を文に自動的に付与することを係り受け解析と呼ぶ.この解析では文脈自由文法が用いられる．文脈自由文法では，句のつながりの規則から係り受け構造を明らかにしようというものである．日本語の係り受け解析のソフトウェアとして Cabocha や KNP が存在する．

意味解析

文の意味は，文を構成する単位である個々の単語の意味と，それらの単語と他の単語との間にある意味的な関係によって決まる．自然言語処理における意味解析の目的は，このような文の意味を形式的に表現することである．意味解析

においては，曖昧さを持たない，解釈や推論の方法を備えている，構文木のような文の形式的な構造から機械的な手順で変換できることが重要となる．前節の構文解析までを完了させることでようやく文章としての意味の解析を行うことができる．

6.2.2 ニューラルネットワークを用いた自然言語処理

近年，様々な分野で機械学習，とりわけニューラルネットワーク，深層学習の利用が注目されている．ニューラルネットワークの利用は自然言語処理においても例外ではなく，その精度の高さから情報抽出のために利用されている．

ニューラルネットワークは生物の神経細胞のふるまいをもとに，モデル化されたアルゴリズムである.1940 年代にはじめて提案され [McCulloch & Walter, 1943],その後も現在に至るまで様々なニューラルネットワークに関するアルゴリズムが提案されてきたものの，これまでその計算量の多さなどからそれほど注目されてこなかった．ところが 2000 年代に入り，情報技術の発展に伴い，マシンパワーの増大や，ビッグデータと呼ばれる大量データの利用が進み，再び注目されるようになった．特に 2010 年代に入り，画像認識，音声認識などの様々なタスクで大きな精度向上が見られたところで大きな脚光を浴びることとなった．自然言語処理においてもニューラルネットワークが 2010 年以降盛んに用いられている．

まず基本的なニューラルネットワークについて述べたのち，自然言語処理で注目されるようになった単語の分散表現についてのべ，自然言語処理においても用いられる再帰型ニューラルネットワークと畳み込みニューラルネットワークについて述べる．

ニューラルネットワーク

ニューラルネットワークにおける基本的な仕組みを図 6-2 に示す．単層のニューラルネットワークは入力を受け付ける N 個の入力ノード（ニューロンとも呼ぶ）とそれを集約して出力する 1 つの出力ノードからなり，入力ノードの

1 つ 1 つが，それぞれ出力ノードと重み w_i $(i = 1 \ldots N)$のエッジでつながっている．入力ノードはそれぞれ入力値 x_i が与えられる．出力ノードではまず入力と重みで線形和として集約し，さらにバイアス項を足し合わせる．これを

$$\boldsymbol{u = \sum_i w_i x_i + b} \qquad \text{(式 6-1)}$$

として定義する．ここで，b はバイアスである．この値 u があるしきい値を超えた場合にのみ出力を行う．この関数を活性化関数と呼ぶ．出力ノードでの出力値 out はこの活性化関数を f として下記であらわされる．

$$\boldsymbol{out = f(u)} \qquad \text{(式 6-2)}$$

活性化関数には様々な種類があるが，のちに説明する誤差逆伝播法を用いる場合には微分可能であることが求められ，シグモイド関数などがよく利用される．

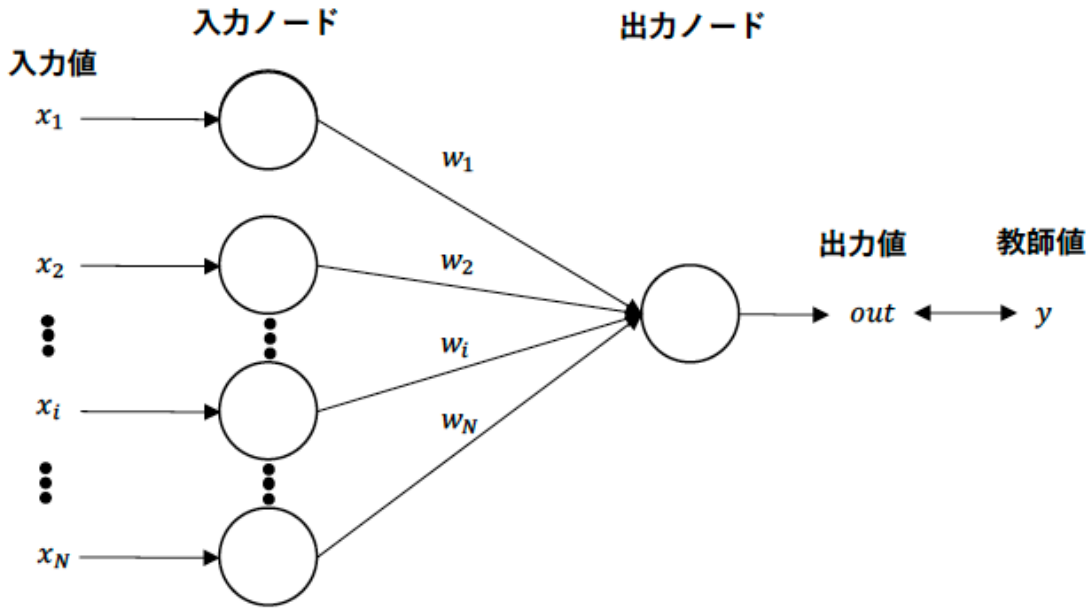


図 6-2　単純なニューラルネットワーク

今日用いられるニューラルネットワークは図 6-2 のように単層ではなく，入

力層と出力層の間に複数の中間層（隠れ層ともよばれる）を含めることが多い．中間層を増やすほどニューラルネットワークの表現力がまし，複雑な処理を行うことが可能となる一方で，その計算量が増大してしまう点には注意が必要である．中間層を含んだニューラルネットワークを図 6-3 に示す．これを順伝播型ニューラルネットワークとも呼ぶ．この場合，層の数は M であり，出力値の数は O となる．各層のノード数は任意であり，かつ各層のノードはその次の層のノードとすべてエッジでつながっているとする．

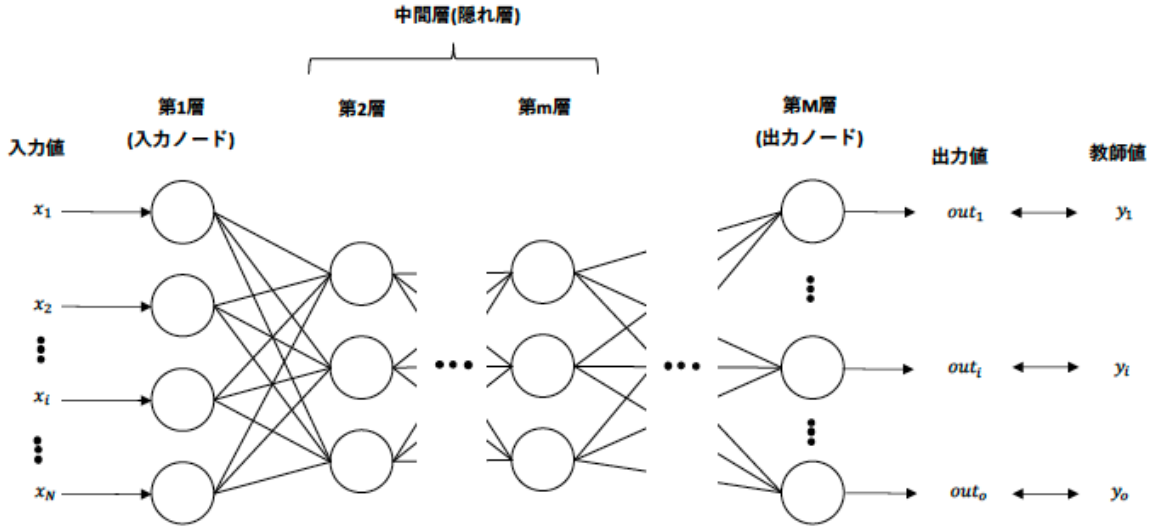


図 6-3　中間層を含んだニューラルネットワーク

ニューラルネットワークの学習において解くべき問題は，入力値 x に対応する教師信号が与えられた際に出力 out が y に近づくように誤差を最小化する最適化問題となる．具体的には，この誤差を最小化するように重み w とバイアス b を求めることになる．簡単のためにバイアスを今後は考慮しない（もしくは常に 1 を出力するノードを各層に加えてバイアスを重みの中に加えてもよい）．

この最適化の解法には誤差逆伝播法（バックプロパゲーション）が用いられることが多い．具体的に図 6-3 での誤差逆伝播法について述べる．誤差を二乗誤差で定義する場合，出力値と教師値の誤差Eは下記で表される．

$$\boldsymbol{E = \frac{1}{2}\sum_{i=1}^{o}(y_i - out_i)^2} \qquad \text{(式 6-3)}$$

勾配降下法を用いて，この E が減少するように各ノード間の重みを更新する．第 $m-1$ 層のノード数が L_{m-1} であり，第 m 層のノード数が L_m であるとする．第 $m-1$ 層の i 番目のノードから第 m 層の j 番目のノードをつなぐエッジの重みを $w_{j,i}^{m,m-1}$ のように表記すると，この重みは以下のように更新される．

$$\boldsymbol{w_{j,i}^{m,m-1(new)} = w_{j,i}^{m,m-1(old)} - \eta \frac{\partial \mathrm{E}}{\partial w_{j,i}^{m,m-1(old)}}} \qquad \text{(式 6-4)}$$

ここで，第二項の偏微分は合成微分によって，m 層の j 番目のノードでの線形和 u_j^m と活性化関数による出力 out_j^m を用いて， 以下のようになる．

$$\boldsymbol{\frac{\partial E}{\partial w_{j,i}^{m,m-1}} = \frac{\partial E}{\partial out_j^m}\frac{\partial out_j^m}{\partial u_j^m}\frac{\partial u_j^m}{\partial w_{j,i}^{m,m-1}}} \qquad \text{(式 6-5)}$$

活性化関数は微分可能であるので，第二項の偏微分は入力値が決まれば値が求まる．ここで，

$$\boldsymbol{u_j^m = \sum_{i=1}^{L_{m-1}} w_{j,i}^{m,m-1} out_i^{m-1}} \qquad \text{(式 6-6)}$$

$$\boldsymbol{out_j^m = f\left(u_j^m\right)} \qquad \text{(式 6-7)}$$

である．よって第三項は，out_j^{m-1} に等しい．第一項の偏微分は，第 m 層のノード j の出力が第 $m+1$ 層のすべてのノードとエッジでつながっていることを思い出すと，偏微分の連鎖率より，さらに，

$$\frac{\partial E}{\partial out_j^m} = \sum_{k=1}^{L_{m+1}} \frac{\partial E}{\partial u_k^{m+1}} \frac{\partial u_k^{m+1}}{\partial out_j^m}$$
$$= \sum_{k=1}^{L_{m+1}} \frac{\partial E}{\partial out_k^{m+1}} \frac{\partial out_k^{m+1}}{\partial u_k^{m+1}} \frac{\partial u_k^{m+1}}{\partial out_j^m}$$
$$= \sum_{k=1}^{L_{m+1}} \frac{\partial E}{\partial out_k^{m+1}} \frac{\partial out_k^{m+1}}{\partial u_k^{m+1}} w_{k,j}^{m+1,m} \quad \text{(式 6-8)}$$

と書くことが可能である．先ほどと同じくこの第二項は入力値が求まれば求まる．第三項は重みそのものである．第一項は再び第 $m+1$ 層と第 $m+2$ 層がつながっているため，同じように書き下すことができる．これは漸化式であるとみなせば，最後の出力層で，誤差を計算すれば，出力層から入力層に向かって誤差が伝播していき勾配を求めることが可能となる．この出力層から入力層への伝播が誤差逆伝播法と呼ばれる所以である．

次に，実際のニューラルネットワークの学習方法について述べる．最終的なモデルを構築するために使用されるデータは，通常，複数のデータセットに分割されうるが，特に 3 つのデータセットへの分割が一般的に用いられる．訓練データセット，検証データセット，テストデータセットである．訓練データセットは，実際の重み更新のための計算に求められる．この訓練データセットを数十から数百のサンプル数からなるミニバッチに分割をし，それぞれのミニバッチでの平均的な誤差を計算し，重みを更新する．これをミニバッチ学習と呼ぶ．ニューラルネットワークの学習ではミニバッチ学習が用いられることが多い．訓練データセットすべてのサンプルを操作することをエポックと呼ぶ．1 エポックにおける全訓練データでの誤差の総和がさがっていることを確認することで，学習が進んでいるかどうかを確認することができる．2 つ目のデータセットである検証データセットは，モデルのハイパーパラメータ（ニューラルネットワークの中間層の数や各層のノードの数など）を調整しながら，訓練データセットで構築したモデルに対して，偏りのない評価を提供することができる．検証データセットを訓練データセットで構築したモデルに入力し誤差を計測する．この誤差が増加する場合，訓練データセットへの過学習の兆候であると考えられるため，検証データセットは学習の停止判定に利用することができる．

最後に，テストデータセットは，訓練データセット上で構築された最終的なモデルに対して，偏りのない評価を提供するために使用されるデータセットである．

ニューラルネットワークは表現力が高いため,過学習が問題となる.すなわち，訓練データセットに対して適応しすぎることにより，訓練データセットに対しては高い精度となるが，他のデータに対しては精度が大きく下がってしまうことがありうる．過学習を抑えるために，検証データセットでの評価を行うのであるが，その外にも過学習を抑える手法にドロップアウトがある．ドロップアウトでは訓練データセットごとにランダムにノードを選び，それらがないものとして学習を行う．これは多数の異なるニューラルネットワークの結果を平均していることがあり，過学習を抑えることができる

Word Embedding

ニューラルネットワークが自然言語処理に活用された成功例として，word embedding がある．Word embedding とは，ニューラルネットワークを用いて，大規模なコーパスから語の意味のベクトル表現（分散表現）を学習したものである．

Mikolov らによって 2013 年に発表された Word2Vec は，最も有名な word embedding のアルゴリズムである（ [Mikolov, et al., 2013a]， [Mikolov, et al., 2013b]）．ここでは word2vec のアルゴリズムについて説明する．

word2vec が行うことを簡潔に述べると大量のコーパス（wikipedia などがよく用いられる）を用いて,似ている文脈に出てくる単語が同士の距離を近づけ，同じ文脈に出てこない単語同士の距離が遠ざけるように，コーパス内の単語を超空間に配置するということである．このとき超空間上の各単語の座標を取り出したものが分散表現ベクトルとなる．このような表現を得るために，2 つの手法が提案されている．

Continues Bag of Words (CBOW)モデルと skip-gram モデルである．CBOW モデルは，注目する単語を中心に与えられたウィンドウサイズ内の周囲の単語を用いて現在の単語を予測するモデルである．一方，skip-gram モデルは，CBOW と

は逆にある単語を用いて周囲の単語を予測するモデルである．図 6-4 に，CBOW と Skip-gram モデルのアーキテクチャを示す．モデル自体は，階層的ソフトマックス法またはネガティブサンプリング法で学習される．

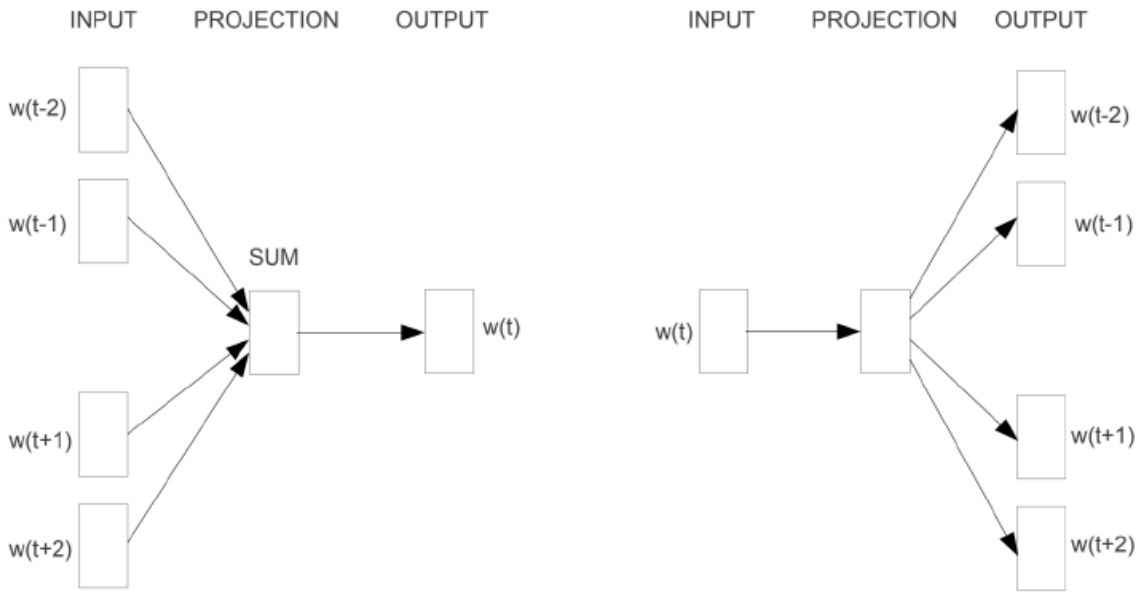


図 6-4　word2vec における CBOW と Skip-gram のアーキテクチャ

[Mikolov, et al., 2013a]より抜粋

密なベクトル表現は，データセット内の単語の同時出現率に基づいて意味関係を抽出する．与えられた 2 つの単語の表現の精度は，モデルがコーパス全体で同じ文脈の中でこれらの単語を何回見たかに依存する．学習中に単語と文脈の共起回数が増えると，隠れた表現が変化し，モデルは将来の予測をより成功させることができ，ベクトル空間における単語と文脈のより良い表現が可能になる．実際に単語がベクトルで表現されることとなるが，異なる単語の類似性などが得られる．よくモデルが学習されていれば

$$\boldsymbol{v_{king} - v_{man} + v_{woman} \cong v_{queen}} \qquad \text{(式 6-9)}$$

という関係が成り立つ．ここで v_{king} は「king」という単語の，v_{man} は「man」という単語の，v_{woman} は「woman」という単語，v_{queen} は「queen」という単語

のベクトル表現である．そのほかにも，似たような意味をもつ単語の分散表現のコサイン距離が近くなるというような関係も成り立つ.このような特徴から，単語のベクトル表現は，自然言語処理分野の様々なタスクで利用されており，word2vec 以外にも，Glove [Pennington, et al., 2014]や fastText [Bojanowski, et al., 2016]といったアルゴリズムが提案されている．

再帰型ニューラルネットワーク

順伝播型ニューラルネットワークでは入力と教師信号のペアは他のペアと独立で，ペア事に中間層の状態をリセットしていた．これに対して，中間層の状態をリセットせずに，次の入力の時に中間層の状態を引き継ぐニューラルネットワークを再帰型ニューラルネットワーク（Recurrent Neural Network: RNN）と呼ぶ．このように前の情報を次の情報へとつなげていくために，時系列情報や連続した情報の解釈に利用される．再帰型ニューラルネットワークのイメージ図を図 6-5 に示す．

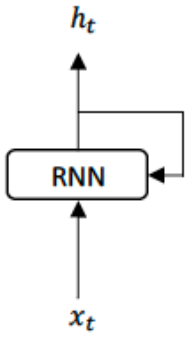


図 6-5　再帰型ニューラルネットワークのイメージ

RNN 層の出力が再び RNN 層の入力になっておりループ構造になっている．このループを時間方向に展開すると，図 6-6 のように右方向に延びるニューラルネットワークとなる．

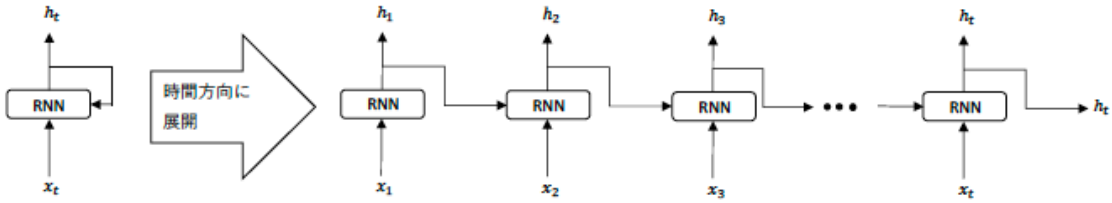


図 6-6　時間方向に展開した再帰型ニューラルネットワークのイメージ

再帰型ニューラルネットワークの伝播は次のように与えられる．

$$\boldsymbol{h_t = f(W_x x_t + W_h h_{t-1} + b)} \qquad \text{(式 6-10)}$$

ここで，h_t は再帰型ニューラルネットワーク層からの出力，W_x は入力値 x に対する重み，W_h は 1 つ前の再帰型ニューラルネットワークからの出力に対する重みであり，b はバイアスとなる．重みは行列，それ以外はベクトルである．さらに f は活性化関数となる．ループを展開した後のネットワークについても，順伝播型ニューラルネットワークと同じような誤差逆伝播法を適用できる．この方法は通時的誤差逆伝播法（back propagation through time，BPTT）と呼ばれる．このような再帰型ニューラルネットワークは，いくつか問題がある．再帰型ニューラルネットワークは一見，浅いニューラルネットワークに見えるものの，実際には，時間方向に展開するため深いネットワークである．したがって，同じ重みに何度も影響されるため，信号や，勾配が消失もしくは爆発してしまためである．この問題を解決するアルゴリズムの 1 つとして，長・短期記憶（Long-Short Term Memory：LSTM）ニューラルネットワークが提案された［Hochreiter & Schmidhuber，1997］．LSTM はゲート付き再帰型ニューラルネットワークとも呼ばれ，一般的な LSTM ユニットは，メモリーセル，入力ゲート，出力ゲート，忘却ゲートからなる．それぞれのゲートは，「入力からメモリーセルへどれくらい情報を通すか」，「メモリーセルからどれくらい出力するか」，「メモリーセルにある情報をどれくらい忘却するか」を制御する．これらのゲートによってメモリーセルの情報を制御し，時系列データの前後関係の重要な情報を長期的に記憶することができる構造となっている．そのほか，同様にゲート付きの再帰型ユニットを用いた GRU（gated recurrent unit）［Cho，

et al., 2014]も存在する．GRU は LSTM でゲートを 1 つ減らした構造となっており，LSTM よりも表現力が低下するものの，計算時間が少ないという特徴がある．

再帰型ニューラルネットワークは時系列情報を扱うのに適しているニューラルネットワークであるが，自然言語も系列であることから，自然言語処理においても，よく利用されるアルゴリズムである．言語モデル学習や，系列ラベリングなど，系列を扱う場合に用いられることが多いが，機械翻訳や文書分類にも利用される．

畳み込みニューラルネットワーク

畳み込みニューラルネットワーク （Convolutional Neural Networks: CNN）とは，全結合していない順伝播型ニューラルネットワークの一種である．動物の視覚野の細胞の反応をベースにした，ニューラルネットワークである．実際にも，動物の視野細胞のニューロンの結合と似たネットワークとなっている．このため，動物の認知と似た性質をもっており，特に画像認識によく利用されるネットワークアルゴリズムである．2012 年のコンピューターによる物体認識の精度を競う国際コンテスト ILSVRC で，畳み込みニューラルネットワークをベースとした AlexNet モデルが優勝し，しかもこれまでの，精度を大幅に更新したとことから,ニューラルネットワークを世に広く知らしめることとなった．

畳み込みニューラルネットワークは，主に，大きくわけて 3 つの層からなる．畳み込み層と，プーリング層と，全結合層である．畳み込み層とは，視覚野のなかでの単純型細胞と呼ばれる細胞を模した役割を果たす．単純型細胞は，ある特定の形状に反応する細胞であり，その種類によって，さまざまな形状に反応する．この単純型細胞が連携して活動することで複雑な形状の物体を認識することができるのである．具体的に畳み込み層は，元の入力に対して，フィルタを作用させ，畳み込み演算を行う．次のプーリング層は，視覚野のなかでの複雑型細胞と呼ばれる細胞を模した役割を果たす．複雑型細胞は視野に入ったものの形状のずれを吸収する作用がある．単純型細胞だけだとある形状が空間的な位置がずれてしまうと，もとの形状と同じであると認識することができな

いが，複雑型細胞があるため，空間的な形状のずれがあっても同一と，認識できるのである．具体的にプーリング層は，畳み込み層の出力をダウンサンプリングする．最大プーリングと呼ばれる手法では，畳み込み層の複数の出力のうちの最大のもののみを選択する．最後の全結合層では，通常の順伝播型ニューラルネットワークと同じく，プーリング層からのすべての出力と，全結合層のノードをつなぐ．これまで 3 次元であったデータを 1 次元に変換する役割を持っている．

畳み込みニューラルネットワークの特徴として，順伝播型ニューラルネットワークと比較して，前の層のノードと次の層のノードが，完全結合しているわけではないため過剰適合となりにくいという利点がある．さらに，再帰型ニューラルネットワークと比較して，並列処理が可能になり，計算効率が高いという点があげられる．

自然言語処理においても，畳み込みニューラルネットワークは利用されている．[Kim, 2014]では，畳み込みニューラルネットワークを用いて，文書の分類タスクを行っている．この論文では文書を画像のように扱っている．例えば，ある文が 9 単語から成り立っており，各単語が word embedding により 6 次元のベクトルで表現されているとすると,9行6列の行列として表記可能となる．これを画像と見立てて畳み込みニューラルネットワークで扱う．画像の場合は畳み込みのフィルタは，水平方向と垂直方向に動作させるが，この文書の行列の場合，行方向が単語の分散表現となるため，列方向，つまり垂直方向下にのみ移動させて，畳み込み計算を行っている．このように処理を行うことで，文脈をとらえることができると考えられる．彼らのモデルの概念図を図 6-7 に示す．

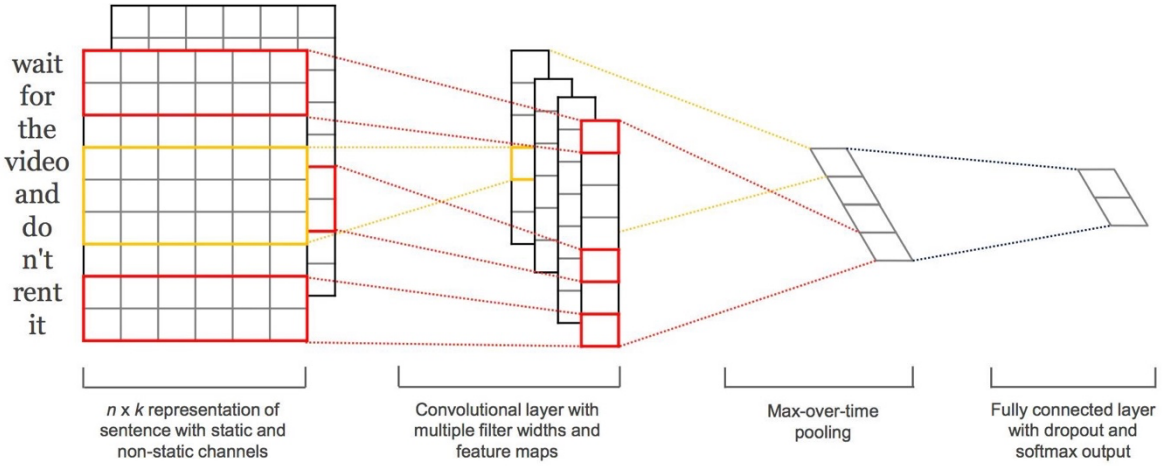


図 6-7 [Kim, 2014]における畳み込みニューラルネットワークの概念図

彼らは，いくつかのタスクで再帰型ニューラルネットワークを含むいくつかの機械学習アルゴリズムの結果を上回る結果となったと報告している.その他，[Zhang, et al., 2015]では，単語ではなく，文字の分散表現を入力値として用いた畳み込みニューラルネットワークのモデルで，いくつかの文書分類の精度が上がると述べている． [Wang, et al., 2016]では，短いテキストに対するセンチメント分析での畳み込みニューラルネットワークと再帰型ニューラルネットワークの合成モデルを提案している.

6.3 SNS を用いた位置情報推定に関する先行研究

SNS を用いて位置情報推定をおこなう研究は多く存在している． [森國, et al., 2015]はツイート投稿位置推定のためのノイズとなる単語の除去手法を提案している．また，推定手法について，投稿内容などを用いて推定をおこなうコンテンツベースの手法とユーザの友人関係などを用いて推定を行うグラフベースの手法の 2 つに大きく分けられると整理している.

コンテンツベースの手法では，コンテンツベースの手法では，推定対象となるエリアごとに単語の出現頻度を学習し，投稿内容に含まれる単語を用いて位置情報を最尤推定することが多く，地理的に狭い範囲にのみ出現する単語を地域語として用いた地域語フィルタ [Cheng, et al., 2010]や TF-IDF [北, et al.,

2002]の概念を用いた地域語フィルタ［三木，et al.，2014]などがある．他にも，Foursquare などの位置情報サービスから投稿されたツイート場所をもとに時間的に近接して発信されたメッセージ中の表現と場所との関係性を学習することを提案したもの［伊川，et al.，2012]がある．

グラフベースの手法では，つながりのあるアカウントの居住地の情報を用いてユーザの居住地を推定する方法があり， Facebook で居住地が記載されたユーザを用いて学習を行い，友人は近くに住んでいるという仮説から，居住地の推定を推定している［Backstrom，et al.，2010]．

その他，位置情報を含むツイートを効率的に発掘する方法として，位置情報サービスから自動投稿されたツイートの収集を提案［服部 & 速水，2011]などがある．

6.4 特定空間に関する Twitter データと特徴分析

本節では，本章において使用するデータの収集方法と抽出したデータおよびそのデータに加えた処理について説明する．

また，特定空間内外における Tweet データの特徴について分析する．

6.4.1 Twitter のデータ収集

特定空間に関連するツイートが該当空間で投稿されたものであるかを分類することに取り組むため，Twitter データを収集する．

Twitter データの収集する対象空間は日本国内プロ野球のスタジアムとし，日本国内プロ野球で使用する全 12 球団のメインスタジアムをまとめて，野球スタジアム（Baseball Stadium）というひとつの空間として扱う．野球スタジアムを特定空間として選んだ理由は，日々，多くの観客，来場者が見込め，多くのツイートがなされているためである．また，研究を母集団推定に発展させる際にも，推定対象となる母集団（試合観客数）が日々公表され，かつ比較的データ数が多いためである．

Twitter データは,Twitter 社が無償で公開している Application Programming

Interface（以下，API）を通じて収集した．Twitterデータの収集はそれぞれの特定空間に対して，ツイート位置の範囲を指定して収集する方法と，特定空間に関係するキーワードを指定して収集する方法の2つを用いた．

一つ目は正解データとなる該当空間でユーザが投稿したTweetデータであり，Twitter データの収集範囲を緯度・経度で範囲指定することで，該当空間で投稿されたツイートのみを収集する．具体的には，検索キーワード「*（ワイルドカード)」，収集対象とする地理的範囲を「スタジアムの中心（各スタジアムの緯度経度で指定）から半径120メートル」という収集条件を設定した．

2つ目が主に不正解データとなる該当空間外でユーザが投稿した Tweet データであり，投稿内容に含まれるキーワードを条件として， Twitterデータを収集する．具体的には収集条件のキーワードを「″ジャイアンツ OR ベイスターズ OR タイガース OR カープ OR ドラゴンズ OR スワローズ OR ライオンズ OR ホークス OR イーグルス OR マリーンズ OR ファイターズ OR バファローズ AND -filter:retweets″」として設定した．

尚，前者の収集方法で集められるツイートは位置情報付きのツイートに限られる．後者の収集方法では位置情報の有無に関わりなく指定したキーワードが投稿内容に含まれるツイートは収集対象となる．しかしながら，位置情報がないツイートは研究において投稿場所の特定ができないことから，位置情報が付与されているTweetデータのみ利用し，位置情報なしのTweetデータについては，最終的に除外する．

6.4.2 Twitterデータの収集結果

ツイート収集の範囲を日本国内プロ野球で使用する全12球団のメインスタジアムに限定し収集したTwitterデータの概要を表6-1に，日本国内プロ野球に関係するキーワードを指定して収集したTwitterデータの概要を表6-2に示す．

表 6-1　範囲を指定して収集した Twitter データの概要

項目	Twitter データ概要
ツイート収集期間	2020/05/31~2020/10/10
ツイート収集条件	緯度経度を用いた 収集範囲の指定
ツイート収集件数	27,013
位置情報付加率（%）	100%
メディア情報添付数（枚数／ツイート）	0.79

表 6-2　キーワードを指定して収集した Twitter データの概要

項目	Twitter データ概要
ツイート収集期間	2020/05/31~2020/10/10
ツイート収集条件	キーワードを用いた 収集対象の指定
ツイート収集件数	2,991,700
位置情報付加率（%）	1.49%
メディア情報添付数（枚数／ツイート）	0.11

表 6-1 と表 6-2 を比較すると，該当空間で発信された Twitter データを集めた場合では，ツイートに対する平均メディア添付枚数は 0.79 枚となり，キーワードで収集した Twitter データと比較して 7 倍ほど高い傾向が見られた．また，Table 2 より，ツイートに対する位置情報が付与されている率については，キーワードで収集した Twitter データでは，1.49%であり，Twitter データ全体における位置情報付与の割合である 0.1% - 0.6%より大きい傾向が見られた．

6.4.3 分類器の学習・評価に用いる基礎データの整備

表 6-1 及び表 6-2 で示したデータ間においては，重複したデータが含まれるため，結合のうえ，重複データを取り除く処理をおこなった．また，位置情報付きデータの中には，位置情報を利用した SNS である Foursquare などから連

携された投稿が含まれる．このようなツイートには「I’m at」や「場所:」などの共通した文字列が含まれており，投稿者が書き込んだ文章をもとに分類する本研究においては，直接的な正解情報となり得ることから，該当の言葉が含まれているツイートをデータから取り除く処理を加えた．データを整備するための処理内容及び処理後のツイート数を表 6-3 に示す．

表 6-3　データ整備の処理プロセスと処理後のツイート数

Item	処理後のツイート数
データ整備対象の初期データ	3,018,713
位置情報無しのツイートの削除	71,592
重複ツイートの削除	69,976
″I’m at.″を含むツイートの削除	66,210
″場所:″を含むツイートの削除	64,273
″(@″を含むツイートの削除	61,800
分析対象とするTwitterデータ(Base data)	61,800

6.4.4 特定空間内外におけるツイートの特徴分析

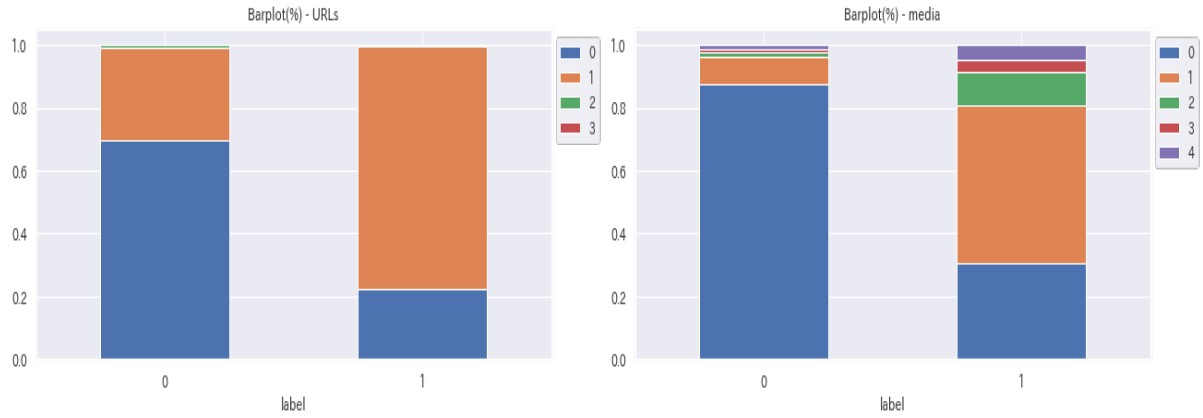
整備した野球スタジアムに関連する 61,800 件の基礎データ（表 6-3 参照）に対して，該当空間から発信されたツイートであるか否かをツイートに付与されている緯度経度と各該当空間（スタジアム）の中心となる緯度経度の情報を用いて，距離を計算し，該当空間内で発信されたものか否かのラベル付けをおこなった．ラベルの定義を表 6-4 に示す．

表 6-4　ラベル定義

ラベル	説明
0	特定空間外から投稿されたツイート
1	特定空間内から投稿されたツイート

特定空間内外でのツイートに対する特徴分析のため，投稿された文章内に含まれている URL の数別の割合及び，ツイートに付与された画像等のメディアの数別の割合を算出し，可視化した．図 6-8(a)に示すとおりツイートに 1 件以上の URL が含まれている割合はスタジアム外の投稿で 30.2%，スタジアム内の投稿で 77.9%であった．また図 6-8(b)に示すとおり，ツイートに 1 件以上のメディアが付与されている割合はスタジアム外の投稿で 12.6%，スタジアム内の投稿で 69.4%であった．

これらの結果は，野球スタジアムという特定空間内からのツイートには投稿者が自身の体験や感想を，自らが撮影した画像などとともに投稿することが多く，特定空間外からのツイートでは，該当空間に関連する情報提供が多く，付与される画像が少ないためと考えることができる．



(a)　　(b)

図 6-8　(a)URL 情報付与率；(b)メディア情報付与率

6.5 分類器の作成と精度評価・要因分析

6.5.1 分析・評価環境

本研究では，ツイートの発信場所が特定空間からであるか否かを推定するための分類器として，自然言語処理モデルの一つである BERT を用いる．BERT とは「Bidirectional Encoder Representations from Transformers (Transformer

による双方向のエンコード表現)」を指し，2018 年 10 月 11 日に Google が発表した自然言語処理モデルである．BERT は Transformer [Vaswani, et al., 2017]をベースとし,まず大規模な生テキストで言語モデルなどの目的関数のもとにモデルを pre-training する．そして各タスクで fine-tuning することにより，様々なタスクで SOTA を更新している．

また，BERT を用いた分類器によるツイート投稿場所の分類精度を評価するとともに，LIME [Ribeiro, et al., 2016]を用いて，分類結果に影響を与えた要因の分析に取り組む．LIME は分類予測の結果に寄与した単語集合を抽出する手法として知られており，テキスト分類だけでなく画像分類や回帰モデルなど様々なタイプのモデルの説明に使われている．LIME では，説明したい予測結果の入力点近傍を説明可能なモデルで近似し，近似モデルの各入力次元の寄与度を用いてモデルの説明をおこなう．数学的には入力空間を Z，説明したい入力座標を z'，説明したいモデルを$f(z)$，説明可能な近似モデルを $g(z)$，入力データの近傍を $\pi_x(z)$ とし，下記の L を最小化する解 $g(z)$ を求める形となる[柳川 & 照井, 2020]， [Ribeiro, et al., 2016]．

$$\boldsymbol{L(f,g,\pi_x) = \sum_{z,z' \in Z} \pi_x(z)\left(f(z) - g(z')\right)^2} \qquad \text{(式 6-11)}$$

BERT によるテキスト分類の予測結果に寄与した特徴量である単語の重要度を数値化し，比較することで，分類予測に影響を与えている要因を明らかにすることができると考える．

6.5.2 学習・評価用データの作成手順

Twitter データを特定空間における投稿であるか否かを分類するための分類器の作成及び評価のためのデータを図 6-9 の手順に従い，作成する．

まず，キーワード指定で収集されたすべてのツイートには，収集のために指定したキーワードのいずれかが含まれている．投稿場所の範囲指定で収集したツイートとの違いになり得ることから，6.4 節で整備した野球スタジアムに関連

する基礎データ(61,800 件)より収集の際にキーワードとして指定した単語を，すべて削除する．

さらに，自然言語処理では一般的にノイズとなる URL の削除処理を施したデータを「data A-1」とし，URL を削除せず，一律に「XXXURLXXX」の文字列に置き換えたデータを「data B-1」とする．URL は一つ一つが文字列として異なり，自然言語処理において扱いが難しいことから，正規表現を用いて，形態素解析処理で分割されない「XXXURLXXX」に置換することで，URL を一つのワードとして処理できる形とした．

「data A-1」及び「data B-1」に対して，各ツイートに付与されているメディア添付数を表 6-5 に示す情報置換を行い，文字列として各ツイートの末尾に加えたデータを「data A-2」及び「data B-2」とする．

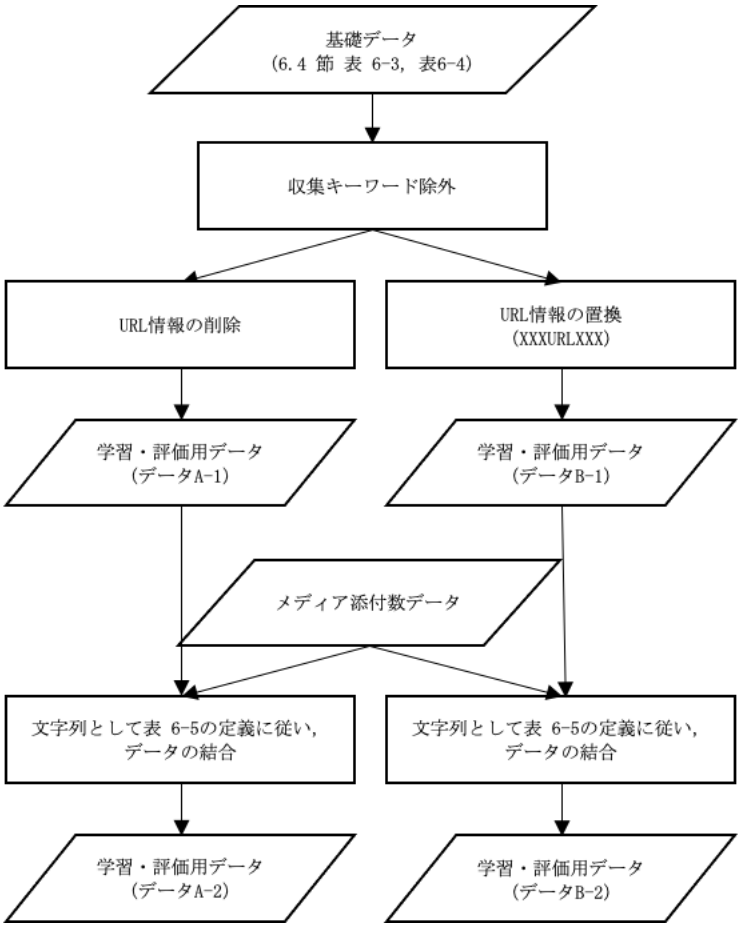


図 6-9　学習・評価用データ作成の手順

表 6-5　ツイートに添付されているメディア数の文章への情報置換

添付メディア数／ツイート	情報置換後の文字列
0	XXXZEROXXX
1	XXXLOWXXX
2	XXXMIDDLEXXX
3 つ以上	XXXHIGHXXX

6.5.3 分類器の作成及び評価の手順

前項で作成した学習・評価用データを用いて，BERT による分類器の作成，分類精度の評価及びLIME を用いた分類結果に影響を与えた要因の分析を,図 6-10 に示す手順でおこなう．野球スタジアムの学習・評価用データを学習データと評価データに分割した結果を表 6-6 に示す.

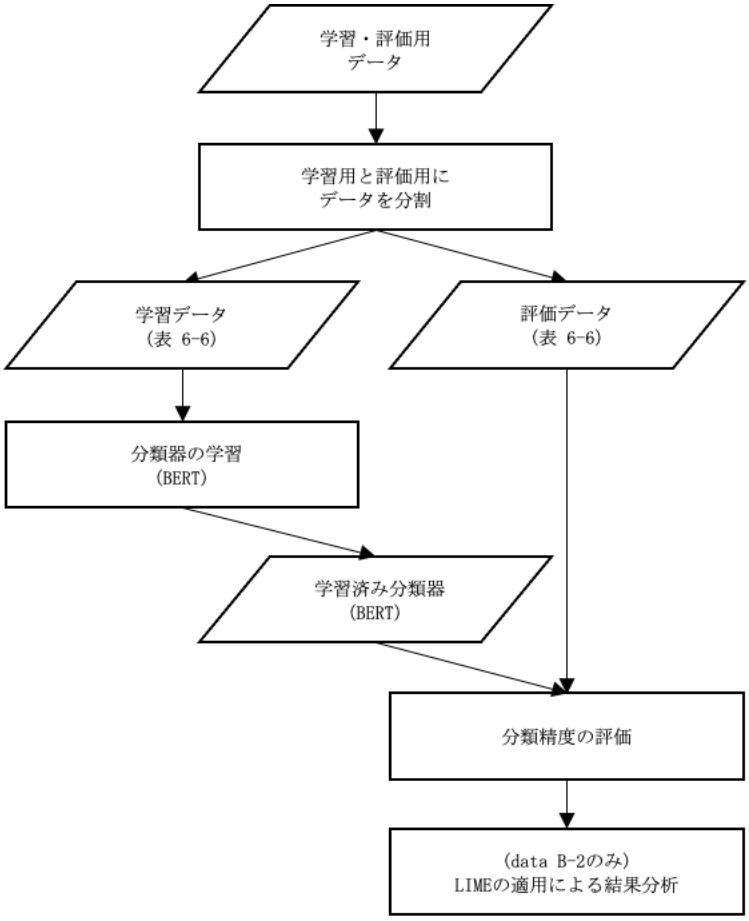


図 6-10　分類器の学習と評価の手順

表 6-6　学習データ数とテストデータ数

Item	ツイート数
学習データ	55,620 件
(breakdown) label = 0	38,068 件
(breakdown) label = 1	17,552 件
評価データ	6,180 件
(breakdown) label = 0	4,230 件
(breakdown) label = 1	1,950 件

6.5.4 分類精度に対する考察

6.4.4 項で示した特定空間内外におけるツイートの特徴である，ツイートあたりの URL 数とメディア数を特徴量（data U/M）とし，機械学習モデルのひとつである「ランダムフォレスト（RF)」を用いて分類した結果と，「data A-1」，「data A-2」，「data B-1」，「data B-2」に対して，図 6-10 で示した手順に従い BERT を用いて分類した結果との比較を表 6-7 およびに図 6-11,図 6-12 に示す.

ランダムフォレストを用いた分類と，ツイート内容を特徴量として組み込み，BERT を用いた「A-1」から「B-2」までの４つの分類では，f1-score [weighted avg]で比較すると，0.8181 から最大で 0.8990 まで結果が向上することを示した．また，ツイート内容を特徴量とした場合でも，ツイートあたりの URL 数とメディア数を文字列としてツイート内容に組み込むことで，分類精度が向上することを示した.

これらの結果より，ツイートの発信元を特定空間内外で分類するために，ツイート内容および 6.4.4 項で示した特定空間内外におけるツイートの特徴であるツイートに付与された URL 数とメディア数の傾向は有効な特徴量あると考えることができる.

表 6-7　特定空間のツイートに対する分類精度

モデル	RF	BERT			
データ（特徴量）	U/M	A-1	A-2	B-1	B-2
accuracy	0.8181	0.8746	0.8985	0.8875	0.8994
f1-score [weighted avg]	0.8167	0.8756	0.8985	0.8864	0.8990

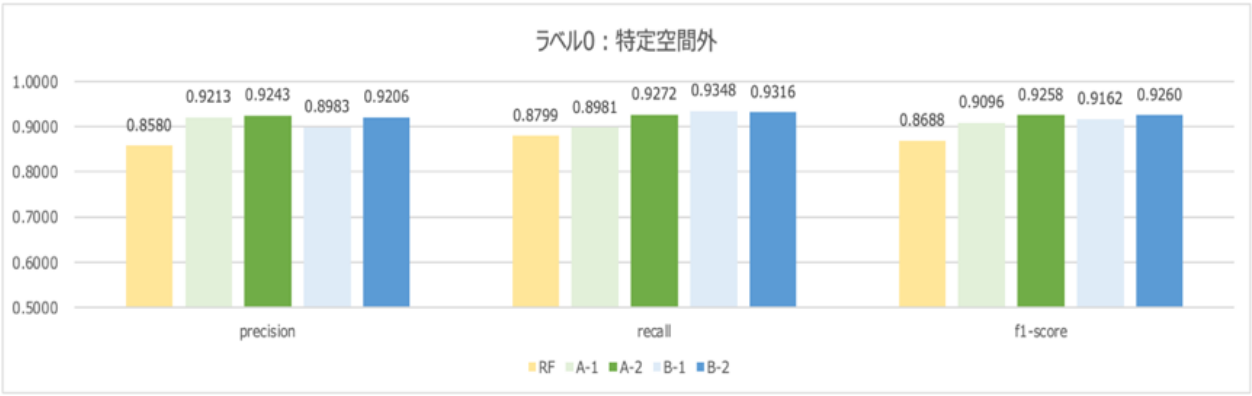


図 6-11　特定空間外のツイートに対する分類精度

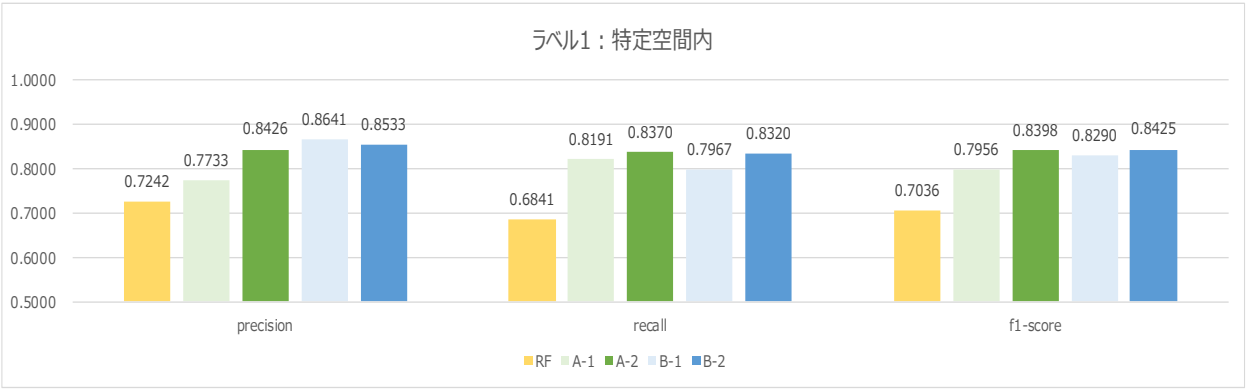


図 6-12　特定空間内のツイートに対する分類精度

6.5.5 分類結果に影響を与える要因の分析

分類結果に影響を与えた要因を分析するため，最も高い分類精度を示した，「data　B-2」を用いて学習した分類器とその評価データ 6180 件のツイートに対して，LIME を用いた判定根拠の可視化を試みた．

各ツイートへの LIME 適用結果から，特定空間内からの投稿であるという予測

にプラスに寄与した重要度の高い単語でかつ品詞が名詞，動詞，形容詞，形容動詞，副詞，感動詞である上位３単語までを取得した．同様にマイナスに寄与した重要度の単語の上位３単語までを取得した．プラス及びマイナスに寄与した単語が 3 つより少ない場合は，得られた単語のみを取得している．

それぞれツイートから得られた単語の出現頻度の合計を求め，LIME を適用したツイート数で割ることで,分類予測に寄与した単語としての出現率を求めた．特定空間内からの投稿であるという予測にプラスに寄与した上位 25 単語および特定空間内からの投稿であるという予測にマイナスに寄与した上位 25 単語を表 6-8 に示す．

特定空間内で投稿されたツイートという判定にプラスに寄与する単語には，「XXXURLXXX, XXXXLOWXXX」といったツイートに付加された URL やメディアの添付を表す単語が上位に見られる．反対に特定空間内で投稿されたツイートという判定にマイナスに寄与する単語では，「XXXURLXXX」の URL を表す単語の出現頻度は小さくなり，「XXXXZELOXXX」というメディアの添付がないことを表す単語が上位に見られた．

これらの結果より，テキスト分類においても,URL 数やメディア数などのツイートの特徴を取り込めていると示唆させる．

表 6-8　LIME 適用結果から導いた分類判定に影響した単語の出現頻度

(Negative/Positive)

Datatype : data B-2					
Negative (−)			Positive (+)		
word	pos	frequency (%)	word	pos	frequency (%)
XXXZEROXXX	名詞	23.71%	XXXURLXXX	名詞	18.54%
し	動詞	6.73%	XXXZEROXXX	名詞	9.21%
XXXURLXXX	名詞	4.64%	0	名詞	8.06%
0	名詞	3.96%	XXXLOWXXX	名詞	7.07%
今日	名詞	2.86%	し	動詞	3.87%
さん	名詞	2.14%	今日	名詞	3.40%
の	名詞	1.84%	ん	名詞	2.35%
てる	動詞	1.75%	ー	名詞	2.31%
ん	名詞	1.72%	さん	名詞	1.86%
ー	名詞	1.54%	い	動詞	1.55%
する	動詞	1.49%	てる	動詞	1.55%
れ	動詞	1.41%	する	動詞	1.26%
い	動詞	1.38%	れ	動詞	1.18%
選手	名詞	1.08%	ここ	名詞	1.13%
見	動詞	1.08%	いい	形容詞	1.12%
絶対	名詞	1.04%	試合	名詞	1.08%
勝っ	動詞	1.02%	選手	名詞	1.08%
て	動詞	1.00%	き	動詞	1.00%
なっ	動詞	0.95%	投手	名詞	0.95%
いい	形容詞	0.92%	くん	名詞	0.89%
勝つ	動詞	0.91%	XXXMIDDLEXXX	名詞	0.83%
いる	動詞	0.87%	見	動詞	0.81%
明日	名詞	0.86%	いる	動詞	0.79%
笑	名詞	0.84%	ない	形容詞	0.76%
よう	名詞	0.83%	来	動詞	0.76%

6.6 結言

本章では，ソーシャルセンサと呼ばれ，ソーシャルメディアの 1 つである Twitter データのテキスト情報や付随する情報を利用し，発信された各ツイートが野球スタジアムという特定空間から発信されたものであるかを分類することに取り組んだ.

分類モデルとして自然言語処理モデルの一つである BERT を用いることで，発信された各ツイートが野球スタジアムという特定空間から発信されたものであるかを f1-score [weighted avg]で最大 0.8990 という高い精度で分類できることを示した.

さらに，特定空間内外でのツイートの特徴を比較し，特定空間内からの投稿ではツイートに付与される URL 数やメディア数が多くなる傾向を明らかにすると

ともに，これら特徴を投稿された文章と組み合わせることで，分類精度が向上することを示した.

また，LIME を用いて分類結果に影響する単語を抽出，比較することで， ツイートに付与されている URL 数やメディア数が分類判定に影響を与えていることを可視化した.

特定空間に関するツイートをキーワードで大量に収集し，その中から同空間内で投稿されたツイートを高い精度で抽出することができれば，該当空間に関する情報の抽出や該当空間におけるツイート数を用いた母集団推定精度の向上[Hara, et al., 2020]， [Hara, et al., 2021a]が期待できる.

第7章　結論

本論文は，ソーシャルセンサと呼ばれ，ソーシャルメディアの１つである Twitter データを特定空間における母集団推定に活用できるかどうかの分析をおこなったものである．

近年，スマートフォンの普及，情報技術の発展により，人々は常時オンラインとなり，様々なデータを発信している．このような状況において，ソーシャルメディアに蓄積されたデータを研究や企業活動に活用することは，不可欠な要素となっている．

そのような中，本論文では，日本国内プロ野球パ・リーグのメインスタジアムで発信された Twitter データを収集，分析することで，該当空間に対する観客数と Twitter データの関係性や特徴を明らかにするとともに，Twitter データを用いて観客数を推定することの有効性を確認している．

また，Twitter データの活用，母集団推定精度の向上に資するため，日本国内プロ野球に関連する Twitter データを収集し，該当空間内での投稿データと該当空間外での投稿データ分離し，それぞれの投稿内容や付加されたデータを比較することで，投稿場所によってツイートに含まれる URL 情報やメディア情報の量が大きく異なることを明らかにした．その上で，明らかにした特徴とディープラーニング（BERT）及び機械学習の解釈手法（LIME）を用いることにより，該当空間でツイートされたものであるかの二値分類を高い精度で実現するとともに，分類に寄与した特徴量を明らかにしている．

本研究は日本国内プロ野球という限定された空間に対しての分析であるものの，これらの結果は Twitter データを活用する上で，Twitter の投稿場所と投稿内容の関係性や特徴を考察する際の重要な情報であり，本研究は実用的な研究成果である．また，Twitter データを用いて，母集団の推定や予測を検証したものはなく，取得が容易であり，ユーザのメッセージが含まれる Twitter データを母集団推定に用いることは，母集団推定だけでなく，その空間に存在する集団の思いや意見など，様々な分析へ拡張できる可能性があり，意義があると考える．

第 3 章では，本研究における母集団推定の全体像と用いるデータについて述

べ,Twitter 及びそのデータの基本について概説した.その後,Twitter データ，プロ野球試合データ及び天候データの収集方法と，それらデータから分析用データを作成する方法について述べた.

Twitter データの収集については緯度経度を指定することで，スタジアム内で出力された Tweet データを正確に取得することができ，また検索キーワードを「*」と設定できることで分析者の主観に基づくキーワード条件を用いないことに意義はあると考えるが，位置情報付きの Tweet データが少なくなることは課題であることを示した.

また，収集したプロ野球データより，観客数の増減は，ホームスタジアムおよび試合開催日が平日か土日祝日であるかの要因が関わっていることを示した.

収集した Twitter データからは，観客数と Tweet 数および観客数と Tweet ユーザ数の相関係数を比較し，前者にくらべて後者の値が大きくなる傾向を示すとともに，Twitter データを用いてプロ野球の試合観客数を推定する場合，Tweet 数よりも Tweet したユーザ数を用いることで推定精度が高くなる可能性を示した.

第 4 章では，第 3 章で収集，加工した Twitter データ，プロ野球試合データ，天候データを用いた，重回帰モデルによる該当空間における母集団推定に取り組んだ．観客数を推定するための説明変数として，ステップワイズ法による変数選択をおこなうとともに，そこから得られた結果を考慮した複数の重回帰モデルを構築した．複数の重回帰モデルの比較，評価をおこない，最良としたモデルでは平均絶対誤差率 13.19502%という観客数の推定精度を得ることを示した.

また，Twitter データである Tweet ユーザ数（Num_of_Users）を特徴量として重回帰モデルに加えた場合と加えなかった場合でのモデル評価をおこない，Tweet ユーザ数（Num_of_Users）を特徴量に加えることでスタジアムの観客数である母集団推定の平均絶対誤差率が 1.58045%向上することを示した.

第 5 章では，第 3 章で収集，加工した Twitter データ，プロ野球試合データ，天候データを用いた，ランダムフォレスト回帰モデルによる該当空間における母集団推定に取り組んだ．ハイパーパラメータの設定値はグリッドサーチにより，探索・選択をおこなった．ランダムフォレスト回帰モデルによる推定精度

は平均絶対誤差率 10.453%であり，重回帰モデルに比べて 2.74%ほど向上することを示した．また，ランダムフォレストの featuer_importances_属性から Tweet ユーザ数が観客数を推定するうえで特徴量として有効であることを示した．

第 6 章では，Twitter データのテキスト情報や付随する情報を利用し，発信された各ツイートが野球スタジアムという特定空間から発信されたものであるかを分類することに取り組んだ．分類モデルとして自然言語処理モデルの一つである BERT を用いることで，発信された各ツイートが野球スタジアムという特定空間から発信されたものであるかを f1-score [weighted avg]で最大 0.8990 という高い精度で分類できることを示した．

さらに,特定空間内外でのツイートを比較し,特定空間外からの投稿に対して，特定空間内からの投稿に付与される URL 情報の割合は 2.6 倍，メディア情報の割合は 5.5 倍と，大きな違いがあることを明らかにするとともに，これら特徴を投稿された文章と組み合わせることで，分類精度が向上することを示した．

また，LIME を用いて分類結果に影響する単語を抽出，比較することで， ツイートに付与されている URL 数やメディア数が分類判定に影響を与えていることを可視化した．

今後の取り組みとして，本研究の一般化，異なる空間に対する適用研究があげられる．そのための一歩として，スタジアム毎に観客数と Twitter データの相関が異なる点に対するさらなる分析が必要と考える．

各スタジアムの観客数と Tweet 数および観客数と Tweet ユーザ数の相関係数は，同じプロ野球の試合であっても図 3-15 で示したとおり，スタジアムごとに違いがあることを示した．また本研究では対象外としていたセ・リーグにおける観客数と Twitter データの関係性を確認するとパ・リーグよりも明らかな違いがあることがわかった．

セ・リーグの観客数をスタジアム別のヒストグラムにしたものが図 7-1 である．第 3 章で示した図 3-6 と比べると，多くのスタジアム／チームにおいてヒストグラムの広がりが小さく，一定の観客数のレンジとなる試合が多くなっていることがわかる．

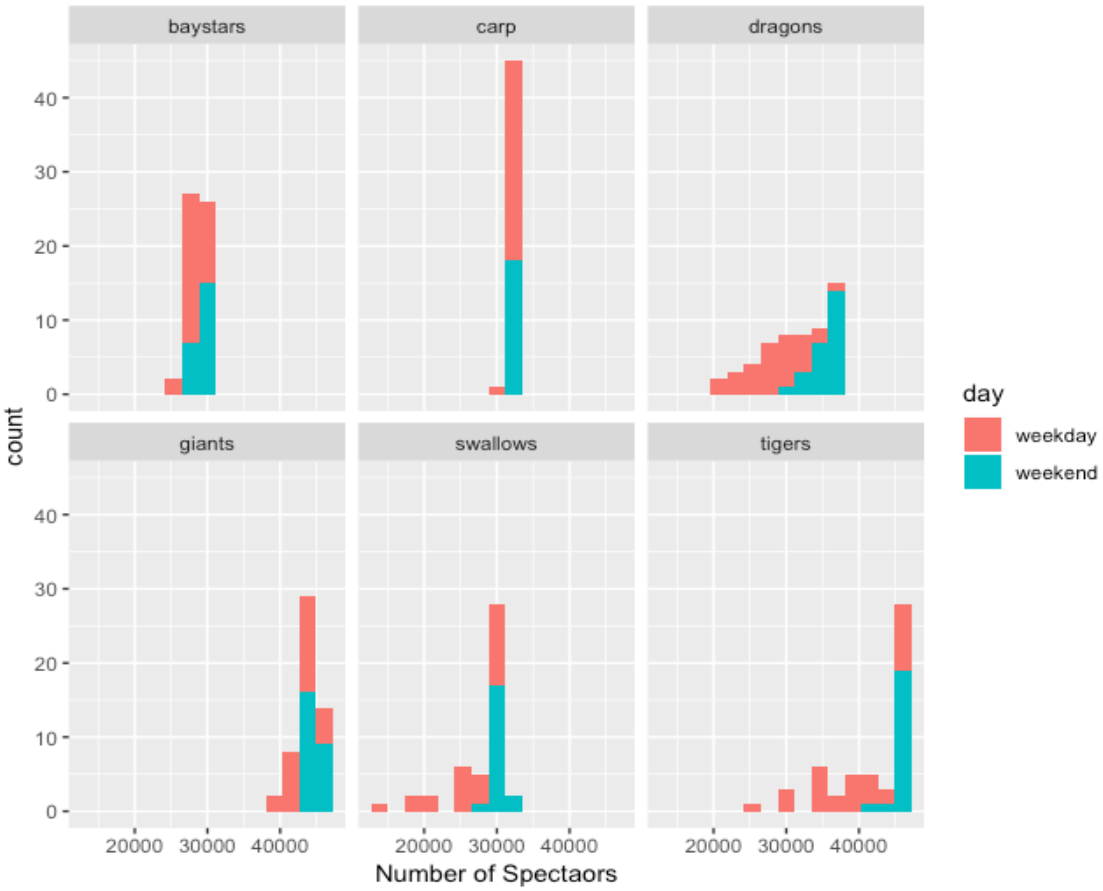


図 7-1　セ・リーグのスタジアム別観客数のヒストグラム

また，セ・リーグの観客数と Tweet 数の散布図，観客数と Tweet ユーザ数の散布図をスタジアム別に示したものが図 7-2，図 7-3 である．阪神甲子園球場でのタイガース戦，ナゴヤドームでのドラゴンズ戦以外は，観客数と Tweet 数，観客数と Tweet ユーザ数のいずれにおいても相関関係がほぼないことがわかる．

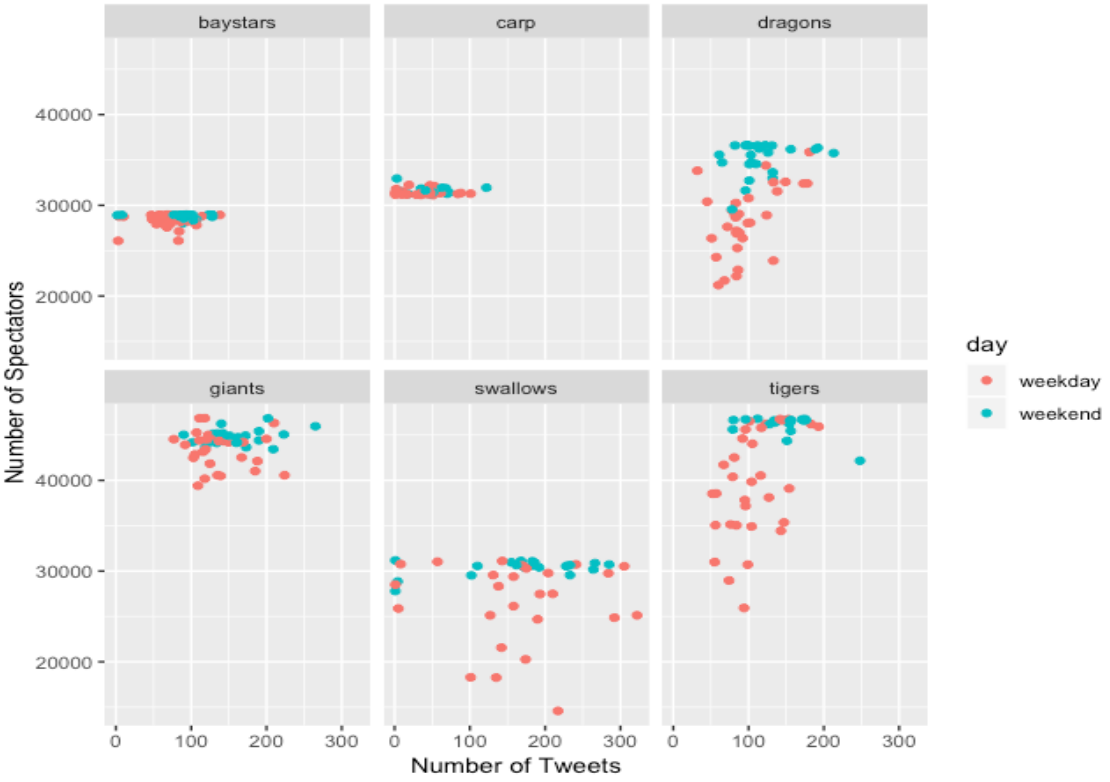


図 7-2　セ・リーグのスタジアム別の観客数と Tweet 数の散布図

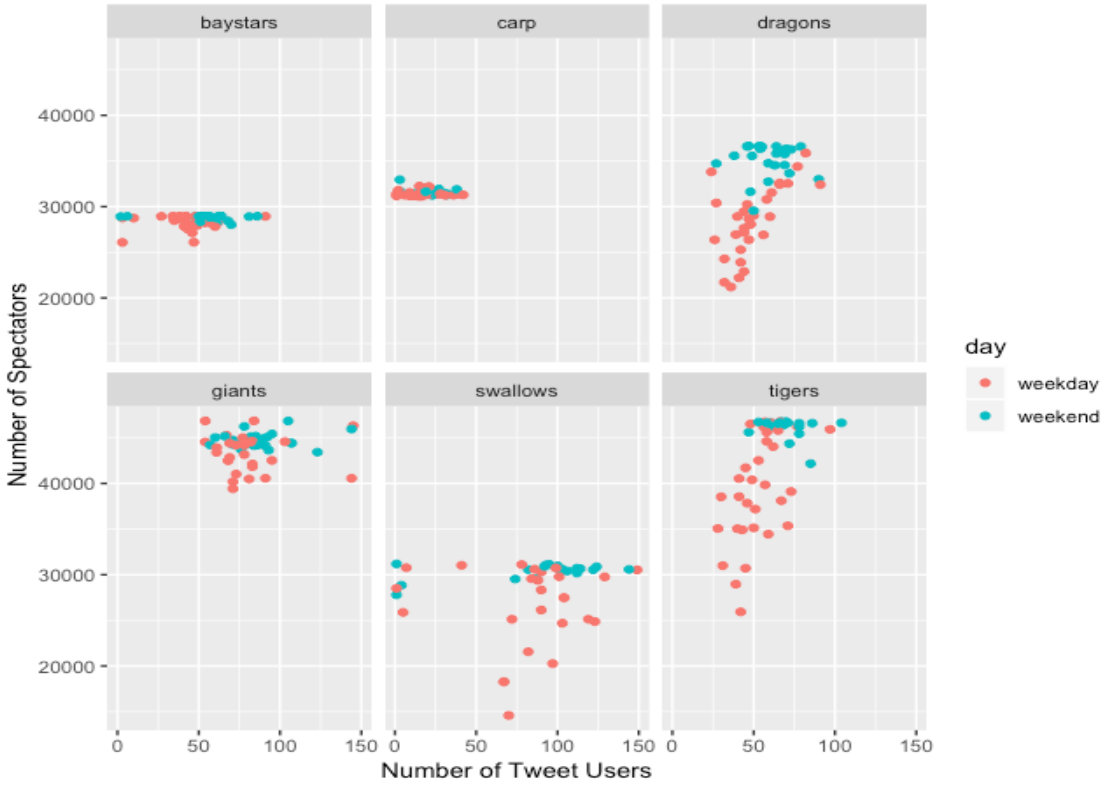


図 7-3　セ・リーグのスタジアム別の観客数と Tweet ユーザ数の散布図

図 7-4 はセ・リーグ，パ・リーグを含めた全球団における観客数と Tweet 数および観客数と Tweet ユーザ数の相関係数をまとめたものである．特にパ・リーグにおいて違いが顕著に現れており，東京ドームでのジャイアンツ戦，明治神宮野球場でのスワローズ戦，横浜スタジアムでのベイスターズ戦など，東京・神奈川の首都圏において，観客数と Tweet 数および観客数と Tweet ユーザ数との相関はいずれも弱いことがわかる．

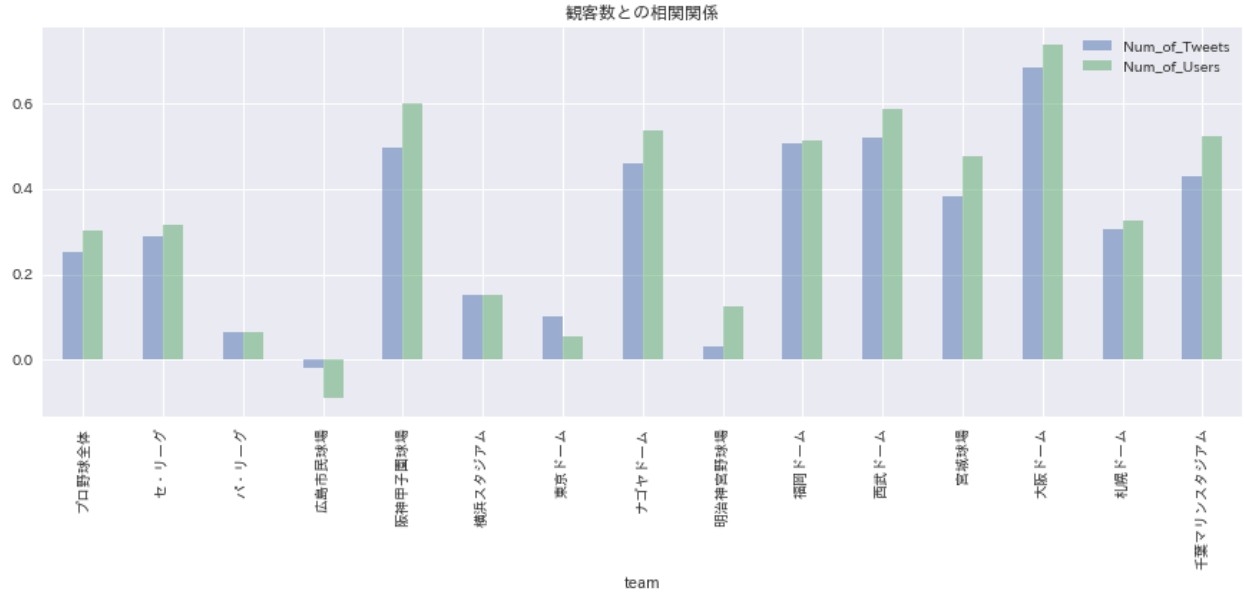


図 7-4　全球団の観客数と Tweet 数および Tweet ユーザ数の相関係数

これは各スタジアムにおける最大収容観客数の違いはあるものの，各プロ野球チームのファンの特徴やスタジアムが存在する地域的な特性などの違いによって観客者の Twitter の利用方法やツイートへの位置情報の付加有無が異なっている結果と推察することできる．

このように，ツイートをする行為はイベント内容や地域性など，様々な要因の影響を受けた集団から出力された結果となる可能性がある．このツイートという行動と空間の関係性を明らかすることができれば，母集団推定の精度向上だけでなく，Twitter データ活用の幅・効果の拡大が期待できる．

謝辞

本論文は著者が，筑波大学大学院システム情報工学研究科リスク工学専攻に在籍中の研究成果をまとめたものである．

研究を進めるにあたっては，同専攻　津田和彦教授には研究方法，進め方など全ての段階において多大なるご指導を賜り，さらに学内だけでなく研究会の紹介など学外からのアドバイスを頂ける機会も用意くださいました．また曜日，時間を問わず相談に乗っていただいたおかげで本論文としてまとめることができました．心より深謝申し上げます．

いつもご指導賜った同専攻　倉橋節也教授，木野泰伸准教授，審査におきましては，伊藤誠教授，新潟大学の木村裕斗准教授にも，貴重なアドバイスを頂戴しましたこと深く感謝申し上げます．

また，帝京大学の藤田昌克教授をはじめ，津田研究室のゼミ生，先輩各位にはデータの抽出方法，論文の作成方法についてアドバイスをいただくとともに，ゼミ時の闊達な助言によって考察を深めることができました．誠にありがとうございました．

最後に様々な専門分野で活躍する同期生，仕事と研究の両立に理解を示してくれた家族や職場の方々に心より感謝申し上げます．

参考文献

[Akaike, 1973]

Hirotogu Akaike (1973). “Information theory and an extension of the. maximum likelihood principle,” Proceedings of the 2nd International Symposium on Information Theory, Petrov, B. N., and Caski, F. (eds.), Akadimiai Kiado, Budapest, pp. 267-281.

[Akcora, et al., 2010]

Cuneyt Gurcan Akcora, Murat Ali Bayir, Murat Demirbas and Hakan Ferhatosmanoglu (2010). “Identifying breakpoints in public opinion,” Proceedings of the First Workshop on Social Media Analytics, July 2010 (SOMA'10), pp. 62-66, NewYork, NewYork, USA, ACMP ress.

[Allan, 2004]

Stephen Allan (2004). “Satellite television and football attendance: the not so super effect,” Applied Economics Letters, Vol. 11, pp. 123-125.

[Aramaki, et al., 2011]

Eiji Aramaki, Sachiko Maskawa and Mizuki Morita (2011). “Twitter. catches the flu: Detecting influenza epidemics using Twitter,” Proceedings of the conference on empirical methods in natural language processing, Association for Computational Linguistics, pp. 1568-1576.

[Asur & Huberman, 2010]

Sitaram Asur and Bernardo A. Huberman (2010). “Predicting the Future. with Social Media,” Predicting the future with social media. 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology.

[Backstrom, et al., 2010]

Lars Backstrom, Eric Sun and Cameron Marlow (2010). “Find me if you. can: improving geographical prediction with social and spatial proximity,” Proceedings of the 19th international conference on World wide web, April 2010 pp. 61-70.

[Benson, et al., 2011]

Edward Benson, Aria Haghighi and Regina Barzilay (2011). “Event. Discovery in Social Media Feeds,” Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, pp. 389-398.

[Bird, 1982]

Peter J. W. N. Bird (1982). “The demand for league football,” Applied Economics, 14, pp. 637-649.

[Bojanowski, et al., 2016]

Piotr Bojanowski, Edouard Grave, Armand Joulin and Tomas Mikolov (2016). ″Enriching Word Vectors with Subword Information,” Facebook AI Research.

[Bollen, et al., 2011a]

Johan Bollen, Huina Mao and Xiaojun Zeng (2011). “Twitter mood. predicts the stock market,” Journal of Computational Science, Vol. 2, No. 1, pp. 1-8.

[Bollen, et al., 2011b]

Johan Bollen, Alberto Pepe and Huina Mao (2011). “Modeling public. mood and emotion: Twitter sentiment and socio-economic phenomena,” Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011), 17-21 July 2011, Barcelona, Spain.

[Borland & Macdonald, 2003]

Jeffery Borland and Robert Macdonald (2003). “Demand for sport,” Oxford Review of Economic Policy, 19, pp. 478-502.

[Breiman, 2001]

Leo Breiman (2001). “Random Forests,” Machine Learning, Vol. 45, pp. 5-32.

[Chakrabarti & Punera, 2011]

Deepayan Chakrabarti and Kunal Punera (2011). “Event summarization. using tweet,” Proc. 5th Int. Conf. on Weblogs and Social Media (ICWSM 2011), pp. 66-73, Barcelona, Spain, AAAI Publications.

[Cheng, et al., 2010]

Zhiyuan Cheng, James Caverlee and Kyumin Lee (2010). “You are where you tweet: A content-based approach to geo-locating twitter uses,” Proceedings of the 19th ACM international conference on Information and knowledge management, pp. 759-768.

[Chew & Eysenbach, 2010]

Cynthia Chew and Gunther Eysenbach (2010). “Pandemics in the age of. twitter: Content analysis of tweets during the 2009 H1N1 Outbreak,” PLoS ONE, Vol. 5, No. 11, p. 13.

[Cho, et al., 2014]

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk and Yoshua Bengio (2014). “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation,” Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, pp. 1724-1734.

[Devlin, et al., 2019]

Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Proc. of NAACL-HLT 2019, pp. 4171-4186.

[Diakopoulos & Shamma, 2010]

Nicholas A. Diakopoulos and David A. Shamma (2010). “Characterizing. debate performance via aggregated twitter sentiment,” Proc. 28th Int. Conf on Human Factors in Computing Systems (CHI 10), pp. 1195-1198.

[Garcia & Rodriguez, 2002]

Jaume Garcia and Placido Rodriguez (2002). “The determinants of. football match attendance revisited: empirical evidence from Spanish Football League,” Journal of Sports Economics, 3, pp. 18-36.

[Ginsberg, et al., 2009]

Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski and Larry Brilliant (2009). “Detecting influenza epidemics using search engine query data,” Nature, Vol. 457, No. 7232, pp. 1012-1014.

[Hara, et al., 2020]

Hiroki Hara, Yoshikatsu Fujita and Kazuhiko Tsuda (2020). “Population estimation by random forest analysis using Social Sensors,” Proceedings the 24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2020), Procedia Computer Science, Volume 176, pp.1893-1902.

[Hara, et al., 2021a]

Hiroki Hara, Yoshikatsu Fujita and Kazuhiko Tsuda (2021). “Population estimation using Twitter for a specific space,” Data Technologies and Applications, Vol. 55, No. 3, pp. 430-445.

[Hara, et al., 2021b]

Hiroki Hara, Tomohiko Harada, Yoshikatsu Fujita and Kazuhiko Tsuda. (2021). “A Method of Classification Twitter Posting Location for a Specific Space,” Proceedings of 25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2021), Procedia Computer Science, Volume 192, pp. 2365-2374.

[Hart, et al., 1975]

Robert A Hart, J. Hutton and Trevor Sharot (1975). “A statistical analysis of association football attendance,” Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 24, No. 1 (1975), pp. 17-27.

[Hecht, et al., 2011]

Brent Hecht, Lichan Hong, Bongwon Suh and Ed H. Chi (2011). “Tweets from Justin Bieber’s heart: The dynamics of the location field in user profiles,” Proc. 2011 Annual Conference on Human Factors in Computing Systems (CHI’11), pp. 237-246, New York, New York, USA, ACM Press.

[Hind, et al., 2019]

Michael Hind, Dennis Wei, Murray Campbell, Noel C. F. Codella, Amit. Dhurandhar, Aleksandra Mojsilović, Karthikeyan Natesan Ramamurthy and Kush R. Varshney (2019). “TED: Teaching AI to Explain Its Decisions,” In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES), 123-129.

[Hochreiter & Schmidhuber, 1997]

Sepp Hochreiter and Juergen Schmidhuber (1997). “Long Short-Term Memory,” Neural Computation (1997), Vol. 9, No. 8, pp. 1735-1780.

[Jansen, et al., 2009]

Bernard J. Jansen, Mimi Zhang, Kate Sobel and Abdur Chowdury (2009). “Twitter Power: Tweets as Electronic Word of Mouth,” Journal of the American Society for Information Science and Technology, Vol. 60, No. 11, pp. 2169-2188.

[Kim, 2014]

Yoon Kim (2014). “Convolutional neural networks for sentence. classification,” Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1746-1751.

[Lander, 2015]

Jared P. Lander (2015). “みんなの R - データ分析と統計解析の新しい教科書 -,” マイナビ.

[Lei, et al., 2016]

Tao Lei, Regina Barzilay and Tommi Jaakkola (2016). “Rationalizing. neural predictions,” Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 107-117.

[Longueville, et al., 2009]

Bertrand De Longueville, Robin S. Smith and Gianluca Luraschi (2009). “OMG, from here, I can see the flames! :A use case of mining location based social networks to acquire spatio-temporal data on forest fire,” Geographic Information Systems, pp. 73-80.

[McCulloch & Walter, 1943]

Warren S. McCulloch and Walter Pitts (1943). “A logical calculus of. the ideas immanent in nervous activity,” The bulletin of mathematical biophysics 5, pp. 115-133.

[Meta, 2020]

Meta, 2020. Facebook 社 2020 年第 3 四半期（7 月－9 月）業績ハイライト. [オンライン]

Available at: https://about.fb.com/ja/news/2020/10/2020-third-quarter-results/

[アクセス日: 31 12 2021].

[Mikolov, et al., 2013a]

Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean (2013). “Efficient estimation of word representations in vector space,” arXiv:1301.3781 [cs.CL].

[Mikolov, et al., 2013b]

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado and Jeffrey Dean (2013). “Distributed representations of words and phrases and their compositionality,” Proceedings of the 26th International Conference on Neural Information Processing Systems, Vol. 2, pp. 3111-3119.

[Noll, 1974]

Roger G. Noll (1974). “Government and the sports business,” Washington, DC, Brookings Institute.

[Pennington, et al., 2014]

Jeffrey Pennington, Richard Socher and Christopher D. Manning (2014). “GloVe: Global Vectors for Word Representation,” Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, pp. 1532-1543.

[Ribeiro, et al., 2016]

Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin (2016). “"Why Should I Trust You?" Explaining the Predictions of Any Classifier,” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2016, pp. 1135-1144.

[Sakaki, et al., 2010]

Takeshi Sakaki, Makoto Okazaki and Yutaka Matsuo (2010). “Earthquake shakes Twitter user: Real-time event detection by social sensors,” Proceedings of the 19th international conference on World wide web. ACM, pp.851-860.

[Schumaker, 2010]

Robert P. Schumaker (2010). “An analysis of verbs in financial news articles and their impact on stock price,” Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media, pp. 3-4.

[Shamma, et al., 2010]

David A. Shamma, Lyndon Kennedy and Elizabeth F. Churchill (2010). “Tweetgeist: Can the twitter timeline reveal the structure of broadcast events?,” Proc. ACM Conf on Computer-Supported Cooperative Work, Savannah, Georgia, ACM.

[Social Media Experience, 2021]

Social Media Experience, 2021. SNSユーザー数（国内／世界）. [オンライン] Available at: https://socialmediaexperience.jp/article/191 [アクセス日: 31 12 2021].

[Song, et al., 2010]

Shuangyong Song, Qiudan Li and Nan Zheng (2010). “A spatio-temporal framework for related topic search in micro-blogging,” Proceedings of the 6th international conference on Active media technology, August 2010, pp. 63-73.

[Tumasjan, et al., 2010]

Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner and Isabell M. Welpe (2010). “Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment,” Proceedings of the International AAAI Conference on Web and Social Media, 4(1), pp. 178-185.

[Twitter, Inc., 2018]

Twitter, Inc., 2018. Search Tweets. ［オンライン］ Available at: https://developer.twitter.com/en/docs/tweets/search/overview ［アクセス日: 25 12 2018].

[Vaswani, et al., 2017]

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin (2017). “Attention is all you need,” In NIPS2017, pp. 5998-6008.

[Wang, et al., 2016]

Xingyou Wang, Weijie Jiang, Zhiyong Luo (2016). “Combination of. convolutional and recurrent neural network for sentiment analysis of short texts,” Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, COLING 2016, pp. 2428-2437.

[Zhang, et al., 2015]

Xiang Zhang, Junbo Zhao and Yann LeCun (2015). “Character-level convolutional networks for text classification,” Proceedings of the 28th International Conference on Neural Information Processing Systems 2015, Vol. 1, pp. 649-657.

[Zhao, et al., 2006]

Qiankun Zhao, Tie-Yan Liu, Sourav S. Bhowmick and Wei-Ying Ma (2006). “Event Detection from Evolution of Click-through Data,” Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, August 2006, pp. 484-493.

[マルティスープ, 2018]

マルティスープ, 2018. Google マップに同心円を描画する. [オンライン]
Available at: https://maps.multisoup.co.jp/blog/1666/
[アクセス日: 25 12 2018].

[ロプロス, 2018]

ロプロス, 2018. プロ野球 Freak. [オンライン]
Available at: https://baseball-freak.com
[アクセス日: 25 12 2018].

[伊川, et al., 2012]

伊川洋平, 榎美紀立, 堀道昭 (2012). “マイクロブログのメッセージを用いた発信場所推定,” 第 4 回データ工学と情報マネジメントに関するフォーラム, DEIM ‘12, 兵庫県神戸市, 日本.

[奥谷 & 山名, 2014]

奥谷貴志, 山名早人 (2014). “メンション機能を利用した Twitter ユーザプロフィール推定,” DBSJ Japanese Journal, vol. 13-J, No. 1, pp. 1-6.

[河合, 2008]

河合慎祐 (2008). “J リーグ観戦需要に関する研究,” 早稲田大学大学院スポーツ科学研究科スポーツ科学専攻スポーツビジネス研究領域, 5007A019-4.

[NTT ドコモ, 2018]
株式会社 NTT ドコモ, 2018. モバイル空間統計. [オンライン]
Available at:
https://www.nttdocomo.co.jp/biz/service/spatial_statistics/
[アクセス日: 19 12 2018].

[NTT ドコモ, 2021]
株式会社 NTT ドコモ, 2021. モバイル空間統計に関する情報 - お客さまのプライバシー保護について -. [オンライン]
Available at:
https://www.nttdocomo.co.jp/corporate/disclosure/mobile_spatial_statistics/#p02
[アクセス日: 30 12 2021].

[気象庁, 2018]
気象庁, 2018. 気象庁|過去の気象データ・ダウンロード. [オンライン]
Available at: https://www.data.jma.go.jp/gmd/risk/obsdl/index.php
[アクセス日: 25 12 2018].

[吉次, 2011]
吉次由美 (2011). “東日本大震災に見る大災害時のソーシャルメディアの役割 -ツイッターを中心に-,” 放送研究と調査, Vol. 61, No. 7, pp. 16-23.

[橋本 & 岡, 2012]
橋本康弘, 岡瑞起 (2012). “都市におけるジオタグ付きツイートの統計,” 人工知能学会誌, Vol. 27, No. 4, pp. 424-431.

[金明哲, 2017]
金明哲 (2017). “R によるデータサイエンス,” 森北出版.

[原田, 2014]
原田智彦 (2014). “口語的な表現を含むテキスト情報の理解支援に関する研究,” 筑波大学 12102 甲第 7275 号.

[榊 & 松尾, 2011]
榊剛史, 松尾豊 (2011). “ソーシャルメデイアからの人物目撃情報抽出システムの試作,” 人工知能学会全国大会 2011 論文集, 人工知能学会.

[榊 & 松尾, 2012]
榊剛史, 松尾豊 (2012). “ソーシャルセンサとしての Twitter - ソーシャルセンサは物理センサを凌駕するか?,” 人工知能学会誌 27 巻 1 号, pp. 67-74.

[榊 & 松尾, 2014]
榊剛史, 松尾豊 (2014). “ソーシャルメディアユーザーの職業推定手法の提案,” 知能と情報 Vol. 26, No. 4, pp. 773-780.

[三田村, 2014]
三田村健史 (2014). “DNS クエリデータにもとづくソーシャルメディア利用者の行動分析,” 筑波大学 12102 甲第 6780 号.

[三木, et al., 2014]
三木翔平, 新田直子, 馬場口登 (2014). “単語の地理的局所性の経時変化を考慮したツイートの発信位置推定,” 第 6 回データ工学と情報マネ ジメントに関するフォーラム.

[山田 & 斉藤, 2010]
山田和貴. 斉藤裕樹 (2010). “マイクロブログサービスの位置情報タグと発言コンテキスト解析を用いた行動推定システムの設計,” 情報処理学会研究報告, Vol. 2010-DBS-151, No. 21, pp. 1-6.

[寺田, 2014]

寺田 雅之 (2014). “モバイル空間統計：携帯電話ネットワークを活用した人口推計技術とその応用,” 日本計算機統計学会大会論文集 28 巻 pp.63-66.

[上里, et al., 2015]

上里和也, 浅井洋樹, 奥野峻弥, 山名早人 (2015). “Twitter ユーザを対象とした属性推定の精度向上-周辺ユーザの属性補完を利用して-,” DEIM Forum 2015, D8-5.

[森國, et al., 2015]

森國泰平, 吉田光男, 岡部正幸, 梅村恭司 (2015). “ツイート投稿位置推定のためのノイズとなる単語の除去手法,” DEIM Forum 2015 G8-1.

[清家, et al., 2011]

清家剛, 三牧浩也, 原裕介, 小田原亨, 永田智大, 寺田雅之 (2011). “まちづくり分野におけるモバイル空間統計の活用可能性に係る研究,” 公益社団法人日本都市計画学会, 都市計画論文集, Vol. 46, No. 3.

[清家, et al., 2013]

清家剛, 三牧浩也, 原裕介, 森田祥子 (2013). “基礎自治体におけるモバイル空間統計の活用可能性に関する研究,” 日本建築学会技術報告集, 第 19 号, 第 42 号, pp. 737-742.

[清家, et al., 2015]

清家剛, 三牧浩也, 森田祥子 (2015). “モバイル空間統計を活用した都市拠点地区の人口特性分析に係る研究,” 日本建築学会計画系論文集, 第 80 巻, 第 713 号, pp. 1625-1633.

[川端，et al.，2018]

川端一光，岩間徳兼，鈴木雅之 (2018). “R による多変量解析入門 データ分析の実践と理論,” オーム社.

[総務省，2017]

総務省 (2017). “平成 29 年版 情報通信白書,” 総務省.

[総務省，2018a]

総務省 (2018). “平成 30 年版 情報通信白書,” 総務省.

[総務省，2018b]

総務省 (2018). “平成 29 年情報通信メディアの利用時間と情報行動に関する調査,” 総務省情報通信政策研究所.

[池田，et al.，2012]

池田和史，服部元，松本一則，小野智弘，東野輝夫 (2012). “マーケット分析のための Twitter 投稿者プロフィール推定手法,” 情報処理学会論文誌 コンシューマ・デバイス＆システム(CDS), Vol.2, No.1, pp.82-93.

[鳥海，2015]

鳥海 不二夫 (2015). “Twitter 上のビッグデータ収集と分析” 組織科学, 48 巻 (2015), 4 号, pp. 47-59.

[那須野 & 松尾，2014]

那須野薫，松尾豊 (2014). “Twitter における候補者の情報拡散に着目した国政選挙当選者予測,” 2014 年度人工知能学会全国大会論文集，第 28 回, pp. 1-4.

[波部, 2012]

波部斉 (2012). “ランダムフォレスト,” 情報処理学会研究報告 Vol.2012-CVIM-182 No.31.

[波部, 2016]

波部斉 (2016). “ランダムフォレストの基礎と最近の動向,” 映像情報メディア学会誌, Vol. 70, No. 5, pp. 788-791.

[服部 & 速水, 2011]

服部哲, 速水治夫 (2011). “位置情報を含むツイートを効率的に発掘するための基本方式の検討,” マルテメディア, 分散, 協調とモバイル (DICOMO2011) シンポジウム, pp. 1526-1530.

[米田 & 前田, 2017]

米田康平, 前田亮 (2017). “ソーシャルメディアユーザのプロフィール推定手法の提案,” DEIM Forum 2017 D8-2.

[北, et al., 2002]

北研, 津田和彦, 獅々堀正幹 (2002). “索引語の抽出と重み付け,” 情報検索アルゴリズム, 共立出版, pp. 33-40.

[柳川 & 照井, 2020]

柳川琢省, 照井文彦 (2020). “言語学的手法を取り入れた機械学習モデルの局所的な説明手法,” The 34th Annual Conference of the Japanese Society for Artificial Intelligence, online.

関連業績

1. 学術論文［査読付き］

(1) Hiroki Hara, Yoshikatsu Fujita, Kazuhiko Tsuda (2021). “Population estimation using Twitter for a specific space,” Data Technologies and Applications, Vol. 55 No. 3, pp. 430-445.

2. 国際会議論文［査読付き］

(1) Hiroki Hara, Yoshikatsu Fujita, Kazuhiko Tsuda (2020). “Population estimation by random forest analysis using Social Sensors,” Proceedings the 24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2020), Procedia Computer Science, Volume 176, pp.1893-1902.

(2) Hiroki Hara, Tomohiko Harada, Yoshikatsu Fujita, Kazuhiko Tsuda (2021). “A Method of Classification Twitter Posting Location for a Specific Space,” Proceedings of 25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2021), Procedia Computer Science, Volume 192, pp. 2365-2374.

3. 学会発表論文［査読無し］

(1) 原大樹, 藤田昌克, 津田和彦 (2019). “ソーシャルセンサを用いた特定空間における母集団推定モデルに関する研究” 電気学会 電子・情報・システム部門 第78回情報システム研究会(2019-05-27), IS-19-024.