

未来のイベントに言及するマイクロブログ投稿を用いた人口変化の予測

塚田涼太郎[†] 詹 浩森[†] 石渡祥之佑^{*††} 豊田 正史^{††} 梅本 和俊^{††}
商 海川^{††,†††} 是津 耕司^{†††}

[†] 東京大学大学院情報理工学系研究科 〒113-8656 東京都文京区本郷 7-3-1

^{††} 東京大学生産技術研究所 〒153-8505 東京都目黒区駒場 4-6-1

^{†††} 情報通信研究機構 〒184-8795 東京都小金井市貫井北町 4-2-1

E-mail: [†], ^{††}{tsukada, hos, ishiwatari, toyoda, umemoto, shang}@tkl.iis.u-tokyo.ac.jp, ^{†††}zetsu@nict.go.jp

あらまし 多くの参加者が集まるイベントは、その発生場所周辺において人口の急増を引き起こす。こうした人口増加に起因する混乱を避けるためには、人口増加の発生時間や規模を事前に知る必要がある。しかしながら、イベントの発生時間や参加者数は日々変化するため、イベントに起因する人口増加をその発生場所の過去の人口情報のみに基づいて予測することは難しい。本研究では、都市における多様な種別のイベント発生場所の数日先までの人口予測のために、未来の日付とイベント発生場所の名称の両方に言及しているマイクロブログ上の投稿の内容およびその場所の過去の人口情報を融合する手法を提案する。混雑統計データと Twitter データを用いた実験の結果、東京・神奈川の主要なスポーツ会場・コンサート会場・展示会場・公園・デモ会場・観光地の 24 時間先までの人口予測について、提案手法が過去の人口情報のみを用いるベースライン手法の予測誤差を 20%から 50%改善することを確認した。さらに、予測に用いた投稿の内容の分析により、予測された人口変化の原因の説明可能性を示した。

キーワード 人口予測, マイクロブログ

1 はじめに

野球の試合や花見などのイベント¹発生時には、参加者が発生場所付近に集まることで突発的に人口が増加する。こうした人口の急増は参加者自身だけでなく、周囲を巻き込んで様々な悪影響を及ぼし得る。例えば、通常は空いている時間帯の電車が大量のイベント帰りの客のため満員になると、乗客は心身の苦痛に加えて乗降時間の増大に伴う遅延により経済的損失まで被る可能性がある。また、イベントによる混雑の発生を知らずに観光地を訪れた観光客にとっては、周遊性が低下するために旅行の満足度が低下する。

イベント発生場所の人口予測は、人口の急増に伴う諸問題を解決する上で重要な役割を果たす。例えば、鉄道の乗客や観光客は予測される混雑を避けて予定を立てることでそのリスクを低減できる。より積極的な利用方法として、コンビニなどの物流ネットワークにおいて人口の急増に伴う需要増加に応じ適切に商品を配置することで、売上を増やすことも考えられる。こうした背景のもと、GPS を搭載した携帯端末から収集されたデータを利用して都市全体の規模で人口を予測する手法が研究されている [2], [3]。しかしながら、これらの手法は数時間先までの予測しか想定していない。予測期間が長期になると、イベ

ントなどの外的要因による非定常的な人の動きの影響が無視できないため、これらの手法を人口の長期予測に適用することは難しい。予測された人口変化への対策を実施するための時間の確保の観点から、人口の長期予測は重要な課題である。

本研究では都市における多種多様なイベント発生場所の人口の数日先までの長期予測の問題に取り組む。とりわけイベント発生場所の周辺ではイベントによる人口増加が不定期に発生するため、過去の人口情報のみに基づく人口変化の長期予測が難しい [4], [5]。この問題を解決するために、マイクロブログ上には未来のイベントに言及している投稿が存在し、イベントの発生時間・場所・規模を予測する上で有用な情報となることに着目する。以下の投稿は、このアイデアを端的に表している。

明日の東京ドームのコンサート、チケット取れなかった

すなわち、これが 3 月 1 日の投稿であった場合、翌 2 日に東京ドームでコンサートが開催されること、およびチケットが取れなかったことから来場者数は満員に近いことが読み取れる。

そこで本研究では、未来のイベントに言及しているマイクロブログ上の投稿とイベント発生場所の過去の人口情報を手がかりとして、多様な種別のイベント発生場所（スポーツ会場・コンサート会場・展示会場・公園・デモ会場・観光地）の数日先までの人口予測に適用可能な手法を提案する（第 3 節）。具体的には、単に未来のイベントに言及している投稿数に着目するだけでは捉えきれないイベントの性質や規模の情報を得るために、投稿の内容のテキストを解析してその場所の過去の人口の時系列のデータと融合して予測を改善することを試みる。実世界の混雑統計データと Twitter データを用いた実験の結果、提

*：現在の所属は Mantra 株式会社である。

1：イベント検出に関する既存の研究 [1] における定義を参考に、本研究で扱うイベントは時空間上のある点で発生する人の自発的な集まりとする。これには、主催者が日程を定めるもの（例：野球の試合、コンサート）ならびに参加者が自然発生的に集合するもの（例：花見、デモ行進）の 2 種類がある。

案手法が過去の人口情報のみを用いるベースライン手法の予測誤差を 20%から 50%改善することを示した (第 4 節). さらにこの手法では, 予測された人口変化の原因を予測に用いた投稿の内容によって説明し得ることを確かめた.

本研究は, 我々のこれまでの研究 [4], [5] を以下の点で発展的に拡張したものである:

- 公園や観光地などの予定表に載らないイベントが発生する場所にも提案手法が適用可能であることを示した.
- ノイズの多い投稿の集合から有用な投稿を区別するために Set Transformer [6] を導入し, 特にスポーツ会場ならびにコンサート会場における予測誤差を削減した.
- 複数のイベント発生場所のデータを用いた同時学習およびファインチューニングによって, 他の場所で得られた知識を用いて予測を改善できることを確認した.
- 自己注意の重みの分析によりモデルが予測時に着目した投稿内の単語を可視化し, 予測結果の説明可能性を示した.

2 関連研究

2.1 位置情報を含むデータを用いた人口予測

イベントの発生によって急速に変化する都市の人口を予測するために, 携帯電話から収集された位置情報を含むデータを未来の人の動きの手がかりとして活用する研究が行われている.

大規模イベントの開催に伴う例外的な人口増加を外れ値として無視してきたという従来の人口予測に関する研究の問題に対して, Fan ら [2] は大規模イベント開催期間における都市全体の人々の動きのモデル化に取り組んだ. Fan らの手法は, 大規模イベントの開催時には会場に早く到着する人と遅れて到着する人がいるという仮定の下で直近の人々の移動経路を含むデータを活用するものであり, 都市全体の規模での人口をイベントの影響下においても予測可能であると報告されている.

未来の人の動きを把握するためのより直接的なアプローチとして, 乗換案内や地図のアプリケーションに入力される検索クエリのログデータを予測に活用する取り組みも行われている. Konishi ら [7] は, 乗換案内アプリケーションの検索クエリログを活用することで, 花火大会などのイベントに伴う都市全体規模の駅の混雑を 1 週間前の時点で予測した. Liao ら [3] は, 地図アプリケーションにおけるイベント会場を目的地とした検索回数がイベント開始時刻直前に増加する現象に着目して, イベント会場周辺の交通速度の数時間先までの予測を行った.

これらの既存研究では特定のサービスの利用者から得られる時間および位置情報を陽に含むデータを未来の人の動きの手がかりとして活用している. 一方本研究では, 未来の日付とイベント発生場所への言及という形で時間と位置情報を含むマイクロブログ上の投稿を, 幅広い利用者から容易に収集可能な手がかりとして用いる. さらに, 本研究の提案手法は投稿の内容のテキストを用いるため, 予測した人口変化の原因について投稿内容による説明を加えられる. 予測された人口変化の原因が分かればより適切な対策の実施が可能になるため, 予測結果の説

明可能性は実用上重要な利点である.

2.2 マイクロブログ上の投稿からのイベント情報の抽出

マイクロブログ上の投稿をソーシャルセンサーとして用い, 実世界で起きているイベントを把握することが試みられてきた.

Yamada ら [8] は, 旅行の計画に役立つローカルなイベント情報を Twitter 上の投稿から抽出する研究を行った. Yamada らの手法は, イベント会場名を含む投稿からイベントの開催期間を取得し会場名の表記揺れを考慮して集約する. この手法により, Yamada らは従来手法より正確なイベント情報抽出ができることを示した. Jatowt ら [9] は, 時間に言及する Twitter 上の投稿に着目することで, 現在だけではなく過去や未来のイベントを可視化して分析するフレームワークを提案した.

これらの研究に着想を得て, 本研究ではマイクロブログ上で未来のイベントに言及している多様な投稿の中から, イベント発生場所の名称と日付に言及している表現の両方を含む投稿に着目し, 未来のイベント参加者数の手がかりとして利用する.

2.3 マイクロブログ上の投稿を用いた人口の推定と予測

マイクロブログ上の投稿は, 実世界の人口を推定・予測するためにも利用されている.

Onishi ら [10] は, 現実空間の人口と仮想空間であるマイクロブログ上の投稿数が相互作用していると仮定して分析を行った. 具体的には, 秋葉原の商業施設においてカメラの映像から取得された人口と, 「秋葉原」という単語が含まれている Twitter 上の投稿数が互いに影響を与えているとするモデルを作成し, モデルが学習したパラメータの大小から人口と投稿数が互いに与える影響の説明を試みた. ただし, Onishi らのモデルは現在の人口を現在の投稿数および過去の投稿数・人口を用いて推定するものであり, 未来の人口の予測には適用できない.

He ら [11] は, Twitter 上の投稿に交通状況への言及があることに着目し, 投稿に含まれる単語の情報をを用いて自動車の交通量を比較的長期にわたって予測する手法を提案した. He らの手法は, 投稿に付与されている位置情報を用い, 交通量の予測対象となる区画内で発信された投稿のみをフィルターして利用している. しかしながら, 2019 年に Twitter は投稿に正確な位置情報 (ジオタグ) を付与する機能を廃止²したため, この手法は現在適用不可能である.

本研究では, これらの研究の流れを前節で述べたマイクロブログ上の投稿からのイベント情報の抽出の研究と合流させ, 未来のイベントに起因する不定期な人口変化の予測を試みる.

3 提案手法

我々のこれまでの研究 [4], [5] で報告したように, 過去の人口情報のみを用いてイベント発生場所周辺の人口を予測することは困難である. これは, 野球の試合やコンサートなどの多くの参加者が集まるイベントの発生日がモデルに与えられていないためである. また, 都市におけるイベント発生場所の種類はス

2: <https://twitter.com/TwitterSupport/status/1141039841993355264>

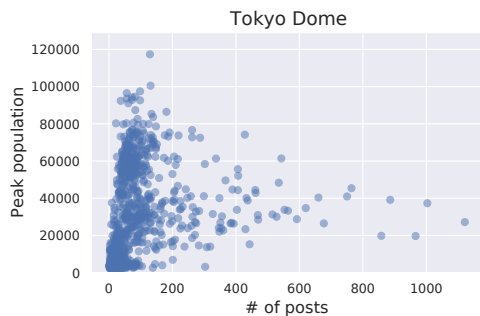


図1 東京ドームにおける1日のピーク人口と言及投稿数.
「混雑統計[®]」©ZENRIN DataCom CO., LTD.

スポーツ会場から公園まで多岐に渡り、それぞれの場所の予測の難しさにも違いが生じている。例えば、スポーツ会場の人口は、数万人規模の観客を動員するスポーツの試合によって急激に変化する。それとは対照的に、公園の人口は花見などの季節性のイベントによって緩やかに変動する。さらに、投稿の量やイベントの頻度などもそれぞれの場所によって異なる。

この問題の単純な解決策としては、イベントの主催者や発生場所の管理者によって公開されるイベント予定表の利用が考えられる。しかしながら、多様なイベント発生場所の中でも公式の予定表が公開されている場所は限られている。さらに、公園での花見や国会議事堂前のデモ行進のように、日程が不明確なイベントも存在する。様々な種類のイベントおよび発生場所に対して予測を行うには、そのような予定表に依存せずに未来のイベント発生情報を自動的に収集できることが望ましい。

そこで本研究では、未来のイベントについての情報がマイクロブログ上に投稿されているという事実に着目する。具体的には、未来のイベントに言及するマイクロブログ投稿ならびにイベント発生場所の過去の人口情報を組み合わせた異種データを用いる手法を提案する(第3.1節)。また、場所ごとの状況の違いに対処するために、提案手法の実装として2種類のモデルを提案する。一方は勾配ブースティングのモデル(第3.2節)で、もう一方はSet Transformerのモデル(第3.3節)である。

3.1 未来のイベント発生情報としてのマイクロブログ投稿

イベント発生場所で将来発生する人口変化を予測する手がかりとして、未来の日付とイベント発生場所の名称の両方に言及しているマイクロブログ上の投稿を用いる。ここで、イベント発生場所 v のある未来の日付 d の人口予測に有用な投稿は、以下の条件(以降、抽出条件)をすべて満たすと考えた:

- d に言及する日付の表現を含む
- イベント発生場所 v の名称を含む
- d の前日以前に投稿されている

詳細は第4.1節で述べるが、抽出条件を満たすイベント関連投稿のカバレッジを上げるために、検索クエリを拡張して日付表現ならびにイベント発生場所の名称の表記揺れをカバーした。

こうした投稿を用いてイベント発生場所周辺の人口を予測する場合、最も単純には投稿数によって対応する日における人口

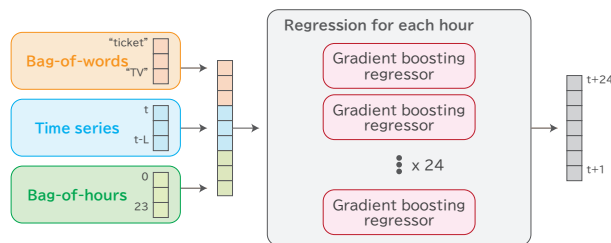


図2 提案手法のモデル(GBR)の概要.

を回帰することが考えられる。しかしながら、図1に示す通り投稿数と人口の間に明確な相関は見られないため、このアプローチによる予測は失敗する。実際に投稿の内容を確認すると、この投稿数と人口の不一致の原因は主に2パターンに分けられた:(1) 実際の人口変化に関係しない大量の投稿(例:過去のコンサートDVDの発売日決定の告知)が抽出条件を満たしている。(2) イベント自体の性質(例:参加者の年齢層)のために他のイベントよりもマイクロブログ上での注目度が低い。そのため、未来のイベント発生情報として利用する場合、投稿数だけではなく投稿の内容に着目することが必要となる。

しかしながら、各予測対象日について収集された投稿の集合には、扱いが困難な性質が2点ある:(1) 投稿数が各予測対象日によって異なる。(2) 投稿の順番は予測対象日の人口変化には関係しない。これは、マルチインスタンス学習[12]の問題に代表される set-input 問題の一例である。以下では、提案手法の2つのモデルにおけるこの問題の性質への対応を説明する。

3.2 勾配ブースティング回帰(GBR)を用いた予測

勾配ブースティング回帰(以降、GBR)[13]を用いる動機を以下に示す:

- GBRは異種データを扱える。この性質により、マイクロブログ投稿の特徴量と過去の人口の時系列の特徴量をスケールや次元数の違いに依らずに混ぜて利用できる。
- GBRは少量のデータでも訓練できる。第4節で述べるように、一部のイベント発生場所について言及している投稿は少ないため、この性質は有用であると考えられる。

図2にGBRモデルの概要を示す。GBR単体ではset-input問題に適用できないため、投稿の集合を集約した表現をモデルの入力とした。具体的には、抽出条件(第3.1節)をすべて満たす投稿から重複(botにより自動生成された投稿など)を取り除き、ユニークな投稿をすべて連結した文書のBag-of-Words表現を特徴量として用いる。Bag-of-Words表現は任意の数の投稿を扱えて、さらに投稿の順序に対して不変である。イベント発生場所の人口変化の時系列における短期的トレンドを捉えるために、Bag-of-Words表現のベクトルと1時間ごとの人口の時系列 $24 \times n$ ステップ分のベクトルを連結して回帰モデルに入力する。ここで、 n は過去の人口情報を何日前まで遡って用いるかを決めるハイパーパラメータである。さらに、様々な表記で現れる時刻を参照する表現を活用するために、Bag-of-Hoursベクトルを入力に連結する。Bag-of-HoursベクトルはBag-of-Words

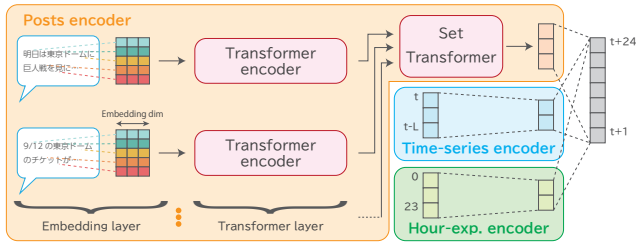


図3 提案手法のモデル (ST) の概要.

ベクトルのアナロジーで、投稿の集合の中で i 時が参照された回数を i 番目の次元が表す 24 次元ベクトルである。ここで、時刻表現の表記は正規表現で網羅した。予測は上述のベクトルを説明変数、対応する時間における人口を目的変数として、24 時間の各時間について別々に回帰することで行う。

3.3 Set Transformer (ST) を用いた予測

Set Transformer (以降、ST) [6] を用いる動機を以下に示す：

- ST は個別の投稿を一つ一つ処理できる。投稿の集合を単一のベクトルで表現する Bag-of-Words のアプローチとは対照的に、ST では集合に含まれる投稿間の複雑な関係性を考慮できることが期待される。
- 各種のデータの特徴を考慮した表現を学習できる。GBR は異なる種類の入力を対等に扱うが、NN モデルは異種データのそれぞれを効果的に扱うための事前知識を反映した個別のエンコーダを使える。データ融合を用いた近年の交通予測の研究 [3], [14] でその有効性が示されている。
- 多くの投稿で言及されている場所については、NN モデルが高い性能を発揮することが期待される。

図3にSTモデルの概要を示す。STは、set-input問題に適用可能な自己注意の機構に基づく順序不変なモジュールである。モデルは、投稿・時系列・時刻表現の3つのエンコーダからなる。投稿エンコーダでは、STが投稿ベクトルの集合を単一のベクトルにエンコードする。ここで、各投稿はTransformer [15]でエンコードされる。各投稿の先頭に付与した<CLS>トークンを、Transformerでエンコードした時の埋め込みをその投稿のベクトルとして用いる。これは、既存の研究 [16]において分類タスクを解く際に文の表現を得るための方法である。ここで、訓練時に出現しない単語には未知語トークン (<UNK>) を割り当てる。時系列エンコーダでは、直近の1時間ごとの人口の時系列 $24 \times n$ ステップ分のベクトルが全結合層に入力される。時刻表現エンコーダでは、投稿中に含まれる時刻を参照する表現を陽に利用するために、Bag-of-Hoursベクトルが全結合層に入力される。データ融合による交通予測の既存のアプローチ [3]を参考に、これら3種類の入力(投稿、時系列、時刻表現)の埋め込み表現を連結し、後続の全結合層に入力する。

第4節で述べるように、一部のイベント発生場所では各予測対象日について収集できた投稿が少なく、投稿エンコーダを訓練させるために必要な訓練例が不足している。そのような場所における予測誤差を削減するために、最初にモデルを同一種別

表1 各イベント発生場所の種別、テストデータ中のイベント日数、訓練データ中の投稿数.

種別	名称	イベント日数	投稿数
スポーツ会場	日産スタジアム	15	6,229
	東京ドーム	131	82,629
	味の素スタジアム	29	15,842
	明治神宮野球場	87	7,842
	横浜スタジアム	72	20,892
	秩父宮ラグビー場	114	2,489
コンサート会場	パシフィコ横浜	53	21,310
	横浜アリーナ	106	52,115
	日本武道館	114	41,507
展示会場	東京国際展示場	80	79,237
公園	上野恩賜公園	21	27,129
	代々木公園	196	3,500
デモ会場	国会議事堂前	111	3,081
観光地	元箱根	100	26,319
	箱根湯本駅	203	26,319

「混雑統計[®]」 ©ZENRIN DataCom CO., LTD.

の場所の全データで訓練させてから、続いて予測対象の場所のデータだけでファインチューニングする。

4 評価実験

提案手法の有効性を実世界データを用いて評価した。実験は3種類行い、それぞれ下記の問いに答えられるように設計した：

- Q1 提案手法のモデルはイベント発生日と非発生日を区別できるか？ (第4.2節)
- Q2 異なるタスク設定において提案手法のモデルはどれほどの精度で人口変化を予測できるか？ (第4.3節)
- Q3 予測に重要な特徴量は何か？ (第4.4節)

4.1 実験設定

データセット. 実験に用いたデータセットは、時空間人口統計データとマイクロログ投稿データからなる。

時空間人口統計データについては、混雑統計データ³を使用した。混雑統計データには、約250m四方のメッシュごとに、そのメッシュ内の滞在人数の推定値を1時間単位で集計した値(混雑度)が含まれる。このうち2014年12月から2018年11月までの4年分について、最初の2年分を訓練、次の1年分を開発、最後の1年分をテストに用いた。表1に示す通り、予測対象のイベント発生場所には、異なる種別のイベントが発生する東京・神奈川の15地点を選んだ。各イベント発生場所についてその場所を囲むメッシュを選び、その混雑度をその場所周辺の人口とした。なお、公園などの単一メッシュに収まらない

3: 「混雑統計[®]」データは、NTTドコモが提供するアプリケーション(※)の利用者より、承諾を得た上で送信される携帯電話の位置情報を、NTTドコモが総体的かつ統計的に加工を行ったデータ。位置情報は最短5分毎に測位されるGPSデータ(緯度経度情報)であり、個人を特定する情報は含まれない。またデータの加工には「非特定化」「集計処理」「秘匿処理」がなされており個人が特定されることはない。※ドコモ地図ナビサービス(地図アプリ・ご当地ガイド)等の一部のアプリ。

表2 イベント有無の予測性能 (Precision, Recall, F1).

イベント発生場所	AR			LSTM			GBR			ST		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
日産スタジアム	0.33	0.07	0.11	0.00	0.00	0.00	0.63	0.80	0.71	0.67	0.80	0.73
東京ドーム	0.87	0.20	0.32	0.80	0.18	0.30	0.86	0.77	0.81	0.81	0.79	0.80
味の素スタジアム	0.00	0.00	0.00	1.00	0.03	0.07	1.00	0.48	0.65	0.78	0.62	0.69
明治神宮野球場	0.68	0.31	0.43	0.87	0.15	0.25	0.77	0.76	0.76	0.79	0.72	0.75
横浜スタジアム	0.62	0.18	0.28	0.80	0.06	0.10	0.95	0.76	0.85	0.92	0.82	0.87
秩父宮ラグビー場	0.68	0.40	0.51	0.64	0.62	0.63	0.89	0.68	0.77	0.49	0.81	0.61
パシフィコ横浜	0.50	0.26	0.35	0.30	0.06	0.10	0.67	0.58	0.63	0.70	0.49	0.58
横浜アリーナ	0.40	0.25	0.31	0.50	0.19	0.27	0.51	0.51	0.51	0.60	0.57	0.58
日本武道館	0.44	0.25	0.32	0.44	0.14	0.21	0.55	0.49	0.52	0.68	0.51	0.58
東京国際展示場	0.48	0.56	0.52	0.56	0.29	0.38	0.44	0.76	0.55	0.46	0.80	0.58
代々木公園	0.00	0.00	0.00	0.00	0.00	0.00	0.59	0.48	0.53	0.50	0.43	0.46
上野恩賜公園	0.76	0.67	0.71	0.83	0.56	0.66	0.74	0.77	0.75	0.74	0.55	0.63
国会議事堂前	0.75	0.71	0.73	0.72	0.76	0.74	0.80	0.67	0.73	0.82	0.67	0.74
元箱根	0.48	0.24	0.32	0.00	0.00	0.00	0.75	0.33	0.46	0.54	0.20	0.29
箱根湯本駅	0.72	0.67	0.69	0.82	0.41	0.55	0.78	0.58	0.66	0.79	0.58	0.66
マクロ平均	0.51	0.32	0.37	0.55	0.23	0.28	0.73	0.63	0.66	0.69	0.62	0.64

「混雑統計[®]」©ZENRIN DataCom CO., LTD.

広い場所では、イベント発生場所として利用される広場などを含む1メッシュを選んだ。ここで、過去の人口変化の時系列に現れる週毎のパターンを捉えるために、第3.2節のハイパーパラメータ n は $n = 7$ とした。この時系列データは、通勤・通学に起因する日毎や週毎の周期的なトレンドを含む。時系列データの解析に先立っては、こうしたトレンドの除去が重要である[17]。そこで、1日の各時間ごとに過去の人口の平均値を時系列データから引くことでこれらのトレンドを除去した。

マイクロブログ投稿データについては、研究室で収集しているTwitterアーカイブを利用した。⁴第3.1節の抽出条件を満たすイベント関連投稿のカバレッジを上げるために、Wikipediaのリダイレクトデータから名寄せ辞書を構築して、イベント発生場所の名称の表記揺れ(通称や旧称など)に対応した。抽出条件を満たした投稿は、JUMAN⁵で分かち書き処理を行いストップワードを除去した。抽出条件を満たす場所の名称と日付表現は、それぞれ特別なトークン(<TARGET_PLACE>および<TARGET_DATE>)で置換することで表記揺れを解消した。

ベースライン手法. 提案手法を2つのベースライン手法と比較した。具体的には、ARおよびLSTMモデルを用いる手法である。LSTMのハイパーパラメータは、PyTorch⁶の実装のデフォルト値とした。これらの手法はマイクロブログ投稿を使わない手法であり、24時間先までの人口を自己回帰的な方法(前の時点の出力を次の時点の入力にする)で予測する。

評価指標. 提案手法の有効性の評価は、イベント有無のレベルと人口変化のレベルの2つの粒度で行った。

イベント有無のレベルの評価では、各手法がイベント発生日と非発生日を分類できるかを確認した。しかしながら、第3節で述べた通り、多くのイベント発生場所でイベント予定表は公開されていない。そこで、場所ごとに1日の人口の時系列に対して k -means クラスタリングを行い、要素数が最大のクラスタ(以降、最大クラスタ)に属する日をイベント非発生日、そうでない日を発生日とした。ここで、クラスタ数 k は $k = 3$ とした。これは、訓練データを用いた予備実験において $k = 3$ とした場合、イベント発生日が属する2つの小さいクラスタと、非発生日が属する1つの大きいクラスタに分かれる傾向にあるという分析に基づく。表1に各イベント発生場所のテストデータ中のイベント日数を示す。テスト時には、回帰である各手法による出力の予測値がどのクラスタの平均値に最も近いかという基準で、その出力をイベント発生日または非発生日のクラスに分類して扱った。具体的には、予測値が最大クラスタの平均値に最も近い場合はイベント非発生日、そうでない場合は発生日とした。この基準により、各手法の性能をイベント有無の2値分類問題としてPrecision, Recall, F1で評価した。

人口変化のレベルの評価では、実際の人口の規模をどの程度正確に予測できるかを確認した。異なるメッシュ間での予測結果を比較可能にするために、評価には式(1)で定義される加重絶対パーセント誤差(WAPE)を用いた。

$$WAPE = \frac{1}{N} \sum_{t=1}^N \left| \frac{\hat{X}_t - X_t}{\bar{X}} \right| \quad (1)$$

ここで、 X_t, \hat{X}_t はそれぞれ時刻 t における実測値、予測値であり、 N は評価対象のデータの総数、 \bar{X} は実測値の平均である。

4.2 Q1: イベント有無の予測

最初に、イベント有無のレベルでの予測結果を表2に示す。AR・LSTMモデルは、多くのイベント発生場所においてRecallが0に近い値になった。このことは、これらのベースライン手

4: このアーカイブは、Twitter API によるクローリングを2011年3月から継続的に行うことで構築されており、約250万人の公開ユーザーのタイムラインからなる。このクローリングは、日本の有名人の30ユーザーから開始して、そのタイムライン上のリツイートとメンションを追っていくことで繰り返しユーザーの集合を拡張してきたものである。

5: <http://nlp.ist.i.kyoto-u.ac.jp/?JUMAN>

6: <https://pytorch.org/>

表3 $n = 7$ の場合の全期間の予測誤差 (WAPE).

イベント発生場所	AR	LSTM	GBR	ST
日産スタジアム	1.00	0.90	0.80	0.79
東京ドーム	0.83	0.82	0.45	0.43
味の素スタジアム	1.18	1.11	0.80	0.75
明治神宮野球場	0.94	0.90	0.65	0.68
横浜スタジアム	0.66	0.60	0.50	0.49
秩父宮ラグビー場	0.40	0.39	0.33	0.38
パシフィコ横浜	0.60	0.58	0.49	0.50
横浜アリーナ	0.49	0.48	0.48	0.46
日本武道館	0.91	1.00	0.79	0.73
東京国際展示場	0.56	0.54	0.51	0.57
代々木公園	0.79	0.87	0.81	0.85
上野恩賜公園	0.54	0.52	0.52	0.56
国会議事堂前	0.43	0.40	0.49	0.53
元箱根	1.24	1.14	1.06	1.02
箱根湯本駅	0.77	0.73	0.71	0.71
マクロ平均	0.76	0.73	0.63	0.63

「混雑統計[®]」 ©ZENRIN DataCom CO., LTD.

法ではほとんどの場所でイベント発生日の検出に失敗したことを示している。また、一部のイベント発生場所ではAR・LSTMモデルともにPrecisionとRecallの値が0になった。こうした異常に低い値は、これらの場所のイベント発生日数の少なさによって部分的に説明できる(表1)。これらの場所では少しのイベントしか発生しないため、入力の時系列は人が集まらない日の系列になることが多く、イベント発生日の時系列を予測するために有用な手がかりがほとんど得られない。一方、提案手法のGBR・STモデルはほぼすべての場所で一貫して高い性能を達成した。未来のイベントに言及するマイクロブログ投稿を追加の手がかりとして活用することで、提案手法のGBR・STモデルはイベント発生日を判別できた。

イベント発生場所の種別ごとにPrecision, Recallを見ると、東京ドーム、明治神宮野球場、横浜スタジアムなどの野球場では比較的高い数値となっているのに対して、横浜アリーナや日本武道館といった主にコンサートが開催される会場においては比較的低い数値となっていることが分かる。原因として、これらの場所における発生イベントの種別の違いが考えられる。例えば、プロ野球の試合は野球場の発生イベントの多数を占めるイベントだが、未来の野球の試合に言及する投稿にはその野球場をホームとするチーム名など特徴的な単語が多く含まれており、学習が容易である。一方、コンサート会場ではコンサートに言及する投稿が出演者の名称など様々な固有名詞を含み、学習が困難であることが考えられる。

4.3 Q2: 人口変化の予測

次に、人口変化のレベルでの予測結果を表3に示す。表3はテストデータの全期間で測定された人口変化の予測誤差を示している。提案手法のGBR・STモデルは、ほぼすべての場所においてベースライン手法のモデルより小さい誤差を達成した。誤差の削減の度合いはコンサート会場と比べて特にスポーツ会場で大きく、第4.2節のイベント有無の予測結果と一致する。

表4 イベント発生日の予測誤差 (WAPE).

イベント発生場所	AR	LSTM	GBR	ST
日産スタジアム	0.81	0.87	0.54	0.54
東京ドーム	0.66	0.76	0.37	0.35
味の素スタジアム	0.86	0.86	0.60	0.47
明治神宮野球場	0.67	0.77	0.45	0.50
横浜スタジアム	0.61	0.69	0.49	0.40
秩父宮ラグビー場	0.81	0.70	0.58	0.58
パシフィコ横浜	0.53	0.58	0.38	0.45
横浜アリーナ	0.56	0.58	0.56	0.49
日本武道館	0.67	0.74	0.62	0.58
東京国際展示場	0.34	0.43	0.28	0.28
代々木公園	0.74	0.70	0.61	0.64
上野恩賜公園	0.44	0.46	0.40	0.48
国会議事堂前	0.58	0.49	0.63	0.67
元箱根	0.82	0.91	0.74	0.78
箱根湯本駅	0.62	0.62	0.60	0.60
マクロ平均	0.65	0.68	0.52	0.52

「混雑統計[®]」 ©ZENRIN DataCom CO., LTD.

表5 予測期間を変化させた場合のイベント発生日のGBRの予測誤差 (WAPE).

イベント発生場所	1日先	2日先	3日先
日産スタジアム	0.54	0.54	0.54
東京ドーム	0.37	0.40	0.41
味の素スタジアム	0.60	0.63	0.62
明治神宮野球場	0.45	0.49	0.50
横浜スタジアム	0.49	0.62	0.61
秩父宮ラグビー場	0.58	0.60	0.64
パシフィコ横浜	0.38	0.45	0.47
横浜アリーナ	0.56	0.57	0.57
日本武道館	0.62	0.65	0.66
東京国際展示場	0.28	0.34	0.37
代々木公園	0.61	0.54	0.53
上野恩賜公園	0.40	0.45	0.45
国会議事堂前	0.63	0.66	0.71
元箱根	0.74	0.78	0.78
箱根湯本駅	0.60	0.62	0.62
マクロ平均	0.52	0.56	0.57

「混雑統計[®]」 ©ZENRIN DataCom CO., LTD.

提案手法のGBR・STモデルの性能を検討するために、以下では同じ予測タスクを異なる設定で実験・評価した。

イベント日. テストデータのほとんどの日は非イベント日であるため、テストデータの全期間での評価は予測誤差を過小評価することになる。そこで、予測誤差をイベント日だけで評価した(表4)。実用上重要なイベント日における人口変化の予測はより難しいタスクだが、その場合も提案手法はベースライン手法と比較してほぼすべての場所で最良の結果となった。

予測期間. 提案手法がさらに長期にわたる人口予測にも適用できるかを検証するために、予測期間を変化させて実験を行った。時間の制約のため、この実験はSTモデルより高速に訓練できるGBRモデルを用いて行った。表5に異なる予測期間における実験結果を示す。「 d 日先」の予測は、予測対象日の d 日前までに入手可能な投稿およびイベント発生場所の人口情報の

表 6 投稿数のサンプリング割合を変化させた場合のイベント発生日の GBR の予測誤差 (WAPE).

イベント発生場所	100%	50%	10%	0%
日産スタジアム	0.54	0.59	0.62	0.81
東京ドーム	0.37	0.36	0.39	0.57
味の素スタジアム	0.60	0.59	0.65	0.83
明治神宮野球場	0.45	0.47	0.46	0.61
横浜スタジアム	0.49	0.47	0.52	0.55
秩父宮ラグビー場	0.58	0.59	0.64	0.76
パシフィコ横浜	0.38	0.38	0.46	0.49
横浜アリーナ	0.56	0.55	0.57	0.54
日本武道館	0.62	0.64	0.65	0.64
東京国際展示場	0.28	0.27	0.28	0.29
代々木公園	0.61	0.56	0.69	0.67
上野恩賜公園	0.40	0.41	0.41	0.41
国会議事堂前	0.63	0.63	0.64	0.42
元箱根	0.74	0.75	0.74	0.78
箱根湯本駅	0.60	0.60	0.60	0.58
マクロ平均	0.52	0.52	0.55	0.60

「混雑統計[®]」©ZENRIN DataCom CO., LTD.

時系列を用いる設定である。

実験の結果、いずれの場所においても予測期間が長いほど予測誤差が大きい傾向を確認できた。特に、「1日先」と「2日先」の実験設定での予測誤差の差が大きかった。このことは、重要な手がかりになる投稿はイベント前日に投稿されやすい可能性や、イベント設営などの準備作業の影響が前日の人口の時系列に手がかりとして出てきている可能性を示唆している。

投稿数. 投稿数の多さが予測精度に寄与するかを調べるために、訓練データとテストデータ中の両方の投稿を一定の割合でランダムにサンプリングする条件で評価を行った。同様に、この実験も GBR モデルのみを用いて行った。投稿数と予測精度との関係を表 6 に示す。使用する投稿の比率を 100%とした実験は表 3 の実験と同じ設定である。比率を 0%とした実験は、イベント発生場所の過去の人口情報の時系列のベクトルだけを GBR の説明変数として用いる設定である。

実験の結果、投稿数が多いほど予測誤差が小さい傾向が明らかになり、マイクロブログ上の投稿はイベント発生場所の人口変化の予測精度向上に寄与する情報を含むことが分かった。

4.4 Q3：特徴量の重要度

最後に、人口予測に有用な特徴量を明らかにする。

GBR. 今回、回帰に用いた GBR のモデルは決定木をベースとするモデルであり、入力の特徴量の予測に対する重要度を計算できる [18]。すなわち、それぞれの時間帯において、どの単語およびどの時点の過去の人口情報が予測に貢献するかを調べられる。東京ドームにおける各時間の特徴量重要度上位 10 件を表 7 に示す。東京ドームを本拠地とするチーム名である「巨人」という単語が各時間の重要度の上位に位置している。これは、東京ドームの年間発生イベントの多くを占める野球の試合を特徴付ける単語であると考えられる。また、12 時から 20 時において最も重要な単語とされた「14」「18」は、試合やコン

サートの開始時刻の数字 (例：18 時開演) であると考えられる。

ST. ST モデルは、各投稿中の単語間で自己注意を適用している。ここで、モデルの振る舞いをより深く理解するために、回帰に用いる <CLS> トークンの埋め込みを得る時に他の単語に掛かっていた自己注意の重みの値を可視化して分析した。この節では、いくつかの場所種別について重みが付いた単語の分析結果を示す。以降の事例は、ST の予測誤差が GBR と同等程度に小さかった事例である。

スポーツの試合が多数開催されるスポーツ会場では、ほとんどの試合の日において試合の開始時刻の数字が重み付けられていた。東京ドームにおいて野球の試合が開催された日の実際の入力の投稿の一例を以下に示す。ハイライトされた単語が自己注意機構において重み付けられた単語である。なお、ハイライトの濃さは重みの大きさに対応している。

```
<CLS> <TARGET_DATE> 木 18 時 巨人 対 阪神 <TARGET_PLACE> 指
定席 2 階 3 塁 2400 円 2 連 kyojin hanshin 巨人 阪神 阪神 タ
イガース giants tigers タイガース 8 <UNK> 発売 中
```

この結果は、GBR モデルの特徴量重要度分析の結果とも一致する。もう一つの興味深い点は、ビジターのチーム名が重み付けられた点である。このことは、試合の観客数がビジターのチームに依存することを示唆している。

予定表にないイベントが発生する公園では、そうしたイベントの関連語に重みが付いた。上野恩賜公園における花見シーズンのある日の実際の入力の投稿の一例を以下に示す。

```
<CLS> <UNK> 予想 東京 埼玉 桜 開花 <UNK> 迎え まし
た <TARGET_PLACE> 来週 後半 <UNK> <UNK> 堂 桜 <UNK>
<TARGET_DATE> 頃 花見 楽しめ そう
```

ここでは、桜に関係する複数の単語が重み付けられた。このことは、ST モデルがマイクロブログ投稿から予定表に載らない未来のイベントさえも検出できることを示している。

ノイズとなる投稿 (例：コンサート DVD の広告) が多く得られるコンサート会場では、その投稿が実際の人口に関係しないことを示す単語に重みが付くという興味深い結果が見られた。日本武道館におけるイベント非発生日の実際の入力の投稿の一例を以下に示す。

```
<CLS> ... KING SUPER LIVE 2017 <UNK> BD <TARGET_DATE> 発
売 3 5 <TARGET_PLACE> 開催 された ライブ イベント 映像 化
```

ここでは、あるコンサートの BD 発売を告知する投稿のうち「映像化」の部分に重みが付いた。この結果は、ST モデルがノイズの多い投稿の中から有用なものを区別できているという仮説を支持するものである。

5 おわりに

本研究では、イベント発生場所周辺の人口予測の問題に取り組んだ。イベント発生場所周辺では、イベントの有無や種類に

表 7 東京ドームの各時間帯の重要度上位 10 位までの特徴量 ([d 日前 h] はその時点の人口).

12 PM	2 PM	4 PM	6 PM	8 PM
14	14	14	18	18
[1 日前 10 AM]	18	18	[DATE]	巨人
ふるさと	kyojin	予告先発	巨人	14
[1 日前 11 AM]	予告先発	[1 日前 4 PM]	当選	チケット
[1 日前 3 PM]	[1 日前 11 AM]	巨人	チケット	行き
日	巨人	指定席	ポールマッカートニー	ポールマッカートニー
jr 東日本	[1 日前 4 PM]	[1 日前 8 PM]	席	サザン
フェスタ	[1 日前 10 AM]	[1 日前 10 AM]	定価	当選
[2 日前 11 AM]	ロス	kyojin	サザンオールスターズ	ポール
[1 日前 8 PM]	祭り	落選	引換	パセラ

「混雑統計[®]」©ZENRIN DataCom CO., LTD.

より人の動きが不規則に変化するため、過去の人口情報だけを特徴量とした人口予測は本質的に不可能である。そこで、予測対象のイベント発生場所の名称と日付に言及するマイクロブログ上の投稿を追加の特徴量とすることで、未来のイベントについての手がかりに基づき人口を予測するモデルを提案した。実世界の混雑統計データと Twitter データを用いた実験によって、提案手法がベースライン手法より優れていることを示した。

特徴量重要度ならびに自己注意の重みの分析により、提案手法のモデルは予測時に入力の投稿の中でも解釈可能な単語に着目していることが分かった、今後の方向性として、こうした分析結果を応用して予測の根拠となる投稿を提示することを検討している。例えば「明日の上野公園は、この投稿で言及されているように桜が見頃なので、普段より 30%以上混雑します」といった形で予測結果を見せられれば、利用者が上野公園に行くべきか判断する材料になるため利便性が高まると考えている。

謝 辞

本研究の一部は、情報通信研究機構の高度通信・放送研究開発委託研究（課題 201：異分野データ連携によるスマートモビリティ基盤の研究開発）の成果です。

文 献

- [1] T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake shakes Twitter users: Real-time event detection by social sensors,” in *WWW*, pp. 851–860, 2010.
- [2] Z. Fan, X. Song, R. Shibasaki, and R. Adachi, “CityMomentum: An online approach for crowd behavior prediction at a citywide level,” in *UbiComp*, pp. 559–569, 2015.
- [3] B. Liao, J. Zhang, C. Wu, D. McIlwraith, T. Chen, S. Yang, Y. Guo, and F. Wu, “Deep sequence learning with auxiliary information for traffic prediction,” in *KDD*, pp. 537–546, 2018.
- [4] 塚田涼太郎, 倉浩森, 石渡祥之佑, and 豊田正史, “マイクロブログおよび携帯電話人口統計を用いた大規模イベント会場における人口変化の長期予測,” in *DEIM*, 2020.
- [5] R. Tsukada, H. Zhan, S. Ishiwatari, M. Toyoda, K. Umemoto, H. Shang, and K. Zettsu, “Crowd forecasting at venues with microblog posts referring to future events,” in *BSD*, 2020.
- [6] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, and Y. W. Teh, “Set Transformer: A framework for attention-based permutation-invariant neural networks,” in *ICML*, pp. 3744–3753, 2019.
- [7] T. Konishi, M. Maruyama, K. Tsubouchi, and M. Shimosaka, “CityProphet: City-scale irregularity prediction using transit app logs,” in *UbiComp*, pp. 752–757, 2016.
- [8] W. Yamada, D. Torii, H. Kikuchi, H. Inamura, K. Ochiai, and K. Ohta, “Extracting local event information from micro-blogs for trip planning,” in *ICMU*, pp. 7–12, 2015.
- [9] A. Jatowt, É. Antoine, Y. Kawai, and T. Akiyama, “Mapping temporal horizons: Analysis of collective future and past related attention in Twitter,” in *WWW*, pp. 484–494, 2015.
- [10] M. Onishi and S. Nakashima, “Mutual interaction model between the number of people in real space and the number of tweets in virtual space,” in *ICPR*, pp. 2073–2078, 2016.
- [11] J. He, W. Shen, P. Divakaruni, L. Wynter, and R. Lawrence, “Improving traffic prediction with tweet semantics,” in *IJ-CAI*, pp. 1387–1393, 2013.
- [12] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, “Solving the multiple instance problem with axis-parallel rectangles,” *Artificial Intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.
- [13] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
- [14] F. Rodrigues, I. Markou, and F. C. Pereira, “Combining time-series and textual data for taxi demand prediction in event areas: A deep learning approach,” *Information Fusion*, vol. 49, pp. 120–129, 2019.
- [15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *NIPS*, vol. 30, pp. 5998–6008, 2017.
- [16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *NAACL*, pp. 4171–4186, 2019.
- [17] Z. Wu, N. E. Huang, S. R. Long, and C.-K. Peng, “On the trend, detrending, and variability of nonlinear and nonstationary time series,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 38, pp. 14889–14894, 2007.
- [18] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC Press, 1984.