イベント発生の不均衡性に適合した人口変動予測

塚田涼太郎† 豊田 正史†† 梅本 和俊†† 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,toyoda,umemoto}@tkl.iis.u-tokyo.ac.jp, ††zettsu@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変動を引き起こす．こうした人口変動に起因する混乱を避けるためには，人口変動の発生時間や規模を事前に知ることが必要である．しかしながら，イベントの発生時間や参加者数は様々であり，またイベント日は非イベント日と比べて少数の事例しかないことから，イベントに起因する人口変動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し， クラスタリングに基づくオーバーサンプリングによって学習データの不均衡性を解消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改善することを確認した．

キーワード 人口変動予測，マイクロブログ，不均衡データ

1 はじめに

都市において野球の試合・コンサート，花見・デモ行進などのイベント[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道路で大量のイベント帰りの客に起因する渋滞が発生すると，その道路を利用しているバスやトラックの交通網に想定外の遅延が引き起こされることになる．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，予定通りの電車に乗れないなどの事態に遭遇して旅行の満足度が低下する．

イベント発生場所の人口変化の予測は，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，道路の利用者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，コンビニなどの小売店の販売網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．こうした動機のもと，我々のこれまでの研究 [2–4] では，マイクロブログ上に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を利用して，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予測の問題に取り組んだ．以下の投稿は，そ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 3 月 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．実際に，このような投稿の内容のテキストを解析して当該会場の過去の人口の時系列のデータと融合するアプローチによって，過去の人口情報のみを用いるベースライン手法より高い精度の予測が可能であることを確かめた．

しかしながら，より高精度な予測を実現するためには，イベント発生に係る以下 2 つの不均衡性が問題になる：

（1） イベント日が非イベント日に比べて非常に少ない
（2） イベントが多い会場と少ない会場が極端に分かれている

そこで本研究では，(1) k-means クラスタリングに基づくオーバーサンプリングによりイベント日の事例を水増しし，(2) 同一種別の会場のデータを利用した同時学習を行う手法を提案する（第 3 節）．実世界の混雑統計 ® データと Twitter（現 X）データを用いた実験の結果，提案手法が当該会場の人口情報のみを用いる場合と比較して予測誤差を最大 30%以上改善することを示した（第 4 節）．

2 関連研究

2.1 位置情報やマイクロブログを用いた人口変動予測

イベントの発生によって突発的に変化する都市の人口を予測するために，携帯電話などから収集された位置情報を含むデー

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．

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．Fan ら [5] 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 [6] 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 [7] 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．

一方で，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試みもなされてきた．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．He ら [9] 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．

これらの研究の流れを合流させ，我々のこれまでの研究 [2–4] 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した．自然言語の形で未来のイベントの位置と時間の情報が記述されているマイクロブログ上の投稿は，幅広い利用者から容易に収集可能であるという利点がある．さらに，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．しかしながら，第 1 節で述べたように，この手法はイベント発生の不均衡性に起因する学習データの品質の制約の影響を受けてしまう．

2.2 不均衡データに対する機械学習

主に分類問題について，一方のクラスのデータが他方のクラスのデータと比べて極端に少ないデータのことを不均衡データと呼ぶ．例えば学習データの中で，予測の対象として注目したい正例が負例よりも極端に少ないような不均衡データの場合，分類結果が負例に偏ってしまい精度が落ちてしまうため，不均衡データに対する機械学習の手法は従来から研究されている．

その手法の一つとして，オーバーサンプリングがある．オーバーサンプリングは，少ない方のクラスのデータを水増しすることで不均衡データのバランスを改善する手法である．SMOTE [10] は，少ない方のクラスのデータの近傍データを補完して作り出すオーバーサンプリングを行う．SMOTE の拡張版も数多く提案されている [11]．Borderline-SMOTE [12] は，クラス境界付近のデータを重点的にオーバーサンプリングすることで，分類を容易にする手法である．ADASYN [13] は，オーバーサンプリング対象となる少数データの近傍のローカルなデータ分布を考慮して補完データを作成する手法である．K-Means SMOTE [14] は，SMOTE によるオーバーサンプリングを行う前に k-means クラスタリングによるクラスター分けを行う手法で，少数データのバリエーションを保ったまま補完データを作成できる．

オーバーサンプリングは，都市における人口変化予測において，イベント発生の事例が通常時の事例と比べて著しく少ない不均衡データになる問題に対しても適用が試みられている．Anno [15] らは，乗換案内アプリケーションの検索クエリログと携帯端末の位置履歴の情報を組み合わせて，非日常的なイベントについても安定的な混雑予測が行えるためのオーバーサンプリング手法を提案した．

3 提案手法

3.1 マイクロブログ投稿の集合と人口の時系列を入力とした予測モデル

予測モデルのアーキテクチャは我々の以前の研究 [4] のモデルから，簡単のために時刻表現エンコーダを取り除いたものである．以下に概要を説明する．

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．

投稿の集合を集約する部分は，Set Transformer（以降，**ST**）[16] を中心とした構造になっている．ST 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

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

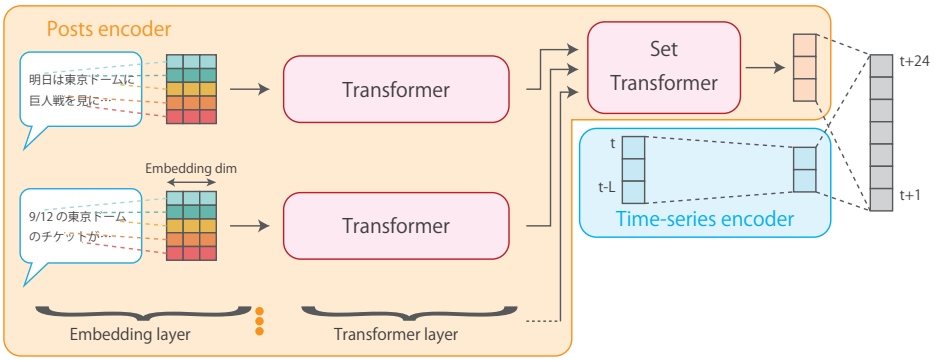

図 1　提案手法のモデルの概要.

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

種別	名称	イベント日数	投稿数
サッカースタジアム	日産スタジアム	15	6,229
	味の素スタジアム	29	15,842
野球場	東京ドーム	131	82,629
	明治神宮野球場	87	7,842
	横浜スタジアム	72	20,892
展示会場	パシフィコ横浜	53	21,310
	東京国際展示場	80	79,237
コンサート会場	横浜アリーナ	106	52,115
	日本武道館	114	41,507

「混雑統計 ®」©ZENRIN DataCom CO., LTD.

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 [7,17] 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 1 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・時系列の 2 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 [18] 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 [19] 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，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，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．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．データ融合による交通予測の既存のアプローチ [7] を参考に，これら 2 種類の入力（投稿，時系列）の埋め込み表現を連結し，後続の全結合層に入力する．

3.2　オーバーサンプリングと同時学習による不均衡性の緩和

表 1 に，本研究で予測対象とするイベント発生場所の年間イベント日数を記載した．この表からは第 1 節で問題提起したイベント発生の不均衡性を確認できる：

（1）　イベント日が非イベント日に比べて非常に少ない
（2）　イベントが多い会場と少ない会場が極端に分かれている

具体的には，(1) 最もイベント日数の多い東京ドームであっても過半数は非イベント日であることが分かる．また，(2) 例えば東京ドームと横浜スタジアムはどちらも野球場であるが，イベント日数としては東京ドームのほうが横浜スタジアムの 2 倍近い値になっている．このような 2 種類のイベント発生の不均衡性の問題は，イベント発生場所の人口変化予測を難しくする要因である．最も単純な解決策は，過去の人口情報を収集する期間を長くすることでイベント日の事例を学習データに多く含めることであるが，特定の場所の人口情報を長期間・一定品質で入手することは実用上困難である．そこで，本研究ではオーバーサンプリングと同時学習という 2 つの学習時の工夫により，この不均衡性の問題に対処する．

オーバーサンプリング：予測の対象として特に注目したいのはイベント日であるため，オーバーサンプリングによってイベント日のデータを水増しすることを考える．ただし，第 3.1 節で述べたように，入力データはマイクロブログ上の投稿の集合と人口の時系列からなる複合的なデータであるため，SMOTE [10] のような離散データに対する補完手法をそのまま適用することは難しい．そこで，人口の時系列に対して k-means クラスタリングを行い，各クラスターから等確率でサンプリングすることで，イベント日と非イベント日が均等に混在した学習データを作成する．具体的には，以下の操作を繰り返して学習データを拡張していく．

（1）　正規化した混雑度データに対して k-means でクラスタリ

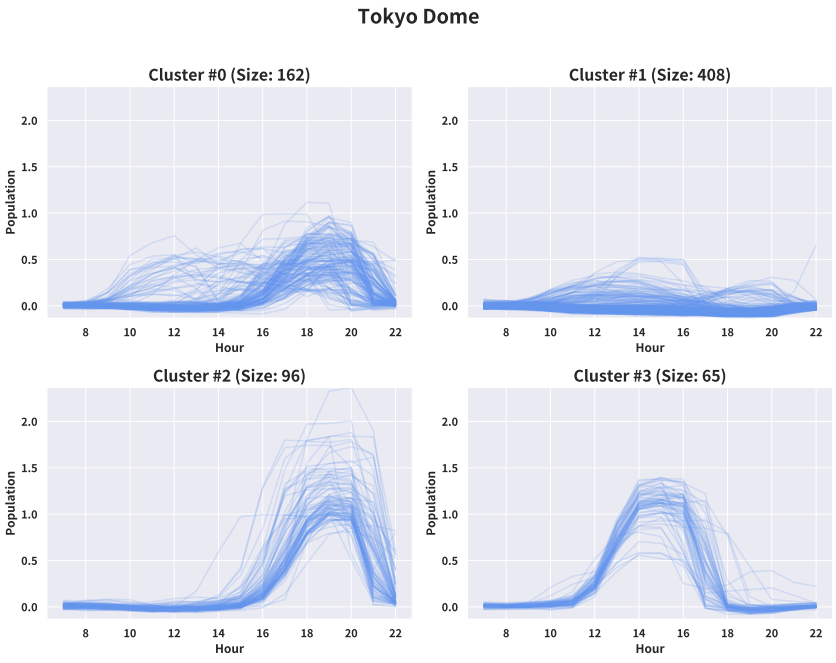

図 2　東京ドームのクラスタリング結果（$k = 4$）．人口は正規化されている．
「混雑統計 ®」©ZENRIN DataCom CO., LTD.

ングする

（2） 各クラスタを $\frac{1}{k}$ の確率で選択する

（3） 選択したクラスタの中から 1 日分をランダムに取り，学習データに追加する

これは，イベント発生場所の人口の時系列に対して k-means クラスタリングを行うと，イベント日と非イベント日のクラスターが分かれるという観察に基づいている．例として，東京ドームにおける $k = 4$ の場合のクラスタリング結果を図 2 に示す．このようにオーバーサンプリング結果の解釈が視覚的に容易であることも，k-means クラスタリングを用いる利点である．ここで，k はハイパーパラメータになる．

同時学習：東京ドームと横浜スタジアムを野球場という種別で見たときに，入力データと予測される人口の時系列との関係は類似していると考えられる．例えば「プロ野球のデーゲームに関する投稿が多数あれば，昼過ぎにピークのある人口変化が予想される」ことは，東京ドームでも横浜スタジアムでも同じであるはずである．こうした仮定のもと，イベント日数の多い会場のデータを有効活用するために，同一種別の会場をまとめて学習する同時学習の手法を導入する．具体的には，以下の手順で予測モデルを学習させる．

（1） 同一種別の全会場の学習データ（オーバーサンプリング済み）を結合して同時学習用の学習データとする

（2） 上記の学習データを利用して同時学習を行う

（3） ターゲット会場のデータのみを使って再学習（ファインチューニング）を行う

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　提案手法はイベント発生の不均衡性の問題をどのように改善しているか？（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ブ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，

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，許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）．

	ベースライン手法						提案手法								
	AR			LSTM			単一・OS なし			単一・OS あり			同時・OS あり		
イベント発生場所	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
日産スタジアム	0.33	0.07	0.11	0.00	0.00	0.00	0.58	0.47	0.52	0.48	0.87	**0.62**	0.73	0.53	0.62
味の素スタジアム	0.00	0.00	0.00	1.00	0.04	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.89	0.57	**0.70**
東京ドーム	0.85	0.24	0.37	0.81	0.20	0.32	0.65	0.66	0.65	0.80	0.71	0.75	0.81	0.77	**0.79**
明治神宮野球場	0.64	0.31	0.42	0.87	0.15	0.26	0.77	0.51	0.62	0.72	0.51	0.60	0.80	0.81	**0.81**
横浜スタジアム	0.65	0.21	0.32	0.83	0.07	0.13	0.97	0.54	0.70	0.90	0.50	0.64	0.92	0.96	**0.94**
パシフィコ横浜	0.39	0.32	0.35	0.33	0.13	0.18	0.67	0.51	**0.58**	0.67	0.51	0.58	0.62	0.51	0.56
東京国際展示場	0.45	0.67	0.54	0.56	0.41	0.47	0.52	0.70	0.59	0.52	0.70	0.59	0.51	0.81	**0.62**
横浜アリーナ	0.43	0.21	0.28	0.55	0.21	0.30	0.54	0.47	0.50	0.54	0.47	0.50	0.54	0.62	**0.58**
日本武道館	0.43	0.27	0.33	0.42	0.15	0.22	0.62	0.52	**0.57**	0.55	0.50	0.53	0.64	0.49	0.55
マクロ平均	0.46	0.25	0.30	0.60	0.15	0.22	0.59	0.49	0.52	0.57	0.53	0.53	0.72	0.67	**0.68**

「混雑統計 ®」©ZENRIN DataCom CO., LTD.

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 9 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．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 3.1 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 [20]．そこで，1 日の各時間ごとに過去の人口から算出した定常人口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．なお，定常人口は平日・祝休日のそれぞれに対して以下の要領で算出した値である．

（1） 各日の人口を $k = 3$ で k-means クラスタリングする
（2） 重心の最大人口が最小のクラスタ重心を定常人口とする

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，JUMAN[5]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，PyTorch[6]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，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．評価には式 (1) で定義される平均絶対誤差（MAE）を用いた．

$$\mathrm{MAE} = \frac{1}{N}\sum_{t=1}^{N} |\hat{X}_t - X_t| \tag{1}$$

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．

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（MAE）.

	ベースライン手法		提案手法		
イベント発生場所	AR	LSTM	単一・OS なし	単一・OS あり	同時・OS あり
日産スタジアム	465.11	415.86	356.00	358.40	**343.80**
味の素スタジアム	723.68	690.36	605.50	558.30	**493.60**
東京ドーム	6,819.39	6,681.91	4,471.80	4,471.80	**2,940.70**
明治神宮野球場	2,041.24	1,926.30	1,419.80	1,482.60	**1,285.90**
横浜スタジアム	1,872.04	1,716.61	1,422.10	1,353.60	**1,058.30**
パシフィコ横浜	2,106.85	1,981.06	1,471.80	1,471.80	**1,467.20**
東京国際展示場	2,471.88	2,361.67	1,911.00	1,911.00	**1,901.60**
横浜アリーナ	1,449.69	1,400.66	1,272.00	1,272.00	**1,240.40**
日本武道館	1,618.61	1,727.46	1,272.70	1,272.70	**1,195.20**
マクロ平均	2,174.28	2,100.21	1,578.08	1,572.47	**1,325.19**

「混雑統計 ®」©ZENRIN DataCom CO., LTD.

表 4　イベント発生日の予測誤差（MAE）.

	ベースライン手法		提案手法		
イベント発生場所	AR	LSTM	単一・OS なし	単一・OS あり	同時・OS あり
日産スタジアム	2,543.48	2,797.59	**1,737.70**	1,945.70	1,848.00
味の素スタジアム	3,811.12	3,813.54	4,141.50	3,126.50	**2,893.40**
東京ドーム	10,124.14	11,877.79	6,465.10	6,465.10	**4,702.50**
明治神宮野球場	4,113.13	4,822.52	3,199.50	2,996.10	**2,594.70**
横浜スタジアム	4,025.14	4,769.67	3,860.60	3,276.20	**2,136.40**
パシフィコ横浜	4,786.35	5,216.44	2,932.00	2,932.00	**2,925.80**
東京国際展示場	3,116.17	4,018.55	2,329.50	2,329.50	**2,216.30**
横浜アリーナ	2,307.84	2,396.94	1,938.70	1,938.70	**1,769.60**
日本武道館	2,405.64	2,596.76	2,151.50	2,151.50	**2,050.90**
マクロ平均	4,137.00	4,701.09	3,195.12	3,017.92	**2,570.84**

「混雑統計 ®」©ZENRIN DataCom CO., LTD.

オーバーサンプリングと同時学習を行った提案手法は，多くの場所で一貫して高い性能を達成した．特に，味の素スタジアムのようなイベント日数が少ない場所では，単一場所のデータをオーバーサンプリングせずに利用する場合と比較して予測精度が大きく向上した．このような場所では極端に少ない数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．提案手法はオーバーサンプリングによりイベント日のデータを水増しし，同時学習により同一種別の他の場所のデータを活用することで，イベント発生の不均衡性に対処して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（野球場），横浜アリーナ・日本武道館（主にコンサートが開催される会場）において，提案手法による予測精度の改善度合いが大きい．原因として，これらの場所における発生イベントが特徴的で，人口の時系列変化も類型化しやすいことが考えられる．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チーム名などの特徴的な単語が頻出する．また，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を含んでいる．そのような情報は会場によらず有用な情報であるため，同時学習の過程で汎化されて，イベント発生日の予測精度向上に寄与したと考えられる．

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．すべての場所において，提案手法は単一場所のデータをオーバーサンプリングせずに利用する場合より小さい誤差を達成した．誤差の削減の度合いは特に野球場で大きく，第 4.2 節のイベント有無の予測結果と同様の傾向である．

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）．実応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．

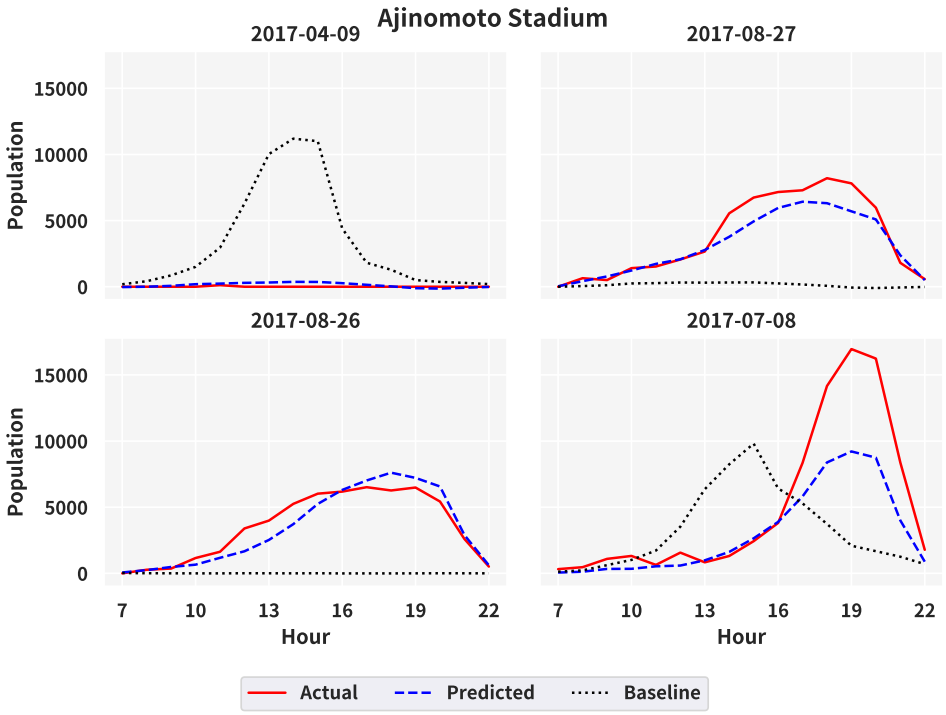

図 3　予測精度の改善幅が上位の日の実測値と予測値（味の素スタジアム）．

「混雑統計 ®」©ZENRIN DataCom CO., LTD.

4.4　Q3：事例分析

最後に，提案手法のオーバーサンプリングと同時学習が，イベント発生の不均衡性に適合できているかを実際の予測事例から分析する．

図 3 は，味の素スタジアムにおける予測精度の改善幅上位 4 例の日について，実測値と予測値（提案手法とベースライン）をプロットしたものである．2017 年 4 月 9 日の例では，実際にはイベントがないにも関わらずベースライン手法ではイベントがあるような時系列を予測しているが，提案手法ではイベントがないことを予測できている．2017 年 8 月 26 日，8 月 27 日の例では，実際にイベントが発生しており，ベースライン手法ではそのことを予測できていないが，提案手法では人口の規模・タイミングともに概ね予測できている．実際，この 2 日間は味の素スタジアムを会場としたコンサート[7]が開催されていた．2017 年 7 月 8 日の例では，提案手法もベースライン手法もイベントがあることを予測できているが，提案手法の方がより実測値に近いタイミングで予測できている．この日は 19 時キックオフの J リーグの試合[8]が開催されていた．

従来の手法は，イベント発生の不均衡性の影響を大きく受けるため，実際にはイベント日だったとしても非イベント日に近いような時系列を予測してしまう傾向にある．その問題が，オーバーサンプリングと同時学習を行う提案手法では改善されていることが分かった．特に，コンサートやサッカーの試合のような他会場の情報を汎化して利用できるイベントについて，想定通り予測精度が改善されていることが確かめられた．

7：https://www.ajinomotostadium.com/schedule/2017/0827_33.php

8：https://www.ajinomotostadium.com/schedule/2017/0708_53.php

5　おわりに

本研究では，イベント発生場所周辺の人口変化の予測における，イベント発生の不均衡性の問題に取り組んだ．イベント発生には (1) イベント日が非イベント日に比べて非常に少ない (2) イベントが多い会場と少ない会場が極端に分かれているという 2 つの不均衡性があり，このことが予測精度を悪化させる原因になっていた．そこで，(1) k-means クラスタリングに基づくオーバーサンプリングによりイベント日の事例を水増しし，(2) 同一種別の会場のデータを利用した同時学習を行うという 2 つのアプローチを提案した．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，提案手法が従来手法より優れていることを示した．

今後の研究の方向性として，より小規模なイベント発生場所についても本研究の提案手法を適用することを検討している．小規模なイベント発生場所ではイベント日自体が少なかったり，イベントに関する投稿が少なかったりすることが予想されるが，本研究の提案手法によりある程度緩和できると考えている．

謝　　辞

本研究成果は，国立研究開発法人情報通信研究機構（NICT）の委託研究（222C02）および，国立研究開発法人情報通信研究機構（NICT）の委託研究（JPJ012368C05401）により得られたものです．

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] 塚田涼太郎, 詹浩森, 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.

[3] 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.

[4] 塚田涼太郎, 詹浩森, 石渡祥之佑, 豊田正史, 梅本和俊, 商海川, and 是津耕司, "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," in *DEIM*, 2021.

[5] 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.

[6] T. Konishi, M. Maruyama, K. Tsubouchi, and M. Shimosaka, "CityProphet: City-scale irregularity prediction using transit app logs," in *UbiComp*, pp. 752–757, 2016.

[7] 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.

[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] J. He, W. Shen, P. Divakaruni, L. Wynter, and R. Lawrence, "Improving traffic prediction with tweet semantics," in *IJCAI*, pp. 1387–1393, 2013.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[11] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," *Journal of Artificial Intelligence Research*, vol. 61, pp. 863–905, 2018.

[12] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in *International Conference on Intelligent Computing*, pp. 878–887, 2005.

[13] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *IEEE International Joint Conference on Neural Networks*, pp. 1322–1328, 2008.

[14] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on K-Means and SMOTE," *Information Sciences*, vol. 465, pp. 1–20, 2018.

[15] S. Anno, K. Tsubouchi, and M. Shimosaka, "CityOutlook+: Early crowd dynamics forecast through unbiased regression with importance-based synthetic oversampling," *IEEE Pervasive Computing*, 2023.

[16] 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.

[17] 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.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, vol. 30, pp. 5998–6008, 2017.

[19] 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.

[20] 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.