

# COVID-19によるイベント開催状況の変化に適合した人口変化の予測

川崎 仁嗣<sup>†</sup> 塚田涼太郎<sup>†</sup> 豊田 正史<sup>††</sup> 是津 耕司<sup>†††</sup>

<sup>†</sup> 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

<sup>††</sup> 東京大学生産技術研究所 〒153-8505 東京都目黒区駒場 4-6-1

<sup>†††</sup> 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: <sup>†</sup>{kawasaki,tsukada,toyoda}@tkl.iis.u-tokyo.ac.jp, <sup>††</sup>zetttsu@nict.go.jp

**あらまし** 新型コロナウイルス感染症 (COVID-19) の流行に伴い、多くのイベントが開催自粛となり、その後も観客数に上限を設ける形でのイベント開催が行われてきたが、今後は上限を解除することが検討されている。上限が解除されたとしても、イベント参加者数を事前に想定し、密とならないようなイベント運営が当面は求められることになる。しかしながら、イベント参加者数の上限等が設けられている場合、それ以前のイベント時の過去データに基づく予測では十分な予測精度が期待できない。本研究では、イベント等の開催自粛期間や、参加者数上限がある期間において、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用いた予測手法をベースに、人口情報などを人口変動パターンごとにクラスタリングし、イベント自粛期間前のデータも含めてクラスターごとにサンプリングしたデータを学習することで予測精度を改善する手法を提案する。東京ドームにおける2020年のプロ野球シーズン期間を通して比較検証を実施した結果、今回の評価を行った期間においては単純に開催状況変化前、または変化後のデータのみで学習したモデルと比較し、本提案手法ではイベント有無の予測精度が20%程度改善されることを確認した。

**キーワード** 時系列人口予測, マイクロブログ, COVID-19

## 1 はじめに

新型コロナウイルス感染症 (以下、COVID-19) のパンデミック的な流行が起こる前までは、ドームやスタジアムで開催されるスポーツの試合やライブなどのイベント<sup>1</sup>実施時には、非常に多くの参加者がイベント会場付近に集まることで突発的な人口増加が発生していた。こうした人口増加は、会場だけでなく、会場までのアクセス手段である公共交通手段なども含め、参加者自身だけでなくイベントとは無関係な交通機関の利用者にも多大な影響を与えていた。

COVID-19の感染流行により、それ以前までに開催されていたような大規模なイベントは一時的にとりやめられた。このようなイベント開催自粛に加え、リモートワークなどの人との接触機会抑制や、ワクチン接種の推進により、COVID-19の感染状況が改善されるとともに、感染が広がりやすい状況などの知見が蓄積されてきたことにより、来場者数の上限等を設ける形でイベント開催も行われるようになった。しかし、昨今でも第6波の感染が広がっているなど、収束したとは言い切れない状況であり、今後の感染症再流行の可能性を踏まえた対策が求められている。

チケットなどの販売数から事前に来場者数の予測を行うこと

はできるものの、複数日に分かれるイベントでその期間内は有効なチケットや、そもそもチケットが不要なイベントの場合、特に混雑する公演日や時間帯を予測するのは容易ではない。また、チケットで日程が指定されているイベントであっても、開演前にグッズ販売が行われるなど、イベント開始時刻よりもかなり早い時間帯に混雑することがあり、混雑平準化や回避等の対策を行うためにはイベント開催時間帯以外も含めた1日の時系列での人口変動予測ができることが望ましいと言える。

人口変動の状況を取得するためにGPSを搭載したスマートフォンなどの携帯端末から収集された位置情報を活用し、年全体の規模で人口予測を行う手法が研究されている[2][3]。しかし、これらの手法は現在の状態や数時間先までの予測を行うことを想定しており、数日先といった長期的な予測を行うにはイベントや天候などの外的要因による非定常的な人口変動が無視できず、過去の人口情報のみに基づく人口変化の長期予測は難しい[4][5]。

数日先の人口変動、とりわけイベント時の人口変動の予測を行うには将来のイベントにおける来場者数を考慮する必要がある。この問題に対して、マイクロブログにおけるイベントに言及する投稿を用いることでイベントの発生時間・場所・規模を予測する手法が提案されている[6]。しかしながら、COVID-19の感染流行に伴うイベント開催自粛や、その後の人数上限を設けたイベント開催時における予測は、それまでの人口変動とは傾向が異なることから、新たにCOVID-19流行後の学習データが蓄積されないと高精度な予測が難しい。また、COVID-19に限らず、開催場所やオンラインとのハイブリッド開催などの

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、主催者が事前に日程を定めるもの(例: 野球の試合, コンサート)だけでなく、参加者が自然発生的に集合するもの(例: 花見, 列車運休時の駅混雑)の2種類がある。

開催方法変更により、今後も過去のデータだけでの予測では精度が十分にでない状況は発生しうると考えられる。

そこで本研究では、マイクロブログにおけるイベントに言及する投稿を用いる手法をベースに、開催状況が大きく変化した場合にも早期に精度が高い予測を実現するための手法を検討する。具体的には、過去の人口変動パターンをクラスタリングにより抽出し、直近の人口変動がクラスタリング結果とは大きく異なった場合に、人口変動パターンごとのデータを適切にサンプリングして学習データとして用いることで、開催状況が変化した際の予測精度劣化を改善することを試みる。実世界の混雑統計データと Twitter データを用いた予備的な実験の結果、提案手法を用いることで有観客でのイベント開催が再開された直後における予測において、イベント日の場合に見られるような人口変動を予測できていることを確認した。

本研究は、我々のこれまでの研究を以下の点で発展的に拡張したものである：

- 2020 年の COVID-19 感染症流行によるイベント開催状況の変化は人口変動のクラスタリング結果の比較により検知できることを示した。
- 人口変動パターンごとのデータ数を適切にサンプリングすることで状況変化発生時の予測精度劣化を改善できることを確認した。

## 2 関連研究

### COVID-19 に関連する特徴量の利用

COVID-19 の感染流行後におけるイベント等での人の移動や人口密度を推定する研究は、十分な移動データが蓄積されていないために、まだ活発には行われておらず、筆者らの知る限りこのような研究は見あたらない。一方で、COVID-19 感染症の感染経路特定や感染地域の予測などにおいて、感染者の発生傾向と人の移動傾向との相関が高いことが報告 [7] [8] されており、将来の人の移動や人数を予測することは重要な取り組みと考える。

また、COVID-19 感染症は株価にも影響を与えており、株価の予測アルゴリズムにおいて COVID-19 感染者数や死者数の特徴量として利用することで予測精度が改善されることが報告 [9] されている。本研究では、COVID-19 感染症に関連した特徴量を用いるのではなく、人口変動やマイクロブログ投稿のデータのみからイベント開催状況の変化を検出しモデル学習に利用するデータセットを適切に選定している。

## 3 提案手法

これまでの研究 [6] で報告しているように、将来のイベント発生時の人口予測を行うには、過去の人口情報のみでは困難であり、イベントに関連する特徴量としてマイクロブログ投稿を用いる手法を提案している。これは、マイクロブログにおいてイベント発生場所  $v$ 、未来の日付  $d$  を含む投稿が、 $d$  の前日以前に投稿されている場合に、イベント発生場所で将来発生する

人口変化を予測する特徴量として利用可能と考え、過去のイベント時における人口変動の時系列データと、マイクロブログの投稿データの Bag-of-Words 表現ベクトルとを学習データとして勾配ブースティング回帰 (GBR) により予測を行う手法、および、マイクロブログの投稿データを Transformer でエンコードした埋め込み表現と人口変動の時系列データ等をエンコードしたベクトルとからニューラルネットワークを用いたモデルで予測を行う手法との 2 つを提案している。

一方で、COVID-19 感染症の流行後においては、イベントにおける人数制限の適用やオンライン配信とのハイブリッド開催など、感染症流行前のイベント開催時とは人口変動が大きく異なっている。人口変動を把握するために、携帯電話の位置情報から推計された人口である混雑統計データ<sup>2</sup>を用いて、感染症流行前後の東京ドームにおける 1 日の人口変動を可視化した。図 1 は感染症流行前 (2019 年 3 月～2020 年 2 月)、図 2 は感染症流行初期 (2020 年 3 月) であり、両者とも横軸は 1 日の各時刻 (0 時～24 時)、縦軸は人口である。通常は 3 月にはプロ野球の開幕戦があるため、数万人の人口ピークが現れるはずであるが、2020 年は開幕戦が 6 月 19 日に変更されたため、3 月のプロットには試合開催時に見られる人口変動が存在しない。また、6 月 19 日～7 月 9 日は観客を入れない形での試合開催であったが、7 月 28 日～31 日に東京ドームで行われた試合では 5,000 人を上限とした人数制限のもとに開催されたため、図 3 には 5,000 人前後のピークとなる人口変動が見られる。

このような開催状況の変化が起こった場合、モデル学習時に用いる過去データには含まれない人口変動パターンであることから、予測精度が劣化する可能性が高い。そこで、予測モデルへの入力特徴量データから開催状況変化を検知して、人口変動パターンごとのデータ数を考慮したモデル学習を行うことで精度劣化を改善する手法を提案する。

### 3.1 開催状況変化の検知

開催状況を知るための単純な手段としては、事前にイベント主催者等が公表しているイベント開催日時や会場と併せて、観客有無やオンライン開催などの情報を収集することが考えられる。しかしながら、多様なイベント発生場所の中でも公式にスケジュールが公開されている会場は限られている。さらに、観客有無や人数制限、オンライン開催有無などの情報が必ずしも一元的に公開されるとは限らず、自動的に情報を収集するのは難しい。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、そのような主催者の公開する情報に依存せずにイベント開催状況の変化を自動的に収集できることが望ましい。

本研究では未来のイベントでの人口予測を行うことが目的であるため、開催状況の変化のうち、人口変動パターンに変化が

2: 「混雑統計<sup>®</sup>」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。

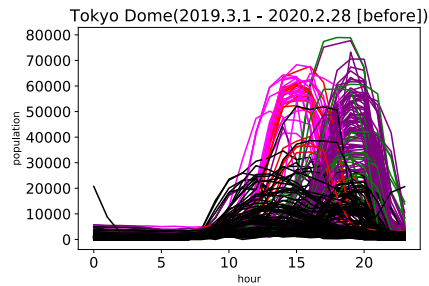


図1 COVID-19 感染症流行前の東京ドームにおける1日の人口変動パターン

「混雑統計<sup>®</sup>」©ZENRIN DataCom CO., LTD.

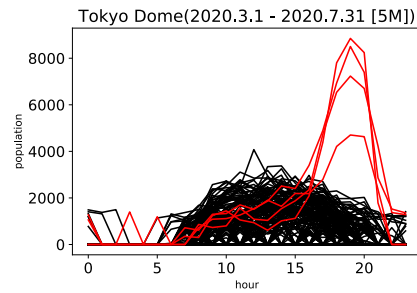


図3 人数制限付でのイベント再開直後（2020年7月）までの東京ドームにおける1日の人口変動パターン

「混雑統計<sup>®</sup>」©ZENRIN DataCom CO., LTD.

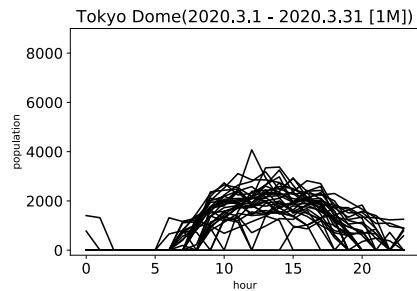


図2 COVID-19 感染症流行直後（2020年3月）の東京ドームにおける1日の人口変動パターン

「混雑統計<sup>®</sup>」©ZENRIN DataCom CO., LTD.

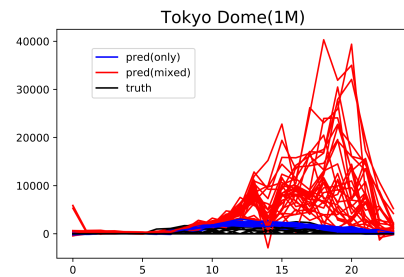


図4 東京ドームにおける2020年4月の人口変動予測結果

「混雑統計<sup>®</sup>」©ZENRIN DataCom CO., LTD.

現れる場合を考慮すれば良い。図1は14時前後に人口がピークとなるパターンと、19時前後に人口がピークとなるパターンがあることが分かる。これはそれぞれ、プロ野球のデイゲームとナイターの試合があった場合の人口変動である。一方で、図2はイベント時に特徴的である急激なピークが現れる人口変動が見られず、日常的（非イベント日）な人口変動とほぼ変わらない人口変動となっている。また、図3では、人数規模は異なるものの図1にも見られた19時前後をピークとした人口変動があることが分かる。これは、人数制限ありの有観客でナイターの試合が開催されたことが人口変動として現れていると考えられる。

このような特徴を鑑みて、人口変動データのクラスタリングを行い、イベントと思われるクラスターの有無や時間帯を比較することで、開催状況に変化があったことを検出できると考えられる。

### 3.2 人口変動パターンごとのデータ数を考慮したモデル学習

COVID-19 感染症流行後のイベント開催自粛期間においては、基本的にイベント自体の開催が自粛されているか、もしくはオンライン開催への移行などにより、図2のようにイベント会場周辺の人口変動は非イベント日と同様になっている。東京ドームの場合は、2020年7月中旬までは有観客の試合は実施されていない。東京ドームでの4月の人口変動を予測した結果を図4に示す。赤線グラフは開催状況が変化する前の2019年3月から2020年2月までのデータで学習したモデルによる予測結果であるが、マイクロブログへの投稿データはイベント開催

の延期や中止に関する投稿が見られるようになるものの、過去の学習データからはイベント中止の投稿と人口変動との相関があまり学習されていないことから、予定通りイベントが開催される前提での予測となってしまう、予測誤差が大きくなってしまった（黒線グラフが真値）。それに対して、青線グラフは開催状況変化後のデータのみを用いて予測を行った結果であり、予測誤差は小さくおさえられている。これは学習データにイベント日が含まれず、予測モデルがイベント日の予測をしなくなったことによるものである。

その後、7月下旬より5,000人を上限とした有観客でのイベント開催が可能となり、東京ドームでは7月28日より試合が開催されるようになっている。この際の人口変動を予測するには、前述した開催状況変化後のデータのみでの予測モデルではイベント日の予測が行われないと考えられるため、開催状況変化前のデータも含めて学習の方が予測精度が向上する可能性が高い。図5の青線グラフは3～6月のデータのみを学習して予測した結果であり、イベントに伴う19時前後をピークとした人口変動の予測がされていない一方で、開催状況変化前のデータも含めた学習モデルの予測結果（図5の赤線グラフ）では人数規模は大幅にずれているもののイベント時に特有の急激な人口変動は予測できている。

開催状況変化後のデータにもイベント開催時の人口変動等が含まれるようになれば、開催状況変化後におけるイベント時の人口変動を予測できると考えられ、図1にあるような14時前後に人口がピークとなるパターンと、19時前後に人口がピークとなるパターンの人口変動が学習データに含まれるように、

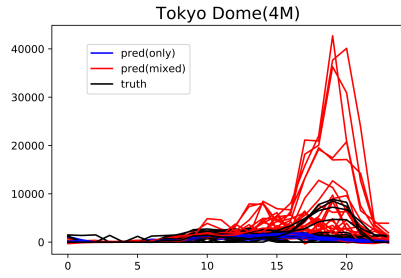


図 5 東京ドームにおける 2020 年 7 月の人口変動予測結果  
「混雑統計<sup>®</sup>」©ZENRIN DataCom CO., LTD.

不足しているパターンの人口変動を開催状況変化前のデータからサンプリングする手法を提案する。これにより、開催状況変化後のデータが不十分な期間は開催状況変化前のデータを用いることでイベント時の人口変動が予測できるとともに、開催状況変化後のデータが十分に蓄積されてくれば開催状況変化後のデータのみでの予測に切り替わることで、人数制限も踏まえた人口変動が予測できるようになる。

具体的には、開催状況変化前の人口変動データ、および、開催状況変化後の人口変動データ（予測モデルの特徴量データ）のそれぞれに対してクラスタリングを適用する。東京ドームの場合、平日、休日のそれぞれで非イベント日、デイゲーム開催日、ナイター開催日の 3 つにクラスタリングしており、図 6 に示すように、開催状況変化後の期間が 7 ヶ月（7M のグラフ）になれば、開催状況変化後のデータだけでも各クラスターの要素数が一定数以上現れるようになった。なお、図 6 のグラフにおける点はクラスタリングを行った際のクラスター重心であり、開催状況変化前のクラスター重心を最大人口の比率で正規化したものである。開催状況変化前後で同一のクラスター同士のデータを母集団としてサンプリングを行い学習データとして用いる。この際、開催状況変化後のデータでクラスターの要素数が少なければ、開催状況変化前のクラスターのデータからサンプリングが行われる。

また、開催状況変化後のデータが蓄積されてくれば、開催状況変化前のデータを用いるよりも、開催状況変化後に合わせた予測が可能であることから、最近のデータを重視してサンプリングを行う。具体的には、図 7 に示すように、サンプリング前のデータを日付順に並べて、最新のデータから過去に遡った際のデータ数を  $d$  としたとき、以下の式 (1) のような指数分布に従う確率密度関数によりサンプリング確率を決定した。なお、式中の  $\lambda$  はパラメータであり、値が大きいほど過去データのサンプリング確率は下がるため、直近のデータのみが重視されるようになる。各クラスター内でのデータは必ずしも連続した日付とは限らないため、データ数と日数は必ずしも一致せず、クラスター内のデータが少ない場合は、より古いデータまでサンプリング対象となる。さらに、クラスターごとに要素数が異なるため、そのまま確率密度関数を適用すると、クラスターごとに最も古いデータのサンプリング確率が異なることとなってしまう。そのため、最も要素数の大きいクラスターに式 (1) の確

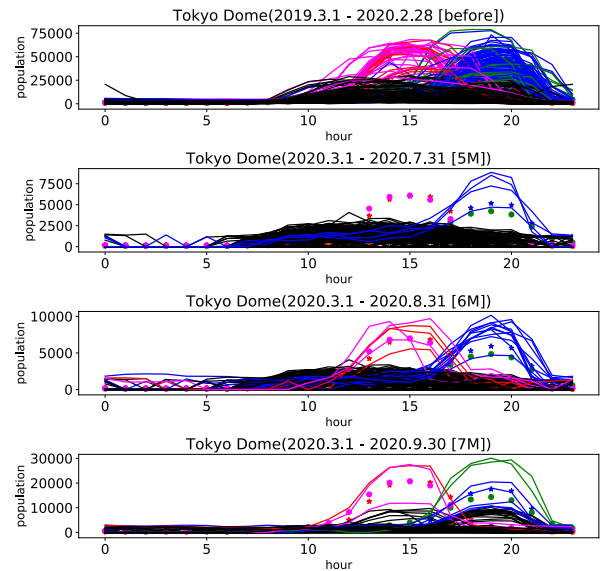
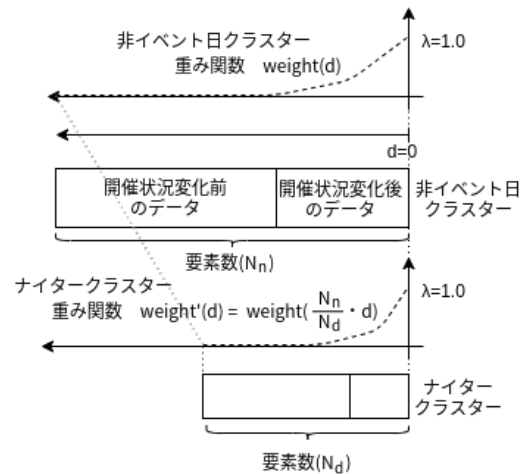


図 6 開催状況変化前とイベント再開後の東京ドームにおける 1 日の人口変動  
(クラスターごとに色分け)

「混雑統計<sup>®</sup>」©ZENRIN DataCom CO., LTD.



※デイゲームクラスターの重み関数も同様に正規化

図 7 データのサンプリング確率

率密度関数を割り当て、その他のクラスターは要素数の比に応じた係数を  $d$  に乗じてステップ幅を変えることで、クラスター内で最古のデータに対するサンプリング確率が同じとなるように正規化した関数（図 7 の  $weight'(d)$ ）を用いる。

$$weight(d) = \lambda e^{-\lambda d} \quad (1)$$

## 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。実験は以下の 2 つの観点で実施した。

- 開催状況変化後のイベント発生日と非イベント日との判別精度

表 1 評価に用いた人口統計データとマイクロブログ投稿データの期間

	学習データ	評価データ
4M	2020.03.01～2020.06.30	2020.07.01～2020.07.31
5M	2020.03.01～2020.07.31	2020.08.01～2020.08.31
6M	2020.03.01～2020.08.31	2020.09.01～2020.09.30
7M	2020.03.01～2020.09.30	2020.10.01～2020.10.31
4M+b 5M+b 6M+b 7M+b	上記期間に加えて、 2019.03.01～ 2020.02.28 を追加	上記期間と同様

- イベント発生日の時系列人口変動の予測誤差

#### 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り、一部データの期間や対象場所に違いはあるものの、基本的には先行研究 [6] での条件を踏襲している。時空間人口統計データについては、3 章で述べた混雑統計データを使用した。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 2019 年 3 月から 2021 年 2 月までの約 2 年分から、表 1 に示すように開催状況変化前のデータとして 2019 年 3 月から 2020 年 2 月まで、開催状況変化後は 1 か月分であれば 2020 年 3 月 1 日から 2020 年 3 月 31 日まで、4 ヶ月分であれば 2020 年 3 月 1 日から 2020 年 6 月 30 日まで、それ以降も同様に 7 ヶ月分（2020 年 3 月 1 日から 2020 年 9 月 30 日）までのデータを学習に用いており、学習データ末日の翌月 1 か月分を評価データとして利用した。

予測対象のイベント発生場所には、プロ野球の試合が開催される東京の 1 施設（東京ドーム）を選んだ。なお、開催状況変化は人数上限等が適用される団体等が主催するイベントで主に発生することから、団体がイベントを主催する施設を対象として選定し、花見などの主催団体が不明瞭となるイベントの実施場所（公園など）は除外した。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、単一メッシュに収まらない広い場所では、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、回帰モデルで利用する 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 [10]。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<sup>3</sup>を利用した。

3：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

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<sup>4</sup>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。

**サンプリング条件.** クラスタリングしたデータからのサンプリング数は、各クラスタにつき 500 サンプルとした。これは今回の評価対象とする学習データ期間のうち最長となる 7M+b の場合においても、データのダウンサンプリングが発生しないように十分なサイズとするためである。サンプリング確率を決定する確率密度関数のパラメータ  $\lambda$  については、0.005～0.1 の範囲で設定した。 $\lambda$  を変化させた際の、日付ごとのサンプリング回数は図 8 に示すとおりであり、非イベント日のクラスターを黒色、イベント日におけるデイゲームやナイターのクラスターを赤および青色のバーで表した。なお、図中の (sampling) は確率分布を一様分布とした場合のサンプリング結果であり、 $\lambda = 0.005$  に近い結果となる。また、 $\lambda$  を大きくすると直近のデータが高頻度でサンプリングされるが、クラスターごとにサンプリングをするため、4M のように開催状況変化後のデータに存在しないクラスターがある場合は開催状況変化前のデータからサンプリングされる。

**ベースライン手法.** 比較する上でのベースラインとして、開催状況変化後（2020 年 3 月以降）のデータだけで学習を行ったモデル、および、開催状況変化前後（2019 年 3 月～2020 年 2 月、および、2020 年 3 月以降）のデータで学習を行ったモデルを用いた。具体的な学習データの期間は表 1 に示したとおりである。また、クラスターからデータをサンプリングする際に、サンプリング確率を指数分布ではなく一様分布としたモデル（図中では `smpl` と表記）についても併せて評価を行った。

**評価指標.** 提案手法の有効性の評価は、イベント有無の予測精度と人口変化の予測誤差の 2 つの粒度で行った。イベント有無の予測精度評価では、各手法がイベント発生日と非イベント日を分類できるかを確認した。しかしながら、3.1 節で述べた通り、必ずしもイベントスケジュールは公開されておらず、またイベントが開催されていてもオンライン配信限定イベントのように人口変動を及ぼさない場合もある。そこで、先行研究 [6] と同様に、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）に属する日を非イベント日、そうでない日をイベント発生日とした。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、非イベント発生日が属する 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。

ザーの集合を拡張してきたものである。

4：http://nlp.ist.i.kyoto-u.ac.jp/?JUMAN



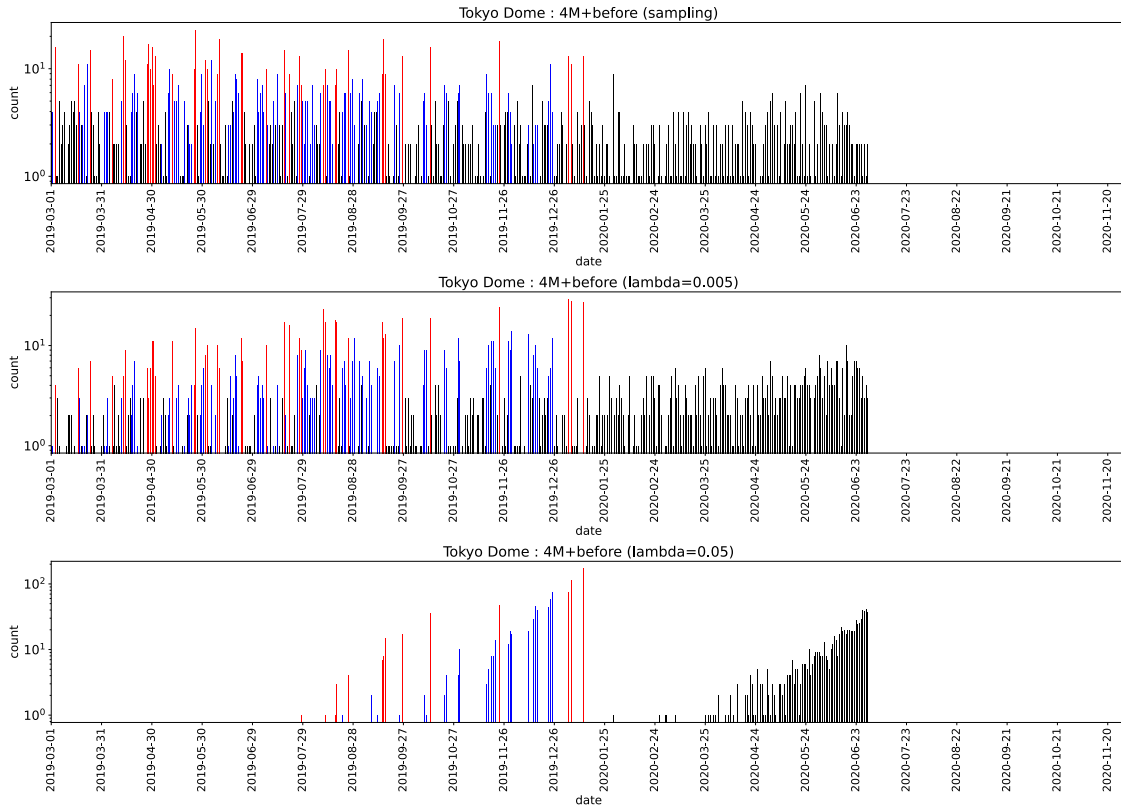


図 8  $\lambda$  値ごとのサンプリング回数ヒストグラム

「混雑統計<sup>®</sup>」©ZENRIN DataCom CO., LTD.

人口変動の移動平均（幅 2）を  $\bar{X}$ ，1 日の人口変動の平均を  $\mu$ ，1 日の人口変動の標準偏差を  $\sigma$  とするとき，以下の式（2）を満たす場合にイベント発生日として時系列人口を予測していると判断する．移動平均の幅については，イベント（今回の場合は野球の試合）の開催時間が最低 2 時間以上はあることから，予測結果も 2 時間の平均で評価をすることとした．これにより，ある時刻だけ極端に大きな人口が予測された場合に，イベント発生として分類されないようにした．提案手法の性能をイベント有無の 2 値分類問題として Precision, Recall, F 値で評価した．

$$\max(\bar{X}) \geq 2\sigma_X + \mu_X \quad (2)$$

人口変動の予測誤差評価は，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，対象時刻  $t(0 < t \leq 24)$  において，人口変動の実測値  $\hat{X}_t$  と予測モデルによる時系列人口変動  $X_t$  とするとき，評価には式（3）で定義される平均絶対パーセント誤差（MAPE）を用いた．なお，評価対象とした時刻範囲は，イベント等による人口変動が発生する朝 6 時以降，当日の 24 時までとした．

$$MAPE = \frac{1}{N} \sum \frac{\sum_{t=6}^{24} |\hat{X}_t - X_t|}{\sum_{t=6}^{24} \hat{X}_t} \quad (3)$$

なお，ここで  $N$  は評価対象の日数である．

## 4.2 イベント有無の予測

各モデルによる予測結果に基づくイベント有無の判定精度を図 9 に示しており，以下の 6 種類の学習データによりそれぞれ学習したモデルを比較した．

- 開催状況変化前のデータのみ : before
- 開催状況変化後のデータのみ :  $xM$   
 $x$  は学習データの期間（4～7）
- 開催状況変化前後のデータ :  $xM+b$
- クラスターごとにサンプリング（一様分布）した開催状況変化後のデータ :  $xM(\text{smpl})$
- クラスターごとにサンプリング（一様分布）した開催状況変化前後のデータ :  $xM+b(\text{smpl})$
- クラスターごとにサンプリング（指数分布）した開催状況変化前後のデータ :  $xM+b(\lambda)$

カッコ内は式（1）のパラメータ  $\lambda$  の値

開催状況変化前のデータのみ（before）では recall が低く，イベント日としての予測がほとんどできていないことが分かる．人数制限によりイベント時の最大人口が低くなっていることから，開催状況変化前における非イベント日の定常的な人口変動幅の範囲に収まってしまい，イベント開催時の人口変動として扱われていないことが要因と考えられる．

人数制限下での有観客の試合が始まった 7 月が評価期間に含まれる 4M, 4M+b での予測では，開催状況変化後の学習データ（4M）にイベント日が含まれないことから，イベント

日の人口変動が予測できていない一方で、開催状況変化前後のデータを用いた場合 ( $4M+b$ ,  $4M+b(\text{smpl})$ ) や、提案手法 ( $4M+b(0.005) \sim 4M+b(0.1)$ ) では、開催状況変化前のイベント時のデータが含まれることから、イベント日の人口変動が予測できていることが分かる。ただし、開催状況変化前のデータを増やすことで recall は改善するものの、precision が悪化しており、必ずしも開催状況変化前のデータが多いほうが良いというわけではなかったが、提案手法での  $\lambda$  が  $0.04 \sim 0.05$  の範囲では最も良い精度となった。

5M~6M の期間は、人数制限下による試合が開催された際のデータが含まれるため、開催状況変化後のみのデータだけで学習を行ったモデル (5M, 6M) でも  $4M+b$  と同程度の精度となっていたが、7M の期間では 5M や 6M の期間と比較しても精度が低下している。7M では評価期間に 10 月が含まれているが、人数制限が 5,000 人から収容人数の 50 % に緩和する方針となり、東京ドームでは 10 月 6 日の試合から人数を 1 万 9000 人とすることが発表 [11] されており、人数制限変化前のデータしか持たない 7M のモデルでは予測精度が低下したと考えられる。一方で、提案手法 ( $7M+b(0.005 \sim 0.1)$ ) や開催状況変化前のデータをサンプリングするモデル ( $7M+b(\text{smpl})$ ) では、5,000 人の人数制限が適用される前のイベント時のデータを学習データとして持っていることから、7M を学習データとするモデルよりも精度が高くなっており、特に提案手法では 5M~6M と 7M の期間とで最高精度に大きな差は見られなかった。

また、6M~7M の期間は、開催状況変化前のデータを加えたモデル ( $6M+b$ ,  $7M+b$ ) の精度が、開催状況変化後のデータのみのモデル (6M, 7M) と比較し精度が落ちているものの、開催状況変化前後のデータでクラスターごとに一様分布でサンプリングを行ったモデル ( $6M+b(\text{smpl})$ ,  $7M+b(\text{smpl})$ ) では  $6M(\text{smpl})$  や  $7M(\text{smpl})$  に近い精度が出ていることから、開催状況変化前のデータが予測精度悪化を引き起こしているというわけではないと考えられる。単純に開催状況変化前のデータを加えた場合、野球のシーズンオフ期間も含まれるため、非イベント日のデータが支配的になってしまうが、クラスターごとにサンプリングを行うことで非イベント日やイベント日のサンプル数が揃えられたことで、相対的にイベント日の予測が行われやすくなり recall が改善したことが要因と考えられる。

さらに、提案手法ではサンプリング時に直近のデータを重視するようにしているため、クラスターごとのサンプル数を揃える効果だけでなく、直近の状況を学習しやすくなると想定され、5M~7M の期間において提案手法が最も高い F 値を示しており、5M では  $0.04 \leq \lambda \leq 0.1$ 、6M では  $0.04 \leq \lambda \leq 0.06$ 、7M では  $0.04 \leq \lambda \leq 0.07$  において一様分布でサンプリングを行ったモデル ( $5M+b(\text{smpl}) \sim 7M+b(\text{smpl})$ ) よりも精度が高くなった。

これらの結果より、提案手法におけるクラスターごとのサンプリングや、直近データを高頻度でサンプリングする手法が予測精度改善に寄与していると考えられる。

### 4.3 時系列人口変動の予測

イベント日における実際の人口変動とモデル予測による人口変動との予測誤差を図 10 に示しており、開催状況変化後のデータが少ない 4M, 5M においては  $\lambda$  の値によっては提案手法の予測誤差が大きい場合もあるものの、単純に開催状況変化前のデータを混ぜた場合 ( $4M+b$ ,  $5M+b$ ) や、開催状況変化前後のデータを一様分布の確率でサンプリングした場合 ( $4M+b(\text{smpl})$ ,  $5M+b(\text{smpl})$ ) よりも誤差を小さくできている。特に 4M の場合は、人数制限等がある形でのイベント開催であるにもかかわらず、イベント時のデータは開催状況変化前のみであるため、人数規模が過大に予測されることにより誤差が大きくなっている。開催状況変化後のデータのみを用いる場合 (4M) は非イベント日としての予測しか行われておらず、人数制限があるイベントはたかだか 5,000 人程度の人口変動であることから、誤差も小さく算出されるものの、イベント時の人口変動が予測できていないため、適切な予測ではないと考える。

また、開催状況変化後のデータが蓄積されてきた 6M~7M においても、提案手法では直近のデータであるほど高頻度でサンプリングしているため、開催状況変化前後のデータを一様分布の確率でサンプリングした場合 ( $6M+b(\text{smpl})$ ,  $6M+b(\text{smpl})$ ) よりも誤差を小さくできしており、特に 6M では開催状況変化後のデータだけで学習した場合 (6M) の誤差よりも  $\lambda = 0.05$  では小さい誤差となった。

## 5 おわりに

本研究では開催状況が大きく変化した場合にも早期に精度が高い予測を実現するための手法検討に取り組んだ。過去の人口変動パターンをクラスタリングにより抽出し、直近の人口変動がクラスタリング結果とは大きく異なった場合に、人口変動パターンのクラスターごとに、開催状況変化前後のデータからより直近のデータを重視して確率的にサンプリングを行い学習データとして用いる手法を提案した。これにより、開催状況が変化した際のイベント日の人口変動予測精度の改善をめざした。実世界の混雑統計データと Twitter データを用いた、東京ドームでの 2020 年 6 月~10 月までの予測精度の比較実験においては、提案手法によりベースライン手法と比較して時系列人口の予測誤差を同程度に抑えつつ、イベント有無の予測精度が 20% 程度改善されることを確認した。

## 謝 辞

本研究成果は、国立研究開発法人情報通信研究機構の委託研究「ウイルス等感染症対策に資する情報通信技術の研究開発」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] Z. Fan, X. Song, R. Shibasaki, and R. Adachi, "CityMo-

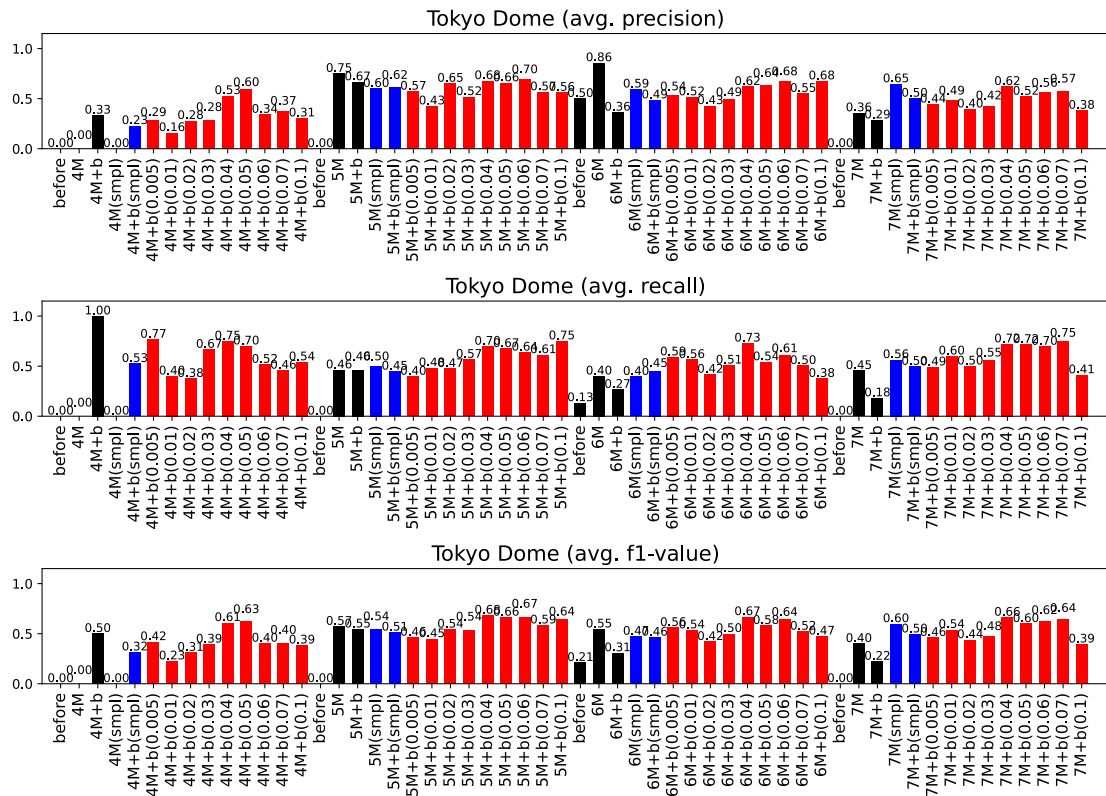


図9 イベント有無予測の評価結果  
「混雑統計」® ©ZENRIN DataCom CO., LTD.

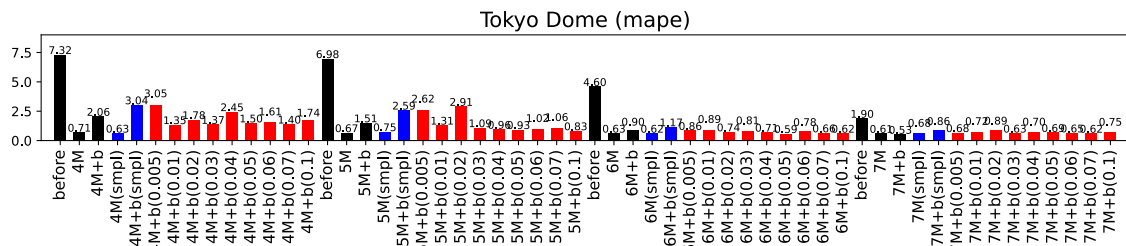


図10 時系列人口変動の平均予測誤差  
「混雑統計」® ©ZENRIN DataCom CO., LTD.

mentum: 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, 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.
- [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] 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.
- [7] R. Jiang, Z. Wang, Z. Cai, C. Yang, Z. Fan, T. Xia, G. Matsubara, H. Mizuseki, X. Song, and R. Shibasaki, “Coun-

trywide Origin-Destination Matrix Prediction and Its Application for COVID-19,” ECML/PKDD (4), pp. 319–334, 2021.

- [8] C. Alket, L. Marco, M. Marco, and Z. Franco, “Sensing and Forecasting Crowd Distribution in Smart Cities: Potentials and Approaches,” IoT, vol. 2, no. 1, pp. 33–49, 2021.
- [9] R. Jindal, N. Bansal, N. Chawla and S. Singhal, “Improving Traditional Stock Market Prediction Algorithms using Covid-19 Analysis,” in ESCI, 2021.
- [10] 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.
- [11] 10月6日～15日の東京ドーム公式戦6試合の入場券販売について  
[https://www.giants.jp/G/gnews/news\\_3915255.html](https://www.giants.jp/G/gnews/news_3915255.html)