GPS位置履歴情報と鉄道の乗換検索履歴を用いた異常混雑事前予測

安納 爽響[1]　坪内 孝太[2]　下坂 正倫[1]

概要：GPS 位置履歴情報を用いた都市における活動人口の予測は，都市特性の把握による災害対策や効率的な出店計画等に役立つ．これに加えて突発的な混雑等，日常には見られない異常な活動人口パターンを予測することは，人々の日常の生活における交通機関の選択から混雑による事故を未然に防ぐための人員配置など，様々な観点で重要な問題である．これまで提案された活動人口の予測手法では，定常的な活動人口の推移を精度よく予測することができる．一方で，突発的な混雑時の混雑度・活動人口の予測においては，精度や，予測可能な異常が最長で 1 時間後という時間的側面で課題が存在した．そこで本研究では，訪問者数を正確に反映する GPS 位置履歴とそれを用いた既存の活動人口予測手法に加えて，乗換検索履歴の持つユーザのスケジュール情報という特性を用い，活動人口の比で定義される混雑度での回帰と頑健性担保のためのオーバーサンプリングを行うことで，時間的制約に縛られない突発的な混雑にも対応可能な活動人口とその混雑度の予測手法を提案する．実際の乗換検索履歴と GPS 位置履歴を用いた大規模実験を行い，提案手法がラグビー W 杯 3 位決定戦が行われた味の素スタジアムでの異常混雑を 1 週間前から，既存手法よりも正確に捉えられることを示す．

1. 序論

都市における人々の生活は，通勤や通学等の日常的な行動だけでなく，コンサート・スポーツ観戦などのイベント等と大きく根付いている．このようなイベントは，会場や会場までの交通機関網に群衆が殺到するなど，予期しない都市動態の変動により不慮の事故の原因になりうる．例えば 2001 年の明石花火大会歩道橋事故では，歩道橋に殺到した人々が折り重なって倒れ，11 人死亡，247 人が重軽傷を負った．このような予期せずに起こる混雑を異常混雑と呼び，その都市異常の詳しい状況を事前に把握することで，安全なイベント計画や警備員の効率的な配置等により，事故防止に貢献することができる [4]．また，この異常混雑状況の予測は，混雑を回避した交通機関の選択を手助けできるなど，人々の日々の生活における意思決定にも役立てられる．

近年，人々の都市における活動からセンシングされたビッグデータ，特に GPS 位置履歴や鉄道の乗換検索アプリの検索履歴を用いた，都市動態の解析が盛んに行われてる [2], [3], [5]，これは都市計画 [10]，災害対策 [6]，公共サービス [7]，異常混雑の検知や予測などに役立てられている．

しかし，異常混雑の事前の予測に焦点を当てた時，従来のセンシングデータに基づくデータ・ドリブンな活動人口予測手法には課題が存在する．

GPS 位置履歴の活動人口に対する時空間的特性を用いて，Shimosaka ら [5] は，天気や曜日等のコンテクストをから，定常的な活動人口推移を予測する手法を提案している．しかし突発的な活動人口の増加に対しては，天気・曜日等に条件づけることができず，予測は不安定になりやすい．一方で Fan ら [2] は，GPS 位置履歴をトラッキングすることで人口流をモデリングし，人口流の変化の検知による異常混雑の発生を予測する手法を提案している．しかし，人口流に変化が現れるのは発生する異常の直前であり，モデルが予測可能な混雑は最大で 1 時間後に限られてしまう．従って，異常混雑予測をより未来まで行うには，未来における活動人口の推移を，混雑まで含め条件づけられるようなコンテクストを得られるデータソースが必要である．

これに対して Konishi ら [3] は，一週間後の異常混雑予測を目的として，乗換検索履歴を用いた手法を提案している．この手法では，乗換検索履歴がユーザのスケジュール情報を表すと同時に，実際の鉄道利用者数の動態に対しても相関があるとの仮定をおいている．しかし GPS 位置履

[1] 東京工業大学 情報理工学院 情報工学系 Department of Computer Science, School of Engineering, Tokyo Institute of Technology
[2] Yahoo! JAPAN 研究所 Yahoo! Japan Corporation

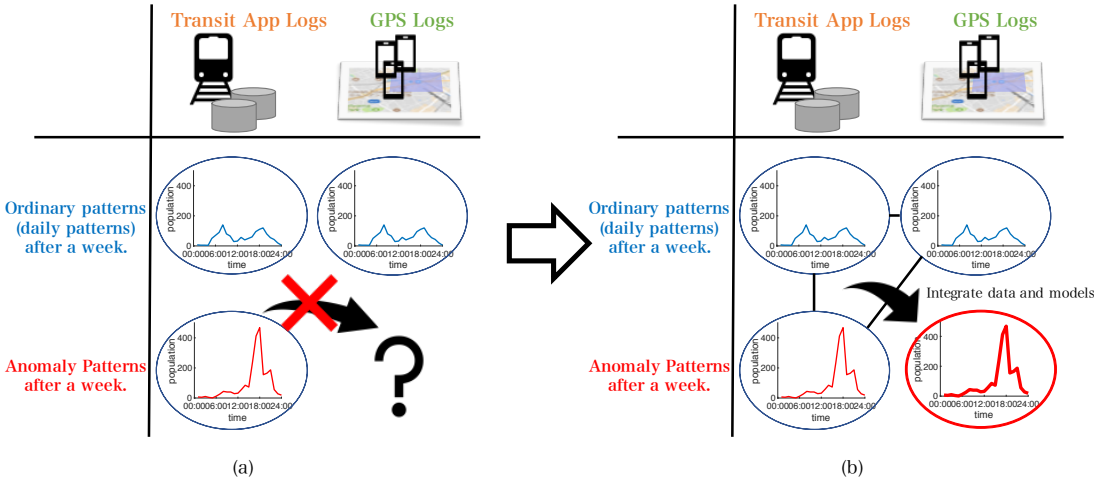

図 1 都市動態予測における既存手法，提案手法の位置付けと課題．(a) 既存研究における都市動態予測手法の位置付けと課題．乗換検索履歴のみを用いた異常混雑予測手法では，乗換検索履歴がもつ活動人口の表現力の限界，鉄道を用いない人々の移動により発生する混雑などを原因として，実際の活動人口に対する予測が不安定である．(b) 本研究における異常混雑予測へのアプローチとその位置付け．GPS 位置履歴と乗換検索履歴，既存の予測手法を統合的に利用することにより，高精度かつ安定した異常混雑予測を実現する．

歴を元にした活動人口で予測性能を評価した場合，予測は不安定であり，混雑状況の詳細なモデリングは困難である．

一週間後の異常混雑状況を安定して予測可能にするためには，これらの問題に対処する必要がある．そこで我々は，GPS 位置履歴を元にした活動人口の突発的な増加を捉えるために，乗換検索履歴をデータソースに用いる．また乗換検索履歴から活動人口への割り戻しを学習可能な規範に落とし込むために，活動人口の比で定義される混雑度での回帰を行う．しかし，混雑の原因になるようなイベントの希少性を原因として，異常混雑を反映するようなデータ点の量は限られてしまい，安直に回帰を適用するだけではモデルは頑健性に欠ける．混雑時に対しても安定した学習・予測を実現するために，このデータセットの不均衡を是正する必要がある．そこで我々は，活動人口の混雑度の回帰に，Synthetic Minority Oversampling を組み合わせた予測手法：CityGuardian を提案する．

本研究の貢献は以下である．

- 都市における異常混雑早期予測のために，GPS 位置履歴と乗換検索履歴を用いた学習の枠組みを提案する．また混雑度の回帰に加えてデータ点の Synthetic Minority Oversampling を用いることで，異常混雑を反映するデータの希少性によるモデルの学習時の不安定性を解消している．
- 実際の携帯電話から得られた GPS 位置履歴と乗換検索履歴を用いた実験により，既存の異常混雑の予測手法では，混雑度や活動人口の予測が不安定であることを示す．
- 同様の実験により提案手法が，既存の予測手法と比較して混雑度の予測において精度改善が見られたことを示す．また，実際に異常混雑が発生したイベントの日時や Point of Interest(POI) において，活動人口のモデリング精度が改善されたことを示す．

以降，同章の後半にて関連研究について述べ，2 章では異常混雑予測の問題設定，3 章では提案手法について述べる．4 章では実際の携帯電話の位置履歴と乗換検索アプリの検索履歴を用いた性能評価実験を行い，5 章で本研究の貢献と将来課題について述べる．

関連研究

ここでは都市活動人口予測のためのモデリングと，モデリングを用いた異常検知・異常予測に焦点を当てた研究について述べる．

都市における定常的な活動人口の予測に焦点を当てた手法に双線型ポアソン回帰 [5] が存在する．この手法は，識別的なアプローチにより，天気や曜日，時刻といった説明変数を用いて活動人口の予測を行う．しかしこの手法では，天気や曜日等のコンテクストに対して条件づけられないような，イベントに起因する突発的な混雑を予測することができない．

都市における異常検知に焦点を当てた研究として Witayangkurn らの手法 [9]，Zhang らの手法 [11] がある．Zhang らの手法 [11] は，異なる地域における都市動態の類似度を計算し，この類似度の時間変化率の変化をもとに異常検知を行う．この研究では，City-Scale な異常検知ではなく，より細かい地域での異常検知に焦点を当てている．地域間の類似度と同時に時間的な情報も取り込むフレームワークを提案している．Witayangkurn らの手法 [9] は，都市動態の時空間的な状態推移を隠れマルコフモデルを用いてモデリングし，グリッドに分割された地域における異

常検知を行うシステムを提案している．しかしこれらの手法は，観測データから普段とは異なるパターンを検出することに焦点を当てた研究であり，将来の異常予測を行うには，この普段とは異なるパターンを事前に予測する必要がある．

都市における異常予測に焦点を当てた研究として，Fan らの手法 [2]，Konishi らの手法 [3] が存在する．Fan らの手法 [2] では，大規模なイベントが起こる直前の人口流の変化，すなわち混雑の「きっかけ」になるようなパターンを検知することで，1 時間後の異常混雑予測を行う手法を提案している．Konishi らの手法 [3] では，より将来の異常混雑を予測可能にするために，鉄道の乗換検索アプリの検索クエリが，活動人口パターンと同時にユーザのスケジュール情報を反映していると仮定し，このデータソースを用いた手法を提案している．しかし，Fan らの手法 [2] は予測可能な混雑に時間的制約が存在する．また Konishi らの手法 [3] は時間的制約を解消しているものの，実際の活動人口に対する予測は不安定であり，混雑状況の詳細な予測には至っていない．

2. 異常混雑予測の問題設定と既存手法におけるアプローチ

我々は，ある領域の 1 日の活動人口の推移と，その混雑度をモデル化する問題を扱う．1 日を T の時間帯に分割し，この時間帯を $t = 1, 2, ..., T$ で表す．地域 l, 日付 d, 時刻 t における活動人口が $y_{d,t}^{(l)}$ と観測されたとする．このとき，同地域，同日時における定常的な活動人口（日常的に観測される活動人口）を $\hat{y}_{d,t}^{(l)}$ とする．活動人口における「混雑度」は人口の比を用いて $\hat{\nu}_{d,t}^{(l)} = \frac{|\hat{y}_{d,t}^{(l)} - y_{d,t}^{(l)}|}{y_{d,t}^{(l)}}$ と定義される．我々の目標は，活動人口 $y_{d,t}^{(l)}$ とその混雑度 $\nu_{d,t}^{(l)}$ をモデル化することである．

2.1 双線型ポアソン回帰による活動人口予測

活動人口予測問題の先行研究として，双線型ポアソン回帰 [5] がある．このモデルでは，活動人口を時間帯 $t = 1, ..., T$ や地域 l，天気や曜日・平日か休日かといったコンテクスト $\mathbf{c}$ を説明変数として，日にち d，時間帯 t における活動人口 $y_{d,t}^{(\mathbf{c},l)}$ を予測する．

このモデルでは，活動人口 $y_{d,t}^{(\mathbf{c},l)}$ がポアソン分布に従うと仮定している．

$$y_{d,t}^{(\mathbf{c},l)} \sim \mathrm{Pois}(\cdot|\lambda_{\mathbf{c},t}^{(l)}). \tag{1}$$

コンテクスト $\mathbf{c}$ と時刻 t に対して，ポアソン分布のパラメータは以下のように表される．

$$\ln \lambda_{\mathbf{c},t}^{(l)} = \varphi(\mathbf{c})^\top W^{(l)} \phi(t), \tag{2}$$

このとき，$\phi(t) \in \mathbb{R}^T$ は時間的特徴量を表し，以下のように定式化される．

$$\phi(t) = \{t_s | t_s = \mathcal{N}(s|\tau, \sigma^2), s = 1, ..., T\}, \tag{3}$$

ここで，$\mathcal{N}(\cdot)$ は正規分布を表し，τ はその平均，σ^2 がその分散である．

また，$\varphi(\mathbf{c}) \in \mathbb{R}^C$ は天気や曜日といった，時間以外のコンテクストを 1-of-K 表現などにより表現したものである．

$W^{(l)} \in \mathbb{R}^{C \times T}$ はこの双線型モデルのパラメータ行列を表し，先行研究 [5] ではパラメータ数を減らし過学習を抑制することを目的として，パラメータ行列 $W^{(l)}$ を $W^{(l)} = U^{(l)} V^{(l)\top}$ なる 2 行列 $U^{(l)} \in \mathbb{R}^{C \times H}$ と $V^{(l)} \in \mathbb{R}^{T \times H}$ の積に分解することで，低ランク化を行う．ここで，$H \ll C, H \ll T$ である．

この手法は，式 2 のように時間的特徴量とそれ以外のコンテクストの組み合わせを考慮し，活動人口のモデリングを行なっている．これにより，平日や休日といった条件の差による活動人口の推移の差をうまく表現することができる．しかし一方で，イベントの開催により発生する突発的な活動人口の増加などは，天気や曜日，時間帯といった特徴量のみでは捉えることができない．

2.2 CityProphet による都市の異常混雑予測

活動人口の異常予測の先行研究として，CityProphet [3] がある．この手法では，乗換検索履歴が，活動人口とユーザのスケジュール情報の両側面を表すとして，この乗換検索履歴を用いて異常混雑予測を行う．

q を，設定日時 d_a, t_a，検索日 d_s，目的地 l_s を持つ乗換検索クエリとし，以下で表す．

$$q = (d_a, t_a, l_s, d_s). \tag{4}$$

このとき，設定日時とは，検索者が目的地 l_s を訪問する際の予定として設定するものであり，一般に検索日と同じかそれ以後であると仮定しても差し支えない．ここで，検索クエリ全体の集合を $Q = \{q\}$ とし，検索日時が d, t，検索日が $d - i$，目的地が l である検索クエリの集合 $Q_{d,t|d-i}^{l}$ は以下で表される．

$$Q_{d,t|d'}^{(l)} = \{q \in Q | d_a = d \wedge t_a = t \wedge d_s = d - i \wedge l_a = l\}. \tag{5}$$

乗換検索量 $x_{d,t|d-i}^{l}$ は，設定日時 d, t，検索日時 $d - i$，目的地 l をもつ乗換検索クエリ数，すなわち，ユーザが目的地 l に訪問する予定日 d の i 日前に検索されたクエリ数で定義される．

$$x_{d,t|d-i}^{(l)} = |Q_{d,t|d-i}^{(l)}|. \tag{6}$$

CityProphet では，2 種類の予測モデルによる予測値：Schedule-based population（SP）と Descriptor-based Population（DP）の差分に着目することで，異常混雑予測を行う．

SP では，設定日 d に対して $i = p_d, p_d + 1..., p_d + p_w$ 日前に検索されたクエリ数を用いて，当日に検索されたクエリ数を，自己回帰を用いて予測する．ここで p_w は，使用する検索クエリの設定日の期間を表す．

$$\hat{x}^{(l,\mathrm{SP})}_{d,t|d} = \sum_{i=p_d}^{p_d+p_w} \sum_{j=-q}^{q} \theta_{i,j} x^{(l)}_{d,t-j|d-i}, \tag{7}$$

ここで，$\theta_{i,j}$ は最適化の対象となるパラメータを表す．また，q は予測の対象となる時間帯における乗換検索クエリ数に影響を与える時間幅である．すなわち，設定時刻が属する時間帯 τ に対し，$\forall \tau, t-q \leq \tau \leq t+q$ である $d-i$ 日前の乗換検索量 $x^{l}_{d,\tau|d-i}$ が，時間帯 t の当日の乗換検索クエリ数 $x^{(l)}_{d,t|d}$ に影響を与えるということを表現する．

DP では，当日に検索されたクエリ数を，双線型ポアソン回帰を用いてモデリングする．

$$\hat{x}^{(l,\mathrm{DP})}_{d,t|d} = \mathbb{E}[\mathrm{Pois}(x^{l}_{d,t|d}|\pi^{l}_{\mathbf{c},t})] = \pi^{l}_{\mathbf{c},t}, \tag{8}$$

双線型ポアソン回帰における活動人口の予測と同様に，$\mathbf{c}$ は，天気や曜日等の時間帯以外のコンテクストを表す．

上記の二種類のモデル；SP model と DP model の予測値 $\hat{x}^{(l,\mathrm{SP})}_{d,t|d}$，$\hat{x}^{(l,\mathrm{DP})}_{d,t|d}$ を用いて，混雑度 $\hat{\mu}^{(l)}_{d,t}$ を算出する．

$$\hat{\mu}^{(l)}_{d,t} = \frac{|\hat{x}^{(l,\mathrm{SP})}_{d,t|d} - \hat{x}^{(l,\mathrm{DP})}_{d,t|d}|}{\hat{x}^{(l,\mathrm{DP})}_{d,t|d}}. \tag{9}$$

この手法は，ユーザのスケジュール情報を表したデータソースとして，鉄道の乗換検索履歴を用いている．また，2 種類の異なるモデルを用いて乗換検索量をモデリングし，その差分に着目することで，混雑度の定量的な指標を与えることができる．しかし実際に GPS 位置履歴に基づいた訪問者数で混雑度を評価した際に，CityProphet は混雑度の予測が不安定である．

3. 提案手法：CityGuardian による異常混雑事前予測

詳細な異常混雑事前予測実現のために，GPS 位置履歴と乗換検索履歴，既存の予測手法を統合的に利用する CityGuardian を提案する．

3.1 活動人口，乗換検索クエリ数の定常的な推移の予測モデル

天気や曜日といった日常的なコンテクストに依存しない異常混雑を定義しこれを予測するには，定常的な活動人口・乗換検索量予測を行う必要がある．先行研究である双線型ポアソン回帰 [5] を用いることにより，曜日や天気といったコンテクストと時間的特徴量を説明変数として，定常的な活動人口・乗換検索クエリ数のモデリングを行う．

$$\hat{x}^{(l,\mathrm{DP})}_{d,t|d} = \mathbb{E}[\mathrm{Pois}(x^{l}_{d,t|d}|\pi^{l}_{\mathbf{c},t})] = \pi^{l}_{\mathbf{c},t}, \tag{10}$$

$$\hat{y}^{(l)}_{d,t} = \mathbb{E}[\mathrm{Pois}(y^{(l)}_{d,t}|\lambda^{(l)}_{\mathbf{c},t})] = \lambda^{(l)}_{\mathbf{c},t}. \tag{11}$$

3.2 活動人口比回帰を用いた混雑度の予測モデル

3.1 章で求められた活動人口・乗換検索クエリ数の定常的な推移を予測するモデルを用いて，人口比で定義された混雑度の回帰を行う．

乗換検索クエリ数における異常度 $\hat{\nu}$ は以下で定義される．

$$\hat{\nu}^{(l)}_{d,t} = \frac{\hat{x}^{(l,\mathrm{SP})}_{d,t|d} - \hat{x}^{(l,\mathrm{DP})}_{d,t|d}}{\hat{x}^{(l,\mathrm{DP})}_{d,t|d}}, \tag{12}$$

ここで，$\hat{x}^{(l,\mathrm{SP})}_{d,t|d}$ は CityProphet [3] と同様に，i 日前に検索された乗換検索クエリ数（$i = p_d, p_d + 1, ..., p_d + p_w$）からの自己回帰で表現する．

$$\hat{x}^{(l,\mathrm{SP})}_{d,t|d} = \sum_{i=p_d}^{p_d+p_w} \sum_{j=-q}^{q} \theta^{l}_{i,j} x_{d,t-j|d-i}. \tag{13}$$

また，活動人口 $y^{(l)}_{d,t}$ における混雑度は以下で定義される．

$$\nu^{(l)}_{d,t} = \frac{y^{(l)}_{d,t} - \hat{y}^{(l)}_{d,t}}{\hat{y}^{(l)}_{d,t}}. \tag{14}$$

以上を用いて活動人口の比で定義された混雑度の回帰を行う．モデルの最適化の対象となる目的関数 $L(\Theta^{(l)})$ は以下で与えられる．

$$\begin{aligned} L(\Theta^{(l)}) &= \sum_d \sum_t (\hat{\nu}^{(l)}_{d,t} - \nu^{(l)}_{d,t})^2 + \gamma ||\Theta^{(l)}||_2^2 \\ &= \sum_d \sum_t \left(\frac{\sum_i \sum_j \theta^{(l)}_{i,j} x^{(l)}_{d,t-j|d-i}}{\hat{x}^{(l,\mathrm{DP})}_{d,t|d}} - \frac{y^{(l)}_{d,t}}{\hat{y}^{(l)}_{d,t}} \right)^2 + \gamma ||\Theta^{(l)}||_2^2. \end{aligned}$$

ここで，$\Theta^{(l)}$ は i, j 成分に $\theta_{i,j}$ を持つパラメータ行列であり，$||\cdot||_2^2$ はフロベニウスノルム，γ は正則化の強さを制御するハイパーパラメータである．

3.3 合成的オーバーサンプリングを用いたデータ不均衡の是正

3.2 章にて，活動人口の混雑度を 2 種類のデータソースを用いた回帰によりモデリングする定式化を行なった．し

かし，イベントやコンサートの希少性ゆえ，異常混雑を反映するようなデータ点が極めて少ないため，単純な回帰を行うだけではモデルは混雑度の低いパターンへ強く最適化を行ってしまう．

この問題に対処するために，混雑度の高いデータ点に対するオーバーサンプリングを施すことを考える．一般的に，異常なパターンを反映するデータ点の単純な複製は過学習を誘発しやすいことが知られている．過学習を起こしにくいオーバーサンプリング手法として，異常なパターンのデータ点間の線分を考え，その線分上の任意の点を人工的な異常データとする SMOTE [1] と呼ばれる合成的オーバーサンプリング手法が提案されている．

本研究では，回帰用の SMOTE である，SMOTER [8] を用いてオーバーサンプリングを行う．SMOTER [8] は，データの空間 $\mathcal{D}$ に対し，事前に定義されたデータ点の重要度（relevance）$\psi : \mathcal{D} \to [0,1]$ を用いて，異常なパターンを反映するデータ点を定量的に選別し，これに対し合成的オーバーサンプリングを施す．υ, ρ^2 を，混雑度のデータ集合 $\mathcal{D} = \{\nu_{d,t}^{(l)}\}$ の標本平均，標本分散とする．データの重要度を決定する関数 ψ をシグモイド関数を用いた以下で定義する．

$$\xi(\nu_{d,t}^{(l)}) = \mathcal{N}(\nu_{d,t}^{(l)} | \upsilon, \rho^2), \tag{15}$$

$$\psi(\nu_{d,t}^{(l)}) = \sigma\left(-\frac{\xi(\nu_{d,t}^{(l)})}{1-\xi(\nu_{d,t}^{(l)})}\right), \tag{16}$$

この重要度を，サンプリングする範囲を制御する閾値 $\bar{\psi}$ を用いることにより，データ点のサンプリングの指標を定量的に与えることができる．この重要度を用いた SMOTER のアルゴリズムを，Algorithm 1，2 に示す．

Algorithm 1 SMOTER Algorithm

Input: $\mathcal{D}, \bar{\psi}, \%u, \%o, k$
// $\mathcal{D}$ - A dataset with continuous target variable Y
// $\bar{\psi}$ - threshold for relevance on Y values
// $\%u$ - percentage of under-sampling
// $\%o$ - percentage of over-sampling
// k - number of nearest neighbours
$\tilde{y} \leftarrow$ median of Y
$rareL \leftarrow \{(\mathbf{x}, y) \in \mathcal{D} : \psi(y) > \bar{\psi} \wedge y < \tilde{y}\}$
$newCasesL \leftarrow \text{genSynthCases}(rareL, \%o, k)$
$rareH \leftarrow \{(\mathbf{x}, y) \in \mathcal{D} : \psi(y) > \bar{\psi} \wedge y > \tilde{y}\}$
$newCasesH \leftarrow \text{genSynthCases}(rareH, \%o, k)$
$newCases \leftarrow newCasesL \cup newCasesH$
$nrNorm \leftarrow \%uof|newCases|$
$normCases \leftarrow$ sample of $nrNorm$ *cases* $\in \mathcal{D} \backslash \{rareL \cup rareH\}$
$new\mathcal{D} \leftarrow newCases \cup normCases$
Output: $new\mathcal{D}$ - A new modified dataset

Algorithm 2 Generating synthetic cases (**function** genSynthCases($\mathcal{D}, \%o, k$))

Input: $\mathcal{D}, \%o, k$
$newCases \leftarrow \{\}$
$ng \leftarrow \%o/100$
for all $case \leftarrow \mathcal{D}$ **do**
  $nns \leftarrow \text{KNN}(k, case, \mathcal{D} \backslash \{case\})$ // k-Nearest Neighbours of case
  for $i \leftarrow 1$ to ng **do**
    $x \leftarrow$ randomly choose one of the nns
    for all $a \in$ attributes of x **do**
      $diff \leftarrow case[a] - x[a]$ // Generate attribute values
      $new[a] \leftarrow case[a] + \text{RANDOM}(0,1) \times diff$
    end for
    $d_1 \leftarrow \text{DIST}(new[a], case[a])$ // Decide the target value
    $d_2 \leftarrow \text{DIST}(new[a], x[a])$
    $new[Target] \leftarrow \frac{d_2 \times case[Target] + d_1 \times x[target]}{d_1 + d_2}$
    $newCases \leftarrow newCases \cup \{new\}$
  end for
end for
Output: $newCases$

4. 性能評価実験

都市の異常混雑早期予測における提案手法の性能を評価するため，携帯端末の位置履歴データ，乗換検索アプリの検索履歴を用いて予測精度の評価実験を行なった．本章では，その性能評価実験の目的やデータセット，評価指標，実験結果等に関して述べる．

4.1 実験データ

本実験では，携帯電話位置履歴データ，乗換検索アプリの検索履歴データをデータセットに用いる．携帯電話位置履歴データには，Yahoo!JAPAN の防災アプリにより，ユーザの同意のもとで取得された携帯端末の位置履歴データ，乗換検索履歴データには，同じく Yahoo!JAPAN の乗換検索エンジンの検索ログ履歴データセットを用いる．どちらもデータの取得はユーザの同意のもとで取得されたデータであり，ユーザ情報をのぞいた，個人を特定することはできない匿名化されたデータである．取得期間は 2019 年 10 月 31 日〜2019 年 12 月 31 日の 3 ヶ月間である．

表 1 に，乗換検索履歴データセットの例を示す．データセットを構成する各データは，表 1 に示すような (検索日時，設定日時，検索条件，出発地，到着地) の 5 要素からなり，携帯端末用アプリは 2019 年 6 月時点で累計 4000 万ダウンロードを超えている．そのため，都市の人口流を織りなす多くのユーザのスケジュール情報を，乗換検索履歴により表現できると言える．到着や出発地は駅に限らず，施設名や一般の地域名も取りうる．検索条件は，（出発，到着，始発電車，最終電車，設定なし）のうちのいずれかをとる．我々はこのうち，検索条件が「到着」に設定された，到着時刻が明示的に指定され検索されたクエリのみを

表 1　乗換検索履歴データの例

検索日時	設定日時	検索条件	出発地	到着地
2019/12/11 08:24:56	2019/12/12 10:20:00	到着	渋谷	新宿
2019/12/11 19:53:13	2019/12/13 15:40:00	出発	大宮	東京
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

用いる．

本実験では，東京周辺のうち，鉄道利用者が多い 2 地点（渋谷，品川），10 月 31 日の渋谷ハロウィンの際の周辺地域として表参道，明治神宮前，10 月 19 日〜11 月 2 日に行われたラグビーワールドカップ日本大会決勝トーナメントの会場として，味の素スタジアム，日産スタジアム周辺を POI（Point of Interest）と定義し，各 POI の鉄道駅周辺や施設周辺を対象領域とした．図 2 に，POI「渋谷」「飛田給」「味の素スタジアム」における位置履歴データの集計範囲を示す．この対象領域内で観測された位置履歴データの数を対象 POI での活動人口とする．表 2 は，各 POI と鉄道駅との対応関係を表したものである．

図 2　POI「味の素スタジアム」における位置履歴データの集計範囲

POI	鉄道駅
渋谷	渋谷
品川	品川
表参道	表参道
明治神宮前	明治神宮前
味の素スタジアム	飛田給

表 2　各 POI と鉄道駅の対応関係

本実験では，1 日を $T = 24$ の時間帯に分割し，0 時 00 分〜23 時 59 分を 1 日あたりのモデルとして表現する．また，異常度算出の際のゼロ割を防ぐため，定常的な活動人口・乗換検索人口が 0 と予測された場合は全て 1 にマスクして実験を行なった．

4.2　性能評価指標

本実験では，提案手法の活動人口とその混雑度の予測に対する性能を評価する．そこで比較手法としては，混雑度の予測に関して平均絶対誤差 (MAE: Mean Abusolute Error)，活動人口の予測に関して平均絶対誤差，混雑度 $\bar{\nu}$ 条件付き平均絶対誤差 ($\bar{\nu}$-MAE) を用いる．これらの評価指標は，従来の都市動態の関連研究で用いられているものである．

混雑度の予測値と観測値をそれぞれ $\hat{\nu}^{(l)}_{d,t}, \nu^{(l)}_{d,t}$，その閾値を $\bar{\nu}$，$d = 1...D, t = 1...T$ と観測されたとする．このとき，混雑度に関する MAE は以下で定義される．

$$\mathrm{MAE} = \frac{1}{DT}\sum_{d=1}^{D}\sum_{t=1}^{T}|\hat{\nu}^{(l)}_{d,t} - \nu^{(l)}_{d,t}|. \tag{17}$$

また，活動人口とその予測値を $y^{(l)}_{d,t}, \hat{y}^{(l)}_{d,t}$，異常度とその予測値を $\hat{\nu}^{(l)}_{d,t}, \nu^{(l)}_{d,t}$，また $d = 1...D, t = 1...T$ とする．このとき，活動人口に関する評価指標は以下で定義される．

$$\mathrm{MAE} = \frac{1}{DT}\sum_{d=1}^{D}\sum_{t=1}^{T}|y^{(l)}_{d,t} - \hat{y}^{(l)}_{d,t}|, \tag{18}$$

$$\bar{\nu}\mathrm{-MAE} = \frac{1}{DT}\sum_{d=1}^{D}\sum_{t=1}^{T}|y^{(l)}_{d,t} - \hat{y}^{(l)}_{d,t}|$$

$$\text{where} \quad \hat{\nu}^{(l)}_{d,t} \geq \bar{\nu}, \tag{19}$$

$$\tag{20}$$

4.3　比較手法

本研究では，実験 1 で混雑度の予測に関する比較を既存手法である CityProphet [3]，実験 2 で活動人口の予測に関する比較を CityProphet，双線型ポアソン回帰 [3] と比較する．CityProphet による活動人口の予測値には，混雑度の定義に従い，混雑度 $\hat{\mu}^{(l)}_{d,t}$ と，双線型ポアソン回帰による活動人口の予測値 $\hat{y}^{(l)}_{d,t}$ を用いて，$(\hat{\mu}^{(l)}_{d,t} + 1)\hat{y}^{(l)}_{d,t}$ で表される量を用いた．

4.4　実験 1．混雑度の予測に関する性能比較実験

ここでは，混雑度に関する既存の予測手法との比較に関して述べる．ここでの比較手法は CityProphet であり，評価指標には MAE を用いる．この比較では，提案手法である CityGuardian が，既存手法と比較して安定した異常度

の予測を実現できていることを確認する．いずれのモデルも 5 交差検証を行い，学習データとして 30 日分，評価データとして 20 日分のデータを用いた．

実験の結果を表 3 に示す．いずれの POI においても，混雑度の予測において，既存手法よりも高精度かつ安定した予測が行えていることがわかり，平均で 89.8%の精度改善を実現した．これにより提案手法が，既存研究における混雑度の過剰評価や過小評価などを抑えられていることがわかる．

表 3　各 POI における異常度の MAE での評価．

POI	CityGuardian	CityProphet
渋谷	**0.70 ± 0.026**	32 ± 19
品川	**0.67 ± 0.022**	27 ± 15
明治神宮前	**0.78 ± 0.057**	2.9 ± 0.49
表参道	**0.98 ± 0.15**	6.6 ± 0.44
飛田給	**0.59 ± 0.072**	9.0 ± 5.1
味の素スタジアム	**0.66 ± 0.10**	9.1 ± 5.1

4.5　実験 2．活動人口の予測に関する性能比較実験

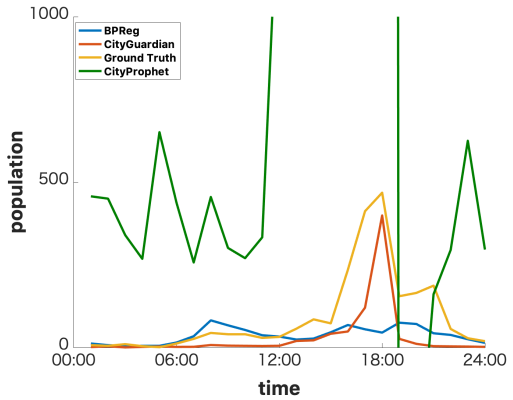

図 3　POI「味の素スタジアム」における 2019 年 11 月 1 日の活動人口とその予測値の推移．BPReg は双線型ポアソン回帰，Ground Truth は GPS 位置履歴による活動人口の真値を表す．

次に，活動人口に関する既存の予測手法との比較に関して述べる．評価の対象は混雑度予測に関する実験の 6 つの POI のうち，データの取得期間内にて異常混雑が観測された POI「味の素スタジアム」とする．比較手法は CityProphet，双線型ポアソン回帰とし，評価指標には MAE と混雑度により条件づけられた $\bar{\nu}$-MAE を用いる．実験 1 と同様にいずれのモデルも 5 交差検証を行い，学習データとして 30 日分，評価データとして 20 日分のデータを用いた．

実験結果を表 4 に示す．まず MAE，$\bar{\nu} = 0.5, 1, 2$ と条件づけた際の $\bar{\nu}$-MAE について，CityProphet の予測精度が既存手法である双線型ポアソン回帰と比較して大きく劣化していることがわかる．また，提案手法が既存手法である双線型ポアソン回帰と同程度の予測性能を持っていることがわかる．この結果から，提案手法においても，定常的な活動人口の予測を既存手法と比較して遜色なく，かつ安定して行うことができる．また，$\bar{\nu} = 3, 4, 5$ と条件づけた際の $\bar{\nu}$-MAE について，双線型ポアソン回帰・CityProphet の両既存手法と比較して，最も良い予測性能となっている．これにより，異常混雑の事前の予測においても提案手法の優位性が示される．

図 3 は，評価対象である POI「味の素スタジアム」において，2019 年 11 月 1 日における活動人口とその予測値の推移を可視化したものである．同 POI では 2019 年 11 月 1 日，ラグビー W 杯の 3 位決定戦が開催され，観戦のために多くの人々が集まった．スタジアムの開場は 15 時，試合開始は 18 時である．活動人口の真値は 15 時付近から徐々に増加し始め，18 時付近にて増加のピークを迎えている．双線型ポアソン回帰による予測値は，この活動人口の突発的な増加に対応できていない．また CityProphet による予測値は，12 時ごろから 18 時ごろのまで間で，突発的な活動人口増加の発生を捉えられているものの，詳細な予測にまでは至っていない．これに対して提案手法では，18 時付近の異常混雑を，混雑時の人数も含めより正確に捉えられていることがわかる．しかし，提案手法は 18 時以降の活動人口の予測に対して過小評価してしまっている．これは，スタジアムに向かう交通手段として鉄道を利用した人々の多くが，乗換検索の到着時刻を試合開始の 18 時に設定して検索していたことにより，それ以降の時刻に設定された検索量が相対的に少なかったことに起因すると考えられる．

5.　結論

本研究では，GPS 位置履歴情報と乗換検索履歴を用いた，都市における活動人口とその混雑度のモデリングを行なった．既存研究における都市動態のモデリング手法では，都市の混雑予測において，混雑状況の正確な把握と，どれくらい早期の予測が可能かという二つの側面にて課題が存在した．そこで，GPS 位置履歴とそれを用いた既存の活動人口予測手法に加えて，乗換検索によるユーザのスケジュール情報を用い，活動人口の比で定義される混雑度での回帰と頑健性担保のためのオーバーサンプリングを組み合わせた手法を提案した．これにより，時間的制約に縛られない，かつ突発的な混雑にも対応可能な活動人口とその混雑度の予測が可能となる．

東京周辺の混雑が発生しやすい，あるいは混雑が発生したと考えられる対象 6POI において実際の携帯電話位置履歴データ，乗換検索履歴データを用いた実験を行い，既存の予測手法である CityProphet，双線型ポアソン回帰と比較し，混雑度の予測においては絶対予測誤差（MAE）にて平均 89.8%の精度改善を実現した．また，味の素スタジアムにおける 2019 年 11 月 1 日のラグビー W 杯 3 位決定戦

表 4 各 POI と対応する鉄道駅における MAE，$\bar{\nu}$-MAE による性能評価．$\bar{\nu} = 0$ は MAE を表す．また，BPReg は双線型ポアソン回帰を表す．

POI	$\bar{\nu}$	CityGuardian	BPReg	CityProphet
味の素スタジアム	0	$\mathbf{25.0 \pm 4.45}$	28.0 ± 1.96	166 ± 109
	0.5	$\mathbf{10.0 \pm 1.13}$	21.4 ± 2.72	69.9 ± 74.6
	1	2.04 ± 1.50	$\mathbf{2.02 \pm 1.68}$	11.8 ± 15.4
	2	$\mathbf{1.84 \pm 0.305}$	2.16 ± 0.781	4.15 ± 4.40
	3	$\mathbf{0.641 \pm 0.693}$	1.45 ± 0.665	2.33 ± 2.57
	4	$\mathbf{0.419 \pm 0.467}$	1.10 ± 0.739	1.42 ± 0.835
	5	$\mathbf{0.419 \pm 0.467}$	1.10 ± 0.739	1.04 ± 0.403

での活動人口の予測において，既存手法よりも混雑状況を正確に捉えることができることを示した．

将来課題として，現在の予測手法は鉄道の利用に関連づけられる地域にのみ対応可能であるため，鉄道駅周辺に限らない地域の異常混雑事前予測が考えられる．実際，田舎の地域では鉄道駅の分布が疎になり，花火大会などのイベントにおいて必ずしも鉄道が利用されるとは限らない．鉄道の乗換検索履歴に加えてバスなど，停車駅がより密に配置された移動手段の乗換検索履歴を用いる，あるいはより一般的な検索履歴等を用いて活動人口動態の予測を行う手法が求められる．

参考文献

[1] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[2] Z. Fan, X. Song, R. Shibasaki, and R. Adachi. Citymomentum: An online approach for crowd behavior prediction at a citywide level. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp ’15, page 559–569, New York, NY, USA, 2015. Association for Computing Machinery.

[3] T. Konishi, M. Maruyama, K. Tsubouchi, and M. Shimosaka. Cityprophet: City-scale irregularity prediction using transit app logs. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp ’16, page 752–757, New York, NY, USA, 2016. Association for Computing Machinery.

[4] Y. Sekimoto, R. Shibasaki, H. Kanasugi, T. Usui, and Y. Shimazaki. Pflow: Reconstructing people flow recycling large-scale social survey data. *IEEE Pervasive Computing*, 10(4):27–35, April 2011.

[5] M. Shimosaka, K. Maeda, T. Tsukiji, and K. Tsubouchi. Forecasting urban dynamics with mobility logs by bilinear poisson regression. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp ’15, page 535–546, New York, NY, USA, 2015. Association for Computing Machinery.

[6] X. Song, Q. Zhang, Y. Sekimoto, T. Horanont, S. Ueyama, and R. Shibasaki. Modeling and probabilistic reasoning of population evacuation during large-scale disaster. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’13, page 1231–1239, New York, NY, USA, 2013. Association for Computing Machinery.

[7] L. Sun, D.-H. Lee, A. Erath, and X. Huang. Using smart card data to extract passenger’s spatio-temporal density and train’s trajectory of mrt system. In *Proceedings of the ACM SIGKDD international workshop on urban computing*, pages 142–148, 2012.

[8] L. Torgo, R. P. Ribeiro, B. Pfahringer, and P. Branco. Smote for regression. In *Portuguese conference on artificial intelligence*, pages 378–389. Springer, 2013.

[9] A. Witayangkurn, T. Horanont, Y. Sekimoto, and R. Shibasaki. Anomalous event detection on large-scale gps data from mobile phones using hidden markov model and cloud platform. In *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication*, UbiComp ’13 Adjunct, page 1219–1228, New York, NY, USA, 2013. Association for Computing Machinery.

[10] J. Yuan, Y. Zheng, and X. Xie. Discovering regions of different functions in a city using human mobility and pois. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’12, page 186–194, New York, NY, USA, 2012. Association for Computing Machinery.

[11] H. Zhang, Y. Zheng, and Y. Yu. Detecting urban anomalies using multiple spatio-temporal data sources. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(1), Mar. 2018.