PWS Cup 2019 データセットについて

PWS Cup 実行委員会

2019 年 8 月 16 日

1　はじめに

PWS Cup 2019 ではデータセットとして，公開データセットである疑似人流データ [1] を基に新たに作成した人工データ（以後，PWSCup2019 用人工データ）を使用する．本資料では PWSCup2019 用人工データの生成法の特徴を説明する． 尚，PWSCup2019 のルールについては，PWS Cup 2019 競技ルール [2] およびルール論文 [3] を参照されたい．

2　PWSCup2019 用人工データ

2.1　概要（ルール論文 [3] の第 3.1 節から抜粋）

疑似人流データは，実データを基に作成した 6 日間（2013 年の 7/1，7/7，10/7，10/13，12/16，12/22）にわたる東京近郊（首都圏）の人工的なトレースの公開データセットである．本コンテストでは，東京中心部（緯度：35.65〜35.75，経度：139.68〜139.8）に対して，図 1 のように均等に $32 \times 32 = 1024$ 個の領域に分割し，左下から右上にかけて領域 ID を割り当てる．各領域の大きさは，緯度 1 度あたり 111km，（東京での）経度 1 度あたり 91km として，縦 347m × 横 341m である．位置情報（領域）の数は $m = 1024$ である．

その後，東京中心部における位置情報が 10 個以上あるユーザ（計 10181 名）のトレースを抽出し，これらを学習データとして，マルコフモデルに基づく生成モデルを学習する．その生成モデルから，チーム番号 $i \in [z]$ およびデータセット番号 $j \in \{1, 2\}$ 毎に異なる $n = 2000$ 名のユーザ集合 $\mathcal{U}^{(i,j)} = \{u_1^{(i,j)}, \cdots, u_{2000}^{(i,j)}\}$ の参照トレース $R^{(i,j)} = (r_1^{(i,j)}, \cdots, r_{2000}^{(i,j)})$ と元トレース $O^{(i,j)} = (o_1^{(i,j)}, \cdots, o_{2000}^{(i,j)})$ を生成する．尚，生成モデルは各ユーザ $u_k^{(i,j)}$（$1 \leq k \leq 2000$）の特徴量を保持しており，それを基に参照トレース $r_k^{(i,j)}$ と元トレース $o_k^{(i,j)}$ を生成する．従って，$r_k^{(i,j)}$ と $o_k^{(i,j)}$ は高い相関を持っており，参照トレースを参考にしながら公開加工トレースに対して ID 識別やトレース推定を行うことが可能である．

時刻については，8 時から 17 時 59 分までを 30 分おきに区切って離散化する．予備戦では，1 日目と 2 日目のトレースを参照トレースに，3 日目と 4 日目のトレースを元トレースとして用いる．即ち，各トレースの長さは $t = 40$ である（時刻 1 は 1 日目の 8 時，時刻 40 は 2 日目の 17 時 30 分，時刻 41 は 3 日目の 8 時，時刻 80 は 4 日目の 17 時 30 分）．但し，本戦では参照トレースと元トレースの日数を（2 日分から）変更する可能性がある．

PWSCup2019 用人工データの概要を表 1 に示す．

表 1　PWSCup2019 用人工データ

ユーザ数	$n = 2000$
対象エリア	緯度：35.65〜35.75，経度：139.68〜139.8
位置情報数	$m = 1024$（32×32 個の領域に分割）
トレースの長さ	予備戦では $t = 40$（8 : 00〜17 : 59 の 2 日分，30 分おき）．本戦では日数変更の可能性あり．

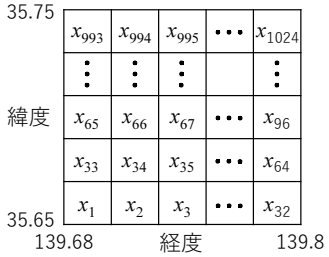

図 1　本コンテストにおける位置情報

尚，生成モデルの詳細は非公開とするが，PWSCup2019 用人工データの生成法は以下のような特徴を持っている．

1. **人口分布の保存**：6 時台から 17 時台までの 1 時間毎の人口分布（$m = 1024$ 個の領域にわたる確率分布）が，元の疑似人流データのそれと近くなるように，人工データを生成する．
2. **遷移行列の保存**：マルコフモデルの遷移行列（1024×1024 の行列）が，元の疑似人流データのそれと近くなるように，人工データを生成する．
3. **家のモデル化**：各ユーザは朝に高い確率で（6-7 時台は約 95%，8 時台は約 30% の確率で）自身の家の領域にいるように，人工データを生成する（家は，人口分布が保存されるようにしつつ，ユーザ毎にランダムに割り当てる）．但し，6-7 時台の位置情報まで参照・元トレースに含めると，攻撃者はほぼ全ユーザの家の領域を知ることになり，最大知識モデルのように仮定が強すぎるため，8 時以降の位置情報のみを用いる．

第 2.2 節において，これらの特徴を詳しく述べる．

2.2　PWSCup2019 用人工データの生成法の特徴

2.2.1　人口分布・遷移行列の保存

PWSCup2019 用人工データの生成法は，6 時台から 17 時台までの 1 時間毎の人口分布と，遷移行列を保存する．このことを定量的に示すために，以下のような評価実験を行った．

まず，疑似人流データにおける計 10181 名のユーザのトレース（第 2.1 節参照）から，6 時台から 17 時台までの 1 時間毎の人口分布と遷移行列を求めた．疑似人流データにおける，k（$6 \leq k \leq 17$）時台における人口分布を $\mathbf{p}_k$（1024 次元の確率分布）とし，その i 番目の要素（$1 \leq i \leq 1024$）を $\mathbf{p}_k[i]$ とする．また，遷移行列を $\mathbf{Q}$（1024×1024 の行列）とし，その (i, j) 番目の要素（$1 \leq i, j \leq 1024$）を $\mathbf{Q}[i, j]$ とする．$\mathbf{p}_k[i]$ は k 時台において領域 x_i における確率，$\mathbf{Q}[i, j]$ は領域 x_i から次の時刻において領域 x_j に遷移する確率を表す．$\mathbf{p}_k$

表 2　人口分布と遷移行列の MAE

	PWS Cup 2019	一様分布
人口分布の MAE	3.70×10^{-4}	9.22×10^{-4}
遷移行列の MAE	9.89×10^{-4}	1.87×10^{-3}

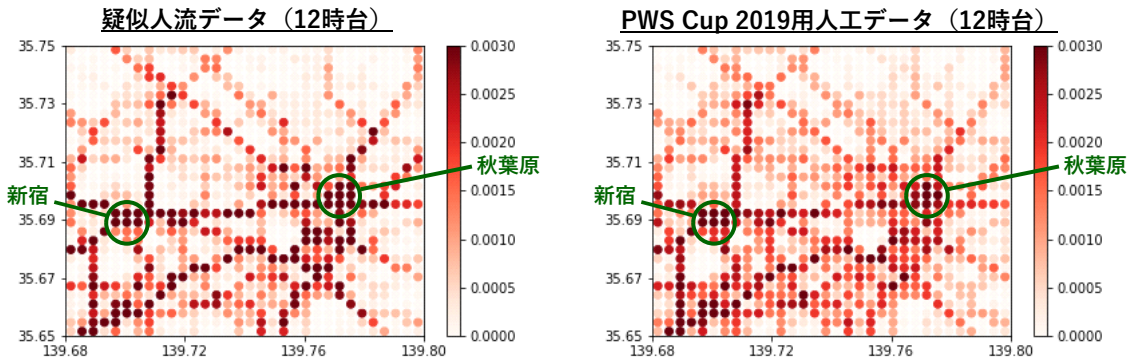

図 2　12 時台の人口分布 $\mathbf{p}_{12}$（左図）と $\tilde{\mathbf{p}}_{12}$（右図）

と $\mathbf{Q}$ は，それぞれ訪問回数と遷移回数を全ユーザに渡ってカウントし，確率値に正規化することで求めた．

次に，予備戦・本戦で使用予定の PWSCup2019 用人工データ（最大 20 チームが参加し得ると想定し，20 チーム分 × ID 識別対策用・トレース推定対策用の 2 つ × 予備戦・本選の 2 つ = 計 80 個のデータセット）に対して同様に，6 時台から 17 時台までの 1 時間毎の人口分布と遷移行列を求めた．PWSCup2019 用人工データにおける，k（$6 \leq k \leq 17$）時台における人口分布を $\tilde{\mathbf{p}}_k$（1024 次元の確率分布）とし，その i 番目の要素（$1 \leq i \leq 1024$）を $\tilde{\mathbf{p}}_k[i]$ とする．また，遷移行列を $\tilde{\mathbf{Q}}$（1024×1024 の行列）とし，その (i, j) 番目の要素（$1 \leq i, j \leq 1024$）を $\tilde{\mathbf{Q}}[i, j]$ とする．$\tilde{\mathbf{p}_k}$ と $\tilde{\mathbf{Q}}$ は，それぞれ訪問回数と遷移回数を PWSCup2019 用人工データ全体（計 80 個のデータセット）に渡ってカウントし，確率値に正規化することで求めたものである．

そして，人口分布 $\mathbf{p}_k$ と $\tilde{\mathbf{p}}_k$ （$6 \leq k \leq 17$）の誤差，および遷移行列 $\mathbf{Q}$ と $\tilde{\mathbf{Q}}$ の誤差を評価した．誤差の評価尺度としては MAE（Mean Absolute Error）を用いた．但し，人口分布については 6 時台から 17 時台までの 12 個の分布に対する MAE の平均を求めた．人口分布 $\mathbf{p}_k$ と $\tilde{\mathbf{p}}_k$ の MAE は，

$$\frac{\sum_{k=6}^{17} \sum_{i=1}^{1024} |\mathbf{p}_k[i] - \tilde{\mathbf{p}}_k[i]|}{12 \times 1024}$$

で表され，遷移行列 $\mathbf{Q}$ と $\tilde{\mathbf{Q}}$ の MAE は，

$$\frac{\sum_{i=1}^{1024} \sum_{j=1}^{1024} \left| \mathbf{Q}[i, j] - \tilde{\mathbf{Q}}[i, j] \right|}{1024^2}$$

で表される．また，比較のため，$\tilde{\mathbf{p}}_k$ と $\tilde{\mathbf{Q}}$ の各行を一様分布としたとき（即ち，各要素を 1/1024 としたとき）の MAE も求めた．

人口分布と遷移行列の MAE を表 2 に示す．人口分布，遷移行列ともに，一様分布より遥かに MAE が小さいことが分かる．また，12 時台の人口分布 $\mathbf{p}_{12}$，$\tilde{\mathbf{p}}_{12}$ を可視化したものを図 2 に示す．12 時台の人口分布 $\mathbf{p}_{12}$ は秋葉原，新宿周辺に集中しており，PWSCup2019 用人工データはこのような時間ごとの人口分布を大体保存している．尚，本実験では PWSCup2019 用人工データ全体（計 80 個のデータセット）における人口

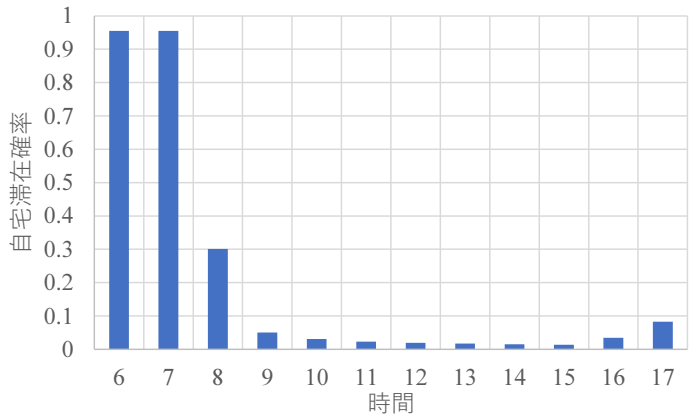

図 3　6 時台から 17 時台までの自宅滞在確率

分布，遷移行列を評価したが，1 つ 1 つのデータセットも同様に人口分布，遷移行列を保存していることを確認している．

2.2.2　家のモデル化

PWSCup2019 用人工データの生成法は，ユーザごとに家を割り当て，6 時台と 7 時台はほとんどの確率で自身の家の領域にいるように人工データを生成する．

6 時台から 17 時台までの各時間帯において，ユーザが自身の家の領域にいる確率（以後，自宅滞在確率）を図 3 に示す．この図は，PWSCup2019 用人工データ全体（計 80 個のデータセット）にわたって家の領域の滞在回数をカウントし，確率値に正規化することで求めたものである．自宅滞在確率は 6 時台と 7 時台で約 95%，8 時台で約 30% であることが分かる（但し，本コンテストでは 8 時以降の位置情報のみを用いる）．また，16 時台から自宅滞在確率が徐々に増加している（17 時台で約 8%）．

参考文献

[1] ナイトレイ，東京大学空間情報科学研究センター (CSIS)，疑似人流データ: https://nightley.jp/archives/1954/

[2] PWS Cup 2019 競技ルール：https://www.iwsec.org/pws/2019/cup19-rule.pdf

[3] 村上隆夫，荒井ひろみ，井口誠，小栗秀暢，菊池浩明，黒政敦史，中川裕志，中村優一，西山賢志郎，野島良，波多野卓磨，濱田浩気，山岡裕司，山口高康，山田明，渡辺知恵美，“PWS Cup 2019: ID 識別・トレース推定に強い位置情報の匿名加工技術を競う”，CSS2019.