全国版擬似人流データ

仕様書

Ver.1.2

令和 5 年 1 月
東京大学　空間情報科学研究センター

■ **改訂履歴**

バージョン	日付	改訂内容
Ver 1.0	2022/04/28	データ提供サービス開始 HP とデータ仕様書を公開
Ver 1.1	2022/10/08	軌跡データの ID 欠損の補正を行った 修正箇所：**集計結果・時間帯別メッシュ人口**データ
Ver 1.2	2023/01/28	データの品質を向上するために改訂を行った．主な改正点は以下のとおり． 1. 目的地選択モデルのパラメータの更新 修正箇所：**活動・トリップ・軌跡**データセット 2. 徒歩と自転車の移動軌跡の遠回りバグの修正 修正箇所：**軌跡**データセット 3. Unixtime 表示ズレの修正 修正箇所：**軌跡**データセット

目次

1. 概要 4

1.1 背景 4
1.2 特徴 4
1.3 データ概要 4
1.4 作成方法 5
1.4.1 人口データ 5
1.4.2 活動データ 5
1.4.3 トリップデータ 6
1.4.4 軌跡データ 7
1.5 提供方法 8

2. データ仕様 9

2.1 人口データ 9
2.2 活動データ 9
2.3 トリップデータ 9
2.4 軌跡データ 10

3. コード一覧 11

4. 利用上の注意 14

4.1 データの精度について 14
4.2 デジタル道路マップ（DIGITAL ROAD MAP, DRM）の扱い方について 14
4.3 データ生成上にパーソントリップ調査データの扱い方について 14

1. 概要

1.1 背景

近年，人間の移動に関連する広範な地理情報データセットの急増により，個人および集団レベルでの日常の移動パターンのメカニズムを解明する機会が提供されています．このような分析は，交通予測，病気の拡散，都市計画，および汚染などの社会問題を解決するために不可欠です．しかし，データ収集の対象となったユーザーのプライバシーに関する懸念から，そのようなデータの公開は制限されています．この課題に対処するために，我々は社会の基礎データの整備を目標として見据え，エージェントベースモデリングとシミュレーション手法を用いて，全国の人口に対して安定した精度を持つ「全国擬似人流データ」を作成し，提供しています．

1.2 特徴

- 擬似人流データの開発では，一般的に入手可能なオープンデータとして公開される統計データと既存のパーソントリップ調査データ（PT 調査）の集計結果，そして建物データ等の低廉に入手可能なデータのみを用いること．
- 上記の集計レベルの調査データを用いることで，合成的人口を作成し，そして人々の典型的な日常の行動を擬似的に再現し，リアルな人の位置情報ではないため，その結果は研究目的で公開可能．
- 位置情報だけでなく，全人口に対する一人一人の人口属性（世帯・年齢・性別・就業状態），1 日の生活活動，活動に伴う発生した移動トリップ（目的・時刻・起点・終点・交通手段），また移動軌跡をすべて提供する事．

1.3 データ概要

本データセットは，オープンな調査データと低廉な価格で入手可能な商業データを活用し，全国の典型的な平日の 1 日中の擬似人流を再現しています．この結果，断片的な位置情報だけでなく，どのような人々が，どのような目的で，いつ，どのような交通手段で，どこからどこへ移動するかといった情報を提供することができます．今回提供する擬似人流データは，以下の 4 種類のデータセットから構成されています．

- 人口データ：すべての人口に対し，一人一人の世帯構成・年齢・性別・就業状況・住所等の情報を表すデータ．
- 活動データ：個々の人間の典型的な 1 日に，「いつ，どこに，何をする」の情報を表すデータ．
- トリップデータ：人の活動に対し，「誰が，いつ，何の目的で，どこからどこへ移動する」の情報を表すデータ．
- 軌跡データ：トリップデータに基づき，数秒ごとに移動者の位置を記録し，GPS データと同じように扱う

1.4 作成方法

1.4.1 人口データ

擬似人口データの生成は，国勢調査から提供する集計レベルの人口統計データに基づき，個人の世帯構成，年齢，性別，就業状況，住所などの情報から推定されます．具体的には，統計データから世帯単位の人口分布を推定し，ゼンリンの建物データに世帯データを割り当てたものであり，建物ごとに世帯数，家族構成，年齢，性別などの属性情報を個人に配分しています[1]．

推定した世帯ごとの人口データに，個人の就業状況に対して，以下の 9 種類を設定する．

① 就業者
② 主夫・主婦と高齢者（65 歳以上の方）
③ 専門学校・大学の学生（18 歳以上の方）
④ 高校生（15 歳以上また 18 歳未満の方）
⑤ 中学生（12 歳以上また 15 歳未満の方）
⑥ 小学生（6 歳以上また 12 歳未満の方）
⑦ 幼稚園児（3 歳以上また 6 歳未満の方）
⑧ 幼児（3 歳未満の方）
⑨ 無職者

その内，①，②，③から④の割合については，国勢調査の就業状態等基本集計の結果より地域・性別・年齢別に決定する．④と⑤の高等教育の学生数について，学校基本調査の結果より都道府県，性別別に決定する．残りの⑥から⑧については，年齢属性より決定した．

1.4.2 活動データ

本研究では，在宅・勤務・通学・買物・食事・通院・その他合計 7 種類の活動を定義し，各時間帯に個々のエージェントは必ずこの 7 種類の中から 1 つの活動を選択する．

次に，各活動間の遷移確率をマルコフモデルによって推定する．

$$p_k(i,j) = P\{X_{K+1} = j | X_K = i\} \tag{1}$$

$p_k(i,j)$は活動 i から j へ遷移する確率．一方，個人の行動パターンは人口属性や地域性による異なるため，上記のマルコフモデルのパラメータについて，2.1 の 10 種類の就業状態に対して，大（50 万人以上）・中（30 万人以上 50 万人未満）・小（30 万人未満）の都市規模ごとにパーソントリップ調査から値を推定した．

---

[1] Kento Kajiwara, Jue Ma, Toshikazu Seto, Yoshihide Sekimoto, Yoshiki Ogawa, Hiroshi Omata, Development of current estimated household data and agent-based simulation of the future population distribution of households in Japan, Computers, Environment and Urban Systems, Volume 98, 2022, 101873, https://doi.org/10.1016/j.compenvurbsys.2022.101873.

1.4.3 トリップデータ

行動先は，2.1 で割り当てた就業状況と移動目的に応じて決定される．具体的な行動先の決定方法は，次の通りである．

【①就業者】

国勢調査の従業地・通学地集計より取得した市区単位の通勤 OD 量に従って，確率的に勤務先の市区を決定する．次に，経済センサスより取得した市区内のメッシュ就業者数に応じて確率的に勤務先メッシュを決定する．一方，対象者の従業値と自宅は同じ市区町村の場合，メッシュの就業者数と目的地までの距離を特徴量としてハフモデルを用いて目的地への移動確率を求める．ハフモデルの式を以下に示す．

$$P_{ij} = \frac{\frac{S_j}{{D_{ij}}^{\gamma}}}{\sum_{i=1}^{n} \frac{S_j}{{D_{ij}}^{\gamma}}} \tag{2}$$

式内のP_{ij}は i から j への移動確率であり，Sjは j の魅力度，D_{ij}は i から j への距離，γは距離抵抗のパラメータを示す．本研究では，距離抵抗のパラメータとして，一般的に使用される 2 の値を設定した．

最後に，ゼンリンの建物データを用いて，メッシュ内の建物の床面積をもとに，確率的に最終的な勤務先建物を決定する．

【③と④15歳以上の学生】

従業者と同様に，国勢調査の従業地・通学地集計より取得した市区単位の通勤 OD量に従って，確率的に通学先の市区を決定する．そして，通学先の市区内の学校のいずれかをランダムに選択する．

【⑤と⑥小中学】

国土数値情報の校区データをもとに，通学先の学校を決定する．ただし，本データには掲載されていない地区が多くあるが，本研究ではそれらを個別に収集する作業は行っていない．

【⑦幼稚園児】

幼稚園児をランダムな順番で，定員をオーバしないように，最寄りの施設を選択するようにした．なお，今回はすべての施設の定員を 300 とした．

【自由行動】

最後に，主夫（婦）・高齢者のすべての行動と就業者及び学生の自由行動の目的地選択に対して，ゾーン（パーソントリップ調査ゾーン）レベルロジットモデルを用いて各メッシュの選択確率を計算して行う．各ゾーンの効用は目的に応じて，該当商業分類の従業員数と事務所数によって算出する．経済センサスから得られる「宿泊業，飲食サービス業」「生活関連サービス業，娯楽」「教育，学習支援業」「医療，福祉」の事業所数を説明変数として事業所種別が与える吸引力への影響度を係数として求める．例えば，通院目的のトリップに対して，各ゾーンの医療機関事務所数と医療従業員数を使っ

てゾーンの効用を算出する．また，効用関数のパラメータは，大（50 万人以上）・中（30 万人以上 50 万人未満）・小（30 万人未満）の都市規模ごとにパーソントリップ調査から値を推定した．移動先のゾーンが決定した後は，就業者の場合と同様に，メッシュ内の建物の床面積をもとに，確率的に最終的な移動先の建物を決定する．

1.4.4 軌跡データ

第一プロトタイプの擬似人流の開発では，ダイクストラ法による最短経路探索により移動経路を算出した．交通手段が徒歩，自転車，自動車の場合には DRM の道路ネットワークデータを，鉄道については金杉ら[2]が開発した鉄道ネットワークデータを使用した．道路ネットワークのコストとしては，道路リンク毎に設定された制限速度とリンク長から算出した移動時間を，鉄道ネットワークについては鉄道のリンク長（路線の長さ）を設定して計算した．なお，作成した軌跡データは，数秒ごとに移動者の位置を記録し，GPS データと同じように扱うことは可能となる．

なお，擬似人流生成のために使用するデータを表 1 に示す．PT 調査ベースのデータは東京大学 CSIS が整備する研究用空間データ基盤 JoRAS からデータを入手することが可能である．パーソントリップ調査データの使用について，プライバシーの問題を十分に考慮する上で，元のデータではなく，調査データの集計結果を用いてエージェントモデルを構築する．ゼンリンが提供する建物データと日本デジタル道路地図協会が提供するデジタル道路地図（DRM）については，有償データではあるが，最低限必要な基盤空間データとして考えて使用した．なお，これらのデータも，研究目的であれば東京大学空間情報科学研究センター共同研究利用システム JoRAS(https://joras.csis.u-tokyo.ac.jp/)より入手可能となっている．

---

[2] 金杉洋，関本義秀，樫山武浩，人々の流動再現へ向けたオープンな鉄道インフラデータの構築，第 22 回地理情報システム学会講演論文集, 2013.

表 1　使用データ一覧

	使用データ
人口データ	• H27 国勢調査 • ゼンリン建物データ（Zmap TOWN Ⅱ）
活動データ	• H28 社会生活基本調査 • 2011 年中京都市圏 PT 調査 • 2015 年全国都市交通特性調査
トリップデータ	• H27 国勢調査 • H28 経済センサス • ゼンリン建物データ（Zmap TOWN Ⅱ） • 国土数値情報：学校・小中学校区 • 2011 年中京都市圏 PT 調査 • 2016 年東駿河湾都市圏 PT 調査
軌跡データ	• DRM 道路ネットワークデータ • 鉄道ネットワークデータ

1.5 提供方法

現在，研究目的の使用に限り，上記の擬似人流データを東京大学 CSIS が整備する研究用空間データ基盤 JoRAS(https://joras.csis.u-tokyo.ac.jp/)からデータを無料に入手することが可能である（利用申請が必要となり，詳細は HP を参照してください）．各データセットは市区町村ごとに csv 形式で整備される．

2. データ仕様

2.1 人口データ

項目	説明
世帯 ID	世帯を識別するユニーク ID
世帯種類	単独・夫婦のみ・夫婦子供等 16 種類（表 2）
市区町村番号	住所の市区町村
住所	住所の経度・緯度
個人 ID	個人を識別するユニーク ID
年齢	対象者の年齢（表 3）
性別	1.男性　2.女性
就業状態	9 種類（表 4）

2.2 活動データ

項目	説明
個人 ID	世帯を識別するユニーク ID
年齢	対象者の年齢（表 3）
性別	1.男性　2.女性
就業状態	9 種類（表 4）
活動内容	6 種類（表 5）
活動開始時間	整数型　0 時からの秒数
活動持続時間	整数型　秒数で表す
活動場所	実数型　経度緯度
市区町村コード	活動場所の市区町村コード

2.3 トリップデータ

項目	説明
個人 ID	世帯を識別するユニーク ID
トリップ ID	トリップを識別する ID
出発時間	整数型　0 時からの秒数
出発場所	実数型　経度緯度
到着場所	実数型　経度緯度
交通手段	4 種類（表 6）
移動目的	6 種類（活動内容と同じ）（表 5）
就業状態	9 種類（表 4）

2.4 軌跡データ

項目	説明
個人 ID	世帯を識別するユニーク ID
Unixtime	整数型　0 時からの秒数
Timestamp	yyyy-MM-dd HH:mm:ss
経度	実数型
緯度	実数型
交通手段	4 種類
移動目的	6 種類（活動内容と同じ）
リンク ID	経度緯度が DRM 上該当するリンク ID

3. コード一覧

表 2　世帯類型コード

コード	世帯類型	配偶者	子供	両親	他の親族
1	夫婦のみ	1	0	0	0
2	夫婦と子供	1	1-9	0	0
3	ひとり親（父）と子供	0	1-9	0	0
4	ひとり親（母）と子供	0	1-9	0	0
5	夫婦と両親	1	0	2	0
6	夫婦とひとり親	1	0	1	0
7	夫婦，子供と両親	1	1-9	2	0
8	夫婦，子供とひとり親	1	1-9	1	0
9	夫婦と他の親族	1	0	0	1-9
10	夫婦，子供と他の親族	1	1-9	0	1-9
11	夫婦，両親と他の親族	1	0	2	1-9
12	夫婦，子供，親と他の親族	1	1-9	1	1-9
13	兄弟姉妹のみ	0	0	0	1-9
14	他に分類されない世帯	0	0	0	1-9
15	非親族を含む世帯	0	0	0	0
16	単独世帯	0	0	0	0

表 3　年齢コード

コード	内　容	コード	内　容
0	0 歳以上 5 歳未満	9	45 歳以上 50 歳未満
1	5 歳以上 10 歳未満	10	50 歳以上 55 歳未満
2	10 歳以上 15 歳未満	11	55 歳以上 60 歳未満
3	15 歳以上 20 歳未満	12	60 歳以上 65 歳未満
4	20 歳以上 25 歳未満	13	65 歳以上 70 歳未満
5	25 歳以上 30 歳未満	14	70 歳以上 75 歳未満
6	30 歳以上 35 歳未満	15	75 歳以上 80 歳未満
7	35 歳以上 40 歳未満	16	80 歳以上 85 歳未満
8	40 歳以上 45 歳未満	17	85 歳以上

表 4　就業状況コード

コード	内容	コード	内容
10	幼児	15	大学生
11	学齢前	16	短期大学（専門学校含む）
12	小学生	21	就業者
13	中学生	23	無職者
14	高校生		

表 5　活動内容コード

コード	内容	コード	内容
1	在宅	200	外食
2	通勤	300	通院
3	通学	400	自由行動
100	買い物	500	業務

表 6 交通手段コード

コード	内容	コード	内容
1	徒歩	3	自動車
2	自転車	4	電車

表 7　都道府県コード

コード	内容	コード	内容	コード	内容
01	北海道	17	石川県	33	岡山県
02	青森県	18	福井県	34	広島県
03	岩手県	19	山梨県	35	山口県
04	宮城県	20	長野県	36	徳島県
05	秋田県	21	岐阜県	37	香川県
06	山形県	22	静岡県	38	愛媛県
07	福島県	23	愛知県	39	高知県
08	茨城県	24	三重県	40	福岡県
09	栃木県	25	滋賀県	41	佐賀県
10	群馬県	26	京都府	42	長崎県
11	埼玉県	27	大阪府	43	熊本県
12	千葉県	28	兵庫県	44	大分県
13	東京都	29	奈良県	45	宮崎県
14	神奈川県	30	和歌山県	46	鹿児島県
15	新潟県	31	鳥取県	47	沖縄県
16	富山県	32	島根県	－	－

4. 利用上の注意

4.1 データの精度について

擬似人流データの評価について，携帯電話データと都市圏のパーソントリップ調査を用いて，異なる規模の都市圏に対して行われました．生成されたデータセットは，以下の 3 つの側面で観測データの統計的性質を再現されています．

人口分布：精度は空間解像度による変わります．行政（市区町村）レベルでは，携帯電話データと比較して相関係数の値が 0.98 であることが示され，擬似データセットが実データの代替として使用できることを示しています．集約レベルの解像度が向上するにつれ，1000×1000m^2 メッシュ上の人口分布は，R^2=0.81 という説得力のある精度を示しており，これはほとんどの実世界のアプリケーションで適切なものと考えられます．

トリップ数：人口分布の評価結果と同じ，生成されたトリップ数の精度もエリアやトリップ目的に依存しています．通勤トリップは，都市圏では地方都市よりも良好に構築されており（R^2=0.58〜0.67），地元の都市では R^2=0.48 です．逆に，レジャー目的（買い物・外食・娯楽など）のトリップの評価結果は，すべての都市圏で R^2 が 0.5 以上と非常に正確でした．

トリップカバレッジ：都市圏の結果は，各目的のトリップ数の誤差がおおよそ 10%であることを示しています．地方都市では，通勤トリップ数が 17%低く推定され，その影響として，帰宅トリップ数が不十分であるとされました．

リンク交通量：現在，本データセットの主体は「人」です．一方，本データセットはできる限り公開できるようにしますので，データ生成する際，個人のプライバシーを含む情報の使用を最大限に避けます．その結果，個人レベルの精度は不十分なところがいくつかあります．例えば，軌跡データを生成する際には，エージェントが最短経路で経路探索を行いますので，散歩・買い物行動において回遊行動，高速道路の料金の影響

詳細は，擬似人流データの論文を参照してください[3]．

4.2 軌跡データと時間帯別リンク交通量のリンク ID の扱い方について

軌跡データと時間帯別リンク交通量データに，リンク ID の欄をついております．このリンク ID は金杉ら[4]日本デジタル道路地図協会が提供するデジタル道路地図（DRM）に基づいて開発した鉄道ネットワークデータですが，元の DRM データのリンク ID と違います．なお，このデータも，研究目的であれば人の流れプロジェクト事務局に(pflow@csis.u-tokyo.ac.jp)に連絡し，提供可能となります．

4.3 データ生成上にパーソントリップ調査データの扱い方について

擬似人流データを作成する際に，PT 調査のマスターデータを使っていません．ただし，エージェントモデルのパラメータを算出するため，人の流れデータを集計した統計情報

3

4 金杉洋，関本義秀，樫山武浩，人々の流動再現へ向けたオープンな鉄道インフラデータの構築，第 22 回地理情報システム学会講演論文集, 2013.

を使っています．具体的に言えば，擬似の人間（エージェント）の活動（例えば在宅・勤務・買い物など）を決める時，各活動の確率を，人の流れデータから統計を取りました．レジャー活動（買い物や食事など）の目的地選択をする用の離散選択モデルのパラメータを，人の流れデータから学習しました．人の流れデータを集計したものですので，実質的には公開しているPTの集計データレベルだと言えると思います．

また，作成するには，人の流れデータとしては全国分ではなく，一部の都市圏の地域のものを使っています．

4.4 人の流れデータと擬似人流データの違いについて

人の流れデータはパーソントリップ調査データ（PTデータ）から起終点の時空間位置をジオコーディングし，最短経路ベースで経路探索を行い，1分ごとの位置を各ネットワークの詳細データをもとに内挿を行うことにより作成したものです．ただ，各トリップの起終点の時空間位置はゾーンレベルで建物面積に応じて配分し，匿名化しました．

擬似人流データは基本的に，国勢調査など公的統計データから作ったモデルから人工的生成したデータです．市区町村レベルやメッシュ単位で集計すれば実際のデータに合いますが，個人レベルの軌跡は実際のPT調査に一致しないです．また，人の流れデータはPT調査が行われた地域しか提供していませんが，擬似人流データは全国規模全人口分を提供しています．