エージェントワークフローを活用した大規模言語モデルによる擬似人流データの生成と精度向上

瀬貫雄介*・龐岩博・関本義秀****

An Agentic Workflow Approach for Pseudo People Flow Data Generation with Large Language Models

Yusuke Senuki*, Yanbo Pang**, Yoshihide Sekimoto**

Abstract: This study explored a method for generating high-precision pseudo people flow data by leveraging Large Language Models (LLMs) and introducing an agent-based workflow. Specifically, it employed an iterative feedback process involving an AI agent that generates synthetic people flow by few-shot, and a separate AI agent that evaluates the generated people flow. This approach resulted in a stepwise improvement in the quality of simulated people flow data. The proportion of flows meeting the criteria increased by 11.8 percentage points compared to conventional methods during the first generation and by 30.0 percentage points during the tenth generation. Furthermore, the frequency of activity appearances in the flows increased by 41.6%, confirming that the generated data includes diverse cases not sufficiently represented by single-generation models. Such data is expected to be utilized in urban planning and commercial analysis scenarios as a method to simulate people flow.

Keywords: 擬似人流データ（pseudo people flow data），大規模言語モデル（large language models），エージェントワークフロー（agentic workflow）

１．はじめに

民間による商業分析や自治体における都市計画・災害対策など，様々な場面で GIS を用いた人流（人の移動）のデータの応用が広まっている．一方で，それらの目的でスマートフォンから得られた位置情報などのデータを直接使用することは，個人のプライバシーを侵害するため許可を得ることが非常に難しい．この問題を解決する手法として，少量の実データや統計的知見を基に，人々の移動パターンをアルゴリズムで生成する「擬似人流データ」の開発が注目されており（Kashiyama et al.,2024），大規模言語モデル（LLM）を基に移動モデルを構築する研究も行われている（Wang et al.，2023）．

その一方で，生成系モデルで大量に活動スケジュールを作ると，イベント回数や時間配分などの外形的な定量指標が許容範囲でも「人間らしくない」「単純すぎる」「現実では起きにくい」サンプルが残るという問題が避けがたい．これは人間行動の複雑さに対して，「人間らしさ」「現実らしさ」の判定を定量化しにくいことに起因する．人間の専門家による定性的評価は可能だが，大量・自動化が難しく，モデル改良の反復に耐える評価基盤になりにくい．

本研究の核心は、このギャップを LLM を評価者（Evaluator）として用いるエージェントフレームワークで埋める点にある．すなわち，生成→評価→再生成の反復ループを導入し，評価エージェントが「矛盾の有無」と「多様性（人間らしい広がり）」の双方をもとに 5 段階で採点し，閾値に満たない場合は改善コメントをもとに再生成する．これにより，これまで人手に頼ってきた曖昧かつ文脈依存の判断を自動かつ大規模に実施できる．実装では生成エージェント・評価エージェントの双方に GPT-4 を用い，15 分スロットの 1 日スケジュール，23 語彙の制約，最大 10 回の反復，5 段階の評価指標を採用した．GPT-4 は文脈理解や推論能力が高く，複雑な人間行動や社会的文脈を考慮した擬似人流の生成が可能である一方で，高い推論能力を持つため，生成された人流データの一貫性や妥当性を評価し，数値スコア（GPTScore）を算出する役割も担うことができる．

本稿では，2 章でエージェントワークフローの詳細やデータの具体的な内容について紹介した上で，3 章で評価エージェント・生成エージェントを稼働させた結果を示した．さらに 4 章でそれらの結果を踏まえ考察した上で，5 章では今後の研究の可能性について記した．

２．提案手法

２.１　全体の構成

本研究では，人流を生成するエージェントと生成され

*　学生会員　東京大学 空間情報科学研究センター（Center for Spatial Information Science at the University of Tokyo）
〒277-8568　千葉県柏市柏の葉 5-1-5，　E-mail：senuki@csis.u-tokyo.ac.jp
**　正会員　　東京大学 空間情報科学研究センター（Center for Spatial Information Science at the University of Tokyo）

表 1　人間による点数ラベルの付与方針

5 点…矛盾点がない上で，買い物や余暇などの多様な行動が含まれる人流．
4 点…矛盾点はないが，やや多様性に欠ける人流．会社員の「家→通勤→会社→帰宅→家」など．
3 点…矛盾点はないが，非常に多様性に欠ける人流．会社員の「家→会社→家」や一日中「家」のみなど．
以下のような矛盾点がある場合は，矛盾の程度に合わせて元の点数から 1 個につき 1~2 点減点．（下限は 1 点）
・通勤や通学が 2 時間以上．あるいは通勤と帰宅・通学と帰宅の時間差が 45 分以上．
・途中に挟まれる不自然な移動．例えば「会社→通勤→会社」など．
・年齢や職業にそぐわない行動．例えば 10 歳の会社など．今回は性別は点数判定の要素としては用いなかった．

表 2　生成エージェントに与えた個人属性・生成エージェントの出力・評価エージェントの出力の例

【与えた個人属性】What is the daily schedule of a female aged 35-39 who works as a Professional/engineering_workers?
【生成エージェント出力】00:00-07:00 House, 07:00-07:30 Commute, 07:30-17:30 Office, 17:30-18:15 Private_Space, 18:15-18:30 Back_Home, 18:30-22:00 House, 22:00-22:30 Go_Eat, 22:30-24:00 House
【評価エージェント出力】スコア: 5　　改善点: 特に改善点なし。属性と行動が一致し、多様な行動が含まれている。

た人流を評価するエージェントのフィードバックの繰り返しによって人流の精度を向上させる．人流を 1 件生成する実行過程は以下①から④であり，これを生成する件数分繰り返す．多様な生成と一貫した評価のため，温度（temperature）は生成時は 0.7, 評価時は 0.0 とした．

① 生成エージェントが，生成する人流の個人属性（性別・職業・年齢層）を決定する．これは訓練データからランダムに選ばれる．
② 生成エージェントが，15 件の few-shot を含む生成用プロンプトに基づいて人流を 1 件生成する．
③ 評価エージェントが，15 件の few-shot を含む評価用プロンプトに基づいて人流を 5 段階で評価する．この際，点数だけではなく何故そのような評価結果になったか説明した短文も 50 文字以内で出力する．
④ 評価が 3 点以下で，なおかつ現在のループ回数が 9 回目以下の場合は，生成エージェントに評価結果の短文を見せ，再生成の指示を送る．逆に，評価が 4 点以上か，あるいは現在のループ回数が 10 回目の場合は，生成された人流を生成結果として書き加える．

2.2　データの内容

本研究で用いられる人流のデータは，0 時から 8 時 45 分が活動 X，8 時 45 分から 9 時 15 分が活動 Y……のように，活動の区切れ目が 15 分ごとに置かれ，個人属性は性別（男性・女性），職業（営業職・技術職・学生など），年齢層（20~25 歳など基本的に 5 歳区切り）の 3 要素から成り立つ．人流において活動を表す語彙は，家・通勤・会社・ショッピングなど全 23 種類が使われている．

本研究では，訓練データとして few-shot 等に用いる人流データとして，Zhang et al.（2024）が LLaMA3 をファインチューニングし，few-shot を組み合わせることで一括生成した人流データを用いた．この一括生成モデルでは元データとして，2008 年に東京都市圏（東京都・神奈川県・埼玉県・千葉県・茨城県南部）を対象に行われた第 5 回東京都市圏パーソントリップ調査の結果が用いられている．この研究では基礎的な人流の大量生成に成功した一方で，多様性にやや欠けるという欠点が見られた．

2.3　スコアリング手法

Zhang の一括生成モデルで生成された人流 1000 件に対して，手動で 5 から 1 の点数ラベルを付与する．5 が最高点，1 が最低点であり，具体的な点数の基準は表 1 に記した．但し，1 点の人流は数が少なかったため，訓練の精度向上のためにデータの一部を表 1 の基準で 1 点になるよう改悪し，1 点が全体の 1 割を超えるようにした．

2.4　擬似人流データの生成・評価

生成エージェントには，以下の 5 種類のプロンプトを与えて擬似人流データを生成させた．

① 基本のプロンプト（「あなたは人流データを生成するエージェントです」で始まる簡単な文章）
② 制約ルールの明文化（「属性と行動の整合性を保ってください」「多様な行動を含めてください」など好ましい生成方法のルールを簡単に書いた文章）
③ 5 点から 1 点の簡単な採点基準
④ スコア例の few-shot（点数ラベル付きの人流について，生成する人物の属性と近いものを，5 点〜1 点のものをそれぞれ 3 個ずつ，計 15 個与える）
⑤ 生成時に使える語彙の種類（全 23 種）

評価エージェントには上記の①に相当する基本のプロンプトのほか．上記の②③④と同じプロンプトを与えた．

出力結果は，表 2 のように個人属性・生成結果・評価結果の 3 要素を含む行が，生成された人流の個数分の行数で結果として保存される．

表 3　一括生成モデルと今回のモデルでの生成された人流の頻出シーケンスの各上位 3 件

【Zhang et al. (2024) の一括生成モデルで生成された人流の頻出シーケンス上位 3 件】
House→Commute→Office→Back_Home→House　157/500 (31.4%)
House→Shopping_Daily→Store Daily→Back_Home→House　30/500 (6.0%)
House→Go_Scholl→School→Back_Home→House　27/500 (5.4%)
【今回の実験で生成された人流の頻出シーケンス上位 3 件】
House→Commute→Office→Go_Eat→Socializing→Back_Home→House　30/500 (6.0%)
House→Commute→Office→Shopping_Daily→Store_Daily→Back_Home→House　18/500 (3.6%)
House→Commute→Office→Go_Eat→Office→Shopping_Daily→Store_Daily→Back_Home→House　17/500 (3.4%)

3．実験と結果

3.1　評価エージェントによる評価結果

まず，評価エージェントの適切な人流評価を確認するための実験を行った．人間が点数ラベルをつけた 1000 件の人流のうち，半分の 500 件を人流と点数がセットになった訓練データとしたうえで，残りの半分をテストデータとした．プロンプトや few-shot の個数を評価エージェントと同じ要領で設定して点数をつけさせ，GPT-4 が適切に点数をつけられているかを確認した．

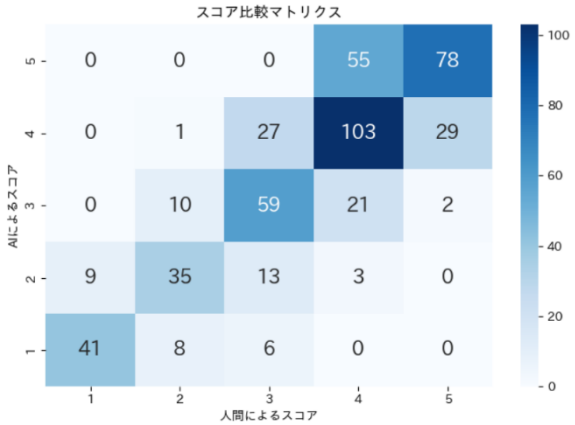


図 1　人間の点数ラベルと評価エージェントによる点数付けの比較マトリクス

図 1 のマトリクスを確認すると，対角線上のセルの値がそれぞれ最も大きく，その次に対角線の一つ隣のセルの値が大きくなっていることから，AI エージェントによる点数の付け方が人間の点数の付け方の傾向とある程度一致していることが読み取れる．

3.2　生成エージェントによる生成結果

次に，エージェントワークフローに基づいて実際に人流を 500 件生成し，生成結果を Zhang による LLaMA3 を用いた一括生成モデル（訓練データ）と比較した．

表 3 からは，本研究による生成では単純な人流の数が減っていることがわかった．Zhang の一括生成モデルでは 30%以上で「家→通勤→職場→帰宅→家」という単純な人流が生成されていたのに比べ，本研究では一番多い種類のものであっても全体の 6%であるように，多様な人流のシーケンスが生成されていることが確認された．

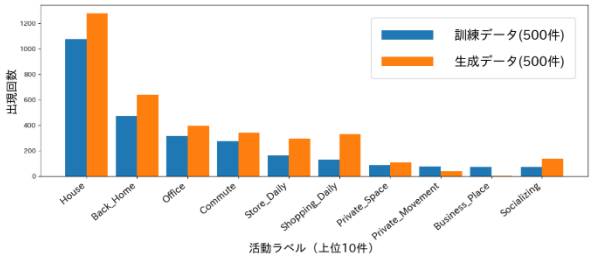


図 2　Zhang の結果(青)と今回の結果(橙)による活動内容の語彙の合計生成回数の比較グラフ

図 2 で示した語彙の出現回数の上位 10 件を見ると，本研究で生成された人流は，Zhang の一括生成モデルと比較して全体的な傾向を保ちつつも人流内の活動の回数を全体的に増やしていることがわかる．今回生成された人流 500 件のうちに語彙は合計 4280 回含まれており，Zhang の一括生成モデルである 3023 回に比べて 41.6%，人流 1 件あたりでは 2.095 回増加していた．特にショッピングが 186.7%増加するなどしており，余暇の活動を中心に従来より多様な人流が生成されていることが確認できる．

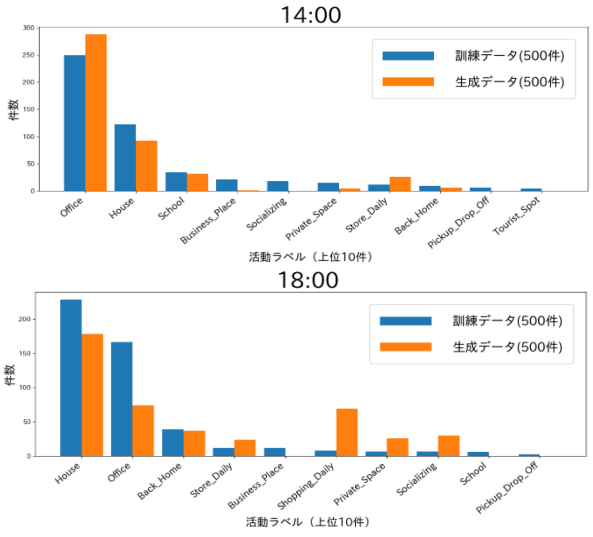


図 3　Zhang の結果(青)と今回の結果(橙)による活動内容の特定の時間(14 時と 18 時)での数の比較グラフ

さらに，図 3 で昼 14 時と夕方 18 時時点に限定した語彙の出現回数を比較すると，18 時には家や職場が減った代わりにショッピングなどが多く生成されている一方で，14 時にはそれらはあまり生成されていないことがわかり，生成エージェントが多様性のある行動をどの時間帯に多く出現させるべきかを考慮できていることが読み取れる．

3.3　再生成による精度向上の結果

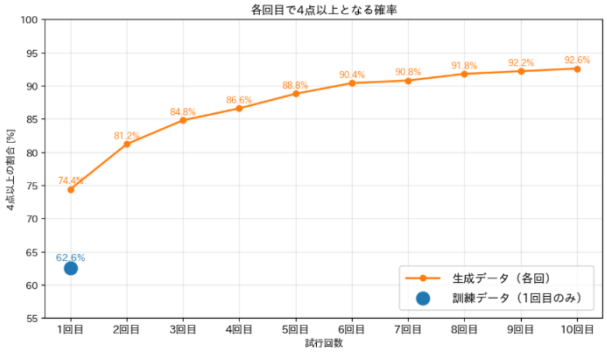


図 4　再生成回数ごとの 4 点以上の割合のグラフ

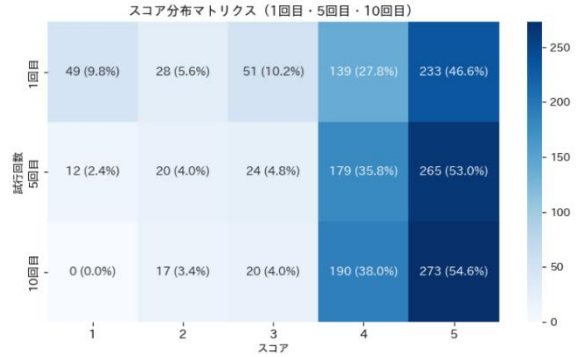


図 5　生成 1，5，10 回目における点数のマトリクス

再生成を繰り返すことでどのように生成結果が向上するかを調べた．図 4 からは，4 点以上の割合が Zhang の一括生成モデル比べて生成 1 回目で 11.8 ポイント増加，生成 10 回目で 30.0 ポイント増加していること，再生成を繰り返すほど 4 点以上の割合が伸びるものの上昇率は逓減してゆくことが示された．また図 5 からは，再生成を繰り返すことで 1 点の生成結果が無くなったものの，2, 3 点はいくつか修正されずに残っていることが見られ，完全に問題がある生成結果よりも僅かに問題がある生成結果を修正する方が相対的に難しいことが示された．

3.4　再生成の閾値変更による効果

4 点となっている閾値を 5 点，3 点に変更した場合の実験も行った．図 6 からは閾値が 5 点の場合は 1~2 回の再生成による効果が大きいこと，閾値が 3 点の場合でも完全に質の低い人流を除去するのは難しいことが示された．また，閾値 5 点の場合の語彙の出現回数は Zhang の一括生成モデルの 3023 回に比べて 47.7%多い 4465 回であり，これは閾値 4 点のときを 6.1 ポイント上回った．

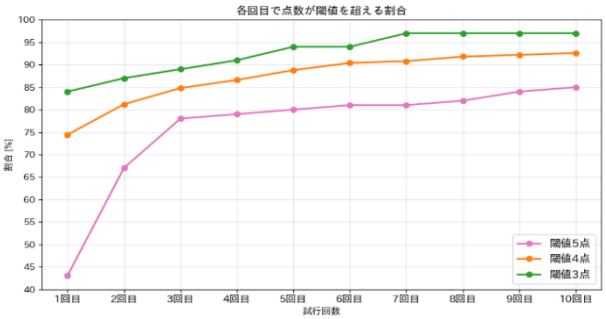


図 6　再生成回数ごとの閾値以上の点の割合のグラフ

4．考察

3 章の結果から，GPT-4 による GPTScore が人流の評価としても適切に作動すること，従来よりも正確かつ多様性のある人流が生成されていること，再生成システムによってワークフロー内で適切に修正が行われていることが示された．その一方で，訓練データでは存在した一部の語彙が極端に生成されなかったことも確認され，多様性をより確保するための他の必要性も見受けられた．

今後は閾値を状況に応じて変化させることで「単純な人流」と「多様な人流」を適切な割合で生成させ現実の人流に近づけることや，一部の語彙を積極的に生成結果に含めるようなプロンプトを状況に応じて用いることでより多様性を確保することが展望として挙げられる．

5．結論と今後

本研究では，生成エージェントと評価エージェントを動かすエージェントワークフローによって擬似人流データを高精度に生成する方法を探求し，従来の一括生成型のモデルに比べてより正確で多様な疑似人流が生成されたことを確認した．今回は人流を行動内容に限定して言語で処理したが，今後の研究では，今回の結果を実際の地図上での人流生成の補助に活用させ，GIS 上で商業分析や都市計画などの研究に役立てることを目指している．

参考文献

Kashiyama, T., Pang, Y., Shibuya, Y., Yabe, T., and Sekimoto, Y.(2024) Nationwide synthetic human mobility dataset construction from limited travel surveys and open data. *Computer-Aided Civil and Infrastructure Engineering,* **39**, 3337–3353.

Wang, Xinglei, Meng Fang, Zichao Zeng, and Tao Cheng (2023) Where Would I Go Next? Large Language Models as Human Mobility Predictors, arXiv:2308.15197.

Zhang, Y., Zhang, K., Pang, Y., and Sekimoto, Y.(2024) Agentic Large Language Models for Generating Large-Scale Urban Daily Activity Patterns. *2024 IEEE International Conference on Big Data (BigData)*, Washington, DC, USA, pp. 6815-6822.