2024 年度 修士論文

ライフコースに基づく広域的居住軌跡モデル

Lifecourse-based Regional Relocation Trajectory Model

東日本圏における過去 20 年間の移動履歴への適用

Application to the 20-year Mobility History in the East Japan Area

37-236021
白井 帆香
Honoka SHIRAI

東京大学大学院 工学系研究科 社会基盤学専攻
主査: 羽藤 英二 教授
2025 年 1 月

付録 B

国勢調査匿名データを用いた人口合成実験

B.1　緒言

本研究では，2.4 節で導入した通り，本研究で構築した居住軌跡モデルを政策シミュレーションに適用することを見据え，そのインプットとなる合成人口の生成を試みた．統計法に基づいて、独立行政法人統計センターから「平成 17 年 国勢調査」「平成 22 年 国勢調査」「平成 27 年 国勢調査」「令和 2 年 国勢調査」（総務省）に関する匿名データの提供を受け，学習データとして用いた．研究を進める過程で，生成データの精度向上や，モデルとの接続が困難であったことから，合成人口に居住地軌跡を投影したシミュレーションを行うには至らなかった．本章では，合成人口の生成に関する試行及び判明した問題点について述べる．

B.2　条件付き合成人口の生成

本研究では，モデルの対象地域（岩手県・宮城県・福島県・茨城県・埼玉県・千葉県・東京都・神奈川県）における指定した年齢・性別の人口を合成するため，調査年・居住都県・年齢・性別の 4 つの属性を条件づけた合成人口の生成を試みた．そのため，学習データとして用いた国勢調査匿名データから，該当する居住地のサンプルを抽出し，表に示すとおり 4 つの属性を条件特徴量，の属性を生成対象の特徴量として，Coditional VAE による生成を試みた．また，併せて Conditional VAE-GAN による訓練も試みた．これは，模擬データを用いた生成実験において，VAE は周辺分布への適合度は高いが組み合わせの多様性が低い一方，WGAN は学習が不安定である傾向が見られたため，両者の長所を組み合わせることで，生成データの精度向上を図った．

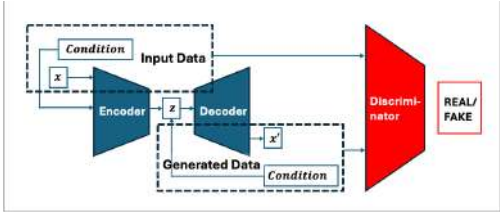

図 B.1: Conditional VAE-GAN の学習構造

表 B.1: 合成人口の生成に用いた属性

	属性	カテゴリ数
条件特徴量	調査年	4
	居住都県	8
	年齢	19
	性別	2
生成対象の特徴量	世帯サイズ	5
	配偶者有無	2
	同居する子どもの人数	3
	世帯内に 15 歳未満が含まれるか否か	2
	5 年前居住地	3 (現住地・同県内・県外)
	職務形態	4
	従業地	3 (自宅・居住自治体内・自治体外)
	正規雇用か否か	2
	大卒か否か	2
	人口集中地区に居住か否か	2
	持ち家か否か	2
	戸建てか否か	2

学習には，対象地域の 8 都県に居住する 1,686,320 人のサンプルを用いた．生成したデータについての各指標を表 B.2 に示す．評価に用いた指標は

- **生成組み合わせ数**: 生成データにおいて，実データと一致するユニークな組み合わせの数.
- **SRMSE (Standardized Root Mean Squared Error)**: 実データとの誤差を標準化したルート平均二乗誤差低いほど実データと生成データの類似度が高いことを示す.
- **Precision**: 生成データのうち，実データと一致するデータの割合．高い値は、生成データの正確性が高いことを示す.
- **Recall**: 実データのうち，生成データに含まれる割合．高い値は、実データの多くを再現できていることを示す.

であり，Recall を除いた指標で VAE-GAN が VAE よりも優れていることが示された．しかしながら，Conditiona-VAE であっても, Precision, Recall ともに 5 割に届かない低い値にとどまいる．Conditional VAE-GAN で生成したサンプルの各特徴量のカテゴリごとのデータ数分布を図 B.2 に示す．データ分布における周辺分布をみなせるものだが，実データと高い精度で一致していることがわかる．これらの結果から，実装した Conditional VAE-GAN は特徴量ごとの周辺分布においては高い生成精度を示しているものの，それらのカテゴリごとの特徴量の組み合わせの再現度は低いことが示唆される．多次元のカテゴリ変数の組み合わせの生成は，機械学習分野においても困難なタスクとされており，本研究でもその困難さが如実に現れる結果となった.

B.3　縦断合成人口の実現に向けて

2.4.3 項で述べた通り，特に本研究が目指す長期間の軌跡のシミュレーションにおいては，断面的な合成人口の生成にとどまらず，多時点にわたる縦断的な合成人口の生成が必要である．前段として，

表 B.2: VAE と VAE-GAN の比較

100,000 サンプル	**VAE**	**VAE-GAN**
生成組み合わせ数 （実データ：62661）	32331	47820
SRMSE	2.045	1.739
Precision	0.352	0.458
Recall	0.374	0.348

Conditional VAE-GAN による人口合成手法を試みたが，この最も直感的な多時点への展開として，離散的な調査年を連続的な時間次元に埋め込みんだ上で，学習を行うことが考えられる．

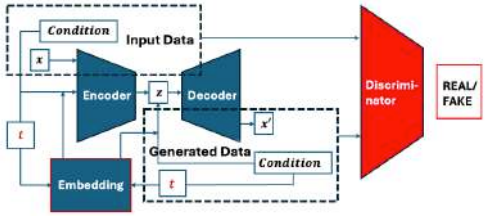

図 B.3: 時間埋め込みを含む Conditional VAE-GAN の学習構造

実データが存在する 2010 年と 2015 年の間を補完するデータを生成することを目指し，図 B.3 に示す学習構造を用いて学習を行い，2010 年-2015 年の 6 年間のデータを生成した．結果について，特に職業及び世帯サイズの周辺分布を図 B.4 に示す．周辺分布の再現度は，特に実データでの頻度が低いカテゴリで著しく下がっており，時間の埋め込みによる精度悪化が見られた．また，特に実データに存在する調査年の翌年と前年（今回のケースでは 2011 年と 2014 年）の周辺分布が著しく期待される分布と異なり，単純な埋め込みネットワークの導入では時間の連続性を考慮した生成が困難であることが示唆された．

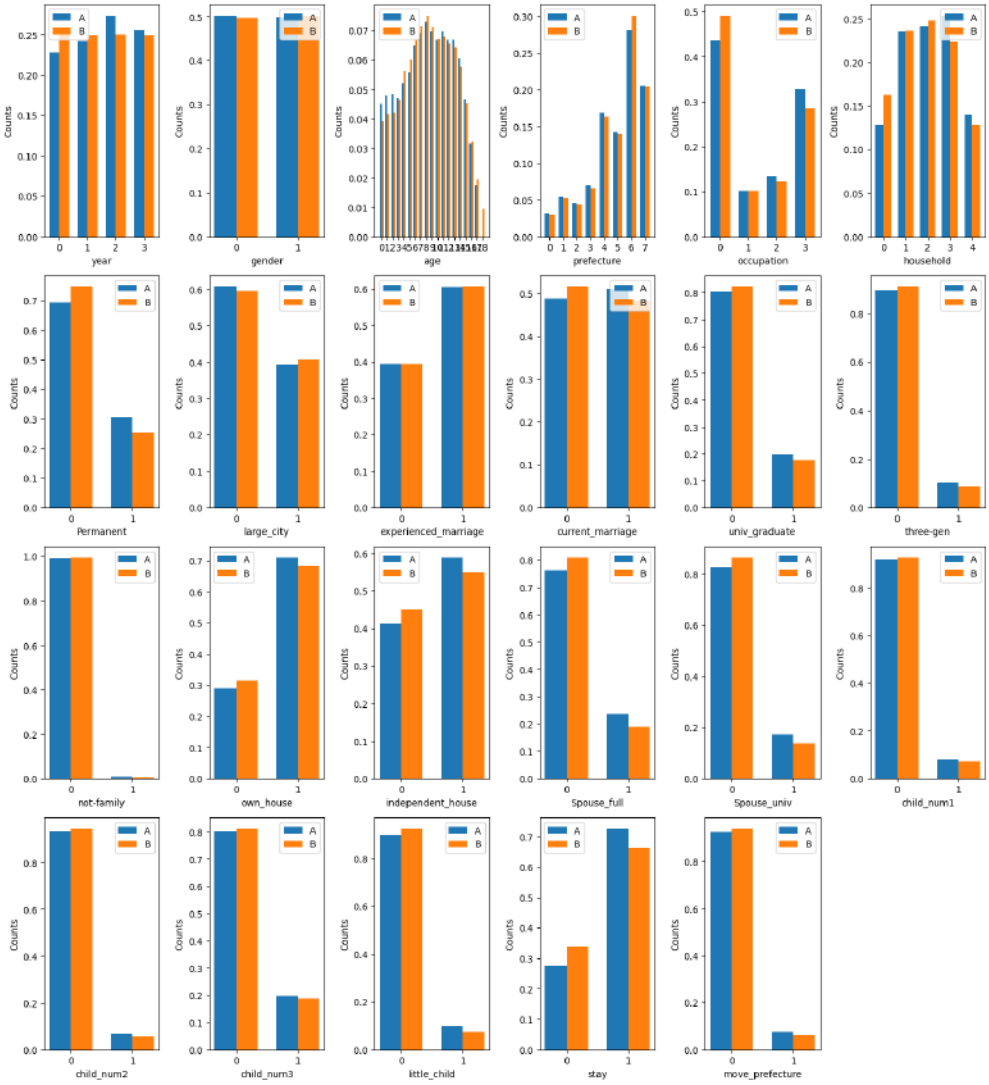

図 B.2: 実データサンプル【青】及び Conditional VAE-GAN で生成したサンプル【橙】の分布

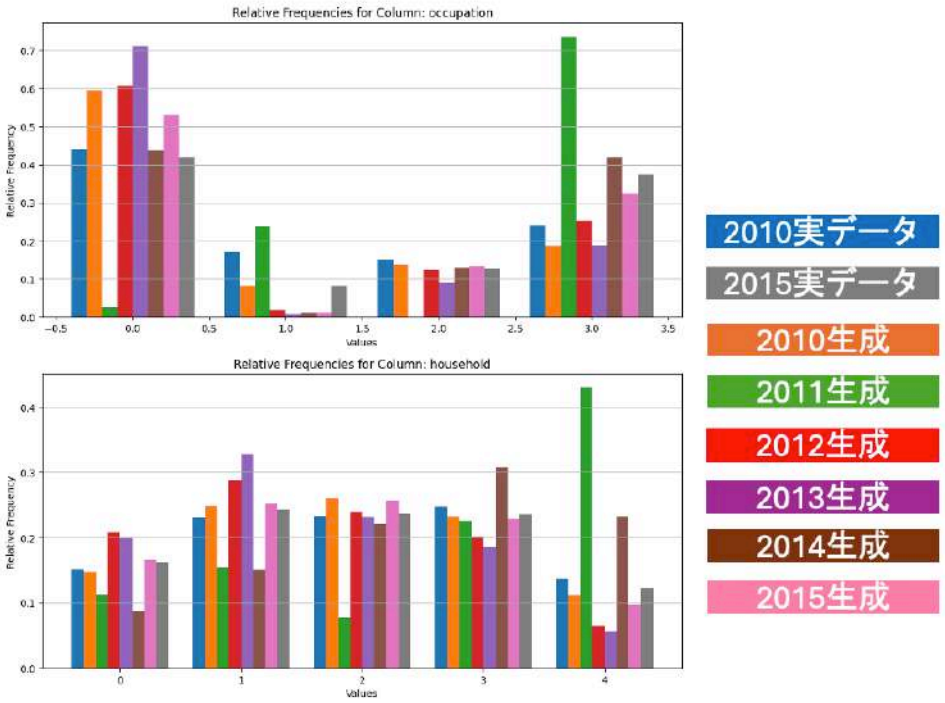

図 B.4: 時間を埋め込んだ生成の周辺分布の確認

B.4　スナップショットデータと縦断ミクロデータの融合に向けて

本研究では成果を出すことができなかったものの，国勢調査匿名データのようなスナップショットデータと，縦断ミクロデータの融合は，人口動態のシミュレーションにおいて有望なアプローチである．国勢調査匿名データは，多時点のデータが存在するものの，各時点のサンプルに連続性はなく，断面的な多次元な属性分布の把握が可能である．調査で収集された全サンプルのうち 1% 程度のサンプルが提供されるため，多次元のカテゴリ変数の組み合わせを大規模に取得することができる．一方で，本研究で実施した Web 調査などから取得される縦断ミクロデータは，こうした個々のサンプルの多次元の属性が時系列でどのように変化するかを把握することができる一方で，大規模なサンプル数を取得することは現実的でない．そのため，縦断ミクロデータから得られる長期間にわたるダイナミクスを，スナップショットデータに投影することで，大規模な時系列変化のシミュレーションを行う手法は非常に有用である．

このようなデータ融合の実現に資する手法としては，付録 A で紹介した SB 問題の適用がある．スナップショットデータから得た 2 時点の分布を両端制約条件としながら，縦断ミクロデータから得た事前知識をラグランジアンとして導入することが考えられる．また，同様の発想として，ODE2VAE[116] のような連続的な時間変化を考慮した生成手法を導入することも有効である．ODE2VAE は動画生成で発達した手法であり，VAE の潜在変数 z に時間的な進化を導入し，エンコーダー・デコーダーの学習と同時に，進化をパラメタライズした常微分方程式 (ODE) の学習を行う．スナップショットデータでエンコーダー・デコーダーを学習しつつ，縦断ミクロデータをエンコードした潜在変数 z の時系列変化を表す ODE を学習することで，スナップショットデータの多時点展開を行うことが可能と考える．