Vol.2022-MPS-138 No.39
Vol.2022-BIO-70 No.39
2022/6/28

並列リンケージ同定を用いた進化計算による合成人口モデルの生成

細川喜生 [1]　棟朝雅晴 [1,2]

概要：近年，社会シミュレーションの発展により，それに用いる市民の年齢，性別，収入，職業，学歴などの属性を統計データから復元することが求められている．しかし，これらの属性はプライバシー等の理由から保護されており，シミュレーションにて利用するためには信頼性の高い復元手法が求められる．疑似焼きなまし法や進化計算による最適化手法が提案されているが，年齢に対して最適化を行うために世帯数や総人口について事前の調整が必要であり，計算高速化のために用いられている並列計算では，統計データの分割による誤差の増加という問題点がある．そこで本論文では，リンケージ同定を用いることで，細かな初期世帯の設定を行わずに世帯間の相互関係を用いて複数の市民属性を統計データからの最適化し，より現実に近い人口データの作成を試みる．また，並列計算によりリンケージ同定を行うことで，データを分割せずに計算の高速化を試みる手法を提案する．

キーワード：進化計算，遺伝的アルゴリズム，リンケージ同定，合成人口

Generation of Synthetic Population Models by Evolutionary Computation Using Parallel Linkage Identification

YOSHIKI HOSOKAWA[†1]　MASAHARU MUNETOMO[†1,2]

1. はじめに

マイクロ・シミュレーション [1] や，エージェント・シミュレーション [2]といった社会シミュレーション技法が，モデル構築の自由度の高さから近年注目されている.これらのモデルでは，モデル化する社会における市民の属性を復元するにあたって，政府や行政が公表している統計データを使用することで再現する．しかし市民の個票レベルの情報は公開されておらず，復元したデータを使用する際も，個人が特定できないようにする等，プライバシーに考慮する必要がある．統計データに基づく市民属性の復元に関する研究は古くから行われており，様々な個票データの復元法が提案されている[3][4].

村田，原田，桝井は，様々な日本の市民属性を再現する手法を提案し，精度の高い復元を実現している[5][6][7].村田らの手法では，年齢に関する統計データの目的関数を最小化するために，それ以外の市民属性について事前に統計データとの調整や最適化を行ったデータを用意している.村田らにより並列計算を用いた世帯復元手法[8]も提案されているが，分割数を多くした際に分割された統計データの不整合により，誤差が増加してしまうという課題がある.

以上を踏まえ本研究では，日本の世帯データに着目し，市民属性の復元を目指す．復元手法として，近年多目的最適化問題などを解くために用いられている遺伝的アルゴリズム (Genetic Algorithm, GA)において，遺伝子間の依存関係を考慮して最適化を行うリンケージ同定に基づいた手法[9]を並列化した並列リンケージ導入し，復元データ内の各要素間の依存関係を調べることによって市民属性の復元を目指す．並列化において，対象のデータを分割するのではなく，依存関係を考慮して探索空間の分割を行うことにより，並列化による誤差を増やすことなく計算を行う手法を提案する．

2. 世帯と人口の復元手法

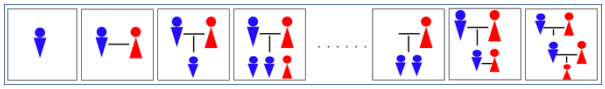

図 1　合成人口データのモデル

Figure 1　A synthetic population data model.

GA にて扱う 1 個体分の合成人口データのモデルは，図 1 のような複数の世帯から構成される世帯集団と含まれる市民の市民属性であり，遺伝操作により更新していく．本研究では先行研究[10]で用いられている 9 種の世帯類型を前提とする．この 9 種類の世帯が日本の約 95%を占めている.

復元するために，村田らの先行研究[6]にて使用されている 21 種の世帯類型別の人口分布や世帯内市民間の年齢差に関する統計データに加え，世帯類型毎の世帯数と,男女別の 1 歳階級の人口分布を使用する.

初期世帯については，扱う世帯数 H を決定し，細かい調整は行わず確率的に各世帯の世帯類型を決定し，その中の市民の属性をランダムに初期設定する．これは,リンケージ同定が多目的最適化に適している点，今後より複雑な属

1 北海道大学大学院情報科学院
Graduate Schools of Information Science and Technology, Hokkaido University

2 北海道大学情報基盤センター
Information Initiative Center, Hokkaido University

Vol.2022-MPS-138 No.39
Vol.2022-BIO-70 No.39
2022/6/28

性が追加された場合にも対応するためである.

本論文では GA にて扱う各個体の適応度として, 以下の式 (1) を計算する. 統計データと合成データの誤差の総和を適応度関数として扱い, 最適化していく.

$$f(A) = \sum_{s=1}^{S}\sum_{j=1}^{G_s} |c_{sj}(A) - Round\left(r_{sj} \times m_{sj}(A)\right)| \quad (1)$$

ここで, A は合成データ, S は統計データの総数, s は統計データの種類, G_s は統計データ s の項目数, c_{sj} は統計データ s の条件 X_{sj} と Y_{sj} を満たす復元データ内の市民数, r_{sj} は統計データ s の項目 j の割合, m_{sj} は条件 X_{sj} を満たす市民数, Round 関数は小数点以下を四捨五入する関数である. これにより, 世帯類型, 性別, 年齢, 世帯内の役割について一度に全て最適化できる. 並列リンケージ同定を用いた GA については図 2 の手順で行う[9].

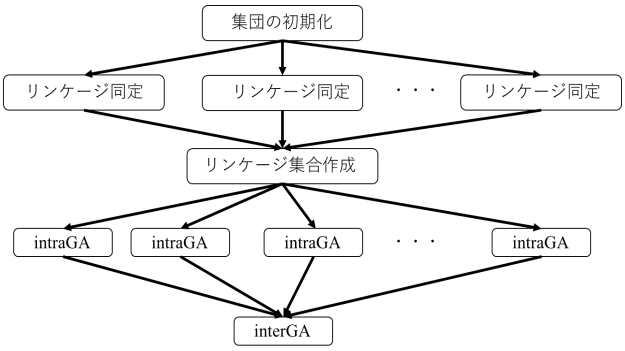

図 2　並列リンケージ同定を用いた GA

Figure 2　GA using parallel linkage identification.

3. 実験

実験については, 原田らの研究[5][8]と同様, 山形県の統計データ(H=500)に対して適合を試みた. これは山形県の平均世帯人員が 3.01 と 47 都道府県で最も高く, 1 世帯あたりの人員が多いほど復元が難しいと考えたからである.

本研究では北海道大学スーパーコンピュータ Grand Chariot を用い, 並列数毎の誤差と実行時間の複数試行における平均値及び標準偏差を求めた. 誤差に関する結果を表 1 に, 並列数毎の実行時間の平均値を図 3 に示す.

表 1　並列数毎の誤差と標準偏差の平均値

Table 1　Average and standard deviation v.s # of processors.

	初期エリート個体		最良個体	
並列数P	平均値	標準偏差	平均値	標準偏差
64	9378.05	121.53	81.10	13.21
32	9362.39	122.91	81.38	13.05
16	9275.55	123.30	82.18	11.05
8	9329.15	127.15	132.13	13.41
4	9302.05	128.72	164.73	16.93
2	9406.50	101.85	**30.50**	14.92
1	9327.89	100.62	7509.67	2223.41

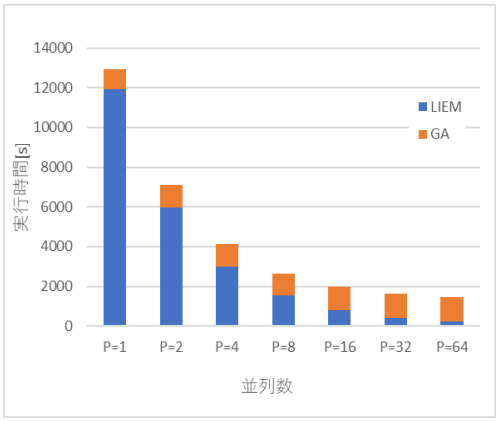

図 3　並列数毎の実行時間の平均値

Figure 3　Average execution time v.s. # of processors.

4. おわりに

本研究では, 並列リンケージ同定を用いて事前の調整を必要としない山形県の世帯・市民データの復元を行った. 誤差については, 最大 99.68%の削減をすることができ, 計算時間についてはリンケージ同定を用いた GA の並列数を増やすことが計算の高速化に有効であることを示した.

一方で今後の課題として, 統計データとの誤差を削減するため, より効果的なリンケージを検出するための市民属性の追加や目的関数の改良, 計算時間を削減するための実装の改良等が挙げられる.

謝辞　本研究を行うにあたり, 有益な助言をいただいた関西大学の村田忠彦教授に対し深く感謝いたします. 本研究は JSPS 科研費 JP20K11967 の助成を受けたものです.

参考文献

[1] 矢田晴那. ”政経分析ツールとしてのマイクロ・シミュレーション, ファイナンス” 35/40 (2010).
[2] 山影進. ”社会科学とマルチエージェントシミュレーションシミュレータ開発と事例提供の課題-” 情報科学 27, 1-10 (2007)
[3] A. G. Wilson, C. E. Pownall. ”A new representation of the urban system for modelling and for the study of micro-level interdependence.” Area, 246-254 (1976)
[4] 池田 心・喜多 一・ 薄田 昌広. ”地域人口動態シミュレーションのためのエージェン ト推計手法” 計測自動車制御学会第 43 回システム工学部会研究会, 11-14 (2010)
[5] 原田 拓弥・村田 忠彦・枡井 大貴. ”家族類型と世帯内の役割を考慮した SA 法による大規模世帯の合成” 計測自動制御学会論文集, 54 巻, 9 号, 705-717 (2018)
[6] 原田 拓弥・村田 忠彦. ”市区町村の統計表を考慮した都道府県単位の仮想個票の合成” 第 15 回社会システム部会研究会 (2018)
[7] 枡井 大貴・ 村田 忠彦. ”統計データからの市民の属性復元のための進化計算と SA による 2 段階最適化” システム制御情報学会論文誌, 30 巻, 6 号, 216-227 (2017)
[8] 原田 拓弥・ 村田 忠彦. ”並列計算を用いた SA 法による都道府県レベルの大規模世 帯の復元” 計測自動制御学会論文集, 54 巻, 4 号, 421-429 (2018)
[9] Masaharu Munetomo. ”Linkage identification based on epistasis measures to realize efficient genetic algorithms” Proceedings of the 2002 Congress on Evolutionary Computation. CEC’02, vol.2, 1332-1337 (2002)
[10] http://www.res.kutc.kansai-u.ac.jp/~murata/synthetic-methods/