市民属性の合成手法における年齢交換による誤差最小化

○枡井大貴 村田忠彦 原田拓弥（関西大学）

Minimizing Errors by Swaping Citizen's Age on Synthetic Population

∗D. Masui, T. Murata and T. Harada (Kansai University)

概要– 現実社会で起きている課題の検証の方法として，社会シミュレーションが注目されている．その中でも，対象地域や具体的な事例を取り上げた社会シミュレーションは，現実的な検証を行うことができる．このようなシミュレーションを実行するためには，対象地域の詳細な環境データと，そこに存在する市民データの個票が必要となる．環境データは GIS データなどが公開されているが，個人情報である市民データは一般的には非公開で，その利用には厳しい制限がある．そこで，池田らが提案した，人口統計データを用いて市民データを合成する手法を用いることで，統計情報の誤差が少ない市民データを得ることができる．本研究では，人口統計データと合成する市民データとの統計情報の誤差を 0 にすることを目的に，先行手法の最適化問題における近傍解生成法の改良を提案する．

キーワード: 社会シミュレーション，人口統計データ，シミュレーテッドアニーリング

1 はじめに

社会シミュレーションは，社会で起きている課題への解決策を提案し，その有効性を検証することができる．特に，対象となる地域を明確にしてエージェントベースシミュレーションを行うことで，より具体的な結果を得ることができる．市川ら [1] は感染症実用シミュレーションで東京都大島町を想定したシミュレーションモデルを構築している．花岡 [2] は京町家の取壊しの事例を取り上げ，京都市西陣地区を対象地域としてモデルを構築している．このようなシミュレーションを行う際，対象地域の環境データと市民データを用いてシミュレーションモデルを構築する必要がある．環境データとは，地理情報や建築物についての情報である．これらは公開されている GIS データから利用することができる．一方，市民データは世帯単位の家族構成や個人の年齢や性別などを表しており，個人情報に該当する．個別の市民データは一般的には非公開で，利用することは厳しく制限されている．そこで，池田ら [3] が提案した，市民属性の合成手法を用いることで必要となる市民データを得ることができる．

池田らが提案した手法は，複数の統計データの割合に従う市民集団をつくりだすための方法である．具体的には，公開されている統計データと市民集団との統計情報の誤差を計算する目的関数を設計して，最小化問題として定式化し，Simulated Annealing を用いて最適化している．池田らの手法では，目的関数値が 1.0 程度であれば，合成した市民集団が統計データの割合とほぼ正確に一致するものとしている．この指標を，対象となっている全ての統計データに対して満たすことは容易ではない．福田ら [4] は，指標の 1.0 を超えた統計データに対して重み付けをして最適化することで，全体の誤差を均衡させて対応している．しかし，合わせるべき全ての統計データについて目的関数値を 1.0 以下にはできていなかった．

そこで本研究では，対象の統計データに対応する全ての目的関数値を 1.0 以下にすることを目的に，最適化問題における近傍解生成の改良を提案する．池田らの手法では，市民の年齢を乱数により変更して近傍解を生成していた．提案手法では，同性の市民の年齢を交換することで近傍解を生成する．複数回試行の最適化実験を行い，提案手法の有効性を示す．さらに，探索回数を増加した実験を行い，提案手法により統計情報の誤差を改善できることを示す．なお，提案手法の近傍解生成を組み込むにあたって，最適化における解の初期生成が重要な処理となる．解の初期生成が重要であることは，著者らがこれまで取り組んできた，合成手法の改良の結果から結論付けたものである．

まず，合成手法の目的を，統計データとの誤差がより小さい市民集団のデータを得ることとして，誤差の値が最小値になっているかを直観的に計算できる目的関数を提案した [5]．なお，この際，統計データの項目数による平均化を式から除き，統計データごとの最適化の偏りを解消した．池田らの目的関数は，誤差の値を合わせるべき統計データの項目数で平均化していたために，項目数の大きい統計データとの誤差が少なく見積もられるという特徴があり，そのような項目数の大きな統計データを優先的に最適化するためには，目的関数における適切な重み付けが必要になっていた．

次に，誤差が小さい解を効率的に探索するために，年齢別人口の統計データとの誤差を考慮して市民の年齢を変更することで，従来よりも目的関数値を改善できる探索手法をを提案した [6]．この手法を用いることで，誤差が最小値の解を 1 つ発見できたことを報告している．しかし，その解を見つけることができたのは 500 回試行の中の 1 試行という結果になっていた．大規模な市民集団のデータを合成する時に，500 回試行の実験を行うことは現実的ではないため，高い精度で誤差が小さい解を求める必要がある．

そして，統計データとの誤差がより小さくなっている結果を分析すると，市民集団のデータに含まれる 9 種類の家族類型別世帯数と総人口が関係していることがわかった [7, 8]．世帯数と総人口を調整することで，統計データとの誤差がより小さい解を高い精度で得ることができた．世帯数と総人口は初期解生成の時点で決定しており，市民の年齢を調整するという最適化の過程ではそれらの値は変化しない．調整する時の指標となる値については，世帯数は統計データの割合に従っていたが，総人口は相関分析で求めたもので，統計データの割合とは関係のない値に調整していた．これは，市

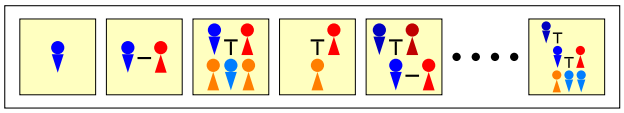

Fig. 1: 合成データのモデル

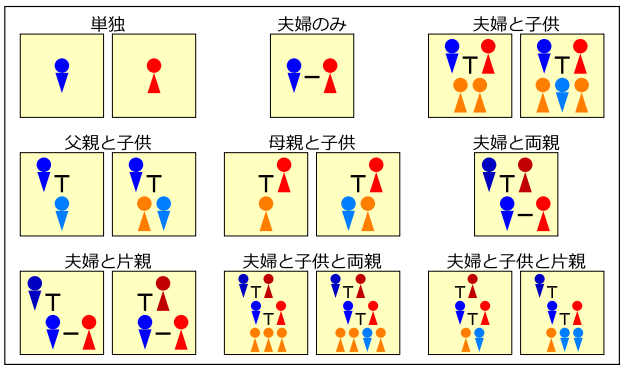

Fig. 2: 9 種類の家族類型

民集団のデータの総人口が統計データの割合に合っていない状態でも，市民の年齢を調整することで目的関数に組み込まれている統計データとの誤差は最小化することができることを意味する．したがって，総人口，男女比，1 世帯の子供の人数などの情報についても統計データの割合に合わせなければならない．これらの情報は，市民集団のデータの世帯数と総人口と同様に，初期解生成の時に決定しており，市民の年齢を調整する最適化では変化しない．以上のことから，より多くの統計情報に一致する市民集団のデータを得るためには，初期解生成の処理が重要となる．

全体の構成については，まず，第 2 章で池田らの手法（以下，先行手法とする）を基本とした市民属性の合成手法について説明する．第 3 章で本稿で提案する近傍解生成法と，最適化における解の初期生成について述べる．その後，第 4 章で提案手法の有効性を確かめるために，複数回試行の実験を行う．最後に，第 5 章で提案手法により得られた市民集団のデータについて考察する．

2　市民属性の合成手法

本研究は，先行手法における近傍解生成を改良し，統計情報との誤差が最小値である市民集団のデータを得ることが目的である．まず，先行手法となる市民属性の合成手法について説明する．

2.1　合成データのモデル

合成データとは，市民集団の個々の属性を合成したデータを表しており，本研究における解である．合成データのモデルを Fig. 1 に示す．合成データは複数の世帯で構成されており，それぞれの世帯には市民が存在している．全ての世帯は Fig. 2 の 9 種類の家族類型のいずれかに分類される．これらの家族類型は，公開されている統計では，日本の全世帯の 95%を包含している．世帯の中の市民は属性として，年齢，性別，親族関係，家族類型，世帯の役割を持っている．親族関係とは，親子や夫婦などの同一世帯内の市民同士の関係性を表現したものである．家族類型の属性とは，自分が所属している世帯がどの家族類型に分類しているかを示す情報で，世帯の役割は父，母，夫，妻，子供などを表すものである．

2.2　合わせる統計データ

現実社会のマクロな人口統計データ [9] と同じ統計情報を持つ合成データを得るために，次の 11 種類の統計データの割合に合わせる．

1. 父子年齢差（人口動態調査，出生，2010 年，中巻 第 8 表）
2. 母子年齢差（人口動態調査，出生，2010 年，中巻 第 8 表）
3. 夫婦年齢差（国勢調査，平成 22 年，第 17 表）
4. 男性の年齢別人口（国勢調査，平成 22 年，第 16–1 表）
5. 女性の年齢別人口（国勢調査，平成 22 年，第 16–1 表）
6. 男性年齢階層別の単独世帯人口（国勢調査，平成 22 年，第 16–1 表）
7. 女性年齢階層別の単独世帯人口（国勢調査，平成 22 年，第 16–1 表）
8. 男性年齢階層別の夫婦のみ世帯人口（国勢調査，平成 22 年，第 16–1 表）
9. 女性年齢階層別の夫婦のみ世帯人口（国勢調査，平成 22 年，第 16–1 表）
10. 夫の年齢別人口（国勢調査，平成 22 年，第 17 表）
11. 妻の年齢別人口（国勢調査，平成 22 年，第 17 表）

池田らが用いている統計データは上記 1–9 の 9 種類である．本研究では，「夫の年齢別人口」と「妻の年齢別人口」の統計データを追加している．2 つの統計データを追加した理由は，合成データに存在する夫と妻が法律上結婚できる年齢に達しているかを考慮するためである．先行研究の 9 種類の統計データにはそのような情報が含まれていないので，合成データの中に 17 歳の夫や 15 歳の妻が存在する可能性がある．追加した統計データに合わせることで，合成データに存在する全ての夫と妻が法律上結婚できる年齢を持つことになる．なお，先行研究の論文では男女ごとの統計データを 1 つにまとめて表記していると考えられるため，本稿の統計データの番号とは異なっている．例えば，上記の「男性の年齢別人口」と「女性の年齢別人口」を 1 つにまとめて「年齢別人口分布」と表記している．

それぞれの統計データの形式を Table 1–6 に示す．統計データの形式と Table の対応は次の通りである．

Table 1： 父子年齢差，母子年齢差
Table 2： 夫婦年齢差
Table 3： 男性の年齢別人口，女性の年齢別人口
Table 4： 男性年齢階層別の単独世帯人口，女性年齢階層別の単独世帯人口
Table 5： 男性年齢階層別の夫婦のみ世帯人口，女性年齢階層別の夫婦のみ世帯人口
Table 6： 夫の年齢別人口，妻の年齢別人口

各 Tbale の「割合（%）」の値が，公開されている統計データから計算した値である．それぞれの行ごとで，条件 X に当てはまる市民の数・組の値を分母，条件 Y に当てはまる市民の数・組の値を分子として計算している．表記の都合上，割合の値は小数点第 4 位までと

Table 1: 父子年齢差

条件 X	条件 Y	割合（%）
父子関係	年齢差 –16	0.00
父子関係	年齢差 17–19	0.41
父子関係	年齢差 20–24	6.61
⋮	⋮	⋮
父子関係	年齢差 45–49	2.04
父子関係	年齢差 50–75	0.69
父子関係	年齢差 76–	0.00

Table 2: 夫婦年齢差

条件 X	条件 Y	割合（%）
夫婦関係	年齢差 – -65	0.00
夫婦関係	年齢差-64 – -55	0.00
夫婦関係	年齢差-54 – -45	0.00
⋮	⋮	⋮
夫婦関係	年齢差-24 – -15	0.07
夫婦関係	年齢差-14 – -5	2.57
夫婦関係	年齢差-4	1.26
夫婦関係	年齢差-3	2.00
⋮	⋮	⋮
夫婦関係	年齢差 9	1.63
夫婦関係	年齢差 10	1.08
夫婦関係	年齢差 11 – 20	2.59
夫婦関係	年齢差 21 – 30	0.21
⋮	⋮	⋮
夫婦関係	年齢差 51 – 60	0.00
夫婦関係	年齢差 61 – 66	0.00
夫婦関係	年齢差 67 –	0.00

Table 3: 女性の年齢別人口

条件 X	条件 Y	割合（%）
女性	年齢 0	0.80
女性	年齢 1	0.80
女性	年齢 2	0.82
⋮	⋮	⋮
女性	年齢 98	0.01
女性	年齢 99	0.02
女性	年齢 100	0.01

なっているが，実際には小数点第 12 位まで計算した値を用いている．

父子年齢の 16 歳以下，母子年齢差の 14 歳以下の項目は統計データに記録されていないので，存在しないものとする．夫婦年齢差は，夫婦の年齢別夫婦数の統計データを用いて（夫の年齢）–（妻の年齢）で計算した値である．実際には 1 歳年齢差ごとに夫婦の数を計算することができるが，存在している割合が少ない年齢差の項目を 10 歳年齢差ごとにまとめて集計した．割合が 1%以上の年齢差は 1 つの項目として扱い，それ以外は 10 歳年齢差ごとにまとめて 1 つの項目として扱っている．

男性・女性の年齢別人口は，本研究で用いている 9 種類の家族類型に該当する人数のみで計算した統計データである．その割合を計算した時に，9 種類以外の家族類型の人数は含まれていない．また，Table 3 の一番最後の項目の年齢 100 歳の割合は，実際の統計データの「年齢 100 歳以上」の項目の値である．本研究では市民の年齢の範囲を 0–100 と設定しており，100 歳以上という条件に該当するのは 100 歳の市民のみであるため，Table 3 のように，一番最後の項目を年齢 100 歳と表記している．

Table 4: 男性年齢階層別の単独世帯人口

条件 X	条件 Y	割合（%）
男性・年齢 0– 4	単独世帯	0.00
男性・年齢 5– 9	単独世帯	0.01
男性・年齢 10–14	単独世帯	0.02
⋮	⋮	⋮
男性・年齢 75–79	単独世帯	11.37
男性・年齢 80–84	単独世帯	12.45
男性・年齢 85–	単独世帯	14.80

Table 5: 女性年齢階層別の夫婦のみ世帯人口

条件 X	条件 Y	割合（%）
女性・年齢 0– 4	夫婦のみ世帯	0.00
女性・年齢 5– 9	夫婦のみ世帯	0.00
女性・年齢 10–14	夫婦のみ世帯	0.00
⋮	⋮	⋮
女性・年齢 75–79	夫婦のみ世帯	31.12
女性・年齢 80–84	夫婦のみ世帯	19.59
女性・年齢 85–	夫婦のみ世帯	6.60

Table 6: 夫の年齢別人口

条件 X	条件 Y	割合（%）
夫	年齢 –17	0.00
夫	年齢 18–25	0.88
夫	年齢 26–30	3.74
夫	年齢 31–35	7.24
⋮	⋮	⋮
夫	年齢 71–75	8.28
夫	年齢 76–80	6.19
夫	年齢 81–84	3.03
夫	年齢 85–	1.93

夫・妻の年齢別人口は本研究で追加した統計データである．その役割は，配偶者を持つ市民の年齢別人口を合わせることと，法律上結婚することができない年齢の夫・妻を誤差として目的関数値に計上するためである．夫の年齢別人口は 2 番目の項目のみ年齢 18–25 の 8 歳階級となっている．年齢 18-20，年齢 21-25 のように項目を分けた場合，年齢 18-20 の割合が 0.04%という極めて少ない割合となった．極端に少ない割合の項目が存在すると，合成データの規模によってはその項目の目標値が 0 になる可能性がある．これを避けるために，年齢 18–25 歳を 1 つの項目として扱った．妻の年齢別人口については，1 番目の項目から年齢 15 歳以下，年齢 16–20，年齢 21–25，… で，その割合は，0.00%，0.1%，1.33%，… となっている．

2.3 目的関数

目的関数は合成データと 11 種類の統計データの誤差を計算する．池田らが提案した目的関数を式（1）に示す．

$$f(A) = \sum_{s=1}^{S} \frac{4}{G_s} \sum_{j=1}^{G_s} \big(c_{sj}(A) - m_{sj}(A) \cdot r_{sj}\big)^2 \qquad (1)$$

各変数の説明は次の通りである．

A： 合成データ

S： 統計データの数（$S = 11$）

G_s： 統計データ s の項目数

m_{sj}： 統計データ s の条件 X_{sj} を満たす，合成データの市民の数・組の値

c_{sj}： 統計データ s の条件 X_{sj} と条件 Y_{sj} を満たす，合成データの市民の数・組の値

r_{sj}： 統計データ s の項目 j の割合の値

式（1）の $m_{sj}(A) \cdot r_{sj}$ の値は $c_{sj}(A)$ に対する目標値である．その差を計算した値が，合成データと統計データの誤差を表している．差が 0 になると，

$$\frac{c_{sj}(A)}{m_{sj}(A)} \quad = \quad r_{sj}$$

となり，合成データの割合が統計データの割合と同じ値になる．ただし，r_{sj} の値によっては，$c_{sj}(A)$ が適切な値になっていたとしても $c_{sj}(A) - m_{sj}(A) \cdot r_{sj}$ には最大で 0.5 の誤差が生じる可能性がある．全ての項目で 0.5 の誤差が生じたと仮定すると，誤差 0.5 の 2 乗を合計して項目数 G_s で平均化しているので 0.25 となる．その値に 4 を乗じることで 1.0 となる．したがって，式（1）の値が 1.0 程度になれば，合成データが統計データに対してほぼ正確に一致しているといえる．

先行手法の式（1）の目的関数は，各項目の誤差の合成を統計データの項目数 G_s で平均化している．11 種類の統計データの項目数 G_s は Table 7 のように，大きく異なっている．つまり，項目数の大きい男性・女性の年齢別人口との誤差が小さく見積もられることになり，

Table 7: 統計データの項目数

統計データ	項目数 G_s
父子年齢差	10
母子年齢差	10
夫婦年齢差	29
男性の年齢別人口	101
女性の年齢別人口	101
男性年齢階層別の単独世帯人口	18
女性年齢階層別の単独世帯人口	18
男性年齢階層別の夫婦のみ世帯人口	18
女性年齢階層別の夫婦のみ世帯人口	18
夫の年齢別人口	15
妻の年齢別人口	16

他の統計データが優先的に最適化される．これを避けるために，本研究では式（2）の目的関数 [5] を用いる．

$$f(A) = \sum_{s=1}^{S} \sum_{j=1}^{G_s} \Big| c_{sj}(A) - Round\big(m_{sj}(A) \cdot r_{sj}\big) \Big| \qquad (2)$$

式（1）との変更点は以下の 2 点である．

- 項目数 G_s による平均化と係数 4 を掛ける部分をなくす
- $m_{sj}(A) \cdot r_{sj}$ を整数値に丸めて，$c_{sj}(A)$ との差の絶対値を計算する

式（2）の目的関数は，合成データの値 $c_{sj}(A)$ と統計データの値 $m_{sj}(A) \cdot r_{sj}$ の差を整数値で計算している．したがって，この目的関数の最小値は理論的には 0 であり，最小値になると市民の年齢調整で小さくできる誤差は全て解消できていることを意味する．

しかし，式（2）の目的関数の最小値が必ずしも 0 にならない状況が存在する．これは，Table 8 のような状態が起こり得るためである．Table 8 は，Table 1，2 のように，条件 X がそれぞれの項目で同じになっている形式の統計データを想定した仮の値であり，実際の値ではない．合成データの値 $c_{sj}(A)$ は最適化によって適切な値になっている状態である．この例では，目的関数の最小値は 1 で，市民の年齢変更で $c_{sj}(A)$ を調整しても 0 にはならない．その理由は，目標値の合計（$= 119$）と $m_{sj}(A)$（$= 120$）が異なっているためである．例えば，誤差が生じている 5 行目の $c_{s5}(A) = 6$ を 1 減少させると，1–4 行目のいずれかの $c_{sj}(A)$ に 1 割り当てることになる．すると，目標値と同じ値になっている 1–4 行目の $c_{sj}(A)$ の，いずれか 1 項目を 1 増加することになり，各項目の誤差の合計は 1 のままである．このように，合成データの世帯数によっては式（2）の目的関数の最小値が 0 にならない可能性がある．Table 8 のような状態は，式（1），（2）のどちらの目的関数を用いるかに関係なく起こり得るため，式（2）の目的関数だけに生じる問題ではない．

何らかの誤差を最小化する時には，誤差の二乗を計算することが極めて一般的である．式（2）では，$m_{sj}(A) \cdot r_{sj}$ を整数値に丸めているため，目的関数値が離散的になり，最適化問題の扱いが困難になる可能性が懸念される．しかし，目的関数の $c_{sj}(A)$ と $m_{sj}(A)$ は，各統計データの条件 X, Y を満たす，合成データの市民の数・組を数えた値なので，目的関数の変数そのものが離散的である．合成データの市民の年齢を調整したとしても，$c_{sj}(A)$ の値を 2.4 増加させたり，3.7 減少させるこ

Table 8: 式（2）の最小値が 0 にならない状態

j	$m_{sj}(A)$	r_{sj}	$c_{sj}(A)$	目標値
1	120	0.22	26	26
2	120	0.36	43	43
3	120	0.27	32	32
4	120	0.11	13	13
5	120	0.04	6	5
合計	—	1.00	120	119

目標値 $= Round\big(m_{sj}(A) \cdot r_{sj}\big)$

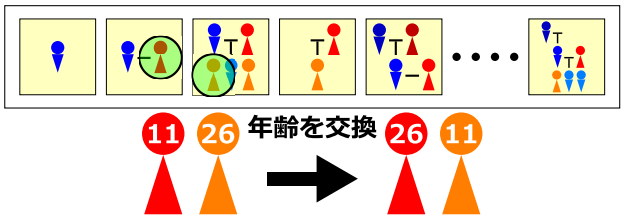

Fig. 3: 市民の年齢交換（提案手法）

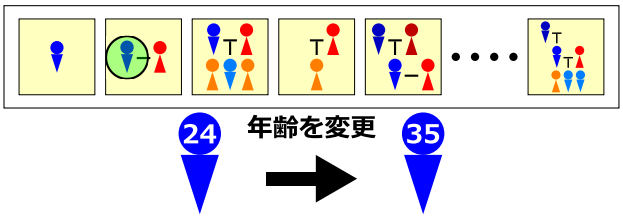

Fig. 4: 市民の年齢変更（先行手法）

とは不可能である．この点から，実数値の $m_{sj}(A) \cdot r_{sj}$ を整数値に丸めても問題ないと考えられるので，本研究では，式（2）の目的関数を用いて最適化を行う．

2.4 最適化処理

目的関数は合成データと統計データの誤差を計算しているので，その値を最小化することで，11 種類の統計データと同じ統計情報を持つ合成データを得ることができる．11 種類の統計データは年齢ごとに項目を分けて集計されている．つまり，市民の年齢を変更することで目的関数値が変化する．本研究では，先行手法と同様に，最適化手法の 1 つである Simulated Annealing (SA) を用いて最適化する．最適化の過程は次の通りである．

Step 1. 初期解生成

Step 2. 終了判定

Step 3. 近傍解生成

Step 4. 解の遷移判定

Step 5. 更新処理

Step 6. Step 2. に戻る

初期解生成は合成データの世帯数 H を設定して，年齢と性別を持った市民をそれぞれの世帯に生成する．合成データの構築には，9 種類の家族類型別の世帯数，総人口，男女比，1 世帯に割り当てる子供の人数などの様々な要素を設定する必要がある．公開されている統計データから計算できる情報については，その割合に従うように全ての要素を設定する．この処理の具体的な内容は文献 [3] では明確にされていなかったため，本稿で詳細に記述する．なお，提案手法を適用するために必要な初期解の生成方法であるため，第 3 章の提案手法で記述する．

終了判定は，SA の規定の探索回数で行う．本研究では，目的関数を用いて解を評価した回数を探索回数と定義した．近傍解生成は，合成データの市民の年齢を変更して新しい解を生成する．解の遷移判定とは，SA の最適化アルゴリズムの特徴である確率的な解の遷移である．受理判定を行う時の確率は Metropolis 基準に従って計算する．更新処理では，探索回数や温度の値などの各パラメータを更新する．これらの処理を繰り返し，規定の探索回数に達した時点で最適化を終了する．

なお，最適化の時，解の評価値は式（2）の目的関数に重み付けをして計算する．重み付けは，統計データ s の r_{sj} の値が 0.0 の項目の誤差を 10^{10} 倍する．これにより，他の項目の誤差よりも重み付けをした項目の誤差を最優先に最小化することになる．対象になる項目は，Table 1 の年齢差 16 以下の項目，Table 5 の女性・年齢 0–4 の項目，Table 6 の年齢 17 歳以下の項目，などである．このような設定を行う理由は，r_{sj} が 0.0 の項目が，その項目に該当する市民が合成データの中に存在してはいけないことを表しているためである．例えば，Table 6 年齢 17 歳以下の項目に該当する市民が存在した場合，法律上結婚できない年齢の夫が存在していることを意味する．そのような解に遷移することを避けるために，重み付けによって対象の項目の誤差を優先的に最適化する．また，Table 8 のように，$c_{sj}(A)$ が適切な値に最適化されていたとしても，いずれかの項目に誤差が生じる可能性がある．その時に，r_{sj} の値が 0.0 の項目に誤差が残らないようにすることも，重み付けの理由の 1 つである．この重み付けは最適化における評価値の計算時にのみ実行するもので，後述する実験結果の時には重み付けされていない評価値で解を評価する．

3 提案手法

本研究の提案手法は，最適化の探索における近傍解生成の改良である．これにより，合成データの市民の年齢の組み合わせを効率的に探索することができる．ただし，この効果を最大限に活かすためには，初期解生成の時点で年齢以外の統計情報を合わせておく必要がある．初期解生成と近傍解生成の 2 つを同時に組み合わせることで効率的な探索を可能にし，先行手法よりも誤差を最小化できることを示す．まず，近傍解生成の処理を明らかにした後で，初期解生成の具体的な手順について説明する．

3.1 近傍解生成

提案手法では，Fig. 3 のように合成データに存在する同性の市民を 2 人ランダムに選択し，それぞれの年齢を交換することで近傍解を生成する．先行研究では，Fig. 4 のように，合成データの 1 人の市民をランダムに選択し，年齢を変更している．

提案手法の大きな特徴は，合成データに存在する男性・女性の年齢別人口が変化しないということである．つまり，初期解生成の時点で，男性・女性の年齢別人口の統計データに一致するよう市民の年齢を設定した後，同性 2 人の市民の年齢交換で近傍解を生成すると，2 つの統計データに一致している状態を保持しつつ，残りの統計データとの誤差が小さくなる解の探索を行うことができる．したがって，提案した近傍解生成を行う時は，初期解生成後で最適化の処理に入る前に，男性・女性の年齢別人口の統計データに合うように市民の年齢を設定する．性別ごとに人数分の年齢を用意して，乱数を用いて年齢の順番をランダムに入れ替えた後，一括で年齢を設定する．

3.2 初期解生成の処理

初期解生成とは合成データを構築する時の処理である．合成データを構成する総世帯数 H を設定した後，公開されている統計データの割合に基づいてさまざま

な要素を決定する．設定する要素の順番は以下の通りである．

Step 1. 家族類型別の世帯数

Step 2. 子供の総人数

Step 3. 家族類型別の子供の人数

Step 4. 1 世帯に割り当てる子供の人数

Step 5. 市民の性別

Step 6. 市民の年齢

本研究では合成データの総世帯数を 5,400 と設定して，全ての実験を行っている．この設定に基づいて，具体的な値を挙げて説明する．なお，統計データ割合の値を用いて計算した実数値を，四捨五入して整数値に丸めている．

Step 1. 家族類型別の世帯数

合成データの総世帯数を 5,400 に設定した後，9 種類の家族類型別の世帯数を統計データの割合に従って決定する．計算した値を Table 9 に示す．9 種類の家族類型には子供が含まれているものが存在する．例えば「夫婦と子供」の場合,「夫婦と子供 1 人」,「夫婦と子供 2 人」，⋯ のように，1 世帯の子供の人数別世帯数を設定する必要がある．しかし，公開されている統計データにはそれらが集計されていないので，複数の統計データを組み合わせてその値を推定する必要がある．

Step 2. 子供の総人数

まず，合成データに存在する子供の総人数を決定する．子供の総人数は次式で計算することができる．

$$(\text{子供の総人数}) = (\text{総人口}) - (\text{子供以外の総人数})$$

世帯当たりの平均構成員人数 2.3725 に従うと，合成データの総人口は 12,812 人（$5,400 \times 2.3725 = 12811.5$）となる．子供以外の総人数は，9 種類の家族類型の規定人数から計算することができる．規定人数とは，家族類型の特徴から存在が確定している市民の人数であり，Table 10 の通りである．家族類型別の世帯数はすでに決まっているので，規定人数を掛けた値の合計が子供以外の総人数となる．これにより，子供の総人数が 3,790 人（$12,812 - 9,022$）となる．

Step 3. 家族類型別の子供の人数

次に，求めた子供の総人数を，統計データの割合に従ってそれぞれの家族類型に分配する．統計データには 9 種類の家族類型別総人数が集計されているので，Table 10 の規定人数を用いた同様の計算で割合を計算することができる．統計データの割合と分配した子供の人数を Table 11 に示す．

Table 9: 家族類型別の世帯数

	家族類型	割合（%）	世帯数
1	単独	32.98	1781
2	夫婦のみ	21.06	1137
3	夫婦と子供	29.68	1603
4	父親と子供	1.37	74
5	母親と子供	7.93	428
6	夫婦と両親	0.48	26
7	夫婦と片親	1.50	81
8	夫婦と子供と両親	1.89	102
9	夫婦と子供と片親	3.12	168
	合計	100.00	5400

Table 10: 家族類型別の規定人数

家族類型	役割	規定人数
1	単身	1 人
2	夫，妻	2 人
3	夫，妻,（子）	2 人
4	父,（子）	1 人
5	母,（子）	1 人
6	父，母，夫，妻	4 人
7	片親，夫，妻	3 人
8	父，母，夫，妻,（子）	4 人
9	片親，夫，妻,（子）	3 人

Table 11: 家族類型別の子供の分配割合と人数

家族類型	割合（%）	人数
1	0.00	0
2	0.00	0
3	69.13	2620
4	2.55	97
5	15.70	595
6	0.00	0
7	0.00	0
8	5.14	195
9	7.47	283
合計	100.00	3790

Step 4. 1 世帯に割り当てる子供の人数

そして，家族類型に分配する子供の人数を決定した後，子供の人数別世帯数を決定する．子供の人数別世帯数とは，前述した「夫婦と子供」の例のように「夫婦と子供 1 人」,「夫婦と子供 2 人」，⋯ に対する世帯数である．この値は,「家族類型と世帯人員別の世帯数」と,「家族類型別の子供の総人数」の 2 つの統計データから計算した割合に従って決定する.「家族類型と世帯人員別の世帯数」は 9 種類の家族類型と世帯人員数別（1 人，2 人，⋯，6 人，7 人以上）の世帯数を集計している．この統計データの世帯人員数別の項目は子供の人数別の項目としても扱うことができる．例えば，夫婦と子供世帯で世帯人員数が 3 人，4 人，5 人，⋯，という分類は，子供の人数が，1 人，2 人，3 人，⋯，であることを表している．これにより家族類型と子供の人数別世帯数割合を計算することができる．

この時，統計データの世帯人員数の 7 人以上の項目を世帯人員数 7 人と想定して子供の総人数を計算すると，Table 11 の割合を計算する時に集計した家族類型別の子供の総人数よりも少ない値となった．これは 8 人以上の世帯人員の世帯が 7 人以上として 1 つの項目にまとめられていることが原因である．そこで，世帯人員が 7 人以上の項目を 7 人と 8 人に分けて，家族類型別の子供の総人数が，Table 11 で集計した子供の総人数と一致するように，世帯人員数が 8 人の世帯を追加した．現実には世帯人員 9 人以上の世帯も存在しているはずだが，統計データ上の根拠がある調整方法が存在

Table 12: 家族類型と子供の人数別の世帯数割合（%）

	子供の人数				
家族類型	1 人	2 人	…	7 人	合計
1	—	—	…	—	—
2	—	—	…	—	—
3	49.09	39.96	…	0.00	100.00
4	73.96	21.65	…	0.01	100.00
5	68.35	25.36	…	0.01	100.00
6	—	—	…	—	—
7	—	—	…	—	—
8	32.05	47.46	…	0.00	100.00
9	47.87	37.96	…	0.00	100.00

しないため，本研究では世帯人員の上限を 8 人と設定した.「家族類型と世帯人員別の世帯数」から計算した「家族類型と子供の人数別の世帯数割合」を Table 12 に示す.

Table 12 の割合に従って，合成データの家族類型と子供の人数別の世帯数を決定する．その割合と正確に一致するように世帯数を決定すると,「夫婦と子供と片親」の合計世帯数が 167 世帯となり，あらかじめ設定した Table 9 の世帯数 168 よりも 1 世帯少ない値となった．Table 12 の割合には 2 つの統計データから推定した値が含まれているが，Table 9 の割合は純粋に統計データのみの値に基づいて計算しており，信頼できる数値であるため Table 9 の 168 世帯に合わせることを優先した．子供の人数が 1–5 人から一様乱数を用いてランダムに 1 つ選択して，選択した子供の人数の世帯数を 1 世帯増加して 168 世帯に調整した．

また，決定した世帯数で合成データの家族類型別の子供の総人数を計算すると，Tabel 11 で決定した 3,790 人よりも 10 人の範囲内で少ない人数となった．これは，Table 12 を計算する時に，世帯人員数の上限を 8 人と設定して世帯人員数が 9 人以上の世帯は存在しないものとしたためである．Table 11 の割合は信頼できる数値であるため，Table 11 の合計人数に一致することを優先した．調整方法は，家族類型別の世帯数が Table 9 から変化しないように，子供の人数別世帯数を増減させた．どの子供の人数の世帯を増減させるかは，一様乱数を用いてランダムに選択した．以上の処理で世帯の生成処理が完了する．

Step 5. 市民の性別

最後に，市民の年齢と性別を設定する．初期の年齢設定には性別の情報を用いるため，先に性別を設定する．夫婦のみ世帯の夫と妻のように，市民が持っている役割から性別が明らかである市民には，その役割に従って性別を設定する．一方，単独世帯に所属する市民と子供・片親の役割を持つ市民は，家族類型や役割の情報からは性別を決定することができない．そこで，公開されている統計データから必要となる男女比（男性を 1 とした時の女性の比率）の割合を計算して，その割合に合うように男女の人数を決定して性別を設定する．

男女比は「家族類型別の男女の合計人数」の統計データを用いて割合を計算する．計算した男女比を Table 13 に示す．全体の男女比は，男女ごとの合計人数から計算した．子供の男女比は，Table 10 の規定人数と家族

Table 13: 家族類型別の全体と子供の男女比

家族類型	全体	子供
1	0.9209	—
2	1.0000	—
3	0.9566	0.9059
4	0.3082	0.7107
5	2.0896	0.7965
6	1.0000	—
7	1.6693	—
8	0.9822	0.9459
9	1.2969	計算不可

類型別の世帯数を用いて計算することができる．例えば,「夫婦と子供」世帯に存在する子供の男性の合計人数の場合は,

$$(\text{男性の合計人数}) - (\text{世帯数}) \cdot 1$$

で求めることができる.「夫婦と子供」世帯は各世帯に夫が 1 人存在しているので，世帯数と同じ数だけ夫が存在していることになる．つまり，男性の合計人数から夫の合計人数を引いた値が子供の男性の合計人数となる．女性についても同様の計算で，子供の女性の合計人数を求めることができる．

ただし,「夫婦と子供と片親」世帯については，性別が不明な市民が子供と片親の 2 人分存在しているので，子供の男女比を計算することができない．そこで本研究では，「夫婦と片親」世帯の片親の市民の男女比 7.0713 を用いて，「夫婦と子供と片親」世帯の片親の男女人数を決定した．そして，Table 13 の「夫婦と子供と片親」の全体の男女比 1.2969 に一致するように子供の男女人数を決定した.「夫婦と片親」世帯の片親の市民の男女比 7.0713 は，女性の比率が高いことが特徴的である．このような特徴的な男女比が，片親の役割を持つ市民に共通しているものと考えて同じ値を適用した．その際，子供の男女比は 0.8377 となり，夫婦の子供の男女比としては少し低い値となったが，全体の男女比は統計データの割合に一致しているため問題はないと考えられる．

子供に性別を設定する時，人数分の性別の情報を用意して順番をランダムに入れ替えた後，一括で設定している．1 世帯に存在する子供の男女配分についての統計情報は分からないので，考慮していない．子供の男女配分とは，例えば 2 人の子供が存在する世帯の時,（男・男）,（男・女）,（女・女）の組み合わせがどのような配分になっているかを意味する．

Step 6. 市民の年齢

性別の設定が完了した後，市民の年齢を設定する．市民の年齢は，年齢別人口の統計データの割合に従って，乱数を用いて確率的に設定する．年齢の範囲は 0–100 としている．男性の市民には男性の年齢別人口，女性の市民には女性の年齢別人口の統計データの割合を用いる．以上の処理をもって合成データの初期解生成の処理が完了する．なお，提案手法の近傍解生成を行う時は，前述したように，この初期解生成の後で最適化の処理に入る前に，男性・女性の年齢別人口の統計データに合うよう年齢を設定しなければならない．

Table 14: 実験時のパラメータ

基礎パラメータ	
世帯数	5,400
探索回数	2 通り
近傍解生成	年齢変更 / 年齢交換
SA の温度設定	
初期温度	1.0
収束温度	0.1
冷却スケジュール	指数冷却
冷却タイミング	探索毎

Table 15: 探索回数 1,281 万 2 千回の結果

	平均値	標準偏差
先行手法	57.23	6.61
提案手法	19.60	3.42

Table 16: 探索回数 1 億 2,812 万回の結果

	平均値	標準偏差
先行手法	21.13	4.43
提案手法	6.33	0.76

4 比較実験

本研究の提案手法の有効性を確かめるために，先行手法と提案手法の 2 つの設定で実験を行う．それぞれの手法で近傍解生成の処理が異なっている．先行手法は 1 人の市民の年齢変更で，提案手法は同性 2 人の市民の年齢交換である．3.2 節の初期解生成は，両方の手法で同じ処理を実行している．ただし提案手法では，初期解生成の後で最適化の処理に入る前に，男性・女性の年齢別人口の統計データに合うように市民の年齢を再設定している．全ての実験は，最適化手法に SA を用いて，式（2）の目的関数で最適化を行った．合成データの世帯数は 5,400 世帯で，総人口は 12,812 人である．探索回数は総人口の 1,000 倍の 1,281 万 2 千回と，その 10 倍の 1 億 2,812 万回の 2 通りで，それぞれ 30 試行の最適化を行った．その他のパラメータを含む実験時の設定を Table 14 に示す．目的関数値の結果を比較した後で，計算時間についても考察する．

4.1 探索回数：1,281 万 2 千回（Table 15）

30 試行分の結果を Table 15 に示す．目的関数値は合わせる統計データと合成データの誤差を計算しているので，その値が小さいほど優れた結果であることを示す．平均値と標準偏差の両方で，提案手法の方が小さな値となっていることから，提案した年齢交換による近傍解生成が高い精度で誤差をより小さくできることがわかる．先行手法と提案手法について t 検定を行ったところ，有意水準 1%で有意な差がみられた．しかしながら，提案手法の 30 試行の結果で，目的関数値が最小値になっている合成データは 1 つも得られなかった．そこで，探索回数を 10 倍にして同様の実験を行った．

4.2 探索回数：1 億 2,812 万回（Table 16）

30 試行分の結果を Table 16 に示す．先行手法と提案手法の両方で，平均値と標準偏差が Table 15 よりも小さくなっているので，探索回数を増加させることが誤差最小化に有効であることがわかる．2 つの手法で，目的関数値が最小値にできているかを確認したところ，提案手法は 30 試行の内の 25 試行で誤差が最小値になっていた．一方，先行手法では誤差が最小値になっている解は 1 つも得られていなかった．なお，この実験で設定した合成データが 5,400 世帯では，目的関数値の最小値は 6 である．市民の年齢の組み合わせの全通りを探索したとしても，目的関数値が 6 以下になることはない．25 試行の中の 1 つの合成データに対して，式（1），(2）の目的関数で解を評価した値を Table 17 に示す．池田らが提案した式（1）の目的関数値が，全ての統計データで 1.0 以下になっていることがわかる．

Table 17: 誤差が最小値の 6 の合成データ

s	式（1）の値	式（2）の値
1	0.32	0
2	0.23	0
3	0.24	1
4	0.35	0
5	0.33	0
6	0.36	0
7	0.37	0
8	0.66	2
9	0.85	2
10	0.53	0
11	0.64	1
合計	4.87	6

また，提案手法によって得られた，誤差が最小値になっていた 25 試行の解の中に，全ての要素において同一の解が存在するかを調べたところ，そのような解の組は 1 つも存在しなかった．これにより，目的関数に組み込まれている 11 種類の統計データとの誤差が最小値の解は，1 つではなく複数存在していることがわかる．

4.3 計算時間

最後に，2 通りの探索回数の実験における計算時間を Table 18，19 に示す．実行環境は以下の通りである．

- CPU：Intel Core i7-4770 3.4GHz
- メモリ：16GB（8GB × 2）
- OS：Microsoft Windows 8.1 64 ビット

それぞれの Table の値は，30 回試行の計算時間の平均値と標準偏差である．両方の探索回数で，先行手法よりも提案手法の方が計算時間が長い結果となった．これは，提案手法の近傍解生成が 2 人の市民の年齢を変更しているためである．先行手法における 1 人分の処理を，提案手法では 2 人の市民に対して処理を行うこ

Table 18: 探索回数 1,281 万 2 千回の計算時間（秒）

	平均値	標準偏差
先行手法	31.78	1.26
提案手法	44.46	0.38

Table 19: 探索回数 1 億 2,812 万回の計算時間（秒）

	平均値	標準偏差
先行手法	317.69	6.91
提案手法	462.40	16.55

とになる．年齢変更の市民の人数に影響する処理は，年齢変更の対象になる市民を選択したり，新しい年齢に変更した市民の情報を目的関数の $m_{sj}(A)$ と $c_{sj}(A)$ に反映させる処理などである．提案手法では，年齢を変更する市民の人数が先行手法よりも多いため，計算時間の増加は避けることができない．

同じ探索回数で計算時間が異なる点を考慮すると，Table 15，16 の実験時のパラメータが提案手法に有利な設定になっている可能性がある．そこで，先行手法の 1 試行の計算時間が約 460 秒になるように探索回数を調整して追加実験を行った．探索回数を 2 億 553 万回に設定した先行手法の実験結果を Table 20 に示す．計算時間の平均値が約 474 秒となっているので，計算時間の観点で，Table 16 の提案手法と公平な実験設定といえる．先行手法同士で比較すると，探索回数の増加によって目的関数値の平均値が，21.13 から 17.10 まで改善していることが確認できる．しかし，提案手法の平均値 6.33 と比較すると，先行手法の誤差最小化が十分でないことがわかる．したがって，提案手法の方が計算時間当たりの探索効率が高いことになる．

以上の結果から，先行手法よりも提案手法の方が本研究の最適化問題に有効であることがわかる．同じ探索回数の下では，提案手法の計算時間が約 1.5 倍になっているが，誤差最小化の性能が高い結果となった．さらに，同じ計算時間の下で比較をするための追加実験により，提案手法の有効性を示すことができた．

Table 20: 先行手法で探索回数 2 億 553 万回

	平均値	標準偏差
計算時間（秒）	473.92	2.33
目的関数値	17.10	3.57

5　まとめ

本稿では，池田らが提案した，統計データを用いた市民属性の合成手法について，効率的な近傍解生成を提案し，複数回試行の実験で，提案手法の有効性を示した．探索回数を合成データの総人口の 1 万倍に設定することで，30 回試行の内の 25 試行で統計データとの誤差が最小値の合成データを得ることができた．また，誤差が最小値の 25 試行の中に同一の解の組が 1 つも存在しなかったことにより，求めるべき解（誤差が最小値の解）が複数存在していることを確認した．今後の課題としては，大規模な世帯数の合成データをつくる時に，本稿の提案手法が有効であるかを検証する必要がある．

社会シミュレーションでは，人の意思決定や環境に起因する不確実性を乱数を用いて表現することが一般的なので，乱数シードを変更して複数回試行の実験を行った上で，結果を検証することが多い．モデルで表現された人に設定するデータが複数存在している場合，それぞれの組み合わせで同様の実験を行う必要性が生じる．設定するデータと乱数シードの組み合わせで，複数回試行の実験が負担になることを避けるため，本研究の市民属性の合成手法で求めるべき解が，比較的少ない数に定まるように最適化問題を設計しなければならない．具体的な方法は，目的関数に組み込む統計データを新たに追加するか，既存の統計データの年齢階級の区分を細かくして項目数を増加させることなどが挙げられる．これにより，最適化問題の難易度は高くなるが，求めるべき解を少ない数に定めることができる．

参考文献

1) 市川学，出口弘：感染症実用シミュレーションにおける仮想都市環境構築法の違いによる結果への影響分析—日常生活スポット内包セル型仮想都市モデルの必要性—，計測自動制御学会論文集，**49**-11，1012/1019（2013）
2) 花岡和聖：動態的な空間的マイクロシミュレーションモデルを用いた社会シミュレーション—京町家の取壊し分析を事例に—，地学雑誌，**118**-4，646/664（2009）
3) 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，計測自動制御学会第 43 回システム工学部研究会，11/14（2010）
4) 福田純也，喜多一：エージェントベースの人口推計モデルにおける属性決定手法の評価，システム制御情報学会論文誌，**27**-7，279/289（2014）
5) 枡井大貴，村田忠彦：SA を用いた統計データからのエージェント属性復元のための目的関数の影響，計測自動制御学会第 5 回社会システム部会研究会，121/126（2014）
6) 枡井大貴，村田忠彦：統計データとの誤差最小化のための SA によるエージェント属性復元，計測自動制御学会第 7 回社会システム部会研究会，47/52（2014）
7) 枡井大貴，村田忠彦：統計データを用いたエージェント属性生成における誤差最小化のための進化計算手法，進化計算学会第 8 回進化計算シンポジウム 2014 講演論文集，196/203（2014）
8) 枡井大貴，村田忠彦：エージェント属性復元における Simulated Annealing を用いた世帯構成の最適化，計測自動制御学会第 8 回社会システム部会研究会，167/172（2015）
9) 総務省統計局 `http://www.e-stat.go.jp/`