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

○枡井大貴　村田忠彦 （関西大学）

Optimizing Household Composition by Simulated Annealing in Reconstructing Agent's Attributes

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

概要– 近年，社会科学の観点から現実社会を捉えようとする研究課題の一つとして社会シミュレーションが注目されている．シミュレーションモデルの構築には現実社会に存在する人々の年齢や性別などの個人情報が必要となる．しかし，個人情報を利用することは政府によって厳しく規制されているため，容易に入手することはできない．そこで，池田らが提案した実統計データから人の属性を復元する手法を用いることで，実統計データに適合するような人の集合を復元することができる．池田らの手法を改良した進化計算による手法がすでに提案されているが，池田らの手法に比べて計算時間が長い問題があった．本研究では，進化計算を用いた手法を Simulated Annealing を用いて再設計し，計算時間の短縮を試みる．また，手法の精度について比較を行いその結果について考察する．

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

1　はじめに

近年，社会科学の観点から現実社会を捉えようとする研究課題の一つとして社会シミュレーションが注目されている [1]．その中でも，人々の行動が引き起こす様々な問題に対する問題解決のアプローチでは，モデル構築の自由度が高いエージェントベースモデルが有効である．シミュレーション上には抽象化された社会の姿を再現することが必要になるが，シミュレーションの信頼性を考えると再現したモデルは現実社会の状態に近づけなければならない．社会の姿をできるだけ精密に記述することは，社会科学の観点から至上命題の一つともされている [2]．つまり，モデルの中に人を表現したエージェントが存在しているならば，そのエージェントは現実社会の人と同様に年齢と性別を持っていることが望ましい．しかしながら，人の年齢や性別や世帯の情報などの個人情報は一般には公開されておらず，それらを入手して利用することは政府によって厳しく規制されている．

そこで，政府が行った国勢調査の人口統計データからその母集団の情報を復元する手法が池田らによって提案されている [3]．母集団は現実社会に存在する人々そのものなので，復元したデータ（以下，復元データと呼ぶ）は年齢や性別を持ったエージェントの集合で形成されている．池田らが提案した手法では，統計データと復元データの誤差を計算する目的関数を設計し，その値を最小化することで統計データとの誤差が小さい復元データを得ることができる．池田らは目的関数値を最小化することを目的に Simulated Annealing（SA）を用いて最適化を行っている．福田ら [4] はその復元手法を人口動態推計の研究に利用している．また，市川らは人口などのデータを含んだモデル設計の際に，そのモデル構築に多くの時間が費やされてしまう問題点を指摘している [5]．その問題点の解決のために仮想都市環境構築システムの構想を提案し，システムには池田らの復元手法を利用することを想定している．

著者らは池田らが提案した復元手法の誤差最小化における精度改良を検討してきた．まず，池田らが提案した手法の目的関数は，項目数が大きい統計データとの誤差が他の統計データに比べて残りやすい性質を持っていた．福田らは各統計データに対する目的関数値に重み付けをして調整を行っている [4]．そこで，既存の目的関数の構造を少し変化させて重み付けをせずに対象となる全ての統計データに対して平均的に誤差を最小化できる目的関数を提案した [6]．また，その目的関数は統計データとの誤差をエージェントの人数，組の単位で示しているので誤差を直観的に理解することができる．次に提案した目的関数を用いた SA による最適化で，探索時の解の変更処理を改良した効率的な探索手法（以下，改良型 SA）を提案した．加えて，探索回数を段階的に増やすことで統計データとの誤差が実質的に 0 の復元データを得ることができた [7]．得られた誤差が 0 の結果をもとに目的関数値が小さくなる要因について分析を行ったところ，復元データの初期生成時の世帯構成が大きく影響していることがわかった [8]．既存の SA による探索では世帯構成を変化させることができないので，世帯の交叉によって世帯構成の変更が可能な進化計算を用いた手法を提案した．進化計算手法の最適化で目的関数値がより最小化できるような状態の世帯構成を形成し，その後で改良型 SA による最適化を行うことで，既存の SA のみの最適化を行うよりも高い精度で目的関数値を最小化することができた [9]．

進化計算を用いた手法 [8, 9] は解を個体して扱い，複数の個体を最適化することになる．解となる復元データはメモリを確保して生成するため，複数の復元データの生成にはより多くのメモリが必要になる．復元データの規模を大きくした時に 1 個体当たりに必要なメモリが大きくなるため，個体の設定値が制限される可能性がある．また，進化計算による手法 [8, 9] は池田らの手法に比べて計算時間が長くなっていた．本稿では，世帯構成を変化させる進化計算を用いた手法を SA を用いて再設計した手法を提案する．提案した手法で実験を行い，計算時間と復元手法の精度について比較を行った．

PG0004/15/0000-0167 © 2015 SICE

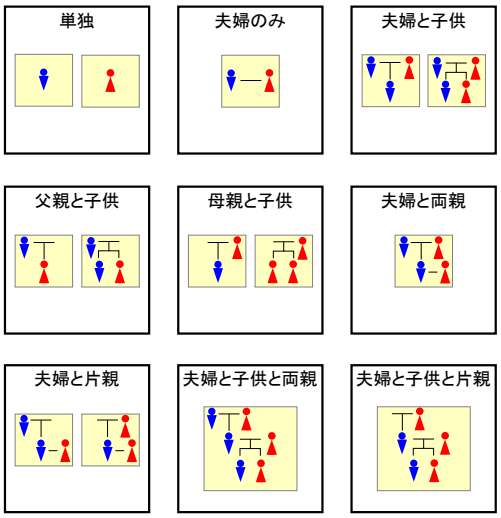

Fig. 1: 世帯の種類

2 復元手法

本研究では池田らが提案した統計データから人の属性を復元する手法について，その精度を向上させた新たな手法を提案した．まず，池田らが提案した基本的な復元手法について，次に著者らがこれまで提案した手法について記述する．その後，本稿で新たに提案する手法について記述する．統計データは基本的に 2010 年の状態を表すデータを用いるが，一部の統計データに関しては同じ年のデータが見つからなかったため近い年のデータで代用する．

2.1 復元データのモデル

復元データは複数の世帯によって構成されている．世帯の種類は，単独，夫婦のみ，夫婦と子供，父親と子供，母親と両親，夫婦と片親，夫婦と片親，夫婦と子供と両親，夫婦と子供と片親の 9 種類を扱う．世帯の種類を Fig. 1 に示す．また，世帯の中には人を表すエージェントが存在しており，年齢，性別，世帯の種類，世帯の役割，親族関係の属性を持っている．親族関係とは自分の父，母，夫，妻，子に該当するエージェントが復元データの中のどのエージェントであるかという情報である．本研究で扱う 9 種類の世帯は統計データ[10](表 7-10，2012 年) に記載されているもので，これらの世帯で全世帯数の 95%を占めている．統計データには他の種類の世帯も存在しているが数が少ないので存在しないものとする．復元データの生成は規定数 H の数だけ世帯を生成することによって行う．この時，統計データの割合に基づいて 9 種類の内のいずれかの世帯を生成する．9 種類の世帯の割合を Table 1 に示す．

また，9 種類の世帯の中で，夫婦と子供，父親と子供，母親と子供，夫婦と子供と両親，夫婦と子供と片親には子供が存在している．しかし，世帯の種類の統計データには子供の数についての分類はないので，政府が行った調査結果のデータ [11, 12] を元に各世帯に存在する子供の数を決定する．子供の数の割合を Table 2 に示す．子供が存在する世帯を生成する時は Table 2 の割合に基づいて子供の数を設定する．夫婦と子供，夫婦と子供と両親，夫婦と子供と片親，の世帯の時は夫婦世帯の割合を用いる．父親と子供の世帯の時は父子世帯を，母親と子供の時は母子世帯の割合を用いる．

Table 1: 9 種類の世帯の割合

世帯の種類	割合 (%)
単独	33.98
夫婦のみ	20.74
夫婦と子供	29.24
父親と子供	1.34
母親と子供	7.81
夫婦と両親	0.47
夫婦と片親	1.48
夫婦と子供と両親	1.86
夫婦と子供と片親	3.07
合計	99.99

Table 2: 世帯の子供の数

	子供の数の割合（%）			
	1 人	2 人	3 人	4 人
夫婦世帯	16.97	59.98	20.70	2.35
父子世帯	54.70	36.00	8.20	1.10
母子世帯	54.70	34.50	8.90	1.90

エージェントの属性については，夫，妻，父，母のいづれかに該当するエージェントは性別を決定し，単独世帯のエージェントや子供のエージェントは性別をランダムに決定する．年齢は人口ピラミッドの統計データ[10]（表 2-3，2012 年）の割合に基づいて設定する．初期生成された復元データの各エージェントは乱数を用いて年齢を設定しているので，現実には存在しないような状態になっている可能性が十分に考えられる．例えば，父子関係にあたるエージェントで父親エージェントよりも子供エージェント方が年上になっていることなどである．そこで，親子や夫婦の年齢関係や人口分布などの統計データに適合させることで，現実に存在するような年齢関係の状態に調整し，統計データの分布に合うような復元データを得ることができる．

2.2 適合させる統計データ

復元データに適合させるための統計データは池田らと同様の以下の 9 種類を用いる．

1. 父子の年齢差（表 4-13，2013 年 [10]）
2. 母子の年齢差（表 4-8，2012 年 [10]）
3. 夫婦の年齢差（表 9-14，2011 年 [13]）
4. 男性の人口分布（表 2-3，2012 年 [10]）
5. 女性の人口分布（表 2-3，2012 年 [10]）
6. ある年齢の男性が単独世帯に属する割合（表 7-28，2013 年 [10]）
7. ある年齢の女性が単独世帯に属する割合（表 7-28，2013 年 [10]）
8. ある年齢の男性が夫婦のみ世帯に属する割合（表 7-28，2013 年 [10]）
9. ある年齢の女性が夫婦のみ世帯に属する割合（表 7-28，2013 年 [10]）

統計データ 1，2 のデータ形式を Table 3 に，統計データ 3 のデータ形式を Table 4 示す．統計データ 4，5 のデータ形式を Table 5 に，統計データ 6-9 のデータ

形式を Table 6 に示す．それぞれの Table の割合は実際の統計データの値であり，条件 X を満たすものの内で条件 Y を満たす数の割合を示している．例えば Table 3 の 2 項目目では，父子関係の人の組の中で年齢差 15〜19 に該当する割合が 0.5469(%) ということを表している．統計データ 1-5 は各項目の条件 X が等しいので割合を計算するときの分母は各項目で同じ値になっている．したがって，統計データ 1-5 では割合の合計はおよそ 100(%) になる．一方，統計データ 6-9 の形式を示す Table 6 では各項目で条件 X が異なっているため，割合を計算するときの分母が各項目で異なっている．したがって，統計データ 6-9 では各項目の割合の合計は 100(%) にはならない．

2.3 目的関数

池田らが提案した目的関数は復元データと適合させる統計データとの誤差を計算している．本研究では，池田らの目的関数の構造を少し変化させた別の目的関数を用いる．まず，池田らの提案した式（1）目的関数と変数を以下に示す．

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

A：復元データ

S：統計データの数（$S = 9$）

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，復元データ内の市民や組み合わせの数

c_{sj}：統計データ s の条件 X_{sj} と条件 Y_{sj} を満たす，復元データ内の市民や組み合わせの数

r_{sj}：統計データ s の項目 j の割合の値

式（1）の c_{sj} は復元データの値で，$r_{sj}\cdot m_{sj}(A)$ は統計データの割合から計算した目標となる値である．したがって，$c_{sj} - r_{sj}\cdot m_{sj}$ の式から復元データと統計データの誤差を計算することができる．この時，r_{sj} は統計データの割合の数値であるので $r_{sj}\cdot m_{sj}$ は実数となる．それに対して復元データの c_{sj} は整数であるので，c_{sj} が適切な値になっていたとしても $c_{sj} - r_{sj}\cdot m_{sj}$ の値は最大で 0.5 になる可能性がある．その値の二乗和を統計データの項目数 G_s で平均化しているので 0.25 となる．最期に 4.0 を掛けることで 1.0 になる．したがって，式（1）が 1.0 程度になれば復元データが統計データに適合していると考えられる．この式（1）は統計データの項目数 G_s で平均化を行っている．統計データ 3，4 の項目数は 0〜100 歳区分の 101 項目となっており，他の統計データに比べて大きな値である．したがって，統計データの項目数が大きいほど誤差の値が小さく見積もられることになる．これに関して福田らは目的関数値に重み付けを行っている [4]．

本研究では式（1）の構造を少し変化させた式（2）の目的関数 [6] を用いる．式（1）と異なるのは，実数の $r_{sj}\cdot m_{sj}$ を $Round$ 関数によって四捨五入して整数に丸めて，c_{sj} との差の絶対値を計算している点である．

Table 3: 父子の年齢差

条件 X	条件 Y	割合 (%)
父子関係	年齢差　〜14	0.0000
父子関係	年齢差 15〜19	0.5469
父子関係	年齢差 20〜24	8.4852
...	...	...
父子関係	年齢差 40〜44	7.9119
父子関係	年齢差 45〜49	2.1035
父子関係	年齢差 50〜	0.1107

Table 4: 夫婦の年齢差

条件 X	条件 Y	割合 (%)
夫婦関係	年齢差〜-4	6.1253
夫婦関係	年齢差　-3	3.1166
夫婦関係	年齢差　-2	4.8194
...	...	...
夫婦関係	年齢差　5	4.7240
夫婦関係	年齢差　6	3.5213
夫婦関係	年齢差　7〜	1.0450

Table 5: 男性の人口分布

条件 X	条件 Y	割合 (%)
男性	年齢　0	0.8669
男性	年齢　1	0.8660
男性	年齢　2	0.8900
...	...	...
男性	年齢　98	0.0110
男性	年齢　99	0.0065
男性	年齢 100	0.0095

Table 6: ある年齢の男性が単独世帯に属する割合

条件 X	条件 Y	割合 (%)
男性・年齢　〜14	単独世帯	0.0116
男性・年齢 15〜19	単独世帯	7.0119
男性・年齢 20〜24	単独世帯	27.9853
...	...	...
男性・年齢 75〜79	単独世帯	10.2981
男性・年齢 80〜84	単独世帯	10.9273
男性・年齢 85〜	単独世帯	11.7366

整数同士の計算なので，c_{sj} が適切な値になっていた時に $c_{sj} - r_{sj}\cdot m_{sj}$ を計算すると 0 になる．したがって，式（2）の目的関数はその値が 0 になれば復元データが統計データに適合していると考えられる．目的関数は復元データと統計データの誤差を計算しており，その値が大きいほど統計データとの誤差が大きいということになる．よって，目的関数値を最小化することで統計データに適合するような復元データを得ることができる．

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

目的関数は 9 つの統計データに対してそれぞれの誤差を計算するので，合計で 9 つの目的関数値を用いることになる．最適化を行うときは式（1），(2）の両方とも 9 つのそれぞれの目的関数値の和を計算した値を

用いる．したがって式（3）で計算した値が最適化で用いる解（復元データ）の目的関数値を表す．

$$Min \sum_{s=1}^{S} f_s(A) \tag{3}$$

2.4 Simulated Annealing を用いた最適化

本研究では池田らが提案している SA の解の変更手法を改良した改良型 SA[7] を用いて最適化を行う．池田らの SA と改良型 SA のアルゴリズムを次に示す．

Step 1. 復元データを初期生成

Step 2. 探索回数が規定数に達していれば探索を終了

Step 3-a. ランダムにエージェントの年齢を変更

Step 3-b. 解の遷移が連続で 5 回行われていない時は目的関数値が小さくなるようにエージェントの年齢を変更，それ以外の時はランダムに年齢を変更

Step 4. 解の遷移判定

Step 5. 探索回数を更新して SA の温度を冷却

Step 6. Step 2 の処理に戻る

池田らの SA と改良型 SA では Step 3 で行う新たな解の生成手順が異なっている．池田らの SA は Step 3-a を表し，新たな解を生成する時にランダムにエージェントを選択して年齢を変更する．改良型 SA は Step 3-b を表し，解の遷移が連続で 5 回行われていない時に統計データ 3，4 に対する目的関数値が必ず小さくなるようにエージェントの年齢を変更する．解の遷移が連続で行われていない回数が 4 回以下の時，または統計データ 3，4 に対する目的関数値がこれ以上改善しない時にはランダムにエージェントを選択して年齢を変更する．いずれの手法もエージェントの選択は一体で，年齢を変更する時は人口分布の統計データに基づいて変更する．この改良型 SA を用いて探索回数を工夫することで式（2）の値が実質的に 0 の復元データを得ることができている [7]．

3 統計データとの誤差最小化のための要素

SA による最適化で 1000 試行の実験から，復元データの目的関数値の最小化には初期生成時の世帯構成が影響していると報告されている [8]．初期生成時の世帯構成とは，9 種類の世帯がそれぞれ存在している数，復元データの総人口，の二つである．まず，9 種類の世帯がそれぞれ存在している数を調整するための式（4）について記述する．

$$\sum_{t=1}^{T} \mid expectation_t - household_t(A) \mid \tag{4}$$

T は世帯の種類を表しているので $T = 9$ となる．$expectation_t$ は復元データの初期生成で，世帯を生成する時に用いる統計データの値から計算したそれぞれの世帯の期待値を表している．本稿では復元データの世帯数を 500 世帯（$H = 500$）と規定した実験のみを行っている．復元データの規模が 500 世帯の時の期待値を Table 7 に示す．$household_t(A)$ は復元データの中で世帯の種類 t が存在している世帯数を表している．式（4）は統計データから計算した各世帯数の期待値との差の絶対値を計算している．したがって，復元データの各世帯数が期待値と異なっているほどその値が大きくなる．世帯の期待値の計算は統計データの割合を計算した後で整数に丸めているので，各世帯の期待値を合計すると 499 となる．つまり，500 世帯の規模の時は式（4）の値が最小で 1 になる．

Table 7: 9 種類の世帯の割合と期待値

世帯の種類	割合 (%)	期待値
単独	33.98	170
夫婦のみ	20.74	104
夫婦と子供	29.24	146
父親と子供	1.34	7
母親と子供	7.81	39
夫婦と両親	0.47	2
夫婦と片親	1.48	7
夫婦と子供と両親	1.86	9
夫婦と子供と片親	3.07	15
合計	99.99	499

Table 8: これまでの結果

	平均値	標準偏差
SA[3]	49.68	13.72
改良型 SA[6]	42.59	15.10
世帯構成 EC+SA[8]	39.15	6.91
世帯構成 EC+改良型 SA[9]	30.44	6.73

次に，復元データの総人口を調整するための式（5）について記述する．$population(A)$ は復元データの総人口を表している．つまり式（5）は，復元データの総人口が 1275 人から離れている大きさを表す．この 1275 という値については，1000 試行の実験で得られた復元データの目的関数値との相関係数が最大になるのが 1275 と設定した時なので，復元データの総人口の目標となる値を 1275 に設定している [8]．適合させる統計データの中に総人口のデータが存在している場合は，そのデータの値を使うことが望ましい．

$$\mid 1275 - population(A) \mid \tag{5}$$

1000 試行の実験で得られた復元データの目的関数値と式（4），(5）の値の和にはある程度の相関あることから，式（4），(5）の値を最小化することで SA による最適化で目的関数値がより小さな値になることが期待できる．しかし，既存の SA で行っているエージェントの年齢変更ではこれらの値は変化しないので，進化計算を用いた新たな最適化手法を考案している [8]．復元データを初期生成した後，まず進化計算を用いた最適化手法で式（4），(5）の値を最小化する．その後，SA を用いた最適化手法で 9 つの統計データとの誤差を最小化する．その結果，SA のみによる最適化よりも目的関数値がより小さな値になっている．これまでの結果を Table 8 に示す [9]．世帯構成 EC とは復元データの世帯構成を変化させるための進化計算を用いた最適化手法を表している．

Table 9: 100 試行の実験結果（計算時間（秒））

	平均値	標準偏差
世帯構成 EC+改良型 SA	50.44	0.97
世帯構成 SA+改良型 SA	11.05	0.07

Table 10: 100 試行の実験結果（目的関数値）

	平均値	標準偏差
世帯構成 EC+改良型 SA	30.44	6.73
世帯構成 SA+改良型 SA	27.87	4.90

4 提案手法

世帯構成を変化させるための進化計算を用いた手法では，解となる復元データを複数保持することで必要なメモリが大きくなることと，計算時間が長くなっていたことの二つの問題があった．復元データを大規模にした時に，これらの問題が顕著になる可能性が考えられる．そこで本研究では，世帯構成を変化させるために SA を用いた手法を提案する．SA は単一の解に対して最適化を行うので，複数の解を保持する必要はなく，単純な処理の繰り返しになるので計算時間の短縮が期待できる．提案したアルゴリズムを次に示す．

Step 1. 復元データを初期生成

Step 2. 探索回数が規定数に達していれば探索を終了

Step 3. 一世帯を新たに生成した世帯と入れ換える

Step 4. 解の遷移判定

Step 5. 探索回数を更新して SA の温度を冷却

Step 6. Step 2 の処理に戻る

Step 3 で新しく生成した世帯との交換を行うことで，9 種類の各世帯数と総人口を変化させることができる．新しく生成する世帯を 9 種類の内のどの世帯の種類にするかは統計データの割合に基づいて決定する．進化計算を用いた世帯構成に対する最適化 [8, 9] では，9 つの統計データとの誤差を計算する式（2）の値，式（4）に 10 を掛けた値，式（5）を二乗した値，の三つの和を目的関数値として扱っていた．この重み付けにより，9 つの統計データとの誤差をある程度考慮しつつ，世帯構成の状態を優先して最適化することができる．本研究では，Table 8 の結果との精確な比較を行うために同じ重み付けの設定で実験を行うものとする．したがって最適化全体の流れは，まず提案した復元データの世帯構成を変化させるための SA による最適化を行う．その後，9 種類の統計データとの誤差を最小化するために改良型 SA を用いて最適化を行う．

実験のパラメータについては，Table 8 と比較できるように設定する．世帯構成に対する最適化の探索回数が 5 万回，9 つの統計データとの誤差に対する最適化の探索回数が 120 万回と設定し，合計で 125 万回の探索回数とする．復元データの世帯数は 500（$H = 500$）とする．SA の初期温度は 1.0 で収束温度は 0.0001 に設定し，冷却スケジュールは毎回の探索時に線型に温度を低下させていく．世帯構成に対する SA の最適化と 9 つの統計データとの誤差に対する SA の最適化で同じ温度設定とする．100 試行の実験を行い，既存の手法と比較する．実験結果を Table 9，10 に示す．

Table 9 は 100 試行の実験の 1 試行ごとの計算時間の平均値と標準偏差である．世帯構成 EC+改良型 SA が既存手法，世帯構成 SA+改良型 SA は提案手法となっている．それぞれの手法で標準偏差が極めて小さいことから，1 試行ごとの計算時間のばらつきはほとんどない．平均値からは提案手法が既存手法に比べて短い時間で実行できていることがわかる．9 つの統計データとの誤差に対して改良型 SA による最適化は両方の手法で共通している．既存手法と提案手法で異なっているのは，世帯構成に対する最適化の部分である．既存手法は進化計算を用いた手法で行い，提案手法は SA を用いた手法で行っている．したがって，進化計算を用いた世帯構成に対する最適化手法は計算時間のコストがかかるといえる．既存手法である進化計算を用いた手法の実験では，世帯数が 2500，親個体が 20，子個体が 20 と設定されていた．1 世代ごとに子個体として 20 個の復元データを生成する．1 個の復元データは 500 世帯なので，1 世代ごとに 1 万世帯（500×20）を生成することになる．世代数は 2500 としているので 1 試行の最適化で 2500 万世帯（10000×2500）を生成することになる．それに対して提案した SA の手法では，1 回の探索で 1 世帯だけを生成するので探索回数と同じ値の 125 万世帯を生成することになる．以上のことから，進化計算を用いた世帯構成に対する最適化では生成する世帯の数が多いために計算時間が長くなると考えられる．

Table 10 は最適化で得られた目的関数値の平均値と標準偏差を示している．この目的関数値は 9 つの統計データとの誤差を計算する式（2）の値のみなので，世帯構成に対する最適化で用いていた式（4），(5）の値は含まれていない．Table 10 から平均値と標準偏差の両方の値で提案手法の方が小さな値となっており，高い精度で最小化されていることがわかる．この 100 試行について平均値の検定を行ったところ，有意水準 5%で有意な差があるといえる結果であった．

4.1 世帯構成の最適化の影響

9 つの統計データとの誤差に対する最適化は，改良型 SA を用いており二つの手法で共通している．したがって，世帯構成に対する最適化手法の違いにより目的関数値の最小化の精度に差が見られたと推測できる．その原因を分析するために，世帯構成に対する最適化の結果を Table 11，12 に示す．

Table 11 は式（4）で計算した統計データに基づいた 9 種類の世帯数の期待値に対するそれぞれの差の総和についての結果を示している．本研究で設定した復元データの世帯数 500 では，Table 7 で示したように統計データの割合から計算した 9 種類の世帯数の期待値の合計が 499 となる．したがって，この値は最小値が 1.0 で，その値が小さいほど統計データの各世帯数の分布に近いことを表す．進化計算を用いた既存手法は多少ばらつきがあるのに対して，提案手法は全ての試行で式（4）の値が最小値となる 1.0 となっていた．したがって，提案手法では各世帯数が十分に最適化されていると考えられる．

Table 12 は式（5）で計算した復元データの総人口が

Table 11: 100 試行の実験結果（式（4）の値）

	平均値	標準偏差
世帯構成 EC+改良型 SA	1.78	1.84
世帯構成 SA+改良型 SA	1.00	0.00

Table 12: 100 試行の実験結果（式（5）の値）

	平均値	標準偏差
世帯構成 EC+改良型 SA	1.03	0.74
世帯構成 SA+改良型 SA	1.55	0.59

Table 13: 復元データの総人口の分布（100 試行）

	世帯構成 EC+改良型 SA	世帯構成 SA+改良型 SA
1272 人	3	4
1273 人	20	48
1274 人	54	47
1275 人	23	1
合計	100	100

1275 人から離れている値の大きさについての結果を示している．この結果に対して平均値の検定を行ったところ，有意水準 5%で有意な差があるといえる結果になった．したがって，復元データの総人口に対しては既存手法の方が高い精度で最適化されているといえる．

また，100 試行の実験で得られたそれぞれの復元データの総人口がどのような分布になっているかを Table 13 に示す．既存手法と提案手法の両方で全ての試行が 1272〜1275 人の間の分布となった．式（5）では 1275 との差の絶対値を計算しているので，Table 12 の結果から考えると復元データの人口が 1276 人以上になっている試行が存在することは十分に考えられる．しかし，実験で得られた結果では人口が 1275 人より大きくなるような試行は存在しなかった．したがって，今回設定した 1275 という値は復元データの総人口の期待値から考えてやや大きい値になっていると考えられる．

以上の分析から，世帯数の期待値に関しては提案手法が，復元データの総人口に対しては既存手法の方が適切に最適化されていたことがわかった．しかし，9 つの統計データとの誤差については提案手法の方がよい結果になっていた．したがって統計データとの誤差を最小化には，復元データの総人口よりも世帯数の期待値との差が大きく影響していると考えられる．

5 まとめ

本稿では統計データから個人の情報を復元する手法について精度改良を試みた．世帯構成に対する最適化を SA による手法で設計した．最適化に用いる解は一つなので単一の復元データのみの生成で最適化を行うことが可能になった．また，計算時間についても短縮することができた．

統計データとの誤差の最小化に対して影響を与えている要素は，他にも存在していると考えられる．実験結果の分析を行うこと い，その要素を一つ一つ解明することで，より統計データとの誤差が小さい復元データを高い精度で得ることが可能になるしかし，本研究の目的は単に統計データとの誤差を最小化することではなく現実に存在するようなデータを復元することである．9 つの統計データに適合させることで統計データの分布には近づいているが，それが実際に存在しているデータであるかという保証や確証はない．また，復元データと実際に存在しているデータとの比較をすることは極めて困難である．今回設定していた復元データの総人口の目標値 1275 は統計データに基づいた値ではなく，目的関数値との相関関係によって導いた値であった．2010 年の状態を表す統計データでは総人口が約 1 億 2800 万人で，総世帯数が約 5195 万世帯となっているから 1 世帯あたりの平均人数は約 2.46 人となる．実験の設定では復元データの規模は 500 世帯なので，統計データの値に基づいた適切な人口は 1230 人となる．したがって，総人口が 1275 人になるように設定して実験を行い，9 つの統計データに対して完全に適合するような復元データが得られたとしても，統計データに基づいた人口は 1230 人なのでその復元データが現実を忠実に再現できているとはいえない．今後は，新たな要素に対して最適化を行うことで現実との乖離を生み出していないかに注意して研究を進めていく．

謝辞

本研究は JSPS 科研費 26380277 の助成を受けたものです．

参考文献

1) N. Gilbert, K. Troitzsch : Simulation for the Social Scientists，Open University Press（2005）
2) 三上達也：マルチエージェントモデルによる社会シミュレーション，政策科学，**14**-3，121/134（2007）
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，計測自動車制御学会第 43 回システム工学部会研究会，11/14（2010）
4) 福田純也，喜多一 : シミュレーテッドアニーリングによるエージェント属性決定手法を用いた人口推計モデルの評価，第 4 回社会システム部会研究会，35/40（2013）
5) 市川学・出口弘 「社会シミュレーションのための仮想都市環境構築システム」 第 7 回社会システム部会研究会，91/94 (2014)
6) 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，第 5 回社会システム部会研究会，121/126（2014）
7) 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，第 7 回社会システム部会研究会，47/52（2014）
8) 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，進化計算シンポジウム 2014，196/203（2014）
9) T. Murata, D. Masui:Designing Simulated Annealing and Evaolutionary Algorithm for Estimating Attributes of Residents from Statistics, CEC2015（投稿中）
10) 国立社会保障・人口問題研究所 : 人口統計資料集，2012 年，2013 年 http://www.ipss.go.jp/syoushika/tohkei/Popular/Popular2013.asp?chap=0
11) 国立社会保障・人口問題研究所 : 第 14 回出生動向基本調査 2-1，2010 年 http://www.ipss.go.jp/ps-doukou/j/doukou14/doukou14.asp
12) 厚生労働省 : 平成 23 年度全国母子世帯等調査結果報告 19，2011 年 http://www.mhlw.go.jp/seisakunitsuite/bunya/kodomo/kodomo_kosodate/boshi-katei/boshi-setai_h23/
13) 厚生労働省 : e-Stat 人口動態調査 9-14，2010 年 http://www.e-stat.go.jp/SG1/estat/List.do?lid=000001101829