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

○枡井大貴　村田忠彦 （関西大学）

Influence of Objective Functions in Attributes Reconstruction from Statistics Using Simulated Annealing

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

Abstract− Social simulation is one of emerging research topics in computational intelligence. However, many applications in social simulations tend to employ too simple model and they can give only abstract lessons from their simulation results. In order to make simulations more concrete, those simulations should employ the real data to specify attributes of simulated citizens and environment. In this paper, we develop a heuristic approach using a simulated annealing and compare two objective functions for reconstructing attributes citizens in social simulation. We employ the real data such as governmental statistics to reconstruct attributes of citizens. Two objective functions are considered to reconstruct them and show demographic pyramids that are reconstructed using the heuristic.

Key Words: 社会シミュレーション，シミュレーテッドアニーリング，統計データ

1　はじめに

社会シミュレーションは計算知能における新しい研究課題の一つである [1]．しかしながら，多くの社会シミュレーションのアプリケーションは単純なモデルを用いる傾向があり，そのような単純なシミュレーションの結果からは抽象的な教訓しか得ることができない．社会シミュレーションをより具体的にするために，社会シミュレーションはシミュレートされた国民や環境の属性あるいは特徴を規定するために現実のデータを用いるべきである．

社会シミュレーションにおいて現実のデータを利用する方法は二つ挙げられる．一つは，シミュレーションで GIS（Geographic Information System）を利用することである [2]．輸送や感染病のための社会シミュレーションは，ある伝染病に感染した患者の町や地域において混雑を予測するために GIS データを用いる．もう一つは，社会シミュレーションにおける国民やエージェントの属性をアンケートや統計データによって定めることである [3–6]．アンケートを使用する時 [4]，社会シミュレーションの設計者はそのアンケートに回答した人々についての詳細な情報を得ることができる．しかし，そのアンケートのサンプルの規模は全ての国民の規模に対して比較的小さいものとなる．そのアンケートの回答者に基づいて，同等の規模の範囲で社会シミュレーションを設計することができる．政府や国立機関によって公表されている統計表を使用する時 [5,6]，統計データから詳細な情報を復元することは困難である．なぜなら，一般に公表されている統計データの中で掲載されているデータは累積されたデータとなっているためである．さらに，近年ではアンケート調査などに回答する時，人々は自身のプライバシーを守ることに敏感になっている．政府や国立機関がアンケート調査を行う時，個人情報保護の観点から，データの二次利用を厳しく制限している．

本研究では，一般に公開されている統計データから詳細な情報を復元する方法について注目する．我々は統計データから国民の属性を復元するために，SA（Simulated Annealing）[7,8] によるヒューリスティックなアプローチ [9] を用いる．我々は，SA を用いて最適化するための性質の異なる目的関数を提案し，提案した目的関数と先行研究 [9] で提案された目的関数を用いた時の，結果を比較しその違いについて考察する．先行研究 [9] では，統計データの数値を超える部分の誤差を求め，その二乗平均を計算する目的関数を提案した．本研究では，推計データと統計データの誤差の絶対値を計算する目的関数を用いる．ヒューリスティックなアルゴリズムを用いて復元された人口動態ピラミッドを示すことで，二つの目的関数を用いた SA の探索性能を比較する．最適化の結果を通して，提案した目的関数が全ての統計データに適合するために，それぞれのエージェントの属性値を的確に調整していることを示す．

2　SA を用いた最適化による属性値の復元

2.1　目的関数

本研究では，現実の統計データに従ってエージェントの属性を復元するために，先行研究の手法 [9] を用いる．最適化に用いる SA は，規定の温度に従ってある解から別の解へ変化することで良い解を見つけ出すような，効果的なヒューリスティックなアプローチであることはよく知られている．その温度は焼きなましの過程を模倣するように，高い温度から低い温度へとスケジュールされなければならない．高い温度の下では一つの解から積極的に別の解へ状態を変化し，低い温度の下ではあまり変化しない．

本研究では，社会における世帯の組み合わせを一つの解と定義する．統計データに従って複数の世帯の種類を作成し，その種類に従って世帯の構成員となるエージェントを生成する．そのエージェントは世帯の中での役割（単身，夫，妻，父，母，子）を持ち，属性値として自身の年齢と性別を保持している．Fig. 1 は本研究で使われている世帯の種類を示している．国立社会保障・人権問題研究所によって公表されている統計データに従って 9 種類の世帯の種類を使用する．世帯の種類は以下のように分類される．

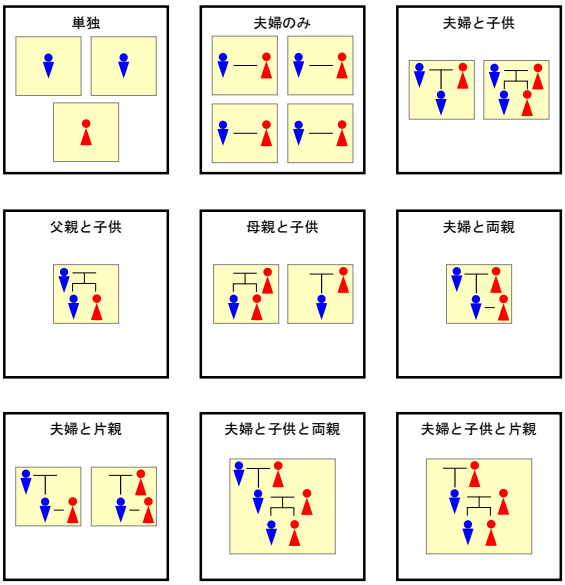

Fig. 1: 世帯の種類

Table 1: 統計データの例

項目	割合
1	3%
2	5%
3	9%
...	2%
C	4%
合計	100%

1. 単独
2. 夫婦のみ
3. 夫婦と子供
4. 父親と子供
5. 母親と子供
6. 夫婦と両親
7. 夫婦と片親
8. 夫婦と子供と両親
9. 夫婦と子供と片親

属性値をランダムに設定して初期生成を行った後，SA を用いて人口動態ピラミッドを復元する．統計データに対して正確に適合するような，各属性値を持つ推計データへと最適化するために，Table 1 のような統計データを用いる．推計データとはこの問題に対する解 x として扱われる．Table 1 のように，統計データ s は $1 \cdots C$ までの項目と割合 r_{sj} を持つ．割合 r_{sj} の数値は現実の統計データの数値によって定まる．ヒューリスティックなアルゴリズムによって発見された推計データを x とし，その x を用いることで，統計データ s の r_{sj} に対する実際の母数となる $m_{sj}(x)$ を計算することができる．例えば，父親と子供の年齢差の統計データに対しては，父と子を構成員に持つ世帯から計算される．「単独」や「母親と子供」等の世帯はこの統計データの対象にはならない．r_{sj} と m_{sj} の積を計算することで，統計データ s の項目 j に対する適切な人数が推計される．それに対して，解 x のエージェントを数え上げることで統計データ s の項目 j に対する実際の値である c_{sj} を計算することができる．これらの変数を用いて，現実の統計データと推計データの誤差の推計するために先行研究 [9] では次の目的関数が提案されている．

$$Min\ f_s^1(x) = \frac{1}{C}\sum_{j=1}^{C} 4 \cdot (c_{sj}(x) - r_{sj} \cdot m_{sj}(x))^2 \quad (1)$$

ここで，$c_{sj}(x)$ は整数値で $r_{sj} \cdot m_{sj}(x)$ は実数値である．r_{sj} は割合の数値であるため，推計データの数値 $c_{sj}(x)$ が適切な値であっても，式（1）を計算した時，最大で 0.5 の誤差が残る可能性がある．それを平方すると，0.25 となり，それから 4 を掛けることで，式（1）の計算結果は 1.0 となる．つまり，もし式（1）が 1.0 以下ならば，現実の統計データと推計データが適合するような値に調整されていることが期待できる．この目的関数は連続関数として設計されている，しかし，直感的に理解するのが困難かもしれない．また，統計データと推計データがエージェントの組・人の数単位でどの程度適合していないかを，その値から推測することは容易ではない．本研究では，現実の統計データと推計データの推計誤差を計算するための以下の目的関数を提案する．

$$Min\ f_s^2(x) = \sum_{j=1}^{C} |c_{sj}(x) - Round(r_{sj} \cdot m_{sj}(x))| \quad (2)$$

この目的関数では $r_{sj} \cdot m_{sj}(x)$ を整数値に丸めているため，統計データ s の各項目における誤差は整数値で計算される．式（2）の目的関数の値は，推計データ上で各統計データに適合していないエージェントの組・人の数を表している．したがって，式（2）が 0 となれば，統計データに対して推計データが適切な値に調整されていると考えられる．

式（1）あるいは式（2）に従って，それぞれの統計データ s に対する評価値を計算することができる．この評価値は目的関数によって計算された誤差を表しているので，その値が小さいほど推計データが統計データにより適合していると考えられる．社会シミュレーションに必要なデータを復元するために，統計データ S を持つとき，以下の目的関数にまとめることができる．

$$Min \sum_{s=1}^{S} f_s^i(x), (i = 1, 2) \quad (3)$$

式（3）は，各統計データに対するそれぞれの目的関数の値の和を計算している．したがって，式（3）の値を小さくすることを目的とし，最小化問題として解くことで各統計データに適合するような推計データを得ることができる．本研究では，SA を用いて推計データに属するエージェントの属性値の最適化を行い，得られた推計データと現実の統計データの数値を比較するためのグラフを示し，式（1）と式（2）の最適化の結果についても比較する．

2.2 シミュレーテッドアニーリング

SA の手順は以下の通りである．

Step 1: ランダムに初期人口の組み合わせを生成する．

Step 2: 目的関数の値を計算する．

Step 3: 推計データ上のエージェント一体の年齢か性別をランダムに変更することで，現在の解から異なる解を生成する．

Step 4: 生成した解の目的関数の値を計算する．

Step 5: もし，新たに生成した解が現在の解より良い値となっている，あるいは SA の温度により確率的に変更が許可された時，新たに生成した解を次の解とする．

Step 6: 規定回数になるまで Step 3 から Step 5 を繰り返す．

Step 7: これらの手順の中で最終的な解を示す．

基本的には，先行研究の手法 [9] を用いて最適化を行う．しかし，エージェントが持つ役割の整合性を保つためにエージェントの属性値の変更に対して制限を設けた．先行研究 [9] では Step 3 でエージェントの年齢・性別をランダムに変更していたが，本研究では Step 3 で「単独世帯に属する」・「未婚の子供の役割を持つ」のどちらかを満たすエージェントにのみ性別の変更を行った．もし，この条件を満たさないエージェントの性別を変更すると,「父親の役割を持つ女性のエージェント」のような現実には存在し得ない状態となる可能性がある．このような状態となることを防ぐために，全てのエージェントに対しては性別の変更を行っていない．

Step 3 では，まず変更対象となるエージェントをランダムに選択し，そのエージェントの性別の変更が可能であるかを判断する．条件を満たしていて性別の変更が可能であれば， 50% の確率で変更を行う．性別の変更が行われなかった場合は年齢を 0 〜 100 の範囲でランダムに変更する．Step 4 での新たな解の目的関数の値を計算する時は，処理時間を減らすために先行研究の手法 [9] と同様の計算方法を用いた．Step 3 ではエージェント一体の属性値のみを変更しているので，変更していない他のエージェントの属性値まで計算する必要はない．現在の解と新たに生成した解の変更点のみを計算するだけで十分である．これにより処理時間を短縮することが可能となり，さらに，世帯数を増加させても実験の処理時間に大きく影響することはない．

3 シミュレーション

3.1 問題設定

国立社会保障・人権問題研究所と厚生労働省によって公表されている，9 種類の統計データ [10] に従ってエージェントの属性を復元する．以上の 9 種類の統計データを用いているので，式（3）の S の値は 9 となる．Tables 2-5 は式（1）と式（2）の目的関数を計算するための統計データの一例である．

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 9-14，2010 年 [11]）
4. 人口ピラミッド（男）（表 2-3，2012 年 [10]）
5. 人口ピラミッド（女）（表 2-3，2012 年 [10]）
6. 単独世帯の人口分布（男）（表 7-28，2013 年 [10]）
7. 単独世帯の人口分布（女）（表 7-28，2013 年 [10]）
8. 夫婦のみ世帯の人口分布（男）（表 7-28，2013 年 [10]）
9. 夫婦のみ世帯の人口分布（女）（表 7-28，2013 年 [10]）

Table 2: 父と子の年齢差

年齢差（父－子）	割合
〜14	0.00%
15〜19	0.55%
20〜24	8.49%
25〜29	26.16%
30〜34	33.79%
35〜39	20.89%
40〜44	7.91%
45〜49	2.10%
50〜	0.11%
合計	100.00%

Table 3: 母と子の年齢差

年齢差（母－子）	割合
〜14	0.00%
15〜19	1.26%
20〜24	10.37%
25〜29	28.59%
30〜34	35.87%
35〜39	20.59%
40〜44	3.24%
45〜49	0.07%
50〜	0.11%
合計	100.00%

Table 4: 夫婦の年齢差

年齢差（夫－妻）	割合
〜-4	6.13%
-3	3.12%
-2	4.82%
-1	9.57%
0	20.23%
1	13.76%
2	9.93%
3	7.74%
4	6.04%
5	4.72%
6	3.52%
7〜	10.43%
合計	100.00%

Table 5: 人口ピラミッド（男）

年齢	割合
0	0.87%
1	0.87%
2	0.89%
3	0.89%
...	...
98	0.01%
99	0.01%
100～	0.01%
合計	100.00%

3.2 シミュレーション結果

二つの目的関数の性能を示すために，それぞれの目的関数を用いて行った実験結果を比較する．SA の最適化では，パラメータは 1000 世帯数で探索回数 1 億回とした．初期温度は 1.0 で収束温度は 0.0001 とし，冷却スケジュールは線形的に温度が下がるように設定した．実験は 1 試行ごとにシード値を変更し，式（1）と式（2）を別々に用いて 30 回の実験を行った．Table 6 は 30 回試行の目的関数の値の平均値と標準偏差を示している．「目的関数」の列は式（1），式（2）のどちらの目的関数を用いて最適化が行われたかを示している．「式（1）の値」の列は，最適化により得られた推計データを式（1）の目的関数を用いて再評価した値である．「式（2）の値」の列は，最適化により得られた推計データを式（2）の目的関数を用いて再評価した値である．Tables 7，8 は 30 回試行の中で，最適化された推計データの目的関数の値が平均値よりも小さく，比較的良い解を探索できたと思われる試行の結果で，各統計データごとの誤差の数値を目的関数別に示している．Table 7 は式（1）で最適化された結果で，Table 8 は式（2）で最適化された結果である．

Table 6: 30 回試行の目的関数の値（平均値と標準偏差）

	目的関数	式（1）の値	式（2）の値
平均値	式（1）	13.10	150.43
	式（2）	73.85	59.77
標準偏差	式（1）	6.92	48.06
	式（2）	63.60	20.88

まず，Table 6 では，縦に並んでいる値同士で比較を行う．横に並んでいる値は異なる目的関数で評価した値なので横の値同士の比較は行わない．「式（1）の値」で見ると，式（1）で最適化された方が小さな値になっている．一方，「式（2）の値」で見ると，式（2）で最適化された方が小さな値になっている．最適化の時と再評価の時の式の番号が，一致している欄の値が良い結果となっている．

次に，Table 6 で式の番号が一致していない欄の値が，一致している欄の値に比べて大きな値となった理由について，Tables 7-8 から考察する．Table 7 では，式（1）の値が各統計データで小さな値となっている．しかし，式（2）の値は人口ピラミッド項目の値が他の項目に比べて大きな値となっている．また，人口ピラミッド（女）の項目については，式（1）の値が目標となる 1.0 以下になっている．しかし，式（2）の値が 25 ということから，式（1）の値が 1.0 以下でも，組・人の数単位では統計データに適合していないエージェントが多数存在している可能性がある．これらは，人口ピラミッドの統計データが 1 歳区切りで 101 項目の統計データであることが原因と考えられる．式（1）では統計データの各項目の誤差の値の平均値を取るために，その誤差の和を項目数 C で割っている．101 項目の人口ピラミッドの統計データでは，エージェント一体の変更から生じる目的関数の値の変化の量が，他の統計データの時に比べて小さな量となる．したがって，SA による最適化の過程では，目的関数の値がより小さい解を探索することになるので，項目数の少ない統計データへの適合が優先されて行われる．結果的に，人口ピラミッドの統計データへの適合が，他の統計データに比べて不十分となる．

逆に，Table 7 では，式（1）の値は 1.0 以下になっていない項目が多いが，式（2）の値は小さな値になっている．また，式（2）で見た時に，特定の統計データに対する適合が不十分になるような最適化は行われていないことが分かる．これらは，本研究で提案した式（2）の目的関数の設計上，エージェント一体の変更によって生じる目的関数の値の量が一定であるためと考えられる．それにより，式（2）で最適化した時よりも，エージェントの組・人の数単位の誤差が小さい解を得ることができる．

Table 7: 式（1）で最適化した時のそれぞれの評価値

統計データ	式（1）の値	式（2）の値
父子の年齢差	0.29	0
母子の年齢差	0.16	0
夫婦の年齢差	0.24	0
人口ピラミッド（男）	2.50	59
人口ピラミッド（女）	0.72	25
単独世帯の人口分布（男）	0.22	1
単独世帯の人口分布（女）	0.67	5
夫婦のみ世帯の人口分布（男）	0.11	0
夫婦のみ世帯の人口分布（女）	0.24	1
合計	5.14	91

Table 8: 式（2）で最適化した時それぞれの評価値

統計データ	式（1）の値	式（2）の値
父子の年齢差	0.89	2
母子の年齢差	1.15	2
夫婦の年齢差	0.24	0
人口ピラミッド（男）	0.64	8
人口ピラミッド（女）	0.44	2
単独世帯の人口分布（男）	9.69	8
単独世帯の人口分布（女）	2.77	5
夫婦のみ世帯の人口分布（男）	1.57	3
夫婦のみ世帯の人口分布（女）	4.00	7
合計	21.39	37

Tables 2-10 は，Tables 7，8 と同じ試行の結果で，最適化で得られた推計データの割合を各統計データ別にプロットしたグラフである．グラフの赤線（statistics）が現実の統計データの値なので，この実線に重なれば推計データが現実の統計データに適合しているといえる．緑線（real func）は式（1）を用いて最適化した結果で，青線（round func）は式（2）を用いて最適化した結果である．Figs. 2-4 はグラフの線が重なっているので，両方の評価関数で適切に最適化されていると考えられる．式（1）を用いて最適化を行った結果については，Table 7 の式（2）の値が 0 であることから，エージェントの組・人の数単位で見ても統計データに完全に適合していることが分かる．式（2）を用いて最適化

を行った結果については，Table 8 の父子と母子の年齢差の統計データの項目の式（2）の値を見ると，その誤差を表す値がそれぞれ 2 であることから，統計データに完全に適合しているとはいえない．したがって，最適化で得られた推計データを正確に評価するためには，Tables 2-10 のように，推計データの適合度を視覚的に比較するだけでなく，式（2）の目的関数を用いて各統計データごとの誤差の数値についても着目しなければならない．Figs. 5，6 では式（1），(2）のどちらの最適化の値も，統計データの値に対して完全には適合していないことが明確に分かる．しかし，人口ピラミッド（男）で赤線とのズレの大きさに着目すると，両方のグラフの 0-20 歳と 80-100 歳までの範囲で，青線よりも緑線の方がそのズレが大きいと読み取ることができる．これは，Tables 7，8 からも分かるように，式（1）の目的関数で最適化を行うと式（2）の値が大きくなるためである．Figs. 7-10 では，式（1）を用いた最適化で得られた推計データの数値である緑線の方が，現実の統計データの数値である赤線により正確に適合していると読み取ることができる．青線と赤線で特にズレの大きさが顕著なのは，0-20 歳と 80-100 歳までの範囲である．これらは，Tables 7，8 の式（2）値からも明らかである．

式（1）を用いて最適化を行うと，単独世帯や夫婦のみ世帯の人口分布の統計データに対しては高い精度で適合するが，人口ピラミッドの統計データに対しては適合が不十分となる．逆に，式（2）を用いて最適化を行うと，人口ピラミッドの統計データに対しては高い精度で適合するが，単独世帯や夫婦のみ世帯の人口分布の統計データに対しては適合が不十分となる．各年齢差の統計データに対しては，二つの目的関数で大きな差は見られない．

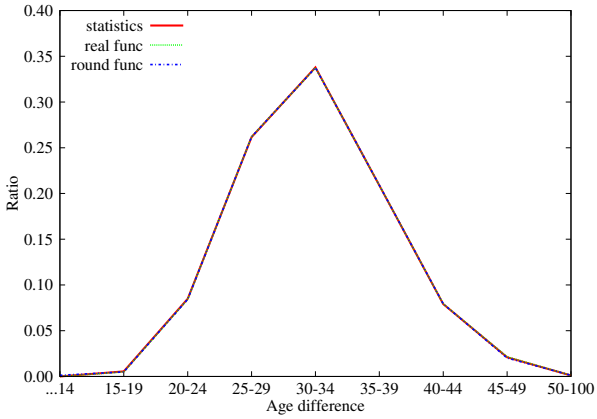

Fig. 2: 父子の年齢差

4 結論

本研究では，SA の最適化に用いる新たな目的関数を提案した．二つの目的関数を用いて最適化を行ったが，両方の目的関数で統計データに完全に適合するような推計データは得られなかった．今後，最適解を発見する効率的な方法をこのエージェント推計に適用する必要がある．また，探索には連続性が必要ではないヒューリスティックなアプローチを用いたので，不連続な値を取る関数を目的関数として利用することができる．このような不連続な値を取る目的関数は，タブー探索や

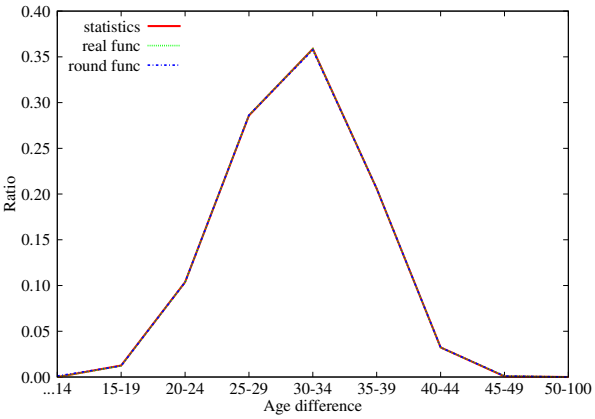

Fig. 3: 母子の年齢差

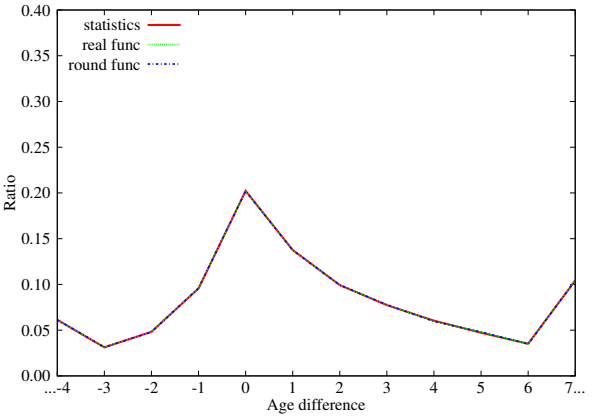

Fig. 4: 夫婦の年齢差

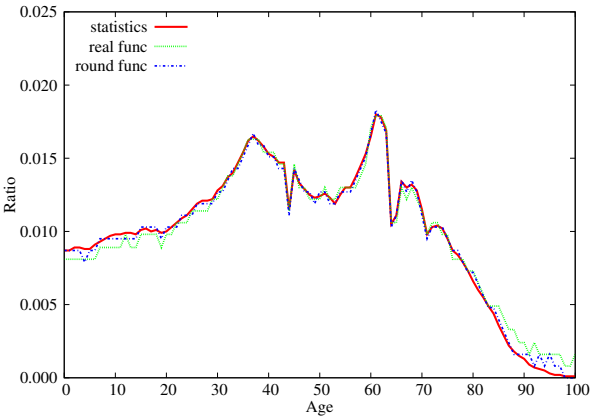

Fig. 5: ピラミッド（男）

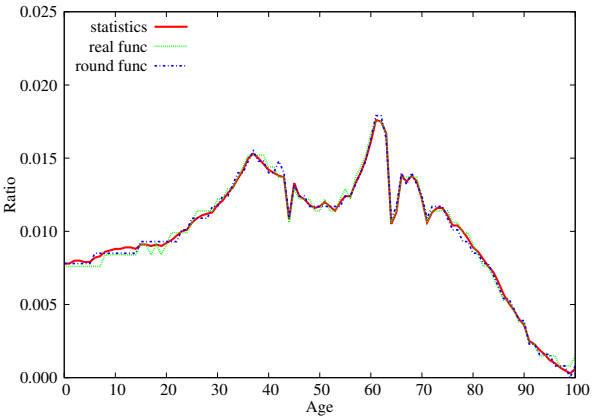

Fig. 6: 人口ピラミッド（女）

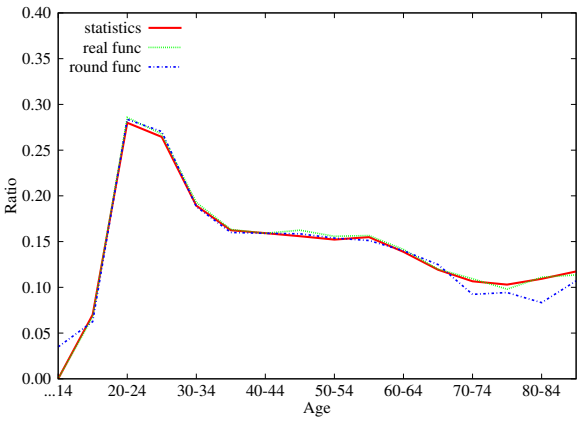

Fig. 7: 単独世帯の人口分布 (男)

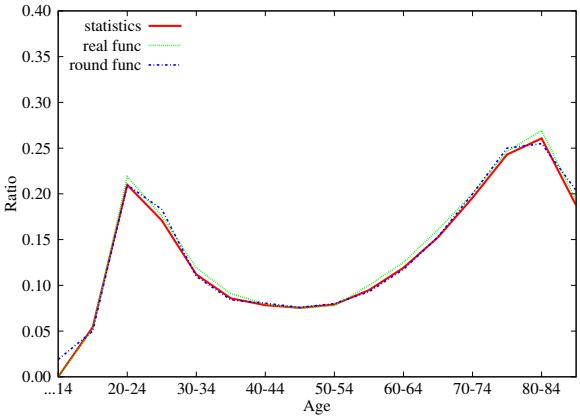

Fig. 8: 単独世帯の人口分布（女）

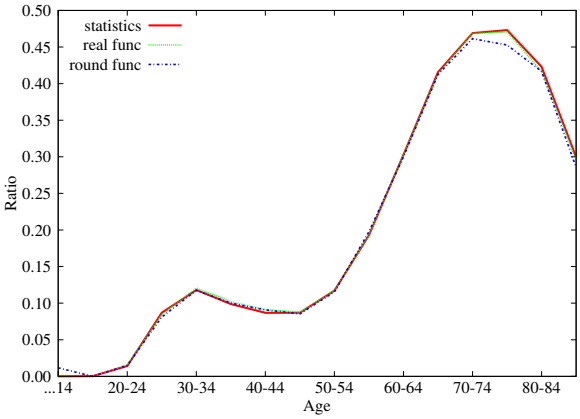

Fig. 9: 夫婦のみ世人口分布（男）

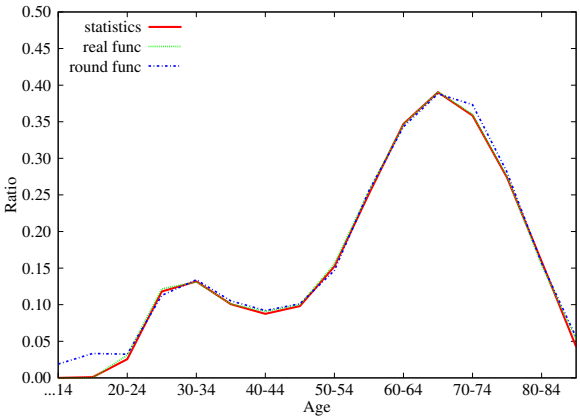

Fig. 10: 夫婦のみ世帯の人口分布（女）

進化アルゴリズムのような探索アルゴリズムに利用されることがある．本研究で扱った問題へのヒューリスティックなアプローチに対する可能性として，この問題に対する進化的手法は既に提案されている [12]．入念なパラメータ調整を行ったが，SA の最適化に比べてわずかに良い解しか見つけることができなかった．この問題に対して交叉や突然変異のような遺伝的操作をより効果的になるよう改良することを検討している．

社会シミュレーションをより現実的にするために，国民や環境に対応する実用的なデータを集めなければならない．しかしながら，GIS データやエージェントシミュレーションのような物理データの統合はまだまだ発展途上である．一つの理由に，データを集めた時と集計した時に時間差が存在する点である．例えば，近年の日本では人口動態調査の結果を集計するためには 1 年が必要とされている．社会シミュレーションの具体性を改良するためには，集計と公開の時間差を無くすことが必要である．しかし，国民の詳細な情報を一般に公開することについては慎重に検討しなければならない．

参考文献

1) N. Gildbert , K. G. Troitzch:Simulation for the Social Scientist, Open University Press（2005）
2) H. R. Gimblett : Integrating Geographic Information Systems and Agent-based Modeling Techniques for Simulating Social and Ecological Processes, Oxford University Press（2002）
3) M. Richiardi, R. Leombruni, N. Saam, M. Sonnessa: A common product for agent-based social simulation, Journal of Artificial Societies and Social Simulation, **9**-1, 15 ページ（2006）
4) 村田忠彦，小西健太：投票シミュレーションツールを用いた投票場所割当のための実用的ポリシー提案の作成，SICE Journal of Control, Measurement, and System Integration, **6**-2, 124/130（2013）
5) 番匠大輔，田村坦之，村田忠彦：世代間公平性と財政的持続可能性を実現するための不確実性下における公的年金の最適計画，システム制御情報学会論文誌，**20**-10, 396/403 (2007)
6) T. Murata, Z. Chen:Agent-Based Simulation for Pension System in Japan, Agent-Based Social System, **10**（Post-Proceedings of Agent-Based Approaches in Economics and Social Complex Systems Ⅶ , Springer Japan）, 183/197（2013）
7) S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi:Optimization by simulated annealing, Science, **220**, 671/680（1983）
8) 喜多一：シミュレーテッドアニーリング，日本ファジィ学会誌，**9**-6，870/875（1997）
9) 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システム工学部会研究会，4 ページ（2010）
10) 国立社会保障・人口問題研究所：人口統計資料集（2012, 2013）`http://www.ipss.go.jp/syoushika/tohkei/Popular/Popular2013.asp?chap=0`
11) 厚生労働省：人口動態調査，e-Stat（2010）`http://www.e-stat.go.jp/SG1/estat/List.do?lid=000001101829`
12) 枡井大貴，村田忠彦：公開統計データを用いた世帯構成員の特徴量の進化計算的推計手法，進化計算シンポジウム，7 ページ（2013）