土地利用マイクロシミュレーションにおける観測マイクロデータ集合と推定集合の適合度評価*

Goodness-of-Fit Evaluation Method between Observed and Estimated Sets of Micro-Data in Land-Use Micro-Simulation*

大谷紀子**・杉木直***・宮本和明****
By Noriko OTANI**・Nao SUGIKI***・Kazuaki MIYAMOTO****

１．はじめに

マイクロシミュレーションは、都市圏における土地利用と交通の詳細な変化の記述手法として、欧米諸国の複数の研究グループによって都市モデルの開発への活用が進められている[1)]。居住立地モデルのような世帯を対象としたマイクロシミュレーションモデルの場合、各世帯には世帯収入、世帯人数、各世帯構成員の年齢、自動車保有、居住地、住宅タイプ等の多くの属性が定義されるが、シミュレーションを実行するためには、全ての世帯に対してこれらの属性を定義したシミュレーション初期年次データを用意する必要がある。しかし、住民基本台帳などから個人や個別世帯に関するデータを入手することは一般的に困難であり、またプライバシー保護の観点からも望ましくない。従って、通常マイクロシミュレーションモデルでは、国勢調査などの入手可能な集計データと追加的に個別世帯の属性情報を提供するサンプル調査を組み合わせて、個別世帯に対して複数の属性の組み合わせを設定したデータ（以降、マイクロデータ）を作成する。マイクロデータの作成手法としては、IPF法やモンテカルロサンプリングによる手法などが提案されているが、推定データ集合と観測マイクロデータ集合間の適合度を評価するための手法が存在しないため、これらの妥当性の検証はなされていない。マイクロデータの観測データ集合は、実際のモデル適用においては入手可能ではないが、推定手法の妥当性はテストデータセットを用いて検証されるべきである。また、マイクロシミュレーションモデルによる結果の有効性自体を評価することができないという課題がある。

そこで本研究では、マイクロデータ推定集合の観測集合に対する適合度の評価手法を構築する。適合度計測自体は、個々の世帯に関する観測および推定マイクロデータ間の距離を定義し、対象地域における全ての世帯の乖離量距離の最小和によってこれらの適合度を評価することにより、比較的単純に定義することが可能である。エージェントの数が30またはそれ以下の場合は、このような手法でも計測可能である。しかし、計算量はエージェント数の階乗に比例して増加するため、一般的な都市モデルにおけるマイクロデータの規模を想定した場合、計算を実行することは現実的に不可能である。よって、一般的なマイクロデータ集合に対して計測を実行可能とするアルゴリズムについても開発を行う。

本稿では、まず初めにマイクロシミュレーションの人口推計に用いられる適合度評価および実行可能な計算を行うために必要なアルゴリズムに関して、既存研究のレビューを行う。その上で、規模が大きい場合でも計測可能なマイクロデータ集合間の適合度評価手法を、近似値の探索手法として遺伝的アルゴリズム（GA）の一手法である共生進化を用いて構築する。少数のエージェントを設定した単純なケースにおいて手法の性能を検証した後、道央都市圏パーソントリップ調査データより抽出された2000世帯のマイクロデータに適用し、構築された適合度評価手法の妥当性を検証する。

２．マイクロデータの適合度評価に関する既存研究

（1）クロスセクション表の適合度

マイクロデータの適合度に関しては、Pritchardら[2)]による研究がなされている。マイクロデータに関する観測データは入手できないことが前提とされているために、観測データについては公表されている属性別人口データよりIPF法を用いて作成したクロスセクション属性の表を用いている。このような集計的なクロスセクション属性の表による観測データの人口特性に対する推定データ集合の適合度を検証しているが、真の観測マイクロデータ集合を知ることができるならば、このような手法では十分な適合度を検証しているとは言えない。

世帯が3つの属性(i, j, k)により区分されると仮定した場合、推定集合$\hat{N}_{ijk}$と妥当性検証のための観測データ集合N_{ijk}間の適合度は、距離ベースの平均平方標準誤差（SRMSE）指標を用いて、式(1)のように評価することが可能である[3)]。

*キーワーズ：マイクロシミュレーション、マイクロデータ、初期データ推計、適合度評価
**博士（情報理工学）、東京都市大学環境情報学部
（〒224-8551横浜市都筑区牛久保西3-3-1、
TEL/Fax 045-910-2938）
***正員、修士(情報科学)、（株）ドーコン総合計画部
****フェロー、工博、東京都市大学環境情報学部

$$SRMSE = \frac{\sqrt{\frac{1}{IJK}\sum_{i,j,k}(\hat{N}_{ijk} - N_{ijk})^2}}{\frac{1}{IJK}\sum_{i,j,k} N_{ijk}} \quad (1)$$

この指標は値が小さいほど、適合度が高いことを示す。各タイプの適合度指標を各観測データ集合に対して順に計算し、これらの平均によって全体の適合度が与えられる。このように属性が3つの場合には、計測に関する計算量の問題は生じない。

（2）共生進化

最適化問題の解法として広く利用されている遺伝的アルゴリズム(Genetic Algorithm, GA)は生物の進化過程を模倣した最適解探索アルゴリズムである。構造が不明確で広大な解空間における最適解探索が可能であり、最適性の定義があいまいな問題にも有効である。学習対象や形態に応じた様々なGAのモデルの1つとして、Moriartyらにより共生進化が提案されている。[4)5)6)]

共生進化では、部分解を個体とする集団と、部分解の組み合わせを個体とする全体解集団を保持し、両集団を並行して進化させる。部分解集団では解の部分的評価を行ない、最適解に含まれ得る多様な部分解を生成する。それらのより良い組み合わせを全体解集団で学習することで、1集団を進化させるGAよりも多様な解候補からの探索を行なうことができる。帰納論理プログラミングや決定木生成への適用手法が提案されており、有用性が確認されている。[7)8)9)]

3．適合度評価問題の定義

（1）定義

適合度評価問題に関する前提は以下のとおりである。

・対象はエージェント集合であり、各エージェントは多変量の属性を持つ。本研究では、特定のゾーンまたは対象地域の世帯マイクロデータセットである。
・属性は全て連続変量とする。本研究では、世帯構成員の年齢、世帯人数などを指す。
・完全な情報を持つ観測データ集合が推定手法の妥当性検証のために入手可能であるものとする。
・推定データはMiyamotoら[10)]によるマイクロデータ推定手法により提供されるものとする。

適合度評価問題は、いずれの推定データセットがより観測データセットに近いかを決定するために各推定データの適合度を算出するものとして定義する。

（2）表記

観測データ A と推計データ E_j は、式(2)および式(3)のように、世帯構成員の年齢を成分とするベクトルで表される。

$$A = \{\mathbf{a}_i = (a_{i1}, a_{i2}, \cdots, a_{iM}) \mid 1 \le i \le N\} \quad (2)$$

$$E_j = \{\mathbf{e}_i^j = (e_{i1}^j, e_{i2}^j, \cdots, e_{iM}^j) \mid 1 \le i \le N\} \quad (3)$$

ここで、M は一世帯の構成人数、N は観測データ数、a_{ik} は観測データにおける i 番目の世帯の k 番目の構成員の年齢、e_{ik}^j は j 番目の推計データにおける i 番目の世帯の k 番目の構成員の年齢を表す。本稿では i を世帯番号と呼ぶことにし、HH IDと表記する。

4．評価指標

（1）定義

2 つのデータ集合の類似度を評価する際、一般には平均と分散が使用される。しかし、全要素の分布ではなく、個々の要素の適合性に基づいて評価するためには、2 要素間距離の和の最小値の使用が有用と考えられる。従って、推計データ集合 E_j と観測データ集合 A との類似度を表す評価値 $Fit(E_j)$ を式(4)で定義する。

$$Fit(E_j) = \min_{\sigma \in S_n} \sum_{i=1}^{N} |\mathbf{a}_i - \mathbf{e}_{\sigma(i)}^j| \quad (4)$$

ここで、S_n は集合 $\{1,2,\cdots,N\}$ から集合 $\{1,2,\cdots,N\}$ へのすべての全単射の集合を表し、$\sigma(i)$ は全単射 σ による i の像を表す。2 ベクトル間の距離にはユークリッド距離を用いる。

（2）計算複雑性

S_n は $N!$ 個の要素を持つので、1 つの推計データ集合の評価値を求めるには、距離和の算出を $N!$ 回繰り返す必要がある。距離和の算出は複雑な処理ではないが、$N!$ は N の増加に従って急速に増加するため、マイクロデータシミュレーションに用いる規模の推計データ集合の評価では計算量爆発の問題が生じる。

評価値計算は、$N!$ 個の全単射から、距離和を最小とするような全単射を探索する問題といえる。すなわち、距離和が最小となるように、観測データ集合の各要素を推計データ集合のいずれかの要素と対応付ける組合せ最適化問題である。さまざまな組合せ最適化問題で有用性が示されている GA の適用により、実時間での評価値算出が期待される。

5．最適化アルゴリズム

（1）共生進化に基づく評価値計算

共生進化に基づいて距離和が最小となるような観測データと推計データの対応付けを決定し、評価値を算出する手法を提案する。本手法では、A と E_j のデータの Lp 組の対応付けを部分解集団の個体として表現する。以降、A と E_j のデータの対応付けを世帯番号ペアと呼ぶ。全体解集団の個体は Lw 組の部分解集団個体の組合せとし、1個体で $Lp \times Lw$ 組の世帯番号ペアを表すようにする。

（2）処理手順

部分解集団の個体の染色体は長さ $Lp \times 32$ のビット列で表す。16 ビットで 1 つの世帯番号を表し、図－1に示すように、1 個体で $Lp \times 2$ 組の世帯番号ペアを表現する。

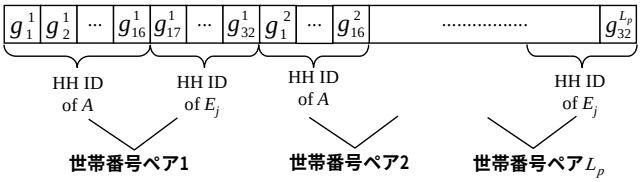
全体解集団の個体の遺伝子は部分解集団の個体を参照するポインタであり、染色体は長さ $Lw = \lceil N / Lp \rceil$ のポインタ列である。全体解集団の個体の例を図－2に示す。全体解集団の個体から世帯番号ペアを生成するアルゴリズムを図－3に示す。本アルゴリズムにより致死遺伝子の生成を回避することができる。

全体解集団の個体の適応度は、世帯番号ペアに基づいて算出される距離和とし、適応度が小さいほど評価は高いものとする。部分解集団の個体の適応度は、当該個体を参照している全体解個体の中で最も評価の高い個体の適応度とする。良い全体解から参照されている部分解ほど良いと判断する。

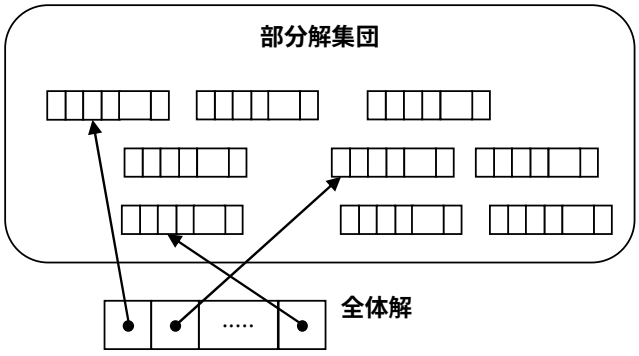
両集団は一般的な GA のオペレータである一点交叉と突然変異により進化する。世代交代は局所解収束の回避に有効な MGG[11]により行なう。

全体の処理の流れは以下の通りである。

1) 初期世代の部分解集団の個体をランダムに生成
2) 初期世代の全体解集団の個体をランダムに生成
3) 全体解集団の個体を評価
4) 部分解集団の個体を評価
5) 次世代の部分解集団を生成
6) 次世代の全体解集団を生成
7) 1)～6)を G 回繰り返す
8) 全体解集団の最良個体の適応度を出力



図－1　部分解集合の個体



図－2　全体解集合の個体

```
makeHHIDpairs() {
    for i := 1 to L_w {
        for j := 1 to L_p {
            fid_A := i番目のポインタが参照している個体の
            g^j_1～g^j_16から算出される世帯番号;
            fid_P := i番目のポインタが参照している個体の
            g^j_17～g^j_32から算出される世帯番号;
            if(fid_Aもfid_Pもまだ世帯番号ペアに使われていない) {
                世帯番号ペア(fid_A, fid_P)を生成;
            }
        }
    }
    for k := 1 to N {
        if(kはfid_Aとして世帯番号ペアに使われていない) {
            fid_P := min(fid_Pとして世帯番号ペアに使われていない数);
            世帯番号ペア(k, fid_P)を生成;
        }
    }
}
```

図－3　全体解集団の個体から世帯番号ペアを生成するアルゴリズム

表－1　パラメータ

パラメータ名	値
全体解集団の個体数	1000
部分解集団の個体数	1000
突然変異確率	0.001
最大世代数 G	5000
部分解個体の染色体の長さ L_p	2

表－2　計算時間の比較

N	全探索 [秒]	提案手法 [秒]
16	1.11	3.07
17	7.54	3.15
18	34.42	3.18
19	147.75	3.30
20	1701.84	3.33

表－3　評価値

データ	平均	標準偏差
E_{1a}	3663.12	113.18
E_{1b}	3681.57	165.17
E_{1c}	3863.49	201.01
E_{1d}	3752.24	163.84
E_{1e}	3754.48	209.03
E_{2a}	5127.76	228.42
E_{2b}	4930.35	146.21
E_{2c}	5021.02	163.23
E_{2d}	4826.07	178.90
E_{2e}	4973.61	140.17
E_{3a}	6376.45	163.06
E_{3b}	6325.69	92.82
E_{3c}	6265.10	225.88
E_{3d}	6435.18	221.18
E_{3e}	6474.74	171.82
E_{4a}	7689.09	156.18
E_{4b}	7557.04	125.38
E_{4c}	7562.18	148.50
E_{4d}	7562.45	111.99
E_{4e}	7469.49	157.51

６．適合度評価手法の検証

（１）全探索との比較

M = 4、N =16〜20 という小規模データ集合を用いて全探索と提案手法を比較する実験を行なった。提案手法で用いたパラメータの値を表－１に示す。本稿の実験に用いたワークステーションのスペックは Intel Xeon 2.5GHz CPU、32GB RAM である。

N=16〜20 のすべてにおいて、全探索と同じ評価値が提案手法でも得られた。各実験における計算時間を表－２に示す。全探索にかかる時間は N の増加とともに急激に増加するが、提案手法では計算時間の増加が非常に小さいことがわかる。

（２）実データによる評価

道央都市圏の世帯に関する観測データ集合 A (M=2, N=2000)、および以下の手順で生成された推計データ集合 E_j を用いて実験を行なった。

1) A からランダムに $j \times 500$ 個のデータを選択する。
2) 選択したデータの半数において、2 番目の構成員の年齢から 5 を減ずる。
3) 残りの半数のデータにおいて、2 番目の構成員の年齢に 5 を加える。

$j = 1$〜4 の各 j に関し、上記の手順を 5 回ずつ繰り返して 5 つの推計データ集合 E_{ja}〜E_{je} を生成した。L_p=5 とし、提案手法による評価値計算を各推計データ集合に関して 10 回ずつ繰り返したときの評価値の平均と標準偏差を表－３に示す。観測データにおける年齢を変化させたデータが多くなるほど、評価値が高くなることがわかる。

７．おわりに

既存研究では、観測データは入手できないと仮定されているために、個々のデータレベルにおける適合度を評価していない。本研究では、マイクロシミュレーションのための個々の世帯の推定手法が存在することを前提とした上で、個々のデータセットレベルでの適合度評価手法を計算手法とともに提案した。既存のマイクロシミュレーション都市モデルにおいて用いられるマイクロデータの多様な属性に対して適用可能な計測を実行するためには更なる研究が必要であるが、単純なケースに対して実行可能な計測手法を提案した。また、本研究の計測手法は都市モデリング以外の他の研究分野にも応用可能なものである。

なお本稿の内容に関しては11th International Conference on Computers in Urban Planning and Urban Management (CUPUM)において発表予定であることを付記する[13]。

本論文は、平成20〜21年度科学研究費補助金（基盤研究（B），課題番号：20360232，研究課題名：詳細属性情報を含む世帯の空間分布予測のためのマイクロシミュレーションシステム）の研究途中成果の一部を取りまとめたものである。ここに記して感謝の意を表したい。

参考文献

1) Wegener, M.: Overview of Land-Use Transport Models, Proceedings of CUPUM ’03, Sendai, CD-ROM, 2003.

2) Pritchard, D. R. and Miller, E.J.: Advances in Agent Population Synthesis and Application in an Integrated Land Use / Transportation Model, 88th Annual Meeting Compendium of Papers, Transportation Research Board, DVD, 2009.

3) Knudsen, D. C. and Fotheringham, A. S.: Matrix Comparison, Goodness-of-Fit, and Spatial Interaction Modelling. International Regional Science Review, Vol.10, No.2, pp.127-147, 1986.

4) Moriarty, D. E. and Miikkulainen, R.: Efficient Learning from Delayed Rewards through Symbiotic Evolution, Proceedings 12th International Conference on Machine Learning, pp.396-404, 1995.

5) Moriarty, D. E. and Miikkulainen, R.: Efficient Reinforcement Learning through Symbiotic Evolution, Machine Learning, Vol.22, pp.11-32, 1996.

6) Moriarty, D. E. and Miikkulainen, R.: Hierarchical Evolution of Neural Networks, Proceedings IEEE World Congress on Computational Intelligence, pp.428-433, 1998.

7) 大谷 紀子, 大和田 勇人: 共生進化に基づく帰納論理プログラミングの予測精度の向上, 人工知能学会論文誌, Vol.17, No.4, pp.431-438, 2002.

8) 大谷紀子, 志村正道: 共生進化に基づく簡素な決定木の生成, 人工知能学会論文誌, Vol.19, No.5, pp.399-404, 2004.

9) 大谷紀子, 貝原巳樹雄, 志村正道: ポリマー判別のための2段階判別決定木, 人工知能学会論文誌, Vol.21, No.3, pp.295-300, 2006.

10) Miyamoto, K., and Sugiki, N.: An Estimation Method of Household Micro-Data for the Base Year in Land-Use Micro Simulation, Proceedings of CUPUM ’09, Hong Kong, CD-Rom, 2009.

11) 佐藤浩, 小野功, 小林重信: 遺伝的アルゴリズムにおける世代交代モデルの提案と評価, 人工知能学会論文誌, Vol. 12,, No. 5, pp.734-744, 1997.

12) 宮本和明, 北詰恵一, 鈴木温: 世界における実用都市モデルの実態調査とその理論・機能と適用対象の体系化, 平成18年度〜19年度科学研究費補助金（基盤研究（C），課題番号：18560524）研究成果報告書, 2008.

13) Otani, N., Miyamoto, K., Sugiki, N.: Goodness-of-Fit Evaluation Method between Observed and Estimated Sets of Micro-Data in Land-Use Micro Simulation, Proceedings of CUPUM ’09, Hong Kong, CD-Rom, 2009.