# 土地利用マイクロシミュレーションにおける質的属性を含むマイクロデータの適合度評価*

Goodness-of-Fit Evaluation Method between Micro-Data Sets which contains Discrete Attributes 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) 2)]。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、自動車保有、居住地、住宅タイプ等の世帯全体に関わるものから、各世帯構成員の年齢および職業等にいたるまでの多くの属性が定義される。マイクロシミュレーションを実行するためには、基準年における各属性値をすべての世帯に対して求める必要がある。しかし、国勢調査や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、わが国はもとより多くの国において禁止されている。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、個別世帯等の一般にエージェントと呼ばれる行動主体に対して複数の属性の組合せを設定したデータ（以下マイクロデータ）を作成する。マイクロデータの作成手法としては、世帯タイプをまず設定した上で各タイプに属する世帯数を推計する IPF 法や、個々のエージェントのマイクロデータを作成する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[3) 4)]。しかし、マイクロデータを要素とする 2 つの集合間の適合度を評価する指標が存在しないため、異なる方法の再現性能の比較評価が行われていない。また、基準年データだけではなく、マイクロシミュレーションモデルによるシミュレーション結果の有効性も評価できないという課題がある。この課題解決を目的に本研究グループは研究開発を進めており、マイクロデータの属性を連続変量に限定した場合に対する評価方法を提案している[5)]。

本研究では、質的な属性を含むマイクロデータに適用可能な評価方法の提案を目的としている。連続変量のみに限定されたマイクロデータの評価方法[5)]を発展させ、世帯構成員の年齢、性別、世帯主との関係を属性とするマイクロデータの評価方法を提案する。

本稿では、はじ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と本研究で用いる計算アルゴリズムの原型、さらには連続変量に限定した先行研究に関して概観する。また、質的属性を含むマイクロデータ間にあらためて適合度を定義する。さらに、エージェント数の規模が大きい場合でも計算可能な近似値の探索手法を遺伝的アルゴリズム（Genetic Algorithm, 以下 GA）の一手法である共生進化に基づき構築する。まず、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 4000 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 らによる研究[3)]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、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いえない。

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、推定データ集合 $\hat{N}_{ijk}$ 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)のように評価できる[6)]。値が小さいほど適合度が高いことを示す。

* キーワード：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
** 正員、博士 (情報理工学)、東京都市大学環境情報学部情報メディア学科（〒224-8551 神奈川県横浜市都筑区牛久保西 3-1-1, TEL : 045-910-2938, E-mail : otani@tcu.ac.jp）
*** 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

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

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、得られた値の平均によって全体の適合度が与えられる。以上のように属性が 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 GA は、生物の進化過程を模倣したアルゴリズムである。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[7) 8) 9)]。共生進化では、部分解を個体とする集団と、部分解の組合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。全体解集団で部分解のより良い組合せを学習することで、1 集団を進化させる GA よりも多様な解候補からの探索が可能となる。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[10) 11) 12)]。

（3）連続変数のみのマイクロデータの評価指標

本研究の先行研究としてマイクロデータが連続変量のみで構成される場合の評価方法を提案している[5)]。まず、マイクロデータ間の距離を定義し、観測データ集合と推定データ集合の全エージェント間距離の総和の最小値を適合度とする。距離の総和は両集合のそれぞれのエージェントの組合せの数だけ存在する。エージェントの数が 20 程度以下の場合は総当たり法で最小値を求めることは容易である。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。そこで、規模が大きい場合でも計算を可能とするために、共生進化に基づいて近似値を探索する手法を提案している。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、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確認している。

## ３．適合度評価問題の定義

（1）定義

観測データ集合により近い推定データ集合を決定するために、観測データ集合に対する推定データ集合の適合度を算出することを適合度評価問題と定義する。適合度評価問題では以下の事項を前提とする。

- 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集合である。
- 属性は世帯構成員の年齢、世帯人数等の連続変量のほか、性別や世帯主との関係等の質的変数を含む。
- 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。
- 推定データは杉木ら[13)]等のマイクロデータ推定手法により提供される。

（2）表記

本稿では、世帯構成員の年齢、性別、世帯主との関係を属性とするマイクロデータを扱う。道央都市圏パーソントリップ調査で得られた 19394 世帯、46500 人分のデータを性別と世帯主の関係ごとに分類した結果、50 人以上が属するカテゴリは以下の 16 カテゴリであった。

| | | |
|---|---|---|
| 本人・男 | 本人・女 | 妻 |
| 息子 1 | 娘 1 | 父 |
| 息子 2 | 娘 2 | 母 |
| 息子 3 | 娘 3 | 息子の妻 |
| 孫・男 | 孫・女 | |
| 兄弟 | 姉妹 | |

いずれのカテゴリにも属さない世帯構成員は、男女とも 1 世帯に最大 2 名まで存在するため、「その他 1・男」「その他 2・男」「その他 1・女」「その他 2・女」の 4 カテゴリを加え、20 カテゴリで世帯構成員の性別と世帯主との関係を表現する。

観測データ集合 $A$ と推定データ集合 $E_j$ の要素は、それぞれ 1 世帯分のマイクロデータであり、式(2)および式(3)のような 20 次元ベクトルで表される。

$$A = \{\mathbf{a}_i = (a_{i1}, a_{i2}, \cdots, a_{i20}) \mid 1 \le i \le N\} \quad (2)$$

$$E_j = \{\mathbf{e}_i^j = (e_{i1}^j, e_{i2}^j, \cdots, e_{i20}^j) \mid 1 \le i \le N\} \quad (3)$$

ここで、$N$ は観測データ数、$a_{ik}$ は観測データ集合における $i$ 番目の世帯の $k$ 番目のカテゴリに属する世帯構成員の年齢、$e_{ik}^j$ は $j$ 番目の推定データ集合における $i$ 番目の世帯の $k$ 番目のカテゴリに属する世帯構成員の年齢を表す。当該カテゴリに属する世帯構成員がいない場合、$a_{ik}$ と $e_{ik}^j$ の値は 999 となる。$i$ を世帯番号と呼ぶ。

## ４．適合度と適合度算出方法

（1）適合度

先行研究[5)]では、推定データ集合 $E_j$ と観測データ集合 $A$ との適合度をデータ間距離和の最小値としていたが、本研究では計算速度向上のため、データ間距離の二乗和の最小値を適合度とする。ただし、計算時のオーバーフ

ローと不在世帯構成員の過度の影響を回避するため、推定データ $\mathbf{e}_k^j$ と観測データ $\mathbf{a}_i$ の距離の二乗 $Dis(\mathbf{a}_i, \mathbf{e}_k^j)$ は、成分の差の二乗の上限値 *DiffMax* を用いて式(4) により算出する。適合度 $Fit(E_j)$ は式(5) で定義される。

$$Dis(\mathbf{a}_i, \mathbf{e}_k^j) = \sum_{l=1}^{20} \min((a_{ik} - e_{kl}^j)^2, DiffMax) \qquad (4)$$

$$Fit(E_j) = \min_{\sigma \in S_n} \sum_{i=1}^{N} Dis(\mathbf{a}_i, \mathbf{e}_{\sigma(i)}^j) \qquad (5)$$

ここで、$S_n$ は集合 {1, 2,…, *N*} から集合 {1, 2,…, *N*} へのすべての全単射の集合を表し、$\sigma(i)$ は全単射 $\sigma$ による *i* の像を表す。

（2）適合度算出方法

式(5)による適合度計算は、*N*!種類の全単射から、距離の二乗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、距離の二乗和が最小となるように、観測データ集合の各要素を推定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。従って、先行研究[5]と同様に、共生進化に基づく手法が有効と考えられる。

先行研究[5]の手法において、全体解個体の進化がデータ間距離和に基づくところを、データ間距離の二乗和に基づいて進化するようにする。また、部分解個体の遺伝子が全体解個体に与える影響の度合を部分解個体の適応度に反映させるため、部分解個体 *p* の適応度 *fitness*(*p*) は式(6) で算出する。*w*→*p* は全体解個体 *w* が *p* を参照していること、*fitness*(*w*)は *w* の適応度、*avail*(*p*, *w*)は *w* の適応度算出に使用された *p* の遺伝子の個数を表す。

$$fitness(p) = \min_{w \to p}(fitness(w) / avail(p, w)) \qquad (6)$$

部分解と全体解の染色体表現、GA オペレータ、進化戦略など、他の部分はすべて同じように処理し、全体解集団の最良個体の適応度をデータ間距離の二乗和の最小値として出力する。

## ６．適合度評価手法の検証

（1）適合度の正確さと計算速度

*N* = 16～20 という小規模データ集合を用いて、提案手法の正確性と迅速性を検証する実験を行なった。各世帯構成員の在否および年齢をランダムに設定してデータ集合を生成し、枝刈りつき全探索と提案手法で適合度を算出した。枝刈りつき全探索とは、距離の二乗和を算出する過程で、確実に最小値をとらない組合せであることが判明した場合、以降の計算を行なわない全探索手法である。なお、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 であり、パラメータの値は**表－1**に示すとおりである。

*N* = 16～20 のすべてのデータ集合において、全探索と同じ適合度、すなわち正しい適合度が提案手法で得られた。各実験で計算に要した時間を**表－2**に示す。全探索では *N* の増加とともに計算時間が急激に増加するが、提案手法では計算時間の増加は微少であることがわかる。

**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 | 5000 |
| 部分解個体の染色体の長さ $L_p$ | 2 |
| 成分差の二乗の上限値 *DiffMax* | 99999 |

**表－2 計算時間**

| *N* | 全探索 [秒] | 提案手法 [秒] |
|---|---|---|
| 16 | 0.30 | 4.47 |
| 17 | 1.66 | 4.81 |
| 18 | 15.24 | 4.98 |
| 19 | 205.05 | 5.14 |
| 20 | 314.52 | 5.42 |

（2）実データによる評価

実データで得られる適合度を検証するための実験を行なった。道央都市圏パーソントリップ調査で得られた 19394 世帯分のデータのうち、4000 データを抽出して観測データ集合 *A* とした。また、*A* の一部を加工して生成したデータ集合 $E_{a1}$～$E_{a5}$、$E_{b1}$～$E_{b5}$ を推定データ集合とした。

データ集合 $E_{aj}$ は、以下のように *A* の世帯構成員の年齢を変更して生成した。

1) *A* からランダムに *j*×400 個のデータを選択する。
2) 選択したデータにおいて、年齢加工対象の世帯構成員をランダムに 1 名決定する。
3) 半数のデータにおいて、年齢加工対象世帯構成員の年齢から 5 を減ずる。
4) 残りの半数のデータにおいて、年齢加工対象世帯構成員の年齢に 5 を加える。

データ集合 $E_{bj}$ は、以下のように *A* の世帯構成員のカテゴリを変更して生成した。

1) *A* からランダムに *j*×400 個のデータを選択する。
2) 選択したデータのうち単身世帯のデータについて、世帯主の性別を変更する。
3) 選択したデータのうち非単身世帯のデータについて、年齢加工対象の世帯構成員を世帯主以外からランダムに 1 名決定し、世帯構成員のいない世帯主以外の属性に変更する。

上記 2 つの手順で *j* = 1～5 としてそれぞれ 5 つずつデータ集合を生成した。*j* が大きいほど観測データ集合からの変更度合が高いデータ集合となる。

提案手法による適応度計算を各推定データ集合に関し

て 10 回ずつ繰り返したときの適合度の平均と標準偏差を**表−3**、**表−4**に示す。この 2 表より、観測データからの変更度合が高くなるほど、高い適合度が算出されていることがわかる。また、その標準偏差は適合度指標値より 2 桁小さいことから、適合度指標としての信頼性が確保されていることがわかる。

**表−3 年齢変更データにおける適合度**

| データ | 平均 | 標準偏差 |
|---|---|---|
| $E_{a1}$ | 319894062 | 3562128 |
| $E_{a2}$ | 360017054 | 1564525 |
| $E_{a3}$ | 394823739 | 1479067 |
| $E_{a4}$ | 402314433 | 1130639 |
| $E_{a5}$ | 416581068 | 1940174 |

**表−4 属性変更データにおける適合度**

| データ | 平均 | 標準偏差 |
|---|---|---|
| $E_{b1}$ | 337396002 | 2592751 |
| $E_{b2}$ | 444136421 | 1548204 |
| $E_{b3}$ | 509090674 | 904534 |
| $E_{b4}$ | 633566939 | 12098938 |
| $E_{b5}$ | 682736551 | 3536506 |

## 7．おわりに

本研究では、属性情報を含むマイクロデータの 2 つの集合間の適合度を評価するための計測手法を提案した。本適合度を用いると、2 つのマイクロデータ集合のうち真の集合により近い集合を判定することができる。本研究で提案した手法は質的情報をも含むことから、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多様な属性を含むマイクロデータに対してもかなりの程度適用可能である。

本研究で開発したマイクロデータ集合間の適合度評価は次の視点から重要である。すなわち、この評価は、マイクロデータ推計手法の段階的な開発において、改良の度合を客観的に判断するための資料を提供するものである。なお、この場合は、開発用に既知のマイクロデータが用意されていることが前提である。そして、このように開発されたマイクロデータ作成手法はその有効性がある程度確認されていることとなる。そのため、実際の都市圏への適用に際して、その手法で求められたマイクロデータの信頼性に対して根拠を与えるもととなる。これは、複数の異なる推計手法が存在した場合の手法の選択においても同様に考えられる。さらに、より広範に考えると、本計算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と考えられる。

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）の研究成果の一部を取りまとめたものである。ここに記して謝意を表したい。

## 参考文献

1) Wegener, M.: Overview of Land-Use Transport Models, Proceedings of CUPUM'03, Sendai, CD-ROM, 2003.
2) 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.
3) 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.
4) 杉木直, 宮本和明, Varameth VICHIENSAN: 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, 土木計画学研究・講演集, 39, CD-ROM, 2009.
5) 大谷紀子, 杉木直, 宮本和明: 土地利用マイクロシミュレーションにおける初期マイクロ世帯データの推定手法, 土木計画学研究・講演集, 39, CD-ROM, 2009.
6) 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.
7) 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.
8) Moriarty, D. E. and Miikkulainen, R.: Efficient Reinforcement Learning through Symbiotic Evolution, Machine Learning, Vol.22, pp.11-32, 1996.
9) Moriarty, D. E. and Miikkulainen, R.: Hierarchical Evolution of Neural Networks, Proceedings IEEE World Congress on Computational Intelligence, pp.428-433, 1998.
10) 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.
11) 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.
12) 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.
13) 杉木直, 宮本和明, 大谷紀子, Varameth VICHIENSAN: 質的属性を含む初期マイクロ世帯データの推定手法, 土木計画学研究・講演集, 40, CD-ROM, 2009.