合成人口データの意義と利用可能性
―仮想都市データの有用性と秘匿性の評価から―

原田　拓弥*
松本　渉†
村田　忠彦‡

Significance and Availability of Synthetic Population: From the Evaluation of the Usability and Confidentiality of Virtual City Data

HARADA Takuya
MATSUMOTO Wataru
MURATA Tadahiko

匿名加工というだけでなく、属性の値の生成を通じて個体情報の秘匿性を確保する擬似的なデータとして、合成データの作成の研究がなされている。一方、合成人口と呼ばれる、合成データとは異なる経緯から開発されてきたデータがある。これは、シミュレーション研究を目的として、統計表として公開されている集計データから数値計算で生成される擬似的なミクロデータである。

合成人口は、合成データや一般用ミクロデータと同様、個票データの情報を秘匿しつつ、もとの集計データの性質をできるだけ維持している。さらに合成人口は、実在する個票データに基づいておらず当然に秘匿性があると考えられてきたため、定量的に評価が示されてこなかった。

そこで、本稿では、仮想都市データをもとに生成した合成人口について、原田他（2022）で示された結果を整理し直すことで有用性を確認し、その上でARD（Absolute Relative Difference）を計算することで秘匿性についての評価を示すことにより、合成人口の意義と利用可能性を主張する。

キーワード：合成人口、合成データ、有用性、秘匿性、ARD

The research of synthetic data examines how to synthesize pseudo micro data to ensure the confidentiality of individual information through the generation of attribute values, not only through anonymous processing methods. On the other hand, synthetic population data has been developed for different purpose from synthetic data. This is pseudo micro data generated by numerical calculations from aggregate data published as statistical tables for the purpose of simulation studies.

Like the synthetic data or Public Use Micro-data, the synthetic population has the same characteristics of the original aggregate data as much as possible while keeping the individual information confidential. Synthetic population has not been quantitatively evaluated because it is not based on real individual data and is considered to be confidential.

In this paper, we confirm the usefulness of the synthetic population generated based on virtual city data by adapting the results of Harada et al. (2022), and then present an evaluation of the confidentiality of the synthetic population by calculating ARD(Absolute Relative Difference), thereby asserting its significance and potential for use.

Keyword: synthetic population, synthetic data, usability, confidentiality, ARD

* 芝浦工業大学システム理工学部　Email: t-harada@shibaura-it.ac.jp
† 関西大学総合情報学部　Email: matsumo@kansai-u.ac.jp
‡ 大阪大学サイバーメディアセンター・大学院情報科学研究科　Email: tadahiko.murata.cmc@osaka-u.ac.jp

1. はじめに

2023 年現在において、公的統計のミクロデータを利用する方法として、オンサイト施設においてアクセスする「調査票情報のオンサイト利用」と調査票情報に匿名化措置が施された「匿名データ(anonymized microdata)」の 2 つがある。この他に、集計表から作成された擬似的なミクロデータである「一般用ミクロデータ」もあるが、対象となるデータは、全国消費実態調査（平成 21 年）と就業構造基本調査（平成 4 年～24 年）に限られている。

匿名データは、実際のミクロデータに基づいて、特定の個人又は法人その他の団体の識別ができないように加工されている。この匿名加工の手法とは別に、属性の値の生成を通じて個体情報の秘匿性を確保する擬似的なミクロデータとして、合成データ(synthetic data)の作成の研究も進められている（Rodríguez (2007)、伊藤 (2018) 、高部 (2022)）。

一方、合成人口（synthetic population）と呼ばれる、合成データとは名称は似ているものの全く異なる研究上の経緯から作成されているデータ（合成人口データ）がある[1]。合成データ作成の目的が実在のミクロデータが持つ分布特性を保持しながら、人工的なミクロデータを提供するにあたって個票の秘匿性を維持することにあるのに対し（高部（2022）、横溝・伊藤（2023））、ここでいう合成人口データとは、シミュレーション研究を目的として、集計表として公開されている国内外のセンサスデータ、あるいは標本データを含めた人口統計に基づいて数値計算によって生成される擬似的なミクロデータである（Wilson and Pownall（1976））。ただし、合成人口データは、一般用ミクロデータ同様、現実の世帯や個人の情報を含んでいない。そのため、合成時に使用していない統計量との整合性は保証できないため、実証研究に用いることはできない。

この点から合成人口は、現状の一般用ミクロデータと似ているが、現状では少なくとも 5 つの点で異なっている。

第一に、作成の目的である。一般用ミクロデータは、統計演習などの教育やデータテスト等の目的で作成されているが、合成人口は、社会シミュレーションの研究を行うために作成されている。社会シミュレーションの研究においては、個人の意思決定モデルを記述するため、性別、年齢、職業といった個人の属性を設定する必要があるという事情がある。

第二に、作成主体が異なる。一般用ミクロデータは、統計センターのような調査主体である公的な機関が作成しているが、合成人口の生成は、そういった社会シミュレーションに関心のある研究者によって手掛けられているのが通常である。

第三に、対象とするデータが異なっている。一般用ミクロデータとして、全国消費実態調査（平成 21 年）と就業構造基本調査（平成 4 年～24 年）が現状では提供されているが、合成人口が扱うのは国勢調査の結果に基づく人口の統計である。杉浦他（2019）のように、国勢調査以外の調査の結果に基づくデータを追加する場合もあるが、その場合でも国勢調査に基づく人口のデータは不可欠である。

この違いは一見本質的でないように思えるかもしれないが、実は四番目の相違点である作成方法の違いにつながっている。一般用ミクロデータは、全国消費実態調査と就業構造基本調査とでその作成方法は異なるが、全国消費実態調査については、相関係数や特化係数を反映しながら主に乱数を用いて作成されている[2]。合成人口についてもこれと共通する発想も見られるが、昨今はかなり異なっている。これは、国勢調査に関連して公表されている集計数値は多岐にわたり、判明している集計表の数値を全て制約条件に用いると複雑になることが関係している。結果として全ての制約条件を満たすように解析的に解くことは困難になるため、少なくとも現状の日本の合

[1] 本稿では、合成データと紛らわしいので、区別する都合上、抽象概念として取り扱う時には、合成人口と呼び、具体的に生成されたデータについては、合成人口データと呼んでいる。

[2] 総務省統計局，独立行政法人統計センター「一般用ミクロデータ利用の手引」
https://www.nstac.go.jp/sys/files/static/services/ippan/ippan_tebiki.pdf

成人口については、数値計算を繰り返し、集計データの数値との差を最小化するような方法で探索的にミクロデータを生成するという方法を採用することが有力となっている。

第五に、上記のような利用目的と作成の経緯から利用方法が異なる。一般用ミクロデータは 1 通りのデータセットとして利用することが可能であるが、合成人口を用いてシミュレーションを行う際には、複数のデータセットを用いることが求められる。

合成人口は、匿名データや一般用ミクロデータと部分的に類似した性質を持つにもかかわらず、主に社会シミュレーションに関心のある研究者によって取り扱われてきたという歴史的経緯があったため、公的統計研究の範疇の外に位置づけられてきた。しかし、昨今高等教育機関における教育だけでなく、オンサイト利用前のプログラム作成のためのテストとしての利用といった目的から合成データの研究が進んでいる現状（高部（2022）、横溝・伊藤（2023））を考慮すると、公的統計における活用可能性を念頭に合成人口の有用性と秘匿性についても確認しておく必要があるのではないだろうか。

合成人口は、合成データと出発点が単に異なっているというだけではない。あくまでその対象は人口の分布にあって、合成データのように広範囲のデータを想定していない。合成データにおいては、事業所や企業のデータを合成することも検討されているが（Chien et al. (2021)、高部（2022）、横溝・伊藤（2023））、合成人口における関心は個人や世帯であり、生成されるのは人口の分布を表す個体の集合である。賃金構造基本統計調査など国勢調査以外の結果を併用する場合もあるが（杉浦他（2019））、国勢調査のミクロデータを擬似的に構成するという目的が基本にある。このような限界はあるものの、合成人口は、合成データや一般用ミクロデータと同様、個票の情報が秘匿されつつ、もとの集計データの性質をできるだけ維持している点では共通する。合成人口の生成がリアルな個票データの再現を目指すことと、匿名データや一般用ミクロデータの作成を始めとするミクロデータの匿名化がリアルな個票データにおいて本来の個票データの性質をできるだけ維持しながら秘匿化処理を進めようとすることは、発想のスタートは異なるものの、実質的には同じようなゴールを目指す技法として位置づけることもできよう。

ただし、合成人口は、一般用ミクロデータと同様、実在するミクロデータに基づいていないだけでなく、作成する主体がシミュレーション研究者など、そもそもミクロデータの取り扱いとはかかわりの無い外部の人間が、入手可能な統計量からミクロデータを生成してきたという経緯がある。そのため、開示リスクがあると考えること自体定義上不自然であり、秘匿性は当然十分にあると想定されてきており、秘匿性についての定量的な評価が示されてきたわけではない。

そこで、本稿では、仮想都市データ[3]をもとに生成した合成人口についてすでに行われてきた有用性の結果を整理して確認するとともに、新たに秘匿性についての評価事例を示すことにより、合成人口の意義と利用可能性を主張する。

なお、一見秘匿性が自明のような状況で、秘匿性の評価を示すことが不思議に思われるかもしれない。この理由を補足しておこう。現在では、調査票情報を直接的に用いない方法により作成した擬似的なミクロデータがあることから、一般用ミクロデータは個別情報の秘匿を気にせずに利用できるようになっているが、真のミクロデータと無縁であるからと言って最初から公表可能であるとされていたわけではない。河野・和田（2018）によれば、当初統計センターが作成した擬似ミクロデータ作成用の高次元の集計表は詳細すぎて、秘匿性の観点から公表が困難であったとされている。高次元の集計表に基づいて作成している点で、擬似ミクロデータの場合と、公開されている集計データから生成される合成人口データでは、状況が大きく異なると考えられるが、合成手法では多数の統計表を重ね合わせて用いることによって、調査票と同様の世帯が生成される可能性が高くなることが知られている。また合成人口データでは、区間の値として同じプロフ

[3] 仮想都市データとは、原田他（2022）で提案された複数の統計表を用いて生成された仮想的な都市のミクロデータである。

ァイルの個人や世帯を増やす方法ではなく、属性値をそのまま表現するようにしているため、居住者のプライバシーが侵害されていないのかという懸念が寄せられたこともある。これが秘匿性評価を改めて示す理由である。

また有用性と秘匿性の評価について、本稿では現実のデータではなく、仮想都市データを利用している。確かに生成された合成人口について、オンサイト施設で有用性と秘匿性の評価を行った方が説得力があるかもしれない。しかし、公的統計のみならず、他の様々な分野のミクロデータへの応用可能性も考慮すると、オンサイト利用によらない方法に基づいて評価を行った方が、ミクロデータへのアクセスが必ずしも可能ではない、民間データなども含む様々なデータへの転用可能性も向上し、研究結果の汎用性がある。そこで、本稿では、現時点ではオンサイト施設での検証を行わず、仮想都市データを用いて合成人口の有用性と秘匿性の議論を行うことにした。

2. 合成人口とは

合成人口（synthetic population）は、合成データとは名称は似ているものの、もっぱら社会シミュレーション研究で利用することを目的として作成されている。合成データも合成人口も擬似的なミクロデータを生成するという点では共通しているが、前述したように、合成データ(synthetic data)が実在するミクロデータを提供するにあたっていかにして秘匿性を保つかという目的の延長上で検討されるのに対し、合成人口は、外部の研究者がシミュレーション研究での利用を目的として、公開されている集計データから実在するミクロデータに接近しようとするものであり、生成の出発点が異なっている。

合成人口の生成方法を大きく分けると、Synthetic Reconstruction Method (SR 法)と組合せ最適化（Combinatorial Optimization、CO）法の 2 つになる。

Synthetic Reconstruction Method (SR 法)は、Wilson and Pownall (1976)が提案した方法であり、国勢調査において必要とする属性のクロス集計から導かれる一連の条件付き確率に基づいて無作為抽出によって構成する。必要な全ての属性の相互依存関係を記述できることはほとんどないので、Iterative Proportional Fitting Procedure (IPFP)（Deming and Stephan (1940)）を使用することにより、国勢調査データの欠測データを他の情報源から導き出し、既存のデータに組み込んで、完全な結合確率分布を形成することになる。Wilson and Pownall (1976)を出発点とする初期の SR 法は、国勢調査の個票の部分母集団として提供されたサンプルに基づく集計量と各セルの値に基づいて、公開されていない国勢調査の各セルの値を推計するというものであった。Barthelemy and Toint (2013)は、それまでの SR 法と CO 法の両方を批判して、サンプルを使わない手法を提案した。特に SR 法について、個人に関する統計と世帯に関する統計のどちらかに適合する合成ができても、両方に適合させることが困難であることを IPFP の弱点であると指摘し、サンプルを使わない SR 法を提案した。同じ年に、Lenormand and Deffuant (2013)もサンプルを用いた場合とサンプルを用いない場合の比較を行い、サンプルを用いない場合の方が個人と世帯の両方に適合した合成ができることを示した。

一方、組合せ最適化法は、Williamson et al. (1998)によって最初に提示された。Williamson et al. (1998)は、複数のアルゴリズムの比較検討の結果、組合せ最適化法の Simulated-Annealing（SA）法が制約条件を満たす合成ミクロデータを生成できることを示した。当時提案された SA 法では、細分化されたデータセットから個票データをサンプルとして無作為に選び、観測されたサンプルに基づく統計量を、対象地域の統計量と比較し、差異を計測する。観測標本の世帯又は個人を入れ替え、差異が小さくなれば入れ替えを認め、差異が大きい場合には確率的に入れ替えの可否を決定する。そして比較する差異が十分小さくなるまで入れ替えを繰り返すのである。Huang and Williamson (2001)は、SR 法と CO 法を比較し、組合せ最適化によって生成されたデータセットのばらつきは、SR 法によって生成されたデータセットのばらつきよりもかなり小さいことを示し

た。しかし、当初、Williamson et al. (1998)によって示されたCO法も、ミクロデータの部分集団としての何らかのサンプルが必要であったし、計算時間が長いというデメリットもあった（花岡(2012)）。

こうした中、Harland et al. (2012)が、決定論的再重み付けアルゴリズム（Smith et al. (2009)、Smith et al. (2011)）、条件付き確率モデル（Birkin and Clarke (1988)）、SA法（Voas and Williamson (2000)、Williamson et al. (1998)）の比較を行うことにより、SA法が多様な規模の集団を合成できる有力な手法であることを示した。さらにChoupani and Mamdoohi (2016) が、IPFPに基づくSR法の問題点と有効性を検討して、（SA法をはじめとする）Simulation-Based Method (SBM) が合成手法として望ましい手法であると結論付けるようになる。

日本においても花岡（2016）が国勢調査のサンプルを用いたSA法を提案している。しかし、サンプルに依存して合成する方法は、そもそもサンプルを使えない国ではその手法は採用できないし、仮に匿名データのサンプルを使えたとしても、合成データの入れ替えが行われており、匿名データの一部がそのまま合成されたデータに残り、ミクロデータそのものを利用していることになる。そのため、日本でも、結局第三者提供に制約が生じてしまうと考えられる[4]。こういった危惧から、日本でもサンプルを用いない合成手法が試みられている。

例えば、池田他（2010）は、早い時期にサンプルを用いないSA法を提案した。公開されている複数の集計データに整合するように探索的解法を用いて人口を合成する方法である。福田・喜多（2014）、枡井・村田（2017）、原田・村田（2018）、原田他（2018）、Murata et al. (2017)によって池田他（2010）の改良を目指した派生的手法も提案されている。

3.　仮想都市と合成人口の生成

本節では、仮想都市で生成される合成人口の評価に先立ち、原田他（2022）の記述に基づき、仮想都市の生成手順の実例とその仮想都市における合成人口の生成の方法を説明する。

3.1 仮想都市の生成

仮想都市の生成では、①世帯の生成と②世帯構成員の属性の設定という2つの手続きにより仮想個票データを生成する。

① 世帯の生成

世帯数 H=500,000 の世帯について、家族類型、世帯人員数の順に決定する。家族類型と世帯人員数の設定にあたっては、家族類型別、世帯人員別世帯数が記載されている国勢調査人口等基本集計[5]表11（東京都）を用いた。このうち9種類の家族類型[6]別、世帯人員別の世帯数を算出し、生成する世帯の家族類型と世帯人員数を確率的に決定する。

② 世帯構成員の属性の設定

世帯の家族類型と世帯人員数の決定後、国勢調査人口等基本集計[7]表16-1（東京都）から

---

[4] 独立行政法人統計センターウェブサイト：「公的統計の二次的利用サービス」＞「匿名データの利用」＞「7各手続きの内容」＞「データの返却・データ消去、匿名データを利用して作成した成果の報告」
（https://www.nstac.go.jp/use/archives/anonymity/#ano06）
「中間生成物（匿名データの個々の情報を判別できるもの。）を速やかに復元できないように消去又は廃棄した上で、すべての提供された匿名データを統計センターに返却します。」

[5] 総務省統計局（2016）. e-Stat 平成27年度国勢調査　人口等基本集計（男女・年齢・配偶関係、世帯の構成、住居の状態など）https://www.e-stat.go.jp/stat-search/files?page=1&toukei=00200521&tstat=000001080615&tclass1=000001089055

[6] 単独世帯、夫婦とひとり親世帯、夫婦と両親世帯、夫婦と子供世帯、夫婦のみ世帯、夫婦と子供とひとり親世帯、夫婦と子供と両親世帯、女親と子供世帯、男親と子供世帯の9つである。この9種類の家族類型で日本全国の約95%の世帯が分類される。

[7] 総務省統計局（2016）. e-Stat 平成27年度国勢調査　人口等基本集計（男女・年齢・配偶関係、世帯の構成、住居の状

家族類型別、男女別人口と家族類型別・男女別、年齢別人口を、又同表 17 から夫婦の年齢差別夫婦の数を算出する。人口動態職業．産業別統計[8]保管表出生表 1、同表 2 から、父の年齢別出生数と母の年齢別出生数を算出する。

単独世帯の構成員の年齢と性別は単独世帯の男女別、年齢別の人口をもとに確率的に設定する。単独世帯以外の世帯は夫婦の両親・ひとり親夫婦、男親・女親、子供の順に年齢を設定する。その際に、ひとり親と子供に該当する構成員は性別も設定する。

世帯に夫婦の両親が存在する場合は夫の父、夫の母の順に年齢を設定する。夫の父の年齢を夫婦のみ世帯における男女別の人口をもとに確率的に設定する。夫の母の年齢は設定した夫の年齢と夫婦の年齢差別夫婦の数をもとに確率的に設定する。世帯に夫婦のひとり親が存在する場合は、当該家族類型の男女別、年齢別、人口をもとに年齢と性別を確率的に決定する。世帯に夫婦とその両親、もしくはひとり親が存在する場合は、夫の母の年齢と母の年齢別出生数をもとに夫の年齢を確率的に設定し、妻の年齢を夫婦の年齢差別夫婦の数をもとに確率的に設定する。

世帯に夫婦の両親が存在しない場合は、夫婦の両親と同様に夫、妻の順に年齢を決定する。一方で世帯に男親もしくは女親が存在する場合は、当該家族類型の男女別、年齢別、人口をもとに年齢と性別を確率的に決定する。世帯内に子供と妻もしくは女親が存在する場合、母の年齢別出生数を世帯に子供と夫もしくは男親が存在する場合、父の年齢別出生数を用いて年齢を設定する。その後、当該家族類型の男女別人口をもとに性別を決定する。世帯構成員の年齢を設定した際に、年齢が 0 歳未満や、101 歳以上となった場合は、構成員 1 人目から再度年齢と性別を設定する。

なお、親より年齢の高い子供や、結婚できない年齢の夫婦が発生しないように、親子年齢差と夫婦年齢差の集計表を用いて、統計上存在する親子や夫婦の年齢差になるように仮想的な個票データを生成している。

以上の手続きから、男性 497,684 人、女性 501,348 人、合計 999,032 人の 500,000 世帯の仮想都市が生成されている[9]。

仮想都市についての合成人口を生成するに先立ち、後述する要素 2 に相当する集計表をあらかじめ作成する。作成される集計表は、父子の年齢差、母子の年齢差、夫婦の年齢差、9 種類の家族類型別の男女別人口分布（18 種類）の合計 21 種類である。

3.2 合成人口の生成

通常、合成人口の生成は、公開されている国勢調査等の統計表における複数の集計表に適合させるため、個人の年齢や親子の年齢差といった集計データと生成された合成人口データから作成する集計データの差を計算し、SA 法を用いて探索的に誤差を最小にする合成人口を生成する。ここでは、公開されている統計表の代わりに、あらかじめ用意した仮想都市についての国勢調査の結果に擬した 21 種類の集計表を用いて、合成人口データを生成する。

合成手法は、1 初期世帯生成法、2 最適化に用いる統計表、3 目的関数、4 最適化の 4 つの要素から構成されている。日本では、池田他（2010）の手法を出発点として、枡井・村田（2017）や原田他（2018）の手法等の派生形が提案された。ここでは、これらの派生の延長にある Murata et al. (2017)の方法を例として概説する。Murata et al. (2017) の手法による合成人口の生成では、仮想都市の生成で用いられた場合と同様、9 種類の家族類型が対象となっている。

態など）https://www.e-stat.go.jp/stat-search/files?page=1&toukei=00200521&tstat=000001080615&tclass1=000001089055

[8] 厚生労働省（2018）. e-Stat 人口動態職業・産業別統計 2015 年度. https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&toukei=00450011&tstat=000001028897&year=20151&tclass1=000001053122

[9] 家族類型別、世帯人員別世帯数の詳細については、原田他（2022）参照のこと。

＜要素 1：初期世帯生成法＞

合成人口データの初期生成では、世帯数 H だけ世帯を生成する。世帯の生成時、家族類型別世帯数を出現確率とする乱数によって、世帯の家族類型を確率的に決定する。世帯の家族類型を決定した後、家族類型別、子供の数別、世帯の割合の集計表を用いて、世帯に居住する子供の数を確率的に決定する[10]。その後、その世帯の構成員の属性を設定する。構成員の年齢については、男女別、年齢別人口の割合に従って確率的に設定する。性別は、まず夫であれば男性、妻であれば女性など、性別を世帯の役割の属性に応じて設定する。次に男女別の人口との相違を発生させないために、対象の家族類型の男性の人口と女性人口から、夫、妻、両親（ひとり親を除く）の人口を差し引き、残された構成員の性別を、この男女比に基づいてランダムに設定する。

＜要素 2：最適化に用いる統計表＞

最適化にあたっては、次の統計表（計 21 種類）との差を最小化する。

（１） 父子の年齢差
（２） 母子の年齢差
（３） 夫婦の年齢差
（４） 9 種類の家族類型別の男女別人口分布（18 種類）

＜要素 3：目的関数＞

以下の目的関数 (1) を用いて最適化を行う。

$$g_s(A) = \sum_{j=1}^{G_s} \left| v_{sj}(A) - Round\left(r_{sj} \times m_{sj}(A)\right) \right| \qquad (1)$$

ここで、A は合成人口データ、S は統計表の数、G_s は統計表 s の項目数、v_{sj} は統計表 s の項目 j の値、r_{sj} は統計表 s の項目 j の割合、$m_{sj}(A)$ は統計値 r_{sj} の補正値である。

＜要素 4：最適化の手続き＞

Step 1　合成人口データを初期生成
Step 2　探索回数が規定数に達すれば探索を終了
Step 3　以下の手順により近傍解を作成する
　Step 3-1　合成人口データ内の個人をランダムに 1 人選択
　Step 3-2　Step 3-1 で選択された個人と同じ家族類型かつ性別の個人をランダムに 1 人選択
　Step 3-3　Step 3-1 と Step 3-2 で選択した個人の年齢を交換
Step 4　解の遷移判定
Step 5　探索回数を更新して SA の探索パラメータを更新
Step 6　Step 2 の処理に戻る

4.　合成人口データの有用性

本節では、前節で説明したような仮想都市において、Murata et al.(2017)の方法に基づいて生成された合成人口が、もとの仮想都市のデータとどの程度近似しているかの比較した手順と結果について、原田他（2022）の記述に若干の加筆修正を補いながら説明する。

[10] 子供の数については、国勢調査人口等基本集計表 11 を活用し、世帯人員数と世帯構成から推定できる場合は、可能な限り確定的に決定している。世帯人員数についても、国勢調査人口等基本集計表 7（世帯人員別世帯数）、表 11（家族類型別、世帯人員数 7 人以上の世帯数）、表 16-1 もしくは表 16-2（家族類型別、人口）の情報をもとに制約条件を定めた上で合成を行っている。

仮想都市の合成人口が、もとの仮想都市のデータとどの程度近似しているかについての比較は、どのような世帯に誤差が発生しているかを明確にしやすくするため、次の手順に従って、家族類型別に年齢差を計測した。

Step 1　合成人口データ（*A*）と仮想個票データ（*B*）の両方について、以下の処理を実施する。
- Step 1-1　各世帯を 9 種類の家族類型、世帯人員数、ひとり親・子供の性別の内訳をもとに分類する。
- Step 1-2　各世帯内の構成員を年齢、性別、世帯内の役割を用いてソートする。

Step 2　合成人口データ（*A*）と仮想個票データ（*B*）において類似する世帯を以下の処理により探索する。
- Step 2-1　探索する世帯間の差 d を 0（世帯間の差なし）と設定する。
- Step 2-2　合成人口データ A の i 番目の世帯 a_i との差が d である世帯 b_j を仮想個票データ B から探す。
- Step 2-3　世帯 a_i との差が d である世帯 b_j が存在する場合、世帯 a_i と b_j をそれぞれの合成人口データから取り除き、i に 1 加算し、Step 2-2 に戻る。
- Step 2-4　探索する世帯間の差 d が既定値以下の場合、d に 1 加算し、i を 0 に設定し、Step 2-2 に戻る。
- Step 2-5　Step 1-1 で分類された全ての分類を探索するまで Step 2-1 に戻る。

Step 1-1 の世帯の分類では、まず、合成対象の 9 種類の家族類型に加えて、ひとり親が存在する「夫婦とひとり親」世帯と「夫婦、子供とひとり親」世帯はひとり親の性別で細分化する。つぎに、子供が存在する家族類型は子供の男女の数によって細分化する。たとえば、「夫婦と子供」世帯において、子供が 2 人存在する世帯は、男性の子供 2 人、男女 1 人ずつ、女性の子供 2 人の 3 種類に細分化する。

Step 1-2 の世帯内の構成員のソートでは、まず、世帯内の役割が夫、妻、夫の父、夫の母、子供となるように並べ替える。つぎに、子供の役割をもつ構成員を性別が男性・女性となるように並べ替える。最後に、子供の役割をもつ構成員を男女別に年齢の降順となるように並べ替える。

Step 2-1 の世帯の比較における d とは、合成人口のとある世帯と仮想都市のある世帯の構成員の役割（夫、妻、子（男性）、子（女性）、夫の父、夫の母）別に年齢の差の絶対値を合算したものである。初期値においては、全構成員において年齢が完全に一致しているので d=0 である。

例えば、以下の世帯 *X* と世帯 *Y* から d を計算する場合は、夫の年齢の差が 1、妻の年齢の差が 0、子 1（男）の年齢差が 1、子 4（女）の年齢差が 1 で合計 3 となる。

世帯 *X*
- 夫：40 歳
- 妻：39 歳
- 子 1（男）：10 歳
- 子 2（男）：9 歳
- 子 3（女）：8 歳
- 子 4（女）：8 歳

世帯 *Y*
- 夫：41 歳
- 妻：39 歳
- 子 1（男）：11 歳
- 子 2（男）：9 歳
- 子 3（女）：8 歳
- 子 4（女）：7 歳

Step 2-2 で用いる世帯間の差は、式(2)で計算する。

$$h(a_i, b_j) = \sum_{m=1}^{M} \left| a_{i,m}^{age} - b_{j,m}^{age} \right| \quad (2)$$

ここで、M は a_i と b_j の世帯人員数、$a_{i,m}^{age}$ は世帯 a_i の m 番目の構成員の年齢である。合成人口データ（A）と仮想個票データ（B）は世帯の家族類型、世帯人員数、ひとり親・子供の性別をもとに分類した後、年齢、性別、世帯内の役割によって世帯構成員がソートされている。そのため、$h(a_i, b_j) = 0$ の場合、本稿が比較対象とする年齢、性別、世帯内の役割、家族類型、世帯人員数については、世帯 a_i と b_j は完全に一致している。Step 2 の操作を Step1-1 の全ての分類に対して実施することで、合成人口データ（A）と仮想個票データ（B）の一致度を世帯間の差 d の値ごとに算出することができる。なお、合成人口データと仮想個票データで Step 1-1 で細分化した世帯数が異なる場合、超過した世帯は比較不可能な世帯として集計した。

このような手順で、個票データと合成人口データで絶対値誤差を計算するが、これを 10 個の合成人口データに対して実施し、差 d ごとの世帯数の平均を計算した結果が、表 1 である。この結果から、世帯間の差の一致割合のうち、d=0 の割合は 76.3%である。既存の合成人口データ同士の比較では、最も高い割合を計上している（原田他 (2022)）。

表 1　合成データと個票データの比較結果

d	平均世帯数	%
0	42379.1	76.3%
1	4538.6	8.2%
2	2272.8	4.1%
3	1210.6	2.2%
4	748.7	1.3%
5	518.5	0.9%
6	388.3	0.7%
7	310.1	0.6%
8	258.8	0.5%
9	222.2	0.4%
10--	2615.5	4.7%
others	92.4	0.2%

この結果の有用性についての絶対的な解釈のための明確な基準はないが、そもそもミクロデータが 100%一致することはありえないことであるし、秘匿性を考慮すればそもそも完全に一致するのも好ましくない。その点からいえば、現在の合成人口で 8 割弱が一致しているということは、もとのデータが十分に再現されており、有用性が認められる結果である。

5.　合成人口データの秘匿性

本節では、仮想都市についての合成人口データの秘匿性を確認する。秘匿性については、原データと合成人口データとの対応が想定されたことがなかったため、これまで評価を行われてこなかった。そこで、合成人口ではなく、合成データの秘匿性の評価の事例にならって評価を試みる。具体的には、Kim et al.(2021)および横溝・伊藤 (2023) 同様、絶対相対差分 (Absolute Relative Difference＝ARD)と呼ばれる属性漏洩に関する評価指標を利用する。なお、ARD は、以下の式 (3) で表される。

$$ARD = \frac{|\hat{L}-L|}{L} \qquad (3)$$

ここで、Lと$\hat{L}$はそれぞれ、原データおよび合成データに関する属性値の最大値である。ARD は、原データと合成データに含まれる属性値の最大値からの乖離を評価する指標である。横溝・伊藤（2023）は、この ARD について単変量から計算するだけでなく、多変量を用いて層ごとの露見リスクを総合的に評価する層化平均 ARD（stratified average ARD）を提案している。キー変数の全組み合わせにおいて、それぞれ原データと合成データの最大値の乖離を計算し、その平均を計算する（横溝・伊藤（2023））。

本稿においては、原田他（2022）で示されている仮想都市データについて、Murata et al. (2017)に基づいて合成人口データ（1 試行）を生成したのちに、次の計算を行う。

0) 合成人口データ全体に対して計算する（0 属性 ARD）
1) 合成に使用した統計ごとに層化して計算する（1 属性 ARD）。
2) さらに 2 つの属性を組み合わせたクロス集計についても層化した計算を行う（2 属性 ARD）。
3) 属性数がそれ以上用いることができる場合、たとえば最大 n 個の属性数が用いることができる場合は、3 属性 ARD、…、n 属性 ARD を計算する。
4) 0 属性（全体）の ARD、1 属性 ARD、2 属性 ARD、…、n 属性 ARD の平均を求める（層化平均 ARD）。

横溝・伊藤（2023）では、売上、付加価値額、給与総額、減価償却費の 4 つの経理項目について上記の層化平均 ARD を計算している。これらは企業固有の値である。合成人口データの場合には、世帯単位での合成を行っていることから、最年長者の年齢、構成員数、夫婦年齢差、父子年齢差、母子年齢差などについて ARD を計算することが考えられる。ただし、単身世帯の場合は、夫婦年齢差、父子年齢差、母子年齢差といった属性は扱うことができない。また所得などのデータも世帯固有の値であるが、仮想都市では所得の設定をしていないので計算できない。

また、合成人口においては、年齢を変更（交換）して合成していることから個人単位での ARD の計算もできる（ただし、計算可能な属性は年齢のみ）。そこで、個人単位での層化平均 ARD を計算すると表 2 のようになる。なお、家族類型、世帯人員数、性別、世帯内の役割の 4 変数で層化を行ってそれぞれの ARD を計算している。

表 2　個人（年齢）についての層化平均 ARD の算出

層化する変数の数	0 属性 ARD	1 属性 ARD	2 属性 ARD	3 属性 ARD	4 属性 ARD	層化平均 ARD
ARD	0.000	0.015	0.058	0.108	0.191	0.074

この ARD の計算では、年齢しか扱っていないため、最大値がたかだか 100 歳である。そのため、家族類型ごとにわけた ARD や層化平均 ARD を計算したところかなり小さな値となった。日本の国勢調査に基づく合成人口の特徴ともいえる世帯内の年齢差などの計算は反映されていない。年齢差などの統計については、その統計値に合わせる形で合成が行われているので、最大値の差だけをみている ARD では、基本的にその差が小さくなる（合成時の最大値の上限と、実際の値の上限値の差が出る程度）と思われる。

そこで、世帯内の年齢差などを反映されている世帯をもとにして、ARD を算出した。その結果が表 3 である。本稿では、年齢、夫婦年齢差（夫の年齢≧妻の年齢、夫の年齢＜妻の年齢）、父子

年齢差、母子年齢差の 5 つの属性について、それぞれ層化平均 ARD を計算している。なお、家族類型、世帯人員数の 2 変数で層化を行ってそれぞれの ARD を計算している。

表 3　世帯についての層化平均 ARD の算出

ARD を計算する属性	0 属性 ARD	1 属性 ARD	2 属性 ARD	層化平均 ARD
年齢（最大）	0.000	0.025	0.104	0.043
夫≧妻の夫婦年齢差（最大）	0.000	0.291	0.410	0.234
夫＜妻の夫婦年齢差（最大）	0.233	0.289	0.305	0.276
父子年齢差（最大）	0.054	0.198	0.201	0.151
母子年齢差（最大）	0.014	0.102	0.109	0.075

年齢について個人の場合よりも層化平均 ARD が小さくなっているが、これは層化に用いる属性の数が減少していることが関係していると思われる。その一方で、夫婦年齢差という統計量では、他の属性項目よりも層化平均 ARD の値が計上されている。特に夫≧妻の場合よりも、妻＞夫の場合の方が高い値となっている。「人口動態統計」（厚生労働省）が示す婚姻状況[11]によれば、夫が妻よりも年上であることの方が、妻が夫よりも年上であることよりも多数派であることが日本社会の現状である。そのため、妻が夫よりも年上であることは、特徴的な現象になりえる。このことを考慮すると、当該属性項目において秘匿性が（相対的ではあるが）強く出ているということは、好ましい性質であり、合成人口データは秘匿性が保たれる必要度が高い属性について秘匿性を保つ性質を持っていると予想される。

ただし、全体として層化平均 ARD が、それほど大きな値をとらないように見えるかもしれない。これは以下のような理由による。

本来、ARD の考え方は、値の大きな個体をどこまで秘匿できているのかの把握にあると思われるが、人口についての統計の場合、狭い地域での値が突出した個体については、X 歳以上という形ですでに秘匿されている。合成人口では、例えば 100 歳以上については、全て 100 歳とするというような形で、合成しているのである。そのため、ある地域に実際に 115 歳の人がいたとすると、本来であれば、その差は 15 歳という形で生じるはずである。ただし、本稿で示したのは仮想都市の事例であり、仮想都市ではそもそも 100 歳までの人しか存在せず、もともと極端な大きな値が生じないのである。したがって、秘匿したという効果が ARD では生じないのである。

ARD についてもう一つ考えておきたいこととしては、統計の対象地域の広さ（粒度）である。これによって、ARD の値は変わりえるし、ARD で秘匿できているものについても変わってくる。企業を対象としたミクロデータにおける秘匿性では、おそらく巨大企業の属性値を推計できないようにする、ということが念頭に置かれているので、最大値の絶対値誤差という指標の意義は大きい。

しかし、個人や世帯を取り扱う合成人口の場合、年齢に関しても最大値はせいぜい 120 歳程度であり、企業のデータベース（売上等）と比較して、そもそも極端に突出した大きな値があまりない。横溝・伊藤（2023）における層化平均 ARD（0.1 を下回る場合もあるが、手法によっては 1.4 程度になる）と比べて小さな値となったのである。層化平均 ARD の値が小さくとも、粒度が大きい（人口が多い）自治体であれば、「どこかにそういう世帯はあるだろう」という程度の話で

[11] 政府統計の総合窓口（e-Stat）（https://www.e-stat.go.jp）「令和４年人口動態統計　上巻　婚姻　第９．１４表　初婚夫妻の年齢差別にみた年次別婚姻件数及び百分率（各届出年に結婚生活に入り届け出たもの）」によれば、2022 年の時点で夫が年上 53.4%、妻が年上 24.3%となっている。

すむ一方で、ある特定の狭い地域に限定されてその世帯がいるということがわかれば、プライバシーが侵害されている（秘匿性が担保されていない）という事態になる。

仮想都市の研究では、50 万世帯（100 万人弱）を対象にしており、比較的粒度が大きい（日本の都道府県のうち 10 県の人口は 100 万人以下）。秘匿性を考えるときには、対象の粒度によってその意義が変わってくると考えられ、そのあたりの粒度付きの秘匿度については、今後の検討課題である。

6. むすびにかえて

本稿では、まず合成人口についての全般的経緯をレビューし、日本の現状に適した手法として、サンプルを用いない SA 法が有力であることを論じた。その上で、仮想都市とそれに基づく合成人口データの生成方法を概説するとともに、合成人口データの研究において示されてきた結果を検討することにより、合成人口の有用性を示した。同時に、これまで全く検討されてこなかった秘匿性について、層化平均 ARD を計算することにより、定量的な評価ができることを示した。

その結果、合成人口データの有用性については、仮想都市データと合成人口データとの間で世帯間の差がない（完全一致する世帯の）割合が（10 回の試行の平均で）8 割程度あることから確認できる。この 8 割という数字を高いと捉えるかは議論があるかもしれない。しかし、もともと社会シミュレーション研究で用いられてきた合成人口データは、事実としてはとらえられていないので実証研究で用いる際のエビデンスにはならないということは当然の前提とされてきた。そのため、社会シミュレーションにおいては、一回の合成に依拠せず、複数の社会を生成することで、どのような未来が起こりえるかをシミュレートしてきた。一種の予測のように聞こえるが、そもそも蓋然性の高いシナリオを予測することを目的としていない。社会シミュレーションの結果に限らず、予測は、予測そのものを社会に示すことで人々の行動が変わり、予言の自己成就（Merton (1936)）を起こす可能性もありえる。そのため、社会シミュレーションにおいては、「複数の仮想的な社会によって、高い確率で起こるシナリオを予測する」ことにあるのではなく、むしろ「最良の未来」と「最悪の未来」が起こりうることに着眼し、悪い未来を回避して、良い未来を導くことにあった。実際、AED の最適配置を目的に合成人口データの活用が検討されており、その主眼は夜間患者の救急搬送問題におけるリスクマネジメントにある（市川 (2018)）。前提としている利用方法からいえば、合成人口の 8 割が固定的に完全一致している現状は、悪い数字とは言えない。無論、合成人口データの有用性を向上させるために、統計表や目的関数、手法などを改良する余地はある。今後の課題としたい。

一方、合成人口データの秘匿性という点については、横溝・伊藤（2023）で定義された層化平均 ARD の値を確認した。合成人口の生成にあたって題材としたデータが今回は実データではなく、仮想都市の人口データであったため、年齢を始めとする属性項目を中心に、全体としては高い値にならなかったが、世帯についての層化平均 ARD の算出結果において、夫婦年齢差、とりわけ夫＜妻の場合の夫婦年齢差については、比較的高い値をとることから、秘匿性が必要な属性については秘匿性を保つ性質があることが期待できることが分かった。

本稿では、仮想都市データという現実にないデータを用いて合成人口データを評価するという手段であったため、大きな値を秘匿するという効果を十分に発揮できなかったように思われる。合成人口の生成結果については、有用性はともかく、秘匿性については、可能な範囲で現実の国勢調査のミクロデータと照合することの必要性を痛感することとなった。現実のミクロデータを用いて秘匿性の評価を行う場合は、地域の変数を層化の変数として用いるなどより精緻な層化平均 ARD を算出することが可能となると考えられるからである。ただし、現実のミクロデータを用いて合成人口の評価を行う場合は、評価後の合成人口の取り扱いが、ミクロデータの持ち出しに近いものとして誤解されないような手順を考える必要がある。

また ARD は最大値に特化した指標であると考えられ、秘匿性の評価としては、十分ではない可能性もある。例えば、「オンサイト利用における分析結果等の提供に関する標準的なチェック内容の解説と例」（p.3）[12]においては、度数表の持ち出しの際に、行計又は列計の 90%超を占めるセルがないこと（加重なし・ありともに）が求められている。特定の地域などの特性を推定させないためであるが、このような観点での秘匿性評価は今後の課題である。

合成人口データの研究は、合成データと異なる経緯で開発が進められてきた。部分合成データ (partially synthetic dataset) (Little(1993))の生成の考え方に見られるように、ミクロデータの存在を前提として、一部のレコード群に含まれるセンシティブな値を秘匿するために欠測を発生させて擬似的な値を補完するという発想は、基本的には補定法（imputation methods）の延長にある。一方、初期の SA 法や SR 法のように、サンプルとして提供されている一部のミクロデータを援用しつつ、集計表からミクロデータ全体を生成する合成人口の手法は、部分合成データの作成方法と極めて似ているがその発想の出発点が異なっている。

しかし、昨今の合成人口の生成においては、サンプルを利用して生成されることなく、集計表の数値だけを利用する。実のところ、Little(1993)の論文が掲載された *Journal of Official Statistics* の同じ号で Rubin(1993)が提唱した完全合成データ (fully synthetic dataset) (Rubin(1993))においては、対象となる全てのレコードの中で欠測値を含む属性群に対して擬似的に値を生成される。またこの補定法が synthetic なデータセットの作成と呼ばれ、英語的にもデータの合成とほぼ同じであるし、Rubin(1993)は、複数の補定データセットの作成を提案しており、（趣旨は異なるが）現在の合成人口データの取り扱いも類似する点がある。合成データは、取り扱う範囲が人口に限定される合成人口よりも、対象となる範囲は広いという点では明確に異なるが、両者の類似点を考慮すると、その知見の交換は有益なはずであるが、盛んであるとは言えない。結果として、合成人口と合成データのそれぞれの研究で培われた知見は十分に整理されていないように思われる。両者において類似する諸概念の整理も今後の重要課題ではないだろうか。

謝辞

本研究では、2020 年度関西大学拠点形成支援経費、学際大規模情報基盤共同利用・共同研究拠点研究課題（jh230011）、HPCI 共用ストレージ（hp230467）、JST 未来社会創造事業（JPMJMI23B1）、JSPS KAKENHI Grant Number JP20K10362、JP22K14442、ならびに 2023 年度関西大学学術研究員制度の御支援をいただきました。また、２人の査読者の御指摘により、本論文の論点をより明確にすることができました。ここに記して感謝申し上げます。

参考文献

［1］　池田心, 喜多一, 薄田昌広 (2010), 「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 回システム工学部会研究会「社会シミュレーション＆ サービスシステム・シンポジウム」』, 11-14.

［2］　市川学 (2018),「医療分野におけるリスクマネジメントー地理情報分析と社会シミュレーション技術を用いた検討ー」, 『計測と制御』, 57(6), 407-412.

［3］　伊藤伸介(2018),「公的統計ミクロデータの利活用における匿名化措置のあり方について」, 『日本統計学会誌』, 47(2), 77-101.

［4］　河野真理子, 和田かず美 (2018),「ミクロデータ分析のための演習用教材の作成方法：一般用ミクロデータ詳細品目版及び擬似ミクロデータによる事例」, 『統計研究彙報』, 75, 総務省統計研修所, 61-80.

[12] https://www.e-stat.go.jp/microdata/sites/default/files/share/data-use/onsite_check.pdf

［5］ 杉浦翔, 村田忠彦, 原田拓弥 (2019), 「賃金構造基本統計調査に基づく合成人口の労働者への就業属性別の所得の割当て」, 『システム制御情報学会論文誌』, 32 (2), 69-78.
［6］ 高部勲 (2022), 「合成データの考え方に基づく公的統計疑似ミクロデータの作成方法の検討」, 『統計研究彙報』, 79, 総務省統計研修所, 111-129.
［7］ 花岡和聖 (2012), 「公的統計「匿名データ」を用いた小地域単位での地理空間分析の可能性―空間的マイクロシミュレーションによる地理的な合成ミクロデータの生成―」, 『人文地理』, 64(3), 195-211.
［8］ 花岡和聖 (2016),「全国版の小地域マイクロデータの構築と災害分析への活用」,『地域安全学会論文集』, 29, 247-255.
［9］ 原田拓弥, 村田忠彦 (2018),「並列計算を用いたSA法による都道府県レベルの大規模世帯の復元」, 『計測自動制御学会論文集』, 54(4), 421-429.
［10］ 原田拓弥, 村田忠彦, 枡井大貴 (2018), 「家族類型と世帯内の役割を考慮したSA法による大規模世帯の合成」, 『計測自動制御学会論文集』, 54(9), 705-717.
［11］ 原田拓弥, 村田忠彦, 高橋真吾 (2022), 「仮想都市の統計情報による合成人口データの評価」, 『計測自動制御学会論文集』, 58 (7), 345-353.
［12］ 福田純也, 喜多一 (2014), 「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.
［13］ 枡井大貴, 村田忠彦 (2017), 「統計データからの市民の属性復元のための進化計算とSAによる2段階最適化」, 『システム制御情報学会論文誌』, 30(6), 216-227.
［14］ 横溝秀始, 伊藤伸介 (2023), 「合成データの生成手法の有効性に関する定量的な評価：事業所・企業系のミクロデータを用いて」, 『統計研究彙報』, 80, 総務省統計研修所, 97-116.
［15］ Barthelemy, J. and Toint, P. L. (2013), Synthetic Population Generation Without a Sample, *Transportation Science*, 47(2), 266–279.
［16］ Birkin, M., and Clarke, M. (1988), Synthesis—A Synthetic Spatial Information System for Urban and Regional Analysis: Methods and Examples, *Environment and Planning A: Economy and Space*, *20*(12), 1645-1671.
［17］ Chien, C. H., Welsh, A. H., and Moore, J. D. (2021), Synthetic Business Microdata: an Australian example, *Journal of Privacy and Confidentiality*, *10*(2), https://doi.org/10.29012/jpc.733
［18］ Choupani, A. A. and Mamdoohi, A. R. (2016), Population Synthesis Using Iterative Proportional Fitting (IPF): A Review and Future Research, *Transportation Research Procedia*, 17, 223-233.
［19］ Deming, W. E. and Stephan, F. F. (1940), On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known, *The Annals of Mathematical Statistics*, 11(4), 427-444.
［20］ Harland, K., Heppenstall, A., Smith, D. and Birkin, M. (2012), Creating realistic synthetic populations at varying spatial scales: a comparative critique of population synthesis techniques, *Journal of Artificial Societies and Social Simulation* 15, 1.
［21］ Huang, Z. and Williamson, P. (2001), A Comparison of Synthetic Reconstruction and Combinatorial Optimisation Approaches To The Creation of Small-Area Microdata, Working Paper, Department of Geography, University of Liverpool.
［22］ Kim, H. J., Drechsler, J. and Thompson, K. J. (2021), Synthetic microdata for establishment surveys under informative sampling, *Journal of the Royal Statistical Society Series A*, 184(1), 255-281.
［23］ Lenormand, M. and Deffuant, G. (2013), Generating a synthetic population of individuals in households: Sample-free vs sample-based methods, *Journal of Artificial Societies and Social Simulation*,16(4), 12.
［24］ Little, R. J. A. (1993), Statistical Analysis Masked Data, *Journal of Official Statistics*, 9(2), 407-426.
［25］ Merton, R. K. (1936), The Unanticipated Consequences of Purposive Social Action, *American Sociological Review*, *1*(6), 894-904.

［26］ Murata, T., Harada, T. and Masui, D. (2017), Comparing Transition Procedures in Modified Simulated-Annealing-Based Synthetic Reconstruction Method without Samples, *SICE Journal of Control, Measurement, and System Integration*, 10(6), 513-519.

［27］ Rodríguez, R. (2007), Synthetic data disclosure control for American Community Survey Group Quarters, paper presented at *Proceedings of the Survey Research Methods Section, American Statistical Association*, 1439–1447.

［28］ Rubin, D. B. (1993), Discussion: Statistical Disclosure Limitation, *Journal of Official Statistics*, 9, 461-468.

［29］ Smith, D. M., Clarke, G. P. and Harland, K. (2009), Improving the Synthetic Data Generation Process in Spatial Microsimulation Models, *Environment and Planning A: Economy and Space*, 41(5), 1251-1268.

［30］ Smith, D. M., Pearce, J. R., and Harland, K. (2011), Can a deterministic spatial microsimulation model provide reliable small-area estimates of health behaviours? An example of smoking prevalence in New Zealand, Health & Place, 17(2), 618-624.

［31］ Voas, D. and Williamson, P. (2000), An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata, *International Journal of Population Geography*, 6, 349-366.

［32］ Williamson, P., Birkin, M., and Rees, P. H. (1998), The Estimation of Population Microdata by Using Data from Small Area Statistics and Samples of Anonymised Records, *Environment and Planning A: Economy and Space*, *30*(5), 785-816.

［33］ Wilson, A. G. and Pownall, C. E. (1976), A New Representation of the Urban System for Modelling and for the Study of Micro-Level Interdependence, *Area*, *8*(4), 246–254.