リサーチペーパー第 61 号　　Research Paper No.61

近年の諸外国の統計調査における欠測値補完の動向及び合成データに関する議論について

Current Trends of Data Imputation Methods for Statistical Surveys and Agendas on Synthetic Data in Foreign Countries

坂下　信之
統計研究研修所統計専門官

SAKASHITA Nobuyuki
SRTI Specialist for Statistics

令和 6 年 11 月
November 2024

総務省統計研究研修所
Statistical Research and Training Institute (SRTI)
Ministry of Internal Affairs and Communications

受理日：令和６年 10 月 29 日

本ペーパーは、総務省統計研究研修所職員である執筆者が、その責任において行った統計研究の成果を取りまとめたものであり、その内容については、総務省統計局又は統計研究研修所の見解を表したものではない。本ペーパーの内容については、執筆者に問い合わせ願いたい。

近年の諸外国の統計調査における欠測値補完の動向及び合成データに関する議論について

坂下　信之

概要

政府統計の精度維持・向上が喫緊の課題となる中で、欠測値や外れ値への対応はその重要な要素である。世界的にも 1980 年代半ばから今日でも参照される文献が現れ、今世紀に入ってからは、国連などの場で盛んに議論されるようになっている。

今回は、アメリカ合衆国（以下「米国」と記す。）におけるインピュテーションをめぐる近年の動向に加えて、合成データについての近年の論点についての情報を収集した。

その結果、米国では、2020 年人口センサスで初めて導入された行政情報の利用結果についてまとめに入るとともに、人口センサスやその他の統計調査で引き続きさまざまな検討が行われていること、合成データについては、データの構造や標本設計と合成データの統計モデルの関係、開示リスクの評価などについて議論があることが分かった。

キーワード：データ・エディティング、欠測値補完、インピュテーション、合成データ

Current Trends of Data Imputation Methods for Statistical Surveys and Agendas on Synthetic Data in Foreign Countries

Nobuyuki Sakashita

Abstract

While maintenance and enhancement of accuracy in official statistics are emerging as urgent issues, treatment of missing data or outliers is their substantial element. Looking around the world, those literatures referenced until today appear from the mid-1980s. Since the beginning of this century, the matter has been actively discussed at the United Nations and other places.

In this paper, we collected, in addition to recent trends surrounding imputation in the United States, information on recent agendas concerning synthetic data.

As a result, we found that, in the United States, the results of the use of administrative records, first introduced in the 2020 Population Census, are now being summarized, and also, various studies continue to be conducted in the Population Census and other statistical surveys.

Regarding synthetic data, we found that such agendas like the structure of the data, the relation between the sample design and the statistical model of the synthetic data, and the evaluation of disclosure risk have emerged in recent years.

Keywords: Data Editing, Imputation of Missing Data, Synthetic Data

０．はじめに

政府統計の精度維持・向上が喫緊の課題となる中で、欠測値や外れ値への対応はその重要な要素である。世界的にも 1980 年代半ばから今日でも参照される文献が現れ、今世紀に入ってからは、国連などの場で盛んに議論されるようになっている。

これまで、入手可能な文献の調査、各国の最新動向や手法の体系がどのように整理されてきたかの観点からの文献の収集・調査、基本的な文献と思われる書籍の収集・調査を行うとともに、一般用ミクロデータを用いたホット・デック法のシミュレーションを行った。

新型コロナウイルス感染症の世界的流行（いわゆるコロナ禍）が始まって以降は、各国からの発信や国際的な情報交換が少なくなり、新たな情報が得にくくなる一方で米国のセンサス局などで継続して行われているプロジェクトもあるため、今回は米国の状況についての情報のほか、インピュテーションと密接に関わる合成データの作成についての情報を収集した。

以下の構成は、１．が米国におけるインピュテーションをめぐる動向、２．が合成データについての近年の論点、３．がまとめとなっている。

１．米国におけるインピュテーションをめぐる動向

（2020 年人口センサス）

坂下 (2021)、坂下(2023) に記したように、米国の 2020 年人口センサスは一部の世帯の集計に行政記録を利用する最初の人口センサスとなったが、それを受けてセンサス局の The Center for Statistical Research & Methodology (CSRM)で行われていた「無回答のフォローアップ」(Nonresponse Follow-Up, NRFU) 及び関連するプロジェクトについては、2021 会計年度からまとめに入っており、CSRM (2022)、CSRM (2023) がその後の状況を伝えている。また、Mulry et al. (2023) は、人口センサスの自計及び NRFU による回答と今回作成された行政記録名簿の比較を記している。

Mulry et al. (2023) では、2020 年の人口センサスにおける行政記録による名簿の作成方法を解説するとともに、自計又は NRFU の回答による名簿と行政記録により作成された名簿が両方存在する世帯を対象として、行政記録による名簿をセンサスへの自計による回答及び NRFU への回答と比較し、さらに、センサスと行政記録が 1 名だけ異なる世帯について、その原因を分析している。その結果、行政記録名簿との世帯人員の一致率は自計の場合は 79.5%、NRFU の場合は 58.7%であり、自計と行政記録名簿のずれは 0 を中心にほぼ対称であるのに対し、NRFU と行政記録名簿の比較では、NRFU の方がやや大きい傾向があるとの結果を得ている。センサスの方が多い世帯は、行政記録名簿が個人所得税申告書 (IRS1040) の申告前に作成された、あるいは離婚した夫婦の子供が行政記録で親権者でない親の住所に登録されている場合などがあり、行政記録の方が多い世帯は、実際には別の場所に住んで

いる大学生の子供などの場合があり、正しい居場所の特定方法は今後の検討課題である、また、一致率は調査期日である 4 月 1 日と結果が得られる時期の差と NRFU の形態に影響され 、この結果から行政情報の利用の仕方に関する知見が得られるとしている。

「行政記録による無回答の補完と支援 (Supplementing and Supporting Non-Response with Administrative Records)」プロジェクトでは、坂下(2023) に記した 7 つの「覚書 (memorandum)」を公表し、居住実態の判定モデル（Administrative Records Modeling Team (2017)、坂下 (2021)）の異常値検出についての追加の覚書の草稿を作成している。さらに 2010 年センサスの行政記録における住居ユニットのサイズを推定する多項ロジスティック回帰モデルを 2020 年の行政記録における住居ユニットデータに適用し、2020 年の結果と比較し、その結果を文書にしている 。

坂下 (2021) に記した「二言語研修の効果の実験 (Experiment for Effectiveness of Bilingual Training)」プロジェクトについては、継続比率ロジットモデル（坂下(2021)、Raim et al. (2020)）を NRFU データに当てはめ、調査員が特に非英語世帯に接触する際に調査員にバイリンガルな研修を行うことの効果を検証するための分析を進めている。

2020 年センサスでは項目欠測に対して、行政記録や過去のセンサスからの割当てを行ったが、2022 年度からは、2030 年センサスに向けて、行政記録と統計モデルを組み合わせる「多変量カテゴリカル特性データのインピュテーションモデル(Imputation Modeling for Multivariate Categorical Characteristic Data in 2030 Census)」のプロジェクトが始まっている。

（コロナ禍による地域社会調査 (American Community Survey, ACS) への影響）

ACS については、コロナ禍でデータ収集が中断したことにより大量に発生した無回答の傾向が社会階層によって異なることが引き起こすバイアスに対処するためのウェイト付けの研究を開始したことが CSRM (2021) で報告され、Rothbaum et al. (2021) では「エントロピー・バランシング」を用いたウェイト調整を試みた（坂下(2023)）。2022 年度から 2023 年度にかけては、この件に関するプロジェクトが引き続き進行しており、2020 年調査の回答に基づいた傾向モデルの開発の開始、ラッソ、ランダムフォレスト、一般化ブースティングなどの機械学習モデルの比較、逆確率ウェイト付けとカリブレーション（レーキングなどの調整）を組み合わせた手法を逆確率ウェイト付けのみのものとの比較などを行い、合同統計会議 (Joint Statistical Meetings, JSM) で議論した上で学術誌に投稿している（Kang et al. (2023)）。

（小売統計におけるビッグデータ利用の研究）

坂下 (2023) で報告した小売統計におけるビッグデータ利用の研究については、サード・パーティーの収集会社のデータに基づく州レベルの小売の売上げを推計する階層的ベイズ・マスインピュテーション手法についての投稿論文を完成させ、また他の調査についてもインピュテーションにおける代替データの利用を構想している。

（統計データの開示抑制のための合成データを用いた統計的推論）

CSRM で行っている、統計データの開示抑制のため実データの代わりに合成データを用いる統計的推論の研究の中で、プラグイン抽出及び事後予測抽出により単一代入された合成データを用いたベイズ統計学による推論手法の多変量正規分布モデルでの検討が予告されていた（CSRM (2021)、坂下(2023)）が、分散や回帰係数を推定するシミュレーションを行った論文が Guin et al. (2023) として公表されている。その結論では、プラグイン抽出の方が事後予測抽出より良い推論ができ、これは有用性とプライバシーのトレードオフを証明しているとしている 。

２．合成データについての近年の論点

米国センサス局における合成データの検討状況については、坂下 (2020) 及び坂下(2021)で経済センサスのインピュテーション及びミクロデータ作成に関して、坂下 (2023) で統計データの開示抑制のための合成データによる推論方法の開発について報告した。

本邦においては、平成 23 (2011) 年に独立行政法人統計センターから合成データの一種と言える教育用疑似ミクロデータの試行提供（山口他(2013)）、平成 28 (2016) 年には総務省統計局及び統計センターからその確定版ともいえる一般用ミクロデータの提供を開始し、高部 (2022)、横溝及び伊藤 (2023) などで合成データについての議論が行われている。

一方、世界的に見ると、合成データについての議論は、Rubin (1993) や Little (1993) を端緒として前世紀末から行われており[1]、今回は近年どのような議論が行われているかを知るため、幅広く文献を収集した。その結果、合成データを巡っていくつかの論点が存在することが分かった。

（合成データの必要性）

Kinney et al. (2011) は、事業所データのミクロデータは、ごく少数の変数から事業所を同定できることが多く、ある企業の業務データを知ることでライバルが優位に立つことが起こりうるが、人口統計の一般利用において機密保護に用いられるトップコーディングやスワッピングは、分布の歪みが大きく、必要な変更（100 パーセントのスワッピングなど）を施すと使えないデータになるため、ほとんどの国の統計機関は事業所のミクロデータを公表していないことを指摘し、対処法の 1 つは、合成データ、すなわち実際のミクロデータの分布を模倣するように設計された統計モデルからシミュレートされたデータを公開することであるとしている。

また、Miranda and Vilhuber (2016) は、米国センサス局の経済動態統計 (Business Dynamics Statistics, BDS)の保護処理は、p パーセントルール（上位２社がセルの合計値の p パーセン

[1] 合成データは、基本的な技術として多重代入法によっているものの、調査客体の秘密保護を目的とすることが多いため、国連などでもデータ・プライバシー（秘匿）に関連して議論されることが多い。

ト以上を占めるセルに一次秘匿を施し、一次秘匿だけでは秘匿セルが計算できる場合など必要に応じて二次秘匿を行う）によっており、公開される表のすべてのセルを分析するため、多くの二次秘匿が必要となっていると指摘する。また、詳細なクロス集計は、データの有用性を高めるように思われるが、実際には秘匿の増加につながり、有用性が低下するとして、合成データと実データを混合した「部分的合成データ」を提案している。

Abowd (2017)[2] は、「データ・プライバシーの社会科学」と題してプライバシーの経済学に関する厳格な分析を行った Stigler (1980)[3] の「法執行の過程で個人情報が取得されるのをどう防ぐかよりも、このような個人情報の利用をどう制限するかが問題になるだろう」という予測、Acquisti et al. (2016) の民間のインターネット大手に対する調査や Acquisti and Varian (2005) の「きめ細かいデータへのニーズと、個人の記録を保護するニーズとのバランスをどう取るかは、経済学者や、統計学者やコンピュータ科学者などが同時に関与する問題である」という指摘を引用し、「ビッグデータ」の時代にあって市民のプライバシーとそのデータの機密保護の問題は複雑に絡み合っていること、統計機関には、収集したデータを何らかの形で公表することと、取得したデータを法律の行使に使用してはならないことの 2 つの法的義務があることを指摘している。

Chien et al. (2021) は、これらの先行する議論を引用しながら、オーストラリアの企業は産業によっては寡占状態となっていて、秘匿や攪乱のような家計データでの既存技術が有効でない可能性があり、オーストラリア統計局 (ABS) では、ビジネス・マイクロデータへのアクセスを強化するために、研究者向けに Chien et al. (2018) に示されたような合成データセットを公開することを検討したと述べている。

また、Kim et al. (2018) は、センサス局は人口調査については最小限の開示処理を施したミクロデータを提供することがあるが、5 年ごとの経済センサスでは、主要データである製造業センサス (CMF) について実際の値を含むマイクロデータの公開を禁止しているとして、エディット&インピュテーションとデータ合成の統合を実装する最初の試みとして、CMF からよく利用される変数の公開用ファイルを作成し、併せてエラー処理を行う方法を紹介している[4]。

これらの文献から窺えることは、特に事業所に関するデータにおいて、分布の歪みや巨大独占企業の存在のため、既存の秘密保護手法ではデータが露見しやすく、その対策として合成データが検討されていることである。この背景には、ミクロデータへの需要が高まったこともあるが、集計表でもクロスを詳細にすると、同様の問題が起きうる。

他方、Manrique-Vallier and Reiter (2012) は、データを公開する統計機関などは機密保護の

---

[2] Abowd, J. M.:コーネル大学教授兼米国センサス局科学主任（2017 年当時）。センサス局で作成している Survey of Income and Program Participation の合成データの手法による一般利用ミクロデータの作成に関与している（Abowd et al.. (2006)）。

[3] Stigler, G. J. :経済学者(1911-1991), ノーベル経済学賞受賞 (1982)。

[4] 坂下 (2020) に記したように、米国センサス局では、Kim et al.(2015)、Kim et al. (2018) に基づいて、経済センサスのエディット、インピュテーション、データの合成を一体化したシステムの開発を進めている。

義務があるが、悪意のある個人が、共通の特性情報（キー）を照合してデータの主体を他のデータベースにリンクするかもしれず、その場合に母集団で一意（キーの同じ組み合わせが他に存在しない）となるデータが問題となるとして、データが一意になる確率に基づく開示リスクの評価方法を提示し、人口センサスへの適用を試みている。この Manrique-Vallier and Reiter (2012) はゼロの多い粗な分割表を念頭に置いて、旧来のモデルに基づく評価方法では歪んだ結果をもたらすとしてその対処を論じている[5]が、構造的ゼロ[6] は考慮していないのに対し、Manrique-Vallier and Reiter (2014) では構造的ゼロを考慮した潜在構造モデル[7]による分布モデルを提案している。Hu et al. (2014) は、旧来の手法で守秘義務を守れないデータとして、事業所データの他に大規模な行政データベースを挙げ、具体的な例として、「所得及びプログラム参加調査（Survey of Income and Program Participation, PIPP)）（Abowd et al. (2006)）と「縦断的ビジネスデータベース (Longitudinal Business Database, LBD)」（Kinney et al. (2011)）で一部の変数を除き、推定されたモデルからシミュレートされた値に置き換えていることを指摘している。さらに、Hu et al. (2018) は、American Community Survey group quarters data（Hawala (2008)）、OnTheMap application（Machanavajjhala et al. (2008)）の一般公開データでも合成データの手法が取り入れられることを指摘している[8]。これらの例は、人口に関するデータにおいても、集計表が疎であったり、変数の関係が複雑である場合には露見の可能性があり、合成データの検討が行われることを示していると考えられる。

（データの性質）

現実の統計データは、変量の単純な羅列ではなく、複雑な構造を取ることが多い。近年、この性質を再現するための統計モデルについて一連の研究が行われており、合成データの作成にも応用されている。

（入れ子構造）

Vermunt (2003)、Vermunt (2008)、Bennink et al. (2016)、Hu et al. (2018) は、世帯の中に個人がいる場合のように、集団と個人が入れ子になって存在するモデルについて論じている。

Vermunt (2003) は、社会科学分析で用いられてきた潜在クラス分析 (latent class analysis)（Lazarsfeld (1950), Goodman (1974)）[9]が観測値が独立している仮定の上に成り立っていたこ

[5] Skinner and Shlomo (2008) を引いて、「表が大きく疎な場合、対数線形モデルからのセル確率の推定値は、表中の多くのランダムなゼロカウントによって歪められる可能性がある。一般的に、これは母数の過大評価となり、その結果、識別情報開示の真のリスクを過小評価することになる。」としている。

[6] 集計表上のセルがゼロとなるもので、論理的にあり得ない変量の組合せによるもの。論理的には可能で、たまたまゼロだったものは「非構造的ゼロ」と呼ぶ。

[7] 観測変数の背後に潜在変数があることを仮定して潜在構造を読み解くモデル。潜在変数を入れることにより、見かけ上交絡因子を有する複雑な変数間の関係を比較的単純な関係に還元することができる。

[8] ただし、Hu et al. (2018) の指摘は、これらのデータには世帯データが含まれず、後述する入れ子構造を反映していないという文脈である。

[9] Dunson and Xing (2009) によると、歴史的に潜在構造分析（latent structure analysis）と呼ばれたアプローチは、近年、潜在クラス・モデリング(latent class modeling)という用語がより一般的に用いられている。

とを指摘し、観測値が階層構造になっている「マルチレベル潜在クラスモデル」で、集団レベル（レベル 2）によって個人レベル（レベル 1）のパラメータ（どの潜在クラスに属するかの分布とその潜在クラスの元で観測値の分布を与える潜在変数）が異なるモデルを提案している。集団の数の増大に伴い、モデルが急激に複雑化し不安定になることに対処するため、具体的なモデルとしては、集団レベルのパラメータに分布を想定し、分布のランダムな要素に正規分布（連続分布）を仮定したパラメトリック・アプローチと、パラメトリック・アプローチでは分布を混合する際の仮定が強すぎるとして、多項分布（離散分布）に置き換えたノン・パラメトリック・アプローチを示し、その上で、組織研究（会社の仕事への意識）、教育研究（中学 2 年生の数学スキルの分析）、国際比較研究（欧州における意識調査の国別比較）への 3 つの適用例を紹介し、各例において明確な潜在クラスのグループ間変動が示されたとしている。

Vermunt (2008) は、応答変数をカテゴリー変数から連続変数や個数に拡張し、病院における患者への抗生物質の処方、家族に属する子供への知能テスト、癲癇の時系列データに対しマルチレベル潜在クラスモデルを適用して、適用するモデルを決定するに際してデータの階層を考慮する必要があること、標準的な潜在クラスモデルと比べて明確な付加価値があること、さまざまな拡張が可能であることを述べている[10]。

Bennink et al. (2016) も集団（マクロレベル）の中に個人（ミクロレベル）が入れ子状に存在するモデルを扱い、既存の個人の変数が一つの場合に対する手法を個人の変数が複数存在する多変量のミクロ・マクロ・モデルに拡張している[11]。拡張に際しての具体的なモデルとしては、複数の個人レベルの離散的な変数が，グループ・レベルの潜在変数の指標として直接使用される「直接モデル」と複数の個人レベルの離散的な変数がグループ・レベルの潜在変数の指標として使用される個人レベルの潜在変数を構築するために使用される「間接モデル」の２種類を提示し、直接モデルについては、2010 年イタリア家計調査、間接モデルについては、小企業への満足度のアンケート調査による実証分析を行い、前者については世帯の型による明確な違いがあるなどの結論、後者については集団レベルで見た結果は個人レベルで見た結果より複雑との結論を得ている。

これらの先行研究に対し、Hu et al. (2018) も集団の中で個人が入れ子になっているデータのモデル化を扱っているが、Vermunt (2003)や Vermunt (2008) ではモデルの選択時にクラスの数が固定されており、またいずれの先行例も「構造的ゼロ」を考慮していないのに対し、「多項分布の積の入れ子データディリクレ過程混合」(Nested data Dirichlet Process Mixture of

[10] Vermunt (2008) は、個人が複数回答である場合は、個人の中に複数の変量があるため実際には 3 レベルモデルであるとして、個人レベル（レベル 2）の潜在変数（パラメータ）と集団レベル（レベル 3）の潜在変数（パラメータ）の離散／連続によりモデルを 4 種類に分類している。また、統計学者はここで議論されている手法を、より専門的な用語である有限混合モデル(finite mixture models) と呼び、すべての応答変数がカテゴリー変数である場合を潜在クラス分析（LC analysis）と呼ぶことが多いと指摘した上で、論文中では「互換的」（同義）に用いるとしている。

[11] Vermunt (2003) からは集団レベルでも結果変数を持つ点が拡張されている。また、潜在変数を用いたモデルの解説書として Skrondal and Rabe-Hesketh (2004) と Bennink et al (2013) を挙げている。

Products of Multinomial distributions, NDPMPM) モデルを用いて対処する方法を論じている[12]。このモデルは、Bennink et al. (2016) の「間接モデル」に類似しているが、「間接モデル」では集団から個人を回帰しているのに対し、同時分布を推定している。この論文では、まず世帯変数と個人変数から成る NDPMPM モデルを解説し、さらに構造的ゼロに対処するための尤度の修正を論じ、2012 年地域社会調査 (American Community Survey, ACS) の公共利用ファイルを用いて、構造的ゼロを考慮していない場合と考慮している場合を例示し、構造的ゼロを考慮していない場合で入れ子構造を無視すると精度が大幅に落ちること、構造的ゼロを考慮する場合は元データとほぼ同じ推定ができることを示している。なお、Hu et al. (2018) ではモデルを推定する際に事前情報を利用していず、潜在クラスモデルに事前情報を取り入れるのは条件分布を歪めない方法で行う必要があり、厄介であるとしている。Schifeling and Reiter (2016) は、入れ子になっていないモデルでこれを行う簡単な方法を示している。

（標本設計の問題）

合成データを生成するための統計モデルについて、多くの統計は母集団の単純な縮図となるような無作為抽出によってはいないため、抽出による原データから合成データで何を再現するのかの観点に係わる議論がある。

Hu et al. (2018) によると、NDPMPM は、多くのジョイントモデルと同様、観測されたデータの分布を反映するため、抽出率が一様でなく、部分母集団によって差のある複雑な標本設計からのデータを用いた多変量分布の推定には適さない。設計に使用した変数がカテゴリカルで、標本だけでなく母集団についても得られる場合は、分析者は、ベイズ有限母集団推論を行うために NDPMPM を使用することができる（Gelman et al. (2013)）が、この情報が得られない場合は、調査のウェイトをモデルに組み込む手法についての合意はないとしつつ、ウェイトを含んだ調査データのベイズ分析についての先行研究として Kunihama et al. (2016)[13] を挙げ、そこで示された抽出されたもののウェイトのみを使用する便利な手法は、入れ子になったカテゴリーデータにも適用可能で、そのようなアプローチは今後の研究課題であるとしている。

Kunihama et al. (2016) は、標本調査では、さまざまな集団が標本に適切に含まれるように、層別抽出を行うのが一般的で、母集団は、含まれる確率が異なる互いに排他的な層に分割されており、層化確率抽出で得られたデータを超母集団からの無作為標本であるかのように分析することには、潜在的に大きなバイアスがあることを指摘し、標本調査のウェイトをベイズ分析に含める提案はいくつかあるが、既存の方法は複雑なモデルを必要とするか、調査ウェイトの基礎となる層別設計を無視しており、推定に層別抽出・設計の調整を含める方法の大部分も、モデルベースの推論、特にノンパラメトリックベイズの枠組みの下では適切で

[12] 入れ子になっていない場合でディリクレ過程混合(DPM)を用いた潜在クラスモデルのノンパラメトリック・ベイズ版を提示した先行研究として Dunson and Xing (2009) を挙げている。
[13] Hu et al. (2018) の文献情報では Kunihama et al. (2014) となっているが、その後改訂された模様。

はないとしている。Kunihama et al. (2016) によると、この問題について、Little (2004) と Gelman (2007) は、モデルベースの分析に調査のウェイトを含めることの重要性を明らかにし、Zheng and Little (2003, 2005) は、「ノンパラメトリックスプラインモデル」を提案し、Zangeneh and Little (2012) は、非標本化ユニットの数が未知であることを許容する修正を提案し、Si et al. (2015) は、調査ウェイトがガウス過程回帰を通して回答とリンクされるノンパラメトリックモデルを提案している。

Kunihama et al. (2016) で引用されている Zheng and Little (2003) は確率比例抽出[14]における合計の推定を念頭に、Horvitz-Thompson（HT）推定量は，結果値の比と選択確率がほぼ一致 (exchangeable) するときはよく機能する[15]が、この仮定が満たされていない場合，非常に非効率になるとして、両者の間に滑らかに変化する関係を仮定し，その関係を「罰則付きスプラインノンパラメトリックモデル」を用いてモデル化する代替法を検討し、一般に、平均二乗誤差で Horvitz-Thompson 推定量と一般化回帰推定量を上回ることが示されたとしている。Zheng and Little (2005) では、さらにその罰則付きスプラインに基づく推定量について、モデルベース、ジャックナイフ、および平衡反復複製 (balanced repeated replicate method, BRR) による分散推定法を開発し、シミュレーションにより、罰則付きスプライン点推定量とそのジャックナイフ標準誤差は、Horvitz-Thompson 推定量や一般化回帰推定量に基づく推論よりも優れた推論を導くことが示されたとしている。

また、Gelman (2007) もウェイトの付いたデータのモデル化について論じ、ニューヨーク市を対象とした世論調査を題材に、調査ウェイトは必ずしも選択確率の逆数と等しくなく無回答の調整を含んでいるため、単純な平均値や比率よりも複雑なものを推定する際にウェイトをどのように使用するかは必ずしも明確ではなく、単純なウェイト付き平均値でも標準誤差の推定は厄介であるとし、ウェイト付けに代わる手法としては抽出あるいは無回答に影響する変数を説明変数とした回帰モデルがあるが、これは多数の交互作用の可能性があって変動しやすいと指摘している。その上で、実務的な解決には至ってないと断りつつ、ウェイトが調査値に影響される階層モデルにおいて、標本と母集団の差異を埋める事後層化と組み合わせた階層回帰を紹介している。

Zangeneh and Little (2012) では、特に確率比例抽出の場合において、抽出されなかった客体の大きさに関する情報は得られないことが多いという問題意識のもとに、スプラインモデルにその重要な特徴である不均一な誤差分散をモデル化するための未知のパラメータを含め、抽出されなかった客体全体の大きさに関する情報を含める改良されたベイズ法を開発し、発展させている。具体的には、第 1 ステップで抽出確率を比例させる大きさを示す変

[14] 総務省統計局の統計調査では、調査区を調査対象数に比例して抽出し、さらに各調査区から調査客体を抽出する多段階抽出で確率比例抽出を用いることが多いが、一連の議論では、従業員数に比例して企業を抽出し、抽出された企業から諸統計量を推定するような一般的な確率比例抽出を想定している。

[15] 確率比例抽出された標本による母集団の合計の推定では、抽出確率の逆数をウェイトとすれば、抽出確率の設定に係わらず理論上不偏推定量になる。一方、誤差は抽出確率と結果値が比例関係に近ければ小さいが、比例から外れるほど大きくなる。

量 X に確率比例抽出用に修正された制約付きベイズ・ブートストラップモデルを適用し、抽出されなかった客体をインピュートした後、第 2 ステップでベイズ罰則付きスプラインモデルにより、抽出されなかった客体の調査対象である変量をインピュートするモデルを提案している。シミュレーションでは、ディリクレ分布に従う説明変数から異なった関数によって導かれた期待値を持つ目的変数から成る 6 種類の人工的なデータを用いて 4 種類の平均値の推定方法を評価し、この方法は従来の Horvitz-Thompson 推定量よりも大きな利点を提供することが示唆されたとし、続いて米国センサス局の ACS の公共利用ミクロデータのデータセットに適用して検証している。その結果は、対象としている変数の分布が極度に歪んでいるため、ベイズ罰則付きスプラインモデルの前提を満たしていないが、Horvitz-Thompson 推定量より効率的な推定を行っているとしている。

Si et al. (2015) も同様にウェイトが標本に含まれる客体についてのみ既知であると仮定しているが、Zangeneh and Little (2012) が抽出されなかった客体の大きさを推定するための制約付きベイズ・ブートストラップにおいて利用可能な標本の大きさにのみ依存して標本設計を考慮していないのに対し、調査ウェイトがガウス過程回帰を通して回答とリンクされるノンパラメトリックモデルを提案している。Si et al. (2015)はその手法を「ベイズノンパラメトリック有限母集団（BNFP）推定」と名付け、ほとんどの標本設計に適用可能であるとして、シミュレーションによりロバスト性のチェックを行い、さらに子どもの幸福度についての追跡調査に適用してモデルの適合度を評価し、標本が少ないときは BNFP による推定は古典的な推定より成績が良いという結果を得て、標本が大きいときは優位は微妙になるが、これは十分な事前情報がなくて大きな標本のもとでの古典的な推定に互する推定を得るという目的に合っているとしている。

以上の先行研究に対して、Kunihama et al. (2016) は、標本に含まれない客体の大きさは、通常の Horvitz-Thompson 推定や Hajek 推定には必要なく、この情報が公共利用データファイルに含まれることはほとんどないとしている。また、標本に含まれない客体を対象とする調査ウェイトのモデル化は、非常に複雑なモデルになる可能性があるとして、選択された標本に対してディリクレ過程混合などの標準混合モデルを適用し、調査ウェイトに基づいて混合ウェイトを調整するという単純な手法を提案し、他の 3 種の方法（Horvitz-Thompson 推定、ランダム要素を持つ多変量回帰、Si et al. (2015) で示されたガウス過程回帰）と比較するシミュレーションを行って、他の手法は密度の多峰性を捉えることに失敗したり、バイアスが生じており、カルバック・ライブラー (KL) 情報量を測定した上で提案した手法は元の分布をよく再現しているとしている。その後、青少年の行動に関する時系列データに適用した検証を行っている。

（開示リスクの評価）

先述の Kim et al. (2018) は、米国センサス局が 5 年ごとに行う経済センサスの主要なデータである製造業センサス (Census of Manufactures, CMF) のデータを用いて、誤データのエ

ディット&インピュテーション (EI) と合成データ (SD) の作成を統合する試みであり、その中で有用性と秘匿性の評価を行っている。それによると、完全合成データ[16]では、個々のケースを識別したと主張することはできないが、個々の施設の機密属性を推測するために合成データを使用する「推測的開示」、例えば、ある産業で最大の給与を支払う事業所を知っている侵入者が公開された合成データを用いて、その最大の給与額を推定することはできるとしている。Kim et al. (2018) では 3 種類の侵入者（(1) 合成データセット以外の情報を持たない、(2) 2 番目に大きな事業所を知っている、(3) 2 番目に大きな事業所と合計の推定値を知っている）を想定し、それぞれの場合に最大値を推測[17]する方法を説明した上で CMF のデータでの実際の値との乖離により、その精度を推定して、(3) の 2 番目に大きな事業所と合計の推定値を知っているケースでは「不運」とは言えないレベルで推測に成功しているとして、合計値あるいは平均値が公表されているデータの合成データを提供することにはリスクがあると結論づけている。

Kim et al. (2018) では、合成データの開示シナリオを論じた先行研究として、部分合成データについては Domingo-Ferrer et al. (2001)、Drechsler (2012)、Drechsler and Reiter (2010)、Drechsler and Reiter (2011)、Hundepool et al. (2012) などがあるが、 完全合成データについては標準的な基準は存在しないと断りつつ、Drechsler et al. (2008) を挙げている。Drechsler et al. (2008) では、母集団のうち標本に選ばれなかったデータを多重代入法による合成データで代替して母集団を再現したものから標本を再抽出する Rubin (1993) のアイデアを応用し、1997 年ドイツ労働市場・職業研究所 (IAB)企業パネルの変数セットを用いて完全合成データセットを生成して、元データを用いた分析結果と合成データで行った分析結果を比較している。具体的には合成母集団の生成を 10 回、それぞれから標本の再抽出を 10 回行い、計 100 系列のデータセットを作成し[18]、このデータセットにおいて、事業所が元のデータセットと一つ以上の新たなデータセットに含まれ、この事業所の元の値とインピュテーション後の値がほとんど等しいという二つの段階を満たした場合に開示リスクが発生するが、結果を評価した結果、両段階においてリスクは低いとしている。

また、Hu et al. (2018) は、補足資料 (supplementary material) で、Hu et al.(2014)で提案されている尺度により開示リスクの評価を行っている。これは、侵入者の知識と攻撃戦略に関する仮定の下で、公開された合成データが与えられた場合に侵入者が機密データから値を学習できる事後確率を定量化するもので、具体的には、特定のデータ以外の元データと元データから作成された合成データが知られているとき、知られていない特定のデータをどの程

[16] 完全合成データと部分合成データについては、坂下(2021)、坂下 (2022) を参照。
[17] 例えば 1 番目の侵入者は合成データの最大値により、2 番目の侵入者は合成データの上位の値の和から現実の 2 番目の値を減算することなどにより推測が可能である。このモデルでは、多重代入法による複数の EI データセットのそれぞれについて、複数の合成データセットを作成しているため、平均を取ることにより、単純にランダムに発生させた単一代入の合成データよりも推測の精度が高くなることが想定される。
[18] 実際には再抽出で選ばれないデータには多重代入（インピュテーション）する必要は無いので、再抽出を先に行ってから選ばれたデータに対してインピュテーションを行っている。

度の精度で推定できるかを判定している。この手法は、「ネストされたカテゴリーデータの統計的開示リスクを評価するための、我々が知る唯一の戦略である」とされているが、構造的ゼロを含む場合と含まない場合の双方を評価した結果、NDPMPM から生成された合成データは開示リスクが低いことが分析から示唆されたとしている。

３．まとめ

今回の調査から、米国では、人口センサスで調査対象に接触できないことの多さへの対策として検討されてきた行政情報の利用が 2020 年人口センサスで初めて導入され、結果についてまとめに入るとともに、引き続きさまざまな検討が行われていること、その他の統計調査でもコロナ禍の影響への対処法やビッグデータの利用、統計データの開示抑制のため、合成データを用いた統計的推論の研究が継続して行われていることが分かった。

合成データについては、合成データの必要性について、分布の歪みや巨大独占企業の存在のため、既存の秘密保護手法ではデータが露見しやすい事業所関係のデータはもとより、人口に関するデータでも集計表が疎であったり、変数の関係が複雑である場合には露見の可能性があり、合成データの検討が行われていること、入れ子関係などのデータの構造や層化確率比例抽出のような抽出率が一様でない標本設計がされている場合の統計モデル、開示リスクの評価などについて議論があることが分かった。

参考文献

[1] 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―」 総務省統計研究研修所『統計研究彙報』, 第 80 号, pp.97-116

[2] 坂下信之 (2020)「統計調査の欠測値補完方法に関する研究動向について（主に米国とオランダ）」、リサーチペーパー第 48 号、総務省統計研究研修所。

[3] 坂下信之 (2021)「近年の諸外国の統計調査における欠測値補完の動向について」、リサーチペーパー第 51 号、総務省統計研究研修所。

[4] 坂下信之 (2023)「近年の諸外国の統計調査における欠測値補完の動向について」、リサーチペーパー第 57 号、総務省統計研究研修所。

[5] 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』第 79 号, pp.111-130.

[6] 山口幸三、伊藤伸介、秋山裕美 (2013）教育用擬似ミクロデータの作成―平成 16 年全国消費実態調査を例として―, 統計学. 2013, No. 104, 1-15。

[7] Abowd, J. M. (2017), ”How will statistical agencies operate when all data are private?” Journal of Privacy and Confidentiality, 7(3), 2017.

[8] Abowd, J., Stinson, M., and Benedetto, G. (2006). “Final report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project”. Tech. rep., U. S. Census Bureau

Longitudinal Employer-Household Dynamics Program.

[9] Acquisti, A. and Varian, H. R. (2005), “Conditioning Prices on Purchase History", Marketing Science, 24: 367-381.

[10] Acquisti, A., Taylor, C., and Wagman, L. (2016), “The Economics of Privacy", Journal of Economic Literature, 54: 442-92.

[11] Administrative Records Modeling Team (2017), “Administrative Records Modeling Update for the Census Scientific Advisory Committee”.

[12] Bennink, M., Croon, M. A., Kroon, B., and Vermunt, J. K. (2016), “Micromacro multilevel latent class models with multiple discrete individual-level variables”, Advances in Data Analysis and Classification, 10(2): 139–154.

[13] Bennink, M, Croon, M. A, Vermunt, J. K. (2013), “Micro-macro multilevel analysis for discrete data: A latent variable approach and an application on personal network data”, Sociological Methods and Research 42(4):431–457.

[14] Chien, C. H., Welsh, A. H., and Moore, J. D. (2018), “Research paper: Synthetic microdata – ‘a possible dissemination tool’”, Report, Australian Bureau of Statistics, 2018.

[15] Chien, C. H., Welsh, A. H., and Moore, J. D. (2021) “Synthetic Business Microdata: An Australian Example,”. Journal of Privacy and Confidentiality 10 (2).

[16] CSRM (2021), “Annual Report of the Center for Statistical Research and Methodology, Research and Methodology Directorate, Fiscal Year 2022”, U.S. Department of Commerce, Economics and Statistics Administration, U.S. CENSUS BUREAU.

[17] CSRM (2022), “Annual Report of the Center for Statistical Research and Methodology, Research and Methodology Directorate, Fiscal Year 2022”, U.S. Department of Commerce, Economics and Statistics Administration, U.S. CENSUS BUREAU.

[18] CSRM (2023), “Annual Report of the Center for Statistical Research and Methodology, Research and Methodology Directorate, Fiscal Year 2023”, U.S. Department of Commerce, Economics and Statistics Administration, U.S. CENSUS BUREAU.

[19] Domingo-Ferrer, J., Mateo-Sanz, J.M., and Torra, V. (2001), “Comparing SDC methods for microdata on the basis of information loss and disclosure risk”, Pre-proceedings of ENK-NTTS, 2001, pp. 807–826.

[20] Drechsler, J. (2012), “New data dissemination approaches in old Europe – synthetic datasets for a German establishment survey”, Journal of Applied Statistics 39, pp. 243–265.

[21] Drechsler, J., Dundler, A., Bender, S., Rässler, S., and Zwick, T. (2008), “A new approach for disclosure control in the IAB establishment panel—multiple imputation for a better data access”, AStA Advances in Statistical Analysis 92 (2008), pp. 439–458.

[22] Drechsler, J. and Reiter, J.P. (2010), “Sampling with synthesis: a new approach for releasing public use census microdata”, Journal of the American Statistical Association 105, pp. 1347–

1357.

[23] Drechsler, J. and Reiter, J.P. (2011), “An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets”, Computational Statistics and Data Analysis 55, pp. 3232–3243.

[24] Dunson, D. B. and Xing, C. (2009), “Nonparametric Bayes modeling of multivariate categorical data”, Journal of the American Statistical Association, 104: 1042–1051.

[25] Gelman, A. (2007), “Struggles with survey weighting and regression modeling”, Statistical Science 22, 153–164.

[26] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013), “Bayesian Data Analysis”, London: Chapman & Hall.

[27] Goodman, Leo A. (1974). “The Analysis of Systems of Qualitative Variables When Some of the Variables Are Unobservable. Part I: A Modified Latent Structure Approach” American Journal of Sociology 79:1179–259.

[28] Guin, A., Roy, A., and Sinha, B. (2023), “Bayesian Analysis of Singly Imputed Synthetic Data under the Multivariate Normal Model”, RESEARCH REPORT SERIES (Statistics #2023-01), CSRM, US Bureau of the Census, Washington, DC.

[29] Hawala, S. (2008). “Producing partially synthetic data to avoid disclosure”, Proceedings of the Joint Statistical Meetings. Alexandria, VA: American Statistical Association.

[30] Hu, J., Reiter, J. P., and Wang, Q. (2014), “Disclosure risk evaluation for fully synthetic categorical data”, In Domingo-Ferrer, J. (ed.), Privacy in Statistical Databases, 185–199. Springer. 185.

[31] Hu, J., Reiter, J. P. and Wang, Q. (2018), “Dirichlet Process Mixture Models for Modeling and Generating Synthetic Versions of Nested Categorical Data”, Bayesian Analysis, 13(1):183–200, 2018.

[32] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S,. Spicer, K., and de Wolf, P. P. (2012), “Statistical Disclosure Control”, JohnWiley & Sons, 2012.

[33] Kang, J., Morris, D. S., Joyce, P., Dompreh, I. (2023), " On calibrated inverse probability weighting and generalized boosting propensity score models for mean estimation with incomplete survey data", Wires Computational Statistics, May 2023.

[34] Kim, H. J., Cox, L.H., Karr. A. F., Reiter, J.P., and Wang, Q. (2015), “Simultaneous edit-imputation for continuous microdata”, Journal of the American Statistical Association 110 (2015), pp. 987–999.

[35] Kim, H. J., Reiter, J. P. , and Karr, A. F. (2018), “Simultaneous edit-imputation and disclosure limitation for business establishment data”, Journal of Applied Statistics, 45(1):63–82, 2018.

[36] Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J, Jarmin, R. S., and Abowd, J. M. (2011), ”Towards unrestricted public use business microdata: The synthetic longitudinal business

database", International Statistical Review, 79(3):362–384, 2011.

[37] Kunihama, T., Herring, A. H., Halpern, C. T., and Dunson, D. B. (2016). "Nonparametric Bayes modeling with sample survey weights." arXiv:1409.5914. 512.

[38] Lazarsfeld, Paul F. (1950), ''The Logical and Mathematical Foundation of Latent Structure Analysis and the Interpretation and Mathematical Foundation of Latent Structure Analysis.'' pp. 362–472 in Measurement and Prediction, edited by S. A. Stouffer et al. Princeton, NJ: Princeton University Press.

[39] Little, R. J. A. (1993), "Statistical analysis of masked data", Journal of Official Statistics, 9, 407-426.

[40] Little, R. J. A. (2004), "To model or not to model? Competing modes of inference for finite population sampling", Journal of the American Statistical Association 99, 546–556.

[41] Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008). "Privacy: Theory meets practice on the map", IEEE 24th International Conference on Data Engineering, 277–286.

[42] Manrique-Vallier, D. and Reiter, J. P. (2012), "Estimating identification disclosure risk using mixed membership models", Journal of the American Statistical Association, 2012 December 1; 107(500): 1385–1394.

[43] Manrique-Vallier, D. and Reiter, J. P. (2014), "Bayesian estimation of discrete multivariatelatent structure models with structural zeros", Journal of Computational and Graphical Statistics, 23: 1061–1079. 189, 194.

[44] Miranda, J. and Vilhuber, L. (2016), "Using partially synthetic microdata to protect sensitive cells in business statistics", Statistical Journal of the IAOS, 32(1):69–80, 2016.

[45] Mulry, M. H., Tello-Trillo, C. J., Mule, T., Keller, A. (2023), " Comparisons of Administrative Record Rosters to Census Self-Responses and NRFU Household Member Responses", RESEARCH REPORT SERIES (Statistics #2023-01), CSRM, US Bureau of the Census, Washington, DC.

[46] Raim, A. M., Mathew, T., Sellers, K. F., Ellis, R., and Meyers, M. (2020), "Experiments on Nonresponse using Sequential Regression Models", RESEARCH REPORT SERIES (Statistics #2020-03), CSRM, US Bureau of the Census, Washington, DC.

[47] Rothbaum, J., Eggleston, J., Bee, A., Klee, M., and Mendez-Smith, B. (2021), "Addressing Nonresponse Bias in the American Community Survey During the Pandemic Using Administrative Data", 2021 AMERICAN COMMUNITY SURVEY RESEARCH AND EVALUATION REPORT MEMORANDUM SERIES # ACS21-RER-05 and SEHSD Working Paper #2021-24.

[48] Rubin, D.B. (1993). "Discussion: Statistical disclosure limitation", Journal of Official Statistics, 9, 461-468.

[49] Schifeling, T. and Reiter, J. P. (2016), "Incorporating marginal prior information in latent class

models.” Bayesian Analysis, 2: 499–518. 197

[50] Si, Y., Pillai, N., Gelman, A. (2015), “Bayesian nonparametric weighted sampling inference”, Bayesian Analysis 10, 605–625.

[51] Skrondal, A., Rabe-Hesketh, S. (2004), “Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models”, Chapman & Hall/CRC Press, Boca Raton, FL.

[52] Stigler, G. J. (1980), “An Introduction to Privacy in Economics and Politics", Journal of Legal Studies, 9: 623-644.

[53] Vermunt, J. K. (2003), “Multilevel latent class models”, Sociological Methodology, 213–239.

[54] Vermunt, J. K. (2008), “Latent class and finite mixture models for multilevel data sets”, Statistical Methods in Medical Research, 33–51.

[55] Zangeneh, S. Z. and Little, R. J., (2012) “Bayesian inference for the finite population total from a heteroscedastic probability proportional to size sample”, Proceedings of the Joint Statistical Meetings 2012.

[56] Zheng, H. and Little, R.J.A. (2003), “Penalized Spline Model-Based Estimation of Finite Population Total from Probability-Proportional-to-Size Samples”. Journal of Official Statistics, 19, 99–117.

[57] Zheng, H. and Little, R.J.A. (2005), “Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model”, Journal of Official Statistics 21, 1–20.