リサーチペーパー第 57 号　　Research Paper No.57

近年の諸外国の統計調査におけるデータ・エディティング及び欠測値補完の動向について

Current Trends of Data Editing and Imputation Methods for Statistical Surveys in Foreign Countries

坂下　信之
統計研究研修所統計研修研究官

SAKASHITA Nobuyuki
SRTI Senior Researcher for Statistical Training

令和 5 年 3 月
March 2023

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５年２月 16 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、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データ・エディティング及び欠測値補完の動向について

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。

今年度は、欠測値補完に先立つデータ・エディティングなどを含めたシステムや米国で長年研究されている合成データの作成についての情報を収集するとともに、質的データについて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。

その結果、アメリカ合衆国では、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研究されてきた行政情報の利用が本格化したこと、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など、欧州などでは、データ・エディティングとインピュテーションのシステムの開発や改良、特に機械学習の適用の検討が進んでいることが分かった。また、ホット・デック法のシミュレーションでは、ドナーとなるデータを適切に選ぶことが他の手法と併用する場合は特に重要であるという知見を得た。

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、人口センサス

Current Trends of Data Editing and Imputation Methods for Statistical Surveys in Foreign Countries

SAKASHITA Nobuyuki

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.

This year, we collected information on systems including data editing, which is conducted prior to imputation of missing values, and the production of synthetic data, which has been studied for many years in the United States. We also exercised numerical simulation of the hot-deck method for categorical data, using General-Use (Synthetic) Microdata.

As a result, we found that in the United States, the use of administrative records, which has been studied to deal with the frequent inability of contacting households in population censuses, is now implemented in practice, that research on statistical inference using synthetic data is being conducted continuously, and that in Europe and elsewhere, data editing and imputation systems are being developed and improved, while the application of machine learning is studied. In the simulation of hot-deck method, we found that the appropriate selection of donor data is particularly important when used in conjunction with other methods.

Keywords: Data Editing, Imputation of Missing Data, Population Census

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、諸外国で行われているデータ・エディティング、特に欠測値補完がどのように行われているかについて、入手可能な文献を調査するとともに、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・調査を行ってきた。

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）が始まってから、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ってい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り、今年度は、欠測値補完に先立つデータ・エディティングなどを含めたシステムや米国で長年研究されている合成データの作成についての情報を収集するとともに、坂下 (2020) で量的データ（全国消費実態調査）について行った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質的データ（就業状況基本調査）について行った。

以下はその結果であり、その構成は、１．が南北アメリカ大陸（アメリカ合衆国、カナダ、メキシコ、チリ）の動向、２．が欧州の動向、３．がシミュレーション、４．がまとめとなっている。

１．南北アメリカ大陸の動向

米国では、センサス局に属する The Center for Statistical Research & Methodology (CSRM) でエディット及びインピュテーションや合成データの作成について継続的な研究が続けられているほか、センサス局の他の部局や統計を実施している他の行政機関も研究・開発を行い、随時報告書を刊行している。カナダは相対的には小国ながら、公的統計の分野の研究及び開発においてさまざまな発信を行っている。また、従来は比較的存在感の低かった、中南米の国々からも、近年は各国の事情についてのいくつかの報告が見られる。

（米国）

（2020 年米国人口センサス）

坂下 (2021) では、米国の人口センサスでの「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報告した。CSRM (2021) はその後の動向を伝えており、それによると、2020 年米国人口センサスは、一部の世帯の集計に行政記録を利用する最初のセンサスとなり、そのための議論が Mulry et al.(2021) として発行されている。その Mulry et al.(2021) によると、1940 年代に行われたセンサスと行政記録の比較においてセンサスの方が良い捕捉率を有しているとは必ずしも言えなかったことから、捕捉率の評価の研究が始まり、長年にわたりさまざまな行政記録との比較を行ってきてい

る。米国には全人口を捕捉する一つの行政記録がないため、2000 年人口センサス以降、センサス仕様の行政ファイルを作成する試みが行われ、2010 年人口センサスではさまざまな業務で行政記録が使用され、さらに 2012 年から 2018 年にかけて行政記録の利用についてさまざまな研究が行われた。Mulry et al.(2021) では、2018 年の 2020 年センサス最終試験の結果を受けて行われた接触戦略の修正[1]、2020 年センサスで用いられた居住状態の判定モデル[2]、当初の行政記録利用計画、コロナ禍を受けて行われた変更について解説した後、2020 年センサスでの行政記録の利用についてまとめている[3]。

NRFU 以外の個別のプロジェクト[4]として、センサスでの行政記録の利用を検討しているものについては、CSRM (2021) によると、「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、異常値検出について「覚書 (memorandum)」、「概要 (outline)」と称するいくつかの文書を作成した。また、2010 年のデータに基づいて世帯人員を予測する多項ロジスティックモデルを作成し、2020 年の行政記録による世帯データに当てはめてセンサスデータと比較し、良好な結果を得た。

「2020 年センサス NRFU 削減目標のための『良い』行政記録を見つける (Identifying "Good" Administrative Records for 2020 Census NRFU Curtailment Targeting)」プロジェクトでは、先住民居留地で居住している住戸及びキャンパスの外にある大学の住宅の特定と計上のためのモデルの適用について文書を作成した。

（小売統計におけるビッグデータ利用の研究）

小売統計でビッグデータを利用する研究については、坂下 (2021) でも報告した。CSRM (2021) は事業所レベルの企業登録データ、州レベルの経済データ、州レベルの空間的ランダム効果を用いて州レベルの小売売上を推定するための階層的ベイズ・インピュテーション・モデルのレポート "Monthly State Retail Sales (MSRS) Technical Documentation" を作成したことを報告し、MSRS を回答者の負担を最小限に抑えつつ、利用者のニーズを満たす、よりタイムリーできめ細かい、適切なデータ製品を提供するための一歩と評価している[5]。このレポート[6]では、情報源として月次小売業調査 (Monthly Retail Trade Survey, MRTS) の標本と第三者機関の POS データを挙げ、州レベルの月次推計手法を解説し、インピュテーションについては、「総額、州、NAICS（北米産業分類）コードから月次小売売上高を予測するために回帰とランダム効果パラメータを使用する線形混合・ベイズモデル」で「事後予測

[1] 具体的な内容は坂下 (2021) 参照のこと。
[2] 坂下 (2021) 参照。
[3] 自己回答率が 65.28%で前回とほぼ同様であり、成功とみられる一方、4.59%が行政記録を利用し、インピュテーションを減らすことができたとしている。
[4] 坂下 (2021) 参照。
[5] 従来の月次小売業調査が調査統計であったのに対し、ビッグデータを利用して拡張し、月次小売高売上 (MSRS) 報告とした。
[6] 解説動画が以下のサイトに示されている。https://www.youtube.com/watch?v=zBe6yQEJ1vQ

分布からの多重代入法により複数の単位 (unit) のある事業所の欠測を推定する」としている。

（地域社会調査 (American Community Survey, ACS) の無回答問題）

コロナ禍による地域社会調査 (American Community Survey, ACS) への影響について、Rothbaum et al. (2021) は、無回答が大幅に増加し、その度合いが社会階層により異なることにより、従来の補正法では推計にバイアスをもたらすと指摘している。そのため、幅広い行政情報、民間のデータ、2010 年センサスのデータを用いて、回答世帯と非回答世帯の特性を解明し、「エントロピー・バランシング」[7]を用いたウェイト調整を試み、住居特性、社会的特性、収入と貧困、労働力と雇用の状況、健康保険への加入などの項目について検証し、2019 年から 2020 年にかけての変化を著しく減少させたとしている。なお、CSRM (2021) では、CSRM と調査部門が共同で ACS でデータ収集が中断したことによる大量の無回答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、2021 年度にはウェイト調整 (calibration weighting) に関する文献調査、ユニット欠測データの規模と性質を変化させた「エントロピー・バランシング」に関する研究計画を作成したとしている。

（住宅関係データのインピュテーションにおける課税記録の利用）

住宅調査 (American Housing Survey, AHS) 及び ACS の敷地面積と築年数のインピュテーションについて、Molfino (2021a) と Molfino (2021b) は、固定資産税の記録を利用する検討を行っている。Molfino (2021a) は、敷地面積について AHS や ACS のデータをマッチングできる場合は固定資産税のデータを代入する「コールド・デック」法が、できない場合は地域ごとの累積分布関数 (local cumulative distribution function) を用いて補完する手法が従来のホット・デック法より優れているとしている。Molfino (2021b) も同様に、築年数についてマッチングできる場合はコールド・デック法の性能がホット・デック法をはるかに上回り、できない場合は累積分布関数を用いる手法がホット・デック法よりも良いとしている。

（米国センサス局における合成データの研究）

CSRM では、坂下 (2020) や坂下(2021) に記したベイズ統計に基づく経済センサスのインピュテーションと合成データの研究の他に、2010 年代前半から継続して、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を行っている。

Klein and Sinha (2016)[8]、Moura et al. (2017)、Klein et al. (2018)、Moura et al. (2018)、Klein et al. (2019)、Guin et al. (2021) などによると、 合成データに基づく推論を行う方法論は、Rubin (1987) [9]の多重代入法の概念を用いて開発されている。合成データには完全合成データと部

[7] 単純な逆確率や回帰に基づく重み付けと比べていくつかの利点があるとしており、参考資料として Hainmueller (2012) を挙げている。
[8] 文献により “Klein and Sinha (2015)” と記されることがあるが同内容。
[9] 坂下 (2019) 参照。

分合成データの 2 種類があり[10]、完全合成データの手法は Rubin（1993）によって提案され、それによって推論を行う方法は Raghunathan et al. (2003) によって開発された。部分合成データの手法は、Little (1993) によって提案され、推論を行う方法は、Reiter (2003)によって開発された。CSRM の一連の研究は、これらの基礎の上に、抽出方法、インピュテーションの方法、分布モデルをさまざまに変えて合成データの作成とそれによる推論を論じたものである。

多重代入法で多く用いられる抽出方法は、パラメータの事前分布の仮定と観測値をもとに事後分布を推定し、そこから導かれる変量の分布から複数のデータセットを抽出する事後予測抽出 (Posterior Predictive Sampling)であるが、Reiter and Kinney (2012) は、部分合成データでは、Reiter (2003) に記されているような事後予測抽出の代わりに、未知のパラメータに標本の観測値から得られる値を挿入（プラグイン）するプラグイン抽出[11]を用いることができるとしており、Klein and Sinha (2016)、Klein et al. (2018) などがこの手法によっている。

CSRM での近年の研究成果として、Klein and Sinha (2016)、Moura et al. (2017)、Klein et al. (2018) などが学会誌に投稿されており、Moura et al. (2018)、Klein et al. (2019)、Guin et al. (2021)、Guin et al. (2022) が CSRM の公式報告書として刊行されている。

学会誌に投稿された論文のうち、Klein and Sinha (2016) は、データが多変量正規分布又は多変量線形回帰モデルに従うという前提のもとで、プラグイン抽出により生成した単一代入による部分合成データによる推定を論じ、仮想データを用いたモンテカルロ法によるシミュレーション及び人口動態調査 (Current Population Survey, CPS) のデータによる実証的な評価を行って、通常の多重代入の場合と比較している。その結果として、このシミュレーションの前提のもとでは、単一代入による部分合成データでも適切な推定が行えると結論づけているが、一方でこの手法はモデルによっており、ただちに一般化できるものではないとしている[12]。Moura et al. (2017) は、多変量回帰モデルに対して、事後予測抽出によって生成された単一代入の合成データ及び新たに提案する固定事後予測抽出 (FPPS)[13]によって生成された多重代入の合成データによる推論を論じている。これらの手法は、推定の信頼区間が大きくなるため、精度は低下するが、データの秘匿性は増すと評価されている。Klein et al. (2018) は、Klein and Sinha (2016) の方法論を多重代入に拡張し、多変量線形回帰モデルのもとで、事後予測抽出及びプラグイン抽出により多重代入で生成した部分合成データによる

[10] 完全合成データと部分合成データについては、坂下(2021) の注釈 12 参照。
[11] 坂下 (2020) 参照。
[12] シミュレーションの結果として、単一代入による推定値の信頼区間は多重代入による場合より大きくなり、分析者が単一代入法によるデータを元データであるかのように単純に分析すれば不適切な推定になると指摘している。また、この結果では多重代入の方が効率的な推定ができるが、合成データ作成の目的である開示抑制の面では単一代入の方がすぐれているとしている。欠測値補完において多重代入法を用いるのは標本の不確実性の他にパラメータ推定の不確実性を反映するためであるが、部分合成データを作成する場合は、必ずしもその必要がなく、またこの例ではデータの分布に強い仮定を置いているために、単一代入でも適切な推定が可能だったものと思われる。
[13] 通常の多重代入のようにデータセットごとのパラメータの事後予測分布からの抽出を行わず、固定したパラメータを用いる手法。

推定を Klein and Sinha (2016) で用いられた有限標本用の推定と通常の多重代入法の推定で行って比較し、新たな推定方法は適切な結果を得るとしている。

CSRM の報告書では、Reiter (2003) や Raghunathan et al. (2003) の推定方法が漸近的なものであったのに対し、Moura et al. (2018) は多変量線形回帰モデルのもとで、プラグイン抽出により単一及び多重代入された合成データの尤度に基づく厳密な推定を論じ、単一代入の方がデータの保護に勝るが、多重代入で提供するデータセットの数が増すほど推定の信頼区間が小さくなるとしている[14]。Klein et al. (2019) もデータ保護の面では単一代入の方が望ましいという問題意識により、多変量正規モデルのもとで、プラグイン抽出により単一代入された合成データの尤度に基づく厳密な推論を論じ、提案する手法は期待される機能を満たしているとしている。Guin et al. (2021) は、多変量線形回帰モデルのもとで、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部分合成データを用いたベイズ統計学による推論手法を開発し、仮想データにより、まずはすべてのデータが秘匿を要すると仮定して分散や回帰係数を推定するシミュレーションを行い、予測通りの結果を得たとしたのち、一部のデータのみが秘匿を要する場合の取扱い[15]を論じ、今後の方針として事前分布の仮定をさまざまに変えること、インピュテーションのモデルと分析モデルが異なる場合の検討、現実のデータに起こりうるさまざまな理想的でない事象の取り込みなどを検討するとしている。一方、Guin et al. (2022) では、多重代入の場合について同様の分析を行う方法を示している[16]。

（センサス局その他）

CSRM は 2021 年度に季節調整と欠測値補完のソフトウェア Ecce Signum のリポジトリのある GitHub サイトを公開し、問い合わせに対応した。

（農業センサス (COA) 等）

農業センサス (COA) を始めとする米農務省農業統計サービス（NASS）の統計のエディットとインピュテーションについては、坂下 (2018) や 坂下 (2019) で報告した。Lipke et al. (2022) によると、NASS ではまた、多くの調査に適用できる汎用システム (IDEAL) を計画しており[17]、その警告処理（対話型の Blaise [18]では警告が出るたびに処理が止まってしまうのを、まとめて処理できるようにする）、全数調査層の無回答への対応、Blaise の自動化

[14] Hawala (2008) によると、複数の部分合成データセットを公開することは、開示リスクを増大させるおそれがあるため、センサス局では単一のデータセットのみ公開している。
[15] プラグイン抽出、事後予測抽出のそれぞれについて、インピュテーションに用いるデータを秘匿を要するもののみとするか、全データを用いるかの 2 通りの手法がある。
[16] なお、CSRM(2021) によると、多変量正規分布モデルによる場合について、同様の検討を行った論文を準備中である。
[17] Lipke et al. (2022) の前半をなすシステムのレビューについては、坂下 (2019) に既出。
[18] Blaise はオランダ統計局で開発されたコンピュータ支援面接調査 (CAPI) のシステムで、NASS では主に中小規模の調査で用いており、COA や大規模調査では内製によるエディットとインピュテーションのシステム PRISM を用いている。（坂下 (2019)）

について報告している。

（カナダ）

カナダ統計局は、エディットとインピュテーションのシステム Banff を開発し、長らくユーザーの意見を受けた改良を続けてきた（坂下 (2018)、坂下 (2019)）。Gray (2022) は 2024 年公開予定の新バージョンについて報告し、新しい Banff は SAS アーキテクチュアに依存せず、モジュール形式となる予定であり、公開後にはさらにユーザーと共同でモジュールを開発したいと呼びかけている。

（メキシコ）

メキシコの 2020 年人口センサスにおける自動データ・エディティングとインピュテーションについて Vielma Orozco (2022) は、データ・エディティングが行われたのが COVID-19 の流行期にあたったため、在宅勤務の必要が発生し、追加のリソースや情報の取扱いについてどこのデータか分からないようにするなどの配慮が必要となったこと、予定が遅れたこと、パンデミック状況下で接触不可能な事例が増えたこと、未成年者の申告漏れが伝統的に多いため、行政記録とセンサスの情報を分析し、7 歳未満の子供のインピュテーションを行うことを決定したことなどを報告している。

また、メキシコの経済センサスについて Hernandez (2022) は、95 パーセント近くが零細事業所である一方、0.5 パーセントに過ぎない大規模事業所が雇用者の 32 パーセントと付加価値の半分以上を占めているなどの特徴を述べた後、大規模事業所と中小零細規模事業所に分けてインピュテーションを論じている。大規模事業所のインピュテーションは、その事業所が月次の抽出調査に含まれればセンサス年の対応するデータを利用し、月次調査に含まれない定性的な情報については最近隣の事業所のデータを利用する。月次の抽出調査に含まれていなければ、税務データ又は公開された損益計算書を利用する。中小零細企業のインピュテーションは、平均値によっている。

（チリ）

チリにおける企業の月次売上高の付加価値税報告を用いたデータセットの外れ値検出について、Vasquez et al. (2022) は従来の手法と機械学習による手法を比較し、選択的エディティングの自動化を提案している。具体的には、従来の手法として四分位 (IQ) 距離による方法、HB 法（Hidiroglou and Berthelot (1986)）、IQ 法と HB 法のハイブリッド（両方で外れ値とされた値を外れ値とする）、機械学習法として DBSCAN（Ester et al. (1996) によって提案されたデータクラスタリングアルゴリズム）の 4 種類を対象とし、DBSCAN は一変数では良い成績をもたらすが、二変数では外れ値を検出し過ぎ、精度が落ちるとの結果を得ている。また、機械学習手法には時間がかかり過ぎる問題があるが、これは機械学習手法の対象とするデータを限定することで改善できるとしている。

２．欧州

欧州においては、国連の欧州経済委員会 (United Nations Economic Commission for Europe, UNECE) などにおける情報の交換・共有や欧州統計局 (Eurostat) と加盟各国の協力体制である欧州統計システム (European Statistical System, ESS) などを通じた共同プロジェクト[19]が盛んに行われている[20]。近年の傾向としては、インピュテーションを行う単一の技法の開発よりも、統計のプロセス全体を見据えたシステム開発、エディットとインピュテーションのための機械学習、エディットとインピュテーションにおける行政データの活用などが中心的な話題となっている。

（イタリア）

イタリア統計局 (ISTAT) の第 7 回農業センサスにおけるエディットとインピュテーション (E&I) 処理の複合システムについて Rosati et al. (2022) が報告している。農業センサスでは、量的変数と質的変数の双方を扱うため、選択的エディティングのための R パッケージ SeleMix により外れ値とエラーを検出し、Banff により量的変数の E&I を行い、残りは R パッケージで対応する。R は当初は試験的な導入だったが、質と量の混合した変数を扱うことができ、大規模で複雑なデータセットを容易に管理できるため、今後は推進していく予定である。

また、ISTAT の E&I システムについて、Buglielli et al. (2022) によると、Fellegi-Holt 法に基づく E&I の自動化システム SCIA[21]は質的データしか扱えなかったが、近年の R パッケージ（インピュテーション部分はオーストリア統計局で開発した VIM[22]を使用）は処理ステップは多くかかるものの、大きく複雑なデータの扱いが SCIA より柔軟で、複数の混合型変数を同時に扱えるため、比較の結果こちらにさらに投資するべきとしている。

坂下 (2019) 及び坂下 (2021) では ISTAT で個人レジスタの最終学歴データの作成を行政記録からの対数線形モデルを用いた「マス・インピュテーション」によって作成していることを報告した。De Fausti et al. (2022a) 及び De Fausti et al. (2022b) によると、近年では多層パーセプトロンのようなニューラル・ネットワークを用いる検討を行い、対数線型モデルと比較して精度は変わらないが効率が向上するとの結果を得ている。また、推定値の分散の推計について Di Zio et al. (2022) は、解析的手法及びモンテカルロ法による推計法の比較を行い、最終学歴のマス・インピュテーションによる分散は非常に小さく、また解析的手法はブ

---

[19] 坂下(2017)、坂下(2018)、坂下(2021) 参照。

[20] ただし、UNECE の専門家会合は、2018 年までは 3 年に 2 回の頻度で行われていたが、その後はコロナによる延期もあって 2 年に 1 回に落ち、最近の 2 回（2020 年及び 2022 年）はオンライン開催となっている。

[21] 伊語 Sistema per il Controllo e l'Imputazione Automatici （自動的なコントロールとインピュテーションのシステム）。

[22] Visualization and Imputation of Missing Values（欠測値の可視化とインピュテーションのための R パッケージ）、坂下(2017)、坂下(2018)、坂下(2019)参照。

ートストラップ法のように時間がかからず、有望であるとしている。

（英国）

機械学習を用いたインピュテーションとして、英国国家統計局 (ONS) で研究している Household Financial Survey (HFS)[23] のランダムフォレスト法によるインピュテーションについて Edward (2022) が報告している。この論文は既存文献のレビュー[24]を含み、その上でランダムフォレスト法には他の手法と比べてオープンソースのパッケージが提供されている、混合データ、相互作用や非線形（回帰）効果を扱うことができるなどの実務的な利点があるとしている。シミュレーションでは、MissForest と MICE の 2 種類のパッケージによるランダムフォレスト法をいくつかの他の手法と比較し、ランダムフォレスト法はパッケージにより異なって性能にばらつきがあり、連続変数では MissForest によるランダムフォレスト法の成績が良く、2 値のカテゴリー変数のインピュテーションでは手法による差は小さいとの結果を得ている。

（オランダ）

オランダ統計局で構想している企業統計を作成するための新たな統合システムについて Vaasen-Otten et al. (2022) 及び Scholtus et al. (2022)が報告している。このシステムでは、早期集計のためにミクロデータが利用でき、前期データや行政記録によるインピュテーションが行われ、回答が得られたときに置き換えられる。また、このシステムではデータの品質が自動的に測定され、スコア関数を用いた判定により選択的な対話型エディティングが行われる。前者では手動エディティングの効果を最大化するための品質指標（スコア関数）について、後者では一部のデータが重複する複数の情報源を用いた場合の自動エディティングについて既存手法のレビューを交えつつ論じ、実証研究について報告している。

ESS で力を入れているデータ検証に関連して、Ten Bosch et al. (2022) は、データからルールを推論する（データ主導型）アプローチの R パッケージ validatesuggest について報告している。これはデータから正値、範囲、無回答、一意性などのルールを生成するもので、生成したルールは人間が確認できるように表示される。また、Van der Loo (2022) は、統計の作成に当たって、扱うテーマについての知識と IT についての知識はなるべく切り離されるべきとの認識のもとに、データ検証、データクリーニング、データ結合などのルールのマネージメントについて論じ、ルール・マネージメントのパッケージ[25]を紹介している。

（スイス）

---

[23] 「家計調査」と訳されることがあるが、日本の家計調査とは異なり、金融を中心とした生活状況に関する調査である。

[24] 冒頭に「今日のインピュテーション手法は統計的インピュテーションと機械学習によるインピュテーションに大別される」と記している。

[25] https://github.com/SNStatComp/rulemanager

スイス連邦統計局が一定のパターンのない「スイス・チーズ型」欠測に対して k-最近隣法を加工した「バランスした」インピュテーションを検討していることについては坂下 (2019) で報告した。Leuenberger (2022) は、所得と生活状況に関する調査 (Survey on Income and Living Conditions, SILC) のデータを用いてこれと機械学習法を比較するシミュレーションを行い、その結果、ニューシャテル大学によって開発された「バランスした」インピュテーションのアルゴリズム[26]は分布の端にある値に対しては良い結果を得るが中央近い値に対しては少し劣るのに対し、対照のために試行した機械学習法の MissForest は中央に近い値で特に良いとの結果を得た。

また、MissForest の SILC への適用については、カテゴリー変数と量的変数を同時に処理することができ、データの構造と分布について特定の仮定をしない機械学習アルゴリズムであると Bianchi (2022) が論じている。その中で、SILC の個人変数を対象としてシミュレーションを行い、欠測の多い物質的・社会的窮乏に関する変数（新しい衣服を買う余裕がある、友人と飲食に行ける、等）も、世帯変数を補助変数とすることにより、高精度のインピュテーションができるなどの結果を得ている。

スイスでは、事業所の売上げを付加価値税 (VAT) によって計測しているが、4 割に上る事業所は VAT を免除されているかグループでまとめて支払っているため、直接の計測ができず、産業部門や従業者数によってインピュテーションを行っていると Saliba (2022) が報告している。Saliba (2022) は、完全に欠測しているものに対しロバストな回帰と MissForest 法の 2 つの手法を産業分類のレベルなどのパラメータを変化させて試みることにより、グループでまとめて支払っているものについては制約条件のある非線形の最適化を伴った調整を行うことにより品質を改善できるとしている。

（スペイン）

スペイン統計局（INE）は近年、選択的エディティングによる効率化に取り組んできた。Barragán and Salgado (2022) は、従来のスコア関数による判定に機械学習を加える手法について、準連続変数（ゼロ又は連続した分布を取る変数）の判定、過去値を用いた収集段階でのインピュテーション、エラーの非対称性（第一種と第二種のエラーに与えるスコアの比）などの観点から論じ、行政記録やデジタルデータなどが大量に入ってくるようになったため、従来の設計に基づく統計にもまして、エディティングと検証が重要になってくること、長期的な研修プログラムが必要であることを指摘している。

（ドイツ）

インピュテーション手法による当てはまりの評価について、Thurow et al. (2022) は、2010 年のドイツの所得構造調査を用いたシミュレーションを報告している。シミュレーションは、単変量と多変量それぞれの場合に、Missing Completely at Random (MCAR) と Missing at

[26] https://github.com/EstherEustache/SwissCheese

Random (MAR) の発生メカニズムについて行われ、単変量の評価では、予測の当てはまりと分布の当てはまりを分け、前者の評価は標準化平方平均二乗誤差 (NRMSE) と誤分類比率 (PFE) により、後者はコルモゴロフ=スミルノフ (KS) 統計量によっている。一方、多変量の評価は、分布を示す 16 の変量の結合によっている。評価の対象としたのは、一般に適用されている 5 つの手法、Amelia、MissRanger、ランダムフォレスト法による MICE、予測平均マッチングによる MICE、通常（ベイジアン）モデルによる MICE (MICE.Norm)である。結果は、単変量では MissRanger による予測の当てはまりが良いが、分布の当てはまりは良くないことがあり、分布については MICE.Norm が良い。多変量についても同様に MICE.Norm が良い結果を出した。

また、Ditscheid (2022) は、ドイツ連邦統計局が行った月次統計の早期化のためのインピュテーションの実験について伝えている。ここでは、製造業における売上高と新規受注の指数のための報告が企業から上がってこない場合にインピュテーションを行うという想定で平方平均二乗誤差による評価を行い、単一変量（前月の値）による手法、多変量でツリーによる手法、ドナーによる手法、回帰による手法を比較して、予備的な実験では回帰による手法が最も良い結果を示したが、適切な説明変数を得るための最適化や多重共線性を回避するための実験を続ける必要があるとしている。

（ポーランド）

ポーランドが 2021 年の人口センサスに向けて、行政レジスタの整備を進めていることについては坂下 (2021) に報告した。Murawski (2022) は、世帯のインターネットへのアクセスに関する調査のインピュテーションに際して行政レジスタからのデータを用いる研究について報告している。ここではプロバイダから提供されるインターネットへのアクセスのデータを住所情報によって 2019 年の情報通信技術利用調査に統合し、統合できなかったレコードについてホット・デック法によってインピュテーションを行っている。

３．一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

坂下 (2020) では、総務省統計局と独立行政法人統計センターが共同で開発し、統計センターが提供している「一般用ミクロデータ」を用いて数値シミュレーションを行った。その目的は、「海外では標準的な手法であるが日本ではあまり用いられていないホット・デック法の数値シミュレーションを行い、その課題について検討する」ことで、「もっとも基本的なデータ」として「全国消費実態調査（平成 21 年）十大費目勤労者世帯」を用いて、「適切なドナー・プールを作成すれば（正しい値からの乖離を）小さく押さえることができると考えられる。」との結論を得ている。ただし、全国消費実態調査は、収支項目ごとに家計簿から一ヶ月平均の値を計算しているため、ミクロデータが既に集計値になっていて調査における実際の欠測値の発生状況は反映していない問題があり、またホット・デック法は経済統計に多い数量データよりも人口統計などの質的（カテゴリー）データに適用されることが多いことから、今回は平成 24 年就業構造基本調査の一般用ミクロデータを用いて質的項目におけるインピュテーションの数値シミュレーションを行った。

（基本数）

シミュレーションの前に、前回同様、基本数の確認を行った。提供側から示されている基本数として、「都道府県別 15 歳以上人口」、「男女別 15 歳以上人口」、「年齢５歳階級別 15 歳人口」」、「産業別 15 歳以上人口」などがあり、これらのクロス集計を行って周辺分布が一致することを確認した。主なものを以下に示す。

表１－１　都道府県別・年齢別 15 歳以上人口

	15～19歳	20～24歳	25～29歳	30～34歳	35～39歳	40～44歳	45～49歳	50～54歳	55～59歳	60～64歳	65～69歳	70～74歳	75歳以上	計
北海道	248,100	250,900	273,200	309,800	378,400	374,400	343,300	344,000	376,600	482,900	368,400	335,500	718,100	4,803,600
青森県	67,300	50,600	60,500	71,500	85,200	87,300	83,600	91,600	99,500	118,700	89,000	83,600	191,500	1,179,900
岩手県	63,600	51,000	60,100	70,000	82,200	81,000	78,100	85,600	94,000	110,800	82,200	81,800	199,700	1,140,100
宮城県	111,700	128,400	134,600	148,400	167,800	159,400	142,400	146,300	158,900	189,400	133,400	122,400	277,500	2,020,600
秋田県	48,300	33,400	44,000	53,700	63,200	62,100	61,300	71,600	82,500	97,500	71,500	70,500	183,900	943,500
山形県	56,400	42,900	53,700	62,900	70,300	68,000	66,600	75,300	83,400	99,200	70,500	68,700	186,600	1,004,500
福島県	101,900	79,200	95,200	108,200	125,300	121,300	118,300	131,400	145,200	170,500	117,200	111,700	282,300	1,707,700
茨城県	146,200	135,600	157,800	176,600	211,200	211,200	182,600	179,100	201,900	249,500	195,700	170,200	335,200	2,552,800
栃木県	95,600	87,700	108,300	122,800	146,400	142,200	124,500	124,800	139,700	169,900	126,200	107,500	228,500	1,724,100
群馬県	100,400	85,800	98,100	115,500	144,000	145,300	124,700	118,200	129,700	165,900	134,400	115,300	246,000	1,723,300
埼玉県	348,500	388,100	406,100	457,200	571,100	588,200	487,700	423,500	436,900	576,100	493,200	428,700	662,900	6,268,200
千葉県	284,000	311,000	343,800	386,700	478,400	498,000	411,100	363,600	380,100	505,400	435,300	374,600	627,300	5,399,300
東京都	526,800	779,600	940,600	1,011,200	1,132,300	1,140,800	970,700	795,100	716,300	901,500	767,500	697,100	1,346,900	11,726,400
神奈川県	412,600	496,600	536,500	603,800	735,500	776,900	657,500	540,200	507,400	666,600	566,800	501,200	879,800	7,881,400
新潟県	114,600	94,800	113,100	130,400	156,400	151,900	141,100	144,400	163,300	203,000	149,500	140,100	349,500	2,052,100
富山県	50,800	40,600	50,700	59,800	78,300	76,700	64,200	62,400	68,400	93,400	78,000	66,300	154,800	944,400
石川県	57,900	56,500	58,800	66,200	85,200	82,100	70,000	68,600	72,000	96,600	78,300	64,900	147,900	1,005,000
福井県	41,200	32,500	39,400	44,500	53,900	52,800	48,400	49,600	51,700	67,700	49,200	44,600	113,700	689,200
山梨県	45,900	40,200	40,500	45,700	56,100	59,900	56,400	53,300	56,500	68,600	53,900	49,600	114,900	741,500
長野県	105,100	76,400	98,100	117,600	147,600	148,200	131,300	127,800	133,900	172,100	140,500	128,100	315,900	1,842,600
岐阜県	103,800	94,100	103,100	116,400	143,600	145,200	127,000	122,700	128,600	171,800	136,300	123,800	259,600	1,776,000
静岡県	178,300	147,400	193,900	219,900	267,700	272,000	238,100	228,600	240,200	310,600	248,000	224,500	459,000	3,228,200
愛知県	361,800	401,600	451,500	489,100	589,800	595,700	491,800	428,300	415,300	549,000	460,400	408,400	722,000	6,364,700
三重県	91,200	80,600	93,800	105,300	131,500	132,500	116,900	111,400	114,400	148,800	120,100	109,000	235,500	1,591,000
滋賀県	72,300	78,900	80,600	89,600	107,300	104,300	87,800	83,300	84,700	108,900	84,000	71,400	150,800	1,203,900
京都府	122,600	163,400	148,300	155,300	190,900	190,600	161,900	146,200	147,700	213,700	177,600	156,700	314,600	2,289,500
大阪府	415,000	470,400	507,000	545,500	679,000	713,200	587,000	498,400	490,100	696,000	601,900	559,200	937,700	7,700,400
兵庫県	270,500	268,600	288,600	322,700	410,600	421,000	360,600	331,300	336,600	449,700	371,500	330,800	652,600	4,815,100
奈良県	71,200	69,900	69,200	74,600	94,800	99,400	87,600	82,400	86,900	119,000	98,200	88,400	167,800	1,209,400
和歌山県	49,100	38,300	44,300	49,400	63,300	66,200	60,900	61,100	64,400	85,200	68,800	64,300	147,600	862,900
鳥取県	28,700	22,500	28,100	32,600	37,800	35,200	33,100	36,000	41,300	50,000	36,300	32,800	89,300	503,700
島根県	34,400	24,100	30,800	36,700	43,900	41,200	38,000	42,900	48,200	62,400	46,900	41,900	122,900	614,300
岡山県	95,200	96,300	102,800	110,000	136,500	131,200	110,100	110,200	117,800	156,500	129,700	112,200	265,400	1,673,900
広島県	137,300	133,600	148,500	165,500	207,100	203,500	170,100	165,900	175,300	231,500	194,000	164,400	360,900	2,457,600
山口県	65,800	56,800	65,500	73,500	93,100	91,300	78,300	82,400	94,000	129,300	104,100	93,700	220,400	1,248,200
徳島県	36,000	31,400	37,000	42,200	50,200	48,700	44,900	47,300	54,600	70,600	50,400	46,600	120,000	679,900
香川県	45,900	38,200	47,200	55,500	70,100	66,600	56,100	58,200	64,300	87,200	67,200	57,900	142,700	857,100
愛媛県	67,800	54,300	66,600	76,800	94,700	91,600	82,900	86,300	94,300	124,900	94,900	86,600	211,100	1,232,800
高知県	35,600	27,800	32,400	38,800	49,000	47,000	42,100	45,600	50,200	67,000	53,800	47,500	125,100	661,900
福岡県	245,200	278,900	290,500	320,200	366,100	351,400	302,900	302,500	328,200	419,600	310,800	280,600	594,100	4,391,000
佐賀県	44,700	35,900	42,800	47,400	53,400	50,900	48,500	53,600	59,600	70,600	49,200	46,600	117,700	720,900
長崎県	71,500	53,000	64,000	72,300	85,400	85,900	82,700	91,800	102,400	123,100	87,000	85,100	207,900	1,212,100
熊本県	90,000	79,700	91,400	101,100	113,100	107,400	105,900	114,200	126,700	149,100	108,400	102,500	267,400	1,556,900
大分県	56,400	49,800	57,900	65,800	77,600	73,100	66,100	71,500	81,100	103,300	78,200	71,600	177,300	1,029,700
宮崎県	57,200	42,500	52,900	61,800	71,200	66,400	63,500	71,500	81,900	98,600	69,000	66,100	165,200	967,800
鹿児島県	83,800	67,600	82,200	92,700	101,300	96,700	97,900	110,000	126,200	142,100	97,200	98,000	261,700	1,457,400
沖縄県	83,600	73,300	84,500	90,700	105,100	98,500	87,400	91,000	95,300	97,700	56,300	61,700	131,800	1,156,900
計	6,041,800	6,240,700	7,018,500	7,813,900	9,402,900	9,454,400	8,193,900	7,671,000	7,948,200	10,241,500	8,200,900	7,394,700	15,191,000	110,813,400

表１－２　年齢別・男女別 15 歳以上人口

	男	女	計
15～19歳	3,093,800	2,948,000	6,041,800
20～24歳	3,182,700	3,058,000	6,240,700
25～29歳	3,563,200	3,455,300	7,018,500
30～34歳	3,954,400	3,859,500	7,813,900
35～39歳	4,760,200	4,642,700	9,402,900
40～44歳	4,772,700	4,681,700	9,454,400
45～49歳	4,116,200	4,077,700	8,193,900
50～54歳	3,833,800	3,837,200	7,671,000
55～59歳	3,943,500	4,004,700	7,948,200
60～64歳	5,020,700	5,220,800	10,241,500
65～69歳	3,931,700	4,269,200	8,200,900
70～74歳	3,439,300	3,955,400	7,394,700
75歳以上	5,800,700	9,390,300	15,191,000
計	53,412,900	57,400,500	110,813,400

表１－３　都道府県別・男女別 15 歳以上人口

	男	女	計
北海道	2,235,000	2,568,600	4,803,600
青森県	545,000	634,900	1,179,900
岩手県	538,500	601,600	1,140,100
宮城県	973,600	1,047,000	2,020,600
秋田県	437,200	506,300	943,500
山形県	476,300	528,200	1,004,500
福島県	823,000	884,700	1,707,700
茨城県	1,265,700	1,287,100	2,552,800
栃木県	850,300	873,800	1,724,100
群馬県	841,900	881,400	1,723,300
埼玉県	3,126,400	3,141,800	6,268,200
千葉県	2,674,000	2,725,300	5,399,300
東京都	5,763,400	5,963,000	11,726,400
神奈川県	3,932,700	3,948,700	7,881,400
新潟県	982,900	1,069,200	2,052,100
富山県	450,800	493,600	944,400
石川県	480,600	524,400	1,005,000
福井県	330,100	359,100	689,200
山梨県	359,200	382,300	741,500
長野県	887,400	955,200	1,842,600
岐阜県	851,300	924,700	1,776,000
静岡県	1,576,600	1,651,600	3,228,200
愛知県	3,163,000	3,201,700	6,364,700
三重県	768,300	822,700	1,591,000
滋賀県	590,300	613,600	1,203,900
京都府	1,084,900	1,204,600	2,289,500
大阪府	3,681,400	4,019,000	7,700,400
兵庫県	2,273,400	2,541,700	4,815,100
奈良県	564,500	644,900	1,209,400
和歌山県	400,900	462,000	862,900
鳥取県	236,800	266,900	503,700
島根県	290,000	324,300	614,300
岡山県	793,400	880,500	1,673,900
広島県	1,170,300	1,287,300	2,457,600
山口県	581,400	666,800	1,248,200
徳島県	319,100	360,800	679,900
香川県	408,600	448,500	857,100
愛媛県	572,100	660,700	1,232,800
高知県	306,900	355,000	661,900
福岡県	2,041,000	2,350,000	4,391,000
佐賀県	334,300	386,600	720,900
長崎県	554,100	658,000	1,212,100
熊本県	719,700	837,200	1,556,900
大分県	479,400	550,300	1,029,700
宮崎県	446,900	520,900	967,800
鹿児島県	669,800	787,600	1,457,400
沖縄県	560,500	596,400	1,156,900
計	53,412,900	57,400,500	110,813,400

表１－４　都道府県別・産業別 15 歳以上人口

	農業，林業	漁業	鉱業，採石業，砂利採取業	建設業	製造業	電気・ガス・熱供給・水道業	情報通信業	運輸業，郵便業	卸売業，小売業	金融業，保険業	不動産業，物品賃貸業	学術研究，専門・技術サービス業	宿泊業，飲食サービス業	生活関連サービス業，娯楽業	教育，学習支援業	医療，福祉	複合サービス事業	サービス業（他に分類されないもの）	公務（他に分類されるものを除く）	分類不能の産業		計
北海道	119,299	33,704	2,036	220,530	231,202	17,643	41,362	145,863	427,228	58,163	48,744	69,713	172,757	113,521	120,807	335,030	34,883	198,228	126,835	90,299	2,195,754	4,803,600
青森県	68,871	12,151	521	62,501	65,513	3,501	7,017	32,138	102,376	12,878	2,944	12,888	34,830	24,043	26,398	81,310	8,998	41,651	31,383	15,249	532,737	1,179,900
岩手県	57,177	5,809	488	63,553	91,820	2,534	7,931	31,698	105,330	12,273	9,698	13,945	36,580	23,447	26,774	72,772	11,257	35,682	22,860	10,237	498,236	1,140,100
宮城県	46,124	4,531	497	121,565	140,277	9,957	23,924	63,904	201,195	26,419	23,329	34,893	64,205	36,386	56,128	118,071	13,445	79,879	45,839	29,493	880,537	2,020,600
秋田県	46,562	999	487	46,612	72,727	1,970	3,439	21,568	80,501	9,708	3,961	11,083	24,284	22,764	20,232	67,423	7,368	27,323	24,427	8,251	441,811	943,500
山形県	55,292	487	0	49,788	108,974	2,614	4,427	20,978	85,964	10,949	9,023	11,520	34,422	22,711	23,277	64,476	6,887	29,431	22,413	10,425	430,441	1,004,500
福島県	61,331	0	1,017	94,481	193,292	7,587	10,124	45,128	128,268	18,773	11,953	22,807	50,343	38,183	40,939	100,590	10,143	47,338	36,399	24,929	764,076	1,707,700
茨城県	85,906	521	0	111,515	322,858	6,112	31,939	82,958	212,107	28,757	20,665	56,171	68,858	58,780	61,720	133,647	11,589	79,781	52,099	49,814	1,077,003	2,552,800
栃木県	50,033	0	1,515	79,297	242,816	4,989	8,557	49,737	137,809	17,544	13,564	39,123	57,188	44,403	42,962	92,859	9,055	48,885	33,072	34,964	715,729	1,724,100
群馬県	54,544	0	516	74,294	243,098	5,063	12,063	48,533	143,878	19,757	14,010	25,836	53,318	35,296	42,336	116,308	9,544	51,768	34,710	33,580	704,847	1,723,300
埼玉県	64,139	481	0	271,912	640,239	12,185	141,943	246,488	602,088	106,600	88,660	133,196	209,736	131,319	160,840	341,763	19,343	241,129	115,305	166,412	2,574,422	6,268,200
千葉県	99,844	3,511	0	252,299	383,732	14,113	131,244	227,612	493,561	95,806	74,569	114,892	179,697	151,017	145,379	297,819	18,152	226,783	124,457	126,059	2,238,754	5,399,300
東京都	27,980	1,014	2,996	418,343	840,258	23,187	574,023	348,687	1,081,216	286,525	254,014	486,043	489,866	275,397	340,918	641,492	36,222	571,192	227,752	380,076	4,419,200	11,726,400
神奈川県	34,475	1,481	2,982	341,442	720,817	20,790	282,757	256,840	680,878	136,344	145,304	213,916	277,300	152,501	225,461	461,098	21,827	325,580	114,383	238,234	3,226,990	7,881,400
新潟県	70,138	501	4,003	114,644	217,140	7,976	12,471	59,393	190,376	22,440	14,973	22,509	61,284	52,917	49,779	131,098	12,990	60,237	46,232	18,933	882,066	2,052,100
富山県	16,085	464	488	42,886	127,888	2,473	11,367	27,791	77,536	14,368	7,797	12,879	32,254	21,945	22,554	67,575	4,876	34,349	14,861	13,648	390,316	944,400
石川県	14,515	2,405	0	48,018	111,234	6,397	14,006	23,815	97,729	13,303	8,809	15,481	40,704	22,078	28,190	71,124	5,459	30,468	21,320	17,209	412,737	1,005,000
福井県	14,276	2,777	0	42,003	88,465	4,960	4,966	13,713	62,660	6,281	4,884	12,337	21,701	17,502	18,969	46,323	4,437	20,142	12,772	12,480	278,151	689,200
山梨県	29,465	0	0	31,490	86,932	1,484	3,030	17,713	64,477	10,310	4,787	11,784	30,034	17,698	19,174	43,893	4,524	22,377	16,392	13,599	312,338	741,500
長野県	101,648	511	518	93,769	231,248	5,045	16,630	45,342	148,972	20,118	14,321	27,264	60,438	37,164	49,309	120,379	13,534	51,559	36,684	25,139	743,008	1,842,600
岐阜県	35,531	0	0	83,977	247,411	5,576	14,592	47,708	156,878	23,629	14,560	29,681	71,429	41,543	46,289	109,799	9,052	55,299	35,211	20,529	727,307	1,776,000
静岡県	83,621	5,993	1,503	147,346	466,996	13,485	29,070	102,865	308,030	38,399	25,024	45,551	118,704	67,381	77,190	173,836	14,471	109,548	48,503	49,669	1,301,015	3,228,200
愛知県	84,036	8,624	509	292,028	1,013,103	19,295	72,021	217,600	579,556	81,577	62,093	113,283	217,240	136,689	164,817	353,189	20,790	219,811	98,968	133,479	2,475,995	6,364,700
三重県	34,614	12,044	508	64,923	216,299	6,013	10,587	54,934	123,784	16,059	7,539	20,512	46,045	33,307	48,903	98,080	7,954	47,031	33,731	43,311	664,822	1,591,000
滋賀県	20,098	482	975	44,256	187,886	3,936	7,685	35,331	95,081	16,322	14,179	17,954	34,551	24,504	39,341	76,261	4,951	37,117	25,873	29,612	487,505	1,203,900
京都府	22,328	486	0	69,641	212,979	5,009	27,483	69,170	202,474	28,381	23,918	45,937	82,985	50,171	87,945	142,477	8,484	81,016	47,744	69,616	1,011,255	2,289,500
大阪府	16,111	1,512	997	297,917	711,110	24,175	119,989	282,215	728,217	109,421	140,885	149,848	253,509	139,825	188,456	473,819	23,696	287,878	101,872	235,566	3,413,383	7,700,400
兵庫県	38,008	3,993	509	159,176	517,275	13,108	51,775	156,475	413,234	67,026	63,707	94,117	148,480	94,303	129,169	300,476	22,674	159,648	73,501	101,522	2,206,923	4,815,100
奈良県	13,659	0	0	40,147	101,885	3,528	14,080	22,122	104,054	16,463	13,909	22,413	36,556	24,106	40,961	73,736	5,901	44,916	26,951	26,345	577,668	1,209,400
和歌山県	38,949	1,952	0	39,858	64,279	3,474	5,770	23,709	65,079	11,164	4,906	12,375	22,217	15,945	25,105	58,626	5,993	22,197	24,020	14,569	402,714	862,900
鳥取県	24,795	964	0	23,660	38,390	446	2,904	11,012	38,103	6,240	2,890	4,723	17,997	7,473	13,786	37,589	7,205	17,112	13,512	9,094	225,805	503,700
島根県	27,133	2,828	0	29,886	47,522	1,425	2,885	14,065	52,482	9,249	5,388	8,223	16,729	9,153	17,719	51,170	5,312	20,324	15,389	9,116	268,302	614,300
岡山県	43,267	2,533	1,019	71,882	176,925	5,466	13,868	48,899	136,974	18,843	12,282	26,229	43,613	34,640	50,369	123,527	8,421	56,460	32,199	28,083	738,400	1,673,900
広島県	44,225	1,005	501	110,793	261,682	11,131	23,675	85,076	222,215	30,061	27,062	40,159	72,725	51,032	68,575	163,264	10,517	76,288	38,717	43,543	1,075,353	2,457,600
山口県	30,093	4,456	497	56,491	110,000	4,470	7,949	35,205	102,638	13,246	5,473	16,949	37,168	23,541	35,379	91,232	11,365	44,169	24,857	17,601	575,419	1,248,200
徳島県	29,450	2,841	0	28,835	56,191	1,002	3,397	11,226	53,133	9,163	5,867	11,792	18,262	9,156	20,794	48,964	4,348	16,064	15,205	11,132	323,080	679,900
香川県	29,001	468	0	38,301	77,748	2,467	4,880	22,243	76,850	11,236	8,107	12,180	21,397	20,810	24,520	63,033	8,903	23,752	16,624	14,087	380,493	857,100
愛媛県	44,214	8,917	0	54,327	99,607	9,281	5,915	33,431	103,070	16,219	11,241	17,705	35,400	24,476	26,939	91,623	8,365	36,994	22,606	18,219	564,251	1,232,800
高知県	35,022	3,327	0	31,726	32,062	1,437	2,538	12,450	51,597	8,599	3,807	7,657	25,122	13,145	19,195	52,890	5,258	16,468	19,576	14,854	305,170	661,900
福岡県	70,673	7,521	511	187,556	306,537	17,064	57,042	142,978	407,954	58,083	51,944	75,575	145,406	90,702	113,727	343,404	16,658	157,439	82,311	88,049	1,969,865	4,391,000
佐賀県	36,766	3,832	0	35,053	66,800	1,972	4,900	20,515	57,409	5,795	3,283	12,712	21,515	11,545	18,812	51,463	6,810	24,661	17,986	4,901	314,169	720,900
長崎県	42,143	16,208	1,001	55,396	76,175	5,475	8,347	26,292	94,620	14,290	9,768	16,253	38,861	20,364	32,853	103,432	5,964	36,454	32,523	15,241	560,440	1,212,100
熊本県	92,835	7,874	0	68,508	108,757	3,503	9,381	33,230	132,186	16,777	16,011	19,817	49,481	31,770	39,229	126,253	10,424	45,407	35,440	15,678	694,340	1,556,900
大分県	36,303	4,751	1,462	49,297	89,645	1,955	7,870	23,015	81,964	14,741	5,363	11,770	34,312	19,265	25,449	80,547	4,007	28,682	21,918	17,907	469,476	1,029,700
宮崎県	51,106	3,435	0	45,019	60,924	2,435	6,407	21,459	86,053	9,401	7,251	15,367	31,086	19,160	26,944	80,755	8,437	28,870	23,553	11,043	429,096	967,800
鹿児島県	71,964	8,869	0	66,892	90,472	3,931	6,891	36,422	127,824	12,448	11,017	17,957	42,434	26,979	38,050	131,962	12,949	34,388	39,822	12,791	663,337	1,457,400
沖縄県	28,390	2,948	0	70,113	34,171	2,517	11,487	31,104	98,348	14,294	16,042	15,933	47,164	27,254	32,781	85,557	7,978	52,145	40,206	22,264	516,204	1,156,900
計	2,242,042	189,207	28,057	4,843,947	10,633,359	334,159	1,876,610	3,430,650	9,823,354	1,604,442	1,349,526	2,230,950	3,730,189	2,367,312	2,975,445	6,992,013	531,417	4,003,921	2,169,013	2,396,849	47,060,937	110,813,400

注：産業の空白は記入がないものであり、無業者を含むがすべてが無業者ではない。
（以下同様）

表１－５　年齢別・産業別 15 歳以上人口

	農業，林業	漁業	鉱業，採石業，砂利採取業	建設業	製造業	電気・ガス・熱供給・水道業	情報通信業	運輸業，郵便業	卸売業，小売業	金融業，保険業	不動産業，物品賃貸業	学術研究，専門・技術サービス業	宿泊業，飲食サービス業	生活関連サービス業，娯楽業	教育，学習支援業	医療，福祉	複合サービス事業	サービス業（他に分類されないもの）	公務（他に分類されるものを除く）	分類不能の産業		計
15～19歳	6,668	487	475	42,546	126,987	2,389	2,944	23,314	235,797	1,849	4,280	4,356	263,051	38,247	39,386	35,550	2,422	23,261	11,227	46,100	5,130,465	6,041,800
20～24歳	40,068	3,417	495	163,985	629,805	20,192	116,113	137,924	756,228	76,794	53,264	69,067	511,397	208,469	217,677	508,895	23,532	176,868	119,961	177,799	2,228,750	6,240,700
25～29歳	64,100	11,474	2,969	294,857	979,046	24,839	278,700	250,614	907,291	207,835	82,719	184,241	285,723	265,319	275,674	795,031	55,809	287,959	209,558	239,217	1,315,525	7,018,500
30～34歳	83,671	8,853	2,013	444,124	1,102,292	29,505	303,071	318,736	919,002	156,626	96,427	259,633	308,422	246,417	268,912	813,743	54,324	354,215	225,510	228,809	1,589,596	7,813,900
35～39歳	92,853	13,513	2,527	607,292	1,370,218	44,346	319,865	429,768	1,129,465	211,266	112,626	285,261	375,865	224,841	311,711	885,717	70,838	433,371	303,162	279,867	1,898,528	9,402,900
40～44歳	104,545	12,025	3,533	627,836	1,471,866	52,431	298,771	464,984	1,165,666	223,514	124,965	305,760	346,402	222,017	345,165	839,049	73,333	407,683	321,415	287,757	1,755,686	9,454,400
45～49歳	108,356	13,476	3,529	496,634	1,266,775	49,329	227,236	424,563	1,017,712	234,014	117,988	254,016	288,545	205,055	368,570	794,127	76,925	372,334	261,439	240,820	1,372,457	8,193,900
50～54歳	151,683	16,906	4,471	498,314	1,048,189	44,194	142,358	386,610	979,109	195,997	110,435	221,132	270,912	183,428	402,984	755,634	68,505	344,670	279,872	199,270	1,366,325	7,671,000
55～59歳	210,934	24,431	2,487	537,132	1,028,932	39,729	95,739	363,242	888,426	148,340	120,544	212,489	323,847	182,875	338,486	655,226	60,288	404,287	261,255	189,059	1,860,450	7,948,200
60～64歳	360,831	32,404	3,552	612,630	881,388	19,706	55,737	387,012	875,130	95,512	180,568	216,023	384,945	248,151	223,692	533,308	34,628	566,190	123,004	219,613	4,187,476	10,241,500
65～69歳	294,099	18,965	1,487	307,742	380,657	5,946	24,257	162,585	449,503	28,786	121,862	111,152	220,689	156,925	103,856	218,861	5,416	362,832	32,721	147,365	5,045,193	8,200,900
70～74歳	294,053	15,544	0	130,936	220,336	550	7,390	60,717	251,733	15,683	90,120	55,147	103,550	121,233	43,529	99,858	4,469	180,549	14,088	76,768	5,608,447	7,394,700
75歳以上	430,182	17,712	518	79,917	126,868	1,002	4,429	20,581	248,293	8,224	133,728	52,675	46,841	64,337	35,804	57,013	927	89,705	5,800	64,405	13,702,038	15,191,000
計	2,242,042	189,207	28,057	4,843,947	10,633,359	334,159	1,876,610	3,430,650	9,823,354	1,604,442	1,349,526	2,230,950	3,730,189	2,367,312	2,975,445	6,992,013	531,417	4,003,921	2,169,013	2,396,849	47,060,937	110,813,400

表１－６　産業別・男女別 15 歳以上人口

	男	女	計
農業，林業	1,349,756	892,286	2,242,042
漁業	139,911	49,296	189,207
鉱業，採石業，砂利採取業	22,617	5,440	28,057
建設業	4,058,220	785,727	4,843,947
製造業	7,423,412	3,209,947	10,633,359
電気・ガス・熱供給・水道業	280,127	54,032	334,159
情報通信業	1,387,160	489,450	1,876,610
運輸業，郵便業	2,764,938	665,712	3,430,650
卸売業，小売業	4,851,432	4,971,922	9,823,354
金融業，保険業	760,292	844,150	1,604,442
不動産業，物品賃貸業	803,986	545,539	1,349,526
学術研究，専門・技術サービス業	1,506,475	724,476	2,230,950
宿泊業，飲食サービス業	1,440,833	2,289,357	3,730,189
生活関連サービス業，娯楽業	963,406	1,403,906	2,367,312
教育，学習支援業	1,308,321	1,667,124	2,975,445
医療，福祉	1,695,037	5,296,977	6,992,013
複合サービス事業	320,011	211,406	531,417
サービス業（他に分類されないもの）	2,447,714	1,556,207	4,003,921
公務（他に分類されるものを除く）	1,567,870	601,142	2,169,013
分類不能の産業	1,307,985	1,088,865	2,396,849
	17,013,397	30,047,540	47,060,937
計	53,412,900	57,400,500	110,813,400

（基本属性ごとの就業状態）

次に、都道府県別、年齢別、男女別の基本的な属性と就業状態のクロス表を作成し、有業率を計算する。

表２－１　都道府県別・就業状態別 15 歳以上人口及び有業率

	有業者	無業者	計	有業率
北海道	2,627,900	2,175,700	4,803,600	54.7%
青森県	657,300	522,600	1,179,900	55.7%
岩手県	654,700	485,400	1,140,100	57.4%
宮城県	1,157,700	862,900	2,020,600	57.3%
秋田県	511,300	432,200	943,500	54.2%
山形県	583,800	420,700	1,004,500	58.1%
福島県	958,100	749,600	1,707,700	56.1%
茨城県	1,488,300	1,064,500	2,552,800	58.3%
栃木県	1,022,200	701,900	1,724,100	59.3%
群馬県	1,028,800	694,500	1,723,300	59.7%
埼玉県	3,713,600	2,554,600	6,268,200	59.2%
千葉県	3,178,400	2,220,900	5,399,300	58.9%
東京都	7,328,100	4,398,300	11,726,400	62.5%
神奈川県	4,682,900	3,198,500	7,881,400	59.4%
新潟県	1,187,300	864,800	2,052,100	57.9%
富山県	564,700	379,700	944,400	59.8%
石川県	603,400	401,600	1,005,000	60.0%
福井県	422,900	266,300	689,200	61.4%
山梨県	441,100	300,400	741,500	59.5%
長野県	1,108,700	733,900	1,842,600	60.2%
岐阜県	1,060,800	715,200	1,776,000	59.7%
静岡県	1,947,000	1,281,200	3,228,200	60.3%
愛知県	3,908,500	2,456,200	6,364,700	61.4%
三重県	939,000	652,000	1,591,000	59.0%
滋賀県	724,300	479,600	1,203,900	60.2%
京都府	1,293,600	995,900	2,289,500	56.5%
大阪府	4,310,100	3,390,300	7,700,400	56.0%
兵庫県	2,622,700	2,192,400	4,815,100	54.5%
奈良県	642,900	566,500	1,209,400	53.2%
和歌山県	470,300	392,600	862,900	54.5%
鳥取県	289,600	214,100	503,700	57.5%
島根県	356,600	257,700	614,300	58.0%
岡山県	948,600	725,300	1,673,900	56.7%
広島県	1,399,800	1,057,800	2,457,600	57.0%
山口県	684,400	563,800	1,248,200	54.8%
徳島県	369,300	310,600	679,900	54.3%
香川県	489,000	368,100	857,100	57.1%
愛媛県	678,700	554,100	1,232,800	55.1%
高知県	368,900	293,000	661,900	55.7%
福岡県	2,444,000	1,947,000	4,391,000	55.7%
佐賀県	424,400	296,500	720,900	58.9%
長崎県	661,900	550,200	1,212,100	54.6%
熊本県	879,200	677,700	1,556,900	56.5%
大分県	571,300	458,400	1,029,700	55.5%
宮崎県	553,400	414,400	967,800	57.2%
鹿児島県	808,500	648,900	1,457,400	55.5%
沖縄県	650,700	506,200	1,156,900	56.2%
計	64,418,700	46,394,700	110,813,400	58.1%

表２－２　年齢別・就業状態別 15 歳以上人口及び有業率

	有業者	無業者	計	有業率
15～19歳	938,900	5,102,900	6,041,800	15.5%
20～24歳	4,061,800	2,178,900	6,240,700	65.1%
25～29歳	5,754,900	1,263,600	7,018,500	82.0%
30～34歳	6,280,500	1,533,400	7,813,900	80.4%
35～39歳	7,565,100	1,837,800	9,402,900	80.5%
40～44歳	7,759,000	1,695,400	9,454,400	82.1%
45～49歳	6,881,100	1,312,800	8,193,900	84.0%
50～54歳	6,363,200	1,307,800	7,671,000	83.0%
55～59歳	6,140,900	1,807,300	7,948,200	77.3%
60～64歳	6,120,300	4,121,200	10,241,500	59.8%
65～69歳	3,201,300	4,999,600	8,200,900	39.0%
70～74歳	1,825,600	5,569,100	7,394,700	24.7%
75歳以上	1,526,100	13,664,900	15,191,000	10.0%
計	64,418,700	46,394,700	110,813,400	58.1%

表２－３　男女別・就業状態別 15 歳以上人口及び有業率

	有業者	無業者	計	有業率
男	36,743,300	16,669,600	53,412,900	68.8%
女	27,675,400	29,725,100	57,400,500	48.2%
計	64,418,700	46,394,700	110,813,400	58.1%

（シミュレーション）

以上の基本的な数値を確認した後に、インピュテーションの数値シミュレーションを行う。

坂下 (2020) では、12 の量的項目 に対しランダムに欠測値を発生させ、「シーケンシャル・ホット・デック法」、「シーケンシャル・ホット・デック法（シャッフル後）」、「全体でのランダム・ホット・デック法」、「分類内でのランダム・ホット・デック法」の 4 種のホット・デック法を順に適用し、結果を比較した。「シーケンシャル・ホット・デック法（シャッフル後）」とは、一般用ミクロデータでは類似したデータが順に並べられているため、単純に次のデータを利用するシーケンシャル・ホット・デック法は実データの場合より当てはまりが良くなり過ぎるので、データを全体の順番をシャッフルしてから適用するように改めたもので、結果は「全体でのランダム・ホット・デック法」と類似したものとなっている。また、欠測値の発生についてはその発生しやすさを考慮せずに、各項目に発生させたものであった。

これらの点を改善するため、今回は欠測値が発生しやすいと考えられる「就業状態」に検討対象を絞り、最初からシャッフルしたデータを用いて「シーケンシャル・ホット・デック法」と「ランダム・ホット・デック法」を区別しないものとする。

まず、「就業状態」について 0.1 の確率で欠測値を発生させ、ランダムなホット・デック法によるインピュテーションを行った。

表３－１　都道府県別・就業状態別 15 歳以上人口及び有業率 10%欠測のランダムなホット・デック法によるインピュテーション、真値からのずれ

	有業者	無業者	計	有業率	真値からのずれ		
					有業者	無業者	有業率
北海道	2,651,258	2,152,342	4,803,600	55.2%	0.9%	-1.1%	0.5
青森県	664,565	515,335	1,179,900	56.3%	1.1%	-1.4%	0.6
岩手県	651,740	488,360	1,140,100	57.2%	-0.5%	0.6%	-0.3
宮城県	1,162,858	857,742	2,020,600	57.6%	0.4%	-0.6%	0.3
秋田県	514,197	429,303	943,500	54.5%	0.6%	-0.7%	0.3
山形県	578,786	425,714	1,004,500	57.6%	-0.9%	1.2%	-0.5
福島県	961,719	745,981	1,707,700	56.3%	0.4%	-0.5%	0.2
茨城県	1,474,766	1,078,034	2,552,800	57.8%	-0.9%	1.3%	-0.5
栃木県	1,014,295	709,805	1,724,100	58.8%	-0.8%	1.1%	-0.5
群馬県	1,038,405	684,895	1,723,300	60.3%	0.9%	-1.4%	0.6
埼玉県	3,697,592	2,570,608	6,268,200	59.0%	-0.4%	0.6%	-0.3
千葉県	3,176,583	2,222,717	5,399,300	58.8%	-0.1%	0.1%	0.0
東京都	7,288,643	4,437,757	11,726,400	62.2%	-0.5%	0.9%	-0.3
神奈川県	4,676,672	3,204,728	7,881,400	59.3%	-0.1%	0.2%	-0.1
新潟県	1,186,719	865,381	2,052,100	57.8%	0.0%	0.1%	0.0
富山県	560,239	384,161	944,400	59.3%	-0.8%	1.2%	-0.5
石川県	601,374	403,626	1,005,000	59.8%	-0.3%	0.5%	-0.2
福井県	422,892	266,308	689,200	61.4%	0.0%	0.0%	0.0
山梨県	440,723	300,777	741,500	59.4%	-0.1%	0.1%	-0.1
長野県	1,105,577	737,023	1,842,600	60.0%	-0.3%	0.4%	-0.2
岐阜県	1,072,579	703,421	1,776,000	60.4%	1.1%	-1.6%	0.7
静岡県	1,934,417	1,293,783	3,228,200	59.9%	-0.6%	1.0%	-0.4
愛知県	3,894,165	2,470,535	6,364,700	61.2%	-0.4%	0.6%	-0.2
三重県	946,947	644,053	1,591,000	59.5%	0.8%	-1.2%	0.5
滋賀県	717,239	486,661	1,203,900	59.6%	-1.0%	1.5%	-0.6
京都府	1,305,186	984,314	2,289,500	57.0%	0.9%	-1.2%	0.5
大阪府	4,316,382	3,384,018	7,700,400	56.1%	0.1%	-0.2%	0.1
兵庫県	2,638,100	2,177,000	4,815,100	54.8%	0.6%	-0.7%	0.3
奈良県	637,366	572,034	1,209,400	52.7%	-0.9%	1.0%	-0.5
和歌山県	476,718	386,182	862,900	55.2%	1.4%	-1.6%	0.7
鳥取県	286,946	216,754	503,700	57.0%	-0.9%	1.2%	-0.5
島根県	356,554	257,746	614,300	58.0%	0.0%	0.0%	0.0
岡山県	946,629	727,271	1,673,900	56.6%	-0.2%	0.3%	-0.1
広島県	1,402,004	1,055,596	2,457,600	57.0%	0.2%	-0.2%	0.1
山口県	687,216	560,984	1,248,200	55.1%	0.4%	-0.5%	0.2
徳島県	374,260	305,640	679,900	55.0%	1.3%	-1.6%	0.7
香川県	488,198	368,902	857,100	57.0%	-0.2%	0.2%	-0.1
愛媛県	690,012	542,788	1,232,800	56.0%	1.7%	-2.0%	0.9
高知県	375,354	286,546	661,900	56.7%	1.7%	-2.2%	1.0
福岡県	2,444,284	1,946,716	4,391,000	55.7%	0.0%	0.0%	0.0
佐賀県	420,617	300,283	720,900	58.3%	-0.9%	1.3%	-0.5
長崎県	675,991	536,109	1,212,100	55.8%	2.1%	-2.6%	1.2
熊本県	886,622	670,278	1,556,900	56.9%	0.8%	-1.1%	0.5
大分県	568,987	460,713	1,029,700	55.3%	-0.4%	0.5%	-0.2
宮崎県	548,858	418,942	967,800	56.7%	-0.8%	1.1%	-0.5
鹿児島県	811,695	645,705	1,457,400	55.7%	0.4%	-0.5%	0.2
沖縄県	653,316	503,584	1,156,900	56.5%	0.4%	-0.5%	0.2
計	64,426,245	46,387,155	110,813,400	58.1%	0.0%	0.0%	0.0

注：真値からのずれは、有業者及び無業者はパーセント、有業率はポイント（以下同様）

表３－２　年齢別・就業状態別 15 歳以上人口及び有業率 10%欠測のランダムなホット・デック法によるインピュテーション、真値からのずれ

	有業者	無業者	計	有業率	真値からのずれ		
					有業者	無業者	有業率
15～19歳	1,173,415	4,868,385	6,041,800	19.4%	25.0%	-4.6%	3.9
20～24歳	4,035,964	2,204,736	6,240,700	64.7%	-0.6%	1.2%	-0.4
25～29歳	5,575,115	1,443,385	7,018,500	79.4%	-3.1%	14.2%	-2.6
30～34歳	6,072,752	1,741,148	7,813,900	77.7%	-3.3%	13.5%	-2.7
35～39歳	7,388,828	2,014,072	9,402,900	78.6%	-2.3%	9.6%	-1.9
40～44歳	7,521,932	1,932,468	9,454,400	79.6%	-3.1%	14.0%	-2.5
45～49歳	6,688,232	1,505,668	8,193,900	81.6%	-2.8%	14.7%	-2.4
50～54歳	6,168,721	1,502,279	7,671,000	80.4%	-3.1%	14.9%	-2.5
55～59歳	5,991,236	1,956,964	7,948,200	75.4%	-2.4%	8.3%	-1.9
60～64歳	6,117,444	4,124,056	10,241,500	59.7%	0.0%	0.1%	0.0
65～69歳	3,361,105	4,839,795	8,200,900	41.0%	5.0%	-3.2%	1.9
70～74歳	2,073,868	5,320,832	7,394,700	28.0%	13.6%	-4.5%	3.4
75歳以上	2,257,633	12,933,367	15,191,000	14.9%	47.9%	-5.4%	4.8
計	64,426,245	46,387,155	110,813,400	58.1%	0.0%	0.0%	0.0

表３－３　男女別・就業状態別 15 歳以上人口及び有業率 10%欠測のランダムなホット・デック法によるインピュテーション、真値からのずれ

	有業者	無業者	計	有業率	真値からのずれ		
					有業者	無業者	有業率
男	36,218,586	17,194,314	53,412,900	67.8%	-1.4%	3.1%	-1.0
女	28,207,659	29,192,841	57,400,500	49.1%	1.9%	-1.8%	0.9
計	64,426,245	46,387,155	110,813,400	58.1%	0.0%	0.0%	0.0

結果をみると、年齢別で乖離が大きくなっているが、これは年齢による有業率の差が大きく、年齢を考慮せずにドナーを選ぶと誤差が大きくなるためである。

より重要な問題は、インピュテーションによって他の項目との矛盾が発生することで、産業と就業状態のクロス表を作成すると次のようになる。

表３－４　産業別・就業状態別 15 歳以上人口元のデータ及び 10%欠測のランダムなホット・デック法によるインピュテーション

	元のデータ			有業率10%欠測後インピュテーション		
	有業者	無業者	計	有業者	無業者	計
農業，林業	2,242,042	0	2,242,042	2,145,616	96,426	2,242,042
漁業	189,207	0	189,207	181,686	7,521	189,207
鉱業，採石業，砂利採取業	28,057	0	28,057	25,532	2,525	28,057
建設業	4,843,947	0	4,843,947	4,654,088	189,859	4,843,947
製造業	10,633,359	0	10,633,359	10,189,387	443,972	10,633,359
電気・ガス・熱供給・水道業	334,159	0	334,159	318,067	16,092	334,159
情報通信業	1,876,610	0	1,876,610	1,788,822	87,788	1,876,610
運輸業，郵便業	3,430,650	0	3,430,650	3,280,457	150,193	3,430,650
卸売業，小売業	9,823,354	0	9,823,354	9,408,561	414,793	9,823,354
金融業，保険業	1,604,442	0	1,604,442	1,538,763	65,679	1,604,442
不動産業，物品賃貸業	1,349,526	0	1,349,526	1,291,956	57,570	1,349,526
学術研究，専門・技術サービス業	2,230,950	0	2,230,950	2,148,257	82,693	2,230,950
宿泊業，飲食サービス業	3,730,189	0	3,730,189	3,576,062	154,127	3,730,189
生活関連サービス業，娯楽業	2,367,312	0	2,367,312	2,266,523	100,790	2,367,312
教育，学習支援業	2,975,445	0	2,975,445	2,851,720	123,725	2,975,445
医療，福祉	6,992,013	0	6,992,013	6,704,814	287,199	6,992,013
複合サービス事業	531,417	0	531,417	514,793	16,623	531,417
サービス業（他に分類されないもの）	4,003,921	0	4,003,921	3,827,013	176,908	4,003,921
公務（他に分類されるものを除く）	2,169,013	0	2,169,013	2,081,074	87,938	2,169,013
分類不能の産業	2,396,849	0	2,396,849	2,292,688	104,162	2,396,849
	666,237	46,394,700	47,060,937	3,340,367	43,720,570	47,060,937
計	64,418,700	46,394,700	110,813,400	64,426,245	46,387,155	110,813,400

前述のように産業に記入がないものが無業者とは限らないが、記入があるものはすべて有業者なので矛盾が生じている。このため、産業に記入があるものは有業者に確定的インピュテーションを行ったのち全体でランダム・インピュテーションを行うように変更すると次のような結果となる。

表３－５　産業別・就業状態別 15 歳以上人口元のデータ及び 10%欠測の確定的インピュテーション及びランダムなホット・デック法によるインピュテーション

	有業者	無業者	計
農業，林業	2,242,042	0	2,242,042
漁業	189,207	0	189,207
鉱業，採石業，砂利採取業	28,057	0	28,057
建設業	4,843,947	0	4,843,947
製造業	10,633,359	0	10,633,359
電気・ガス・熱供給・水道業	334,159	0	334,159
情報通信業	1,876,610	0	1,876,610
運輸業，郵便業	3,430,650	0	3,430,650
卸売業，小売業	9,823,354	0	9,823,354
金融業，保険業	1,604,442	0	1,604,442
不動産業，物品賃貸業	1,349,526	0	1,349,526
学術研究，専門・技術サービス業	2,230,950	0	2,230,950
宿泊業，飲食サービス業	3,730,189	0	3,730,189
生活関連サービス業，娯楽業	2,367,312	0	2,367,312
教育，学習支援業	2,975,445	0	2,975,445
医療，福祉	6,992,013	0	6,992,013
複合サービス事業	531,417	0	531,417
サービス業（他に分類されないもの）	4,003,921	0	4,003,921
公務（他に分類されるものを除く）	2,169,013	0	2,169,013
分類不能の産業	2,396,849	0	2,396,849
	3,452,357	43,608,580	47,060,937
計	67,204,820	43,608,580	110,813,400

データの矛盾はなくなるが、有業者が不自然に多くなり、無業者が少なくなっている。年齢別の結果は次のようになる。

表３－６　年齢別・就業状態別 15 歳以上人口及び有業率 10%欠測の確定的インピュテーション及びランダムなホット・デック法によるインピュテーション、真値からのずれ

	有業者	無業者	計	有業率	真値からのずれ		
					有業者	無業者	有業率
15～19歳	1,226,228	4,815,572	6,041,800	20.3%	30.6%	-5.6%	4.8
20～24歳	4,196,670	2,044,030	6,240,700	67.2%	3.3%	-6.2%	2.2
25～29歳	5,836,077	1,182,423	7,018,500	83.2%	1.4%	-6.4%	1.2
30～34歳	6,362,755	1,451,145	7,813,900	81.4%	1.3%	-5.4%	1.1
35～39歳	7,670,070	1,732,830	9,402,900	81.6%	1.4%	-5.7%	1.1
40～44歳	7,863,103	1,591,297	9,454,400	83.2%	1.3%	-6.1%	1.1
45～49歳	6,954,028	1,239,872	8,193,900	84.9%	1.1%	-5.6%	0.9
50～54歳	6,452,087	1,218,913	7,671,000	84.1%	1.4%	-6.8%	1.2
55～59歳	6,246,618	1,701,582	7,948,200	78.6%	1.7%	-5.8%	1.3
60～64歳	6,364,906	3,876,594	10,241,500	62.1%	4.0%	-5.9%	2.4
65～69歳	3,514,439	4,686,461	8,200,900	42.9%	9.8%	-6.3%	3.8
70～74歳	2,160,516	5,234,184	7,394,700	29.2%	18.3%	-6.0%	4.5
75歳以上	2,357,323	12,833,677	15,191,000	15.5%	54.5%	-6.1%	5.5
計	67,204,820	43,608,580	110,813,400	60.6%	4.3%	-6.0%	2.5

各年齢で有業率にプラス方向へのバイアスが生じている。このような結果になるのは、ホット・デック法によるインピュテーションを行ったのは、産業に記入がないものであるにも関わらず、ドナーを全体から選んだためである。バイアスを避けるためには、データ全体を産業に記入があるもの（就業状態が欠測していても確定的インピュテーションが可能なもの）と記入がないもの（確定的インピュテーションが不可能なもの）に分け、後者の中でドナーを選ぶ必要がある。

また、産業に記入がないものであっても、雇用者であるもの、正規又は非正規就業者であるものはすべて有業者であり、就業希望、求職に記入があるものはすべて無業者であるので、これらも確定的インピュテーションの対象として、まとめてホット・デック法の対象及びドナー・プールから外した上でシミュレーションを行うと次のようになる。

表４－１　都道府県別・就業状態別 15 歳以上人口及び有業率 10%欠測の確定的インピュテーションの後、ランダムなホット・デック法によるインピュテーション、真値からのずれ

	有業者	無業者	計	有業率	真値からのずれ		
					有業者	無業者	有業率
北海道	2,627,909	2,175,691	4,803,600	54.7%	0.0%	0.0%	0.0
青森県	656,883	523,017	1,179,900	55.7%	-0.1%	0.1%	0.0
岩手県	654,211	485,889	1,140,100	57.4%	-0.1%	0.1%	0.0
宮城県	1,157,702	862,897	2,020,600	57.3%	0.0%	0.0%	0.0
秋田県	511,300	432,200	943,500	54.2%	0.0%	0.0%	0.0
山形県	583,800	420,700	1,004,500	58.1%	0.0%	0.0%	0.0
福島県	958,100	749,600	1,707,700	56.1%	0.0%	0.0%	0.0
茨城県	1,487,798	1,065,002	2,552,800	58.3%	0.0%	0.0%	0.0
栃木県	1,020,709	703,391	1,724,100	59.2%	-0.1%	0.2%	-0.1
群馬県	1,028,800	694,500	1,723,300	59.7%	0.0%	0.0%	0.0
埼玉県	3,713,090	2,555,110	6,268,200	59.2%	0.0%	0.0%	0.0
千葉県	3,179,405	2,219,895	5,399,300	58.9%	0.0%	0.0%	0.0
東京都	7,330,554	4,395,846	11,726,400	62.5%	0.0%	-0.1%	0.0
神奈川県	4,682,901	3,198,499	7,881,400	59.4%	0.0%	0.0%	0.0
新潟県	1,187,300	864,800	2,052,100	57.9%	0.0%	0.0%	0.0
富山県	564,219	380,181	944,400	59.7%	-0.1%	0.1%	-0.1
石川県	603,400	401,600	1,005,000	60.0%	0.0%	0.0%	0.0
福井県	423,300	265,900	689,200	61.4%	0.1%	-0.2%	0.1
山梨県	441,592	299,908	741,500	59.6%	0.1%	-0.2%	0.1
長野県	1,108,700	733,900	1,842,600	60.2%	0.0%	0.0%	0.0
岐阜県	1,060,800	715,200	1,776,000	59.7%	0.0%	0.0%	0.0
静岡県	1,947,000	1,281,200	3,228,200	60.3%	0.0%	0.0%	0.0
愛知県	3,908,994	2,455,706	6,364,700	61.4%	0.0%	0.0%	0.0
三重県	938,521	652,479	1,591,000	59.0%	-0.1%	0.1%	0.0
滋賀県	724,794	479,106	1,203,900	60.2%	0.1%	-0.1%	0.0
京都府	1,294,565	994,935	2,289,500	56.5%	0.1%	-0.1%	0.0
大阪府	4,311,103	3,389,297	7,700,400	56.0%	0.0%	0.0%	0.0
兵庫県	2,623,203	2,191,897	4,815,100	54.5%	0.0%	0.0%	0.0
奈良県	643,357	566,043	1,209,400	53.2%	0.1%	-0.1%	0.0
和歌山県	470,300	392,600	862,900	54.5%	0.0%	0.0%	0.0
鳥取県	289,073	214,627	503,700	57.4%	-0.2%	0.2%	-0.1
島根県	356,600	257,700	614,300	58.0%	0.0%	0.0%	0.0
岡山県	949,076	724,824	1,673,900	56.7%	0.1%	-0.1%	0.0
広島県	1,400,295	1,057,305	2,457,600	57.0%	0.0%	0.0%	0.0
山口県	683,916	564,284	1,248,200	54.8%	-0.1%	0.1%	0.0
徳島県	370,269	309,631	679,900	54.5%	0.3%	-0.3%	0.1
香川県	488,533	368,567	857,100	57.0%	-0.1%	0.1%	-0.1
愛媛県	679,686	553,114	1,232,800	55.1%	0.1%	-0.2%	0.1
高知県	369,384	292,516	661,900	55.8%	0.1%	-0.2%	0.1
福岡県	2,443,498	1,947,502	4,391,000	55.6%	0.0%	0.0%	0.0
佐賀県	423,886	297,014	720,900	58.8%	-0.1%	0.2%	-0.1
長崎県	661,408	550,692	1,212,100	54.6%	-0.1%	0.1%	0.0
熊本県	879,707	677,193	1,556,900	56.5%	0.1%	-0.1%	0.0
大分県	571,300	458,400	1,029,700	55.5%	0.0%	0.0%	0.0
宮崎県	552,898	414,902	967,800	57.1%	-0.1%	0.1%	-0.1
鹿児島県	808,004	649,396	1,457,400	55.4%	-0.1%	0.1%	0.0
沖縄県	650,203	506,697	1,156,900	56.2%	-0.1%	0.1%	0.0
計	64,422,045	46,391,355	110,813,400	58.1%	0.0%	0.0%	0.0

表４－２　年齢別・就業状態別 15 歳以上人口及び有業率 10%欠測の確定的インピュテーションの後、ランダムなホット・デック法によるインピュテーション、真値からのずれ

	有業者	無業者	計	有業率	真値からのずれ		
					有業者	無業者	有業率
15～19歳	938,900	5,102,900	6,041,800	15.5%	0.0%	0.0%	0.0
20～24歳	4,061,800	2,178,900	6,240,700	65.1%	0.0%	0.0%	0.0
25～29歳	5,755,325	1,263,175	7,018,500	82.0%	0.0%	0.0%	0.0
30～34歳	6,281,495	1,532,405	7,813,900	80.4%	0.0%	-0.1%	0.0
35～39歳	7,563,570	1,839,330	9,402,900	80.4%	0.0%	0.1%	0.0
40～44歳	7,759,857	1,694,543	9,454,400	82.1%	0.0%	-0.1%	0.0
45～49歳	6,880,103	1,313,797	8,193,900	84.0%	0.0%	0.1%	0.0
50～54歳	6,364,201	1,306,799	7,671,000	83.0%	0.0%	-0.1%	0.0
55～59歳	6,138,441	1,809,759	7,948,200	77.2%	0.0%	0.1%	0.0
60～64歳	6,120,809	4,120,691	10,241,500	59.8%	0.0%	0.0%	0.0
65～69歳	3,203,747	4,997,153	8,200,900	39.1%	0.1%	0.0%	0.0
70～74歳	1,825,602	5,569,098	7,394,700	24.7%	0.0%	0.0%	0.0
75歳以上	1,528,197	13,662,803	15,191,000	10.1%	0.1%	0.0%	0.0
計	64,422,045	46,391,355	110,813,400	58.1%	0.0%	0.0%	0.0

表４－３　男女別・就業状態別 15 歳以上人口及び有業率 10%欠測の確定的インピュテーションの後、ランダムなホット・デック法によるインピュテーション、真値からのずれ

	有業者	無業者	計	有業率	真値からのずれ		
					有業者	無業者	有業率
男	36,743,118	16,669,782	53,412,900	68.8%	0.0%	0.0%	0.0
女	27,678,927	29,721,573	57,400,500	48.2%	0.0%	0.0%	0.0
計	64,422,045	46,391,355	110,813,400	58.1%	0.0%	0.0%	0.0

このように、ほぼ元データと同じ結果となっている。これは、今回欠測値を発生させたのが就業状態のみで、それ以外の項目は元データを残したので、多くの欠測データで確定的インピュテーションが可能だったためである。

以上の結果から分かるのは、ホット・デック法で良い結果を得るには、ドナーとなるデータを適切に選ぶ（ドナー・プールを適切に作成する）ことは当然ながら、特に確定的インピュテーションのような他の手法と併用する場合は、ホット・デック法に用いるデータの選定を適切に行う必要があることである。

４．まとめ

今年度の調査対象となった文献から、アメリカ合衆国では、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研究されてきた行政情報の利用が本格化したこと、その他の統計調査でもインピュテーションにかかわる行政記録やビッグデータの利用が検討されていること、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。

また、欧州などでは、データ・エディティングとインピュテーションのシステムの開発や改良、特に機械学習の適用の検討が進んでいることが分かった。

一般用ミクロデータを用いた質的変数を対象とするホット・デック法のシミュレーションでは、ドナーとなるデータを適切に選ぶこと、特に確定的インピュテーションのような他の手法と併用する場合は、ホット・デック法に用いるデータの選定を適切に行うことが重要であるという知見を得た。

参考文献

[1] 坂下信之 (2017) 「諸外国の公的統計における欠測値補完（インピュテーション）の現状～文献調査～」、リサーチペーパー第 40 号、総務省統計研究研修所。

[2] 坂下信之 (2018)「諸外国における統計調査の欠測値補完方法の動向と手法の体系について」、リサーチペーパー第 43 号、総務省統計研究研修所。

[3] 坂下信之 (2019)「統計調査の欠測値補完方法に関する基本的文献と諸外国の動向について」、リサーチペーパー第 44 号、総務省統計研究研修所。

[4] 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。

[5] 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。

[6] Barragán, S. and Salgado, D. (2022), “Improving statistical data editing with Machine Learning: first use cases in Statistics Spain (INE)”, Expert Meeting on Statistical Data Editing, Conference of European Statisticians, United Nations Economic Commission for Europe, October, 2022.

[7] Bianchi, B. (2022), “Application of the MissForest algorithm for imputation in the Survey on Income and Living Conditions”, Expert Meeting on Statistical Data Editing, Conference of European Statisticians, United Nations Economic Commission for Europe, October, 2022.

[8] Buglielli, M. T., Filippini, R., and Rosati, S. (2022), “The SCIA system implementing the Fellegi and Holt methodology compared to the recent R packages”, Expert Meeting on Statistical Data Editing, Conference of European Statisticians, United Nations Economic Commission for Europe, October, 2022.

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

Statistics Administration, U.S. CENSUS BUREAU.

[10] De Fausti, F., Di Zio, M., Filippini, R., Toti, S., and Zardetto, D. (2022a). Multilayer perceptron models for the estimation of the Attained level of Education in the Italian Permanent Census. Statistical Journal of the IAOS,38, pp. 637–646.

[11] De Fausti, F., Di Zio, M., Filippini, R., Toti, S., and Zardetto, D. (2022b), "The imputation of the 'Attained Level of Education' in the base register of individuals through Neural Networks using sampling weights.", Expert Meeting on Statistical Data Editing, Conference of European Statisticians, United Nations Economic Commission for Europe, October, 2022.

[12] Ditscheid, J. (2022), "Experimental Short-Term Statistics based on Data Imputation Methods", Expert Meeting on Statistical Data Editing, Conference of European Statisticians, United Nations Economic Commission for Europe, October, 2022.

[13] Di Zio, M., Filippini R., and Toti S. (2022), "Variance estimation for the mass imputation of the "Attained level of education" in the Italian Base Register of individuals: A comparison between analytical and MonteCarlo estimates", Expert Meeting on Statistical Data Editing, Conference of European Statisticians, United Nations Economic Commission for Europe, October, 2022.

[14] Edward, M. (2022), "Machine Learning Imputation for Social Surveys: Random Forest imputation of ONS' Household Financial Survey", Expert Meeting on Statistical Data Editing, Conference of European Statisticians, United Nations Economic Commission for Europe, October, 2022.

[15] Ester, M., Kriegel, H. P., Sander, J., and Xu, X. (1996), "A density-based algorithm for discovering clusters in large spatial databases with noise", Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96).

[16] Gray, D. (2022), "Banff's next step: an open-source data editing system for advanced tools and collaboration", Expert Meeting on Statistical Data Editing, Conference of European Statisticians, United Nations Economic Commission for Europe, October, 2022.

[17] Guin, A., Roy, A., and Sinha, B. (2021), "Bayesian Analysis of Singly Imputed Partially Synthetic Data Generated by Plug-in Sampling and Posterior Predictive Sampling Under the Multiple Linear Regression Model.", RESEARCH REPORT SERIES (Statistics #2021-02), CSRM, US Bureau of the Census, Washington, DC.

[18] Guin, A., Roy, A., and Sinha, B. (2022), "Bayesian Analysis of Multiply Imputed Synthetic Data Under the Multiple Linear Regression Model", RESEARCH REPORT SERIES (Statistics #2022-02), CSRM, US Bureau of the Census, Washington, DC.

[19] Hainmueller, J. (2012), "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies" Political Analysis 20(1), 25-46.

[20] Hawala, S. (2008), "Producing partially synthetic data to avoid disclosure", Proceedings of the Joint Statistical Meetings, American Statistical Association, 1345-1350.

[21] Hernandez, J. L. M. (2022), “The use of administrative records for the imputation of data from the Economic Censuses of Mexico”, Expert Meeting on Statistical Data Editing, Conference of European Statisticians, United Nations Economic Commission for Europe, October, 2022.

[22] Hidiroglou, M., and Berthelot, J. (1986), “Statistical editing and imputation for periodic business surveys”, Survey Methodology, 73-83.

[23] Klein, M. and Sinha, B. (2016), “Likelihood based finite sample inference for singly imputed synthetic data under the multivariate normal and multiple linear regression models”, Journal of Privacy and Confidentiality 7 (1), 43-98.

[24] Klein, M., Zylstra, J., and Sinha, B. (2018), “Finite Sample Inference for Multiply Imputed Synthetic Data under a Multiple Linear Regression Model”, Calcutta Statistical Association Bulletin.

[25] Klein, M., Moura, R., Sinha, B. (2019), "Multivariate Normal Inference based on Singly Imputed Synthetic Data under Plug-in Sampling", RESEARCH REPORT SERIES (Statistics #2019-06), CSRM, US Bureau of the Census, Washington, DC.

[26] Leuenberger, M. (2022), “Application of the "SwissCheese" algorithm for the imputation of partial non-response in the Survey on Income and Living Conditions”, Expert Meeting on Statistical Data Editing, Conference of European Statisticians, United Nations Economic Commission for Europe, October, 2022.

[27] Lipke, M., Miller, D., Wagner, V., Brown, K., and Agnihotri, V. (2022), “Growing a Modern Edit and Imputation System”, Expert Meeting on Statistical Data Editing, Conference of European Statisticians, United Nations Economic Commission for Europe, October, 2022.

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

[29] Molfino, E. (2021a), “Imputing Lot Size with Property Tax Data”, U.S. Department of Housing and Urban Development Office of Policy Development and Research, Washington, DC 20410-6000.

[30] Molfino, E. (2021b), “Imputing Year Built with Property Tax Data”, U.S. Department of Housing and Urban Development Office of Policy Development and Research, Washington, DC 20410-6000.

[31] Moura, R., Klein, M., Coelho, C. A., Sinha, B. (2017), “Inference for Multivariate Regression Model based on Synthetic Data generated under Fixed-Posterior Predictive Sampling: Comparison with Plug-in Sampling”, REVSTAT – Statistical Journal, 15 (2), April 2017, 155-186

[32] Moura, R., Klein, M., Coelho, C. A., Zylstra, J., Sinha, B. (2018), "Inference for Multivariate Regression Model based on Synthetic Data generated using Plug-in Sampling", RESEARCH REPORT SERIES (Statistics #2018-02), CSRM, US Bureau of the Census, Washington, DC.

[33] Mulry, M. H., Mule, T., Keller, A. K., and Konicki, S. (2021), “Administrative Records Modeling

in the 2020 Census." 2020 CENSUS PROGRAM MEMORANDUM SERIES: 2021.10.

[34] Murawski, P. (2022), "Data imputation for the purposes of statistical research with the use data from administrative registers", Expert Meeting on Statistical Data Editing, Conference of European Statisticians, United Nations Economic Commission for Europe, October, 2022.

[35] Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003), "Multiple Imputation for Statistical Disclosure Limitation", Journal of Official Statistics, 19-1, 1-16.

[36] Reiter, J. P. (2003), "Inference for partially synthetic, public use microdata sets", Survey Methodology 29, 181–189.

[37] Reiter, J. P. and Kinney, S. K. (2012), "Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary", Journal of Official Statistics 28, 583-590.

[38] Rosati, S., Buglielli, M. T., and Tosco, L. (2022), "Multiple software systems for the editing and imputation process of the 7th General Census of Agriculture", Expert Meeting on Statistical Data Editing, Conference of European Statisticians, United Nations Economic Commission for Europe, October, 2022.

[39] 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.

[40] Rubin, D. B. (1987), "Multiple Imputation for Nonresponse in Surveys". John Wiley & Sons, New York.

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

[42] Saliba, J. (2022), "Robust regression, MissForest and calibration combined with non-linear optimization with constraints to impute VAT turnover", Expert Meeting on Statistical Data Editing, Conference of European Statisticians, United Nations Economic Commission for Europe, October, 2022.

[43] Scholtus, S., De Jong, W., Vaasen-Otten, A., and Aelen, F. (2022), "Towards a new integrated uniform production system for business statistics at Statistics Netherlands: automatic data editing with multiple data sources", Expert Meeting on Statistical Data Editing, Conference of European Statisticians, United Nations Economic Commission for Europe, October, 2022.

[44] Ten Bosch, O., De Jonge, E., Van der Loo, M. (2022), "Discover the hidden validation rules in your data with 'validatesuggest'", Expert Meeting on Statistical Data Editing, Conference of European Statisticians, United Nations Economic Commission for Europe, October, 2022.

[45] Thurow, M., Dumpert, F., Ramosaj, B., Pauly, M. (2022), "Goodness (of fit) of Imputation Methods", Expert Meeting on Statistical Data Editing, Conference of European Statisticians,

United Nations Economic Commission for Europe, October, 2022.

[46] Vaasen-Otten, A., Aelen, F., Scholtus, S., and De Jong, W. (2022), “Towards a new integrated uniform production system for business statistics at Statistics Netherlands: quality indicators to guide top-down analysis”, Expert Meeting on Statistical Data Editing, Conference of European Statisticians, United Nations Economic Commission for Europe, October, 2022

[47] Van der Loo, M., De Jonge, E., Ten Bosch, O. (2022), “Rule Management”, Expert Meeting on Statistical Data Editing, Conference of European Statisticians, United Nations Economic Commission for Europe, October, 2022.

[48] Vasquez, B., Quintana, O., Larrain, J. (2022), “Automatic selective editing approach using machine learning: an application to VAT data”, Expert Meeting on Statistical Data Editing, Conference of European Statisticians, United Nations Economic Commission for Europe, October, 2022.

[49] Vielma Orozco, E. (2022), “Automatic Data Editing and Imputation. Experience in the 2020 Mexican Census”, Expert Meeting on Statistical Data Editing, Conference of European Statisticians, United Nations Economic Commission for Europe, October, 2022.