リサーチペーパー第 51 号　　Research Paper No.51

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

Current Trends of Research in 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 3 年 9 月
September 2021

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３年９月 15 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。

本稿では、アメリカ合衆国及び欧州各国の近年の動向を把握するとともに、アメリカ合衆国については人口センサスに関して継続して行われているプロジェクトの状況を調査した。

その結果、調査対象に接触できないことの多さが問題となっているアメリカ合衆国の人口センサスでは、行政記録の利用が調査対象への接触の判断とデータ補完の両面で課題となっており、その際、インピュテーションが実査時点の課題にもなっていること、カナダや欧州でも、コロナ禍などの影響で研究発表は低調になっているものの、インピュテーションに関する行政記録の利用、機械学習の適用、システムの構築・共有などさまざまなプロジェクトが進められていることが分かった。

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、人口センサス

Current Trends of Imputation Methods for Statistical Surveys 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 investigated the recent trends in the United States and European countries, and for the United States, we studied the status of the projects which are continuously taking place on population census.

As a result, we found that, in the United States where the amount of the case unable to contact census targets is a problem, use of administrative records is becoming a issue, both in terms of contact strategy and supplement of the data for uncontactable targets, and imputation is an issue at the time of field enumeration. In Canada and European Countries, although research outcome have been scarce due to the effects of the coronavirus pandemic, various projects are ongoing in the area such as the use of administrative records on imputation, the application of machine learning, and the development and sharing of systems.

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。

そのため、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。

本年は、昨年に続きコロナ禍などの影響で新たな情報は少なかったが、米国センサス局において継続して行われているプロジェクトや欧州を中心とする国際会議の情報を収集した。

以下はその結果であり、その構成は、１．がアメリカ合衆国及びカナダの動向、２．が欧州の動向、３．がまとめとなっている。

１．アメリカ合衆国及びカナダの動向

（米国の人口センサス関連プロジェクト）

米国の人口センサスでは、実地調査で接触できなかった住居を対象に「無回答のフォローアップ」(Nonresponse Follow-Up, NRFU) を行っており[1]、関連して、センサス局に属する The Center for Statistical Research & Methodology (CSRM) においていくつかのプロジェクトが進行している。そのうち、インピュテーションに関わるものについては、今までも各年の状況について部分的に紹介しているが[2]、今回はその実施状況を過去に遡って調査した。

概観的な文献としては、統計方法論に関する連邦委員会 (Federal Committee on Statistical Methodology, FCSM) 研究会議で 2020 年センサスの計画と行政記録の利用について説明した Vitrano and Chapin (2012)、国際公的統計協会 (IAOS) で報告された Keller (2016)、行政記録モデルチームからセンサス科学諮問委員会 (Census Scientific Advisory Committee) に提出された Administrative Records Modeling Team (2017) などがある。Vitrano and Chapin (2012) では、2010 年人口センサスのコスト要因を分析し、調査段階での対象者の参加意識が低いため NRFU の現場スタッフが大量に必要であること、回答を得る戦略に適切な時期についての要件がないこと、調査前の住所フレームのアップデートが膨大であることがコストを押し上げており、コストと品質の目標を達成するには人口センサスの設計、企画と実施に根本的な変更を加える必要があるとして、住所フレームの改善、回答を得る戦略の改善、行政記

[1] 米国の人口センサスは郵送調査を基本としており、回収されなかった世帯を対象に NRFU が行われる。以下に記すようにこの活動が人口センサス最大のコスト要因であるため、空家及び接触困難世帯の見極めと、接触できない場合の補完方法が重要な課題となっている。

[2] 坂下 (2018)、坂下 (2019)。

録利用可能性の評価、センサスのカバレッジ評価の改善などの行政記録関連プロジェクトを列挙している。Keller (2016) は、2010 年人口センサスでは、居住の状態と世帯人員についてインピュテーションが行われた割合が低く、NRFU の業務が膨大であったため、センサス局では 2020 年センサスにおいて、(1) 無回答世帯における行政記録の利用の可能性、(2) 無回答世帯への訪問回数の削減 について調査しているとしている。Administrative Records Modeling Team (2017) は、2010 年センサスの NRFU 活動の内容と 2015 年試験調査、2018 年の最終 (end-to-end) 試験、2020 年センサスでの方針について述べている。NRFU で接触できない住戸は、「居住」、「空き家」、「不存在または削除」(住所地に実際には家が存在しない) に分類され、「居住」と判断した住戸に再接触を試みるか、インピュテーションが行われるが、その判定及びインピュテーションに用いる行政記録やその利用方法、品質評価についての解説が主な内容である。なお、Deaver (2020) が、2018 年最終試験の結果を受けてまとめられた、NRFU を含む人口センサス全体において用いる行政記録とその利用方法の 2020 年 5 月時点での解説となっている。

2012 会計年度に開始され、現在も続いている「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、品質を維持しつつ人口センサスのコストを大幅に削減するために、無回答のフォローアップの企画、準備、実施において行政記録を利用する方法を研究している。Vitrano and Chapin (2012) によると、このプロジェクトの焦点は、行政記録を利用して NRFU の活動を支援することによるコストと品質への影響を調べることで、NRFU 活動における接触の試みを減らすために行政記録を用いることも含まれる[3]。プロジェクトでは、2013 年度までにフォローアップ困難 (Hard To Followup, HTF) 指標の変量の候補の検討[4]、2014 年度には 2015 年センサス試験調査を行うアリゾナ州マリコパ郡の NRFU データを用いた傾向 (propensity) モデルの検討を行い、2016 年度は、NRFU で付された「記載住所に配達不能」(Undeliverable As Addressed, UAA) フラグの数と傾向モデルの検討で良い共変量とされた内国歳入庁 (Internal Revenue Service, IRS) の個人所得税申告 (IRS1040) の世帯人員数を用いて、行政記録を用いた世帯人員数の予測は条件付きで有望との知見を得た[5]。2017 年度以降は、IRS1040 データの追加、アメリカ先住民や低所得者のデータの利用などの検討を行っている。

2014 年度から 2017 年度にかけて行われた「NRFU の聞き取り回答と行政記録を比較するための 2010 年センサスのカバレッジ評価データの活用 (Using 2010 Census Coverage

[3] 2010 年センサスでは最大 6 回の接触を試みたが、NRFU 改善のためのプロジェクトでは接触困難世帯を行政記録からモデルにより予測することで効率化を図り、困難世帯の欠側データへのインピュテーションにも行政記録の利用することを検討している。
[4] 人口密度の対数が最も重要な説明変数で、世帯の大きさも重要との結果を得ている。
[5] このプロジェクトでは行政記録をそのまま代替 (substitution) データとするのではなく、インピュテーションの一つの手がかりとして用いることを検討している。

Measurement Data to Compare Nonresponse Follow-up Proxy Responses with Administrative Records)」プロジェクトは、無回答世帯への接触方法を見直すとともに、行政記録を利用することで聞取り調査 (proxy response) より正確な結果を得られるかどうかを検討した。このプロジェクトでは、2014 年度に 2010 年センサスカバレッジ評価データ (CCM) の結果を用いて NRFU の回答の質を、利用可能な 2010 年すべての内国歳入庁及びメディケアの記録と比較するための方法論を開発し、2015 年度には行政記録と無回答フォローアップを比較して行政記録の品質を評価して合同統計会議 (Joint Statistical Meetings, JSM) で報告している（Mulry and Keller (2015)）。2016 年度の研究では、データが内国歳入庁やメディケアから来ているにもかかわらず、聞取り調査の計数が正しい値となる割合が行政記録よりも高いこと、評価できない情報の割合は行政記録の方が高いことが結論となっており、すべての行政記録を用いる計画から、最良の行政記録をモデルによって特定することに変更する必要性が指摘されている（CSRM(2016)、Mulry and Keller (2017)、坂下(2018)）。

2013 年から行われている「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、接触を試み続ける世帯と、行政記録を利用する世帯を見分けるために、「良い」行政記録を特定することを研究している。プロジェクトではまず分類方法のプロトタイプを開発してセンサスへの有用性に応じて行政記録を分類し、合同統計会議 (Joint Statistical Meetings, JSM)[6] で Morris (2014) を発表した。そこではロジスティック回帰、分類木、ランダムフォレストの 3 手法が比較されており、ロジスティック回帰とランダムフォレストはほぼ同じ予測力を示し、分類木は少し劣るが解釈が容易とされている。さらに試験調査の行われる 2015 年度にはマリコパ郡の 2010 年データの分析を中心に試験調査の戦略を決定するための研究を行い、線形計画法により空き家を見分けて NRFU を効率化する手法を JSM に報告するとともに査読誌に論文を投稿し、この論文は翌年 Morris et al. (2016) として刊行された[7]。2016 年度には世帯員の居住状況を判断するため行政記録を用いるモデルを論じた Morris (2017)[8] を作成し（翌年改訂・刊行）、行政記録の最適な選定のためにベイジアンモデルを用いることを論じた Thibaudeau and Morris (2016) を JSM で報告した。ここでは、調査世帯による自己回答が望ましいとしつつも、調査費用との兼ね合いで必ずしも成功するまで接触を試み続けることが最適とは言えないとして、損失関数を用いて行政記録の利用を最適化するモデルを構築している。2017 年度からは Keller et al. (2018) [9] を Journal of Official Statistics に投稿するとともに、民間企業の不動産データを用いた住宅の空室率と物件の状況の関係の評価などに着手し、2019 年度以降は実査と行政記録の乖離

[6] アメリカ統計学会 (American Statistical Association, ASA) 等が集まって毎年開催される統計家の会議。坂下 (2017) 参照。
[7] 坂下 (2018)。
[8] 坂下 (2018)。
[9] 坂下 (2019)。

についての洞察を得るための試験調査による住戸の状況の潜在クラス分析、NRFU が遅れた時や所得税申告期限の延長の緊急対応計画の探索的分析、先住民居留地での居住している住戸の特定と計上、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改造を行っている。

これらが行政記録の利用に焦点を当てているのに対し、2019 年度に開始された「二言語研修の効果の実験 (Experiment for Effectiveness of Bilingual Training)」プロジェクトは、特に英語を話さない世帯への再訪問の効率性の測定方法を検討している。そこでは、調査員の 2020 年調査から導入された世帯とコミュニケーションを取るための教材による研修受講の有無、調査世帯への訪問回数による回答率の違いを継続比率ロジット (CRL) モデルにより分析するために必要な標本の大きさを考察し、シミュレーションを行い、研究レポート Raim et al. (2020) を刊行している。

（米国のその他の動向）

坂下 (2017) に記したように、センサス局の The Center for Statistical Research & Methodology (CSRM) では、小売統計で地域レベルの推計や月次の売上推計の改善のため商用のビッグデータを利用する調査研究を行っている[10]。最近の検討内容について記述した Hutchinson (2019) では、商用ビッグデータを用いて、インピュテーションを含む月次小売業調査（Monthly Retail Trade Survey, MRTS）の検証を行っており、CSRM (2020) によると、2020 年度には、事業所レベルのデータにより州別、北米産業分類システム (North American Industry Classification System, NAICS) コード 3 桁の月次データを作成するためのインピュテーションモデル及び州レベルの売上高を推計する階層的ベイズインピュテーションモデルの研究を行い、調査データ、行政データ、第三者のデータから合成とインピュテーションによって作成され、MRTS よりも地理的に詳細な早期の推定値を得る試験的なデータを開発した。

坂下 (2020) に記した「経済センサスのベイジアンによるインピュテーションと合成データ」の研究について、CSRM (2020) は、エディティングと多重代入法をデータ合成と統合し、一般に共有するのに適した完全な合成データのジェネレータを実装し、研究報告書を作成したと記している[11]。この報告書は、Thompson et al. (2020) のことと思われるが、そこではエディティングとインピュテーションの比較的詳細な説明を含む米国の経済センサス、完全な合成データ及び部分的な合成データ[12]の作成手法について解説している。

---

[10] ただし、2017 年時点ではインピュテーションについては言及なし（坂下 (2017)）。

[11] 2019 年度までは経済センサスの「エディティング及びインピュテーション」と「合成データの開発」は別項目の研究となっており、坂下 (2020) に記したように前者の知見を後者に応用した形だったが、2020 年度は両者が統合され、合成データをインピュテーションの一手法と位置づける方向にあるとみられる。

[12] 完全な合成データは、完全にモデルによって作成された合成値であり、元のデータと対応しないのに対し、部分的な合成データは元データとの対応を持つ。

同様に坂下 (2020) にも記した季節調整と欠測値補完の新たなソフトウェア Ecce Signum を完成し、各方面に配布するとともに文書を作成中である。また、多重線形回帰モデルの下で事後予測分布及びプラグイン・サンプリングを用いて単一代入によって発生させた合成データによる推定手法について、ベイジアンによる分析を準備しており、多重線形回帰モデルの下で事後予測分布及びプラグイン・サンプリングを用いて多重代入によって発生させた合成データによる推定手法については、頻度論及びベイジアンによる分析を完了し、報告書を準備中である（以上 CSRM (2020)）。

（カナダ）

Gray (2020) は、Haziza (2003)[13] や Stelmack (2018)[14] を引き継いだインピュテーション手法の評価ツール ImpACT について論じている。ImpACT は、「無回答」、「インピュテーション」、「分析」の 3 つのモジュールから成り、与えられたデータについて指定した欠測パターンで欠測値を発生させ、それをいくつかのインピュテーション手法で補完して結果を評価している。評価の対象は、(a) 分布の正確性（周辺分布及びより高い次元の分布）、(b) 推定の正確性（低次のモーメント）、(c) 予測の正確性（個々の値） である。この報告では、カナダの小売商品調査 (Retail Commodity Survey, RCS) で実際に行われているインピュテーションによったデモンストレーションを行い、特に履歴インピュテーション（前月又は前年の値を用いるインピュテーション）で外れ値の影響が大きいこと、シミュレーション用データでは履歴インピュテーションでほとんどのデータが補完されるが、現実には履歴インピュテーションを行えるデータは 20 パーセント未満であるとの結果を得ている。個々のインピュテーション手法の評価では、各手法において値が大きくなるとインピュテーションの結果に下方バイアスが生じること、比率によるインピュテーションでは、小さな値の時に系統的な上方バイアスが見られることが指摘されている。

２．欧州

（オランダ）[15]

Scholtus and Daalmans (2020a) はオランダの「バーチャル・センサス」の「マス・インピュテーション」[16]による分散の推定を扱っている。先行するディスカッションペーパー（Scholtus (2018)）と同様に分析的手法とブートストラップの 2 種類の方法を比較し、加えて分析的手法が可能な条件を考察している。また、この課題について新たな文書（Scholtus and Daalmans (2020b)）を準備中であるとしている。

---

[13] 2003 年時点で使用されていたシミュレーションシステム GENESIS (Generalized Simulation system) について解説している。
[14] 坂下 (2019) 参照。
[15] 欧州統計システム (ESS) の事業についてのオランダ統計局 (CBS) の報告を含む。
[16] 坂下 (2020) 参照

坂下 (2018) に記した欧州統計システム (European Statistical System, ESS) のデータ妥当性の検証プロジェクト ValiDat Integration について Ten Bosch and van der Loo (2018) は、その重要な概念とプロトタイプの使用例を示している。ESS では、データ検証の原則として、(1) 早いほど良い、(2) 信頼するが、検証する、(3) 十分に文書化され、適切に伝達された検証ルール、(4) 十分に文書化され、適切に伝達された検証エラー、(5) ルールに準拠していなければ説明する、(6) 「十分良い」ことが新しい「完全」 の 6 つの原則を掲げ、特に (4) は、明確に文書化された一般的な検証レポートの必要性を意味しているとしている。また、検証レポートは機械によって読めるものと人間が読めるものの両方の形態で出力され、人間が読むための出力には対話式ダッシュボードと静的なレポートの 2 形式が実装されている。詳細な設計と技術的標準については Ten Bosch and van der Loo (2017) に記されている。

また、Ten Bosch et al. (2020) によると、ESS が各国統計局とやりとりするデータのチェックに多大な労力がかかっているため、欧州統計局が各国の検証ルールを調査して 21 の「主要な検証ルール」を特定し、一般化して、自然言語と検証と変換用の言語 (Validation and Transformation Language, VTL) で記述し、これに基づいてオランダ及びポルトガルの統計局が、「主要な検証ルール」を国内のシステムに実装する実験を行った。論文は、実験の実際を報告し、(1) VTL で表現されたルールを自動的に実行する、(2) 分野固有の表現に変換して適用する の 2 種類のアプローチが考えられるが、少なくとも後者の方法は可能であると結論づけている。

（フランス）

Babet (2020) は、フランスの労働力調査の賃金の深層学習 (Deep Learning: DL) によるインピュテーションのシミュレーションを行い、その中で、労働力調査の仕組み、無回答の発生状況、インピュテーション手法を報告している。シミュレーションの結果として、賃金の DL によるインピュテーションは、現行のインピュテーションモデル (SALRED: salaire redressé、修正された賃金) や伝統的な賃金方程式 (Mincer type wage equations、ミンサー型賃金方程式) より良い結果を得たとしている。

（ノルウェー）

Jentoft (2020) は、機械学習を用いた予測モデルによって欠測値の補完を行う際の学習データのエディティングについて論じ、雇用統計のフルタイム換算係数の予測を例として、外れ値を検出する方法を見直すことにより、2018 年から 2019 年にかけて予測モデルに使用できる学習データが増加したと報告している。また、分類ごとの学習データ数が偏っていると、まれな分類の予測を正しく行うことが難しくなるので、その対処法としてリサンプリングにより分類ごとのデータ数を調整する手法を紹介している。ただし、これを行うと学習データの当てはまりは良くなるものの、予測するデータの二乗平均誤差は増加する傾向があり、過学習が起きているのではないかと考察している。

Pekarskaya and Zhang (2020) は、Jentoft and Zhang (2018) で紹介した二段階学習 (two-phase learning)[17] のテスト結果について報告している。この手法は、データ全体を学習データ、テストデータ、最終テストデータに３分割し、第１段階で学習データによる学習を行い、第２段階でテストデータを用いて誤差の評価を行ってモデルを修正する。この論文では第１段階で用いるモデルや第２段階で用いる補助データについていくつかのパターンでシミュレーションを行い、この手法により予測誤差の平均及び分散の不均一を捉える枠組みが提供され、個別データレベルでの予測誤差の評価を改善できると結論している。

（英国）

Leather (2020) によると、英国国家統計局 (Office for National Statistics, ONS) では、最近隣法に基づく新たなインピュテーションシステム RBEIS (Rogers & Berriman E&I System) を開発した。これは、大規模調査用に開発された CANCEIS のような旧来のシステムが小規模調査ではインピュテーションによる分散を大きくするリスクがあったため、個別のレコードではなく欠測値の集合全体にインピュテーションを行うことにより対処したもので、2017 年以降の ONS の社会調査のエディティングとインピュテーションに成功裏に用いられているとしている。

また、Sthamer (2020) によると、ONS において生活費・食料 (Living Cost and Food, LCF) 調査のエディティングとインピュテーションは職員による手作業で行われており、時間がかかりすぎていると思われる一方、家計関係の他の調査で行われている外れ値検出システムに LCF のデータをかけると修正されたデータの 10 パーセントしか検出されないため、このシステムの精度に疑問があると示唆されている。このため、ONS では機械学習を用いて LCF で修正されるデータを予測するシステムを検討した。論文では、機械学習に用いる項目、モデルや閾値の設定についての実験で 22 項目を用いたランダムフォレスト・モデルが良い成績を収めたこと、純粋な好奇心で始めた研究だが実用化も視野に入れていることを報告している。

（ドイツ）

Dumpert (2020) は、UNECE で進められているエディティングとインピュテーションへの機械学習によるアプローチについて検討するプロジェクトの報告である。それによると、このアプローチのインピュテーションについての過去の文献は豊富だが、エディティングについては数少ない。参加しているのは、オーストラリア、ベルギー、イタリア、ポーランド、スイス、英国、ドイツの統計局であり、現在エディティングについては 2 つ、インピュテーションについては 4 つの予備研究 (pilot study) が進行中である。パラメトリック・モデルなどと比べて仮定が少なくて済み、他の手法と比べて少ない人間の介入で他の手法に匹敵する結果を得られるなどの肯定的な結果が得られており、2020 年に報告を計画していると記

[17] 坂下 (2019) 参照。

述している。

Lange, K. (2020) によると、ドイツ連邦統計局では、現状では多くの統計で結果への影響を考慮することなく手作業で行われているエディティングとインピュテーションの過程の自動化と標準化のため、新たに設けられる「所得のデジタル構造調査」を対象として 3 種のツールをテストし、その中ではカナダ統計局の CANCEIS が最良の結果をもたらした。このため、CANCEIS を導入し、将来的には結果を比較し違いを分析するために R パッケージの missForest を含めたいとしている。

（イタリア）

Rocci et al. (2020) によると、イタリアのサービス業の売上についての短期調査 (STS FAS) は、Eurostat の新たな指令により四半期ごとから毎月の報告に移行するため、より短い時間で結果を公表するために必要な新たなプロセスを検討するプロジェクトを開始した。現状分析の結果、現行のエディティングとインピュテーションは対話型のエディティングにより莫大なコストがかかっているため、対話型の処理を要する影響の大きいエラーを特定する選択的エディティングに優先順位が置かれた。従来の外れ値検出法に加えて SeleMix R パッケージを用いた選択的エディティングを導入すると、対話型のエディティングを要するデータを 3 割に抑えることができたが、なおコストがかかる。従来の手法の一部にランダムフォレストによる機械学習を加えた手法が有望であることが、最初の結果として得られている。

Di Zio et al. (2020) は、Di Cecco et al. (2018)[18] が記していた個人レジスタにおける最終学歴データのマス・インピュテーションについてのその後の検討を報告している。ここでは、さまざまなデータソースから得られたイタリア基礎個人レジスタ (BRI) のデータを、2011 年の人口センサス、2018 年から始まった恒久 (permanent) センサスのための抽出調査、教育大学研究省の行政データに含まれているか否かでグループに分け、それぞれについて最新の最終学歴を推測（prediction、この文献で「マス・インピュテーション」と同義とされている）している。手法は Di Cecco et al. (2018) で推奨された対数線型モデルに基づいており、毎年の BRI を推計するためのモデルの変更、推計の改善のための情報の追加、2021 年センサスで利用するための分類の細分化が必要と結論づけている。

その 2018 年恒久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) の処理については、Bianchi et al. (2020) が報告している。これは、毎年 2,800 の自治体に居住する約 140 万世帯を対象として標本調査を行い、関連する行政データや登記データを統合して作成するものである。その E&I 戦略では、課題を分割してそれぞれに単純化し、それぞれについて解決法を見いだすために、調査事項を 3 つのグループ（1.住居と建物、2.性別、出生等、3.教育、経済活動等）に分け、ある程度独立して行われている。各グループにおいて主要なインピュテーション手法は、演繹的手法又は確率的手法だが、第 3 グループの教育と雇用に

[18] 坂下 (2019) 参照。

は行政データ、交通手段にはビッグデータも用いられている。また、データに基づいた (data driven) インピュテーション・ソフトウェア DIESIS をローマ大学と共同で開発した。また、DIESIS を含む E&I 全体のマネージメントシステムとして DEIS (Data Editing and Imputation System) がある。

（ポーランド）

Długosz (2020) によると、ポーランド統計局では、伝統的な統計の作成過程が分野ごとに分かれ、共有されていなかったのを見直し、UNECE、Eurostat 及び OECD で開発した汎用統計ビジネスプロセスモデル (Generic Statistical Business Process Model, GSBPM) に基づいた統計作成プロセスモデルを導入し、プロセス指向でメタデータに基づいた統計作成への移行を進めている。ここでは統計作成の過程を、収集、処理、分析、普及の 4 段階に分割し、各段階におけるデータとメタデータを分けて保存することとしている。論文では、特にメタデータの重要性が強調されている。

また、Dygaszewicz (2020) によると、ポーランド統計局は過去のセンサスにおいて世界に先駆けて行政及び非行政記録の統合、紙の調査票の全廃など、コンピュータを使用した先進的な手法を導入しており、2021 年人口センサスに向けて、センサス用データベースとして利用するレジスタのフル・センサスへの拡張、調査区ベースから座標ベースへの位置情報の精緻化などを進めている。ビッグデータの導入には現時点では明確なガイドラインと十分な専門家がいないとして慎重であるが、将来的な可能性は排除していない。

（オーストリア）

Kowarik et al. (2020) は、欧州統計システム (ESS) のプロジェクト ESSnet Big Data II の一環として行った、ウェブスクレイピング（ウェブサイトから情報を抽出する技術）によって企業の情報通信技術 (ICT) に関する特性を収集し、調査を置き換え、又は検証する試みについて報告している。欧州統計局では毎年、ICT と電子商取引についての調査を行っているが、調査内容が拡大する中で回答の負荷が大きく、またウェブについては隔年でしか調査されず、新たな項目は含まれていない。ウェブスクレイピングは、この制限を克服するために、代替的な情報源として検討されたもので、時間の短縮と対象の拡張も期待されている。論文では、サイトの有無、ソーシャルメディアへのリンク、ネット販売の有無のウェブスクレイピングによる調査を検討した結果を報告し、ウェブスクレイピングの精度は良いが、調査を完全に置き換えるには至っていないとしている。

３．まとめ

今年度の調査対象となった文献から、アメリカ合衆国の人口センサスでは、調査対象に接触できないことの多さが大きな問題となっており、その対策として、行政記録の利用が調査対象への接触の判断とデータ補完の両面で課題となっていることが分かる。企画、実

査、製表が分かれている日本の組織ではインピュテーションは製表における課題とされるが、人口センサス全体が連邦政府のセンサスの直轄事業となっている米国では実査時点の課題にもなっていることが興味深い。カナダや欧州でも、コロナ禍などの影響で研究発表は低調になっているものの、インピュテーションに関する行政記録の利用、機械学習の適用、システムの構築・共有などさまざまなプロジェクトが進められていることが窺える。

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

[6] Babet, D. (2020), “Wage Imputation with Deep Learning in the French Labor Force Survey”, Conference of European Statisticians, United Nations Economic Commission for Europe, Geneva, April 2020.

[7] Bianchi G., Filippini R., Lipsi R.M., Pezone A., and Scalfati F. (2020), “An overview of the editing and imputation process of the 2018 Italian Permanent census”, Conference of European Statisticians, United Nations Economic Commission for Europe, Geneva, April 2020.

[8] CSRM (2016), “Annual Report of the Center for Statistical Research and Methodology, Research and Methodology Directorate, Fiscal Year 2016”, U.S. Department of Commerce, Economics and Statistics Administration, U.S. CENSUS BUREAU.

[9] CSRM (2020), “Annual Report of the Center for Statistical Research and Methodology, Research and Methodology Directorate, Fiscal Year 2020”, U.S. Department of Commerce, Economics and Statistics Administration, U.S. CENSUS BUREAU.

[10] Deaver, K. D. (2020), “Intended Administrative Data Use in the 2020 Census”, May 1, 2020. 2020 Census Planning Documents, 2020 Census Memorandum Series.

[11] Di Cecco D., Di Laurea D., Di Zio M., Filippini R., Massoli P., and Rocchetti G. (2018), “Mass imputation of the attained level of education in the Italian System of Registers”, Workshop on Statistical Data Editing, United Nations Economic Commission for Europe, Neuchâtel, September 2018.

[12] Di Zio M., Filippini R., and Rocchetti G. (2020), “An imputation procedure for the Italian attained level of education in the register of individuals based on administrative and survey data”, Conference of European Statisticians, United Nations Economic Commission for Europe, Geneva, April 2020.

[13] Długosz, A. (2020), “Modern, process oriented and metadata driven statistical production”, Conference of European Statisticians, United Nations Economic Commission for Europe, Geneva, April 2020.

[14] Dumpert, F. (2020), "The UNECE High-Level-Group for the Modernization of Official Statistics Machine Learning Project: A report of the Editing & Imputation Group", Conference of European Statisticians, United Nations Economic Commission for Europe, Geneva, April 2020.

[15] Dygaszewicz, J. (2020), "Use of administrative data and alternative data for census when applying modern technologies", Conference of European Statisticians, United Nations Economic Commission for Europe, Geneva, April 2020.

[16] Gray, D. (2020), "Evaluating Imputation Methods using ImpACT: First Case Study", Conference of European Statisticians, United Nations Economic Commission for Europe, Geneva, April 2020.

[17] Haziza, D. (2003), "The Generalized Simulation System (GENESIS): A Pedagogical and Methodological Tool", 2003 Joint Statistical Meetings – Section on Survey Research Methods.

[18] Hutchinson, R. J. (2019), "Improving Retail Trade Data Products Using Alternative Data Sources", Big Data for Twenty-First Century Economic Statistics, National Bureau of Economic Research.

[19] Jentoft, S. (2020), "Data editing for machine learning prediction models", Conference of European Statisticians, United Nations Economic Commission for Europe, Geneva, April 2020.

[20] Jentoft, S. and Zhang, L. C. (2018), "Two-phase and double machine learning for data editing and imputation", Workshop on Statistical Data Editing, United Nations Economic Commission for Europe, Neuchâtel, September 2018.

[21] Keller, A. (2016), "Imputation Research for the 2020 Census", Statistical Journal of the International Association of Official Statistics, 32 (2016): 189-198.

[22] Keller, A., Mule, V. T., Morris, D. S., and Konicki, S. (2018), "A Distance Metric for Modeling the Quality of Administrative Records for Use in the 2020 U.S. Census", Journal of Official Statistics, 34(3): 599–624.

[23] Kowarik, A., Gussenbauer, J., Mikesa, L., Weinauer, M., Peterbauser, J., and Rannetbauer, W. (2020), "Webscraped data for replacing and validating survey questions", Conference of European Statisticians, United Nations Economic Commission for Europe, Geneva, April 2020.

[24] Lange, K. (2020), "Automation of E&I processes", Conference of European Statisticians, United Nations Economic Commission for Europe, Geneva, April 2020.

[25] Leather, F. (2020), "RBEIS: A robust nearest neighbour donor imputation system implemented in SAS", Conference of European Statisticians, United Nations Economic Commission for Europe, Geneva, April 2020.

[26] Morris, D. S. (2014), "A Comparison of Methodologies for Classification of Administrative Records Quality for Census Enumeration," Joint Statistical Meetings.

[27] Morris, D. S. (2017), "A Modeling Approach for Administrative Record Enumeration in the Decennial Census", Public Opinion Quarterly: Special Issue on Survey Research, Today and Tomorrow, 81(S1): 357-384.

[28] Morris, D. S., Keller, A., and Clark, B. (2016), "An Approach for Using Administrative Records to Reduce Contacts in the 2020 Census", Statistical Journal of the International Association for Official Statistics, 32(2): 177-188.

[29] Mulry, M. H. and Keller, A. (2015), "Are Proxy Responses Better Than Administrative Records?", Joint Statistical Meetings.

[30] Mulry, M. H. and Keller, A. (2017), "Comparison of 2010 Census Nonresponse Follow-Up Proxy Responses with Administrative Records Using Census Coverage Measurement Results", Journal of Official Statistics, 33(2): 455–475.

[31] Pekarskaya, T. and Zhang, L. C. (2020), "Two-Phase Learning", Conference of European Statisticians, United Nations Economic Commission for Europe, Geneva, April 2020.

[32] 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.

[33] Rocci, F., Varriale, R., and Coppola, S. (2020), "ML to identify patterns behind errors in STS statistics", Conference of European Statisticians, United Nations Economic Commission for Europe, Geneva, April 2020.

[34] Scholtus, S. (2018), "Variances of Census Tables after Mass Imputation", Discussion paper, Statistics Netherlands, The Hague.

[35] Scholtus, S. and Daalmans, J. (2020a), "Variance estimation after mass imputation with an application to the Dutch population census", Conference of European Statisticians, United Nations Economic Commission for Europe, Geneva, April 2020.

[36] Scholtus, S. and Daalmans, J. (2020b), "Variance Estimation after Mass Imputation based on Combined Administrative and Survey Data".

[37] Sthamer, C. (2020), "Editing of LCF (Living Cost and Food) Survey Income data with Machine Learning", Conference of European Statisticians, United Nations Economic Commission for Europe, Geneva, April 2020.

[38] Stelmack, A. (2018), "On the Development of a Generalized Framework to Evaluate and Improve Imputation Strategies at Statistics Canada", Workshop on Statistical Data Editing, United Nations Economic Commission for Europe, Neuchâtel, September 2018.

[39] Ten Bosch, O. and van der Loo, M. (2017), "Design of a generic machine-readable validation report structure", Technical report, Statistics Netherlands, 2017.

[40] Ten Bosch, O. and van der Loo, M. (2018), "A generic Validation Report for the ESS", Workshop on Statistical Data Editing, United Nations Economic Commission for Europe, Neuchâtel,

September 2018.

[41] Ten Bosch, O., van der Loo, M., and Quaresma, S. (2020), “Implementing main types of International validation rules in national validation processes”, Conference of European Statisticians, United Nations Economic Commission for Europe, Geneva, April 2020.

[42] Thibaudeau, Y. and Morris, D. S. (2016), “Bayesian Decision Theory to Optimize the Use of Administrative Records in Census NRFU”, Joint Statistical Meetings.

[43] Thompson, K. J., Kim, H., Bassel, N., Bembridge, K., Coleman, C., Freiman, M., Garcia, M., Kaputa, S., Riesz, S., Singer, P., Valentine, E., White, T. K., and Whitehead, D. (2020), “Final Report: Economic Census Synthetic Data Project Research Team”, ADEP WORKING PAPER SERIES, Working Paper ADEP-WP-2020-05, October 2020, Associate Directorate for Economic Programs, U.S. Census Bureau.

[44] Vitrano, F. A. and Chapin, M. M. (2012), “Possible 2020 Census Designs and the Use of Administrative Records: What is the impact on cost and quality?”, 2012 Federal Committee on Statistical Methodology (FCSM) Research and Policy Conference.