リサーチペーパー第 40 号　　Research Paper No.40

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～

Current Condition in Imputation of Missing Data in Official Statistics outside Japan
― Review of Literatures

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

平成 29 年 7 月
July 2017

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：平成 29 年 6 月 23 日

本ペーパーは，総務省統計研究研修所統計研修研究官が，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。

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～

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、データ・エディティング、特に欠測値補完（インピュテーション）は避けて通れない事項であり、海外でも国際会議や統計機関の発表するワーキング・ペーパーなどでの議論が活発に行われるようになってきている。このため、諸外国におけるデータ・エディティング、特に欠測値補完の状況について、公開情報として入手可能な文献を調査した。

その結果、現時点で行われているインピュテーションはホット・デック法、比率代入法などの伝統的な手法が多く、回帰代入法や平均値代入法を用いている国（機関）も多い一方で、高度な分布モデルに基づいた手法も検討されていること、行政情報の統計作成への利用の進展に伴って、そのためのインピュテーションの必要性や、行政情報による統計のインピュテーションについての報告も見られること、欧州統計システム(ESS)や一部の国では手法の標準化が試みられていること、インピュテーションのロバスト化手法については、外れ値を除去してからのホット・デック法の適用、平均値に代わる中央値やトリム平均の使用、winsorization の採用などが見られることが分かった。

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、精度

Current Condition in Imputation of Missing Data in Official Statistics outside Japan — Review of Literatures

SAKASHITA Nobuyuki

Abstract

While maintenance and enhancement of accuracy in official statistics are emerging as urgent issues, data editing, including imputation of missing data, is recognized as an inevitable matter and active discussions in international conferences and working papers are taking place outside Japan. In this concern, we reviewed foreign literatures, which can be accessed publicly, on current condition in imputation of missing data in official statistics.

We found that current methods of imputation are mostly conventional, such as hot-deck or ratio imputation while many countries and organizations are using regression or mean imputation method as well, whereas advanced methods based on distribution models are under examinations, that utilization of administrative data for statistics lead to necessity of imputation for that purpose and/or imputation of statistics using administrative data, that ESS (European Statistical System) and some countries are seeking for standardization of methodology, and that hot-deck method after removing outliers, use of median or trimmed mean in place of conventional arithmetical mean, and winsorization are used to obtain robustness of imputation.

Keywords: Data Editing, Imputation of Missing Data, Accuracy

０．はじめに

日本の公的統計において、これまで欠測値補完（インピュテーション）を含むデータ・エディティングについての大規模な議論は行われてこなかった。これは、話題の性質上、現状を公開しにくいなどの事情があったものと思われるが、政府統計の精度維持・向上が喫緊の課題となる中で、データ・エディティング、特にインピュテーションは避けて通れない事項であり、海外でも国際会議や統計機関の発表するワーキング・ペーパーなどでの議論が活発に行われるようになってきている。

このため、日本における検討の端緒とすることを目的として、諸外国におけるデータ・エディティング、特に欠測値補完（インピュテーション）の状況について、ネットなどで公開情報として入手可能な文献を調査した。調査対象としたのは、各国統計機関が公表した論文や国際会議用のプロシーディングなどで、近年のものを中心に、必要に応じて過去に遡る方法を取った。

公的統計の欠測値補完（インピュテーション）は、わが国では長らく実務的な課題と考えられ、学術的な議論にさらされていなかったこともあり、宇都宮・園田(2001)や平川・鳩貝(2012)に記されているように、方法の整理についてはっきりとしたコンセンサスはなく、日本語の定訳も存在しない[1]。このため、本稿では宇都宮・園田(2001)に倣って、用語などは基本的に原文に従い、必要に応じて原語や注釈を付記することとした。

構成は、１．が国内の先行サーベイ、２．がアメリカ合衆国、３．が欧州諸国、４．がその他、５．がまとめとなっている。

１．国内の先行サーベイ

平成 24 年度に内閣府が野村総合研究所に委託して行った「統計データの補完推計に関する調査」[2]では、米国センサス局、労働統計局 (BLS) 及びカナダ統計局にヒアリングを行い、米国の人口センサス、地域社会調査 (American Community Survey, ACS)（以上センサス局）、職業雇用統計 (Occupational Employment Statistics, OES)（労働統計局）、労働力調査 (Labour Force Survey, LFS)（カナダ統計局）について、その第 4 章にまとめている。

その結果、米国では人口センサスの項目非回答[3]について Assignment、Allocation、代替 (Substitution) の 3 つの手法を組み合わせた対応、ACS の項目非回答について Assignment、Allocation の 2 つの手法を組み合わせた対応を取っていること、OES では、ユニット非回答の事業所の従業員構造について最近隣法に基づくホット・デック法[4]による代入を行った

[1] “imputation” の訳語としては（独）統計センター及びその周辺では「補定」が使用されているが、以下の方針に基づき、本稿では「欠測値補完（インピュテーション）」又は単に「インピュテーション」を用いる。
[2] 野村総合研究所（2013)。
[3] 調査対象の一部項目の回答が得られないことを項目非回答 (item nonresponse) と言い、全項目の回答が得られないユニット非回答 (unit nonresponse) と区別している。
[4] 広い意味の「ホット・デック法」は、インピュテーションのためのデータを同じ調査から得る手法全般を指すが、以下の議論を通じて、あるデータの変量の欠測した部分に、観測された部分が一致する他のデータ群から代入する手法を指しているようである。これは、日本で言う「窓法」や「世帯類型データベー

後、ユニット非回答及び項目非回答の事業所に賃金ごとの従業員数分布の代入を行っていること、カナダでは LFS において、項目非回答についてはドナーを用いたランダム・ホット・デック法による代入、前月からの横置き (Carry-Forward)、推定による代入を行い、ユニット非回答については回答履歴の状況に応じて、ホット・デック法あるいはウェイト付けによる補完を行っていることが分かったと記述している。

ここで、Assignment はその回答者の他の回答項目から推測される値を代入する手法、Allocation は類似した他の回答者の回答から推測される値を代入する手法(「最近隣法等の統計学的な処理を含む手法の総称で、具体的にはホット・デック法が用いられている」との記述あり。)、「代替 (Substitution)」は「世帯内のすべての対象者に関する回答が欠測している場合、近隣の世帯の回答に基づいて補完する」とされている。

また、高橋 (2016) は、選択的 (Selctive) エディティングを中心に政府統計データのエディティングに関する国際的動向を論じているが、付録として過去 2 回[5]の国連の欧州経済委員会 (United Nations Economic Commission for Europe, UNECE) 欧州統計家会議 (Conference of European Statisticians, CES) の「統計データ・エディティングに関するワークセッション」(Work Session on Statistical Data Editing) の報告論文の概要を掲載している[6]。この会合は、2000 年以降ほぼ 1 年半に 1 回各国の統計家が集まってインピュテーションを含むデータ・エディティングについて各国から報告し、議論しているもので[7]、各会合には欧州各国の他、カナダ、ニュージーランド、米国、日本なども参加している。会合のアジェンダ及び発表内容（プレゼン資料及び論文) は UNECE のサイト上に公開されており、各国の事情を知ることができる。

高橋(2017) [8]は、政府統計ミクロデータの作成・提供に焦点を当て、欠測値の処理を論じているが、その中で、UNECE 会合の参加 23 機関に対して、インピュテーションの手法として回帰代入法、比率代入法、平均値代入法、ホット・デック法の使用状況についてアンケートを行い、87%の回答を得ている。その結果、ホット・デック法は 100%、他の手法も 95%の機関で使用されており、よく使用されている手法としてはホット・デック法 (65%)、次いで比率代入法 (60%) が挙げられ、統計の種別には事業所・企業統計で比率代入法 (80%)、世帯統計でホット・デック法 (80%) の使用が多くなっていると記述している。

なお、前述の宇都宮・園田 (2001) では、Bailar and Shapiro (1981) を引用する形で、「米国のセンサス局では（中略）『人口統計の分野では、似通った特性を持つ回答者を代入する Hot Deck を用いる』一方、『経済統計の分野では、同一標本や同業種の過去値を基礎とする』といった大枠の考え方を示している。」と記述している。

ス」に近い手法であるが、両者は前回の調査データを用いているためむしろコールド・デックに属する。

[5] 2014 年のパリ会合及び 2015 年のブダペスト会合。

[6] ただし、あくまでも本論の参考としての要約で、必ずしも各国のインピュテーションの現状に言及していないため、本稿を書くに当たって改めて元報告を参照した。

[7] 最新の会合は 2015 年 9 月にハンガリーのブダペストで開かれている。次回は 2017 年 4 月にオランダのハーグで開かれる予定。

[8] 経済統計学会の学会誌「統計学」創刊 60 周年記念事業特集に寄稿された論文である。

２．アメリカ合衆国

分散型の統計制度を有する米国では、センサス局やその他の官庁の統計部局が統計に関する研究を行い、さまざまな機会に発表を行っている。ここではセンサス局の公式の業務報告、センサス局に属する The Center for Statistical Research & Methodology (CSRM) [9]の研究報告書、センサス局のその他の公表論文、センサス局以外の公表論文に分けて記述する。

（センサス局報告書）

米センサス局では、2000 年人口センサス以降、人口センサスの項目非回答についての報告書を作成しており。Zajac (2003)、Rothhaas et al. (2012)などではその発生状況の他、上記各補完手法の解説、適用状況の分析などを公開している。なお、Rothhaas et al. (2012)では、2000 年人口センサスでの項目非回答の報告書として Norris (2003) を挙げており、これは報告書番号から見て Zajac (2003)と対を成すものと思われるが、ネット上に発見できない[10]。

また、人口センサスのロング･フォーム調査を独立させた地域社会調査のインピュテーション手法について、Slud (2015) は、1990 年人口センサスのロング・フォームで行われた手法（ホット・デック法）を踏襲していると記述している。この報告では、2 値変量 (binary outcome variables) のインピュテーションを分布モデルに基づいた手法で置き換えることを試み、かなり異なったものとなると評価している。

（CSRM）

CSRM では、センサス局の業務に関する調査研究及び分析業務を行っており、その業務には新たな手法やソフトウェアの開発が含まれている[11]。ここで作成した報告書は Research Report Series としてまとめられ、ネットから入手できる。インピュテーションに関する報告書には、Thibaudeau (1999)、Thibaudeau et al. (2006)、Chen (2007)、Winkler (2008)、Kim et al. (2008)、Erdman and Nagaraja (2010)などがある。

このうち、Thibaudeau (1999) は、ホット・デック法を多変量変数に適用した場合、変量間の相関を実際以上に高めてしまうとの問題意識から、サクラメント市におけるドレス・リハーサルでの人口学的な項目の項目非回答につき、最尤推定に基づいた階層的対数線形モデル (hierarchical log-linear model) を用いた結果をシーケンシャルなホット・デック法と比較したものであり、その時点で一般的な手法は、固定セル (fixed-cell) 、シーケンシャル又は最近隣 (nearest neighbor) によるホット・デック法であると記述し、モデルに基づいた

[9] 2011 年に統計研究課(Statistical Research Division)の大部分を改組して発足。この際、Research and Methodology Directorate と the Center for Survey Measurement and the Center for Disclosure Avoidance Research は独立した(CSRM, 2016a)。

[10] センサス局サイトのセンサスの評価報告書を集めたページに見当たらないものであるが、他の報告でもホームページからのリンクが見つからないものがあり、公表を前提としていないか整理が良くないと思われる。

[11] https://www.census.gov/srd/csrm/

アプローチは極めて豊か (rich) であると結論している。

Thibaudeau et al. (2006) は、Survey of Income and Program Participation (SIPP)（所得及びプログラム参加調査）における資産と負債のインピュテーションに用いられていた単変量の後入れ先出し(last-in, first-out)スタックによるホット・デック法が、変数間の相関を弱める結果をもたらしている可能性があるとの問題意識から、代替手法として多変量のホット・デック法(Joint Hot-Deck Imputation)と予測平均法(Predictive Mean Imputation)を検討し、著しく改善するとの結果を得ている。

Chen (2007) は、カナダで開発された人口センサス用のエディット訂正システムの CANCEIS を合衆国センサスの 2006 年のテストデータに適用し、合衆国で用いられている if-then-else 形式のエディット・ルールと必ずしも同じではないものの、先住民関係の一部のデータを除いてほぼ同じ結果となることを示している。

Winkler (2008) は、従来のホット・デック法は、観測された変量の一致するドナーを見つけるために一致する変量を統合（collapse: 観測された変量のすべてを一致させられるドナーがないため一部の変量だけで一致させること）せねばならないことが多く、そうでなくても変量間の同時分布が保存されないとの問題意識から、制約条件のある離散データ一般に用いることができる対数線形モデルに基づいたインピュテーションの手法を論じ、制約条件を満たしつつ同時分布を保存でき、インピュテーションによる分散の推計も容易であると結論している。

Kim et al. (2008) は、人口センサスのロング･フォームデータ（現在は ACS で代替）で用いられていた数量項目の最近隣法によるインピュテーション (nearest neighbor imputation, NNI) の誤差分散の推定法を提案している。

Erdman and Nagaraja (2010)は、ACS における小地域推計のための Group Quarters（日本で言う施設等世帯）のインピュテーション手法を論じている（ACS は抽出調査なので、小地域推計では標本の存在しない地域が発生するためで、欠測値の補完と言うよりも一種の合成データ作成のためのインピュテーションと思われる。）。

なお、CSRM の年度及び四半期の業務報告 CSRM (2016)a、CSRM (2016)b、CSRM (2016)c によると、インピュテーションに関する共同研究プロジェクトとして、ソフトウェア[12]の開発と評価、ACS のインピュテーションに関する研究開発、月次卸売調査 (Monthly Wholesale Trade Survey, MWTS) のインピュテーション手法に関する研究、経済センサスの生産物データに関する欠測値補完についての研究が行われている。このうち、MWTS については CSRM (2016)a で、高度に歪んだ分布や経済センサスから得られた説明変数の欠測値などの課題があって、現在の Horvitz-Thompson (HT) とランダムなグループに基づいた推計値が標本理論の示唆するほど良くなく、代替案として、(1) 変数間の関係を保つためのパラメトリック（多変量正規）モデル、(2) 各変数の周辺分布を保つためのノンパラメトリックな単変量の密度

[12] センサス局では、エディティング、 インピュテーションと開示抑制の汎用ソフトウェア TEA を開発している。

推定、(3) (1)と(2)を連結するためのコピュラ・モデルから推計値を得ることを提案し、2015年の合同統計会議（後述）で構想 (idea) を発表した[13]と記述している。さらに、CSRM (2016)bでは母集団のデータを満たすためのインピュテーション手法を改良したとされ、CSRM (2016)c では全数調査層、抽出層それぞれについて、欠測値に対して、ランダムフォレストを条件モデルとする連鎖方程式による多変量インピュテーション(MICE, Multivariate Imputation by Chained Equations)[14]を適用したと記述している。また、一般的な研究として、CSRM (2016)a に、順次ロジスティック回帰 (sequential logistic regression) に基づいた品目と度数のインピュテーション手順を実装したこと、空き家を特定し居住している住戸の一部の計数を決定するために行政記録を利用する手法を開発したとの記述がある。

（センサス局その他）

報告書としてまとめられたものの他にも米センサス局は随時、アメリカ統計学会 (American Statistical Association, ASA) 等が集う合同統計会議 (Joint Statistical Meetings, JSM) や UNECE のワークセッション等で発表を行っている。その報告の中でインピュテーションを扱ったものとしては、Thibaudeau et al. (1997)、Thibaudeau (2002)、Thibaudeau et al. (2007)、Sands (2013)、Garcia et al. (2014) 、Garcia et al. (2015) などがある。

Thibaudeau et al. (1997) は、2000 年人口センサスのショート･フォームの人口学的項目のインピュテーションに、多変量の予期せぬ状況 (multivariate contingencies) に対応可能なシステムを考察したものであり、その中で、1990 年センサスのインピュテーションの仕様は複雑で長大であるが、もっとも重要なのは一種のホット・デック法、最後に観察された値で置き換えるもの（上記のシーケンシャルなホット・デック法を指すと思われる）で、この手法はインピュテーションの過程で変量の統計的な性質が失われる可能性があると記述している。

Thibaudeau (2002) は、Thibaudeau (1999)で得られた知見をカナダ統計局で発行している論文誌 Survey Methodology 用にまとめたものである（固定セル、シーケンシャル又は最近隣によるホット・デック法についての説明及び文献紹介有り）。

Thibaudeau et al. (2007) は、全米科学財団 (National Science Foundation, NSF) がスポンサーとなってセンサス局が実施している研究開発調査 (The Survey of Research and Development in Industry, SRDI)[15], の全数調査層 (certainty strata)[16]について、欠測値のインピュテーションを行う方法と乗数補正 (calibration) を行う方法を比較している。その中で、全数調査層にお

---

[13] 論文等未公表のため詳細不明だが、経済センサスを母集団として、標本調査で得られたデータを展開して推計値を得ている模様。MWTS と対を成す小売統計については、ビッグデータの利用の調査研究を行っているとしているものの、インピュテーションについては言及なし。

[14] MICE については、高橋・伊藤(2014)に解説されている。

[15] http://www.nsf.gov/statistics/srvyindustry/sird.cfm なお、この調査は Business R&D and Innovation Survey(BRDIS)に引き継がれた。https://www.census.gov/manufacturing/brdis/

[16] 背景に全数調査層でインピュテーションを行ったものの合計値が標本抽出層全体の合計値に匹敵するとの問題意識がある。

ける R&D の無回答の推計には ad-hoc なホット・デック法が用いられており、その時点でそれを支援する公式の統計的な仮定が存在していないと記述している。当時行われていた手法は、欠測値を前回或いはそれ以前の値を全体の増加率によって補正したもので置き換える手法で、論文では、この手法は一定の条件下で乗数補正を行うことに等しいことを示し、誤差評価を試みている。

Sands (2013)[17] は、2010 年人口センサスの 3 項目（世帯主との関係、年齢、性別）に対し「統計的な原理に基づいたランダムなインピュテーション」(statistically principled random imputation) を試みている。2010 年センサスで実際に用いられた手法は「レガシー・システム」とされ、50 年以上の歴史を持ち、最近隣法によると記述する一方、新たに提案する手法は、飽和対数線形モデル (saturated loglinear model) により多変量のインピュテーションを行うもので、レガシー・システムに匹敵する結果を得るものの、年齢と性別については改善の余地があると記述している。

Rastogi et al. (2014) は、人種とヒスパニック系に関する項目非回答について、行政データの利用を論じている。センサス局では従来、最近隣あるいは類似の性質を持ったユニットからの代入によるホット・デック法を用いていたが、レコード・リンケージ技術により 2000 年センサスその他の調査からの回答を利用して 2010 年センサスの欠測値に人種とヒスパニック系の回答を割り当てる (assign) ことができるようになり、ホット・デック法を要するデータの割合を人種については 50%、ヒスパニック系については 60%軽減することができたと記述している。

Garcia et al. (2014) は、Thibaudeau et al. (2006) でも取り上げられた SIPP の主に収入について論じ、その中で現行の SIPP の欠測値のインピュテーションは主に（確定的）ホット・デック法によっており、全世帯が欠測している場合はウェイトによる調整、世帯の一部の構成員が（全量）欠測している場合は人口特性の一致する観測値によるホット・デック、項目非回答については、回答されている項目からの類推（前記 Assignment に当たると思われる）又はホット・デックを行っていると記述している。当該論文では、従来の手法に代えて、TEA（脚注 12 を参照）によるランダム化されたホット・デック法及びモデルに基づいた順次回帰多変量法 (sequential regression multivariate imputation, SRMI) の 2 種類の多重代入法を検討し、モデルに基づいた手法は実現可能 (feasible) であること、モデルに基づく手法の分散はランダムなホット・デック法より小さくなること、モデルによる手法にはモデルに補助情報を取り込める可能性や多重代入法により誤差評価ができる利点があることなどを結論している。

Garcia et al. (2015) は、米国の経済センサスの調査内容（分類）が次回より北米生産物分類体系 (North American Products Classification System, NAPCS) に合わせて大幅に変わり、生産物と産業のリンクがなくなるとの問題意識から、比率 (Ratio) によるインピュテーション

[17] 著者は CRSM ではなく、人口センサスに関する研究開発を所管する部局(Decennial Statistical Studies Division)に所属している。

と SRMI を検討している。現行の手法は、分野ごとに異なり、鉱工業では総売上高との差を「特定されず」としてインピュテーションを行わず、建設業では最近隣法によるホット・デック、サービス業では総売上高からの比率によるインピュテーションを行っていると記述している。比率法と SRMI では、前者の方はエラーが少ないが分布の広い後者の方が欠測情報の割合 (The Fraction of Missing Information, FMI) が小さいとしている。なお、比率法の説明の中で、当てはまりに影響するかもしれない幾つかの (a few) 外れ値の存在について言及しているが、特に対処法を示してはいない[18]。

（米国その他）

分散型の統計制度を有する米国では、センサス局以外のさまざまな官庁も統計に関する研究発表を行っている。

Osburn (2013) は、内閣府／野村総研の報告書でも触れられている職業雇用統計 (OES) の賃金のインピュテーションを論じている。現行の方法は、同じ時期、都市統計地域 (Metropolitan statistical area, MSA)、産業（北米産業分類体系 (North American Industry Classification System, NAICS) で 4 ないし 5 桁)、企業規模のドナーを定め、十分な回答がない場合は統合 (collapse) したのちに分布の平均を代入する[19]が、この手法は州内格差が大きい場合、統合により推計値への影響が大きくなり、小地域での公表が限られると記述している。提案している手法は、賃金のレベルに基づいたより大きな 3 段階の区分を用い、OES の地域の賃金レベルと各企業の賃金レベルの正確な推計に焦点を当てたもので、OES のデータを雇用・賃金プログラム四半期センサス (Quarterly Census of Employment and Wages Program, QCEW) から得た企業の賃金についての補助データと併せて、誤差が入れ子になった多変量の線形モデル (nested error linear model for the case of multivariate data) を用いたインピュテーションを行うものである。ここで用いられている LP 法（Lohr and Prasad (2003)）の主な作業は、分散共分散行列の推計であるが、共分散の要素の推計値は、外れ値に対して極めて敏感であるため、ロバストな手法によって推計されなければならないとしている。ここでは、標準化したデータに winsorization[20]を適用し、M-推定量[21]を得ている。

なお、Kaminski and Kapani (2009) によると、QCEW における雇用と賃金のインピュテーションは、対象企業の過去一年のトレンドを延長して行っている。当該論文は、この方法では情勢の変化を取り込めないとの問題意識により、セル全体或いは最近隣の企業のトレンドを用いる方法（それぞれ「比率法」、「最近隣法」と命名）の評価を試みているとともに、除外すべき異常値の判定方法を論じ、「最近隣法」は誤差が大きく「比率法」の方が妥

18 経済センサスの比率インピュテーションにおける外れ値の対処法については、(独) 統計センターにおいて研究中である。

19 平均値を用いているが、代入する値を同じ調査から得ているため、野村総研の報告書ではホット･デック法の一種とされているものと思われる。

20 トリミングやウェイト付与により分布の両端の影響を小さくするロバスト化手法。

21 最尤法のように、外れ値に影響されにくい推定量。

当であるとしている。

Barboza et al. (2014) は、全米農業統計サービス (National Agricultural Statistics Service, NASS)[22] で行っている農業資源経営調査 (Agricultural Resource Management Survey, ARMS) 第 3 フェーズ（アラスカとハワイを除く全米で毎年行われる、郵送、面接等複数のモードにより、NASS 及び地域の両フレームを用いる、約 3 万 5 千の農場を対象とした抽出調査、調査の一部は 51 ページにのぼり、800 の変量があるとの記述あり。）のインピュテーションについて論じている。現行の手法では、欠測値には、地域、農場の種類、売上げ規模などに基づいたグループの乗率のない平均を代入している。それぞれの均一な農場のグループは、少数の大／小規模或いは珍しい業者のために代入する値がバイアスを受けないように上下双方の外れ値が除かれるが、各グループには最小で 10 個の観測値が必要で、満たない場合はなるべく均一性を保持するべく定められた優先順位でグループ統合 (collapse) を行う。第 1 段階のグループ化でほぼ 75%のインピュテーションが達成され、最終的にインピュテーションができないものは 1%の半分以下である。論文では、この手法には、データの分散が人為的に小さくなることや多変量解析において変量間の関係を保つ十分大きな変量の集合を保てないなどの短所があるとして、多変量同時分布を一連の線形モデルに分解し、回帰に基づいた手法を試みている。この手法は繰返し順次回帰 (iterative sequential regression, ISR) と呼ばれ、一連の線形モデルとインピュテーションのパラメータの推定値はマルコフ連鎖モンテカルロ法を用いた反復によって得られる。ISR は重要な関係や回答者の分布が保存されることが期待されるが、乗率補正 (calibration) との相互作用や外れ値評価の欠如により両手法を直接比較することは困難であるとし、初期の結果は有望であり近い将来その違いがより詳細に探求されるであろうと結論している[23]。

また、Miller and Young (2015) は、NASS の諸調査におけるインピュテーションについて報告しており、農業センサスでは、(1) 決定論理表 (decision logic tables, DLTs) の評価に基づいて得られる値（合計が欠測している場合など）、(2) 以前の調査から得られる値、これは他のさまざまな NASS の調査や前回センサスなどから組み立てられる。(3) ドナーを用いたインピュテーションの順番で行っていると記述している。また、前述の ARMS 第 3 フェーズの 2014 年調査で ISR を採用したとしている。

３．欧州

（欧州統計システム(European Statistical System, ESS)）

欧州統計局(Eurostat）と加盟各国の協力体制である ESS は、統計品質向上のためにさまざまなプロジェクトを実行している。

Scholtus and Willenborg (2014) は、2011 年から 2014 年まで行われた Memobust プロジェクトについて報告している（概略については高橋(2016)も参照のこと）。Memobust の目的の一

[22] 農務省 (USDA)の統計部門。
[23] なお、論文中に 2014 年の JSM により詳細な論文を提出すると予告しているが、未入手である。

つにビジネス統計作成の ESS ガイドラインを作成することがあるが、このガイドラインにおいて、インピュテーションは 7 節に分かれ、インピュテーションに関する特定の手法に属さない話題の他に、一般的な手法（演繹的インピュテーション、モデルによるインピュテーション、ドナーによるインピュテーション）、特定のテーマ（時系列データのインピュテーション、Little and Su 法、制約条件下のインピュテーション）について記述している。

Di Zio et al. (2015) は、データ妥当性検証方法の一般化 (generic approach to data validation) を目指すプロジェクト (ValiDat Foundation Project) について報告し（同じく高橋 (2016) も参照のこと）、その中の方法論に関するパッケージ (WP2) において、データ妥当性検証の定義が考えられていた以上に難しく、特にデータ・エディティングとインピュテーションの関係を明確にする必要があったと記述している。また、Giessing (2015) は、ValiDat Foundation Project はその一環として ESS 加盟国にアンケートを行い、人口センサス、物価、構造的ビジネス統計 (Structural Business Statistics, SBS) の分野ではデータ・エディティングとインピュテーションを分離処理しているところが多く、農業及び労働力統計では少なかったと記述している。

（英国）

Da Silva and Zhang (2014) は、英国の 2011 年人口センサスにおけるインピュテーションにより発生する分散の推定を論じているが、その中で、2011 年時点のインピュテーション方法は CANCEIS に実装されたドナーによる手法で、複数のモジュールの系列を通していると記述している。代入値を提供するドナーは、CANCEIS ver.4.5 ユーザーガイドで変化をほぼ最小にするインピュテーション動作 (near minimum change imputation actions, NMCIAs)と呼ばれている集団からランダムに選ばれている。

Gaughan (2015) は、英国統計局 (ONS) と国税庁が企業の売上や収益などの収集を重複して行っているため、これらの収集を統計調査の代わりに税務データにエディティングとインピュテーションを施して利用することの検討について報告している。その内容は、税務データと統計調査の報告単位が異なっているため、カンパニーの税務データを企業 (establishment) ごとに統合し、雇用データによって報告単位に按分する[24]ことを基本とし、カンパニーの一部に欠測値がある場合は欠測値に対し中央値によるインピュテーションを行い、カンパニーが丸ごと欠測している場合は、報告単位に対し直接インピュテーションを行っている。報告単位を直接インピュテーションする場合の手法としては、i) 層内の中央値によるインピュテーション、ii) 同じくトリム平均によるインピュテーション、iii) ONS のビジネスレジスタから得られる補助変数による比率インピュテーションを試験し、現在結果をまとめている所である。

（ドイツ）

[24] 「カンパニー」は日本で言う社内カンパニーで、「報告単位」が事業所に相当すると思われる。

Spies et al. (2014) は、ドイツ連邦統計局の 2010 年農業センサス[25]データにおける欠測値のインピュテーションによる結果の変動係数への影響の多重代入法による評価を試みたものである。インピュテーションのモデルとしてはホット・デック法と予測平均値マッチング (Predictive Mean Matching, PMM) を採用し、ホット・デックによるインピュテーションは極端な値になることがある一方、PMM は現実的な結果となったと報告している。

Spies (2015) は、初めてレジスタ･ベースで行われたドイツの 2011 年人口センサスの評価について論じている。2011 年人口センサスは、自治体の人口レジスタに 10%抽出調査を加味して行われ、抽出調査において実際に行われたインピュテーションは、コールド・デック法、演繹的 (deductive) インピュテーション、最近隣法の結合であると記述している。まず、人口レジスタ又は調査員が調査中に作成した世帯の基本情報に基づいたコールド・デックによるインピュテーションを行い、次に可能な所には変数間の関係性に基づいた演繹的インピュテーションを行い、それでも欠測している値には単変数の最近隣インピュテーションを行っていると記述している。この論文では、多重代入法によってインピュテーションの評価を行っている。

（オーストリア）

Kowarik et al. (2014) は、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 VIM の GUI 機能を備えた新バージョンについて報告している。その中で、VIM で扱える手法は「ホット・デック法のような old-fashioned な手法から順次ステップワイズ・ロバスト回帰法 (iterative step-wise robust regression imputation)[26] のような極めて洗練 (sophisticated) された手法まで多岐に渡っているとし、具体的には他に k 最近隣法と個別の回帰法が挙げられている。

Froehlich (2015) は、オーストリアでは産業及び建築の短期統計における欠測値のインピュテーションは、専門家が税務報告や社会保険会計など様々な行政上の資源にアクセスすることにより、ルーチンで処理していると述べている。この報告では、より早期の公表が求められるため、X12-Arima を用いた外れ値検出及びインピュテーションの自動化を試み、評価している。

（オランダ）

Pannekoek et al. (2014) は、企業統計のエディティングの自動化とその評価について論じている。その中でインピュテーションについて、保育所データでは演繹的インピュテーションと最近隣インピュテーション、農作物卸売のデータでは演繹的インピュテーションと回帰によるインピュテーションを示している。

Hoogland (2015) はオランダ統計局の短期統計 (Short Term Statistics, STS) システムのマ

[25] 高橋(2015)参照。
[26] Templ et al. (2011)。

クロエディンティングについて論じており、その中で欠測値は比率によるインピュテーション、具体的には補助変数に比率を乗ずることで行い、補助変数の候補は、前期値、前年同期値、就業者数の３つで、比率の算出に用いる回答者と補助変数はルールに基づいて決まり、外れ値のウェイトは低く置いて算出すると記述している。

（フランス）

Deroyon and Gros (2015) は行政情報を結合して統計を作成する INSEE の年次企業統計細密化システム (ESANE) で行っている外れ値対処のための winsorization（脚注 20 参照）について論じている。ユニット無回答に対しては外れ値を winsorization によって処理した後に乗数調整 (calibration) を行っている。財務データや社会保険の欠測値に対してはインピュテーションが行われており、財務データでは主に以前の年のデータに基づいている。

（スイス）

Kilchmann (2015) は、スイスの人口センサスの構造調査のデータ作成過程の評価について論じている。スイスの人口センサスは 2010 年にそれまでの古典的なセンサスから行政データと標本調査を結合したものに移行し、毎年約 250,000 人（抽出率約 3%）を対象に構造調査が行われている。確定的インピュテーションと最近隣インピュテーションを行っていること、最近隣インピュテーションを行う SAS マクロを開発したことを記述している。

（フィンランド）

Ollila (2015) はフィンランド統計局におけるエディティングの品質評価について論じており、その中でモニタリングの結果としての改善の一例として、平均値によるインピュテーションから中央値によるインピュテーションへの変更を挙げている。

（ノルウェー）

Foss and Seierstad (2015) は、ノルウェー統計局における行政データに基づく統計のエディンティングと評価について報告している。その中で、数値変数の欠測値は前回の値でインピュテーションを行い、カテゴリー変数は重要なものであるので、専門家による手作業のエディティングを行っていると述べている。

（イタリア）

Luzi (2015) は、イタリアのビジネス構造統計が従来の報告中心から行政データの徹底的な使用と限定的な標本調査へ移行する中での変化について記述している。従来の統計は中小企業の標本調査と大企業の全数調査からなっており、前者については外れ値を見出だすための分布の分析とエラーの発見を行った後に、主な経済変数については（財務諸表や部門別データベースの）行政記録によるインピュテーション、そしてセル内の最近隣ドナー

によるインピュテーションを行っていた。後者については、極めて大きな企業については注意深くフォローアップを行い、それ以外については中小企業と同じ方法によっている。このように行政記録を補助的に用いる旧来の方法に対して、新たな方法では、行政記録を最初の情報源とし、中小企業及び大企業のデータを補助的に用いている。この中でも、最近隣法や予測平均法のような古典的なインピュテーション手法が用いられている。なお、このような複数の手法を用いたインピュテーションについては 2 年前に報告された Zio (2014)に述べられている。

Pichiorri et al. (2015) は、農業物価指数のモデルに基づいた選択的エディテング処理について述べている。モデルに応じて指数の 1 ヶ月前の値、12 ヶ月前の値、前年に対する比で伸ばした値が予測値となり、欠測値に対しては予測値によるインピュテーション又は手作業による修正、影響の大きい値に対しては手作業による修正を行っている。

（スペイン）

Revilla (2015) は、選択的エディティングを最適化するためのモデルについて述べており、その中で、スペイン統計局は時系列データ（産業の短期指数）の外れ値検出及びインピュテーションに REGARIMA モデルを用いていると記述している。

（ハンガリー）

Andics (2015) によると、ハンガリー統計局はデータ提供者への自動的なリマインダーシステムを有しているが、電話や文書による督促を行う必要がしばしば生ずる。そのため、データ収集の費用対効果の最適化を検討している。専門家によって設定された主要変数については、事前の予測値が計算され、エディティングとインピュテーションは主要変数に集中して行われる。また、推計値へのインパクトが大きい回答者も前回の調査データや他の情報により事前に定義されている。エディティング後の処理のために統合情報処理システムを開発し、インピュテーション、外れ値処理などを行っている。伝統的なインピュテーションは回帰法、最近隣法や EU の提唱する他の手法（MEMOBUST (2014)）によっているが、それらの妥当性の検証も予定している。

(スロヴェニア)

Seljak (2014) は、スロヴェニア統計局のデータ処理近代化システムについて論じており、そこで例示しているインピュテーションの手法は、ドナーを用いたホット・デック法である。当初のドナーは同じ自治体から選ばれ、インピュテーションのためのセルには最低 10 のデータが必要であるが、それで代入できなかった場合は対象を同じ地方に拡大し、セル内の最低のデータ数は 5 に下げられる。

４．その他

（カナダ）

Guertin (2014) は、カナダの 2011 年人口センサスにおける CANCEIS の利用とその改善点について報告している。その中で、CANCEIS のドナーによるインピュテーションは、良く知られた Fellegi-Holt 法[27]ではなく（as an alternative to）、最近隣法 (Nearest Neighbour Imputation Method, NIM)[28] に基づいていると記述している。また、履歴による (historical) 或いは回帰によるインピュテーションを直接には行わないが、ドナーを見つける際に焦点の変数に相関のある（履歴上のあるいは回帰モデルの）変数を補助的なマッチング変数として用いることによって間接的に取り扱うことができる。

Saint-Pierre (2015) によるとカナダ統計局では経済統計のシステムを根本的に見直し、Integrated Business Statistics Program (IBSP) の下に統合しようとしており、一つの調査が一つの組織で完結していた従来型の体制から機能別に組織を横断して行う体制へ移行しつつある。その中で、エディティングとインピュテーションについても、一般化したシステムの使用を義務付けることが求められている。この 4 年前に提出された Godbout et al. (2011)では、影響度 (Measure of Impact, MI) のスコアを用いた評価と督促を一定の品質が達成されるまで繰り返す繰返し推計 (Rolling Estimates, RE) のモデルについて解説している。

（ニュージーランド）

Cox (2014) は、ニュージーランド統計局が進めているデータ処理の移行プログラムの中で開発している大規模調査のための非回答インピュテーション・パッケージの課題について、2008 年から導入されたミクロ経済プラットフォーム (micro-economic platform, MEP) で最初に統合された年次企業調査 (Annual Enterprise Survey, AES) を例にとって論じている。AES の情報源は、税務データ(IR10)、郵送調査、政府データ（公営企業）、慈善団体のデータの 4 種類で、税務データと慈善団体のデータの欠測値は収集時にインピュテーションを行い、郵送調査の欠測値は統合後にインピュテーションを行うと記述している。エディティングとインピュテーションは Banff[29] を含む AES エディティング＆インピュテーションシステムによって行われる。代入値を計算するに当たって、疑わしい値や外れ値は除外され、AES で用いられる主なインピュテーションは、消費税 (Goods and Services Tax, GST) による比率インピュテーション、平均雇用者数 (Rolling Mean Employment, RME) による比率インピュテーション、履歴によるインピュテーション、平均値インピュテーションの 4 種類であり、その優先順位は各変数についてさまざまな補助変数との相関や利用可能性によって決定される。

また、Zabala (2015)a によると、ニュージーランド統計局では社会調査のデータを処理するための世帯処理プラットフォーム(Household Processing Platform, HHP)を開発し、世帯経済

[27] Fellegi and Holt (1976)。
[28] Bankier et al. (1999) 及び Bankier (2011)。なお、Fellegi-Holt 法及び NIM については、国連(2000)も参照。
[29] カナダ統計局が開発した主に数量データ用のエディティングとインピュテーションのシステム。野村総合研究所(2013)第 4 章に解説あり。

調査(Household Economic Survey, HES)の収入データを HHP に移行中である。この中で、データ・エディティングとインピュテーション処理は長足の進歩を遂げたと記述している[30]。また、HES のエディティングについて、外れ値に優先順位をつけるため、選択的エディティングを採用すること、結果に重大な影響のあるものは手作業でエディットし、他はインピュテーションを行うことなどを勧告している。収入票の欠測値、支出票のいくつかの変数、世帯としては回答している所の個人ベースの家計簿の欠測に対しては、マッチング変数の多くがカテゴリー変数であるため、主に CANCEIS に実装された最近隣ドナーインピュテーション法が使用され、いくつかの支出の変数のインピュテーションには Banff も用いられると記述している。

５．まとめ

以上の文献調査から得られる傾向をまとめると、対外的に公表する論文では先進事例が紹介されることが多いことから来るバイアスはあるとしても、現時点で行われているインピュテーションはホット・デック法、比率代入法などの伝統的な手法が多く、回帰代入法や平均値代入法を用いている国（機関）も多い一方で、高度な分布モデルに基づいた手法は米国やオーストリアなどで検討されている。また、行政情報の統計作成への利用の進展に伴って、そのためのインピュテーションの必要性や、行政情報による統計のインピュテーションについての報告も見られる。欧州統計システム(ESS)、オーストリア、フィンランド、ニュージーランドでは手法の標準化が試みられている。また、インピュテーションのロバスト化手法については、外れ値を除去してからのホット・デック法の適用、平均値に代わる中央値やトリム平均の使用、winsorization の採用などへの言及が見られる。

参考文献

[1] 宇都宮浄人・園田桂子(2001)「『全国企業短期経済観測調査』における欠測値補完の検討」,日本銀行ワーキング・ペーパー。

[2] 高橋将宜・伊藤孝之(2014)「様々な多重代入法アルゴリズムの比較～大規模経済系データを用いた分析～」『統計研究彙報』第 71 号、総務省統計研修所。

[3] 高橋将宜(2015)「公的統計における欠測値補定の研究：多重代入法と単一代入法」『製表技術参考資料 30』独立行政法人統計センター（平成 27 年 6 月）。

[4] 高橋将宜(2016)「政府統計データのエディティングに関する国際的動向：選択的エディティングの理論とソフトウェア」『製表技術参考資料 31』独立行政法人統計センター（平成 28 年 3 月）。

[30] それらの処理における機械学習の導入についての議論を Zabala (2015)b が報告している。主な内容は、最近隣法を中心に用いるのは従来と同じだが、収入源を 3 つのモジュール（仕事、政府所得移転、投資）に分けること、インピュテーションのクラスを定めるのに決定木を用いることなどである。

[5] 高橋将宜(2017)「諸外国の公的統計における欠測値対処法 ：集計値ベースと公開型ミクロデータの代入法」『「統計学」創刊 60 周年記念事業特集：政府統計ミクロデータの作成・提供における方法的展望』経済統計学会。
[6] 国連(2000)『統計データ・エディティングに関する用語集（対訳）』ヨーロッパ統計家会議方法論資料、独立行政法人統計センター製表関連国際用語集 No.1(平成 17 年 1 月)。
[7] 野村総合研究所（2013)『統計データの補完推計に関する調査報告書』（平成 25 年 3 月）。
[8] 平川貴大・鳩貝淳一郎(2012)「ビジネスサーベイにおける欠測値補完の検討――全国企業短期経済観測調査（短観）のケース――」日本銀行ワーキング・ペーパー。
[9] Andics, A.. and Horváth, G. (2015), “Data collection optimization – first attempt”, Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Budapest, September 2015.
[10] Bailar, B.A. and Shapiro, G.M. (1981), “Survey research at the bureau of the census”, in Current Topics in Survey Sampling (D.Krewski, R.Platek, and J.N.K.Rao, Eds) , Academic Press 1981.
[11] Bankier, M. (2011), “Imputing Numeric and Qualitative Variables Simultaneously”, A Technical Report Detailing the Methodology of CANCEIS, Internal report, Statistics Canada.
[12] Bankier, M., Lachance, M., and Poirier, P. (1999), “A Generic Implementation of the Nearest Neighbour Imputation Method”, Proceedings of the Survey Research Methods Section, American Statistical Association, 548-553.
[13] Barboza, W., Miller, D. and Cruze, N. (2014), “Assessing the Impact of a New Imputation Methodology for the Agricultural Resource Management Survey”, Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Paris, April 2014.
[14] Chen, Bor-Chung (2007), “CANCEIS Experiments of Edit and Imputation with 2006 Census Test Data”, STUDY SERIES (Computing #2007-1).
[15] Cox V. (2014), “Migration of a large survey onto a micro-economic platform", Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Paris, April 2014.
[16] CSRM (2016)a, “Annual Report of the Center for Statistical Research and Methodology, Research and Methodology Directorate, Fiscal Year 2015”, U.S. Department of Commerce, Economics and Statistics Administration, U.S. CENSUS BUREAU.
[17] CSRM (2016)b, “CENTER FOR STATISTICAL RESEARCH AND METHODOLOGY, FY 2016 FIRST & SECOND QUARTERS REPORT –October 2015 through March 2016 –”, U.S. Department of Commerce, Economics and Statistics Administration, U.S. CENSUS BUREAU.
[18] CSRM (2016)c, “CENTER FOR STATISTICAL RESEARCH AND METHODOLOGY, FY 2016 THIRD QUARTER REPORT –April 2016 through June 2016 –”, U.S. Department of Commerce, Economics and Statistics Administration, U.S. CENSUS BUREAU.
[19] Da Silva, D. N. and Zhang, L. C. (2014), “ESTIMATION OF THE VARIANCE DUE TO

IMPUTATION FOR THE 2011 UK CENSUS", Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Paris, April 2014.

[20] Deroyon, T., and Gros, E. (2015), "Output editing based on winsorization in the French SBS multisource system Esane", Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Budapest, September 2015.

[21] Di Zio M., Fursova N., and Quensel-von Kalben L., Ten Bosch O. (2015), "Towards a generic approach to validation: the ValiDat foundation project", Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Budapest, September 2015.

[22] Di Zio, M., Guarnera, U., and Varriale, R. (2014), "Imputation with multi-source data: the case of Italian Structural Business Statistics", Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Paris, April 2014.

[23] Erdman, C., and Nagaraja, C. H. (2010), "Imputation Procedures for American Community Survey Group Quarters Small Area Estimation", RESEARCH REPORT SERIES (Statistics #2010-09).

[24] Fellegi, I.P. and Holt, D. (1976), "A Systematic Approach to Automatic Edit and Imputation", Journal of the American Statistical Association, March 1976, Volume 71, No. 353, 17-35.

[25] Foss, A. H., and Seierstad, A. (2015), "Editing and evaluation of statistics based on administrative microdata – example by Norway", Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Budapest, September 2015.

[26] Froehlich, M. (2015), "Flash Estimates for Short Term Indicators - Data cleaning with X12 Arima", Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Budapest, September 2015.

[27] Garcia, M., Erdman, C., and Klemens, B. (2014), "Multiple Imputation Methods for Imputing Earnings in the Survey of Income and Program Participation", Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Paris, April 2014.

[28] Garcia, M., Morris, D.S., and Diamond, L.K. (2015), "Implementation of Ratio Imputation and Sequential Regression Multivariate Imputation on Economic Census Products", Proceedings of the Joint Statistical Meetings.

[29] Gaughan, C. (2015), "An assessment of the feasibility of editing and imputing administrative tax return data to provide a substitute for survey data", Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Budapest, September 2015.

[30] Giessing, S. and Walsdorfer, K. (2015), "The ValiDat Foundation Project: Survey on the Different Approaches to Validation Applied Within the ESS", Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Budapest, September 2015.

[31] Godbout, S., Beaucage, Y., and Turmelle, C. (2011), "Achieving Quality and Efficiency Using a Top-Down Approach in the Canadian Integrated Business Statistics Program", Work Session on

Statistical Data Editing, United Nations Economic Commission for Europe, Ljubljana, May 2011.

[32] Guertin, L., Bureau, M., and Morel, J. (2014), “Editing the 2011 Census data with CANCEIS and options considered for 2016”, Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Paris, April 2014.

[33] Hoogland, J. (2015), “Changes in macro-editing and score functions for Dutch STS”, Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Budapest, September 2015.

[34] Kaminski, M. W. and Kapani, V. (2009), “Empirical Evaluation of Imputation Methods on Quarterly Census of Employment and Wages (QCEW) Data”, Joint Statistical Meetings 2009.

[35] Kilchmann, D. (2015), “Analysis of the data preparation process of the structural survey of the federal population census”, Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Budapest, September 2015.

[36] Kim, J. K., Fuller, W. A., and Bell, W. R. (2008),”Variance Estimation for Nearest Neighbor Imputation for U.S. Census Long Form Data”, RESEARCH REPORT SERIES (Statistics #2008-13).

[37] Kowarik, A., Templ. M., and Schopfhauser, D. (2014), “NEW FEATURES OF VIM - VISUALIZATION AND IMPUTATION OF MISSING VALUES”, Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Paris, April 2014.

[38] Lohr, S and Prasad, N.G.N. (2003). “Small Area Estimation With Auxiliary Survey Data,” Canadian Journal of Statistics, Vol.31, No.4., pp. 383-396.

[39] Luzi, O. (2015), “Managing changes in the E&I strategy of the Italian SBS”, Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Budapest, September 2015.

[40] MEMOBUST (2014), “Handbook on Methodology of Modern Business Statistics“.

[41] Miller, D. and Young, L. J. (2015), “Imputation at the National Agricultural Statistics Service”, Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Budapest, September 2015.

[42] Norris, Sherri (2003), Analysis of Item Nonresponse Rates for the 100 Percent Housing and Population Items from Census 2000, Census 2000 Evaluation B.1.b, September 23, 2003.

[43] Ollila, P. (2015), “Editing process and its quality regarding design and production phases using process metadata and calculation modules”, Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Budapest, September 2015.

[44] Osburn, J. G. (2013), “Wage Imputation in the OES Survey: A Model-Assisted Approach Incorporating Data from the Quarterly Census of Employment and Wages”, Joint Statistical Meetings 2013.

[45] Pannekoek, J., van der Loo, M., and van den Broek, B. (2014), “IMPLEMENTATION AND EVALUATION OF AUTOMATIC EDITING”, Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Paris, April 2014.

[46] Pichiorri T., Ichim D., Ferraro M.L., and Guarnera U. (2015), “Model-based selective editing procedures for agricultural price indices”, Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Budapest, September 2015.

[47] Rastogi, S., Fernandez, L., Noon, J., Zapata E,, and Bhaskar, R., (2014), “Exploring Administrative Records Use for Race and Hispanic Origin Item Non-Response”, Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Paris, April 2014.

[48] Revilla, P. (2015), “Developing a theoretical framework for selective editing based on modelling and optimization”, Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Budapest, September 2015.

[49] Rothhaas, C., Lestina, F., and Hill, J.M. (2012), “2010 Decennial Census: Item Nonresponse and Imputation Assessment Report”, 2010 CENSUS PLANNING MEMORANDA SERIES No. 173, February 8, 2012.

[50] Saint-Pierre, E. (2015), “Redefining roles and responsibilities in a new harmonized statistical production process: opportunities and challenges”, Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Budapest, September 2015.

[51] Sands, R. D. (2013), “Simplified Census Edit and Imputation Based on Statistical Principles”, Contributed Paper, 2013 Joint Statistical Meetings.

[52] Scholtus, S. and Willenborg, L. (2014), “Editing and Imputation in the Memobust Handbook on Methodology of Modern Business Statistics”, Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Paris, April 2014.

[53] Seljak, R. (2014), “Metadata driven application for data processing – from local toward global solution”, Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Paris, April 2014.

[54] Slud, E. V. (2015), “Impact of Mode-Based Imputation on ACS Estimates”, 2015 AMERICAN COMMUNITY SURVEY RESEARCH AND EVALUATION REPORT MEMORANDUM SERIES #ACS15-RER-07.

[55] Spies, L. (2015), “Evaluation of Census 2011 survey estimates”, Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Budapest, September 2015.

[56] Spies, L., Schmiedel, L., Schmidt, K., (2014), “Simulating Multiple Imputation of Water Consumption in the German Agricultural Census 2010”, Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Paris, April 2014.

[57] Templ, M., Kowarik, A., and Filzmoser, T. (2011) “Iterative stepwise regression imputation

using standard and robust methods", Comput Stat Data Anal, 55(10):27932806, 2011.

[58] Thibaudeau, Y. (1999), "Model Explicit Item Imputation for Demographic Categories for Census 2000", Research Report # RR-99-02.

[59] Thibaudeau, Y. (2002), "Model Explicit Item Imputation for Demographic Categories", Survey Methodology, 28(2), 135-143.

[60] Thibaudeau, Y., Gottschalck, A., and Palumbo, T. (2006), "The Predictive-Mean Method of Imputation for Preserving Coupling Between Assets and Liabilities", RESEARCH REPORT SERIES (Computing #2006-1).

[61] Thibaudeau, Y., Shao, J., and Mulrow, J. (2007), "A Study of Basic Calibration Estimators in Presence of Nonresponse", Proceedings of the American Statistical Association.

[62] Thibaudeau, Y., Williams, T., and Krenzke, T. (1997), "MULTIVARIATE ITEM IMPUTATION FOR THE 2000 CENSUS SHORT FORM", Proceedings for the Section on Survey Research Methods, American Statistical Association.

[63] Winkler, W. E. (2008), "General Methods and Algorithms for Modeling and Imputing Discrete Data Under a Variety of Constraints", RESEARCH REPORT SERIES (Statistics #2008-08).

[64] Zabala, F. (2015)a, "Getting commitment to a new editing strategy", Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Budapest, September 2015.

[65] Zabala, F. (2015)b, "Let the data speak: Machine learning methods for data editing and imputation", Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, Budapest, September 2015.

[66] Zajac, Kevin J. (2003), "Analysis of Imputation Rates for the 100 Percent Person and Housing Unit Data Items from Census 2000", Census 2000 Evaluation B.1.b, September 25, 2003.

参考　諸外国の公的統計における主な欠測値補完方法

米国

調査部局	調査名	補完手法	備考
センサス局	人口センサス	Assignment（項目非回答につき回答された項目から推測） Allocation （Assignmentができない場合、他の世帯員又は近隣世帯から代入、具体的には何種類かのホット･デック法による） Substitution（全世帯が欠測した場合、近隣世帯により代替） 人種とヒスパニック系に関する項目非回答については、過去のセンサスその他の調査からの回答を利用しての割当ても行っている。	離散データについて､対数線形モデルに基づいた手法を検討中。 2010 年センサスより。
	地域社会調査(ACS)	人口センサスの long form の手法を踏襲し、Assignment 及び Allocation を適用。	多変量モデルに基づいた手法を検討中。
	Survey of Income and Program Participation (SIPP)	資産と負債について、単変量のスタックを用いたホット・デック法（後入れ先出し法） 主に収入について、Assignment 又は（確定的）ホット・デック法（人口特性の一致する	多変量のホット･デック法(Joint Hot-Deck インピュテーション)及び予測平均法(Predictive Mean インピュテーション)を検討中。 ランダム化されたホット･デック法及びモデルに基づいた順次回帰多変量法 (sequential regression

		観測値を代入)。	multivariate imputation, SRMI)による多重代入法を検討中。
	研究開発調査(The Survey of Research and Development in Industry: SRDI)（2007 年時点）	Ad-hoc なホット・デック法（前回或いはそれ以前の値を全体の増加率によって補正したもので代替)。	調査は” Business R&D and Innovation Survey”に引き継がれた。
	経済センサス	分野ごとに異なり、鉱工業では生産物の合計と総売上高との差を「特定されず」としてインピュテーションを行わず、建設業では最近隣法によるホット・デック、サービス業では総売上高からの比率により行っている。	調査内容（分類）が次回から大幅に変わり、生産物と産業のリンクがなくなるため、比率(Ratio)によるインピュテーションと SRMI を検討している。
	月次卸売調査(Monthly Wholesale Trade Survey, MWTS)		母集団のデータを満たすためのインピュテーション手法を改良。ランダムフォレストを条件モデルとする連鎖方程式による多変量インピュテーション(MICE: Multivariate インピュテーション by Chained Equations)の適用について研究中。

労働統計局(BLS)	職業雇用統計(OES)	賃金について、同じ時期、都市統計地域(MSA)、産業、企業規模のドナーを定め、十分な回答がない場合は統合(collapse)したのちに分布の平均を代入（ホット・デック法の一種)。	多変量の線形モデルを用いたインピュテーションを検討。モデルに用いる共分散の推計においては、雇用・賃金プログラム四半期センサス(QCEW)から得た企業の賃金についての補助データも使用。 共分散の推計は外れ値に敏感であるため、winsorizeしたロバストな手法によって推計。
	雇用・賃金プログラム四半期センサス(QCEW)	雇用と賃金について、対象企業の過去一年のトレンドを延長。	セル全体或いは最近隣の企業のトレンドを用いる方法及び除外すべき異常値の判定方法を検討。
農務省(USDA)全米農業統計サービス(NASS)	農業センサス	以下の順番で行う。 (1) 決定論理表(DLT)の評価に基づいて得られる値（合計が欠測している場合など） (2) 他の調査、前回センサスなど以前の調査から得られる値 (3) ドナーを用いたインピュテーション	

	農業資源経営調査(ARMS)第3フェーズ	地域、農場の種類、売上げ規模などでグループを作り、impute する値がバイアスを受けないようにグループごとに上下双方の外れ値を除いた後の乗率のない平均を欠測値に代入（各グループには最小で10個の観測値が必要で、満たない場合はなるべく均一性を保持するべく定められた優先順位でグループ統合(collapse)を行う）。	繰返し順次回帰(ISR)による手法を検討し、2014 年調査で採用（多変量同時分布を一連の線形モデルに分解し、回帰する手法。一連の線形モデルとインピュテーションのパラメータの推定値はマルコフ連鎖モンテカルロ法を用いた反復によって得られる）。

欧州

調査部局	調査名	補完手法	備考
総説（UNECE会合参加 23 機関）		回帰代入法、比率代入法、平均値代入法、ホット・デック法をそれぞれ 95%以上の統計機関で使用。事業所・企業統計で比率代入法、世帯統計でホット･デック法の使用が多い。	高橋(2017)による。
欧州統計システム(ESS)		Memobustプロジェクトで作成されたガイドラインでは、演繹的インピュテーション、モデルによるインピュテーション、ドナーによるインピュテーション、時系列データのインピュテーション、Little and Su 法を記述。	ESS は欧州統計局と加盟各国の協力体制で、統計品質向上のためにさまざまなプロジェクトを実行している。

英国統計局	人口センサス	CANCEIS によるドナーを用いたインピュテーション。	
	企業統計における税務データの利用	カンパニーの欠測値については中央値による補完 企業がまるごと欠測している場合は、報告単位に対し、i) 層内中央値による補完 ii) 層内トリム平均による補完 iii) ビジネスレジスタから得られる補助変数による比率による補完を検討中	現在検討中。
ドイツ連邦統計局	2011 年人口センサス	コールド・デック法、演繹（deductive）インピュテーション、最近隣法の結合。	レジスタ･ベースの人口センサスの抽出調査部分のインピュテーション。人口レジスタ等によるコールド･デック法、変数間の関係性に基づいた演繹的インピュテーション、単変数の最近隣インピュテーションの順に適用。インピュテーションの評価を多重代入法によって行う。
	2010 年農業センサス	ホット・デック法、予測平均値マッチング。	インピュテーションによる結果の変動係数への影響を多重代入法により評価。
オーストリア統計局	一般	VIM により、ホット・デック法、モデルに基づいた繰り返しロバスト法、k 最近隣法、回帰法などが扱える。	VIM は、欠測値の可視化とインピュテーションのための R パッケージで、最新バージョンは GUI 機能を備える。

	産業及び建築の短期統計	専門家による税務報告や社会保険会計などを用いたインピュテーション。	X12-Arima を用いた外れ値検出及びインピュテーションの自動化を試行、評価している。
オランダ統計局	企業統計の自動エディティング	演繹的インピュテーション、最近隣インピュテーション（保育所） 演繹的インピュテーション、回帰によるインピュテーション（農作物卸売）	
	短期統計 (Short Term Statistics, STS)システム（経済統計システムの一部）	比率によるインピュテーション	比率の算出対象とする補助変数は、前期値、前年同期値、就業者数の３つを候補として、ルールに基づいて決める。外れ値のウェイトは低く置いて算出。
フランス国立統計経済研究所(INSEE)	年次企業統計細密化システム (ESANE)	ユニット無回答に対しては外れ値を winsorization によって処理した後に乗数調整(calibration)を行う。 財務データや社会保険の欠測値に対してはインピュテーションを行う。財務データでは主に以前の年のデータに基づいている。	
スイス連邦統計局	人口センサス	確定的インピュテーション及び最近隣インピュテーション。	行政データによる人口センサスの抽出調査部分(抽出率約 3%) 最近隣インピュテーションを行う SAS マクロを開発

フィンランド統計局	品質評価の一環として。インピュテーション法のモニタリング	平均値によるインピュテーションから中央値によるインピュテーションへの変更。	
ノルウェー統計局	行政データに基づく統計	数値変数の欠測値は前回の値でインピュテーションを行う。 カテゴリー変数は専門家による手作業のエディティング。	
イタリア統計局	ビジネス構造統計	最近隣補定法、予測平均法。	報告中心の統計から行政データの徹底的な使用と限定的な標本調査へ移行。従来は行政記録によるインピュテーション、セル内の最近隣ドナーによるインピュテーションを行っていた。
	農業物価指数	予測値によるインピュテーション又は手作業による修正。	予測値はモデルに応じて指数の 1 ヶ月前の値、12 ヶ月前の値、前年に対する比で伸ばした値のいずれか。
スペイン統計局	産業短期指数	REGARIMA モデルによる外れ値検出及びインピュテーション。	
ハンガリー統計局		回帰法、最近隣法など。	妥当性の検証を予定。
スロヴェニア統計局		ドナーを用いたホット・デック法。	データ処理近代化システムの例示。 セルには最低 10 のデータが必要、それで代入できなかった場合は対象地域を拡大し、セル内の最低のデータ数を 5 とする。

その他

調査部局	調査名	補完手法	備考
カナダ統計局	2011 年人口センサス	ドナーによるインピュテーション(CANCEIS)。旧来の Fellegi-Holt 法に代えて、最近隣法を使用。	履歴や回帰によるインピュテーションは間接的に取り扱い可能。
	労働力調査	項目非回答については、ホット・デック法による代入、前月からの横置き(Carry-Forward)、推定による代入を行う。 ユニット非回答については、回答履歴の状況に応じて、ホット･デック法あるいはウェイト付けによる補完を行っている。	前月の回答を含め、ホット･デックのドナーとする Longitudinal Hot-deck という手法を導入。 2005 年 1 月以前はユニット非回答の横置きも行っていたが､月次変化を過小に評価すること､クロスセクションでの Hot-deck についても過大に評価することから、現在の Longitudinal Hot-deck を用いることとした。
ニュージーランド統計局	年次企業調査(Annual Enterprise Survey, AES)	比率インピュテーション、履歴によるインピュテーション、平均値インピュテーション。	ミクロ経済プラットフォーム (micro-economic platform: MEP) に最初に統合されたものとしての例示。手法の優先順位は各変数についてさまざまな補助変数との相関や利用可能性によって決定される。

	世帯経済調査 (Household Economic Survey, HES)	最近隣ドナーインピュテーション法。	CANCEIS に実装されている。いくつかの支出の変数のインピュテーションには Banff も使用。 外れ値に優先順位をつけるため選択的エディティングを採用。