リサーチペーパー第 44 号　　Research Paper No.44

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

Study on Basic Literatures and Current Trends of 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

Nobuyuki Sakashita
SRTI Senior Researcher for Statistical Training

令和元年８月
August 2019

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：令和元年７月 30 日

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

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。

このため、本リサーチペーパーでは今までの調査を受け、引用されることが多く、欠測値補完についての基本的文献と思われる書籍の収集・調査及び各国の最新動向の調査を行った。

その結果、リサーチペーパー第 43 号で見たように共通の体系的了解が必ずしも存在しているわけではなく、また出版物には多重代入法などの理論的な書籍が多い傾向があるものの、1980 年以降、議論の根拠となるような多数の文献が提供されていることが分かった。また、近年の動向としては、個別のインピュテーション手法よりも、行政情報の利用に伴うインピュテーションの必要性の発生への対応、総合システムの開発、商用あるいは他国で開発したシステムの利用に関する話題が多くなっており、国によっては機械学習についての検討も進んで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、精度、ホット・デック法

Study on Basic Literatures and 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 concern, succeeding surveys so far, we collected and reviewed literatures that are often cited and considered to be basic documents about the treatment of missing values, and surveyed the latest trends of foreign countries.

As a result, although there is not necessarily a common systematic understanding as we saw in Research Paper No.43 and publications tend to be theoretical rather than practical, with preference for, say, multiple imputation, those literatures that give bases for discussion are provided since 1980. As for trends in recent years, topics dealing with the necessity of imputation accompanying the use of administrative information, development of integrated system, use of commercial systems or those developed in other countries are increasing, and studies on machine learning are also in progress in several countries.

Keywords: Data Editing, Imputation of Missing Data, Accuracy, Hot-deck Method

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。

2018 年度は、各国の最新動向や欠測値補完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・調査を行った。その結果、実務においてはなおホット・デック法、比率代入法などの伝統的な手法が採用されることが多く、先進的な手法を検討したうえでホット・デック法を採用した例もあること、カナダやオーストリアでは、システムの改良が続けられており、他国での採用例も見られること、手法の体系については、必ずしも共通の了解が存在しているわけではないが、1980 年以降豊富な文献の蓄積があり、特に 90 年代末からは、統計を作成している機関自ら発信することも増え、意見の交換が行われていることが分かった（坂下 (2018)）。

本リサーチペーパーでは、今までの調査から、引用されることが多く、欠測値補完についての基本的な文献と思われる書籍の収集・調査及び各国の最新動向の調査を行った。

以下はその結果であり、その構成は、１．が基本的文献、２．がアメリカ合衆国及びカナダの動向、３．欧州その他の動向、４．がまとめとなっている。

１．基本的文献

今までの調査から、引用されることが多く、欠測値補完についての基本的文献と思われる以下の書籍を入手し、内容を吟味した。このうち、もっとも古い Rubin (1987) は統計図書館（総務省統計局）の蔵書、他は統計研究研修所で新規に購入したものである。

(1) Rubin, D.B. (1987), Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, New York.
(2) Little, R.J.A. and Rubin, D. B. (2002), Statistical Analysis with Missing Data (second edition). John Wiley & Sons, New York.
(3) De Waal, T., Pannekoek, J., and Scholtus, S. (2011), Handbook of Statistical Data Editing and Imputation. John Wiley & Sons, New York.
(4) Van Buuren, S. (2018), Flexible Imputation of Missing Data (second edition), Chapman & Hall/CRC, Boca Raton, Florida.

（Rubin (1987)）

Rubin (1987) は、その書名のとおり多重代入法を論じたものだが、導入部において具体的な調査に即して欠測値の状況を記述し、それぞれへの対応を述べた上で、インピュテーションが持つべき性質を挙げ、現行の無回答に対する対応ではこれらを満たすことができないと記述している。

その後は、統計学的な背景、根底にあるベイズ理論、ランダム化に基づく評価、ignorable な無回答の処理、nonignorable な無回答の処理を論じており、途中で欠測パターンが単調 (monotone) な例として人口動態調査 (Current Population Survey, CPS) の社会保障項目を例としてシミュレーションを行っているが、基本的には理論的な学習書である。附録１に概要を記す。

（Little and Rubin (2002)）

Little and Rubin (2002) は、欠測値を含むデータの分析に主眼を置いた書である。導入部では、欠測パターン、欠測メカニズムについて論じており、今日欠測値について語る際の標準的な記述となっている。その後、最小二乗法を初めとして、ウェイト補正を含む完全ケース分析 (complete case analysis)、利用可能ケース分析 (available case analysis)、単一代入法、インピュテーションに伴う不確実性の評価（直接推定、ブートストラップ、ジャックナイフなどのリサンプリング手法、多重代入法など）などの欠測値を含むデータの分析法を挙げている。附録２に概要を記す。

（De Waal et al. (2011)）

De Waal et al. (2011) は、国家統計局 (NSI) などの公的統計機関で行われるデータ・エンティティングとインピュテーションを主題としており、今回のリサーチしたものの中では国内業務の参考とするのに短期的には最も適した内容と思われる。構成は、エディティングとインピュテーションの一般的なプロセス、演繹的修正手法、連続データの Fellegi-Holt 法などによる自動エディティング、カテゴリー・データへの拡張（Fellegi-Holt 法、NIM 法など）、選択的エディティング、インピュテーション、エディット制約下のインピュテーション、インピュテーション済データの修正と、実務に沿って進んでいる。インピュテーションについては、無回答の発生、補助変数の選定、外れ値などの一般的な事項について述べた後、一般化したモデルにより詳細に分類している。附録３に概要を記す。

（Van Buuren (2018)）

Van Buuren (2018) は、今回対象とした中では最新のものである。著者は不完全データの統計分析を研究するユトレヒト大学教授で、多重代入法の代表的なアルゴリズムである MICE の開発者である。このような背景もあって本書は MICE を用いた欠測値への対処を目的とし、多重代入法とそれを用いたデータ分析を中心に記述している。インピュテーション自体の手法については、軽く触れている程度であるが、数式を最小限にとどめる一方で R によるプログラミングの説明が豊富であり、多重代入法の入門には適していると思われる。附録４に概要を記す。

各文献の概要について附録１から附録４に記す。

２．アメリカ合衆国及びカナダの動向

米国では、センサス局に属する The Center for Statistical Research & Methodology (CSRM) などで、欠測値補完に関するさまざまな研究が続けられている。また、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。

（センサス局）

米国の人口センサスについては、CSRM の業務報告（CSRM (2018)）によると、モデルに基づいたセンサス・カバレッジの小地域推定についての研究を行い、センサスのカバレッジ測定 (Census Coverage Measurement, CCM) についての助言と共同作業を行っている。具体的には、無回答の修正とインピュテーションの研究計画の実装と解釈について実務部署を援助している。また、坂下 (2018) にも記した 2020 年センサスに向けての「良い」行政情報の研究も継続中で、用いる歳入庁 (IRS) データの拡張や、行政情報を用いることができるか、接触を試みるべきかの判定のため、従前の手法に加え距離関数を用いる検討を行い、Keller et al. (2018) として Journal of Official Statistics 誌に投稿されている[1]。

月次卸売調査 (Monthly Wholesale Trade Survey, MWTS) についての、より現実的な卸売業の母集団の合成と推計値の改善の研究も継続中である。現在の研究内容は坂下 (2017) に記したが、2018 年度も合成された MWTS のデータを繰り返し引き出すことに使える現実的で人工的な母集団の開発を継続し、いくつかの変数に観察されるゼロ値の比率を修正する処理を開発している[2]。なお、月次小売販売速報 (Advance Monthly Retail Trade Survey, MARTS) について、統計方法論に関する連邦委員会 (Federal Committee on Statistical Methodology, FCSM) の会合で、センサス局から傾向スコアマッチングを加味したホット・デック法が報告されている（Thompson et al. (2018)、プレゼン資料のみ）。

CSRM のエディット及びインピュテーションに関する一般的な研究としては、人口及び経済プロジェクトをサポートするコンピューターシステムを開発し、プロトタイプシステムの実行、手法の検討を行うプロジェクトが進行中である。具体的には、対数正規モデルをセルの中に該当するデータがない場合の推計を行うように拡張し、人口動態調査 (Current Population Survey, CPS) のグロスのフローの推定に適用して結果を発表する予定である。また、シンシナティ大学統計学部の Hang Kim 助教授が開発したエディットとインピュテーションのソフトウェアを用いた経済センサスの合成データを作成するプロジェクトが進行中である。

（経済分析局 (BEA)）

---

[1] CSRM (2018) に記載されている論文名が本文と書誌情報で異なっているが、前者は 2018 Joint Statistical Meetings における発表時のタイトルであり、論文名としては後者が正しい。

[2] 坂下 (2017) に記したとおり詳細不明だが、価格が収集されていない対象をインピュテーションで補うことにより、母集団を擬似的に再構成しているものと思われる。

BEA は坂下 (2018) にも記したカナダ統計局のエディットとインピュテーションのシステム Banff を用いたインピュテーション自動化の検討をさらに進め、「米国における海外直接投資の年次調査」(Annual Survey of Foreign Direct Investment in the United States, BE-15) のデータを用いて手法の優先順位とインピュテーションに用いるデータについてのシミュレーションを行った。その結果、いくつかの項目については、従来のドナーを用いる手法を推定値による手法に優先させる順序を入れ替えた方が良いこと、また当該年次だけではなく過去のデータも加えてインピュテーションを行った方が良いとの結果を得て、上記 FCSM の会合でその評価を報告している（Terrie (2018)）。

（労働統計局(BLS)）

労働統計局 (Bureau of Labor Statistics, BLS) は、国連の欧州経済委員会 (United Nations Economic Commission for Europe, UNECE) の会合における手法を分類整理する要望に応じて、文献のテキスト分析を用いて統計手法の分類を作成することを提案し、UNECE に提出された論文を 2 種類のクラスター分析によって分類した結果を UNECE 会合及び合同統計会議 (Joint Statistical Meetings, JSM)に報告している（Martinez and Savitsky (2018)a, Martinez and Savitsky (2018)b)。

（農業統計サービス (NASS)）

全米農業統計サービス (National Agricultural Statistics Service, NASS) は、商用オフザシェルフ（Commercial-off-the-Shelf, COTS: 市販の既製品）ソフトウェアの利用を評価し、UNECE 会合に報告している（Miller (2018)）。その記述によると、農業センサス (COA) 及び NASS の大きな統計調査では内製したエディットとインピュテーションのシステム PRISM を用い、他の多くの調査では Blaise の対話型システムを用いている。現状では農業センサスでは坂下 (2018) に記した IVEware の他に SAS のプロシージャを用いている一方で、農業センサス以外の PRISM を用いた調査はモジュラー化し、インピュテーションの過程が独立しているため IVEware の導入が急速に進んでいる。他方、Blaise を用いた調査では修正は手作業で行われており、自動化は進んでいない[3]。今後の方向性としては、可能なところでは COTS を利用すること、手作業によるエディットとインピュテーションの削減、モジュラー化の推進などが示される一方、現行プロシージャを Banff で置き換えることはしないとしている。また、坂下 (2017) の時点で繰返し順次回帰 (iterative sequential regression, ISR) を採用した農業資源経営調査 (Agricultural Resource Management Survey, ARMS) については、COTS の利用に置き換えることを計画している。

（カナダ）

[3] 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 (Computer-Assisted Personal Interviewing, CAPI)のシステムである。

Gray (2018) によるとカナダでは、坂下 (2018) にも記したユーザーの意見を受けた上での Banff の改良が続けられており、その中のプロジェクトとして (A) 方法論の拡大、(B) モジュール化・自動化されたデータ・エディティングツールとしての Banff の改良、(C) 一般化された評価ツールの開発 がある。

また、カナダでは、CANCEIS を用いたエディットとインピュテーションを評価する枠組みの作成が試みられている。Stelmack (2018) では、CANCEIS の基本的なパラメータ設定（k-最近隣法によるドナーの候補となるデータの数、補助変数と距離関数の選定、同じドナーを使用できる回数の上限など）を解説した後、さまざまに変えてシミュレーションを行い、将来へ課題として Banff との結合、現在はモンテカルロ法に頼っている枠組の一般化を挙げている。

３．欧州その他の動向

（英国）

Leather et al. (2018) によると、英国の 2011 年人口センサスのインピュテーションは CANCEIS に搭載されたドナーに基づいた手法[4]であったが、2021 年センサスは主にオンラインで行われ、行政情報で補われることから、国家統計局 (ONS) は行政情報を利用したインピュテーションの検討を行っている。そこで引用されている Blum (2006) では、インピュテーションの補助として行政情報を用いるには、コールド・デック、モデルのスペック改善、継続的な品質保証などの手段があるが、このうちコールド・デック（行政情報の値をセンサスに代入する方法）にはセンサスのデータを外部の値で置き換えることにより不整合が生じるリスクがあるとしているため、ONS の今回の検討では、既存のホット・デック法の中で行政情報を補助的に用いることに焦点を当て、年齢のインピュテーションにおいて各レコードに「行政データの年齢」変数をリンクしたものを補助情報として用いたものを、従来のホット・デック法と比較し、著しい改善が見られたとしている。

また、Davies (2018) によると、英国では付加価値税などの行政データによって調査を置き換える[5]際のエラー発見手法の検討が行われている。発見されたエラーは比率インピュテーションにより修正されている。

（イタリア）

Di Cecco et al. (2018) は、イタリアの統計に使用する個人レジスタで欠測値の多い最終学歴データのマス・インピュテーション[6]について報告している。その中で、対数正規モデルが伝統的なホット・デック法より優れていること、次に労働力統計と行政情報の結合を行い、人口センサスにつなげたいとしている。

---

[4] 坂下 (2017) に記述あり。
[5] VAT 等データの利用については坂下信之 (2017) 参照。
[6] この論文では、時期の関係で得られていない情報を一括して予測することを指して「マス・インピュテーション」と呼んでいる。De Waal et al. (2011) 参照。

また、Luzi et al. (2018) によるとイタリア統計局 (ISTAT) では、経済分野での統計作成プロセスを支援する環境のインフラ SINTESI (Integrated Business Statistics System) の開発が決定され、短期経済統計 (Short-term business statistics, STS) のデータ・エディティングとインピュテーション、特に選択的エディティングから適用することとなった。ここで行う選択的エディティングは混合モデルに基づき、Istat で開発された Selemix R パッケージ[7]に実装されている。

（オーストリア）

Gussenbauer et al. (2018) によると、オーストリア統計局は、自ら開発したインピュテーションのための R パッケージ VIM[8] の中の k-最近隣法について、距離を計算する複数の手法（ウェイトなし、ランダムフォレスト法によりウェイトを生成、予測値を距離計算に含める等）によるインピュテーション及びランダムフォレスト法のみによるインピュテーションを世帯レジスタのデータを用いて比較検討し、順序付き変数（教育）ではランダムフォレスト法がウェイトなし及びウェイト付きの k-最近隣法より正確で、準連続変数（収入など）ではその逆などの結果を得ているが、より明確な結果を得るためには多数の変数についてより多くのテストが必要だとしている。

（ドイツ）

Spies and Lange (2018) によると、ドイツ連邦統計局では、統計システムのデジタル化や政策的な背景から、公的統計における処理の自動化のため、人工知能と機械学習の導入を検討している。現状では、多くのインピュテーションはまだ手作業で行われているため、この検討において、異なった機械学習の手法と対応する R や Python のソフトウェア、さらにスタンフォード大学やワーテルロー大学で開発されたソフトウェアをテストし、それらの長短を理解することを目指している。初期段階の検討として、マイクロセンサス（1% 抽出世帯調査）で無回答が多く MCAR でない（回答率が調査方法により異なる）出産経験に関する質問においてロジット・モデル、予測平均値マッチングなどの従来の手法が十分満足できる結果になっていないとして、機械学習法の最初の適用を試みている。

（オランダ）

Pannekoek (2018) によると、オランダでは、De Waal et al. (2011)（第 1 章及び附録 3 参照）にも記述され、経済統計に良く用いられている比率によるインピュテーションにおける外れ値の影響を是正するためのロバスト統計と機械学習（ブースティング）の利用について検討している。前者では、外れ値に影響されやすいという短所を是正するため、構造経済統計の変数の比率によるインピュテーションについて、日本でも行われているウェイト付けに

[7] 高橋 (2013) 参照。
[8] 坂下 (2017)、坂下 (2018) 参照。

よるロバスト化を行うものである。Huber 及び Tukey のウェイト[9]を従来の手法、比率の中央値を用いる手法と比較して、ウェイト付けによるロバスト化を行うと誤差が改善するが、もともとの誤差が小さい場合はその改善は小さく、中央値を用いると誤差が増大するとの結果を得ている。後者は、通常の比率法がただ一つの補助変数と比率を用いるのに対し、複数の補助変数や比率を用いるもので、その結合を gradient boosting （勾配ブースティング）によって行う。補助変数の追加は、通常の比率法（合計同士の比率）がモデル

$$\frac{y_i}{\sqrt{x_i}} = \theta\sqrt{x_i} + \varepsilon_i$$

に最小二乗法を適用したものであるため、そのままでは複数の補助変数に拡張できないことから、実測値と補助変数から推計した値との残差にさらに比率法を適用して、他の補助変数を加えるものである。また、比率の追加は、ロバストな比率に、補助変数の大きさにより変わる比率を加えることにより当てはまりを良くするものである。この論文では、最大 4 つの要素で比率推定することにより、誤差が改善することを示している。

（ノルウェー）

Jentoft and Zhang (2018) によると、ノルウェーでは、データ・エディティングとインピュテーションの自動化のために、教師あり機械学習を発展させて、二段階機械学習 (two-phase machine learning) と二重機械学習 (double machine learning) を労働力調査のデータを用いて試行している。二段階機械学習では、部分休業 (partial absence) の変数について、第一段階でランダムフォレスト法によるインピュテーション、第二段階では別のランダムフォレスト法によるその不確実性の評価を行い、聞き取り調査や再調査などの他の手法の法が良いかを判断する。二重機械学習では、第一段階で 5 種類の変数について別々にランダムフォレスト法を適用し、第二段階でこれらを結合する。

（スイス）

Vallée and Tillé (2018) によると、スイスでは、対策の少ない「スイス・チーズ型」欠測パターン[10]にバランスした K-最近隣法で対処する手法を検討している。バランスした K-最近隣法とは、Hasler and Tillé (2016)[11] で提案されたもので、アイテム無回答に k-最近隣法を適用する際に、ドナーの候補となる k 個の観測値それぞれの採用確率を、欠測していない変量をドナーで置き換えた場合に、その変量の合計の期待値が元の観測値を用いた場合の合計値と一致するように調整 (calibration) するものである。Vallée and Tillé (2018) では、これにより、ランダムなインピュテーションに伴う合計値の変動を押さえることがで

[9] 和田 (2012)、和田・野呂 (2019) 参照。
[10] 単変量や単調（モノトーン）でない一般的な欠測パターンのこと。Andridge and Little (2010)、坂下 (2018) 参照。欠測パターンについては高橋・渡辺 (2017) 参照。
[11] Vallée and Tillé (2018) では小文字の k を別の所で添え字で使用しているため、大文字で「K-最近隣」(K-nearest neighbor) と記しているが、元文献では k は小文字。

きるとしている。

４．まとめ

基本的文献の調査については、2018 年度に見たようにインピュテーションの体系についての共通の了解が必ずしも存在しているわけではなく、また出版物には多重代入法などの理論的な書籍が多い傾向があるものの、1980 年以降、議論の根拠となるような文献が提供されていることが分かる。

また、近年の動向としては、個別のインピュテーション手法そのものの開発というよりは、行政情報の利用に伴うインピュテーションの必要性の発生への対応、総合システムの開発、商用あるいは他国で開発したシステムの利用に関する話題が多くなっている。国によっては機械学習についての話題もあるが、インピュテーションそのものを機械学習で行うというより、インピュテーション手法の選択に機械学習を用いることが多いようである。

附録１　Rubin (1987) の概要

著者（当時の肩書き）

Donald B. Rubin, ハーバード大学統計学部 (Department of Statistics, Harvard University)

出版社

John Wiley & Sons

内容

序文に「何らかの理論的背景を有する応用調査統計家向け」とあり、その書名のとおり多重代入法を論じたものだが、第１章は導入で当時の状況を紹介した上で理論を概観し、第２章は統計的背景、第３章が根拠となるベイズ理論、第４章はランダム化に基づく評価、第５章は無視可能 (ignorable) な無回答の処理、第６章は無視不可能 (nonignorable) な無回答の処理となっている。

本書の思想的な部分は第１章にほぼ書かれている。最初に、調査における無回答の発生事例として、(1) 教育試験サービス (Educational Testing Service, ETS) が 1971 年に行った学校調査、(2) センサス局の人口動態調査 (Current Population Survey, CPS)、(3) 1970 年センサスの汎用データベースの職業分類、(4) 1982 年にボストンの復員軍人援護局が行った飲酒習慣の調査 を挙げ、それぞれの発生状況、処理方法（(1) はリストワイズ除去、平均値など、(2) は比較的複雑なホット・デック法[12]、(3) はホット・デック法及びロジスティック回帰[13] (4) は無回答者に対するフォローアップ調査）を紹介した上で、インピュテーションが持つべき性質として、(1) 完全データに対する標準的な手法を用いることができること、(2) 無回答者と回答者の違いを反映した推定値や標本の減少を反映した標準誤差をもたらす妥当な推計を行うことができること、(3) 無回答に対するさまざまなモデルに対する推計の感度を反映することが求められるとし、現行の無回答に対する対応ではこれらを満たすことができないとしている。

その後、単一代入法と多重代入法の概説があり、単一代入法の長所として、(1) （欠測値を含んだままのデータを分析するのは専門的なプログラムを要するのに対し）分析に完全

[12] ただし、本書はこの手法に懐疑的で、そのことが多重代入法を提唱する動機ともなっている。
[13] これについても、インピュートした値を真の値であるかのように扱うため、変動を過小評価する問題があるとしている。

データに対する標準的な手法が使えること、(2) 一般利用のデータベースの視点からは、作成者がその知見をもって一度だけインピュテーションを行えば良いことを挙げつつ、短所として、ただ一つの値を代入するために実際の値の標本変動も代入モデルによる追加的な変動も反映されないことを挙げている。具体的な単一代入法としては、単純無作為標本の場合の最良予測値によるインピュテーションは変動を過小評価し、これを避けるための分布から値を得るインピュテーションはデータの分布を近似的に保つが変動は正しい値にならないとしている。多重代入法は、単一代入法の長所を保ちつつ、(1) 分布を保つインピュテーションにおいて推定の効率性が増す、(2) 追加的な変動も反映した適切な推計が完全データに対する推計を結合することによって得られる、(3) 完全データに対する手法を単純に繰り返すことで、推計の安定性を調べることができる 長所があるとしている。一方、多重代入法の短所として、(1) 作成により多くの労力を要する、(2) データセットの補完により多くの容量を必要とする、(3) 分析に要する労力が単一代入法より多い ことを挙げ、その後、初歩的な例を用いて、代入内分散と代入間分散の説明を行っている。

第２章は共変量と結果変数、標本に含まれるか、回答されたかの指標などの統計的な基礎から条件付確率分布などの概念や統計量の推計を解説し、ベイズ理論への橋渡しとなっている。後段では、ベイズ的な事後確率と従来型の無作為標本の仮定に基づく統計量の範囲の推定を比較し、漸近的に等しいとしている。第３章ではさらに進んで、多重代入法は理論的な根拠と作成・分析の手法をベイズ的な視点から得ているとして、根拠となるベイズ理論を解説し、統計量の事前分布、事後分布という観点から、複数回のインピュテーションによる範囲の推定について論じている。

これに対し、第４章では、多重代入法を従来型の無作為標本に基づく推定から評価し、無回答がない場合に妥当な完全データの推計法であり、インピュテーション手法が適切であれば、多重代入法は無作為標本に基づく推計の視点からも妥当な推定をもたらすとしている。具体的には、ランダムなホット・デック法を繰り返すホット・デック法の多重代入バージョンは、代入間の変動を過小評価し、妥当な方法ではないが、正規分布あるいはブートストラップ法に基づくベイズ的な多重代入法は妥当な結果をもたらすとしている。また、インピュテーションに伴うパラメータの不安定性を取り入れて修正したホット・デック法も妥当とされている。

具体的な多重代入法のプロセスは第５章及び第６章に記されている。第５章では、ignorable[14] な欠測への対処法として修正されたホット･デック法や最小二条法による回帰を取り上げ、単変量の場合、多変量だが欠測パターンが単調 (monotone) な場合について細かく説明した上で、欠測パターンが単調な場合の実例として、CPS[15] の社会保障項目を例とし

[14] 「無視できる」という意味だが、欠測値についての用語としては、欠測メカニズムが Non Ignorable でないこと、すなわち観測された値を用いて対処できる(MCAR 又は MAR)ことを意味している。
[15] ここで、CPS についてのかなり詳しい説明がある。

てシミュレーションを行い、多重代入法の優位性を示している。また、単調でない欠測パターンについてはいくつかの対処法を示している。第 6 章では、nonignorable な欠測について、ベイズ統計的には回答／無回答について仮定することで、ignorable な場合に準じて扱えることを示し、第 1 章で紹介した学校調査、CPS 及び飲酒習慣の調査を用いてシミュレーションを行っている[16]。

[16] 飲酒習慣の調査については、フォローアップの結果を活用している。

附録２　Little and Rubin (2002) の概要[17]

著者（当時の肩書き）

Roderick J. A. Little, ミシガン大学公衆衛生大学院 (School of Public Health, University of Michigan)

Donald B. Rubin, ハーバード大学統計学部 (Department of Statistics, Harvard University)

出版社

John Wiley & Sons

内容

本書は全 3 部 15 章から成るが、主眼は欠測値を含むデータの分析にある。古典的な欠測値の補完法については第 1 部（第 1〜5 章）にまとめられており、今日における欠測値についての標準的な記述に従って、欠測パターン、欠測メカニズムから説き起こしている。第 1 章では、欠測値への対処法について、(1)（欠測を含むデータを無視して）完全データのみで分析する方法、(2) 欠測データの存在に応じてウェイトを補正する方法、(3) インピュテーションによる方法、(4) モデルに基づいて尤度又は事後分布を想定する方法 に分類し、第 2 章では結果変数が一つの場合の欠測値の最小二乗法による推定、第 3 章ではウェイト補正、完全ケース分析 (complete case analysis)、利用可能ケース分析 (available case analysis)について述べ、第 4 章で単一代入法、第 5 章でインピュテーションに伴う不確実性の評価を論じている。

単一代入法は、大きく明示的モデル (explicit modelling) と非明示的モデル (implicit modelling) に分類され、それぞれの例として、明示的モデルでは、(a) 平均によるインピュテーション、(b) 回帰によるインピュテーション、(c) 回帰による確率的なインピュテーション（(b)に残差を加えたもの）、非明示的モデルでは、(d) ホット・デック法、(e) 代替 (substitution)、(f) コールド・デック法（ここでは、「前回の値のような外部の情報源からの規定値で置き換える」こととされている）、(g) 複合的な手法（回帰で得られた予測値に任意の観測値の残差部分を加えるなど）を挙げ、さまざまな手法の性質を論じたのち、インピュテーションは一般に、(a) 無回答によるバイアスを軽減し精度を向上させ、欠測値と観測値

[17] 今回入手したのは第 2 版だが、2019 年中に第 3 版が出る予定である。

の関係を保存するため、観察された値に基づき (conditional)、(b) 欠測値相互の関係を保つために多変量であり、(c) 幅広い推定値の有効性のために平均よりは予測分布から得られる値であるべきとしている。

また、インピュテーションに伴う不確実性の評価方法としては、(1) 欠測を含むデータに対する明示的な計算式で直接推定する方法、(2) 補完済みデータから適切な標準誤差を計算できるようにインピュテーションを修正する方法、(3) 不完全データからブートストラップ、ジャックナイフなどのリサンプリング手法により再抽出したデータに対してインピュテーションと分析を繰り返し適用する方法、(4) インピュテーションによる不確実性を反映するために複数回インピュテーションを行ったデータを作成する方法を示している。多重代入法などを挙げ、第 2 部以降につないでいる。それぞれの評価方法のうち、(1) についてはごく単純な仮定の下以外では明示的な推計値が得られるか疑わしいとしてこれ以上論ぜず、(2) と (3) については第 1 部で簡単に紹介している。(4) は多重代入法であり、第 1 部で紹介した後、続く部で詳細に論じている。

附録３　De Waal et al. (2011) の概要

著者（当時の肩書き）

Ton de Waal, Jeroen Pannekoek, Sander Scholtus, オランダ統計局 (Statistics Netherlands)

出版社

John Wiley & Sons

内容

本書は、国家統計局 (NSI) などの公的統計機関で行われる実務をターゲットとしており、今回調査した中では最も目的に合った内容と思われる。全 11 章から成り、導入となる第 1 章では実務書らしく、統計のプロセスを整理した上で、データ・エンティティングとインピュテーションについて、両者は密接に関連しており、どこで前者が終わって後者が始まるか厳密には区別できないものの、この本の大部分では別々のプロセスとして扱うと述べている。その後は、エラーの種類、無回答の種類（欠測メカニズム)、エンティティングとインピュテーションの基本手法とプロセスについて述べている。

続く章は、第 2 章がシステマティックなエラー（位取り、正負の誤りなど）の演繹的修正、第 3 章が連続データの自動エディティング（Fellegi-Holt 法など)、第 4、5 章が自動エディティングのカテゴリー・データや整数データへの拡張（Fellegi-Holt 法、NIM 法など)、第 6 章が選択的エディティングに当てられている。インピュテーションについては、基本的に第 7 章以降で扱われているが、最近隣インピュテーション（NIM 法）については、それが Fellegi-Holt 法に対する批判と改善として出現した経緯があり、エディティングとインピュテーション手法が分離できない性格[18]のため、第 4 章に初出している。

第 7 章では、インピュテーションを行うに当たっての一般的な事項（母集団の分割、ウェイト補正、マス・インピュテーション、補助変数の選択、外れ値）について触れた後、代表的なインピュテーション手法である回帰によるインピュテーション、比率によるインピュテーション、平均値によるインピュテーション、ホット・デック法（ランダム、シーケンシャル、最近隣）について、オランダの世帯調査（回帰)、公共図書館の調査（非線形回帰)、

[18] Fellegi-Holt 法では、修正すべきデータの変量を特定する作業（「エラーの局所化」(localization)と言う。）において、修正する変量の数を最小に抑えるが、これが必ずしも現実的な結果とならないことから、その改善策として NIM 法が提案された。これはホット・デック法を前提としているため、エディティングとインピュテーションにまたがる手法といえる。

経済構造調査（比率、平均値）、オートメーション調査（平均値）、住宅需要調査、所得構造調査、農業及び園芸の機械化調査（以上ホット・デック法）などの実例を用いて解説し、紹介している。その後で、Kalton and Kasprzyk (1986)[19] に範を取った一般化モデルにより詳細な分類を行い、全体を大きく確定的インピュテーションと確率的なインピュテーションに分けている。その上で、

- 確定的インピュテーションとしてはプロキシ（代替値）によるインピュテーション（前回の値の横置き、夫の国籍に妻の国籍を代入するなど）、バランスを用いた演繹的インピュテーション、平均によるインピュテーション、グループ内の平均によるインピュテーション、比率によるインピュテーション、グループ内の比率によるインピュテーション、回帰によるインピュテーション（残差なし）、最近隣インピュテーション、予測平均値マッチング、シーケンシャルなホット・デック法
- 確率的なインピュテーションとしては回帰によるインピュテーション（パラメータから得られる残差あり）、回帰によるインピュテーション（観測値の残差）、ランダムなホット・デック法、グループ内のランダムなホット・デック法

を挙げ、手法による性質の違いを論じている。時系列データのインピュテーションについては、他と異なった特徴があるとして節を改めて論じられており、オランダ経済調査、建設産業調査、欧州共同体世帯パネル調査(European Community Household Panel Survey, ECHP)[20] の例が示されている。

以後は第 8 章が多変量のインピュテーション、第 9 章が制約下のインピュテーション。第 10 章がインピュートされたデータの修正、第 11 章がアプリケーションとなっている。第 9 章及び第 10 章は、インピュテーションの理論面からはあまり注目されてこなかったものの、実務的には興味を持たれる課題であり、坂下 (2018) で紹介した De Waal (2017) もこの問題を扱ったサーベイである。

[19] 坂下 (2018) において紹介した。
[20] De Waal (2000)、坂下 (2018) を参照。

附録４　Van Buuren (2018)[21] の概要

著者（当時の肩書き）

Stef van Buuren, オランダ応用科学研究機構 (the Netherlands Organization for Applied Scientific Research TNO), ユトレヒト大学不完全データ統計分析教授

出版社

Chapman and Hall/CRC

内容

今回対象とした中では最新のものである。著者は不完全データの統計分析を専門とするユトレヒト大学教授で、多重代入法の代表的なアルゴリズムである MICE の開発者である。このような背景もあって本書は MICE を用いた欠測値への対処を目的とし、多重代入法とそれを用いたデータ分析を中心に記述している。インピュテーション自体の手法については軽く触れている程度であるが、数式を最小限にとどめる一方で R によるプログラミングの説明が豊富であり、多重代入法の入門には適していると思われる。

全体は 4 部 13 章から成り、第 1 章において、基本的な欠測値への対処法について欠測メカニズムなどと共に簡潔にまとめている。具体的には、リストワイズ消去法（完全ケース分析）、ペアワイズ消去法（利用可能ケース分析）、平均値によるインピュテーション、回帰によるインピュテーション、回帰による確率的インピュテーション、横置き (carry forward) （前回値及びベースライン）などを紹介した上で、多重代入法の手順、用いる理由、実例を簡単に紹介している。多重代入法について詳しく紹介している第 2 章では、冒頭でインピュテーションの歴史を概観し、19 世紀には土地や住宅の帰属収入(imputed income) の概念が用いられ、統計に関する文献では 1957 年のセンサス局の資料に「データを埋める」意味で現れる一方で、欠測値への対処法として広く用いられるようになったのは 1983 年に開催された ”Panel on Incomplete Data” 以降であること、多重代入法については、Rubin (1987) が足場を築いて以降文献が指数関数的に増え続けていることを指摘している。第 2 章の残りの部分では不完全データの概念、多重代入法の意義等について詳しく説明し、多重代入法が必要な場合を論じている。第 1 部の以後の章では、単変量の場合、

[21] 今回入手したのは第 2 版だが、現時点で多く引用されているのは 2012 年の初版である。

多変量の場合、補完されたデータの分析方法と続き、以下の部でさらに深めていくが、多重代入法によるデータ分析の詳細については本稿の範囲を超えるのでここでは省略する。

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] 高橋将宜 (2013)「公諸外国における最新のデータエディティング事情～混淆正規分布モデルによる多変量外れ値検出法の検証～」『製表技術参考資料 30』独立行政法人統計センター（平成 25 年 8 月）。

[4] 高橋将宜・渡辺美智子 (2017) 「欠測データ処理 R による単一代入法と多重代入法」統計学 one point 5、共立出版、2017 年 12 月。

[5] 和田かず美 (2012)「多変量外れ値の検出～繰返し加重最小二乗 (IRLS) 法による欠測値の補定方法～」」、『統計研究彙報』第 69 号、総務省統計研修所。

[6] 和田かず美・野呂竜夫 (2019) 「ロバスト回帰推定へのウェイト関数や残差尺度の影響について」、『統計研究彙報』第 76 号、総務省統計研究研修所。

[7] Andridge, R. R. and Little, R. J. A. (2010), “A Review of Hot Deck Imputation for Survey Nonresponse”, International Statistical Review 78, pp. 40-64.

[8] Blum, O. (2006), “Evaluation of editing and imputation supported by administrative records”, Statistical Data Editing Volume 3: Impact on Data Quality, UNECE, pp300-309.

[9] CSRM (2018), “Annual Report of the Center for Statistical Research and Methodology, Research and Methodology Directorate, Fiscal Year 2018”, U.S. Department of Commerce, Economics and Statistics Administration, U.S. CENSUS BUREAU.

[10] Davies, K. (2018), “Investigating methods of efficient detection of errors in VAT data”, Workshop on Statistical Data Editing, United Nations Economic Commission for Europe, Neuchâtel, September 2018.

[11] De Waal, T. (2000), “A Brief Overview of Imputation Methods Applied at Statistics Netherlands”, Report, Statistics Netherlands.

[12] De Waal, T. (2017), “Imputation Methods Satisfying Constraints”, Work Session on Statistical Data Editing, United Nations Economic Commission for Europe, The Hague, April 2017.

[13] De Waal, T., Pannekoek J., and Scholtus, S. (2011), “Handbook of Statistical Data Editing and Imputation”, John Wiley & Sons, New York.

[14] 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.

[15] Gray, D. (2018), “The Evolution of Banff in the Context of Modernization”, Workshop on Statistical Data Editing, United Nations Economic Commission for Europe, Neuchâtel,

September 2018.

[16] Gussenbauer, J., Kowarik, A., and Meraner. A. (2018), “Using Random Forest to Improve Single Imputation Precision - A Case Study”, Workshop on Statistical Data Editing, United Nations Economic Commission for Europe, Neuchâtel, September 2018.

[17] Hasler, C. and Tillé, Y. (2016), “Balanced k-nearest neighbor imputation”. Statistics, Volume 50, Issue 6, 2016.

[18] 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.

[19] Kalton, G. and Kasprzyk, D. (1982), “Imputing for Missing Survey Responces”, Proceedings of the Section on Survey Research Methods, American Statistical Association, 22-31.

[20] 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, Vol.34, No.3, 2018. pp. 599–624.

[21] Leather, F., Sharp, K., and Rogers, S. (2018), “Towards an integrated census-administrative data approach to item-level imputation for the 2021 UK Census”, Workshop on Statistical Data Editing, United Nations Economic Commission for Europe, Neuchâtel, September 2018.

[22] Little, R. J. A. and Rubin, D. B. (2002), “Statistical Analysis with Missing Data (second edition)”, John Wiley & Sons, New York.

[23] Luzi, O., Manzari, A., Pichiorri, T., Rocci, F., Rosati, S., and Varriale, R. (2018), “Selective editing in the Integrated Business Statistics System (SINTESI)”, Workshop on Statistical Data Editing, United Nations Economic Commission for Europe, Neuchâtel, September 2018.

[24] Martinez, W. L. and Savitsky, T. (2018)a, “Towards a Taxonomy of Statistical Data Editing Methods”, Workshop on Statistical Data Editing, United Nations Economic Commission for Europe, Neuchâtel, September 2018.

[25] Martinez, W. L. and Savitsky, T. (2018)b, “Creating a Taxonomy of Statistical Methods using Text Analysis”, Proceedings of the Joint Statistical Meetings.

[26] Miller, D. (2018), “Using Commercial-off-the-Shelf (COTS) Software at the National Agricultural Statistics Service”, Workshop on Statistical Data Editing, United Nations Economic Commission for Europe, Neuchâtel, September 2018.

[27] Pannekoek, J. (2018), “Improvements of ratio-imputation using robust statistics and machine learning-techniques”, Workshop on Statistical Data Editing, United Nations Economic Commission for Europe, Neuchâtel, September 2018.

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

[29] Spies, L. and Lange, K. (2018), “Implementation of artificial intelligence and machine learning

methods within the Federal Statistical Office of Germany", Workshop on Statistical Data Editing, United Nations Economic Commission for Europe, Neuchâtel, September 2018.

[30] 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.

[31] Terrie, L. (2018), "Assessing the Automated Imputation of Missing and Erroneous Survey Data: A Simulation-Based Approach", Proceedings of the 2018 Federal Committee on Statistical Methodology (FCSM) Research and Policy Conference.

[32] Thompson, K. J., Bechtel, L., and Czaplicki, N. (2018), "Evaluating Hot Deck with Propensity Score Matching For the Advance Monthly Retail Trade Survey (MARTS)", 2018 Federal Committee on Statistical Methodology (FCSM) Research and Policy Conference.

[33] Vallée, A. and Tillé, Y. (2018), "Balanced Imputation for Swiss Cheese Nonresponse", Workshop on Statistical Data Editing, United Nations Economic Commission for Europe, Neuchâtel, September 2018.

[34] Van Buuren, S. (2018), "Flexible Imputation of Missing Data (second edition)", Chapman & Hall/CRC, Boca Raton, Florida.