リサーチペーパー第 48 号　　Research Paper No.48

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）

Current Trends of Research in Imputation Methods for Statistical Surveys in Foreign Countries (the United States and the Netherlands)

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 2 年 9 月
September 2020

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２年９月 17 日

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）

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。

本リサーチペーパーでは、各国で継続して研究されてきたテーマに焦点を当て、テーマを特定した継続的な研究の見られるアメリカ合衆国及びオランダについて重点的に論文を収集した。また、一般用ミクロデータを用いて、諸外国では頻繁に用いられているが我が国では適用例の少ないホット・デック法の数値シミュレーションを試みた。

その結果、少なくとも今回対象にした２カ国については、インピュテーションシステム、ベイズモデル、制約条件下のインピュテーションなどの継続的な課題を設定し、豊富な過去の蓄積の上で新たな検討を行っていることが分かった。また、我が国の統計データにおける欠測値への対処にホット・デック法を適用することは可能であると考えられるが、ドナーの選び方などの具体的な方法は、適用する調査に応じて子細かつ実務的に検討する必要がある。

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、ホット・デック法

Current Trends of Research in Imputation Methods for Statistical Surveys in Foreign Countries (the United States and the Netherlands)

SAKASHITA Nobuyuki

Abstract

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

In this paper, we focused on the themes that have been continuously researched in foreign countries, and collected documents of the United States and the Netherlands, where continuous research on specific themes have been held. We also exercised numerical simulation of the hot-deck method, which is frequently used in other countries but rare in Japan, using General-Use (Synthetic) Microdata, provided by the National Statistics Center.

As a result, we found that, at least for those two countries, new researches are held on continuing themes such as imputation system, Bayesian model, and imputation under constraints, based on abundant storage of the past. Although hot-deck method seems to be feasible for application in imputation of the missing data in our statistical surveys, specific methods such as selection of donors need to be carefully and practically examined in conformity with the survey applied.

Keywords: Data Editing, Imputation of Missing Data, 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。

そのため、平成 28 年以降、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。

令和元年からの調査でも新たな情報の収集を行う予定だったが、米国センサス局の報告書が遅れ新たな内容も乏しいことや、欧州諸国が中心となって行われる予定だった欧州統計家会議 (Conference of European Statisticians, CES) の会合が新型コロナウィルス COVID-19 蔓延の余波で延期されたことなどにより、新たな情報が少なかったため、各国で継続して研究されてきたテーマに焦点を当て、テーマを特定した継続的な研究の見られるアメリカ合衆国及びオランダについて重点的に論文を収集した。また、一般用ミクロデータを用いて、諸外国では頻繁に用いられているが我が国では適用例の少ないホット・デック法の数値シミュレーションを試みた。

以下はその結果であり、その構成は、１．がアメリカ合衆国、２．が欧州（オランダ統計局)、３．が一般用ミクロデータを用いたホット・デック法の数値シミュレーション、４．がまとめとなっている。

１．アメリカ合衆国

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。

（インピュテーションの歴史と変遷）

CSRM は毎年何編かの研究レポートを公表しており、その中の一つ、2018 年に公表された Winkler (2018) は、統計に行政情報を用いるために必要なレコード・リンケージとエディティング／インピュテーションを扱っている。論文自体は行政情報の利用に関するものであるが、その中で初期の手書きシステムから Fellegi-Holt 法、さらに統計モデルに基づいた近年の手法に至るデータ・エディティング／インピュテーションの歴史が論じられている。

それによると、初期のシステムは数百から数千の ”if-then-else” ルールにより構成されていたが、そのような古典的な手法は、指示の論理的な誤りやコーディングの誤り、マッチング・ルールの効率性などの困難、同時分布が保存されることの保証がないなどの問題があったのに対し、Fellegi and Holt (1976) で発表された Fellegi-Holt 法は、修正の容易な表形式で記述され、論理的な一貫性がデータ入力前にチェックでき、修正後のデータはエディット・ルールを満たすなどの長所があったとしている。

さらに、Fellegi-Holt 法に基づいたシステムは、5 カ国（カナダ、スペイン、イタリア、オランダ、アメリカ合衆国）の統計機関が独自に開発しているとして、米センサス局による連続的な経済データ用の SPEER、離散データ用の DISCRETE、カナダ統計局による連続データ用の GEIS[1]、離散データ用の CANCEIS、オランダ統計局による離散及び連続データ双方に用いられるハイブリッドシステムを挙げている[2]。

その上で、Winkler (2008)[3] を引用して、エディティングに Fellegi-Holt 法、インピュテーションにホット・デック法を採用しても、同時分布を効果的に保存するのは困難であり、Little and Rubin (2002) の第 13 章などで示された新しいインピュテーションの考えに明らかな長所があるとしている。ここで引用されている Winkler (2008) は、従来型のホット・デック法が多変量の同時分布を維持しないことから、多変量対数線形モデルの優位性を述べたものであり、Little and Rubin (2002) の第 13 章[4]は欠測値のあるデータの分割表の対数線形モデルによる推計を論じたものである。

なお、本論文は、Chun and Larson (2020) の一章として構想したものと見られるが、当該書籍はまだ刊行されていない[5]。また、2011 年に開催された合同統計会議 (Joint Statistical Meetings)の会合 JSM2011 で同じ題の報告（Winkler (2011)）がされており、その内容を拡充したものとなっている。

（米センサス局及び各国のエディティング／インピュテーションシステム）

上記論文の著者である William E. Winkler 自身、長らくセンサス局で SPEER 及び DISCRETE の開発に携わっており、関連する文献を多数著している。Winkler and Draper (1996) は、SPEER は、1980 年代前半に製造業調査のために開発されたシステムを源流としているとし、その中で、手法と機能を解説する文献として Greenberg and Surdi (1984) や Greenberg and Petkunas (1990) を挙げているが、前者では SPEER の名称はまだ現れず ”core edit”と呼ばれており、後者で「SPEER システムは、センサス局で開発された多目的のエディット及びインピュテーションシステムであり、比率エディティングのもとで連続データに対して用いられる」と紹介されている。

SPEER は 1990 年代後半に改訂され、新たなシステムについて Draper and Winkler (1996) が解説している。新たな SPEER システムは旧システムが 2 つのモジュールから成っていたのに対し、補助的な 2 つを含む 4 つのモジュールから成り、比率インピュテーションのほかに部分的なバランス調整（内訳の和を合計に一致させるなど）を行い、アルゴリズムが簡潔で高速であるなどの特徴があるとされている。

---

[1] 現在数カ国で採用・検討されている Banff の前身に当たる（野村総合研究所（2013））。
[2] 他の国については Kim et al. (2014) の所で述べる。
[3] 坂下(2017) 参照。
[4] “Models for Partially Classified Contingency Tables, Ignoring the Missing Data Mechanism”, Chapter 13, Little and Rubin (2002). 坂下 (2019) に記したように本書は 2019 年に第 3 版が出版されているが、同名の章は継続して存在している。
[5] 当初 2020 年中の刊行予定だったところ、2021 年に延期された模様。

離散データ用の DISCRETE も Winkler によって開発されたもので、解説した文献として Winkler (1997) がある。また、Winkler (2018) は、DISCRETE は、「SPEER システムよりもはるかに難しい一般的な整数プログラミング手法を使用している」としている。Winkler and Chen (2002)、Winkler (2003)、Winkler (2008) は、DISCRETE に関連する離散データにおける制約条件下の補完[6]について論じている。

なお、Kim et al. (2014) によると、Fellegi-Holt 法に基づいて開発された各国のシステムは、米センサス局の SPEER（Draper and Winkler (1997)）、カナダ統計局の GEIS（Whitridge and Kovar (1990)）、スペイン国家統計局の DIA（Garcia-Rubio and Villan (1990)）、オランダ統計局の CherryPi[7]（De Waal (2000)）、イタリア国家統計局の SCIA（Manzari (2004) : 本体未入手、Abstract のみ）である。これらは、いずれもエディット・ルールによってエラーのあるデータを発見し（狭義のエディティング）、データの中の修正の必要な値を最小の変更で済むように特定し（エラーの局所化 localization）、値を変更する（インピュテーション）という Fellegi-Holt 法の手順に則っているが、スペイン国家統計局の DIA はシステマティック・エラー（位取りの誤りなど一定の法則に従って発生するエラー）に対してはインピュテーション・ルールによる確定的（deterministic）なインピュテーションを行う混合システムとなっている[8]。

（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）

Kim et al. (2014) は、シンシナティ大学の Hang J. Kim 博士のグループが経済センサスを対象に「線形制約のある欠測値又はエラーのあるデータの多重代入法のための完全なベイジアンによるジョイントモデリング・アプローチ」を研究していると記しているが、CSRM (2019)[9]では、経済センサスのミクロデータのために Kim et al. (2015) 、Kim et al. (2018)[10]に基づいたシステムの開発を進めているとしている。このシステムは、ノンパラメトリックなベイズ法によりデータのエディット、（多重）インピュテーション、合成を行うもので、

---

[6] これは De Waal (2017) が指摘するように特に厄介な問題である。エディティング・ルールに反する値を修正する際に、その値が直接反していたルールにのみ注目すると、修正後の値が別のルールに反してしまうことが起こり得る。これを避けるためには、直接示されているルール（明示的 (explicit) ルール）から導かれるが当初は記述されていなかったルール（暗黙の (implicit) ルール）を予め抽出する必要があり、ルールが増えるにつれてこの作業が膨大になる。Winkler and Chen (2002) によると、この問題が CANCEIS でも採用されている NIM 法（Bankier (1991) : 未入手、Bankier et al. (1997) : 未入手、 Bankier et al. (2000)）開発の背景となっている。Bankier et al. (2000) によると Fellegi-Holt 法と NIM 法の主な違いは、Fellegi-Holt 法では、まず修正する値の数が最小になるように修正箇所を決定するのに対し、NIM 法ではまず最近隣のドナーを探し、次にこれらのドナーに基づいて変化が最小になるインピュテーションを決定することである。

[7] 文献により CherryPi 又は CherryPie と表記に揺れが見られるが、De Waal のオリジナル論文では CherryPi である。

[8] Fellegi-Holt 法ではインピュテーションに独自のルールを設けず、インピュテーションのルールはエディット・ルールから自動的に導かれる。

[9] 実際の発刊は遅れて 2020 年 2 月末だったが、過去の報告との整合性のため略称を CSRM (2019)とする。

[10] CSRM (2019) では”Kim et al. (2017)”と表記しているが、書誌情報がなく、該当する論文が見当たらない。内容的に Kim et al. (2018) のことと思われる。

真のデータと同じ結果表を得る合成データを作成することにより、一般的に利用できるミクロデータを提供することを目指すとされている。

現時点で経済センサスのインピュテーションはホット・デック法で行われており（坂下(2018)）、企業調査の合成データを扱っている CSRM の研究レポート Kim et al. (2019) で合成データは「もともと経済センサスを契機としている」として、その手法の基礎に Kim et al. (2018)を挙げているので、経済センサスのインピュテーションの研究から得られた知見を合成データの作成に活用しているものと思われる。

（その他の近況）

CSRM(2019) に記されたセンサス局の最近の動向としては、他に以下の報告がある。

(1) 坂下(2017)、坂下(2018)、坂下(2019) で報告した CSRM における月次卸売調査 (Monthly Wholesale Trade Survey, MWTS) のインピュテーションに関する研究は、2019 年度は手法の性能を評価するツールを解説する原稿に着手したが、スタッフ不足により中断したとのことである。
(2) 地域社会調査 (ACS) の人口に係わる項目のインピュテーションに行政記録を利用することの研究と評価を行っている。
(3) 1990 年代から 2000 年代初頭にかけて開発されたまま放置されていた汎用システム BigMatch の再評価、多次元分割表の下でのパラメータ空間の状況を追跡するように設計された階層的対数線形モデルの研究を行っている。
(4) 幅広い予測と外れ値修正が可能な多変量の季節調整と欠測値補完の新たなソフトウェア Ecce Signum を開発した。
(5) 欠測値の補完手法、モデルによる推定、小地域推計の評価を目的の一つとして、シミュレーションと統計モデルの研究を行っており、現在文書を作成中である。
(6) 元データが多変量正規分布をしている時に、「プラグイン・サンプリング」によって発生させた合成データによる推定手法の開発に取り組んでいる（Klein et al. (2019)）[11]。

２．欧州（オランダ統計局）

欧州ではオランダ統計局 (CBS) が継続的にデータ・エディティングやインピュテーションに関する研究・開発を行い、専門家会合、ディスカッション・ペーパー、学会誌などでの発表を行っている。古いものは 1990 年代まで遡ることができる[12]が、近年のものでは、Pannekoek et al. (2009)、Pannekoek (2009)、Hoogland et al. (2011)、Israëls et al. (2011)、Van der Loo et al. (2011)、De Waal et al. (2011b) 、Pannekoek et al. (2013)、De Waal et al. (2015)、De Waal and Coutinho (2017)、Daalmans (2017)、De Waal et al. (2018)、Scholtus (2018)などが注目され

[11] 「プラグイン・サンプリング」は、未知の母集団分布のパラメータに標本から得られた推定値を代入（プラグイン）したものからサンプリングする手法で、補完した値から推計を行う多くの取組のような多重代入法ではなく、単一代入に拠っている（Klein and Sinha (2015)）。
[12] 一部は坂下 (2018) に既述。

る。

(スループット・プログラム)

Pannekoek (2009) は、「スループット・プログラム」の名前のとおり、オランダ統計局の自動化されたエディティングとインピュテーションの過程について、費用対効果と品質の両面での改善を目的として、原データをエディティングにかける「入口」からインピュテーションを終えた「出口」までのプロセス全体の処理に焦点を当てて整理したディスカッション・ペーパーである。具体的には、この時点で進行している 6 つのプロジェクト (1) 検出可能な原因のあるエラーの（演繹的）修正、(2) ランダムな（＝系統的でない）エラーのモデルによる局所化（＝項目同士が矛盾している際、エラー箇所を特定すること)、(3) 複雑な経済データのインピュテーション、(4) 調整された（＝既知の合計値に一致する）インピュテーション、(5) 回答確率の推定によるインピュテーションの改善、(6) データと推計値の品質へのエディットとインピュテーションの影響の指標 のそれぞれについて解説している。(1) は、過程の最初に行うもので、これまでの所、符号のエラー、変数の取り違え (interchange) エラー、丸めエラーの 3 種について検出と修正のアルゴリズムの開発に成功し、他の種類のエラーに拡張する研究を進めている。(2) は、Fellegi-Holt 法では複数の解が存在することがあるため、モデルを用いるか項目にウェイトをつけることで、修正すべき項目を特定するものである。(3) は、当時のオランダ統計局による回帰によるインピュテーションが項目ごとに単変量で行うもので、エラーが多い場合にあまり正確でないと思われたため、多変量モデル又は順次回帰によるインピュテーションを研究したものである。(4) は、レジスタベースの統計を作成する場合に別に知られている合計値と結果を合わせる手法で、欠測値の問題をウェイト付けで解決しようとした場合に不整合が起こるため、インピュテーションによる方法を検討しているものである。(5)は、従来から知られている無回答への対処方法であるウェイト付けとインピュテーションに加えて、両者を結合した手法を検討するものである。(6) は、体系的エラーの発見、ルール設定の適切さの検討などを含めたプロセス管理、品質の検証を行うものである。

（演繹的インピュテーションの R パッケージ）

Van der Loo et al. (2011) は、オランダ統計局で開発した演繹的インピュテーションを行う R パッケージ deducorrect について解説している。このパッケージは、数値データ、カテゴリー・データの双方について演繹的インピュテーションが可能なものであり、与えられた制約条件の下で可能な値の空間を与える solSpace、具体的なインピュテーションを行う imputes、カテゴリー・データにおいて単一のインピュテーション値が定まる項目を特定する deductiveLevels などの関数から成る。

（データ・エディティングとインピュテーションの手法解説）

Hoogland et al. (2011) と Israëls et al. (2011) は、それぞれデータ・エディティングとインピュテーションの手法解説書である。いずれも坂下 (2019) でレビューした De Waal et al. (2011a) と同時期のものであるが、よりオランダ統計局の業務に特化しており、CherryPie やその親プログラムの SLICE に言及している（同様の部分で De Waal et al. (2011a) では各国のシステムを紹介）。また、オランダ統計局では、NIM 法を適用するために CANCEIS を用いているとのことである。

（データ・エディティング概論）

De Waal et al. (2011b) はデータ・エディティングに関する概論であるが、章立ては 1.前書き、2.不整合のチェック：エディットルール、3.自動エディティング、4.インピュテーション：欠測データとランダムなエラーの修正、5.選択的エディティング、6.データ・エディティング戦略、7.結語 となっており、一章がインピュテーションに割かれている。ここでも De Waal et al. (2011a) を「包括的な説明」として、当文献は「簡単な概要」と位置づけられている。一方で、結語では近年の状況を反映して、過去数十年の間に全エラーデータの手作業による修正から、大量の自動検出・修正と選択的な人手修正へと移行してきたこと、統計機関がデータを収集する方法が変化しており、異なる情報源による不整合への対応が必要であることが述べられている。

（自動エディティングと人手によるエディティング）

Pannekoek et al. (2013) は Journal of Official Statistics の選択的エディティング特集に寄稿されたもので、自動エディティングと人手による（選択的）エディティングの関係について、古典的な視点では重要なエディティングは人手によって行われるべきで、自動エディティングは影響の少ないものに限るべきであるとされていたが、むしろ人手によるエディティングを影響が大きく「かつ」十分な品質によって自動化できないものに限るべきとの視点で書かれている。ここでは、エディティングを一般的なシステマティック・エラーの修正、分野特有の修正ルール、エラーの局所化、欠測値やエラーとされた値のインピュテーション、整合性維持のための値の修正などの機能 (function) に分解して解説し、エラーを含むデータが各機能を通過するにつれてどのように変化するかを 2 つのデータ（オランダの保育園のデータと卸売業についての経済構造統計のデータ）から得た数値例を用いて図示している。結論としては、各機能の品質評価を開発する必要があること、今後の研究は自動化に向かない部分が重要になる（非定型情報の利用など）ことが述べられている。

（制約条件下のインピュテーション）

Pannekoek et al. (2009)、De Waal et al. (2015)[13]、De Waal and Coutinho (2017) はオランダ統

[13] Pannekoek et al. (2009)、De Waal et al. (2015) の２編は、ディスカッション・ペーパーとして刊行された後にそれぞれ”Annals of Applied Statistics” 及び”Journal of Survey Statistics and Methodology”に掲載されている。

計局が継続的に課題にしている制約条件下のインピュテーション[14]を扱っている。Pannekoek et al. (2009) は、インピュテーションを行った結果が、数値項目の線形制約、及び変量の合計値が既知の値に一致する制約を満たさない場合の対策として、2 種類の「修正された予測平均インピュテーション」とマルコフ連鎖モンテカルロ (MCMC) 法について、合成されたデータセット及び 2005 年のイスラエルの収入調査の数値例を用いて評価し、MCMC 法の結果は他の手法より悪いが、さらなる研究が必要としている。De Waal et al. (2015) は、同じく数値データの制約条件に応じた修正手法を扱うが、ホット・デック法を基本としている[15]。具体的には、制約条件を満たすドナーの選び方を、最近隣法とランダム・ホット・デック法で、それぞれウェイト付きの合計を保持するかウェイトなしで考えるかの計 4 種の手法について、経済構造統計の実データと極度に困難な場合を想定した合成データによって評価を行っている。その結果は、制約条件を満たすことはできたが、数値の統計的性質は必ずしも良くなく、特に極度に困難なデータについてはより高度なインピュテーション手法が必要であるとしている。各手法の中では、ウェイト付きの最近隣法が、中央値を除いて相対的に良い結果を与えており、中央値についても代入法により改善できる可能性があるとしている。また、変数間の相関関係や分散の維持が今後の課題であること、項目を順番に補完していく手法はすべての項目について最適にならない問題があるが、多変量の補完を同時に行うのはなかなか困難であることを指摘している。De Waal and Coutinho (2017) は、先行研究がまず統計分布を描いてからエディット規則により刈り込んでいたのに対し、まず規則で許された領域を頂点で表し、その領域に対して分布を描くことによって補完を行う試みで、パラメトリック及びホット・デック法に似たノンパラメトリックな手法を検討しているが、旧来の最近隣法と比べてあまり良くない結果となっており、補助情報を十分に活用していないことが原因ではないかとしている。

（バーチャル・センサスのマス・インピュテーション）

Daalmans (2017)、De Waal et al. (2018)、Scholtus (2018) は、オランダのセンサスの「マス・インピュテーション」を取り扱っている。Daalmans (2017) によれば、オランダの人口・住宅センサスは、複数のレジスタや労働力統計などの抽出調査を組み合わせた「バーチャル・センサス」として行われており、2011 年のセンサスでは、重要な変数である最終学歴はレジスタから得られないため、労働力統計調査を用いている。一方、このデータはより包括的な学歴ファイルからも得られるため、2021 年センサスに向けてこのファイルが利用できないかの検討を進めている。このデータは現在整備中だが、新しいものであるため、80 年代以前に教育を終えたものは含まれない。このような部分的なデータから全体を推定するためには、ウェイト付けとマス・インピュテーションの 2 種類の方法があり、マス・インピュテーションには個別データでの項目間の関係が保たれないおそれがあるので一般には

[14] 坂下 (2018) で紹介した De Waal (2017) がこの問題のサーベイとなっている。
[15] 冒頭で Pannekoek らの手法は複雑で実装が難しいと指摘している。

ウェイト付けの方が好まれるが、オランダの人口センサスにおいては、ウェイトを他の調査と整合させる方法がはっきりしないことや、集計の簡易さなどの点からマス・インピュテーションが魅力的な選択肢だとしている。推計に用いるデータは、学歴ファイルに情報があるものはそれを用い、情報がないものは、その中で労働力統計調査から情報が得られるものから、学歴ファイルに情報がないもの全体を推計する方法を採っている。推計手法は、ホット・デック法とロジスティック回帰の 2 種類が考えられるが、ランダム・ホット・デック法ではドナーが見つからない可能性があること、最近隣法では時間がかかりすぎることなどのため、ここでは層化されたロジスティック回帰を適用し、かなり良い結果を得たと伝えている。De Waal et al. (2018) は、同じ研究の手法の詳細やインピュテーションによる分散の評価の方法を解説し、統合したレベルではマス・インピュテーションは EU のセンサスよりも変動係数が小さく、繰り返しウェイティングとの違いも大きいが、個別の区分ではこの違いは小さいとしている。Scholtus (2018) はマス・インピュテーションにより発生する分散の評価を分析的手法とブートストラップの 2 種類の方法で行ったもので、ほぼ妥当な結果を得たが、分析的手法が採用している近似計算のため、ブートストラップ法の方がいくらか正確であるとしている。

３．一般用ミクロデータを用いたホット・デック法のシミュレーション

ここでは、独立行政法人統計センターが提供している一般用ミクロデータを用いて、海外では標準的な手法であるが日本ではあまり用いられていないホット・デック法の数値シミュレーションを行い、その課題について検討する。

（ホット・デック法についての文献）

ホット・デック法については、海外の書籍でインピュテーションを扱ったものではたいてい言及されているが、入手しやすい論文としては、Andridge and Little (2010) が詳しい。その内容は、坂下 (2018) で紹介したように、基本的な手法から得られた推定値の性質、インピュテーションによる誤差の推定まで多岐にわたっているが、日本での経験が少ないこともあり、今回はシーケンシャル・ホット・デック法、ランダム・ホット・デック法などの基本的な手法を適用し、その課題について考察することとする。

（使用するデータ）

シミュレーションのためのデータとして、総務省統計局と独立行政法人統計センターが共同で開発し、統計センターが提供している「一般用ミクロデータ」の「全国消費実態調査（平成 21 年）十大費目勤労者世帯」を用いる。これは、「集計表から作成するなど、調査票情報を直接的に用いない方法により作成する擬似的なミクロデータ」[16]で、広く一般的に活用することを目的として、平成 28 年から提供されている。一般用ミクロデータとして

[16] （独）統計センター「一般用ミクロデータの利用」https://www.nstac.go.jp/services/ippan-microdata.html

提供されているデータには、他に、同じ調査の「十大費目全世帯」、「詳細品目全世帯」や就業構造基本調査（平成 4 年～24 年）があるが、今回は課題の発掘を目的とすることから、もっとも基本的なデータを用いることとする。提供側から示されている基本数（世帯主年齢階級別（5 分類）の集計世帯数及び世帯数分布＝集計用乗率の合計、十大費目別消費支出）は次のとおりである。

表１－１　集計世帯数及び世帯数分布

世帯主年齢階級	集計世帯数	世帯数分布
総数	26,239	18,171,954
30 歳未満	1,022	729,806
30～39 歳	6,076	4,267,250
40～49 歳	7,257	5,003,801
50～59 歳	7,624	5,122,436
60 歳以上	4,260	3,048,661

表１－２　十大費目別消費支出

世帯主年齢階級	年間収入（千円）	消費支出（円）	食料	住居	光熱・水道	家具・家事用品	被服及び履物	保健医療	交通・通信	教育	教養娯楽	その他の消費支出
総数	7,106	319,750	69,949	19,281	18,962	9,465	13,048	12,237	50,975	21,275	31,957	72,600
30歳未満	4,437	242,124	45,901	41,185	13,769	7,213	11,644	8,556	41,182	4,712	20,455	47,507
30～39歳	5,861	275,268	60,123	25,292	16,276	8,740	12,450	11,417	47,609	13,433	29,961	49,968
40～49歳	7,508	327,350	73,766	15,913	19,791	8,995	13,659	12,135	50,456	34,522	34,338	63,775
50～59歳	8,573	364,165	76,152	13,881	21,309	9,861	13,748	12,645	58,271	27,881	32,123	98,295
60歳以上	6,364	313,491	72,774	20,225	18,662	11,128	12,042	13,751	46,623	3,374	33,316	81,596

（シミュレーションの内容）

「全国消費実態調査（平成 21 年）十大費目勤労者世帯」の一般用ミクロデータには、4 つの質的項目（カテゴリー変量）とウェイト以外の 12 の量的項目（数量変量）がある。今回のシミュレーションでは、12 の量的項目[17]に対しランダムに欠測値を発生させ、シーケンシャル・ホット・デック法、ランダム・ホット・デック法などの手法を適用して結果を比較した。欠測値の発生については、10%と 20%の 2 つの発生率を想定し、各項目に独立に発生させた。

[17] うち１項目（消費支出）は合計であり、他の項目との関係にエディティング制約があるが、今回は考慮していない。

（シミュレーションの結果）

シミュレーションにおいて用いる分類は、基本数が示されている世帯主年齢階級別の 5 分類とする。まず、欠測値の発生率 10%のものと 20%のものに、続くデータで欠測していないもので補うシーケンシャル・ホット・デック法[18]を適用すると次のようになった。

表２－１　シーケンシャル・ホット・デック法（欠測率 10%）

世帯主年齢階級	年間収入	消費支出										
			食料	住居	光熱・水道	家具・家事用品	被服及び履物	保健医療	交通・通信	教育	教養娯楽	その他の消費支出
	（千円）	（円）										
総数	7,109	319,528	69,893	19,238	18,956	9,537	13,036	12,255	50,893	21,268	32,011	72,386
30 歳未満	4,445	238,667	45,871	41,151	13,739	7,113	11,575	8,692	40,792	4,837	20,790	48,044
30～39 歳	5,875	276,218	60,138	25,409	16,343	8,904	12,459	11,394	46,947	13,610	30,077	50,070
40～49 歳	7,511	326,824	73,437	16,080	19,800	9,007	13,560	12,148	50,927	34,415	34,433	63,223
50～59 歳	8,572	363,521	76,225	13,745	21,291	9,887	13,746	12,697	58,606	27,758	32,156	97,967
60 歳以上	6,355	313,612	72,839	19,768	18,555	11,287	12,142	13,743	45,821	3,439	33,183	81,507

表２－２　シーケンシャル・ホット・デック法（欠測率 20%）

世帯主年齢階級	年間収入	消費支出										
			食料	住居	光熱・水道	家具・家事用品	被服及び履物	保健医療	交通・通信	教育	教養娯楽	その他の消費支出
	（千円）	（円）										
総数	7,120	319,256	69,875	19,136	18,989	9,494	12,989	12,192	50,756	21,291	32,027	72,181
30 歳未満	4,475	239,530	45,750	40,592	13,823	6,876	11,774	8,909	39,939	4,860	20,028	47,389
30～39 歳	5,892	276,978	60,099	25,234	16,372	8,842	12,479	11,419	47,089	13,686	29,857	49,931
40～49 歳	7,529	326,524	73,339	15,901	19,864	9,021	13,516	12,026	51,211	34,793	34,419	62,836
50～59 歳	8,566	363,491	76,479	13,380	21,296	9,909	13,575	12,588	58,600	27,407	32,370	98,827
60 歳以上	6,368	311,267	72,552	20,445	18,574	11,109	12,142	13,669	44,549	3,434	33,434	79,828

欠測のない場合からのずれは、以下のとおりとなる。

[18] 最後のデータに行き着いた場合は最初に戻る。

表２－３　真の値からのずれ：シーケンシャル・ホット・デック法（欠測率 10%）

世帯主年齢階級	年間収入	消費支出										
			食料	住居	光熱・水道	家具・家事用品	被服及び履物	保健医療	交通・通信	教育	教養娯楽	その他の消費支出
総数	0.0%	-0.1%	-0.1%	-0.2%	0.0%	0.8%	-0.1%	0.1%	-0.2%	0.0%	0.2%	-0.3%
30 歳未満	0.2%	-1.4%	-0.1%	-0.1%	-0.2%	-1.4%	-0.6%	1.6%	-0.9%	2.7%	1.6%	1.1%
30～39 歳	0.2%	0.3%	0.0%	0.5%	0.4%	1.9%	0.1%	-0.2%	-1.4%	1.3%	0.4%	0.2%
40～49 歳	0.0%	-0.2%	-0.4%	1.0%	0.0%	0.1%	-0.7%	0.1%	0.9%	-0.3%	0.3%	-0.9%
50～59 歳	0.0%	-0.2%	0.1%	-1.0%	-0.1%	0.3%	0.0%	0.4%	0.6%	-0.4%	0.1%	-0.3%
60 歳以上	-0.1%	0.0%	0.1%	-2.3%	-0.6%	1.4%	0.8%	-0.1%	-1.7%	1.9%	-0.4%	-0.1%

表２－４　真の値からのずれ：シーケンシャル・ホット・デック法（欠測率 20%）

世帯主年齢階級	年間収入	消費支出										
			食料	住居	光熱・水道	家具・家事用品	被服及び履物	保健医療	交通・通信	教育	教養娯楽	その他の消費支出
総数	0.2%	-0.2%	-0.1%	-0.8%	0.1%	0.3%	-0.5%	-0.4%	-0.4%	0.1%	0.2%	-0.6%
30歳未満	0.9%	-1.1%	-0.3%	-1.4%	0.4%	-4.7%	1.1%	4.1%	-3.0%	3.1%	-2.1%	-0.2%
30～39歳	0.5%	0.6%	0.0%	-0.2%	0.6%	1.2%	0.2%	0.0%	-1.1%	1.9%	-0.3%	-0.1%
40～49歳	0.3%	-0.3%	-0.6%	-0.1%	0.4%	0.3%	-1.0%	-0.9%	1.5%	0.8%	0.2%	-1.5%
50～59歳	-0.1%	-0.2%	0.4%	-3.6%	-0.1%	0.5%	-1.3%	-0.5%	0.6%	-1.7%	0.8%	0.5%
60歳以上	0.1%	-0.7%	-0.3%	1.1%	-0.5%	-0.2%	0.8%	-0.6%	-4.4%	1.8%	0.4%	-2.2%

この結果は、真の値からのずれが欠測率 20%では多少大きくなるものの、全体的には実用上十分小さくなっているように見えるが、一般用ミクロデータは質的項目によって分類したセルごとに並べられているため、当てはまりが良くなっている可能性がある。欠測を発生させた同じデータをランダムにシャッフルしてから同じ手法を適用すると、次のようになる。

表３－１　シーケンシャル・ホット・デック法・シャッフル後（欠測率 10%）

世帯主年齢階級	年間収入	消費支出										
			食料	住居	光熱・水道	家具・家事用品	被服及び履物	保健医療	交通・通信	教育	教養娯楽	その他の消費支出
	（千円）	（円）										
総数	7,115	319,682	69,890	19,275	18,988	9,504	13,054	12,303	51,309	21,394	32,064	72,362
30 歳未満	4,697	246,614	48,900	39,903	14,259	7,804	11,647	9,125	42,928	6,726	21,697	50,392
30～39 歳	5,988	280,782	61,216	24,736	16,585	8,862	12,543	11,555	48,516	14,387	30,199	52,459
40～49 歳	7,479	326,834	73,041	16,553	19,763	9,032	13,581	12,185	51,004	33,256	34,125	64,013
50～59 歳	8,432	359,733	75,513	14,282	21,102	9,826	13,586	12,660	58,419	27,287	32,142	95,515
60 歳以上	6,462	312,592	72,439	19,550	18,658	11,043	12,347	13,705	45,778	5,344	33,640	80,281

表３－２　シーケンシャル・ホット・デック法・シャッフル後（欠測率 20%）

世帯主年齢階級	年間収入	消費支出										
			食料	住居	光熱・水道	家具・家事用品	被服及び履物	保健医療	交通・通信	教育	教養娯楽	その他の消費支出
	（千円）	（円）										
総数	7,111	319,321	69,917	19,007	19,014	9,482	13,077	12,322	51,517	21,661	31,874	72,311
30 歳未満	5,060	254,220	50,981	37,553	14,630	8,019	12,060	9,501	44,875	7,861	21,682	54,322
30～39 歳	6,133	285,436	62,167	23,667	16,832	8,902	12,718	11,697	48,877	15,258	30,287	54,700
40～49 歳	7,442	326,002	72,913	16,715	19,809	9,063	13,575	12,274	50,895	32,317	33,550	64,732
50～59 歳	8,244	353,654	74,915	14,654	20,889	9,763	13,558	12,613	57,875	26,979	32,012	93,500
60 歳以上	6,525	313,680	71,984	19,122	18,661	10,859	12,199	13,463	47,140	7,503	33,556	78,106

表３－３　真の値からのずれ：シーケンシャル・ホット・デック法・シャッフル後

（欠測率 10%）

世帯主年齢階級	年間収入	消費支出										
			食料	住居	光熱・水道	家具・家事用品	被服及び履物	保健医療	交通・通信	教育	教養娯楽	その他の消費支出
総数	0.1%	0.0%	-0.1%	0.0%	0.1%	0.4%	0.0%	0.5%	0.7%	0.6%	0.3%	-0.3%
30 歳未満	5.9%	1.9%	6.5%	-3.1%	3.6%	8.2%	0.0%	6.7%	4.2%	42.7%	6.1%	6.1%
30～39 歳	2.2%	2.0%	1.8%	-2.2%	1.9%	1.4%	0.7%	1.2%	1.9%	7.1%	0.8%	5.0%
40～49 歳	-0.4%	-0.2%	-1.0%	4.0%	-0.1%	0.4%	-0.6%	0.4%	1.1%	-3.7%	-0.6%	0.4%
50～59 歳	-1.6%	-1.2%	-0.8%	2.9%	-1.0%	-0.4%	-1.2%	0.1%	0.3%	-2.1%	0.1%	-2.8%
60 歳以上	1.5%	-0.3%	-0.5%	-3.3%	0.0%	-0.8%	2.5%	-0.3%	-1.8%	58.4%	1.0%	-1.6%

表３－４　真の値からのずれ：シーケンシャル・ホット・デック法・シャッフル後

（欠測率 20%）

世帯主年齢階級	年間収入	消費支出										
			食料	住居	光熱・水道	家具・家事用品	被服及び履物	保健医療	交通・通信	教育	教養娯楽	その他の消費支出
総数	0.1%	-0.1%	0.0%	-1.4%	0.3%	0.2%	0.2%	0.7%	1.1%	1.8%	-0.3%	-0.4%
30 歳未満	14.0%	5.0%	11.1%	-8.8%	6.3%	11.2%	3.6%	11.0%	9.0%	66.8%	6.0%	14.3%
30～39 歳	4.6%	3.7%	3.4%	-6.4%	3.4%	1.9%	2.2%	2.5%	2.7%	13.6%	1.1%	9.5%
40～49 歳	-0.9%	-0.4%	-1.2%	5.0%	0.1%	0.8%	-0.6%	1.1%	0.9%	-6.4%	-2.3%	1.5%
50～59 歳	-3.8%	-2.9%	-1.6%	5.6%	-2.0%	-1.0%	-1.4%	-0.3%	-0.7%	-3.2%	-0.3%	-4.9%
60 歳以上	2.5%	0.1%	-1.1%	-5.5%	0.0%	-2.4%	1.3%	-2.1%	1.1%	122.4%	0.7%	-4.3%

このように、世帯主が 30 歳未満のような標本数の少ない区分や欠測率を 20%に上げた場合のずれが大きくなっている。なお、世帯主が 30 歳未満の場合の教育費で特にずれが大きいのは、もともとこの区分では教育費がほとんどかかっていない世帯が多い中に、少数の多額の教育費を支出している世帯が存在する構造になっていて（子供の有無が関係していると思われる)、欠測値のインピュテーションによる元データからの乖離が大きくなるためである。

ところで、この手法はシャッフルをランダムに行っているので、標本全体をドナーとしたウェイトのないランダム・ホット・デック法とほぼ同じと考えられる。同じ欠測データに全体でのウェイトのないランダム・ホット・デック法を適用すると、以下のように似た

結果となる。

表４－１　全体でのランダム・ホット・デック法（欠測率 10%）

世帯主年齢階級	年間収入	消費支出										
			食料	住居	光熱・水道	家具・家事用品	被服及び履物	保健医療	交通・通信	教育	教養娯楽	その他の消費支出
	（千円）	（円）										
総数	7,107	319,473	69,893	19,166	18,996	9,513	13,074	12,281	51,414	21,666	31,963	72,607
30 歳未満	4,700	247,667	48,548	39,375	14,420	7,495	11,842	8,919	42,371	7,452	21,727	52,120
30～39 歳	6,008	280,488	61,112	24,423	16,581	8,899	12,517	11,500	47,663	14,095	30,350	52,270
40～49 歳	7,456	326,878	73,114	16,363	19,755	9,070	13,589	12,082	51,306	33,802	33,883	64,378
50～59 歳	8,433	358,220	75,594	14,366	21,110	9,821	13,666	12,715	58,313	27,709	32,205	96,186
60 歳以上	6,417	313,971	72,430	19,634	18,674	11,065	12,307	13,778	47,412	5,592	33,111	79,867

表４－２　全体でのランダム・ホット・デック法（欠測率 20%）

世帯主年齢階級	年間収入	消費支出										
			食料	住居	光熱・水道	家具・家事用品	被服及び履物	保健医療	交通・通信	教育	教養娯楽	その他の消費支出
	（千円）	（円）										
総数	7,094	319,067	69,949	18,979	19,024	9,495	12,990	12,271	51,202	21,657	31,994	72,968
30 歳未満	5,004	254,249	50,643	37,799	14,681	7,641	11,903	9,414	43,697	7,890	21,968	53,704
30～39 歳	6,125	284,354	62,100	23,758	16,841	8,922	12,592	11,673	48,143	15,071	30,599	54,760
40～49 歳	7,425	324,807	72,997	16,372	19,753	9,108	13,519	12,210	50,818	32,571	33,539	65,965
50～59 歳	8,219	354,274	74,824	14,420	20,977	9,784	13,475	12,592	57,905	26,786	32,128	93,415
60 歳以上	6,519	314,595	72,367	19,721	18,638	10,892	12,123	13,350	46,649	7,642	33,588	80,203

表４－３　真の値からのずれ：全体でのランダム・ホット・デック法（欠測率 10%）

世帯主年齢階級	年間収入	消費支出										
			食料	住居	光熱・水道	家具・家事用品	被服及び履物	保健医療	交通・通信	教育	教養娯楽	その他の消費支出
総数	0.0%	-0.1%	-0.1%	-0.6%	0.2%	0.5%	0.2%	0.4%	0.9%	1.8%	0.0%	0.0%
30 歳未満	5.9%	2.3%	5.8%	-4.4%	4.7%	3.9%	1.7%	4.2%	2.9%	58.1%	6.2%	9.7%
30～39 歳	2.5%	1.9%	1.6%	-3.4%	1.9%	1.8%	0.5%	0.7%	0.1%	4.9%	1.3%	4.6%
40～49 歳	-0.7%	-0.1%	-0.9%	2.8%	-0.2%	0.8%	-0.5%	-0.4%	1.7%	-2.1%	-1.3%	0.9%
50～59 歳	-1.6%	-1.6%	-0.7%	3.5%	-0.9%	-0.4%	-0.6%	0.6%	0.1%	-0.6%	0.3%	-2.1%
60 歳以上	0.8%	0.2%	-0.5%	-2.9%	0.1%	-0.6%	2.2%	0.2%	1.7%	65.7%	-0.6%	-2.1%

表４－４　真の値からのずれ：全体でのランダム・ホット・デック法（欠測率 20%）

世帯主年齢階級	年間収入	消費支出										
			食料	住居	光熱・水道	家具・家事用品	被服及び履物	保健医療	交通・通信	教育	教養娯楽	その他の消費支出
総数	-0.2%	-0.2%	0.0%	-1.6%	0.3%	0.3%	-0.4%	0.3%	0.4%	1.8%	0.1%	0.5%
30 歳未満	12.8%	5.0%	10.3%	-8.2%	6.6%	5.9%	2.2%	10.0%	6.1%	67.4%	7.4%	13.0%
30～39 歳	4.5%	3.3%	3.3%	-6.1%	3.5%	2.1%	1.1%	2.2%	1.1%	12.2%	2.1%	9.6%
40～49 歳	-1.1%	-0.8%	-1.0%	2.9%	-0.2%	1.3%	-1.0%	0.6%	0.7%	-5.7%	-2.3%	3.4%
50～59 歳	-4.1%	-2.7%	-1.7%	3.9%	-1.6%	-0.8%	-2.0%	-0.4%	-0.6%	-3.9%	0.0%	-5.0%
60 歳以上	2.4%	0.4%	-0.6%	-2.5%	-0.1%	-2.1%	0.7%	-2.9%	0.1%	126.5%	0.8%	-1.7%

このように、シーケンシャル・ホット・デック法（シャッフル後）、全体でのランダム・ホット・デック法ともに乖離が 40%を超えるセルが発生しており、このままでは実用に向かず、改善が必要と考えられる。

Kalton and Kasprzyk (1986)[19] では、ホット・デック法の種類として、シーケンシャルなホット・デック法の他、「全体でのランダムなインピュテーション」、「分類内でのランダムなインピュテーション」[20]を挙げている。これに従い、用いた一般用ミクロデータで提供されている質的項目 4 種（産業 3 分類、職業 3 分類、企業規模 4 分類、世帯主の年齢階級 5 分類）の組合せで分類（ドナー・プール）を作り[21]、この中でランダムなインピュテーション

[19] 坂下(2018)参照。
[20] Kalton and Kasprzyk (1986) ではランダムな手法に（おそらく「処理中のカードの束」という原義に忠実に従ったため）「ホット・デック」の呼称を用いていないが、今日では通常ホット・デック法に含める。
[21] 単純に計算すると 180 の分類ができるが、「職業」と「企業規模」の中で「第 1 次産業」は他と区別した分類となっていて、「産業」とのクロスで「構造的ゼロ」が発生するため実際の分類数はより少ない。

を行うと、乖離を小さく押さえることができる。

表５－１　分類内でのランダム・ホット・デック法（欠測率 10%）

世帯主年齢階級	年間収入（千円）	消費支出（円）	食料	住居	光熱・水道	家具・家事用品	被服及び履物	保健医療	交通・通信	教育	教養娯楽	その他の消費支出
総数	7,100	319,566	69,863	19,048	18,999	9,537	13,070	12,296	51,241	21,414	31,983	72,555
30 歳未満	4,451	240,042	46,070	40,729	13,798	7,130	11,694	8,558	41,752	4,773	21,020	48,241
30～39 歳	5,854	276,858	60,184	25,475	16,302	8,870	12,446	11,476	47,374	13,391	30,214	50,366
40～49 歳	7,503	327,746	73,599	15,791	19,872	9,059	13,662	12,119	50,966	34,911	34,174	63,562
50～59 歳	8,574	362,858	75,998	13,652	21,365	9,860	13,705	12,726	59,098	28,017	32,220	98,431
60 歳以上	6,343	312,215	72,667	19,278	18,607	11,289	12,232	13,905	46,175	3,377	33,091	80,721

表５－２　分類内でのランダム・ホット・デック法（欠測率 20%）

世帯主年齢階級	年間収入（千円）	消費支出（円）	食料	住居	光熱・水道	家具・家事用品	被服及び履物	保健医療	交通・通信	教育	教養娯楽	その他の消費支出
総数	7,120	318,022	69,903	18,975	19,040	9,540	12,993	12,246	51,121	21,609	31,995	72,903
30 歳未満	4,457	238,509	45,981	42,056	13,770	7,053	11,651	8,774	42,169	4,918	20,266	48,248
30～39 歳	5,903	274,381	60,180	25,173	16,310	8,809	12,463	11,441	47,161	13,470	29,645	49,754
40～49 歳	7,550	327,726	73,622	15,915	19,949	9,042	13,597	12,047	50,920	34,875	34,247	64,350
50～59 歳	8,552	360,588	76,027	13,434	21,409	9,984	13,623	12,534	59,749	28,497	32,159	99,461
60 歳以上	6,350	310,694	72,848	19,107	18,650	11,228	12,011	14,046	44,643	3,654	34,124	80,622

表５－３　真の値からのずれ：分類内でのランダム・ホット・デック法（欠測率 10%）

世帯主年齢階級	年 間 収入	消費支出	食料	住居	光熱・水道	家具・家事用品	被服及び履物	保健医療	交通・通信	教育	教養娯楽	その他の消費支出
総数	-0.1%	-0.1%	-0.1%	-1.2%	0.2%	0.8%	0.2%	0.5%	0.5%	0.7%	0.1%	-0.1%
30 歳未満	0.3%	-0.9%	0.4%	-1.1%	0.2%	-1.2%	0.4%	0.0%	1.4%	1.3%	2.8%	1.5%
30～39 歳	-0.1%	0.6%	0.1%	0.7%	0.2%	1.5%	0.0%	0.5%	-0.5%	-0.3%	0.8%	0.8%
40～49 歳	-0.1%	0.1%	-0.2%	-0.8%	0.4%	0.7%	0.0%	-0.1%	1.0%	1.1%	-0.5%	-0.3%
50～59 歳	0.0%	-0.4%	-0.2%	-1.6%	0.3%	0.0%	-0.3%	0.6%	1.4%	0.5%	0.3%	0.1%
60 歳以上	-0.3%	-0.4%	-0.1%	-4.7%	-0.3%	1.4%	1.6%	1.1%	-1.0%	0.1%	-0.7%	-1.1%

表５－４　真の値からのずれ：分類内でのランダム・ホット・デック法（欠測率 20%）

世帯主年齢階級	年 間 収入	消費支出	食料	住居	光熱・水道	家具・家事用品	被服及び履物	保健医療	交通・通信	教育	教養娯楽	その他の消費支出
総数	0.2%	-0.2%	-0.1%	-0.8%	0.1%	0.3%	-0.5%	-0.4%	-0.4%	0.1%	0.2%	-0.6%
30 歳未満	0.9%	-1.1%	-0.3%	-1.4%	0.4%	-4.7%	1.1%	4.1%	-3.0%	3.1%	-2.1%	-0.2%
30～39 歳	0.5%	0.6%	0.0%	-0.2%	0.6%	1.2%	0.2%	0.0%	-1.1%	1.9%	-0.3%	-0.1%
40～49 歳	0.3%	-0.3%	-0.6%	-0.1%	0.4%	0.3%	-1.0%	-0.9%	1.5%	0.8%	0.2%	-1.5%
50～59 歳	-0.1%	-0.2%	0.4%	-3.6%	-0.1%	0.5%	-1.3%	-0.5%	0.6%	-1.7%	0.8%	0.5%
60 歳以上	0.1%	-0.7%	-0.3%	1.1%	-0.5%	-0.2%	0.8%	-0.6%	-4.4%	1.8%	0.4%	-2.2%

（結果の考察と課題）

以上のシミュレーションから、ホット・デック法のドナーを全データからランダムに選ぶと乖離が大きくなるが、適切なドナー・プールを作成すれば小さく押さえることができると考えられる。ただし、用いたデータの性質から、この結果にはいくつかの留保が必要である。

一つは、元になっている全国消費実態調査は、調査世帯に３ヶ月にわたって家計簿をつけてもらい、収支項目ごとに一ヶ月平均の値を計算しているものであるため、ミクロデータが既に集計値になっていて、今回の方法は調査における欠測値の発生状況は反映していないということである。もう一つは、一般用ミクロデータは、元データを集計して分布のパラメータを作成し、そこから逆に個別データを発生させるという手法を採っているので、

外れ値のような統計の実務上で遭遇する問題を有さず、分布を用いた手法[22]の当てはまりが実際以上に良くなる可能性があることである。今回のシミュレーションは、あくまでもホット・デック法によるインピュテーションの技術的な可能性を調べるものである。

今後の課題としては、以下のようなものが考えられる。

- ウェイトの取り扱いについての検討。今回はランダム・ホット・デック法の適用に当たってウェイトを考慮しなかったが、標本の分布をもって母集団の分布とみなすのであれば、考慮するのが適当であると考えられる。一方で、ウェイトが不均等な場合、ウェイトの高い特定の標本ばかりドナーに採用される可能性がある。ウェイトに関する問題は先行文献でも明確な結論が出ておらず[23]、個別の検討を要する。
- MCAR 以外の発生メカニズムへの対応。今回は欠測値の発生をランダムに行ったが、これは MCAR (Missing Completely at Random) の発生メカニズムを前提としていることになる。現実には欠測変数或いは他の変数により発生確率が変わることが考えられ、対応の検討を要する。
- エディティング制約への対応。今回は合計項目も他の項目と同様に扱ったが、これは現実にはエディティング制約下にあると考えられ、ホット・デック適用以前に演繹的インピュテーションを行うなどの処理の検討を要する[24]。
- 質的項目における欠測の処理。今回は質的項目をフェイス項目のように扱って、欠測はないものとしたが、質的項目で欠測が発生した場合には複数の対応法が考えられる（質的項目だけを考慮してインピュテーションを行う、量的項目も手がかりとする等)。
- 複数項目が欠測した場合への対応。今回は項目ごとに独立したインピュテーションを行っているが、これは項目間の相関を毀損する可能性があるため、同一のドナーを用いるなどの対処の検討を要する。
- ドナーが足りなくなった場合への対処。今回の欠測率 20%までのシミュレーションでは発生しなかったが、欠測率を上げていくとドナー・プール内のドナーが足りなくなることが起こりうる。また、分類を細かくしていくと分類の中のデータがすべて欠測することも起こりうる。このような場合は通常、ドナー・プールを統合することで対応するが、その方法の検討を要する。
- ドナーの選び方に関する検討。今回のホット・デック法はドナーをシーケンシャル或いはランダムに選定したが、最近隣法などの手法も考えられる。
- インピュテーションによる誤差の評価。今回は単純に集計表における真値からのずれを評価しているが、真の値が分からない場合にどのように評価するかの検討。ランダムな代入を繰り返すだけではモデルの安定性を評価できないので、ブートストラップ

[22] ホット・デック法は、標本の分布が母集団の分布を反映しているとの仮定に基づいているため、非明示的とは言え分布を用いた手法の一種と言うことができる（Little and Rubin (2002)）。

[23] Andridge and Little (2010) 第 4 節など。

[24] シミュレーション上の課題であって、現実の全国消費実態調査では合計項目は下位の項目の積算として計算されているので、このようなことは起こらない。

法などの多重代入的な手法が必要で、その場合のウェイトの取り扱いも課題となる。

４．まとめ

文献の調査については、少なくとも今回対象にした２カ国については、インピュテーションシステム、ベイズモデル、制約条件下のインピュテーションなどの継続的な課題を設定し、豊富な過去の蓄積の上で新たな検討を行っていることが分かる。

数値シミュレーションについては、我が国の統計データにおける欠測値への対処にホット・デック法を適用することは可能であると考えられるが、ドナーの選び方などの具体的な方法は、適用する調査に応じて子細かつ実務的に検討する必要がある。

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] 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)。

[5] 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.

[6] Bankier, M. (1991), "Alternative Method of Doing Quantitative Variable Imputation", Statistics Canada Memorandum.

[7] Bankier, M., Houle, A.-M., Luc, M., and Newcombe, P. (1997), "1996 Canadian Census Demographic Variables Imputation", American Statistical Association, Proceedings of the 1997 Section on Survey Research Methods, 389-394.

[8] Bankier, M., Lachance, M., and Poirier, P. (2000), "2001 Canadian Census Minimum Change Donor Imputation Methodology", U.N. Economic Commission for Europe Work Session on Statistical Data Editing, Cardiff, UK, October 2000.

[9] Chun, A.Y. and Larson, M. (Eds) (2020), "Administrative Records for Survey Methodology", New York, NY: Wiley.

[10] CSRM (2019), "Annual Report of the Center for Statistical Research and Methodology, Research and Methodology Directorate, Fiscal Year 2019", U.S. Department of Commerce, Economics and Statistics Administration, U.S. CENSUS BUREAU.

[11] Daalmans, J. (2017), "Mass Imputation for Census Estimation.", Discussion paper (201704), Statistics Netherlands, The Hague.

[12] De Waal, T. (1996), "CherryPi: A Computer Program for Automatic Edit and Imputation", Report, Statistics Netherlands, Voorburg.

[13] De Waal, T. (2000), "A Brief Overview of Imputation Methods Applied at Statistics Netherlands", Netherlands Official Statistics, 15, 23–27.

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

[15] De Waal, T. and Coutinho, W. (2017), "Imputation of Numerical Data under Edit Restrictions: The Vertices Approach", Discussion paper (201702), Statistics Netherlands.

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

[17] De Waal, T., Pannekoek, J., and Scholtus , S. (2011b), "The editing of statistical data: methods

and techniques for the efficient detection and correction of errors and missing values", Discussion Paper (201132), Statistics Netherlands.

[18] De Waal, T., Coutinho, W., and Shlomo, N. (2015). "Calibrated Hot Deck Imputation for Numerical Data under Edit Restrictions", Discussion Paper (201520), Statistics Netherlands.

[19] De Waal, T., Daalmans, J., and Linder, F. (2018), "Mass imputation for Census Estimation: Methodology", Report, Statistics Netherlands, The Hague.

[20] Draper, L. R., and Winkler, W. E. (1997), "Balancing and Ratio Editing With the New Speer System", Research Report RR97/05, Statistical Research Division, US Bureau of the Census, Washington, DC.

[21] Fellegi, I. P., and Holt, D. (1976), "A Systematic Approach to Automatic Edit and Imputation", Journal of the American Statistical Association, 71, 17-35.

[22] Garcia-Rubio, E., and Villan, I. (1990), "DIA system: Software for the Automatic Editing of Qualitative Data", in Proceedings of 1990 Annual Research Conference, US Bureau of the Census, Arlington, VA, pp. 525–537.

[23] Greenberg, B. G. and Surdi, R. (1984), "A Flexible and Interactive Edit and ImputationSystem for Ratio Edits", Research Report RR84/18, Statistical Research Division, U. S. Bureau of the Census, Washington, DC.

[24] Greenberg, B. G., and Petkunas, T. (1990), "Overview of the SPEER System", SRD report RR-90/15, U.S. Bureau of the Census, Washington, D.C., USA.

[25] Hoogland, J., Van der Loo, M. Pannekoek, J., and Scholtus , S. (2011), "Data editing: Detection and correction of errors", Statistical Methods (201110), Statistics Netherlands.

[26] Israëls, A., Kuyvenhoven, L., Van der Laan, J., Pannekoek, J., and Schulte Nordholt , E. (2011), "Imputation", Statistical Methods (201112), Statistics Netherlands.

[27] Kalton, G. and Kasprzyk, D. (1986), "The Treatment of Missing Survey Data", Survey Methodology 12, pp. 1-16.

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

[29] Kim, H. J., Drechsler, J., and Thompson, K. J. (2019), "Synthetic Microdata for Establishment Surveys Under Informative Sampling", RESEARCH REPORT SERIES (Statistics #2019-07), CSRM, US Bureau of the Census, Washington, DC.

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

[31] Kim, H. J., Reiter, J. P., Wang, Q., Cox, L. H., and Karr, A. F. (2014), "Multiple Imputation of Missing or Faulty Values Under Linear Constraints", Journal of Business & Economic Statistics 32 (2014): 375-386. DOI: 10.1080/07350015.2014.885435.

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

[33] Klein, M. and Sinha, B. (2015), "Likelihood based finite sample inference for singly imputed synthetic data under the multivariate normal and multiple linear regression models", Journal of Privacy and Condentiality 7 (1), 43-98.

[34] Little, R. A., and Rubin, D. B., (2002), "Statistical Analysis with Missing Data (2nd Edition)", New York, N.Y., John Wiley.

[35] Manzari, A. (2004), "Combining Editing and Imputation Methods: An Experimental Application on Population Census Data", Journal of the Royal Statistical Society, Sereis A, 167, 295–307.

[36] Pannekoek, J. (2009), "Research on edit and imputation methodology: the throughput programme", Discussion Paper (09022), Statistics Netherlands.

[37] Pannekoek, J., Scholtus , S., and Van der Loo, M. (2013), "Automated and manual data editing: a view on processdesign and methodology", Discussion Paper (201309), Statistics Netherlands.

[38] Pannekoek, J., Shlomo, N., and De Waal, T. (2009), "Calibrated Imputation of Numerical Data under Linear Edit Restrictions", Discussion Paper (09016), Statistics Netherlands.

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

[40] Van der Loo, M. and Jonge, E. (2011), "Deductive imputation with the deducorrect package", Discussion paper (201126), Statistics Netherlands, The Hague.

[41] Whitridge, P., and Kovar, J. G. (1990), "Applications of the Generalized Edit and Imputation System at Statistics Canada", in American Statistical Association Proceedings of the Survey Research Method Section, pp. 105–110.

[42] Winkler, W. E. (1997), "Editing Discrete Data", American Statistical Association, Proceedings of the Section on Survey Research Methods, 108-113.

[43] Winkler, W. E. (2003), "A Contingency Table Model for Imputing Data Satisfying Analytic Constraints", RESEARCH REPORT SERIES (Statistics #2003-07), Statistical Research Division, US Bureau of the Census, Washington, DC.

[44] 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), Statistical Research Division, US Bureau of the Census, Washington, DC.

[45] Winkler, W. E. (2011), "Cleaning and Using Administrative Lists: Enhanced Practices and Computational Algorithms for Record Linkage and Modeling/Editing/Imputation", Section on Survey Research Methods – Joint Statistical Meetings 2011.

[46] Winkler, W. E. (2018), "Cleaning and Using Administrative Lists: Enhanced Practices and

Computational Algorithms for Record Linkage and Modeling/Editing/Imputation", RESEARCH REPORT SERIES (Statistics #2018-05), CSRM, US Bureau of the Census, Washington, DC.

[47] Winkler, W. E. and Chen, B.-C. (2002), "Extending the Fellegi-Holt Model of Statistical Data Editing", RESEARCH REPORT SERIES (Statistics #2002-02), Statistical Research Division, US Bureau of the Census, Washington, DC.

[48] Winkler, W. E., and Draper, L. R. (1996), "Application of the SPEER Edit System", Research Report RR96/02, Statistical Research Division, US Bureau of the Census, Washington, DC.