資料９

諸外国における欠測値補完について

令和 2 年 1 月 27 日
総務省統計研究研修所
統計研修研究官

１　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、特に欠測値補完の問題は避けて通れない事項であり、海外においては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いる。総務省統計研究研修所では、諸外国の統計機関におけるデータ・エディティング、特に欠測値補完の状況について、各国の現状、研究や議論の動向、書籍や論文などの基礎的な文献について情報収集を行っている[1]。ここでは現在までの結果からその要点を解説する。

２　諸外国の公的統計における主な欠測値補完方法

高橋(2017)[2]によると、国連欧州経済委員会(UNECE)の会合に参加している 23 機関に対するアンケートの結果、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・デック法のうち、ホット・デック法は 100%、他の手法も 95%の機関で使用されている。一般的に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している。

具体的な手法は調査に応じてさまざまであるが、調査で判明したものを参考に挙げる[3]。これによると

- 現時点で行われている手法は、ホット・デック法、比率代入法などの伝統的な手法が多い。
- 分布モデルを用いるような高度な手法は検討課題とされていることが多い（米国、イタリアなど)。
- 行政情報の統計作成への利用の進展に伴って、そのために必要となった欠測値補完や、欠測値補完に行政情報を反映させる検討も見られる。
- 国を超えた手法の共有化として、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 CANCEIS が英国、ドイツ、ニュージーランドで採用又は採用を検討

---

[1] 現在までの結果は、総務省統計研究研修所のリサーチペーパー第 40 号、第 43 号、第 44 号としてまとめている。
[2] 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。
[3] リサーチペーパー第 40 号に基づき、分かるものについては更新。

されている。CANCEIS は他にイスラエルが採用しているほか米国でも研究用に用いられている。

- 経済統計向けのシステム Banff は表に記載したニュージーランドのほか米国の経済分析局 (BEA) でも採用を検討している。

などの点が注目される。

3　欠測値補完方法の分類

欠測値を補完する方法の分類は、必ずしも定まった体系があるわけではない。しかしながら、1980 年以降豊富な書籍や研究論の蓄積があり、それらを踏まえて行われる議論で混乱をきたすことはないものと思われる。よく引用される論文としては、

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

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.

などがある[4]。

Kalton and Kasprzyk (1986) は、欠測値の補完手法として、

(a) 演繹的 (deductive) インピュテーション
(b) 全体の平均による (Overall mean) インピュテーション
(c) 分類内の平均による (Class mean) インピュテーション
(d) 全体でのランダムな (Random overall) インピュテーション
(e) 分類内でのランダムな (Random within classes) インピュテーション
(f) シーケンシャルなホット・デック (Sequential hot-deck) インピュテーション
(g) 階層的なホット・デック (Hierarchical hot-deck) インピュテーション
(h) 回帰による (Regression) インピュテーション
(i) 距離関数によるマッチング (Distance function matching)[5]

を挙げ、残差を加えるか（加えるならばどのような形の残差か）などの視点で分類している。また、回帰によるインピュテーションの変種として、回帰による推定値の代わりに、実際のデータが取った値の中で推定値に最も近いものを代入する予測平均値マッチング (Predictive mean matching) を挙げる。

Andridge and Little (2010) は、「実地には広範囲に用いられているが、他の手法に比べて

[4] リサーチペーパー第 43 号。
[5] 通常「最近隣法」と呼ばれるもので、ホット・デック法の一種とされることが多い。

背景の理論が開発されていない」ホット・デック法について論じている。

４　欠測値補完方法についての書籍

リサーチの中で、欠測値補完の分野において、海外で出版されている書籍について諸論文の参考文献などを手がかりに、以下の 4 冊を収集・調査した[6]。

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

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

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

Van Buuren, S. (2018), “Flexible Imputation of Missing Data (second edition)”, Chapman & Hall/CRC, Boca Raton, Florida.

これらのうち Rubin (1987) は、その書名のとおり多重代入法を論じた古典的書籍で、Van Buuren (2018) によると、これ以降多重代入法に関する文献が指数関数的に増え続けている[7]。Little and Rubin (2002) と Van Buuren (2018)も多重代入法や欠測値を含むデータの分析についての理論的な書籍という面が強いが、De Waal et al. (2011) は公的統計の実務における欠測値の補完に焦点を絞った内容となっている[8]。

このように、海外では商業出版されている書籍においても、多重代入法などの理論的な内容が多い傾向があるものの、1980 年以降、公的統計の欠測値補完の分野で議論の根拠となるような文献が提供されている[9]。

５　統計機関による補完方法の解説

オランダ統計局では、統計調査の欠測値補完方法の解説を公開しており、少なくとも 90 年代まで情報が遡れる。また手法に関するディスカッション・ペーパーなども盛んに公表している。

具体的な文献としては、

---

[6] リサーチペーパー第 44 号。

[7] 多重代入法についての文献が増えている背景には、技術的にはコンピュータの進歩により反復シミュレーションが行いやすくなったこと、ニーズ面からは統計の精度評価やミクロデータの提供からの要請があるものと考えられる。

[8] この書籍での補完法の分類は大方 Kalton and Kasprzyk (1986) に拠っている。

[9] 岩波書店の「欠測データの統計科学」（星野崇宏・岡田謙介編）や共立出版の「欠測データ処理」（高橋将宜・渡辺美智子著）のように日本にも出版物がないわけではないが、前者は自然科学分野も含んだものであり、後者は「統計学 One Point」シリーズの一冊として多重代入法に焦点を当てたもので、公的統計分野で網羅的なものは乏しい。

Schulte Nordholt, E. (1997), “Imputation in the New Dutch Structure of Earnings Survey (SES)”, Work Session on Statistical Data Editing, Statistical Commission and Economic Commission for Europe, Prague, October 1997.
Schulte Nordholt, E. (1998), “Imputation: Methods, Simulation Experiments and Practical Examples”, International Statistical Review, 66, 157-180.
De Waal, T. (2000), “A Brief Overview of Imputation Methods Applied at Statistics Netherlands”, Report, Statistics Netherlands.
Hoogland, J., Van der Loo, M. Pannekoek, J., and Scholtus , S. (2011), “Data editing: Detection and correction of errors”, Statistical Methods (201110), Statistics Netherlands.
Israëls, A., Kuyvenhoven, L., Van der Laan, J., Pannekoek, J., and Schulte Nordholt , E. (2011), “Imputation”, Statistical Methods (201112), Statistics Netherlands.
De Waal, T., Pannekoek, J., and Scholtus , S. (2011), “The editing of statistical data: methods and techniques for the efficient detection and correction of errors and missing values”, Discussion Paper (201132), Statistics Netherlands.
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.

などがある。

米国でもセンサス局では、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、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。また、人口動態調査 (CPS) については、 欠測値の補完を含む手法全体の解説が公開されている。具体的な文献としては、米国医療研究・品質調査機構(AHRQ)でも医療費支出パネル調査(MEPS) の欠測値補完手法についての報告を公開している。
具体的な文献としては、

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.a, September 25, 2003.
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.
U. S. Bureau of Labor Statistics and U. S. Census Bureau (2006), “Design and Methodology, Current Population Survey, TP66”, Tech rep. Vol. 66, 2006. Technical paper 66.
Agency for Healthcare Research and Quality (2007), “Overview of Methodology for Imputing Missing Expenditure Data in the Medical Expenditure Panel Survey”, Methodology Report #19.

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.

などがある。

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（世帯員全員が欠測した場合、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の研究が行われていた。 聞き取り調査と行政情報利用の比較検討、用いる行政情報の選定方法の検討を行っている。 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、商用データによる予測を検討中。
	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	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

		（人口特性の一致する観測値を代入）。	量法(sequential regression multivariate imputation, SRMI)による多重代入法を検討中。 （SIPP についてはさまざまな研究プロジェクトが行われている。）
	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 を検討していたが、2 種類のホット･デック法（ランダム､最近隣）のいずれかを用いることになった。
	月次卸売調査(Monthly Wholesale Trade Survey, MWTS)		継続的な研究プロジェクトが進行しており、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(Multivariate Imputation by Chained Equations)や順次回帰(Sequential Regression)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	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 を利用。 新たに加わった調査項目に対しては他の統計調査ともに、ミシガン大学で開発された IVEware を用いている

	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）。 現在、COTS（市販の既製品）ソフトウェアで置き換えることを検討中。

欧州

調査部局	調査名	補完手法	備考
イギリス国家統計局	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（ホット・デック法）。	次期 2021 年センサスでは行政情報を緩用したインピュテーションの検討を行っている。ただし、行政情報の値をそのまま持ってくるのではなく、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もの。
	企業統計における税務データの利用	カンパニー（社内の単位）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	現在検討中。

		から得られる補助変数による比率による補完を検討中	
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。 次回に向けて CANCEIS 及び多重代入法を検討中。
	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	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 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。	k-最近隣法について、ウェイト作成に機械学習を適用する研究を行っている。
	人口センサス	2011 年センサスではホット･デック法、次回は未定。	2011 年以降レジスタ・ベースで行われている。職業の欠測が多く精度が低い。

	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。	
イタリア国家統計局	人口レジスタ		最終学歴データのマス・インピュテーションについて、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。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ることを検討。
	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。	
ハンガリー中央統計局		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	一般	エディティング、補完システムとして人口センサス用の CANCEIS、経済統計用の Banff を開発。他の統計局でも採用ないし採用を検討されている。	
	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。