諸外国の国勢調査におけるインピュテーション方法

北原　昌嗣†
寺垣内　雅子†

The Imputation Method on the Population Census in Foreign Countries

KITAHARA Masatsugu
TERAGAUCHI Masako

国勢調査のような大規模調査データに欠測値や項目間における矛盾が生じた場合、それら全てのデータに対して調査回答者に確認を取ることは、現実的には極めて困難であるため、その対処方法として、尤もらしい整合性のあるデータを当該項目に代入して訂正（補正）をする、いわゆるインピュテーションを実行することで、各国統計局は結果を公表している。現在、このインピュテーション方法には、いくつかの種類があり、各国統計局が自国のデータに適合したインピュテーションを実施しているところである。多くの国では、同じ調査データの中から、欠測（又は矛盾）した箇所（値）に極めて類似していると推定される値を検索し、それを欠測箇所へ代入するホットデック法が利用されているが、ホットデック法も細分化すると、様々なやり方、手順が存在し、各国統計局はより良いインピュテーションを実行できるように創意工夫をしている。公表日が定まっている統計調査では、インピュテーション処理に多くの時間を費やすことはできず、手順を変更するだけでも処理速度の向上が見込めるインピュテーション処理には、試行錯誤をしていく価値が十分にあると考えられる。

本稿では、国勢調査で使われるインピュテーション方法について、国連統計部が発行しているガイドラインにおける記述を紹介した上で、各国統計局へのヒアリング調査結果を比較し考察する。

キーワード：国勢調査、補定処理、国際比較

In case of clear inconsistencies, or when the item is blank in the large-scale survey data such as the Population Census, it is practically extremely difficult to confirm to respondents on all such data. As the solution, the National Statistical Organizations (NSOs) publish results corrected by substituting similar data into their items, so-called “Imputation.” Presently, there are several types of this imputation method, and NSOs use imputation method that is appropriate for own country’s data. Many countries use the Hot-Deck Technique, which looks for value that is estimated to be similar to the missing (or inconsistent) value and substitutes it into the missing value. Various methods and procedures exist for the Hot- Deck Technique, and the NSOs are devising it so that better imputation can be implemented. It is not possible to spend much time on imputation work for statistical surveys with a fixed publication date. Therefore, the imputation processing, which can be expected to improve processing speed simply by changing the procedure, is thought to be well worth studying.

This paper introduces the description of imputation methods used in the Population Census from the guidelines issued by the United Nations Statistics Division, and then compares and considers the results of interviews with NSOs.

Key words: Population Census, Imputation, International Comparison

† 総務省統計局統計調査部国勢統計課

Ⅰ　はじめに[1]

インピュテーション(Imputation)とは、データに欠測値や項目間に矛盾が生じている場合に、その欠測値等に尤もらしい整合性のある推定値を代入する方法であり、日本では補定や補完などと呼ばれている。インピュテーションは、近年の調査環境の悪化等による未回答データに対処するための国際的に認められている方法であり、各国の国勢調査などで利用されている。

生データには、調査票回答者や調査員によって生じる多種多様なエラーが含まれている可能性がある。こういったエラーは回答者本人や調査員に確認し解消するのが一番確実な方法であるが、国勢調査のような全数調査や大規模な調査の場合、全てのエラーを当人に確認するのは時間も費用もかかり至極困難でほぼ不可能である。そのため、そういったエラーを訂正する場合においては、データエディティングにより、データの論理的な矛盾や欠測値を検出して、真値である可能性が高いと思われる代替値を代入（インピュテーション）し、データ内における欠落や不整合をなくしているのである。また、膨大なデータを人手でチェックするのには限界があり、コンピュータの助けがあってこそのデータエディティングである。

各国の国勢調査で主に使われているインピュテーション方法は、欠測値やエディットルール（Edit Rule）により論理的矛盾が明らかとなった値に対して、ドナー（Donor)と呼ばれる代替値を見つけ、それを代入するドナーインピュテーション方法であるが、インピュテーションを行う手順についてはそれぞれの国で特色があり、各国のデータにおける特徴に応じた工夫がなされている。これは、国によって、データ処理数や集計項目、慣行の違いがあり、また、利用できる行政記録情報が存在しているかどうかにもよることから、効率的な処理手順は各国で異なってくるのである。カナダ統計局では、データ駆動(data-driven)に優れた国勢調査のためのエディット・インピュテーションシステムが開発されており、そのシステムを各国と共有することで、多くの国で同じシステムが使われているという国際的な協力関係も見られる。

本稿では、統計調査結果の作成に大変重要な役割を果たしているデータエディティングのうち特にインピュテーションに焦点を当て、国連統計部におけるデータエディティングに関する見解や、諸外国の国勢調査におけるインピュテーション方法について各国統計局にヒアリング調査を実施した結果を紹介する。

図１　データエディティングの説明図

[1] 本稿の作成に当たっては、永井恵子氏（総務省統計局）、関野秀峰氏（総務省統計局）、都筑貴史氏（総務省統計局）、野崎政志氏（独立行政法人統計センター）及び塚本大器氏（内閣府経済社会総合研究所）から大変貴重な助言をいただいた。また、総務省統計局総務課国際係の皆様からは、海外情報の提供をしていただいた。ここに改めて感謝の意を表したい。なお、本稿の内容は個人的見解も含まれており、筆者の所属する機関の公式見解を示すものではない。

Ⅱ　国連統計部におけるデータエディティングの見解

国連統計部（United Nations Statistics Division）は、世界の統計制度の発展に尽力している組織であり、各種統計の国際的な基準を策定している。国勢調査に関して言えば、2017 年に「国勢調査に関する原則及び勧告　改訂３版（Principles and Recommendations for Population and Housing Censuses Revision 3）」を発行している。この原則及び勧告は、各国が国勢調査を実施する際のガイドラインとして必要不可欠なものであると同時に、各国統計局がより効率的で費用対効果の高い国勢調査を計画し実施できるよう支援するものである。また、国勢調査のデータエディティングについて詳細な説明がなされている「国勢調査エディティングに関するハンドブック　改訂２版（Handbook on Population and Housing Census Editing Revision 2）」も 2019 年に発行している。これら資料において、データエディティングがどのように記載されているのかをここでは紹介する。

１　国勢調査に関する原則及び勧告

国勢調査に関する原則及び勧告（改訂３版）では、データエディティング処理について、項目番号【3. 189】【3. 190】に以下の記述が見られる。

3.189. Since for large censuses manual correction is rarely economically feasible, the conditions for such corrections are usually specified in specially designed computer programs for automatic error scrutiny and imputation based on other information for the person or household or for other persons or households. Whenever imputation is used, a flag should be set so that analysts are able to distinguish between reported information and that imputed by the editing system. For cases where sufficient information is unavailable for the specific persons or household to correct apparent errors, imputation methods can be used such as the hot deck approach. This technique uses information obtained from previously processed persons, families or households with similar characteristics as the "best suited" value in replacing missing values or values that have failed processing edits. However, this technique requires careful programming work, considering that the search for appropriate information in the census database would slow down computer program execution.

日本語訳[2]

3. 189　大規模な国勢調査のための手動による訂正は経済的にほぼ実現不可能であるため、そのような訂正の条件は通常その人及びその世帯、又は他の人及び他の世帯の情報に基づいて自動エラー検査及びインピュテーションのための特別に設計されたコンピュータプログラムで指定される。インピュテーション時には、必ずアナリストが生データとインピュテーションデータを区別できるようにフラグを立てておく必要がある。特定の人物や世帯について、明らかなエラーを訂正するのに十分な情報が利用できない場合には、ホットデックなどのインピュテーション方法を使用することができる。この技術は、欠測値やエディットで不合格だった数値を「最も適した」数値に置き換える際に、前もって処理された同様の特性を持った人物、家族又は世帯のデータから得られた情報を利用する。しかし、国勢調査データベースから適切な情報を検索することは、コンピュータの処理速度を低下させるため、慎重なプログラミング作業を必要とする。

国勢調査のようなデータ数の多い調査におけるエラーの検出及び訂正を人手で対処することはほぼ不可能である。そのため、国連統計部ではコンピュータによるエラーの検出及び訂正を勧告している。また、特定の人物や世帯について明らかなエラーを訂正するための他の情報[3]が存在しない場合、ホットデック法などのインピュテーション方法を使用することができるとあり、訂正が必要な整合性がとれないデータや欠測値については、同じ調査で得られた類似性の最も高い

[2] 本稿における日本語訳は筆者による仮訳

[3] 調査票に記入された情報による論理チェックで訂正できることもあるが行政記録情報を利用することもある。ただし、行政記録情報では、時点の違いや用語の定義が異なっていることがあるため、利用には注意が必要である。

データに置き換えることで整合性をとることができると考えられている。また、このほかにも、過去に得られたデータ、例えば同一人物が過去に回答したデータをインピュテーションするコールドデック法を利用することもできる。こういった方法でデータを訂正することにより、整合性がとれたデータを公表することができるのである。ただし、類似したデータを膨大なデータの中から探してくるのは容易なことではなく、検索方法やプログラミングのやり方によっては膨大な処理時間がかかってしまう。そのため、慎重なプログラミング作業が必要であると記述されている。

人による目視チェック等ではヒューマンエラーや個人による判断の違いが存在するため、大容量データをチェックするにはコンピュータの力を借りることが効率的であり、精度向上の面でも期待できる。

3.190. In some cases, the best solution will be to move out-of-range or clearly inconsistent values into a special category, prior to deciding how such cases should be edited and classified. In this way, the pitfalls of introducing statistical biases are considerably reduced. But precautionary measures should also be defined and set for the fact that overambitious automatic editing programs may cause the so-called "corrected" data to be significantly flawed. In this respect, it would make sense to have an acceptable cut-off value for error rates at the enumeration area level. If a data scrutiny program finds that more than a certain percentage of the records in a particular batch have one or more serious problems, the whole batch should be rejected and subjected to human or fieldwork verification.

日本語訳

3.190 場合によっては、範囲外の値や明らかに不整合な値を特別なカテゴリに移動させた後に、このような事例について、どのようにエディットや分類をするべきかを決定することが最適解になるだろう。そうすることで、統計的バイアスという落とし穴をかなり回避することができる。しかし、行き過ぎた自動エディティングプログラムは、いわゆる「補正」データに重大な欠陥を生じさせる危険を伴っているという事実を考慮して予防策を講じておくべきである。この点において、集計地域レベルでの誤差率に対する許容できるカットオフ値を設定することは理にかなっている。データ精査プログラムが、特定のバッジ内に一つ以上の重大な問題があるレコードを一定割合以上発見した場合、そのバッチ全体の利用を中止し、人手又は実地調査による検証を受けるべきである。

項目番号【3.190】では、過度な自動エディティングにより、「補正」されたデータに重大な欠陥が生じる可能性を指摘しており、その防止策として、誤差率に対するカットオフ値の設定や、人手や実地調査による検証を行うべきとの記述が見られる。自動エディティングは、便利ではあるものの、過信しすぎてはいけないということである。

2　国勢調査エディティングに関するハンドブック

同じく国連統計部の発行する「国勢調査エディティングに関するハンドブック（改訂２版）（以下、ハンドブック）」では、国勢調査におけるデータエディティングについて詳細に説明がされており、各国統計局のデータエディティングの指南書と呼べるものである。各国統計局では、長い間、統計調査データには欠測値や不整合などの問題があることを認識しており、これら問題に対応するため様々な手段を講じてきた。しかし、国勢調査は実施が５年や 10 年ごとといった周期調査であるため、次回実施までの間隔が長く、その間にデータエディティングに関するノウハウが失われてしまい、新しい国勢調査のために、以前のデータ収集活動で使用された手順を作り直さなければならなかったのである。国連統計部では、それを未然に防ぐために、データエディティングの経験を文書化しハンドブックにすることで、各国へ提供しているのである。このハンドブックは、各国が自らのデータエディティングの経験を失うことなく蓄積できるように、現在の国勢

調査で実施された活動を記録し、将来の作業における重複を回避することを目的として作成されている。また、このハンドブックは、各国が自国の現在の統計状況に最も適したエディティングを実施することができるような内容になっており、立場の違う専門家同士が、エディティングプログラムを開発・実行する際に、より良いコミュニケーションを取ることができるようにも作られている。

このハンドブックではデータエディティングについて、どのような記述が見られるのかを以下で紹介したい。

（１）不詳の取扱

不詳の取扱については、Ⅱ章-D-３-項目番号【147】に以下の記述が見られる。

147. The editing team must decide early in census planning how to handle "not stated" or unknown cases. As noted earlier, columns or rows of unknowns in tables are neither informative, nor useful, so planners in most countries prefer these data imputed. Without treatment of unknowns, many users distribute the unknowns in the resulting tables in the same proportions as the known data, thus imputing the unknowns after the fact. The editing team needs to decide how to deal with the unknowns systematically.

日本語訳

147. エディティングチームは、国勢調査の計画の初期段階で、「未記載」又は「不詳」をどのように扱うかを決めなければならない。表中の不詳の列や行には情報がなく、有用でもないので、ほとんどの国の企画担当者はこれらのデータをインピュテーションさせることを希望している。不詳を（データエディティング内で）処理しない場合、多くのユーザーは不詳データを既知データと同じ割合で表中に按分させるため、不詳データは事後的にインピュテーションされる。エディティングチームは、不詳データを体系的にどのように扱うか決める必要がある。

日本の国勢調査では、不詳については不詳項目を作ることで「不詳数」を表章しているが、2015 年国勢調査からは参考値として、主な項目の集計結果に含まれる「不詳数」を既知データと同じ割合で表中に按分させる「不詳補完値」を、結果利用者の利便性向上を図るために公表しているところであり、項目番号【147】文章中にあるように事後的に不詳データをインピュテーションしている。

（２）手動によるエディティング VS コンピュータによる自動エディティング

Ⅲ章-B-項目番号【233】では、国勢調査のデータエディティングでは手動と自動のどちらが良いかについての記述が見られる。これは言うまでもなく、コンピュータによる自動エディティングが手動によるエディティングよりも有利であり推奨される手段である。

233. Manual correction inevitably lowers quality and consistency unless the respondent is contacted. It takes more time, and it costs more. Computers do not tire and are faster; they do not have personal problems that might interfere with maintaining quality or consistency; and, in most cases, they make processing cheaper. Most countries now use some kind of automatic correction.

日本語訳

233. 手作業による訂正は、回答者に連絡しない限り、どうしても品質や整合性が低下してしまう。時間もかかり、コストもかかる。コンピュータは疲弊しないし迅速である。品質や整合性の維持の妨げになるような人的な問題もなく、ほとんどの場合、処理にかかるコストが割安になる。現在では、ほとんどの国で何らかの自動訂正が行われている。

（3）妥当性及び整合性チェック

エディティングにおけるデータの妥当性及び整合性を検証する方法については、Ⅲ章-D-１-項目番号【248】【249】及びⅢ章-D-２-項目番号【254】にメインとなる記述が見られ、トップダウンエディット法（Top-down editing approach）と多変量エディット法（Multiple-variable editing approach）が紹介されている。

1. Top-down editing approach

248. The top-down editing approach starts with the first item to be edited (the "top"), which is usually the first variable on the questionnaire, and then moves through the items in sequence, until completing the edit of all items. The usual approach is to first take into consideration the response rates and the relative importance of the various items. Because of their importance, particularly in dynamic imputation, the edits usually start with sex and age. While the top-down approach does not completely preserve the relationships among the data items, it does provide an adequate framework to complete the edit.

249. During the editing process, some edits change the value for an item more than once. This procedure can introduce one or more errors into the dataset. An imputed value may be inconsistent with other data. Even when variables are dealt with sequentially, a particular variable should be edited against all other variables concurrently, if possible. For example, a child's age, imputed on the basis of the mother's age, may be inconsistent with the child's reported years of school or years lived in the district. In this instance, the age will be re-imputed until it is consistent. An imputed age is an intermediate variable until final assignment. In creating the edits, imputed intermediate variables should not be recorded as changes until the final assignment.

日本語訳

１．トップダウンエディット法

248. トップダウンエディット法は、エディットされる最初の項目（「トップ」:通常調査票上の最初の変数）から始まり、全ての項目のエディットを完了するまで項目順に検証する。通常の方法は、まず回答率と各項目の相対的な重要度を考慮することである。特に動的インピュテーション（ホットデック）では、その重要性から、エディティングは通常、性別と年齢から開始される。トップダウンエディット法はデータ項目間の関係を完全に保持するものではないが、エディティングを完了するための適切なフレームワークを提供する。

249. エディティングの過程で、ある項目の値を複数回変更するエディットがある。これにより、データセットに一つ又は複数のエラーが発生する可能性がある。インピュテーションされた値が他のデータと矛盾している可能性がある。変数が順次処理される場合でも、可能であれば他の全ての変数（項目）と同時にエディットする必要がある。例えば、母親の年齢を基にインピュテーションされた子供の年齢についてみると、報告された学年やその地区に住んでいる年数と整合性がとれない可能性がある。この場合、子供の年齢は、整合性がとれるまでインピュテーションを繰り返し実行するだろう。インピュテーションされた年齢は決定値までの中間値なのである。エディティングをする際には、インピュテーションされた中間値は決定値に至るまでの変更として記録するべきではない。

2. Multiple-variable editing approach

253. The "top-down" approach to census and survey editing which is the procedure that was introduced in Section 1 above, may not always give the best results—those that come closest to the real distribution of the variables. As indicated, the top-down approach, if applied without proper precautions, frequently causes problems in the edit.

254. Another approach is multiple-variable editing, which is based on the Fellegi-Holt system. This approach requires more computing expertise and computer power but probably obtains results that are closer to "reality"….In the multiple-variable editing system it is necessary to determine a set of positive statements to

test the relationship between the variables. Then, the edit tests each statement against the data in the household to see whether all statements are true. For any false statement, the edit will keep track, on an item-by-item basis, of invalid entries or inconsistencies. After all tests, the editing and imputation system must assess how best to change the record so that it will pass all edits. Editing teams usually use a minimum-change approach and change the smallest possible number of variables to obtain an acceptable record.

日本語訳
２．多変量エディット法
253. 国勢調査や他の統計調査のエディティングについては、トップダウンエディット法により、必ずしも変数の実際の分布に最も近い結果が得られるとは限らない。適切な予防策を講じずにトップダウンエディット法を適用するとエディットでしばしば問題が生じる。
254. もう一つの方法はフェレギ-ホルト方式に基づく多変量エディット法である。この方法は、より多くの計算の専門知識とコンピュータの力を必要とするが、おそらく「現実」に近い結果を得ることができるだろう。（中略）多変量エディットシステムでは、変数間の関係を検証するために、一連の肯定的なステートメントを定める必要がある。そして、エディットは全てのステートメントが真であることを確認するために世帯内のデータに対して各ステートメントを検証する。誤ったステートメントに関して、エディットは項目ごとに、無効な入力や矛盾を追跡する。全ての検証後、エディット・インピュテーションシステムは、全てのエディットに合格するために、レコードをどのように変更するのが最善かを評価しなければならない。エディティングチームは通常、最小限の変更方法を使用して合格レコードを得るために可能な限り少ない数の変数を変更する。

トップダウンエディット法とは、エディットされる最初の項目、調査票でいえば、最初に問われている質問項目から始まって、全ての項目のエディットを質問項目順で完了する方法である。つまり、項目は収集された順番にエディットされ、例えば、最初の調査項目が年齢であった場合、まずはそれをエディットし、その項目に基づいて次の調査項目である性別をエディット、更に次の調査項目である続柄は、年齢と性別の両方を使ってエディットをすることができるのである。トップダウンエディット法は従来よく使われていた方法であるが、項目順にデータエディティングをしていくことで、項目間での整合性がとれなくなってしまうことがあるため、同時に項目をエディティングすることが推奨されている。また、トップダウンエディット法は、項目番号【253】において必ずしも最良の結果が得られるとは限らないと記されている。

次に、今最も多くの国で採用されている方法である、多変量エディット法がある。ハンドブックにおいても、より多くの計算の専門知識とコンピュータの力を必要とするが、おそらく「現実」に近い結果を得ることができるだろうと紹介されている。多変量エディット法は、名前のとおり、変数（変量）間における条件式を定めて、多変数を同時にエディティングする方法である。現在、各国の国勢調査で実際に利用されている多変量エディット法には、メインとなるものが３つあり、それが最近隣法、フェレギ-ホルト法及び新インピュテーション法[4]である。

（４）インピュテーション方法

ハンドブックの付録Ⅶには、各種インピュテーション方法の説明が記述されており、その中から現在世界的に主流である最近隣法、フェレギ-ホルト法及び新インピュテーション法

[4] 最近隣法（Nearest-neighbor Imputation Method）及び新インピュテーション法（New Imputation Methodology）は両者ともに略称としてNIM と呼ばれることがある。これら３つ以外のインピュテーション方法としては、合格ユニットの平均値を不合格ユニットにインピュテーションする平均インピュテーション法（Overall Mean Imputation）、類似度を持ったレコードグループを作るために定義された各階級内における合格ユニットの平均値を不合格ユニットへインピュテーションする階級平均インピュテーション法（Class mean imputation）、合格ユニットの値を利用してインピュテーションが必要な変数を回帰分析によって求めるモデルベースインピュテーション法（Model-Based Imputation）などがある。詳細については、参考資料［8］United Nations Statistics Division (2019) “Handbook on Population and Housing Census Editing Revision 2 ANNEX VII - IMPUTATION METHODS”を参照されたい。

について紹介する。

11. Nearest-neighbor imputation or distance function matching assigns an item value for a failed edit record from a "nearest" passed edit record where "nearest" is defined using a distance function in terms of other known variables. This method can be applied within imputation classes. It is usually considered appropriate for continuous variables but can also be applied with non-numeric variables.

日本語訳

11. 最近隣法（Nearest-neighbor Imputation Method）又は距離関数マッチング（Distance Function Matching）は"最も近しい"合格レコードを不合格レコードの項目値にインピュテーションする方法である。"最も近しい"の判断は、他の既知の変数による距離関数を用いて行う。この方法は、インピュテーション階層内で適用することができる。通常は、連続変数に適用されるが、非数値変数にも適用が可能である。

20. The Fellegi-Holt edit and imputation method (Fellegi and Holt, 1976) considers all edits concurrently. A key feature of the Fellegi-Holt edit and imputation method is that the imputation rules are derived from the corresponding edits without explicit specification. For each failed edit record, it first proceeds through a step of error localization in which it determines the minimal set of variables to impute as well as the acceptable range(s) of values to impute and then performs the imputation. In most implementations, a single donor is selected from among passed edit records by matching on the basis of other variables involved in the edits but not requiring imputation. The method searches for a single exact match and can be extended to take account of other variables not explicitly involved in the edits. Occasionally no suitable donor can be found and a default imputation method must be employed.

日本語訳

20. フェレギ−ホルトのエディット・インピュテーション（Fellegi and Holt, 1976）は全てのエディットを同時に検討する。 フェレギ−ホルトのエディット・インピュテーションの主な特徴は、明示的な指定なしで、対応するエディットからインピュテーション規則を導き出すことである。エディットに対して不合格だった各レコードに関しては、まずエラーの場所特定を通して、インピュテーションする数値の許容範囲だけでなくインピュテーションする変数の最小セットを決定し、インピュテーションを実行する。多くのインピュテーションでは、エディットに関係するがインピュテーションを必要としない他の変数に基づいてマッチングすることにより、一つのドナーがエディットに合格したレコードから選定される。この方法は完全に一致するドナーを検索するが、エディットに明示的には関係しない他の変数を考慮するように拡張することもできる。適切なドナーが見つからない時は、既定のインピュテーション方法を使わなければならない。

21. The New Imputation Methodology (NIM) (Bankier, Luc, Nadeau and Newcombe 1996; Bankier, Lachance and Poirier, 1999) is similar to the Fellegi-Holt method in that it considers all edits concurrently, does not explicitly specify imputation actions and imputes from a single donor. For each failed edit record it identifies minimum-change imputation actions conditional on the potential donors available. This guarantees that a donor will be available. Unlike Fellegi-Holt, NIM first searches for donors and then determines minimum-change imputation actions. NIM searches for donors by matching, using all variables (including those potentially to be imputed) involved in the edits, and can be satisfied by near matches for numeric variables plus matches for most, but not necessarily all, other variables. Imputation actions based on each potential donor are determined and those that are minimum-change imputation actions are identified. The method also considers near minimum-change imputation actions; these can sometimes yield more plausible imputed records. Finally, one of the minimum-change and near minimum-change imputation actions is selected at random and the imputation is performed.

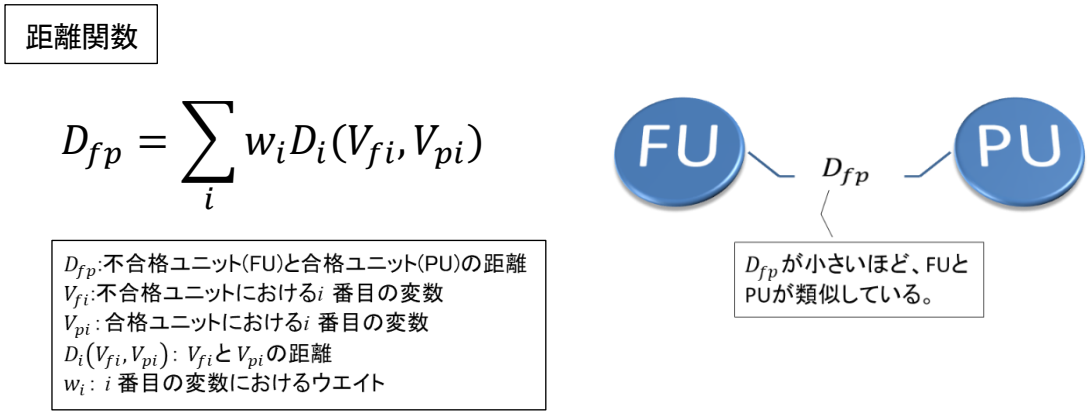
日本語訳

21. 新インピュテーション法（Bankier、Luc、Nadeau and Newcombe 1996; Bankier、Lachance and Poirier、1999）

は、一つのドナーからインピュテーションをし、インピュテーションアクション（IA）[5]を明示的に指定せず、全てのエディットを同時に検討するという点においてフェレギ-ホルト法に似ている。エディットに対して不合格だったレコードに関して、利用可能なドナー候補を条件として最小限の変更を行う IA を特定する。最小変化量の IA を特定することで、ドナーが利用可能であることを保証している。フェレギ-ホルト法とは異なり、新インピュテーション法は最初に利用可能なドナーを検索し、次に最小変化量の IA を決定する。新インピュテーション法は、エディットに関係する全ての変数（潜在的にインピュテーションされる可能性のある変数を含む）を使ってマッチングすることによってドナーを検索し、数値変数のほぼ一致と必ずしも全てではないがほとんどの他の変数の一致によって満たされる。各潜在的ドナーに基づいて IA は決定され、最小変化量の IA が特定される。この方法では、ほぼ最小変化量である IA も考慮される。これらの IA がより妥当なインピュテーションレコードを生成することがある。最後に、最小変化量及びほぼ最小変化量の IA のうちの１つがランダムに選択され、インピュテーションが実行される。

項目番号【11】に記述されている最近隣法（Nearest-neighbor Imputation Method）は、多くの国で利用されているインピュテーション方法である。最近隣法は、不合格ユニット（FU）に対してインピュテーションする際に選択される合格ユニット（PU）を、距離関数を用いて決定する方法である。距離関数の変数にはユニットの持つ特徴が現れるような項目をいくつか任意で選定し距離D_{fp}を算出する。この距離D_{fp}の数値が小さいものほど比較したユニットが類似していると見なすことができるため、一番値が小さいものをドナーとして、インピュテーションするのである。

図２　距離関数の説明図



項目番号【20】のフェレギ-ホルト法は、全てのエディティングを同時に検討し、各不合格ユニットに対して、インピュテーションする変数の最小セットと数値の許容範囲を決定し、インピュテーションする方法である。簡単に言えば、二つの項目にインピュテーションをしてエラーを解消するよりも、一つの項目のみにインピュテーションをしてエラーが解消されるのであれば後者を優先するという方法である。オリジナルのデータからの変更がより少ない方のインピュテーションを選択するのである。ただし、フェレギ-ホルト法にも欠点があり、その点については後述する。

項目番号【21】の新インピュテーション法は、全てのエディティングを同時に検討し、利用可能なドナー候補をまず選定して、それを不合格レコードへインピュテーションする方法である。フェレギ-ホルト法と異なる点は、最初に利用可能なドナーを検索することである。

[5] 付録【用語解説】④参照

この方法は、カナダ統計局によって開発されたインピュテーションシステム CANCEIS[6]で初めて世に出された。項目番号【22】では、特に大容量データにおけるエディット・インピュテーションでフェレギ−ホルト法よりも新インピュテーション法が優れていると評している。

22. Although both Fellegi-Holt and NIM are computationally demanding, efficient algorithms are available so that their implementation and application are feasible with modern computers. This is particularly true for NIM, which can readily handle somewhat larger editing and imputation problems than can the Fellegi-Holt method.

日本語訳
22. フェレギ−ホルト法と新インピュテーション法はどちらも計算量が多くなるが、効率的なアルゴリズムが利用できるため、最新のコンピュータでそれらの実装とアプリケーションを実行できる。特に新インピュテーション法は、大容量データにおけるエディット・インピュテーションの問題をフェレギホルト法よりも容易に処理できる。

実際にインピュテーションを実行する場合、人手では困難であるため、インピュテーションシステムを構築してコンピュータによって行うのだが、そのシステムについても項目番号【25】で言及している。それによるとほとんどの国家統計局のインピュテーションシステムでは、インピュテーション方法を組み合わせて使用しており、シーケンシャルホットデック法とフェレギ−ホルト法が主流であるとのことである。フェレギ−ホルト法を使っている国も新インピュテーション法に移行している、又は移行を検討している段階にあるようである。

25. In most imputation systems a mix of imputation methods is used; typically, deductive imputation is used where possible and is followed by one or more other procedures. Most national statistical offices use some form of dynamic imputation method for census editing and imputation. Sequential hot deck imputation and the Fellegi-Holt method are currently the most commonly used. Of the national statistical offices presently using the Fellegi-Holt method, one is changing to NIM and a number of others are considering it….

日本語訳
25. ほとんどのインピュテーションシステムでは、インピュテーション方法を組み合わせて使用している。通常、可能な場合には演繹的インピュテーションを使用し、その後に１つ以上の他のインピュテーション方法が続く。ほとんどの国の統計局は、国勢調査のエディット・インピュテーションに何らかの形式の動的インピュテーション法（ホットデック）を使用している。シーケンシャル・ホットデック法とフェレギ−ホルト法は、現在最も一般的に使用されている。現在、フェレギ−ホルト法を採用している国家統計局のうちの１つは新インピュテーション法に変更中であり、他のいくつかの国家統計局が新インピュテーション法への変更を検討している。（後略）

（５）データエディティングに関するコンピュータソフトウェア及びアプリケーション

データエディティングを実行するには、コンピュータソフトウェア（以下ソフトウェア）が必要不可欠であり、ハンドブックの付録Ⅷで説明がなされている。ソフトウェアを利用する利点として、全ての入力に対して標準的で信頼性の高い方法でデータを出力することができると記述があり、また、ソフトウェアを新たに開発するよりも既存の汎用性のあるソフトウェアを使うことの方が効率的で、開発で必要だった人員をエディティング作成に回すことができると項目番号【４】で記述されている。

ソフトウェアについては、一つのソフトウェアで全てのエディティングを完了させる必要はなく、例えばカナダのように、インピュテーションでは別のソフト（CANCEIS）を使って

[6] CANadian Census Edit and Imputation System の略称

いる国もある（項目番号【６】）。また、フェレギ-ホルト法が出現するまでは、ほとんど全てのエディットがトップダウンエディット法であったようである（項目番号【９】）。

4. One advantage of using editing software is that when properly used, data will be output in a standard, reliable fashion for all inputs. This may aid subsequent steps such as tabulations. Generic software packages such as SAS and SPSS, or other high-level languages, can be used to write editing programs in the customized fashion noted above. Or, a country can choose to use software applications written specifically for some stage of editing of census and survey data. For most countries, using an already developed and generalized edit software may be more efficient than developing a custom application, and allow expertise to be more fully focused on editing decisions as opposed to software development.

日本語訳

４．エディットソフトウェアを利用する利点として、適切に利用すれば、全ての入力に対して標準的で信頼性の高い方法でデータが出力されることが挙げられる。これは、集計などの後続のステップに役立つ可能性がある。SAS や SPSS などの一般的なソフトウェアパッケージ、又はその他の上位言語を上述のカスタマイズされた方法でエディットプログラムを書くために使用することができる。あるいは、国勢調査や他の調査のデータエディットのために特別に書かれたソフトウェアアプリケーションを利用することもできる。多くの国にとって、既に開発され汎用化されたエディットソフトウェアを利用することは、カスタムアプリケーションを開発するよりも効率的であり、専門知識をソフトウェア開発ではなくエディティングの作成により集中させることができるかもしれない。

6. Any editing software that a country might consider for use will need to meet a variety of requirements across entire edit process. For instance, it will need to produce reports for the various checks, tests and imputations required for editing census data. It should be noted that not all requirements need be satisfied by a single edit software package. In Canada, for example, edits are applied during the data capture phase by one software application, but imputation is carried out by another (CANCEIS, discussed below).

日本語訳

６．国が利用を検討するエディットソフトウェアは、エディットプロセス全体にわたって様々な要件を満たす必要がある。例えば、国勢調査データのエディティングに必要な様々なチェック、テスト、インピュテーションに関するレポートを作成する必要があるだろう。ただし、一つのエディットソフトウェアで全ての要件を満たす必要はない。例えばカナダでは、データ取得段階でのエディットはあるソフトウェアで行われるが、インピュテーションは別のソフトウェア（CANCEIS）で実施される。

9. As noted in the text, until the Fellegi/Holt (1976) method and its follow-ups, almost all editing used a top-down approach. That is, items were edited in order, usually – but not always – in the same order as they were collected. The first population item, for example, is usually relationship, so it would edited, then sex would be edited on the basis of that item, then age could use both sex and relationship, and so forth.

日本語訳

９．本文で述べたように、フェレギ-ホルト法（1976)が出現するまでは、ほとんど全てのエディットはトップダウン方式であった。つまり、通常（いつもではないが)、項目は収集された順番にエディットされたのである。例えば、最初の人口項目は通常、続柄であり、それをエディットし、次にその項目に基づいて性別をエディット、更に続く年齢は性別と続柄の両方を使ってエディットすることができるのである。

項目番号【10】～【13】では、カナダ統計局が開発し、利用しているインピュテーションシステム CANCEIS の優れた機能等が紹介されており、国連統計部は CANCEIS を非常に評価の高いインピュテーションシステムとして位置づけているようにみえる。以下に、その内容について紹介する。一部詳細については、Ⅱにおけるカナダの紹介で後述するが CANCEIS は、カナダ統計局の Mike Bankier 氏が開発した新インピュテーション法を利用したシステム

であり、汎用性やデータ処理能力の高さ、カナダ統計局における国勢調査での有効性が証明されたことなどから現在では各国統計局で利用されている。

前述した（4）にあるように、新インピュテーション法は現在、最も優れた方法だと考えられているが、ドナーとして用いることができるユニットがデータに十分存在していない場合や、他のインピュテーション方法との組み合わせが必要な場合には、引き続きフェレギ-ホルト法を利用した方が良いようである。

10. In the last few decades, several minimum change imputation systems based on Fellegi/Holt have been developed at Statistics Canada, each incorporating incremental improvements to either function or performance, and there have been several "precursors" to the current system used at Statistics Canada. These (and other) tools were developed based on another approach to minimum change (the NIM – new imputation methodology), by Mike Bankier at Statistics Canada. NIM was first implemented in CANEDIT then replaced by the CANadian Census Edit and Imputation System (CANCEIS) (Bankier 2005, Chen 2007). CANCEIS has been in use at Statistics Canada since 2001, and is used to process data after certain edits during collection and data capture have been applied such as validity edits in electronic questionnaires and checks for record duplication.

日本語訳

10. 過去数十年の間に、カナダ統計局では、フェレギ-ホルト法に基づく最小変化量インピュテーションシステムがいくつか開発され、それぞれに機能又は性能の漸進的な向上が盛り込まれ、現在、カナダ統計局で利用されているシステムの「前身」となっている。これらの（及び他の）ツールは、カナダ統計局の Mike Bankier 氏 による最小限の変化量に対する別の方法（新インピュテーション法）をベースに開発された。新インピュテーション法は、まずCANEDIT に実装され、その後 CANCEIS へと引き継がれた（Bankier 2005, Chen 2007）。CANCEIS は 2001 年からカナダ統計局で利用されており、収集やデータ取得時に一定のエディット（電子調査票の妥当性エディットやレコードの重複チェックなど）を行った後のデータ処理に利用されている。

11. In a review of the CANCEIS, the Canadian authors summarize: "CANCEIS, with its highly efficient editing and imputation algorithms, shows great promise for solving very general imputation problems involving a large number of edit rules and a large number of qualitative and quantitative variables when minimum change donor imputation is appropriate. The Fellegi/Holt minimum change edit and imputation algorithm, however, should still be the method of choice for smaller imputation problems if there may not be sufficient donors available or if it is more appropriate to use another method to perform imputation". (Bankier, Lachance and Poirier 2000. p.10) Since then, CANCEIS has proven itself within the Canadian Census context and is now used by other agencies as well.

日本語訳

11. CANCEIS のレビューにおいて、カナダの著者は次のようにまとめている。「CANCEIS は、その非常に効率的なエディット・インピュテーションアルゴリズムにより、最小変化量ドナーインピュテーションが適切な場合に、多数のエディットルールと多数の質的・量的変数に関係する非常に一般的なインピュテーション問題の解決に大きな可能性を示している。しかし、フェレギ-ホルト法の最小変化量エディット・インピュテーションアルゴリズムは、利用できるドナーが十分でない場合やインピュテーションを行うために他の方法を用いることがより適切である場合、より少量のデータのインピュテーション問題を扱う場合には、依然として選択されるべき方法である。」と述べている。CANCEIS はカナダの国勢調査でその有効性が証明され、現在では他の機関でも利用されている。

12. The 1996 Canadian Census (and others) used a different approach, called Nearest Neighbor (NIM). The 1996 version imputed responses for age, sex, marital status and relationship for all persons in a house simultaneously (Bankier 1999). The method was improved and expanded for the 2001 and subsequent Canadian statistical activities (Bankier et al, 2000, 2001) and CANCEIS now processes all Census variables for the Canadian Census of population (from both short and long form populations).

日本語訳
12. 1996 年のカナダの国勢調査（及びその他の調査）では、最近隣（Nearest-Neighbor）と呼ばれる別の方法が用いられた。1996 年版では、１つの住宅に住む全ての人の年齢、性別、婚姻関係、続柄に関する回答を同時にインピュテーションした（Bankier 1999）。この方法は、2001 年以降のカナダの統計活動において改良・拡張され（Bankier et al, 2000, 2001）、現在、CANCEIS はカナダの国勢調査の全ての変数を処理している（ショートフォーム及びロングフォームの両方）。

13. The NIM approach searches for nearest-neighbor donors first and then determines the minimum change imputation actions based on these donors. While the Fellegi/Holt method involves imputing the fewest variables and preserving the integrity of the subpopulations, the NIM, which reverses the order of the operation – starting with donors and then moving to minimum variables to change – provides computational advantage and is data driven. But the NIM can only carry out donor nearest neighbor imputation while Fellegi/Holt can be used with other methodologies (like top-down). Statistics Canada incorporated the NIM into its Canadian Census Edit and Imputation System (CANCEIS) for the 2001 and it has been part of its overall editing strategy since then.

日本語訳
13. 新インピュテーション法は、まず最近隣のドナーを検索し、次にこれらのドナーを基に最小変化量のインピュテーションアクションを決定するものである。フェレギ−ホルト法は、一番少ない変化量となる変数をインピュテーションし、部分母集団の整合性を保持するが、ドナー検索から始めて、最小変化量の変数を選定するという作業の手順を逆転させた新インピュテーション法は、計算上の利得をもたらすデータ駆動型である。しかし、フェレギ−ホルト法が他の方法論（トップダウンエディット法のような）と併用できるのに対し、新インピュテーション法は最近隣ドナーインピュテーションしか実行できない。カナダ統計局は、2001 年に新インピュテーション法を CANCEIS に組み込み、それ以来、全体的なエディット戦略の一部になっている。

Ⅲ　諸外国の国勢調査におけるインピュテーション方法

Ⅱを踏まえて、諸外国の国勢調査におけるインピュテーション方法について、各国統計局へメールにてヒアリング調査を行った。ここでは、ヒアリング調査の結果内容だけでなく、補足情報が必要だと思われるものについては、各国統計局の HP 等で公開している情報も含めて、インピュテーション方法について記述している。この調査によりわかったことは、多くの国家統計局で、やり方に違いはあるもののホットデック法を使うことで欠測値や不詳データの解消を実施しているということである。また、近年では欠測値や不詳データが存在しない行政記録情報を調査に使うことで、インピュテーションの必要性がかなり低くなっていると謳っている国もあった。近年の国勢調査の調査方法の変更により、インピュテーション方法の動向にも変化が現れてくるかもしれない。

1　アメリカ

アメリカのセンサス局（USCB）では、CSpro[7]と呼ばれる国勢調査などの統計調査のエディティングに特化したシステムを開発しており、現在、アメリカ以外の国でも利用されている。CSpro はセンサス局の HP から無料でダウンロードすることができるだけでなく、センサス局スタッフによる技術的な支援が行われており、多くの国で使われているシステムである。しかし、センサス局では国勢調査のエディティングに CSpro は使っておらず、センサス局で開発された別のホットデックシステムを使っており、名称も特に付けているわけではない。アメリカの国勢調査でエディット・インピュテーションに CSpro を使わない理由としては、CSpro にはコンピュータ支援による Web インタビュー機能[8]がないことを挙げている。現状では回答しなかった住宅には、まず調査員が来訪し、在宅の場合には携帯電話によるインタビューを実施している。

アメリカのインピュテーション方法は、どの情報が欠落しているかによって、インピュテーションを大きく２つに分けている。それが計数インピュテーション（Count Imputation）と特性インピュテーション（Characteristic Imputation）である。国勢調査データは、自己回答（自計申告）、無回答世帯への現地調査によるフォローアップ、その他様々な手続きを経て集められた後に、データ処理がされる。データ処理の過程で、住宅[9]の状態を、①有効居住住宅（a valid occupied housing unit）、②有効空住宅（a valid vacant housing unit）、③存在しない（無効）（nonexistent, that is, not a valid, livable housing unit）の３つに区分している。③には誰も居住していない商業目的の住宅や新築でまだ居住していない住宅などを含んでいる。①に関しては、もしその住宅に居住者がいれば、何人居住しているかを確認しようとする。ほとんどの国勢調査回答では通常 99%以上で状態や数が入力されている。しかし、少ない割合ではあるものの、状態や数を決定するのに十分なデータが存在しないことがある。この場合に、計数インピュテーションを実行するのである。計数インピュテーションはエディット前のファイル（Census Unedited File）作成時に実行され、居住していると確認された住宅（有効居住住宅）の人数に関する欠測情報を他のデータを用いて置き換える。計数インピュテーションを実行した後も、年齢や性別、人種などといった１つ又は複数の個々人の内容に関するデータが欠落しているレコードが存在し、中には住宅全体の全ての人に関する情報が欠落していることもある。それについては特性インピュテーションを実行するのである。

特性インピュテーションは、エディットされたファイル（Census Edited File）作成時に実行され、これは人口数が確定した後に行われる。そのため、人口数がこれ以降で増加することはない。また、特性インピュテーションは最近隣法を使用することで実行され、行政記録情報も品質向上の

[7] Census and Survey Processing System の略称 (URL https://www.census.gov/data/software/cspro.html)
[8] CAWI (Computer Assisted Web Interviewing)
[9] 住宅(Housing Unit)には、老人ホーム、兵舎、大学学生寮などを含む。

ため利用される。

特性インピュテーションには大きく分けてアサイメント（Assignment）とアロケーション（Allocation）の２種類が存在する。

アサイメントは、回答が欠落もしくは他の回答との矛盾があり、かつ、欠落項目の値が同一人物又はその世帯からの情報に基づいて決定できる場合に行われる。行政記録情報の利用はアサイメントで行われ、行政記録情報の値をそのままインピュテーションすることもある。全てのデータには何らかのエラーが含まれている可能性があり、行政記録情報もその例外ではないと考えられるが、行政記録情報における潜在的なエラーを処理する手段は今のところ存在しないということである。しかし、利用する行政記録情報と国勢調査の回答データの比較研究においては、高い一致率が既に検証されており、その研究では行政記録情報と国勢調査データの紐付けに個人 ID 番号[10]を利用したとのことであった。

アサイメントが終わると、次にアロケーションが実行される。アロケーションは、回答が欠落もしくは他の回答との矛盾があり、かつ、欠落項目の値が同一人物からの情報に基づいて決定できない場合に行われる。その場合、同じ住宅内の他の人や近隣の住宅の人からの回答が用いられる。アロケーションのうち、特別な方法としてサブスティテューション（Substitution）がある。サブスティテューションは、住宅内の全ての人の内容が欠落している場合に用いられる方法であり、まず、行政記録情報を利用して、住宅全体のデータを見つけ、もしデータが無い場合はホットデック法を用いる。その際、ドナーには当年国勢調査における近隣の回答データが用いられる。

センサス局のインピュテーション担当者へのヒアリングでは、センサス局のエディットは統計的な文献（最近隣法やフェレギ-ホルト法などの論文、以下、文献と呼ぶ。）が世に出る前に開発しており、それらとは異なるアプローチを取っているということであった。文献的には、エディットとインピュテーション[11]を別々に論じているが、センサス局では、エディットと言えばこの二つをセットにして考えている。それは国勢調査のエディットが複雑であり、多くのレコードに適用されるエディットの条件が多いためである。そのため、どのエディットを利用するかを決定する際にハンドブックを参考にすることや、統計モデリングによって最小変化量を勘案することもない。文献では、エディットに失敗しないために、エディットによる変更（数値の変化）を最小限に抑えようと考えているが、センサス局ではどんなケースにおいても変化量は意識していないということであった。また、極めて稀にではあるが、コールドデック法も使用しているとのことである。

図３　アメリカのインピュテーション手順

表　アメリカのインピュテーション方法

インピュテーション名	内容
計数インピュテーション(Count Imputation)	住宅に住む人の数に対するインピュテーション
特性インピュテーション(Characteristic Imputation)	住宅に住む人の内容（年齢、性別、ヒスパニック系等）に対するインピュテーション
アサイメント(Assignment)	同一人物又は同一世帯の値を用いてインピュテーション
アロケーション(Allocation)	住宅内の別人物又は近隣住宅の値を用いてインピュテーション
サブスティテューション(Substitution)	住宅内の全ての人の内容が不明の場合、近隣住宅の値を用いてインピュテーション

10 個人 ID 番号とは、マスター住所ファイル ID 番号（Master Address File Identification Numbers (MAFIDs)）と個人識別検証システムによって割り当てられたセキュリティ対策が講じられた識別キー（Protected Identification Keys (PIKs)）のことである。国勢調査と行政記録情報をマッチングするためには、MAFIDs と PIKs が両データに付与されている必要がある。詳細については、参考資料[10] United States Census Bureau (2012), “2010 Census Match Study” を参照されたい。

11 エディット及びインピュテーションの解説については、付録【用語解説】①及び②参照

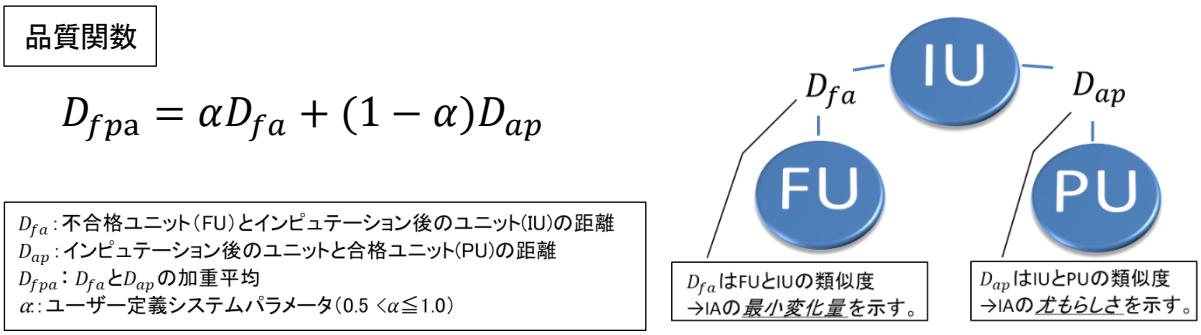
2　カナダ

カナダ統計局(StatCan)では、独自開発した CANCEIS というインピュテーションシステムを1996 年[12]より利用することで国勢調査のインピュテーションを実行している。CANCEIS はデータ駆動性に優れ大容量データであっても効率的に処理が可能であり、また、Windows 環境で使えるなど汎用性も高く、Excel やテキスト形式でモジュール[13]作成が可能であり、ユーザーの要望を満たすための新機能も迅速に追加可能であることなどから、世界各国で利用されている。

CANCEIS と他のインピュテーションシステムの違いは、処理の手順を通常と逆にしたことによる優れたデータ駆動にある。以前、カナダの国勢調査で利用されていたシステムでは、フェレギ−ホルト法を利用し、最小限の変更条件を満たすために、不合格ユニットにおけるインピュテーションすべき変数を最初に特定していた。どの変数をインピュテーションするかを決めた後に初めて、ドナー候補を検討するのである。この方法の問題点は利用可能なドナープールを考慮すると、最初のステップで特定された変数にインピュテーションするための適したドナーを見つけることがとても困難であり時には不可能なことである。対照的に CANCEIS は最近隣法により、まずドナー候補を検索して、不合格ユニットに最も類似したドナーを見つけ出し、そのドナーに対してインピュテーションアクションを検討するのである。２つのステップを逆にすることで、より優れたデータ駆動を実行することができ、計算上の大きな利得を得るのである。

さらに CANCEIS は、ドナー選定にも独自の原理を追加している。ドナー選定には通常、距離関数を用いて、不合格ユニット（以下、FU）に類似した合格ユニット（以下、PU）を検索して、距離D_{fp}の小さい PU をインピュテーション後のユニット（以下、IU）として採用している（図２）。それに加えて、CANCEIS では、インピュテーションの品質を担保するために、FU と IU の距離、IU と PU の距離を加重平均した値D_{fpa}を加味することで（品質関数と呼ぶ。）、ドナー選定を実行しているのである（図４）。この値が小さい IU をインピュテーションに使用することで、インピュテーションアクションにおける最小変化量や尤もらしさを担保でき、より優れたインピュテーションアクションを実行することができるようになるのである。CANCEIS は距離関数D_{fp}と品質関数D_{fpa}の二つの値を参照することで最良のインピュテーションアクションを実行しているのである。

図４　CANCEIS の品質関数



[12] 国連統計部発行のハンドブックには、CANCEIS はカナダ統計局で 2001 年から利用されているという記述がみられたが、カナダ統計局へのメールによるヒアリングでは、1996 年から利用しているとのことであった。

[13] モジュール(module)とは、CANCEIS を動かすために設定するエディットルール等を定めたインプットデータのことである。

３　フランス

フランス国立統計経済研究所（INSEE）では国勢調査のインピュテーション方法にシーケンシャルホットデック法によるシステムを長年にわたって利用している。このシステムでは、住所変数でソートしたファイルを使って、その並び順に地理的特徴がより近似であるドナーを探してインピュテーションするのである。メリットとしては、未回答の割合が低い場合や、データファイルがまとまっている場合に効率的なことである。デメリットとしては、ファイルが適切にソートされていない場合、地理的に相違のあるドナーを選択する可能性があり、インピュテーションの品質に影響を与えることである。現在のところ、インピュテーション方法について当該方法から変更する予定はないとのことである。

４　イギリス

イギリス国家統計局（ONS）では、国勢調査のインピュテーションに2011年からCANCEISを利用している。2011年と2021年の国勢調査におけるCANCEISの利用には膨大な検討がなされたとのことであった。CANCEIS は、ライセンスは必要であるものの、自由に利用できるソフトウェアであり、2001年にONSで利用していたソフトウェアよりも2011年の国勢調査へ向けたインピュテーション戦略の目的と目標をより良く満たすということで、ワークショップや外部レビューを通じて、利用が合意された。別のONS独自ソフトウェアの開発に割当てていたリソースを、CANCEIS プラットフォームで構築済の多くの統計的機能に移植することで代用できるという明確なメリットもあった。

また、CANCEISのプラットフォーム上で実装された方法論と処理戦略は、2021年国勢調査のために提案されたインピュテーション戦略の目的と目標を支持するフェレギ-ホルト法の原理を自然拡張したものであり、従来のトップダウン法よりも優れた多変量エディットシステムの例示として、国連統計部の発行するハンドブックにも掲載されていた。

CANCEISでは、ドナーインピュテーションに先立って、複雑な国勢調査のエディットルールを簡素化する独自のコンピュータ処理方法を用いている。それは、①重複を除くことで大きく処理速度を向上させる、②複数の独立したエディットが暗示的に組み合わされることによって、インピュテーションを不可能にする箇所を特定することができることである。インピュテーションの実行にあたり、CANCEISは広い範囲のユーザーが定義するパラメータを備えており、特定の分析目的や優先順位に沿うように基準となる値を調整したり、異なる調査から得られる異なったデータ様式に対応できるようにシステムを調整したりすることができる。与えられたパラメータ内において、CANCEISが他のシステムと比較して優れている点は、欠測値の置き換えや不整合の解消に求められる最小となるデータ変数を特定する多変量インピュテーションタスクの最適解を自動的に検出できる点である。

以上がメリットであるが、デメリットもあり、それは、①複雑なプログラムであるため、適切に利用するためには専門的な知識が必要であること、②（国勢調査に限らず）CANCEIS の利用は、全体として他の方法よりも時間がかかること（処理速度とインピュテーションの品質の関係がトレードオフ）、③大規模調査や国勢調査にCANCEISの利用を考えたとき、多くの変数の特有な組み合わせがあることで、全データからドナーを見つけ出すことが困難になることが挙げられる。③のような場合にはモジュール化して、同時に変数の部分集合のみでインピュテーションを実行するが、データの同時分布を歪めるリスクがあるとしている。

イギリス統計理事会（UK Statistics Authority）の国勢調査に関する方法論保証再調査研究班[14]で提出されたワーキングペーパーでは、2021 年国勢調査においてもインピュテーションに

[14] Methodological Assurance Review panel – Census

CANCEIS を用いるかどうかの検討がなされている。社会調査における SAS を利用した CANCEIS の機能代替プロジェクトでは、１年で実働可能な解決案（a working solution）に辿り着いたようだが、CANCEIS と比較するとその機能は限定的であり、チェック機能未実装、大規模データセットに必要な機能もほとんどないものであった。そのため、原理から自前で CANCEIS に代わるシステムを開発・設計するよりも 2021 年にもう一度 CANCEIS を利用し、2011 年の経験を行かすことの方が遙かに賢い選択であると結論づけたようである。

５　ドイツ

ドイツ連邦統計局（FSO）では、国勢調査（人口センサス）のインピュテーションに 2022 年[15]から CANCEIS を利用している。既に 2011 年の住宅センサスでは CANCEIS を利用していたのだが、当時、人口センサスには技術的な理由により適用されなかった。ドイツの人口センサスの調査方法は、行政記録情報を利用する複合型（Combined Census）であり、サンプルによる踏査を実施している。2011 年の人口センサスでは、その踏査を実施している 10%サンプル調査において、インピュテーションが実行された。インピュテーションの方法は、コールドデック法、演繹的（Deductive）インピュテーション法[16]及び最近隣法を組み合わせている。

2011 年人口センサスにおけるインピュテーションの手順を見ていくと、ステップ１で「性別」や「生年月日」といった変数のあり得ない値や欠測値について外部から入手した情報を用いてインピュテーション（コールドデック）、ステップ２では、演繹的インピュテーションを実行する。これは例えば、学校の種類とクラスレベルについて尤もらしい情報を持つ全ての回答者について、国勢調査の基準週に学校に行ったかどうかの質問項目における欠測値を「Yes」とした。最後のステップ３では、ステップ１及び２の後においても欠測値であるものについて、最近隣法によるインピュテーションを実行した。なお、インピュテーションツールは、ノルトライン・ヴェストファーレン州の中央統計・IT サービスプロバイダーである IT-NRW[17]によって開発・作成され、データの収集はドイツ政府によって行われるが、エディットとインピュテーションのプロセスの全てについては、IT-NRW の中央サーバーで行われた。

2022 年の国勢調査では人口センサスにおいても CANCEIS が利用されることになった。CANCEIS の大きな利点は、エディットとインピュテーションの同時実行と、処理速度の速さをもたらす効率的なアルゴリズムにあり、特にカテゴリ変数に対してうまく動作するということであった。しかし、調査やデータの状況によっては他のインピュテーションシステムの方が良い場合もあるとも言っていた。CANCEIS 以外にも、インピュテーションには R パッケージや SAS プロシジャを利用しているとのことだった。CANCEIS によるインピュテーションでは、該当データが少なくドナーとして利用できるユニットが存在しないなどの理由からインピュテーションが不可能である事例も出てくるため、それに対応するためにこれらソフトを利用していると考えられる。

今後については、様々なインピュテーション方法を使用していくとのことだった。インピュテーション方法の選定は、調査のデータ特性（欠測値の数、欠測パターン、エディットルールなど）に基づいて行っていくとのことである。既存の方法やシステムを継続しつつも、適切なツールが開発された際には、新しい方法を取り入れ、使えるツールを増やしていきたいとのことであった。

15 ドイツの国勢調査は、2021 年に実施予定だったが、新型コロナウイルス感染症の影響により 2022 年に延期された。
16 演繹的インピュテーション法とは、決定論的インピュテーションに含まれ、欠測値や一貫性のない値を確実に推定する方法である。多くの場合、調査票に記入がある他の項目の回答に基づいている。
17 Information und Technik Nordrhein-Westfalen の略称。IT-NRW はノルトライン・ヴェストファーレン州にあるドイツの公営企業

６　イタリア

イタリア国立統計研究所（Istat）では、国勢調査のインピュテーションにローマ大学と共同で開発したインピュテーションシステム DIESIS[18]を 2001 年国勢調査から利用している。DIESIS は、世帯レベルと個人レベルにおいて質的変数と量的変数の同時処理を可能にした。DIESIS には、「First donors then fields[19]」と「First fields then donors[20]」の二つのアルゴリズムによる 「データ駆動」と「(理論上の) 最小変化量」の二つのエディット方法が実装されており、これら二つのアルゴリズムが集められた情報の保持とインピュテーションアクションの妥当性のバランスを取るために併用されている。デフォルトでは「First donors then fields」アルゴリズムが設定されているが、ある不合格ユニットに対して出力された変更数が、「First fields then donors」アルゴリズムと比較して非常に大きかった場合、「First fields then donors」アルゴリズムに切り替えるというシステムになっている。

また、Istat では、国勢調査ではないが死因調査（Causes of death survey）において、バイタルソーシャル変数のチェックと訂正に CANCEIS を 2003 年から利用しているということであった。CANCEIS の利用に至った背景としては、より変化に柔軟で、より透明性と質の高いデータを提供することができ、CANCEIS に実装されている最近隣法が多くの方法のなかで最も自分たちの目的に合致していたため、自動チェック・訂正の新しいソフトウェアとして CANCEIS の利用を決定したとのことである。また、彼らの目標が、欠測値をインピュテーションし、不整合なデータを取り除くことであり、できる限り信頼のできる数値は変更しないというものであった。最近隣法を採用している CANCEIS は、これを満たしてくれるソフトウェアであり、生データと CANCEIS によるインピュテーションで得られたデータを比較しても、観測値の分布を歪めないことがわかったということであった。これはインピュテーション数が多い場合においても、同様であったようである。

７　オランダ

オランダ中央統計局（CBS）では、レジスターベースセンサス（Register-Based Census）[21]を実施しており、多くの変数が行政記録情報に基づくため、インピュテーションの必要性は、現在より大規模な調査を行っていた 25 年前よりもはるかに低いとのことであった。ただし、国勢調査の学歴項目は例外であり、行政記録情報が整備されていないため、多項ロジスティック回帰モデル（Multinomial Logistic Regression Model）を用いて欠測値をインピュテーションすることを計画している。調査データのインピュテーションが必要な場合は、集計の過程で固有のプログラムを作成しているとのことである。

８　オーストラリア

オーストラリア統計局（ABS）では、国勢調査のインピュテーションに自前で開発した最近隣法を実装したオラクルベースのシステムを使っている。このシステムを用いた国勢調査のインピュテーションは以前から実施されており、インピュテーション方法論もより良いものに改善し続けている（例えば、欠測値又は不整合値とドナーの適合基準)。また、このシステムは他のデータ処理システムと統合している。

インピュテーションは、①調査票が返ってこない場合、②一部記入のある調査票が返ってきた

18 Data Imputation and Editing System - Italian Software の略称

19 First donors then fields とは、最初に潜在的なドナーの部分集合を特定し、そしてそれらのドナーに基づいてインピュテーションする変数の最小量を決定するアルゴリズムである。

20 First fields then donors とは、最初にインピュテーションする変数の最小数を決定し、その後、潜在的なドナーを決定するアルゴリズムである。

21 付録【諸外国の国勢調査におけるインピュテーション方法一覧表】の（注）参照

場合に実行する。①については、主要な人口統計的変数（年齢、性別、婚姻状況、常住地及び従業地）の全てをインピュテーションする。残りの変数については「記述無し（not stated）」とする。②については、主要な人口統計的変数のうち回答のなかった部分のみインピュテーションを実行する。例えば、年齢を除く全ての主要変数への回答があった場合、年齢のみがインピュテーションされる。

9　ニュージーランド

ニュージーランド統計局（Stats NZ）では、国勢調査のインピュテーションに 15 年以上前から CANCEIS を採用しているということであった。CANCEIS 採用の理由としては、フェレギ-ホルト法よりもデータ駆動性が優れていることを挙げている。また、カナダ統計局による継続的な公式サポートが存在し、継続的に改良されている点を挙げていた。

インピュテーションに CANCEIS を利用するメリットとしては、汎用性とカスタマイズ性が高いことを Stats NZ は評価しており、同一モジュール内で、あらゆる種類の変数（例えば、数値変数、文字変数など）が含まれる大規模データセットのインピュテーションアクションが可能であることや、ほとんどのパラメータ（ドナー選択、距離関数、ドナーを特定するための決定ロジックなど）について、利用する調査の要件に合わせて調整が可能であることを挙げている。また、利用したドナーやインピュテーションされたデータの詳細が記述されたレポートがアウトプット時に出力されるため、インピュテーション処理の品質評価がしやすいというメリットもある。デメリットとしては、CANCEIS を利用すると、短期間で多くを学ぶ必要があることや、大規模なデータセットでは実行に長時間（数時間）かかることを挙げているが、これらに関しては、データセットの事前調整や、パラメータの調整で緩和ができる場合もあるとしている。

今後のインピュテーション方法については、CANCEIS を標準的なツールの一つとして、使い続けていく可能性が高いが、多重インピュテーション法（Multiple Imputation）[22]を実装することも検討しているとのことであった。

10　中国

中国国家統計局（NBS）は、2020 年国勢調査（第 7 回国家人口センサス）において、初めて自己回答を導入[23]し、携帯端末で QR コードを読み取ることで回答できるようにしたことや、WeChat と呼ばれるアプリを利用したりするなど、ICT の活用に力を入れていた。また、調査員が各世帯を訪ねて情報を登録する際も、タブレットやスマートフォンを活用し、さらに、国民識別番号を調査票の記入事項とすることで、戸籍記録との照合ができるようになり、品質管理対策として利用した。

NBS では、データ収集基盤システムを開発することで、国勢調査の進捗状況をリアルタイムで監視することができるようになり、回答データを行政記録情報と照合し、データのエラーや欠測値を効果的に削減できるようにしたとのことである。また、データ収集のプログラムにデータ検証基準を搭載したことで、欠測値やエラーが大幅に減少し、欠測値に関して言えば、例えば住宅面積では、回答世帯の近隣における住宅面積の平均をインピュテーションに使うだろうということだった。

[22] 多重インピュテーション法（Multiple Imputation）とは、インピュテーションによる分布の歪みを防ぐことを目的として、インピュテーションを要する各値に対して、複数回インピュテーションを実行することで、歪みの少ないデータセットを選定し対処する方法である。

[23] 従来は、調査員が調査票に記入する他計申告であったが、2020 年国勢調査では自計申告が導入された。

Ⅳ　国勢調査におけるインピュテーション方法の国際比較

国勢調査におけるインピュテーション方法については、ヒアリング調査したほとんどの国でホットデック法（最近隣法）を利用していることがわかった。また特筆すべきなのが、カナダ統計局の開発した CANCEIS であり、その汎用性や性能の高さから多くの国で利用されていることがわかった。

CANCEIS を利用していない国の中で特色のある国としては、イタリアが CANCEIS に似たようなシステムである DIESIS を開発しており利用している。オランダは、調査方法がレジスターベースセンサスへ移行しており、行政記録情報を利用した調査であるため、インピュテーションの必要性は低いという回答であった。また、中国については、データ収集プログラムにチェック機能を組み込んだことで欠測値やエラーの発生が大幅に削減されたとのことであった。

アメリカについては、CSpro というエディット・インピュテーションシステムを開発してはいるものの、自国の国勢調査では利用しておらず、フェレギ−ホルト法などが出現する前から独自に開発したホットデックインピュテーション法を用いているとのことであった。そのため、最小変化量などを考慮するよりもエディットルールを適用したいと考えており、担当者は、「Our approach is procedural in nature.（我々の方法は事実上、手続き的なものである。）」と言っていた。

【諸外国の国勢調査におけるインピュテーション方法一覧表】

	アメリカ	カナダ	フランス	イギリス	ドイツ	イタリア	オランダ	オーストラリア	ニュージーランド	中国
調査方法	伝統的センサス	伝統的センサス	ローリングセンサス	伝統的センサス	複合型センサス	複合型センサス	レジスターベースセンサス	伝統的センサス	伝統的センサス	伝統的センサス
調査周期	10年	５年	毎年	10年	10年	毎年	10年	５年	５年	10年
インピュテーション方法（システム）	ホットデック法（最近隣法）	CANCEIS	シーケンシャルホットデック法	CANCEIS	CANCEIS	DIESIS	必要に応じて適切なインピュテーション方法を採用	ホットデック法（最近隣法）	CANCEIS	必要に応じて適切なインピュテーション方法を採用
詳細内容	▸欠測の内容によって、インピュテーション処理を計数インピュテーション（人数）と特性インピュテーション（内容）に分けている。 ▸可能な限りのデータを収集した後、わずかな欠測値、無効データ、不整合データに対してエディットや特性インピュテーションを実施する。 ▸エディットでは無効データや不整合データを発見し、特性インピュテーションで欠測値をインピュテーションする。 ▸総人口数が確定した後にエディットや特性インピュテーションを実施するため、総人口数には影響しない。 ▸エディットルールを適用する際に最小変化量は意識していない。 ▸未回答住戸には調査員を派遣し、在宅の場合には、携帯電話を使って聞き取り調査を実施	▸CANCEISは大規模データのエディットインピュテーションを効率的に処理可能にしたシステム ▸CANCEISは、まずドナー候補を探索し、そのドナー候補に対しインピュテーションアクションを検討することで優れたデータ駆動を実行することができ計算上の大きな利得を得る。 ＜今後の展望＞ ▸国勢調査のため、引き続きCANCEISを利用していく。 ▸CANCEISは今後もエラーが見つかればその都度改善することで性能を向上させ、新機能も導入していくだろう。	▸住所変数でソートされたファイルから順に地理的により類似の特徴を持ったドナーを探索する。 ＜メリット＞ ▸未回答率が低い場合やデータファイルが住所変数によって正確にソートできる場合には効率的 ＜デメリット＞ ▸データファイルが正確にソートできない場合は、地理的にはるか遠くのドナーを選択する可能性があり、インピュテーションの質に影響を与える。 ＜今後の展望＞ インピュテーション方法の変更は、現状では計画にない。	＜CANCEIS採用理由＞ ▸ライセンスは必要だが大規模データに特化して設計されたインピュテーションシステムを無料で利用可能 ▸CANCEISの機能及び基礎となるインピュテーション方法が、2001年以前の国勢調査においてイギリス国家統計局が設計・開発したシステムよりも優れていた。 ▸国連発行の国勢調査エディティングハンドブックでも従来のトップダウン法より優れた多変量エディットシステムとしてCANCEISが例示されている。 ＜今後の展望＞ ▸国勢調査でのCANCEISの利用は将来の調査計画や評価段階における意思決定過程の一部である。	＜CANCEIS採用理由＞ ▸CANCEISを採用した一番の理由は、エディットとインピュテーションの同時実行が可能、処理速度も許容範囲で効率的なアルゴリズムであるため。 ＜今後の展望＞ ▸今後も様々なインピュテーション方法を利用していく。 ▸インピュテーション方法の選定は調査データに依存する。 ▸現状のインピュテーション方法やシステムを維持した上で、適切なツールが開発されれば、それを取り入れていくだろう。	▸DIESISでは、世帯レベルと個人レベルにおいて、質的変数と量的変数を同時に処理可能 ▸DIESISには、"First donors then fields"と"First fields then donors"の２つのアルゴリズムによる「データ駆動」と「（理論上の）最小変化量」の2つのエディット方法が実装されている。 ▸これら二つのアルゴリズムが、データ保持とインピュテーションアクションの妥当性とのバランスを取るために、併用されている。	▸現在では多くの変数が行政記録情報由来であるため、大規模調査をしていた25年前よりもインピュテーションの必要性がかなり低い。 ▸行政記録情報から情報をほとんど得ることが可能であるが、唯一の例外として、「学歴」に関する行政記録情報は未完成である。 ＜今後の展望＞ ▸「学歴」については、多項ロジスティック回帰モデルを使ってインピュテーションをする計画がある。	▸最近隣法を備えた独自のオラクルベースのインピュテーションシステムを利用している。これまでの国勢調査でも利用されおり、メソドロジーも経時的に改良されている（例えば、インピュテーション対象とドナーの適合基準など）。 ▸このインピュテーションシステムはオーストラリア統計局の他のデータ処理システムと統合されている。 ＜今後の展望＞ ▸2026年国勢調査のインピュテーション計画については、まだ何も決まっていない。	＜CANCEIS採用理由＞ ▸CANCEISのインピュテーション処理はフェレギ-ホルト法よりもはるかにデータ駆動性が優れている。 ▸CANCEISはカナダ統計局による公式なサポートがあり、継続的な改善がある。 ▸汎用性とカスタマイズ性が優れている。 ＜今後の展望＞ ▸CANCEISを標準的なツールの一つとして今後も使用し続ける可能性が高いが、多重インピュテーションに関する研究もしているところである。	▸第７回国家人口センサス（2020年）においては、デジタル手法を用いてデータを収集 ▸データ収集プログラムにデータチェック機能を組み込んだことで、データ欠測やエラーの発生が大幅に削減された。 ＜今後の展望＞ ▸例えば、住宅面積の欠測値については、近隣の回答世帯の平均値をインピュテーションに利用するだろう。

（注）調査方法の種類

1．伝統的センサス：調査員調査のことであり、紙媒体及び/又はインターネット調査票を使用した実地調査に基づく調査方法である。登記情報や行政記録情報を活用することもあるが、あくまで補助的な使用であり、直接的に調査項目を把握するために使用するものではない。

2．レジスターベースセンサス：登記情報や行政記録情報のみに基づき、踏査を実施しない調査方法。ただし、国勢調査を目的としていない既存の調査結果を活用することは可能である。（例：労働力調査の結果を国勢調査に利用）

3．複合型センサス：登記情報や行政記録情報から直接的に調査項目を把握するほか、悉皆又はサンプルによる踏査を実施し、調査項目を把握する調査方法である。

4．ローリングセンサス：累積的な連続したサンプル調査のことであり、長期間に渡って全国全てを網羅する調査方法である。

Ⅴ　おわりに

コンピュータ技術が今よりも進歩すれば、さらに複雑なインピュテーションアクションが可能となり、処理速度の向上も見込めるようになる。統計調査では調査から公表までの期間が定められているため、それに間に合うようにデータエディティングを計画していかなければならず、時間的な制約が課されてしまっている。時間的な制約があることにより、精度向上において必要性の高い処理は優先して実施するものの、必要性の低い処理はどうしても後回しになってしまい、場合によっては時間切れになってしまうこともある。逆に言えば、データエディティングの必要性があることで、国勢調査のような大規模データを扱う調査では、公表までの時間をかなり要するのである。統計調査結果のニーズは、調査から時間が経てば経つほど、世間の関心が薄れて低くなるため、鮮度の高いうちに公表する必要がある。コンピュータの技術革新、いわゆるブレークスルーが起きれば、例えば量子コンピュータのような処理速度が現行のコンピュータに比べて飛躍的に向上したものが使えるようになれば、国勢調査結果の更なる公表の早期化は夢ではない。

2020 年に実施された日本の国勢調査は、調査が始まってちょうど 100 年目に当たる節目の年でもあった。100 年前の国勢調査はどうだったかというと、調査事項が「氏名」「世帯主との続柄」「男女の別」「出生の年月日」「配偶の関係」「職業及び職業上の地位」「出生地」「民籍または国籍」の８項目であり、現在も「出生地」以外は、調査事項として存在している[24]。第１回国勢調査は 1920 年に実施されたのだが、その実施の根拠となる法律「国勢調査ニ関スル法律」は 18 年も前の 1902 年に制定されていた。法律制定から実施までに 18 年の年月を要した原因は、日露戦争や第一次世界大戦の影響もあったが、そのほかに日本には元々、明治政府による国家的な戸籍制度が 1872 年以降存在していたため、わざわざ莫大な予算を使ってまで国勢調査を実施することに疑問を抱いていた人たちが少なからずいたことにある。しかし、当時の戸籍による推計からでは年齢や職業などを把握することができず、また、届け出の間違いなどもあり正確な人口を捉えることができないという問題もあった。実際に第１回国勢調査を実施してみると、戸籍人口の方が国勢調査人口よりも 196 万人ほど多く、調査員による踏査の重要性がはっきりとわかったのである。

今現在、世界各国の国勢調査の方法は、新型コロナウイルス感染症流行の影響もあり、非接触で実施できる行政記録情報を用いた調査に関心が移り変わりつつある。そのため、インピュテーションの必要性は今後低下していくのかもしれないが、上記のように調査員調査の重要性は 100 年も前から指摘されていたことであり、現在住んでいる場所を把握できる常住地ベースの調査は、調査員調査でないとなかなか難しい。行政記録情報では、一時的な移動（住民票を移さずに何らかの理由で転居）を把握することが難しく、行政記録情報による住所に実際には当人が居住していなかったケースも存在する。既に 100 年前の国勢調査から、戸籍のような登記情報（行政記録情報）と大規模調査員調査の違いは指摘されており、登記情報よりも大規模調査員調査の方が、有用性が高かったからこそ 100 年もの間、調査員による国勢調査が継続されてきたと考えられる。調査環境の悪化や費用対効果を追求するあまり精度を犠牲にした調査方法の転換はするべきではないと考えるが、これについても技術革新や新制度の導入が助け船となり、より良い方法で国勢調査が実施できるようになる日が来るかもしれない。

[24] 出生地は 1955 年調査で調査事項から削除され、「出生の年月日」は年月に 1965 年調査で変更された。2020 年調査の調査事項は 19 項目であり 100 年の時を経ておよそ倍になっている。

付録

【用語解説】

主に本文中で使用した用語について解説する。

① データエディティング（Data-Editing）

データエディティングとは、データの誤りを検知し、訂正することを目的とした作業のことである。エディットとは、データ中の誤りを検出し処理することをいうが、訂正処理であるインピュテーションとは分けて論じられることが多い。国連統計部の発行する「国勢調査に関する原則及び勧告」において、インピュテーションはデータエディティングの項目で紹介されていることからインピュテーションはデータエディティングに含まれると考えられる。

② インピュテーション（Imputation）

インピュテーションは、エディットにより検知された欠測値や、無効又は一貫性のない回答に関する問題を解決するための処理である。不整合な値又はユニット[25]に、整合性のある値（ドナーインピュテーションの場合は、ドナー[26]の値）を代入する作業のことであり、エディット中のレコード（一つ又は複数）にある欠測値や回答の一つ又は複数を代入により置換することで、レコード内で尤もらしい整合性のとれたデータにするための作業である。回答者への接触や調査票の人手による解析は、処理の早い段階で欠測値や不整合などの問題を解決するが、回答者の負担、費用及び適時性の問題があるため一般的にその実現は不可能である。インピュテーションの種類は下表のとおりである。

表　インピュテーションの種類

インピュテーションの種類	内容
ディタミニスティックインピュテーション （Deterministic imputation）	一つしか正しい値が存在しない場合、例えば、法定婚姻年齢未満の場合は配偶の関係が「未婚」に限られるように、同じ調査票から真の回答を導き出すことができることで使われるインピュテーション
モデルベースインピュテーション （Model based imputation）	平均値、中央値、回帰式などで算出した値をインピュテーション
デックインピュテーション/ ドナーインピュテーション （Deck imputation/Donor imputation）	欠測値を解消するためにドナーが使われるインピュテーション。ドナー探索には、類似性の高いドナーを探索する最近隣法がしばしば利用される。
ミックスインピュテーション （Mixed imputation）	上記インピュテーション法を混合したインピュテーション。例えば、最初にディタミニスティックインピュテーションを実行し、それがうまくいかなかった場合、デックインピュテーションを実行する。

（引用文献）United Nations Statistics Division (2019) “Handbook on Population and Housing Census Editing Revision 2”

③ エディットルール（Edit Rule）

チェックルール（Checking Rule）とも呼ばれ、データグループ又はデータ項目の値が正しいと判断するために必要な閾値や論理的な条件のことである。単にエディットを通過などという場合は、エディットルールに合格した値を指す。

④ インピュテーションアクション（Imputation Action）

インピュテーションアクションとは、インピュテーションを実行する際の処理作業のこと

[25] ユニット(Unit)とは、構成単位、一郡という意味があり、例えば国勢調査の場合、１レコードには個人の続柄、性別、国籍などが入っており、１個人に属するレコードの項目が１つであることが少ないため、これら全ての項目を含める呼び名が必要であるためユニットと言っている。

[26] ドナー（Donor）とは、代入する元となる値やユニットのことである。ドナーに対してレシピエント（Recipient）という単語も使われることがあるが、これはドナーがインピュテーションする値であるのに対し、インピュテーションされる側の値又は欠測値のことである。

である。例えば、ドナー候補を検索して、どのドナーを欠測値又は不整合な値にインピュテーションするかを決定する作業をインピュテーションアクションと言う。

⑤ 合格ユニット（Passed Unit）

合格ユニットとは、エディットルールに合格した（エディットルールを通過した）値（ユニット）を意味する。

⑥ 不合格ユニット（Failed Unit）

エディットルールに不合格だった（エディットルールを通過できなかった）値（ユニット）を意味する。

⑦ ホットデック法（Hot-Deck Technique or Dynamic Imputation）

コンピュータによるインピュテーション方法の一つであり、欠測値、誤値、無効及び不整合のある項目に対して、同じデータセットから整合性のとれる値をインピュテーションする方法である。インピュテーションされるたびにデータセットが更新されることから動的インピュテーションとも呼ばれ、現在、最も利用されているインピュテーション方法である。インピュテーションに使われる整合性のとれた値（ユニット）はドナーと呼ばれ、欠測値や不整合である項目に最も類似した値（ユニット）である（最近隣法）。また、ホットデック法のうちシーケンシャルホットデック法（Sequential Dynamic Imputation）とは、データファイルをレコード順に走査していく中で、エディットに合格した値を保持しておき、エディットに失敗したエラーデータに対しては保持している直近の合格値を用いてインピュテーションする方法である。

⑧ コールドデック法（Cold-Deck Technique or Static Imputation）

コールドデック法とはホットデック法と同様、コンピュータによるインピュテーション方法の一つであり、静的インピュテーションとも呼ばれる。これは、エディティングプログラムにより、あらかじめ決められたデータセットから欠測値に対して特定の回答を割り当てるか、又は有効な回答の分布を使って比例的にインピュテーションする方法である。インピュテーションに使うデータが固定され、時間とともに変化しないので静的インピュテーションとも呼ばれる。しかし、ホットデック法のように、最近隣法を使うわけではないので、矛盾した項目や無効な項目に使用することが困難であり、また、過去の国勢調査データや信頼できる参照データがない限り実行することができない。

【参考資料】

［1］坂下信之（2017 年）　「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，2017 年 7 月

［2］坂下信之（2018 年）　「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，2018 年 7 月

［3］坂下信之（2019 年）　「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，2019 年 8 月

［4］坂下信之（2020 年）　「統計調査の欠測値補完方法に関する研究動向について（主に米国とオランダ）」，リサーチペーパー第 48 号，2020 年 9 月

［5］総務省統計局(2018 年)「国勢調査 100 年のあゆみ」

［6］I. P. Fellegi and D. Holt (1976) “A Systematic Approach to Automatic Edit and Imputation,” Journal of the American Statistical Association, Vol.71, No.353, pp.17-35

［7］United Nations Statistics Division (2017) “Principles and Recommendations for Population and Housing Censuses Revision 3”
https://ec.europa.eu/eurostat/cros/content/handbook-sdc_en

［8］United Nations Statistics Division (2019) “Handbook on Population and Housing Census Editing Revision 2”
https://unstats.un.org/unsd/publication/SeriesF/seriesf_82rev2e.pdf

［9］United States Census Bureau (2021), “How We Complete the Census When Demographic and Housing Characteristics Are Missing”
https://www.census.gov/newsroom/blogs/random-samplings/2021/08/census-when-demographic-and-housing-characteristics-are-missing.html

［10］United States Census Bureau (2012), “2010 Census Match Study”
https://www.census.gov/programs-surveys/decennial-census/decade/2010/program-management/cpex/2010-cpex-247.html

［11］Statistics Canada (2020), “CANCEIS USER’S GUIDE Version 5.4”

［12］UK Statistics Authority (2020), National Statistician’s Advisory Committees and Panels, Methodological Assurance Review panel – Census, Papers, EAP110-2021-Census-Editing-and-Imputation-Strategy.docx
https://uksa.statisticsauthority.gov.uk/the-authority-board/committees/national-statisticians-advisory-committees-and-panels/methodological-assurance-review-panel-census/papers/

［13］Australian Bureau of Statistics (2021) “How the data is processed”
https://www.abs.gov.au/census/guide-census-data/census-methodology/2021/how-data-processed

［14］Lydia Spies(2017) “Possible imputation procedures for the Census 2021,” Work Session on Statistical Data Editing, CONFERENCE OF EUROPEAN STATISTICIANS, UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE, 24-26 April 2017
https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2017/mtg2/Paper_10_Possible_imputation_procedures_for_the_Census_2021.pdf

［15］OECD “GLOSSARY OF STATISTICAL TERMS”
https://stats.oecd.org/glossary/

［16］UNECE “Censuses of the 2020 round,”　Plans and practices of UNECE countries for the population and housing censuses of the 2020 round
https://statswiki.unece.org/display/censuses/Censuses+of+the+2020+round（2022 年 10 月 1 日現在）