国勢調査の補完方法に関する一考察

北原　昌嗣†

A Consideration of Imputation Methods for Population Census

KITAHARA Masatsugu

2020 年国勢調査では、結果利用者の利便性を図るために、主な項目の集計結果に含まれる「不詳」をあん分等によって補完した「不詳補完値」を算出し、参考表として公表しているが、「あん分」による不詳の補完が可能な統計表は一部に限定され、不詳を補完できない統計表が残るなどの課題がある。

また、不詳への対応は、各国の国勢調査（人口センサス）共通の課題であるが、主要国で「あん分」により不詳を補完している国は、我が国のみであり、我が国を除く主要国においては、不詳を補完するソフトウェアを集計プロセスに実装することで、原則として不詳のない結果が提供されており、我が国においても同様の対応が可能か否かを検討している。

本稿では、主要国で用いられている補完方法のうち、カナダ統計局の開発した補完システム「CANCEIS」を利用して、不詳を補完した結果を 2020 年国勢調査のデータを用いて試算し、その補完結果をあん分等による不詳補完値と比較することで考察した。

キーワード：国勢調査、補完方法

In the 2020 population census, in order to improve the convenience of users, the "Imputed Values" were calculated by imputing the "unknown" included in the tabulation results of major items by proportional distribution, etc., and published as reference tables. However, there are only a limited number of statistical tables that can impute unknowns by proportional distribution, and there are still issues such as the remaining statistical tables that cannot impute unknowns.

Furthermore, dealing with unknowns is a common issue for population censuses in all countries, but Japan is the only major country that imputes the unknown by proportional distribution. In major countries other than Japan, results with no unknowns are provided in principle by implementing software in the tabulation process that imputes unknowns, and we are currently considering whether a similar approach is possible in Japan.

In this paper, we used "CANCEIS" developed by Statistics Canada, which is one of the imputation systems used in major countries and calculated imputed values by using this system with data from the 2020 population census. And then, we compared and validated the imputation results by proportional distribution, etc. and by CANCEIS.

Key words: Population Census, Imputation method

† 総務省統計局統計調査部国勢統計課　　Email: mkitahara@soumu.go.jp

1　はじめに[1]

近年、単身世帯や共働き世帯の増加、オートロックマンションの増加などにより、調査員が調査世帯に容易に会えないことで、調査票の回収に支障をきたすケースが増えてきている。国勢調査では、調査票未回収世帯については、近隣等の住民へ聞き取りを行うことで、その世帯を把握していたが、それでは把握できる調査項目が限定されてしまうため欠測値が増加し、集計結果の不詳が増える。不詳が増えると結果の利用にも悪影響が出てくるため、結果利用者の利便性向上のためにも 2015 年国勢調査から、不詳を除いたデータ構成比を利用した「あん分」により不詳を補完[2]した結果を参考表として提供している。

2020 年国勢調査では、不詳をあん分等[3]により補完する項目を 2015 年国勢調査よりも拡大している。2015 年国勢調査では、人口等基本集計における年齢・国籍項目の不詳を補完した一部結果表のみ作成し、参考表として公表していたが、2020 年国勢調査では、他の集計区分の項目においても、不詳を補完した結果を公表している。さらに、2020 年国勢調査では、あん分の際に利用するデータ構成比の偏りを解消するため、個票レベルでホットデックや行政記録情報（在留外国人登録データ）等を利用して推定が可能な年齢・国籍をあん分処理前に部分的に補完している[4]。

あん分を用いて不詳を補完する方法は、国連統計部の発行する「人口・住宅センサスの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）（Handbook on Population and Housing Census Editing Revision 2）」にも、データエディティング内で補完を実施しなかった場合の事後的な補完方法として紹介されており、一般的な不詳を補完する方法である。また、その補完結果においても、原数値[5]と比較して、行政記録情報に近似した数値が算出されている。例えば、図 1-a の人口ピラミッドでは、年齢階級別に国勢調査人口と住民基本台帳人口の数値を比較しているが、若年層を中心に国勢調査人口と住民基本台帳人口の数値に差が見られる。

図 1-a　人口ピラミッド（原数値）

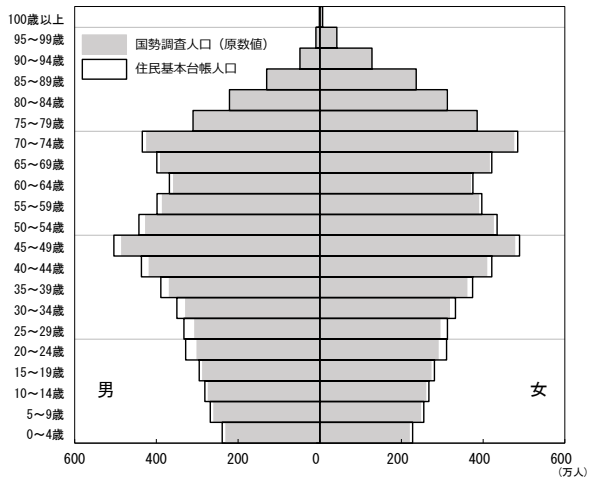
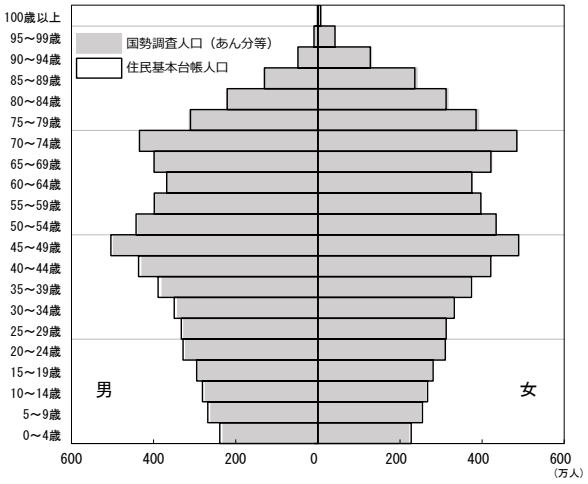


図 1-b　人口ピラミッド（あん分等）



---

[1] 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。

[2] 「補完」とは、データに欠測値の存在があり、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、「補定」や「インピュテーション（Imputation）」とも呼ばれている。なお、本稿では、結果表レベルのあん分による事後的な不詳の処理についても「補完」と称している。これは、①結果表レベルのあん分による不詳の処理、②個票レベルの欠測値の処理という２つの性格の異なる処理について、どちらも英語では「Imputation」が用いられているためである。

[3] 2020 年国勢調査では、あん分の事前処理として、ホットデック等による部分補完を実施しているため、「あん分等」としている。

[4] 詳細な作成方法については、総務省統計局（2021）を参照されたい。

[5] 本稿では、国勢調査の公表値を原数値としている。

一方の図 1-b では、国勢調査のあん分等による不詳補完値[6]と住民基本台帳人口の比較になっており、その差が縮小していることがわかる。両者とも全国結果による比較である。

国勢調査人口と住民基本台帳人口が異なる要因は、調査時点及び調査方法にある（表１）。調査時点（住民基本台帳では、集計時点）については、１月１日と 10 月１日という違いがあり、調査方法については、国勢調査は、調査員が実地調査をしているのに対し、住民基本台帳は、市区町村への届け出を集計したものとなっている。そのため、住所において、住民基本台帳では、住民が移動届を出さずに移動している可能性もあり、短期の移動では、住民票を移さずに生活している可能性もある。例えば、その可能性があるのは、通学のために親元を離れて生活する学生や単身赴任者等である。

このような違いから国勢調査人口と住民基本台帳人口を比較する際には注意が必要であり、時点を線形補間などによって揃えることや、住民基本台帳人口では住民が移動している可能性があるため、比較する範囲を全国レベルにするなどの対応が必要になる[7]。

表１　国勢調査人口と住民基本台帳人口の比較

	国勢調査	住民基本台帳※1
調査（集計）周期	５年に１度	毎年
調査（集計）時点	10月1日	1月1日
住所	【常住地】 同一の場所に３か月以上にわたって住んでいるか、又は３か月以上にわたって住む予定になっている場所	【生活の本拠】 民法第22条と同様、各人の生活の本拠をいう。「生活の本拠」とは私的生活の中心地を意味する。
調査方法	調査員による実地調査	市区町村に届け出された住民票の集計
共通する集計事項	男女 出生年月 国籍	男女 出生年月日 国籍※2

※１　住民基本台帳に基づく人口、人口動態及び世帯数（総務省自治行政局）
※２　平成 25 年調査から「外国人住民」の区分を追加し、「日本人住民」と「外国人住民」の計を「総計」として記載

上記はあん分等による不詳補完値と行政記録情報を比較した一つの例であるが、一般的に不詳補完値と行政記録情報を比較すると、原数値よりもお互いが近接した結果になる。それは国勢調査結果に存在していた不詳が、あん分等により、適した項目に配分された結果であり、行政記録情報と近接した結果になるということは、あん分等による不詳補完値の妥当性を担保するものである。

一方で、あん分等による不詳補完値では、結果表レベルでのあん分による集計になっているため、①粒度に限界があること、②集計区分間の整合性が取れないこと、③個票レベルの補完ではないため二次的利用に使えないことが課題として残っている。これら課題を解決する方法として、主要国で実施されているソフトウェアによる個票レベルでの補完が挙げられる。本稿では、カナダ統計局が開発した補完システムである CANCEIS（キャンサイス）を利用することで、個票レベルでの補完処理を実施し、CANCEIS による不詳補完値を試算、そして、その補完値とあん分等による不詳補完値との比較検証を行った。

[6] 不詳を補完した値を不詳補完値という。
[7] 図 1-a 及び図 1-b における国勢調査人口と住民基本台帳人口の比較に当たっては、①住民基本台帳人口を線形補間により 10 月１日時点に調整、②地理的範囲を広範にするため全国結果で比較した。①については、調査（集計）時点が異なるためであり、②については住民基本台帳人口が登録人口であることから、住民の移動を考慮し、人口移動の影響を最小にした（国外移動については、考慮していない）。

2 カナダ統計局の補完システム

カナダ統計局の人口センサス補完システムである CANCEIS[8]は、カナダ統計局に所属していた Mike Bankier 氏によって開発されたシステムで、30 年以上も前に開発された補完システムであるにも関わらず未だ現役で利用されている。そして、その利用はカナダだけに留まらず、イギリスやドイツなどの人口センサスにも利用されており、ハンドブックにも補完システムの実例として記載が見られる。動作原理は、最近隣法（Nearest-neighbor Imputation Methodology (NIM))であり、欠測値のあるユニット[9]へ、距離関数により算出された距離[10]が最も近い欠測値のないユニットを代入する。システムの特徴としては、①数値、カテゴリ変数及び英数字のドナー補完を同時に実行できること、②データ駆動性に優れ大容量データであっても効率的処理が可能であること、③調査データ側の問題に対する解決策を容易で正確に定義可能であること、④ソフトウェアが Windows 環境で使用できるなど高い汎用性、⑤Excel やテキスト形式でモジュール[11]の作成が可能であること、⑥システムは動的であり、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が挙げられる。

CANCEIS は前述したように補完方法に最近隣法を利用しているが、一般的な補完システムでは、フェレギ-ホルト法(The Fellegi-Holt method)を利用している。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、欠測値を持ったユニットに対して、まず補完対象となる変数（調査項目）の最小セットと数値の許容範囲を決定し、ほとんどのケースではマッチング代入法により補完を実行している。すなわち補完対象となる変数や数値の許容範囲を始めに決定することで、補完による変化を最小限にしているのである。両者ともそれぞれに弱点があり、最近隣法では、ドナーとなるデータが不十分な場合、当たり前ではあるが補完処理が困難になる。一方のフェレギ-ホルト法では、始めに補完する変数を特定してしまうため、最適なドナーを見つけることが困難になってしまう場合がある。このフェレギ-ホルト法の弱点を CANCEIS の最近隣法では、まずドナー候補を検索することから始め、その後に補完箇所を特定するという処理順を逆にすることで克服し、より優れたデータ駆動を実現している。

さらに、CANCEIS では、ドナーを選定するときに、単に最近隣法を実行するだけでなく、フェレギ-ホルト法のように最小変化量も考慮した上で、ドナー選定を行っており、一般的な最近隣法では、欠測値のない整合性の取れたユニット(Passed Unit (PU))と欠測値のある不整合なユニット(Failed Unit (FU))の距離を見て、最も近い距離のドナー(PU)を選定してくるだけなのだが、CANCEIS では、補完後に実現したユニットとの距離も考慮することで、最小変化量や補完処理の尤もらしさ（いわば補完の品質）も検討し、ドナー選定を実施している[12]。CANCEIS については、そのシステムの完成度の高さから、世界各国で利用されている。

CANCEIS による不詳の補完では、FU 全体を最短距離により選定された PU 全体で置き換えることで補完をするのか、それとも FU 内の不詳項目だけを補完するために、PU から該当項目だけを取り出して補完するのかを設定することが可能であり、今回の試算では、最小変化量を考慮するためにも後者を設定している。

---

[8] CANCEIS は、CANadian Census Edit and Imputation System の略。

[9] ここでいうユニット(Unit)とは、世帯（単身を含む）単位のデータ群のことである。CANCEIS による補完処理で利用する１世帯当たりのデータを１ユニットと呼ぶ。また、サブユニット(Subunit)とは個人単位のデータ群のことであり、１ユニットの中に世帯人員分のサブユニットがあるということになる。

[10] ここでいう距離とは、日常で用いるようなユークリッド距離とは異なり、データ項目間の相関関係を考慮した距離となっている。

[11] モジュール (Module)とは、CANCEIS を実行するために必要な変数や距離関数、エディットルール等を Excel シートやテキスト上で設定したもの。

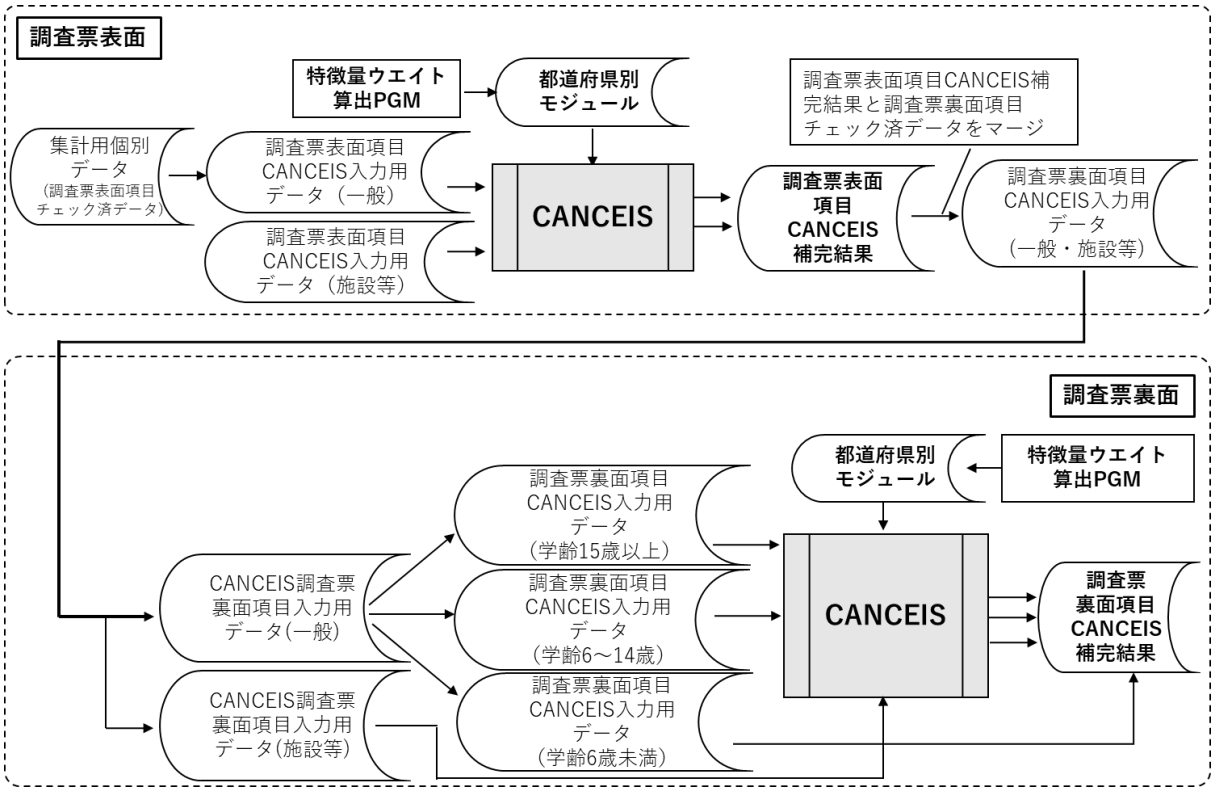
[12] 詳細な説明については、北原・寺垣内（2022）参照。

３ CANCEIS による不詳補完手順

CANCEIS を実行するに当たってのジョブフローは、図２のとおりである。全国の国勢調査のデータを一度に補完するともなると、データ容量が大きすぎるため、実行は困難である。そこで、都道府県ごとのデータに分割し、CANCEIS を各々実行することとした。また、それにより、各々で特色の違いがある都道府県内での補完というある種のグルーピングもでき、ドナー検索上の利得にもなる。CANCEIS による補完では、同じ種類のデータをグルーピングし、実行することが推奨されており、適切なインプットデータの作成が、質の高い補完結果を算出するために必要となる。さらに今回、一回の補完処理での対象補完項目数が多いと、処理速度低下の要因にもなることから、調査票表面項目（以下、表面と呼ぶ。）と調査票裏面項目（以下、裏面と呼ぶ。）を分割し、それぞれに CANCEIS を実行することで、CANCEIS への負荷を分散している。

なお、CANCEIS による補完を表面裏面別々に実行する理由は他にもあり、それは、①表面だけでも十分な距離関数に使う変数が得られる、②裏面は主に就業状態に関する項目であり表面と調査内容が異なる、③表面裏面を同時に補完処理することになるとドナーとして利用できるユニットが減少する[13]などである。

図２ CANCEIS ジョブフロー



表面の CANCEIS による補完のジョブフローは図２の上段であり、まず始めに都道府県別の集計用個別データ（調査票表面項目チェック済データ）をグルーピングのため一般世帯と施設等世帯に分割、一般世帯に関しては、さらに世帯人員ごとにデータを分割しグルーピングを行っている。そして、それぞれのデータを CANCEIS インプットデータ用に変換[14]する。一般世帯は世帯単位の補完、施設等世帯は個人単位の補完にしている。CANCEIS 実行にはインプットデータだけで

[13] 表面と比較して、裏面の不詳（欠測値）は多く、ユニットに１つでも欠測値があるとドナーとして利用することができないため、結果として、補完不可能ユニットが増加する。

[14] CANCEIS による補完処理に必要な調査項目を集計用個別データから抽出し、CANCEIS で扱える属性変数の型に変換する。CANCEIS で扱える属性変数の型はカテゴリ変数、数値型（離散型及び連続型）、英数字文字列型の３種類。

なく、補完によって不整合データや外れ値が生じないように、制御をするためのモジュールが必要である。モジュールには、ドナー選択時に必要な距離関数で利用する特徴量ウエイトも設定する。特徴量ウエイトとは、具体的にいえば、距離関数の項として使用している変数間距離の係数である（式(1)）。その特徴量ウエイトには、国勢調査の集計用個別データを用いて、R パッケージの randomForest により算出された特徴量重要度を利用している [15]。randomForest は機械学習アプリケーションであり、本家のカナダ統計局でも CANCEIS の距離関数に用いる特徴量ウエイトには、機械学習アプリケーションにより算出した値を利用している [16]。

距離関数

$$D_{fp} = \sum_{i} w_i D_i(V_{fi}, V_{pi}) \quad \cdots \quad (1)$$

D_{fp}: FU と PU の距離　V_{fi}: FU における i 番目の変数　V_{pi}: PU における i 番目の変数
$D_i(V_{fi}, V_{pi})$: V_{fi} と V_{pi} の距離　w_i: i 番目の変数における特徴量ウエイト

一方の裏面の補完では、表面の補完結果をインプットデータとして用いた（図２下段）。表面と同様に、始めに一般世帯と施設等世帯にデータを分割し、その上で、一般世帯では、「学齢 15 歳以上」「学齢６～14 歳」及び「学齢６歳未満」にデータ分割した。この分割は、裏面が就業状態を調査項目としているためであり、原則的に就労可能である学齢 [17] とそれ以外の学齢にグルーピングすることで、インプットデータの段階で効率的にドナー検索ができるようにしている。そして、「学齢６歳未満」については、幼稚園又は保育園などに通っている学齢となり、その場合、調査票上の質問項目である「９月 24 日から 30 日までの１週間に仕事をしましたか」という問に対しては、回答選択肢の「その他」に記入することになり、そこで回答は終了となることから、それ以降の質問に答える必要がない。そのため、「学齢６歳未満」については、CANCEIS による補完を実行する必要がなく、演繹的に「その他」に補完することができる。さらに、「学齢 15 歳以上」については、世帯人員別にデータ分割をし、グルーピングをしている。この場合の世帯人員とは、学齢 15 歳以上に限定した世帯人員のことである。例えば、夫婦と子１人の３人世帯の場合、子が 15 歳未満の場合は、２人世帯となり、子は学齢に応じて「学齢６～14 歳」「学齢６歳未満」のどちらかのグループに入る。また、「学齢 15 歳以上」は世帯単位での補完、「学齢６～14 歳」「学齢６歳未満」及び「施設等世帯」は個人単位での補完とした。

一般世帯を「学齢 15 歳以上」「学齢６～14 歳」及び「学齢６歳未満」にデータ分割したことで、一般世帯のグループにおける世帯人員が減少するため、世帯内の特徴量ウエイトは当然のことな

[15] 一部の特徴量ウエイトについては、特徴量の性質から判断することで値を設定している。例えば、同じ市区町村からドナーを持ってきた方が良いと考えられる場合は、地理に関係する特徴量ウエイトを他の特徴量ウエイトと比較して、相対的に大きくしている。

[16] カナダ統計局の CANCEIS 担当者へヒアリングしたところ、多くの場合、特徴量ウエイトの選択は、専門知識によって決定され、いくつかの項目には特定のモデルがあるとのことであった。そして、ウエイトの算出には、randomForest に似た ReliefF という機能選択アルゴリズムを使用しており、局内における検証では、均一のウエイトやモデリングなしで提案されたウエイトを使うよりも、機械学習によって得られたウエイトを使用した方が、精度、正確性及び再現性の点で、大幅なプラスの影響があることがわかったという。例えば、全てのウエイトを 1 に設定したとすると、最終的な人口分布が類似していたとしても、個人レベルでの補完結果は悪化するため、機械学習により算出された特徴量重要度（ウエイト）を使用することは、CANCEIS ドナー補完の特徴量ウエイトを取得するためのより有望な方法の１つであるとのことであった。

[17] ここでいう学齢とは、調査年である 2020 年から出生の年を引いた値（出生の年が 2020 年を除く）であり、出生月=4～12 の場合には、さらに-1 の処理を行っている。例えば、2000 年 1～ 3 月生まれの場合は、学齢 20 歳、 2000 年 4～12 月生まれの場合は、学齢 19 歳となる。

がら減少する。特徴量ウエイトが減少すると、距離関数によるドナー選定に影響がでるため、それを補うために、「学齢 15 歳以上」では、世帯内における「学齢 15 歳未満の人数」を、「学齢６歳未満」では、「世帯主の職業」などを特徴量として追加している。また、裏面では、「従業地・通学地」を調査しているため、「単位区中心点座標」を特徴量に追加することで、ユークリッド距離による影響も距離関数として取り入れている。

また、モジュールには、DLT（Decision Logic Table）という不整合データや外れ値が増加しないよう制御するためのコードを入力できる機能が備わっており、今回の試算では、この DLT に 2020 年国勢調査のチェックで用いられた主なエラー条件（エディットルール）を組み込むことで、不整合データ等、要するに適当でないデータが補完によって生成されないようにしている。

なお、インプットデータは、前述のとおりにデータ分割をしただけではなく、最終的に CANCEIS に入力する前にユニット（世帯）内で男女や続柄等の項目でソートをかけたものを使用している[18]。CANCEIS による補完処理では、ユニット間の類似性を検索するために、ユニット内での並び順を見ており、それが整列されていないと類似したユニットではないと CANCEIS は結論づけてしまうのである。例えば、４人世帯において、続柄が「１（世帯主），２（世帯主の配偶者），３（子），３（子）」の並び順のユニットと「２，３，３，１」の並び順のユニットでは、世帯構成は同一であるにも関わらず、CANCEIS は両者が類似したユニットであると判断できないのである。後者をソートにより「１，２，３，３」という並び順にすることで初めて、CANCEIS は両者が類似したユニットであると判断できるようになる。

次節では、上記手順により試算した CANCEIS による不詳補完結果を紹介する。

４　CANCEIS による不詳補完結果[19]

2020 年国勢調査結果を用いて、全ての集計区分[20]で CANCEIS による不詳補完値を主に愛知県[21]の国勢調査データを利用して試算した結果について、原数値及びあん分等による不詳補完値と比較することで CANCEIS による不詳補完値を検証した。

4.1　人口等基本集計

人口等基本集計は、市区町村別の人口、世帯及び住居に関する結果並びに外国人、高齢者世帯等に関する結果について集計している。

4.1.1　年齢

図 3-a 及び図 3-b は、男女・補完方法・各歳別人口のグラフであり、原数値、あん分等による不詳補完値（以降、グラフや表上では「あん分等」）及び CANCEIS による不詳補完値（以降、グラフや表上では「CANCEIS」）を比較している。このグラフを見ると、男女ともに、「原数値」と比較して、「あん分等」と「CANCEIS」が近接した数値になっていることがわかる。

---

[18] どの項目をソートするかは、表面と裏面のインプットデータで変更している。インプットデータの構成や、補完対象がユニットかサブユニットかなども考慮した上で、ソートをする項目を決定している。国勢調査データを確認することにより得られたデータの特性を利用して、CANCEIS がユニット間の類似性を見出せるような並び順にしている。

[19] 本節の CANCEIS による不詳補完値は、総務省統計局「国勢調査」の調査票情報を用いて、CANCEIS version5.4 により独自に算出し、集計したものである。

[20] 全ての集計区分とは、人口等基本集計、就業状態等基本集計、従業地・通学地集計及び人口移動集計のことである。抽出詳細集計は、抽出による推計値であるため除いている。また、試算に使用した調査項目は、簡易調査項目に限定している（教育や利用交通手段を補完対象から除いている。）。

[21] 主に愛知県で試算、検証を行った理由は、三大都市圏であることや、データ数も東京ほど多くなく、産業・職業データについても比較的バランス良く得られるためである。4.2.3、4.4.2 及び 4.4.3 を除き、データは全て愛知県の国勢調査データを利用して試算している。

図 3-a　補完方法・各歳別人口（男性）

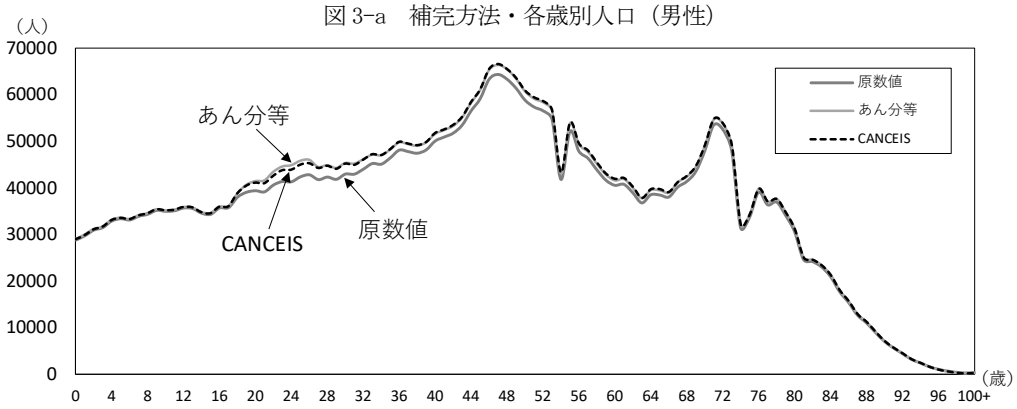
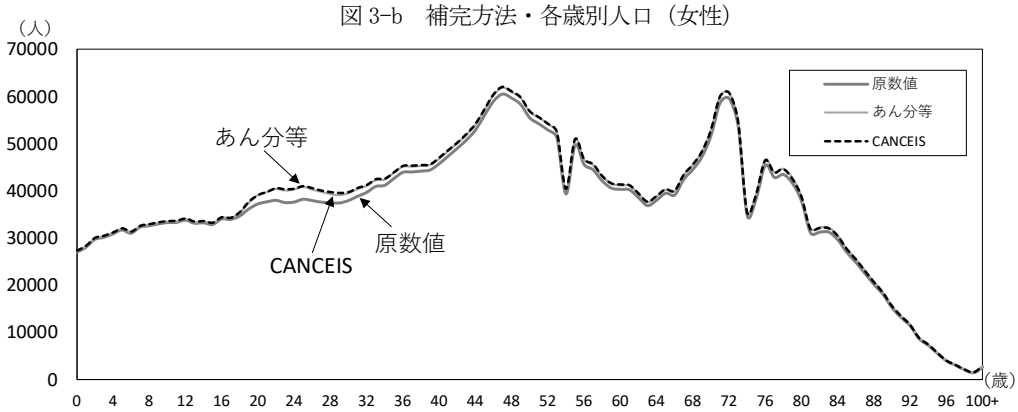


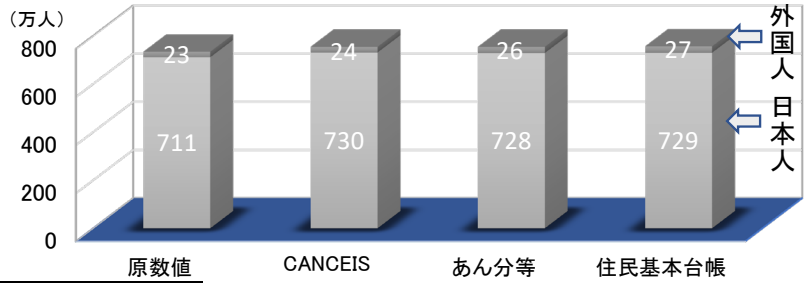
図 3-b　補完方法・各歳別人口（女性）



4.1.2　国籍

図４は、補完方法・国籍別人口のグラフである。これを見ると、「CANCEIS」では、「あん分等」や「住民基本台帳」の数値と比較すると、日本人が増加し、外国人が減少していることがわかる。これは外国人に補完すべき不詳が日本人へ補完されてしまった結果と推測できる。このような事象が起きてしまう原因として考えられるのが、元々の国勢調査データが、日本人が大多数を占める不均衡データだということである。要するに、不均衡データであるがために、ドナーとして外国人を選定する確率が低くなってしまっているのである。そのため、対応策として考えたのが、CANCEIS による補完処理前に部分的な補完を、行政記録情報を利用して実施することであった。それにより、ドナーとして不足していると考えられる外国人データを補填することができる。こういった事前補完は、「あん分等」でも実施している手法であり、イギリスにおいても、CANCEIS による補完が困難な世帯に対して、事前に部分的な補完をしてから、CANCEIS を実行している[22]。

図４　補完方法・国籍（日本人・外国人）別人口



[22] イギリス国家統計局(ONS(2022))によると、世帯の続柄は、CANCEIS において重要な変数である上、高い確率で回答エラーが発生すると予想されており、①少数派人口における信頼性の低下、②ドナー不足による CANCEIS 補完不可能ユニットの増加が懸念点として挙げられていた。そのため、CANCEIS 実行前に続柄項目を補完した上で、CANCEIS を実行している。

今回、外国人において、CANCEIS による補完の弱さが見られたことから、行政記録情報である在留外国人統計データを利用して、推定が可能な単身の外国人について、部分的な補完を事前に実施した上で、CANCEIS による補完を実施した。その結果が図５、図 6-a 及び図 6-b である。

これらの図を見ると、「CANCEIS」は部分補完により、「あん分等」と同程度の結果になることがわかった。年齢別にみてもそれは顕著である。ESRI (2017) では、基本的に CANCEIS のような最近傍マッチング代入法（最近隣法）を原理とする補完方法では、原数値が MNAR (Missing Not at Random)の場合、補完をしたとしても、欠測バイアスを解消することができないと言及しており、その影響がでたものと考えられる。今回、それを解消するために行政記録情報を利用し、推定が可能な単身の外国人については、CANCEIS 実行前に事前補完を実施したのである[23]。

図５　補完方法・国籍（日本人・外国人）別人口

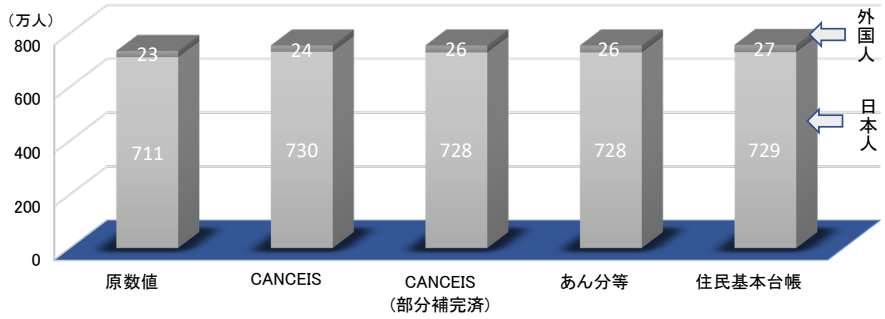


図 6-a　補完方法・各歳別人口（外国人・男性）

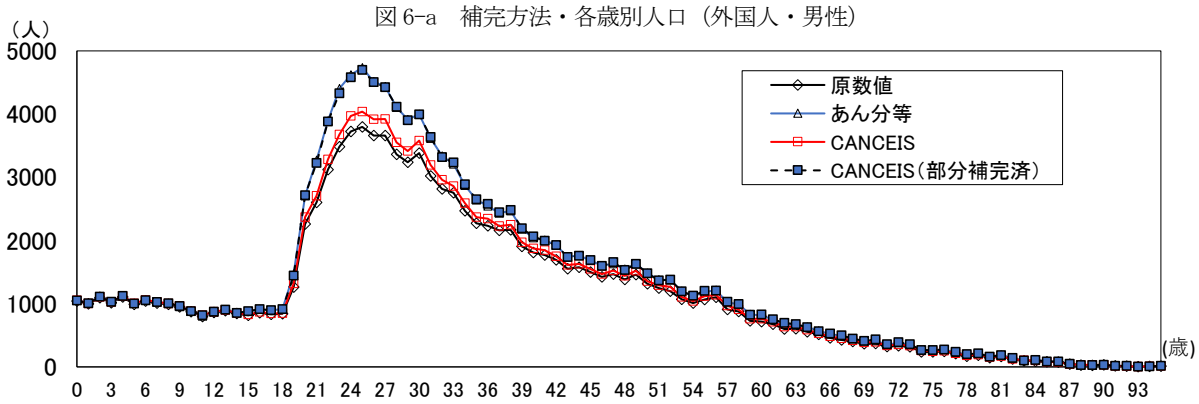
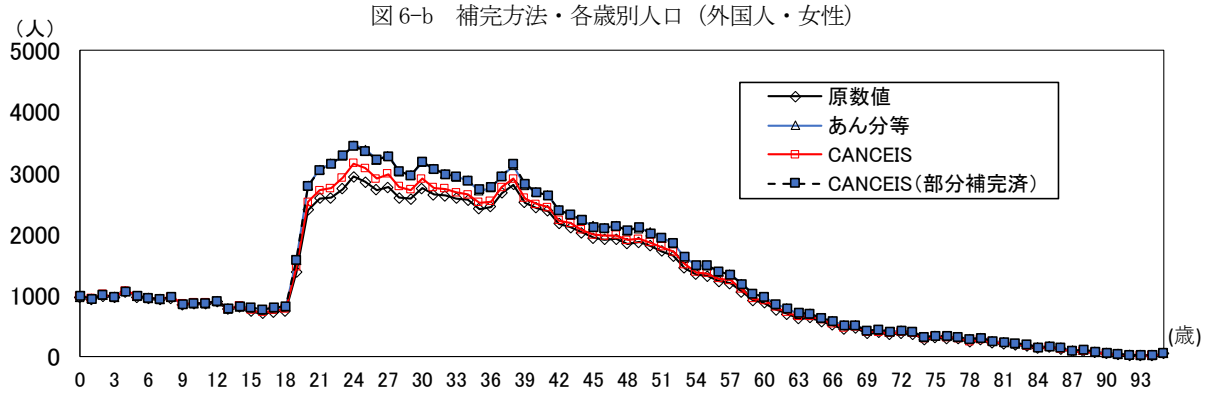


図 6-b　補完方法・各歳別人口（外国人・女性）



[23] 国勢調査の年齢や国籍の不詳は、事後調査などの検証結果より、若年層や外国人といった属性に多いとされることから、国勢調査結果は MNAR と考えることができる。CANCEIS は、前述したようにモジュールによるドナー補完の制御機能が実装されており、整合性や外れ値を考慮して補完を実行するため、単純な最近傍マッチング代入法とはいえないことから、若年層への欠測バイアスについては、大きな影響を受けなかったものと考えられる。

4.1.3　配偶関係

図 7-a 及び図 7-b は、男女・補完方法・配偶関係別人口である。配偶関係においても、「CANCEIS」は「原数値」と比較して、「あん分等」に近接した数値が算出されていることがわかる。男性では、「あん分等」と同様に他の項目と比較して、未婚の人口が多く増加する結果となった。女性では、未婚以外に死別も「あん分等」と同様、増加する結果となった。

図 7-a　補完方法・配偶関係別人口（男性）

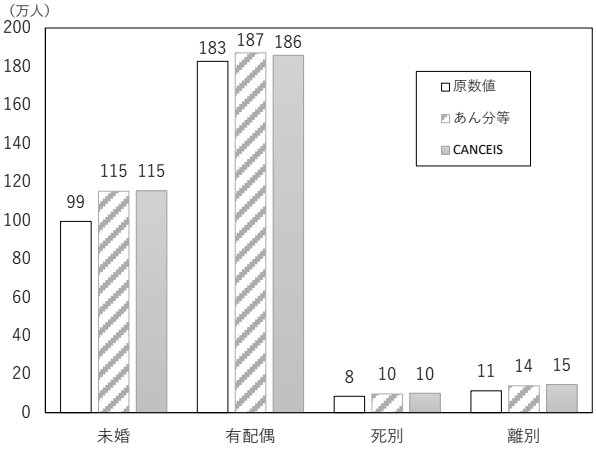
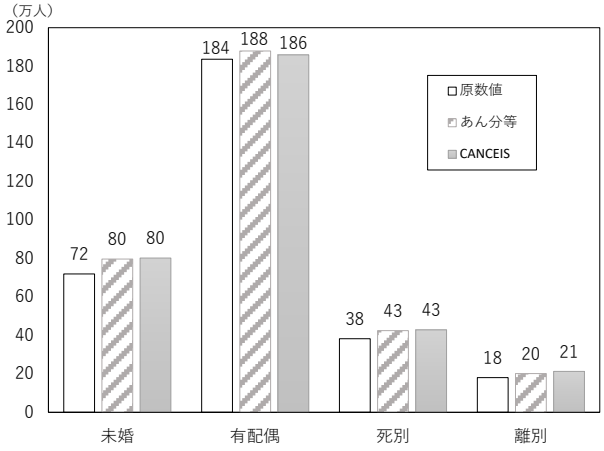


図 7-b　補完方法・配偶関係別人口（女性）



4.1.4　続柄

表２は、一般世帯における補完方法・世帯主との続柄別人口及び構成比である。人数については、続柄の不詳が補完されている分だけ「原数値」よりも「CANCEIS」の方が多くなっているが、構成比でみれば、ほぼ等しい値になっていることがわかる。2020 年国勢調査で、参考値として公表している、「あん分等」による不詳補完値では、続柄については作成していない。続柄の場合、区分数が多いことなどからあん分による作成が困難であったことが要因として考えられる。一方の「CANCEIS」では、個票データの不詳を補完しているため、このような事例においても、補完結果を算出することが可能であり、それが優れた点でもある。

表２　一般世帯における補完方法・世帯主との続柄別人口及び構成比

世帯主との続柄	人数			構成比(%)		
	原数値	CANCEIS	一致率[注](%)	原数値	CANCEIS	一致率(%)
世帯主	3,233,126	3,233,126	100.0	43.6	43.6	100.0
配偶者	1,682,459	1,690,676	99.5	22.7	22.8	99.6
子	2,029,607	2,043,616	99.3	27.4	27.6	99.3
子の配偶者	52,842	53,458	98.8	0.7	0.7	100.0
世帯主の父母	157,284	158,687	99.1	2.1	2.1	100.0
世帯主の配偶者の父母	35,457	35,596	99.6	0.5	0.5	100.0
孫	115,945	116,997	99.1	1.6	1.6	100.0
祖父母	2,198	2,215	99.2	0.0	0.0	100.0
兄弟姉妹	35,540	35,996	98.7	0.5	0.5	100.0
他の親族	10,979	11,151	98.5	0.1	0.2	50.0
住み込みの雇人	908	942	96.4	0.0	0.0	100.0
その他	29,214	29,631	98.6	0.4	0.4	100.0
世帯主との続き柄「不詳」	26,532	0	-	-	-	-

注）一致率は、次の式で算出　【公表値／CANCEIS×100】

4.1.5 家族類型

表３は、「原数値」と「CANCEIS」について、一般世帯の家族類型世帯数及び構成比を比較したものである。家族類型も続柄と同様、「あん分等」による不詳補完値は作成していない。世帯数については、不詳を補完した分だけ、「原数値」よりも「CANCEIS」の方が増加していることがわかる。単身世帯では、「原数値」と「CANCEIS」で世帯数の違いは見られないが、これは単身世帯の場合、続柄が不詳だったとしても世帯主に一意的に定まるためである。続柄と同様、構成比についても、「原数値」と「CANCEIS」でほぼ等しい数値になっている。

表３　一般世帯の補完方法・家族類型別世帯数及び構成比

家族類型	世帯数			構成比(%)		
	原数値	CANCEIS	一致率(%)	原数値	CANCEIS	一致率(%)
核家族世帯	1,794,260	1,805,119	99.4	55.7	55.8	99.8
夫婦のみの世帯	623,565	626,991	99.5	19.4	19.4	100.0
夫婦と子供から成る世帯	900,894	905,388	99.5	28.0	28.0	100.0
男親と子供から成る世帯	41,751	42,885	97.4	1.3	1.3	100.0
女親と子供から成る世帯	228,050	229,855	99.2	7.1	7.1	100.0
夫婦と両親から成る世帯	9,022	9,054	99.6	0.3	0.3	100.0
夫婦とひとり親から成る世帯	33,059	33,227	99.5	1.0	1.0	100.0
夫婦，子供と両親から成る世帯	34,024	34,099	99.8	1.1	1.1	100.0
夫婦，子供とひとり親から成る世帯	61,819	62,305	99.2	1.9	1.9	100.0
夫婦と他の親族（親，子供を含まない）から成る世帯	5,746	5,774	99.5	0.2	0.2	100.0
夫婦，子供と他の親族（親を含まない）から成る世帯	20,676	20,805	99.4	0.6	0.6	100.0
夫婦，親と他の親族（子供を含まない）から成る世帯	3,433	3,456	99.3	0.1	0.1	100.0
夫婦，子供，親と他の親族から成る世帯	9,833	9,894	99.4	0.3	0.3	100.0
兄弟姉妹のみから成る世帯	15,698	15,866	98.9	0.5	0.5	100.0
他に分類されない世帯	28,421	28,986	98.1	0.9	0.9	100.0
非親族を含む世帯	28,887	29,320	98.5	0.9	0.9	100.0
単独世帯	1,175,221	1,175,221	100.0	36.5	36.3	100.6
世帯の家族類型「不詳」	13,027	0	-	-	-	-

注１）一致率は、次の式で算出　【公表値／CANCEIS×100】
注２）原数値の構成比は、「不詳」を除いて算出
注３）CANCEISの家族類型は、CANCEISにより補完した続柄等を用いて算出

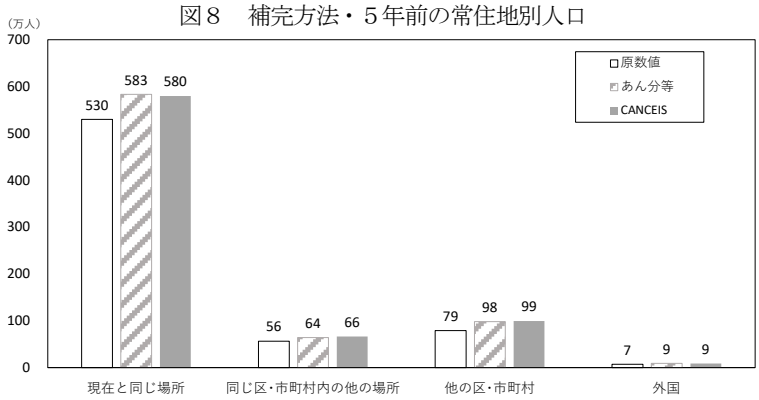
4.2 人口移動集計

人口移動集計は、2015 年国勢調査時の常住地（５年前の常住地）と 2020 年国勢調査時の常住地を比較することにより、人口の転入・転出状況について、男女・年齢別等に集計した結果である。

4.2.1 ５年前の常住地人口

図８は、愛知県の５年前の常住地別人口について、補完方法別に「原数値」「あん分等」及び「CANCEIS」を比較したグラフになっており、「あん分等」と「CANCEIS」が近接した数値となっている。

図８　補完方法・５年前の常住地別人口



4.2.2　転入元市区町村[24]

表４は、愛知県名古屋市千種区における補完方法別転入元市区町村ランキングであり、順位が多少前後する市区町村はあるものの、「原数値」「あん分等」及び「CANCEIS」において、傾向はほぼ同じになった。人数については、「あん分等」と「CANCEIS」が近接した数値となった。

表４　補完方法別転入元市区町村ランキング（愛知県名古屋市千種区）

順位	原数値			あん分等			CANCEIS		
	転入元市区町村	人数	転入元総数に対する割合（%）	転入元市区町村	人数	転入元総数に対する割合（%）	転入元市区町村	人数	転入元総数に対する割合（%）
1	東京都世田谷区	299	1.8	東京都世田谷区	396	1.8	東京都世田谷区	410	1.8
2	岐阜県岐阜市	279	1.7	岐阜県岐阜市	379	1.7	岐阜県岐阜市	384	1.7
3	石川県金沢市	241	1.5	石川県金沢市	327	1.5	石川県金沢市	333	1.5
4	三重県津市	223	1.4	三重県津市	307	1.4	三重県津市	307	1.4
5	東京都練馬区	200	1.2	東京都練馬区	278	1.3	三重県四日市市	293	1.3
6	三重県四日市市	200	1.2	三重県四日市市	273	1.3	東京都練馬区	276	1.2
7	東京都大田区	177	1.1	東京都大田区	243	1.1	東京都大田区	246	1.1
8	兵庫県西宮市	177	1.1	兵庫県西宮市	235	1.1	兵庫県西宮市	237	1.1
9	静岡県浜松市中区	152	0.9	静岡県浜松市中区	217	1.0	東京都江東区	228	1.0
10	大阪府吹田市	144	0.9	大阪府吹田市	201	0.9	静岡県浜松市中区	218	1.0
11	東京都江東区	142	0.9	静岡県静岡市駿河区	198	0.9	大阪府吹田市	200	0.9
12	東京都杉並区	140	0.9	東京都江東区	192	0.9	東京都杉並区	184	0.8
13	静岡県静岡市駿河区	136	0.8	東京都杉並区	188	0.9	東京都板橋区	183	0.8
14	東京都新宿区	135	0.8	東京都江戸川区	183	0.8	東京都江戸川区	183	0.8
15	東京都江戸川区	132	0.8	東京都新宿区	182	0.8	静岡県静岡市駿河区	174	0.8
16	千葉県船橋市	128	0.8	千葉県船橋市	171	0.8	千葉県市川市	170	0.8
17	大阪府豊中市	124	0.8	東京都板橋区	165	0.8	千葉県船橋市	169	0.8
18	東京都板橋区	119	0.7	大阪府豊中市	165	0.8	東京都新宿区	167	0.7
19	千葉県市川市	115	0.7	千葉県市川市	156	0.7	神奈川県横浜市港北区	160	0.7
20	東京都品川区	108	0.7	東京都品川区	153	0.7	大阪府豊中市	152	0.7

注）他県移動のみ集計

4.2.3　転出先市区町村[25]

表５は、愛知県名古屋市千種区における補完方法別転出先市区町村ランキングであり、転入元市区町村ランキングと同様、各補完方法でほぼ同じ傾向になることがわかった。また、人数についても、「あん分等」と「CANCEIS」で近接した数値となっている。

表５　補完方法別転出先市区町村ランキング（愛知県名古屋市千種区）

順位	原数値			あん分等			CANCEIS		
	転出先市区町村	人数	転出先総数に対する割合（%）	転出先市区町村	人数	転出先総数に対する割合（%）	転出先市区町村	人数	転出先総数に対する割合（%）
1	東京都世田谷区	391	2.6	東京都世田谷区	612	3.0	東京都世田谷区	623	3.0
2	東京都杉並区	272	1.8	東京都杉並区	435	2.1	東京都大田区	378	1.8
3	東京都大田区	253	1.7	東京都大田区	395	1.9	東京都杉並区	375	1.8
4	東京都江東区	229	1.5	東京都江東区	352	1.7	東京都練馬区	337	1.6
5	東京都練馬区	223	1.5	東京都練馬区	348	1.7	東京都江東区	335	1.6
6	大阪府吹田市	208	1.4	東京都品川区	331	1.6	東京都品川区	316	1.5
7	東京都品川区	207	1.4	大阪府吹田市	292	1.4	大阪府吹田市	270	1.3
8	岐阜県岐阜市	194	1.3	東京都文京区	268	1.3	東京都文京区	267	1.3
9	石川県金沢市	176	1.2	石川県金沢市	239	1.2	岐阜県岐阜市	234	1.1
10	東京都文京区	167	1.1	岐阜県岐阜市	239	1.2	石川県金沢市	224	1.1
11	大阪府豊中市	166	1.1	東京都江戸川区	226	1.1	東京都港区	215	1.1
12	三重県津市	156	1.1	大阪府豊中市	212	1.0	東京都江戸川区	215	1.1
13	兵庫県西宮市	148	1.0	東京都港区	209	1.0	東京都目黒区	212	1.0
14	東京都江戸川区	141	1.0	東京都新宿区	209	1.0	大阪府豊中市	207	1.0
15	神奈川県横浜市港北区	134	0.9	東京都中野区	204	1.0	東京都中野区	198	1.0
16	神奈川県川崎市中原区	127	0.9	東京都目黒区	202	1.0	三重県津市	197	1.0
17	神奈川県横浜市青葉区	124	0.8	兵庫県西宮市	200	1.0	神奈川県横浜市港北区	188	0.9
18	千葉県市川市	123	0.8	三重県津市	198	1.0	東京都豊島区	184	0.9
19	東京都目黒区	122	0.8	神奈川県横浜市港北区	192	0.9	東京都新宿区	182	0.9
20	静岡県浜松市中区	119	0.8	東京都板橋区	187	0.9	兵庫県西宮市	181	0.9

注）他県移動のみ集計

4.3　就業状態等基本集計

就業状態等基本集計は、人口の労働力状態、従業上の地位、産業・職業大分類別等に関する結果について集計している。労働力状態は、調査年の９月 24 日から 30 日までの１週間に「仕事を

[24] 令和２年 10 月１日時点で、当該市区町村に常住する人の５年前の常住地を集計。
[25] 全都道府県別市区町村の令和２年 10 月１日時点における５年前常住地が当該市区町村である人を集計。

したかどうかの別」により、区分したものをいう。

4.3.1 労働力状態

図９は、補完方法・労働力状態別人口のグラフであるが、数値の大きい順に「CANCEIS」＞「あん分等」＞「原数値」となっている。この順序は、就業状態等基本集計の他の項目でも同様であった。また、「原数値」と比較して、「あん分等」と「CANCEIS」が近接した値となっている。「あん分等」より、「CANCEIS」の人口が多くなっている要因は、「CANCEIS」では、年齢補完も実行されており、それにより 15 歳以上人口が「あん分等」よりも増加、それを含めて労働力状態を補完したためである。

表 6-a 及び表 6-b は、男女・補完方法・労働力状態別構成比であるが、「原数値」「あん分等」及び「CANCEIS」ともに類似した構成比となっていることがわかる。

図９　補完方法・労働力状態別人口

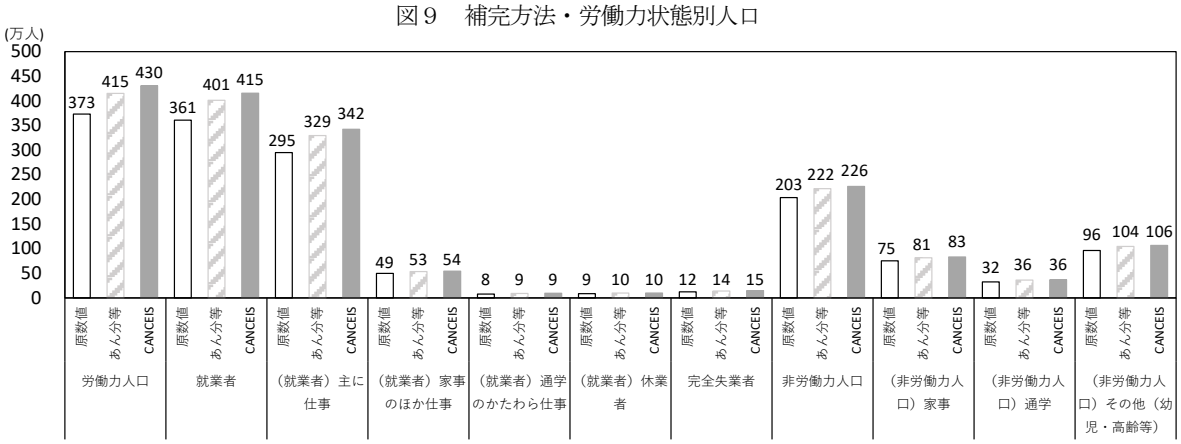


表 6-a　補完方法･労働力状態別構成比(男性)

労働力状態（男性）	原数値	あん分等	CANCEIS
(就業者)主に仕事	67.5	67.9	68.1
(就業者)家事のほか仕事	1.4	1.4	1.4
(就業者)通学のかたわら仕事	1.3	1.4	1.4
(就業者)休業者	1.4	1.4	1.4
完全失業者	2.6	2.7	2.8
家事	2.6	2.6	2.6
通学	5.9	5.9	5.8
その他　(幼児･高齢など)	17.2	16.8	16.6

表 6-b　補完方法・労働力状態別構成比（女性）

労働力状態（女性）	原数値	あん分等	CANCEIS
(就業者)主に仕事	35.4	35.8	36.5
(就業者)家事のほか仕事	15.5	15.3	15
(就業者)通学のかたわら仕事	1.3	1.4	1.4
(就業者)休業者	1.6	1.6	1.6
完全失業者	1.7	1.7	1.7
家事	23	22.8	22.6
通学	5.4	5.4	5.3
その他　(幼児･高齢など)	16.1	15.9	15.8

4.3.2 労働力率

図 10-a 及び図 10-b は、男女・補完方法・年齢階級別の労働力率のグラフである。「原数値」「あん分等」及び「CANCEIS」において、年齢階級別の労働力率は 15～19 歳で 0.8～0.9%ポイントほどの差が見られたもののほぼ等しい数値になることがわかった。

図 10-a　補完方法・年齢階級別労働力率（男性）

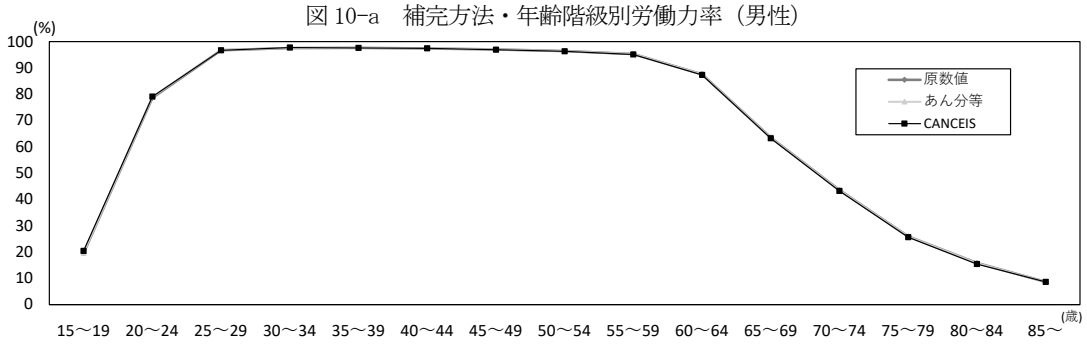
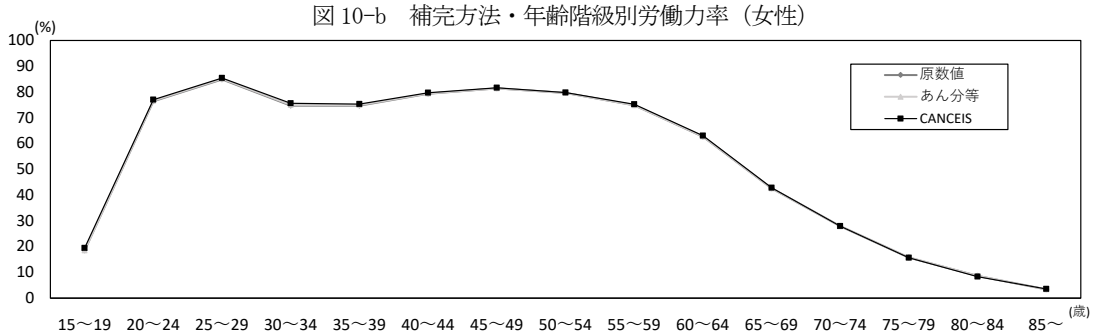


図 10-b　補完方法・年齢階級別労働力率（女性）



4.3.3　従業上の地位

図 11 は、補完方法・従業上の地位別人口のグラフであるが、「原数値」と比較して、「あん分等」と「CANCEIS」が近接した数値になっている。「CANCEIS」の方が「あん分等」より多くなっている要因は、労働力状態の補完と同様に、年齢補完により 15 歳以上人口が増加したことによるものである。

表 7-a 及び表 7-b は、男女・補完方法別に従業上の地位の構成比を示しているが、「あん分等」「CANCEIS」ともに原数値と類似した構成比になっていることがわかる。

図 11　補完方法・従業上の地位別人口

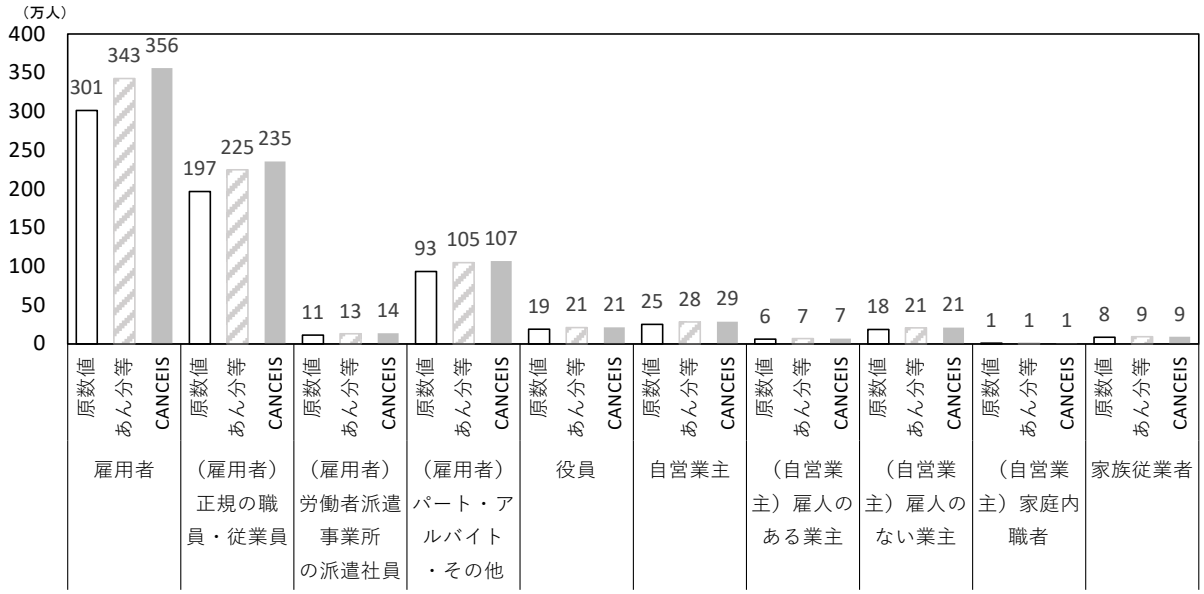


表 7-a　補完方法・従業上の地位別構成比（男性）

従業上の地位（男性）	原数値	あん分等	CANCEIS
正規の職員・従業員	68.7	68.8	69.3
労働者派遣事業所の派遣社員	2.6	2.7	2.7
パート・アルバイト・その他	11.7	11.8	11.7
役員	7.1	6.9	6.8
雇人のある業主	2.5	2.4	2.4
雇人のない業主	6.6	6.6	6.4
家族従業者	0.8	0.8	0.7
家庭内職者	0	0	0

表 7-b　補完方法・従業上の地位別構成比（女性）

従業上の地位（女性）	原数値	あん分等	CANCEIS
正規の職員・従業員	38.9	39.3	40.2
労働者派遣事業所の派遣社員	3.9	3.9	4
パート・アルバイト・その他	45.2	44.9	44.1
役員	3.1	3	3
雇人のある業主	0.7	0.7	0.7
雇人のない業主	3.4	3.4	3.4
家族従業者	4.4	4.3	4.2
家庭内職者	0.5	0.5	0.5

4.3.4 産業

図 12-a 及び図 12-b は、男女・補完方法・産業大分類別人口であるが、「原数値」と比較して、「あん分等」と「CANCEIS」が近接した数値となっている。「あん分等」よりも「CANCEIS」の人口が多いのは、他の就業状態等基本集計の調査項目同様に、年齢補完により 15 歳以上人口が増加した影響である。

表８は、男女・補完方法・産業大分類別人口の構成比を示しており、「原数値」「あん分等」「CANCEIS」ともに類似した構成比になっていることがわかる。

図 12-a 補完方法・産業大分類別人口（男性）

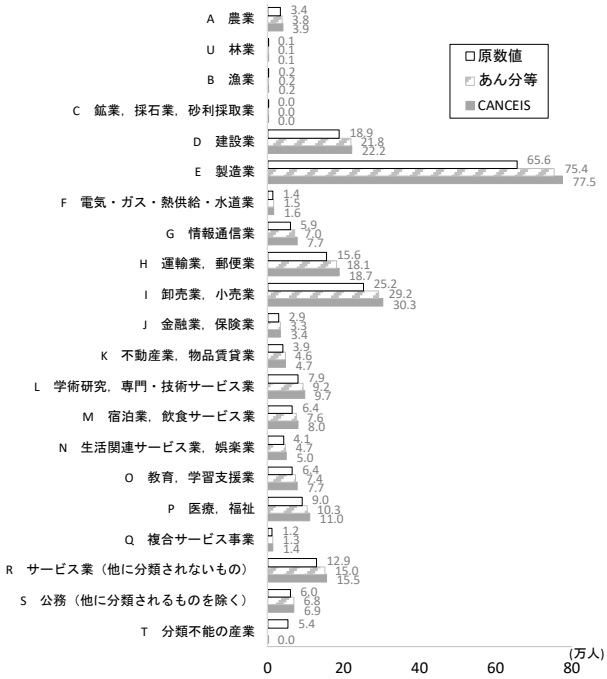


図 12-b 補完方法・産業大分類別人口（女性）

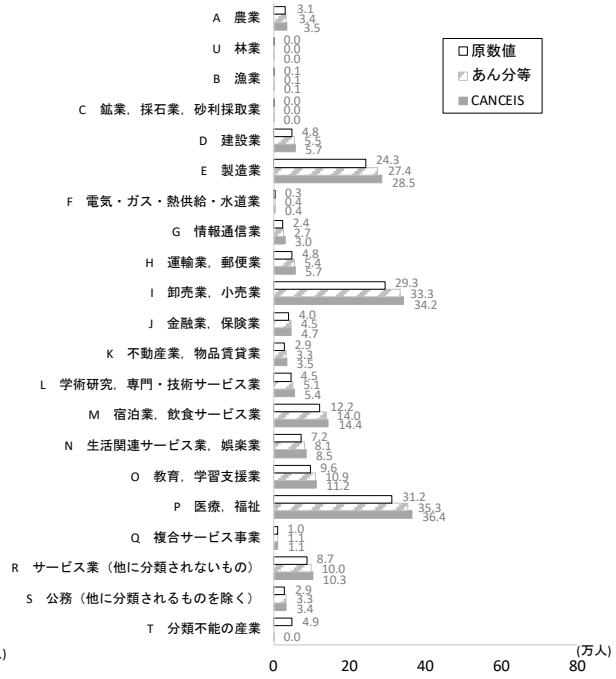


表８ 男女・補完方法・産業大分類別構成比

（%）

産業大分類	男性			女性		
	原数値	あん分等	CANCEIS	原数値	あん分等	CANCEIS
農業	1.7	1.7	1.6	2.0	1.9	2.0
林業	0.0	0.0	0.0	0.0	0.0	0.0
漁業	0.1	0.1	0.1	0.1	0.0	0.0
鉱業，採石業，砂利採取業	0.0	0.0	0.0	0.0	0.0	0.0
建設業	9.6	9.6	9.4	3.1	3.1	3.2
製造業	33.3	33.2	32.9	15.8	15.8	15.8
電気・ガス・熱供給・水道業	0.7	0.7	0.7	0.2	0.2	0.2
情報通信業	3.0	3.1	3.3	1.5	1.6	1.7
運輸業，郵便業	7.9	8.0	8.0	3.1	3.1	3.2
卸売業，小売業	12.8	12.8	12.9	19.1	19.1	19.0
金融業，保険業	1.5	1.4	1.5	2.6	2.6	2.6
不動産業，物品賃貸業	2.0	2.0	2.0	1.9	1.9	1.9
学術研究，専門・技術サービス業	4.0	4.1	4.1	2.9	3.0	3.0
宿泊業，飲食サービス業	3.3	3.3	3.4	8.0	8.0	8.0
生活関連サービス業，娯楽業	2.1	2.1	2.1	4.7	4.7	4.7
教育，学習支援業	3.3	3.3	3.3	6.3	6.3	6.2
医療，福祉	4.6	4.5	4.7	20.3	20.3	20.2
複合サービス事業	0.6	0.6	0.6	0.6	0.6	0.6
サービス業（他に分類されないもの）	6.5	6.6	6.6	5.7	5.7	5.7
公務（他に分類されるものを除く）	3.0	3.0	2.9	1.9	1.9	1.9

4.3.5　職業

図 13-a 及び図 13-b は、男女・補完方法・職業大分類別人口であるが、「原数値」と比較して、「あん分等」と「CANCEIS」が近接した数値となっている。「あん分等」よりも「CANCEIS」の人口が多いのは、他の就業状態等基本集計の調査項目同様に、年齢補完により 15 歳以上人口が増加した影響である。

表 9 は、男女・補完方法・職業大分類別人口の構成比を示しており、「原数値」「あん分等」「CANCEIS」ともに類似した構成比になっていることがわかる。

図 13-a　補完方法・職業大分類別人口（男性）

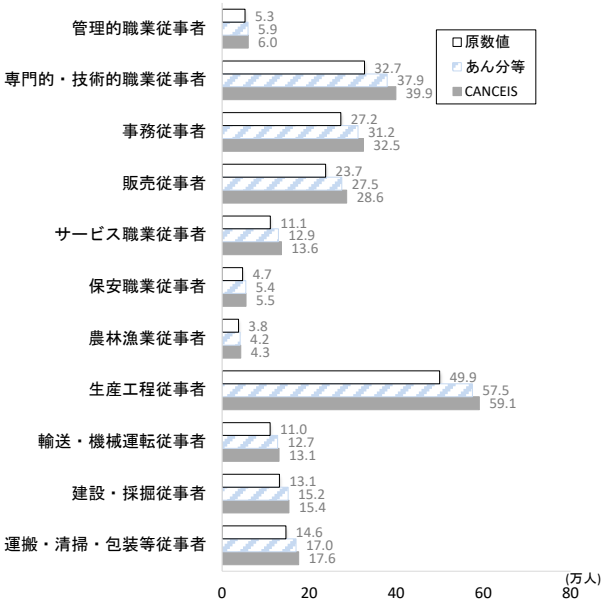


図 13-b　補完方法・職業大分類別人口（女性）

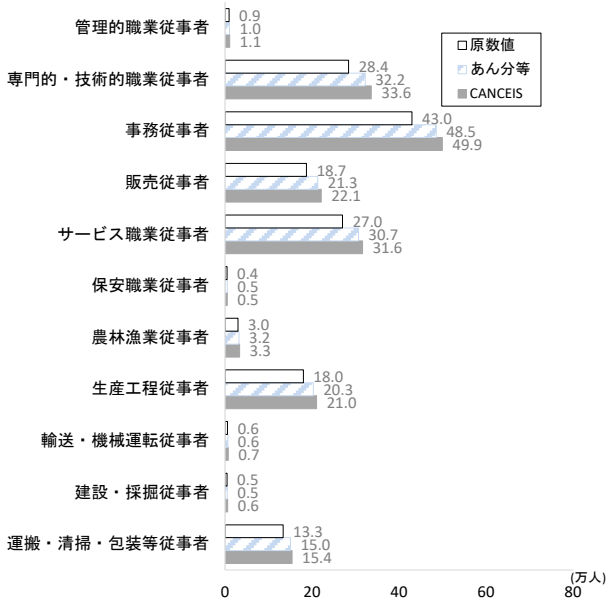


表 9　男女・補完方法・職業大分類別構成比

（%）

職業大分類	男性			女性		
	原数値	あん分等	CANCEIS	原数値	あん分等	CANCEIS
管理的職業従事者	2.7	2.6	2.5	0.6	0.6	0.6
専門的・技術的職業従事者	16.6	16.7	16.9	18.5	18.5	18.7
事務従事者	13.8	13.7	13.8	27.9	27.9	27.7
販売従事者	12	12.1	12.2	12.2	12.2	12.3
サービス職業従事者	5.6	5.7	5.8	17.5	17.6	17.6
保安職業従事者	2.4	2.4	2.3	0.3	0.3	0.3
農林漁業従事者	1.9	1.8	1.8	1.9	1.8	1.9
生産工程従事者	25.3	25.3	25.1	11.7	11.7	11.7
輸送・機械運転従事者	5.6	5.6	5.6	0.4	0.4	0.4
建設・採掘従事者	6.7	6.7	6.5	0.3	0.3	0.3
運搬・清掃・包装等従事者	7.4	7.5	7.5	8.7	8.7	8.6

4.4　従業地・通学地集計

従業地・通学地集計は、従業地・通学地による人口の構成や、現在住んでいる市区町村と従業地・通学地の市区町村との関係などを集計しており、昼間人口なども把握することができる。「従業地・通学地」とは、就業者が仕事をしている場所又は通学者が通学している学校の場所をいう。

4.4.1 従業地・通学地別人口

表 10 は、補完方法・従業地・通学地別人口及び構成比を示している。従業地・通学地別人口については、「原数値」と比較して、「あん分等」と「CANCEIS」が近接した数値となっている。構成比については、「原数値」「あん分等」「CANCEIS」ともに類似した数値となっていることがわかる。

表 10 補完方法・従業地通学地別人口及び構成比

（人）

	常住地による人口（夜間人口）	従業も通学もしていない	自市区町村で従業・通学	自宅で従業	自宅外の自市区町村で従業・通学	他市区町村で従業・通学	自市内他区で従業・通学	県内他市町村で従業・通学	他県で従業・通学	従業・通学市区町村「不詳・外国」	従業地・通学地「不詳」
原数値	7,542,415	2,245,089	2,367,625	301,952	2,065,673	1,960,164	463,232	1,376,889	88,668	31,375	969,537
不詳を除いた構成比（%）	-	34.3	36.2	4.6	31.6	30.0	7.1	21.0	1.4	-	-
あん分等	7,542,415	2,565,900	2,728,433	340,860	2,387,573	2,248,082	570,721	1,576,051	101,310	0	0
構成比（%）	100.0	34.0	36.2	4.5	31.7	29.8	7.6	20.9	1.3	-	-
CANCEIS	7,542,413	2,457,945	2,771,517	344,505	2,427,012	2,312,841	594,198	1,613,686	104,938	19	110
不詳を除いた構成比（%）	-	32.6	36.7	4.6	32.2	30.7	7.9	21.4	1.4	-	-

注１）「常住地による人口（夜間人口）」では、「CANCEIS」が「原数値」や「あん分等」と比較して、２人少ない値となっているが、これは施設等世帯において、エディットルールを満たさなかったことでエラーとなったユニットが２件存在したことによるものである。

注２）従業地・通学地「不詳」が「CANCEIS」では、110 人存在しているが、これはドナーが見つからなかったことによる補完不可能ユニットが存在したことを意味している。エディットルールは満たしているため、人口総数（表 10 における「常住地による人口（昼間人口）」）には影響しない。

4.4.2 従業地・通学地市区町村ランキング

表 11 は、愛知県名古屋市千種区における従業地・通学地市区町村ランキングである。「原数値」と「あん分等」は、上位 20 位まで全ての順位が同じになっている。また、「CANCEIS」についても「原数値」「あん分等」とほぼ同じ順位となっている。総数に対する割合についても、ほぼ同様の傾向となっている。

表 11 補完方法別従業地・通学地市区町村ランキング（愛知県名古屋市千種区）

（人）

順位	原数値			あん分等			CANCEIS		
	従業地・通学地市区町村	人数	総数に対する割合（%）	従業地・通学地市区町村	人数	総数に対する割合（%）	従業地・通学地市区町村	人数	総数に対する割合（%）
1	愛知県名古屋市千種区	38,038	43.0	愛知県名古屋市千種区	46,717	43.0	愛知県名古屋市千種区	47,900	42.7
2	愛知県名古屋市中区	12,417	14.0	愛知県名古屋市中区	15,544	14.3	愛知県名古屋市中区	16,256	14.5
3	愛知県名古屋市東区	5,430	6.1	愛知県名古屋市東区	6,718	6.2	愛知県名古屋市中村区	7,046	6.3
4	愛知県名古屋市中村区	5,288	6.0	愛知県名古屋市中村区	6,627	6.1	愛知県名古屋市東区	6,980	6.2
5	愛知県名古屋市昭和区	3,173	3.6	愛知県名古屋市昭和区	3,922	3.6	愛知県名古屋市昭和区	4,072	3.6
6	愛知県名古屋市名東区	2,528	2.9	愛知県名古屋市名東区	3,064	2.8	愛知県名古屋市名東区	3,245	2.9
7	愛知県名古屋市守山区	1,625	1.8	愛知県名古屋市守山区	1,947	1.8	愛知県名古屋市守山区	1,969	1.8
8	愛知県名古屋市西区	1,500	1.7	愛知県名古屋市西区	1,858	1.7	愛知県名古屋市西区	1,941	1.7
9	愛知県名古屋市瑞穂区	1,318	1.5	愛知県名古屋市瑞穂区	1,600	1.5	愛知県名古屋市瑞穂区	1,670	1.5
10	愛知県名古屋市北区	1,280	1.4	愛知県名古屋市北区	1,545	1.4	愛知県名古屋市北区	1,625	1.4
11	愛知県名古屋市港区	998	1.1	愛知県名古屋市港区	1,219	1.1	愛知県名古屋市港区	1,214	1.1
12	愛知県春日井市	980	1.1	愛知県春日井市	1,177	1.1	愛知県春日井市	1,213	1.1
13	愛知県名古屋市天白区	968	1.1	愛知県名古屋市天白区	1,171	1.1	愛知県名古屋市天白区	1,200	1.1
14	愛知県豊田市	955	1.1	愛知県名古屋市熱田区	1,150	1.1	愛知県名古屋市熱田区	1,160	1.0
15	愛知県名古屋市熱田区	947	1.1	愛知県豊田市	1,143	1.1	愛知県豊田市	1,147	1.0
16	愛知県長久手市	846	1.0	愛知県長久手市	1,030	0.9	愛知県長久手市	1,032	0.9
17	愛知県名古屋市中川区	763	0.9	愛知県名古屋市中川区	919	0.8	愛知県名古屋市中川区	966	0.9
18	愛知県小牧市	611	0.7	愛知県小牧市	728	0.7	愛知県小牧市	746	0.7
19	愛知県名古屋市南区	583	0.7	愛知県名古屋市南区	697	0.6	愛知県名古屋市南区	730	0.7
20	愛知県日進市	568	0.6	愛知県日進市	689	0.6	愛知県日進市	700	0.6

注）原数値の「総数に対する割合（%）」については、総数である分母から不詳を除いて算出している。

4.4.3　市区町村結果の検証（埼玉県→大島町）

埼玉県の従業地・通学地集計を見てみると、埼玉県から東京都の離島である大島町への従業・通学の数値を確認することができる。こういったレアケースと思われるデータについて、「CANCEIS」による補完値を確認してみた。

表 12-a は、埼玉県から東京都大島町への従業・通学の補完方法別人数を示しているが、「CANCEIS」による補完で極端な増加は見られなかった。また、表 12-b は、当該従業者・通学者がどの産業に属しているかを示しているが、「CANCEIS」による補完で極端な増加は見られなかった。一方、表 12-c については、どの職業に属しているかを示しているが、これについても、極端な増加は見られなかった。そして、予想通り、人数が元々多い分類に、補完がされていることがわかった。この結果は、元々の人数が多いほど、ドナーとして見つかる可能性が高いことからも想像に難くない。産業において、「建設業」「製造業」「運輸業」及び「郵便業」が増加していることも、島しょ部への移動を伴う産業として推定できるのではないだろうか。「あん分等」では、ここまでの粒度の結果を算出することが困難であるため、表 12-b と表 12-c については、「あん分等」の数値が存在しない。

表 12-d は、東京都大島町における昼間人口を示している。この数値は、全国の国勢調査データの従業地・通学地を算出することで初めて集計できる数値であるため、全国の「CANCEIS」による補完結果の妥当性も確認することができる。その結果を見ると、「あん分等」と「CANCEIS」が極めて近い数値になっていることがわかる。

表 12-a　埼玉県（常住地）→東京都大島町

（人）

地域名（常住地）	原数値	あん分等	CANCEIS
さいたま市大宮区	1	1	2
さいたま市見沼区	1	1	1
さいたま市浦和区	1	1	1
さいたま市緑区	1	1	1
さいたま市岩槻区	1	1	1
川越市	1	1	1
川口市	4	5	6
行田市	1	1	1
所沢市	2	2	2
狭山市	1	1	1
鴻巣市	2	2	2
上尾市	1	1	1
草加市	2	2	3
越谷市	1	1	2
志木市	1	1	1
新座市	1	1	1
八潮市	2	2	4
富士見市	1	1	1
三郷市	2	2	3
松伏町	1	1	1

表 12-b　埼玉県（常住地）→東京都大島町（産業大分類）

（人）

産業大分類	原数値	CANCEIS
建設業	5	7
製造業	4	6
情報通信業	1	1
運輸業，郵便業	3	5
卸売業，小売業	5	5
金融業，保険業	1	1
不動産業，物品賃貸業	1	1
学術研究，専門・技術サービス業	1	1
宿泊業，飲食サービス業	2	3
医療，福祉	2	2
サービス業（他に分類されないもの）	2	2
通学	1	2

表 12-c　埼玉県（常住地）→東京都大島町（職業大分類）

（人）

職業大分類	原数値	CANCEIS
管理的職業従事者	3	5
専門的・技術的職業従事者	4	4
事務従事者	9	12
販売従事者	6	6
サービス職業従事者	1	2
保安職業従事者	1	1
生産工程従事者	1	2
建設・採掘従事者	1	1
運搬・清掃・包装等従事者	1	1
通学	1	2

表 12-d　昼間人口（東京都大島町）

（人）

東京都大島町	原数値	あん分等	CANCEIS
昼間人口	7,276	7,332	7,330

５　おわりに

本稿では、CANCEIS による不詳補完値（以下、CANCEIS 補完値と呼ぶ。）を試算し、あん分等による不詳補完値（以下、あん分等補完値と呼ぶ。）と比較検証をした結果、年齢補完の影響が見られる集計区分を除いて、どの集計区分においても、両者が近似した数値になることを確認した。また、両者は構成比も近似しており、構成比においては、原数値とも近似していることがわかった。本稿に記載した CANCEIS 補完値は主に愛知県の結果であったが、本検証以降に、全都道府県で CANCEIS 補完値を試算しており、愛知県における CANCEIS 補完値の特性は、他の都道府県でも同様に見られることがわかった。CANCEIS による補完では、ドナーが見つからなかったことによる補完不可能ユニットがいくつかあったが、調査票表面の補完では、それは全都道府県で 100 ユニット未満であった。調査票裏面の補完では、若干増加した印象ではあったものの、大きな変化は見られなかった。これらの CANCEIS による補完が不可能だったユニットについては、CANCEIS 実行後に手動で補完をするなどの別の方法による補完が必要になってくるかもしれないが、結果をマクロでみるのであれば、補完不可能ユニットの存在はないに等しい。

あん分等補完値における課題であった①粒度に限界があること、②集計区分間の整合性が取れないこと、③個票レベルの補完ではないため二次的利用に使えないことについては、CANCEIS による個票レベルの補完により、一応の解決が図られている。①については、例えば、あん分等補完値を作成する際に用いる構成比の算出において、市区町村レベルなどの細かい区分の結果表を利用したい場合、算出しようにも集計項目に数値が存在せず、構成比を算出することが困難な事例が存在する。こういった事例では、あん分により、数値のない項目の存在を「有」にするのは不可能である（実際に「無」の場合もある。）。一方の CANCEIS 補完値は、個票データを補完しており、その補完された個票データを用いて集計するため、こういった問題は起こらない。②と③についても、①と同様、個票レベルの補完により解決できた。ただし、③については、当該数値が CANCEIS により補完された数値か否かを明示するかどうかも含めて、二次的利用における集計用個別データの提供方法には議論の余地がある。

CANCEIS 補完値の算出においては、インプットデータの段階で、CANCEIS がユニット間の類似性を識別できる世帯構造を、ソートやグルーピングを適用することで、いかにして作り上げることができるかが結果を左右する。例えば年齢等を補完する調査票表面の場合、続柄のソートは特に重要で、もし、続柄がユニット内で全て不詳であった場合、そのユニットが補完不可能となる可能性が高くなる。CANCEIS はユニット内の続柄の構成により、類似ユニットか否かを判断しているといっても過言ではない。CANCEIS による補完における続柄項目の重要性については、イギリス国家統計局の人口センサスにおける CANCEIS 利用の検討においても言及されており、国を問わず CANCEIS で重要な特徴量であることがわかる。また、データのグルーピングも重要であり、現在、主に世帯人員でグルーピングをしているところであるが、これは続柄が重要なことからも自明であり、基本的に続柄の構成は、世帯人員に依存している面が強いためでもある。さらに、CANCEIS のドナー検索処理では、ユニットの並び順で、近傍のユニットからドナーを選択してくるという性質がある。ドナー検索は無制限に実行されるのではなく、ある程度距離が収束して、これ以上の収束が見込めなくなったと CANCEIS が判断したときに検索処理を停止する。そのため、もし、最良のドナーが FU から遠くにあった場合、途中で検索を切り上げてしまう可能性もある。これを避けるためにも、なるべくグルーピング[26]やソートをすることによって、類似したユニットをインプットデータの段階で、近傍に集めているのである。

今後の CANCEIS による補完における課題は、個票レベルの整合性検証である。DLT により補完に利用するドナーを制御してはいるものの、表面と裏面を別々に CANCEIS により補完してい

[26] 大容量データを補完するため、演算時間が膨大になることを防ぐという目的もある。

ることから、意図せず不整合や外れ値が生じてしまっている可能性がある。また、特に小地域における CANCEIS 補完値については、妥当性について引き続き検証をしていく必要があると考える。

現在、主要国の人口センサスにおける補完方法の潮流は、「あん分」による補完よりも、最近隣法によるソフトウェアシステムを用いた補完である。確かに、実際に CANCEIS を実行してみると、「あん分」よりも「CANCEIS」の方が、妥当性のある数値が算出できていると考えられる。あん分等補完値は、マクロで見れば妥当性があると確認できるデータであっても、ミクロで分析をすると、その補完方法の性質上、綻びが生じてしまう。一方の CANCEIS 補完値は、個票データを補完していることから、そういった綻びが生じることはなく、ユーザーの使い勝手も向上する。

一方で、最近隣法も万能ではなく、ドナーが存在しているからこそできる補完方法であり、ドナーが見つからなければ不詳のままである。それ故に、サンプル調査でこの補完方法を実行するのは困難であり、ドナーが潤沢に存在する大規模（全数）調査でこそ能力を発揮するのである。データに不詳が多い場合も同様であり、いくら補完方法が進歩したとしても、不詳が多くてはドナーを見つけることができず、妥当性のある数値を推定するのは困難だろう。そのため、補完システムがあるから不詳はあっても問題ないという理屈にはならず、調査票上の不詳を減少させることもまた重要なのである。

謝辞等

本稿は、日本人口学会第 75 回及び第 76 回における報告内容に加筆・修正を行ったものである。本研究の遂行に当たり、総務省統計局統計調査部国勢統計課の中村英昭氏及び宮下佳孝氏、総務省統計局統計情報システム管理官開発企画第一担当の塚本大器氏の皆様方から大変有益なコメントをいただいた。また、外部有識者の菅幹雄教授（法政大学）、加藤久和教授（明治大学）及び西郷浩教授（早稲田大学）の皆様方からも大変貴重なご意見をいただいた。そして、独立行政法人統計センター統計編成部人口統計編成課の野崎政志氏、相馬輝氏及び會澤豪大氏の皆様方からは、CANCEIS を実行する上での困難な演算について、技術的にもご協力いただいた。

さらに、本稿の執筆に当たっては、２名の匿名査読者から大変有益なコメントをいただき、内容を改善することができた。

この場を借りて、ご協力いただいた皆様方へ感謝の意を表したい。

【参考文献】

［1］北原昌嗣・寺垣内雅子（2023）「諸外国の国勢調査におけるインピュテーション方法」，『統計研究彙報』第 80 号，2023 年３月，pp137-162

［2］総務省統計局（2021）「不詳補完結果（参考表）に関する解説及び参考資料（令和２年及び平成 27 年国勢調査に関する不詳補完結果（参考表）について）」，
https://www.stat.go.jp/data/kokusei/2020/kekka.html，2021 年 11 月

［3］内閣府経済社会総合研究所(ESRI)景気統計部（2017）「欠測値補完に関する調査研究報告書」，2017 年３月

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

［5］Andrew Stelmack (2019) “Improving Edit and Imputation Strategies through Feature Selection,” 2019 Joint Statistical Meetings, 27 July - 1 August 2019
https://ww2.amstat.org/meetings/proceedings/2019/data/assets/pdf/1199678.pdf

［6］Office for National Statistics (ONS) (2022) “Item editing and imputation process for Census 2021, England and Wales”
https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/itemeditingandimputationprocessforcensus2021englandandwales

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

［8］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/