資料２

令和７年国勢調査有識者会議企画WG（第１回）

CANCEISによる補完方法の概要と試算

令和５年７月

総務省統計局

概要

◯本資料では、CANCEISの概要・実行原理と、CANCEIS補完の試算結果を紹介

※国勢調査（人口等基本集計）の愛知県分の結果について、CANCEISによる補完を実施し、（原数値又は不詳補完値）とCANCEIS適用後の人口等を比較

CANCEISの概要

正式名称	**CAN**adian **C**ensus **E**dit and **I**mputation **S**ystem (CANCEIS)
作成者	Mike Bankier（カナダ統計局）
利用開始	1996年（1992年に開発）
動作原理	最近隣法 (Nearest-neighbor Imputation Methodology (NIM))
主な特徴	▸ **数値、カテゴリー及び英数字データのドナー補完を同時に実行可能** ▸ **大量のデータを効率的に処理可能** ▸ **調査データの問題や希望する解決策を容易で正確に定義可能※**

※　決定論的又はドナー利用により解決

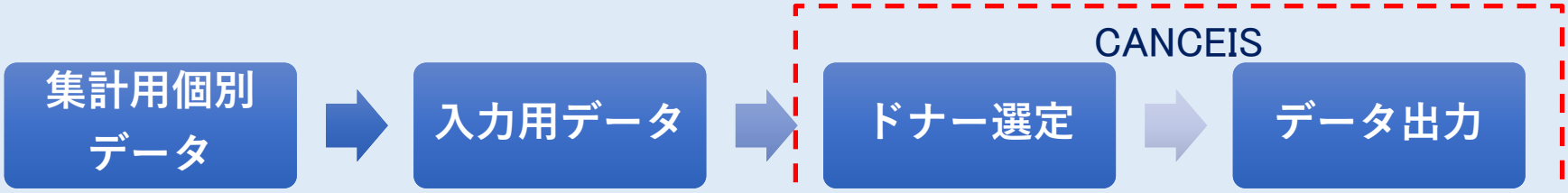
【その他、以下の特徴】

- 既存の変数に基づいて新しい変数を導出（作成）可能
- ソフトウェアは、Windows 環境で使用するための移植が可能
- システムは、新規ユーザーが直感的に使用可能
- Excel又はテキスト形式でモジュールを使う（作る）ことができる柔軟性
- 上級ユーザーは一連のパラメーターを介して補完方法を緻密に制御可能
- システムは動的であり、ユーザーの要望を満たすため新機能を迅速に追加可能

CANCEISの実行原理

◯ CANCEISでは、元データ（集計用個別データ）に適切な前処理をした入力用データから、
- 入力データの各ユニットを**合格**と**不合格**のいずれかに分類し、
- 不合格ユニットの問題を修正するデータを提供できるドナーを選定（通常は合格ユニットを選定）

※　**合格ユニット（passed units）・・・すべての変数に有効な値があり、変数間の不整合もないユニット**
不合格ユニット（failed units）・・・インピュテーションを必要とする一つ以上の問題を持つユニット



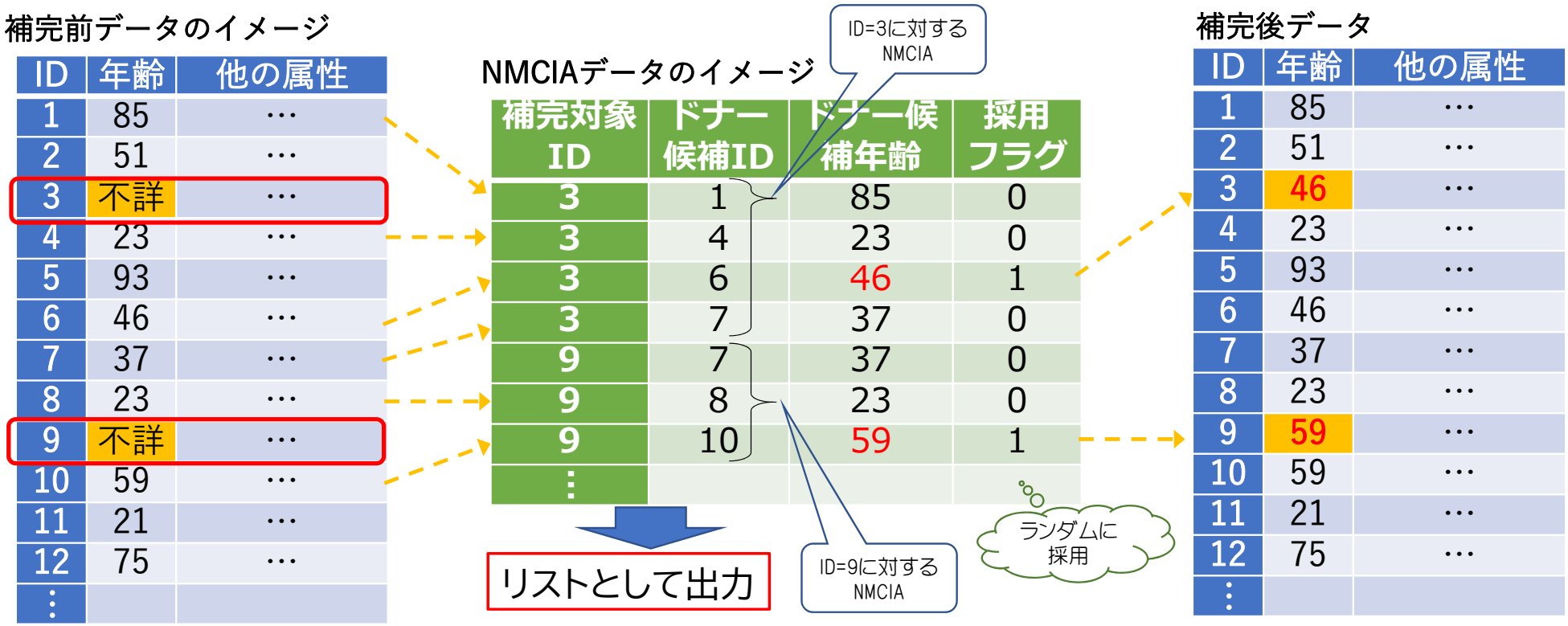
合格、不合格ユニットのイメージ

ID	年齢	配偶関係	合格/不合格
1	38	配偶者あり	**合格**
2	36	配偶者あり	**合格**
3	13	ブランク	不合格
4	0	配偶者あり	不合格
5	51	配偶者あり	**合格**
6	53	配偶者あり	**合格**
7	62	未婚	**合格**
8	27	ブランク	不合格

合格したID１，２，５，６，７は、
不合格のID３，８のブランクを代替し、
ID４の矛盾を解消するためのドナー候補となる。

CANCEISのドナー選択方法

○CANCEISは、不合格ユニットのドナー候補として、不合格ユニットと類似のユニットを複数探す。
○その際、不合格ユニットとドナー候補との相違を測定する方法として**距離関数**を使用し、**最短距離に近い(※1)ドナー候補（「NMCIA」と呼ばれる）(※2)を複数選択し、その中からドナーをランダムに選択採用する。**

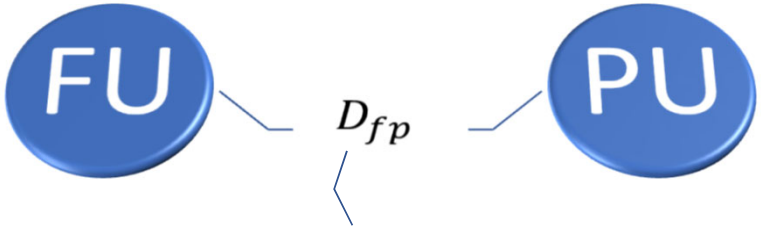


(※1)デフォルトでは最短距離×1.1以下（「1.1」はシステムパラメータで変更可能）
(※2)デフォルトでは最小1個、最大10個（最小値、最大値ともシステムパラメータで変更可能）

CANCEISドナー選定の原理（距離関数）

距離関数

$$D_{fp} = \sum_i w_i D_i(V_{fi}, V_{pi})$$



D_{fp}:不合格ユニット(FU)と合格ユニット(PU)の距離
V_{fi}:不合格ユニットにおけるi 番目の変数
V_{pi} : 合格ユニットにおける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ウエイト

D_{fp}が小さいほど、FUとPUが類似している。

具体例

$$D_i = \begin{cases} 0, & if \quad V_{fi} = V_{pi} \\ 1, & otherwise \end{cases}$$

不合格ユニットの文字列	ドナーユニットの文字列	Di の戻り値
Hello	Hello	0
HELLO	hello	1
Ah a	A ha	1

CANCEISドナー選定の原理（インピュテーションの品質）

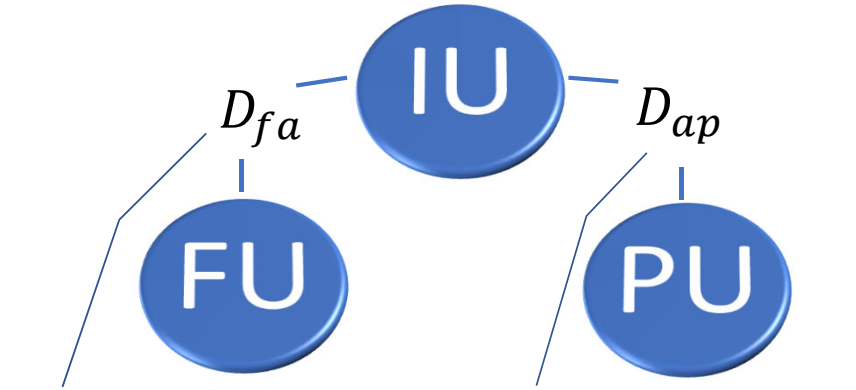
○CANCEISのドナーインピュテーションは以下の目的で実行され、その品質は下式で評価される。
- 不合格ユニットと極めて類似したユニットが選定されたか
- 複数人のドナーからではなく、可能な限り1人のドナーからのデータが得られたか

※ ドナー候補ユニットとの類似度が高く、不合格ユニットからの変化が最小のユニットをドナーとして選定
※ 実際のドナー選定は、不合格ユニットに対して検査できる潜在的なドナーの最大数を決定する等により効率的に実施

インピュテーションの品質

$$D_{fpa} = \alpha D_{fa} + (1-\alpha) D_{ap}$$

D_{fa}：不合格ユニットとインピュテーション後のユニット(IU)の距離
D_{ap}：インピュテーション後のユニットと合格ユニットの距離
D_{fpa}：D_{fa}とD_{ap}の加重平均
α：ユーザー定義システムパラメータ（$0.5 < \alpha \leqq 1.0$）



D_{fa}はFUとIUの類似度
→IA※の*最小変化量*を示す。

D_{ap}はIUとPUの類似度
→IAの*尤もらしさ*を示す。

※IA：インピュテーションアクション

D_{fp} 及びD_{fpa}が最小のユニットをドナーとして選定

CANCEIS補完の試算　ジョブフロー

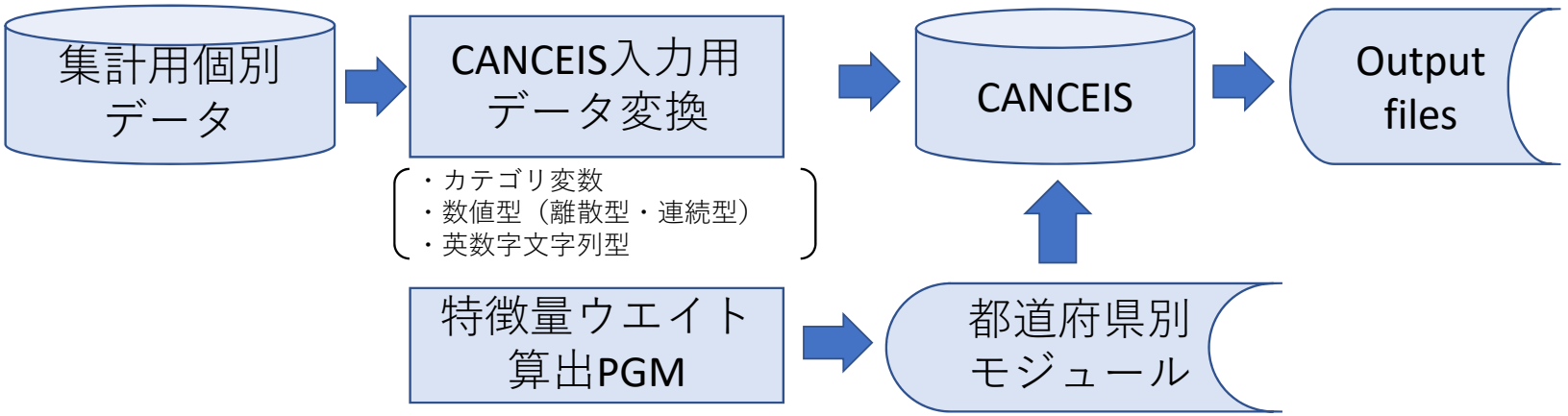
CANCEIS補完結果算出手順

①国勢調査の集計用個別データからCANCEIS入力用データに変換
②都道府県別にCANCEISの動力源となるモジュールを作成
　・都道府県別に、ドナー選択時の距離関数に必要な特徴量ウエイトを
　Random　Forest※により算出し、モジュールへ入力

※　多数の決定木モデル予測を組み合わせる機械学習手法

③CANCEIS入力用データと都道府県別モジュールをセットし、CANCEISを実行
④CANCEIS補完結果が出力される。

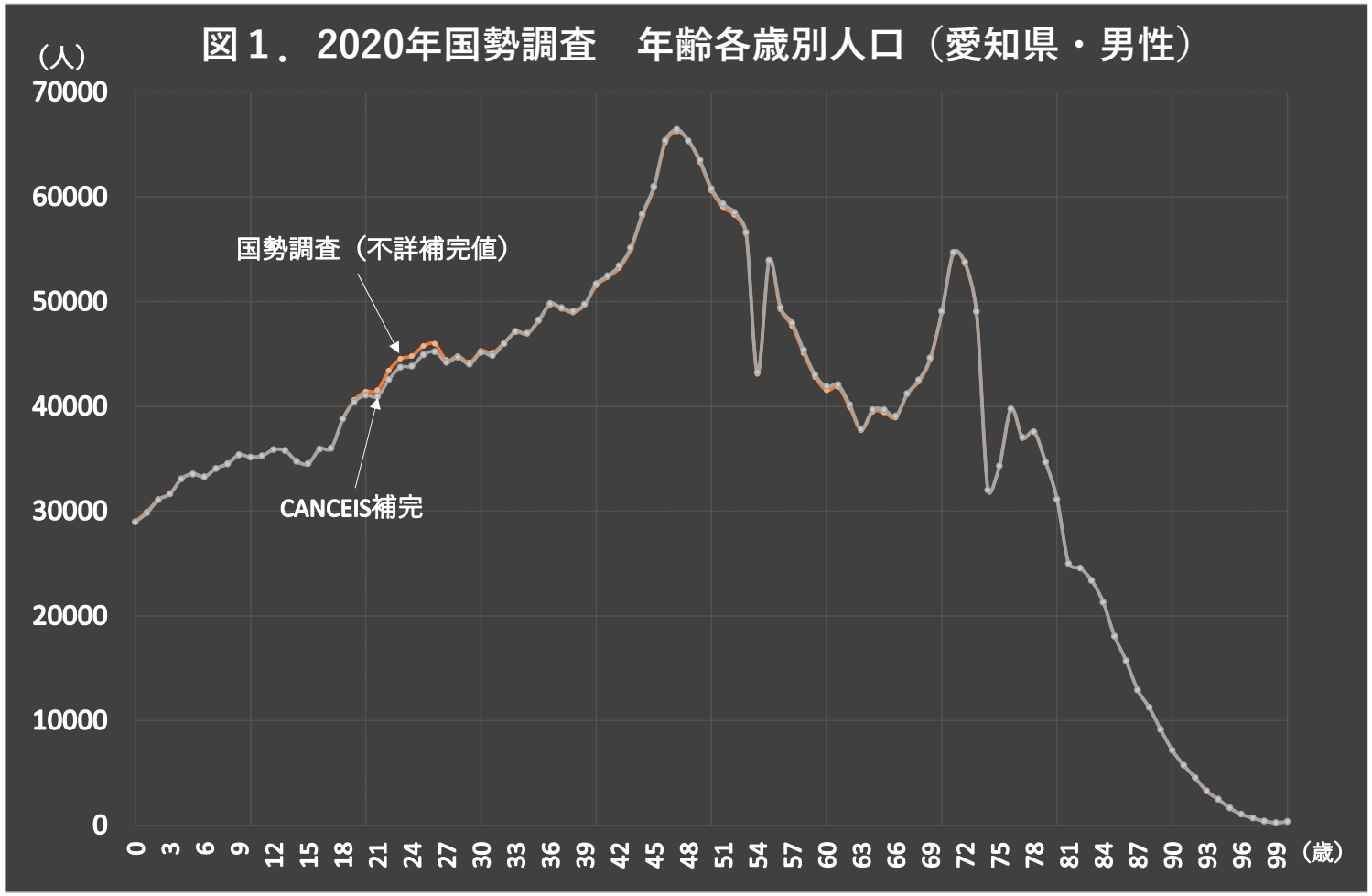
【都道府県別】



CANCEIS補完の試算結果①（男性・年齢）

◆ 年齢各歳別人口（CANCEIS補完結果と不詳補完値の比較）

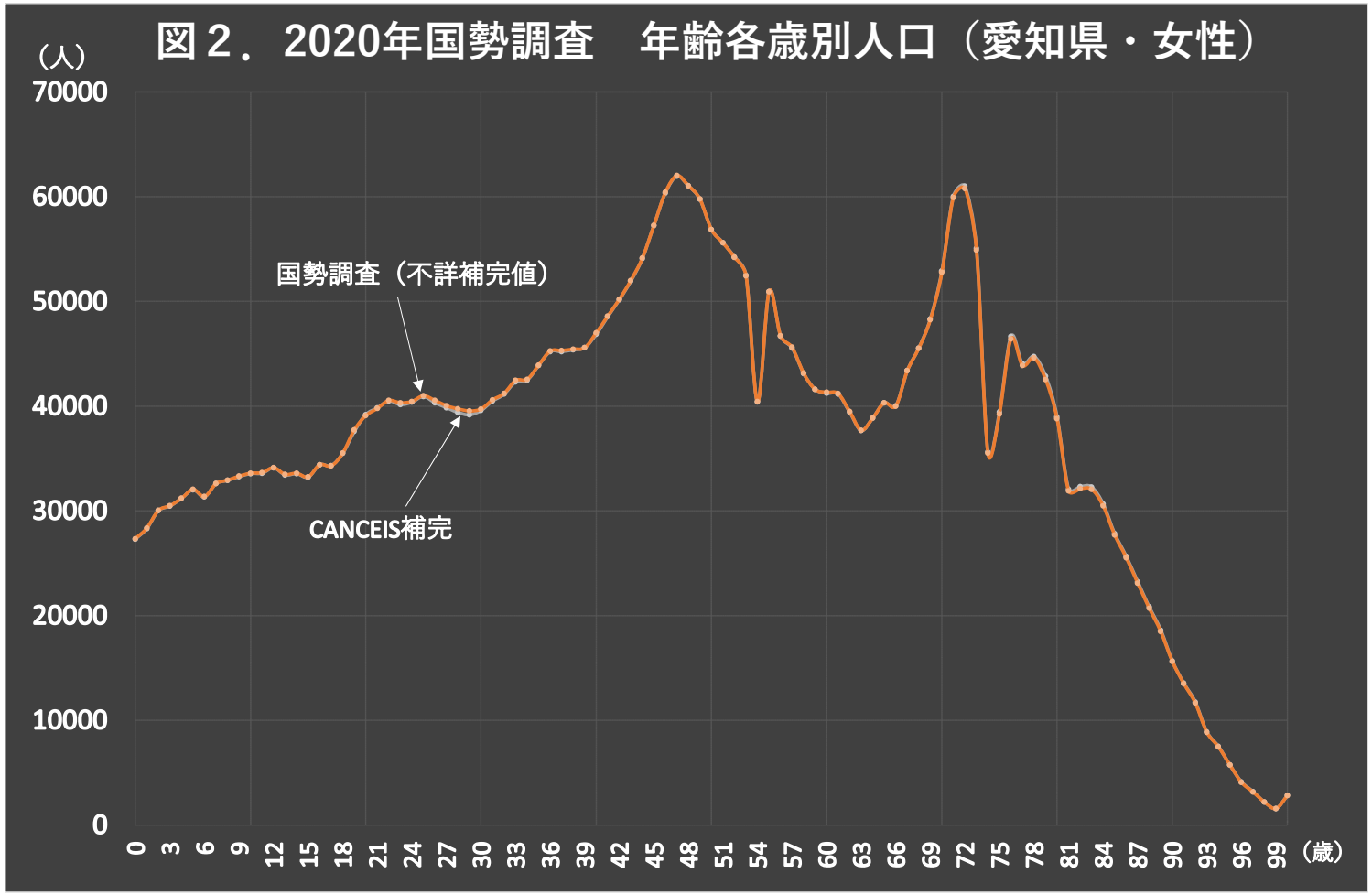
➡ CANCEIS補完結果と不詳補完値は、お互いが近接した数値となった。



CANCEIS補完の試算結果②（女性・年齢）

◆ 年齢各歳別人口（CANCEIS補完結果と不詳補完値の比較）

➡ CANCEIS補完結果と不詳補完値は、お互いが近接した数値となった。

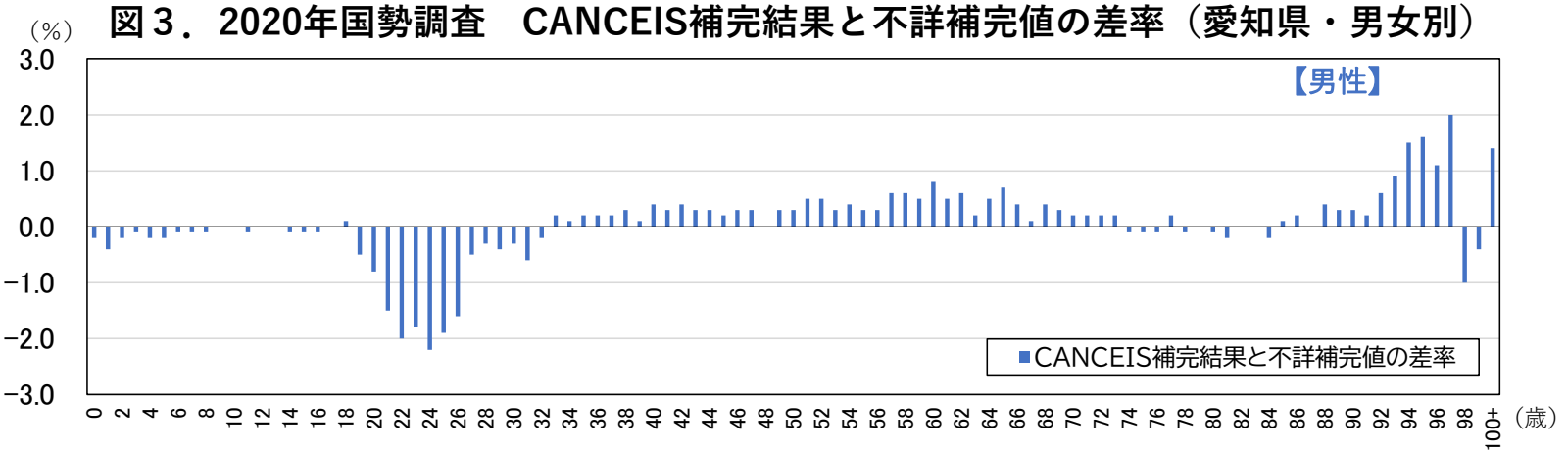


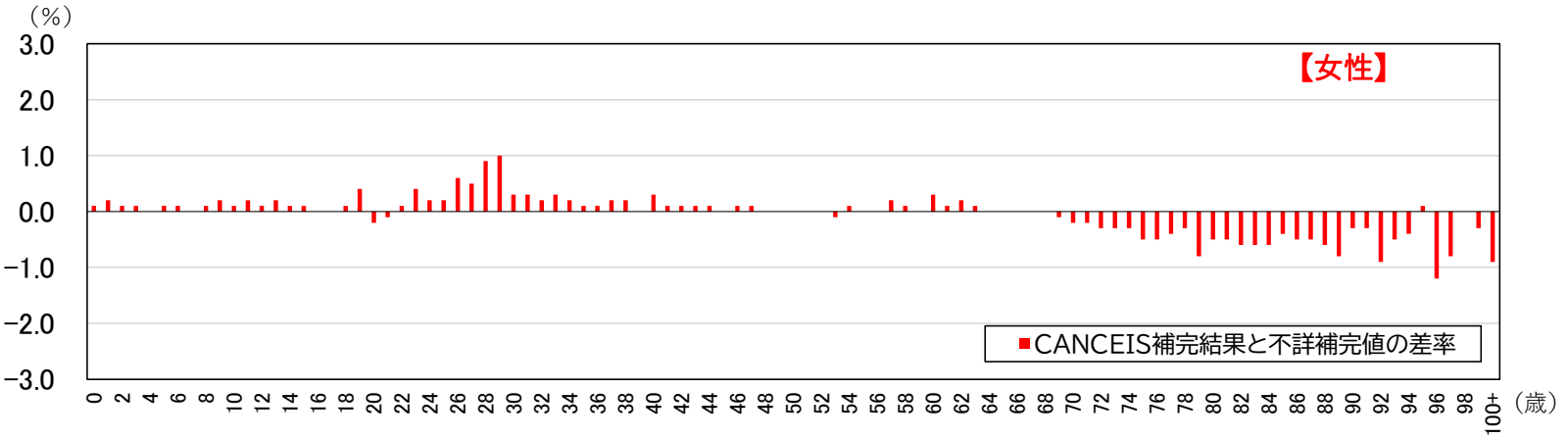
CANCEIS補完の試算結果と不詳補完値の差率

◆ 年齢各歳別人口におけるCANCEIS補完結果と不詳補完値の差率

➡ 男女ともに差率は±２％以内

図３．2020年国勢調査　CANCEIS補完結果と不詳補完値の差率（愛知県・男女別）





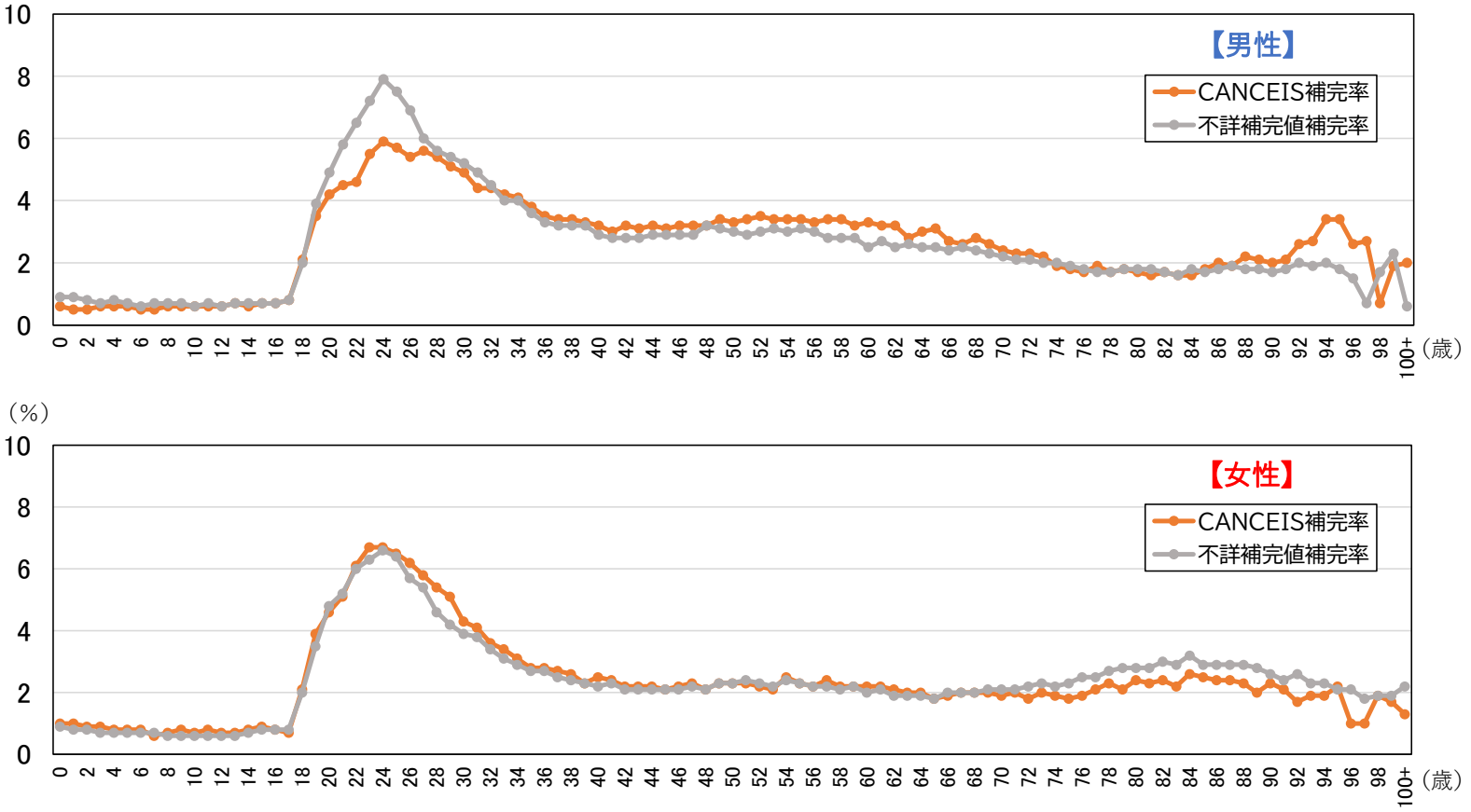
注）　差率（%）は、（CANCEIS補完結果ー不詳補完値）／CANCEIS補完結果×100により算出

CANCEIS補完の試算結果と不詳補完値の補完率

◆ 年齢各歳別補完率（CANCEIS補完結果と不詳補完値）

➡ 男性では、20代及び90代で補完率に差が見られるが、女性ではほぼ同率

図４．2020年国勢調査　CANCEIS補完結果と不詳補完値の補完率（愛知県・男女別）



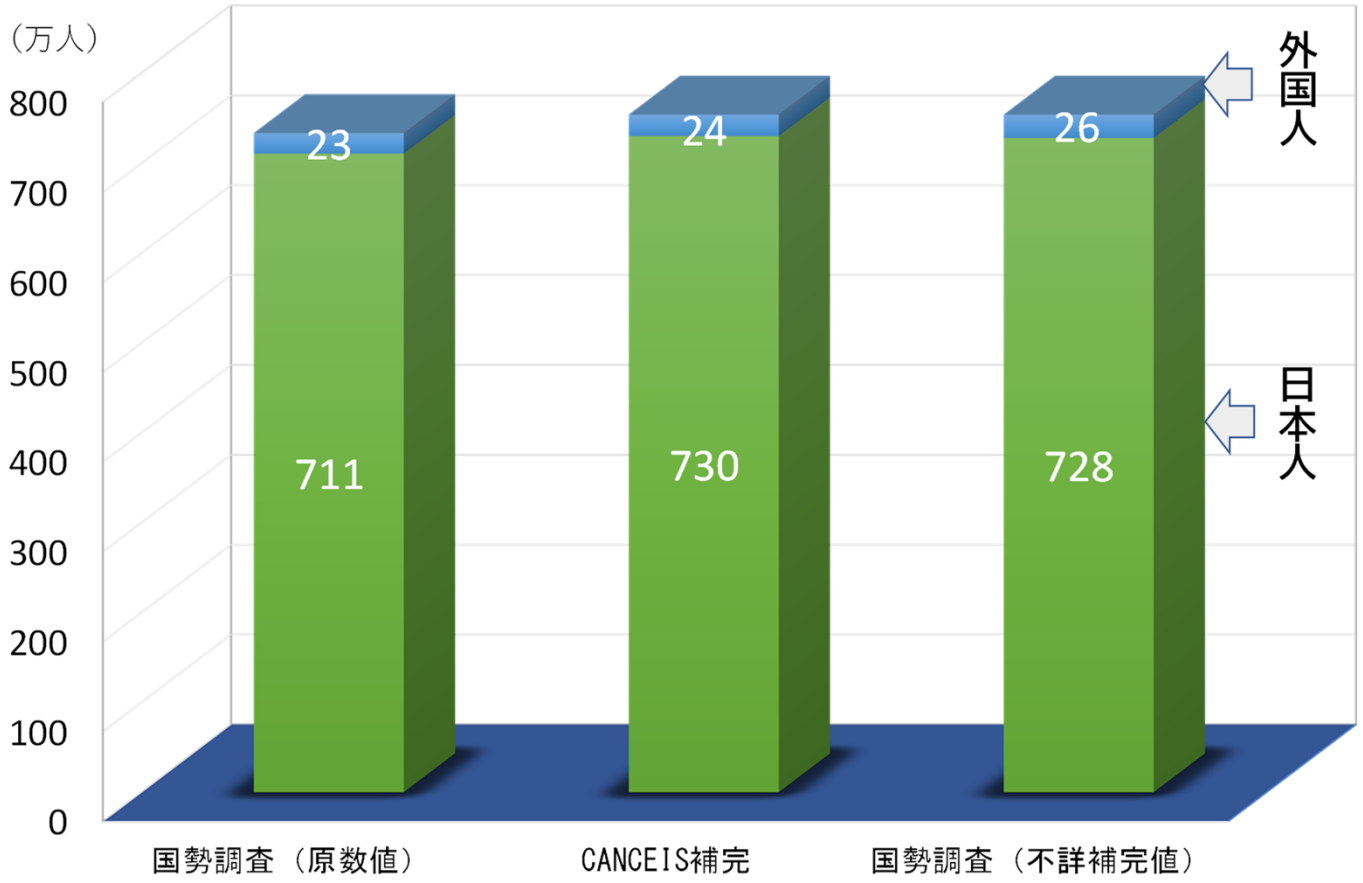
注１）　CANCEIS補完率（%）は、（CANCEIS補完結果ー原数値）／ CANCEIS補完結果×100により算出
注２）　不詳補完値補完率（%）は、（不詳補完値ー原数値）／不詳補完値×100により算出

CANCEIS補完の試算結果（国籍）

◆国籍（日本人・外国人）別人口

➡CANCEIS補完結果は、不詳補完値に近接した数値

図５．2020年国勢調査　国籍（日本人・外国人）別人口（愛知県）



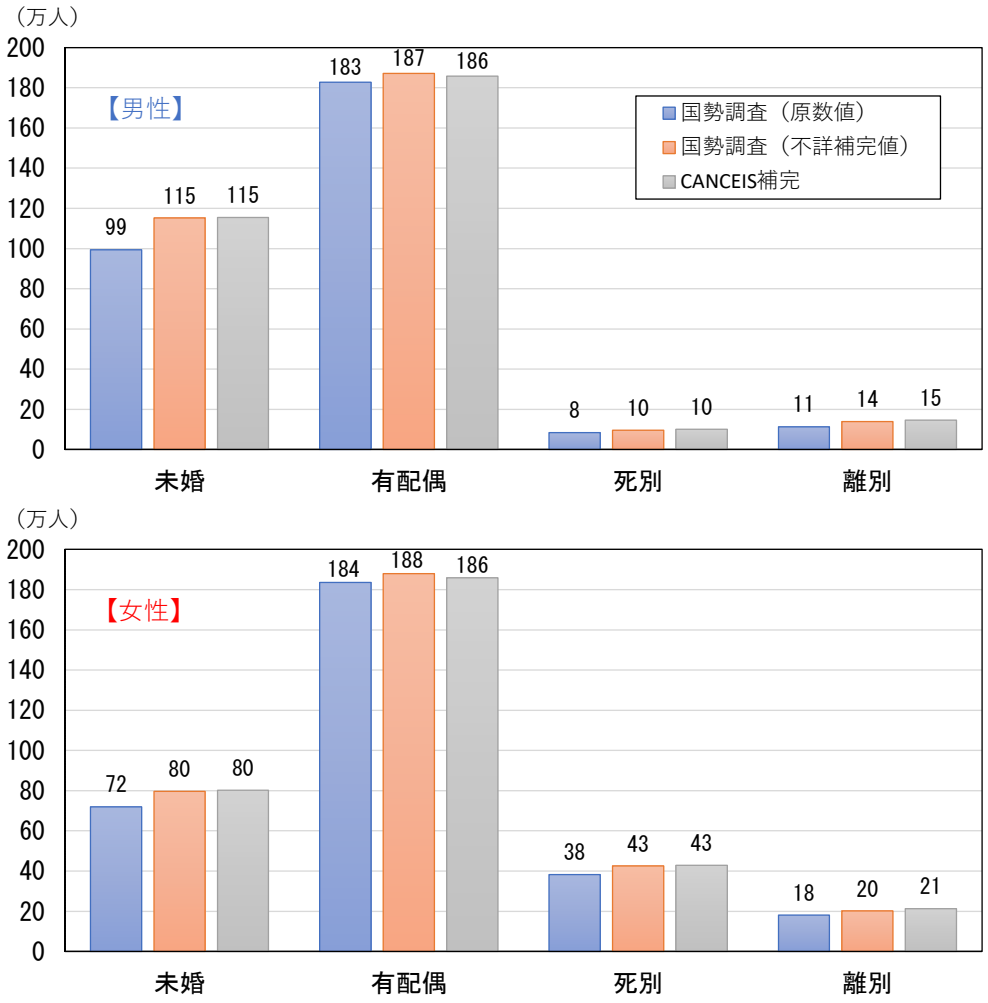
CANCEIS補完の試算結果（配偶関係）

◆ 配偶関係別人口

➡ CANCEIS補完結果は不詳補完値に近接した数値

- 男性では、不詳補完値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
- 女性では、未婚以外に死別も不詳補完値と同様、顕著に増加

図６．2020年国勢調査　男女・配偶関係別人口（愛知県）



試算結果のまとめと今後の検証事項

◯CANCEIS補完値を試算したところ、令和２年国勢調査の不詳補完値に近い結果を得たことから、我が国の国勢調査においても、CANCEISの適用により年齢・国籍等の不詳を補完することが可能と考えられる。

◯今回は、一部地域に限定して試算を行ったため、今後、

- 試算対象地域を拡大するとともに、市区町村レベルや小地域レベルでの補完が可能か
- 調査事項を拡大し、就業状態等についてもCANCEISが適用可能か

などについて検証を行う。