資料４（参考）

令和７年国勢調査におけるCANCEIS補完の実装に向けて（参考資料）

令和６年３月

総務省統計局

単純なホットデック法における課題

○**ホットデック法**は、欠測値や矛盾値を**同一のデータセット内のデータから一定の方法で補完する方法**

○**単純なホットデック法の運用では、エディットルールを充足する保証がない**

（補完前後でデータが変わり、補完前にパスしたチェックを補完後にパスしない可能性がある）

例）年齢、配偶関係、産業が不詳の場合　⇒　慎重にエディットしないと、年齢を14歳に補完した場合などに配偶関係や産業に矛盾が生じる

＜ホットデック法で性別のみ補完する例＞

ID	世帯主との続き柄	性別	年齢	性別の補完列
1	1 世帯主	1　男	39	性別＝1
2	2 配偶者	**2　女**	35	**性別＝2**
3	3 子	1　男	13	性別＝1
4	3 子		10	性別＝1
5	4 その他の親族	**2　女**	40	**性別＝2**
6	4 その他の親族	1　男	ブランク	性別＝1
7	4 その他の親族	2　女	13	性別＝2
8	5 親族以外		ブランク	性別＝2
9	5 親族以外	1　男	44	性別＝1
10	5 親族以外	2　女	36	性別＝2

男女はほぼ同頻度で出現するため、
最終的な補完回数は男女で半数ずつになると想定

＜ID順に処理＞

①　処理の都度、性別の補完列を置換

※　例えば、ID２に到達した際、ID２の「性別」欄のデータ（性別２）に性別の補完列を置換

※　３人目は男性なので、性別の補完列は再び１になる。

②　ブランクのレコードに遭遇した際、その直前の性別の補完列の値を補完

※　４人目の性別はブランクのため、性別の補完列の値（この場合は「１男性」）から、ブランクに「１男性」を補完

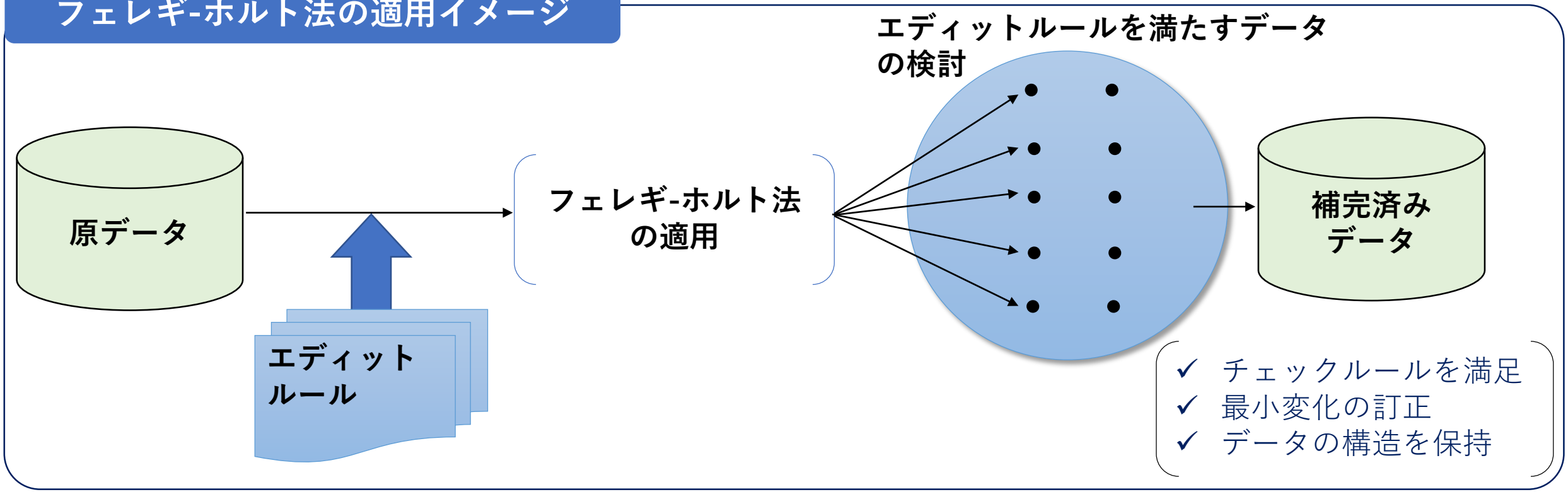
③　５人目は女性なので性別の補完列を「２女性」に置換

従前の補完方法（フェレギ・ホルト法）

○ フェレギ・ホルト法は、補完により変更されたレコードがエディットに失敗しないことを保証する全体的なモデルであり、以下の原理により、データの欠測値を修正するためのアルゴリズムを提供
- ✓ 各レコードが全てのチェックルールを満足する
- ✓ できるだけ少ない変更で訂正が達成される
- ✓ データの構造を保持する補完手順となっている

○ これにより、上記の利点を保持した上で、ドナー候補を探索し、欠測値を補完可能
○ ただし、エディットする変数を特定し、ドナーを探索する流れはデータ処理上、非効率

フェレギ-ホルト法の適用イメージ



CANCEISのドナー検索・補完の流れ

○ CANCEISは、**要補完ユニット**のドナー候補として、要補完ユニットと**距離が近く、類似性の高い複数のユニット（NMCIA（注1））をリスト化し、その中から最終ドナーをランダムに選択して補完**

注1）NMCIA･･･Near Minimum Change Imputation Actions

ドナー検索・補完の流れ

※　単身世帯の例

要補完ユニット（ID＝１）

ユニット(世帯)番号	世帯員番号	年齢	配偶関係
1	1	25	

第１ステージ

✓ 要補完ユニットの近隣の500ユニットからドナーを検索

ユニット(世帯)番号	世帯員番号	年齢	配偶関係
5	1	34	離別
8	1	27	未婚
⋮	⋮	⋮	⋮

✓ ドナー候補の中から、NMCIAを10ユニット順次選定

ユニット(世帯)番号	世帯員番号	年齢	配偶関係
20	1	23	未婚
31	1	25	未婚
⋮	⋮	⋮	⋮

✓ NMCIAの中で最適な補完を実現する補完アクション（IA）を特定

ユニット(世帯)番号	世帯員番号	年齢	配偶関係
31	1	25	未婚

第２ステージ（検索範囲の拡大）

✓ 第１ステージとは別の500ユニットからドナーを検索

ユニット(世帯)番号	世帯員番号	年齢	配偶関係
510	1	54	離別
513	1	47	未婚
⋮	⋮	⋮	⋮

更新

✓ NMCIAを更新（第１ステージより優れたドナーがある場合）

ユニット(世帯)番号	世帯員番号	年齢	配偶関係
520	1	25	未婚
31	1	25	未婚
⋮	⋮	⋮	⋮

✓ IA より有意に（注2）優れたIAがなければ検索を終了し、第２ステージ終了時点のNMCIAからランダムに最終ドナーを選定

✓ そうでなければ第３ステージに進み、1,000ユニットを検索

✓ 以下同様に最大10ステージまで検索（３ステージ以降、検索数はステージごとに倍増）

注2）第１ステージの最適な IA に比べた第２ステージの IA の品質（P8の品質評価式の値）の改善が10%未満の場合、それ以上のステージの追加は無意味と判断し、検索を終了

「距離が近い」ユニット（距離関数）

○ CANCEISの距離関数は、あるユニット（世帯）の調査票への回答内容と、別のユニットの回答内容との近さを定量化するための指標であり、年齢、国籍、世帯主との続き柄などの質的な相違も加味して算出※

※　ユニット間の地理的な近さも加味

D_{fp}:要補完ユニット(FU)と通過ユニット(PU)の距離

V_{fi}:要補完ユニットにおけるi番目の変数

V_{pi}:通過ユニットにおけるi番目の変数

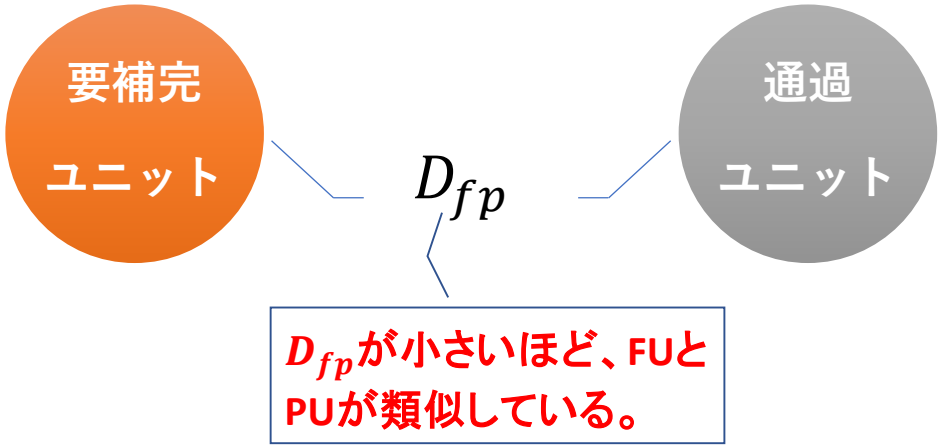
$D_i(V_{fi}, V_{pi})$: V_{fi}とV_{pi}の距離

w_i: i番目の変数におけるウエイト

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

具体例

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



※ V_{fk}とV_{pk}の値が一致するかしないかにより、 $D_k(V_{fk}, V_{pk})$を０～１の数値に設定する。

$$D_{fp} = w_1 D_1(V_{f1}, V_{p1}) + w_2 D_2(V_{f2}, V_{p2}) + \cdots + w_n D_n(V_{fn}, V_{pn})$$

ユニット	年齢	国籍	・・・	世帯の種類
要補完ユニット	V_{f1}	V_{f2}	・・・	V_{fn}
通過ユニット	V_{p1}	V_{p2}	・・・	V_{pn}

✓ 例えば、国籍が両ユニットとも日本人ならD$_2$は０で、それにウエイトを掛ける。

距離のイメージ

○ ユニット番号１を要補完ユニットとすると、ドナー候補は、世帯員数が同一のユニット４及び５がドナー候補
○ これを要補完ユニットとの距離で表すと、ユニット４との距離（０）の方がユニット５との距離（７）より近いと判断

※　簡単化のため、いずれのウエイトも１とした

回答データ　　**距離関数で出した距離**

ユニット(世帯)番号	世帯員番号	年齢	国籍	続き柄	配偶関係	年齢	国籍	続き柄	配偶関係
1	1	45	日本人	世帯主	配偶者あり	-	-	-	-
1	2	46	日本人	配偶者		-	-	-	-
1	3	18	日本人	子	未婚	-	-	-	-
1	4	16	日本人	子		-	-	-	-
2	1	55	日本人	世帯主	配偶者あり	-	-	-	-
2	2	20	日本人	配偶者	配偶者あり	-	-	-	-
3	1	19	日本人	世帯主	未婚	-	-	-	-
4	1	45	日本人	世帯主	配偶者あり	0	0	0	0
4	2	46	日本人	配偶者	配偶者あり	0	0	0	0
4	3	18	日本人	子	未婚	0	0	0	0
4	4	16	日本人	子	未婚	0	0	0	0
5	1	45	アメリカ	世帯主	配偶者なし	0	1	0	1
5	2	46	アメリカ	兄弟姉妹	配偶者なし	0	1	1	0
5	3	18	アメリカ	子	未婚	0	1	0	0
5	4	16	アメリカ	他の親族	未婚	0	1	1	0

ユニット１と４の距離
$D_{f4}=0$

ユニット１と５の距離
$D_{f4}=7$

CANCEISの特徴

ドナー候補による補完の品質（Quality of Imputation Action）

○ CANCEISにおける各ドナー候補による補完の品質は、(1)**原データからの最小変化性(D_{fa}：③と①の距離)**及び(2)**実在可能性(D_{ap}：③と②の距離)**について下式で定義され、ドナー検索の結果、下式の値が小さい10ユニットがNMCIAに登録される。

○ NMCIAの中からランダムに１ユニットが、最終的なドナーに選定される。

※ なお、我が国のように、CANCEIS補完において欠測値補完のみ行う場合は、原データからの最小変化性は担保されるため、D_{fa} はゼロとなり、D_{fpa}は D_{ap} のみに依存する。

＜品質評価式＞

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

D_{fpa}：加重平均
D_{fa}：①要補完ユニットと③仮補完ユニットとの距離
D_{ap}：③仮補完ユニットと②通過ユニットとの距離

＊） α：ユーザー定義システムパラメータ（$0.5 < \alpha \leqq 1.0$）

単身世帯のユニット1（①要補完ユニット）に対し、近隣の単身世帯500ユニットの品質を確認する事例

①補完前のデータ（要補完ユニット）

ユニット（世帯）番号	世帯員番号	年齢	住宅の建て方	性別	続き柄	配偶関係
1	1	25	共同住宅	男	世帯主	
5	1	34	一戸建て	男	世帯主	離別
8	1	27	一戸建て	男	世帯主	未婚
12	1	35	一戸建て	女	世帯主	配偶者あり
15	1	28	共同住宅	女	世帯主	未婚
20	1	23	共同住宅	男	世帯主	未婚
：	：	：	：	：	：	：

②実在するデータ（通過ユニット）
※要補完ユニットとのレコードの位置が近接する500ユニット

③各ドナー候補の配偶関係を仮に①に補完したデータ（仮補完ユニット）から、D_{ap}を算出

D_{fa}：最小変化性 ③と①の距離	D_{ap}：実在可能性 ③と②の距離
0.00	2.00
0.00	1.10
0.00	3.00
0.00	1.15
0.00	0.10
：	：

D_{fa}とD_{ap}の加重平均D_{fpa}が最小となる10ユニットがNMCIAに入る

※） 年齢の距離は、両者の年齢差が1～2歳程度であれば、ほぼゼロとなり、年齢差が6歳以上であれば1となる関数

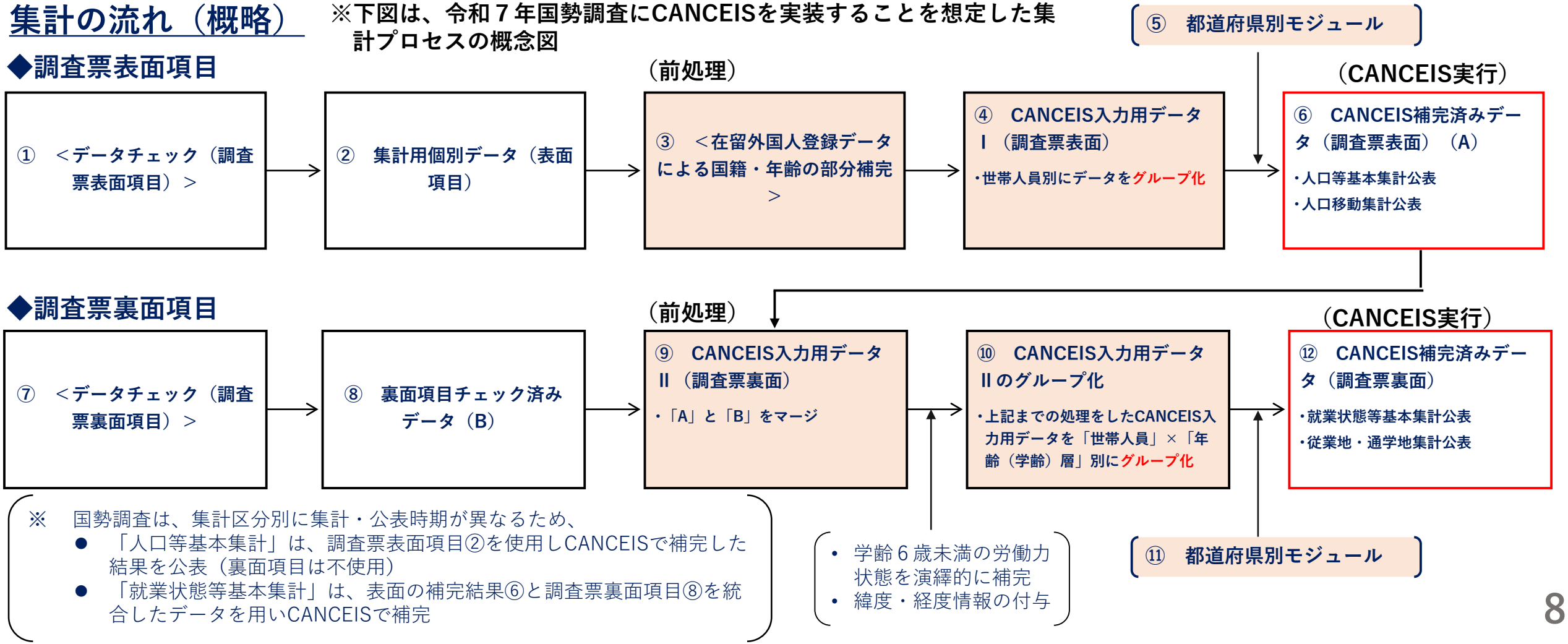
CANCEISの実装

既存の集計プロセスと整合したCANCEIS補完の実装

○ CANCEISは、国勢調査における既存のデータチェックと演繹的補完の拡張として、ドナー補完を体系的に導入するためのプログラムとして位置付けることが可能
○ 従前のデータチェックと整合したエディットルールをCANCEISのモジュール（⑤、⑪）に組み込むことにより、既存の集計プロセスと整合した集計プロセスの一環としてCANCEISを実装することが可能

集計の流れ（概略）

※下図は、令和７年国勢調査にCANCEISを実装することを想定した集計プロセスの概念図



国籍データの事前補完

○ **国籍データは、人口構成比（日本人の比率と外国人の比率）が不均衡であり、事前処理なしにCANCEISを実行すると、外国人の過少補完となるため、CANCEIS実行前の事前補完が必須**
○ 事前処理は、令和２年調査の不詳補完値作成方法に準じつつ、CANCEISで代替可能な処理（A. 二人以上の世帯及び C. 単身世帯のうち民営賃貸共同住宅に居住している年齢不詳の者に係る処理）は除外

令和２年不詳補完値における部分補完の方法

A　二人以上の世帯
小地域別、男女・世帯人員の構成別、住宅の建て方別に、**基本項目不詳世帯以外の世帯をドナーとしたホットデック法により、世帯員の年齢及び国籍の不詳を補完**

B　単身世帯で国籍不詳の者
小地域別、男女別に、在留外国人登録データを活用したコールドデック法により国籍及び年齢の不詳を補完

C　単身世帯のうち、民営賃貸共同住宅に居住している年齢不詳の者
市（区）町村別、男女別に年齢を確率的に補完

CANCEISにおける補完処理で代替可能のため、Ｒ２年のような事前の部分補完は不要

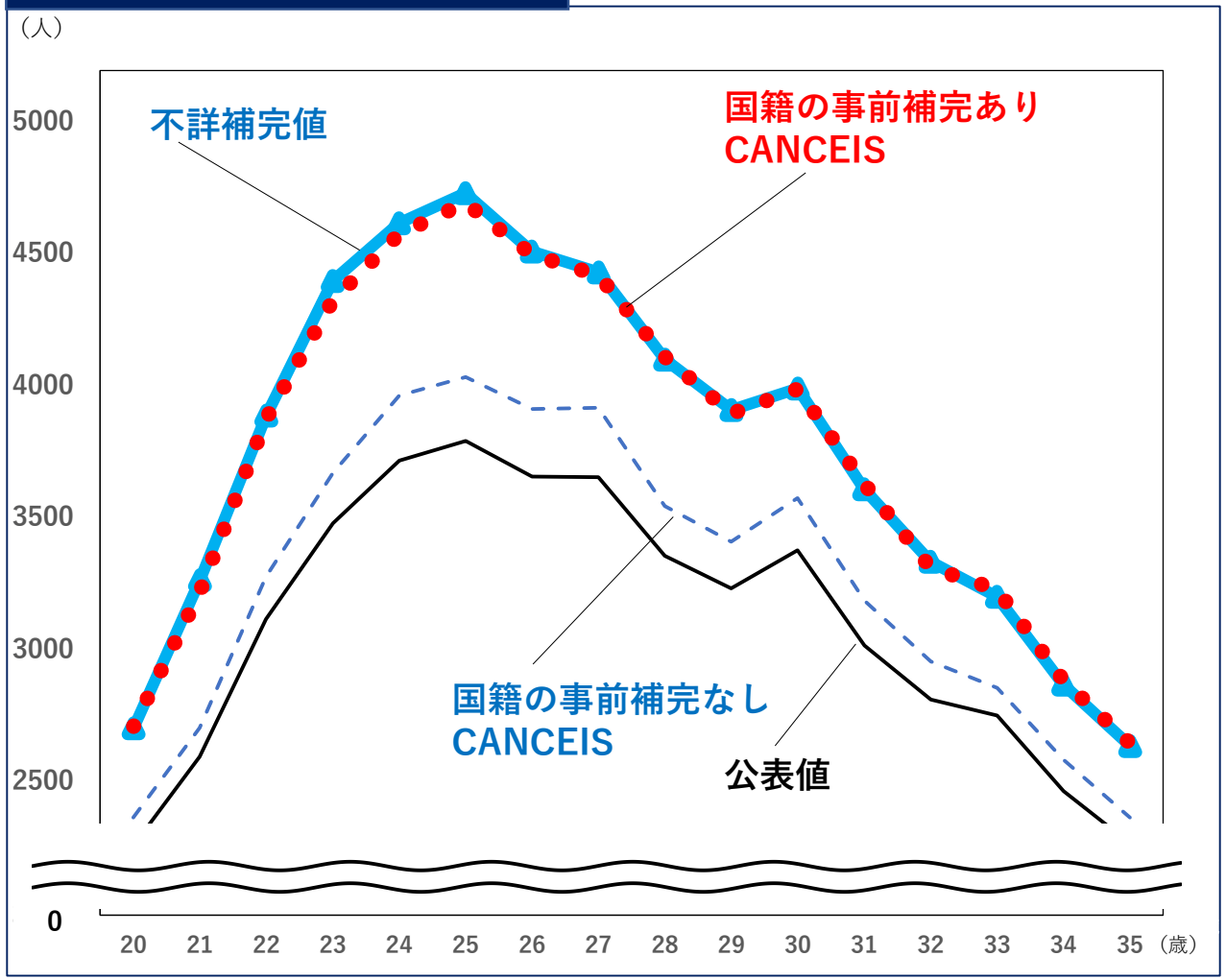
令和７年国勢調査の集計におけるCANCEIS適用の前処理として継承

CANCEISにおける補完処理で代替可能のため、Ｒ２年のような事前の部分補完は不要

国籍の事前補完を行ったCANCEIS試算結果

○ 下図は、令和２年国勢調査の人口等基本集計におけるCANCEISの適用結果を国籍の事前補完の有無別に示したもの。国籍の事前補完ありのCANCEIS適用結果は、国籍の事前補完なしのCANCEIS適用結果よりも不詳補完値に近い結果となった。

年齢別人口(愛知県)・外国人男性



年齢別人口(愛知県)・外国人女性