資料３（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２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

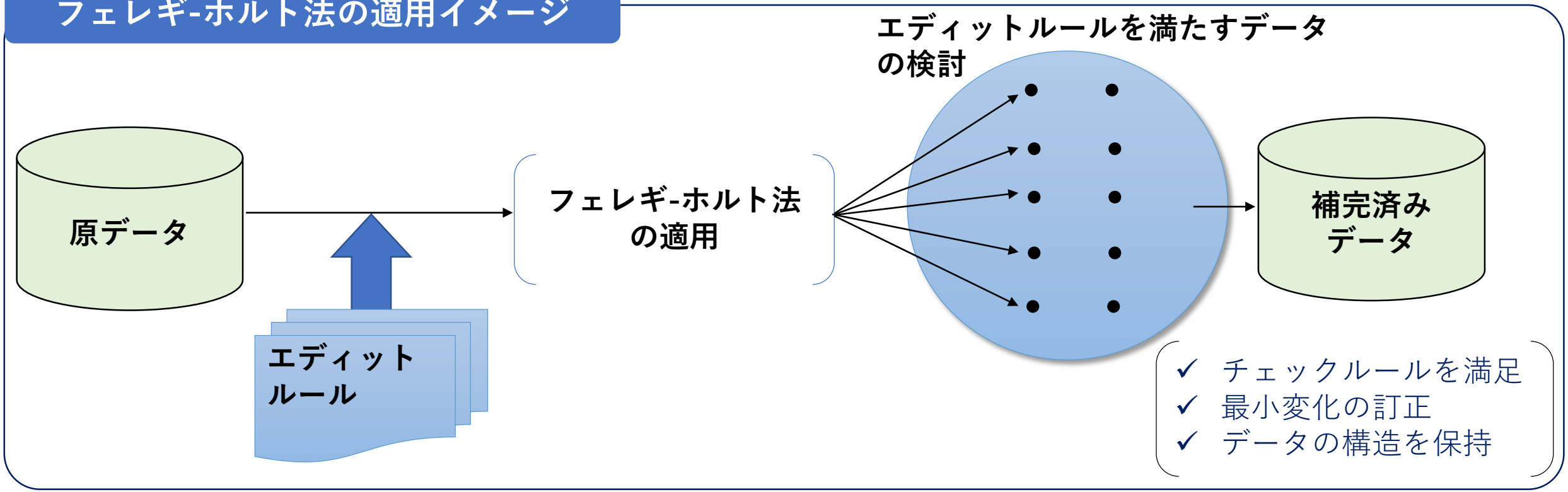
③　５人目は女性なので性別の補完列を「２女性」に置換

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

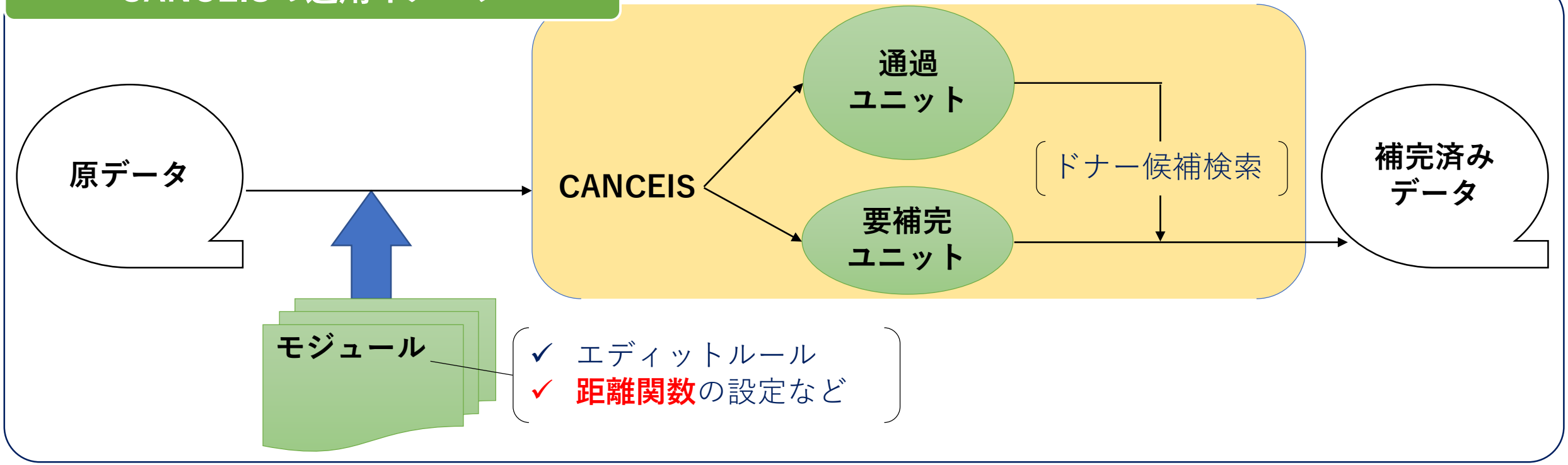


CANCEISの特徴

○ CANCEISは、フェレギ・ホルト法の利点を取り込んだ上で、データ駆動の効率化を図る手法
○ 欠測値があるユニット（**要補完ユニット**）を特定し、エディットルールを考慮した上、チェックをパスしたユニット（**通過ユニット**）の中から要補完ユニットに類似したユニットを**ドナー候補ユニット**として選定し、その中からランダムに特定のドナーを選定

○ 補完範囲を最小限に限定し、かつ、それ以外のデータを極力変更しないことで、原データの分布構造を最大限保持
○ 個別データレベルで補完するため、統計表間・集計区分間の整合性や二次利用の観点で利点がある

CANCEIS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

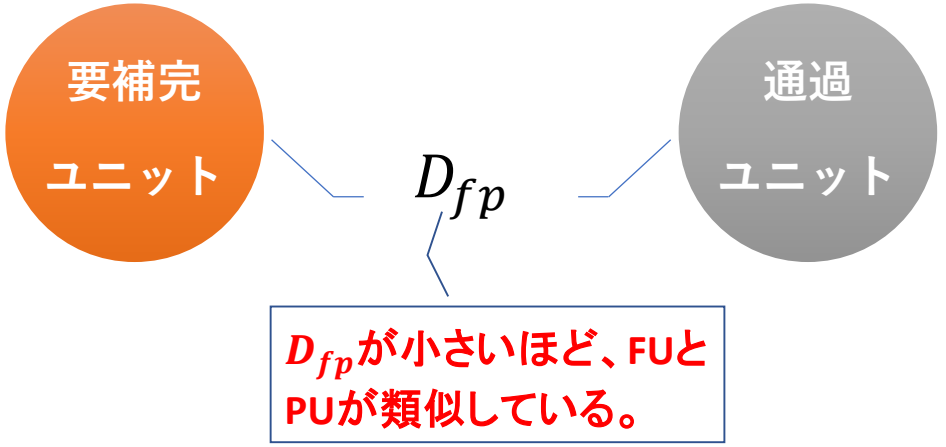
$D_i(V_{fi}, V_{pi})$: V_{fi}とV_{pi}の距離

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補完の品質評価指標

○ CANCEISのドナー補完の品質は、補完後のデータの（1）**原データからの最小変化性**及び（2）**実在可能性**をもっとも満たすデータをドナー選定することにより担保される。
（1）**最小変化性**：①補完前のデータと③仮補完したデータとで、データの変化（距離）が最小であること
（2）**実在可能性**：③仮補完したデータが、②実在するデータと類似しているか（尤もらしいか）

CANCEIS補完の品質評価指標

D_{fp} と D_{fpa}が最小となるユニット20がNMCIAに入る

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	

D_{fp}
（①と②の距離）

D_{fa}：最小変化性
①と③の距離

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）

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	D_{fa}	D_{ap}
1	1	25	共同住宅	男	世帯主	離別	0.00	2.00
1	1	25	共同住宅	男	世帯主	未婚	0.00	1.10
1	1	25	共同住宅	男	世帯主	配偶者あり	0.00	3.00
1	1	25	共同住宅	男	世帯主	未婚	0.00	1.15
1	1	25	共同住宅	男	世帯主	未婚	0.00	0.10

（仮補完）

②実在するデータ（通過ユニット）

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	D_{fp}
5	1	34	一戸建て	男	世帯主	離別	2.00
8	1	27	一戸建て	男	世帯主	未婚	1.10
12	1	35	一戸建て	女	世帯主	配偶者あり	3.00
15	1	28	共同住宅	女	世帯主	未婚	1.15
20	1	23	共同住宅	男	世帯主	未婚	0.10

※） ①と②の年齢の距離は、両者の年齢差が1～2歳程度であればほぼ0、6歳以上であれば1となる関数。

D_{ap}：実在可能性
②と③の距離

＜品質評価式＞

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

加重平均　①要補完ユニットと③仮補完ユニットとの距離　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（$0.5 < \alpha \leqq 1.0$）

令和７年国勢調査にCANCEISを実装するために必要な処理（ポイント）

1. 既存の集計プロセスと整合したCANCEIS補完の実装（p10）

- 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であるため、既存の集計プロセスと整合した実装が必要

2. 国籍データの事前補完（p11～12）

- 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

3. 都道府県別モジュールの作成（p13～14）

- CANCEISの稼働に不可欠な変数の定義や距離関数の指定などのため、モジュールの作成が必要

4. データのグループ化（p15～16）

- 適切な補完を行うため、データのグループ化や並び替えが必要

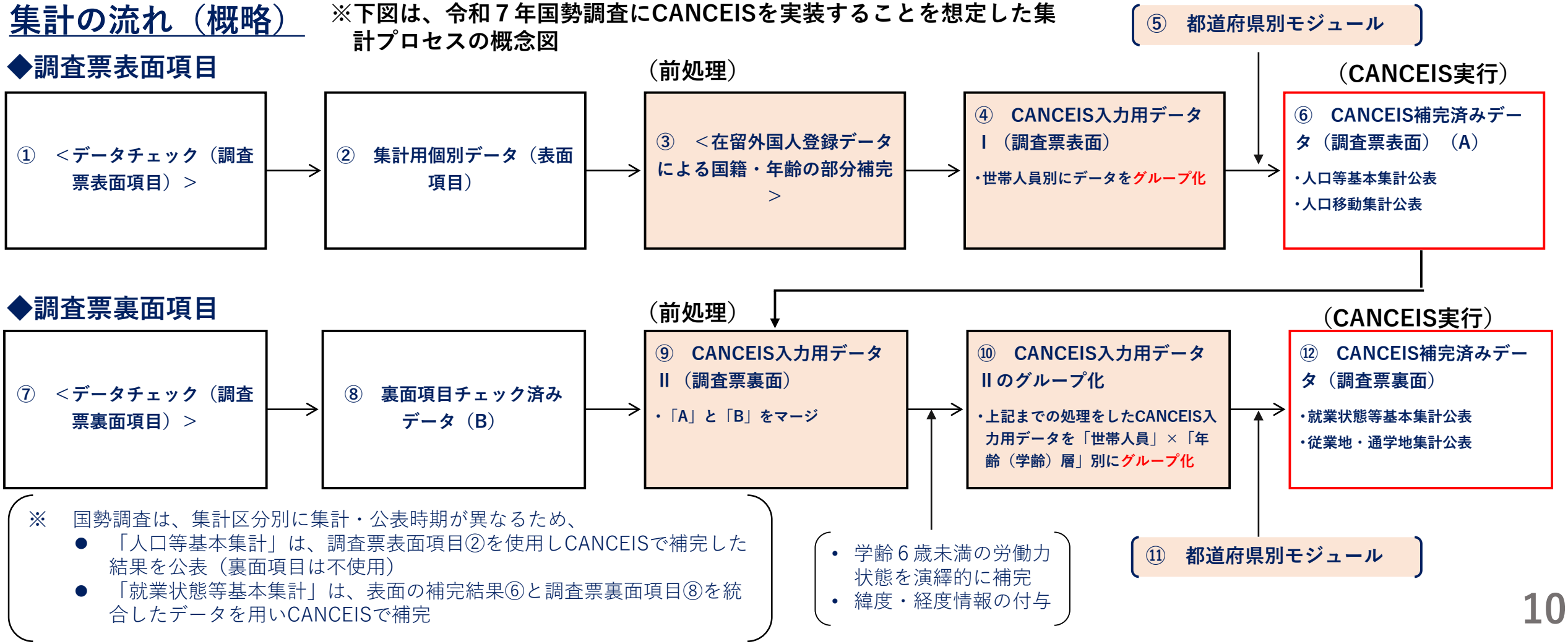
CANCEISの実装

1. 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



2. 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

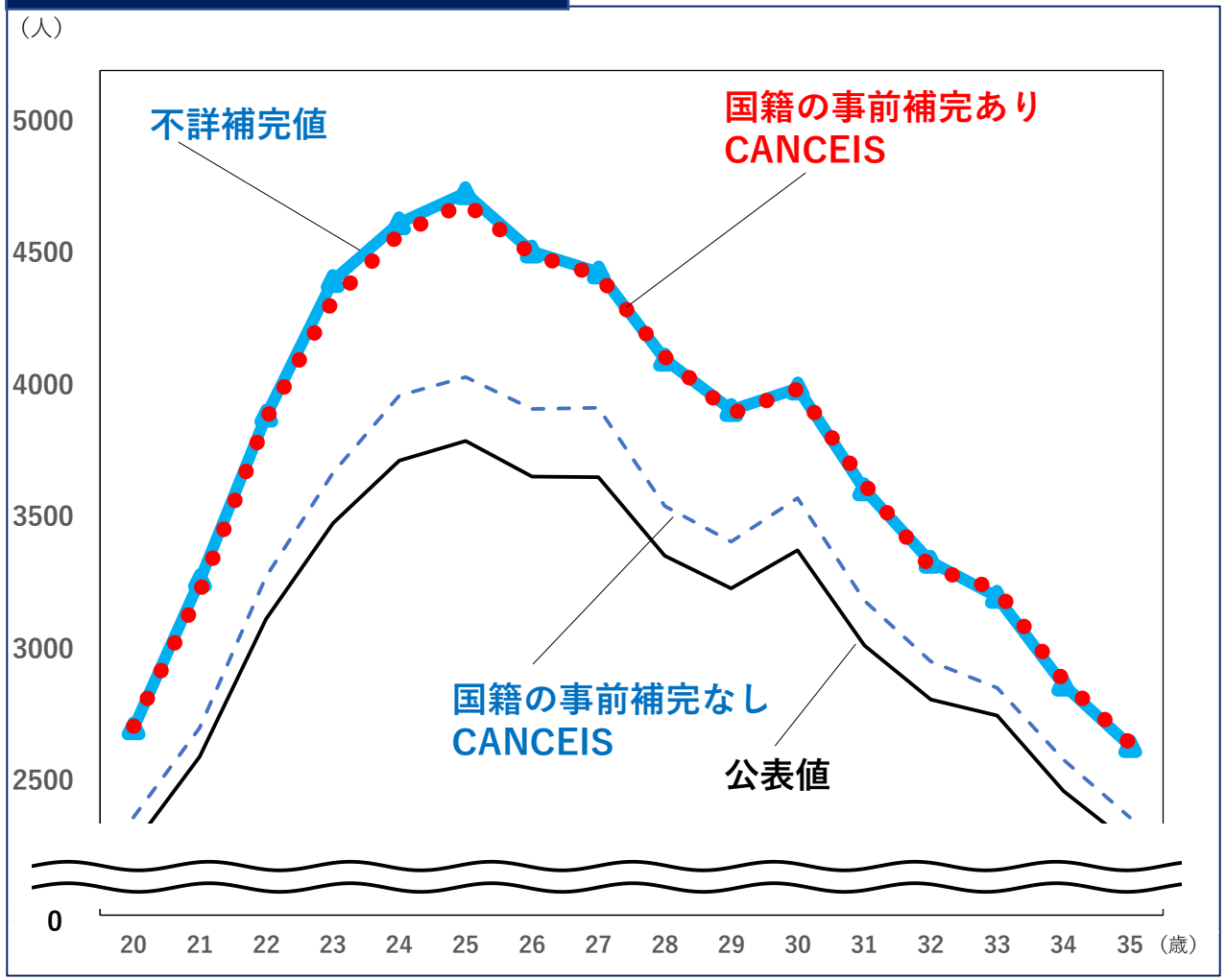
令和７年国勢調査の集計におけるCANCEIS適用の前処理として継承

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

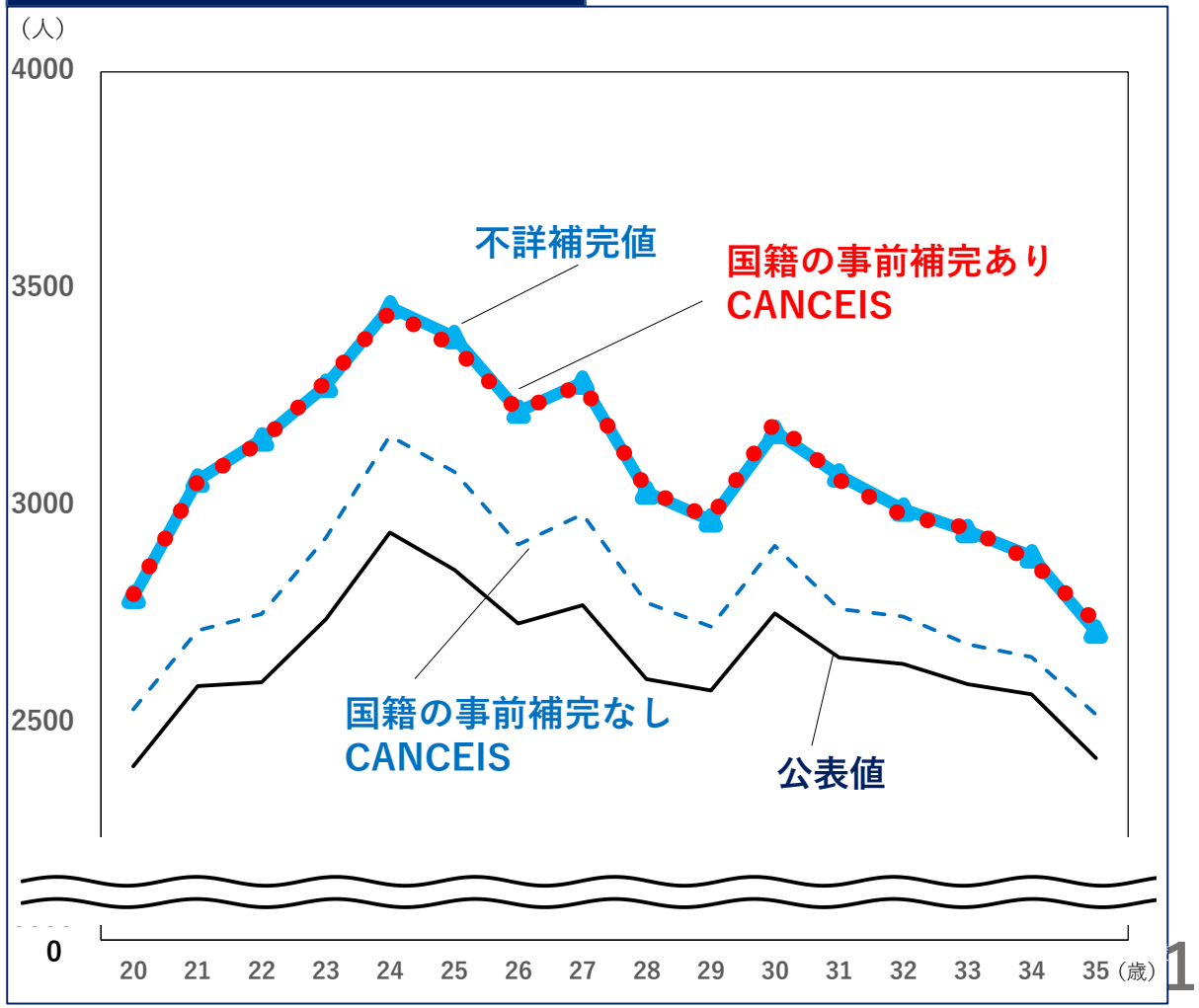
（参考）国籍の事前補完を行ったCANCEIS試算結果

○ 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



3. CANCEIS実装に必要な処理：⑪都道府県別モジュールの作成

○ CANCEISのモジュールは、入力データの読込方法、データの処理方法、処理後のデータやレポートの書き出し方法等を指定する機能群であり、CANCEISの稼働に不可欠な変数の定義だけでなく、補完が必要なユニットを特定し、それに最も近いドナーを割り当てるための距離関数の指定など、的確なCANCEIS補完に必須の要素。
○ 本試算では、特に以下の３点に留意してモジュールを作成した。

（１）DLT（Decision Logic Table）の設定

- **DLTは、各ユニットの整合性を判定するためのルールを設定するもの**
- 「⑨CANCEIS入力用データⅡ（調査票裏面）」は、調査票表面項目の補完済みデータと調査票裏面項目のチェック済みデータをマージしたものであり、マージ後のデータのチェックは行っていないため、本来データチェックで行うチェック項目のうち、主要なものをDLTとして取り込むことで、データの不整合を抑制

（２）学齢カテゴリ変数の設定

- **労働力状態や従業地・通学地が特定の学齢カテゴリ**（６～11歳：小学生、12～14歳：中学生、15～17歳：高校生、・・などの通学年齢層）**で不連続に変化するため、学齢カテゴリを設定することで、学齢カテゴリ間を跨いだ不詳とドナーの組合せにならないよう制限。**これにより、例えば「23,4歳の非就業者」や「18歳未満の就業者」が不自然に増加することを抑制

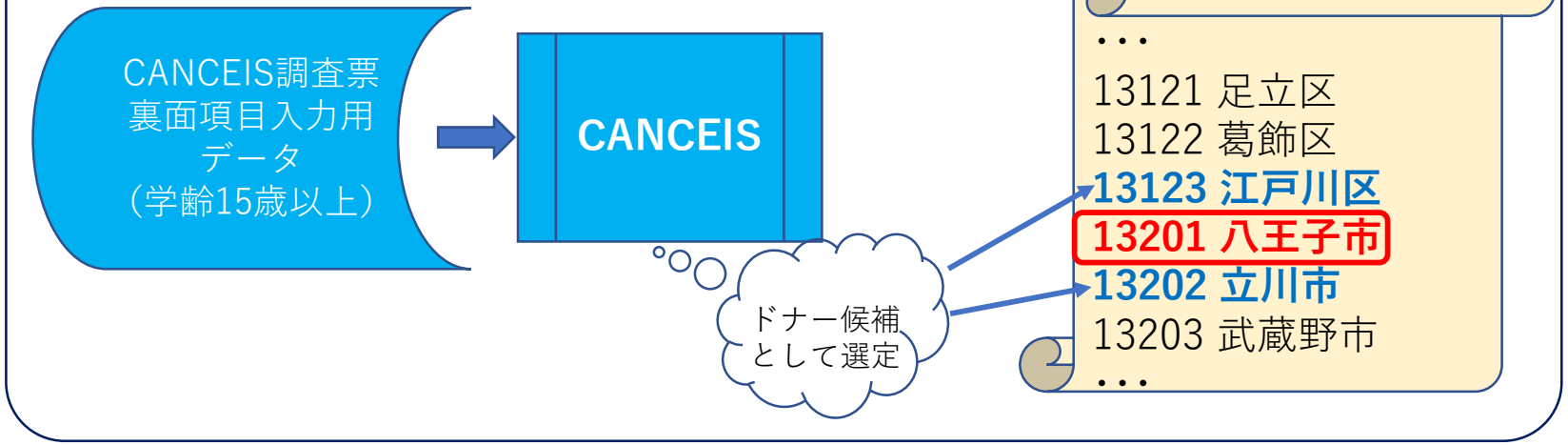
3. CANCEIS実装に必要な処理： ⑪都道府県別モジュールの作成

＜入力用データのうち、学齢15歳以上のデータ＞

（３）緯度経度情報を利用した距離関数の設定

・**従業地・通学地不詳の補完で市区町村を跨いでドナーを探索する場合に、（昼間移動の観点で）関係性の低い市区町村からドナーが選ばれるのを防ぐ**ため、不詳世帯の常住市区町村の**15歳以上就業者・通学者総数に占める従業地・通学地別割合が一定割合※を超える市区町村のみ許容されるようにウエイト及び距離関数を設定。**これにより、例えば「八王子市から江戸川区周辺への通勤通学者」が不自然に増加することを抑制 ※ **本試算では便宜５％を基準とした。**

デフォルトで八王子市のドナーを選定すると、立川市と江戸川区から同等の確率で選定される可能性



市区町村番号が近いと、データセット内のレコード配置が近いため、ドナー探索の順序も前の方になり、ドナーに選ばれやすい。

逆に市区町村番号が遠いとレコード配置も遠くなり、ドナー探索の順序も後ろの方になるため、距離の小さい正データがあったとしても、そこにドナー探索が行き着く前にNMCIAのプール数上限に達して探索が打ち切られ、当該正データが採用されなくなる可能性が高い。

4. CANCEIS実装に必要な処理：④調査票表面項目に係るデータのグループ化

○ CANCEISの補完精度を向上させるため、元データの適切なグループ化が必要となる。
○ 調査票表面項目へのCANCEIS補完の前処理として、世帯人員数ごとにグループ化した上、世帯データを世帯主との続き柄でソートする。

【世帯人員が４人のグループについて、世帯データを世帯主との続き柄でソートした場合の例】

- **世帯番号１のデータに不詳があった場合、世帯番号３、４、８の世帯が世帯番号１のドナーとなる可能性が高い**

	世帯員１	2	3	4
世帯番号１	**世帯主**	**配偶者**	**子**	**子**
2	**世帯主**	**配偶者**	**子**	**子の配偶者**
3	**世帯主**	**配偶者**	**子**	**子**
4	**世帯主**	**配偶者**	**子**	**子**
5	**世帯主**	**子**	**世帯主の父母**	**世帯主の父母**
6	**世帯主**	**配偶者**	**子**	**子の配偶者**
7	**世帯主**	**子**	**世帯主の父母**	**世帯主の父母**
8	**世帯主**	**配偶者**	**子**	**子**
9	**世帯主**	**子**	**世帯主の父母**	**世帯主の父母**

4. CANCEIS実装に必要な処理：⑩調査票裏面項目に係るデータのグループ化

○ **就業関係の項目が多い調査票裏面項目の補完は、就業可能年齢が概ね15歳以上であることを考慮**し、世帯内の15歳以上の者が的確に補完されるよう、**各世帯を15歳以上の世帯人員数ごとにグループ化した上、同一グループ内の15歳以上世帯員の中からドナーを選定**（※１）

○ 15歳未満の者については、就学年齢（６～14歳）と就学年齢未満（６歳未満）でグループ化した上、６～14歳については同一グループ内の者をドナーとして補完（※２）。６歳未満は演繹的に補完（※３）

※１　学齢15歳以上：特徴量として、学齢15歳**未満**人員数を各ユニットに追加
※２　学齢６～14歳：特徴量として、学齢15歳**以上**のCANCEIS補完結果から、「世帯主の職業」及び「世帯主の従業上の地位」を追加
※３　学齢６歳未満：労働力状態を「その他（幼児や高齢など）」に補完し、産業・職業及び従業地・通学地は一律にブランクに補完

【15歳以上世帯人員が２人のグループの例】

※本処理は世帯データを年齢、男女の順でソートした上で行う

- **世帯番号１の世帯員１・２のデータに不詳があった場合、距離関数が近い世帯番号３の世帯がドナーとなる可能性が高い**

	世帯員１（男）	2（女）	3	4	5	世帯人員数	15歳以上世帯人員数	15歳未満世帯人員数
世帯番号１	40歳（男）	40歳（女）				2	2	0
2	50歳（男）	53歳（女）	14歳（男）	10歳（女）	3歳（男）	5	2	3
3	39歳（男）	40歳（女）				2	2	0
4	55歳（男）	60歳（女）	12歳（男）	7歳（女）	2歳（男）	5	2	3

15歳未満の世帯人員数を特徴量として保持

世帯主の職業と従業上の地位を特徴量として保持

（参考）CANCEISで利用する特徴量一覧

特徴量	調査票表面 一般世帯	調査票表面 施設等世帯	調査票裏面 一般世帯(15歳以上)	調査票裏面 一般世帯(6歳～14歳)	調査票裏面 施設等世帯
市区町村	○	-	○	○	○
町字	○	-	-	-	-
調査区番号（後置番号）	-	○	-	-	-
単位区中心点座標 （緯度・経度）	-	-	○	○	-
住宅の建て方	○	-	○	○	○
世帯の種類	-	○	-	-	○
男女の別	○	○	○	○	○
続柄	○	-	○	○	-
配偶関係	○	○	○	○	○
国籍	○	○	○	○	○
居住期間	○	-	-	-	-
前住地(５年前常住地)	○	-	-	-	-
労働力状態	-	-	○	○	○
従業地・通学地	-	-	○	○	○
産業大分類	-	-	○	○	○
職業大分類	-	-	○	○	○
年齢	○	○	-	-	-
学齢	-	-	○	○	○
世帯ID	-	○	-	-	○
学齢15歳未満人員数	-	-	○	-	-
世帯主の職業	-	-	-	○	-
世帯主の従業上の地位	-	-	-	○	-

※赤枠は調査票裏面項目のCANCEIS補完で新たに取り入れた特徴量