

諮問第 44 号の答申
国勢調査に係る匿名データの作成について（案）

本委員会は、総務省が作成を予定している国勢調査（以下「本調査」という。）に係る匿名データの作成方法の計画について審議した結果、下記の結論を得たので答申する。

記

1 計画の適否

本計画については、これにより作成される匿名データにおいて、本調査の調査客体の匿名性及び学術研究等における有用性がおおむね確保されるものと認められることから、適当である。

ただし、以下の「2 理由等」で指摘した事項については、計画の修正が必要である。

2 理由等

本調査は、我が国の人口の状況を明らかにするため、本邦内に常住している者を対象にほぼ5年ごとに実施している統計調査であり、10年ごとの大規模調査とその中間年の簡易調査とに大別される。今回、総務省は、大規模調査年と簡易調査年とで調査項目が異なること及び時系列での比較を考慮して、大規模調査である平成12年調査と簡易調査である平成17年調査の2か年分の匿名データを作成することを計画している。

また、その作成に当たって、これまで用いられてきた、情報の削除、一定の値を上限値としそれを上回る場合に上限値以上でまとめる措置（以下「トップコーディング」という。）や、分類の程度を粗いものにする措置（以下「リコーディング」という。）等の匿名化措置以外に、攪乱的な秘匿手法の一つである調査票情報のレコードの一部を別のレコードのものと入れ替える措置（以下「スワッピング」という。）を新たに導入する計画である。

これらの匿名データの具体的な作成方法に関する計画の適否等については、以下のとおりである。

（1）情報の削除

ア 地域区分

地域区分については、「都道府県」及び「人口50万以上の市区」とし、人口50万未満の市区町村コードは削除する計画である。

このうち、「都道府県」及び「人口50万以上の市区」を提供することについては、本調査が悉皆調査であり、地域レベルでの分析が可能であることが特徴の一つであるため、地域分析に対するニーズは高いと考えられること、本調査の結果では従来から人口50万以上の市区について詳細な統計表が提供されていることから、適当である。

また、人口50万未満の市区町村コードを削除することについては、人口50万未満の市区町村が明らかになると、外観識別性の高い属性情報と組み合わせた場合に調査客体が特定される可能性が生じることから、適当である。

イ レコードのサンプリング

本調査の匿名データの作成に当たっては、全ての調査票情報のレコードから、全世帯を母集団として1%を抽出することとし、世帯の種類（「一般世帯」及び「施設等の世帯」）ごとに、

市区町村及び世帯人員等で並べ替えた上で、「一般世帯」については世帯単位で、また「施設等の世帯」については個人単位で抽出処理を行い、これらを統合したものをを用いる計画である。また、抽出された世帯は、乱数により並べ替える計画である。

これについては、次の理由から適当である。

サンプリングは、匿名データの中に特定の調査客体が含まれるか否かの判断を困難とする措置であること

サンプリングに当たっては、各市区町村における世帯数や各世帯の世帯人員等による分布を反映した抽出処理が行われるとともに、その抽出率は、都道府県のみならず、人口50万以上の市区など一定の人口規模以上の市区についての集計・分析が可能なデータ量が確保できるものとなっていること

当該方法により作成された匿名データによる統計と既に公表されている本調査の全レコードから作成された統計（以下「公表統計」という。）との間で、男女別、年齢5歳階級別、世帯主との続き柄別、配偶関係別等でみた分布に大きな乖離は無く、当該匿名データの有用性が確保されていること

ただし、世帯の種類によって抽出単位が異なることから、分析等において支障が生じないように利用者に対して利用上の注意を示す必要がある。また、主要な項目に関して、匿名データによる統計と公表統計の結果を比較できるような情報を提供するなど、利用者の利便性の向上を図る必要がある。

ウ 直接的な識別情報の削除

本調査のレコードに含まれる情報のうち、調査区番号、世帯番号・調査区内連番など、調査客体が直接識別できる情報を削除する計画である。

これらについては、調査客体の特定や探索を防止するために効果的な措置であることから、適当である。

エ 出現頻度が低い又は特徴的な値があるレコードを含む世帯の削除

本調査のレコードのうち、表1に示す出現頻度が低い又は特徴的な値があるレコードを含む世帯を削除する計画である。

表1

対象となる世帯	具体的な措置の内容
世帯人員が多い世帯	地域区分ごとの出現頻度により、世帯人員が7人以上～9人以上いる世帯を削除
父子世帯	未婚、死別又は離別の男親と、その未婚の20歳未満の子供のみからなる世帯を削除
年齢差の大きい夫婦のいる世帯	年齢差が25歳以上の夫婦のいる世帯を削除
年齢差の大きい又は小さい親子のいる世帯	年齢差が45歳以上の親と子、年齢差が14歳以下の親と長子、又は19歳以下の親と末子のいる世帯を削除
世帯主又は配偶者のいずれか一方若しくは双方が外国人で子供の数が多い世帯	地域区分ごとの出現頻度により、子供の数が3人以上～7人以上の外国人世帯を削除

これらについては、次の理由から適当である。【P】

世帯員数、性別、年齢等は、世帯の外部から比較的容易に把握可能な属性であり、トッ

ブコーディングやリコーディング等の匿名化措置を行ったとしても、地域情報等の他の属性情報と組み合わせた場合に出現頻度が低くなる世帯構成の場合は、調査客体が特定される可能性が生じること

地域によって出現頻度に大きな違いが見られる「世帯人員が多い世帯」や「世帯主又は配偶者のいずれか一方若しくは双方が外国人で子供の数が多い世帯」については、有用性の観点から地域区分ごとに削除基準を変える措置が講じられていること

複数の変数の組合せをみる必要がある「年齢差の大きい夫婦のいる世帯」、「年齢差の大きい又は小さい親子のいる世帯」、「世帯主又は配偶者のいずれか一方若しくは双方が外国人で子供の数が多い世帯」については、実数値による分布に基づいて削除数を可能な限り減らす措置が講じられていること【P】

オ 公表統計により母集団一意又は二意であることが判明しているレコードを含む世帯の削除

公表統計により、全国において、個体が識別されやすい属性情報を組み合わせた場合に、母集団の中で個体が1つしかないような母集団一意の状態であるか、又は同じ組合せを持つ個体が2つしかないような母集団二意の状態であることが判明しているレコードが含まれる世帯については、これを削除する計画である。加えて、公表統計のうち、特に外観識別性が高いと考えられる項目が含まれている統計表については、地域（「都道府県」及び「人口 50 万人以上の市区」）において、母集団一意又は二意であることが判明しているレコードが含まれる世帯も削除する計画である。

これらについては、国勢調査が悉皆調査であることや、詳細な集計表が外部参照情報として利用可能であることなどの特性を踏まえて、全国において母集団一意又は二意である場合や、地域において、特に外観識別性が高いと考えられる項目が含まれている統計表で母集団一意又は二意である場合は、調査客体が特定される可能性が生じることから、適当である。

(2) 識別情報の分類区分の再編等

ア 世帯員に関する項目の再編等

(ア) 年齢

世帯員の年齢については、85歳以上をトップコーディングすることとし、トップコーディングを行う高齢者を除く0～84歳をリコーディングして5歳階級別とする計画である。

これらについては、各歳別の年齢や出現頻度の低い高齢者の年齢が明らかになると、今回提供される詳細な地域情報や他の外観識別性の高い属性情報との組み合わせにより調査客体が特定される可能性が生じることから、やむを得ない措置である。

(イ) 世帯主との続き柄、国籍、5年前の住居の所在地等

世帯主との続き柄、国籍、5年前の住居の所在地（平成12年国勢調査）、従業上の地位、職業（大分類）及び常住地による従業地・通学地については、それぞれ表2に示す分類区分の再編等を行う計画である。

表2

項目	措置の内容
世帯主との続き柄	「他の親族」、「住み込みの雇人」及び「その他」を統合
国籍	「日本人」及び「外国人」の2区分とし、外国籍の内訳は提供しない

5年前の住居の所在地 (平成12年国勢調査)	「他県から」及び「国外から」を統合
従業上の地位	「雇人のある業主」、「雇人のない業主」及び「家庭内職者」を統合
職業(大分類)	「保安職業従事者」、「農林漁業作業者」及び「運輸・通信従事者」を統合
常住地による従業地・通学地	「県内他市区町村で従業・通学」及び「他県で従業・通学」を統合

これらについては、出現頻度の低い分類区分を統合するか又は提供しないことによって、地域情報や他の外観識別性の高い属性情報との組み合わせによる調査客体の特定を防ぐことから、適当である。

なお、5年前の住居の所在地における「他県から」の場合の調査項目である5年前の常住都道府県等については、「国外から」との統合のため提供されないこととなるが、これは統合に際して「国外から」の匿名性を確保する必要があることから、適当である。

(ウ) 労働力状態

労働力状態については、「家事などのほか仕事」及び「通学のかたわら仕事」を統合する計画である。

これについては、・・・【P】

(エ) 就業時間

就業時間については、実数及び公表統計と同じ階級(14区分及び10区分)で提供することとし、実数については90時間以上をトップコーディングする計画である。

このうち、出現頻度が低い長時間労働をトップコーディングすることについては、これにより、他の属性情報との組み合わせによる調査客体の特定を防ぐことから、適当である。ただし、トップコーディングを行った当該変数については、その地域区分別の平均値等を匿名データの提供に併せて提供することにより、利用者の利便性向上を図る必要がある。

また、公表統計と同じ階級で提供することについては、公表統計との整合性を図ることにより匿名データの利便性が向上することから、適当である。

(オ) 産業(大分類)

産業については、大分類で提供することとし、以下の分類区分の統合を行う計画である。

「農業」、「林業」及び「漁業」を統合

「鉱業」及び「建設業」を統合

「製造業」及び「電気・ガス・熱供給・水道業」を統合

「複合サービス業」及び「サービス業(他に分類されないもの)」を統合

これらについては、「農業」、「林業」、「漁業」、「鉱業」、「電気・ガス・熱供給・水道業」及び「複合サービス業」は、地域別にみると出現頻度が低くなる場合があり、他の属性情報との組み合わせによって調査客体を特定される可能性が生じること、統合に当たっては、できるだけ類似性のある項目を統合する方がよいと考えられることから、基本的に適当である。

ただし、「農業」、「林業」、「漁業」については、これらを統合したとしても秘匿が不

十分な場合があることから、特定化の危険性が高いレコードを含む世帯を削除することにより、匿名性を確保する必要がある。

(カ) 利用交通手段（平成 12 年国勢調査）

利用交通手段については、「利用交通手段が 1 種類」に関して、「勤め先・学校のバス」、「ハイヤー・タクシー」、「オートバイ」及び「その他」を統合する計画である。

これについては、利用交通手段が 1 種類の場合は外観識別性が高いと考えられる一方、複数の場合には 2 種類目以降の交通手段を識別することが難しいこと、「勤め先・学校のバス」、「ハイヤー・タクシー」及び「オートバイ」は、地域情報や他の属性情報と組み合わせた場合に出現頻度が低くなり、調査客体を特定される可能性が高くなることから、・・・【P】

【案 1】 適当である。

【案 2】 基本的に統合による匿名性の確保が必要である。

ただし、統合に当たっては、できるだけ交通手段として類似性のあるものを統合する方が望ましく、「オートバイ」と「自転車」に関しては、産業や職業別の就業者割合等をみても、分布状況が似ていることから、「オートバイ」及び「自転車」を統合することとし、残りの「勤め先・学校のバス」、「ハイヤー・タクシー」及び「その他」を統合するように変更する必要がある。

イ 世帯に関する項目の再編等

(ア) 世帯の種類、世帯人員、世帯の家族類型等

世帯の種類、世帯人員、世帯の家族類型、住居の種類・住宅の所有の関係、住宅の建て方については、それぞれ表 3 に示す分類区分の再編等を行う計画である。

表 3

項目	措置の内容
世帯の種類	「一般世帯」及び「施設等の世帯」の 2 区分を提供することとし、「施設等の世帯」の内訳は提供しない
世帯人員	「施設等の世帯」については提供しない
世帯の家族類型	公表統計と同じ 6 区分により提供
住居の種類・住宅の所有の関係	「住宅に住む一般世帯」の以下の分類区分を統合することとし、「住宅以外に住む一般世帯」の内訳（「会社等の独身寮」及び「その他」）は提供しない 「公営の借家」及び「都市機構・公社の借家」を統合 「給与住宅」及び「間借り」を統合
住宅の建て方	「長屋建」及び「その他」を統合

これらについては、出現頻度の低い分類区分を統合するか又は提供しないことによって、地域情報や他の外観識別性の高い属性情報との組み合わせによる調査客体の特定を防ぐこと、世帯の家族類型は製表の過程で作成されたものであるが、国勢調査の特徴を表す分類の一つであり利用ニーズは高いと考えられ、公表統計との整合性を図ることにより匿名データの利便性が向上することから、適当である。

(イ) 家計の収入の種類（平成 12 年国勢調査）

家計の収入の種類については、これを提供しない計画であるが、家計の収入の種類は、調

査に際し世帯の忌避感が大きい調査項目であり、今後の国勢調査の実施に対する影響を懸念する指摘はあるものの、利用者にとっては有益な情報であり、出現頻度の低い分類区分を統合することにより匿名性を確保することが可能なことから、以下のような匿名化措置を講じた上で提供するように計画を変更する必要がある。

- 「賃金・給料が主な世帯」のうち、「農業収入もある世帯」及び「その他」を統合
- 「農業収入が主な世帯」及び「農業収入以外の事業収入が主な世帯」を統合し、その内訳は提供しない
- 「内職収入が主な世帯」及び「その他の収入が主な世帯」を統合し、その内訳は提供しない

(ウ) 住宅の床面積、建物全体の階数及び世帯が住んでいる階

住宅の床面積、建物全体の階数及び世帯が住んでいる階については、それぞれ表4に示す分類区分の再編を行う計画である。

表4

項目	措置の内容
住宅の床面積	実数ではなく、基本的に公表統計に合わせた階級で提供することとし、この際、「200～249㎡」及び「250㎡以上」を統合
建物全体の階数	実数ではなく、基本的に公表統計に合わせた階級で提供することとし、地域区分ごとの出現頻度により、必要に応じて、「6～10階建」、「11～14階建」、「15階建以上」の区分のうち、「6～10階建」以上、あるいは「11～14階建」以上の区分を統合
世帯が住んでいる階	実数ではなく、基本的に公表統計に合わせた階級で提供することとし、地域区分ごとの出現頻度により、必要に応じて、「3～5階」、「6～10階」、「11～14階」、「15階以上」の区分のうち、「3～5階」以上、「6～10階」以上、あるいは「11～14階」以上の区分を統合

これらについては、住宅の床面積及び階数は外観識別性が高く、実数が明らかになると、今回提供される詳細な地域情報や他の外観識別性の高い属性情報との組み合わせにより調査客体が特定される可能性が生じること、出現頻度の低い高層階は地域による差が大きく、有用性の観点から地域区分ごとに統合する階数区分を変える措置が講じられていることから、適当である。

(3) その他の匿名化措置

ア スワッピング

本調査の匿名データの作成に当たっては、新たな匿名化措置としてスワッピングを行うこととし、一部世帯を他の地域の類似世帯と入れ替える計画である。

このスワッピングの導入及び方法については、次の理由から適当である。

今回、地域区分として「都道府県」及び「人口50万以上の市区」を提供することとしており、地域情報が強力な識別情報となり得ること、また、国勢調査は悉皆調査であることから、既に公表されている詳細な集計表自体が有力な外部参照情報として利用可能であり、これらから想定される範囲で情報の削除やリコーディング等の措置を講じることに加え、更に攪乱的な秘匿手法を用いることにより、属性情報を様々に組み合わせた場合において

も匿名性が確保されるように措置することはやむを得ないと考えられること

本計画の方法によるスワッピングの前後で、影響を受ける項目の分布にあまり差異はなく、有用性を損なう程のものではないと考えられること

ただし、利用者に対して・・・の措置を講じる必要がある。【P】

イ 匿名データの提供時期

本計画では、匿名データの作成対象調査を調査実施後5年以上経過したものとしている。

これについては、有用性の観点からは、最新のデータに基づく匿名データの作成に対する強いニーズがあり、匿名データの提供時期を早められないかという指摘はあるものの、匿名性の観点からは、調査実施後5年以上経過することで属性の変化により個人を特定し難くする効果があること、及び本調査が本邦内に常住している者を対象とした悉皆調査であり、実査への影響も考慮した場合に直近の調査の匿名データは提供し難いことから、本調査に関する匿名データの提供時期の短縮化は困難であり、やむを得ない措置である。

3 今後の課題

(1) トップコーディングを行う高齢者の年齢の検討

本計画では、世帯員の年齢については、85歳以上でトップコーディングを行うこととしている。しかし、トップコーディングの上限値については、近年の急速な高齢化の進展や高齢者に関する分析の重要性等を踏まえ、今後、匿名データの作成対象年次を拡大する際には、当該年次の人口構成に応じた検討が必要である。

(2) 複数の匿名データの作成の可能性に関する検討

今回の匿名データの作成に当たっては、悉皆調査の調査票情報を世帯単位で、また、地域区分を「都道府県」及び「人口50万以上の市区」として提供することを前提に考えられていることから、情報の削除や識別情報の分類区分の再編等において、厳格な匿名化措置を講じていることはやむを得ない。

しかし、匿名データの利用者のニーズについては様々なものが考えられ、例えば、世帯員の年齢を各歳別とすること、あるいは世帯の削除や産業等の分類区分の統合を緩和することなどに対するニーズが指摘されている。これについては、地域区分を全国のみとすることや提供データを個人単位とすることなどにより、匿名化措置を緩和できる可能性も考えられることから、今後、利用者のニーズを踏まえて、匿名化措置の内容や組合せ、抽出単位の異なる複数の匿名データの作成の可能性について検討する必要がある。