資　料

諮問第 13 号の答申
全国消費実態調査、社会生活基本調査、就業構造基本調査及び
住宅・土地統計調査に係る匿名データの作成について（案）

本委員会は、総務省が平成 21 年に作成を予定している全国消費実態調査（指定統計第 97 号を作成するための調査）、社会生活基本調査（指定統計第 114 号を作成するための調査）、就業構造基本調査（指定統計第 87 号を作成するための調査）及び住宅・土地統計調査（指定統計第 14 号を作成するための調査）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　計画の適否とその理由等

(1)　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、全国消費実態調査等４調査（以下「作成対象４調査」という。）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、一部修正を行うことを前提に適当なものと考える。

この判断の理由及び修正点は以下のとおりである。

(2)　理由及び修正点

ア　情報の削除

(ｱ)　レコードのリサンプリング

匿名データの作成に当たっては、作成対象４調査の全ての標本のレコード（調査客体）から、世帯単位により、全国消費実態調査、社会生活基本調査及び就業構造基本調査の３調査については 80%を、また、住宅・土地統計調査については 10%を、無作為または各レコードに付された乗率の大きさに基づく確率比例で再抽出（以下「リサンプリング」という。）したもの（以下「サブサンプル」という。）を用いる計画である。

これについては、次の理由等から適当である。

①　リサプリングは、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別を困難とする措置であること

②　特に、今回のリサンプリングにおいては、無作為抽出を基本としつつ、各レコードが持つ集計用乗率に抽出地域との一定の対応関係がある場合、当該乗率から抽出地域が特定されてしまうことを防ぐための措置を採っていること

③　世帯単位による抽出は、匿名データの利用者のニーズが高い世帯収支等世帯に着目した分析が可能となるため、個人単位による抽出よりも当該データの有用性が高まること

④　サブサンプルの抽出率は、各調査の母集団の大きさやそれに含まれる情報の内容等を踏まえ設定しているものであり、当該抽出率によりリサンプリングされたサブサンプルから作成された匿名データによる統計と全レコードから作成された公表統計（以下「公表統計」という。）との間で、代表的な項

目の平均値や標準偏差に大きな乖離はなく、当該データの有用性が確保されていること

(ｲ)　識別情報の削除等

作成対象４調査のサブサンプル中のレコードに含まれる情報のうち調査区番号等の識別情報は、これを削除するとともに、当該レコードを乱数により並び替える計画である。

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等から、適当である。

(ｳ)　裾切りによるレコード削除

a　世帯人員８人以上等の世帯

作成対象４調査のサブサンプル中のレコードのうち世帯人員８人以上の世帯及び三つ子以上のいる世帯に係るものは、匿名データから削除する計画である。

これについては、世帯員の人数等の情報は世帯の外部から比較的容易に把握可能な属性であり、それが極端に大きい場合は調査客体が特定される可能性が生じること等から、適当である。

b　年収等が高額な世帯

全国消費実態調査のサブサンプル中のレコードのうち年収、貯蓄及び借入金が一定金額以上の高額な世帯に係るものは、匿名データから削除する計画である。

これについては、①匿名データの信頼性の確保の観点からサブサンプルの削除は必要最小限に留めるべきであること、②年収等の情報は世帯外からの把握可能性が低いこと、③世帯収支等の経済分析に対する研究者のニーズが非常に高いこと等から、年収等が高額な世帯のレコードを全面的に削除することは適当でない。したがって、当該世帯のレコードについても、提供する情報を年収等の総額のみに限定し、かつ全体に占める構成比が 0.5%未満の変数について、一定の水準を上限値とし、こ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。）等の匿名化措置を講じた上で、匿名データに残すことが必要である。

イ　識別情報の階級区分の統合

(ｱ)　トップコーディング及びボトムコーディング

a　高齢者の年齢

４調査の匿名データの各レコード上の個人の年齢について、一定年齢を上限値とし、それを上回る高齢者の場合、トップコーディングを行うこととし、当該上限値は、全国消費実態調査及び住宅・土地統計調査は 75 歳以上、社会生活基本調査は 85 歳以上、就業構造基本調査は 80 歳以上とする計画である。

これについては、トップコーディングは、それにより極めて高齢であるという特殊な属性をまとめられ、調査客体の判別を困難とすることから適当であるが、上限値については、近年の高齢化の進展状況、年齢を用いた就業行動や家族関係の分析の重要性等を踏まえ、４調査とも 85 歳以上にすることにより、匿名データの有用性の向上を図ることが必要である。

b　住宅の規模等

全国消費実態調査及び住宅・土地統計調査の匿名データの各レコード上の住宅の規模等に係る数値（延べ床面積、敷地面積、家賃・間代等）について、トップコーディング及び一定の水準を下限値としこれを下回る場合に下限値以下でまとめる措置（以下「ボトムコーディング」という。)を講じる計画である。

これについては、トップコーディング及びボトムコーディングにより、住宅の規模等が極端に大きい（または小さい）という特殊な属性をまとめられ、住宅の特徴を通じそこに居住する調査客体の判別を困難とすること等から、適当である。

(ｲ)　リコーディング（分類区分の再付与）

ａ　地理的情報（地域区分）

匿名データの各レコードに付与する地理的情報については、その分類の程度を粗いものとする措置（以下「リコーディング」という。）を講じることとし、住宅・土地統計調査では 47 都道府県別に、また、全国消費実態調査、社会生活基本調査及び就業構造基本調査では全国６ブロック別とする計画である。

このうち、住宅・土地統計調査の地域区分については、当該調査の標本規模が非常に大きく、かつサブサンプルの抽出率が 10%と低いこと等により、地域区分と他の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極めて低いこと等から、適当である。

一方、全国消費実態調査等３調査の地域区分については、①３調査のサブサンプルの抽出率が 80%と高いこと、②３調査は調査客体である個人の職業、配偶者との年齢差等多くの属性情報を有しており、これと詳細な地理的情報を組み合わせると調査客体の特定の可能性が生じること、③地域区分を６ブロックとしても、公表統計と照合することにより都道府県の別が明らかになるケースが一部あること等から、地域区分を「３大都市圏」及び「その他の地域」の２区分とすることにより、調査客体の匿名性の確保を十分に図るよう、万全を期すことが必要である。

b　個人の年齢

４調査の匿名データの各レコード上の個人（トップコーディングを行う高齢者を除く。）の年齢については、15 歳以上の者は、リコーディングとして、５歳階級別とする一方、15 歳未満の者は、リコーディングを行わず各歳別とする計画である。

このうち、15 歳以上の者については、各歳別のデータ提供に比べ、匿名データの有用性が低下するものの、各歳別の年齢が明らかになると、個人の職業等他の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。

これに対して、15 歳未満の者については、調査客体の特定に利用可能な属性情報が限定されているため、各歳別の年齢を明らかにしても判別を困難とする観点からは適当である。

2　今後の課題

本計画については、①政府における匿名データの作成は今回の総務省によるものが初めてであり、調査客体の匿名性の確保に慎重を期する必要があること、②本年４月の統計法の全面施行に合わせて、匿名データの提供を速やかに開始する必要があること、③これまで匿名データの利用ニーズが必ずしも十分に把握されていないこと等から、調査客体の匿名性を確保するために厳格な匿名化措置を講じていることはやむを得ない。

しかしながら、匿名データの利用者ニーズ等については様々なものが考えられることから、以下の課題等について検討を進め、当該データのより一層の充実に努める必要がある。

(1)　本計画では、匿名性を確保するため、個人の年齢等調査客体の特定につながる可能性がある重要かつ基本的な属性情報については厳格な匿名化措置を講じることとしている。

しかしながら、調査客体の匿名性は、一つの匿名化措置のみで確保される訳ではなく、複数の匿名化措置により全体として確保されるものであるため、匿名化措置の内容や組合せを変えることにより、同一の調査について複数の匿名データを作成することが可能であると考えられる。【以下の点線部分はＰ】例えば、就業構造基本調査について、世帯主の年齢を各歳別とする一方、職業、産業等の分類区分を大括り化した匿名データの作成についてのニーズも指摘されている。

こうした観点から、今後、複数の匿名データのマッチングによる調査客体の特定の危険性に関する研究等の結果や匿名データの利用者のニーズを踏まえて、匿名化措置を課す情報及びその程度が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。

(2)　本計画では、匿名データの作成対象調査を平成以降に実施したものであり、かつ調査実施後５年以上を経過したものとしている。

しかし、経済・社会事象に関する研究には、長期の時系列分析が不可欠であり、また、近年、経済・社会の状況がめまぐるしく変化していることから、直近の統計に基づく当該研究の重要性も増している。

こうした観点から、今後、作成対象調査を、平成より前に実施したものに拡張することについて検討するとともに調査実施後５年以上経過したものを提供するという基準を緩和することについて検討する必要がある。

(3)　匿名データの分析手法としては、集計値の分析のほかに回帰分析がある。しかし、本計画により作成された匿名データの各レコード上の変数のうち、トップコーディング、ボトムコーディング及びリコーディング（年齢等の階級化等）が行われている変数については、集計値の分析には大きな問題がないものの、回帰分析へは必ずしも十分利用することができない。

こうした観点から、今後、トップコーディング等が行われた変数についても回帰分析に十分利用できるよう、当該変数の平均値等をメタデータとして利用者に提供する等の措置に関して、運用後のニーズ等の状況を踏まえ、検討する必要がある。【以下の点線部分はＰ】また、年齢についても、回帰分析に利用目的を限定した上で各歳別データの利用を可能とするなど、匿名データの利用環境の整備を引き続き検討する必要がある。