参　考

第８回匿名データ部会　議事概要

１　日　　時　平成23年７月４日（月）　15:31～17:18

２　場　　所　中央合同庁舎４号館２階　共用第３特別会議室

３　出　席　者

椿広計部会長、津谷典子委員、廣松毅委員、伊藤伸介専門委員、稲葉由之専門委員、黒田祥子専門委員、安田聖専門委員、総務省（政策統括官（統計基準担当））、文部科学省、厚生労働省、農林水産省、経済産業省、国土交通省、日本銀行、東京都、千葉県

【諮問者（総務省統計局統計調査部）】

高田調査企画課調査官、北林調査企画課長補佐、横内調査企画課二次利用推進係長、佐藤国勢統計課労働力人口統計室主任研究官

【事務局（内閣府統計委員会担当室）】

若林参事官、谷道参事官補佐

４　議事次第　（１）労働力調査に係る匿名データの作成について

（２）その他

５　議事概要

（１）労働力調査に係る匿名データの作成について

冒頭、事務局から前回の部会審議等を踏まえて修正された資料「『労働力調査に係る匿名データの作成について』の論点」の説明が行われ、諮問者より論点への回答について説明を受けた後、論点の項目ごとに審議が行われた。

各委員等の主な意見等は次のとおり。

ア　リサンプリングの方法等

・　地域11ブロック及び組符号８区分による層化を行い、リサンプリング率80%で抽出することに関して、総務省より、完全失業率等の代表的な比率は公表結果と大きな乖離はないという説明、また、乗率の再付与・再計算について、数値を1.25倍することにより元の値とほぼ一致することから乗率の再付与は行わないという説明があったところであり、妥当と判断する。

イ　地域区分、事業の種類（産業）及び本人の仕事の種類（職業）

・　前回から議論があったところではあるが、地域区分については全国１区分とし、産業及び職業区分については報告書の表章区分に併せてリコーディングすることについては、仮に地域区分を付与した

場合、産業分類や職業分類を大くくり化したとしても個人が特定されるリスクがあるのであれば、やむを得ない。

ウ　世帯人員

・　世帯人員が８人以上の世帯については、0.5%基準に基づいてレコードを削除することは、妥当と判断する。

エ　同一年齢の子供の数

・　同一年齢階級に３人以上の子供のいる世帯のレコードを削除することについて、年齢階級を０～３歳、４～６歳とした場合と０～６歳とした場合について整理してもらったところ、後者の場合は、子供が３人の世帯は削除されないで済むようになり、こうした世帯の分析が可能となるが、０～３歳、４～６歳という細かい情報は落ちてしまい、トレードオフの関係となる。

・　同一年齢階級に３人以上の子供のいる世帯の割合は、すべての世帯を対象とすれば、高齢者、単独世帯などがあるため非常に小さくなるが、20～49 歳の再生産年齢の有配偶女性から見たときにその割合がどの程度変わるかを懸念した。今回提示された資料を見ると、０～３歳、４～６歳とした場合と、０～６歳とで比較した場合で、有配偶者女性の世帯から見てもほとんど割合が減らない結果となっており、安心材料ではある。しかし、原案では、同一年齢階級に３人以上の子供のいる世帯のすべての情報が削除されてしまうということを考慮すると、判断が非常に難しい。

・　先ほどのリサンプリングの議論で、新たに乗率の再付与を行わないという結論が出されたが、その場合、本来であればレコードの削除を極力行わない方が望ましいと思われる。一方、乗率で復元した労働力人口等が、削除対象となるレコードを除いた母集団と元の母集団とでは、それほど変わっていないことが考えられる。そうであれば、０～３歳、４～６歳という区分において３人以上の子供のいる世帯のレコードを削除した場合においても、その影響は大きくはないのではないか。

≪部会長のまとめ≫

・　分析側のニーズがどこにあるかという問題で、例えば、年齢の幅を０～３歳、４～６歳となっているものを０～６歳にするということは、ユーザーにとっては情報量が落ちると考えられる。このため、今回は原案の方法を妥当とするものの、今後、総務省でも、どのようなユーザーの研究ニーズがあるかということについて、注意深く調べていただきたい。

オ　15 歳以上の世帯員の年齢

・　労働経済学の観点からすると、年齢を５歳刻みにグルーピングするのは少々粗い。統計局が労働力調査ミニトピックスにおいて、過去６年間で 61 歳の男性の就業率が 10 ポイント以上上がったという結果を出されていた。これは、2006 年の高齢者雇用安定法の改正が大きく寄与しているのではないかと考えられるし、今後も 65 歳の定年制の義務化が検討されていることを考えると、高齢者の１歳刻みの就業状態がどうなるかということは、国民にとっても非常に関心が高いと思われる。地域も出さないのであれば、特定化はほとんどできないのではないか。他の匿名性とのトレードオフの兼ね合いがあるとは思うが、１歳刻みというのは非常にニーズが高いということを指摘しておく。

・　匿名データとしては難しいならば、オーダーメード集計ではどの程度のことが利用できる状況なのか。

≪総務省からの回答≫

・　産業・職業と年齢のクロスを考えると、年齢を各歳別にした場合は個人が特定化されるリスクがあると考えている。匿名性を破られるということを厳格に考えているため、各歳別にというのは、非常に難しいと考えている。ただ、非常にニーズがあるということについては認識している。このような分析には、統計法第 33 条で匿名化されていないものを使っていただく等の対応も考えられる。産業も職業も付けずに、就業状態だけを知りたいというニーズも考えられることから、年齢についてもっと詳しい分析ができるよう複数種類の匿名データを準備することも、将来的に考えさせていただきたい。また、労働力調査については、昨年度にオーダーメード集計の受付を開始したが、年齢については５歳階級で提供している。

≪部会長のまとめ≫

・　各歳別に対する実際の研究上のニーズが大きいということを認識した。今回は匿名性を重視し、５歳階級別で提供するということを認めることとしたいが、将来的には、年齢を細かく分析できるような匿名データの検討もお願いしたい。

カ　月末１週間に仕事をした時間

・　労働時間の外観識別性が本当にあるかについては若干気になるところ。また、前回議論のあった、労働時間がどれくらいバイアスを持っているかということについては、真の値と回答値との間にどの程度乖離があるか、超長時間労働の数値がどの程度信頼性の低いものなのかは必ずしも明らかではないので、トップコーディングして使用するかどうかはユーザーの判断に委ねれば良いのではないか。本部会では、あくまでも匿名性を担保できるかということを議論した上で、それ以外の部分はユーザーが自由に使えるように、できる限り情報を出してユーザーに判断してもらうのがよいのではないか。

・　匿名性の担保という点に関しては、個票のデータを出すため、大変慎重にならなければならない。ただ、匿名性だけでなく有用性も考えた場合に、よくわかっている人ばかりが利用するわけではないということでデータの信頼性についても考慮すべきと考える。90 時間以上で線を引いたのは、匿名性、秘匿性を破られないようにするために 0.5%基準に基づいてトップコーディングするという考えの他に、信頼性の低いものを除くためにトップコーディングするという考えもできるのではないか。

≪部会長のまとめ≫

・　外観識別性に関しては議論のあるところではあるが、現時点では外観識別についてリスクがあるという意見がある以上、0.5%基準により 90 時間以上をトップコーディングすることは、データ分析的に極端な値に引きずられないようにすることも考慮して、妥当ではないかと判断したい。

キ　前月欄の情報

・　前回の議論を踏まえて、産業分類を大くくり化すること等により前月欄の他の情報の秘匿をもう少し緩和できないかということに関する検討結果を説明していただいたが、産業分類を大分類としたとしても個人が特定化されるリスクが残っていることから、経時的な変化の情報の匿名データ化は大変難しいということを改めて理解した。前月欄の情報については、個人の特定化につながる恐れがあることから、「月末１週間に仕事をしたかどうかの別」以外は、匿名データに含めないとしたことについては、妥当と判断する。

ク　その他の匿名化措置

・　本調査は住戸をベースに調査対象者を選んでおり、死亡・転出のレコードを削除することになって

いるが、新たに入居した場合は、データとしてはどのような形で保持されているのか。

≪総務省からの回答≫

・　２ヶ月目については、調査票内に転入、転出の確認をする項目があり、そこにマークが付けられるため、新しく転入された人に対して「転入」というフラグが付くようになる。

≪部会長のまとめ≫

・　15 歳未満の男女別総数を男女の区別をしないで総数に置き換えること、自衛官、受刑者、並びに死亡・転出等のレコードを削除することについては、妥当と判断したい。また、転入のレコードの扱いについても妥当と考える。

ケ　匿名化措置を予定していない事項

・　15 歳未満の世帯員については、２～４歳階級別に調査されており、就業状態等の情報もないことから、有用性の観点も含めて、更にリコーディングする等の措置を講じる必要はなく、妥当と判断する。また、他の外観識別可能な事項は存在しないということで、妥当と判断させていただきたい。

コ　他の情報との関係

・　今回匿名化された内容であれば、外部の情報との対応関係から世帯・個人を特定される危険性があるものはないと考えられることから、妥当であると判断させていただきたい。

サ　匿名データの提供時期及び作成対象年

・　既に匿名データが提供されている総務省４調査のときも、平成以降の調査を対象とするということであったが、平成元年以降という基準は、どういう理由によるものか。

≪総務省からの回答≫

・　平成元年より前の電子データにはさまざまな符号が入っており、クリーニングが必要であることから、データ整備に時間やリソースを要する。

・　総務省４調査のときにも議論されたと思うが、例えば社会生活基本調査などは時期によって断層があり、うまくつなげることができないということもあって、現在の技術的な問題も含めて、平成以降ということで開始することとした。ただ、過去の遡及に関しては、ある程度代表的なものが出揃ってから、概念の調整や技術的な問題の解決も含めて、次のステップとしてやっていただこうという整理にしたと思う。

・　試行的提供の時に５年周期の調査を対象としており、マンパワーとの関係もあって、直近の調査結果と一つ前の調査結果を対象としたため、結果的に平成元年になったということはある。

・　過去に遡るほどデータの整備は大変になるが、逆に労働経済学などでのニーズがある程度わかっていると、過去のデータの作成についても促進できるのではないか。

・　５年周期の構造調査と、速報性を重視した月次調査とでは、匿名データのニーズはかなり違うと思う。したがって、目的が違い、調査方法も違うといったときに当然提供する匿名データも変わってくると思う。

・　時間の経過により、過去のデータを電子化するのは難しくなっていくことも考えると、過去の調査も貴重な財産であることから、できる限り電子化すべきと考える。

≪部会長のまとめ≫

・　過去に遡るほど整備が大変ということのようだ。今回は平成元年以降ということであるが、平成よ

り前のデータについても必要な時間やリソースがあるときに対応できるよう、情報の保護に努めてもらいたい。

シ　トップコーディングが行われた変数

・　前回の国民生活基礎調査の場合は、トップコーディングの対象が世帯所得などであり、対象レコード数が少なく、値が大きく振れてしまうため、トップコーディングをした部分の平均値等を出すことはミスリードするという結論だった。一方、今回は、変数が就業時間であるため、極端な値が出てくるものではなく、レコード数、平均値、標準偏差を見てみても比較的安定している印象を受ける。

・　トップコーディングが行われた変数の扱いについては、調査や変数ごとに状況が異なるため、ある程度、実際のものを見て判断するということになるのではないか。

≪厚生労働省からの回答≫

・　国民生活基礎調査については、現在、匿名データを作成中であり、それが終われば今後の課題として検討したいが、トップコーディングが行われた変数の扱いについては、各省バラバラよりは統一的な基本があった方がよいのではないか。

≪総務省からの回答≫

・　今回、ある程度レコード数もあるため、トップコーディングした部分の平均値と標準偏差を出したいと考えている。前回の国民生活基礎調査の場合は、対象レコード数が少なく、結果が安定しないということがあるため、調査によって提供の仕方が異なることはあると解釈している。

≪部会長のまとめ≫

・　トップコーディングをした変数については、最低限、変数全体の基本統計量は出していただきたい。ただし、今回のようにトップコーディングした部分の平均値や標準偏差がある程度の安定性を持って公表でき、特に誤解を招かないということであれば、利用者からすれば、変数全体の平均値よりもトップコーディングした部分の平均値の方が便利であるため、そのような提供の仕方も認めることとしたい。提供の仕方については、各省の判断となるのではないか。

ス　特定調査票の匿名データ化

・　特定調査票については、総務省の回答でも今後の課題とされているが、匿名データ化を進めて行く方向で考えられていることから、期待したい。

セ　同一世帯のマッチング

・　同一世帯のマッチングによる匿名データは、現段階では困難であるということだが、一方で、本調査では同一世帯のマッチングについては検討中の課題とされていることから、そういった研究が進んだ場合、将来的には匿名化されたパネル化データの可能性について検討の余地はあるのか。

≪総務省からの回答≫

・　前月からの異動状況を考えると、個人を特定される可能性が高い。今の匿名データ制度、法律上では個人が絶対特定されてはならないという形になっているため、現状では匿名データとしては提供できないと考える。

・　今の匿名化の方式を前提とする場合、パネル化はなかなか難しいが、オンサイト拠点という形でセキュリティを確保した上で、分析自体にコントロールがかかっているような状況の中であれば、制度的には目的外申請ということになるだろうが、可能かもしれない。あるいは、いろいろな匿名化技法

を駆使して、ほとんどイミテーションに近いデータをつくるということはあるかもしれない。

・　労働力調査は２ヶ月調査を行って翌年また同じ調査対象を調査することになっているが、そのデータをつなげることに関して、かなり根本的な問題があるように思える。今の労働力調査の調査方法でパネル化するということについては、そもそも原理的に難しいのではないか。

・　労働力調査は、抽出単位が世帯ではなく住戸であるため、その意味では誤解を招く「パネル化」という言葉を使ってよいのかということも含めて、整理をしておいたほうがよい。

≪部会長のまとめ≫

・　同一世帯のマッチングが可能ではないかということで、学会の期待もあるところではあるが、労働力調査はきちんとしたコンプリートパネルができるというわけではなく、パネルと呼べるかどうかも難しい。前月欄の情報提供でさえも個人の特定化のリスクがある中で、同一世帯のマッチングを行うことは、現時点では困難と言わざるを得ない。

ソ　匿名化技法の検討

・　ノイズの付与やスワッピング等による匿名化技法については、個票にあるデータをそのまま公表するという制約の下で匿名データをつくるのか、ノイズ等を付与した擬似的なデータを許容するのかということに関して、この部会の中でも両論があり、相当な研究が必要である。基本的には、今後の研究課題と捉えたい。

（２）その他

次回の匿名データ部会は８月１日（月）に開催することとなった。

以上

＜文責 内閣府大臣官房統計委員会担当室 速報のため事後修正の可能性あり＞