参考２

第15回 匿名データ部会 議事概要

１　日　　時　平成26年10月３日（金）　13:00～14:50

２　場　　所　内閣府庁舎３階　特別会議室

３　出 席 者
（部　会　長）北村　行伸
（委　　員）川崎　茂、津谷　典子
（専 門 委 員）伊藤　伸介、川口　大司、村田　磨理子
（審議協力者）総務省（政策統括官（統計基準担当））、財務省、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、東京都、千葉県
（諮　問　者）総務省統計局：植山　克郎調査企画課長、江刺　英信労働力統計室長ほか
（事　務　局）内閣府統計委員会担当室：伊藤　由樹子室長、佐々木　健一企画官ほか

４　議　　事
（１）社会生活基本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
部会長から部会長代理として川崎委員が指名された後、社会生活基本調査の匿名データ作成について、了承された論点に従い審議された。委員等の主な意見は以下のとおり。

Ⅰ　匿名性及び有用性の確保
（１）匿名化措置をしている事項
ア　ファイルの種類
世帯単位のファイルとしていることについては、適当と判断した。
イ　地域区分
・大都市圏とそれ以外の地域で生活時間は違うので、秘匿性が大きな問題にならないのであれば、調査票Ａと同様に三大都市圏、それ以外の地域区分の情報があった方がユーティリティーは上がる。
・有用性の観点からみると、都道府県レベルの情報を追加してほしい。都道府県によって異なる政策環境を持つケースについて分析をする際の有用性が非常に高まる。標本サイズが小さいので標準誤差が大きくなるが、標準誤差を見ればどの確度で推定されているかが分かるので、サンプルサイズが小さいから細かい地域区分を出せないということにはならない。また、環境変数を都道府県情報を使って統合した分析を行う際には、環境変数は連続変数として取り扱われることになり、必ずしも分析精度が下がるとは限らない。
・調査票Ｂが全国一本で結果表章する前提で調査設計の面から精度保証されてきたとすれば、匿名データの分析結果が実態と合っているかどうかをユーザー側が確認す

ることも考えると、全国一本で提供するのは止むを得ない。

・三大都市圏、それ以外の地域区分の情報を提供しても問題ないのではないか。秘匿性の観点から細かい地域区分を出すのは問題だが、大都市か否かでは、かなり地域が広くなるので、特異な属性を持った人がいても、特定される可能性は低い。地域区分があれば分析が深まるのであれば、地域区分を追加するのは好ましい。公表されている結果表の中に入っていないという理由で否定するのは難しい。

・調査票Ａの匿名データを作る際の議論では、地域区分情報を都道府県単位で表章すると匿名化が破られる危険性があったため、三大都市圏か否かでまとめられた。今回も同様で良いのではないか。精度については問題ないと思う。

・有用性と精度について、実際に調査票Ａのデータを使った立場から話すと、三大都市圏か否かという地域区分は分析に活かすことは出来なかった。統計表を見た限りでは、三大都市圏別々と三大都市圏以外の４区分は欲しい。地域区分が細かくなると標準誤差が大きくなって使いづらくなるため、地域区分を入れるのであれば、秘匿性保持の観点と、標準誤差がどのくらいあるのかということを実際の数字を見て判断したい。

・精度の保証を作成者がどこまで責任を持つかは、もう少し考えても良いのではないか。匿名データを利用した者が研究結果を公表した際、一義的な責任は利用者にあり、作成者に責任を負わせるのは酷だと思う。

（部会長のまとめ）

・地域区分については意見が分かれており、次回に論点を整理して回答をいただきたい。

ウ　リサンプリングの方法

・80％は妥当だと思うが、原データとリサンプリング済みデータのかい離の度合いを見ると、差があるかないかを判断するのは微妙であり、平均値自体は小さくても差が大きい項目もある。80%のリサンプリングをするときにどのような抽出をするのか、ばらつきが大きい項目については層化して抽出した方が、80％のサンプルの安定度が高まるのではないか。いま技術的な提示を行うことは出来ないが、検討した方が良いのではないか。

・以前はコンピュータのデータ処理能力に問題があったので、データの信用性との兼ね合いでリサンプリング率を80％にしたのだと思うが、現在は処理能力の問題はほとんどないので、もう少し高くしても良いのではないか。今回は元データの規模が小さいので、臨機応変に考えてみてはどうか。

（部会長のまとめ）

・リサンプリング率80％は絶対ではないので、今後検討すべき課題だと思うが、今回は適当だと判断する。

エ　情報の削除

ｉ）直接的な識別情報の削除

調査地域を特定する調査区番号などの実査用の識別番号や調査客体を直接識別でき

る情報は削除することについては、適当と判断した。

ⅱ）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

・出現頻度の分布を見るにあたって、全国レベルで特異なレコードを確認して削除したということであれば、地域区分を細かくすると削除するレコードが増えることになる。サンプル数が少ないので、削除する数は出来るだけ少ないことが望ましい。

・データの精度を考えると、削除されるデータ数を示してほしい。世帯員の数と子供の数が多い世帯について、親の年齢や住宅形態のクロスをとって決めるということだが、クロスの作成条件を明文化した上で、その条件のとおり削除するとどの程度の数になるのかということを見てから決定するべきではないか。

・削除数は、ユーザーにも分かるような対応をお願いしたい。

（部会長のまとめ）

・削除されるデータ数が分かる資料を準備して次回、審議したい。特定の世帯を削除することは、該当の世帯が稀なケースであっても政策的に意味があることがあるので、慎重に行わなければならない。生活時間の集計バイアスについて、簡単に提示できるものがあれば対応をお願いしたい。

オ　分類区分の再編（世帯員に関する項目）

・年齢を各歳で全て表章する必要はないと思うが、学校卒業時の就職年齢を見たいときなど、一律に５歳階級でまとめてしまうと有用性が落ちるのではないか。また、定年退職した後の生活時間は大きく変わると思うが、55歳から59歳、60歳から64歳を一括りにして良いのだろうか。80代の後半は我が国で人口の増加が急激に進んでいる年齢層なので、トップコーディングを85歳以上で行う場合と、90歳以上で行う場合ではどれだけ違うのかということをデータで提示していただき、慎重に判断したい。

・データの有用性について見ると、５歳階級区分は残念に思う。日本の各種の政策が年齢を基準にして行われることがあり、対象年齢前後の行動を調べる研究を行う際に、５歳階級区分は研究を大きく阻害する要因になっている。各歳情報を提供できるような方法を検討願いたい。

・社会生活基本調査の調査票Ｂの年齢階級の話に限ってみれば、85歳のトップコーディングというのは特に支障を感じない。また、末子の年齢については、10歳未満の世帯員について教育や年齢の情報が入っているし、末子でも10歳以上であれば情報は更にたくさん入っているので、そこを組み合わせればかなりの情報は使えると思う。

・末子の年齢についてはこれで良いのではないかと思う。５歳階級の途中で　色々な現象が起こるという指摘があったが、例えば、大学を卒業して就職する年齢はかなりばらつきがあるので、年齢の要因で説明するよりも、就業状態あるいは就学状態で説明する方が合理的だと思う。今回は元データのサンプルが小さいことを考えればやむを得ないのではないかと思う。また、この調査は生活行動を詳細に記入するものであり、85歳以上の生活行動の記入精度を考えると原案のとおり

で良いのではないかと思う。
末子の年齢については、秘匿のためのトップコーディングとは違うのではないか。大事なことは、末子の年齢に相当するような世帯員の情報が各レコードの中に埋め込まれているかどうかである。

・年齢が各歳で提供されると分析の可能性が広がるが、年齢は外観識別性を表す変数の一つなので、秘匿性の観点から十分に注意して議論する必要がある。今後の課題として、匿名データの提供可能性に関する議論がなされると思うが、各歳提供の可能性を改めて検討した方が良いのではないか。

（部会長のまとめ）

・トップコーディングを90歳で行った場合、どの程度のサンプルになるのか次回提示していただき、もう一度議論する。末子年齢の12歳以上のトップコーディングについては適当と判断する。年齢を各歳で提供することは、今後の提供の仕方に関わってくると思われるので、次回以降の部会で議論したい。

（2）匿名化措置をしていない事項

匿名性の確保については適当と判断する。

Ⅱ　次回の議題について

地域区分、出現頻度の低いデータの削除、85歳以上のトップコーディングについては次回に再度議論をすることになった。

また、前回答申の「今後の課題」に係る審議も次回とした。

6　その他

・次回は、10月17日（金）13時30から中央合同庁舎８号館８回特別中会議室で開催することとされた。

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＞