国勢調査の匿名化ミクロデータの作成可能性について

―地域区分に着目して―*

中央大学経済学部　伊藤 伸介**
(独)統計センター　星野 なおみ
総務省統計局　阿久津 文香

1. はじめに

諸外国では、公的統計(政府統計)のミクロデータに関して、①匿名化ミクロデータ(個票データに匿名化処理が施されたデータ)の提供、②個票データの提供、③オーダーメイド集計、④オンデマンド型の提供サービス(リモート集計)といったさまざまな形態による提供が進められてきた。一方、我が国では、平成 19 年に統計法（平成 19 年法律第 53 号）が成立してから、様々な形態で公的統計のミクロデータの提供が進められてきた。統計法に基づき策定された「公的統計の整備に関する基本的な計画」（平成 21 年 3 月 13 日閣議決定、以下「基本計画」と略称）においては、「二次的利用に係るガイドラインに基づき、平成 21 年度から、秘密の保護に配慮しつつ、二次的利用に係る事務処理を適切に開始し、平成 22 年度以降、順次、二次的利用の対象となる統計調査やサービスを拡大する」ことが求められている。こうした「基本計画」に基づいて、匿名データの作成・提供が開始された。

このような二次的利用に関する統計法制度に基づいて、総務省統計局は、6 種類の公的統計の匿名データの提供を行っている。国勢調査に関しては、平成 25 年 12 月と平成 26 年 3 月にそれぞれ、平成 12 年と 17 年の匿名データの提供が開始されている。現在提供されている国勢調査の匿名データ(以下「提供済匿名データ」と呼称)は、世帯単位に基づいて、サンプリング率が 1%で抽出されるだけでなく、リコーディング(再符号化)、トップコーディング、レコード削除等の匿名化措置が適用されている。なお、リコーディングやトップコーディングにおいては、いわゆる「0.5%基準(単変量において母集団の 0.5%を下回る区分を統合すること)」が適用されてきた。

一方、我が国の国勢調査の提供済匿名データでは、都道府県及び人口 50 万以上の市区が最小の地域区分となっているが、より詳細な地域区分を用いた分析が可能なミクロデータに対するニーズが存在すると思われることから、国勢調査の匿名データに関しては、より詳細な地域区分の提供可能性が検討されてよいと考える。他方で、ミクロデータに含まれる個体情報の保護と利用者のニーズを勘案した上で、ミクロデータに対する匿名化措置を検討することが求められる。本稿では、秘匿性と有用性の両面から、詳細な地域区分の作成可能性を検討する。

2. 「地域の人口規模の閾値」に基づく秘匿性の検証

匿名化ミクロデータの作成・提供においては、秘匿性に関する閾値を設定した上で、その

---

* 本稿は、伊藤・星野・阿久津(2016b)に基づいている。なお、本稿の内容は個人的な見解を示すものであり、統計センターの見解を表すものではないことに留意されたい。

** (独)統計センター非常勤研究員

閾値を超えない形で、様々な匿名化技法が適用されることが考えられる。秘匿性に関する閾値については、主に２つの考え方が存在する。第１は、閾値に関する設定可能性であって、例えば、イギリスの 1991 年人口センサスの匿名化標本データ(Samples of Anonymised Records=SARs)を作成する上で、閾値ルール(thresholding rule)に基づく地域区分、個人・世帯属性の分類区分とサンプリング率の関係についての定式化 (Dale(1995), Marsh *et al.*(1994), 伊藤(2011))[1]が議論された。第 2 は、既存の匿名化ミクロデータの秘匿性と同レベルの秘匿性を有する匿名化ミクロデータの作成可能性であって、イギリスの 2001 年人口センサスでは、地域区分が詳細な小地域ミクロデータ(Small Area Microdata=SAM)の作成に関する研究において、Tranmer 等が 1991 年 SARs における露見リスクを基準とした SAM の露見リスクに関する相対評価を試みている。

一方、地域の人口規模に関する閾値との関連で、秘匿性の程度を定量的に明らかにした研究も存在する。Hawala(2001)は、アメリカ人口センサスの Public Use Microdata Sample の作成において用いられる 10 万人という地域区分の閾値に関して、その秘匿性に関する事後検証を行うために、母集団一意(population unique)の比率を用いて地域の人口規模と秘匿性の指標との関連性を明らかにしている。

他方、伊藤ほか(2016a)は、提供されている平成 12 年国勢調査の提供済匿名データを作成した上で、より詳細な地域区分において「許容可能な」[2]リコーディングの組み合わせとサンプリング率との関係を検討した。具体的には、提供済匿名データについて秘匿性に関する定量的な評価を行い、その秘匿性の評価値を基準値とした上で、その基準値を超えない形で、様々なリコーディングとサンプリングが施されたデータを対象に、秘匿性に関する相対的な評価を行っている。

ところで、わが国では、匿名データの作成において、様々な匿名化手法が用いられるが、主要な方法の１つは、リコーディング(再符号化)である。わが国では、リコーディングやトップコーディングにおいて「0.5%基準」[3]が用いられてきたが、この 0.5%基準に基づくリコーディングが匿名データの有用性および秘匿性に及ぼす影響については、これまでも議論の対象となってきた。他方で、ある特定の秘匿性の閾値(例えば、地域の人口規模に関する閾値)を設定し、その閾値を満たすように、個人・世帯の属性に関する区分統合を行うことも考えられる。このような観点から、地域の人口規模の閾値を設定することができれば、地域情報の秘匿を考慮した上で、地域区分が詳細な匿名データの作成も可能になる。

本研究では、平成 22 年国勢調査の A 県の調査票情報(個票データ)を用いて、キー変数における分類区分のリコーディングの可能性を探る。具体的には、個人単位の提供を前提に、

---

[1] 閾値ルールとは，SARs において許容されている最小の地理的区域において，各変数における分類区分に関する期待度数(expected count) が，標本レベルにおいて 1 以上になることである(Marsh *et al.*(1994, p.43), 伊藤(2011))。なお, 集団における期待度数に関する定式化は、下記の式で示される(Dale(1995, p.8))。

$$C = \frac{1}{X} * \frac{Y}{Z}$$

ここで C: 母集団における期待度数
X: 標本抽出率(個人 SAR の場合 1/50 世帯 SAR の場合 1/100)
Y: イギリス全土の総人口（約 5600 万人）
Z: 地理的区域における最小人口規模(ex. 個人 SAR の場合は 120000 人，世帯 SAR については約 190 万人が適用)

[2] 本研究における「許容可能」とは，匿名化ミクロデータにおける秘匿性が、本研究において設定した秘匿性の閾値を超えないことを意味している。

[3] 0.5%基準については、「匿名データの作成・提供に係るガイドライン」（改正 平成 28 年 1 月 22 日）「匿名化処理の技法」を参照。

地域の閾値を変更した場合の区分統合の可能性を探ることにしたい。

最初に、秘匿性に関する第 1 の研究として、伊藤・星野(2014, 2015)に基づき、リコーディングを行ったデータに対して母集団一意(population unique)の比率を計測した。本研究において、母集団一意の計測のために使用するキー変数は、次の 10 変数である。

① 住宅の建て方
② 住居の種類
③ 性別
④ 配偶者の有無
⑤ 国籍
⑥ 労働力状態
⑦ 従業上の地位
⑧ 年齢
⑨ 産業
⑩ 職業

上記の 10 変数の中で、①住宅の建て方、②住居の種類、③性別、④配偶者の有無、⑤国籍、⑥労働力状態、⑦従業上の地位については、提供済匿名データの区分を利用するが、本研究では、⑧年齢、⑨産業と⑩職業に着目し、以下のようなリコーディングおよびトップコーディングを施した。

⑧年齢
(1)各歳年齢区分でトップコーディングなし
(2)各歳年齢区分でかつ 85 歳以上トップコーディング
(3)各歳年齢区分でかつ 90 歳以上トップコーディング
(4)各歳年齢区分でかつ 95 歳以上トップコーディング
(5)5 歳年齢区分でかつ 85 歳以上トップコーディング
(6)5 歳年齢区分でかつ 90 歳以上トップコーディング
(7)5 歳年齢区分でかつ 95 歳以上トップコーディング
(8)10 歳年齢区分でかつ 90 歳以上トップコーディング
(9)10 歳年齢区分でかつ 100 歳以上トップコーディング

⑨産業
(1) 大分類(21 区分)
(2) 鉱業・採石業・砂利採取業と建設業のリコーディング、農業、林業と漁業のリコーディング、電気・ガス・熱供給・水道業と製造業のリコーディングおよびそれ以外の産業については大分類(0.5%基準を考慮)(17 区分)
(3) 平成 17 年提供済匿名データに合わせた区分での大分類のリコーディング(14 区分)

⑩職業
(1)大分類(12 区分)
(2)保安職業従事者、農林漁業従事者と輸送・機械運転従事者のリコーディング＋それ以外の大分類(0.5%基準を考慮)(10 区分)
(3)平成 17 年の提供済匿名データの職業のリコーディング(8 区分)

本研究では、上記の全 81 パターンについて、母集団一意の計測を行った。

つぎに、地域の人口規模と属性の分類区分との関連性を実証的に明らかにするために、A 県を対象にし、地域の閾値については、(1)人口 20 万人以上地域(県庁所在市に該当する地域も含む)、(2)人口 10 万人以上地域、(3)人口 5 万人以上地域、(4)人口 3 万人以上地域、(5)人口 2 万人以上地域、(6)人口 1 万人以上地域、(7)人口 5000 人以上地域、および(8)人口 1000 人以上地域の 8 パターンを設定し、それに該当する 20 地域を選定した。表 1 は、選出された 20 地域（それぞれ「地域 A」～「地域 T」）とその人口規模を示している。

付表 1 はそれぞれ、地域 A のそれぞれにおける年齢、産業と職業の様々なパターンにおける母集団一意の比率を示したものである。例えば、付表 1 を見ると、地域 A において、年齢、産業と職業が原区分である場合の母集団一意の比率は約 18.43%であるが、年齢のみを 5 歳区分に統合した場合、母集団一意の比率が 8.83%に減少することがわかる。その一方で、産業についてのみ、匿名データと同様の区分でリコーディングを行っても、母集団一意の比率は 17.47%であり、大きな変化は見られない。職業に関しても、原区分から匿名データと同様の区分でリコーディングを行った場合、母集団一意の比率は 17.64%となっている。さらに、産業と職業のいずれも、0.5%基準を踏まえてリコーディングを行った場合の母集団一意の比率はそれぞれ、18.21%と 18.33%となっている。このことから、母集団一意の観点から見た場合、産業と職業に関しては、現行より詳細化された区分で提供できる可能性がある。

つぎに、図 1 は、地域の人口規模ごとに年齢、産業と職業のあらゆる組み合わせについて算出された母集団一意の平均値を用いた場合の地域の人口規模と母集団一意の比率の関係をそれぞれ示している。本図によれば、地域の人口規模と母集団一意の比率との間にトレードオフの関係が実証的に確認できる。すなわち、人口が多いほど、母集団一意の比率は減少傾向にあることがわかる。したがって、年齢、職業、産業と分類区分が設定された場合に、母集団一意の比率に関する閾値を適切に定めることができれば、秘匿の観点から、提供可能な地域の人口規模の閾値を導出することが可能である。

そこで、年齢 5 歳階級 85 歳以上トップコーディング、産業パターン(3)および職業パターン(3)(提供済匿名データと同じ区分)を用いた場合の地域の人口規模と母集団一意の比率の関係についてグラフを作成した上で、A 県において世帯単位および個人単位に基づいて作成されたテストデータを対象に母集団一意の比率を算出し、その比率を適用すると、表 2 のように、個人単位においては、約 20 万人と算出することができる[4]。現在、匿名データの作成・提供に関するガイドラインでは、「地域の人口規模が人口 50 万人以上であること」が匿名化の目安になっているが、本分析結果に基づけば、この目安とする「50 万人以上」という閾値を緩和することができるかもしれない。

つぎに、秘匿性に関する第 2 の研究として、年齢、産業、職業と地域の人口規模に秘匿処理を施した場合、母集団一意の比率に対する影響の程度を明らかにするために、重回帰分析を行った。具体的には、母集団一意の比率を被説明変数とし、説明変数に関しては、年齢のリコーディングおよびトップコーディングに関する 9 パターン、産業のリコーディングに関する 3 パターン、職業のリコーディングに関する 3 パターンのそれぞれに関するダミー変数(パターンに該当する場合には 1、そうでない場合には 0)、および対数変換された地域の人口規模をモデルに組み込んでいる。

回帰分析に用いた変数の定義は、表 3 で示されている。また、回帰分析の結果は、表 4 で

[4] 世帯単位に基づいて作成されたテストデータは、世帯主のレコードのみから構成されている。そのため、世帯単位のテストデータにおける母集団一意の比率は、個人単位に基づいて作成されたテストデータにおけるそれと比較して小さくなっている。

表１　本研究で用いる地域の名称と地域の人口規模に関する一覧表

地域の名称	地域の人口規模の閾値
地域　A	200，000人以上
地域　B	100，000人以上
地域　C	50，000人以上
地域　D	50，000人以上
地域　E	30，000人以上
地域　F	30，000人以上
地域　G	20，000人以上
地域　H	20，000人以上
地域　I	20，000人以上
地域　J	10，000人以上
地域　K	10，000人以上
地域　L	10，000人以上
地域　M	10，000人以上
地域　N	5，000人以上
地域　O	5，000人以上
地域　P	5，000人以上
地域　Q	5，000人以上
地域　R	1，000人以上
地域　S	1，000人以上
地域　T	1，000人以上

図１　地域の人口規模と母集団一意の比率との関係―地域ごとの平均値

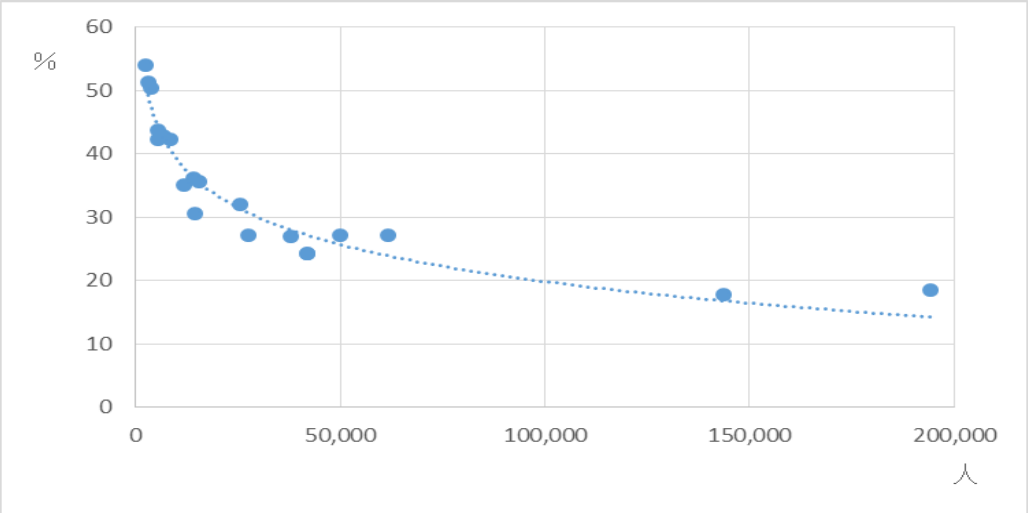


表２　地域の人口規模の閾値

	母集団一意の比率	地域の人口規模の閾値（人）
世帯単位	4．20%	164，194
個人単位	2．98%	197，758

示されている。年齢、産業と職業については原区分をリファレンスグループとしている。分析結果を見ると、全般的にマイナスに有意になっていることから、秘匿処理を施すことによって、母集団一意の比率が有意に低減することがわかる。また、 年齢については、区分統合の程度を高めるにしたがって、回帰係数の絶対値が相対的に大きくなっていることも確認できる。さらに、標準化偏回帰係数(Beta)に着目すると、地域の人口規模の係数の絶対値が最も大きいことから、本分析結果から、年齢と比較した場合、地域の人口規模が母集団一意の低減に及ぼす影響が最も高いことが明らかになった。

表 3 回帰分析で用いられた説明変数の一覧表

変数の名称	変数の定義
年齢カテゴリー 1	年齢パターンが各歳年齢区分でトップコーディングなしに該当する場合には1。そうでない場合には0。
年齢カテゴリー 2	年齢パターンが各歳年齢区分でかつ85歳以上トップコーディングに該当する場合には1。そうでない場合には0。
年齢カテゴリー 3	年齢パターンが各歳年齢区分でかつ90歳以上トップコーディングに該当する場合には1。そうでない場合には0。
年齢カテゴリー 4	年齢パターンが各歳年齢区分でかつ95歳以上トップコーディングに該当する場合には1。そうでない場合には0。
年齢カテゴリー 5	年齢パターンが5歳年齢区分でかつ85歳以上トップコーディングに該当する場合には1。そうでない場合には0。
年齢カテゴリー 6	年齢パターンが5歳年齢区分でかつ90歳以上トップコーディングに該当する場合には1。そうでない場合には0。
年齢カテゴリー 7	年齢パターンが5歳年齢区分でかつ95歳以上トップコーディングに該当する場合には1。そうでない場合には0。
年齢カテゴリー 8	年齢パターンが10歳年齢区分でかつ90歳以上トップコーディングに該当する場合には1。そうでない場合には0。
年齢カテゴリー 9	年齢パターンが10歳年齢区分でかつ100歳以上トップコーディングに該当する場合には1。そうでない場合には0。
産業カテゴリー 1	産業パターンが21区分に該当する場合には1。そうでない場合には0。
産業カテゴリー 2	産業パターンが17区分に該当する場合には1。そうでない場合には0。
産業カテゴリー 3	産業パターンが14区分に該当する場合には1。そうでない場合には0。
職業カテゴリー1	職業パターンが12区分に該当する場合には1。そうでない場合には0。
職業カテゴリー2	職業パターンが10区分に該当する場合には1。そうでない場合には0。
職業カテゴリー3	職業パターンが8区分に該当する場合には1。そうでない場合には0。
地域規模の対数	地域の人口規模の対数

表 4　回帰分析の結果

説明変数	係数	標準誤差	t値	Beta	有意性
年齢パターン〈年齢カテゴリー 1〉					
年齢カテゴリー 2	-0.007	0.003	-2.232	-0.018	**
年齢カテゴリー 3	-0.003	0.003	-1.121	-0.009	
年齢カテゴリー 4	-0.001	0.003	-0.478	-0.004	
年齢カテゴリー 5	-0.140	0.003	-46.274	-0.368	***
年齢カテゴリー 6	-0.139	0.003	-45.956	-0.365	***
年齢カテゴリー 7	-0.139	0.003	-45.829	-0.364	***
年齢カテゴリー 8	-0.184	0.003	-60.597	-0.481	***
年齢カテゴリー 9	-0.184	0.003	-60.533	-0.481	***
産業パターン〈産業カテゴリー 1〉					
産業カテゴリー 2	-0.003	0.002	-1.645	-0.011	
産業カテゴリー 3	-0.008	0.002	-4.368	-0.030	***
職業パターン〈職業カテゴリー 1〉					
職業カテゴリー 2	-0.001	0.002	-0.564	-0.004	
職業カテゴリー 3	-0.007	0.002	-3.873	-0.027	***
地域規模の対数	-0.072	0.001	-120.441	-0.718	***
定数	1.044	0.006	164.414		***
Adj.R^2	0.943				
F値	2043.248				
N	1620				

注　***・・・1%有意、**・・・5%有意、*・・・10%有意をそれぞれ表している。また、< >は、リファレンスグループを示している。

3. 情報量損失に基づいた有用性の定量的な評価

つぎに、本研究では、様々な地域を対象に有用性の検証を行った。ミクロデータにおける有用性の定量的な評価方法については、クラーメルの V といった関連性の指標の算出や原データからの絶対距離の平均値(average absolute distance)の計測等を行うことが考えられる(伊藤・星野(2014))。また、伊藤ほか(2016a)においては、情報量損失に関する指標の 1 つであるエントロピー[5]を用いて、秘匿の観点から許容可能な分類区分の組み合わせに関する情報量損失の計測を行った[6]。具体的には、関連性の指標として、原区分から分類区分の統合を行った場合のセルごとのエントロピーを計測し、区分の統合を行った場合におけるエントロピーの総計と該当する度数の総計の積を算出することによって、情報量損失の指標を作成した[7]。一方、本研究では、原区分からリコーディングを行った場合の距離を計測することによって、情報量損失の計測を行った。原区分と統合区分におけるクロス表の差に関する指標を作成し、原区分と統合区分におけるクロス表の差の検証を行う。具体的には、区分統合を行った場合、リコーディング後の度数をリコーディングの対象となった区分数で除することによって、度数の按分を行う。つぎに、リコーディング前のクロス表とリコーディング後に按分済みの度数が入力された表を用いて、情報量損失(IL)を算出する。情報量損失は、以下の(1)式で計測された。

$$IL = \frac{\sum_{c} \left| T^{R}(c) - T^{O}(c) \right|}{n_T} \quad \cdots (1)$$

ここで、IL は、個票データを用いて作成したクロス表におけるセルの度数$T^{O}(c)$とリコーディング済データを用いて作成したクロス表におけるセルの度数をリコーディングの対象となった分類区分数で除した$T^{R}(c)$の差の絶対値の合計を集計表におけるセルの数n_Tで割った数値である。以下の例は、年齢 5 歳区分と産業（農業、林業、漁業）のクロス表を対象に、80 歳以上と農林漁業に区分統合した場合の情報量損失を計測したものである[8]。以下のように、リコーディング後に按分された度数とリコーディング前のクロス表に含まれる度数が用いられる。

表 5 は、地域 A から地域 T の 20 地域を対象に、81 の変数のパターンにおいて算出される情報量損失値に関する基本統計量を示したものである。人口規模が小さい地域については、情報量損失値の平均値が相対的に小さくなっているだけでなく、標準偏差も小さいことが確認できる。このことから、地域の人口規模が小さい場合には、区分統合しても情報量損失の変化が小さいことが推察される[9]。また、図 3 は、地域ごとの情報量損失に関する平均

---

[5] 本研究で用いるエントロピーの説明については、伊藤ほか(2010, 7 頁)を参照されたい。

[6] 質的属性に関する有用性の定量的な評価については、シャノン情報量の概念に基づいた「情報エントロピー(entropy-based measures)」を用いて情報量損失を評価することが提案されている(Kooiman *et al.*(1998), Domingo Ferrer and Torra(2001), 竹村(2003))。

[7] 本研究では、De Waal and Willenborg(1999)の研究と同様に、リコーディングを用いて作成した匿名化ミクロデータを対象に、情報エントロピーを用いて情報量損失の計測を行った。

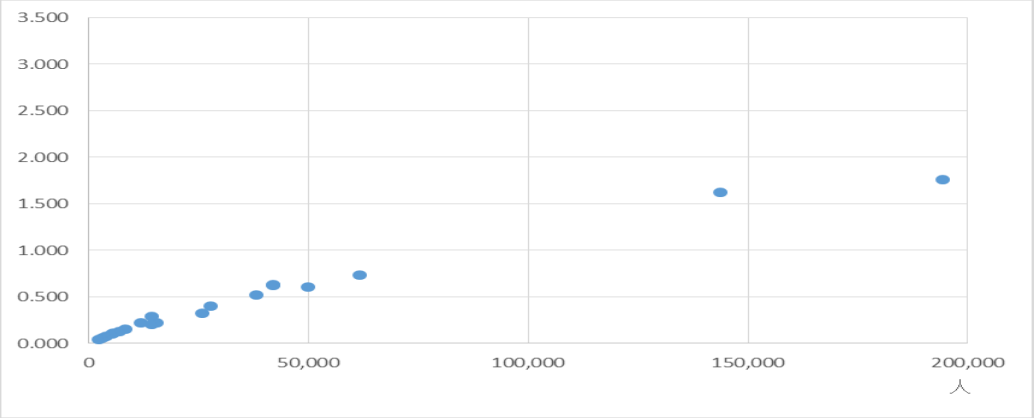
[8] 区分統合されているセルに度数 0 が含まれる場合でも平均値を入れていることに留意されたい。

[9]本研究では、地域 A から地域 T に関する情報量損失を計測している。本分析結果から、年齢、産業と職業の区分を統合するほど情報量損失が大きくなることが確認される。また、地域の人口規模が小さい場合には、情報量損失値は相対的に小さくなっているだけでなく、区分統合に伴う情報量損失の変化は小さくなっている。これについては、表 5 で示されるように、地域の人口規模が小さい場合、セルに含まれる度

表 5　地域別に計算された情報量損失の基本統計量

地域の名称	地域の人口規模の閾値	平均値	標準偏差	最小値	最大値	N
地域 A	200,000人以上	1.764	0.721	0	3.042	81
地域 B	100,000人以上	1.620	0.636	0	2.704	81
地域 C	50,000人以上	0.732	0.287	0	1.198	81
地域 D	50,000人以上	0.610	0.238	0	0.989	81
地域 E	30,000人以上	0.631	0.239	0	1.015	81
地域 F	30,000人以上	0.624	0.231	0	0.979	81
地域 G	20,000人以上	0.521	0.198	0	0.836	81
地域 H	20,000人以上	0.401	0.160	0	0.654	81
地域 I	20,000人以上	0.322	0.134	0	0.527	81
地域 J	10,000人以上	0.220	0.089	0	0.355	81
地域 K	10,000人以上	0.290	0.104	0	0.444	81
地域 L	10,000人以上	0.204	0.088	0	0.339	81
地域 M	10,000人以上	0.221	0.078	0	0.340	81
地域 N	5,000人以上	0.151	0.060	0	0.236	81
地域 O	5,000人以上	0.127	0.050	0	0.199	81
地域 P	5,000人以上	0.113	0.045	0	0.174	81
地域 Q	5,000人以上	0.099	0.040	0	0.156	81
地域 R	1,000人以上	0.076	0.033	0	0.120	81
地域 S	1,000人以上	0.058	0.026	0	0.093	81
地域 T	1,000人以上	0.046	0.022	0	0.074	81

図 3　地域の人口規模と情報量損失との関係―地域ごとの平均値



値と地域の人口規模との関係を表している。地域の人口規模が大きくなるほど、情報量損失が大きくなる傾向にあることを定量的に明らかにすることができる[10]。その一方で、本図から、地域の人口規模が相対的に小さい場合、情報量損失における大きな違いが見られないことを確認することができる。

4. むすびにかえて

本稿では、国勢調査の匿名化ミクロデータを用いて、地域の人口規模に基づく閾値を設定した場合の母集団一意の比率と情報量損失の検証を行った。本研究の結果を踏まえると、個人単位で抽出した匿名化ミクロデータにおいては、地域の人口規模と母集団一意の比率と

数が小さくなることから、リコーディングを行っても、度数の変化が相対的に小さいことが考えられる。

10 本研究では、情報量損失の指標としてエントロピーを用いた場合の地域の人口規模と情報量損失の関係も実証的に明らかにしている。さらに、有業者に該当するレコードに限定した場合の地域の人口規模とエントロピーの関係も追究した。本分析結果によれば、有用性の指標としてエントロピーを用いたとしても、基本的には、図 3 と同様の結果が見て取れることがわかっている。

の間にトレードオフの関係があることを実証的に明らかにすることができた。さらに、情報量損失に関する分析結果からは、地域の人口規模が大きいほど、情報量損失が大きくなること、秘匿処理が情報量損失に及ぼす影響は、属性によって異なることがわかった。

また、本稿では、年齢、産業、職業といった変数におけるリコーディングの程度を変えることによって、提供済匿名データよりも詳細な地域区分の提供可能性について実証的な研究が行われた。本分析結果によれば、年齢、産業、職業を含むキー変数を用いて算出された母集団一意の比率をもとに、秘匿性の閾値を適切に設定することができれば、提供済匿名データにおいて定められている地域の人口規模 50 万以上という区分より細かな地域区分の設定についても、議論することが可能であると思われる。さらに、本分析結果においては、0.5%基準を用いたリコーディングを行っても、母集団一意の比率が大きく変わらないことが定量的にも確認されたことから、母集団一意の比率を評価指標とするのであれば、産業と職業に関しては、提供済匿名データよりも詳細な区分で提供できる可能性についても検討する余地が出てくるだろう。本研究で議論された、秘匿性と有用性の両面からの実証研究を踏まえた上で、複数ファイルにおける地域詳細化データや年齢等の属性に関する詳細化データの作成可能性を模索する必要があるように思われる。

参考文献

Dale, A(1995) “Samples of Anonymised Records from the 1991 Census for Great Britain” *IASSIST Quarterly*, pp.5-12.

De Waal, T. and Willenborg, L. (1999) “Information Loss Through Global Recoding and Local Suppression”, *Netherlands Official Statistics (special issue on SDC)*, Vol.14, pp.17-20.

Domingo-Ferrer, J. and Torra, V. (2001) “Disclosure Control Methods and Information Loss for Microdata”, Doyle *et al.*(eds.) *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, Elsevier Science,Amsterdam, pp. 91-110.

Hawala, S.(2001) “Enhancing the "100,000 rule" On the Variation of the Per Cent of Uniques in A Microdata Sample and the Geographic Area Size Identified on the File”, Proceedings of the Annual Meeting of the American Statistical Association

伊藤伸介・高野正博・秋山裕美・後藤武彦(2010)「ミクロデータにおける有用性と秘匿性の定量的な評価に関する研究」,『製表技術参考資料』No.14, 1～40 頁

伊藤伸介(2011)「わが国におけるミクロデータの新たな展開可能性について―イギリスにおける地域分析用ミクロデータを例に―」, 明海大学『経済学論集』Vol.23, No.3, 36～54 頁

伊藤伸介・星野なおみ(2014)「国勢調査ミクロデータを用いたスワッピングの有効性の検証」『統計学』107 号, 2014 年 9 月 30 日, 1～16 頁

伊藤伸介・星野なおみ(2015)「ミクロデータにおける匿名化の誤差の評価に関する研究―国勢調査を例に―及びスワッピングの適用可能性に関する評価研究―国勢調査ミクロデータを用いて―」No.28, 1～43 頁

伊藤伸介・星野なおみ・阿久津文香(2016a)「国勢調査における匿名化ミクロデータの有用性と秘匿性の定量的な評価」『製表技術参考資料』No.32, 1～33 頁

伊藤伸介・星野なおみ・阿久津文香(2016b)「国勢調査ミクロデータに対する匿名化措置の可能性に関する研究」『製表技術参考資料』No.34, 1～59 頁

Kooiman, P., L. Willenborg and J. Gouweleeuw (1998) “PRAM: A Method for Disclosure Limitation of Microdata”, Research Paper, No. 9705, Statistics Netherlands, Voorburg.

Marsh, C., Dale, A., Skinner, C.(1994) “Safe Data versus Safe Settings: Access to Microdata from the British Census”, *International Statistical Review*, Vol.62, No.1, pp.35-53.

竹村彰通(2003)「個票開示問題の研究の現状と課題」『統計数理』第 51 巻 第 2 号,241～260 頁

Tranmer, M., Pickles, A., Fieldhouse, E., Elliot, M., Dale, A., Brown, M.(2005) “The Case for Small Area Microdata”, *Journal of Royal Statistical Society A*, Vol.168,pp.29-49.

付表１　母集団一意率　地域Ａ

年齢									産業			職業			母集団一意率
各歳トップなし	各歳85トップ	各歳90トップ	各歳95トップ	5歳85トップ	5歳90トップ	5歳95トップ	10歳90トップ	10歳100トップ	21区分	統合17区分	統合14区分	12区分	統合10区分	統合8区分	
＊									＊			＊			18.43%
＊									＊				＊		18.33%
＊									＊					＊	17.64%
＊										＊		＊			18.21%
＊										＊			＊		18.11%
＊										＊				＊	17.41%
＊											＊	＊			17.47%
＊											＊		＊		17.35%
＊											＊			＊	16.63%
	＊								＊			＊			18.24%
	＊								＊				＊		18.14%
	＊								＊					＊	17.45%
	＊									＊		＊			18.02%
	＊									＊			＊		17.92%
	＊									＊				＊	17.22%
	＊										＊	＊			17.28%
	＊										＊		＊		17.16%
	＊										＊			＊	16.43%
		＊							＊			＊			18.37%
		＊							＊				＊		18.26%
		＊							＊					＊	17.58%
		＊								＊		＊			18.15%
		＊								＊			＊		18.04%
		＊								＊				＊	17.34%
		＊									＊	＊			17.40%
		＊									＊		＊		17.28%
		＊									＊			＊	16.56%
			＊						＊			＊			18.41%
			＊						＊				＊		18.31%
			＊						＊					＊	17.62%
			＊							＊		＊			18.19%
			＊							＊			＊		18.09%
			＊							＊				＊	17.39%
			＊								＊	＊			17.45%
			＊								＊		＊		17.33%
			＊								＊			＊	16.60%
				＊					＊			＊			8.83%
				＊					＊				＊		8.75%
				＊					＊					＊	8.27%
				＊						＊		＊			8.69%
				＊						＊			＊		8.61%
				＊						＊				＊	8.13%
				＊							＊	＊			8.13%
				＊							＊		＊		8.04%
				＊							＊			＊	7.56%
					＊				＊			＊			8.86%
					＊				＊				＊		8.78%
					＊				＊					＊	8.30%
					＊					＊		＊			8.72%
					＊					＊			＊		8.64%
					＊					＊				＊	8.16%
					＊						＊	＊			8.16%
					＊						＊		＊		8.07%
					＊						＊			＊	7.59%
						＊			＊			＊			8.87%
						＊			＊				＊		8.79%
						＊			＊					＊	8.31%
						＊				＊		＊			8.73%
						＊				＊			＊		8.64%
						＊				＊				＊	8.17%
						＊					＊	＊			8.17%
						＊					＊		＊		8.08%
						＊					＊			＊	7.60%
							＊		＊			＊			6.45%
							＊		＊				＊		6.38%
							＊		＊					＊	6.01%
							＊			＊		＊			6.34%
							＊			＊			＊		6.27%
							＊			＊				＊	5.91%
							＊				＊	＊			5.88%
							＊				＊		＊		5.80%
							＊				＊			＊	5.45%
								＊	＊			＊			6.45%
								＊	＊				＊		6.38%
								＊	＊					＊	6.02%
								＊		＊		＊			6.34%
								＊		＊			＊		6.27%
								＊		＊				＊	5.91%
								＊			＊	＊			5.88%
								＊			＊		＊		5.81%
								＊			＊			＊	5.45%