事業所・企業系のミクロデータにおける匿名化措置の有効性の評価
－経済センサス - 活動調査を例として－

横溝　秀始†

伊藤　伸介‡

An Assessment of the Effectiveness of Anonymization Methods for Business Microdata: The Example of Economic Census for Business Activity

YOKOMIZO Shuji
ITO Shinsuke

海外では、イタリア統計局やドイツ連邦統計局といった統計作成部局が、事業所・企業系の匿名化ミクロデータの作成・提供を行ってきた。それに対して、わが国では、事業所・企業系の統計調査に関する匿名データの作成は現状では実現していない。しかしながら、事業所・企業系の匿名化ミクロデータについては、学術研究だけでなく高等教育のための利用に対するニーズが期待される。そこで本稿では、「経済センサス - 活動調査」の個票データを用いて、製造業を対象にした上で、ミクロアグリゲーション等の匿名化措置の有効性の定量的な評価を行った。本研究では、各種の匿名化措置が施されたミクロデータを対象に、R-U マップによる有用性と秘匿性の検証を行うだけでなく、相対的な露見リスクについても探索的に追究した。本分析結果から、事業所・企業系のミクロデータの特定化のリスクを低減するために、歪みを持つ分布特性や特異値を考慮した上での匿名化措置が必要なことが実証的に確認できた。

キーワード　事業所・企業系のミクロデータ、匿名化措置、ミクロアグリゲーション

Several National Statistical Institutions including the Italian National Institute of Statistics and the German Federal Statistical Office are creating and releasing anonymized business microdata. On the contrary, Japan has not yet released anonymized microdata for business statistical research despite expected benefits towards both research and education. This paper conducts an empirical assessment of the effectiveness of anonymization methods based on individual data from the “Economic Census for Business Activity” with particular focus on the manufacturing industry. Using R-U confidentiality maps, we compare data usability and data confidentiality for business microdata created using different kinds of anonymization methods, and conduct an exploratory investigation into relative disclosure risk. The results of this empirical analysis establish that application of anonymization methods which reflect the characteristics of business microdata such as skewed distribution and outliers is necessary in order to reduce the disclosure risk for such microdata.

Keywords: Business microdata, Anonymization methods, Microaggregation

---

† 総務省統計局統計データ利活用センター
‡ 中央大学経済学部　Email: ssitoh@tamacc.chuo-u.ac.jp

1. はじめに

わが国では、国勢調査、全国消費実態調査、社会生活基本調査、就業構造基本調査、住宅・土地統計調査、労働力調査、国民生活基礎調査といった 7 種類の世帯・人口系の統計調査が匿名データとして作成・提供されているが、事業所・企業系の統計調査の匿名データは提供されていないことが知られている。

それに対して、海外では、事業所・企業系の統計調査の匿名化ミクロデータ(anonymized microdata)が作成されているが、それは、Eurostat(欧州統計局)、イタリア、ドイツなどにおける作成事例に限られる。そして、事業所・企業系の統計調査のミクロデータの提供は、欧米諸国においてはオンサイト利用やリモートアクセスによるデータの利用サービスによって近年展開されてきた(伊藤(2018))。

しかしながら、事業所・企業系の匿名化ミクロデータの利用目的に関して、学術研究に対する利用や高等教育のための活用が考えられる。具体的には、学術研究目的の利用においては、研究者が匿名化ミクロデータを用いて、探索的な研究を行うことが想定される。事業所・企業系の匿名データが作成・提供されれば、公的統計ミクロデータの利用の起点として、匿名データが利用され、それがオンサイト利用の促進にもつながることから、公的統計の二次的利用のさらなる推進を図ることも期待できる。

高等教育目的での利活用については、わが国では、世帯・人口系の調査である全国消費実態調査や就業構造基本調査の一般用ミクロデータが作成・利用されているが、事業所・企業系の統計調査において、一般に利用可能な公開型ミクロデータは、作成の対象となっていない。近年、統計教育の重要性が叫ばれているが、教育目的を指向した事業所・企業系の匿名化ミクロデータは、その一助になると考えられる。

2018 年の「公的統計の整備に関する基本的な計画(第Ⅲ期基本計画)」では、厚生労働省所管の「賃金構造基本統計調査」の匿名データの提供可能性を指摘している。具体的には、社会・経済情勢の変化を的確に捉える統計の整備の一環として、働き方の変化等をより的確に捉える統計の整備の必要性も指摘されている。このことから、労働供給サイドだけでなく、労働需要側の観点から見た実証分析を行うために、事業所・企業系の匿名データの作成方法に関する追究の方向性が示唆されている。

このように、わが国においても、事業所・企業系の匿名化ミクロデータのニーズは存在すると考えられる。こうした状況を踏まえて、海外における匿名化ミクロデータの作成状況も考慮した上で、わが国における代表的な事業所・企業系の統計調査である「経済センサス - 活動調査（以下、「経済センサス」と略称)」を例に、匿名化ミクロデータの作成可能性を検討する。具体的には、経済センサスの個票データを対象に、非攪乱的手法だけでなく、攪乱的手法の適用可能性を追究する。さらに、試行的に作成した各種の匿名化ミクロデータに含まれる質的属性と量的属性のそれぞれに対して、秘匿性や有用性、またそれらを共に考慮した定量的な評価を行い、その結果を可視化する。

2. 事業所・企業系の匿名化ミクロデータの作成状況

本節では最初に、事業所・企業系の匿名化ミクロデータの特徴を明らかにした上で、海外における事業所・企業系の匿名化ミクロデータの作成状況について概観する。

2.1　事業所・企業系の匿名化ミクロデータの特徴

表 1 は、個人・世帯を対象にしたミクロデータと、事業所・企業に関するミクロデータの

表 1 個人・世帯に関するデータの特徴と企業に関するデータの特徴

	個人・世帯に関する ミクロデータ	企業に関する ミクロデータ
レコード数	多い	少ない
母集団となる個体が 標本として抽出される可能性	特定の個人が含まれる 確率は低い	**大規模企業**は常に含まれる 中規模企業はしばしば含まれる 小規模企業が含まれる確率は低い
属性の数	多い	少ない
属性の種類	ほとんどが質的変数	ほとんどが量的変数
属性の分布	-	分布特性の歪みが大きい 変数間の相関性が高い
外れ値	稀	ほとんどの属性で **大規模企業**は外れ値
精度の高い外部情報の取得	難	易 （**大規模企業**は公表義務あり）
露見（disclosure）に 伴うリスク	小	大

出所　O'Keefe and Shlomo (2014)、 Lenz *et al.* (2006)

特徴を整理したものである。世帯・人口系のデータと比較した場合、事業所・企業系のデータの特性としては、①サンプルサイズが小さいこと、②大企業の場合、標本抽出の対象として、悉皆で捕捉される調査客体も少なくないことから、母集団となる個体が標本として抽出される可能性が非常に高くなること、そして③大企業における属性値のほとんどが外れ値であることが指摘できる。したがって、事業所・企業系の匿名化ミクロデータの作成においては、大企業に含まれる属性情報に対してどのような秘匿措置を施すかが、論点になると言える。

また、世帯・人口系のデータと比較して、事業所・企業系のデータに関しては、秘匿性だけでなく、有用性の確保も難しいことが指摘されている。事業所・企業系のデータについては、母集団が相対的に小さく、量的変数についても歪みがあるため、特異値となる個体をより多く含む分布特性を備えていることから、調査客体が標本一意である場合に母集団一意にもなる可能性が高くなる。また、侵入者(データの利用者、intruder)にとっては、上場企業の有価証券報告書のような精度の高い外部情報が入手可能であることから(Lenz *et al.*(2006), Franconi and Ichim(2007), 星野(2010), Ritchie *et al.*(2019))、こうした外部情報とのマッチングを通じて、ミクロデータの提供によって個体が特定される危険性を表す露見リスク(開示リスク、disclosure risk)(竹村(2004), 伊藤(2010))が高まることが指摘できる。なお、ミクロデータに対する匿名化技法の適用はこうした露見リスクに影響を与えることから、原データと各種の匿名化技法が適用されたミクロデータとのリンケージによって、秘匿性の相対的な程度(秘匿性の強度)を定量的に評価することも考えられる(伊藤(2010))。

2.2 事業所・企業系の匿名化ミクロデータの作成事例―イタリアとドイツ

ISTAT(イタリア国立統計研究所)は、企業のイノベーション活動の調査であるCIS(=Community Innovation Survey)の匿名化ミクロデータを提供している。CISはEU内での比較可能性を考慮した標本調査であり、主な変数として、経済活動(産業分類)、地理的区分、従業員数、売上高、研究費等が捕捉される。CISについて、学術研究用ファイル(Scientific Use File = SUF)や一般公開型ファイル(Public Use File = PUF)が現在作成・提供されている。Ichim(2007)は、1998年から2000年の間に調査されたCIS3を例に、SUFの作成手順を体系的に示している。それによれば、以下のステップを踏むことが推奨されている(図 1)。

図 1 CIS の SUF 作成手順

出所　Ichim(2007)

第 1 の手順は、露見に関するシナリオを設定することである。具体的には、CIS の場合、外部参照情報(external register)に含まれる産業分類、地域、従業員数、売上高といった識別情報に基づくリンケージや大規模な売上高等からの偶発的な個体特定 (spontaneous identification)が行われることが露見シナリオとして想定されている。第 2 の手順は、変数の前処理として、産業分類、地域、従業員数といったキー変数に対してグローバルリコーディングを行うことである。第 3 の手順は、キー変数で層化を行った上で、リスクの高いレコードの特定を行うことである。その場合、類似したレコードが少なければ露見リスクが大きいとみなした上で、密度ベースのアルゴリズム[1]を用いて、攪乱の対象となるレコードを選定する。第 4 の手順は、ミクロデータに含まれる変数値に攪乱を行うことである。事業所・企業系のミクロデータについては、例えば、リスクが高いと判断された売上高の外れ値に対しては最近傍のクラスターからの補完(the nearest clustered unit imputation)や、分布の右裾については個別ランキング法によるミクロアグリゲーション(microaggregation)[2]が適用される。第 5 の手順は、情報量損失と情報量保護に関する指標を設定することである。具体的には、産業分類ごとの、売上高の分散の変化率や変数間の相関係数が考慮される。そして第 6 の手順は、公開するミクロデータファイルの説明に関する資料を作成し、各変数について攪乱の有無に関する情報やイノベーション変数の比率や売上高の変化率等、データの有用性に関する指標を資料の中で明示することである。

ドイツの連邦統計局でも、事業所・企業を対象とした複数の匿名化ミクロデータが、PUF や SUF、さらには高等教育用で匿名化の強度が高い campus file(CF)の形で提供されている。ド

---

[1] 1998～2000 年に調査された CIS3 では、密度ベースのクラスタリングアルゴリズムの一種である DBSCAN(Density-based spatial clustering of applications with noise)(Ester(1996))が、2002～2004 年の CIS4 では、密度ベースの外れ値検出アルゴリズムである局所外れ値因子法（local outlier factor = LOF）(Breunig *et al.*(2000))がそれぞれ検討された。

[2] ミクロアグリゲーションとは、「ミクロデータ(個別データ)を k 個(k は閾値(threshold))のレコードを有する同質的なレコード群にグループ化した上で、そのレコードにおける個々の属性値を平均値等の代表値に置き換える」匿名化技法である(Domingo-Ferrer and Mateo-Sanz (2002, p.190), 伊藤(2009, p.201))。ミクロアグリゲーションの手法として、個別ランキング法や MDAV(maximum distance to average vector)法等が存在する。個別ランキング法とは、量的属性ごとにソート化とグループ化を行った上で、グループ内の量的属性値の各々を平均値に置き換える方法である(伊藤(2009, p.204))。それに対して、MDAV 法とは、Domingo-Ferrer and Mateo-Sanz(2002)で議論された多変量固定サイズのミクロアグリゲーションに基づいて、Hundepool *et al.*(2003)で実装されたアルゴリズムを用いる手法であって(Domingo-Ferrer and Torra(2005))、複数の量的属性の平均ベクトルを求めた上で、探索的にグルーピングが行われる。

イツの匿名化には、事実上の匿名性(factual anonymity)という概念が存在する。これは、「著しく大きな時間、経費および労力の支出によって、当事者に関連づけることができない」こと(濱砂(1999))であり、連邦統計法に規定されている、学術研究用ファイル（SUF）を作成する上での重要な概念である(伊藤(2020))。

ドイツでは、2000年代に、研究者に対して事業所・企業系のミクロデータの利用の促進を図ることを目的とした2つのプロジェクトが行われた。第1は、「企業ミクロデータに関する事実上の匿名化」プロジェクト(Factual Anonymisation of Business Microdata, 2002～2005年)である(Lenz *et al.*(2006))。本プロジェクトでは、SUF作成における匿名化手法の適用可能性が検討され、ミクロアグリゲーションやノイズ付与[3]等の有効性に関する検証がなされた。検証結果によれば、情報量損失の観点からは、ミクロアグリゲーションの中でも特に個別ランキング法(individual ranking methods)がSUFの作成に適していると判断された。

第2は、「企業パネルデータに関する事実上の匿名化」プロジェクト(Business Statistics Panel Data and Factual Anonymisation, 2006～2008年)である(Brandt *et al.*(2008))。本プロジェクトでは、匿名化に関する研究実績のある年次ベースの事業所・企業のデータを縦断的にリンケージするパネルデータの作成が指向された。匿名化手法としては、ミクロアグリゲーション、乗法ノイズ、多重代入法(multiple imputation)が検討された。また、ドイツの賃金構造に関する統計調査であるGerman Cost Structure Survey(1999-2002)を用いた検証実験では、記述統計量や属性値、相関係数を用いた有用性の評価や、リンケージ技法を用いた秘匿性の強度の確認が行われた。

イタリア・ドイツの事例で注目すべき点としては、以下のように整理できる。露見シナリオについては、SUF作成を前提に、偶発的な個体特定や外部情報を用いたマッチングに重点が置かれている。露見シナリオを考慮しつつ、秘匿性の定量的な評価指標と有用性に関する指標を踏まえた上で、最小限の攪乱的措置を抑えていることも特徴的である。そのための匿名化手法としては、グローバルリコーディングといった非攪乱的手法だけでなく、ミクロアグリゲーション等の攪乱的手法が採用されている。ミクロアグリゲーションの中でも特に、元データとの近似性が相対的に高い個別ランキング法がSUF作成に適していることが明らかになっている。今後、わが国で事業所・企業の匿名化ミクロデータの作成を検討する上で、海外の事例は参考になると考えられる。

3. 経済センサスのミクロデータを用いた秘匿性と有用性の評価研究

本節では、前節におけるイタリアやドイツにおける事業所・企業系のデータの匿名化の作成状況を参考にしつつ、基幹統計のひとつである経済センサスをもとに、各種の匿名化技法を用いて試行的に作成したミクロデータを用いて実証研究を行う。本研究においては、産業大分類E（製造業）の事業所レコードについて、諸条件を満たす約36万レコードの中から、無作為抽出した10,000レコードをテストデータとして使用した。秘匿性の観点から従業者合計(男女計)が1人以上1,000人未満のレコードを対象とした。また、製造業経理項目が集計対象外であることから個人票（01票）のレコードは対象外としている。分析対象となる項目には、集計結果表で用いられている集計事項を中心に、属性情報の入手可能性等を考慮し

[3] ノイズ付与とは、量的属性に対してランダムなノイズを付加することで攪乱を行う手法である(Duncan & Pearson(1991))。加法ノイズ(additive noise)や乗法ノイズ(multiplicative noise)が存在する。なお、加法ノイズの数理的な特徴については、伊藤他(2014)を参照。

て外部参照情報になりうるキー変数、露見リスクが大きいと考えられるセンシティブな属性、匿名化ミクロデータとして分析上有用と思われる属性を以下のように選定した(表 2)。なお、属性によっては欠測値を含むこともあるが、補完処理については行っていない。

3.1 記述統計量および分布特性

本研究では、最初に選定した調査項目について記述統計量や分布特性を確認した。表 3 は、量的属性の記述統計量を示したものである。未記入または不詳の事業所は除外されている。また、製造業の一部の属性には、従業者数に応じた記入条件が存在していることから、実験条件における記入率を参考情報として掲載している。

表より、売上(収入)金額、現金給与合計、付加価値額といった経理項目については、平均値と中央値との間に大きな差が生じていることがわかる。すなわち、分布に大きな歪みがあると言える。資本金額についてはその傾向がより顕著である。なお、属性によっては記入条件が存在しており、従業者数の要件を満たさないものは未記入となる。そのため、実験条件を満たすレコードの中でも、属性によって記入率が異なることに注意が必要である。レコードの抽出条件や属性ごとの記入条件の影響は、4 節で述べる。

図 2 は、対数化した売上(収入)金額のヒストグラムを示している。秘匿上の観点から目盛りを省略している。量的属性は概ね対数正規分布に従っていることが確認できる。図 3 は、量的属性について相関係数行列を算出した結果を示している。従業者合計と現金給与合計の間には 0.91、売上（収入）金額と原材料使用額等の間には 0.95 と非常に強い相関が存在している。その他の経理項目も 0.5 前後の相関関係を有している。このことから、従業者合計と一部の経理項目や、主要な経理項目同士には比較的強い相関があることが確認できる。

表 2 分析対象項目

項目名	備考
都道府県	
従業者合計	
資本金額	
産業中分類	
売上（収入）金額	
現金給与合計	
原材料使用額等	
年末在庫合計	従業者29人以下の事業所は記入対象外
付加価値額	
有形固定資産年末現在高	除従業者9人以下の事業所は記入対象外
（有形固定資産）投資総額	除従業者29人以下の事業所は記入対象外

表 3 分析対象項目における量的属性の要約統計量［万円］

	事業所数	平均値	中央値	標準偏差	1%点	99%点	実験条件記入率
売上（収入）金額	246,933	94,706	9,936	866,557	182	1,501,152	100.00%
従業者合計	246,933	26	9	64	1	318	100.00%
資本金額	237,615	55,079	1,000	804,010	100	969,823	96.23%
現金給与合計	244,924	11,064	2,680	35,740	77	156,936	99.19%
原材料使用額等	246,070	57,389	3,428	628,886	26	933,355	99.65%
年末在庫合計	79,407	26,557	2,826	157,587	16	407,848	32.16%
付加価値額	246,727	30,937	4,899	204,937	-886	471,031	99.92%
有形固定資産年末現在高	102,834	57,993	11,364	261,588	-1,200	784,961	41.64%
（有形固定資産）投資総額	32,050	20,739	3,673	84,605	6	289,841	12.98%

図 2　対数化した売上(収入)金額の分布

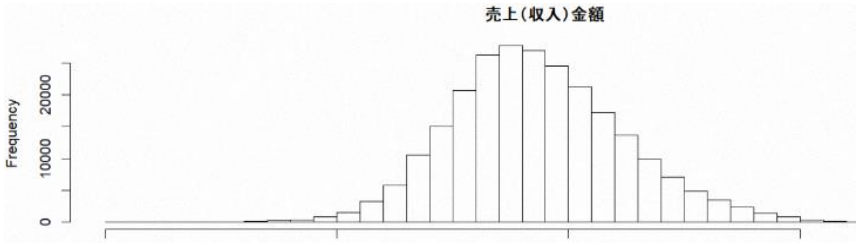


図 3 量的変数における相関係数行列

	従業者合計	資本金額	売上（収入）金額	現金給与合計	原材料使用額等	年末在庫合計	付加価値額	年末現在高有形固定資産	投資総額（有形固定資産）
従業者合計	1.00	0.29	0.40	0.91	0.36	0.38	0.48	0.47	0.37
資本金額		1.00	0.30	0.39	0.27	0.31	0.27	0.37	0.25
売上（収入）金額			1.00	0.46	0.95	0.68	0.55	0.62	0.41
現金給与合計				1.00	0.41	0.44	0.52	0.53	0.42
原材料使用額等					1.00	0.69	0.32	0.60	0.38
年末在庫合計						1.00	0.38	0.56	0.40
付加価値額							1.00	0.35	0.33
有形固定資産年末現在高								1.00	0.55
（有形固定資産）投資総額									1.00

3.2 質的属性の匿名化

本節では、分布特性をもとに、質的属性の匿名化に関する実証研究を行う。匿名化手法としては、イタリアやドイツの事業所・企業系の匿名化ミクロデータの作成において適用され、わが国の匿名データの作成において適用されているリコーディング(区分統合)を用いた。外部参照情報からの個体特定のリスクや分析上の有用性を考慮し、本研究では地域、産業、従業者合計、資本金額の４属性を中心に、リコーディングに基づく匿名化を行った。なお、これらは経済センサスにおける結果表においても分類区分の統合が行われて

表 4 本研究で用いた変数におけるリコーディングのタイプ

項目	分類	内訳
地域	8区分	北海道, 東北, 関東, 中部, 近畿, 中国, 四国, 九州・沖縄
	3区分	東日本, 中日本, 西日本
産業分類	24区分	09~32
	11区分	09_10, 11, 12_13_14, 15, 16_17_18_19, 20_32, 21, 22_23_24, 25_26_27, 28_29_30, 31
従業者規模	13区分	1, 2, 3, 4, 5~9, 10~19, 20~29, 30~49, 50~99, 100~199, 200~299, 300~499, 500~999, 1000人~
	5区分	1~4, 5~9, 10~29, 30~99, 100~
資本金規模	11区分	~300万, 300万~500万, 500万~1000万, 1000万~3000万, 3000万~5000万, 5000万~1億, 1億~3億, 3億~10億, 10億~50億, 50億~, 以外
	5区分	~1000万, 1000万~1億, 1億~10億, 10億~, 以外

※「~300 万円」は「~300 万円未満」を表す。以降の図表も同様。

いる属性でもあることから、その分類区分を参考にした上で、リコーディングの程度が異なる複数のリコーディングの組み合わせを設定した。原則として、特定の分類区分の構成比が小さくなりすぎないように配慮している。

表 4 は、本研究で用いた変数におけるリコーディングのタイプを示している。地域については、都道府県の地域区分をもとに、地域ブロック 8 区分(北海道、東北、関東、中部、近畿、中国、四国、九州・沖縄)と 3 区分(東日本、中日本、西日本)を採用した。産業については、製造業における産業中分類をもとにリコーディングを行った。平成 19 年就業構造基本調査の匿名データでは、製造業における産業中分類が、原区分をより粗くした区分になっているため、これを参考に 24 区分から 11 区分への統合を行っている。従業者合計については、結果表における 15 区分に基づいて、13 区分(出向・派遣従業者のみおよび従業者数 1,000 人以上は実験条件で除外している)に区分統合した上で、従業者規模とした。従業者規模については、より粗い区分で統合した 5 区分もさらに設定した。資本金額については、結果表における 10 区分を参考に、「以外(未記入または不詳)」を含めた 11 区分を資本金階級として採用した。また、それをより粗い形で区分統合した 5 区分も新たに設定した。

3.3 質的属性の秘匿性と有用性の定量的評価

伊藤他(2014)では、質的属性の秘匿性の評価にあたって、クロス集計表による評価方法についての議論が展開されている。それは、データに含まれる複数の質的属性を対象に、クロス集計表における分布特性を比較することによって、秘匿性の強度を評価する手法である。具体的には、原データと匿名化ミクロデータの間で度数が 1 となるセルの総数を比較し、度数 1 となるセル数の変化の確認を行っている。

本実験では、地域、産業、従業者規模、資本金階級の 4 項目を用いて、その組み合わせによって事業所数の度数 1 または 2 となるレコード数(事業所数)がどのように変化するかを確認した。経済センサスの結果表では、事業所数 1 または 2 の場合に一次秘匿の対象となり、売上(収入)金額等の経理項目が秘匿されるため、本研究でもその基準に基づいている。度数 1 または 2 となるレコード数を確認することは、k-匿名性[4]の概念に基づいて地域、産業、従業者規模、資本金階級の 4 属性で形成された層ごとに 3-匿名性を満たさない(以下、「3-匿名性違反」と呼称)レコードの数を確認することと同等であると考えられる。表 5 は、地域 3 区分、産業 11 区分、従業者規模 13 区分と資本金階級 5 区分の条件で層別に事業所数を計測した場合のイメージを示している。強調されたセルに該当する事業所が 3-匿名性に違反するリスクの高い事業所と判定される。このような事業所の数に焦点を当てた上で、本実験を行う。

図 4 は、分類区分を変更したキー変数の組み合わせ別の 3-匿名性違反のレコード数の割合を示している。図 4 の下表は、各 index における キー変数の組み合わせを示す（以下の図でも同様)。index1 は、最も細かい分類区分を用いているため、結果として層の種類は最も多くなる。なお、計算上は 8×24×13×11＝27,456 通りの層が存在することになるが、実際に事業所の存在しない層も多数存在する。層の数が増えるほどひとつの組み合わせに含まれる事業所数は少なくなる一方で、index 16 は最も分類区分が粗く、層の数が少ないことから、3-匿名性違反のレコード数は比較的小さな値となる。図 4 からも明らかなように、全体を通じて、キー変数のリコーディングが粗くなるほど秘匿性が強くなる傾向が明確である。

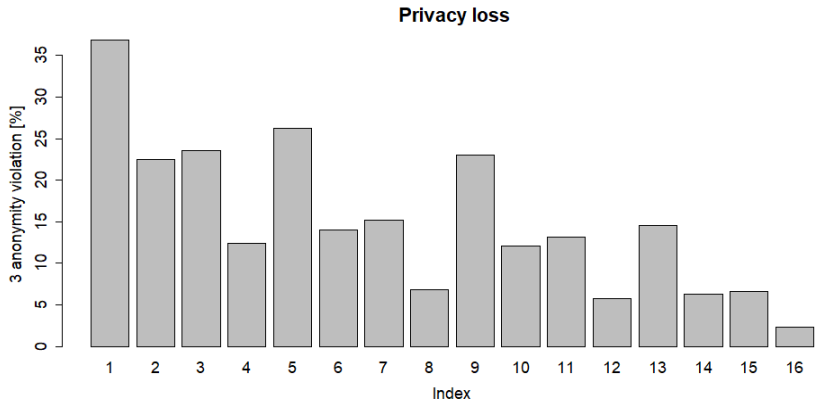
---

[4] 提供対象となっているデータが、準識別子（氏名、住所、電話番号といった明示的な識別情報以外でも再識別を可能にする属性）のすべての組み合わせによって、少なくとも k 個の個体を識別することができない場合、k-匿名性を持つと定義される(Samarati and Sweeney(1998))。

表 5 層別の事業所数のイメージ

地域	産業	従業者規模	資本金階級	事業所数
1東日本	09_10	1~4人	1,000万円未満	12
1東日本	09_10	1~4人	1,000万円～1億円未満	56
1東日本	09_10	1~4人	1～10億円未満	1
1東日本	09_10	1~4人	10億円以上	0
1東日本	09_10	1~4人	以外	16
1東日本	09_10	5~10人	1,000万円未満	23
1東日本	09_10	5~10人	1,000万円～1億円未満	42
1東日本	09_10	5~10人	1～10億円未満	0
1東日本	09_10	5~10人	10億円以上	0
1東日本	09_10	5~10人	以外	21
⋮	⋮	⋮	⋮	⋮
3西日本	31	100~999人	1,000万円未満	7
3西日本	31	100~999人	1,000万円～1億円未満	28
3西日本	31	100~999人	1～10億円未満	2
3西日本	31	100~999人	10億円以上	0
3西日本	31	100~999人	以外	8

図 4 質的属性の秘匿性評価(3-匿名性違反のレコード数の割合)



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
地域	8区分	8	8	8	8	8	8	8	3	3	3	3	3	3	3	3
産業	24区分	24	24	24	11	11	11	11	24	24	24	24	11	11	11	11
従業者規模	13区分	13	5	5	13	13	5	5	13	13	5	5	13	13	5	5
資本金階級	11区分	5	11	5	11	5	11	5	11	5	11	5	11	5	11	5

注　表に示される index（インデックス番号）とそれに対応する各属性における分類区分の数は、地域、産業、従業者規模と資本金階級を対象に、区分統合された分類区分数の組み合わせの一覧を示している。以下同様。

なお、本実験では 10,000 レコードを対象としているが、レコード数によって 3-匿名性違反のレコード数の割合は大きく変化しうることに注意が必要である。予備的に行った実験では、サンプリング前の全レコードを使用すると 3-匿名性違反のレコード数は多くの index で 3-匿名性違反のレコード数の割合は 1%を切った。統計実務上の観点では、標本の大きさを考慮してキー変数のリコーディングを考える必要があると考えられる。

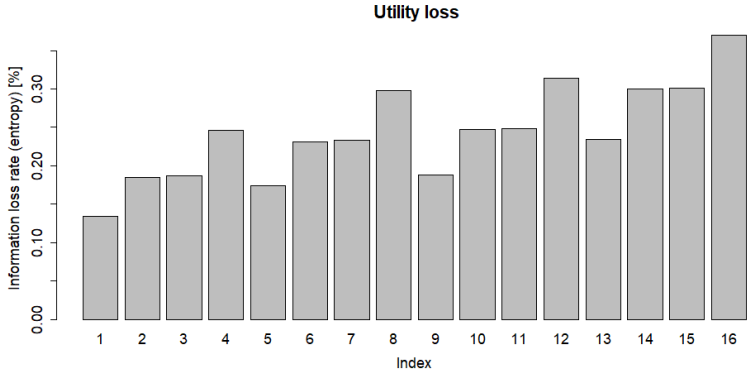
伊藤他(2014)では、質的属性の有用性評価手法のひとつとして、情報エントロピーに基づいて情報量損失を計測する手法について検討が行われた。稀少な状態が生じたことを表す情

報(確率の低い情報)ほど大きくなるシャノン情報量の期待値である情報エントロピーを求めることで、リコーディングの前後によって変化する質的属性の有用性を評価することが可能である。匿名化技法の適用によって属性値が変化する推移確率(transition probability)を用いて情報エントロピー[5]を算出した上で、情報エントロピーが計測された対象となるリコーディング前のレコード数を乗じることによって、情報量損失が求められる。情報量損失をその最大値で除することで、情報量損失率を算出する[6]。図５は、情報エントロピーに基づく情報量損失率を示している。キー変数に対するリコーディングが粗くなるほど、情報量損失率が増加していることを確認することができる。

以上の結果を踏まえて、質的属性を対象に、秘匿性と有用性をもとに R-U マップ(R-U confidentiality map)(Duncan and Pearson(2001))を作成した。横軸が Risk(秘匿性)を、縦軸が Utility(有用性)を表している。具体的には、秘匿性には総レコード数に占める 3-匿名性違反のレコード数の割合を、有用性には情報エントロピーに基づく情報量損失率を用いた。図６から、秘匿性が増大するほど有用性が相対的に低下するトレードオフの関係にあることがわかる。例えば、最も細かい分類区分の組み合わせである index 1 は、図中右下に位置しており、秘匿性は低く、有用性は高い領域に位置している。それに対して、最も粗い分類区分の組み合わせである index 16 は、図中左上に位置しており、秘匿性は高く、有用性は低い位置に存在している。

図中で左下の領域にプロットされた匿名化技法の組み合わせの場合、データの有用性と秘匿性のいずれも相対的に高くなることから、匿名化ミクロデータとしてはより望ましいと判断できる。ただし、秘匿性と有用性はトレードオフの関係にあるため、統計実務の観点から見れば、許容できる秘匿性の範囲で有用性を最大にする index を選択するのが望ましいと言える。

図 5 質的属性の有用性評価(情報エントロピーに基づく情報量損失率)



5

$$情報エントロピー = -\sum_{i=1}^{n} p_i \log p_i$$

6

$$情報量損失率 = -\frac{情報量損失}{情報量損失の最大値} \times 100$$

なお、ここでは、各属性の全ての区分を１つに統合したときに算出される情報量損失の値を仮想的な情報量損失の最大値と設定している。

図 6 質的属性の R-U マップ

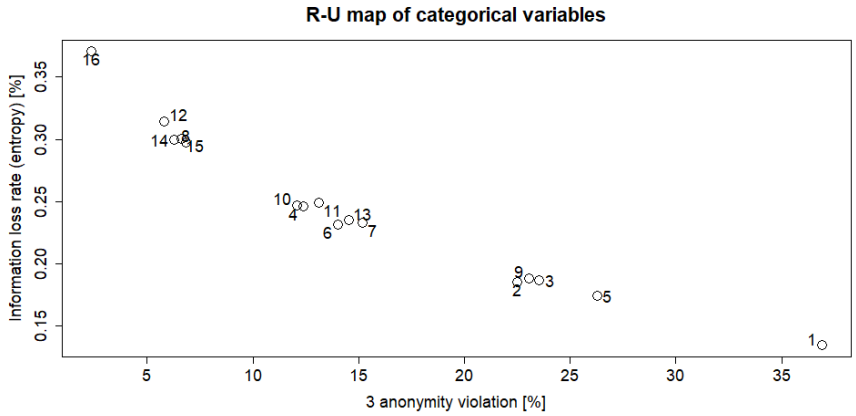


表 6 ある分類区分の組み合わせの売上(収入)金額の要約統計量［万円］

	事業所数	平均値	標準偏差	中央値	歪度	尖度	標準誤差	1%点	99%点
総数	293	1,300.31	8,820.79	500.00	16.60	278.11	515.32	11.00	4,775.60
上位5%	15	13,912.60	37,914.63	3,358.00	3.11	8.31	9,789.52	2,613.24	131,261.08
上位5%以外	278	619.79	535.43	457.50	1.26	1.32	32.11	11.00	2,277.43

3.4 量的属性の匿名化の検討

本節では、量的属性のうち、センシティブな属性である売上（収入）金額、現金給与合計、原材料使用額等、年末在庫合計額、付加価値額、有形固定資産年末現在高、投資総額の匿名化を検討する。表 6 は、ある分類区分の組み合わせにおける、売上(収入)金額の総数、上位5%の事業所、上位 5%以外の事業所のそれぞれについての実数値の要約統計量を示したものである。上位 5%では平均が 13,912.60 に対し、上位 5%以外では 619.79 と、上位 5%に大きく分布が偏っていることが読み取れる。一般に、平均値よりも分布の歪みの影響を受けにくいとされる中央値においても、上位 5%が 3,358.00 に対して上位 5%以外は 457.50 と一定の歪みが見られる。匿名化にあたって、上位 5%のような露見リスクの大きい事業所を削除する非攪乱的な手法も考えられるが、これらの分布特性を考慮すると、レコード削除によって生じる分布への影響は無視できない。そのため、分布の右裾の事業所については安易にレコード削除を行うのではなく、平均値等の統計量を維持できる攪乱的手法が適切であると考えられる。

そこで、センシティブや量的属性に対する匿名化技法として、イタリアやドイツにおける匿名化ミクロデータ作成の実務においても採用実績のあるミクロアグリゲーションを適用した。本実験では、閾値は経済センサス結果表の一次秘匿の基準に揃えて 3-匿名性を確保し、代表値には平均値を採用した。ミクロアグリゲーションの手法には、近年研究例が多く、匿名化ツール sdcMicro(Templ *et al.*(2015))でも microaggregation コマンドのデフォルトの手法になっている MDAV 法を選択した。

3.5 量的属性の秘匿性と有用性の定量的評価

量的属性の秘匿性評価については、伊藤他(2014)を参考に、距離計測型リンケージを用いた。距離計測型リンケージは、原データと匿名化ミクロデータにおけるレコード同士の距離を計算し、その距離の大きさに基づいて、２つのデータが対応付け可能かを判定する方法である(伊藤(2010))。本実験では、最初に、匿名化ミクロデータのレコードから原データの各レコードへのユークリッド距離を計測し、次に、最も距離が短くなるレコードが、原データ

の元のレコードに一致し、かつ同じ距離となるレコードが他に存在しない場合に、そのレコードは真のリンクであると判定した。リンケージを行うためのリンクキー変数としては、ミクロアグリゲーションによって攪乱される売上（収入）金額、現金給与合計、原材料使用額等、年末在庫合計額、付加価値額、有形固定資産年末現在高、投資額の７つのセンシティブな量的属性を用いた。距離計測型リンケージで使用する距離には、水準やばらつきをそろえるために属性値を標準化したユークリッド距離を選択した。この条件のもとで、原データから最も距離の近い攪乱済みのレコードが真のリンクである確率(true link rate)を求めた。

図 7 は、MDAV 法を例に、距離計測型リンケージに基づく真のリンクとなる比率を示したものである。いずれもミクロアグリゲーションを行う際のキー変数の分類区分が細かくなるほど、真のリンクとなる比率が増加する傾向にあることがわかる。

さらに、量的属性の有用性評価を行った。ミクロデータに含まれる量的属性に対して有用性の相対的な程度を評価する手法として、統計指標を用いた有用性の評価が考えられる。原データと匿名化ミクロデータについて、属性値の差、分散共分散行列、相関係数行列に見られるデータ構造の変化によって情報量損失の計測を行うことができる(伊藤他(2014))。情報量損失の大きさについては、平均絶対誤差(mean absolute error)や平均変化率(mean variation)といった尺度が選択されるが、ミクロアグリゲーションを行う際のキー変数の分類区分が粗くなるほど平均絶対誤差や平均変化率が大きくなり、原データの性質が失われていることが確認されている。本研究では、匿名化ツール sdcMicro の dUtility コマンドにおける IL1s[7] (Yancey *et al.*(2002))を使用した。IL1s は、平均変化率を求めるにあたって、分母の値に原データの属性値ではなく、原データの属性ごとの標準偏差を用いる評価指標である。複数の属性、レコードに対して、攪乱の前後でどの程度属性値の変化があったかを定量的に評価することができる。

MDAV 法を対象に、IL1s を用いて評価した有用性評価の結果が図８である。図 7 と比較すると、キー変数における分類区分のリコーディングの組み合わせについては、図 7 で秘匿性の強度が高い組み合わせが IL1s については概ね反対の傾向を示していることが確認できる。

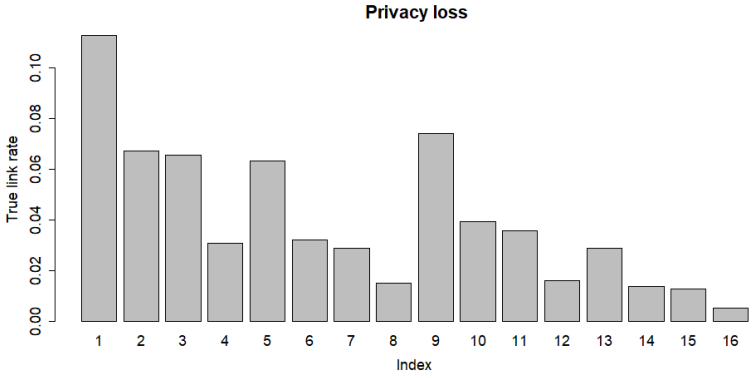
以上の結果を踏まえて、量的属性についても、秘匿性と有用性をもとに R-U マップを作成した(図 9)。横軸が秘匿性として距離計測型リンケージによる true link rate を、縦軸には有用性として IL1s に基づく情報量損失率を用いた。本 R-U マップより、秘匿性が増大するほど有用性が低下するトレードオフの関係にあることがわかる。最も細かい分類区分の組み合わせである index 1 は、図中右下の秘匿性は低く、有用性は高い位置に存在している。一方、最も粗い分類区分の組み合わせである index 15 や 16 は、図中左上の秘匿性は高く、有用性は低い位置に存在している。図中で左下の領域にある index ほど秘匿性と有用性のバランスが図られているが、本実験の結果では該当する index は存在しない。このような R-U

7

$$IL1s = \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\frac{|x_{ij}-y_{ij}|}{\sqrt{2}S_j}.$$

m: 属性数
n: レコード数
x_{ij}: レコード i 属性 j の攪乱前の属性値
y_{ij}: レコード i 属性 j の攪乱後の属性値
S_j: 攪乱前の属性 j の標準偏差

図 7 量的属性の秘匿性評価(距離計測型リンケージに基づく true link rate)、MDAV 法



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
地域	8区分	8	8	8	8	8	8	8	3	3	3	3	3	3	3	3
産業	24区分	24	24	24	11	11	11	11	24	24	24	24	11	11	11	11
従業者規模	13区分	13	5	5	13	13	5	5	13	13	5	5	13	13	5	5
資本金階級	11区分	5	11	5	11	5	11	5	11	5	11	5	11	5	11	5

図 8 量的属性の有用性評価(IL1s)、MDAV 法

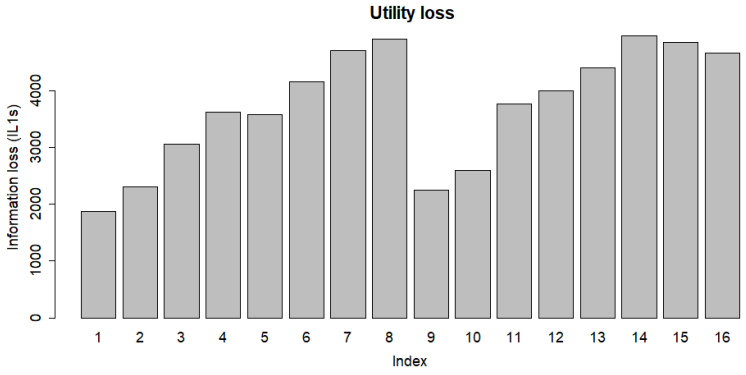
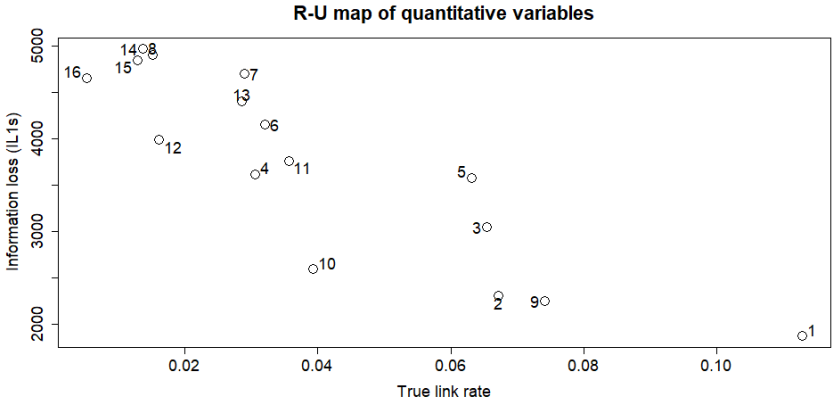


図 9 量的属性の R-U マップ



マップに示される秘匿性と有用性のバランスを図りながら、適切なしきい値に基づいてリコーディングにおける区分統合の程度やミクロアグリゲーション技法のような攪乱的手法の適用に関する選択について議論することが、統計実務上も有益であると思われる。

4. 経済センサスにおける事業所の分布特性の把握と露見リスクの評価に関する探索的な検証

前節では、経済センサスのミクロデータを対象に、先行研究に基づいた匿名化処理の有効性を検証するだけでなく、秘匿性と有用性に関する評価手法を適用することによって、匿名化ミクロデータの特性を明らかにした。本節では、経済センサスのデータ特性を踏まえた上で、露見リスクの評価方法をさらに追究するために、経済センサスのミクロデータを用いて、露見リスクの定量的な評価に関する探索的な実験を行う。

4.1 経済センサスにおける事業所の分布特性

匿名化ミクロデータの作成を行う上で、露見リスクを最小限に抑えることは不可欠である。露見リスクについては、偶発的な個体識別だけでなく、外部参照情報とのマッチングの可能性が存在する。後者を考える際、どのような外部参照情報が存在するのか、そのすべてを検討することは統計実務面から見ても現実的な困難だと言える。そのため、前節では、外部参照情報とのマッチングキーとして、特に重要であると考えられる地域、産業、従業者規模、資本金階級の 4 属性をキー変数として実験を行った。しかしながら、その他にも、売上（収入）金額、現金給与合計、原材料使用額等、年末在庫合計額、付加価値額、有形固定資産年末現在高、投資総額といったセンシティブな経理項目も、それ自体が準識別子として外部参照情報に用いられることによって、露見リスクを高めるケースもある。

前節ではリコーディングを行うにあたって、ミクロデータを用いた実証研究を行う場合の有用性を考慮して、リコーディングの区分を設定した。分類区分ごとの構成比を一定以上に高めれば露見リスクを相対的に小さくできると考えられるが、一部の属性については区分を粗くするためにより大きな情報量損失が発生する可能性や、区分を細かくすることによって秘匿性が確保されない可能性が存在する。経済センサスのデータ特性を踏まえて、どのようなタイプの事業所群にどういった匿名化措置が必要となるのかを把握することで、秘匿性と有用性の両方を十分に考慮した匿名化措置を検討することができる。

そこで、本研究では、事業所ごとの秘匿性を評価するための露見リスクに関する指標を新たに設定した。本実験では、個票データのサンプルから抽出されたテストデータを対象に、k 個の変数(例えば k=2)の組み合わせによってクロス表を作成した上で、そのクロス表に含まれる度数が閾値を下回った場合にそれに該当するレコードを「露見リスク」のあるレコードとみなし、そのリスクの程度を相対的に評価するための指標を算出した。その意味で、本実験から得られた露見リスクの結果は、母集団に含まれる事業所を特定化するリスクを表していないことに留意されたい。本研究では、このようなリスクを定量的に評価した上で、露見リスクが相対的に高いと考えられる事業所と低いと考えられる事業所を類別し、その特徴を明らかにする。

4.2 経済センサスを用いた探索的な検証

本研究では、製造業の事業所から 10 万レコードを無作為抽出してテストデータを作成した。なお、本実験では前節と異なり、従業者合計の条件は考慮していない。具体的な実験方法として、地域 47 区分、産業 24 区分、従業者規模 14 区分、資本金階級 11 区分に加えて、

売上（収入）金額階級８区分、現金給与合計階級８区分、原材料使用額等階級９区分、年末在庫合計額階級８区分、付加価値額階級10区分、有形固定資産年末現在高階級10区分、投資総額階級10区分に対し、2属性ずつクロス集計を行った。なお、経理項目については、集計表で公開されている売上（収入）金額を参考にリコーディングを行っている。負の値を持つ属性について１つの階級区分とした。

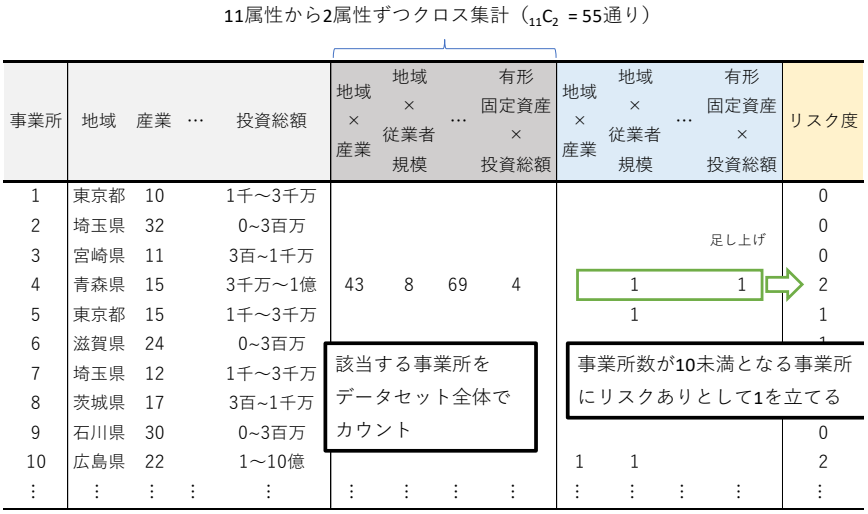
表 7 は、上記を 2 属性ずつクロスさせた場合に 10-匿名性を満たさない事業所数の一覧（事業所数でソート済み）を示している。最初の行は、地域と産業でクロス集計を行った結果を示している。具体的には、地域と産業のおのおののカテゴリーにおけるクロスの組み合わせにしたがってグループ化された階層に含まれる事業所数が 10 未満となるような事業所の数を集計した結果、合計で 2483 事業所あったことを意味している。地域×産業や、地域×従業者規模でリスクが高いと判定された事業所数が多いのは、地域、産業、従業者規模は分類区分の粒度が他の属性と比較して細かいことが、その理由の１つとして考えられる。このような分類区分の粒度は、外部参照情報との照合におけるリンクキーとしての精度と関連すると考えられる。そのため、本実験では他の属性と分類区分の構成比を揃えるような補正については行っていない。

つぎに、上記の検証と同様に、11 の属性に対して 2 属性ずつクロス集計を行い、その個々の分類区分に当てはまる事業所数が 10 未満となった場合に、本研究では、該当する事業所

表 7 2 属性のクロス集計で 10-匿名性を満たさない事業所数

属性1	属性2	事業所数
地域	産業	2483
地域	従業者規模	1192
地域	資本金階級	1048
地域	有形固定資産年末現在高階級	990
地域	年末在庫合計額階級	941
地域	投資額階級	849
地域	付加価値額階級	623
地域	現金給与合計階級	553
地域	売上（収入）金額階級	538
地域	原材料使用額等階級	523
産業	投資額階級	480
産業	資本金階級	474
産業	従業者規模	454
産業	有形固定資産年末現在高階級	317
産業	年末在庫合計額階級	297
産業	付加価値額階級	296
産業	売上（収入）金額階級	213
産業	現金給与合計階級	195
産業	原材料使用額等階級	192
従業者規模	資本金階級	164
資本金階級	投資額階級	132
⋮		
売上（収入）金額階級	年末在庫合計額階級	30
現金給与合計階級	投資額階級	29
従業者規模	現金給与合計階級	28
売上（収入）金額階級	投資額階級	23

図 10 リスク度評価のイメージ



を「露見リスクが相対的に高くなるレコード」と判定した。それぞれの事業所に対して、2 属性の組み合わせによって、リスクの判定に用いられるパターン数は 55 である。事業所レベルで見た場合、露見リスクが相対的に高い事業所は、その中の複数のパターンに該当すると考えられる。この該当するパターンの総計値を「リスク度」としてランク付けすることで、複数の準識別子を考慮して、露見リスクが相対的に高いレコードを探索する(図 10)。

つぎに、本研究では、リスク度 1 以上を「高リスク事業所」とリスク度 0 を「低リスク事業所」とそれぞれ設定し、高リスク事業所と低リスク異業所の特性を明らかにする。具体的には、本研究では、2 属性ごとにクロス集計を行い、それぞれの分類区分別に含まれる事業所および高リスク事業所の割合をバブルチャートで表示した。図 11 はその一例を示したものである。なお、バブルの大きさは事業所数を示しており、バブルの色は、高リスク事業所の割合が小さいほど薄く、高いほど濃くなるように表されている。

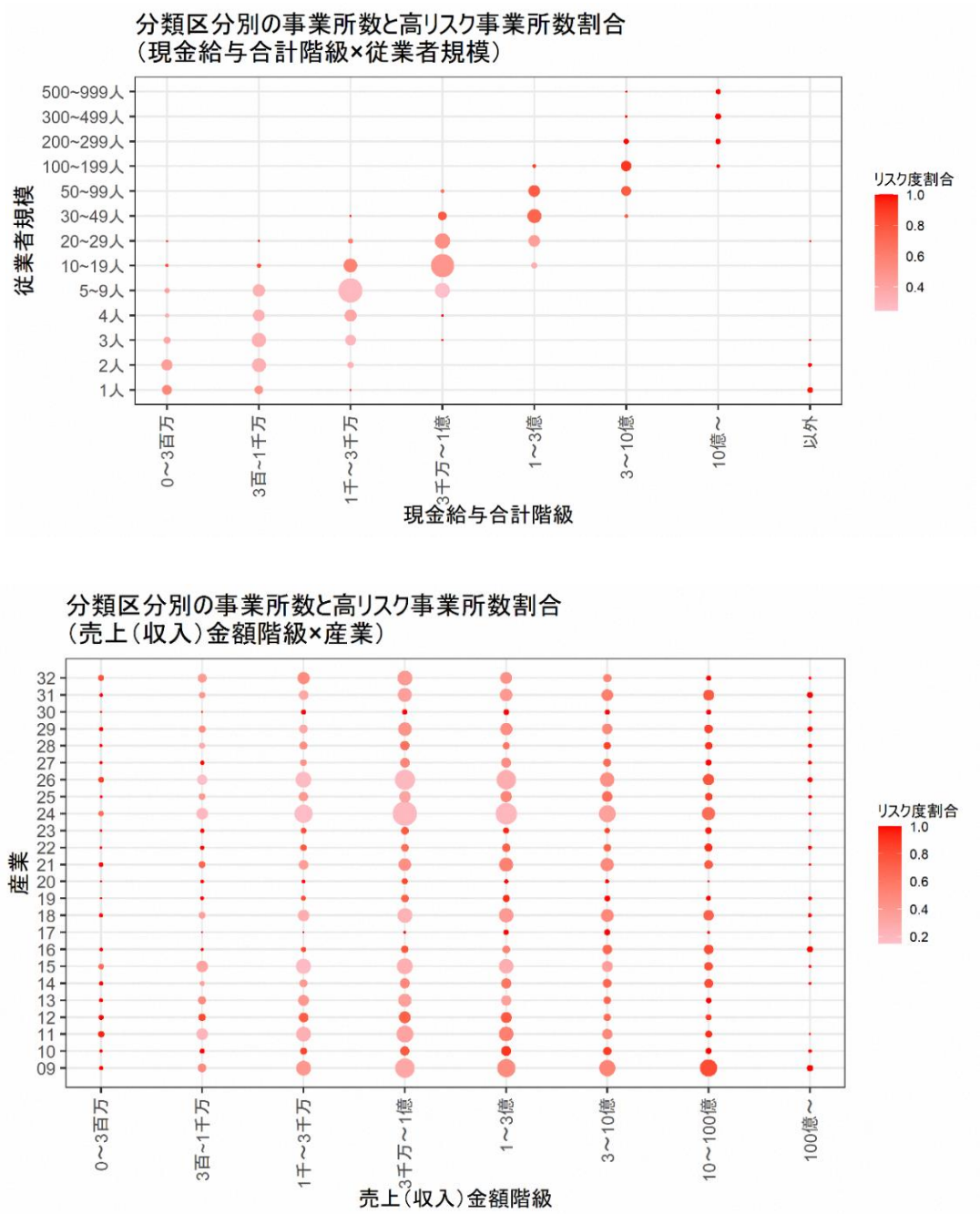
図 11 における現金給与合計階級×従業者規模について見ると、いずれの属性も小～中規模である階級区分のバブルは大規模な階級区分よりも相対的に大きく、また薄く表示されている。これはその分類区分に該当する事業所が多く、またリスク度も高くないため、相対的に露見リスクの小さな事業所であると考えられる。一方、現金給与合計階級や従業者規模の階級区分が大きくなるにつれて、バブルの大きさが小さくなるだけなく、色も濃くなっていることから、相対的に露見リスクが大きくなっていることが読み取れる。なお、分類区分の組み合わせによっては該当する事業所が存在しない場合バブルの大きさは 0 になるため、リスク度の割合も算出できない。しかしながら、リサンプリング等で少数の事業所がカウントされるケースも考えられるため、リスクが相対的に大きいものと判断する必要がある。このような理由から、バブルが濃く見えるエリアだけでなく、バブルが存在しないエリアも匿名化において注意を要すると解釈することができる。

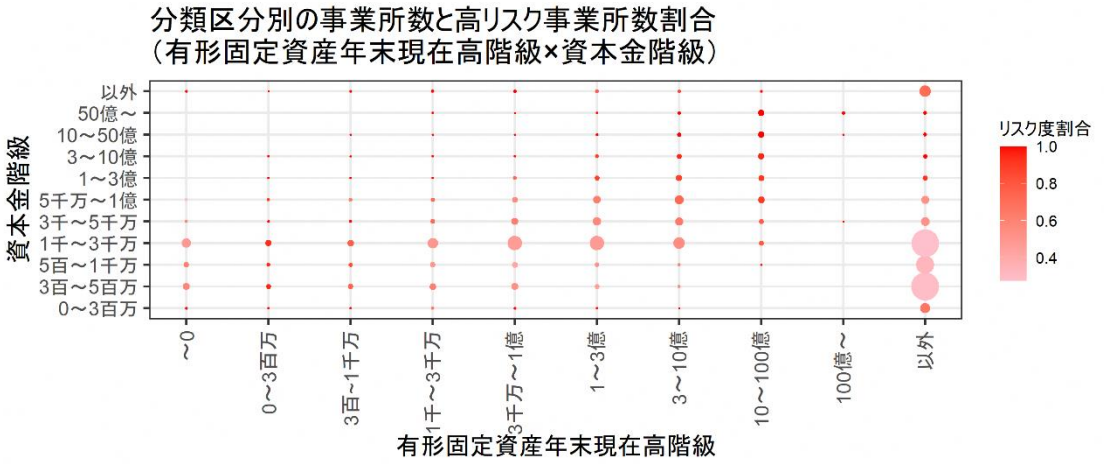
また、経理項目の組み合わせによっては相関性にも留意する必要がある。現金給与合計階級×従業者規模の場合、顕著な相関性が現れていることが確認できる。この相関性を表す分布から外れるような事業所については、例えば、一方の属性が小規模な階級区分であり、もう一方の属性が中規模以上の階級区分が該当する。相関性の観点から特異な傾向を示す事業

所については、露見リスクが大きくなる可能性がある。こうした特異な事業所に匿名化措置を施す場合、属性ごとのリコーディングだけでは十分ではない場合が考えられる。

つぎに、売上（収入）金額階級×産業の事例を見ると、売上（収入）金額の規模の大きさがバブルの大きさや色の濃さに影響を与えていることがわかる。しかしながら、売上（収入）金額が小さい場合でも、露見リスクが小さくなるとは限らない点について注意が必要である。その理由としては、0～300 万円未満の階級区分に該当する事業所が少ないことが指摘できる。これは、本実験の条件において、経理項目が記入対象外となっている個人票を除外していることが大きな理由であると考えられる。個人票の多くは規模の小さい事業所であること

図 11 分類区分別の事業所数と高リスク事業所数の分布





から、該当する事業所を除外することによって、売上（収入）金額の小さい事業所の数も大きく減少する。そのために、露見リスクが相対的に低いと考えられる売上（収入）金額の小さい事業所も、相対的に露見リスクが大きくなっているように見える。

有形固定資産年末現在高階級×資本金階級でも、同様の傾向が見られる。これについては、有形固定資産年末現在高の記入条件が従業者９人以下を除くこととなっていることから、小規模な事業所の数が少なく、リスク度割合も大きくなっている。このことは、露見リスクを検討するにあたって、統計調査の設計も関わることを示している。さらに、有形固定資産年末現在高階級においては、負の値にも留意する必要がある。本実験では、負の値にはすべて一括してリコーディングを施したが、もし正の値と同様に細かくリコーディングを行うのであれば、その区分の統合の仕方を検討することが求められよう。

本実験では、統計実務の視点に立った上で、外観識別性の観点からどのような属性が外部参照情報との準識別子になるかを検討し、選定された属性を対象にリスク度の定量的な評価を行っている。また、属性の中でどのような分類区分を用いるかによって、リスクが相対的に高いとされる事業所の分布傾向は変化する可能性がある。統計実務的にはこれらの考慮は不可欠であり、より慎重な検討と評価が求められる。

本実験で得られた研究成果は、事業所・企業系の匿名化ミクロデータの作成を指向する場合、キー変数の決定やリコーディングの分類区分の定め方、量的変数の攪乱の際の層化の基準など、様々な点で応用可能ではないかと考えられる。さらに、規模が大きい事業所は相対的にリスク度が高い傾向があり、相関性を見る上でも、リスク度が高い事業所の取り扱いに留意することが求められることも、本研究から確認された。

5. むすびにかえて

本稿では、イタリアやドイツの事例を中心に海外における事業所・企業系の匿名化ミクロデータの作成の現状を明らかにした上で、経済センサスの個票データをもとに各種の匿名化手法を適用した。試行的に作成した各種の匿名化ミクロデータについて、その有用性と秘匿性に関する定量的な評価を行うだけでなく、バブルチャートも用いながら、事業所の属性をクロスした場合の露見リスクの強度を定量的に評価した。

事業所・企業系のデータについては、把握される量的変数の分布に歪みがあるだけでなく、特異な分布が存在する。また、事業所・企業系の統計調査の中で一部抽出の調査の場合、標

本抽出の対象となるレコード数は企業規模ごとに大きく異なっており、規模の大きな事業所・企業については、悉皆で抽出されることも少なくない。これらの大規模な事業所・企業においては、有価証券報告書といった外部に開示される企業情報も存在することから、データの利用者が精度の高い外部情報を容易に取得できる場合がある。このことから、事業所・企業系のデータの露見に伴うリスクは、個人・世帯の調査における露見リスクと比較して相対的に大きいことが知られている。

そこで、本稿では、イタリアやドイツの先行事例を参考にした上で、経済センサスの個票データを用いた実証研究を行った。具体的には、リコーディングのような非攪乱的手法だけでなく、各種のミクロアグリゲーション技法による攪乱的手法を適用して作成した匿名化ミクロデータを対象に、クロス表による評価方法やリンケージ技法等を用いて作成した有用性と秘匿性に指標に基づいて、R-U マップによる定量的な評価を行った。さらに、経済センサスのミクロデータを対象に露見リスクに関する相対的な高さを評価するために、本稿では「リスク度」に関する指標を設定し、相対的な露見リスクの探索的な検証を行った。本研究から、事業所・企業系のデータ特有の歪みを持つ分布特性や特異値の存在を考慮した上での匿名化措置が必要なことが明らかになった、本研究では、ミクロアグリゲーションを適用したが、今後は攪乱的手法としてのノイズの付加の適用についても検討の必要性があると考える。さらに、経済センサスのような多数の経理項目が存在する事業所・企業系の統計調査においては、属性間の相関性のような有用性の指標を確認する場合でも、分布の歪みや特異値といったデータ特性に留意する必要があることがわかった。

ところで、経済センサスでは、有形固定資産年末現在高や投資総額といった一部の属性群は、関連する複数の属性項目を合計した上で算出されている。このことから、匿名化においては、総計と内訳との関係性を保持しつつ、どのように匿名化を行うかが課題となる。このような内訳に対する匿名化措置は、事業所・企業系の統計調査だけでなく、世帯・人口系の統計調査においても、詳細な内訳項目の提供を指向するための匿名化の論点になりうることから、さらなる検討が求められる。これについては、今後の研究課題としたい。

謝辞

本研究では、統計法の規定に基づき、「経済センサス － 活動調査」に係る調査票情報を使用した。なお、匿名の２名の査読者より貴重なコメントをいただいたことについて、深謝いたします。

参考文献

[1] 伊藤伸介（2009）「匿名化技法としてのミクロアグリゲーションについて」熊本学園大学『経済論集』第 15 巻第 3・4 号合併号, pp.197-232

[2] 伊藤伸介 (2010) 「ミクロデータにおける秘匿性の評価方法に関する一考察」，明海大学『経済学論集』第 22 巻第 2 号，pp.1-17.

[3] 伊藤伸介, 村田磨理子, 高野正博 (2014)「ミクロデータにおける匿名化技法の適用可能性の検証」 総務省統計研究研修所『統計研究彙報』，第 71 号, pp.83-124.

[4] 伊藤伸介 (2018)「公的統計ミクロデータの利活用における匿名化措置のあり方について」『日本統計学会誌』第 47 巻第 2 号，pp.77-101.

[5] 伊藤伸介（2020）「諸外国における公的統計と行政記録データの二次利用に関する展開方向」『経済学論纂(中央大学)』第 61 巻第 2 号，pp.1～16 頁.

[6] 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

[7] 濱砂敬郎 (1999)「ドイツ連邦統計法におけるミクロデータ規定の匿名化措置」 法政大学日本統計研究所『研究所報』No.25，pp.69-99.

[8] 星野伸明 (2010)「公的統計ミクロデータ提供制度の課題」『日本統計学会誌』 シリーズJ 40(1), pp.23-45.

[9] Brandt M., Lenz R., Rosemann M. (2008), Anonymisation of Panel Enterprise Microdata – Survey of a German Project, Domingo-Ferrer J., Saygın Y. (eds) Privacy in Statistical Databases PSD 2008 Lecture Notes in Computer Science, vol 5262 Springer, Berlin, Heidelberg.

[10] Breunig M.M., Kriegel H.-P., Ng T.R., SanderJö. (2000), LOF: identifying density-based local outliers, ACM sigmod record (pp. 93--104).

[11] Domingo-Ferrer J., Mateo-Sanz J.M. (2002), Practical data-oriented microaggregation for statistical disclosure control, IEEE Transactions on Knowledge and Data Engineering, 14(1):189–201.

[12] Domingo-Ferrer J., Torra V. (2005), Ordinal, Continuous and Heterogeneous :math:`k`-anonimity through Microaggregation, Data Mining and Knowledge Discovery 11(2), pp. 195-212.

[13] Duncan T.G., Pearson W.R. (1991), Enhancing Access to Microdata While Protecting, Statistical Science, Vol.6, pp.219-239.

[14] Duncan, G., Keller-McNulty, S. A., & Stokes, S. L. (2001). Disclosure Risk vs. Data Utility: The R-U Confidentiality Map. Carnegie Mellon University. Journal contribution.

[15] Ester M . (1996), A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, Proceedings of the second ACM International Conference on Knowledge Discovery and Data Mining (KDD), pp.226-231.

[16] Franconi L., Ichim D. (2007), Community Innovation Survey: comparable dissemination,

[17] Hundepool A., de Wetering A.V., Ramaswamy R., Franconi L., Capobianchi A., DeWolf P.-P., DomingoFerrer J., Torra V., Brand R., Giessing S. (2003), μ-ARGUS version 3.2 Software and User's Manual, Statistics Netherlands, Voorburg NL. http://neon.vb.cbs.nl/casc://neon.vb.cbs.nl/casc.

[18] Ichim D. (2007), Microdata anonymisation of the Community Innovation Survey data: a density based clustering approach for risk assessment, Documenti Istat, 2.

[19] Lenz R., Rosemann M., Vorgrimler D., Sturm R. (2006), European Data Watch: Anonymising Business Micro Data – Results of a German Project, Schmollers Jahrbuch : Journal of Applied Social Science Studies / Zeitschrift für Wirtschafts- und Sozialwissenschaften, Duncker & Humblot, Berlin, vol. 126(4), pp. 635-651.

[20] Yancey, W. E., Winkler, W. E., & Creecy, R. H. (2002). Disclosure risk assessment in perturbative microdata protection. In Inference control in statistical databases (pp. 135-152). Springer, Berlin, Heidelberg.

[21] O'Keefe C.M., Shlomo N. (2014), Applicability of Confidentiality Methods to Personal and Business Data. Domingo-Ferrer J. (eds) Privacy in Statistical Databases PSD 2014 Lecture Notes in Computer Science, vol 8744 Springer, Cham.

[22] Ritchie F., Hafner H., Lenz R. (2019), User-focused threat identification for anonymised microdata. Statistical Journal of the IAOS, 35(4), 703-713.

[23] Samarati P., Sweeney L. (1998), Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, Carnegie Mellon University Journal contribution.

[24] Templ M., Kowarik A., Meindl B. (2015), Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro, Journal of Statistical Software, 67(4), 1 - 36.