＜ 修 士 論 文 ＞

公的統計ミクロデータの利活用の促進に向けた統計的開示抑制の検討

―事業所・企業の匿名化ミクロデータの作成に資する基礎研究―

滋 賀 大 学 大 学 院
デ ー タ サ イ エ ン ス 研 究 科
デ ー タ サ イ エ ン ス 専 攻

修了年度：２０２０年度
学籍番号：6019122
氏　　名：横溝　秀始
指導教員：竹村　彰通
提出年月日：２０２１年１月２０日

本研究は、2019 年度行政官国内研究員制度（修士課程コース）の一環として行われた。研究にあたっては、指導教員・副指導教員だけでなく、伊藤伸介特任教授との共同研究という形を取っている。本研究における海外の事業所・企業系の匿名化ミクロデータの作成状況については、横溝他（2020）や 伊藤・横溝（2020b）を、経済センサスを用いた実証研究については伊藤・横溝（2020a）をベースとしている。

目次

1　はじめに

本研究では、公的統計ミクロデータの利活用の促進に向けた統計的開示抑制の検討の一環として、事業所・企業の匿名化ミクロデータの作成に資する基礎研究を行うことを目的とする。

匿名化ミクロデータとは、調査対象の秘密の保護が図られた、世帯単位や事業所単位といった集計する前の個票形式のデータのことを指す。名称や電話番号といった直接的な識別情報を削除する、属性ごとに公開する分類事項の程度を粗くする、データに偽の要素を混ぜ込んで攪乱するといった匿名化処理を行うことで、秘匿性の担保されたデータセットを作成し、学術研究等に活用されるものである。

この匿名化ミクロデータについて、わが国では、現在 7 種類の世帯・人口系の統計調査が統計法第 35 条および 36 条に基づく匿名データとして提供されているが、事業所・企業系の統計調査については未提供となっている。海外においても、イタリア、ドイツ、Eurostat において事業所・企業系の匿名化ミクロデータが作成された事例があるものの、事業所・企業系については、オンサイト利用やリモートアクセスの形で利用されているのが現状である（伊藤(2018a)）。その一方で、事業所・企業系の匿名化ミクロデータには、学術研究の利用だけでなく、高等教育のための利用や、オンサイト利用等でプログラムを作成するためのテストデータとしての利用も考えられる。また、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、賃金構造基本統計調査の匿名データの作成の可能性が指摘されている。こうした点を踏まえると、事業所・企業系の匿名化ミクロデータへのニーズはわが国でも存在すると思われる。

そこで、本研究では、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した上で、海外における現状を踏まえて事業所・企業系の匿名化ミクロデータの作成に関する論点を整理する。その上で、経済センサスの個票データを用いて、わが国での事業所・企業系の匿名化ミクロデータの匿名化措置の可能性を追究する。2 章では、まず前段として、公的統計における匿名化ミクロデータの制度の概要や、統計的開示抑制における基本的な概念をサーベイした。続く 3 章では、その中でも特に、事業所・企業における匿名化ミクロデータの先行研究や先行事例に焦点を当てて詳しいサーベイを行った。イタリアやドイツにおける学術研究への利用を目的とした匿名化ミクロデータ作成の論点を整理することが中心となっている。4 章では、わが国の事業所・企業系の統計調査の中でも最も規模の大きい経済センサスについて、オンサイト利用を通じて匿名化処理の実証研究を行った。先行研究に基づき、秘匿性と有用性の評価を様々な角度から行っている。さらに 5 章では、同じく経済センサスについてそのデータ特性を探り、より個別具体的な匿名化を検討するための探索的な試行も行った。事業所単位での相対的なリスクを評価し、匿名化にあたって特に注意しなければならない事業所や属性を洗い出した。最後に 6 章で本研究をまとめる。

本研究は、事業所・企業系のミクロデータを対象とした試論的な基礎研究である。事業所・企業のデータ特性を踏まえた匿名化手法について、統計実務の観点も踏まえつつ、さらなる検討を進めていきたいと考えている。

2　公的統計における匿名化ミクロデータ

2.1　公的統計データの提供

わが国を含む多くの国々では、公的統計データについて統計表以外にも様々な形態でのミクロデータ（個々の回答者、若しくは、経済主体についての情報を含んだレコードのセット）の提供が行われている。国内、海外、それぞれの観点からその概要を整理する。

2.1.1 国内

わが国の公的統計データの**二次的利用**制度とは、統計調査により集められた情報を、既存の調査結果（集計表・報告書等）のほかに、秘密の保護を図った上で、新たな統計作成や統計的手法を利用した学術研究等のために活用するものである（永島（2018））。

昭和 22 年に制定された旧統計法では、原則として統計調査の目的に沿った利用（一次利用）のみが認められており、それ以外の利用は禁止されていた。一方、平成 19 年に大きく改定された新統計法では、統計法の定める特別の場合には例外的に二次的な利用が認められることとなった。これにより、「行政のための統計」から「社会の情報基盤としての統計」へと転換が図られ、統計調査で集められた情報がより広範に活用されることとなった。二次的利用には、調査実施者以外の者による統計データを活用した学術研究等が可能になることや、新たに別の統計調査を行う必要性が減り、調査実施者・調査対象の負担軽減につながるといったメリットが存在する。平成 29 年の「統計改革推進会議 最終取りまとめ」では、「EBPM（= Evidence Based Policy Making、証拠に基づく政策立案）」が謳われているが、公的統計データはその根幹を担う役割を果たすことから、二次的利用制度にも大きな注目が集まっている。

二次的利用の利用形態については、大まかに調査票情報の利用、調査票情報の提供、オーダーメイド集計、匿名データが存在する。

調査票情報の二次利用：統計第 32 条に基づく、調査を実施した各府省等（行政機関、独法等）自身が利用する場合の利用形態である。対象が国民ではない点が他の二次的利用制度とは一線を画する。

調査票情報の提供：統計法第 33 条に基づく、統計調査により集められた情報（統計調査の回答内容とほぼ同等な情報）を、公的機関や公的機関が認めた者に提供する形態である。具体的な利用方法としては、**オンサイト利用**（統計センターと連携する大学や行政機関等に設置された情報セキュリティを確保したオンサイト施設において調査票情報を利用する）と**磁気媒体による提供**（調査実施者である行政

機関等が磁気媒体により必要な範囲において調査票情報を提供する）とが存在する。 ミクロデータ利用ポータルサイト miripo（2020）によれば、2020 年 8 月現在（以下同様）、オンサイト利用については 7 府省の 56 の統計調査が、磁気媒体では 9 府省の 215 の統計調査が利用可能となっている。以下に述べる匿名データよりも利用要件は厳しいが、その分より研究目的に適した調査票情報を取り扱うことができるという違いがある。

オーダーメード集計：統計法第 34 条に基づく、利用者からの委託（オーダー）を受けて、利用者の分析目的に対応した集計表を新たに作成する仕組みである。所管省庁で行うオーダーメード集計とは、既存の統計調査で得られた調査票データを活用して、調査実施機関等が申出者からの委託を受けて、そのオーダーに基づいた新たな統計を集計・作成し、提供するものである。学術研究の発展に資することと有料であることを条件に、一般の者に認められている。10 府省等（日本銀行含む）の 31 の統計調査が対象となっている。

匿名データ：統計法第 35 条、第 36 条に基づき、行政機関等が行う統計調査によって集められた調査票情報を、特定の個人又は法人その他の団体の識別（他の情報との照合による識別を含む。）ができないように加工したデータを提供する形態である。匿名化措置に当たっては、安全性（調査客体の匿名性）に加え、データ分析の有用性にも配慮がなされている。匿名データの作成にあたっては、外部有識者を交えた研究会等により匿名データの作成方法の検討を重ねるとともに、さらに、基幹統計調査（重要性が特に高いと位置付けられているもの）に係る匿名データの作成方法については、統計委員会において審議も必要となっている。そのため、他の利用形態に比べて提供できる統計調査の数は多くなく、総務省および厚生労働省の管轄する 7 調査のみが利用可能となっている。匿名データは、オーダーメード集計と同じく、学術研究の発展に資することと有料であることを条件に、一般の者に提供が認められている。なお、現在利用可能な調査はすべて世帯・人口系の統計調査であり、事業所・企業に関するものは存在していない。本論文では、以降の章で事業所・企業系調査の匿名化ミクロデータの作成可能性を検討する。

これらの利用形態の主な利用条件等を一覧にしたものが表 1 である。

表 1 利用形態別の主な利用条件（永島（2018）より）

<table>
<tr><th>利用形態</th><th>根拠</th><th>利用できる者</th><th>利用目的</th></tr>
<tr><td>①調査票情報の二次利用</td><td>法第32条</td><td>調査を実施した各府省等（行政機関、独法等）自身が利用する場合</td><td rowspan="2">統計の作成
統計的研究
調査名簿の作成</td></tr>
<tr><td rowspan="2">②調査票情報の提供</td><td>法第33条第１号</td><td>公的機関（行政機関等＋会計検査院、地方独法等）が利用する場合</td></tr>
<tr><td rowspan="3">法第33条第２号</td><td>公的機関が委託又は共同して調査研究を行う者</td><td rowspan="4">統計の作成
統計的研究</td></tr>
<tr><td>公的機関が公募の方法により補助する調査研究を行う者</td></tr>
<tr><td>行政機関等（行政機関＋地方公共団体、独法等）が政策の企画・立案、実施又は評価に有用であると認める統計の作成等を行う者</td></tr>
<tr><td>③オーダーメード集計</td><td>法第34条</td><td rowspan="2">一般の者

※学術研究の発展に資するなどが条件
・研究等の目的に限定
・研究成果等の公表義務
※有料（法第38条）
・手数料（実費を勘案し設定）を納付</td></tr>
<tr><td>④匿名データ</td><td>法第35条、法第36条</td></tr>
</table>

また、図 1 は、わが国における統計作成と二次的利用制度を用いて作成される匿名化ミクロデータに対する匿名化の関係性のフローを示している（伊藤(2018b)）。集計計画に従って調査客体から収集された記入済みの調査票情報（個票データ）は、通常の集計結果表としての公表だけでなく、利用申請に応じて匿名データ、調査票情報の提供、オーダーメイド集計といった二次的利用の枠組みを通じて提供が行われていることがわかる。

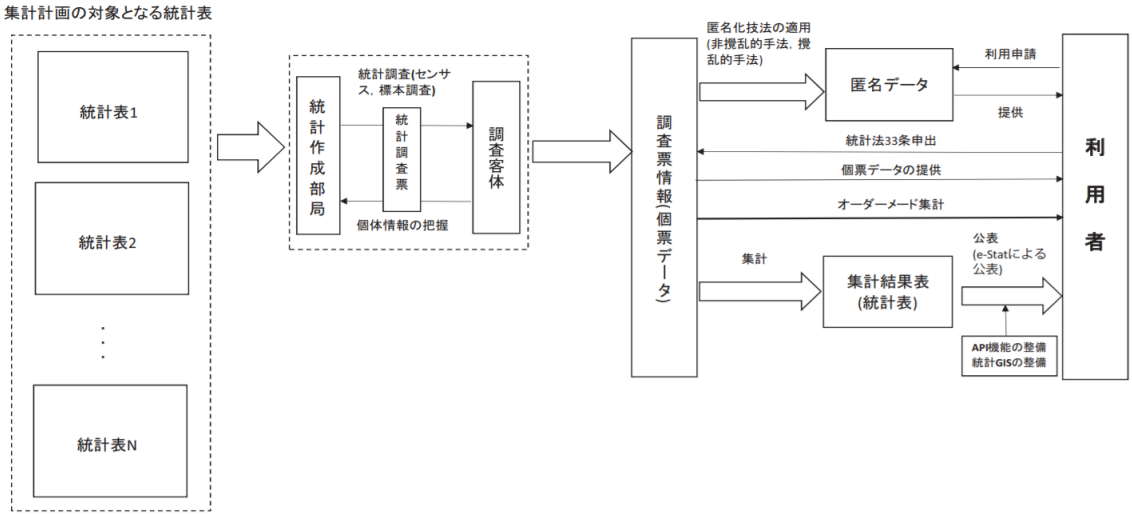


図 1 統計作成とミクロデータに対する匿名化との関係（伊藤（2018）より）

なお、上記の二次的利用制度とは別に、**一般用ミクロデータ**（独立行政法人統計センター（2020a））という枠組みも存在している。一般用ミクロデータとは、集計表から作成するなど、調査票情報を直接的に用いない方法により作成する擬似的なミクロデータで、統計演習など教育用に広く一般に利用されることを目的としている。疑似的なミクロデータであるため、匿名データとは異なり実証研究には適さない。一方で、利用要件の制約が緩く、個別情報の秘匿を気にする必要がないというメリットがある。2020 年 8 月現在、利用可能な調査は全国消費実態調査と就業構造基本調査のみとなっている。一般用ミクロデータやそれ以前の試みである教育用疑似ミクロデータの教育利用については河野・和田（2018）が詳しいが、その中では、調査客体が世帯のみでなく企業や事業所にも拡充され、時系列やパネルデータの形でのデータセットが作成・公表されることへの期待が述べられている。本論文は、その可能性について考察する。

2.1.2 海外

海外の公的統計における二次的利用のシステムは、国ごとに様々である。各国の法制度に基づきながら、利用目的、利用対象、利用場所、および利用の仕方に即して、様々な形態でミクロデータの作成・提供がなされている。伊藤（2016）や伊藤（2018a）では、公的統計ミクロデータの提供形態は、①学術研究目的のための利用を対象にした個票データ（deidentified data）の提供サービス、②リモートエグゼキューション（remote execution、オンデマンドによる集計サービス（リモート集計）も含む）に基づく分析結果の提供、③主として学術研究のために作成される匿名化ミクロデータ（anonymized microdata）の作成・提供、および一般公開型ミクロデータ（public use microdata）の公開に大別されている。これらをベースにし

て、伊藤（2020）ではさらに、個票データの提供サービスが、①オンサイト施設による提供、②磁気媒体による提供、③リモートアクセスによる提供に類別されている。また、リモートエグゼキューションに基づく分析結果の提供サービスは、①プログラム送付型のリモートエグゼキューションによる提供と②オンデマンドシステムによる提供に区別されている。これらを整理したものが表 2 である。国ごとにその形態は様々であることがわかる。

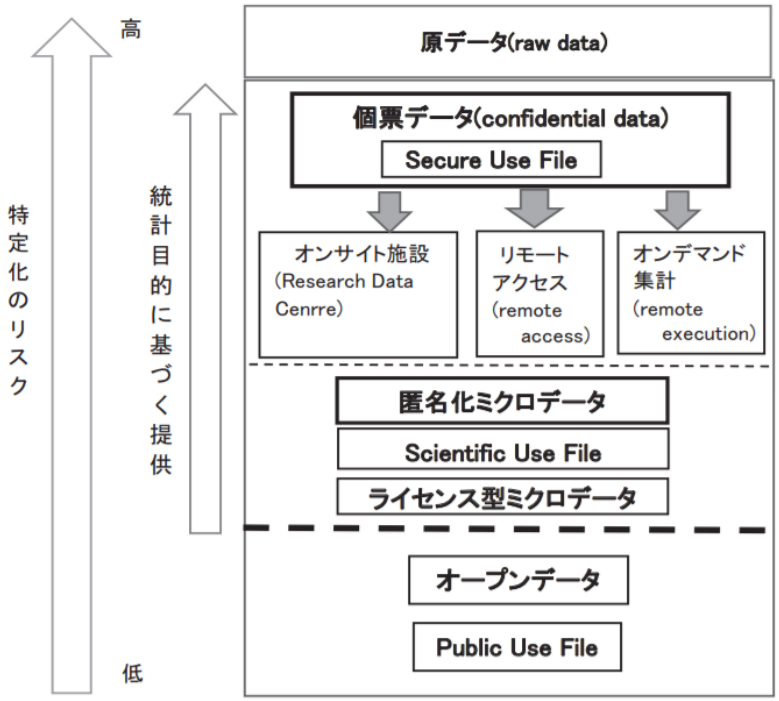
表 2 諸外国の統計作成部局におけるミクロデータの提供形態
（伊藤（2020）より）

データの種類/提供形態 各国の統計作成部局	個票データのオンサイト施設による提供	個票データの磁気媒体による提供	個票データのリモートアクセスによる提供	プログラム送付型のリモートエグゼキューションによる提供	オンデマンドシステムによる提供	匿名化ミクロデータの提供	一般公開型ミクロデータによる提供
Eurostat	○				●	○	○
イギリス国家統計局	○		○		●	○(UKDSから提供)	○
ドイツ連邦統計局	○			○	○	○	○
フランスINSEE			○			○	○
オランダ統計局	○		○	△	○	○	○
デンマーク統計局			○		○		
フィンランド統計局	△		○			○	
アメリカセンサス局	○				○		○
カナダ統計局	○			○	○		○
オーストラリア統計局	○		○	△	○	○	
ノルウェー統計局	○	○		○			○(NSDから提供)
総務省統計局	○	○				○	

注：○…運営されている。●…計画中である。△…運営中であるが、活動を休止している（2020年２月時点）。
資料：赤谷・荒川・伊藤（2014）、伊藤（2016）、伊藤（2017）、伊藤（2018）、伊藤・谷道・小島（2018）、小林（2011）、Zayatz（2007）に基づき作成。

特定化のリスクの観点から見たミクロデータの位置関係を図示すると、図 2 のように整理される（伊藤（2018b））。氏名や住所といった直接的な識別子を含む原データ（raw data）は一部の統計作成部局の職員のみがアクセス可能であり、利用者に提供されることはない。原データから直接的な識別子のみが削除された個票データは、個体が特定化されるリスクが高いミクロデータであるため、オンサイト施設やリモートアクセス（remote access）、あるいはオンデマンドの集計システムといった制御された環境でのみアクセスすることができる。個票データに匿名化処理が施された匿名化ミクロデータやライセンス型ミクロデータ（end user license data）[1]は、後述するように学術研究に主に用いられる。個票データやライセンス型ミクロデータについては、原則的に統計目的に対して提供が行われている。オープンデータと呼ばれる一般公開型ミクロデータは教育用目的に用いられることが多く、最も特定化のリスクが低いという特徴がある。

[1] イギリスのエセックス大学の UK Data Service を通じて web 上で取得できるデータである。ライセンスを取得することによって、学術研究目的のための利用が可能となる。



資料：伊藤（2016b，7 頁）を加筆修正

図 2 個票データ、匿名化ミクロデータと Public Use File との関係
（伊藤（2018b）より）

上記の個票データから様々な秘匿処理が施されたデータをもう少し整理する。伊藤（2018a）や伊藤（2020）によると、これらは匿名化の強度によって、①匿名化ミクロデータと②一般公開型ミクロデータに類別することができる。

匿名化ミクロデータ：匿名化ミクロデータはヨーロッパの多くの国々で提供されており、**学術研究用ファイル（Scientific Use File＝SUF）**とも呼ばれる。多くの場合匿名化ミクロデータは世帯・人口系の調査のものであるが、イタリアやドイツでは事業所・企業系の統計調査の SUF の例も存在する。学術研究を志向していることから、原データの性質を損なわないようにすることが重視されており、秘匿の程度は比較的小さい。反面、秘匿性が強いとは言えないため、利用要件が制限されていることが多い。

一般公開型ミクロデータ：一般公開型ミクロデータは、**一般公開型ファイル（Public Use File ＝ PUF）**とも呼ばれる。一般公開型ミクロデータが広範に提供されているのはアメリカとカナダであり、人口センサスや標本調査の PUF が提供されている。また、ヨーロッパでは教育目的やテストデータの利用のために PUF が公開されている。匿名化ミクロデータに比べて、公開する範囲が広いことから秘

匿性が高く、安全に利用することができる。一方で原データの性質は相対的に大きく損なわれているため、学術研究に利用することは推奨されない。

このように、SUF と PUF は相互補完的な役割を果たしている。

2.1.3 国内外の比較

わが国には、2.1.1 で示した通り、調査票情報の提供、オーダーメイド集計、匿名データといった二次的利用制度が存在する。これらに類似する仕組みは、2.1.2 のように海外にも存在している。一方で、わが国にはないリモートアクセスやプログラム送付型のリモートエグゼキューション、オンデマンドシステムといった取り組みもなされている。その動向を把握することは重要である。

また、2.1.1 では、わが国には統計法上で匿名データや調査票情報の区別があることを、2.1.2 では、海外では PUF と SUF の分類があることを示した。この PUF や SUF の公的統計に関する法律における条文上の位置付けは諸外国で異なり、わが国においてもその位置づけが明確でない。現行の統計法では、「匿名データ」と「調査票情報」が法第 2 条で定義されているが、PUF や SUF が統計法においてどのように解釈されうるかについては議論の余地があることに注意が必要である。

2.2 統計的開示抑制

本節では、匿名化ミクロデータを作成する際の方法論である統計的開示抑制や、それに関連するサーベイを中心にまとめる。

2.2.1 統計的開示

統計の公開によって、データの外部ユーザーが、機密情報について確度の高い推定値を得ることが可能になった場合、すなわち個体情報の**開示**または**露見（disclosure）**が発生した場合、**統計的開示（statistical disclosure）**が起きたと言われる。ミクロデータは研究や教育に有益なものであるが、それは調査客体の機密情報が保護されることが前提である。いかに有用性の高いミクロデータであっても、統計的開示が発生するものは個人情報や機密情報の観点から利用に供することができない。ミクロデータ作成にあたって、この統計的開示をいかに抑制するかが重要となる。独立行政法人統計センター（2005）によれば、**統計的開示抑制（statistical disclosure control＝SDC）**[2]とは、個人、企業あるいはその他の機関に関する情報が露見されるリスクを抑制するための方法と定義されるものである。

[2] 日本語では統計的開示制御、英語では statistical disclosure limitation と表現されることもある。

この種の方法は公表段階のみに関係し、通常、公開するデータ情報量を制限するかあるいは修正する方法である。

2.2.2 統計的開示の事例

過去に実際に匿名化ミクロデータの露見が発生したことがあるのか、また露見が発生した場合にはどのような問題が発生するのかについて述べる。

2020 年現在、公的統計の分野において匿名化ミクロデータの露見が大きな問題となった事例は知られていない。Hafner *et al.* (2019) によれば、学術研究用ファイルやオンサイト施設のような管理化にある管理アクセスファイルが、悪意のある侵入者によって露見された例は、10 年間の管理の中では一度もなかった。約 1 万人のユーザーが訪れた中で、3 回の意図的な誤用と 10 回ほどの意図的ではない誤用はあったが、意図的な誤用はすべて研究者が自分たちの都合のためにプロセスを再編成しようとした結果であり、データを再特定しようとした結果ではなかった。約 40 年間にわたり、学術研究用ファイルや管理アクセスファイルにおける制度面等による非データ的な管理手段は成功を収めて来たと述べている。さらに、一般公開型ファイルについても、侵入者が存在することは否定できないが、強い匿名化が施されている一般公開型ファイルを攻撃することにあまり価値はないとも述べられている。その他、公的統計の分野ではワーストケースのシナリオばかりが考慮されており、過剰な秘匿性を担保するためにミクロデータとしての有用性が損なわれているという主張も展開されている。

このような傾向が現れる理由は大きく二点あると考えられる。第一に、公的統計の場合、たった一度の露見でも統計局の信頼が失墜しかねないという点である。仮に露見が発生した場合、それによる調査客体への実質的な被害が微少であったとしても、センセーショナルな報道への対応等で実務担当者に間接的な被害が発生すると考えられる。また、続く調査での回答率が低下するという問題も発生する。そのようなリスクに備えるため、慎重にならざるを得ないという考え方である。第二に、実務担当者がリスクを取るインセンティブが小さいことである。ミクロデータの有用性を高めて評価されるメリットよりも、秘匿性を弱めて上記のような万が一のリスクを負うデメリットのほうが大きい。以上のような理由から、公的統計の匿名化ミクロデータは、世界的・歴史的に見て有用性よりも秘匿性が重視されてきた経緯があると考えられる。

一方で、公的統計以外については、ミクロデータの露見事例がいくつか存在する。内閣官房（2013）では、3 種類の事例が紹介されている。一つ目は、1997 年に米国マサチューセッツ州が公開した医療データから州知事の情報が特定された事例である。マサチューセッツ州は医療データから氏名等を削除して公開していたが、その中には性別、生年月日、郵便番号が含まれていた。既に公開（販売）されてい

る投票者名簿とマッチングしたところ 1 人に特定される事態が発生した。二つ目は、2006 年の米国インターネットサービス企業 AOLL（America Online）が検索履歴の公表を中止した例である。氏名や IP アドレスを匿名化した上で 65 万人のユーザーの 3 か月の検索履歴のリスト 2,000 万件を公表していたが、職業や検索内容、住居といった準識別子から個人が特定される事態となった。三つ目は、2006 年に米国の映画レンタル・サービスの Netflix が映画推薦アルゴリズムコンテストを中止した例である。匿名化したユーザーの視聴履歴データを用いたコンテストが行われたが、他の映画情報サイトとリンケージを行うことで一部の個人が特定されることが発覚し、コンテストは中止に追い込まれた。いずれも社会的な関心を集め、自治体や企業の評価面や金銭面にも小さくない損害を被ったと考えられる。このような事態は、公的統計においても避けなければならない。

その他、近年では、医療や社会科学の分野においてミクロデータの個体の再識別モデルを作成して推定を行った結果、15 の人口統計の属性を使用することでアメリカ人の 99.98％の再識別に成功したとする研究例も存在する（Rocher *et al.* (2019)）。公的統計の匿名化ミクロデータ作成においても、医療統計や社会科学統計の露見事例やその対策に学ぶところは多いと考えられる。

2.2.3 統計的開示の種類

Truta *et al.* (2003) では、統計的開示について、一般的に ID 開示と属性開示の 2 種類が存在すると述べられている。

ID 開示（identity disclosure）：ID 開示とは、個人や機関などの実体が特定されることである。侵入者が既知のレコードを、公開されたミクロデータレコードに対応付けを行うことによって発生する。特異な属性やその組み合わせを持っているレコードは、外部参照情報から個体の特定が行われやすい。ID 開示を目的とする侵入者は一般的に、リンケージというアプローチを取る。既知のレコードの属性とミクロデータ上のレコード間の距離を算出することにより、もっともそれらしいレコードの特定を行う手法である。このリンケージによる容易な対応付けを許さない匿名化が、ミクロデータ作成では重要となる。

属性開示（attribute disclosure）：属性開示とは、侵入者がその実体について何か新しいことを発見することである。公開されたミクロデータの属性に基づいて個体の非公開の属性を入手することによって発生する。ID 開示が個体を特定するものであるのに対し、属性開示は必ずしも個体の特定を必要としない。例えば、20~24 歳のすべての男性患者が癌であるというミクロデータが公開された場合、その属性に当てはまる不特定多数に対して、不利益な情報が開示されることになる。なお、属性開示は集計表に対しても発生しうるため、ミクロデータ固有の露見リスクというわけではない（竹村（2003））。

推論開示（Inferential disclosure）：IHSN（2019）ではさらに、3 種類目の開示として推論開示があげられている。侵入者が、リリースされたミクロデータを使用して、個体の属性推定できる場合に発生する。たとえば、精度の高い回帰モデルによって、公開済みの属性（例えば地域、職業、年齢等）から機密の属性（年収等）を推論できた場合、推論的開示が起こったと言える。ミクロデータはもともと多変量の関係性を知るためのものであり、また実際の個体の情報ではなく、あくまでも一般的な傾向でしかないため、ミクロデータ作成においては推論開示を問題視しないことが多い。なお、合成データによる補完という匿名化手法は、この性質を利用していると考えることができる。

2.2.4 属性の分類

ミクロデータは一般に複数の属性（変数）を持っているが、これらは機密性や外観識別性によってその性質が異なる。IHSN（2019）や Domingo-Ferrer & Torra (2005) を参考に、以下のように分類することができる。

識別子（identifiers）：個体の識別を可能にする外観識別性の高い属性である。これは直接的な識別子と準識別子の 2 種類に分類できる。

- **直接的な識別子（direct-identifiers）**：個体の身元を明確にする、氏名や詳細な住所などの情報である。ミクロデータ作成において、これらを削除することは大前提であり、不可欠なステップである。しかしながら、直接的な識別子の削除だけでは、以下で説明する準識別子からの露見を防ぎきれないことが多い。
- **準識別子（quasi-identifiers）**：単体では個体の特定には至らないが、複数組み合わせることでそれが可能になるような属性を、準識別子と呼ぶ。例えば、大まかな地域、性別、世帯人員、資本金、産業分類といった識別性の高い属性がそれにあたる。ミクロデータの利用上、準識別子は非常に有用であるため、直接的な識別子のように無条件に削除することはできない。その代わり、準識別子の組み合わせることによって露見に繋がるようなユニークなレコードを作らないように匿名化する必要がある。匿名化にあたってはこの準識別子の取り扱いが重要であり、また、外部参照情報との鍵の役割を果たすことから、キー変数（key variables）と呼ばれることも多い。
- **非識別子（non-identifiers）**：個体の識別に用いることのできない外観識別性の低い変数である。識別子に比べれば露見リスクは小さいが、外部参照情報によっては、思いもよらぬ属性が識別子となる可能性もある。明確に識別子と非識別子を区別することは容易ではない。

さらに、機密属性と非機密属性に分類することもできる。

- **機密属性（sensitive attributes, confidential outcome attributes）**：個体が特定された場合に不利益につながるような機密性の高い属性である。例えば、資産、売上（収入）金額、健康状態などが挙げられる。分析上非常に有用な属性であるため、これもまた削除以外の匿名化を考えることになる。
- **非機密属性（non-sensitive attributes, non-confidential outcome attributes）**：個体に関する機密性の低い属性である。

上記を構造的にまとめた図が図 3 である。

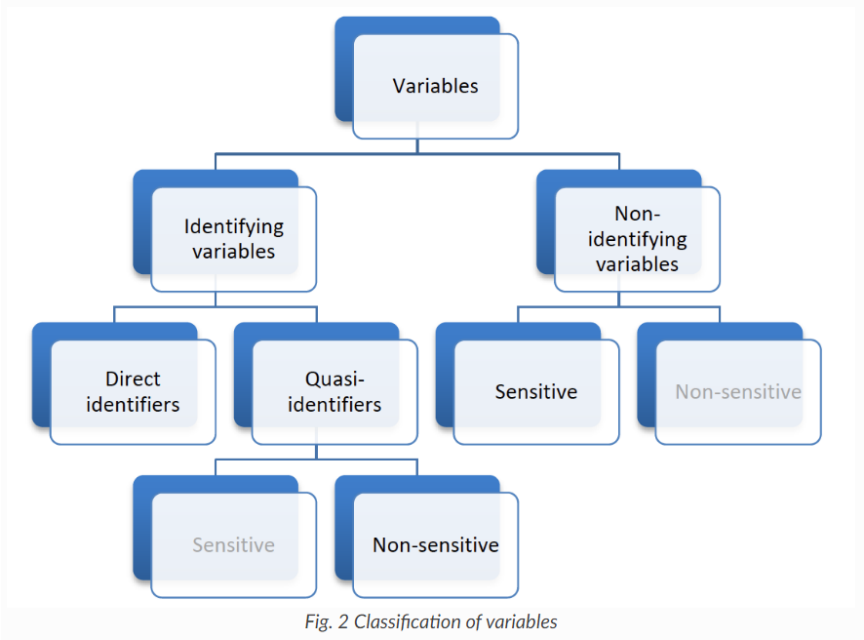


図 3 属性の分類（IHSN (2019) Fig.2 より）

識別性や機密性の分類は、線引きが難しいことに注意が必要である。例えば、住所について、番地まで含まれていれば個体の特定は容易であるため識別子になりうるが、それが市町村までである場合は準識別子止まりかもしれない。国単位の情報しかなければ、非識別子と言える。他の準識別子との組み合わせや外部参照情報によってその匿名性は変化するため、匿名化ミクロデータ作成にあたっては個別具体的な検討を行った上で評価基準を定める必要がある。

また、識別性と機密性の概念は必ずしも排他的ではないことに注意が必要である。例えば、職業は準識別子として扱われることが多いが、それが特殊な公安職であった場合、同時に機密属性にもなりうる。逆に収入のように、多くの場合それだけでは識別子になりえない機密属性も、極端に収入額が大きい場合には準識別子としての性質を有することがある。このような侵入者の知識を予測することの難

しさ、準識別子と機密属性を区別することの難しさは、Elliot & Dale (1999) や Orooji & Knapp (2018) が詳しい。

2.2.5 露見シナリオ

露見シナリオ（disclosure scenario）とは、どのような侵入者によってどのように個体情報の露見が行われるのか、そのシナリオを規定したものである。

Elliot & Dale (1999) では、侵入者の攻撃を構成する要素を動機、手段、機会、攻撃の種類、キー（マッチング）変数、ターゲット変数、データの誤差の影響など、11 種類に分類した考察が行われている。そのうち攻撃の種類については、以下の 5 種類に分類されている。また、マッチングに際してはキー変数の選定が重要となることが述べられている。

1. データベースクロスマッチ（database cross match）：外部データベースからのマッチング。
2. 単一の特定の個人を対象としたマッチング（match for a single specific individual）：特定個人の個体情報を得ることを目的としたマッチング。
3. 任意の個人へのマッチング（match for an arbitrary individual）：不特定多数に対する攻撃そのものが目的のマッチング。
4. 個人の特定のグループ（specific group of individuals）：3 の代替手段であり、属性開示を攻撃の目的とする。
5. 逆マッチング（reversed matching）：外部データベースからのマッチングではなく、外部データベースに対するマッチング。

Hundepool *et al.* (2020) によれば、機密情報の露見は、回答者が特定された後に発生する可能性があるため、個体の特定を防ぐことが重要である。識別における重要な概念はキー変数であり、侵入者が個体を識別するために使用することが想定される。個体の識別は、特定のキー変数の値の組み合わせに関して母集団の中で稀な場合に発生する可能性がある。そのため、母集団内で数の少ない特定のキー変数の組み合わせには注意が必要である。侵入者が利用すると想定したキー変数と実際に利用されるキー変数が一致するとは限らないことから、キー変数の組が母集団内で 1 レコードのみを避けたとしても、それが必ずしも安全であるとは限らない。2 レコード以上のものにも匿名化の必要がある。

露見シナリオの実用例としては、例えば、ドイツの事業所・企業における学術研究用の匿名化ミクロデータの作成を指向した Lenz *et al.* (2006) がある。ここでは、データベースクロスマッチや単一の特定の個人を対象としたマッチングを防

止することに重点が置かれている。信頼できる研究者の利用が前提であるため、任意の個人へのマッチングや個人の特定のグループの露見、逆マッチングについてはメインシナリオとはなっていない。逆に、これが一般公開型の場合、上記についても対策が必要になると考えられる。このように、どのような匿名化ミクロデータを作成にするか、どのような形で公開するかによって、重視すべき露見シナリオは変化することに注意が必要である。

2.3 匿名化手法

匿名化ミクロデータ作成のための匿名化手法を概説する。匿名化手法は**非攪乱的手法**と**攪乱的手法**に大別される。非攪乱的手法は、データ構造を歪めることなく、特定の値の区分の再編や秘匿、削除によってデータの詳細をマスクする手法である。一方の攪乱的手法は、値を変更（攪乱）して不確実性を付与することにより、露見リスクを制限する手法である。これらはそれぞれ、主に質的属性に対して行われるもの、量的属性に対して行われるもの、それらの区別なく行われるものが存在する。

公的統計のミクロデータにおける匿名化手法について、海外ではこれまで多くの研究が行われ、実用化されてきた。一方、わが国でのミクロデータ作成においては、非攪乱的手法が主に用いられてきた。2020 年 12 月現在、匿名データにおける攪乱的手法の活用例は、国勢調査におけるスワッピングに限られる（独立行政法人統計センター (2014) ）。研究例については伊藤他（2014）や稲葉（2017）、Ito *et al.* (2018) などがあるが、その数は多いとは言えない。ミクロデータとして使用される統計調査のデータ特性や公開方法によっては、非攪乱的手法だけでは秘匿性と有用性の両立に限界があることから、諸外国では攪乱的手法が広く活用されている。特に、事業所・企業系の匿名化ミクロデータにおいてはその重要性は大きいと考えられる。そこで、本研究では、攪乱的手法の積極的な活用を検討する。

以下では、代表的な匿名化手法を紹介する。一般的に用いられる匿名化手法は汎用ツールに実装されていると考えられるため、μ-ARGUS (Hundepool *et al.* (2020)) および sdcMicro (Templ *et al.* (2020)、 IHSN (2019)) のマニュアルに記載のある手法を中心に述べる。なお、これらのツールの詳細は 2.6 にて説明する。また、わが国の事例である総務省政策統括官（2018）、独立行政法人統計センター（2020b）、内閣官房 (2013) も参考にした。補足的に近年提案されているアプローチについても触れる。

2.3.1 非攪乱的手法

リサンプリング（resampling）：調査票のレコードのすべてを用いず、一部を抽出するレコード単位の処理である。仮に特定のレコードが標本内においてキー変数の条件により一意に定まる（標本一意）としても、母集団内で一意に定まる（母集団一意）わけでなければ秘匿性が保たれる。リサンプリングにあたっては単純無

作為抽出、層化抽出、集落抽出などの方法がある。なお、一般にリサンプリングは非攪乱的な手法として用いられるが、攪乱的手法としての利用法も存在する(Domingo-Ferrer & Torra (2001a))。

特異なレコードの削除：露見リスクの特に大きい特徴的な属性値を持つレコードを、レコードごと削除する。例えば国勢調査の匿名データでは、世帯人員の多い世帯、父子世帯、年齢差の大きい夫婦のいる世帯などのレコードは削除の対象として考慮されている。どのようなレコードが特異であるかは母集団の分布によるほか、その組み合わせや特徴的な属性の外観識別性にも大きく影響を受けるため、調査ごとの個別具体的な検討が重要となる。また、削除するレコードによっては要約統計量が大きく変化する可能性があることに注意が必要である。

識別情報の削除：対応関係を特定する危険性の高い識別情報である、世帯や居住地を直接的に特定できるような情報（氏名、住所、世帯員数、性別、住宅の大きさ等）を削除する。

ミクロデータのソート：ミクロデータの配列順を並べ替えることでランダムにし、対応関係を探り出すことができないようにする。

局所秘匿（local suppression）：リスクの大きい特定の属性値の組み合わせによるセンシティブな属性値を局所的に秘匿する。例えば、あるレコードの職業が、特定の地域、性別において非常に稀であるなら、その職業は局所秘匿される（地域や性別が考慮されることもある）。わが国でのミクロデータではレコード削除やリコーディングが選択されるが、例えばドイツの匿名化ミクロデータには活用例がある。

トップ・ボトムコーディング（top and bottom coding）：分布の上部や下部のような、対応関係を特定できる可能性が高くなる特殊な属性をまとめる方法である。順序関係が必要となるため、連続変数や順序尺度の属性値が対象となる。値の大部分が分布の中心にあり、裾部分のレコード数が少ない場合には、有用性を損なわずに秘匿性を高められるため、特に有用である。年齢を「85 歳以上」や「15 歳未満」でまとめる例や年収額の上位をまとめる例は、様々な調査で見られる。わが国の個人・世帯系の匿名データでは、トップコーディングにおいては母集団全体の0.5%程度（少数の特定の集団を対象とする場合は 3~5%程度）が目安となっている。

グローバルリコーディング（global recoding）：特定の値をグループ分けして階級区分に変更する手法である。大域的再符号化（再格付け）とも呼ばれる。トップ・ボトムコーディングは分布の裾が対象であったが、グローバルリコーディングは分布を問わず再編の対象となる。市区町村を県に、年齢各歳を年齢 5 歳階級に、職業分類において農林漁業をひとまとめにするなどの例がある。わが国でも海外でも頻繁に使用されている。

ローカルリコーディング（local recoding）：グローバルリコーディングはすべてのレコードで共通して属性の区分を荒くするが、ローカルリコーディングは特定のレコード群の属性値のみリコーディングの対象とする。局所的再符号化（再格付け）とも呼ばれる。局所秘匿が属性値を削除するのに対し、こちらは幅を持たせて開示する。わが国の匿名データで利用例はないが、研究例は存在する（Takemura (2002)）。

2.3.2 攪乱的手法

事後ランダム化法（post randomization method = PRAM）：PRAM は Gouweleeuw *et al.* (1997) によって開発された。伊藤他 (2018) によると、PRAM は、個票データの各セルの値をあらかじめ決められた遷移確率行列に基づいて遷移させる「攪乱」と、攪乱された個票データから原データが持つ分布を推定する「再構築」と呼ばれる 2 つのステップから構成される。オランダやデータベースの分野で研究事例がある。リコーディングが質的属性の階級を荒くするのに対し、PRAM は質的属性の値を一定の確率に基づいて攪乱するという違いがある。

（データ）スワッピング（data swapping）：データスワッピングとは、ミクロデータに含まれるレコードあるいは属性の組み合わせ同士で属性値群を入れ替える手法であり、データスワッピングは PRAM の一種と見なすこともできる (Willenborg (2001))。 伊藤他 (2018) では、露見リスクが相対的に高いレコードに絞ってスワッピングを行うターゲット・スワッピング (targeted data swapping) と、スワッピングの対象となるレコードを無作為に選んだ上でスワッピングを適用するランダム・スワッピング (random data swapping) の比較が行われている。世帯・人口系のミクロデータに対し、地域の入れ替えが行われることが多い。アメリカやイギリスだけでなく、わが国の国勢調査でも利用実績があり、2020 年現在、匿名データ作成において唯一使用されている攪乱的手法である。

ノイズ付加（adding noise）：ノイズ付加とは、量的属性に対してノイズ成分を加算または乗算することで攪乱を行う手法である (Duncan & Pearson (1991))。ノイズの平均が 0 であれば、ノイズ付加後の平均を保ったまま外部参照情報からのリンケージを防止できる。一方、ノイズの大きさによっては、特に外れ値に対して近似的なマッチングが可能になる可能性や、属性値の分散や属性間の相関等に影響を与える可能性に注意が必要である。ドイツの匿名化ミクロデータ作成では、特にパネルデータに対して乗法ノイズが用いられているケースがある。

ランクスワッピング（rank swapping）：Moore (1996) によって開発されたランクスワッピングは、データスワッピングの一種であり、量的属性や順序尺度に対して定義される。属性値に順序（ランク）付けを行い、ランクの変化率が一定の範囲を超えないようにランダムに入れ替えることで攪乱を行う。データスワッピン

グのスワッピング基準にさらにランクの概念を加えることにより、多変量の相関等の維持を可能としている。

シャッフリング（shuffling）：シャッフリングは Muralidhar & Sarathy (2006) によって提案された、回帰モデルを使用して属性のスワッピングを行う手法である。原データの属性のランク付けと、回帰モデルによって得られた属性の推定値のランク付けを照らし合わせてシャッフルを行う。ランクスワッピングは確率的な手法であるが、シャッフリングは決定論的手法である。周辺分布を維持するが、計算量はやや大きいという性質がある。

ミクロアグリゲーション（microaggregation, micro-aggregation）：ミクロアグリゲーションとは、 Defays & Nanopoulos (1993) によって提案された、ミクロデータを閾値 k 個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や中央値といった代表値に置き換える手法である。 伊藤 (2009) によると、量的属性によるミクロアゲリゲーションには、ソートキーとなる特定の量的属性に着目する単一軸法、主成分分析を応用して多変量のソートキーを第一主成分とする第 1 主成分法、標準化された属性値群の総計値（Z スコア総計値）に基づいてソートする Z スコア総計法、量的属性の各々について個別にソートする個別ランキング法[3]、個別データの分布特性に即した形でグループのレコード数を探索的に設定する Ward の階層区分法といった手法が存在する。また、質的属性については、順序変数に対して個別ランキング法を適用したスネーク法や、ソートの尺度にエントロピーの計測を用いる方法が挙げられている。質的属性のミクロアグリゲーションにおいては、中央値を用いる手法も提案されている（Torra (2004) ）。

近年では、MDAV (maximum distance to average vector) がよく用いられている。MDAV とは、Domingo-Ferrer & Mateo-Sanz (2002) で述べられた多変量固定サイズのミクロアグリゲーションをもとに、 Hundepool *et al.* (2003) で実装されたアルゴリズムである（Domingo-Ferrer & Torra (2005))。複数の量的属性の平均ベクトルを求め、探索的にアグリゲーションを行うヒューリスティックなミクロアグリゲーションの一手法である。質的属性や平均以外の演算子、また様々な距離関数で計算可能な MDAV-generic (Domingo-Ferrer & Torra (2005) ）、I-多様性を考慮した MDAV（Ting-ting *et al.* (2008) ）、合成データを活用した MDAV (Domingo-Ferrer & González-Nicolás (2010) ）、名義尺度に対する MDAV (Martínez *et al.* (2012) ）、準識別子の分布を考慮した MDAV (Abidi *et al.* (2020)) などの応用も多数考案されている。また、距離測定にユークリッド距離ではなく、

[3] 個別ランキング法は、イタリアやドイツの事業所・企業系匿名化ミクロデータの作成でも用いられている、本研究で注目すべき手法のひとつである。詳細は 3.3 で述べる。

多変数間の相関を考慮したマハラノビス距離を用いる MDAV は、外れ値への頑健性が高いという研究例もある（Templ & Meindl (2008) ）。

イタリアやドイツの匿名化ミクロデータ作成においてはミクロアグリゲーションが実際に活用されている。先行事例や先行研究が多いことから、本研究ではこのミクロアグリゲーションを中心とした攪乱を考察する。

ラウンディング（rounding）：ラウンディングは、量的属性に対して切り捨てや四捨五入等を行う手法である。シンプルな手法であるため、他の攪乱的手法と組み合わせて使われることが多い。イタリアの CIS の SUF 作成で利用実績がある。

JPEG (joint photographic experts group)：画像圧縮技術を応用して、ミクロデータを非可逆的な JPEG 圧縮・解凍プロセスによって変換された画像と見なし、攪乱を行う手法である。Domingo-Ferrer & Torra (2001a) によって開発され、Jiménez *et al.* (2014) でも研究されている。

合成データ（synthetic data）：原データの特定の性質を保存した合成データを生成する方法する手法である（Muralidhar & Sarathy (2008) ）。事業所・企業系の調査のように秘匿性と有用性のバランスを取ることが難しいデータには、この合成データを用いたアプローチが取られることもある。アメリカでは Synthetic Longitudinal Business Database (= SynLBD) （Vilhuber *et al.* (2013) ）、オーストラリアではリモートアクセスへの応用（O'Keefe & Shlomo (2012) ）などの例がある。

これらの匿名化手法を様々な観点で比較した研究例も存在する。Domingo-Ferrer & Torra (2001a) では、量的属性については加法ノイズ、確率分布によるデータの歪み（Data distortion by probability distribution）、攪乱的手法としてのリサンプリング、ミクロアグリゲーション、JPEG、ランクスワッピングが、質的属性についてはトップコーディング、ボトムコーディング、グローバルリコーディング、PRAM といった匿名化手法の総合評価が行われ、ランクスワッピングやミクロアグリゲーション、トップコーディングが有望であることが示されている。

Mateo-Sanz *et al.* (2004) では、量的属性の外れ値に対する匿名化手法の比較として、JPEG、ランクスワッピング、加法ノイズ、リサンプリング、ミクロアグリゲーションが評価され、リサンプリングや加法ノイズの結果は思わしくなく、逆にランクスワッピングやミクロアグリゲーションは有望であるという実験結果を得ている。

Templ & Meindl (2008) では、ノイズ、ランクスワッピング、ミクロアグリゲーション、シャッフリングといった量的属性の攪乱的手法について、特に外れ値に対する頑健性に着目して評価を行い、マハラノビス距離を用いた MDAV ミクロアグリゲーションである RMDM (robust mahalanobis distance based

microaggregation）や、クラスター化されたデータに対する主成分分析法に基づくミクロアグリーションである clustpppca、MDAV ミクロアグリゲーション、頑健なシャッフリング（robust shuffling）といった手法から良好な結果が得られたとしている。

2.4 評価手法

ミクロデータに対する匿名化技法の適用可能性を検証するためには、匿名化されたミクロデータの秘匿性が守られているのか、あるいはどの程度有用性を保っているのか、その定量的な評価が必要である。こうした定量的な評価によって、ミクロデータの作成における判断材料として有益な数量情報を提示することが可能となる。なお、評価手法の良し悪しはミクロデータのデータ特性に左右されるため、適切な使い分けや複数の評価指標での比較が重要である。本節では評価手法の代表的な例をサーベイする。

2.4.1 秘匿性評価

匿名化ミクロデータの秘匿性を評価する指標としては、**k-匿名性（k-anonymity）**（Samarati & Sweeney (1998)）がよく用いられる。同じ属性値の組み合わせを持つレコードが、どの組み合わせについても必ず k 個以上存在する時、そのミクロデータは k-匿名性を満たすと言う。閾値である k の値を変化させることで、匿名化の強度を変更することが可能である。k-匿名性は ID 開示と関わりの深い概念であり、匿名化ミクロデータの作成においては、秘匿性の観点から k-匿名性に違反しないようにリコーディングが行われることがある（Domingo-Ferrer & Torra (2005)、 Ichim (2007)）。

k-匿名性は個体が特定されないことへの保証を与えるが、同様のキー変数を持つ個体がすべて同じ属性を持っている場合、センシティブな属性の露見が発生する可能性がある。例えば同一の性別、年齢、地域の属性を持つ個体が 3 人存在し、いずれもガンの病歴があった場合、3-匿名性は守られてもセンシティブな病歴の属性は保護されないことになる。これを考慮する概念が、**l-多様性（l-diversity）**である（Machanavajjhala *et al.* (2006)）。同一の属性値の組み合わせにおいて、l 種類以上のセンシティブな属性が存在している時、l-多様性が満たされる。l-多様性は属性開示に関わる概念である。l-多様性に着目した匿名化ミクロデータ作成の研究例としては、Li *et al.* (2007)や Ting-ting *et al.* (2008) などがある。

伊藤他（2014）では、様々な匿名化技法を用いて作成した秘匿処理済データの秘匿性の定量的な評価手法のサーベイや実証実験が行われている。その中でも代表的なものとして、レコードリンケージとクロス集計表による評価方法を以下に整理する。

量的属性群に対して行われた匿名化の秘匿性の強度を定量的に評価する方法として、まず**レコードリンケージ（record linkage）**による秘匿性評価があげられる。レコードリンケージとは、原データのレコードと秘匿処理済データのレコードとの間で対応付け（真のリンク）が可能かどうかを判定することによって、秘匿性の強度を定量的に評価する手法である。一般に真のリンクとなるレコードの割合が秘匿処理済データにおける秘匿性評価のための指標として用いられる。

レコードリンケージの一手法である**確定的リンケージ（deterministic record linkage）**とは、対応付けを行うためのキーとなる属性群（リンクキー変数）を用いて、原データと秘匿処理済データに含まれるそれぞれのレコード同士が１対１で照合するかどうかを判定する方法である。原データのレコードと秘匿処理済データのレコードにおいてリンクキー変数の属性値がすべて一致した場合、そのレコードは真のリンクであると判定される。

一方、原データと秘匿処理済データにおけるレコード上の属性値が完全に一致しない場合でも、２つの属性値における近似の程度を判定することによって秘匿性を評価する手法もある。**interval disclosure**（Domingo-Ferrer & Torra (2001a)）とは、秘匿処理済データにおいてレコードの属性値を中心とした一定の区間（interval）を設定し、原データにおいて対応するレコードの属性値が、設定した区間の範囲内に存在するかどうかを確認する方法であり、順位統計量に基づいた区間（rank-based intervals)と標準偏差に基づいた区間（standard deviation-based intervals）といった例が存在する。

確定的リンケージに対して、**距離計測型リンケージ（distance-based record linkage）**は、原データと秘匿処理済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。最初に、秘匿処理済データのレコードから原データの各レコードへの距離を計測し、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であり、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される。距離の算出にあたっては、属性ごとの平均や分散の違いを考慮して標準化される。ユークリッド距離や相関を考慮したマハラノビス距離といった距離が用いられることが多い。

さらに、**確率的リンケージ（probabilistic record linkage）**と呼ばれる手法もある。これは、原データと秘匿処理済データの全てのレコードの組み合わせ（ペア）を考え、各ペアがリンクされる集合またはリンクされない集合のどちらに属するかを、属性値の一致基準及び確率値にしたがって分類する方法である。原データにおけるレコードと秘匿処理済データに含まれるレコードの全てのペアを対象に、２つのレコード間における各属性値の一致の程度に関する情報に基づいて真のリ

ンクであるレコードを判定する。確率的リンケージでは、量的属性と質的属性のいずれについても、秘匿性の相対的な評価を行うことが可能なことが特徴的である。

質的属性については、レコードリンケージ以外にも、**クロス集計表**による秘匿性の評価も考えられる。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ことを指向しており、クロス集計表を用いることによって、原データと秘匿処理済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を確認することができる。度数 1 となるセルに含まれる個体は特定化リスクが高いと考えられるため、この減少率を把握することで秘匿性の評価を行う。

2.4.2 有用性評価

匿名化ミクロデータにおける有用性は、①ミクロデータの公開情報（基本統計量等）、②匿名化手法の相対評価のための指標、③ユーザーの使い勝手やニーズ、といった複数の意味合いが考えられる。ここでは、②匿名化手法の相対評価のための指標に焦点を当てて有用性評価について述べる。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、 伊藤他（2014）が詳しい。量的属性に対する統計指標、質的属性に対する距離の計測、情報エントロピーを用いた有用性の評価について、伊藤他（2014）のサーベイを参考に、その他の研究例も交えつつ以下にまとめる。

量的属性に対する有用性の評価手法としては、**統計指標**を用いた例がある。平均、分散等の基本統計量、分布上の特性、情報量損失について比較・検証を行う(Domingo-Ferrer & Torra (2001b)) 。情報量損失は、秘匿処理済データが原データと比べてどの程度情報量を失っているかを算出したものであり、原データと秘匿処理済データに含まれる属性値の差や、分散共分散行列や相関係数行列に見られるデータ構造の変化によって、情報量損失の計測が行われる。情報量損失の大きさについては、平均平方誤差(mean square error)、平均絶対誤差(mean absolute error)、平均変化率(mean variation)といった尺度で評価が行われる。情報量損失の値が 0 に近いほど、原データと秘匿処理済データは近似しており、有用性が相対的に高いと判断される。なお、平均変化率は分母に攪乱前の値を取るため、撹乱前の値が小さい場合は相対的に平均変化率が大きく評価される、あるいは 0 の場合は計算できないという問題がある。それらの対策として、分母に属性値の標準偏差を用いた IL1s という評価指標も考案されている（Mateo-Sanz, *et al.* (2004)]）。また、Domingo-Ferrer & Torra (2001a) のように、属性値、分散共分散行列、相関係数行列等の差の平均変化率を一本の式でまとめて考慮した情報量損失の指標を活用している事例もある。

属性値間の距離（distance for categorical variables）を定義して計測を行う手法が考えられる。Domingo-Ferrer & Torra (2001b) では、質的属性の場合、順序尺度については質的属性値の変化幅を属性値の分類区分の数で除した値で、名義尺度に関しては属性値が変化した場合の質的属性値間の距離を 1、属性値が変化しなかった場合の距離は 0、とそれぞれ定義されている。Takemura (2002) では、順序尺度と名義尺度の両方の質的属性が含まれるデータについては、上記で定義した距離を結合した上で、相対的な属性の重要度を考慮し、それに応じた重みを付けた指標を作成する手法も検討されている。また、ミクロアグリゲーションのようにクラスタリングが必要となる匿名化手法については、クラスター内の二乗誤差を表す SSE (sum of squares errors) やデータセット全体の二乗誤差を表す SST (total sum of squares) といった指標が用いられることがある（Domingo-Ferrer & Mateo-Sanz (2002) ）。SSE を SST で除した値は、情報量損失として有用性評価に用いることが可能である。

質的属性に対しては、**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有用性の評価手法も考案されている（Kooiman *et al.* (1998) ）。リコーディングといった匿名化技法の適用によって属性値が変化する際の移行確率（transition probability）を、シャノン情報量の期待値を用いて情報エントロピーとして算出し、情報エントロピーが計測された対象となるレコード数を乗じることによって、情報量損失を求める手法である（De Waal & Willenborg, (1999) ）。なお、リコーディングは量的属性に対しても行うことが可能であるため、量的属性・質的属性のいずれにおいても評価が可能である。

その他、モデルベースでの有用性評価も考えられる。佐野・服部（2020）では、グローバルリコーディングによって秘匿されたデータに対しても評価可能な、モデルの判別精度にもとづいた複数の情報損失評価指標を提案している。適合率 (precision)、再現率（recall）、F-値（F-value）、正確度（accuracy）といった指標で有用性の評価が行われている。

2.4.3 総合評価

匿名化ミクロデータを作成するにあたり、秘匿性と有用性のバランスを評価するための総合評価が行われることがある。

複数の匿名化を実施し、それらの相対的な秘匿性と有用性を視覚的に確認したい場合、**R-U マップ（(R-U confidentiality map)**（Duncan *et al.* (2001) ）が用いられる。R-U マップとは、risk (秘匿性) および utility (有用性) について何らかの客観評価尺度それぞれを軸に取り、匿名化手法やその強度を変化させることで、秘匿性と有用性がどのように変化するかをプロットしたものである。R-U マップを解釈するにあたってはまず、ミクロデータを公開する機関が許容できる秘匿性

の閾値を定め、その閾値を満たす匿名化手法の中でも最も高い有用性を示すものが選択されることとなる。R-U マップの活用例としては、例えば伊藤他（2014）では、risk には度数 1 の減少率、utility には情報量損失率を取ることで、8 種類の匿名化技法の組合せを R-U マップにプロットし、その評価を行っている。その他、risk と utility について、現在ポートフォリオ理論における効率的フロンティアの概念を援用する研究例も存在する（Li & Li (2009)、Kim *et al.* (2015)）。

また、秘匿性と有用性をひとつの計算式でまとめ、スコアとして算出して総合的に評価する総合評価指標も存在する。Domingo-Ferrer & Torra (2001a) では、量的属性に対しては、情報量損失、距離計測型リンケージ、確率的リンケージ、インターバルディスクロージャーに重みづけしたスコア指標を、質的属性に対しては、確率的リンケージ、属性値、クロス表、情報エントロピー基準のランクを考慮したスコア指標が提案されている。このスコアの考え方は、Nin *et al.* (2008) や Jiménez *et al.* (2014) といったのちの研究でも活用されている。

2.5 匿名データ作成の流れ

総務省政策統括官（2018）や総務省統計局（2017）では、主に世帯・人口系の匿名データ作成の一般的な流れがまとめられている。ここでは、わが国の匿名化ミクロデータの一形態である匿名データの作成方法について、その概要をまとめる。

匿名データの作成及び提供は、学術研究の発展や、高等教育の発展に資することを目的に、統計法第 35 条及び第 36 条の規定に基づいて行われる。総務省統計局では、統計調査を通じて得られた情報を、特定の個人・法人等が識別されないように匿名化処理を行って提供している。

匿名化処理の考え方として、基本的には、調査単位とミクロデータの対応関係を特定されないようにするということが前提となる。そのためには、調査単位とミクロデータの対応関係の特定の可能性を高めるような識別情報を削除・低減する。特定の試みを防ぐためには、利用目的を限定し、データの管理を適正に行わせることを義務付けることも重要である。匿名化処理技法としては、以下のようなものがあげられているが、論理的な可能性だけでなく、実際には、秘匿の必要性や利用面も考慮して現実的な判断の下で決定していることが述べられている。

- 識別情報の削除
- 匿名データの再ソート（配列順の並べ替え）
- 識別情報のトップ（ボトム）・コーディング
- 識別情報のグルーピング（リコーディング）
- リサンプリング
- スワッピング
- 誤差の導入

匿名化の基準については、調査票情報の特性は統計調査ごとに異なることから、各統計調査について一律に設定することは困難であるとされている。その一方で、その大まかな目安は示されている。その概略を表 3 にまとめる。世帯・人口系の匿名データ作成における基準が事業所・企業系のミクロデータにそのまま適用できるとは限らないが、その作成方法の目安や背後にある考え方を把握しておくことは重要である。

表 3 匿名化処理の目安（総務省統計局（2017）別紙 3 より要約）

1 地理的情報について
(1)地域内に最小でも人口50万人以上
(2)サンプリング情報も同基準に適合
(3)人口50万人未満の地理的情報を提供する場合は匿名化の程度を高める
(4)外部情報による特定の種類の施設の特定を防ぐ
2 個人・世帯の識別情報について
(1)直接的に識別できる情報は削除
(2)間接的な識別子はトップコーディング、グルーピングまたは削除
トップコーディングにおいては、母集団全体の0.5％が目安
(3)少数の特定の集団を対象とする場合、トップコーディングの基準は3～5％が目安
(4)トップコーディングした項目は、平均値や中央値の公開が望ましい
(5)世帯単位の特定を防ぐ匿名化を必要に応じて
3 誤差（ノイズ）
(1)適当な匿名化技法がない場合、有用性を損なわない範囲で誤差を付加
(2)誤差を加える場合、ランダムノイズ、スワッピング、ブランク置換または補定
4 リサンプリング
ミクロデータを全て提供するとリスクが高まるため、一部のみの提供を考慮すべき
5 外部ファイルとのマッチングの可能性
(1)外部の既存ファイルとの突合を防ぐ措置を取らなければならない
(2)調査の標本フレームが提供されている場合、回避措置をとらなければならない
6 その他の問題
(1)一連番号や並び順による推測を防ぐため、削除、付替え又は並べ替えをするべき
(2)地理的情報以外でも特定の地域や集団が明らかになる場合は削除すべき
(3)秘匿の必要性の高い調査項目はそれ自体のグルーピング、削除等が必要
(4)提供時期は調査時点から最低限2年は離すべき

2.6 匿名化ツール

匿名化ミクロデータの作成にあたって、複雑な処理や繰り返しの試行が必要となることが多い。この際、統計実務の観点から、匿名化における定型処理と臨機応変な非定型処理をワンストップに行えるツールの利用が望ましい。わが国ではそういったツールが実務で利用されているか明確でないが、海外では特定のツールの利用を明示して

いるケースも存在する。そこで、諸外国で用いられている匿名化ミクロデータ作成用ツールのサーベイとして、代表的なものである μ-ARGUS、sdcMicro 等の概要を紹介する。

2.6.1 μ-ARGUS

μ-ARGUS とは、オランダ統計局が開発した安全なミクロデータの作成を支援するソフトウェアパッケージである（Hundepool *et al.*（2020））。ARGUS は、"Anti-Re-identification General Utility System"の頭文字であるが、ギリシャ神話に登場する 100 の目を持つ巨人アルゴスになぞらえて、ミクロデータの安全性を守るものという意味合いが込められている。

開発は Windows 7、JAVA 7、SPSS 22 で行われており、第 4 次フレームワーク SDC プロジェクトや第 5 次フレームワーク CASC（統計的機密性の計算側面）プロジェクト、2006 年の CENEXSDC プロジェクトや ESSNet-SDC プロジェクトの一部として更新が行われている。CENEX と ESSNet プロジェクトは、いずれも Eurostat の支援を受けている。CASC プロジェクトや μ-ARGUS 開発の経緯は瀧敦弘（2003）が詳しい。

μ-ARGUS は、図 4 に示すように、メタデータを適用してキー変数の組み合わせごとの頻度の取得、個々のリスクやデータセット全体のリスクの測定、各種非攪乱的手法・攪乱的手法、レポート作成の機能などを備えている。このほか、世帯系ミクロデータ作成のために世帯をキー変数として取り扱う機能も有している。GUI を有しているため、統計的開示抑制手法以外に求められる専門的な知識は少ない。

μ-ARGUS は匿名化ツールとしては最も有名で広く普及している。オランダ統計局等での利用実績があるため、あらかじめ実装されていないような複雑な匿名化処理や高速なリアルタイム処理を求められない限りは、わが国の公的統計のミクロデータ作成にも活用できる可能性があると考えられる。なお、集計表作成用ツールである τ-ARGUS についてはわが国でも過去に研究報告がなされている（独立行政法人統計センター（2006））。

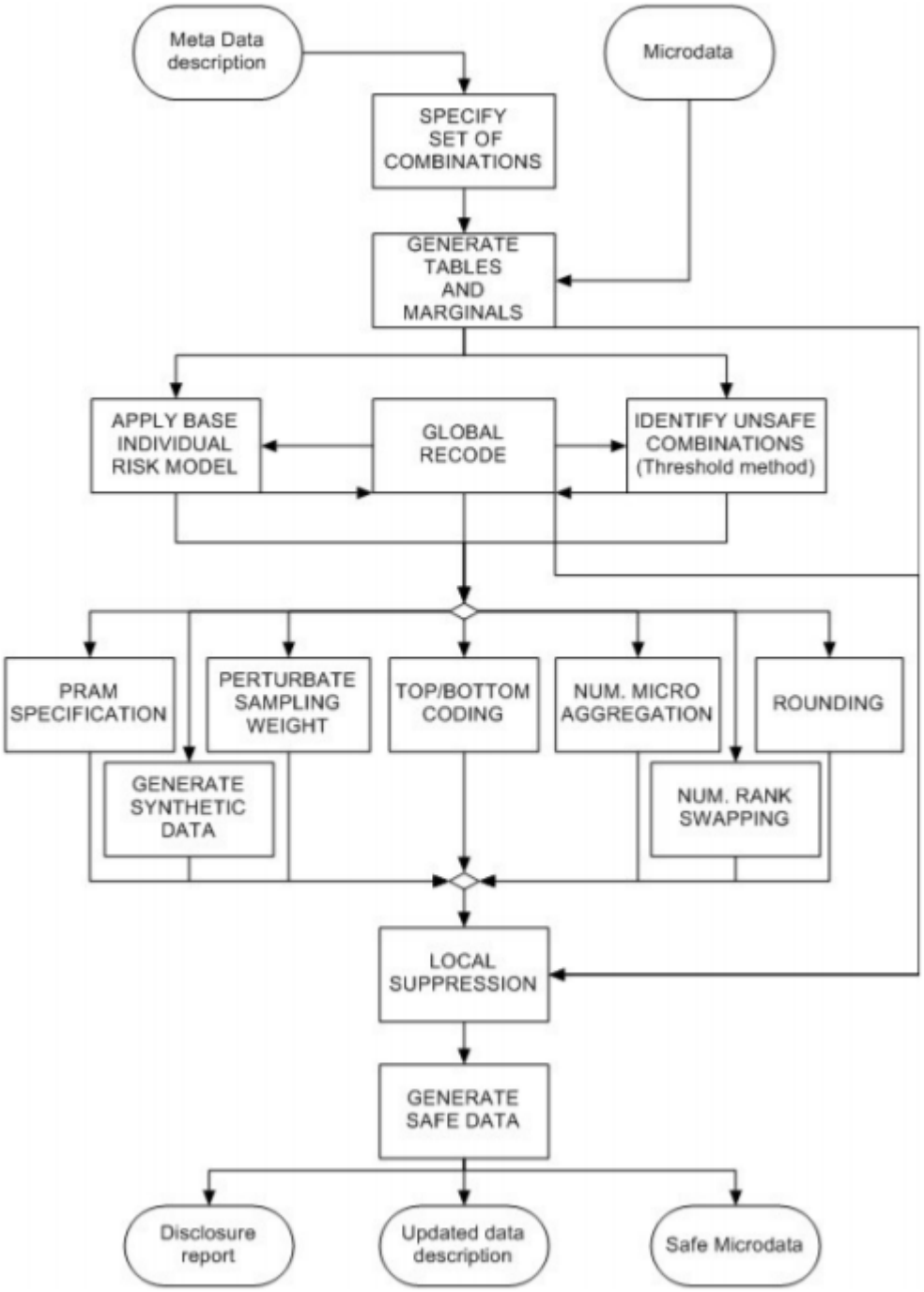


Figure 15. Functional design of μ-ARGUS

図 4 μ-ARGUS の機能デザイン（Hundepool *et al.*（2020）より）

2.6.2 sdcMicro

sdcMicro とは、研究利用や公共利用に向けた匿名化ミクロデータを生成するための R 言語ベースのフリーのオープンソースパッケージである（IHSN (2020b) ）。世界銀行が International Household Survey Network（IHSN）やオーストリア統計局等と連携して開発を行っている。さまざまな非攪乱的手法・攪乱的手法や、秘匿性評価・有用性評価のためのメソッドが実装されている。派生のパッケージ sdcMicroGUI では、sdcMicro のさまざまなメソッドをグラフィカルなユーザーインターフェイスで提供している。また、simPop (Templ *et al.* (2017)) では、合成データの作成も支援している。

sdcMicro の紹介については、Templ (2007) や Templ *et al.* (2015) が詳しい。本パッケージは、SUDA2 アルゴリズム、個別リスクアプローチ、対数線形モデルを用いたリスク推定など、リスク推定のための一般的な統計的開示手法が実装されている。また、グローバルリコーディング、ローカルリコーディング、PRAM、ミクロアグリゲーション、相関ノイズの追加、シャッフリングなどの匿名化手法も実装されている。これらは探索的、対話的かつユーザーフレンドリーに利用可能である。

すべての演算結果は、全関連情報を含む構造化された S4 クラスオブジェクトに保存される形式となっている。クラス 'sdcObMicroj' のオブジェクトは、匿名化メソッドが適用されるたびに自動更新される。Print メソッドと Summary メソッドが定義されており、露見リスクとデータの有用性の状態を確認可能である。さらに、R の機能を用いてユーザーの定義した評価基準でオブジェクトに保存することができる。その他、匿名化とそのデータの品質とリスクに対する効果を要約したレポートを自動的に生成することも可能である。多くのメソッドは演算速度の速い C++で効率的に実装されており、μ-Argus に比べて大規模なデータを高速で処理することが強みのひとつである。表 4 では、μ-Argus や IHSN が提供するアプリケーションと比較した、sdcMicro のバージョンごとの機能の差異がまとめられている。sdcMicro1.0.0 の段階では機能に限りがあったが、sdcMicro4.3.0 以降では一般的な匿名化手法や評価手法の多くが利用可能となっている。

本論文では、経済センサスを用いた実証実験において、この sdcMicro の一部の機能を活用することとした。上記のように演算速度が速いことと、個別ランキング法や MDAV 法といったミクロアグリテーションの各種手法が実装済みであること、R の元々の機能やその他のパッケージを併用することもできることがその理由である。汎用性が高いため、わが国の統計機関でもミクロデータ作成への活用が期待される。

表 4 統計的開示抑制ツールの比較（Templ *et al.*（2015）より）

Method \ Software	μ-Argus 4.2	sdcMicro 1.0.0	sdcMicro > 4.3.0	sdcMicroGUI > 1.1.0	IHSN
Frequency counts	✔	(✔)	✔	✔	✔
Individual risk	✔	(✔)	✔	✔	✔
Individual risk on households	✔		✔	✔	✔
l-diversity			✔	✔	✔
SUDA2			✔		✔
Global risk	✔		✔	✔	✔
Global risk with log-lin mod.			✔		
Recoding	✔	(✔)	✔	✔	(✔)
Local suppression	(✔)	(✔)	✔	✔	(✔)
Swapping	(✔)	(✔)	✔		✔
PRAM	✔	✔	✔	✔	✔
Adding correlated noise		✔	✔	✔	✔
Micro-aggregation	✔	(✔)	✔	✔	✔
Shuffling			✔	✔	
Utility measures	(✔)	✔	✔	✔	
GUI	(✔)			✔	
CLI		✔	✔		✔
Missing values	✔		✔	✔	✔
Cluster designs	✔		✔	✔	✔
Large data			✔	✔	(✔)
Reporting	✔		✔	✔	
Platform independent		✔	✔	✔	✔
Free and open-source		✔	✔	✔	✔

Table 1: List of methods supported by different statistical disclosure control software. Ticks are in brackets when only limited support is provided to a method. A comparison to version 1.0.0 of **sdcMicro** (released May 29, 2007; published in Templ 2008) is given to indicate the progress of the new complete reimplementation of the package.

2.6.3 IHSN

IHSN (2020a) は、ミクロデータファイルの匿名化を目的とした露見リスクの評価や削減のための実用的なガイドラインとツールを作成している。統計的処理用ツールである Stata、SPSS、および SAS 向けに文書化された専用のプログラムを作成することで、(1) 開発とサポートの持続可能性の懸念、(2) 不十分なドキュメント、(3) ユーザーフレンドリーの欠如、(4) パフォーマンスの問題と大規模な調査データセットとの関連性といった問題に対処してきた。IHSN では、最適なパフォーマンスをサポートするために、Stata 8、9、10、SPSS 16+、および Windows / Linux で動作する C ++用のプラグインが開発されている。このプラグインは、近年、学術研究や公的統計の分野で R 言語の高速化に貢献する R 言語のパッケージである sdcMicro と連携しており、関連メソッドの共有化等が図られている。

2.6.4 ARX

ARX（2020a）は、主に生物医学の分野で使用される個人データを匿名化するための包括的なオープンソースソフトウェアである。商用のビッグデータ分析プラットフォーム、研究プロジェクト、臨床試験データの共有、学習用等の用途が想定されている。

Prasser & Kohlmayer (2015) によると、ARX は、(1) 再識別リスクを分析するためのモデル、(2) リスクベースの匿名化、(3) k-匿名性、l-多様性、t-近接性、δ-存在性などの構文規則的なプライバシー基準、(4) 有用性の自動・手動評価、(5) 一般化、秘匿およびミクロアグリゲーションを使用した直感的なコーディングモデルを提供する。市販のハードウェアでも数百万のレコードを含むデータセットを匿名化できることや、包括的なグラフィカルユーザーインターフェイスを提供していることも強みとして挙げられている。

上記のように、公的統計のミクロデータ作成は想定されておらず、海外の文献でも利用実績は見当たらなかった。そのため、本論文では ARX は活用していない。なお、公式サイトから関連する匿名化ソフトウェアの一覧を確認することも可能である（ARX (2020b) ）。

3　事業所・企業系の匿名化ミクロデータ

本章では、事業所・企業系の匿名化ミクロデータに特化したサーベイの結果を概説する。はじめに事業所・企業系の匿名化ミクロデータの現状について述べたのち、世帯・人口系との差異をまとめる。また、イタリアやドイツでの研究例や実際の提供例を紹介する。最後に、それらの知見から得られた事業所・企業系の匿名化ミクロデータ作成における要点をまとめる。

3.1　事業所・企業系の匿名化ミクロデータの現状

海外において、事業所・企業系の匿名化ミクロデータについては、Eurostat やイタリア、ドイツなどわずかな作成事例しか存在しない（イタリアやドイツの事例については後述する）。これは、秘匿性を保ったまま有用性を保つことが技術的に容易ではないことや法制度面の問題等が考えられる。近年ではオンサイト利用やリモートアクセスにシフトしている傾向にあり、事業所・企業系のミクロデータもそちらで取り扱われることが多い（伊藤（2018a））。

わが国においても、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提供されているが、事業所・企業系の統計調査はこれに含まれておらず、未提供となっている。

しかしながら、事業所・企業系の匿名化ミクロデータにも需要は存在すると考えられる。消費活動や生産活動の多くは事業所や企業が担っており、経済的事象を分析するにあたって、世帯・人口系調査からでは得られない様々な知見が得られるからである。その具体的な用途としては、学術研究への活用や高等教育への活用などが考えられる。

研究利用では、研究者がミクロデータを活用し、探索的な研究を行うことが想定される。現状、事業所・企業系のミクロデータを取り扱う方法は極めて限定されており、公的統計調査の個票を活用したい場合、原則として要件の厳しい、統計法 33 条に基づく調査票情報の提供を受けなければならない。一方、調査票情報ではなく、統計法 35 条、36 条に基づく匿名データであれば、利用要件の緩和が期待できる。匿名データの利用を入口にした、オンサイト利用の促進にもつながることが期待される。

さらに、高等教育目的での活用が考えられる。秋山他（2012）では、事業所・企業系の統計調査のミクロデータを利用する機会が乏しいことが将来の課題として示されている。2020 年現在でも、わが国で高等教育に活用できる国単位の公的な事業所・企業系の匿名化ミクロデータは存在しない。独立行政法人統計センターが提供する一般用ミクロデータとして、世帯・人口系調査である全国消費実態調査や就業構造基本調査が利用できるが、事業所・企業系の統計調査はやはり未提供である。また、自

治体や民間企業が提供できる情報も限定的であり、教材としての事業所・企業系の匿名化ミクロデータは貴重な存在となっている。昨今、統計教育の重要性が叫ばれているが、事業所・企業系の教育用目的のミクロデータの存在は、その一助になると考えられる。

さらに、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、厚生労働省主幹の事業所・企業の調査である賃金構造基本統計調査の匿名データの提供可能性について言及がなされており、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も論点にあげられている。匿名データ化の手法が確立している世帯調査の手法を準用できる可能性のある個人票の提供を優先的に検討するとされているが、対となる事業所票の匿名データ化の検討も今後行われる可能性がある。

以上のように、わが国の公的統計において事業所・企業系の匿名化ミクロデータは作成されていないが、その需要は存在すると考えられる。

3.2 事業所・企業系と世帯・人口系調査の差異

前述のように、事業所・企業系の匿名化ミクロデータにも需要が存在する一方で、国内では実用化されず、世界的に見ても稀なのは、事業所・企業の統計調査の特有の難しさがあるからである。以下に、サーベイで得られた知見を述べる。

O'Keefe & Shlomo (2014) では、個人に関するミクロデータの特性と企業に関するミクロデータの差異を以下のようにまとめている（表 1）。事業所・企業のデータはサンプルサイズが小さいこと、調査に大企業が含まれていること、そしてその大企業の属性値のほとんど外れ値であるという事実から、大企業に含まれる情報の秘匿性が問題であるという指摘がなされている。

表 5 世帯・個人に関するデータの特徴と企業に関するデータの特徴
(O'Keefe Shlomo (2014) Fig.1 を参考に再編)

	個人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
レコードの対象	個人	企業
母集団に含まれる個体が 標本にも含まれている可能性	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
属性間の分布	-	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企業は外れ値

また、Lenz *et al.* (2006) では、個人・世帯系のミクロデータよりも、事業所・企業系のミクロデータの方が秘匿性や分析妥当性の維持が難しいことが説明されている。事業所・企業系の調査では、一般に母集団が小さく、個々のグループに含まれるレコード数（セルに含まれる度数）は小さい。量的属性の分布は極端に不均質である。また、サンプリングの対象となるレコード数は企業規模ごとに大きく異なり、サンプリングにあたっては悉皆で抽出される層も存在する。企業にはデータの公表義務があるため、侵入者（intruder）は精度の高い入手可能な外部情報の取得が容易である。さらに事業所・企業系のデータの露見（disclosure）に伴うリスクは、個人・世帯の調査における露見リスクより大きいことが述べられている。

その他、Hafner *et al.* (2019) は、事業所・企業系のミクロデータの場合、最も重大なリスクは偶発的に外れ値が特定されることにあるとしている。星野 (2010) では、大きな企業や事業所は全数調査されるために調査された事が既知となることや、外れ値として目立つことがビジネスデータの匿名化の難しさとして挙げられている。Franconi & Ichim (2007) でもサンプリングの問題が指摘されているほか、売上高や輸出のような変数は一般的に非常に歪んだ分布を持っており、外部情報と照合されるリスクが高まることを説明している。

3.3 先行事例・先行研究

これらの課題を踏まえ、先進事例や先行研究ではどのようなアプローチが取られてきたのかを概説する。

3.3.1 イタリア

ISTAT（イタリア国立統計研究所）のサイトでは、ミクロデータの分析フェイズにおける開示管理（disclosure control）の記載が存在する（Istat (2020a)）。Elementary data は、調査の設計、実施、監督、修正の段階を経て、統計調査の最終成果物として定義され、これを基にしてミクロデータの作成を行う。ミクロデータの保護の方法としては、グローバルリコーディング、局所的な秘匿、データの攪乱を用いることが明記されている。匿名化されたミクロデータの公開方法には、研究用ミクロデータファイル（Microdata Files for Research = MFR）、一般公開用ファイル PUF (mIcro.STAT)、さらに、elementary data の分析のための機密保持が確保された施設（ADELE）を用いる等の方法があげられている。

ISTAT では 2020 年現在、CIS (Community Innovation Survey) の PUF (Istat 2020b)) および MFR (Istat (2020c)) が提供されている。Franconi & Ichim (2007)、 Ichim (2007)、Ichim (2008)、Ichim (2009) では、CIS の研究用ミクロデータの作成方法やその普及方法について考察されている。ここではこれらの

サーベイに基づき、SUF 作成の場合のミクロデータ作成や普及方法の概要を紹介する。

CIS は、企業のイノベーション活動を調査する EU 内のサンプリング調査である。各々の企業について、経済活動（NACE）、地理的位置（NUTS）、従業員数（EMP）、売上高（TURN）、イノベーションと研究への支出（RTOT）といった項目のほか、イノベーションを決定する要因や阻害要因、高等教育を受けた従業員の数、登録特許の数など、 イノベーションの様々な側面についても調査を行っている。この調査では、他国との比較可能性（適用される定義の違いが、地理的・時間的な統計の比較に与える影響）を考慮しており、事前に定められている閾値を満たす限りは、サンプリングデザイン、データの収集方法、重み計算、補完といった手法は国ごとの判断に任されている。CIS のミクロデータのリリースにあたってもこの原則が意識されている。

まず、1998 年から 2000 年の間に調査された CIS3 の SUF 作成手順が体系的に示されている Ichim (2007) を紹介する。大まかに以下のステップを踏むことが薦められている（図 5）。

図 5 CIS の SUF 作成手順

露見シナリオの定義（definition of the disclosure scenario）：研究目的の公開を前提とするため、研究者自身が個々の事業所・企業に関する内部情報を保有していること、意図的に外部参照情報を照合することは考慮していない。代わりに、外部参照情報（external register）と接続可能性や、特徴的な変数から偶発的に個体が特定されるケース（spontaneous identification）に注意を払っている。CIS における露見シナリオの主要な変数は、産業分類（NACE）、地域（NUTS）、従業員数（EMP）、売上高（TURN）であるが、外部参照情報との接続可能性を低減するために、いずれもリコーディング、攪乱、削除といった手法が取られている。また、イノベーションに対する総支出（RTOT）、輸出、研究開発に関与した人の数などの情報によって、偶発的な個体特定が行われる可能性があるため、調査の専門家（survey expert）のチェックが重要である。

変数の前処理（preliminary work on variables）：まず、直接的識別子になりうる名称、住所、本所所在地、初年度の売上高や従業員数は削除される。また、CISにおいては輸出額も削除されている。続いて、イタリアと EU の区分の違いを考慮したグローバルリコーディングを行い、産業分類は中分類相当まで、従業員数は 5 区分まで、地域は国全体で 1 区分にまとめられる。さらに、ミクロデータの公開日と調査日が近すぎる場合は、秘匿性を高めるために売上高（TURN）に平均を保存したランダム丸めを適用するケースもあるとされる。

リスク評価：リスクの高いレコードの特定（risk assessment: identification of units at risk）：あるレコードが他のレコードと混同された場合には識別が困難であるという前提に基づき、産業分類ごとに、対数化された売上高が類似したレコードが近傍に存在するかどうかを基準とする（国や調査によっては従業員数も含めて層化する必要であることに留意）。この判断のために、密度基準にもとづいた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) ）が用いられた。DBSCAN には、距離ベースではクラスタリングの難しい、歪んだ分布に対しても頑健であるという特徴がある。DBSCAN では、eps（基準となる半径）と minPts（密であると見なす、半径の中の最小近傍数）の 2 つのパラメータをもとに計算が行われる。クラスタリングから漏れた孤立したレコードが、相対的にリスクの高いレコードとして攪乱の対象となる。売上高は分布の歪みが大きいことから、クラスタリングにあたっては対数変換が行われた。多くの場合分布の両裾に外れ値が現れたが、層によっては中央部分にも外れ値が現れることもあった。このようなレコードの特定が可能になる点も密度ベースのアルゴリズムの強みとして挙げられている。なお、本ケースでは DBSCAN の対象となったのは売上高という単一の属性であったが、複数の変数でも適用可能であることが示されている。また、距離関数の選択が任意であることや、2 種類のパラメータを調整することで外れ値とするレコードの数を調整することが可能であることも述べられている。さらに、偶発的な識別には様々な可能性が考えられるため、高度な知識を持つ調査専門家の確認が必要であることも強調されている。

ミクロデータの保護（microdata protection）：有用性の観点から最小限の攪乱を行う。すべてのレコードに対してではなく、リスクの高いレコードのみを攪乱することが前提となる。グローバルリコーディング、ランダム丸めについては先に述べた通りである。ここではまず、最近傍のクラスタリング済みのレコードからの補完（the nearest clustered unit imputation）を行う。データの有用性の観点から最小限の攪乱を行うために、クラスタリングから漏れた外れ値の売上高を最近傍のレコードから補完を行うことで、孤立したレコードの攪乱を実施する。こちらも多変量への拡張は容易である。分布の歪みから最近傍クラスターからの補完だけ

では情報損失が大きくなる可能性があること、また k-匿名性の担保という観点から、ミクロアグリゲーションも使用される。本ケースでは単一の変数であるため、ミクロアグリゲーションの中でも相対的に撹乱の程度の小さい個別ランキング法が採用された。k-匿名性および攪乱を最小限に抑えるという観点から k=3 が選択された。すべてのレコードに対して一律にミクロアグリゲーションするだけでは安全とは言い難いが、キー変数の組み合わせごとに層化することでリスクを低減できる点が重要であると述べられている。さらに、公表済みの結果表との整合性のために、右裾の最も孤立したレコードおよび産業分類と従業者数の組み合わせに対して、売上高の加重合計の調整も行われた。以上のミクロデータの保護を通じて、売上が攪乱されるのは特定化リスクが相対的に高い孤立したレコードのみであり、クラスリングされたレコードはすべて未攪乱の値が提供されることとなる。最後に、レコードの攪乱においても、調査専門家の確認が重要であることが強調されている。

情報量損失の評価（information loss assessment）：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ている。また、売上高を分母にとったいくつかの変数の比率を使用して、データの有用性を評価した。

公開するミクロデータファイルの説明（description of the microdata file to be released）：研究者にミクロデータを公開するにあたって、それぞれの変数が未攪乱、攪乱済み、削除済みなのかを明示することが望ましい。

Ichim (2008)、Ichim (2009) では、CIS3 に続く CIS4 (2002 年から 2004 年に実施) の SUF における普及方法や、そのための手法について言及されている。大まかな内容は前述の Ichim (2007) と同様であるが、以下の点が補足されている。

研究可能性（research potential）：露見シナリオを定めるにあたって、CIS が実際にどのように利用されているのか、CIS のミクロデータの利用例についてサーベイが行われている。CIS を分析にするにあたっては、国単位のデータを用いて NACE2 桁レベル（産業分類中分類相当）で行われるのが一般的であり、これ以上のリコーディングはミクロデータとしての有用性を大きく損なうことが指摘されている。また、経済指標にあらわれる属性の相関や比率が特に重要であること、大部分の分析に加重平均が関与していることも述べられている。匿名化後にこれらを確認するだけでなく、はじめからこれらの要素を大きく乱さないような匿名化手法を講じることが重要であると考えられる。

特定（identification）：Ichim (2007) ではリスクの高いレコードの特定のために DBSCAN が用いられたが、Ichim (2008)、Ichim (2009) では新たに外れ値検出アルゴリズムの一種である局所外れ値因子法（local outlier factor = LOF）(Breunig *et al.* (2000)) が紹介されている。LOF もまた密度ベースのアルゴリズ

ムであり、半径や最小近傍数といった概念を DBSCAN と共有している。あるレコードの局所到達可能密度（local reachability density）をその近傍群の局所密度と比較することで、周囲と比べて相対的に密度が高い点や低い点を特定することができる。LOF の特徴は、レコードごとの LOF、いわば相対的な孤立度の度合いを定量的に評価できる点にあり、カットオフポイント α を定めることで、再識別のリスクのあるレコードを選択することが可能である。こうして選び出したレコードを中心に攪乱を行う点は、Ichim (2007) と同様である。

その他、レコードリンケージの実験や、EU 内でのミクロデータの普及のための提案等にも触れられている。

3.3.2 ドイツ

ドイツの連邦統計局はミクロデータの提供を行っている（Research Data Centre of the Federal Statistical Office (2020)）。そのサイト上でドイツの匿名化ミクロデータを考える上で重要な概念が記載されているため、以下にその概要を整理する。

ドイツの匿名化において、原則として公的統計のミクロデータは厳格な機密保持の対象となる。ただし、連邦統計法（BStatG）の特別規定により、特定の要件が満たされている場合は統計分析の目的でミクロデータを提供できるものとされている。ドイツの匿名性の程度には 3 段階ある。第一に、**絶対的な匿名性（absolute anonymity）**である。これは完全に匿名化されたデータであり、強力な匿名化が施されている public use file（PUF）や、高等教育のためにさらに強い匿名化が施された campus file (CF) の形で提供されている。第二に、**事実上の匿名性（factual anonymity）**である。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状態を指す（濱砂（1999））。連邦統計法に従い、事実上の匿名化がなされたデータは、科学プロジェクトの目的でのみ利用が可能となっている。データの利用にあたっては、上記の PUF や CF よりも撹乱の程度の小さい scientific use file (SUF) や、オンサイト施設での利用という選択肢がある。第三に、**形式上の匿名性（formal anonymity）**である。直接的な識別子等が削除されているのみで最も原データの性質を残しているため、リモートアクセスやオンサイト施設での利用のみが許可されている。

これらのドイツの匿名性の考え方を、日本の制度と比較したのが表 6（小林, 2011）である。わが国の匿名データに対応するミクロデータは、ドイツにおいては事実上の匿名性に基づく SUF に相当することが示されている。

表 6 日本とドイツのミクロデータの比較（小林（2011）の表 1 より）

表１　匿名化の程度からみた我が国のミクロデータ ― ドイツとの比較

日本のミクロデータ		ドイツのミクロデータ				
ミクロデータの種類	利用目的	匿名化の程度	情報損失	情報の有用性	ミクロデータの種類	利用目的
調査票情報	(注1)	非匿名化	低	高	アクセス不可	
		形式的な匿名化(注2)	↑	↑	CRDPによる利用(注3)	科学研究目的
匿名データ（国外での利用可）	・学術研究目的 ・高等教育目的 ・国際比較統計利活用事業目的	事実上の匿名化	↓	↓	SUF（国外からの購入不可）	科学研究目的
－		絶対的な匿名化	高	低	PUF（国外からの購入可） CAMPUS Files	一般汎用目的 教育目的

（注１）　新統計法第32条及び第33条によれば，行政機関等での二次利用（調査実施者内部で調査票情報を二次的に利用すること）及び行政機関等での利用と同等の公益性を有すると認められる学術研究目的に利用することができる。
（注２）　氏名，住所のような直接的識別子をなくしたもの
（注３）　CRDP（Controlled Remote Data Processing）は，我が国のオーダーメード集計に相当

以下では、前述の事実上の匿名性の考えに基づいた、ドイツにおける事業所・企業系の匿名化ミクロデータ作成について述べる。Lenz *et al.* (2006) によると、2002 年から 2005 年にかけて統計局が科学者と協力して、ドイツ連邦教育研究省 (BMBF) が後援した「**企業ミクロデータに関する事実上の匿名化**」プロジェクトが行われた。これは、企業のミクロデータ用にドイツのデータインフラストラクチャを拡張し、事業所・企業のデータを研究者が使用できるようにするものである。その試行の結果、横断的な企業ミクロデータの事実上の匿名化が達成可能であることが示されている。情報量と利用者の関心のある分析を考慮した結果、匿名化手法としてはミクロアグリゲーション、加法および乗法ノイズ、ラテン超立方体サンプリング[4]、リサンプリング、PRAM、データスワッピング等が候補として選定された。その中でも特に、確定的なミクロアグリゲーションについて詳細な検討が行われている。具体的には、異なる属性をすべてひとまとめでグルーピングする MA_COM (単一軸法) 、単一の属性ごとに個々にグルーピングする MA_IND (個別ランキング法) 、さらに、相関を元に数値属性の集合を最初にグループに細分化し、

[4] ラテン超立方体サンプリング（Latin hypercube sampling）とは、属性ごとの低次のモーメントを正確に再現した合成データを作成する手法である（Dandekar *et al.* (2001) ）。属性間の関係についてはピアソンの相関や順位相関を考慮する。

グループ内でのみまとめてミクロアグリゲーションを行う MA_GR といった手法である。これらの匿名化手法の効果を記述的統計的手法や計量経済学的手法で理論的に導出し、モンテカルロシミュレーションで結果を確認することで、SUF 作成への適正の評価が行われた。その結果、乗法ノイズに加えて、個別ランキング法である MA_IND が SUF 作成においては最も有望であることが示されている。

さらに、2006 年から 2008 年にかけて、連邦統計局の研究データセンターは各種機関と連携して、BMBF が後援するプロジェクト「**企業パネルデータに関する事実上の匿名化**」を実施してきた。これは、縦断的に（時系列的に）ミクロデータを接続し、パネルデータとして活用できる匿名化ミクロデータ作成を試行するものであり、企業のパネルデータを匿名化するにあたり、どの程度まで情報を失わずに匿名化できるかを検証することが重要な目的であるとされている。Brandt *et al.* (2008)、Lenz (2008)、Lenz & Zwick (2009) の概要を以下にまとめる。

対象調査例：匿名化する調査は、すでに年次ベースでの匿名化の研究実績があり、研究需要高いものから選ばれた。Brandt *et al.* (2008) では、製造業の事業者や賃金を調べる Monthly Reports, Survey of Investments and Survey of Small Units、製造業の生産物や付加価値を調べる Cost Structure Survey、売上税の統計である Turnover Tax Statistics、雇用動向を見る IAB Panel of Local Units の 4 種類が紹介されている。また Lenz & Zwick (2009) では、小売業を対象にした German Retail Trade Statistic、所得を調べる German Structure of Earnings Survey、職業訓練の調査である Second European Continuing Vocational Training Survey 2000 についても紹介されている。

匿名化手法：具体的な匿名化の手順には触れられていないが、事業所・企業系のミクロデータの匿名化にあたっては、非攪乱的手法と攪乱的手法の組み合わせが推奨されている。各種手法の評価には、計量経済学的の見地からモンテカルロシミュレーション等を行って検証された。その結果、ミクロアグリゲーション、加法および乗法ノイズ、ラテン超立方体サンプリング、リサンプリング、PRAM、データスワッピング等の手法が有望であるとされた。その中でも特に、分散の低下を補う個別ランキング法によるミクロアグリゲーションや、混合分布を変形した乗法ノイズの概要が示されている。また、公開されている主要な変数のみ、多重代入法によって攪乱する手法も紹介されている。ミクロアグリゲーションの考察については、Lenz (2006) が詳しい。

匿名性の測定：匿名性については、外部参照情報からのリンケージに重点を置いている。まず、データベースのクロスマッチシナリオを多基準代入問題として数学的にモデルリングし、適切なパラメータ化によって、最小化すべき目的関数を持つ代入問題に変換される。次に、この目的関数の最適な係数を選択する。具体的なア

プローチとしては、従来の距離ベース（conventional distance based approach）、相関ベース（correlation based approach）、分布ベース（distribution based approach）、共線性（collinearity approach）の 4 種類がある。このうち、従来の距離ベース以外はパネルデータとして時系列を考慮している。

Lenz (2008) では、これらの複数のアプローチではそのミクロデータの特性を部分的にしか測定できない欠点を補うために、各指標に係数で重み付けをして和を取る hybrid matching、各指標を and 条件や or 条件で定式化して算出する composite matching で評価する手法が提案されている。

なお、Lenz & Zwick (2009) では、ミクロデータの活用において必要とされる機密性の程度は、主に利用者が決定するデータアクセスの方法に依存することが指摘されている。オンサイト利用、オフサイト利用（SUF、PUF、CF 等）、リモートアクセスといった利用手段が存在するが、それらは一般に相互に排他的なものではなく、むしろアプローチの適切な組み合わせにより、特定の要件に応じて匿名化手法を適応させることができる。

以上のような試行を経て、ドイツでは現在、実際に事業所・企業系の匿名化ミクロデータの提供が行われている。表 7 に、Research Data Centre of the Federal Statistical Office (2020) で検索可能であった事業所・企業系の匿名化ミクロデータの例をまとめた。TOPICS はその統計調査のジャンルを表している。ミクロデータの名前については、ドイツ語表記と参考までに英語表記を併記した[5]。CF、PUF、SUF のうち、いずれかが提供されているものをまとめている。この表からわかるように、ドイツでは農業、教育、金融、工業など、様々な領域の統計調査について、その性質に応じて CF、PUF、SUF といった様々な形式での提供が行われている。また、調査によっては単年度のミクロデータだけでなく、複数の年度を経時的に接続したパネルデータの提供が行われていることも特徴的である。

[5] 元がドイツ語表記しか見当たらない調査名については、機械翻訳を用いて英語名を補足している。

表 7 ドイツで提供されている事業所・企業系の匿名化ミクロデータの例

TOPIC	Microdata (English, German)	CF	PUF	SUF
Agrarian	AFiD panel agricultural structure AFiD-Panel Agrarstruktur	○		
	Agricultural Census - Main Survey Landwirtschaftszählung - Haupterhebung	○		
Education	European survey on continuing vocational training Europäische Erhebung zur beruflichen Weiterbildung	○		○
Finances	Annual debts of the core budgets of the municipalities / municipal associations Jährliche Schulden der Kernhaushalte der Gemeinden/Gemeindeverbände			○
	Accounting results of the core budgets, the cameral/ double booking extra households and the cameral/ double-entry other public funds, institutions and companies Rechnungsergebnisse der Kernhaushalte, der kameral/ doppisch buchenden Extrahaushalte und der kameral/ doppisch buchenden sonstigen öffentlichen Fonds, Einrichtungen und Unternehmen			○
Other economic statistics	Structure of earnings survey Verdienststrukturerhebung	○		○
	Salary and wage structure survey in the manufacturing industry and in the service sector Gehalts- und Lohnstrukturerhebung im Produzierenden Gewerbe und im Dienstleistungsbereich	○		○
Taxes	Sales tax statistics (advance notifications) Umsatzsteuerstatistik (Voranmeldungen)			○
Environment	Survey of investments for environmental protection Erhebung der Investitionen für den Umweltschutz	○		
Manufacturing	Cost structure survey in the field of manufacturing, mining and quarrying of stones and earth Kostenstrukturerhebung im Bereich Verarbeitendes Gewerbe, Bergbau und Gewinnung von Steinen und Erden	○		○
	Panel of the cost structure survey in the field of manufacturing, mining and quarrying of stones and earth Panel der Kostenstrukturerhebung im Bereich Verarbeitendes Gewerbe, Bergbau und Gewinnung von Steinen und Erden			○
	AFiD panel industrial companies AFiD-Panel Industrieunternehmen	○		

表中では省略したが、ドイツで現在提供されている匿名化ミクロデータの中には、個々の調査、年度、提供形式ごとにどのような匿名化を行ったか、その概要が公開されているものもある。例えば、企業における雇用者の賃金構造の調査である Verdienststrukturerhebung（Structure of earnings survey）では 2010 年に SUF と CF がそれぞれ作成されているが、匿名化の概要も個別に文書として用意されている。

SUF の場合（Research Data Centre of the Federal Statistical Office (2013)）、冒頭でまず、研究目的の利用を前提として、事実上の匿名化の概念に基づいて匿名

化が行われていることが明記されている。約 32,000 社と約 190 万人の従業員からデータセットが構成されており、地域情報と産業についてはリコーディングが行われている。また、従業員数については、少なくとも 500 人の従業員を抱えるすべての企業と、各地域の産業の中で最大の 3 社に対して、ミクロアグリゲーションが実施されている。また、企業と従業員が紐づくというデータセットの特殊な構造のため、企業の事業部門の露見を避けるために、従業員が行う活動の種類を匿名化することが必要な場合があったとされている。以上を踏まえて、5 つの地域、44 の産業、53 の職業グループをミクロデータとして利用することが可能である。このほかにも、年間総収益が大きい場合にはトップコーディングする、一部の経理項目は条件次第で金額ではなく割合のみを提供する、従業員の年齢はトップ・ボトムコーディングを行うなど、主要なキー変数以外にも匿名化が施されている。その他、パネルデータという枠組みではないが、2001 年や 2006 年の同調査の SUF との時系列分析にも適用可能であることが示されている。

一方、CF の場合（Research Data Centre of the Federal Statistical Office (2016)）、絶対的な匿名化の概念に基づいて大学での教育用に特別に設計されていることが明記されている。SUF と類似した匿名化が行われているが、SUF で実施された匿名化が損なわれないようにいくつかの差異がある。まず、CF の場合は層化二段抽出によるサンプリングが行われている。地域、産業、従業員数で層化した上でまず一段階目の抽出を行い、その後企業ごとに従業員数をランダムサンプリングする。二段階目の抽出の際には、秘匿性に考慮してサンプルサイズを一定にはしていない。また、地域は SUF の 5 区分に対して 2 区分まで、産業は SUF の 44 ある中分類相当の区分から 14 の大分類相当の区分まで荒くリコーディングされている。さらに従業員数については、SUF と同様のミクロアグリゲーションが行われるだけでなく、SUF との重複の可能性を回避するために、従業者規模としてリコーディングする前およびサンプルが抽出される前にミクロアグリゲーションが実施されている。その他、職業については SUF の 53 区分から 20 区分にリコーディング、週当たりの労働時間については SUF の時間表記から階級値にリコーディング、いくつかの項目が追加・削除されるなどの違いも存在している。

以上のような提供形態ごとの匿名化の考え方の違いや属性ごとの匿名化の具体的な手法は、わが国で学術研究目的や高等教育目的の匿名化ミクロデータを検討するにあたって、貴重な参考資料になると考えられる。

3.4 事業所・企業系の匿名化に向けた考察

以上のサーベイから、事業所・企業の匿名化ミクロデータの作成を考える上でいくつかの論点に整理できる。

最も重要な論点は、**大規模な事業所・企業の秘匿性**である。世帯・人口データは原則として数が多く、またレコードごとの差異が相対的に大きくないため、サンプリングが前提の処理となる。一方、事業所・企業系のデータは比較的レコード数が少なく、分布の偏りが大きい。また、無作為抽出では、規模の大きい事業所・企業系は疎らにしか抽出されないが、それらの事業所・企業は多くの場合平均や分散に大きな影響を持つため、どの事業所・企業が抽出されるかが全体の統計量に大きな影響を与える。そのため、事業所・企業系の調査では、悉皆で抽出する、規模ごとに層化抽出する、規模の大きい事業所や企業はデータの対象から除くなどの措置を講じる必要がある。加えて、**分布の歪みや外れ値**をどう取り扱うかという問題も存在する。規模別の事業所数の観点では、小規模な事業所・企業が大多数を占めるため、大規模な事業所・企業が外れ値として評価されやすい。一方、売上の観点から見ると、多くの大規模な事業所・企業が占める割合は非常に大きく、外れ値はむしろ小規模な事業所・企業となる。大規模な事業所・企業の存在は分析上の価値や、社会的な影響力も大きく、事業所数の観点からのみ外れ値と判断することには困難を伴うことが予想される。

また、**外部データとの接続可能性**も大きな課題となる。世帯・個人系の調査の場合、収入や病歴といった特定個人のセンシティブな情報が一般に公開されているケースは多くない。仮に知る者がいるとすれば、当事者と社会的・距離的に近しい人間であることが推察される。すなわち、潜在的な侵入者の数は限られている。一方、事業所・企業の情報は、売上、資本金といった情報が一般に公開することを義務付けられている。わが国においては、東洋経済新報社から発行される会社四季報（2020）、日本経済新聞社の NEEDS-FinancialQUEST（2020）といったデータサービスだけでなく、それぞれの企業のサイトの企業情報や会社概要から容易に閲覧できるケースも存在する。これは潜在的な侵入者が膨大に存在しうるということを示す。外部参照情報は事業所・企業を特定する大きな手掛かりとなるため、特定化リスクを高める結果となる。

地域情報にも注意が必要である。地理的な情報が個体の露見に繋がるケース自体は世帯・人口系のミクロデータにも存在するが、事業所・企業系の場合は特に地域と産業が深く結びついているケースが多く、従業者規模等の情報も相対的に識別性が高い。さらに、特定の地域に支社や支所として事業所を保有するような企業は、事業所の地域情報からだけでも特定化のリスクが高まりやすい。このような事情から、事業所・企業系の地域情報は、世帯・人口系よりもより一層慎重な匿名化が求められる。

さらに、**属性の数や種類**は、匿名化手法や評価手法に影響を与える。世帯・人口系の調査のように変数の数が少なく、質的属性が多い場合は、特定のキー変数に対するリコーディングやスワッピングが主に行われる。秘匿性の評価にあたっては母集団や標本に対する一意性の確認が中心となる。一方、属性の数が多く、量的属性が多く含まれる事業所・企業系のデータの場合は、量的属性に対しても匿名化を考える必要がある。秘匿性の評価にあたっては、複数の量的属性の相関性にも注意を払わなければならない。

最後に、**侵入者の動機**や**露見リスクの大きさ**にも違いがある。世帯・人口系のデータにおいては、侵入者の目的は基本的に興味を満たすためのものである。特定の個人や世帯を探し当てることや、個人情報を露見させることで自らの技術を誇示することが主な目的として考えられる。一方、事業所・企業系のデータの場合は、それらに加えて、関連企業の情報を握って悪用する、あるいはその情報を売買するという金銭的メリットが存在する。競合他社や取引先の個体情報を握ることは、競争や取引において優位性が生まれる可能性があるからである。これは、侵入者のモチベーションが強くなる可能性や、露見した際のリスクが大きくなることを示唆する。

以上のような事業所・企業系のミクロデータの特性を考慮すると、事業所・企業系の統計調査に対する匿名化をわが国でも検討しようとすれば、量的属性については、海外の事例でも見られるミクロアグリゲーションやノイズの付加といった攪乱的手法の適用可能性を追究する必要が出てくると考えられる。匿名化ミクロデータの対象となる産業や従業者規模の範囲、キー変数となる属性の選定やセンシティブな属性への対応、特異値(外れ値)の形で示される属性値の取り扱いなど、海外の事例を踏まえつつ、匿名化の対象となるレコードや属性について、データ特性に即した個別具体的な検討が必要になるであろう。

4　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

4.1　使用するデータ

事業所・企業の統計調査のうち、基幹統計のひとつである平成 28 年経済センサス - 活動調査（以下、「経済センサス」という）を用いて、事業所・企業系における匿名化ミクロデータ作成の実証研究を行う。産業大分類 E（製造業）の事業所レコードについて、従業者合計（男女計）が 1 人以上 1000 人未満等の条件[6]を満たす 414,258 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。分析対象項目には、外部参照情報になりうるキー変数、露見リスクの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中心に、以下の項目を選定した（表 8）。項目番号や項目名等の情報は、経済センサスの個別データ符号表に従っている。

表 8 分析対象項目

項目番号	項目名	変数名	符号	備考
3	都道府県番号（所在地）	K_KEN	01--47, NULL	都道府県番号
77	［事］７従業者合計（男女計）	MTX_JI_TTOTAL	0--999999, NULL	
124	補正__４給与総額	MTX_URIAGE_4	0--99999999999, NULL	
127	補正__７減価償却費	MTX_URIAGE_7	0--99999999999, NULL	
159	補正__有形固定資産（土地を除く）	MTX_YUKEI	0--9999999999, NULL	
160	補正__無形固定資産（ソフトウェア）	MTX_MUKEI	0--9999999999, NULL	
166	資本金額	KC_SHIHON	0--99999999999, NULL, V	
177	[事]産業中分類	KC_JSANGM	09--32, NULL	産業中分類番号
181	［集計用］売上（収入）金額	MTX_URIAGE	0--99999999999	
184	［事］付加価値額（円単位）	MTX_JI_FUKAKACHI	-999999999999999--9999999999999999, NULL	

4.2　記述統計量および分布特性

まず、選定した主な分析対象項目の記述統計量や分布特性を調査した。量的属性の記述統計量を表 9 に示す。計算の都合上、未記入 NULL または不詳 V の事業所は計算に含めていない。付加価値額については、他の経理項目と同じ万円表章に補正している。

売上（収入）金額、給与総額、減価償却費、付加価値額といった経理項目は、平均値と中央値との間に大きな差が生じている。また、歪度や尖度からも、分布に大きな歪み

[6] その他、結果表における売上集計対象および付加価値集計対象をいずれも満たすレコード。

があることが明らかである。資本金額についてはその傾向がより顕著であり、非常に大きな歪度や尖度を持っている。有形固定資産や無形固定資産については、複数事業所の調査事項にはなっていないほか、そもそも固定資産を持たないケースもあるため、0 が多く見られた。そのため、中央値も 0 となっている。

表 9 分析対象項目における量的属性の要約統計量

	平均値	標準偏差	中央値	歪度	尖度	標準誤差	1%点	99%点
従業者合計	18.46	54.58	5.00	8.75	99.88	0.55	1.00	248.05
資本金額	68,388.91	991,247.47	1,000.00	32.10	1,261.02	11,868.89	100.00	1,282,818.72
売上（収入）金額	60,872.70	403,633.67	3,809.00	22.34	788.82	4,036.34	0.00	1,019,616.59
給与総額	2,466.93	7,126.66	681.50	14.94	396.94	81.90	0.00	25,621.58
減価償却費	364.55	1,782.18	37.00	22.67	793.70	20.48	0.00	5,547.15
付加価値額	11,978.85	64,333.00	1,580.50	18.43	547.14	643.33	-1,096.45	184,480.48
有形固定資産	232.04	1,471.68	0.00	13.12	233.45	16.91	0.00	5,383.02
無形固定資産	4.38	58.95	0.00	20.19	471.97	0.68	0.00	70.00

続いて、質的属性、量的属性のそれぞれについて棒グラフおよびヒストグラムを作成した。質的属性として代表的な、都道府県や製造業における事業所産業中分類の事業所の度数を棒グラフとしてプロットしたものが図 6 である（符号と名称の対応は、表 10 および表 11 を参照）。都道府県番号については、東京（13）、愛知（23）、大阪（27）といった主要都市は事業所数が相対的に多いのに対して、鳥取（31）のように事業所数の少ない都道府県も存在する。地域ごとにばらつきが大きいため、県単位の情報も露見リスクに繋がることが予想される。産業中分類についても同様の指摘が可能である。金属製品製造業（24）や食料品製造業（09）の事業所数は比較的多いのに対し、石油製品・石炭製品製造業（17）や情報通信機械器具製造業（30）が占める割合は非常に小さい。これらもそのままミクロデータとして公開するにあたっては特定化リスクが強くなる恐れがあり、匿名化の必要性があると考えられる。

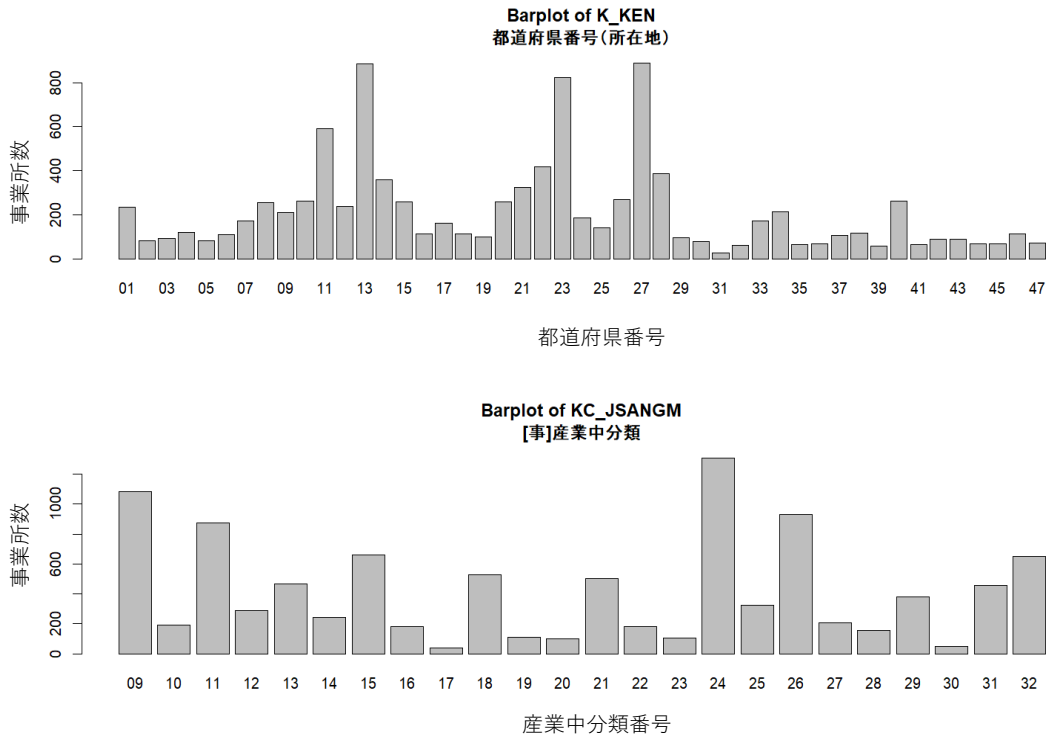


図 6 都道府県番号（上）、産業中分類（下）の棒グラフ

次に、量的属性のヒストグラムを作成した（図 7、図 8）。量的属性の多くは分布の歪みが大きいことから、横軸は対数を用いて表示している。また、秘匿上、目盛りは割愛した。従業者合計は対数軸を用いても右裾に長い分布となっている。資本金額については、会計上の都合からか、ピークが複数存在する点が特徴的である。売上（収入）金額、給与総額、減価償却費、付加価値額は概ね対数正規分布に従っていると考えられる。なお、付加価値額については、負の値を取ることもあることに注意が必要である（対数軸であるため、図中では負の付加価値額は省略）。最後に、有形固定資産および無形固定資産については、前述のように 0 や未記入が多く存在するため対数軸では頻度が小さく、また分布もまばらとなっている。

さらに、量的属性について相関係数行列および散布図行列を求めた。図 9 より、従業者合計と給与総額には 0.88、従業者合計と付加価値額には 0.78 と非常に強い相関が存在している。また、売上（収入）金額は、給与総額、減価償却費、付加価値額との間にも 0.6 を超える相関を有している。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と言える。

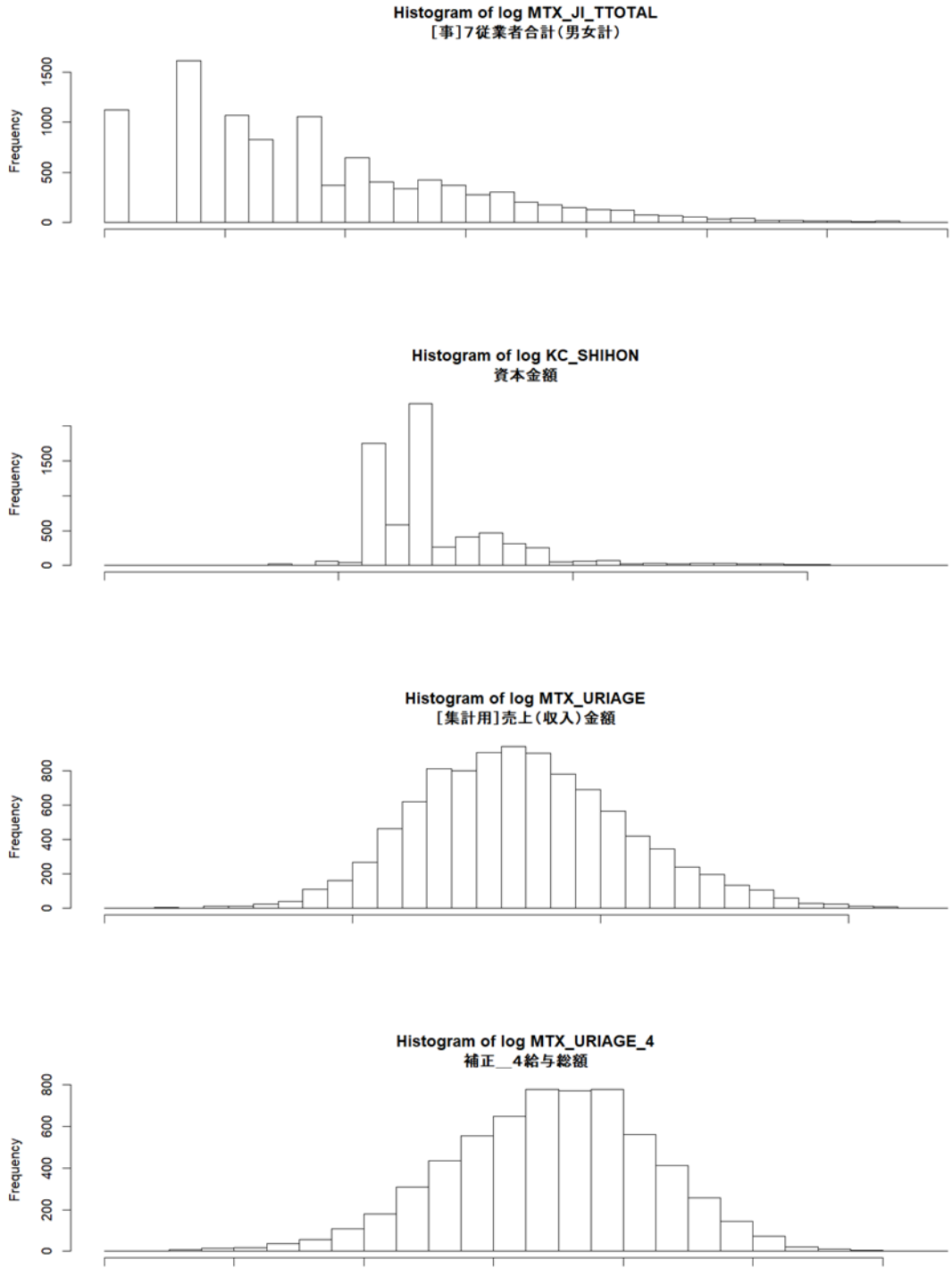


図 7 従業者合計、資本金額、売上(収入)金額、給与総額のヒストグラム

※秘匿上、横軸（対数）の目盛りは省略

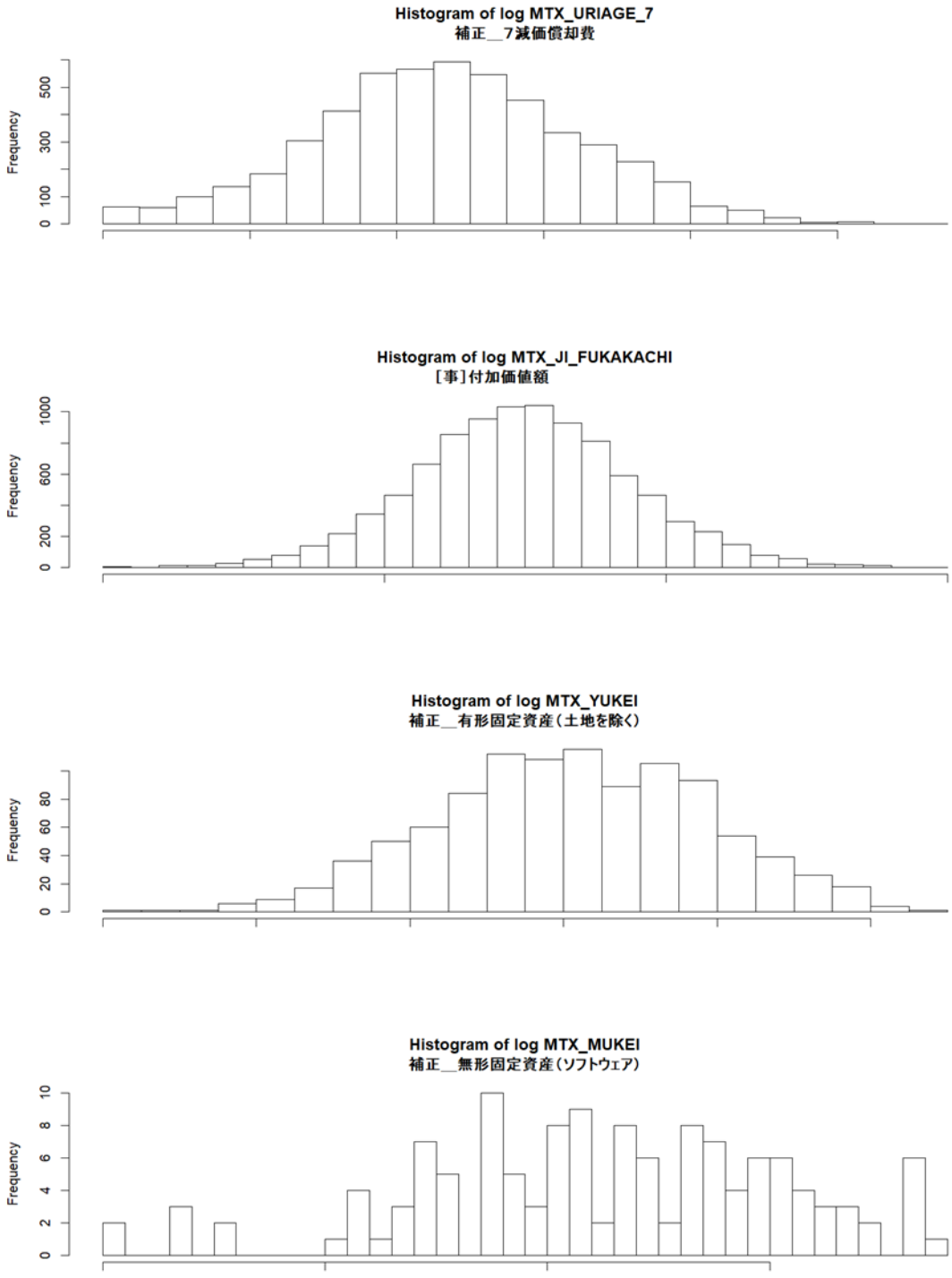


図 8 減価償却費、付加価値額、有形固定資産、無形固定資産のヒストグラム

※秘匿上、横軸（対数）の目盛りは省略

	従業者合計	資本金額	売上（収入）金額	給与総額	減価償却費	付加価値額	有形固定資産	無形固定資産
従業者合計	1.00	0.25	0.63	0.88	0.51	0.78	0.47	0.19
資本金額		1.00	0.27	0.37	0.32	0.27	0.29	0.11
売上（収入）金額			1.00	0.71	0.64	0.65	0.43	0.16
給与総額				1.00	0.50	0.69	0.45	0.20
減価償却費					1.00	0.09	0.56	0.13
付加価値額						1.00	0.33	0.16
有形固定資産							1.00	0.21
無形固定資産								1.00

図 9 量的属性の相関係数行列および散布図行列

4.3 質的属性のリコーディング

分布特性をもとに、質的属性の匿名化を考える。匿名化手法には、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作成で実績があり、わが国の結果表作成や匿名データ作成にも用いられているグローバルリコーディング（以下、「リコーディング」という）を選択した。リコーディングは、分類区分を粗くすることで秘匿性を高める手法である。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、本稿では地域、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。従業者合計および資本金額は本来量的属性であるが、リコーディング後は質的属性として取り扱うことができるため、便宜上ここに含めた。なお、これらは経済センサスにおける結果表において複数の表や区分でリコーディングが行われている属性でもあるため、その分類区分を参考にリコーディングの程度を変更したいくつかのパターンを用意した。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。

地域（表 10）については、47 都道府県をもとに、それらをまとめる区分として一般的によく用いられる 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（表 11）については、製造業における産業中分類（09～32）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、製造業について産業中分類をより粗くした区分を行っているため、これを参考に 24 区分から 11 区分への統合を図っている。従業者合計（表 12）については、結果表における 15 区分を参考に 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）にまとめ、従業者規模とした。また、それらをより荒い基準でまとめた 5 区分も用意している。資本金額（表 13）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、それらをさらに粗くまとめた 5 区分を用意した。

表 10 地域のリコーディング

47区分	度数	構成比	8区分	度数	構成比	3区分	度数	構成比
01北海道	234	2.34%	1北海道	234	2.34%	1東日本	3681	36.81%
02青森県	81	0.81%	2東北	653	6.53%			
03岩手県	92	0.92%						
04宮城県	119	1.19%						
05秋田県	81	0.81%						
06山形県	110	1.10%						
07福島県	170	1.70%						
08茨城県	255	2.55%	3関東	2794	27.94%			
09栃木県	209	2.09%						
10群馬県	261	2.61%						
11埼玉県	590	5.90%						
12千葉県	236	2.36%						
13東京都	884	8.84%						
14神奈川県	359	3.59%						
15新潟県	259	2.59%	4中部	2570	25.70%	2中日本	4613	46.13%
16富山県	113	1.13%						
17石川県	162	1.62%						
18福井県	111	1.11%						
19山梨県	100	1.00%						
20長野県	259	2.59%						
21岐阜県	323	3.23%						
22静岡県	419	4.19%						
23愛知県	824	8.24%						
24三重県	185	1.85%	5近畿	2043	20.43%			
25滋賀県	140	1.40%						
26京都府	270	2.70%						
27大阪府	889	8.89%						
28兵庫県	388	3.88%						
29奈良県	94	0.94%						
30和歌山県	77	0.77%						
31鳥取県	25	0.25%	6中国	535	5.35%	3西日本	1706	17.06%
32島根県	61	0.61%						
33岡山県	171	1.71%						
34広島県	213	2.13%						
35山口県	65	0.65%						
36徳島県	68	0.68%	7四国	349	3.49%			
37香川県	107	1.07%						
38愛媛県	116	1.16%						
39高知県	58	0.58%						
40福岡県	262	2.62%	8九州・沖縄	822	8.22%			
41佐賀県	65	0.65%						
42長崎県	89	0.89%						
43熊本県	90	0.90%						
44大分県	66	0.66%						
45宮崎県	68	0.68%						
46鹿児島県	112	1.12%						
47沖縄県	70	0.70%						

表 11 産業分類のリコーディング

産業中分類	24区分	度数	構成比	11区分	度数	構成比
食料品製造業	09	1081	10.81%	09_10	1273	12.73%
飲料・たばこ・飼料製造業	10	192	1.92%			
繊維工業	11	874	8.74%	11	874	8.74%
木材・木製品製造業（家具を除く）	12	286	2.86%	12_13_14	993	9.93%
家具・装備品製造業	13	467	4.67%			
パルプ・紙・紙加工品製造業	14	240	2.40%			
印刷・同関連業	15	659	6.59%	15	659	6.59%
化学工業	16	182	1.82%	16_17_18_19	857	8.57%
石油製品・石炭製品製造業	17	36	0.36%			
プラスチック製品製造業（別掲を除く）	18	527	5.27%			
ゴム製品製造業	19	112	1.12%			
なめし革・同製品・毛皮製造業	20	100	1.00%	20_32	749	7.49%
その他の製造業	32	649	6.49%			
窯業・土石製品製造業	21	500	5.00%	21	500	5.00%
鉄鋼業	22	182	1.82%	22_23_24	1596	15.96%
非鉄金属製造業	23	106	1.06%			
金属製品製造業	24	1308	13.08%			
はん用機械器具製造業	25	324	3.24%	25_26_27	1461	14.61%
生産用機械器具製造業	26	930	9.30%			
業務用機械器具製造業	27	207	2.07%			
電子部品・デバイス・電子回路製造業	28	154	1.54%	28_29_30	583	5.83%
電気機械器具製造業	29	378	3.78%			
情報通信機械器具製造業	30	51	0.51%			
輸送用機械器具製造業	31	455	4.55%	31	455	4.55%

表 12 従業者合計のリコーディング

13区分	度数	構成比	5区分	度数	構成比
1人	1120	11.20%	1~4人	4624	46.24%
2人	1609	16.09%			
3人	1070	10.70%			
4人	825	8.25%			
5~9人	2078	20.78%	5~9人	2078	20.78%
10~19人	1437	14.37%	10~29人	2119	21.19%
20~29人	682	6.82%			
30~49人	484	4.84%	30~99人	860	8.60%
50~99人	376	3.76%			
100~199人	181	1.81%	100~999人	319	3.19%
200~299人	61	0.61%			
300~499人	43	0.43%			
500~999人	34	0.34%			

表 13 資本金額のリコーディング

11区分	度数	構成比	5区分	度数	構成比
300万円未満	176	1.76%	1,000万円未満	2689	26.89%
300～500万円未満	1764	17.64%			
500～1,000万円未満	749	7.49%			
1,000～3,000万円未満	2816	28.16%	1,000万円～1億円未満	3739	37.39%
3,000～5,000万円未満	469	4.69%			
5,000万円～1億円未満	454	4.54%			
1～3億円未満	188	1.88%	1～10億円未満	332	3.32%
3～10億円未満	144	1.44%			
10～50億円未満	102	1.02%	10億円以上	215	2.15%
50億円以上	113	1.13%			
以外	3025	30.25%	以外	3025	30.25%

4.4 質的属性の秘匿性と有用性の定量的評価

伊藤他（2014）では、質的属性の秘匿性を評価する方法として、クロス集計表による方法が提案されている。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、原データと秘匿処理済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う。

本実験では、地域、産業、従業者規模、資本金階級の 4 項目をキー変数として扱い、その個々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、本稿でもその基準に則ることとした。また、度数 1 と度数 2 をまとめて考慮するため、セル数ではなくレコード数（＝事業所数）の割合を算出している。なお、度数 1 または 2 となるレコード数の確認することは、k-匿名性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-匿名性に反するレコード数を確認することに等しい。表 14 に、地域 3 区分、産業 11 区分、従業者規模 13 区分、資本金階級 5 区分の条件で層別に事業所数をカウントした場合のイメージを作成した（事業所数は説明のための疑似的な値）。強調されたセルに含まれる事業所が 3-匿名性に反するリスクの高い事業所である。このような事業所の数に焦点を当てて以下の実験を行う。

表 14 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

表 15 に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す。index 1 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存在するため、3,741 通りとなっている。層の数が増えるほどひとつひとつの組み合わせに含まれる事業所数は少なくなるため、3-匿名性違反のレコード数は全体の 33.75%と大きな値となる。逆に、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全体の 2.28%と比較的小さな値となる。表 15 を棒グラフとして図示した図 10 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。

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 414,258 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。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。

表 15 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

index	地域	産業	従業者規模	資本金階級	分類区分の組み合わせ	3-匿名性違反のレコード数[%]
1	8区分	24区分	13区分	11区分	3,741	33.75
2	8区分	24区分	13区分	5区分	2,797	22.20
3	8区分	24区分	5区分	11区分	2,600	22.05
4	8区分	24区分	5区分	5区分	1,741	12.29
5	8区分	11区分	13区分	11区分	2,881	23.94
6	8区分	11区分	13区分	5区分	1,993	13.65
7	8区分	11区分	5区分	11区分	1,882	14.32
8	8区分	11区分	5区分	5区分	1,139	6.38
9	3区分	24区分	13区分	11区分	2,614	21.62
10	3区分	24区分	13区分	5区分	1,746	11.47
11	3区分	24区分	5区分	11区分	1,667	12.35
12	3区分	24区分	5区分	5区分	992	5.32
13	3区分	11区分	13区分	11区分	1,829	12.83
14	3区分	11区分	13区分	5区分	1,102	5.34
15	3区分	11区分	5区分	11区分	1,081	6.40
16	3区分	11区分	5区分	5区分	582	2.28

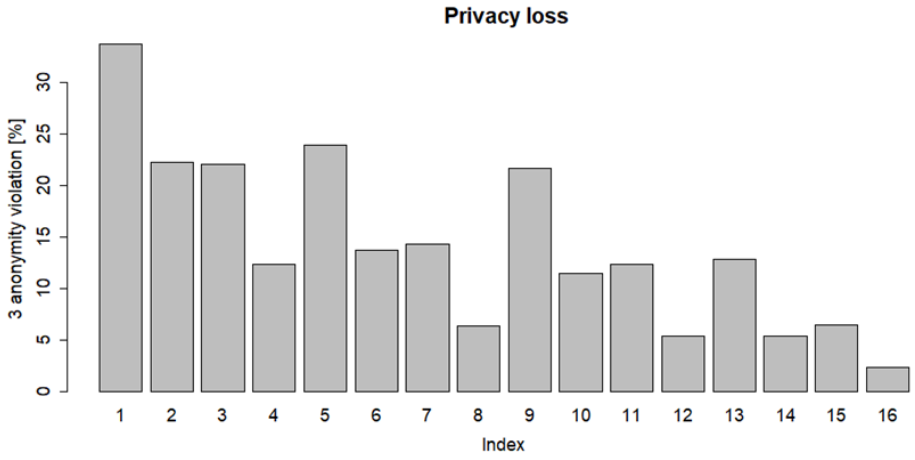


図 10 質的属性の秘匿性評価（3-匿名性違反のレコード数の割合）

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た情報量損失の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移行確率（transition probability）を用いて情報エントロピーを算出したのち、情報エン

トロピーが計測された対象となる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。さらに、情報量損失の最大値を分母に取ることで情報量損失率を算出できる。図 11 より、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が視覚的にわかる。

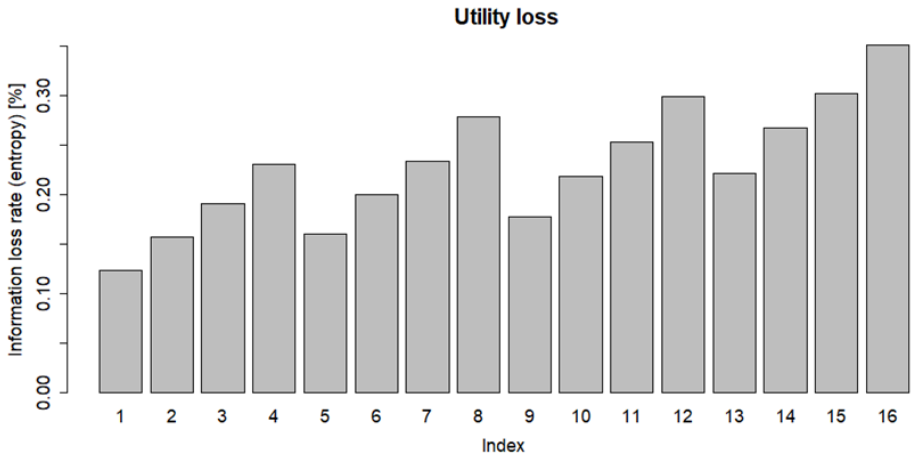


図 11 質的属性の有用性評価（情報エントロピーに基づく情報量損失率）

以上の結果を踏まえて、質的属性について、秘匿性と有用性をもとに R-U マップ（R-U confidentiality map）（Duncan & Pearson (1991)）を作成した（図 12）。横軸が秘匿性（risk の低さ）を、縦軸が情報量の損失（utility の低さ）を表しており、横軸は右にいくほど秘匿性が高く、縦軸は上にいくほど情報量の損失が大きくなる。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 12 から、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、有用性は高い領域に位置している。一方、最も荒い分類区分の組み合わせである index 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ほど秘匿性と有用性の両立できていることになるが、本実験の結果では概ねひとつの曲線上に乗っており、特定の index がそのバランスに優れているという結果は得られていない。実務においては、この有用性とのバランスを考慮しつつ、許容できる秘匿性の基準を満たす index を選択することが想定される。

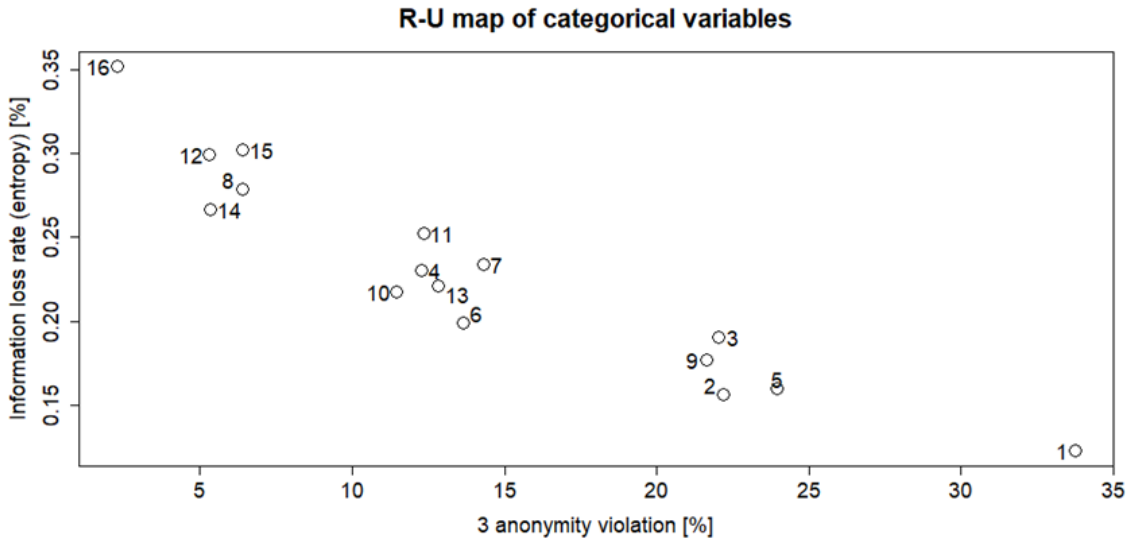


図 12 質的属性の R-U マップ

(3-匿名性違反のレコード数の割合×情報エントロピーに基づく情報量損失率)

4.5 量的属性の匿名化の検討

続いて、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、給与総額、減価償却費、付加価値額の匿名化を検討する。匿名化の前段として、より細かい分布の確認を行った。図 13 はリコーディング済みの分類区分を用いたとある層における、対数化された売上（収入）金額の分布の一例である。対数化されているため、もともと売上（収入）金額が 0 の事業所はこれに含まれていない。これより、層化を行っても売上の分布にはばらつきが見られることが多く、最大値付近のレコードが疎らであることがわかる。特にランク上位 5%のレコードが分布の範囲（range）に大きな影響を与えている。なお、これらは対数軸であるため、非対数の場合はより顕著な分布の歪みが現れることに注意が必要である。

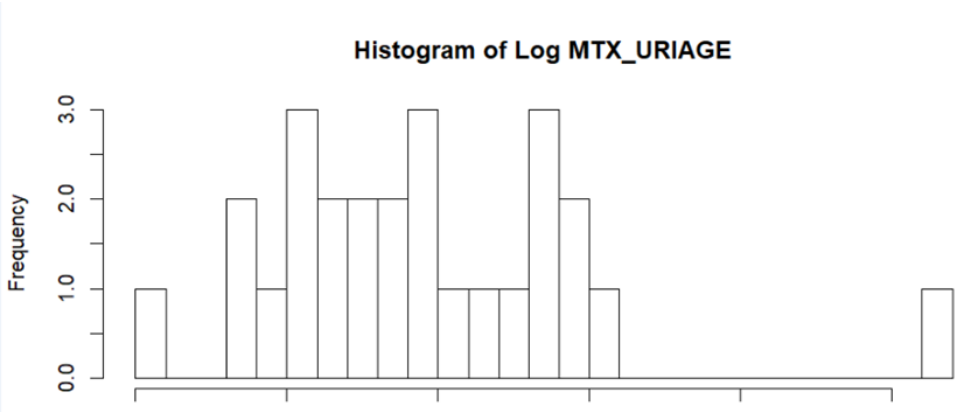


図 13 あるキー変数の組の売上（収入）金額のヒストグラムの例

※秘匿上、横軸（対数）の目盛りは省略

表 16 は、上記とはまた別の層における売上（収入）金額の、総数、上位 5%の事業所のみ、上位 5%以外の事業所のみのそれぞれについての非対数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。なお、キー変数ごとに層化した上で 3-匿名性を満たさないレコードを削除する手法については、付録 A で考察した。

表 16 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（非対数）

	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

そこで、センシティブな量的属性に対する匿名化技法として、イタリアやドイツの事例でも採用実績のあるミクロアグリゲーションを選択した。ミクロアグリゲーションとは、最初にレコード群に含まれる質的属性を用いてレコードを層ごとに分け、層内のレコードについて特定のレコード数（あるいは特定の閾値）にしたがってグループ化を行い、グループ内の量的属性値を平均値等の代表値に置き換える方法である。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、代表値には平均を維持するために平均値を採用した。ミクロアグリゲーションの手法には、イタリアやドイツで前例のある個別ランキング法と、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となっている MDAV 法の 2 種類を選択した。

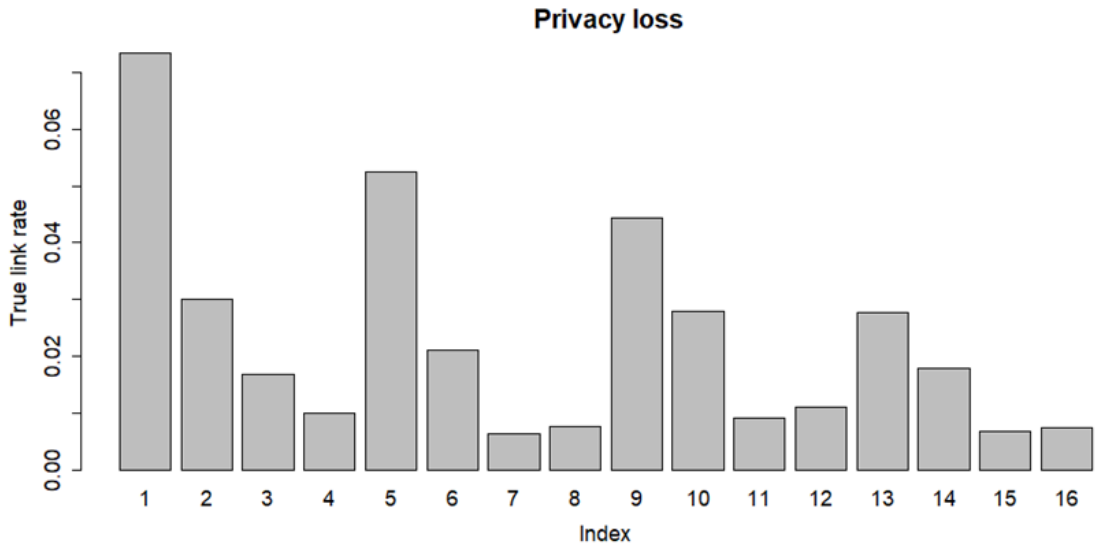
4.6 量的属性の秘匿性と有用性の定量的評価

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、原データと秘匿処理済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））。具体的には、最初に、秘匿処理済データのレコードから

原データの各レコードへの距離を計測し、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される。

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、給与総額、減価償却費、付加価値額の 4 つのセンシティブな量的属性を用いた。なお、量的属性の評価が目的であるため、質的属性であるキー変数は距離の計算に含めていない。距離計測型リンケージで使用する距離には、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、秘匿処理済データのレコードから最も距離の近い原データのレコードが真のリンクである確率（true link rate）を求めた。

個別ランキング法と MDAV 法のそれぞれについて、その結果を図 14 に示す。横軸は表 15 の index に準拠している。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、true link rate が減少する傾向にあることがわかる。個別ランキング法のほうがやや true link rate の水準は低いが、大きな差は見られなかった。



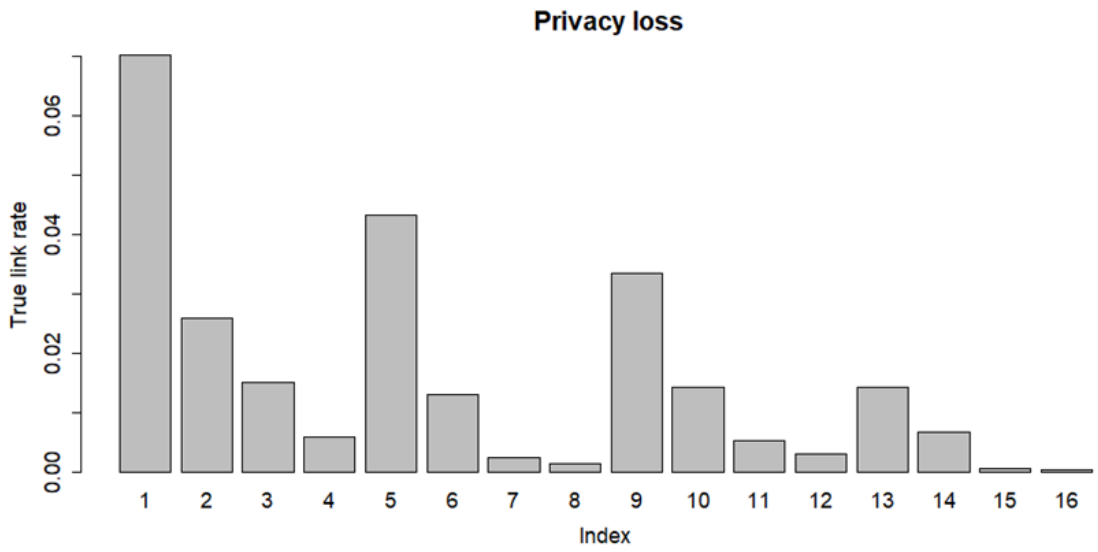


図 14 量的属性の秘匿性評価（距離計測型リンケージに基づく true link rate）
（上：個別ランキング法、下：MDAV 法）

続いて、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、伊藤他（2014）をもとに統計指標を用いた有用性の評価を用いた。原データと秘匿処理済データについて、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った。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）といった尺度を選択した。なお、平均二乗誤差（mean square error）は、平均絶対誤差と本質的に変わらないこと、桁数が多く見づらいことから割愛した。その計算式を表 17 に示す（伊藤他（2014）表 1 より）。

表 17 平均平方誤差、平均絶対誤差と平均変化率による情報量損失の算定式
（伊藤他（2014）表 1 より）

	平均平方誤差 (Mean square error)	平均絶対誤差 (Mean absolute error)	平均変化率 (Mean variation)
属性値の差	$\frac{\sum_{j=1}^{k}\sum_{i=1}^{n}(x_{ij}-x'_{ij})^2}{nk}$	$\frac{\sum_{j=1}^{k}\sum_{i=1}^{n}\lvert x_{ij}-x'_{ij}\rvert}{nk}$	$\frac{\sum_{j=1}^{k}\sum_{i=1}^{n}\frac{\lvert x_{ij}-x'_{ij}\rvert}{\lvert x_{ij}\rvert}}{nk}$
相関係数行列の差	$\frac{\sum_{j=1}^{k}\sum_{1\le i\le j}(r_{ij}-r'_{ij})^2}{\frac{k(k-1)}{2}}$	$\frac{\sum_{j=1}^{k}\sum_{1\le i\le j}\lvert r_{ij}-r'_{ij}\rvert}{\frac{k(k-1)}{2}}$	$\frac{\sum_{j=1}^{k}\sum_{1\le i\le j}\frac{\lvert r_{ij}-r'_{ij}\rvert}{\lvert r_{ij}\rvert}}{\frac{k(k-1)}{2}}$
分散共分散行列の差	$\frac{\sum_{j=1}^{k}\sum_{1\le i\le j}(v_{ij}-v'_{ij})^2}{\frac{k(k+1)}{2}}$	$\frac{\sum_{j=1}^{k}\sum_{1\le i\le j}\lvert v_{ij}-v'_{ij}\rvert}{\frac{k(k+1)}{2}}$	$\frac{\sum_{j=1}^{k}\sum_{1\le i\le j}\frac{\lvert v_{ij}-v'_{ij}\rvert}{\lvert v_{ij}\rvert}}{\frac{k(k+1)}{2}}$

n：原データと秘匿処理済データにおけるレコードの総数
k：原データと秘匿処理済データに含まれる属性の数
x_{ij}：原データ上の i 番目のレコードにおける j 番目の属性の値
x'_{ij}：秘匿処理済データ上の i 番目のレコードにおける j 番目の属性の値
r_{ij}：原データにおける i 番目の属性と j 番目の属性に関する相関係数
r'_{ij}：秘匿処理済データにおける i 番目の属性と j 番目の属性に関する相関係数
v_{ij}：原データにおける i 番目の属性と j 番目の属性に関する分散ないしは共分散
v'_{ij}：秘匿処理済データにおける i 番目の属性と j 番目の属性に関する分散ないしは共分散

表 18 に、個別ランキング法と MDAV 法のそれぞれについて、属性値の差、分散共分散行列、相関係数行列の平均絶対誤差と平均変化率を算出した。原則として、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増加している。これは原データの性質が失われていることを示唆する。個別ランキング法と MDAV 法とでは、属性値の差については個別ランキング法のほうが原データの性質を残す結果となった。一方で、相関係数行列や分散共分散行列には顕著な差は見られなかった。

表 18 分類区分を変更したキー変数の組み合わせ別の平均絶対誤差と平均変化率
（上：個別ランキング法、下：MDAV 法）

					属性値の差		相関係数行列		分散共分散行列	
index	地域	産業	従業者規模	資本金階級	平均絶対誤差	平均変化率	平均絶対誤差	平均変化率	平均絶対誤差	平均変化率
1	8区分	24区分	13区分	11区分	3614	NaN	0.008	0.12	2,593,599,473	0.15
2	8区分	24区分	13区分	5区分	4854	NaN	0.013	0.20	2,948,889,683	0.22
3	8区分	24区分	5区分	11区分	5501	NaN	0.016	0.25	3,846,385,813	0.38
4	8区分	24区分	5区分	5区分	6740	NaN	0.017	0.24	4,471,561,300	0.39
5	8区分	11区分	13区分	11区分	4320	NaN	0.010	0.15	2,731,139,890	0.17
6	8区分	11区分	13区分	5区分	5585	NaN	0.017	0.27	3,348,146,315	0.28
7	8区分	11区分	5区分	11区分	7092	NaN	0.023	0.30	5,529,749,569	0.41
8	8区分	11区分	5区分	5区分	8073	NaN	0.023	0.36	5,429,077,780	0.49
9	3区分	24区分	13区分	11区分	4518	NaN	0.013	0.13	3,024,181,841	0.11
10	3区分	24区分	13区分	5区分	5388	NaN	0.018	0.24	3,366,445,157	0.20
11	3区分	24区分	5区分	11区分	7168	NaN	0.018	0.22	4,889,050,572	0.28
12	3区分	24区分	5区分	5区分	7577	NaN	0.018	0.22	5,137,402,099	0.38
13	3区分	11区分	13区分	11区分	5603	NaN	0.021	0.64	3,438,962,222	0.59
14	3区分	11区分	13区分	5区分	6194	NaN	0.033	0.59	3,986,367,738	0.56
15	3区分	11区分	5区分	11区分	8655	NaN	0.041	0.91	6,610,230,709	0.80
16	3区分	11区分	5区分	5区分	8270	NaN	0.038	0.70	5,782,304,591	0.74

					属性値の差		相関係数行列		分散共分散行列	
index	地域	産業	従業者規模	資本金階級	平均絶対誤差	平均変化率	平均絶対誤差	平均変化率	平均絶対誤差	平均変化率
1	8区分	24区分	13区分	11区分	3781	NaN	0.008	0.12	2,595,087,548	0.15
2	8区分	24区分	13区分	5区分	5168	NaN	0.012	0.20	2,953,962,619	0.23
3	8区分	24区分	5区分	11区分	5813	NaN	0.016	0.25	3,852,600,487	0.38
4	8区分	24区分	5区分	5区分	7266	NaN	0.020	0.25	4,477,920,605	0.35
5	8区分	11区分	13区分	11区分	4572	NaN	0.010	0.15	2,732,652,054	0.16
6	8区分	11区分	13区分	5区分	6059	NaN	0.019	0.27	3,357,131,412	0.30
7	8区分	11区分	5区分	11区分	7579	NaN	0.024	0.28	5,541,498,757	0.39
8	8区分	11区分	5区分	5区分	8795	NaN	0.033	0.40	5,349,639,660	0.40
9	3区分	24区分	13区分	11区分	4820	NaN	0.013	0.14	3,028,961,194	0.12
10	3区分	24区分	13区分	5区分	5902	NaN	0.018	0.22	3,375,099,694	0.18
11	3区分	24区分	5区分	11区分	7639	NaN	0.021	0.23	4,896,341,073	0.26
12	3区分	24区分	5区分	5区分	8362	NaN	0.026	0.34	5,216,013,885	0.39
13	3区分	11区分	13区分	11区分	6006	NaN	0.022	0.64	3,445,366,274	0.59
14	3区分	11区分	13区分	5区分	6895	NaN	0.034	0.59	3,992,662,032	0.57
15	3区分	11区分	5区分	11区分	9259	NaN	0.046	0.95	6,559,192,075	0.80
16	3区分	11区分	5区分	5区分	9361	NaN	0.054	0.79	5,731,747,197	0.70

なお、いずれにおいても属性値の差の平均変化率が NaN（非数）になっているのは、原データの度数に 0 がひとつでも存在すれば、計算式上分母が 0 となって発散するためである。また、0 にならないまでも分母となる原データの度数が小さい場合には、情報量損失率が過大に評価されてしまうという問題もある。

この問題に対処するため、匿名化ツール sdcMicro の dUtility コマンドにおける IL1s メソッド（Mateo-Sanz *et al.* (2004)）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度数ではなく、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。そのため、上記のような平均変化率の問題を解消している。属性が d 個ある i 番目のレコードの場合、標準偏差 S を用いて以下のように定義される。

$$IL1s = \frac{1}{d}\sum_{j=1}^{d}\frac{|x_{ij} - x'_{ij}|}{\sqrt{2}S_j}$$

IL1s を用いて評価した有用性評価の結果が図 15 である。横軸は表 15 の index に準拠している。個別ランキング法と MDAV 法を比較した場合、わずかに個別ランキング法のほうが情報量損失は小さい傾向にある。分類区分を変更したキー変数ごとの差異は、全体の傾向に大きな差異はなかった。

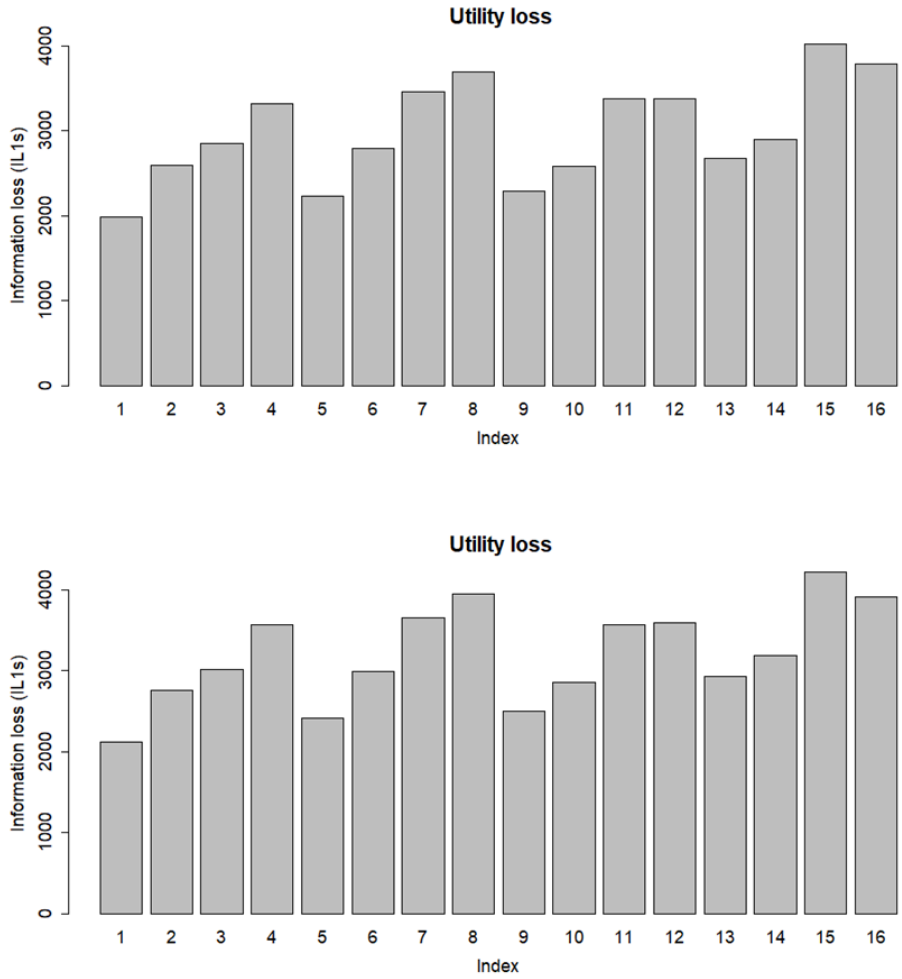


図 15 量的属性の有用性評価（IL1s）
（上：個別ランキング法、下：MDAV 法）

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。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。図 16 から、個別ランキング法、MDAV 法のいずれにおいても、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、最も荒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両立できていることになるが、本実験の結果では大きな差異は存在していない。実務にあたっては、質的属性の R-U マップと同じく、それぞれのバランスを総合的に考慮してリコーディングやミクロアグリゲーションの細部を決定していくことが重要であると考えられる。

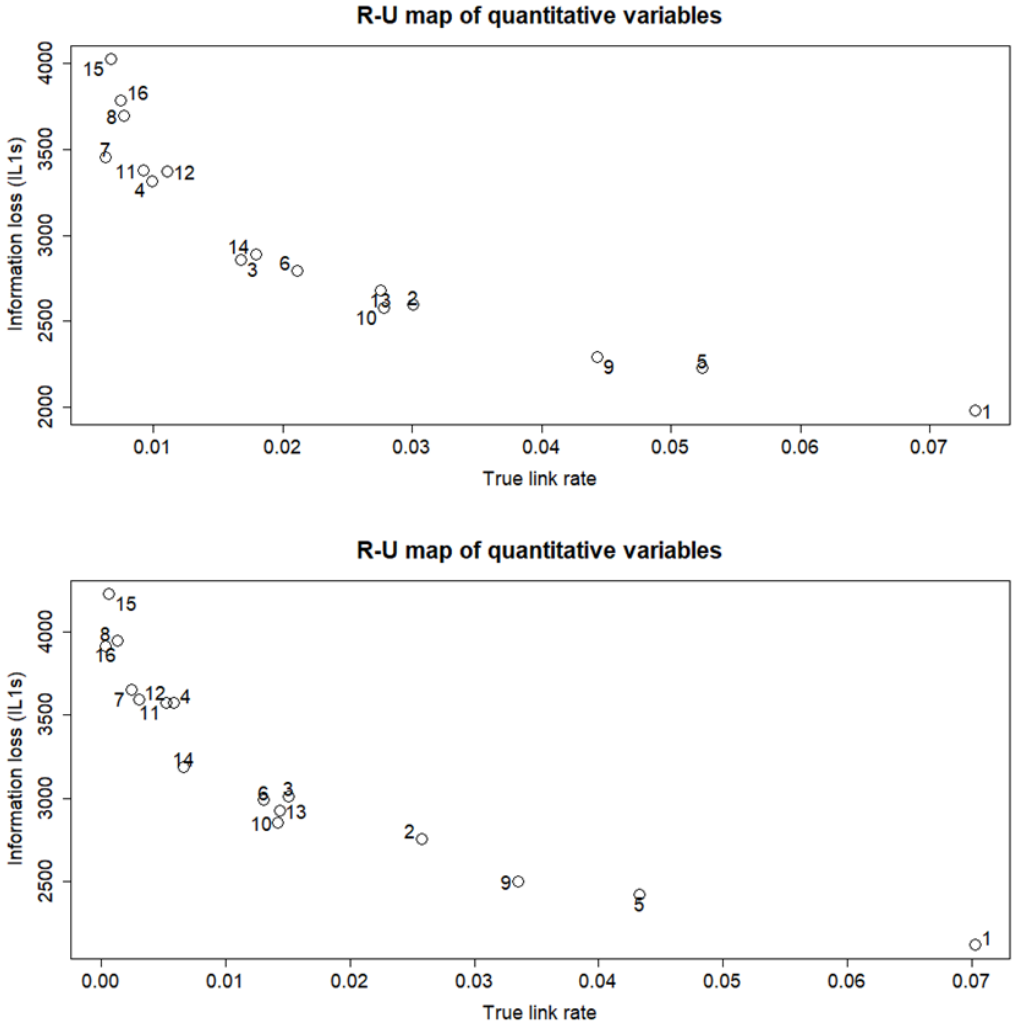


図 16 量的属性の総合評価（R-U マップ）

（上：個別ランキング法、下：MDAV 法）

5　経済センサスにおける事業所の分布特性の把握と探索的な検証

前章では、経済センサスのミクロデータに対して、先行研究に基づいた匿名化処理や評価手法を適用した。本章では、経済センサスの匿名化ミクロデータ作成に向けて、経済センサスのミクロデータとしての特性をより深く探るために探索的な実験を行う。

5.1　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。露見には、偶発的な個体の特定・識別のほかに、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現実的に困難である。そのため、前章では、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、その他にも、経営組織、単独・本所・支所の別、開設時期といった属性も準識別子になる可能性がある。また、売上（収入）金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露見リスクを高めるケースも考えられる。しかしながら、分類区分の異なる属性の組み合わせごとの事業所数を評価するにあたって、そこまでの考慮は行っていない。

さらに、前章ではリコーディングを行うにあたって、ミクロデータとしての使い勝手を優先してリコーディングの区分を用意した。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小さくできると考えられるが、一方で、一部の属性についてはリコーディング幅が粗く、必要以上に情報量損失が発生している可能性や、リコーディング幅が細かすぎて必要な秘匿性を担保できていない可能性が存在する。経済センサスのデータ特性を知り、特にどのような事業所に対して匿名化が必要となるのかを把握することで、より秘匿性と有用性の高い匿名化を検討することができると考えられる。そこで、異なる属性を基準にして事業所ごとの露見リスクを評価し、それをもとに露見リスクが相対的に高いと考えられる事業所、低いと考えられる事業所の差異を分析する。

5.2　経済センサスを用いた探索的な検証

産業大分類 E（製造業）の事業所である 414,258 レコードを母集団とし[7]、その中から 10 万レコードを無作為抽出してテストデータとした。本実験では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、売上（収入）金額階級 8 区分、経営組織 5 区分、単独・本所・支所の別 3 区分、開設時期 16 区分、に対し、2 属性ずつクロス集計を行った。それぞれの属性の区分ごとの度数と構成比は、以下の通りである（表 19）。

[7] その他、結果表での売上集計対象および付加価値集計対象をいずれも満たすレコード。

表 19 各属性の分類区分別の度数と構成比

都道府県	度数	構成比	都道府県	度数	構成比
01北海道	2396	2.40%	25滋賀県	1299	1.30%
02青森県	710	0.71%	26京都府	2841	2.84%
03岩手県	854	0.85%	27大阪府	9054	9.05%
04宮城県	1235	1.24%	28兵庫県	3990	3.99%
05秋田県	787	0.79%	29奈良県	1012	1.01%
06山形県	1135	1.14%	30和歌山県	882	0.88%
07福島県	1605	1.61%	31鳥取県	318	0.32%
08茨城県	2410	2.41%	32島根県	566	0.57%
09栃木県	2023	2.02%	33岡山県	1555	1.56%
10群馬県	2431	2.43%	34広島県	2328	2.33%
11埼玉県	5778	5.78%	35山口県	727	0.73%
12千葉県	2445	2.45%	36徳島県	599	0.60%
13東京都	8981	8.98%	37香川県	923	0.92%
14神奈川県	3836	3.84%	38愛媛県	1024	1.02%
15新潟県	2587	2.59%	39高知県	528	0.53%
16富山県	1206	1.21%	40福岡県	2604	2.60%
17石川県	1614	1.61%	41佐賀県	671	0.67%
18福井県	1240	1.24%	42長崎県	889	0.89%
19山梨県	1040	1.04%	43熊本県	948	0.95%
20長野県	2469	2.47%	44大分県	659	0.66%
21岐阜県	3102	3.10%	45宮崎県	676	0.68%
22静岡県	4485	4.49%	46鹿児島県	1101	1.10%
23愛知県	8089	8.09%	47沖縄県	651	0.65%
24三重県	1697	1.70%			

産業中分類	度数	構成比
09 食料品製造業	10632	10.63%
10 飲料・たばこ・飼料製造業	1760	1.76%
11 繊維工業	8801	8.80%
12 木材・木製品製造業（家具を除く）	3047	3.05%
13 家具・装備品製造業	4861	4.86%
14 パルプ・紙・紙加工品製造業	2544	2.54%
15 印刷・同関連業	6300	6.30%
16 化学工業	1937	1.94%
17 石油製品・石炭製品製造業	347	0.35%
18 プラスチック製品製造業（別掲を除く	5171	5.17%
19 ゴム製品製造業	1135	1.14%
20 なめし革・同製品・毛皮製造業	1032	1.03%
21 窯業・土石製品製造業	4755	4.76%
22 鉄鋼業	1993	1.99%
23 非鉄金属製造業	1207	1.21%
24 金属製品製造業	13146	13.15%
25 はん用機械器具製造業	3221	3.22%
26 生産用機械器具製造業	9142	9.14%
27 業務用機械器具製造業	1998	2.00%
28 電子部品・デバイス・電子回路製造業	1694	1.69%
29 電気機械器具製造業	3772	3.77%
30 情報通信機械器具製造業	582	0.58%
31 輸送用機械器具製造業	4455	4.46%
32 その他の製造業	6468	6.47%

従業者規模	度数	構成比
1人	11529	11.53%
2人	16912	16.91%
3人	10610	10.61%
4人	7916	7.92%
5~9人	20639	20.64%
10~19人	13913	13.91%
20~29人	6407	6.41%
30~49人	4909	4.91%
50~99人	3894	3.89%
100~199人	1920	1.92%
200~299人	569	0.57%
300~499人	430	0.43%
500~999人	247	0.25%
1000人～	105	0.11%

資本金階級	度数	構成比
300万円未満	1836	1.84%
300～500万円未満	17247	17.25%
500～1,000万円未満	7335	7.34%
1,000～3,000万円未満	27316	27.32%
3,000～5,000万円未満	4845	4.85%
5,000万円～1億円未満	4663	4.66%
1～3億円未満	1904	1.90%
3～10億円未満	1436	1.44%
10～50億円未満	1033	1.03%
50億円以上	1167	1.17%
以外	31218	31.22%

売上（収入）金額階級	度数	構成比
300万円未満	11641	11.64%
300～1,000万円未満	16229	16.23%
1,000～3,000万円未満	18519	18.52%
3,000万円～1億円未満	20711	20.71%
1～3億円未満	14677	14.68%
3～10億円未満	10267	10.27%
10～100億円未満	6830	6.83%
100億円以上	1126	1.13%

経営組織	度数	構成比
1個人経営	30066	30.07%
2株式会社・有限会社・相互会社	68504	68.50%
3合名会社・合資会社	496	0.50%
4合同会社	127	0.13%
5会社以外の法人	807	0.81%

単独・本所・支所の別	度数	構成比
単独事業所	76690	76.69%
本所・本社・本店	8481	8.48%
支所・支社・支店	14829	14.83%

開設時期	度数	構成比
昭和59年以前	53819	53.82%
昭和60年～平成6年	19193	19.19%
平成7～16年	13627	13.63%
平成17年	580	0.58%
平成18年	1559	1.56%
平成19年	1575	1.58%
平成20年	1528	1.53%
平成21年	1273	1.27%
平成22年	1135	1.14%
平成23年	1002	1.00%
平成24年	1287	1.29%
平成25年	1194	1.19%
平成26年	966	0.97%
平成27年	732	0.73%
平成28年	390	0.39%
不詳	140	0.14%

表 20 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である。8 属性から 2 属性ずつ選択されるため、都合 ${}_8C_2 = 28$ 通りのパターンが形成されている。例えば、一番上の行は、地域と産業でクロス集計を行うことで北海道×食料品製造業、東京×化学工業など様々な層を作成し、ひとつの層でカウントされる事業所数が 10 未満となるような事業所の数を集計した結果、合計で 898 事業所あったことを意味している。地域×産業や、地域×開設時期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、開設時期の分類区分の数が他に比べて細かいことがその理由のひとつとして考えられる。逆に、分類区分が 3 しかない単独・本所・支所の別は、どの属性と組み合わせてもリスクの高い事業所はほとんど出てきていない。この分類区分の粒度はそのまま外部参照情報との照らし合わせにおけるリンクキーとしての精度に繋がると考えられる。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は行っていない。

表 20 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	898
地域	従業者規模	371
地域	資本金階級	292
地域	売上（収入）金額階級	95
地域	経営組織	348
地域	開設時期	1,151
地域	単独・本所・支所の別	0
産業	従業者規模	188
産業	資本金階級	40
産業	売上（収入）金額階級	26
産業	経営組織	120
産業	開設時期	293
産業	単独・本所・支所の別	0
従業者規模	資本金階級	48
従業者規模	売上（収入）金額階級	80
従業者規模	経営組織	44
従業者規模	開設時期	191
従業者規模	単独・本所・支所の別	3
資本金階級	売上（収入）金額階級	33
資本金階級	経営組織	40
資本金階級	開設時期	66
資本金階級	単独・本所・支所の別	3
売上（収入）金額階級	経営組織	32
売上（収入）金額階級	開設時期	37
売上（収入）金額階級	単独・本所・支所の別	0
経営組織	開設時期	100
経営組織	単独・本所・支所の別	7
開設時期	単独・本所・支所の別	0

続いて、属性単位ではなく、事業所単位での考察を行った。上記と同じく、8 つの属性に対して 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 属性の組み合わせである 28 パターンのリスクの判定がある。事業所単体で見た時、露見リスクの高い事業所は、このうちの複数のパターンでカウントされると考えられる。これらを足し上げて「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定量的に露見リスクが相対的に高いレコードを探索的に発見することができる（図 17）。

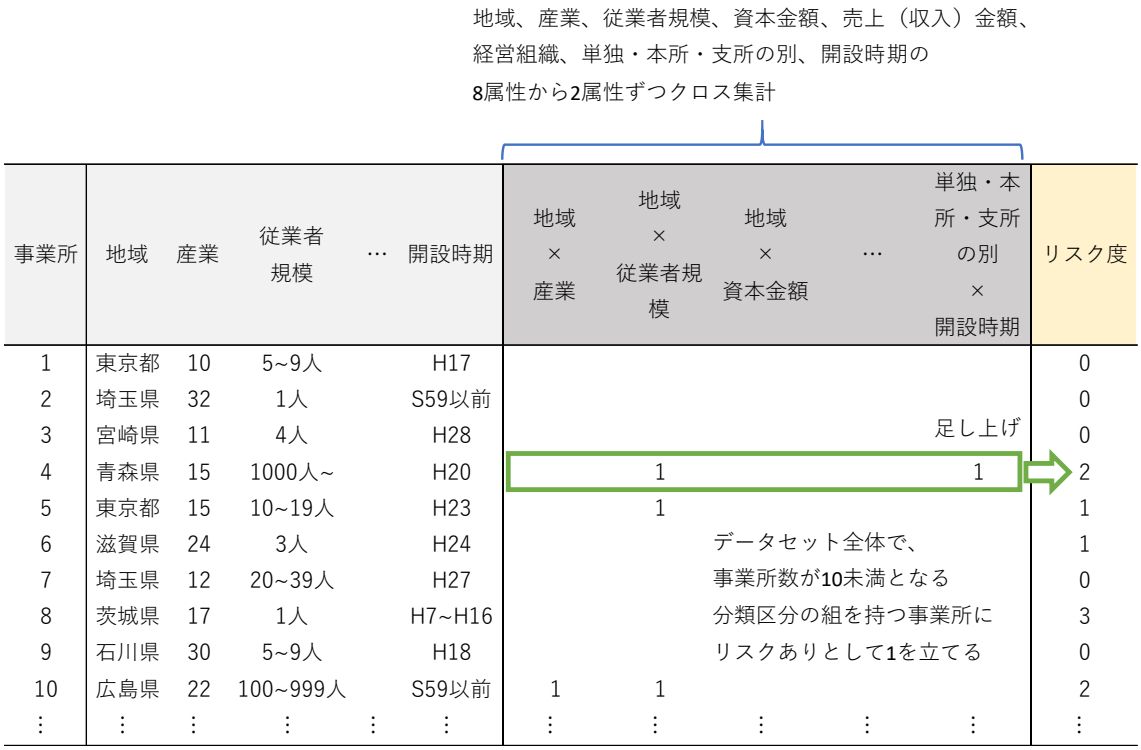


図 17 リスク度評価のイメージ

事業所ごとに何回リスク大と判断されたかをリスク度として集計し、層別に量的属性の要約統計量をまとめたものが表 21 である。本実験条件では、リスク度は 0 から 7 まで存在したが、秘匿上の問題からリスク度 6 とリスク度 7 の詳細は省略している。リスク度 0 は、売上（収入）金額や従業者合計においては 96,589 レコードと最も事業所数が多く、平均値等の値は最も小さかった。一方で、リスク度が上がるにつれてレコード数は減少し、平均値等の統計量の値は増加する傾向にあることがわかる。このことから、複数の属性を用いた評価においても、規模の大きい事業所が相対的に高い特定化リスクを秘めている可能性が推察される。

表 21 リスク度別の量的属性の要約統計量

リスク度	属性	レコード数	平均値	標準偏差	中央値	1%点	99%点
0	売上（収入）金額	96,589	58,960	2,761,050	3,500	0	738,016
	従業者合計	96,589	16	109	5	1	179
	資本金額	66,073	46,253	870,217	1,000	100	659,339
1	売上（収入）金額	2,703	654,135	4,449,081	9,161	0	10,877,842
	従業者合計	2,703	89	345	10	1	937
	資本金額	2,085	470,606	2,493,460	3,000	20	11,942,566
2	売上（収入）金額	456	1,582,457	5,503,532	25,689	0	23,372,664
	従業者合計	456	231	505	16	1	2,234
	資本金額	395	1,615,131	5,561,653	10,000	10	25,915,254
3	売上（収入）金額	164	2,562,453	7,447,916	132,392	0	44,433,442
	従業者合計	164	330	522	65	1	2,296
	資本金額	149	1,181,915	3,853,345	35,000	10	23,178,873
4	売上（収入）金額	53	2,352,191	6,243,454	28,602	0	26,797,508
	従業者合計	53	536	805	40	2	3,156
	資本金額	48	1,992,845	6,330,495	40,000	10	28,150,032
5	売上（収入）金額	25	3,343,831	8,800,848	1,123,633	156	36,337,285
	従業者合計	25	418	466	342	2	1,749
	資本金額	24	1,545,090	3,557,905	32,500	223	13,342,540

※秘匿上の問題から、リスク度 6、7 の詳細は省略した。

さらに、高リスク事業所（リスク度 1 以上）と低リスク事業所（リスク度 0）に層化を行い、それぞれについて各々の属性の分類事項の構成比の差異を調べた（図 18）。地域（都道府県）の場合、低リスク事業所については、東京都が占める割合は 9.19%と比較的大きい。一方、高リスク事業所については 3.14%と、東京都が占める割合は小さくなっている。これは、東京都という分類区分はその事業所数の多さから他の属性と組み合わせても露見リスクが高まりづらいことを意味している。一方で、沖縄県は 0.56%から 3.17%になるなど、元の構成比の小さい事業所はリスクが高まる可能性を示している。他の属性についても同様の傾向が見られる。従業者規模、資本金額、売上（収入）金額階級などでは、規模が大きいほど高リスク事業所になりやすい。また資本金額は例外的に、300 万円未満の事業所にも高リスク事業所が多く存在していることが特徴的であった。

図18-1 地域（都道府県）

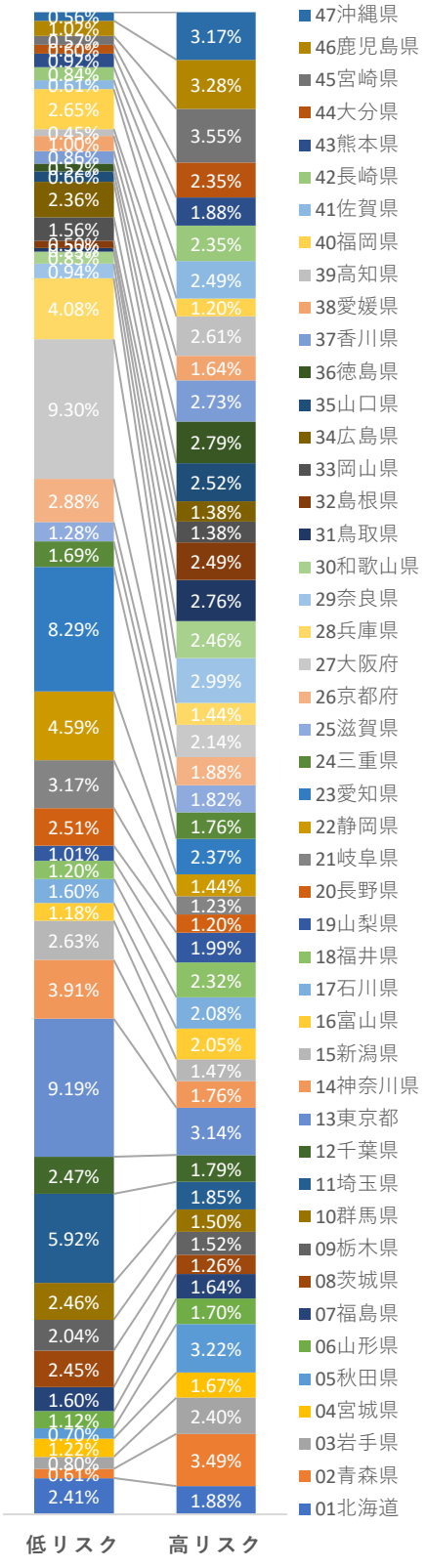


図18-2 産業（中分類）

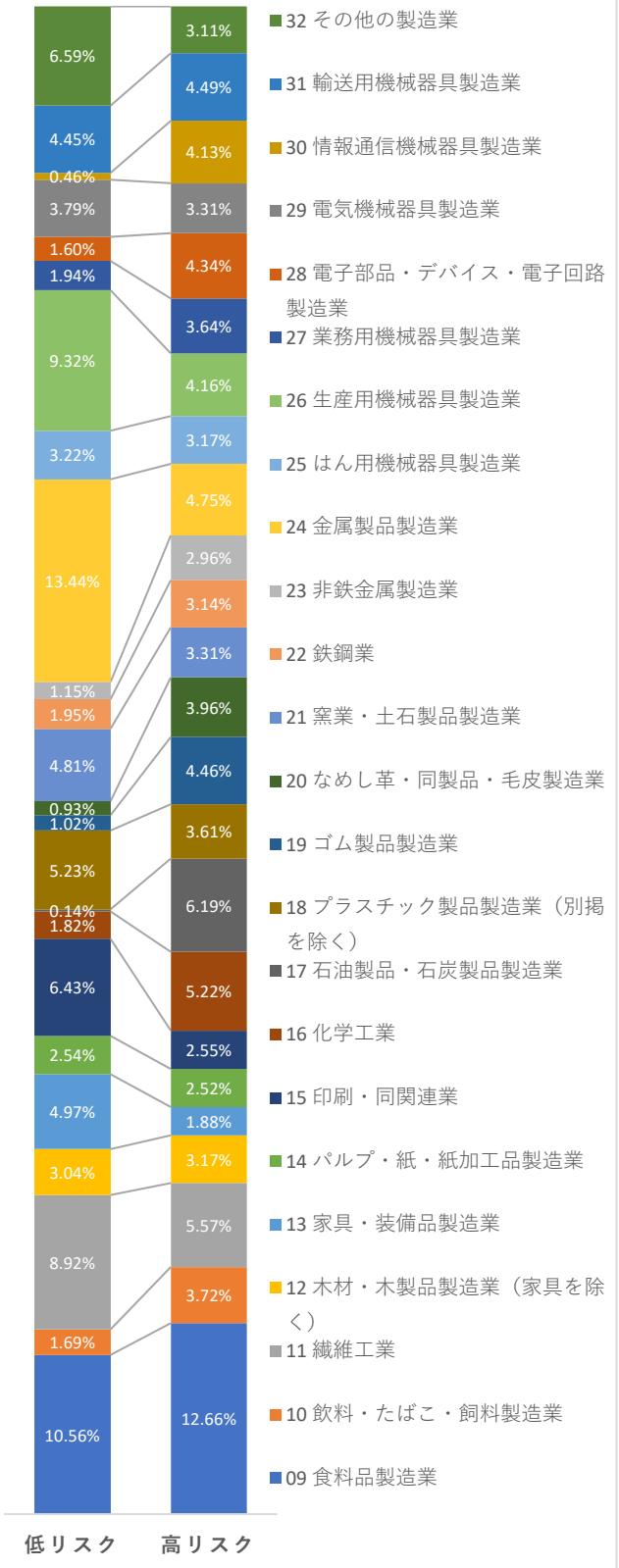


図18-3 従業者規模

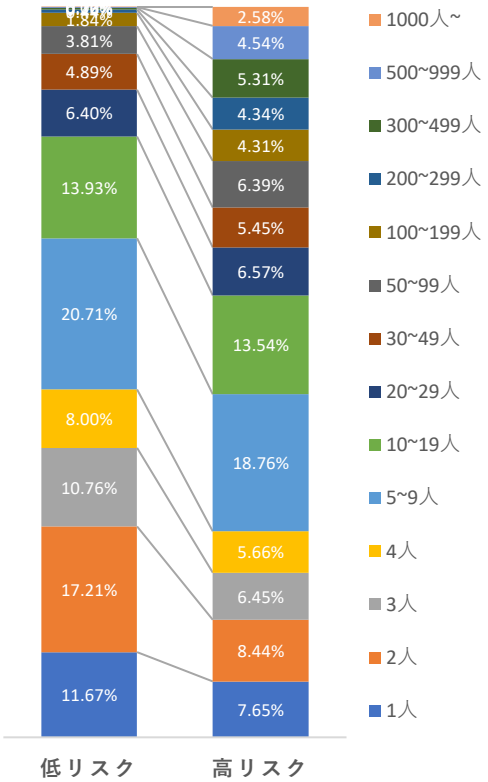
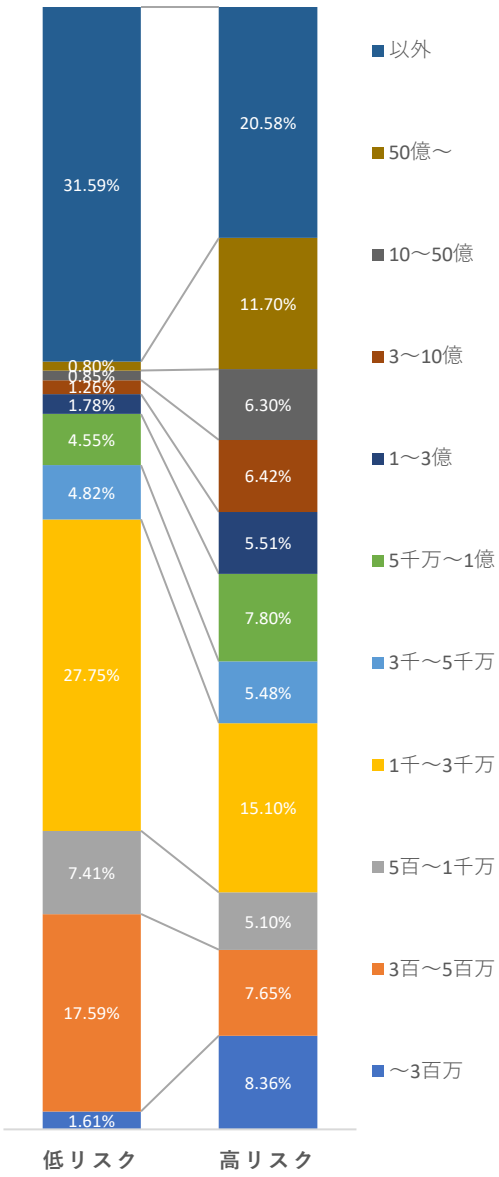
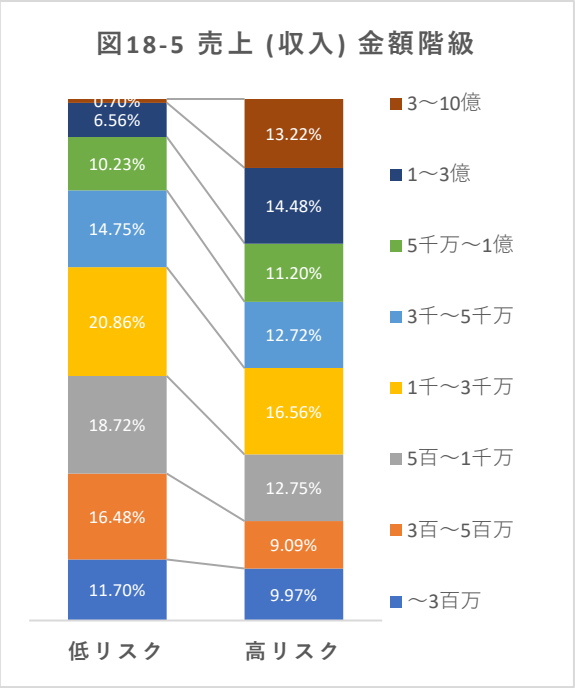
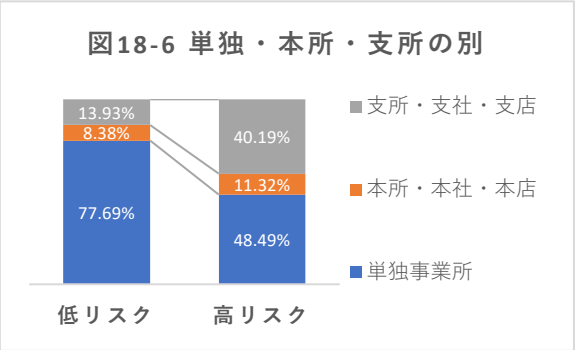
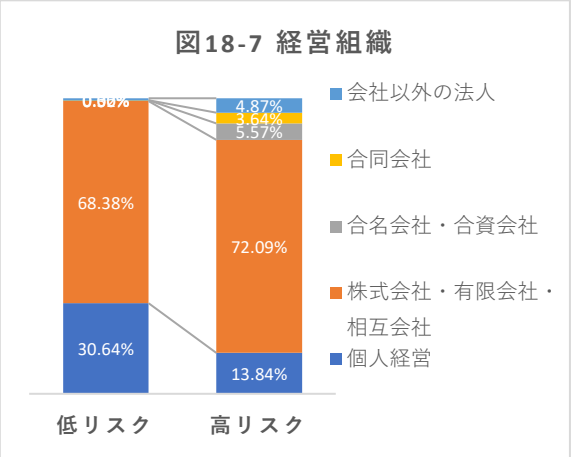


図18-4 資本金階級









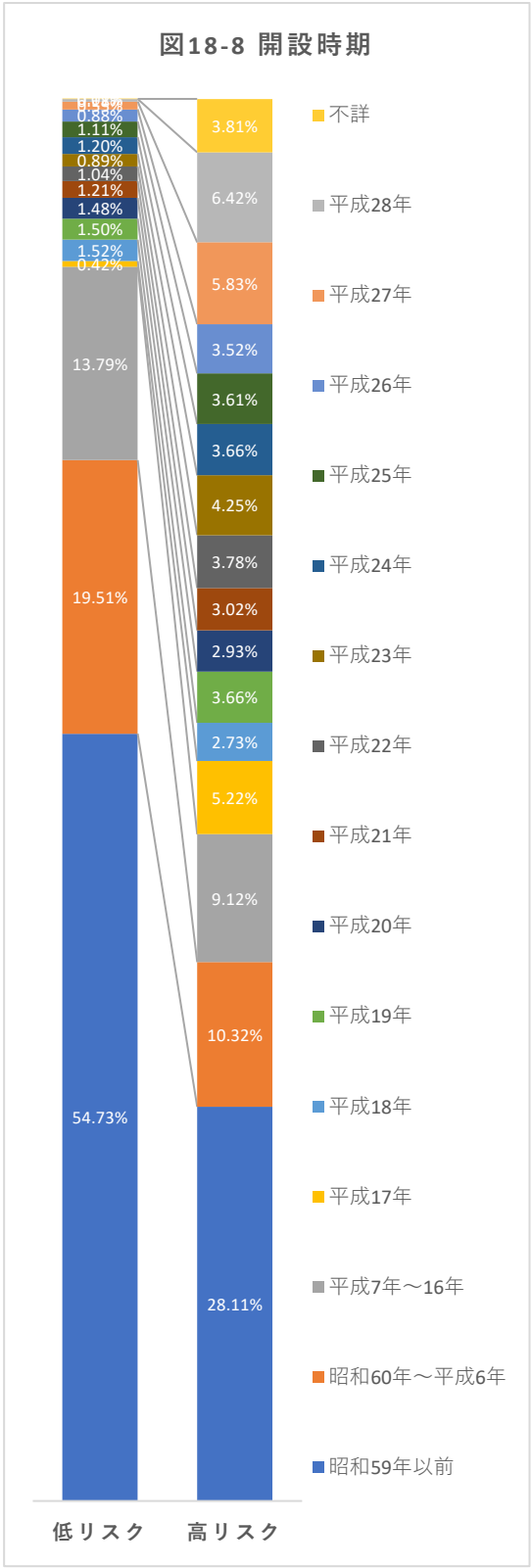


図 18 高リスク事業所と低リスク事業所の分類区分の構成比の比較

最後に、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と、高リスク事業所の割合をバブルチャートとして表示した（図 19）。バブルの大きさは事業所数を示しており、色は高リスク事業所の割合が小さいほど白く、高いほど黒く表現されている。なお、分類区分の組み合わせによっては事業所がひとつも存在しないため、バブルの大きさは 0 になり、また割合も計算できない。しかし、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、リスクは大きいものと判断する必要がある。そのため、背景には白ではなく黒を使用し、高リスク事業所が多い時だけでなく、該当する事業所数が少ない場合にも、その分類区分の組み合わせが暗く表示されるように工夫した。逆に、明るく見える分類区分の組み合わせは相対的にリスクが小さいと見なすことができる。なお、紙面の都合上、経営組織、単独・本所・支所の別、開設時期を含む組み合わせについては付録 B に掲載した。

例えば、図 15-1 の地域×産業の場合、01 北海道×産業 09（食料品製造業）の組み合わせはバブルが大きく色も白いため、この分類区分の組み合わせにおいては、あまりリスクは大きくないと考えられる。一方で、01 北海道×産業 19（ゴム製品製造業）のセルはバブルが小さく色も暗いため、これに該当する事業所は高リスク事業所であると考えられる。このような分類区分の組み合わせは、優先的に匿名化の対象とする必要がある。地域×産業の一覧を見渡すと、地域については西日本が全般的に、産業については、産業 17（石油製品・石炭製品製造業）、19（ゴム製品製造業）、30（情報通信機械器具製造業）などの特定の産業は事業所が少なく、高リスク事業所が多いことが読み取れる。このように、特定の分類区分の組み合わせに着目するだけでなく、行または列単位で事業所の露見リスクを大まかに評価することも可能である。逆に、行または列単位での傾向が見られない飛び地的な分類事項の組み合わせは、グローバルリコーディング以外の攪乱的手法の適用可能性を検討することも考えられる。

複数のバブルチャートを概観した結果、本実験では、従業者規模の大きい事業所の露見リスクが特に大きいことがわかった。次いで、資本金階級や売上（収入）金額階級の大きい事業所のリスクが大きく、産業も一部の中分類については注意が必要である。地域については、やや西日本のリスクは大きいと考えられるが、前述の項目ほど極端な傾向は現れなかった。また、図 15-9 従業者規模×売上（収入）金額階級などのように、相関が比較的高い量的属性同士の場合、バブルチャートでもその相関の傾向が現れている。原則として従業者規模が小さいほど売上（収入）金額階級も小さく、その逆もまた然りである。規模の大きい事業所だけでなく、複数の量的属性でその程度に大きな差異のある事業所も、比較的リスクが大きくなると考えられる。これらへの対処も、事業所・企業の匿名化では重要になると考えられる。

なお、本実験では、特定の属性に絞ったうえで重みづけを行わずに事業所数の観点からのみリスク度の評価を行っている。現実にはどのような属性が外部参照情報との準識別子になるか定かではなく、その外観識別性の程度も評価は難しい。また、どのよう

な分類区分を用いるかによっても、リスクが相対的に高いとされる事業所の傾向は変化する可能性がある。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と考えられる。本実験で得られた知見はあくまでも一例ではあるが、今後、事業所・企業系の匿名化ミクロデータの作成を試行するにあたって、キー変数の決定やリコーディングの分類区分の定め方、量的属性の攪乱の際の層化の基準など、さまざまな点に応用できる基礎的な考え方ではないかと思われる。

図 19 分類区分別の事業所数と高リスク事業所数割合

（バブルの大きさは事業所数、バブルの濃淡は高リスク事業所割合）

図 19-1 地域×産業

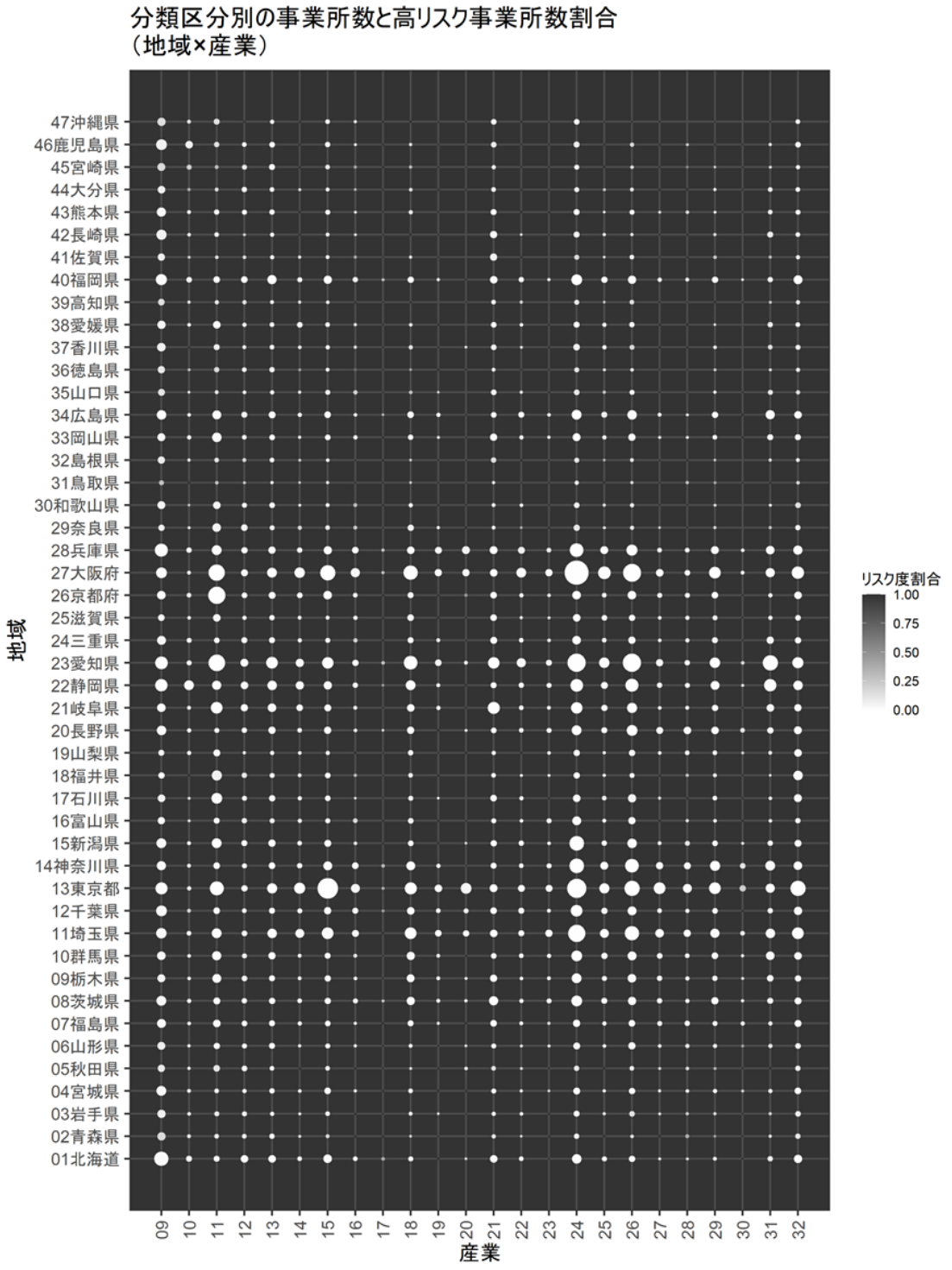


図 19-2 地域×従業者規模

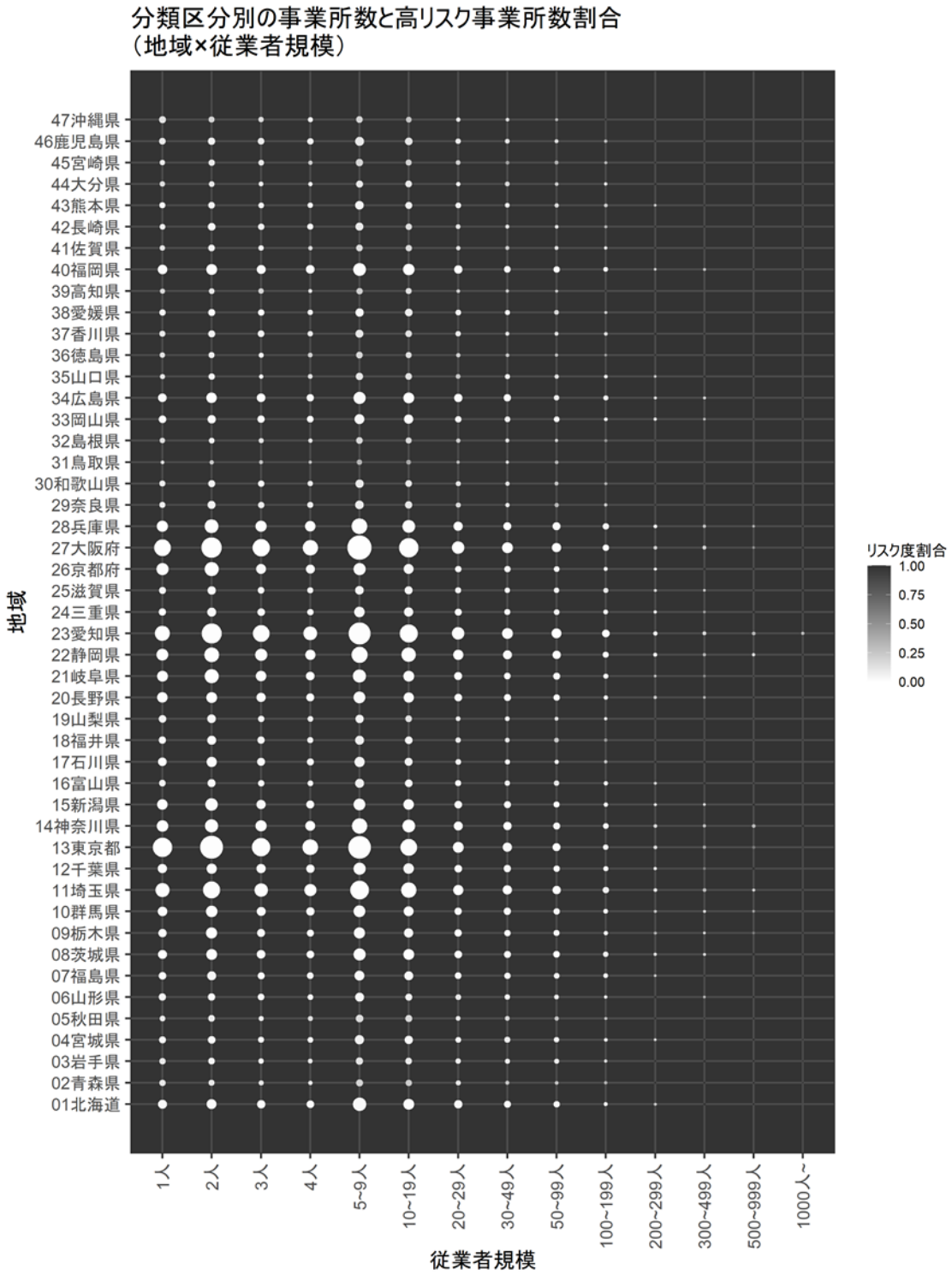


図 19-3 地域×資本金階級

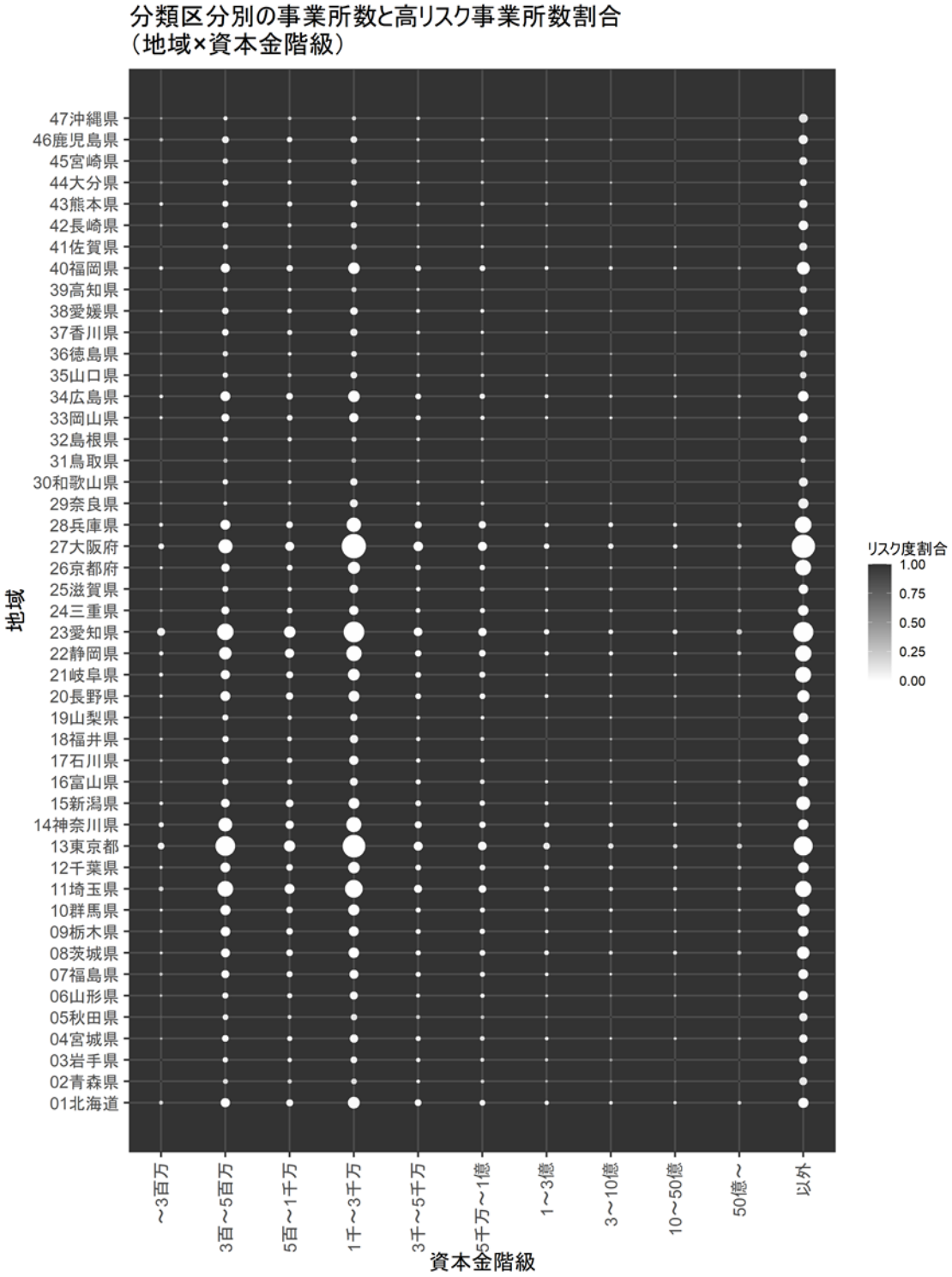


図 19-4 地域×売上（収入）金額階級

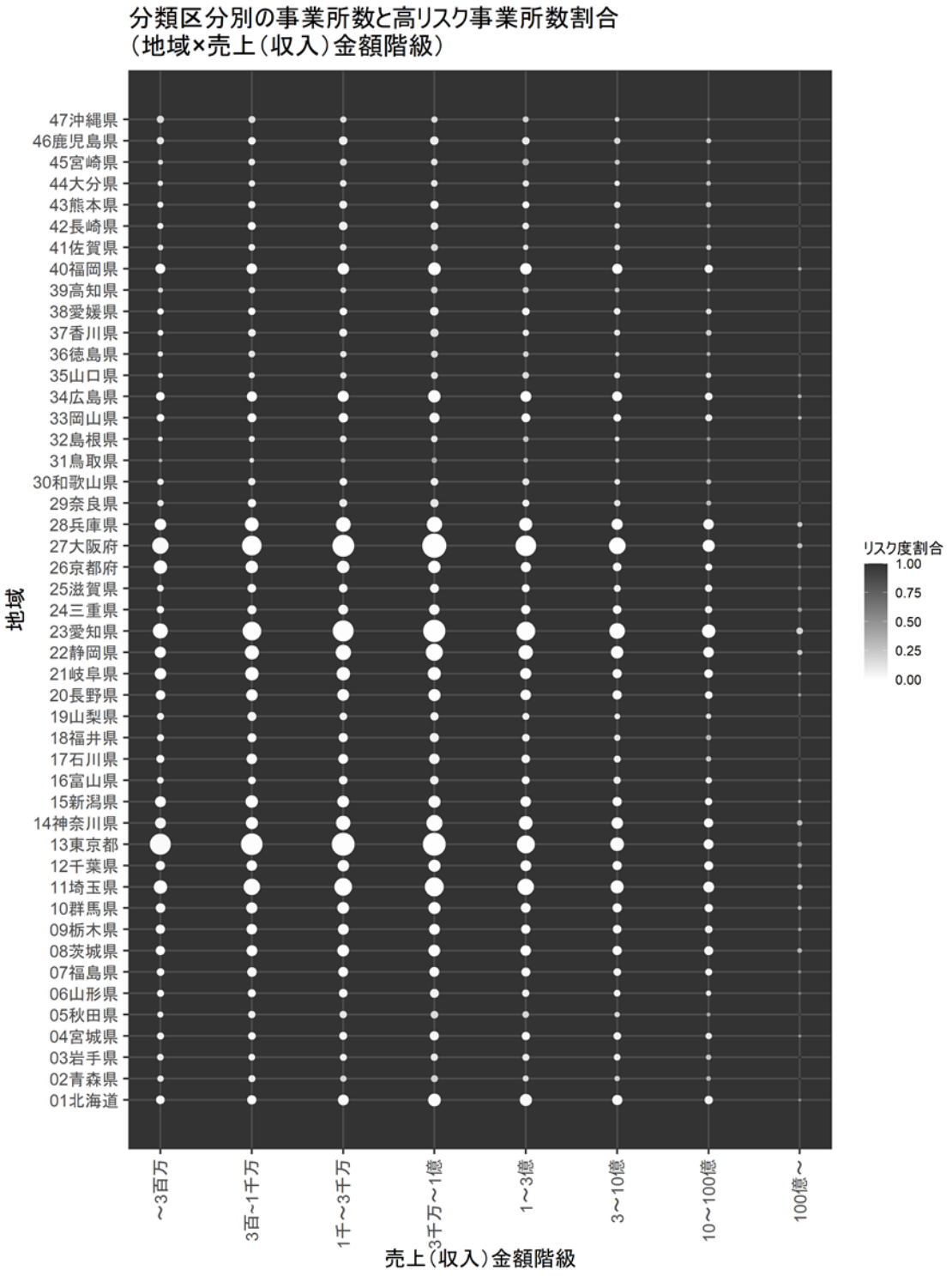


図 19-5 産業×従業者規模

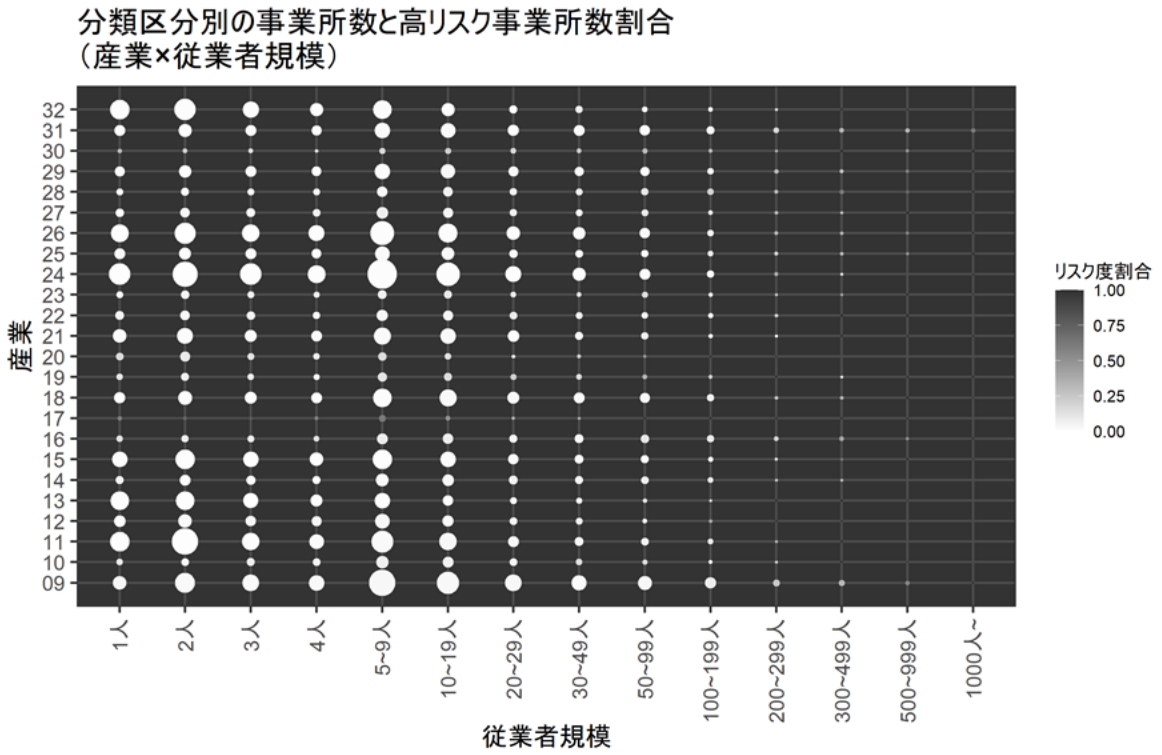


図 19-6 産業×資本金階級

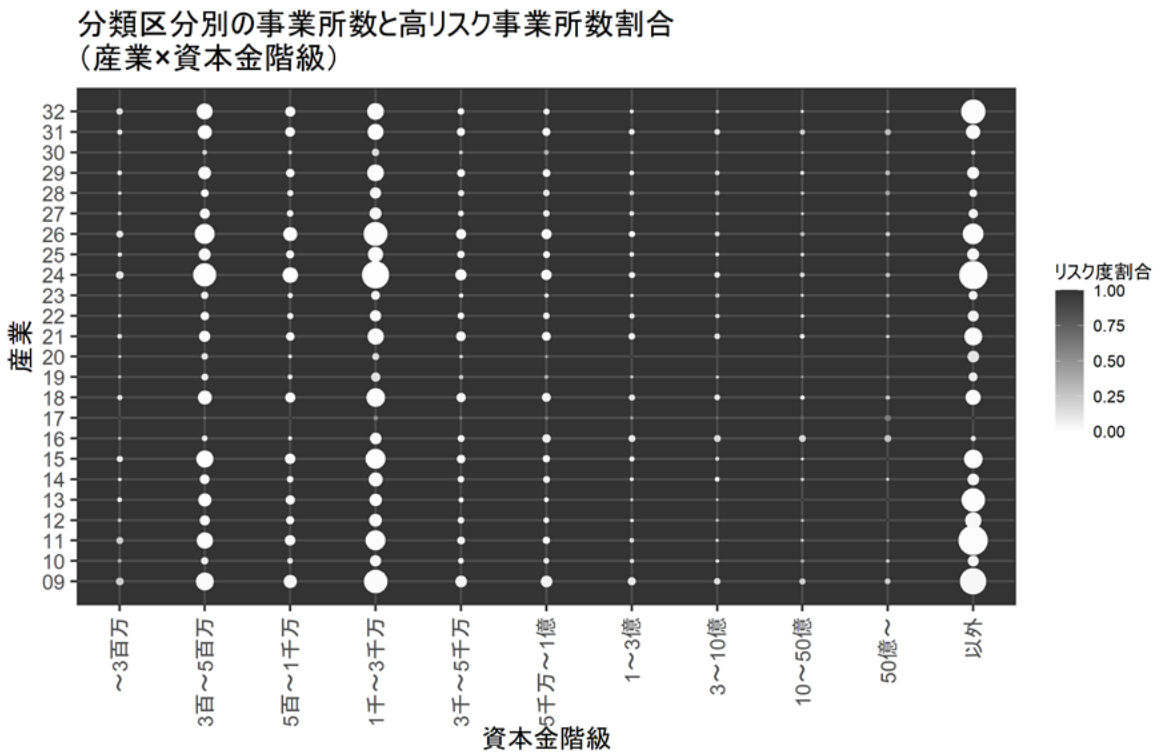


図 19-7 産業×売上（収入）金額階級

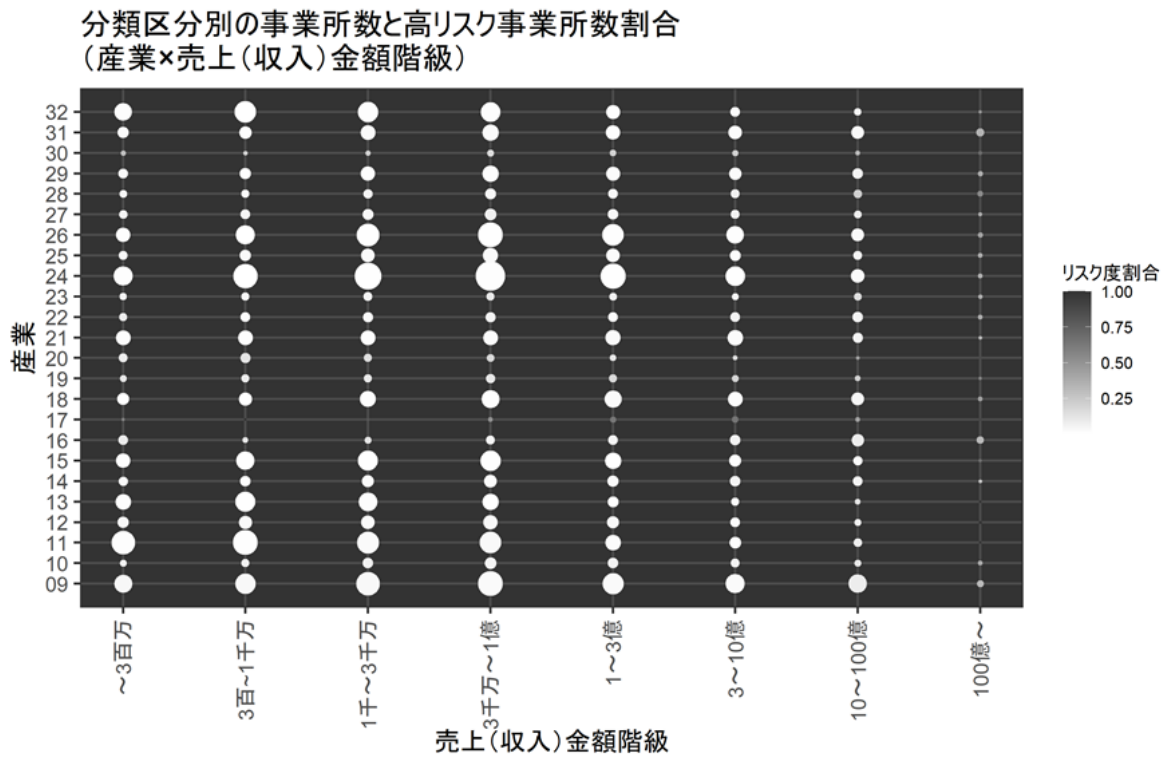


図 19-8 従業者規模×資本金階級

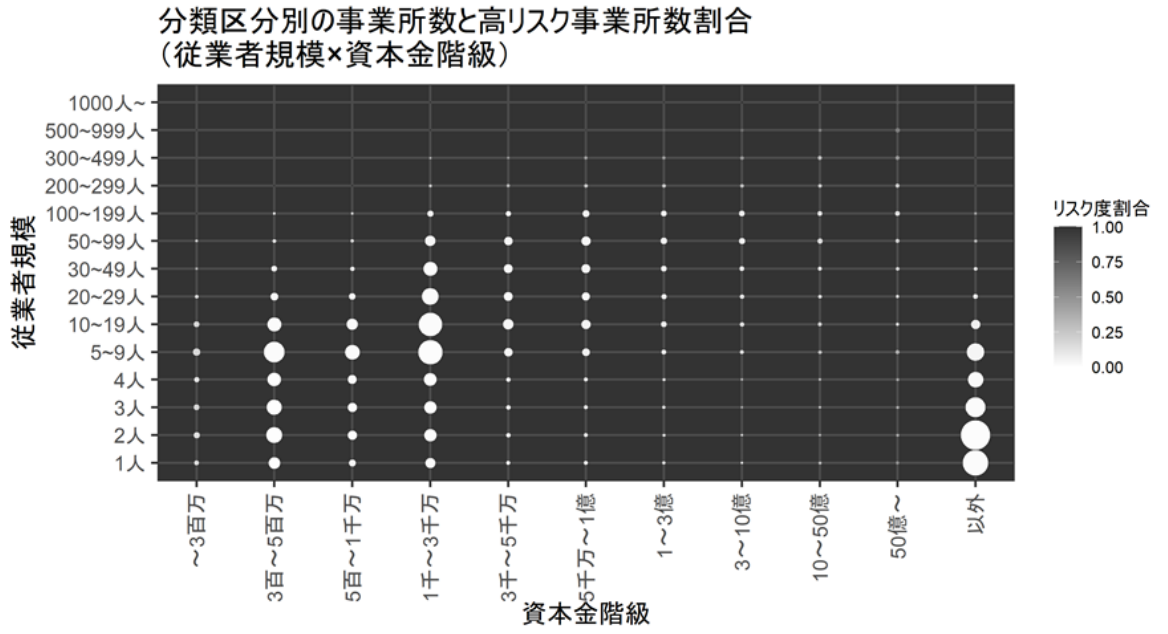


図 19-9 従業者規模×売上（収入）金額階級

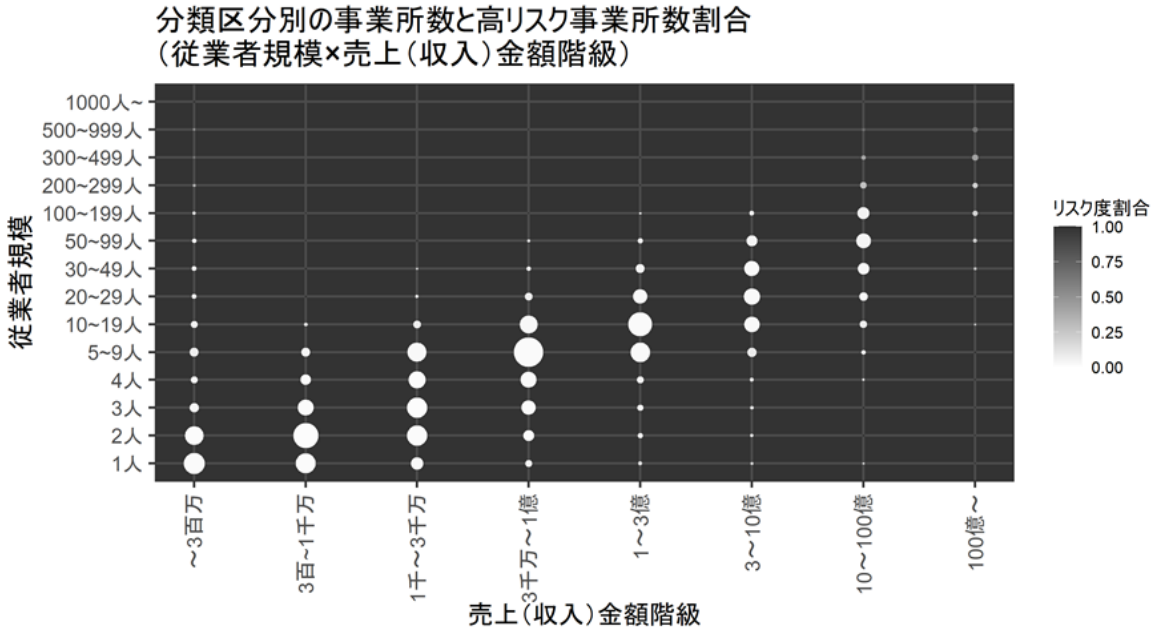
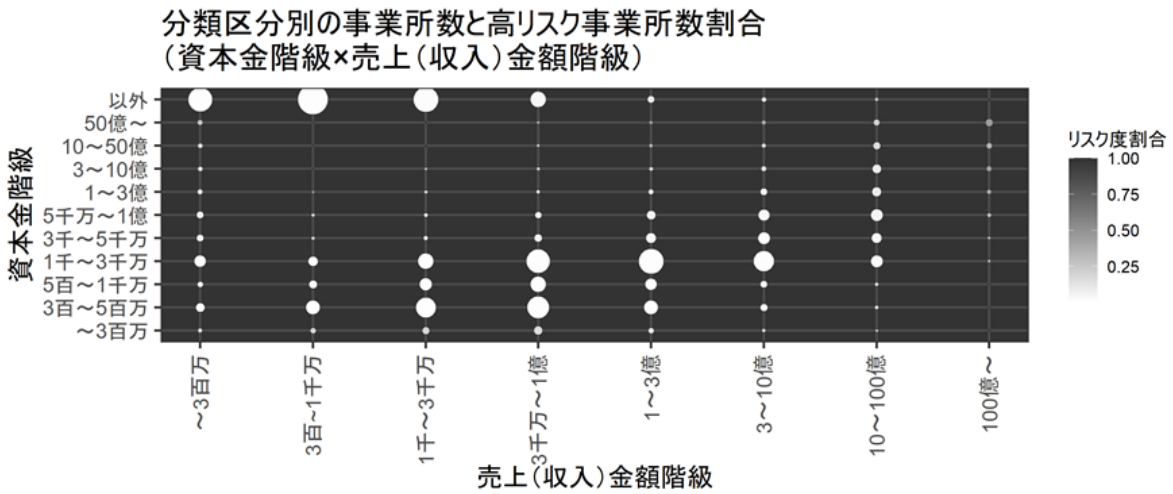


図 19-10 資本金階級×売上（収入）金額階級



6 むすびにかえて

本研究では、公的統計ミクロデータの利活用の促進に向けた統計的開示抑制の検討の一環として、事業所・企業の匿名化ミクロデータの作成に資する基礎研究を行った。公的統計における近年のミクロデータ作成のサーベイを行ったのち、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、および事業所・企業系のミクロデータに対する匿名化手法の概要を整理した。その上で、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て作成された匿名化ミクロデータの有用性と秘匿性に関する定量的な評価や、より経済センサスに適した匿名化を指向したデータ特性の探索的な分析を行った。

本研究の成果は、大きくわけて二点ある。一つは、わが国で実用化しておらず、また研究事例も少ない事業所・企業系の匿名化ミクロデータ作成について、諸外国のサーベイを行った点である。先行研究や先行事例を通じて、量的属性の分布は極端に不均質性、企業規模の異なる事業所のサンプリングの難しさ、財務情報などの外部に開示される企業情報などから、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それよりも大きいといった問題を整理した。また、イタリアやドイツにおける匿名化ミクロデータの作成事例を考察し、以下の三点の知見を得た。第一に、学術研究用のファイルの作成を前提に、偶発的な個体特定や外部情報を用いたマッチングを行うことに重点が置かれている露見シナリオをもとに、定量的な評価基準に基づいて最小限の攪乱に留めている。第二に、匿名化手法に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、特に情報量損失が相対的に低い個別ランキング法といった攪乱的手法が採用されている。第三に、匿名化手法の適用にあたっては、統計調査ごとのデータ特性や統計調査の実務担当者の助言も考慮することが強調されている。

本研究のもう一つの成果は、これらの先行事例を踏まえて攪乱的手法を使用し、経済センサスのオンサイト利用を通じて、実データを用いた実証研究を行った点である。売上（収入）金額、地域、産業分類、従業者合計、資本金額といった属性に着目して探索的なリコーディングを行い、露見リスクが相対的に高くなると判断されるレコードを発見した。量的属性の匿名化にあたってはミクロアグリゲーションを適用し、クロス表による評価方法やリンケージ技法等を用い攪乱済みデータの有用性、秘匿性、そしてその総合評価を定量的に行った。また経済センサス特有の分布特性等を探るため、相対的にリスクが高いと考えられるレコードや属性の分析も行った。

本研究はわが国での事業所・企業系の匿名化ミクロデータ作成のための基礎的な研究である。実務への適用を目標とした場合、課題は下記のようなものがあげられる。

- 諸外国の事業所・企業系の匿名化ミクロデータ作成手法のさらなるサーベイ（特にドイツにおける統計調査ごとやファイル形式ごとの差異の分析）

- 分布特性を考慮した外れ値の評価方法の追求
- 匿名化にあたって、質的属性のリコーディングにおける分類区分の幅の決定や攪乱的手法の適用可能性の検討、量的属性における属性間の相関性の考慮
- 有用性の観点から、経済センサスのミクロデータ利用者の利用事例や分析手法の把握
- 秘匿性の観点から、外部参照情報の入手可能性や接続可能性を考慮したリンケージ実験の評価
- 経済センサスにおける、製造業以外の産業についての匿名化ミクロデータ作成の検討
- 経済センサスにおける、企業と傘下事業所の関係性も含めた匿名化ミクロデータ作成の検討
- パネルデータ作成を想定した経済センサス基礎調査と経済センサス活動調査のリンケージ実験
- 経済センサス以外の事業所・企業系の統計調査の分析・ミクロデータ作成の検討

これらについて、引き続き研究を続けていく所存である。

謝辞

修士論文作成にあたって貴重なご助言をくださった指導教員の竹村彰通先生に感謝の言葉を申し上げます。また、研究面でのご指導やオンサイト利用の申請でお世話になりました、共同研究者の伊藤伸介先生に心より感謝いたします。オンサイト利用の手続きにてお世話になりました総務省統計研究研修所、統計データ利活用センター、滋賀大データサイエンス教育研究センターの担当者の皆様にも御礼申し上げます。さらに、行政官国内研究員制度を通じて滋賀大学をご紹介いただき、進学を後援していただいた独立行政法人統計センター、総務省統計局、人事院の関係者の皆様にも感謝の念に絶えません。最後に、データサイエンス研究科修士課程で様々な刺激をいただいた先生や院生の皆様に感謝の言葉を述べて、謝辞とさせていただきます。誠にありがとうございました。

参照文献

Abidi, B., Ben Yahia, S., & Perera, C. (2020). *Hybrid microaggregation for privacy preserving data mining.* J Ambient Intell Human Comput 11, 23–38 (2020). https://doi.org/10.1007/s12652-018-1122-7.

ARX. (2020a). *Data Anonymization Tool*. Retrieved from https://arx.deidentifier.org/

ARX. (2020b). *Related software*. Retrieved from https://arx.deidentifier.org/overview/related-software/

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. https://doi.org/10.1007/978-3-540-87471-3_12.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). *LOF: identifying density-based local outliers.* ACM sigmod record (pp. 93-104). https://doi.org/10.1145/335191.335388.

Dandekar, R., Cohen, M., & Kirkendall, N. (2001). *Applicability of Latin Hypercube Sampling to Create Multivariate Synthetic Micro Data.* Proceedings of ETK-NTTS, Eurostat, Luxemburg, 839-847.

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

Defays, D., & Nanopoulos, P. (1993). *Panels of enterprises and confidentiality: the small aggregates method.* Proc. of 92 Symposium on Design and Analysis of Longitudinal Surveys, pp. 195-204. Statistics Canada, Ottawa.

Domingo-Ferrer, J., & González-Nicolás, Ú. (2010). *Hybrid microdata using microaggregation.* Information Sciences. 180. 2834-2844. 10.1016/j.ins.2010.04.005. doi:10.1016/j.ins.2010.04.005.

Domingo-Ferrer, J., & Mateo-Sanz, J. (2002). *Practical data-oriented microaggregation for statistical disclosure control*. IEEE Transactions on Knowledge and Data Engineering, 14(1):189–201. DOI: 10.1109/69.979982.

Domingo-Ferrer, J., & Torra, V. (2001a). *A quantitative comparison of disclosure control methods for microdata.* L.J.I. Doyle P., Theeuwes J.J.M., Zayatz L.V. (Ed.) Confidentiality, disclosure, and data access: Theory and practical applications for statistical agencies, Elsevier, 2001, pp. 111-134.

Domingo-Ferrer, J., & Torra, V. (2001b). *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.

Domingo-Ferrer, J., & Torra, V. (2005). *Ordinal, Continuous and Heterogeneous `k`-anonimity through Microaggregation.* Data Mining and Knowledge Discovery 11(2), pp. 195-212. DOI: 10.1007/s10618-005-0007-5.

Domingo-Ferrer, J., & Torra, V. (2005). *Ordinal, Continuous and Heterogeneous k-Anonymity Through Microaggregation.* Data Min Knowl Disc 11, 195–212 (2005). https://doi.org/10.1007/s10618-005-0007-5. https://doi.org/10.1007/s10618-005-0007-5.

Duncan, G., & Pearson, R. (1991). *Enhancing access to microdata while protecting confidentiality: prospects for the future.* Statistical Science 6, 219-239.

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.

Elliot, M., & Dale, A. (1999). *Scenarios of attack: the data intruder's perspective on statistical disclosure risk.* Netherlands Official Statistics, 6-10.

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), 226-231.

Franconi, L., & Ichim, D. (2007). *Community Innovation Survey: comparable dissemination.*

Gouweleeuw, J., Kooiman, P., Willenborg, L., & de Wolf, P. (1997). *Post randomisation for statistical disclosure control: Theory and implementation.* Technical report, Statistics Netherlands. Research paper no. 9731.

Hafner, H., Ritchie, F., & Lenz, R. (2019). *User-focused threat identification for anonymised microdata.* Statistical Journal of the IAOS, 35(4), 703-713. https://doi.org/10.3233/SJI-190506.

Hundepool, A., de Wetering, A., Ramaswamy, R., Franconi, L., Capobianchi, A., DeWolf, P.-P., . . . 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.

Hundepool, A., de Wetering, A., Ramaswamy, R., Franconi, L., Polettini, S., & Capobianchi, A. (2020). *μ -Argus. User Manual. Version 5.1.* Retrieved from http://neon.vb.cbs.nl/casc/Software/MUmanual5.1.3.pdf

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

Ichim, D. (2008). *Community Innovation Survey: a Flexible Approach to the Dissemination of Microdata Files for Research.*

Ichim, D. (2009). *Disclosure Control of Business Microdata: A Density-Based Approach.* International Statistical Review / Revue Internationale De Statistique, 77(2), 196-211. Retrieved June 25, 2020, from www.jstor.org/stable/27919722.

IHSN. (2019). *SDC Practice Guide*. Retrieved from https://sdcpractice.readthedocs.io/en/latest/index.html

IHSN. (2020a). *Statistical disclosure control (anonymization) - Software Development*.

IHSN. (2020b). *Statistical Disclosure Control (sdcMicro)*. Retrieved from https://ihsn.org/software/disclosure-control-toolbox.

IHSN, (2019). *SDC Practice Guide*. Retrieved from SDC Practice Guide: https://sdcpractice.readthedocs.io/en/latest/index.html

Istat. (2020a). *ANALYSE PHASE*. Retrieved from https://www.istat.it/en/methods-and-tools/methods-and-it-tools/analyse

Istat. (2020b). *ITALIAN INNOVATION SURVEY: PUBLIC USE MICRO.STAT FILES*. Retrieved from https://www.istat.it/en/archivio/87787

Istat. (2020c). *THE ITALIAN INNOVATION SURVEY (COMMUNITY INNOVATION SURVEY, CIS): MICRODATA FOR RESEARCH PURPOSES*. Retrieved from https://www.istat.it/en/archive/35223

Ito, S., Yoshitake, T., Kikuchi, R., & Akutsu, F. (2018). *Comparative Study of the Effectiveness of Perturbative Methods for Creating Official Microdata in Japan.* In: Domingo-Ferrer J., Montes F. (eds) Privacy in Statistical Databases. PSD 2018. Lecture Notes in Computer Science, vol 11126. Springer, Cham. DOI: 10.1007/978-3-319-99771-1_14.

Jiménez, J., Navarro-Arribas, G., & Torra, V. (2014). *JPEG-Based Microdata Protection.* In: Domingo-Ferrer J. (eds) Privacy in Statistical Databases. PSD 2014. Lecture Notes in Computer Science, vol 8744. Springer, Cham. https://doi.org/10.1007/978-3-319-11257-2_10.

Kim, H., Karr, A., & Reiter, J. (2015). *Statistical Disclosure Limitation in the Presence of Edit Rules.* Journal of Official Statistics, Vol. 31, No. 1, 2015, pp. 121–138. DOI: 10.1515/jos-2015-0006.

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

Lenz, R. (2006). *Measuring the Disclosure Protection of Micro Aggregated Business Microdata. An Analysis Taking as An Example the German Structure of Costs*

Survey. Journal of official statistics. 22. 681-710.

Lenz, R. (2008). *Risk assessment methodology for longitudinal business microdata .* Wirtsch Sozialstat Arch 2, 241–257.

Lenz, R., & Zwick, M. (2009). *Business Microdata in Germany: Linkage and Anonymisation.* Journal of Applied Social Science Studies / Zeitschrift für Wirtschafts- und Sozialwissenschaften. DOI: https://doi.org/10.3790/schm.129.4.645.

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), pages 635-651.

Li, N., Li, T., & Venkatasubramanian, S. (2007). *t-Closeness: Privacy Beyond k-Anonymity and l-Diversity.* IEEE 23rd International Conference on Data Engineering, 2007. ICDE 2007. pp. 106–115. 10.1109/ICDE.2007.367856.

Li, T., & Li, N. (2009). *On the tradeoff between privacy and utility in data publishing.* Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 517-526. https://doi.org/10.1145/1557019.1557079.

Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam., M. (2006). *diversity: Privacy beyond k-anonymity.* InProc. 22nd Intnl. Conf. Data Engg. (ICDE), page 24. https://doi.org/10.1145/1217299.1217302.

Martínez, S., Sánchez, D., & Valls, A. (2012). *Semantic adaptive microaggregation of categorical microdata.* Comput. Secur., 31, 653-672. https://doi.org/10.1016/j.cose.2012.04.003.

Mateo-Sanz, J., Sebé, F., & Domingo-Ferrer, J. (2004). *Outlier Protection in Continuous Microdata Masking.* In: Domingo-Ferrer J., Torra V. (eds) Privacy in Statistical Databases. PSD 2004. Lecture Notes in Computer Science, vol 3050. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-25955-8_16.

Moore, R. (1996). *Controlled data swapping techniques for masking public use microdata sets.* U.S. Bureau of the Census, Statistical Research Division Report 96/04.

Muralidhar, K., & Sarathy, R. (2006). *Data Shuffling: A New Masking Approach for Numerical Data.* Management Science, 52(5), 658-670. Retrieved July 15, 2020, from www.jstor.org/stable/20110544. https://doi.org/10.1287/mnsc.1050.0503.

Muralidhar, K., & Sarathy, R. (2008). *Generating sufficiency-based non-synthetic perturbed data.* Transactions on Data Privacy 1(1), 17-33.

Nin, J., Herranz, J., & Torra, V. (2008). *On the disclosure risk of multivariate microaggregation.* Data Knowl. Eng., 67, 399-412.

https://doi.org/10.1016/j.datak.2008.06.014.

O'Keefe, C., & 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. https://doi.org/10.1007/978-3-319-11257-2_27.

O'Keefe, C., & Shlomo, N. (2012). *Comparison of Remote Analysis with Statistical Disclosure Control for Protecting the Confidentiality of Business Data.* Transactions on Data Privacy. 5. 403-432.

Orooji, M., & Knapp, G. (2018). *A Novel Microdata Privacy Disclosure Risk Measure.* IISE Annual Conference. Proceedings. Institute of Industrial and Systems Engineers (IISE).

Prasser, F., & Kohlmayer, F. (2015). *Putting Statistical Disclosure Control into Practice: The ARX Data Anonymization Tool.* Gkoulalas-Divanis A., Loukides G. (eds) Medical Data Privacy Handbook. Springer, Cham. https://doi.org/10.1007/978-3-319-23633-9_6.

Research Data Centre of the Federal Statistical Office. (2013). *Scientific-Use-File zur Verdienststrukturerhebung 2010 - Anonymisierungskonzept und Metadaten -.* Retrieved from https://www.forschungsdatenzentrum.de/sites/default/files/vse_2010_suf_ak_md.pdf

Research Data Centre of the Federal Statistical Office. (2016). *CAMPUS-File zur Verdienststrukturerhebung 2010 - Anonymisierungskonzept -.* Retrieved from https://www.forschungsdatenzentrum.de/sites/default/files/vse_2010_cf_ak.pdf

Research Data Centre of the Federal Statistical Office. (2020). *Research Data Centre of the Federal Statistical Office*. Retrieved from https://www.forschungsdatenzentrum.de/de#

Rocher, L., Hendrickx, J., & de Montjoye, Y. (2019). *Estimating the success of re-identifications in incomplete datasets using generative models.* Nat Commun 10, 3069 (2019). https://doi.org/10.1038/s41467-019-10933-3.

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

Takemura, A. (2002). *Local recoding and record swapping by maximum weight matching for disclosure control of microdata sets.* Journal of Official Statistics, 18, 275-289. 2002.

Templ, M. (2007). *sdcMicro: A new flexible R-package for the generation of anonymised*

microdata - design issues and new methods. In to appear in:Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality. Monographs of Official Statistics.

Templ, M., & Meindl, B. (2008). *Robustification of Microdata Masking Methods and the Comparison with Existing Methods.* Domingo-Ferrer J., Saygın Y. (eds) Privacy in Statistical Databases. PSD 2008. Lecture Notes in Computer Science, vol 5262. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-87471-3_10.

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. 10.18637/jss.v067.i04.

Templ, M., Meindl, B., & Kowarik, A. (2020). *Package 'sdcMicro' Version 5.5.1.* Retrieved from Statistical Disclosure Control Methods for Anonymization of Data and Risk Estimation: https://cran.r-project.org/web/packages/sdcMicro/sdcMicro.pdf

Templ, M., Meindl, B., Kowarik, A., & Dupriez, O. (2017). *Simulation of synthetic complex data:The R-package simPop.* Journal of Statistical Software, 1–38. 10.18637/jss.v079.i10.

Ting-ting, C., Hui-qun, Y., & Jian-min, H. (2008). *An Improved V-MDAV Algorithm for l-Diversity.* in 2010 Third International Symposium on Information Processing, 2008 pp. 733-739. 10.1109/ISIP.2008.110.

Torra, V. (2004). *Microaggregation for Categorical Variables: A Median Based Approach.* Domingo-Ferrer J., Torra V. (eds) Privacy in Statistical Databases. PSD 2004. Lecture Notes in Computer Science, vol 3050. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-25955-8_13.

Truta, T., Fotouhi, F., & Barth-Jones, D. (2003). *Disclosure risk measures for microdata.* Proceedings of the 15th International Conference on Scientific and Statistical Database Management, Cambridge, MA, pp.15–22.

Vilhuber, L., Miranda, J., Kinney, S., & Reiter, J. (2013). *Cross-National Longitudinal Business Database : A Synthetic Data Approach.* Comparative Analysis of Enterprise Data Conference.

Willenborg, L. a. (2001). *Elements of Statistical Disclosure Control.* Lecture Notes in Statistics, 155 . Springer Verlag, New York (2001).

ミクロデータ利用ポータルサイト miripo. (2020 年 8 月). ミクロデータ利用ポータルサイト miripo. 参照先: https://www.e-stat.go.jp/microdata/

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. 2009, vol. 15, no. 3/4, p. 197-232.

伊藤伸介. (2016). 諸外国における政府統計データの提供の動向について. (中央大学経済研究所 Discussion Paper No. 267) 1-13 ページ.

伊藤伸介. (2018a). 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, 77〜101 頁.
伊藤伸介. (2018b). 公的統計ミクロデータの利活用の動向とわが国における課題. 『統計』, 2018 年 6 月号, p.13-18.
伊藤伸介. (2020). 海外における公的統計と行政記録の二次活用の現状. 『統計』 2020 年 8 月号, p10-15.
伊藤伸介, 横溝秀始. (2020a). 経済センサスのミクロデータを用いた匿名化手法の適用可能性に関する実証研究. 総務省統計研究研修所, リサーチペーパー第 49 号.
伊藤伸介, 横溝秀始. (2020b). 事業所・企業系の統計調査に対する匿名化措置の可能性について. 経済統計学会第 64 回（2020 年度）全国研究大会報告要旨集.
伊藤伸介, 星野なおみ, 阿久津文香, 菊池亮. (2018). 匿名化された公的統計ミクロデータの作成における攪乱的手法の有効性の評価. 経済学論纂（中央大学）第 59 巻第 1 ・ 2 合併号（2018 年 9 月）.
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. 統計研究彙報, 第 71 号, 2014 年 3 月.
稲葉由之. (2017). 攪乱的方法を用いて作成する匿名データに関する基礎研究 . 総務省統計研究研修所, リサーチペーパー第 39 号 .
永島勝利. (2018). 公的統計データの二次的利用. 総務省統計委員会担当室, http://www.iwsec.org/pws/pwscup/PWSCUP2017_files/PWSMeetup_8_nagashima.pdf. 参照先: http://www.iwsec.org/pws/pwscup/PWSCUP2017_files/PWSMeetup_8_nagashima.pdf
横溝秀始, 伊藤伸介, 竹村彰通. (2020). 事業所・企業系の匿名化ミクロデータの作成可能性に関する一考察. 2020 年度統計関連学会連合大会公演報告集, 255 頁.
河野真理子, 和田かず美. (2018). ミクロデータ分析のための演習用教材の作成方法〜一般用ミクロデータ詳細品目版及び擬似ミクロデータによる事例〜. 統計研究彙報, 第 75 号, 2018 年 3 月, (61〜80).
佐野夏樹, 服部雄太. (2020). モデルの判別精度によるグローバルリコーディングの有用性評価. 統計研究彙報第 77 号 2020 年 3 月（1-14）.
秋山裕美, 山口幸三, 伊藤伸介, 星野なおみ, 後藤武彦. (2012 年 7 月). 教育用擬似ミクロデータの開発とその利用 〜平成 16 年全国消費実態調査を例として〜. 統計センター 製表技術参考資料 16 (2012 年 7 月). 参照先: https://www.nstac.go.jp/services/pdf/sankousiryou2407.pdf
小林良行. (2011). 匿名データの教育目的利用に関する一考察. 『統計学』第 100 号 2011 年 3 月.
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), 23-45.
総務省政策統括官. (2018 年 6 月). 匿名データの作成・提供に関するガイドライン. 参照先: https://www.soumu.go.jp/main_content/000631450.pdf
総務省統計局. (2017 年 1 月). 総務省統計局における匿名データの作成・提供の概要. 参照先: https://www.soumu.go.jp/main_content/000462274.pdf
瀧敦弘. (2003). 集計表におけるセル秘匿問題とその研究動向. 『統計数理』, 第 51 巻, 第 2 号, 2003 年, pp.337-350 .
竹村彰通. (2003). 個票開示問題の研究の現状と課題. 統計数理, 51, 241-260.
東洋経済新報社. (2020 年 6 月). 会社四季報. 参照先: https://shikiho.jp/
独立行政法人統計センター. (2005 年 8 月). 統計データ開示抑制に関する用語集改訂版（対訳） 製表関連国際用語集 No.2. 参照先: https://www.nstac.go.jp/services/pdf/skk-yogosyu2.pdf
独立行政法人統計センター. (2006 年 3 月). 汎用秘匿処理ソフトウェア τ -ARGUS による集計表の秘匿処理 順次 LP 法と τ -ARGUS 搭載の集計表秘匿処理法の比較結果及び最近の集計表開示抑制法の研究動向. 独立行政法人統計センター 製表技術参考資料 4. 参照先: https://www.nstac.go.jp/services/pdf/sankousiryou1803.pdf
独立行政法人統計センター. (2014 年 10 月). 国勢調査における匿名データの作成とその検証. 参照先: https://www.nstac.go.jp/services/pdf/sankousiryou2610.pdf
独立行政法人統計センター. (2020a 年 8 月). 一般用ミクロデータの利用. 参照先: https://www.nstac.go.jp/services/ippan-microdata.html
独立行政法人統計センター. (2020b 年 7 月). 匿名データの利用に関する FAQ (回答) 6 匿名データは、どのようなファイル構成となっていますか？ 参照先: https://www.nstac.go.jp/services/faq-a-anonymity.html#Q07
内閣官房. (2013 年 12 月). 技術検討ワーキンググループ報告書（第 5 回 パーソナルデータに関する検討会) . 参照先: https://www.kantei.go.jp/jp/singi/it2/pd/dai5/siryou2-1.pdf
日本経済新聞社. (2020 年 6 月). NEEDS-FinancialQUEST. 参照先: http://www.nikkei.co.jp/needs/fq/
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, 69～99 頁.

付録

付録 A レコード削除の検討

4.4 では質的属性の客観評価を行い、リコーディングの荒さを変えたキー変数別の 3-匿名性を満たさないレコード数とその割合を算出した。この際、3-匿名性を満たさないレコードを削除せずに評価したが、実務においてはこれらのレコードをそのまま公開することは事業所の特定化リスクを高めることとなる。その対策のひとつとして、最もシンプルな手法にレコード削除がある。レコード削除とは、リスクの高い事業所を文字通りデータセットから削除することで秘匿性を高める手法であり、わが国や諸外国でも一般に広く用いられている。そこで、経済センサスにおいてレコード削除が適用可能かどうかの検討を行った。

具体的には、リコーディングの種類を変えたキー変数別の 3-匿名性を満たさないレコードを削除し、前後の要約統計量の確認を行った。以下の表 A に、分類区分を変更したキー変数別の、売上(収入)金額、従業者合計、資本金額における平均値、標準偏差、中央値の変化率を一覧にまとめた。index1 は最も分類区分が細かいため、ひとつひとつの層に含まれるレコード数が少なく、レコード削除が発生しやすい。そのため、レコード数は原データから 33.8%削除されている。その結果、売上(収入)金額の平均値では 77.5%、資本金額の平均値にいたっては実に 98.4%もの低下が発生してしまっている。最も荒い分類区分の組である index 16 においては、レコード削除は 2.3%程度と比較的少ないが、それでも売上(収入)金額の平均値で 10.6%、資本金額の平均値で 23.7%の低下と、やはり無視できないレベルの要約統計量の変化が発生している。少ないレコード削除からでも要約統計量に大きな差異が発生するのは、削除対象となっているレコードが相対的に大きな売上(収入)金額や資本金額を有しているからであると考えられる。これは、表 10〜表 13 で示した属性ごとの構成比とも整合的である。

表A　レコード削除による要約統計量の変化率

						売上（収入）金額		従業者合計		資本金額	
index	地域	産業	従業者規模	資本金階級	レコード数	平均値	標準偏差	平均値	標準偏差	平均値	標準偏差
1	8区分	24区分	13区分	11区分	-33.8%	-77.5%	-89.3%	-54.6%	-76.7%	-98.4%	-99.8%
2	8区分	24区分	13区分	5区分	-22.2%	-62.9%	-73.5%	-39.8%	-50.5%	-89.6%	-80.2%
3	8区分	24区分	5区分	11区分	-22.1%	-60.2%	-33.1%	-43.7%	-47.7%	-67.6%	-29.1%
4	8区分	24区分	5区分	5区分	-12.3%	-43.1%	-25.6%	-29.4%	-30.6%	-57.8%	-30.1%
5	8区分	11区分	13区分	11区分	-23.9%	-71.1%	-83.0%	-50.0%	-73.8%	-94.0%	-83.2%
6	8区分	11区分	13区分	5区分	-13.7%	-56.4%	-69.1%	-35.3%	-49.1%	-82.5%	-73.3%
7	8区分	11区分	5区分	11区分	-14.3%	-44.5%	-18.1%	-34.3%	-37.0%	-44.3%	-13.3%
8	8区分	11区分	5区分	5区分	-6.4%	-21.1%	-8.0%	-14.7%	-10.9%	-34.9%	-16.1%
9	3区分	24区分	13区分	11区分	-21.6%	-71.6%	-85.0%	-48.7%	-70.5%	-90.9%	-82.8%
10	3区分	24区分	13区分	5区分	-11.5%	-55.5%	-65.1%	-33.8%	-43.7%	-81.9%	-74.9%
11	3区分	24区分	5区分	11区分	-12.4%	-44.3%	-26.2%	-29.9%	-28.3%	-51.0%	-27.3%
12	3区分	24区分	5区分	5区分	-5.3%	-31.2%	-22.2%	-18.6%	-19.6%	-48.1%	-30.0%
13	3区分	11区分	13区分	11区分	-12.8%	-61.1%	-69.9%	-41.1%	-63.3%	-75.1%	-55.5%
14	3区分	11区分	13区分	5区分	-5.3%	-45.3%	-53.0%	-28.1%	-40.0%	-64.3%	-54.7%
15	3区分	11区分	5区分	11区分	-6.4%	-27.0%	-11.5%	-19.0%	-18.3%	-31.4%	-14.3%
16	3区分	11区分	5区分	5区分	-2.3%	-10.6%	-2.6%	-7.0%	-4.4%	-23.7%	-15.4%

以上のことから、世帯・人口系の匿名化ミクロデータ作成では一般によく用いられるレコード削除も、事業所・企業系においては慎重に取り扱う必要があると言える。削除対象となるレコードを最小限に絞り込む、残存レコードの要約統計量と照らし合わせて削除レコードを決定する、レコード削除後に残存レコードに補正をかけるなどの方策が考えられるが、いずれにおいても精査が必要である。

また、レコード削除を行わずに別の匿名化手法を検討する可能性も考えられる。例えば、3-匿名性を満たさないレコードは、類似した別の層に質的属性を攪乱する、あるいは類似した別の層とまとめて量的属性に対するミクロアグリゲーションを行うなどの手法である。質的属性の攪乱については、PRAM (post randomization method) （Kooiman *et al.* (1997)）や質的属性のミクロアグリゲーション（Torra (2004)）の先行研究があるため、これらの適用を検討することも今後の課題のひとつである。

付録 B 分類区分別の事業所数と高リスク事業所数割合

図 B-1 地域×経営組織

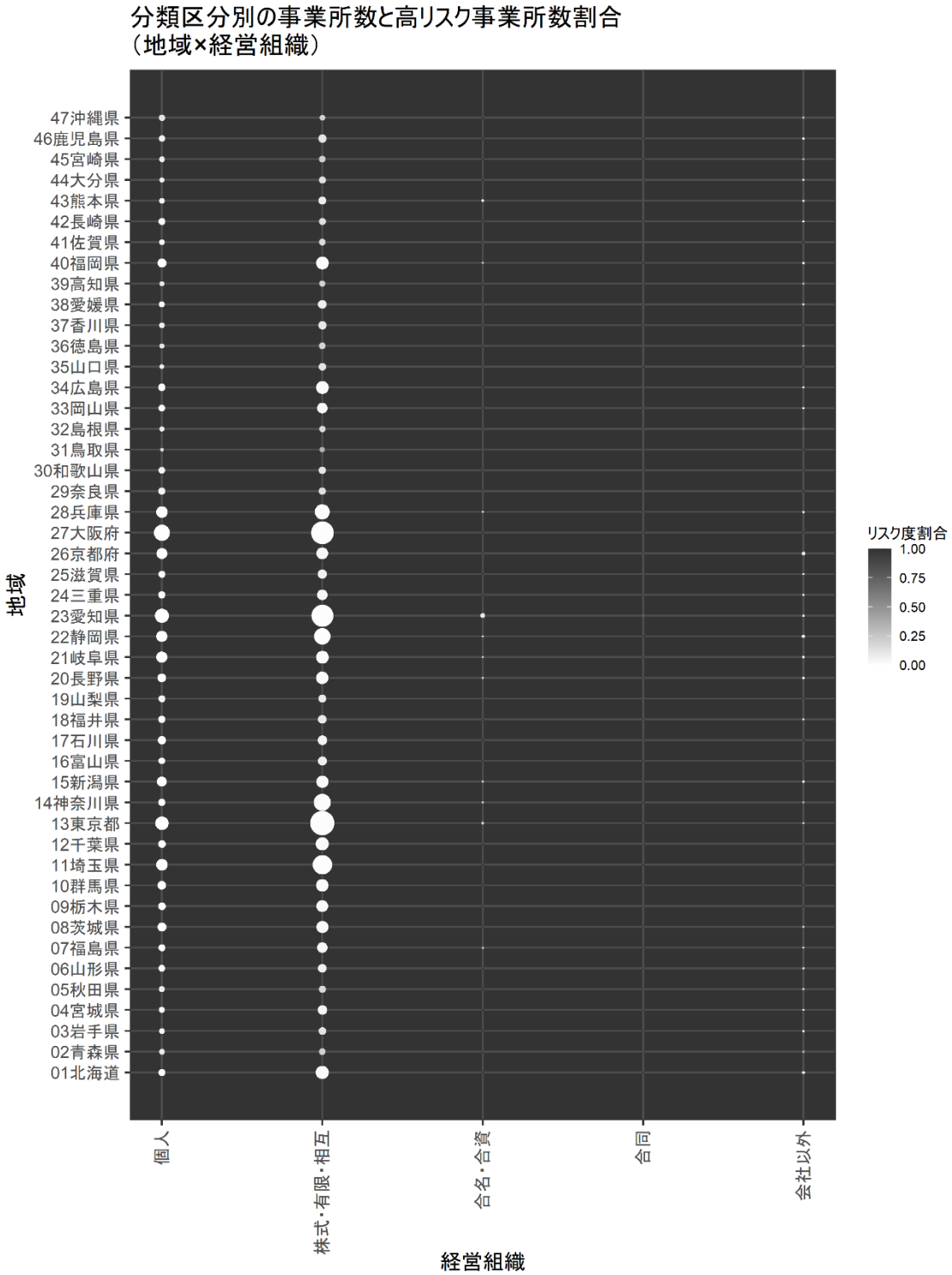


図 B-2 地域×開設時期

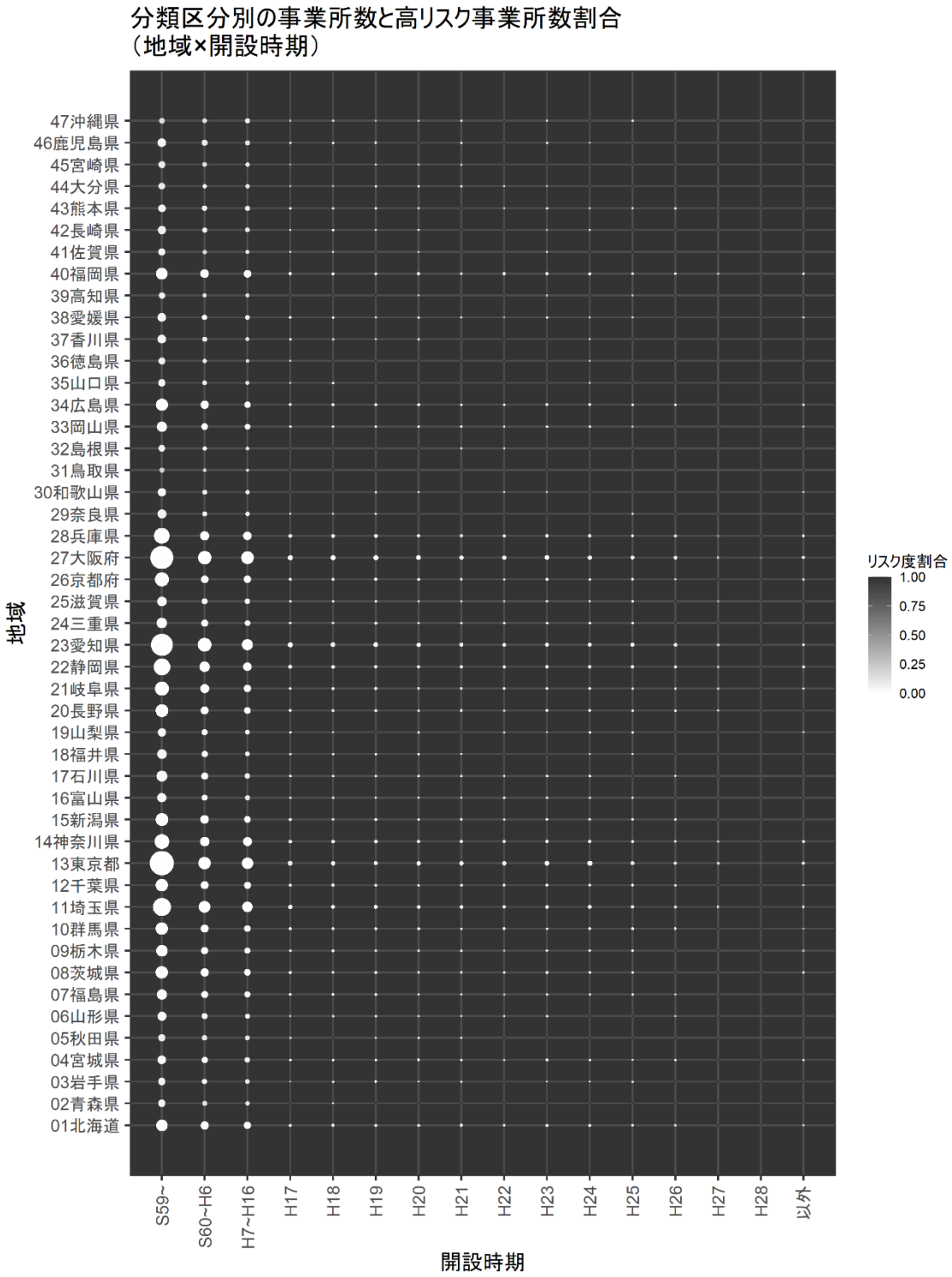

分類区分別の事業所数と高リスク事業所数割合
（地域×開設時期）
47沖縄県
46鹿児島県
45宮崎県
44大分県
43熊本県
42長崎県
41佐賀県
40福岡県
39高知県
38愛媛県
37香川県
36徳島県
35山口県
34広島県
33岡山県
32島根県
31鳥取県
30和歌山県
29奈良県
28兵庫県
27大阪府
26京都府
25滋賀県
24三重県
23愛知県
22静岡県
21岐阜県
20長野県
19山梨県
18福井県
17石川県
16富山県
15新潟県
14神奈川県
13東京都
12千葉県
11埼玉県
10群馬県
09栃木県
08茨城県
07福島県
06山形県
05秋田県
04宮城県
03岩手県
02青森県
01北海道
地域
S59~
S60~H6
H7~H16
H17
H18
H19
H20
H21
H22
H23
H24
H25
H26
H27
H28
以外
開設時期
リスク度割合
1.00
0.75
0.50
0.25
0.00

図 B-3 地域×単独・本所・支所の別

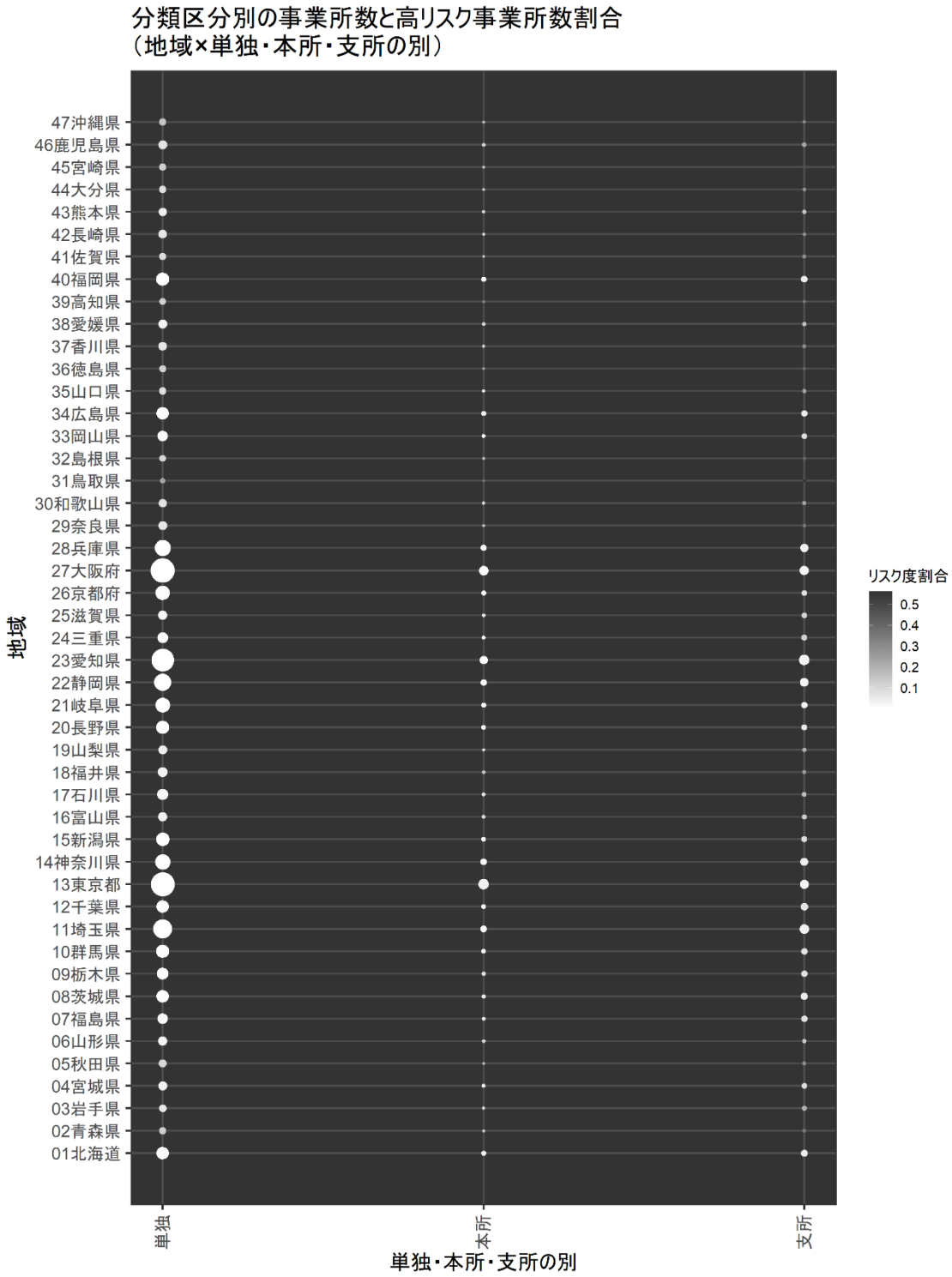


図 B-4 産業×経営組織

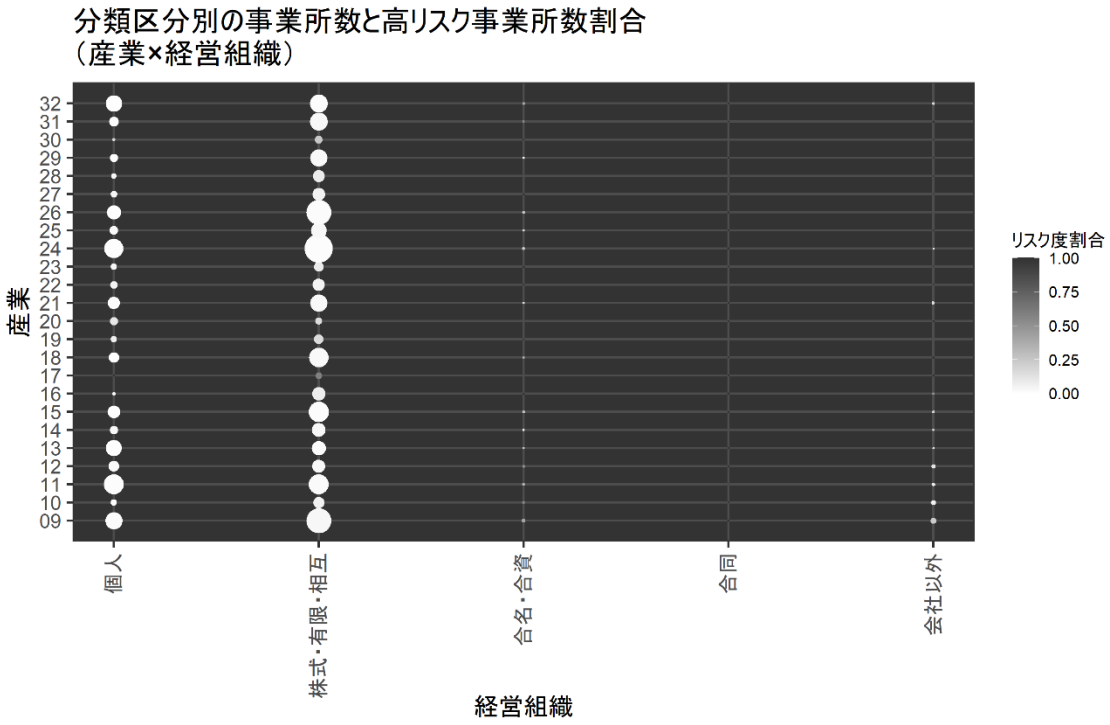


図 B-5 産業×開設時期

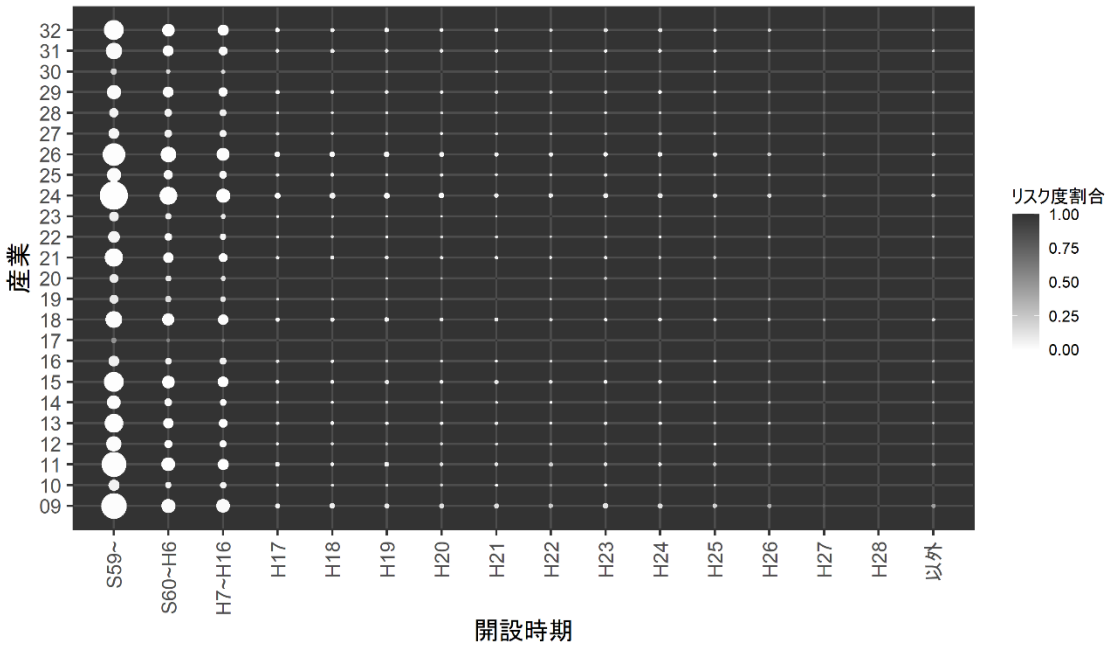


図 B-6 産業×単独・本所・支所の別

図 B-7 従業者規模×経営組織

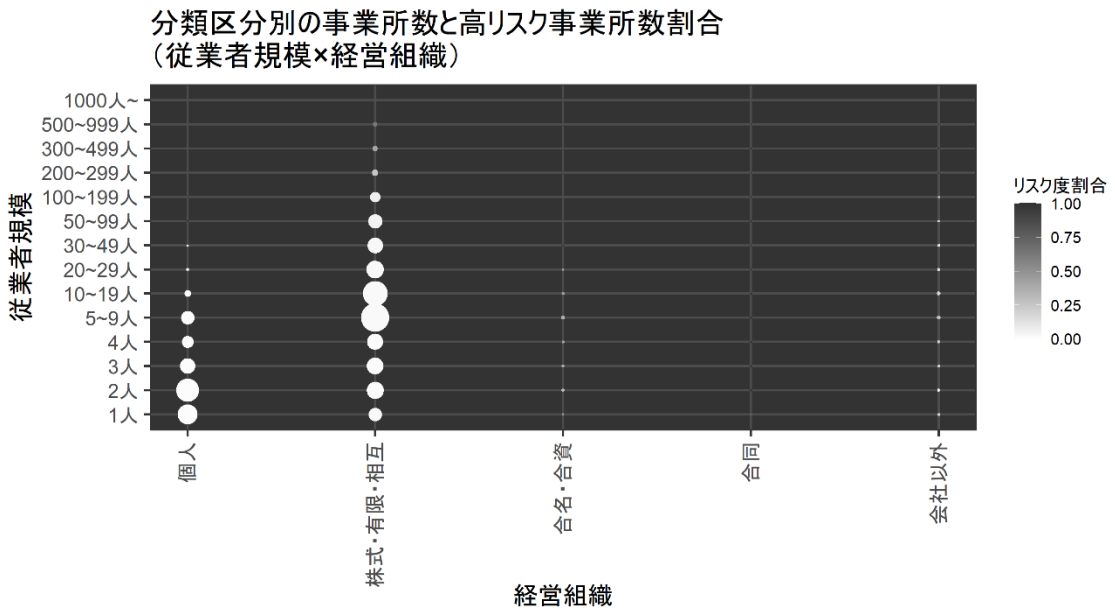


図 B-8 従業者規模×開設時期

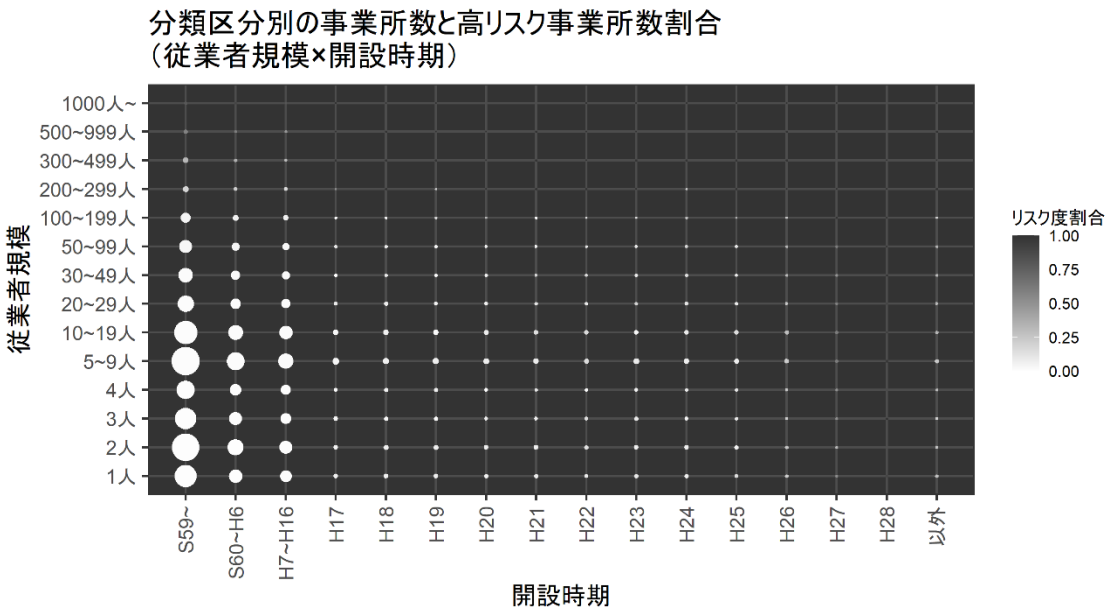


図 B-9 従業者規模×単独・本所・支所の別

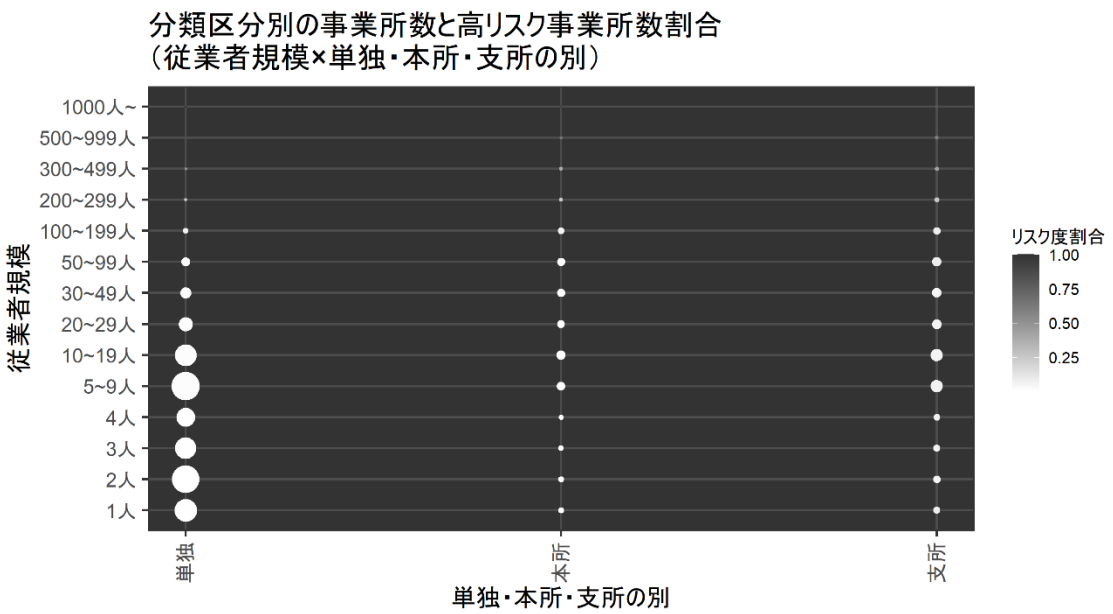


図 B-10 資本金階級×経営組織

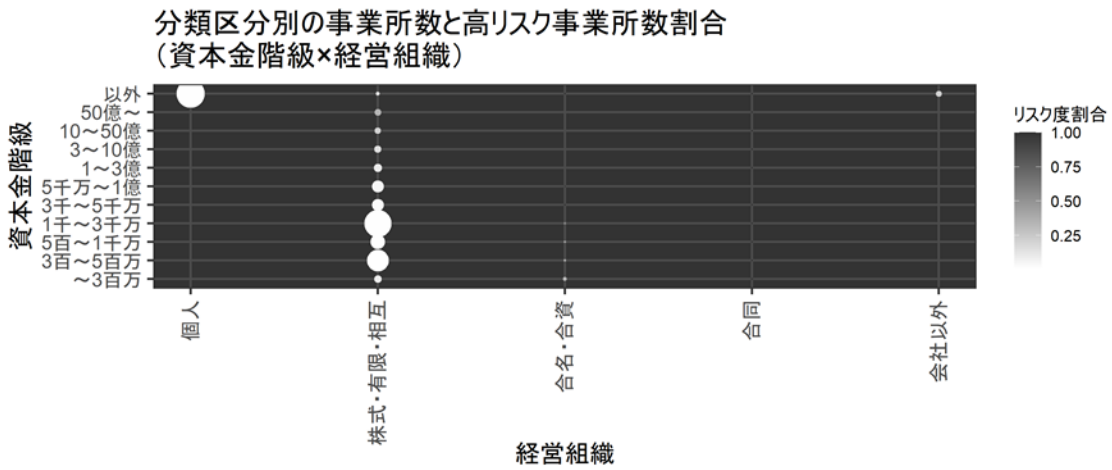


図 B-11 資本金階級×開設時期

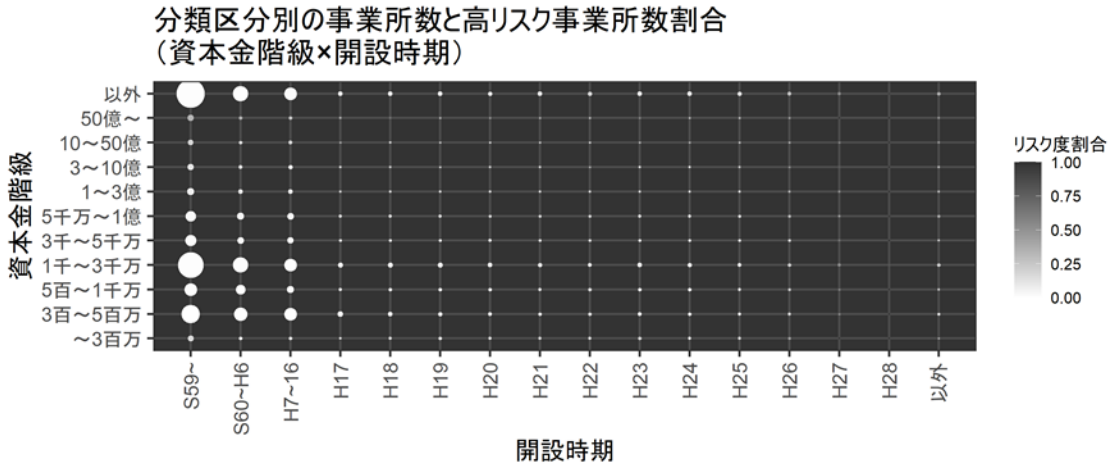


図 B-12 資本金階級×単独・本所・支所の別

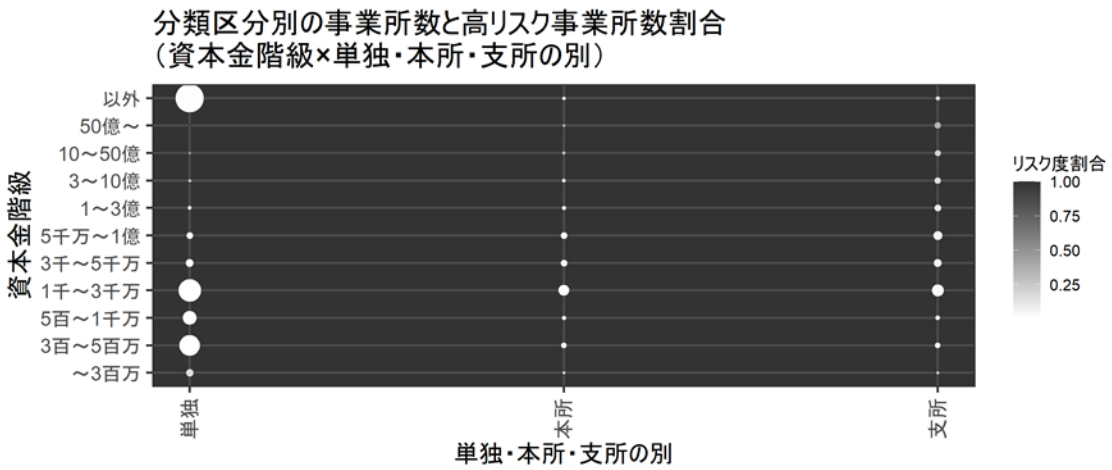


図 B-13 売上（収入）金額階級×経営組織

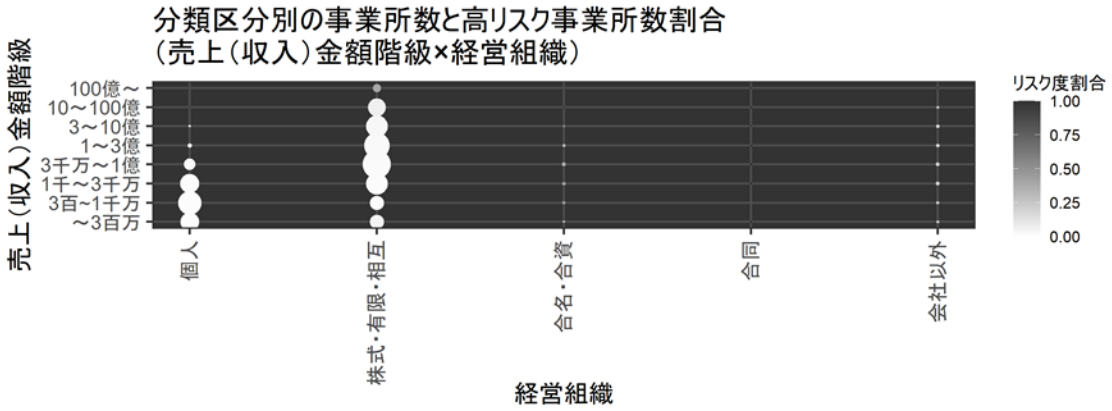


図 B-14 売上（収入）金額階級×単独・本所・支所の別

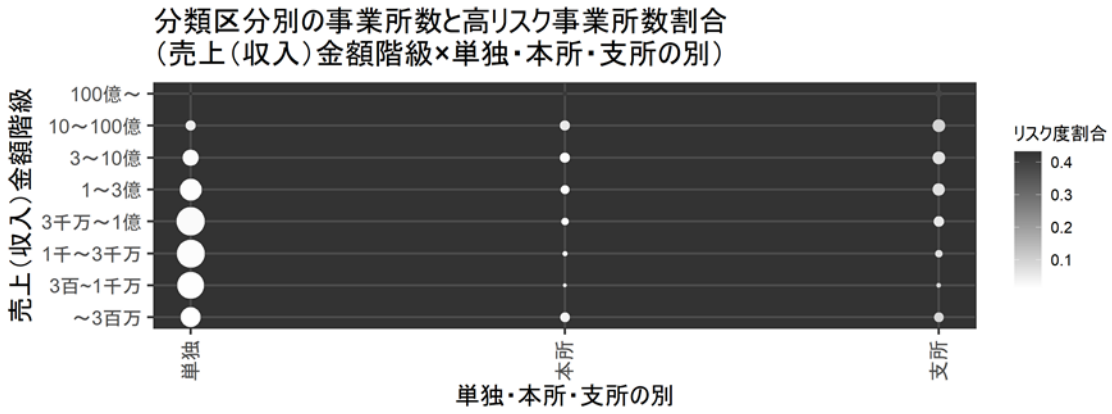


図 B-15 売上（収入）金額階級×開設時期

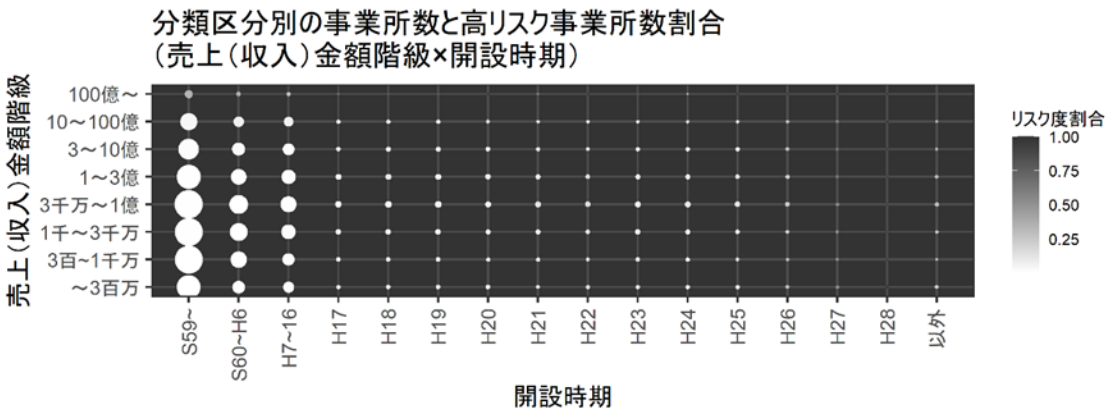


図 B-16 経営組織×単独・本所・支所の別

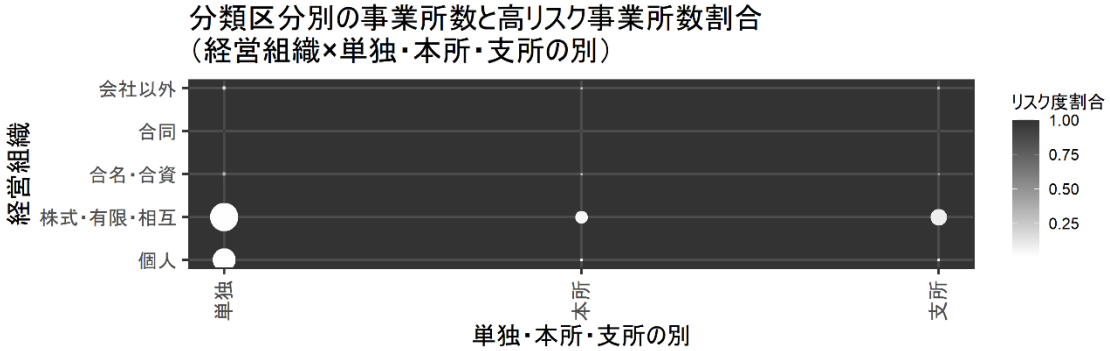


図 B-17 経営組織×開設時期

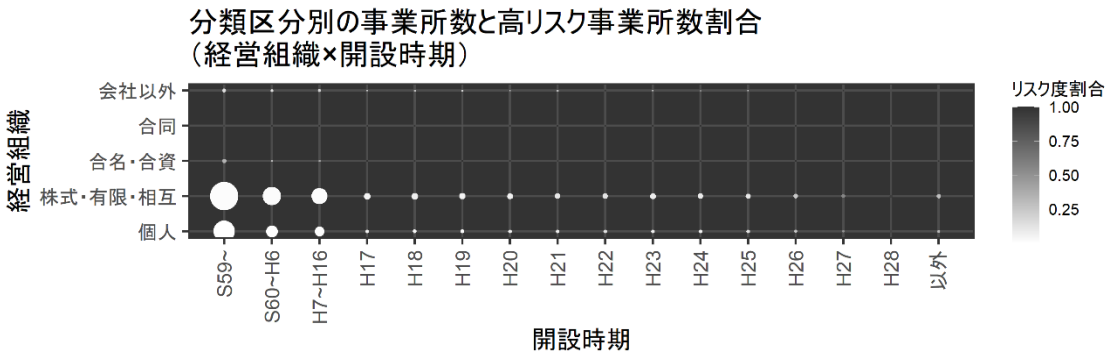


図 B-18 開設時期×単独・本所・支所の別