リサーチペーパー第 49 号　　Research Paper No.49

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

Empirical Study on the Applicability of Disclosure Limitation Methods Based on Microdata from the Japanese Economic Census

伊藤 伸介
統計研究研修所客員教授
中央大学経済学部教授
ITO Shinsuke
SRTI Guest Professor
Professor, Faculty of Economics, CHUO University

横溝 秀始
統計局総務課
滋賀大学大学院データサイエンス研究科修士課程
YOKOMIZO Shuji
General Affairs Division, Statistic Bureau
Master’s Course, Graduate School of Data Science, SHIGA University

令和 3 年 1 月
January 2021

総務省統計研究研修所
Statistical Research and Training Institute (SRTI)
Ministry of Internal Affairs and Communications

受理日：令和 2 年 12 月 24 日（一部修正：令和 3 年 2 月 3 日）
本ペーパーは、総務省統計研究研修所の客員教授及び総務省統計局職員である執筆者が、その責任において行った統計研究の成果を取りまとめたものであり、その内容については、総務省統計局又は統計研究研修所の見解を表したものではない。本ペーパーの内容については、執筆者に問い合わせ願いたい。
本研究では、統計法(平成 19 年法律第 53 号)第 32 条の規定に基づき、経済センサス基礎調査及び活動調査に係る調査票情報を使用した。

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

伊藤 伸介
横溝 秀始

概要

わが国では、現在 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・企業系の統計調査については、匿名データの作成は行われていない。海外においては、イタリア、ドイツ、Eurostat において事業所・企業系の匿名データが作成された事例があるものの、海外の多くの国々では、事業所・企業系のミクロデータは、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。

本稿では、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した。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した。

事業所・企業系のデータについては、把握される量的変数の分布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っており、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。さらに、企業に関する財務情報など、外部に開示される企業情報も存在することから、侵入者(intruder)は精度の高い外部情報を容易に取得できる場合がある。このことから、事業所・企業系のデータの露見(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知られている。

こうした事業所・企業系のデータ特性を踏まえて、本稿では、特定の露見シナリオを想定した上で、非攪乱的手法だけでなく、攪乱的手法を用いた上で、経済センサスを例に、各種の匿名化手法を適用することによって作成し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を行った。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が、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。

キーワード：匿名化ミクロデータ、経済センサス、匿名化手法、事業所・企業系データ、
有用性、秘匿性

Empirical Study on the Potential for Disclosure Limitation Methods Based on Microdata from the Japanese Economic Census

ITO Shinsuke
YOKOMIZO Shuji

Abstract

Anonymized business microdata is a valuable resource that can be used not only for academic research but also educational purposes and as test data for creating various programs. While the Japanese government currently creates and provides access to anonymized data from multiple types of household and demographic statistical surveys, it does not provide access to anonymized data from business surveys. This contrasts with the situation in other countries such as Germany and Italy, as well as institutions such as Eurostat, who are already creating anonymized business microdata. However, the data is available only via on-site access and/or remote access.

This paper provides an overview of how anonymized business data are prepared in other countries and then organizes issues to be discussed regarding the preparation of anonymized business data in Japan. This paper also examines possible anonymization procedures for business microdata in Japan using respondent-level data from the Economic Census.

In business data, the distributions of variables exhibit extreme variations, and the number of records to be sampled varies significantly depending on firm size and in sampling there may be layers subject to complete enumeration. Furthermore, because the data include company-specific information (such as financial information, that is intended for external disclosure) system intruders are potentially able to obtain highly detailed external information. This makes the risks associated with providing access to business data potentially more serious than those associated with providing access to individual and household survey data.

This paper uses and compares different anonymization methods and employs both perturbative and non-perturbative methods in order to quantitatively evaluate usability and confidentiality of microdata. This paper conducts exploratory basic research on business microdata, and further examination is needed regarding anonymization methods that takes into account not only the characteristics of business data, but also practical statistical procedures.

Keywords : Anonymized Microdata, Economic Census, Anonymization Methods, Business Data, Usability, Confidentiality

1　はじめに

海外では、世帯・人口系の統計調査を対象に、個票データ(deidentified data)に対して匿名化措置を適用した匿名化ミクロデータ(anonymized microdata)の提供が行われている。それに対して、Eurostat、イタリア、ドイツといった国々では、事業所・企業系の統計調査についても、匿名化ミクロデータが作成されていることが注目される。

事業所・企業系の統計調査のミクロデータの提供は、欧米諸国においてはオンサイト利用やリモートアクセスによるデータの利用サービスによって近年展開されてきた(伊藤(2018))。わが国では、国勢調査、全国消費実態調査、社会生活基本調査、就業構造基本調査、住宅・土地統計調査、労働力調査、国民生活基礎調査といった 7 種類の世帯・人口系の統計調査が匿名データとして提供されているが、事業所・企業系の統計調査は、現在匿名データの作成の対象になっていないことが知られている。

しかしながら、事業所・企業系の匿名化ミクロデータにおいてもニーズは存在すると考えられる。その理由としては、事業所や企業の経済活動に焦点を当てて、社会経済の実態について匿名化ミクロデータを用いた実証分析を行うことができれば、世帯・人口系の調査からでは得られない様々な知見が得られることが指摘できる。

事業所・企業系の匿名化ミクロデータの利用目的に関しても、学術研究に対する利用や高等教育のための活用が考えられる。具体的には、学術研究目的の利用においては、研究者が匿名化ミクロデータを用いて、探索的な研究を行うことが想定される。事業所・企業系のミクロデータを利用するためには、統計法 33 条に基づく調査票情報の利用申出が必要になっている。その一方で、統計法 36 条に基づいて匿名データが提供されれば、公的統計ミクロデータの利用の起点として、匿名データが利用され、それがオンサイト利用の促進にもつながることから、公的統計の二次的利用のさらなる推進を図ることも期待できる。

他方、高等教育目的での活用については、例えば、秋山他(2012)で指摘しているように、事業所・企業系の統計調査に関するミクロデータを教育目的のために利用することは困難だと言える。それは、わが国において大学等の授業で利用可能な事業所・企業系の匿名化ミクロデータが存在しないからである。独立行政法人統計センターは、一般用ミクロデータとして、世帯・人口系の調査である全国消費実態調査や就業構造基本調査を提供しているが、事業所・企業系の統計調査は、一般用ミクロデータの作成・提供の対象となっていない。近年、統計教育の重要性が叫ばれているが、教育目的を指向した事業所・企業系の匿名化ミクロデータは、その一助になると考えられる。

2018 年の「公的統計の整備に関する基本的な計画(第Ⅲ期基本計画)(以下、『基本計画』という。)」では、厚生労働省所管の事業所・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いる。『基本計画』では、社会・経済情勢の変化を的確に捉える統計の整備の一環として、働き方の変化等をより的確に捉える統計の整備の必要性も指摘されている。このことは、労働供給側だけでなく、労働需要側からもミ

クロレベルで実証分析を行う上で、事業所・企業系の匿名データの作成方法を追究することの必要性を示唆している。

以上のように、わが国においても、事業所・企業系の匿名化ミクロデータのニーズは存在すると考えられる。このような状況を踏まえて、本稿では、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を踏まえた上で、わが国における事業所・企業系の匿名化ミクロデータの可能性を検討する。具体的には、経済センサスの個票データに対して、攪乱的手法を含む各種の匿名化手法を適用した上で作成した匿名化ミクロデータの有用性と秘匿性を定量的に把握することによって、事業所・企業系のミクロデータに対する匿名化技法の適用可能性を追究する。

2　事業所・企業系の匿名化ミクロデータの特徴

本節では最初に、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状況について概観する。次に、海外での状況を踏まえて事業所・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。

2.1　事業所・企業系の調査と世帯・人口系の調査に関するデータ特性

O'Keefe and Shlomo(2014) では、個人・世帯を対象にしたミクロデータと、事業所・企業に関するミクロデータの特徴を以下のように整理している(表 1)。事業所・企業系のデータの特性としては、①サンプルサイズが小さいこと、②調査対象に大企業が含まれていること、そして③大企業における属性値のほとんどが外れ値であることが指摘されている。したがって、事業所・企業系の匿名化ミクロデータの作成においては、大企業に含まれる属性情報に対してどのような秘匿措置を施すかが論点になると言える。

表 1 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 and 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確保が難しいことが指摘されている。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)は小さく、量的変数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るため、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。さらに、わが国で上場している企業について有価証券報告書の開示義務があるように、侵入者(データの利用者、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。そして、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。

Franconi and Ichim(2007)は、サンプリングの問題が指摘されているほか、事業所・企業系のミクロデータにおける売上高や輸出額のような属性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述べている。星野(2010)は、大きな企業や事業所は全数調査されることから調査された事が既知となること、外れ値の存在がビジネスデータの匿名化を難しくすることを指摘している。さらに、Ritchie *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な外れ値が認識されることであることを論じている。

2.2　イタリアやドイツに関する先行事例

ISTAT(イタリア国立統計研究所)では 2020 年現在、企業のイノベーション活動の調査である CIS(=Community Innovation Survey)の匿名化ミクロデータが公開されている。CIS は EU 内での比較可能性を考慮した標本調査であり、主な変数は、経済活動(産業分類)、地理的区分、従業員数、売上高、研究費等があげられる。この CIS について、学術研究用ファイル(Scientific Use File=SUF)や一般公開型ファイル(Public Use File=PUF)が現在提供されている。Ichim(2007)では、1998 年から 2000 年の間に調査された CIS3 を例に、SUF 作成手順を体系的に示している。Ichim(2007)によれば、以下のステップを踏むことが推奨されている(図 1)。

図 1 CIS の SUF 作成手順

最初に、①露見シナリオの定義を行う。CIS では、外部参照情報(external register)に含まれる識別情報をもとにリンケージされることへの対策と、非常に大きな売上等から偶発的な個体特定(識別)(spontaneous identification)が行われることへの対策が中心となっている。主要な識別変数は、産業分類、地域、従業員数、売上高である。次に、②変数の前処理として、産業分類、地域、従業員数といったキー変数に対してグローバルリコーディングを行う。その後、キー変数で層化を行ったのち、③リスクの高いレコードの特定を行う。この際、類似したレコードが少なければ露見リスクが大きい

と見なし、密度ベースのアルゴリズム[1]で攪乱対象のレコードを選択している。その後、④ミクロデータの攪乱においては、リスクが高いと判断された売上高の外れ値には最近傍のクラスターからの補完(the nearest clustered unit imputation)が行われるが、分布の右裾については個別ランキング法によるミクロアグリゲーション[2]が行われる。さらに、⑤情報量損失と情報量保護では、産業分類ごとの、売上高の分散の変化率や変数間の相関係数が考慮される。最後に、⑥公開するミクロデータファイルの説明に関する資料として、それぞれの変数について攪乱の有無や、イノベーション変数の比率や売上高の変化率がデータの有用性の尺度として明示されている。なお、いくつかのステップでは、統計調査の実務担当者の助言を考慮することの重要性が強調されている。

ドイツの連邦統計局でも、事業所・企業を対象とした複数の匿名化ミクロデータが、PUF や SUF、さらに高等教育のために強い匿名化が施された campus file(CF)の形で提供されている。ドイツの匿名化には、事実上の匿名性(factual anonymity)という概念が存在する。これは、「著しく大きな時間、経費および労力の支出によって、当事者に関連づけることができない」こと(濱砂(1999))を指し、連邦統計法に規定されている、学術研究用ファイル (SUF) を作成する上で重要となる概念である(伊藤(2020))。

ドイツでは 2002 年から 2005 年にかけて、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ことを目的とした「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(Factual Anonymisation of Business Microdata)が進展した(Lenz *et al.*(2006))。本プロジェクトでは、SUF 作成における匿名化手法の適用可能性が評価され、ミクロアグリゲーション(microaggregation)やノイズ付与(加法ノイズ(additive noise)および乗法ノイズ(multiplicative noise)等が有用であることが確認された。本実験の結果、情報量損失の観点からミクロアグリゲーションの中でも特に個別ランキング法(individual ranking methods)が SUF の作成に適していると判断された。

さらに、2006 年から 2008 年にかけて、「企業パネルデータに関する事実上の匿名化」プロジェクト(Business Statistics Panel Data and Factual Anonymisation)が展開された(Brandt *et al.*(2008))。このプロジェクトでは、匿名化に関する研究実績のある年次ベースの事業所・企業のデータを縦断的にリンケージするパネルデータの作成が試行されている。匿名化手法としては、ミクロアグリゲーション、乗法ノイズ、多重代入法(multiple imputation)が検討されたほか、マッチングリスクについては、従来

---

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

[2] ミクロアグリゲーションの方法的特徴については、伊藤(2009)を参照されたい。

の距離ベースだけでなく、時系列を考慮した相関ベース、分布ベース等の尺度の組み合わせを用いた評価が行われた。ドイツの賃金構造に関する統計調査である German Cost Structure Survey(1999-2002)を用いた検証実験では、記述統計量や属性値、相関係数を用いた有用性の評価や、リンケージ技法を用いた秘匿性の強度の確認が行われた。

以上のイタリア・ドイツの事例で注目すべき点としては、以下のように整理できる。まず露見シナリオについては、SUF 作成を前提に、偶発的な個体特定や外部情報を用いたマッチングに重点が置かれている。秘匿性については露見シナリオを考慮した定量的な評価基準に基づいて、また、有用性に関しては実用例のサーベイを基に複数の指標を考慮して、攪乱を最小限に抑えていることも特徴的である。そのための匿名化手法としては、グローバルリコーディングといった非攪乱的手法だけでなく、ミクロアグリゲーション等の攪乱的手法が採用されている。ミクロアグリゲーションの中でも特に、元データとの近似性が相対的に高い個別ランキング法が SUF 作成に適していることが明らかになっている。最後に、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に言及されている。今後、わが国で事業所・企業の匿名化ミクロデータの作成を検討する上で、これらの知見は重要になってくると考えられる。

2.3　事業所・企業系の匿名化に向けた考察

本節におけるサーベイから、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を指摘することができる。

最も重要な論点は、大規模な事業所・企業における秘匿処理である。世帯・人口系のデータの場合、世帯間の属性値における異質性が相対的に大きくないことから、サンプリングを前提とした秘匿処理が行われる。一方、事業所・企業系のデータは、規模の大きな事業や企業におけるレコード数が少なく、分布の歪みが大きい。また、無作為抽出では、規模の大きい事業所・企業系データ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。事業所数を例えば従業者規模別で見た場合、従業者が少ない事業所・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、売上高の観点から見ると、大規模な事業所・企業が全体の売上高に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、事業所の規模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、外部参照情報の入手可能性と、外部参照情報とのマッチングの危険性も大きな

課題となる。世帯・人口系の統計調査の場合、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。もしそのようなセンシティブな情報を有す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、大規模な事業所・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、会社四季報(東洋経済新報社)やNEEDS Financial QUEST(日経新聞社)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、特定化の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、匿名化手法の適用、さらには秘匿性や有用性に関する定量的な評価方法に影響を与える。世帯・人口系の調査のように変数の数が少なく、質的変数が多い場合は、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、量的変数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(伊藤他(2014))。

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。

3　経済センサスのミクロデータを用いた秘匿性と有用性の評価研究

3.1　使用するデータ

本節では、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、試行的に作成した匿名化ミクロデータを用いた実証研究を行う。本研究では、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[3]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。分析対象となる項目には、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 2)。なお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 2 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	

3.2　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 3 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は、複数事業所の調査事項に該当しない場合があるほか、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

[3] 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。

表 3 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

見られた。そのため、中央値も 0 となっている。

つぎに、質的属性、量的属性のそれぞれについて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ヒストグラムとしてプロットしたものが図 2 である(符号と名称の対応は、表 4 および表 5 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。これらの属性値については、特定化のリスクの相対的な高さを考慮した上で、匿名化の必要性があると考えられる。

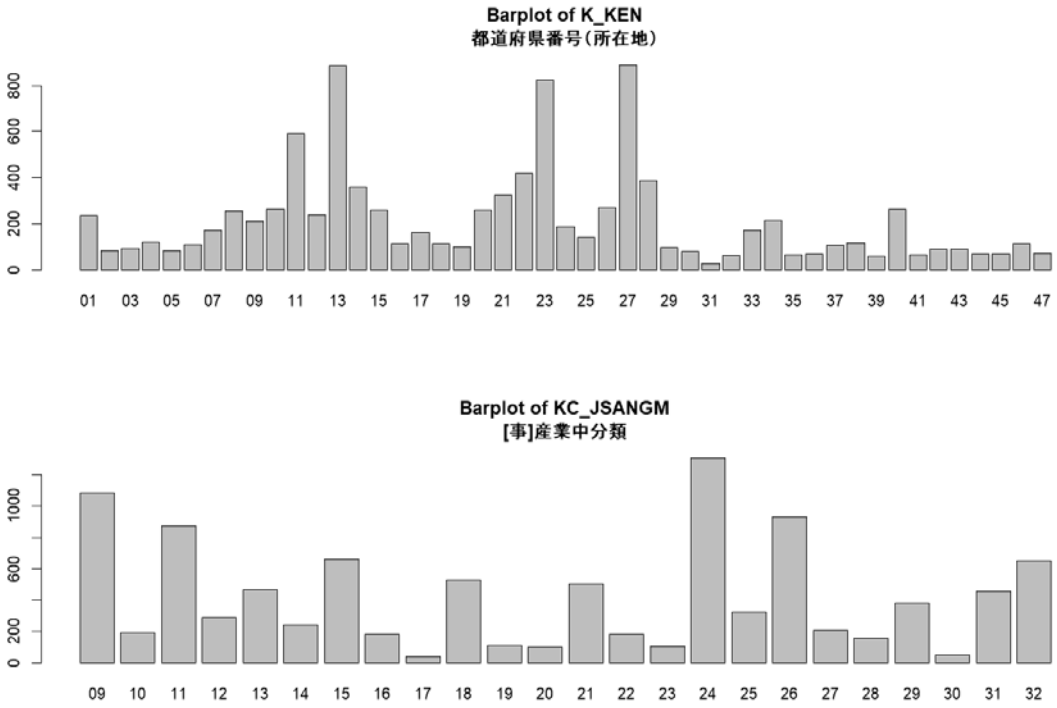


図 2 都道府県番号、産業中分類のヒストグラム

次に、量的属性のヒストグラムを作成した(図 3、図 4)。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、秘匿の観点から、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、会計上の理由のためか、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、度数が小さくなっている。

さらに、量的属性について相関係数行列を求めた。図 5 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。

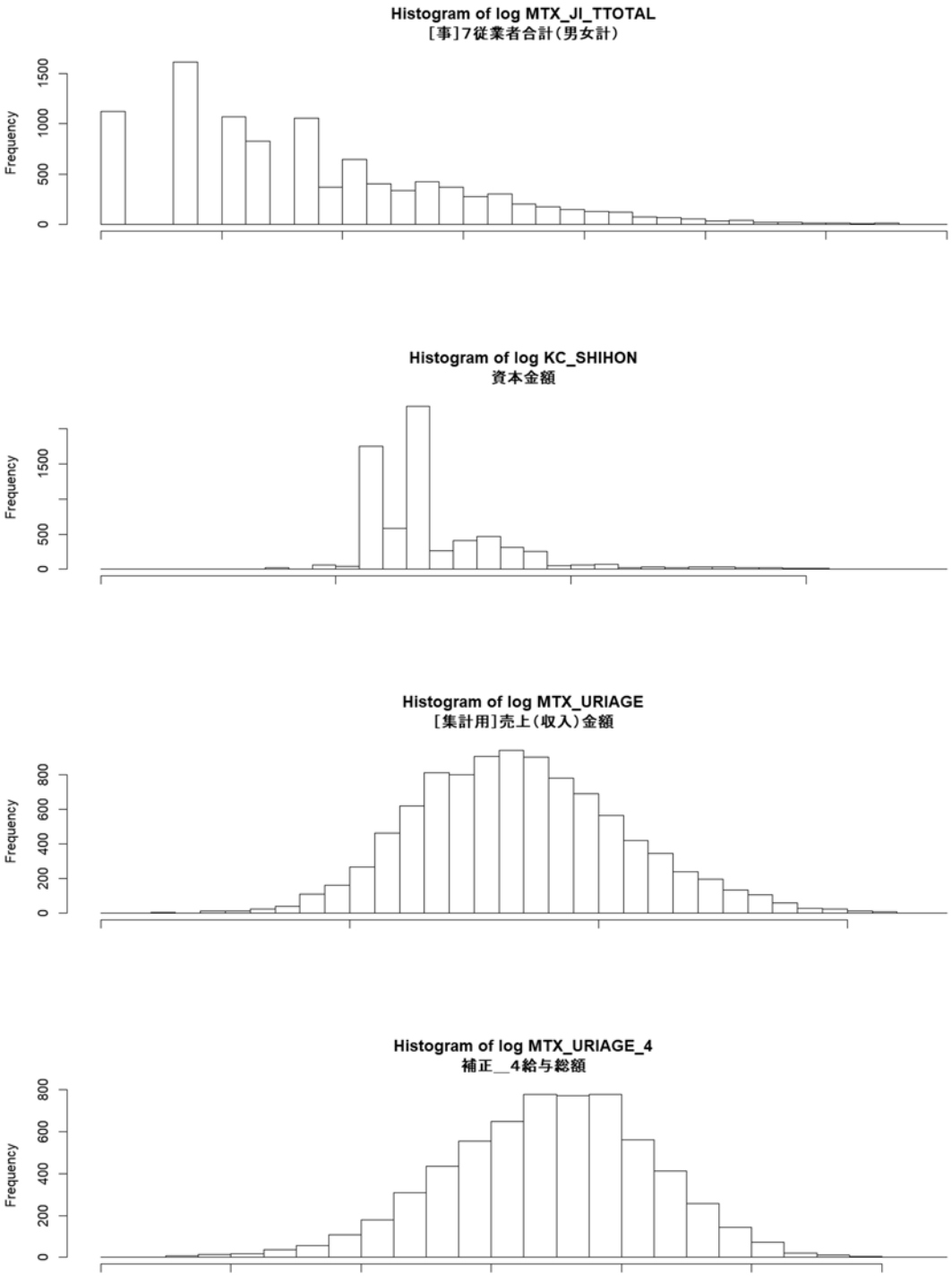


図 3 従業者合計、資本金額、売上(収入)金額、給与総額のヒストグラム

※秘匿上、横軸（対数）の目盛りは省略

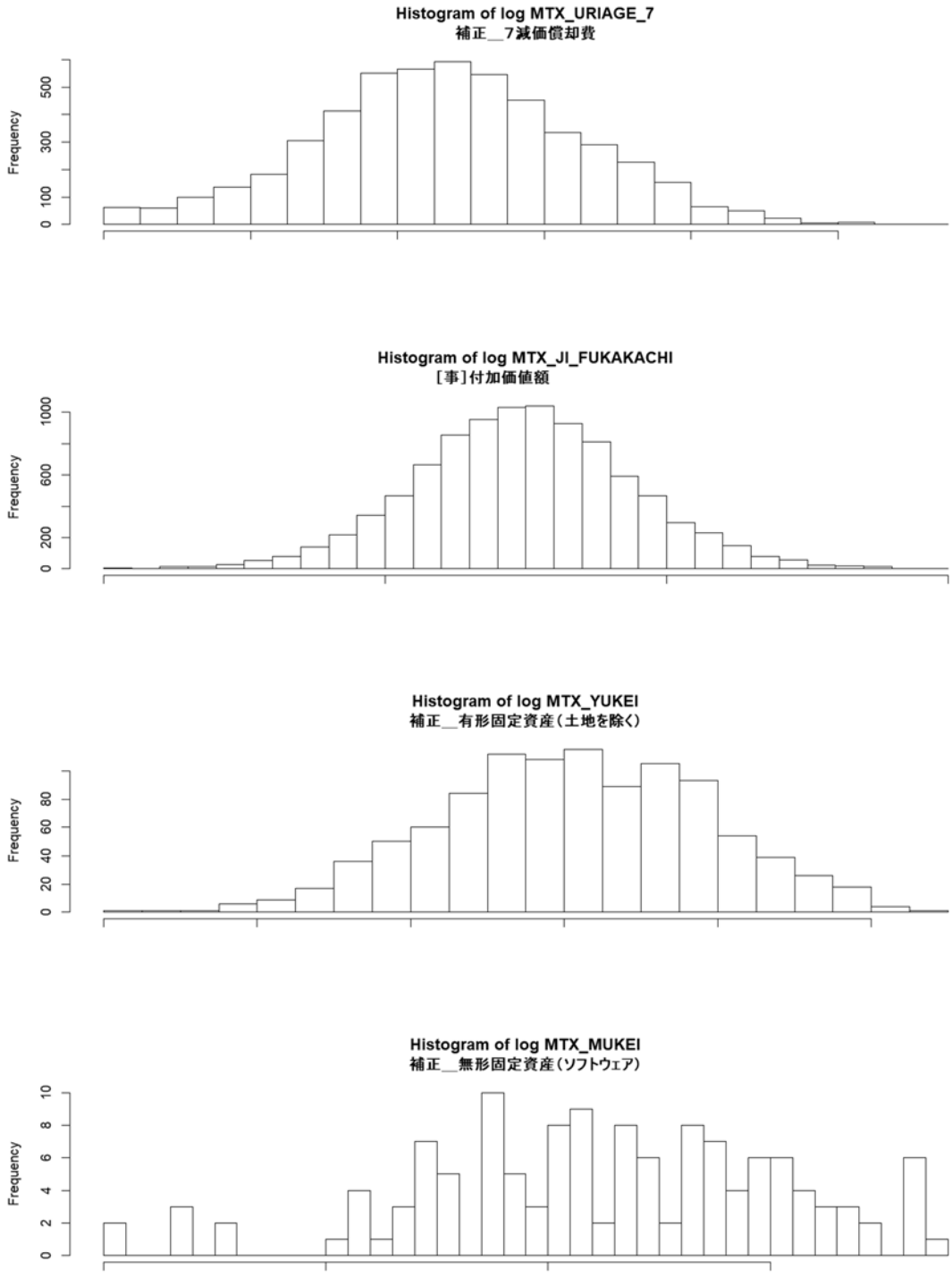


図 4 減価償却費、付加価値額、有形固定資産、無形固定資産のヒストグラム

※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 5 量的属性の相関係数行列

3.3 質的属性のリコーディング

本節では、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の作成において適用され、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の強度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、本研究では地域、産業、従業者合計、資本金額の 4 属性を中心に、リコーディングに基づく匿名化を行った。なお、これらは経済センサスにおける結果表において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ことから、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 4)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、地域ブロック 8 区分(北海道、東北、関東、中部、近畿、中国、四国、九州・沖縄)と 3 区分(東日本、中日本、西日本)を採用した。産業(表 5)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、製造業における産業中分類が、原区分をより粗くした区分になっているため、これを参考に 24 区分から 11 区分への統合を行っている。従業者合計(表 6)については、結果表における 15 区分に基づいて、13 区分(出向・派遣従業者のみおよび従業者数 1,000 人以上は実験条件で除外している)に区分統合した上で、従業者規模とした。また、それらをより粗い区分で統合した 5 区分も設定した。資本金額(表 7)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、それらをさらに粗い形で区分統合した 5 区分も設定した。

表 4 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 5 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%

※分類統合の都合上、産業中分類 32 を 20 の直後に置いた。

表 6 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 7 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%

3.4　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、原データと匿名化ミクロデータの間で度数が 1 となるセルの総数を比較し、度数 1 となるセル数の変化の確認を行う。

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、本研究でもその基準に基づいている。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[4]の概念に基づいて地域、産業、従業者規模、資本金階級の 4 属性で形成された層ごとに 3-匿名性を満たさない(以下、「3-匿名性違反」という。)レコード数を確認することと同等であると考えられる。表 8 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。

[4] 提供対象となっているデータが、準識別子のすべての組み合わせによって少なくとも k 個の個体を識別することができない場合、k-匿名性を持つと言う(Samarati and Sweeney(1998))。

表 8 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 9 は、分類区分を変更したキー変数の組み合わせ別の 3-匿名性違反のレコード数の割合を示している。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 1 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。図 6 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。統計実務上の観点では、標本の大きさを考慮してキー変数のリコーディングを考える必要があると考えられる。

表 9 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

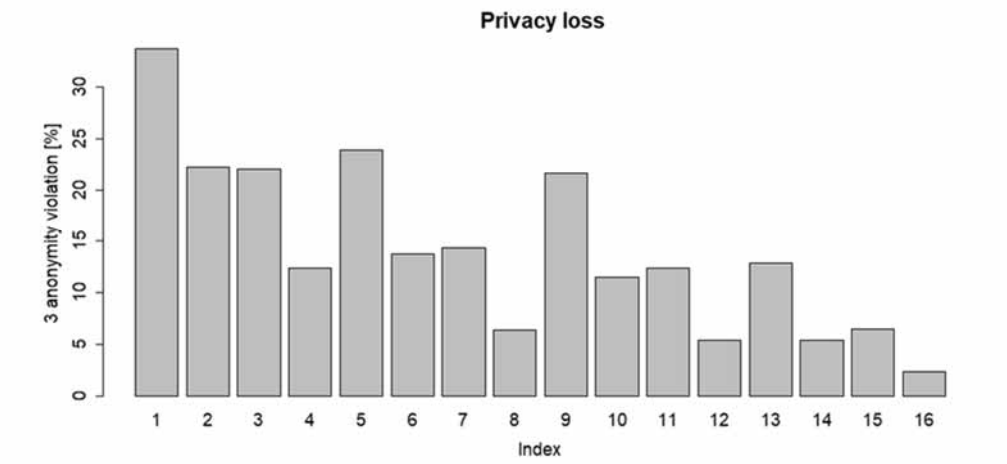


図 6 質的属性の秘匿性評価(3-匿名性違反のレコード数の割合)

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で除することで情報量損失率を算出できる。図 7 は、情報エントロピーに基づく情報量損失率を示している。図 7 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。

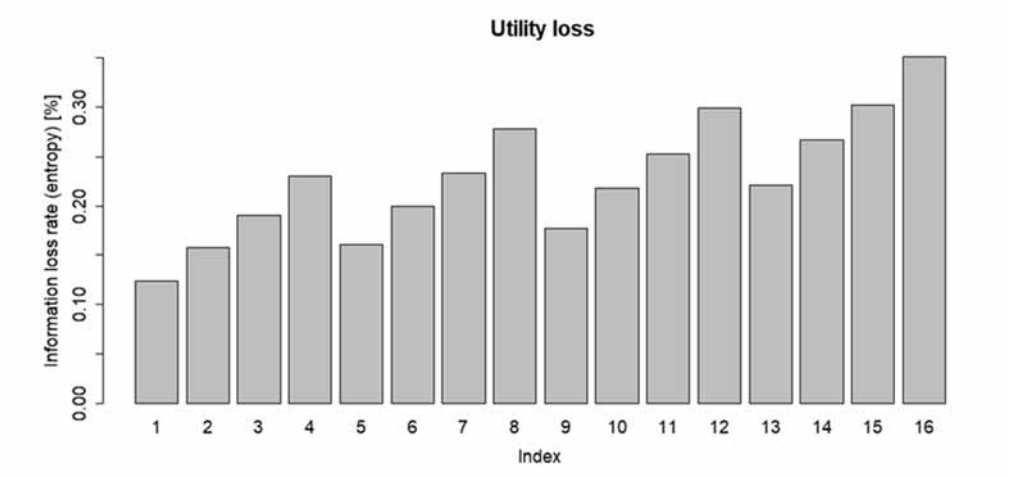


図 7 質的属性の有用性評価(情報エントロピーに基づく情報量損失率)

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 and Pearson(1991))を作成した。横軸が Risk(秘匿性)を、縦軸が Utility(有用性)を表している。具体的には、秘匿性には総レコード数に占める 3-匿名性違反のレコード数の割合を、有用性には情報エントロピーに基づく情報量損失率を用いた。図 8 から、秘匿性が増大するほど有用性が相対的に低下するトレードオフの関係にあることがわかる。例えば、最も細かい分類区分の組み合わせである index 1 は、図中右下に位置しており、秘匿性は低く、有用性は高い領域に位置している。それに対して、最も粗い分類区分の組み合わせである index 16 は、図中左上に位置しており、秘匿性は高く、有用性は相対的に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バランスが取れているということができるが、本実験の結果では概ね曲線に沿って位置していることから、特定の index が、秘匿性と有用性の観点からバランスに優れているという結果は得られていない。統計実務の観点から見れば、有用性を考慮しつつ、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。

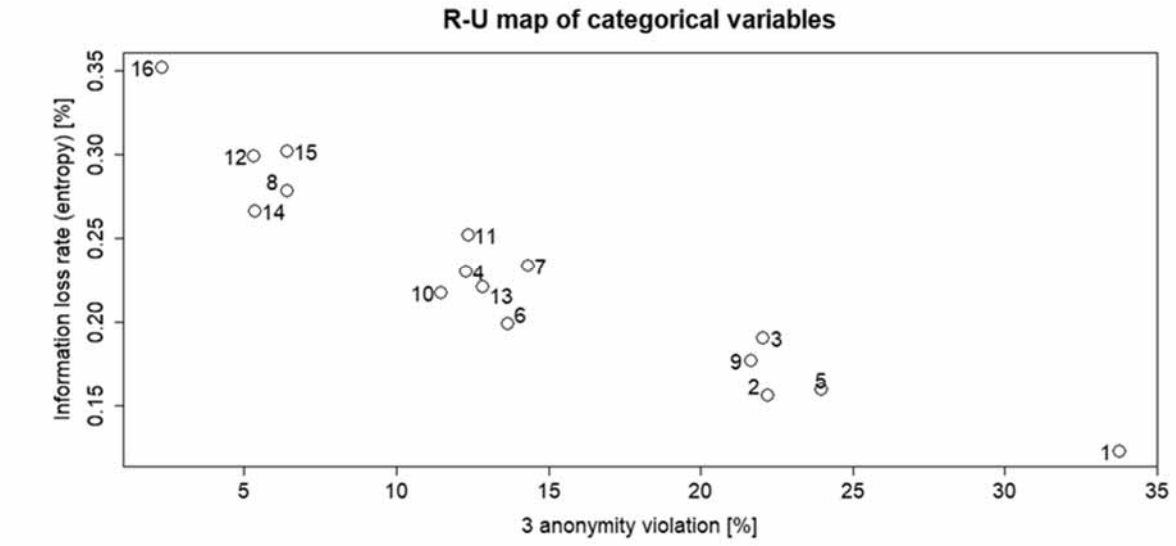


図 8 質的属性の R-U マップ

(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)

3.5　量的属性の匿名化の検討

本節では、量的属性のうち、センシティブな属性である売上(収入)金額、給与総額、減価償却費、付加価値額の匿名化を検討する。匿名化を行うための準備作業として、より細かい分布の確認を行った。図 9 はリコーディング済みの分類区分を用いた層における、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留意する必要がある。

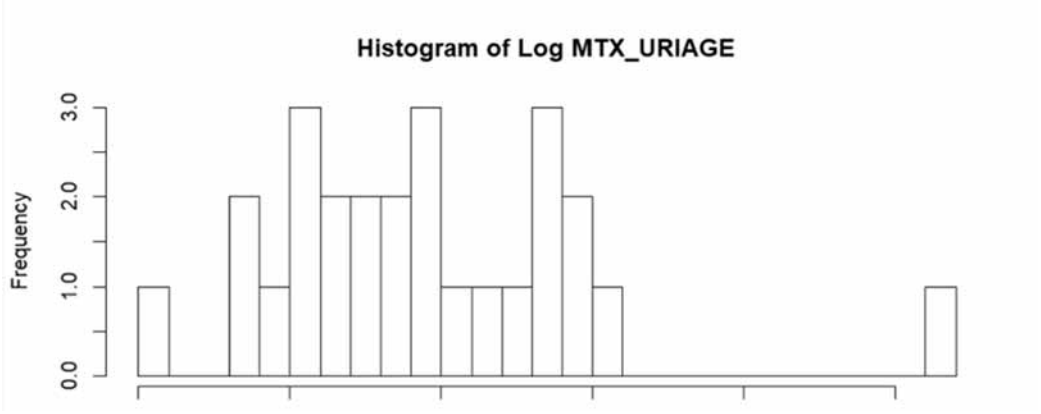


図 9 あるキー変数の組の売上(収入)金額のヒストグラムの例

※秘匿上、横軸（対数）の目盛りは省略

表 10 は、上記とは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実数値の要約統計量を示したものである。上位 5%では平均が 13,912.60 に対し、上位 5%以外では 619.79 と、上位 5%に大きく分布が偏っていることが読み取れる。一般に、平均値よりも分布の歪みの影響を受けにくいとされる中央値においても、上位 5%が 3,358.00 に対して上位 5%以外は 457.50 と一定の歪みが見られる。匿名化にあたって、上位 5%のような露見リスクの大きい事業所を削除する非攪乱的な手法も考えられるが、これらの分布特性を考慮すると、レコード削除によって生じる分布への影響は無視できない。そのため、分布の右裾の事業所については安易にレコード削除を行うのではなく、平均値等の統計量を維持できる攪乱的手法が適切であると考えられる。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 10 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(実数値)

	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、センシティブや量的属性に対する匿名化技法として、イタリアやドイツにおける匿名化ミクロデータ作成の実務においても採用実績のあるミクロアグリゲーションを選択した。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[5]の 2 種類を選択した。

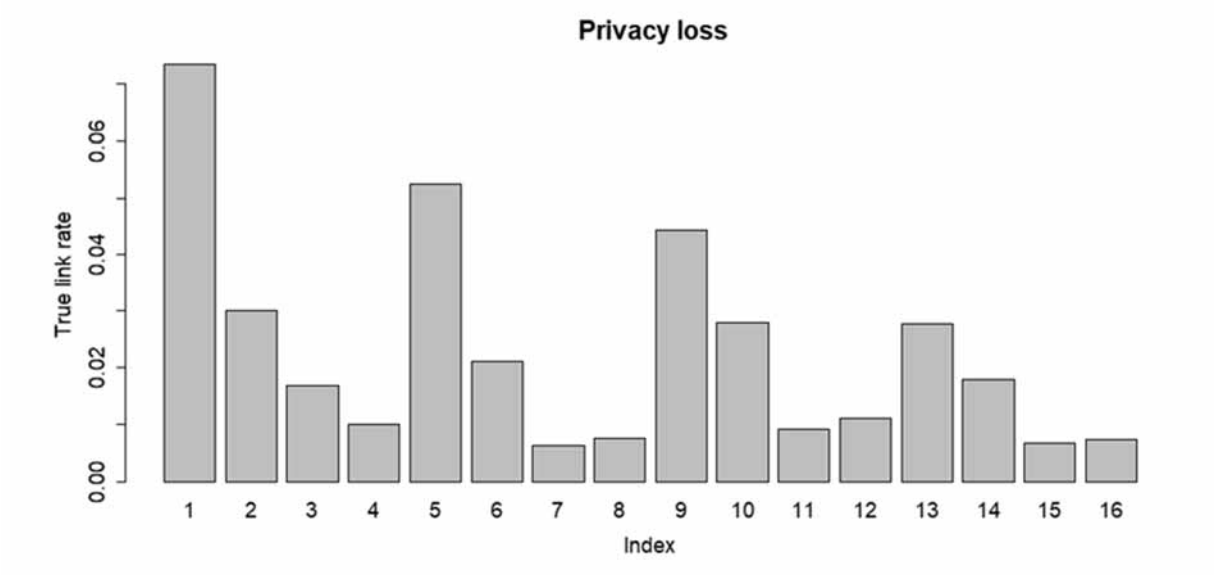
[5] MDAV(maximum distance to average vector)法とは、Domingo-Ferrer and Mateo-Sanz(2002)で述べられた多変量固定サイズのミクロアグリゲーションをもとに、Hundepool *et al.*(2003)で実装されたアルゴリズムである(Domingo-Ferrer and Torra(2005))。複数の量的属性の平均ベクトルを求め、探索的にグルーピングを行うヒューリスティックなミクロアグリゲーションの一手法である。

3.6　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、原データと匿名化ミクロデータにおけるレコード同士の距離を計算し、その距離の大きさに基づいて、２つのデータが対応付け可能かを判定する方法である(伊藤(2010))。具体的には、最初に、匿名化ミクロデータのレコードから原データの各レコードへの距離を計測し、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。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、原データから最も距離の近い攪乱済みのレコードが真のリンクである確率(true link rate)を求めた。

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 10 に示す。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。



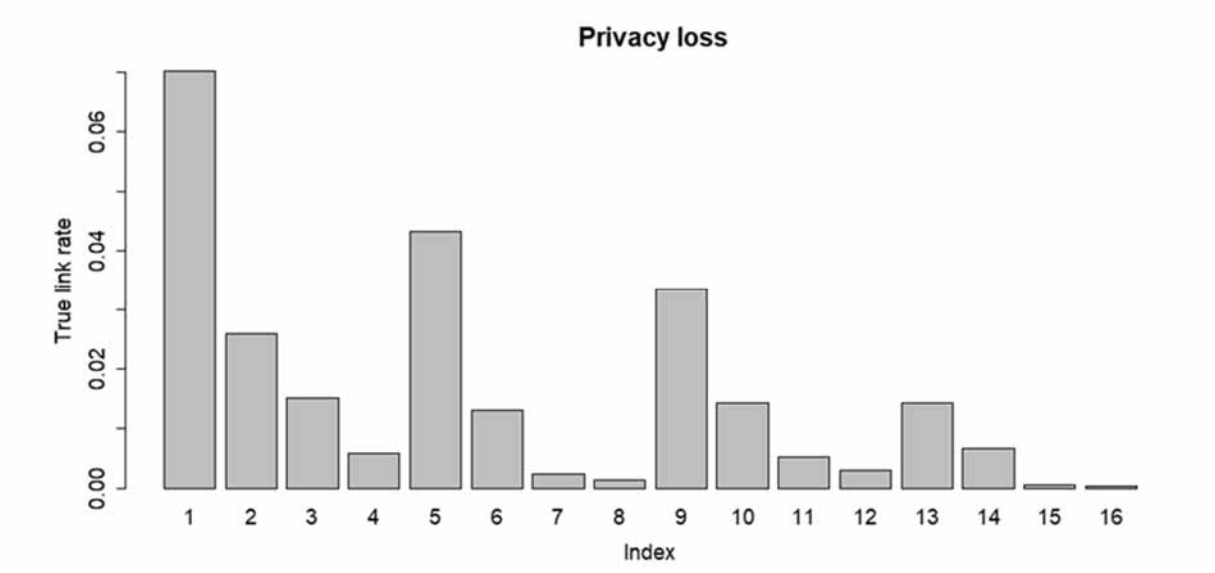


図 10 量的属性の秘匿性評価(距離計測型リンケージに基づく true link rate)
(上：個別ランキング法、下：MDAV 法)

さらに、量的属性の有用性評価を行った。ミクロデータに含まれる量的属性に対して有用性の相対的な程度を評価する手法として、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。原データと匿名化ミクロデータについて、属性値の差、分散共分散行列、相関係数行列に見られるデータ構造の変化によって情報量損失の計測を行った。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 11 に示す(伊藤他(2014)表 1 より)。

表 11 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 12 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、データ特性の観点からより原データに近似する結果となった。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 12 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。

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

IL1s を用いて評価した有用性評価の結果が図 11 である。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より小さくなる傾向にある。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。

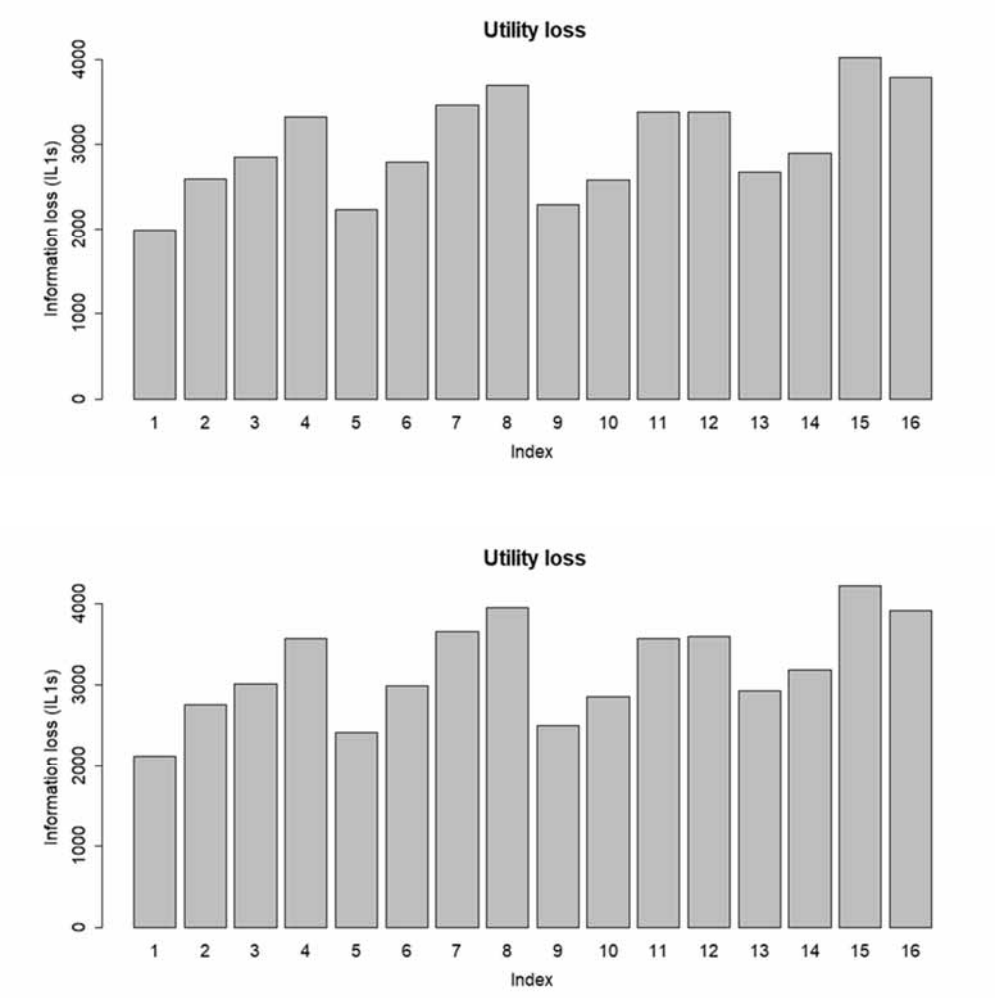


図 11 量的属性の有用性評価(IL1s)
(上：個別ランキング法、下：MDAV 法)

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 12 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、最も粗い分類区分の組み合わせである index 15 や 16 は、図中左上の秘匿性は高く、有用性は低い位置に存在している。図中で左下の領域にある index ほど秘匿性と有用性のバランスが図られているが、本実験の結果では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。

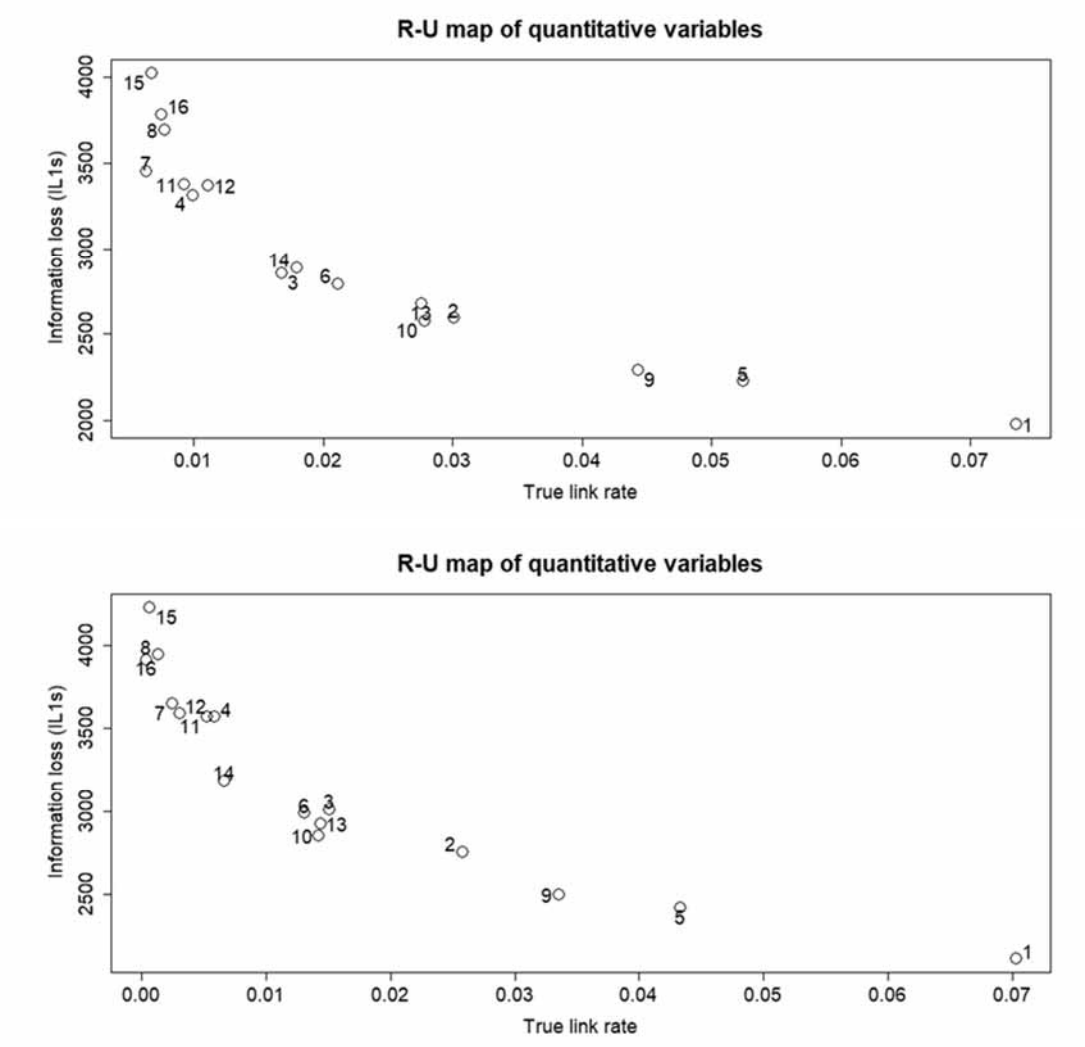


図 12 量的属性の総合評価(R-U マップ)

（上：個別ランキング法、下：MDAV 法）

4　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を対象に、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さらに追究するために探索的な実験を行う。

4.1　経済センサスにおける事業所の分布特性

匿名化ミクロデータの作成を行う上で、露見リスクを最小限に抑えることは不可欠である。露見には、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。

さらに、前節ではリコーディングを行うにあたって、ミクロデータを用いた実証研究を行う場合の有用性を考慮して、リコーディングの区分を設定した。分類区分ごとの構成比を一定以上に高めれば露見リスクを相対的に小さくできると考えられるが、一方で、一部の属性についてはリコーディング幅が粗くなるために、より大きな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いために、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両面について十分に考慮された匿名化を検討することができると考えられる。そこで、本研究では、異なる属性を基準にして事業所ごとの秘匿性の強度としての「露見リスク」[6]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。

4.2　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

---

[6] 本実験における「露見リスク」とは、経済センサスのミクロデータに基づくテストデータを用いて算出された秘匿性の相対的な程度に関する指標として定義されている。したがって、本実験から得られた露見リスクの結果は、母集団に含まれる事業所を特定化するリスクを表していないことについて留意されたい。

[7] 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。

10 万レコードを無作為抽出してテストデータを作成した。本実験では従業者合計の条件は考慮されていない。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 13)。

表 14 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を示している。なお、8 属性から 2 属性ずつ選択されるため、選択される属性の組については 28 のパターンが存在する。表 14 を見ると、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×化学工業の組み合わせにしたがって様々な階層が設定され、各層に含まれる事業所数が 10 未満となるような事業所の数を集計した結果、合計で 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、外部参照情報との照合におけるリンクキーとしての精度と関連すると考えられる。そのため、本実験では他の属性と分類区分の構成比を揃えるような補正は行われていない。

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 属性の組み合わせのよって、リスクの判定に用いられるパターン数は 28 である。事業所レベルで見た場合、露見リスクが相対的に高い事業所は、その中の複数のパターンに該当すると考えられる。この該当するパターンの数を総計した上で、それを「リスク度」としてランク付けすることで、複数の準識別子を考慮して、露見リスクが相対的に高いレコードを探索的に発見することができる(図 13)。

表 13 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%

表 13 続き

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 14 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

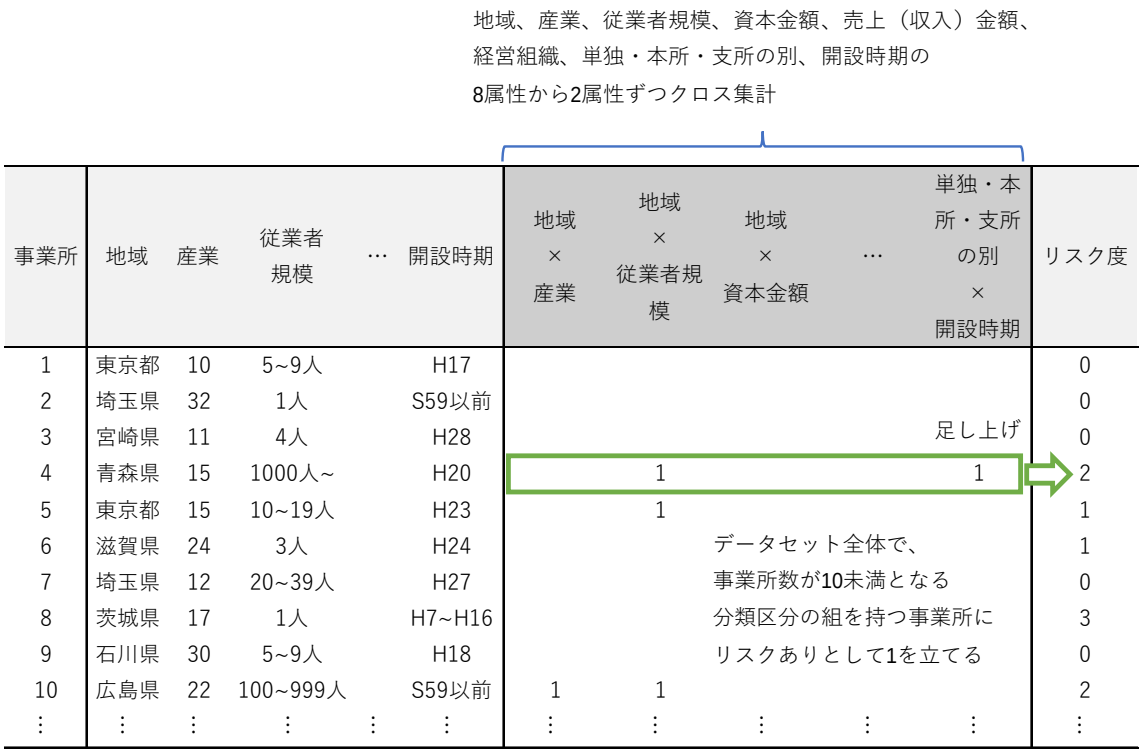


図 13 リスク度評価のイメージ

表 15 は、事業所ごとに何回リスク大と判断されたかを「リスク度」として計測し、層別に量的属性の要約統計量をまとめたものである。本実験条件では、リスク度は 0 から 7 まで設定されているが、秘匿の観点からリスク度 6 とリスク度 7 の詳細については本稿には載せていない。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のリスクを持つ可能性が推察される。

表 15 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、図 14 は、リスク度 1 以上を「高リスク事業所」とリスク度 0 を「低リスク事業所」とそれぞれ設定し、高リスク事業所と低リスク異業所で層化を行った上で、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ものである。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。それに対して、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、相対的な露見リスクは上昇しにくいことを意味している。一方で、沖縄県では、数値が 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ことが確認できる。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14-1 地域（都道府県）

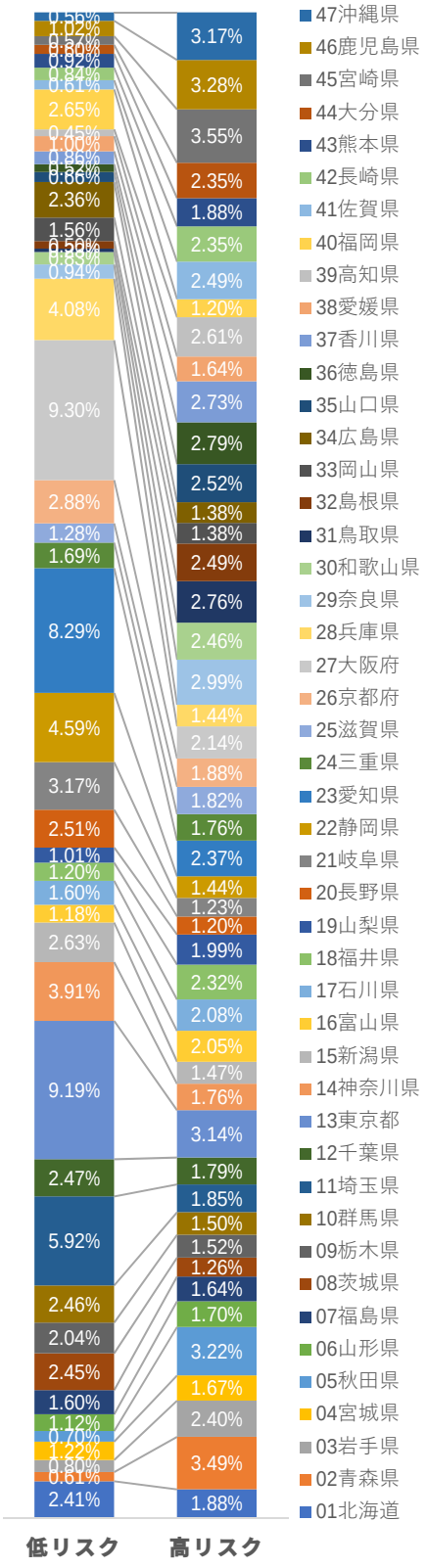


図14-2 産業（中分類）

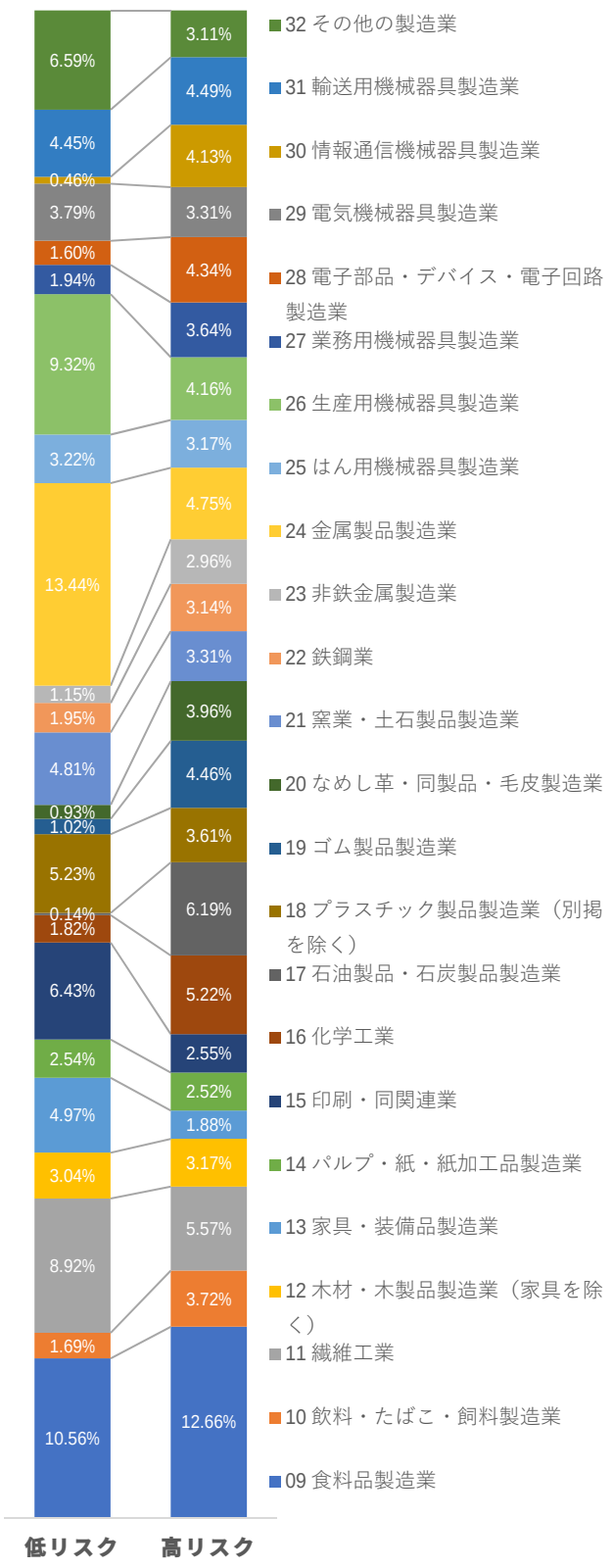


図14-3 従業者規模

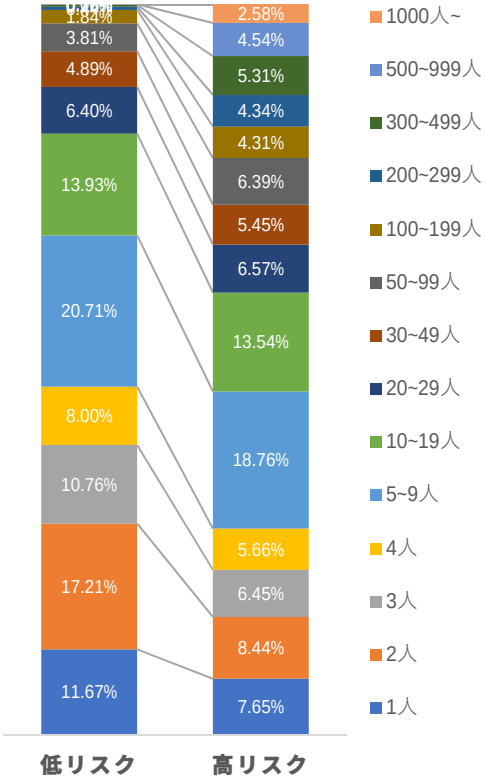
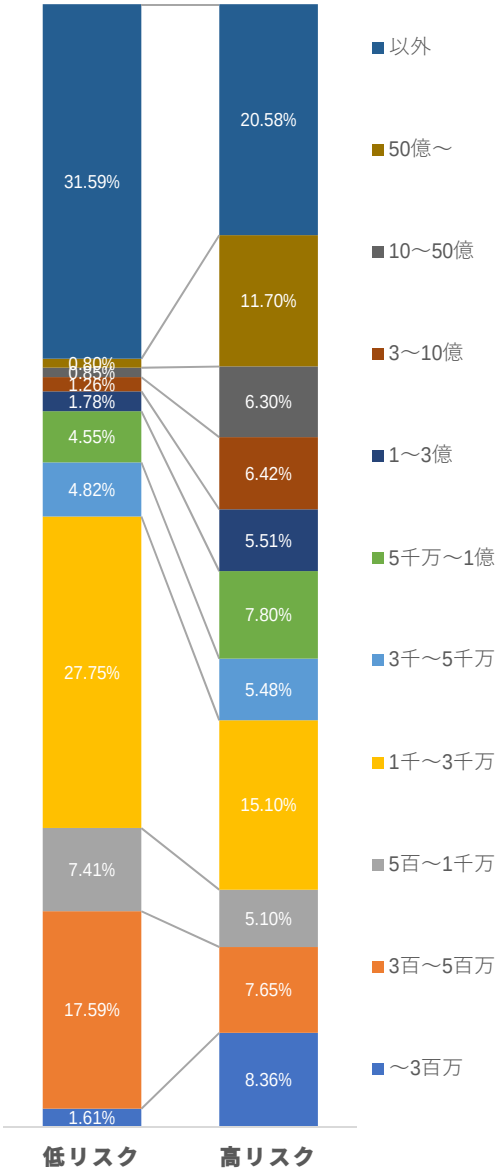
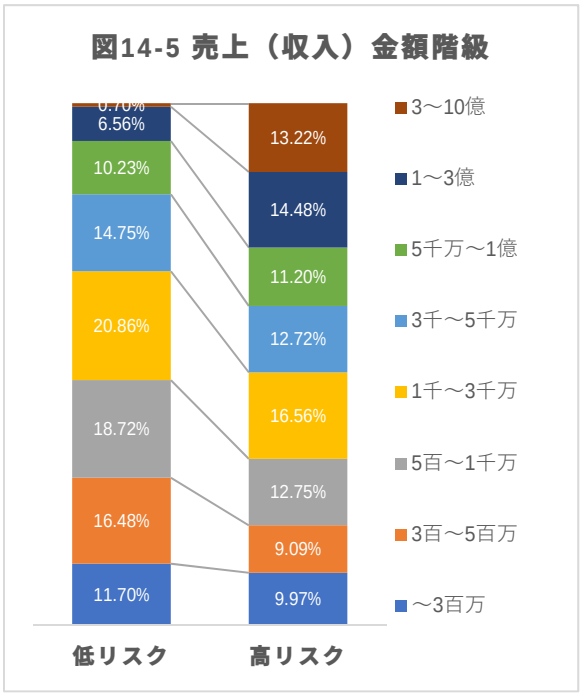
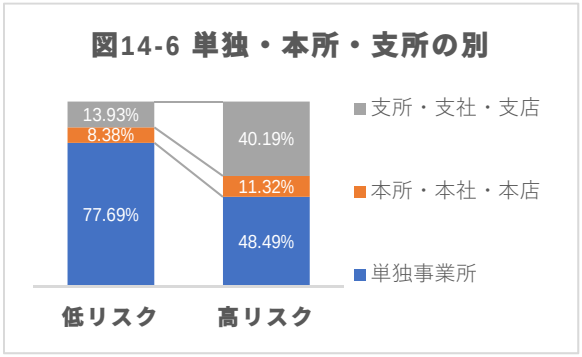
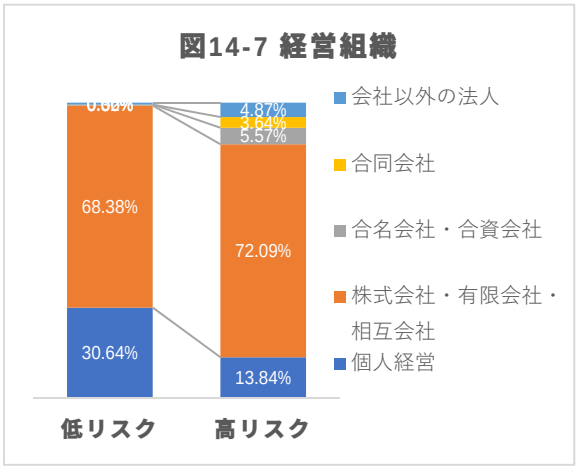


図14-4 資本金階級









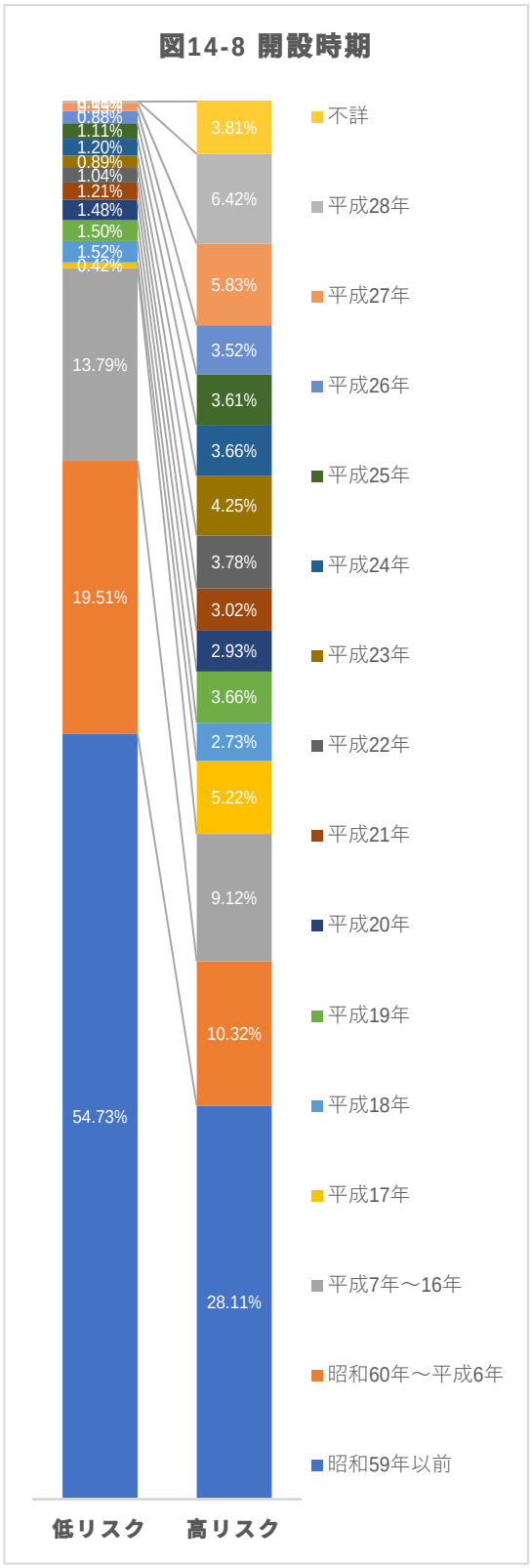


図 14 高リスク事業所と低リスク事業所の分類区分の構成比の比較

最後に、本研究では、2 属性ごとにクロス集計を行い、それぞれの分類区分別に含まれる事業所と、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 15)。バブルの大きさは事業所数を示しており、バブルの色は、高リスク事業所の割合が小さいほど白く、高いほど黒くなるように表されている。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場合、バブルの大きさは 0 になり、また比率も計算できない。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、リスクは相対的に大きいものと判断する必要がある。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。本実験によれば、こうした分類区分の組み合わせは、優先的に匿名化の対象になると思われる。地域×産業の一覧を確認すると、地域については西日本、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)といった特定の産業を対象にした場合、全般的に事業所が少なく、高リスク事業所が多いことが読み取れる。このように、本分析結果から、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相対的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相対的に大きいことが実証的に確認できた。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より大きくなり、産業も一部の中分類についても、リスクが高くなっている。地域については、西日本のリスクは相対的に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では、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、その逆も当てはまることが確認できた。さらに、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っている。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。統計実務の観点からどのような属性が外部参照情報との準識別子になるかを検討しなければ、リスク度に関する定量的な評価は困難

である。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、事業所・企業系の匿名化ミクロデータの作成を指向する場合、キー変数の決定やリコーディングの分類区分の定め方、量的変数の攪乱の際の層化の基準など、様々な点に応用できるのではないかと考えている。

図 15 分類区分別の事業所数と高リスク事業所数割合

図 15-1 地域×産業

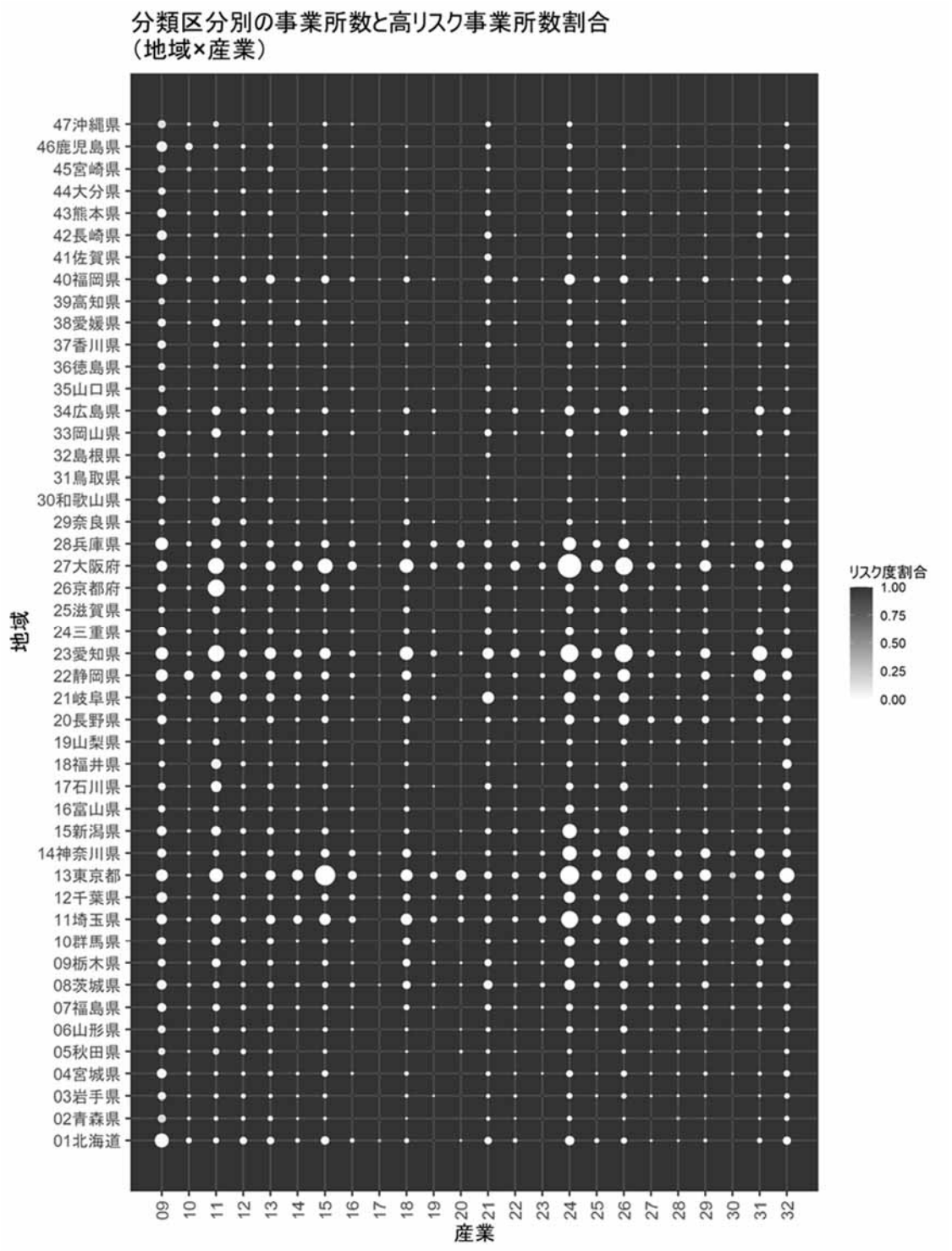


図 15-2 地域×従業者規模

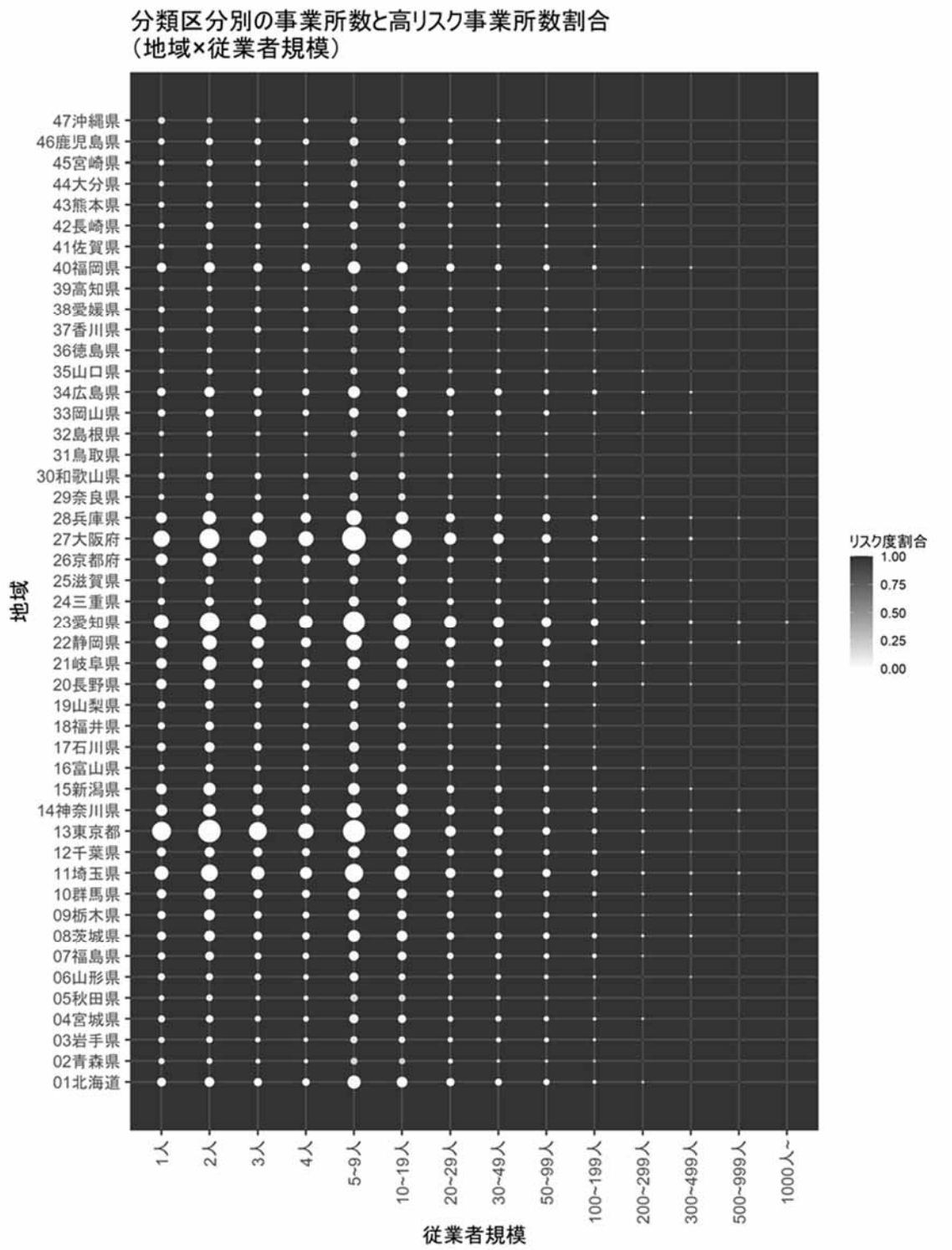


図 15-3 地域×資本金階級

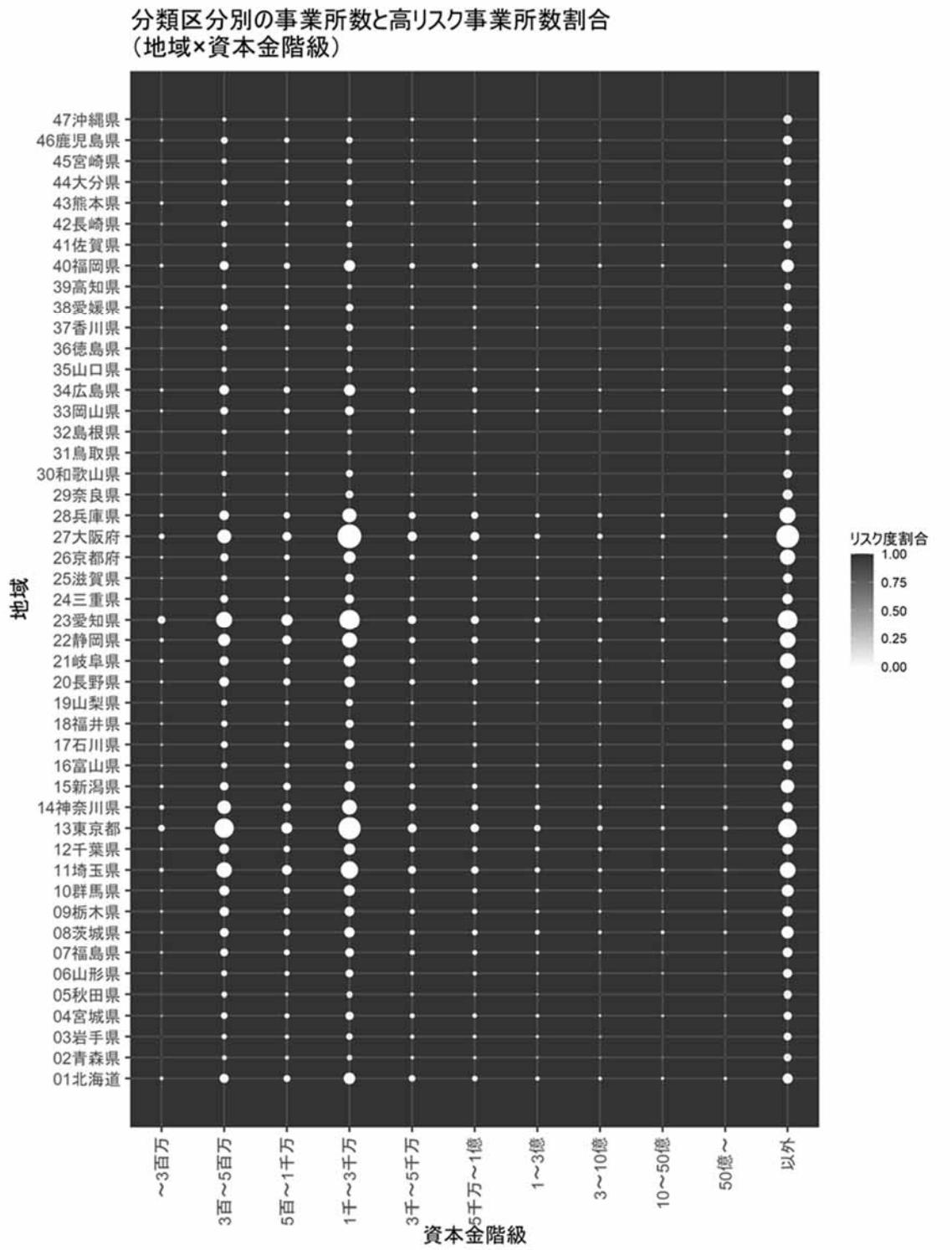


図 15-4 地域×売上（収入）金額階級

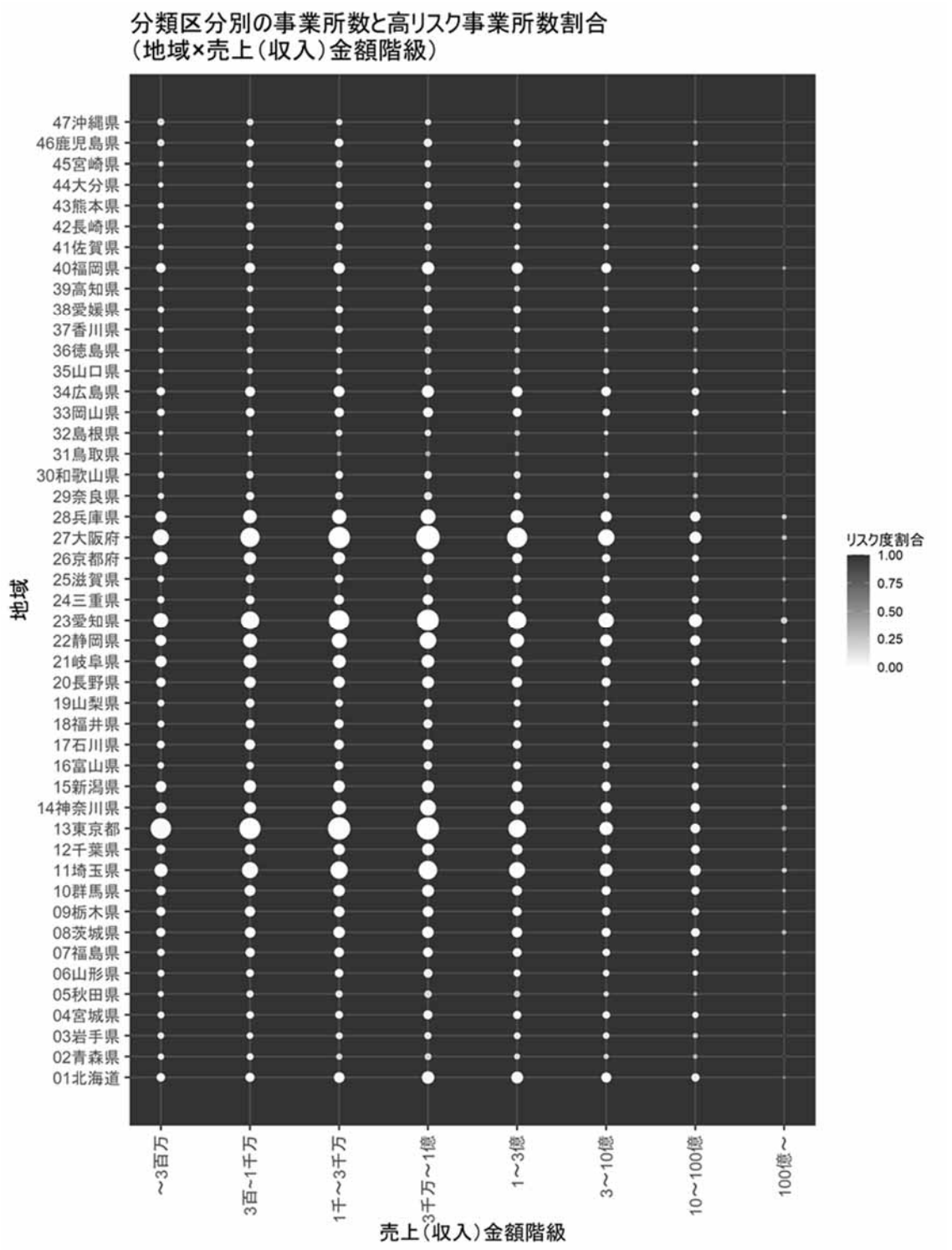


図 15-5 産業×従業者規模

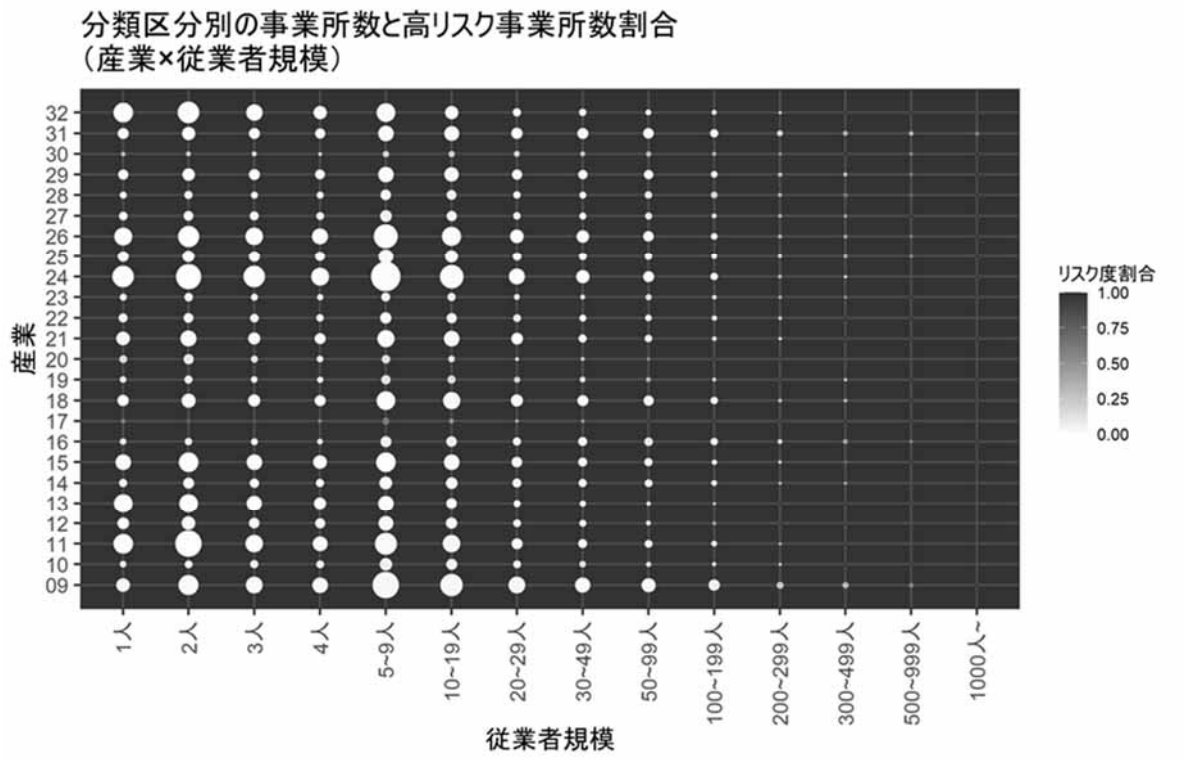


図 15-6 産業×資本金階級

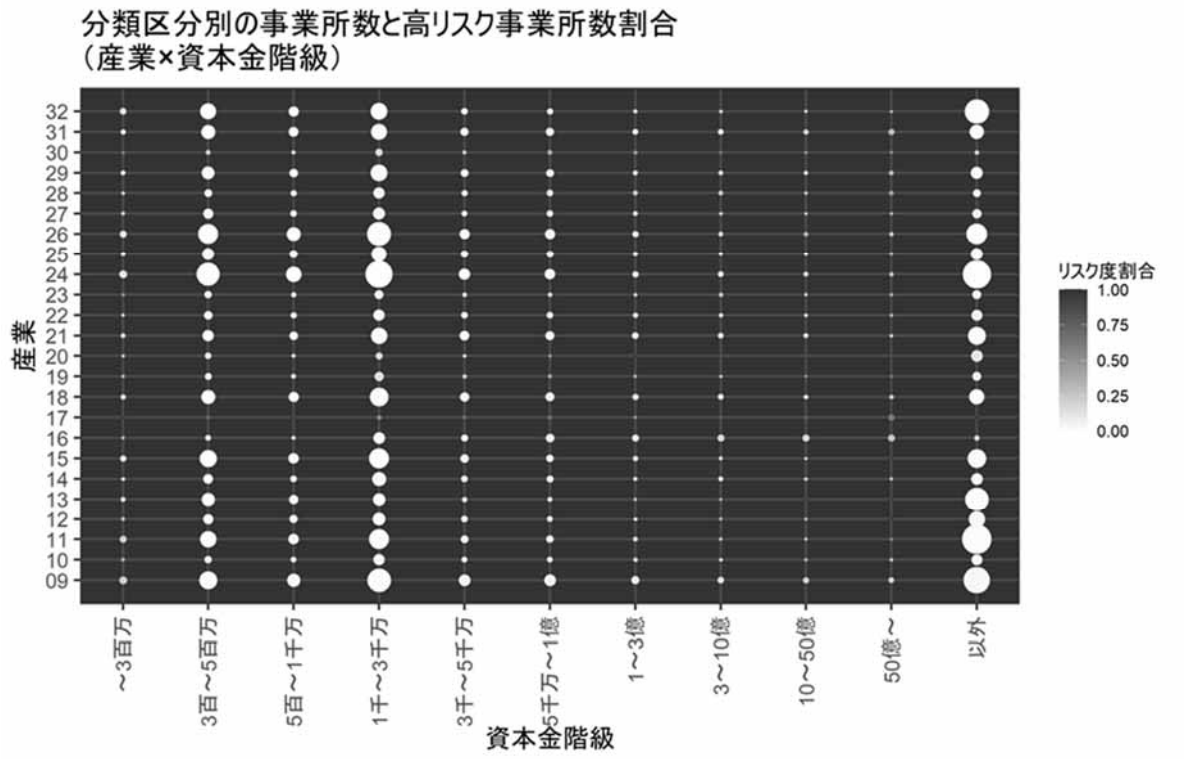


図 15-7 産業×売上（収入）金額階級

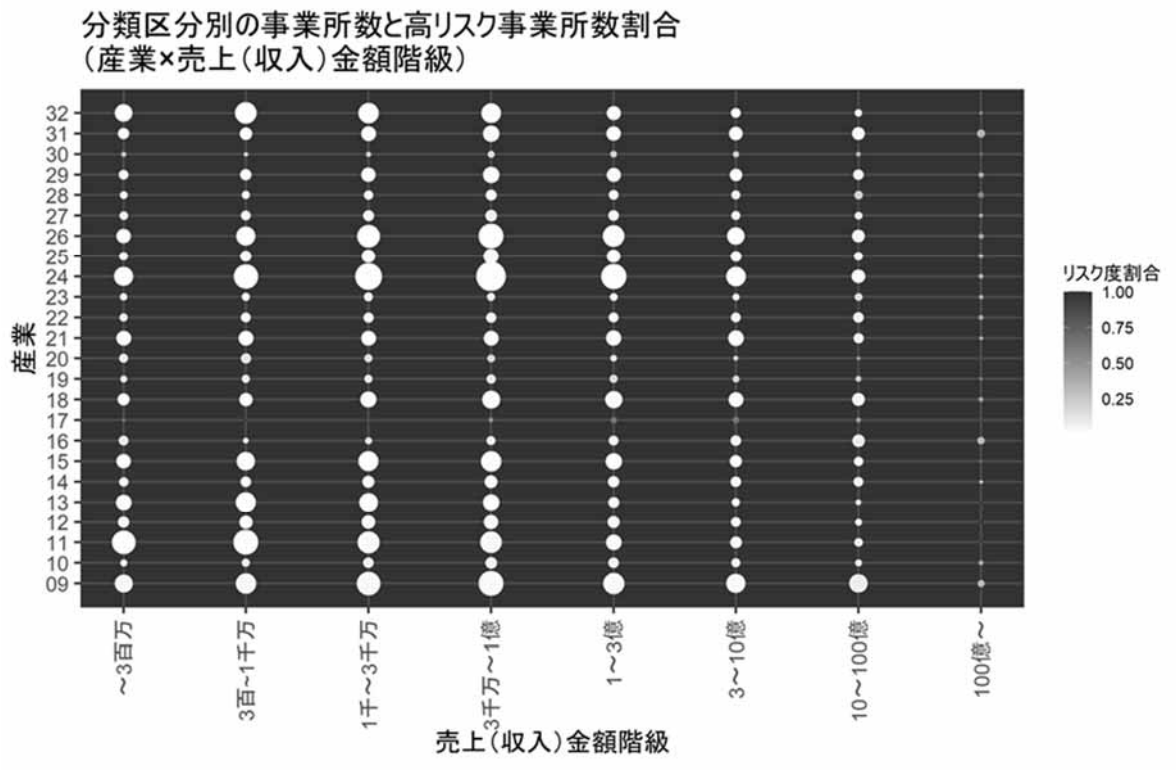


図 15-8 従業者規模×資本金階級

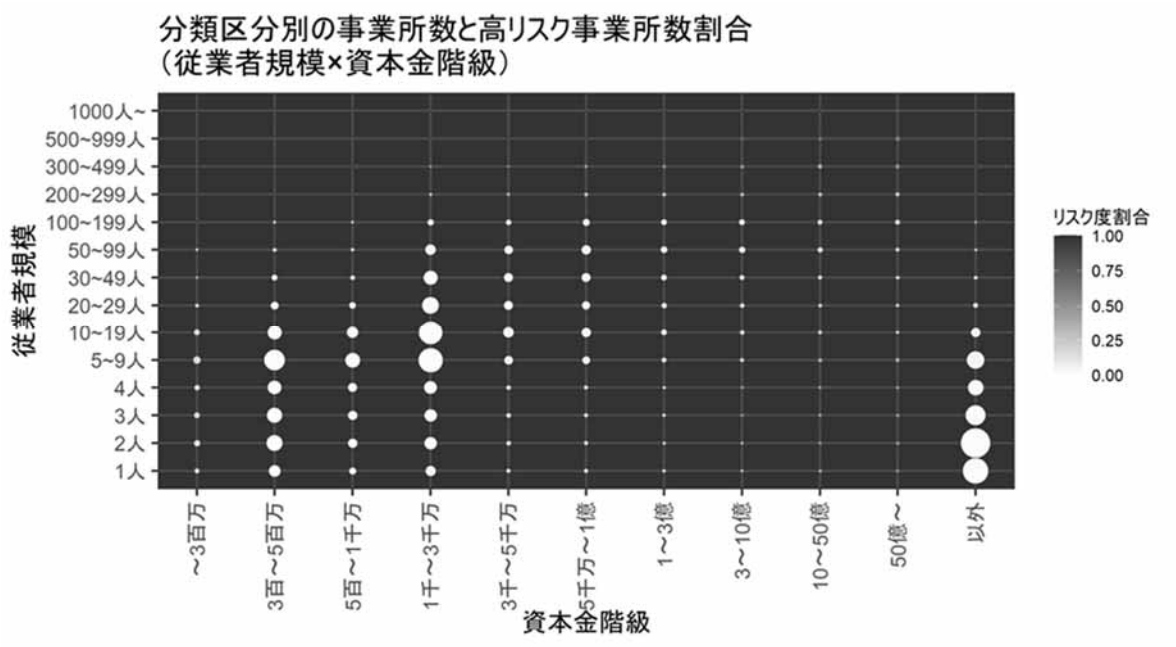


図 15-9 従業者規模×売上（収入）金額階級

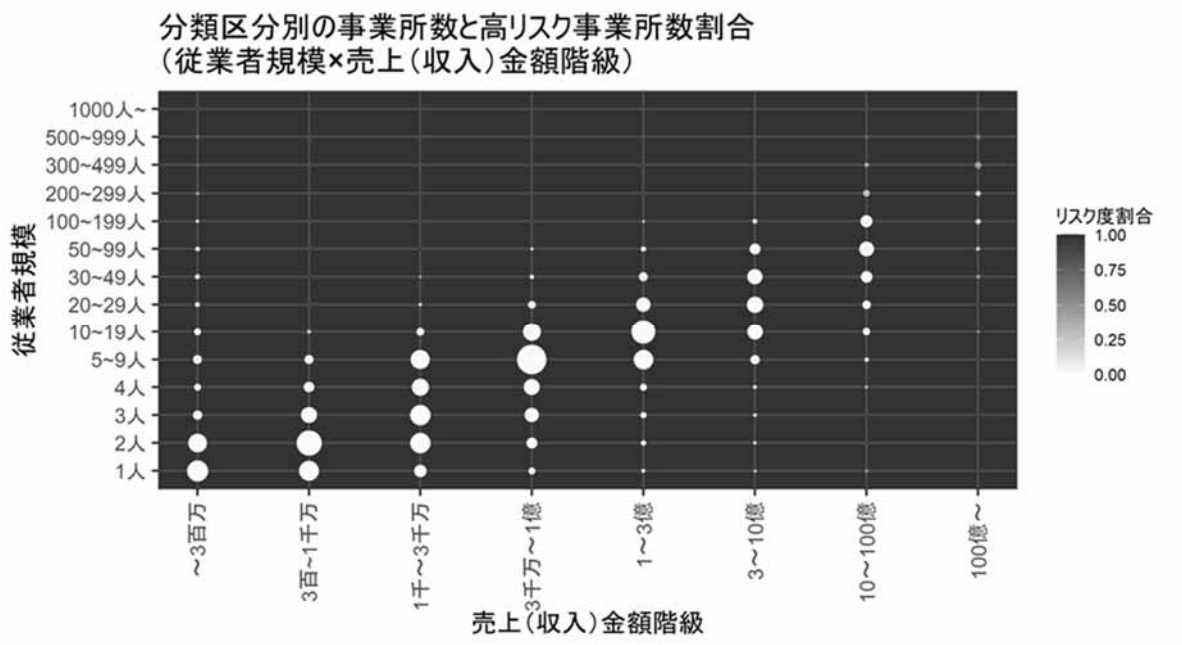
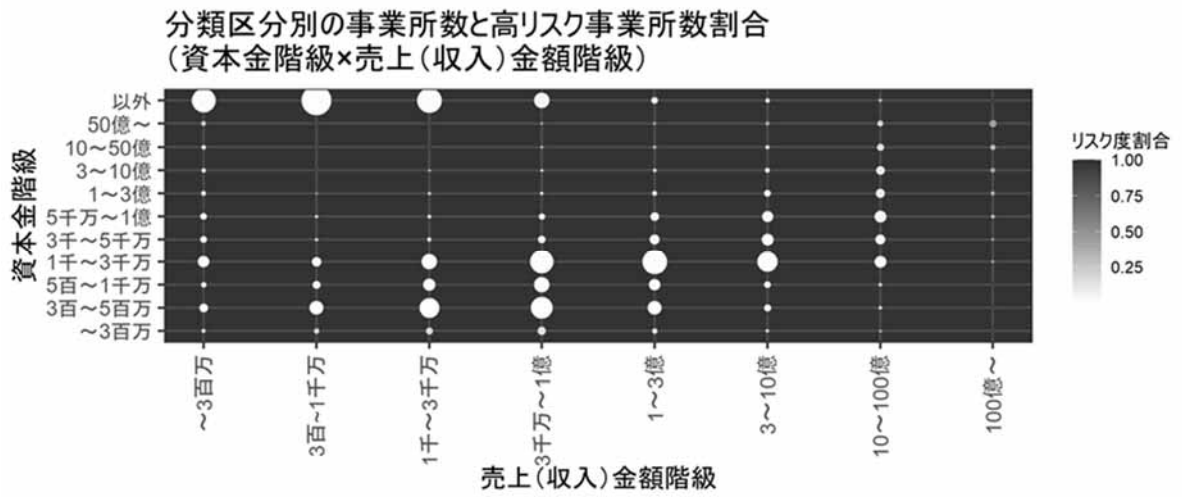


図 15-10 資本金階級×売上（収入）金額階級



5　むすびにかえて

本稿では、イタリアやドイツの事例を中心に、海外における事業所・企業系の匿名化ミクロデータの作成の現状、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述べた上で、経済センサスの個票データをもとに、各種の匿名化手法を適用して作成し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を行った。

本稿で明らかになったように、事業所・企業系のデータについては、把握される量的変数の分布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っており、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。さらに、企業に関する財務情報など、外部に開示される企業情報も存在することから、侵入者は精度の高い外部情報を容易に取得できる場合がある。このことから、事業所・企業系のデータの露見に伴うリスクは、個人・世帯の調査における露見リスクより大きいことが知られている。

そこで、本稿では、イタリアやドイツにおける匿名化ミクロデータの作成事例を考察することによって、以下の 3 点を明らかにした。第 1 に、露見シナリオについては、基本的には、学術研究用ファイルの作成を指向することを前提に、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。また、匿名化ミクロデータの秘匿性については露見シナリオを考慮した定量的な評価基準に基づいて、有用性に関しては実用例のサーベイを基に複数の指標を検討することによって、攪乱的手法の適用を最小限に抑えている。第 2 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等の攪乱的手法も採用されるが、原データとの情報量損失が相対的に低い個別ランキング法の適用可能性が追究されている。第 3 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。

これらのドイツやイタリアの先行事例を踏まえて、本研究においては、特定の露見シナリオを想定した上で、非攪乱的手法だけでなく、攪乱的手法を用いた上で、経済センサスを例に、各種の匿名化手法を適用して試行的に作成し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を行った。具体的には、売上高、資本金、地域、産業といった属性に着目して、探索的なリコーディングを行った上で、分布特性を把握するだけでなく、本研究で用いたテストデータにおいて露見リスクが想定的に高くなると判断されるレコードを発見した。その上で、各種のミクロアグリゲーションを適用して、クロス表による評価方法やリンケージ技法等を用いてミクロアグリゲートデータの有用性と秘匿性の定量的な評価を行い、R-U マップによる各種の匿名化技法による比較・検討を行った。海外では、個別ランキング法の適用の有効性が提案されたが、本稿における実証分析の結果においても、匿名化ミクロデータにおける有用性を重視するのであれば、個別ランキング法が追究できることが分かった。このように経済センサスを用いて攪乱的手法の適用可能性を追究したことは、

本研究の大きな成果と言えよう。

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。

参考文献

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.

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

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

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

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

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

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

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

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

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

Mateo-Sanz J.M., 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.

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

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

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

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.

秋山裕美, 山口幸三, 伊藤伸介, 星野なおみ, 後藤武彦 (2012), 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～, 独立行政法人統計センター『製表技術参考資料 16』.

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

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

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

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

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

東洋経済新報社 (2020 年 6 月), 会社四季報 参照先: 会社四季報: https://shikiho.jp/

日本経済新聞社 (2020 年 6 月), NEEDS-FinancialQUEST, 参照先: NEEDS-FinancialQUEST: 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, pp.69-99.

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

付録 A レコード削除の検討

3.4 節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比率を算出した。本研究では、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、事業所の特定化のリスクを高める。その対策の１つとしては、リスクの高い事業所をデータセットから削除する非攪乱的手法であるレコード削除が考えられる。レコード削除は、わが国や諸外国でも一般に広く用いられている匿名化技法であることから、本研究では、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１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)金額の平均値と資本金額の平均値はそれぞれ 77.5%、98.4%低下している。最も粗い分類区分の組である index 16 においては、レコード削除は約 2.3%と比較的少ないが、売上(収入)金額の平均値については 10.6%、資本金額の平均値に関しては 23.7%の低下が生じており、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 4～表 7 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の作成では一般によく用いられるレコード削除も、事業所・企業系のデータにおいては慎重に取り扱う必要があると言える。具体的には、①削除するレコードの範囲を最小限に絞り込む、②残存するレコードの要約統計量を踏まえて削除対象となるレコードを決定する、③レコード削除後に残ったレコードに補正を行う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(伊藤(2009))、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の割合

図 B-1 地域×経営組織

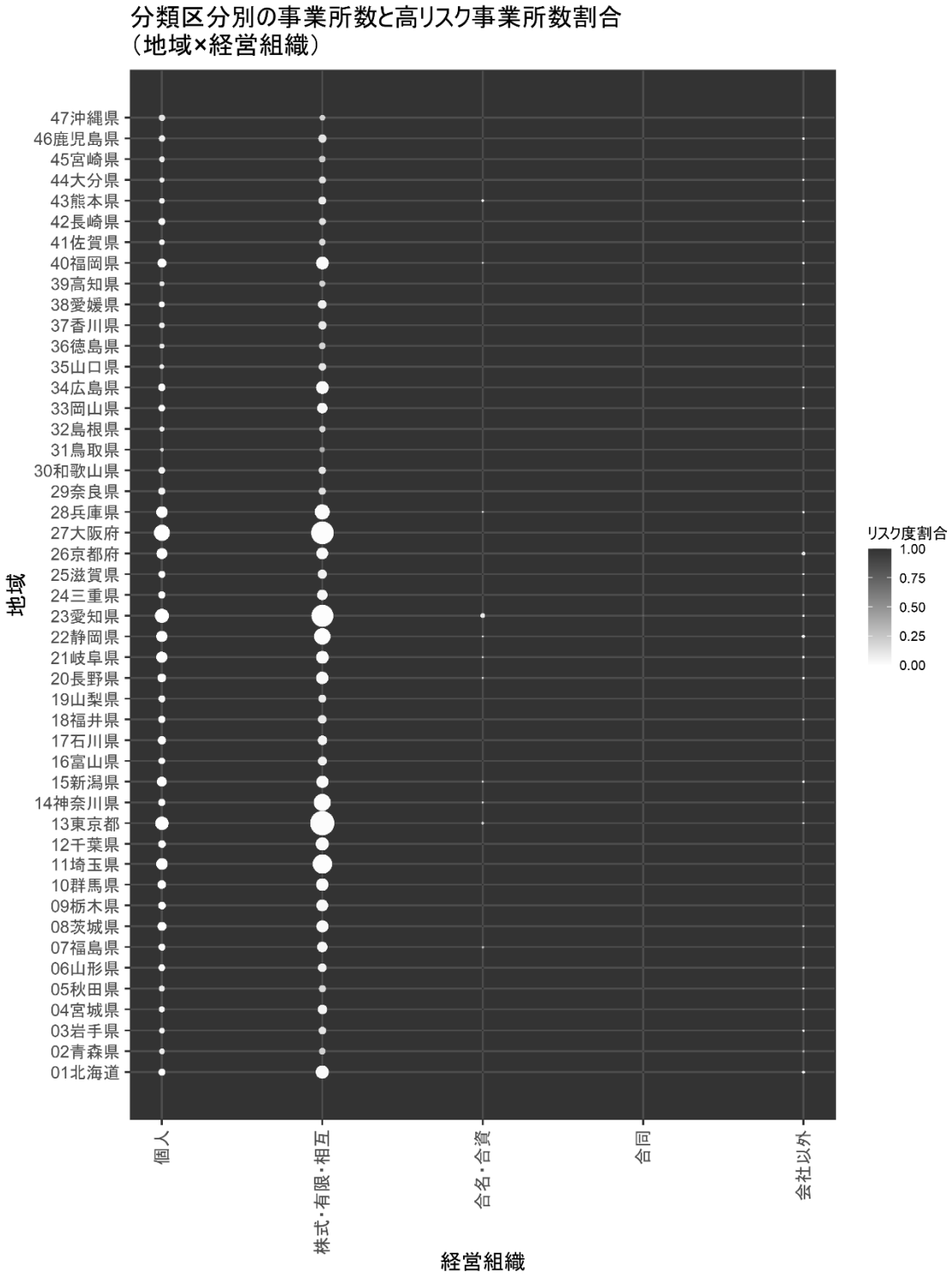


図 B-2 地域×開設時期

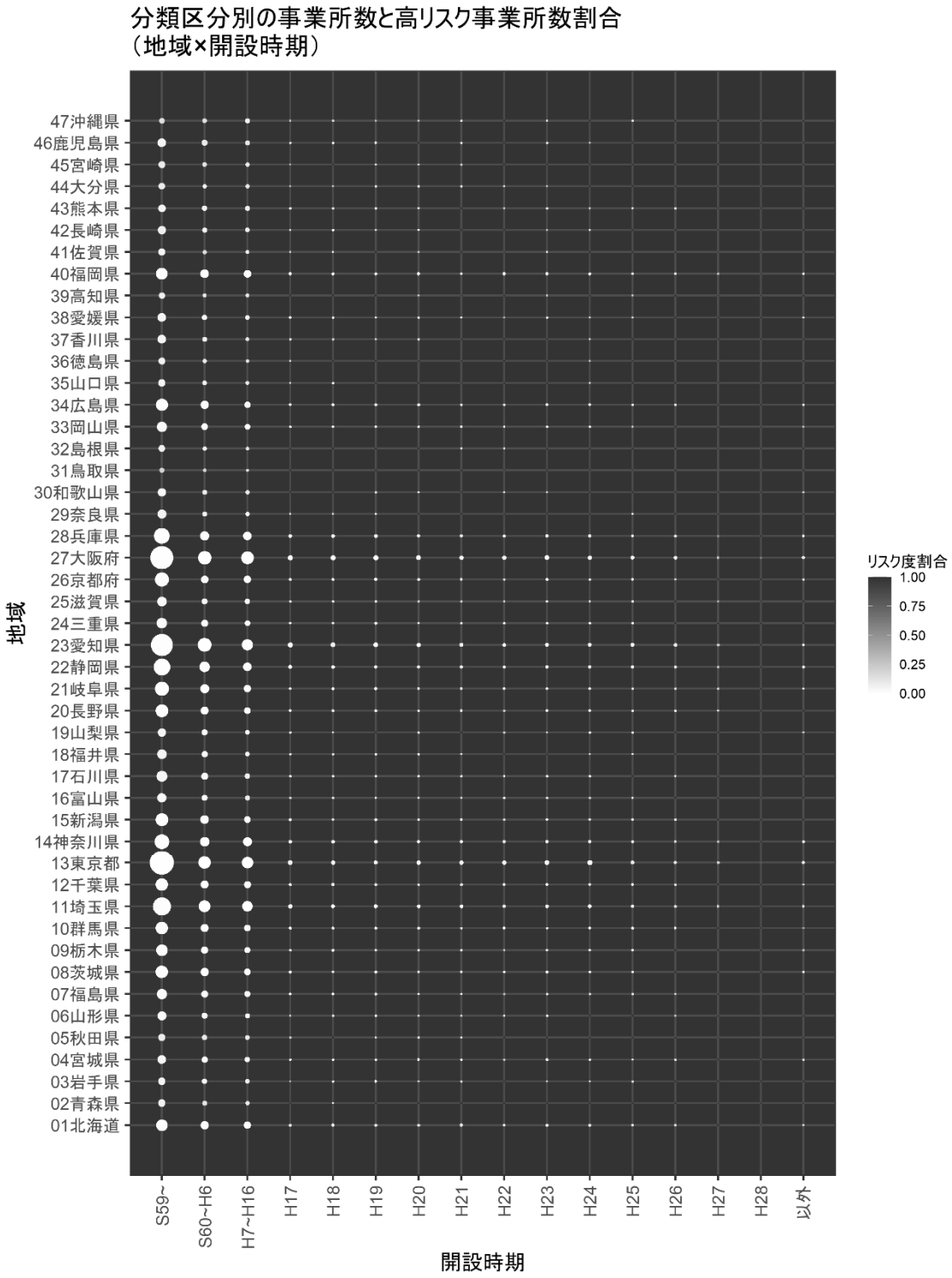

分類区分別の事業所数と高リスク事業所数割合
（地域×開設時期）
地域
47沖縄県
46鹿児島県
45宮崎県
44大分県
43熊本県
42長崎県
41佐賀県
40福岡県
39高知県
38愛媛県
37香川県
36徳島県
35山口県
34広島県
33岡山県
32島根県
31鳥取県
30和歌山県
29奈良県
28兵庫県
27大阪府
26京都府
25滋賀県
24三重県
23愛知県
22静岡県
21岐阜県
20長野県
19山梨県
18福井県
17石川県
16富山県
15新潟県
14神奈川県
13東京都
12千葉県
11埼玉県
10群馬県
09栃木県
08茨城県
07福島県
06山形県
05秋田県
04宮城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

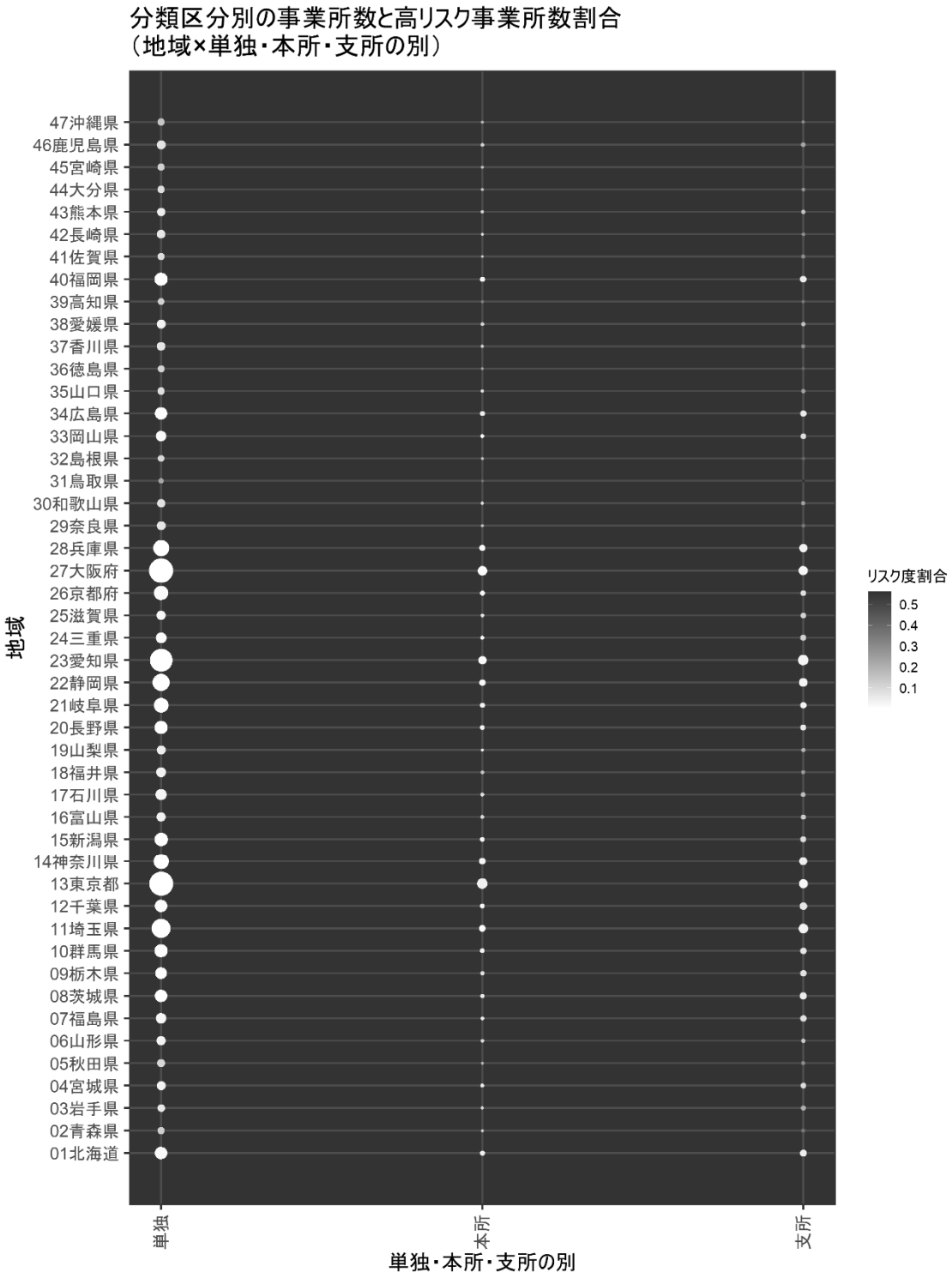


図 B-4 産業×経営組織

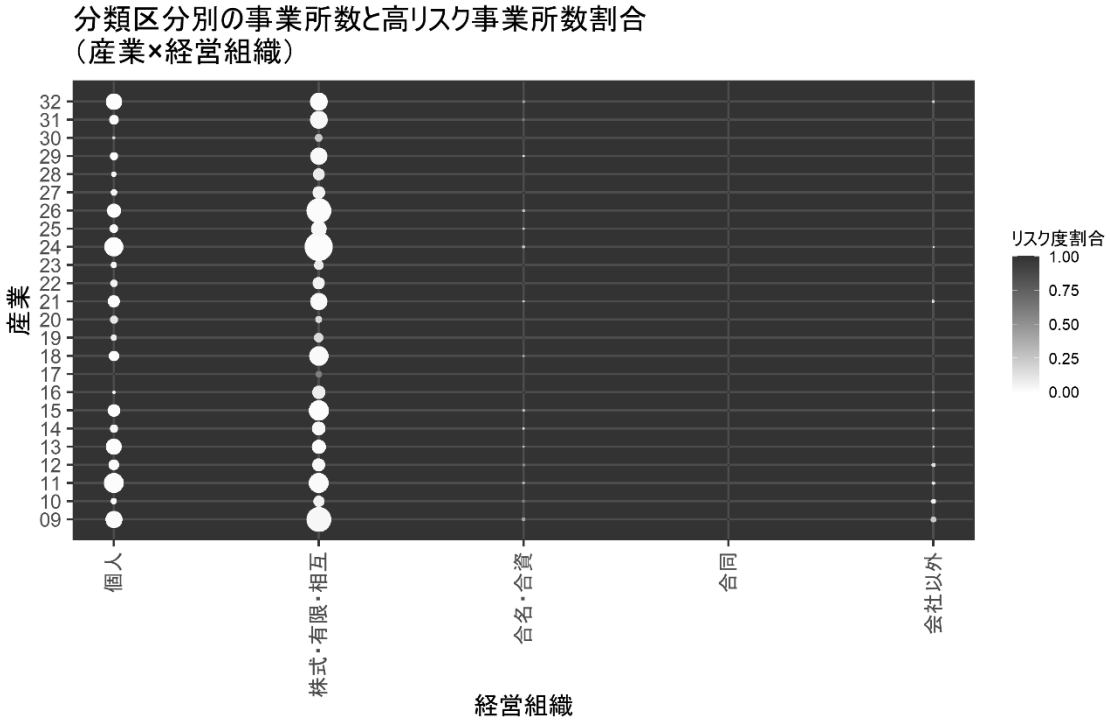


図 B-5 産業×開設時期

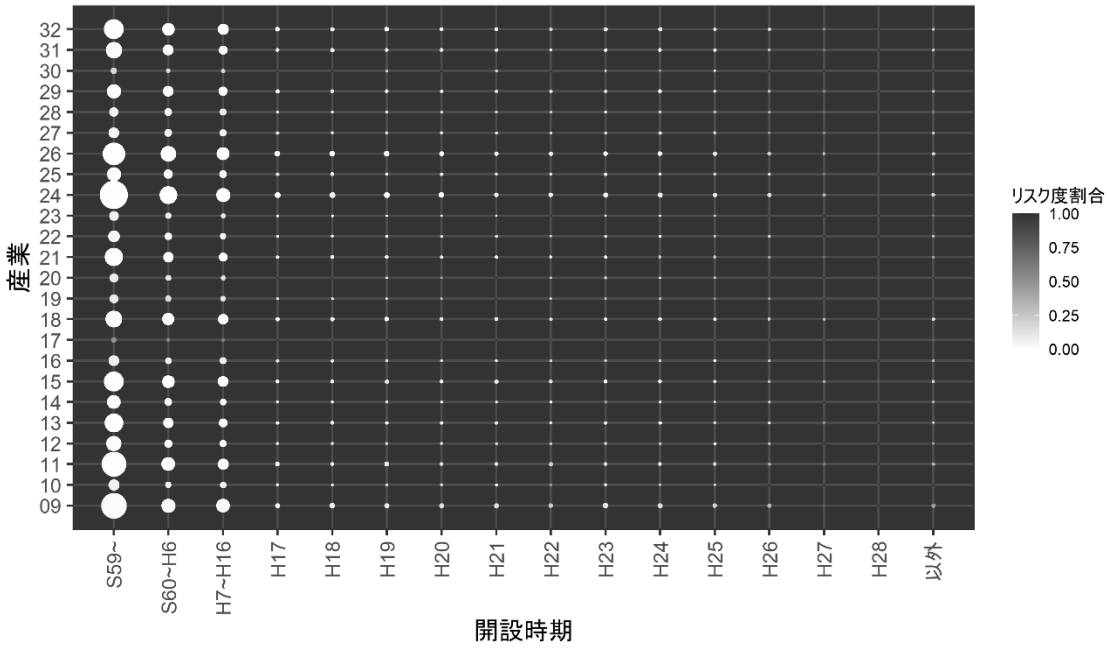


図 B-6 産業×単独・本所・支所の別

図 B-7 従業者規模×経営組織

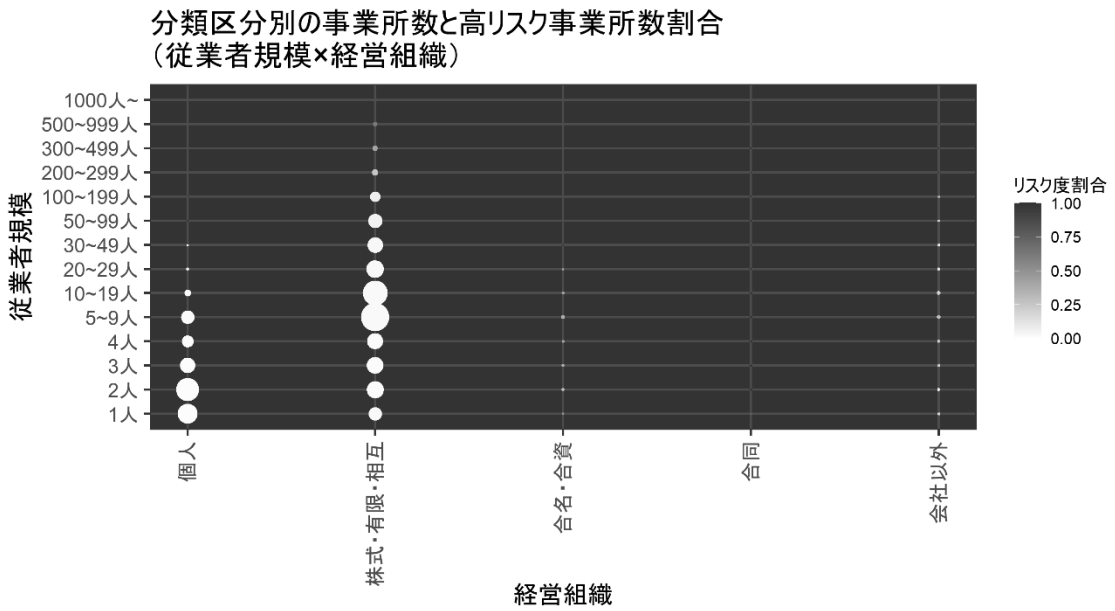


図 B-8 従業者規模×開設時期

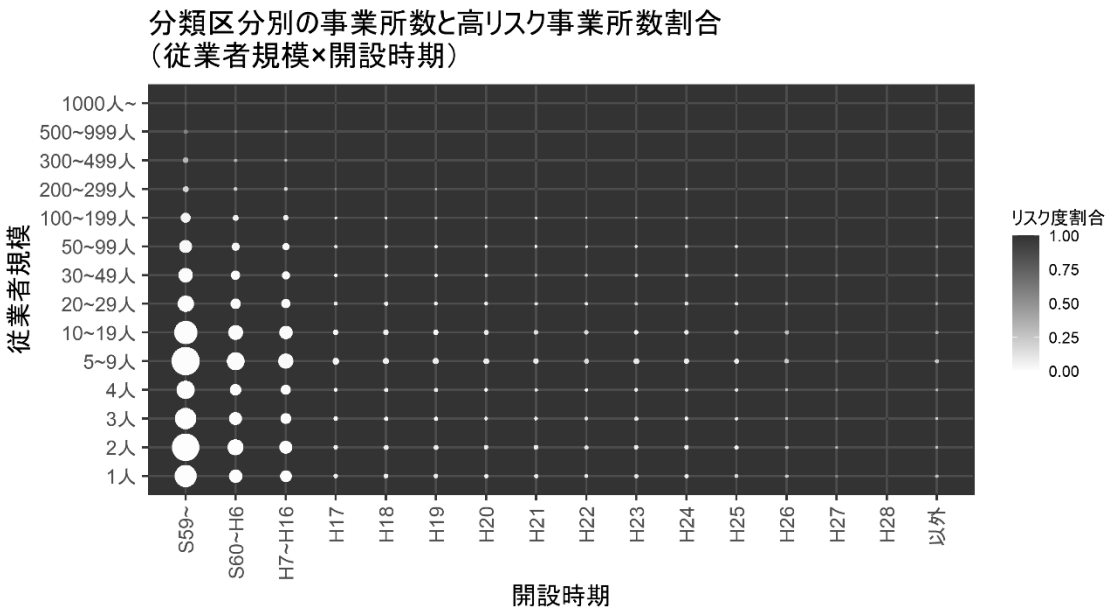


図 B-9 従業者規模×単独・本所・支所の別

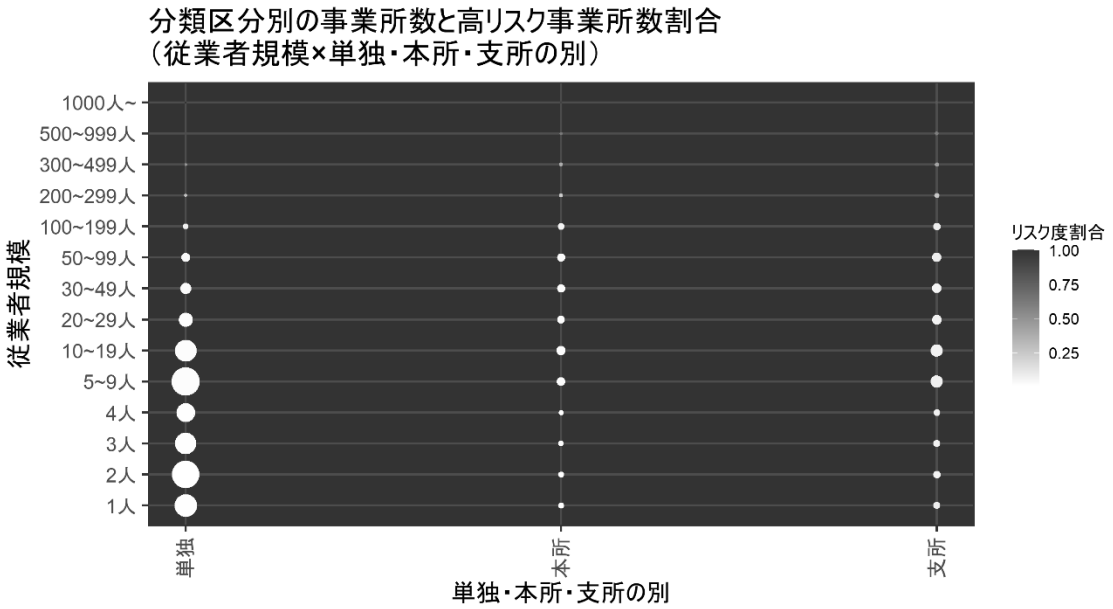


図 B-10 資本金階級×経営組織

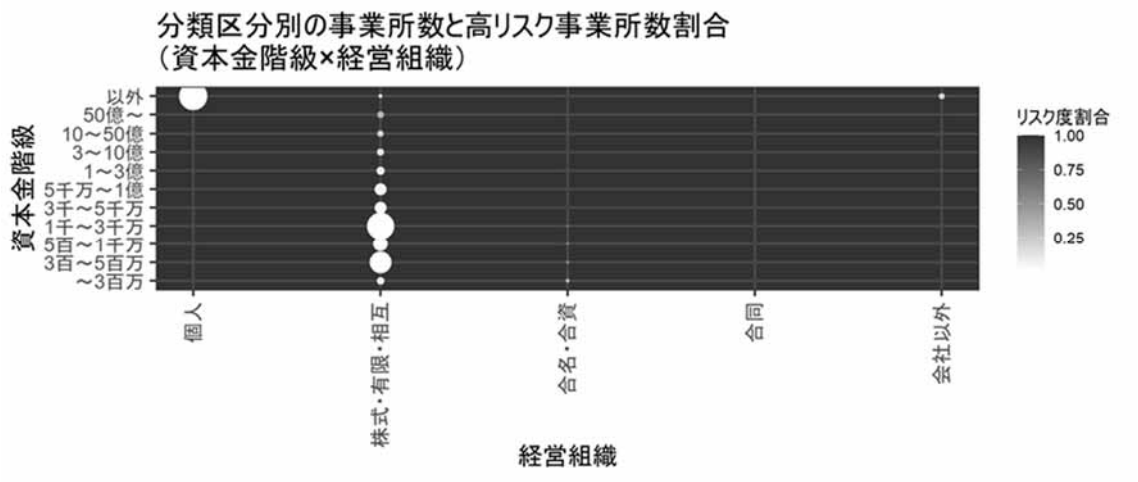


図 B-11 資本金階級×開設時期

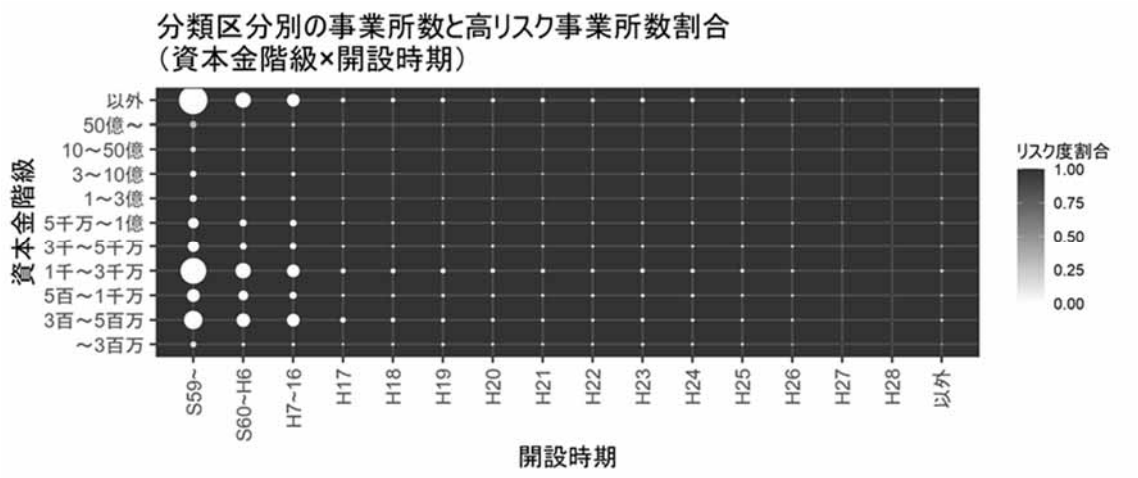


図 B-12 資本金階級×単独・本所・支所の別

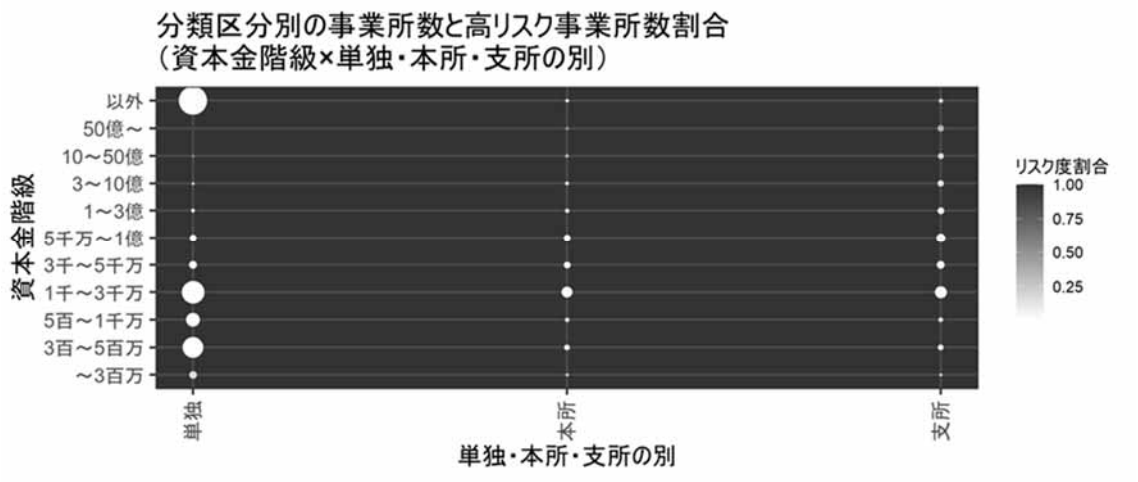


図 B-13 売上（収入）金額階級×経営組織

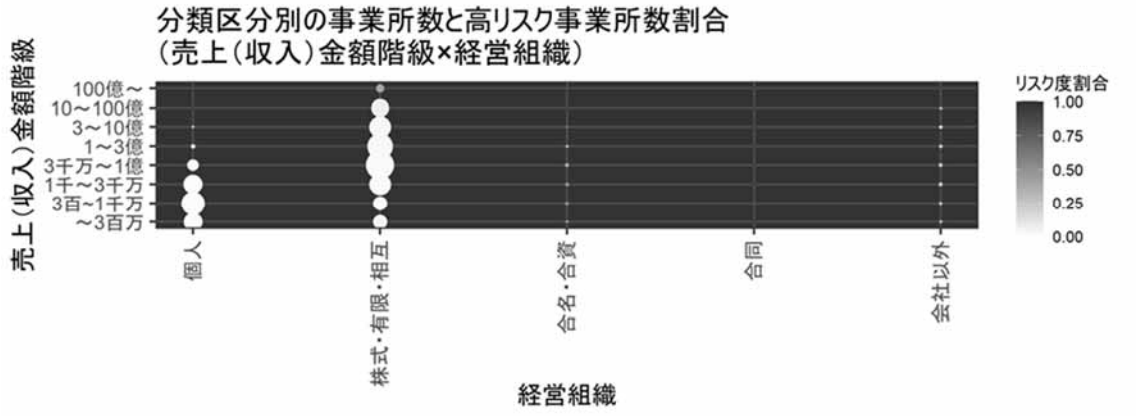


図 B-14 売上（収入）金額階級×単独・本所・支所の別

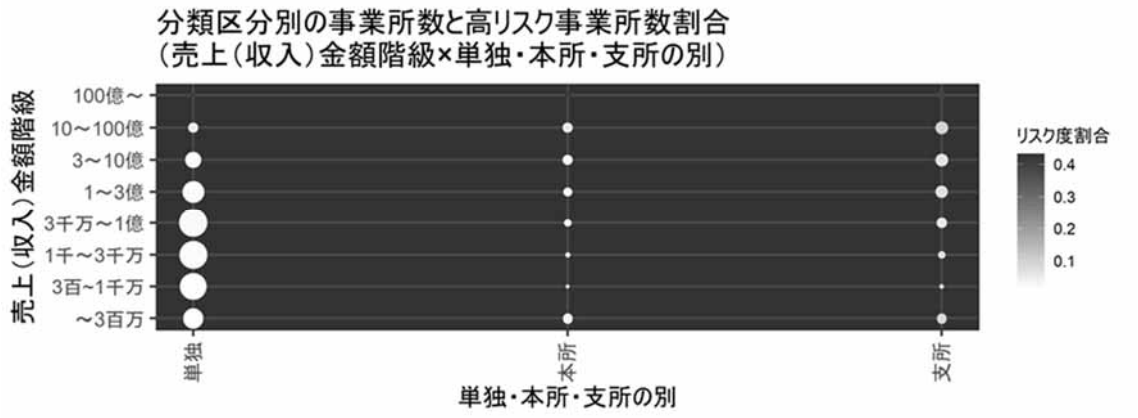


図 B-15 売上（収入）金額階級×開設時期

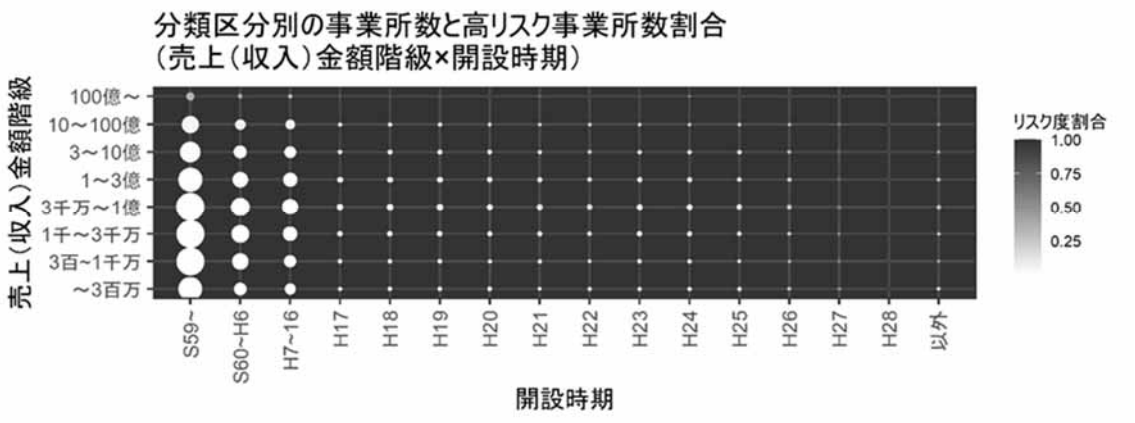


図 B-16 経営組織×単独・本所・支所の別

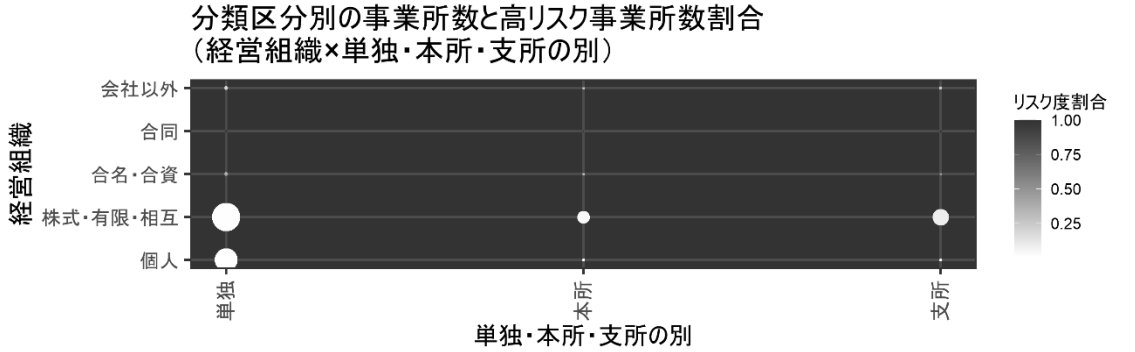


図 B-17 経営組織×開設時期

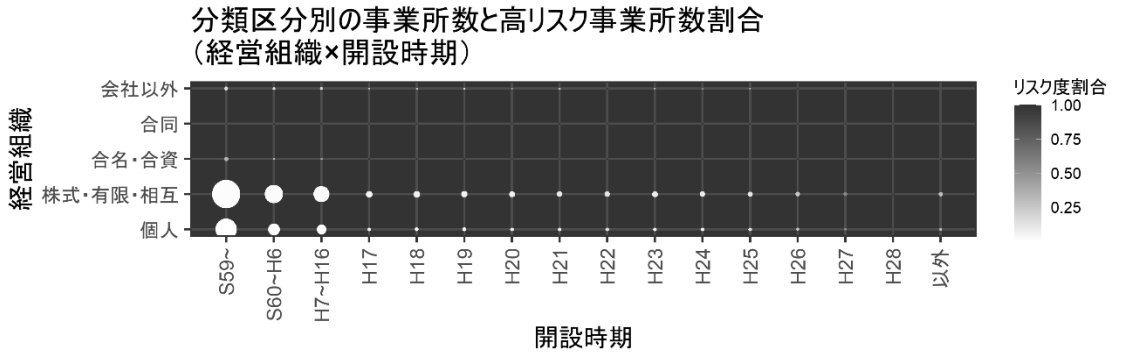


図 B-18 開設時期×単独・本所・支所の別