

< 修 士 論 文 >

公的統計マイ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・企業の匿名化マイクロデータの作成に  
資する基礎研究—

(要 旨)

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

## 1. 研究背景

匿名化マイ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であり、海外においても作成事例は多くない（伊藤（2018））。しかし、学術研究の利用だけでなく高等教育への需要も存在するほか、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るなど、わが国でも事業所・企業系の匿名化マイクロデータへの需要は存在すると考えられる。そこで本研究では、公的統計マイ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・企業の匿名化マイクロデータの作成に資する基礎研究を行う。

## 2. 事業所・企業系の匿名化マイクロデータ

諸外国の事業所・企業系の匿名化マイクロデータのサーベイをもとに、はじめに事業所・企業系の匿名化マイクロデータの現状について整理し、世帯・人口系との差異をまとめた。また、イタリア（Ichim（2007））やドイツ（Lenz *et al.*（2006））での研究例や実際の提供事例のサーベイから、第一に、学術研究用ファイルにおいては偶発的な個体特定や外部情報を用いたマッチングへのリスクを定量的に評価すること、第二に、匿名化手法には従来の非攪乱的手法だけでなくマイクロアグリゲーションといった攪乱的手法が有望であること、第三に、匿名化手法の適用にあたって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重要であるという知見を得た。

## 3. 経済センサスのマイクロデータを用いた秘匿性と有用性の評価研究

わが国の代表的な事業所・企業の統計調査である平成 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マイクロデータ作成の実証研究を行った。実験には、諸条件を満たす産業大分類 E（製造業）の事業所 1 万レコードをテストデータとして使用し、先行研究（伊藤他（2014））で用いられている手法を中心に評価を行った。

まず、地域（都道府県）、産業中分類、従業者合計、資本金額、売上（収入）金額といった記述統計量および分布特性を調査し、特に量的属性について分布の歪みが非常に大きいことを確認した。

続いて、地域、産業、従業者合計、資本金額のそれぞれの分布を考慮しながら、分類区分を荒くまとめるグローバルリコーディングを行った。リコーディング後の質的属性の組み合わせごとに、事業所の度数が 3 未満となるレコード数の割合で秘匿性評価を、情報エントロピーに基づく情報量損失率で有用性評価を、またそれらを用いた総合評価を行った。

さらに、量的属性である売上（収入）金額を対象に、前述のリコーディング済みの質的属

性で層化を行って匿名化を検討した。レコード削除では要約統計量が大きく変化することから、攪乱的手法のひとつであるマイクログリゲーションを選択した。秘匿性評価には距離計測型リンケージに基づくリンクが真である確率を、有用性評価には度数の平均変化率に基づく指標を、さらにそれらを合わせた総合評価を行った。実務への適用にあたっては、許容できる秘匿性を満たす中で有用性が大きくなる匿名化手法の選択が推奨される。

#### 4. 経済センサスにおける事業所の分布特性の把握と探索的な検証

前章では先行研究に基づいて匿名化手法の評価を行った。本章では、経済センサスの匿名化マイクロデータ作成に向けて、経済センサスの特性をより深く探るための探索的な検証を行った。実験には、諸条件を満たす産業大分類 E（製造業）の事業所 10 万レコードを使用した。地域、産業、従業者合計、資本金額だけでなく、売上（収入）金額、また、経営組織、単独・本所・支所の別、開設時期といった属性も含め、どのような属性の組み合わせが外部参照情報とのマッチングキーになる可能性が高いかを検討した。

さらに、属性単位ではなくレコード単位で、どのような事業所が相対的に高い特定化リスクを有するかを調べた。属性ごとに分類区分別の事業所数の観点から「リスク度」を定義してランク付けすることで、複数の準識別子を考慮して相対的に露見リスクが高いレコードを探索した。高リスク事業所と低リスク事業所に層化して比較した結果、量的属性の多くは規模が大きいほど高リスク事業所になりやすいこと、また、資本金額は例外的に小規模な事業所にも高リスク事業所が多く存在することがわかった。さらに、従業者合計と売上（収入）金額のような相関が高い量的属性同士の場合、その規模の程度に差がある場合にも相対的にリスクが高くなることがわかった。現実にはどのような属性が準識別子になりうるか、外観識別性をどう評価するなど今後の検討課題である。本研究の成果は、経済センサスの匿名化マイクロデータの作成を試行するにあたっての基礎資料になることが期待される。

#### 参考文献

- Ichim, D. (2007) *Microdata Anonymisation of the Community Innovation Survey Data: A Density Based Clustering Approach for Risk Assessment*. Dokumenti 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.
- 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.
- 伊藤伸介(2018)「公的統計マイ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.