2021年度統計関連学会連合大会
一般講演　経済・経営統計 (2)

独立行政法人
統計センター

事業所・企業系のミクロデータを用いた匿名化手法の適用可能性の検討

独立行政法人統計センター　横溝秀始
中央大・経済　伊藤伸介

1 はじめに：事業所・企業系の匿名化ミクロデータの現状

わが国の公的統計では、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

海外においてもEurostat、イタリア、ドイツなどわずかな作成事例しかなく、
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（伊藤(2018)）

一方……

事業所・企業系の匿名化ミクロデータには、**学術研究**だけでなく**高等教育**への需要大

また、匿名データ有識者会議の匿名データ作成方法ワーキンググループでは、2020年から事業所・企業系の調査である**賃金構造基本統計調査**に関する議論も行われている

わが国においても、事業所・企業系の匿名化ミクロデータへのニーズは存在するのではないか

1 はじめに：本発表の流れ

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

- 事業所・企業系の匿名化ミクロデータの作成に関する海外の現状
- 経済センサスのデータ特性の把握と匿名化措置の可能性の追究
- 先行研究に基づく経済センサスの個票データを用いた匿名化技法の有効性の検証

2 サーベイ：世帯・人口系のデータと事業所・企業系のデータの特性の比較

個人に関するミクロデータと企業に関するミクロデータの特性　　O'Keefe (2014) Fig. 1を和訳

	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

⇨ **情報の秘匿性が問題になるのは、特に大企業**

2 サーベイ：世帯・人口系のデータと事業所・企業系のデータの特性の比較

企業に関するデータの匿名化の難しさ（個人に関するデータとの比較）

Lenz (2006)

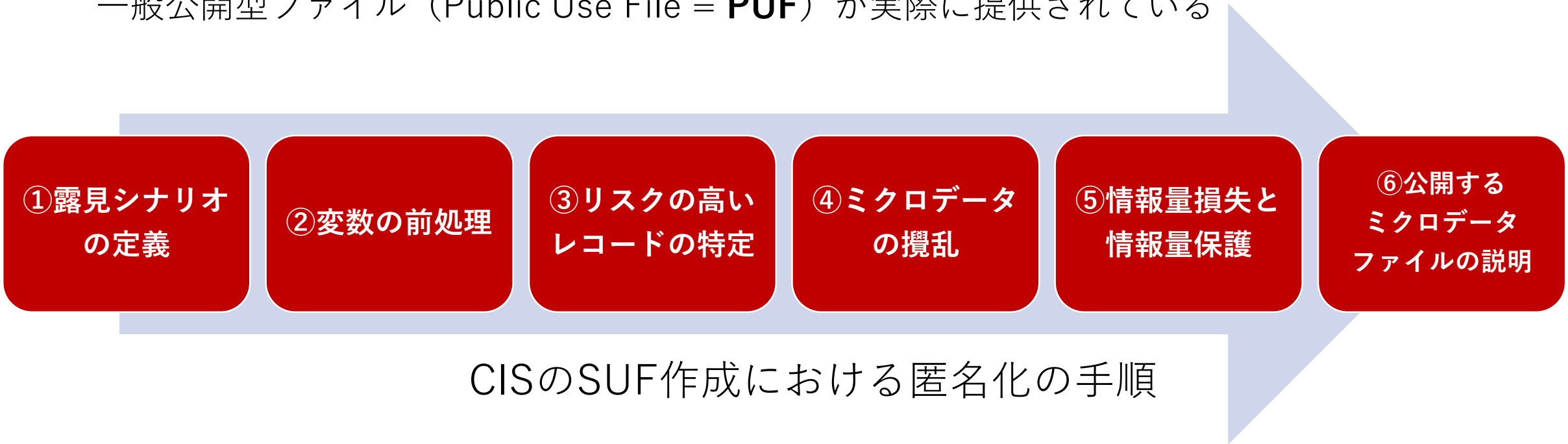
- 母集団が小さい場合、個々のグループに含まれるレコード数(セルに含まれる度数)も小さくなる
- 量的変数の分布は極端に不均質
- サンプリングの対象となるレコード数は企業規模ごとに大きく異なり、サンプリングにあたっては悉皆で抽出される層も存在する
- 企業にはデータの公表義務があるため、侵入者（intruder）は精度の高い**外部情報の取得**が容易
- 事業所・企業系のデータの**露見（disclosure）に伴うリスク**は、個人・世帯の調査における露見リスクより大きい

2 サーベイ：イタリアの先行事例・先行研究

ISTATにおけるSUF作成事例（Ichim (2007)、Ichim (2008) ）

CIS (Community innovation survey)

- EUの企業のイノベーション活動の調査（日本の科学技術研究調査に類似）
- 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
- 学術研究用ファイル（Scientific Use File = **SUF**）および
  一般公開型ファイル（Public Use File = **PUF**）が実際に提供されている



CISのSUF作成における匿名化の手順

2 サーベイ：ドイツのミクロデータの匿名化に関する研究の動き

事実上の匿名性
（factual anonymity）

- **「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)）

2002～2005年
「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）

2006～2008年
「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）

いずれも連邦教育研究省（BMBF）が後援
（Lenz *et al.* (2009)）

Bundesministerium
für Bildung
und Forschung

事業所・企業系の学術研究用ファイル（Scientific Use File）や教育用ファイル（Campus File）を10調査以上提供

2 サーベイ：イタリア・ドイツの事例で注目すべき点

露見シナリオは、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

秘匿性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、
有用性は実用例のサーベイを基に複数の指標を考慮して評価する

匿名化手法には、**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**等の攪乱的手法を採用し、元データとの近似性を重視

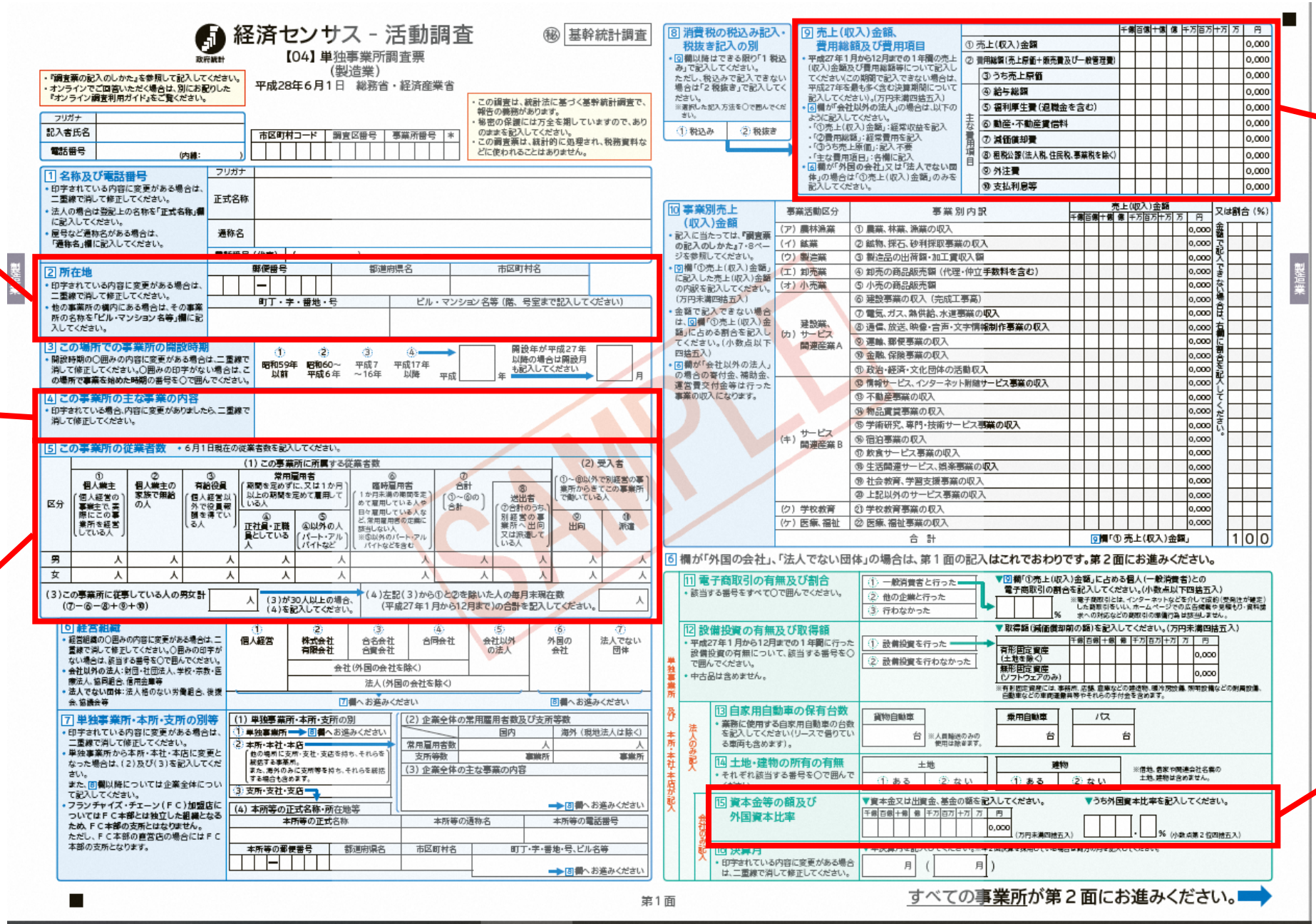
匿名化手法の適用にあたっては、
統計調査ごとのデータ特性や統計調査の**実務担当者の助言**も考慮

3 経済センサスを用いた評価：経済センサスの概要

平成28年経済センサス - 活動調査

- 調査目的：**全産業分野**の**経理項目**を網羅的に把握し、
  わが国における事業所・企業の経済活動を明らかにするとともに、
  事業所及び企業を対象とした各種統計調査の**母集団情報**を得ること
- 調査対象：一部例外を除くわが国**すべての事業所・企業**
- 調査事項：所在地、従業者数、経営組織、売上（収入）金額、費用総額等

3 経済センサスを用いた評価：調査票（産業共通部）



地域

産業

従業者

売上

資本金

3 経済センサスを用いた評価：実験における条件

実験における条件

- 産業大分類は**製造業**を対象
- 実験で用いる個票データは、**事業所**を対象にしたレコードに限定
- 従業者合計（男女計）が1人以上1000人未満に限定
- 結果表における売上集計の対象および付加価値集計の対象が該当
- 上記を満たす414,258レコードの中から10,000レコードを無作為抽出

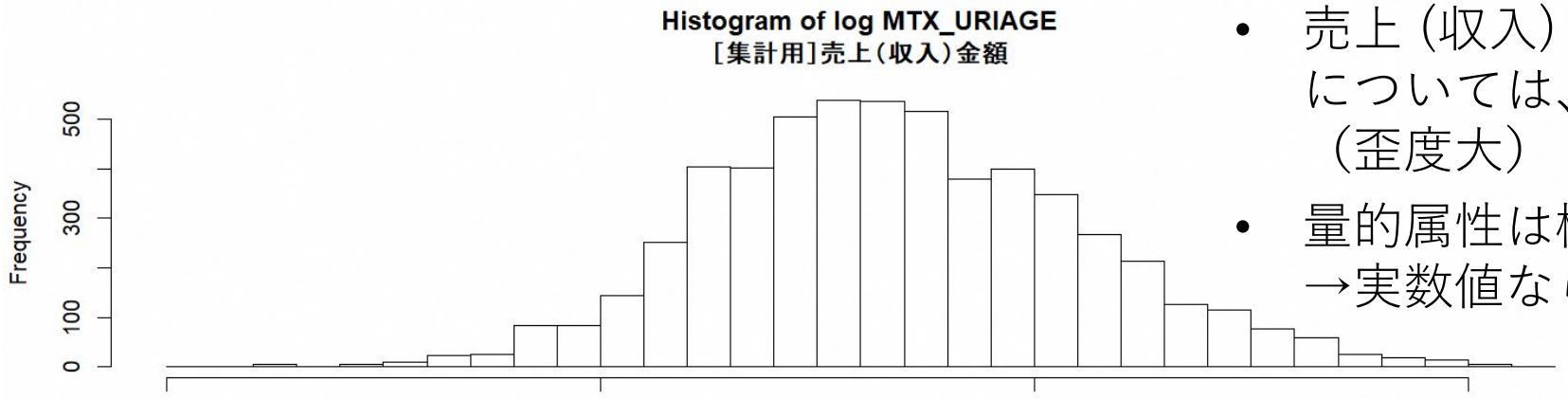
海外の事例や、ミクロデータの評価手法の先行研究（伊藤他（2014））をもとに、経済センサスにおける匿名化ミクロデータの作成可能性を検討する

3 経済センサスを用いた評価：記述統計量および分布特性

記述統計量

	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

量的属性の分布の例（対数）



※秘匿上の観点から目盛りは省略

- 売上（収入）金額、付加価値額などの経理項目については、平均値と中央値の差が大きい（歪度大）
- 量的属性は概ね対数正規分布に従っている。
→実数値ならば非常に右裾の長い分布

⇒分布の歪みに注意が必要

3 経済センサスを用いた評価：質的属性の匿名化

市区町村レベルの情報や具体的な資本金額などを公開すると、
容易に事業所が特定される可能性がある

⇒**リコーディング**：分類区分を粗くした上で秘匿性を高める匿名化手法

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

- 特定化のリスクを大きく高めると考えられる4つの属性（キー変数）を
  リコーディングして匿名化
- 特定の内訳の構成比が小さくなりすぎないことと、ユーザーの使い勝手を考慮

⇒**どのようなリコーディングの組み合わせがよいか？**

3 経済センサスを用いた評価：質的属性の秘匿性評価

k-匿名性 Samarati & Sweeney (1998)

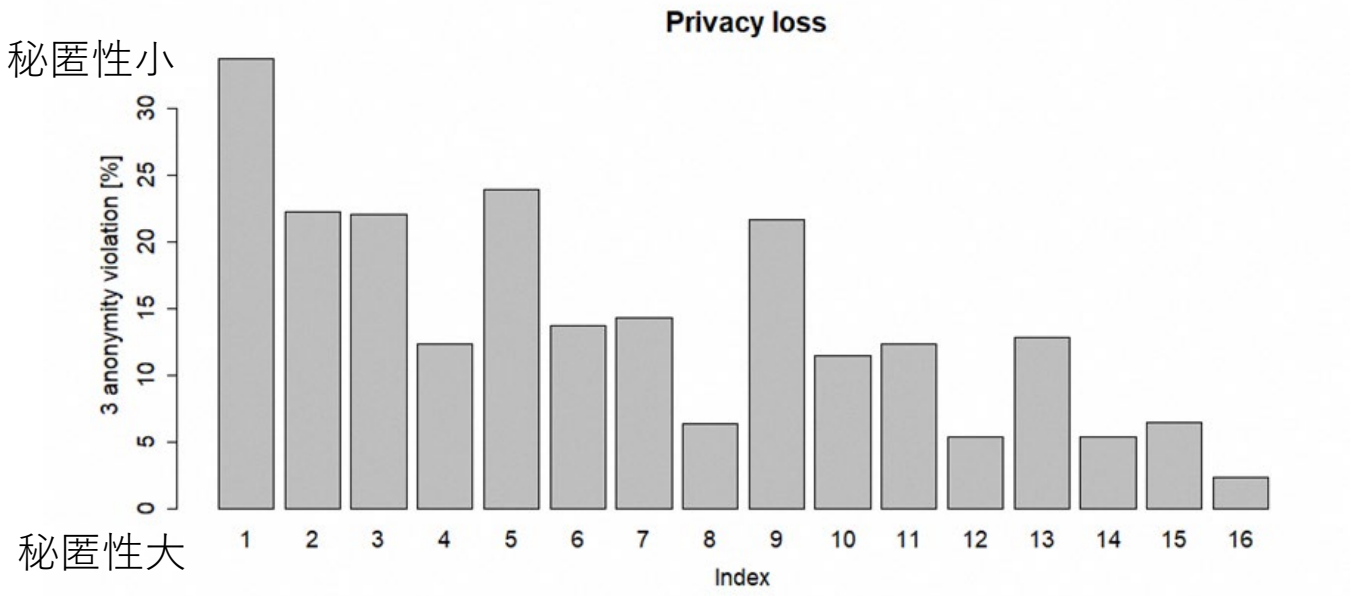
同じ属性値の組み合わせを持つレコードが、どの組み合わせについても必ずk個以上存在すること

例：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

3-匿名性違反

質的属性の秘匿性評価
(3-匿名性違反のレコード数の割合)



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

リコーディングの区分が粗いほど秘匿性は大きくなる

3 経済センサスを用いた評価：質的属性の有用性評価

情報エントロピー　　Kooiman *et al.* (1998)

ある事象の組み合わせで表される系で、
各事象の情報量の平均のこと

$$\text{シャノン情報量} = -\log p (0 \leq p \leq 1)$$

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

n：事象の数

p_i：i番目の事象が起こる確率

例

区分	世帯人員	件数
01	5人	600
02	6人	300
03	7人以上	100

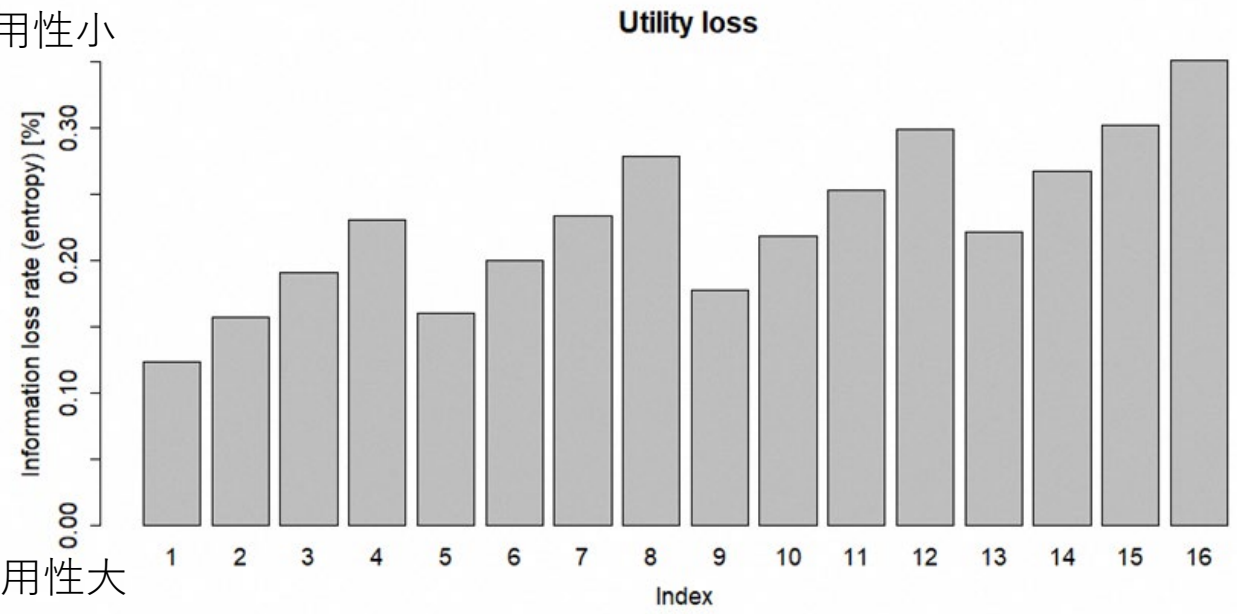
リコーディング →

区分	世帯人員	件数
01	5人	600
02	6人以上	400

$$\text{情報エントロピー} = -\frac{300}{400}\log_2\frac{300}{400} - \frac{100}{400}\log_2\frac{100}{400} = 0.81128$$

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）

有用性小



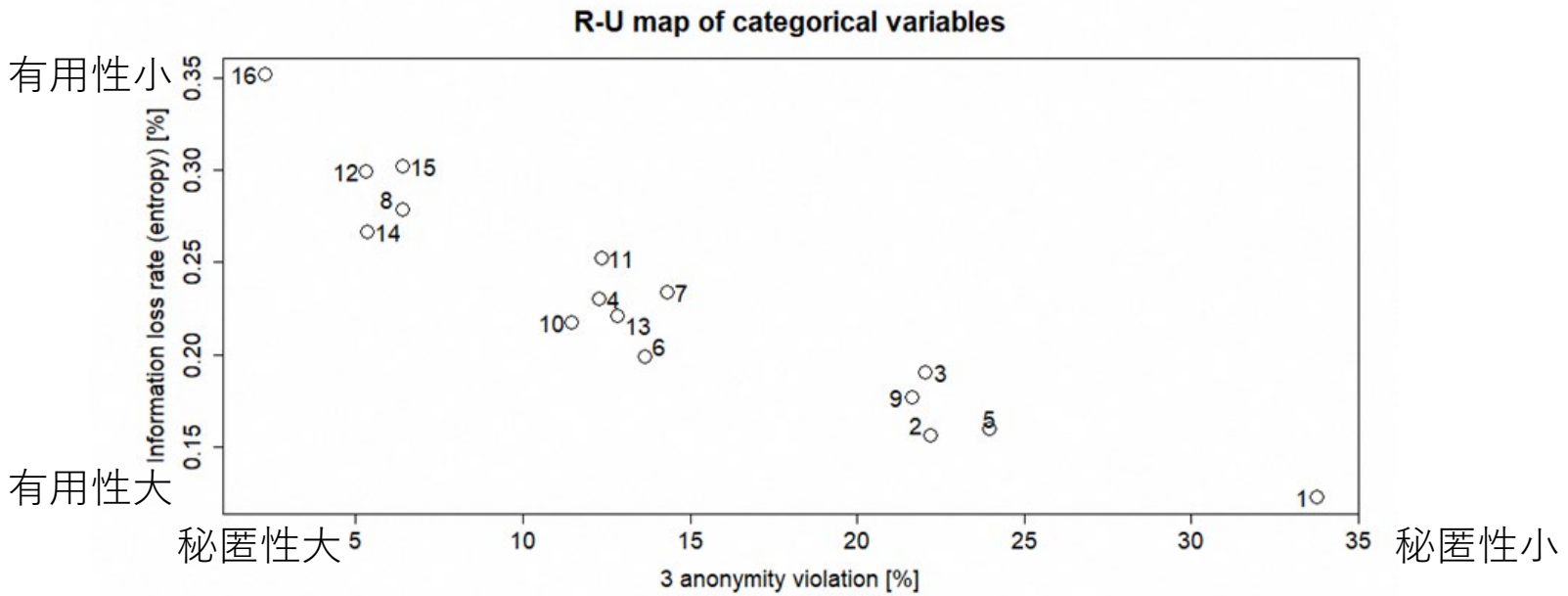
有用性大

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

リコーディングの区分が粗いほど有用性は小さくなる

3 経済センサスを用いた評価：質的属性の総合評価

質的属性の**R-Uマップ** (Risk-Utility confidentiality map)　　Duncan *et al.* (2001)



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

- 左下の領域にあるデータセットほどミクロデータとして優秀
- 秘匿性と有用性はトレードオフの関係にあるため、
  許容できる秘匿性の範囲で有用性を最大にするデータセットを選択する

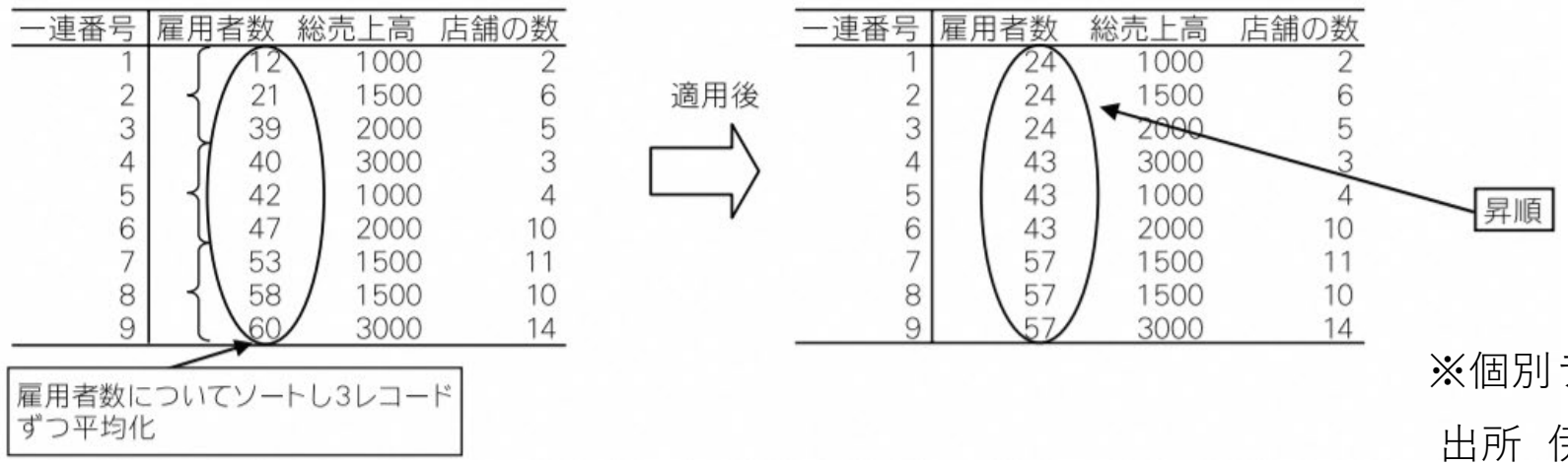
3 経済センサスを用いた評価：量的属性の匿名化

極端に大きな売上（収入）金額などは、事業所の特定化のリスクが高い
しかし、安易なレコード削除は分布を大きく歪ませる

⇒ **ミクロアグリゲーション** (microaggregation)　Defays & Nanopoulos (1993)

同質的なレコード群にグループ化した上で、個々の属性値を平均値等の代表値に置換（攪乱）

例



※個別ランキング法の場合

出所　伊藤(2009)

- 本実験では、 売上（収入）金額、給与総額、減価償却費、付加価値額を対象に匿名化技法を適用
- リコーディングされた質的属性で層化し、それぞれの層の中でミクロアグリゲーションを実行

3 経済センサスを用いた評価：量的属性の秘匿性評価

距離計測型リンケージ
(distance-based record linkage)
※標準化ユークリッド距離

$$d_{il}^2 = \sum_j \left(\frac{x_{ij} - \bar{x}_j}{\sigma(x_j)} - \frac{X_{lj} - \bar{X}_j}{\sigma(X_j)} \right)^2$$

原データと秘匿処理済データにおけるレコード間の距離を基準に、正しいレコードに対応付けされる割合 (true link rate)

例

原データ

一連番号	年齢	収入
001	40	520
002	50	480
003	60	500

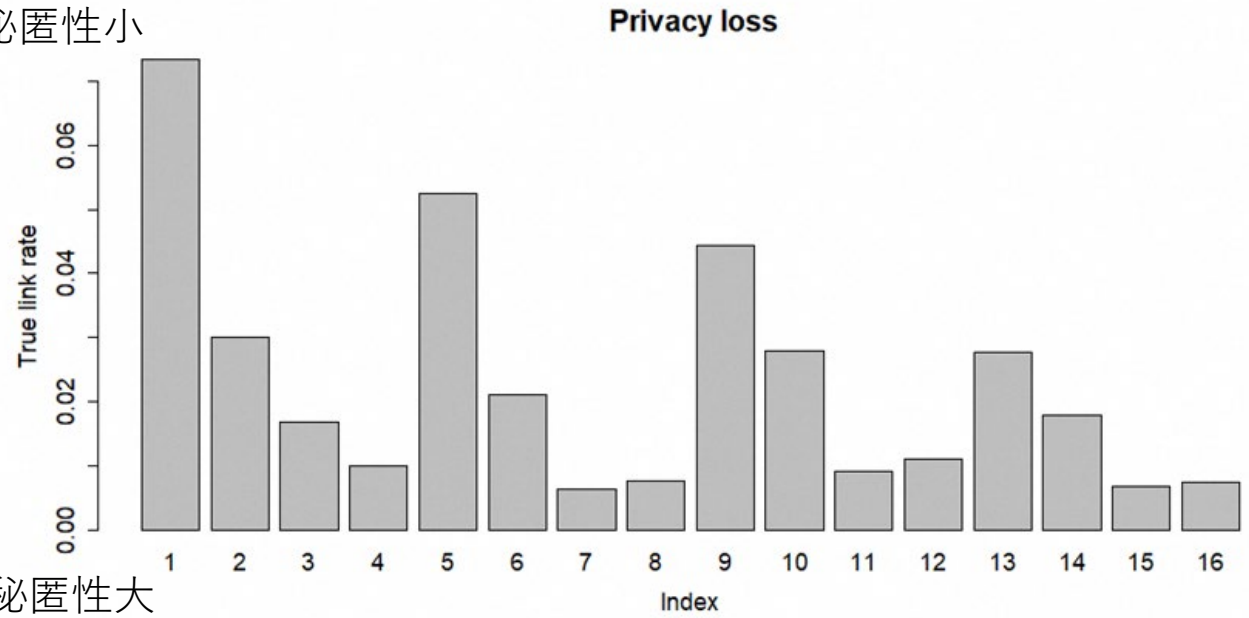
距離の計測

秘匿処理済データ

一連番号	年齢	収入
001	40	500
002	60	500
003	50	500

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)

秘匿性小



秘匿性大

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

リコーディングが粗いほど秘匿性は大きくなる

3 経済センサスを用いた評価：量的属性の有用性評価

IL1s

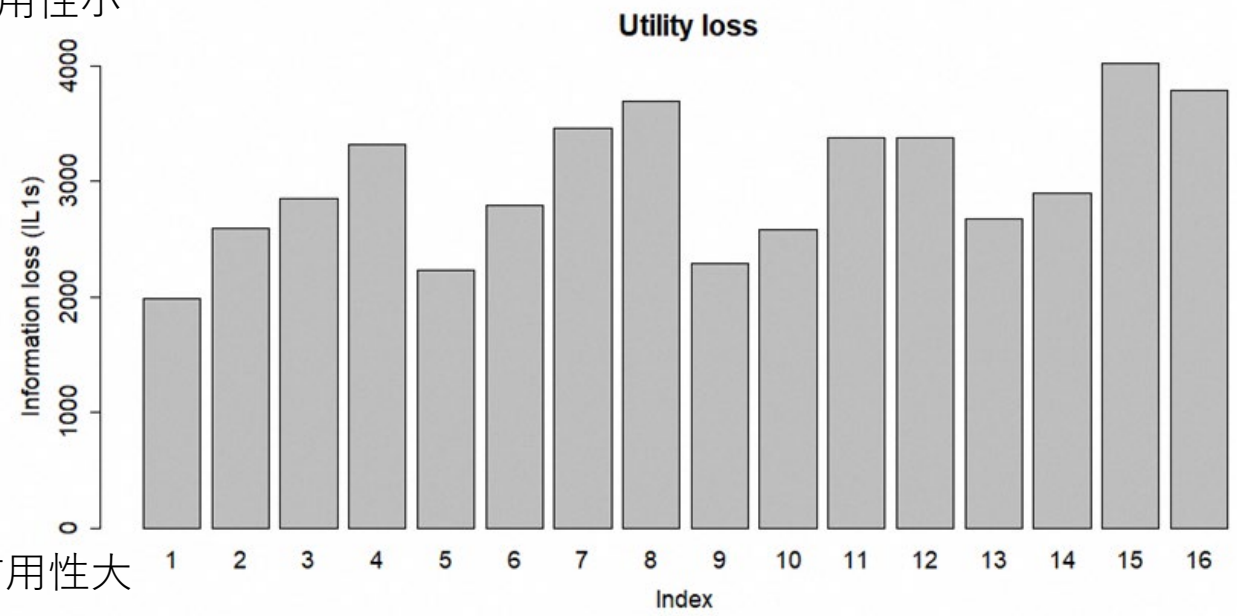
Mateo-Sanz, *et al.* (2004)

攪乱の前後で属性値がどの程度変化したかを、標準偏差を考慮して表した指標

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

量的属性の有用性評価 (IL1s)

有用性小



有用性大

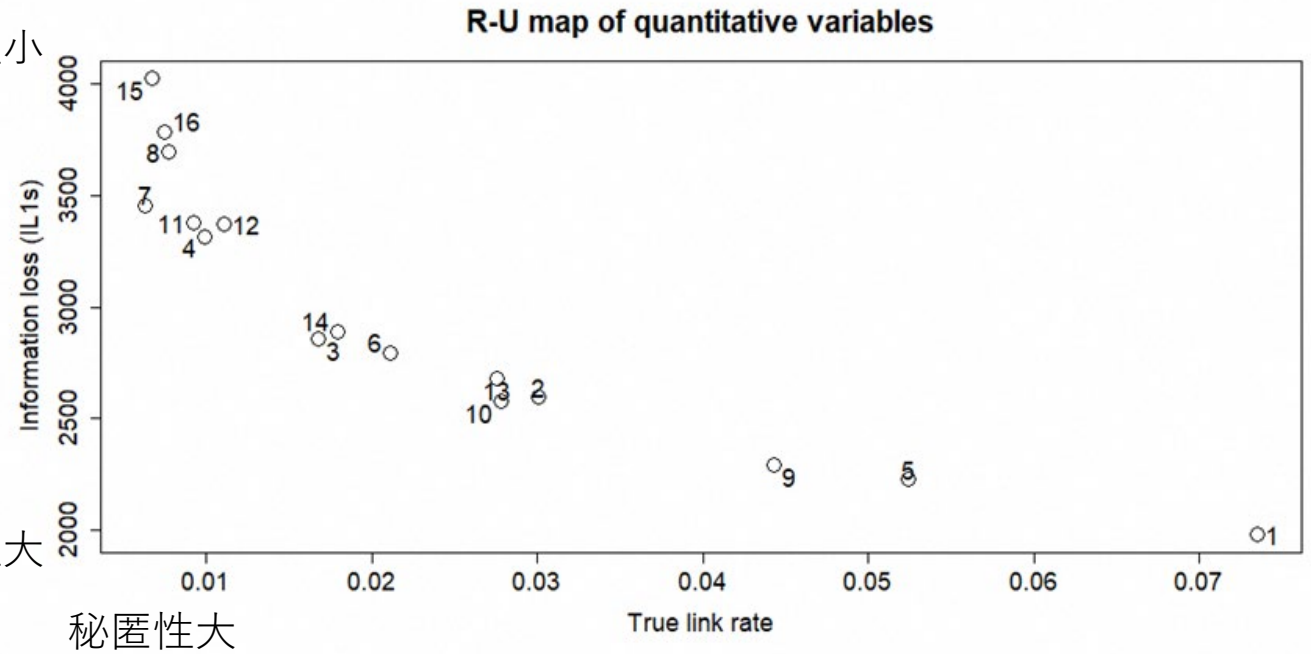
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
地域	8区分	8	8	8	8	8	8	8	3	3	3	3	3	3	3	3
産業	24区分	24	24	24	11	11	11	11	24	24	24	24	11	11	11	11
従業者規模	13区分	13	5	5	13	13	5	5	13	13	5	5	13	13	5	5
資本金階級	11区分	5	11	5	11	5	11	5	11	5	11	5	11	5	11	5

リコーディングが粗いほど有用性は小さくなる

3 経済センサスを用いた評価：量的属性の総合評価

量的属性の総合評価(R-Uマップ)

有用性小



有用性大

秘匿性大　　秘匿性小

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

- 質的属性の総合評価と同様に、**秘匿性と有用性はトレードオフの関係**

4 経済センサスの分布特性の把握と探索的検証

これまで先行研究をもとに有用性と秘匿性の定量的な評価を行ってきたが、本研究においては、経済センサスの分布状況も考慮しながら、さらなる実証研究を行う

- 露見リスクが相対的に高い事業所の特性を把握する必要性
- 地域、産業、従業者規模、資本金階級だけでなく、売上（収入）金額、経営組織、単独・本所・支所の別、開設時期といった属性もキー変数になりうる

⇨ これらを総合的に考慮した上で、本研究では、事業所あたりの相対的な露見リスクの大きさを「**リスク度**」として定義して、比較・検討を試みる

4 経済センサスの分布特性の把握と探索的検証：リスク度

リスク度の考え方

8属性から2属性ずつクロス集計（${}_8C_2 = 28$通り）

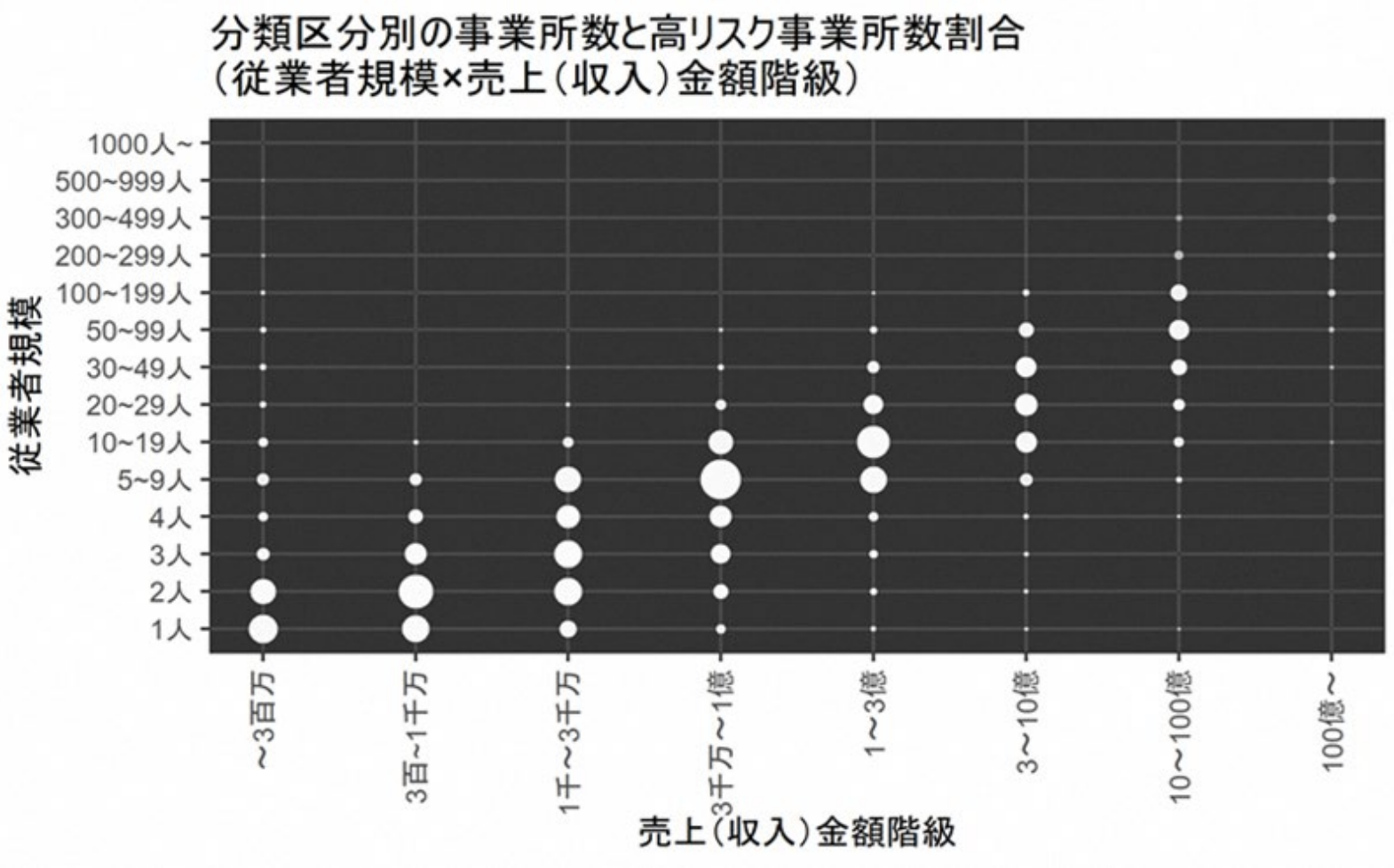
事業所	地域	産業	従業者規模	…	開設時期	地域×産業	地域×従業者規模	地域×資本金額	…	経営組織×開設時期	**リスク度**
1	東京都	10	5~9人		H17						0
2	埼玉県	32	1人		S59以前						0
3	宮崎県	11	4人		H28					足し上げ	0
4	青森県	15	1000人~		H20		1			1	2
5	東京都	15	10~19人		H23		1				1
6	滋賀県	24	3人		H24						1
7	埼玉県	12	20~39人		H27						0
8	茨城県	17	1人		H7~H16						3
9	石川県	30	5~9人		H18						0
10	広島県	22	100~999人		S59以前						2
⋮	⋮	⋮	⋮	⋮	⋮						⋮

データセット全体で、
事業所数が**10**未満となる
分類区分の組を持つ事業所に
リスクありとして**1**を立てる

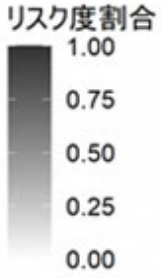
リスク度≧1：高リスク事業所

リスク度0：低リスク事業所

4 経済センサスの分布特性の把握と探索的検証：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
- バブルの色はリスク度割合
  （暗いほどリスク度が高くなる）



- 規模が大きい事業所は相対的にリスク度が高い傾向
- しかし、規模が小さければ低リスクとは限らない
- 相関性の考慮が必要であり、暗く見えるエリアに特に注意が必要

まとめと今後の課題

事業所・企業の匿名化ミクロデータの作成に関する論点の整理を行った

海外における事業所・企業系の匿名化ミクロデータの作成の現状を明らかにした

経済センサスの個票データをもとに、先行研究に基づく匿名化技法の定量的な評価、経済センサスのデータ特性を踏まえた匿名化技法の有効性を検証を行った

今後の実証研究の方向性としては、経済センサスにおける製造業特有の経理項目に着目した評価研究等が考えられる

参考文献

- ミクロデータ利用ポータルサイトmiripo, 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, 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―全国消費実態調査と家計調査を用いて― 」 統計研究彙報 第71号 2014年3月 （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, 77 101頁
- 伊藤伸介・横溝秀始(2021)「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, 1～61頁
- 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
- Brandt, M., Lenz, R. and Rosemann, M. (2008) Anonymisation of Panel Enterprise Microdata – Survey of a German Project, in Privacy in Statistical Databases, LNCS 5262 (Domingo-Ferrer et al. eds.), 139–151, Springer, Berlin
- 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.
- 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.

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
- Ichim, D. (2008). Community Innovation Survey: a Flexible Approach to the Dissemination of Microdata Files for Research.
- Kooiman, P., Willenborg, L., & Gouweleeuw, J. (1998). PRAM: A Method for Disclosure Limitation of Microdata. Research Paper, No. 9705, Statistics Netherlands, Voorburg.
- 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.
- Lenz, R. & Zwick, M. (2009). Business Microdata in Germany: Linkage and Anonymisation. Schmollers Jahrbuch : Journal of Applied Social Science Studies / Zeitschrift für Wirtschafts- und Sozialwissenschaften. 129. 645-653. 10.3790/schm.129.4.645.
- 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.
- 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.
- Samarati, P., & Sweeney, L. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Carnegie Mellon University. Journal contribution.

ご清聴ありがとうございました

本報告資料は、以下でダウンロード可能です

統計センター　学会発表

https://www.nstac.go.jp/services/society.html

参考資料

匿名化ミクロデータ

（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（ミクロデータ利用ポータルサイトmiripo）

元データ（調査票情報）

連番	名称	地域	産業	従業者	売上[万]
1	○○興行（株）	愛知県名古屋市	設備工事業	42	6,000
2	有限会社××組	滋賀県長浜市	職別工事業	189	128,000
3	△△製造所	滋賀県彦根市	食品製造業	4	1,200
4	△△金属□□支店	滋賀県彦根市	鉄鋼業	5	1,600
5	△△金属◇◇営業所	滋賀県大津市	鉄鋼業	7	2,200
6	△△金属☆☆営業所	京都府宇治市	鉄鋼業	13	3,200
7	株式会社▽▽繊維	京都府京都市	繊維工業	425	980,000

匿名化

匿名化ミクロデータ

連番	地域	産業	従業者	売上[万]
1	愛知県	建設業	30~	5,000
2	滋賀県	建設業	30~	140,000
3	滋賀県	製造業	1~4	1,700
4	滋賀県	製造業	5~9	900
5	滋賀県	製造業	5~9	1,800
6	京都府	製造業	10~29	3,900
7	京都府	製造業	30~	880,000

- 直接的な識別情報（名称等）の削除
- リコーディング（公開する分類事項の程度を粗くする）
- 量的変数の攪乱（データに偽の要素を混ぜ込む）　etc...

わが国の匿名化ミクロデータ

公的統計の二次的利用制度の例（出典：miripo）

現在提供中の匿名データ
- 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

匿名データ（統計法36条）

- 調査票情報を、特定の個人又は法人その他の団体の**識別ができない**ように加工したもの
- 学術研究及び高等教育の発展に資すると認められる場合に提供される

匿名化手法：類型

手法	匿名化手法	匿名化手法の内容
非攪乱的手法	データのリサンプリング	統計調査の調査票のレコードのすべてを用いず、一部を抽出
	識別情報の削除等	直接識別できる情報をすべて削除し、無作為に並べ替え
	特異なレコードの削除	特徴的な値のレコードを削除（例：世帯人員が８人以上など）
	トップコーディング ボトムコーディング	極端に大きな（小さな）値は、上限値（下限値）を設ける （年齢を「85歳以上」や「15歳未満」でまとめるなど）
	リコーディング	分類事項の程度を粗くする（年齢５歳階級など）
攪乱的手法	ノイズ付加	加法ノイズや乗法ノイズを付与する
	スワッピング	ミクロデータに含まれるレコード同士で属性値を入れ替える
	ミクロアグリゲーション	同質的なレコード群にグループ化した上で、個々の属性値を平均値・中央値等の代表値に置き換える
	PRAM (post randomization method)	マルコフ推移確率行列に基づき，確率的にレコードの値を置き換える

伊藤ら(2014)、統計センター(2020)を元に作成

ミクロデータの匿名化にあたり、諸外国では攪乱的手法である、
「**ミクロアグリゲーション(microaggregation)**」の研究や実用化が進んでいる

匿名化手法：ミクロアグリゲーション / 単一軸法の例

- 同質的なレコード群にグループ化した上で、個々の属性値を平均値等の代表値に置き換える
- ソートやグルーピングの基準によっていくつかの手法が存在する

単一軸法（single axis method）

ある単一の属性あるいは統計指標でソートし、まとめてグループ化

伊藤（2009）図2より

個別データ

一連番号	雇用者数	総売上高	店舗の数
1	12	1000	2
2	21	1500	6
3	39	2000	5
4	40	3000	3
5	42	1000	4
6	47	2000	10
7	53	1500	11
8	58	1500	10
9	60	3000	14

グループ化

ミクロアグリゲーション済のデータ

一連番号	雇用者数	総売上高	店舗の数
1	24	1500	4
2	24	1500	4
3	24	1500	4
4	43	2000	6
5	43	2000	6
6	43	2000	6
7	57	2000	12
8	57	2000	12
9	57	2000	12

平均値

注　閾値を3に設定

複数の属性をまとめて攪乱するため、**攪乱の程度が比較的大きい**

匿名化手法：ミクロアグリゲーション / 個別ランキング法の例

①雇用者数の攪乱

一連番号	雇用者数	総売上高	店舗の数
1	12	1000	2
2	21	1500	6
3	39	2000	5
4	40	3000	3
5	42	1000	4
6	47	2000	10
7	53	1500	11
8	58	1500	10
9	60	3000	14

雇用者数についてソートし3レコードずつ平均化

適用後

一連番号	雇用者数	総売上高	店舗の数
1	24	1000	2
2	24	1500	6
3	24	2000	5
4	43	3000	3
5	43	1000	4
6	43	2000	10
7	57	1500	11
8	57	1500	10
9	57	3000	14

昇順

②総売上高の攪乱

②総売上高

一連番号	雇用者数	総売上高	店舗の数
1	24	1000	2
5	43	1000	4
2	24	1500	6
7	57	1500	11
8	57	1500	10
3	24	2000	5
6	43	2000	10
4	43	3000	3
9	57	3000	14

総売上高によって一連番号の順番が変わっている

総売上高についてソートし3レコードずつ平均化

適用後

一連番号	雇用者数	総売上高	店舗の数
1	24	1167	2
5	43	1167	4
2	24	1167	6
7	57	1667	11
8	57	1667	10
3	24	1667	5
6	43	2667	10
4	43	2667	3
9	57	2667	14

③店舗の数の攪乱（スペースの都合で省略）

個別ランキング法（individual ranking method）

量的属性の各々について、
個別にソートとグループ化

ミクロアグリゲーションの結果

一連番号	雇用者数	総売上高	店舗の数
1	24	1167	3
2	24	1167	7
3	24	1667	7
4	43	2667	3
5	43	1167	3
6	43	2667	12
7	57	1667	12
8	57	1667	7
9	57	2667	12

属性を個々に攪乱するため、
攪乱の程度は比較的小さい