合成データの生成手法の有効性に関する定量的な評価
―事業所・企業系のミクロデータを用いて―

横溝　秀始*

伊藤　伸介†

A Quantitative Evaluation of Generation Methods for Synthetic Data: Using Business Microdata

YOKOMIZO Shuji
ITO Shinsuke

本稿では、「経済センサス - 活動調査」の個票データを用いて、事業所・企業系の統計調査を対象にした合成データの作成可能性を追究するたけでなく、有用性と秘匿性に関する各種の評価方法に基づいて、合成データの生成手法の有効性に関する定量的な評価を行った。本実験では、回帰や決定木に基づいて作成された各種の合成データの生成手法について、ミクロアグリゲーション等の攪乱的手法に適用された評価指標と同様の指標を用いて、有効性に関する比較・検証を行った。本研究の結果から、個票データに対して合成データの手法である CART やバギングを適用した場合、ミクロアグリゲーションの MDAV 法と比較しても、有用性を維持したまま秘匿性を高められる可能性があることがわかった。

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

This paper investigates the potential of synthetic data for statistical analysis using individual data from the Economic Census for Business Activity and quantitatively evaluates the effectiveness of generation methods for synthetic data based on data usability and data confidentiality. Specifically, comparative quantitative research is conducted to establish the effectiveness of generation methods of synthetic data such as regression and decision tree using the same indicators as for perturbative methods such as microaggregation. This results show that applying CART and Bugging as synthetic methods to individual data enhances the degree of data confidentiality with the same level of data utility as the microaggregation method MDAV.

Keywords: Business microdata, Anonymization methods, Synthetic data, Microaggregation, CART

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

1. はじめに

わが国の公的統計では、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関しては匿名データの作成・提供はなされていない。また、全国消費実態調査と就業構造基本調査については一般用ミクロデータが公開されているが、事業所・企業系の統計調査は現時点では作成の対象となっていない。その一方で、高等教育機関における教育目的の利用だけでなく、オンサイト利用を行う前のプログラム作成用のテストデータへのニーズが存在することから、近年海外では、統計実務の観点から、合成データ(synthetic data)の作成と実用化に向けた研究も進められている[1]。

合成データとは、元になるデータからその分布特性が近似するように、データに含まれる属性の値を新たに生成することによって作成され、個体情報の秘匿性が確保されたミクロレベルの擬似的なデータである(Templ(2017, p.157))。その意味では、合成データは、個票データに含まれる個体情報の特定につながりうる変数・レコード群を対象に、攪乱的手法を含む各種の匿名加工の手法を施して作成される匿名化ミクロデータ(anonymized microdata)とは異なると言える[2]。また、合成データの作成は、統計モデルに基づくパラメトリックな手法の適用と、CART(＝Classification And Regression Tree)等によるノンパラメトリックな手法の適用に大別される(Reiter(2005)等)。さらに合成データは、対象となるすべてのレコードの中で欠測値を含む属性群に対して擬似的に値を生成する完全合成データ(fully synthetic dataset) (Rubin(1993))、一部のレコード群に含まれるセンシティブな属性にのみ擬似的な値によって補完を行う部分合成データ(partially synthetic dataset) (Little(1993))に類別できる(Drechsler(2011), 高部(2022))。完全合成データの場合、例えばパラメトリックな手法であれば、対象となるすべてのレコードに含まれる特定の属性値をそれ以外の属性で説明する統計モデルを設定し、モデルの推定によって値を擬似的に生成することが考えられる。それに対して部分合成データの場合には、一部のレコード群を対象に，外れ値に該当するような特定の属性値群を欠測値と見なした上で、生成の対象となる属性値群については、統計モデルのようなパラメトリックな手法によって値を生成するか、あるいはクラスタリングの手法を用いて、元データにおける該当する属性値群の分布特性に近似するように確率的に生成する方法が適用される[3]。

世帯・人口系の統計調査だけでなく、事業所・企業系の統計調査についても、このような合成データを作成することができれば、高等教育やテストデータといったニーズへの対応が期待できる。そのためには、事業所・企業系の統計調査をもとに、様々な手法を用いて合成データを作成した上で、合成データの生成手法の有効性を検証することが求められる。

本稿では、「経済センサス - 活動調査」(以下、「経済センサス」と略称)の個票データを例に、事業所・企業系の統計調査を対象にした合成データの生成手法を追究するだけでなく、その有効性について定量的な評価を行う。具体的には、生成された合成データに関する秘匿性や有用性に関する指標を用いた評価研究を行う。さらにそれらの評価指標に基づいて、攪乱的手法が適用された匿名化ミクロデータと各種の合成データの比較・検証を試みることに

---

[1] 例えば、Bates et al. (2019)は、イギリス国家統計局における合成データの生成手法に関する検討状況を議論している。また、2017 年 4 月～6 月のイギリス労働力調査(Labour Force Survey)のミクロデータをもとに、後述する R のパッケージである synthpop 等、合成データ作成用の複数のパッケージを比較・検討した上で、統計実務の観点から合成データの作成可能性を追究している。

[2] アメリカセンサス局が 2010 年に作成した人口センサスの Public Use Microdata Sample のように、匿名化ミクロデータの作成において合成データの方法論が用いられる場合もある。それは、匿名化技法の 1 つとして合成データの手法が適用される事例である(伊藤(2022))。

[3] 高部(2022)では、部分合成データの作成を指向しており、一部のレコード群に含まれる特定の属性値群に対して、パラメトリックな統計モデルを用いて対象となる数値を生成する方法について議論している。

よって、わが国における公的統計の合成データの作成可能性について議論していきたい。

2.　公的統計における合成データの作成状況―海外の事例とわが国の動向―

諸外国では、合成データの作成方法に関する研究やその適用が進められてきた(伊藤(2018))。本節では、公的統計を対象にした合成データに関して、海外での作成の事例とわが国における動向について見ていくことにしたい。

2.1.　公的統計の合成データ作成に関する海外の現状

海外では、公的統計の合成データに関する生成手法に関して、具体的な作成事例も含め、様々な調査研究がなされている。合成データの生成に関する研究については、公的統計や大規模データを対象にしたプライバシー保護に関する国際会議である Privacy in Statistical Databases 2022 においても、多くの研究論文が発表されている[4]。また、海外では、公的統計を対象にした合成データの方法論の適用事例が存在する。

第 1 の事例は、アメリカセンサス局(以下「センサス局」)における合成データの方法論の適用である。これは、2010 年の一般公開型ミクロデータサンプル(Public Use Microdata Sample)に部分合成データの手法を適用したものである。また、地理的なクエリ応答システムである On the Map において、The Longitudinal Employer-Household Dynamics および Origin-Destination Employment Statistics が、合成データという形で設定されている(伊藤・寺田(2020))。

第 2 の事例は、欧州統計局(Eurostat)の一般公開型ファイル(Public Use File=PUF)の作成のための取り組みである。Eurostat は、EU-SILC(=European Union Statistics on Income and Living Conditions)において、合成データの方法論を用いた一般公開型ファイル(Public Use File)を作成・公開している(伊藤(2018))。具体的には、原データの分布に基づいて統計的なモデルをもとにシミュレーションによる合成データの手法を適用している。

第 3 の事例は、エディンバラ大学を中心として実施されているスコットランド縦断調査(Scottish Longitudinal Study、以下 SLS と略称)を対象にした、合成データの作成の取り組みである。SLS は、ONS がオンサイト施設で提供サービスを行っているイギリス国家統計局縦断調査(ONS Longitudinal Study、以下「LS データ」と略称)と同様のデータ特性を有している。LS データは、1971 年人口センサスの個票データのサンプルに政府保健中央レジスターの行政記録情報と突合した上で、そのサンプルを対象に人口センサスの個票データを連結した縦断的なデータ構造を備えている。SLS も同様に、スコットランドの人口センサスの個票データに医療データがリンケージされた縦断的なデータである。SLS に対して、R の合成データ作成用のパッケージである synthpop (Nowok et al.(2016))を用いて、人口センサスの合成データが作成されている。

なお、統計作成部局による事業所・企業系のデータを対象にした合成データの作成に関する興味深い研究事例としては、オーストラリア統計局による事業所・企業用の合成データの研究がある。具体的には、Chien らによって、事業所・企業用の合成データの研究が進められてきた(Chien et al.(2021))。

Hang et al. (2021) は、調査客体が特定されるリスクの観点から見た事業所・企業系のデータの特徴を以下のように整理している。第 1 は、事業所・企業系のデータに含まれる変数の

[4] 2022 年 9 月に開催された Privacy in Statistical Databases 2022 (PSD2022)では、合成データに関するセッションが設置されており、所得税に関する合成データの生成方法や、合成データの有用性に関する実証研究といった論文についての報告がなされている。

多くは大きく歪んでおり、最大規模の事業所については、個体識別リスクは著しく高くなることである。第 2 は、事業所・企業系のデータにおいては変数間の相関性が高い場合に、一部の事業所について属性漏洩(attribute disclosure)のリスクが相対的に高まることである。第 3 は、規模が大きな事業所や企業が多く把握される事業所・企業系の統計調査の場合、これらの大規模な事業所・企業がサンプルとして抽出される、または悉皆調査される可能性が極めて高いことである。第 4 は、企業に関するより多くの情報が公開されており、特に株式公開の対象企業や大企業はその経済活動に関連する情報の公表を法律で義務付けられていることである。そして、第 5 は、競合する企業が特定されることによって、機密情報が入手できれば、それがもたらす利益が享受できることから、個体識別のインセンティブが個人・世帯系のデータよりも高くなることである。

こうした事業所・企業系のデータの特性を考慮した上で、Chien らは、オーストラリアの場合、事業所・企業は、産業によっては市場が寡占あるいは複占の状態になっていることから、情報の削減や攪乱的手法といった従来の匿名化手法が世帯・人口系ほど有効でない可能性を指摘している。そこで、攪乱的手法との比較・検証を定量的に行った上で、秘匿性や有用性のバランスの観点から、合成データの方法論が事業所・企業系のミクロデータに対して適用可能な手法として追究されている。

2.2.　わが国における公的統計の合成データの作成について

わが国でも、合成データの作成方法に対する関心が高まっている。例えば、南(2022)は、深層学習の方法論が適用された GAN(=Generative Adversarial Networks、敵対的生成ネットワーク)を用いた合成データの実装化の可能性に言及している。また、千田他(2022)は、海外における民間のデータを対象にした合成データの方法論の実用化の動向を踏まえて、プライバシー保護型合成データの可能性を議論している。さらに、最近のわが国における合成データの作成方法に関する実証研究としては、例えば高部(2022)による公的統計の擬似ミクロデータの作成を指向した統計モデルを用いたパラメトリックな手法の検討を指摘することができる。

現在公開されている全国消費実態調査の一般用ミクロデータの作成については、2011 年 8 月から(独)統計センターで試行提供された教育用擬似ミクロデータ(Synthetic Microdata)の作成に関する手法が、その方法的な基礎になっている。具体的には、全国消費実態調査の個票データから高次元の集計表を作成した上で、高次元の集計表の中のセルに含まれる平均や標準偏差だけでなく、変数間の相関性も考慮した上で、多変量の正規乱数の生成が行われている(山口他(2013))。これは、全国消費実態調査における高次元の集計表の分布特性に近似させる形で乱数を生成することを指向している点では、合成データの方法論の適用事例の１つとみなされうる。

わが国にもこのような作成事例があるものの、海外と比較すると公的統計を対象にした合成データの生成手法に関する調査研究は多くない。したがって、海外における研究動向を踏まえて、パラメトリックおよびノンパラメトリックの両面から、わが国の公的統計に対する合成データの方法論の適用可能性を模索することは、有益であると考える。とくに事業所・企業系の統計調査に関して、わが国で匿名データが作成・提供されていない現状を踏まえると、こうした事業所・企業系の統計調査に合成データの生成手法を適用することによって、例えば個票データの分析で使用するプログラムコード作成用のテストデータへの展開も期待できる[5]。そこで、次節以降では、経済センサスの個票データを用いて合成データの生成手

[5] 合成データは擬似的なデータであり、公的統計の分野では、教育目的の利用かあるいはプログラムコードの作

法を追究した上で、その有効性について定量的な評価を行うことにしたい。

3.　使用するデータ

本研究では、平成 28 年の経済センサスの個票データを用いて実験を行う。テストデータとして、産業大分類 E(製造業)の事業所レコードについて、諸条件[6]を満たす約 36 万レコードの中から単純無作為抽出した 10,000 レコードを使用した。分析対象となる項目には、集計結果表で用いられている集計事項を中心に、属性情報の入手可能性等を考慮して、外部参照情報になりうるキー変数として地域、産業(中分類)、従業者規模、資本金階級を採用した。また、露見リスクが大きいと考えられるセンシティブな属性や分析上有用と思われる属性として、売上(収入)金額、付加価値額、給与総額、減価償却費を選定した。キー変数については、秘匿性および分析上の有用性を考慮してあらかじめリコーディングを行った(表 1)。地域については、47 都道府県を 8 区分に、産業は 23 区分ある産業中分類を 11 区分に統合し、量的属性である従業者合計および資本金額はそれぞれ 5 区分に階級化した。

表 1　キー変数とリコーディング[7]

属性名	区分数	分類区分
地域	8区分	北海道, 東北, 関東, 中部, 近畿, 中国, 四国, 九州・沖縄
産業（産業中分類）	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	5区分	1~4, 5~9, 10~29, 30~99, 100~
資本金階級	5区分	~1000万, 1000万~1億, 1億~10億, 10億~, 以外

注　「～1000 万円」は「～1000 万円未満」を表す。

表 2 は原データにおける経理項目間の相関係数である。売上(収入)金額と給与総額の間には 0.8 を超える正の相関が確認できる。ミクロデータを用いて様々な計量分析が行われることを勘案すると、原データにおける相関関係が合成データにおいても保持されることが期待される。

表 2　量的属性間の相関係数

	売上（収入）金額	付加価値額	給与総額	減価償却費
売上（収入）金額	1.00	0.60	0.81	0.62
付加価値額	0.60	1.00	0.52	0.33
給与総額	0.81	0.52	1.00	0.50
減価償却費	0.62	0.33	0.50	1.00

原データにおける量的属性の要約統計量を表 3 に示す。経理項目は mean(平均)と median(中央値)の差が大きく、右裾の長い分布になっている。また、at_1%(1%点)を見ると、0 や負の値

成・チェックのための利用が指向されている。合成データの場合、学術研究目的のために直接利用することは、一般に想定されておらず、仮に実証分析を行ったとしても、その結果数値の信頼性は担保されない。

[6] 結果表における売上集計対象および付加価値集計対象の両方を満たすレコードのうち、秘匿性の観点から従業者合計(男女計)が 1 人以上 1,000 人未満のレコードを対象とした。また、製造業の経理項目が集計対象外であることから個人票のレコードは対象外としている。

[7] 製造業における産業中分類は、「09 食料品製造業」～「32 その他の製造業」の 22 種類が存在する。該当する事業所数が多い中分類としては、「24 金属製品製造業」や「26 生産用機械器具製造業」、「15 印刷・同関連業」等があげられる。産業分類の詳細は以下を参照。
https://www.stat.go.jp/data/e-census/2016/kekka/bunrui.html

を含む属性も存在している。このことは、対数変換への制約をもたらすことを意味している。対数化を行わない場合、合成データの生成や定量的な評価を行うにあたり、外れ値となるような極端な値を持つレコードの影響を大きく受ける可能性がある[8]。そこで本研究では、量的属性として取り扱う売上(収入)金額、付加価値額、給与総額、減価償却費の 4 つの経理項目に対して、neg-log 変換(negative logarithmic transformation)(高部(2017))を適用した[9]。neg-log 変換は主にファイナンスや信用リスク分析の分野で多く用いられる変換である。0 や負の値も含めて対数化することで、事業所・企業系特有の極端な分布の歪みの影響が緩和されることが期待される。図 1 は、neg-log 変換前後の付加価値額のヒストグラムを表示している。左側(neg-log 変換前)は非常に偏りの大きい分布を示している。一方、右側(neg-log 変換後)には、概ね対数正規分布に従う大きな山が出現している。また、正の値だけでなく負の値にも対数正規分布に近い小さな山が確認できる点が特徴的である。

表 3　原データにおける量的属性の要約統計量[10]

変数名	n	mean	sd	median	at_1%	at_99%
従業者合計	10,000	17	47	5	1	226
資本金額	6,843	80,119	1,200,141	1,000	84	1,192,520
売上（収入）金額	10,000	53,768	427,885	3,500	0	859,421
付加価値額	10,000	10,640	55,503	1,459	-823	154,092
給与総額	7,618	2,430	6,034	600	0	24,622
減価償却費	7,618	371	1,607	34	0	5,847

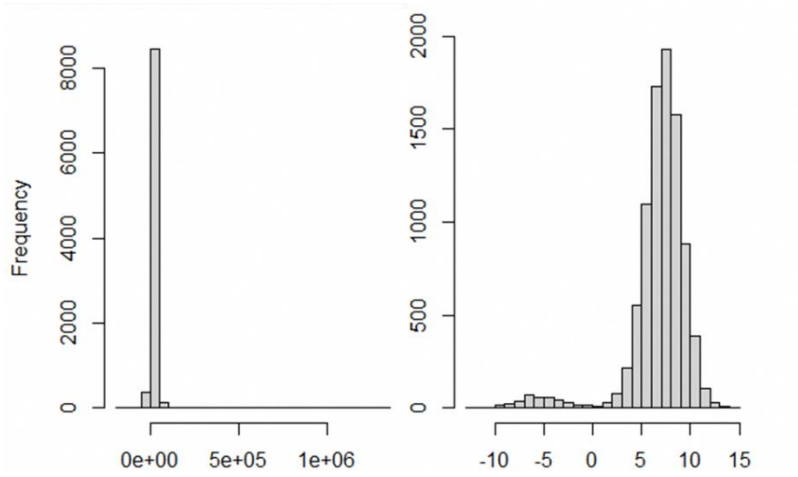

図 1　neg-log 変換後の付加価値額のヒストグラム(左：neg-log 変換前、右：neg-log 変換後)

4.　秘匿性と有用性に関する評価指標について

本節では、秘匿性と有用性のそれぞれについて定量的な評価を行うための指標について述

[8] 例えば、複数のレコード間で距離を計算し、その平均等を評価するような手法の場合、極端に大きな値のレコードが含まれていると対数化を行うか否かで結果に大きな差が生じる懸念がある。

[9] $Y_n = sgn(X_n) \times ln(|X_n| + 1)$　(高部(2017) p.33)

[10] 秘匿性の観点から 0%点(最小値)および 100%点(最大値)は記載していない。

べる。具体的には、攪乱的手法と合成データ生成手法の両方について有効性を評価することが可能な指標について議論する。

4.1.　秘匿性の評価指標

匿名化ミクロデータを対象にした場合、その秘匿性に関する定量的な評価方法は、(1)外部情報とミクロデータのマッチング、(2)母集団一意に関する指標の計測、(3)特殊な一意の分析(special uniques analysis)、(4)レコードリンケージによるリスク評価、(5)クロス集計表によるリスク評価に類別することができる(伊藤(2019))。

匿名化ミクロデータに関する秘匿性の評価指標に比較して、合成データの秘匿性評価指標の先行研究は多くない。その理由としては、合成データが、原データに対して直接匿名化措置が施されたデータではないことから、原データと合成データにおけるレコードの対応付けの可能性が追究されず、相対的に安全性の高いデータであると想定されていることが考えられる。

合成データにおける秘匿性の評価指標としては、完全合成データにおける属性漏洩を評価するために提案された差分属性正当確率(Differential Correct Attribution Probability = DCAP)(Taub et al. (2018))や、原データと合成データのすべてのレコードのペアのリンク確率を確率的に分類する確率的リンケージ(probabilistic record linkage) (Chien et al.(2021))が存在する。Hang et al. (2021)もまた、合成データにも属性漏洩によるリスクが存在することを指摘しており、合成データの秘匿性の評価については、定式化が複雑でなく、統計実務への適用が容易な、絶対相対差分(Absolute Relative Difference = ARD)と呼ばれる属性漏洩に関する評価指標を利用している。ARD は以下の式で表される。

$$ARD = \frac{\left|\hat{L} - L\right|}{L} \tag{1}$$

Lと$\hat{L}$はそれぞれ、原データおよび合成データに関する属性値の最大値を表している。ARD は、原データと合成データに含まれる属性値の最大値からの乖離を評価する指標である。レコードリンケージのように原データのレコードと匿名化措置を行ったレコードにおける 1 対 1 の対応を必要としないため、合成データだけでなく、攪乱した匿名化ミクロデータの評価を行うことも可能である。

完全合成データの場合、合成したレコードが原データのレコードにリンケージされることは原理的にないが、売上等のセンシティブな属性において最大値となる属性の情報が漏洩する可能性がないわけではない。侵入者が取りうる最も基本的な攻撃手段は、キーとなる属性の層ごとのセンシティブな属性の最大値を推定することである。特定の地域や産業において最も規模の大きい事業所は一般に知られていることが多いため、例えば、合成データ内の特定の地域や産業の中で最も大きな事業所の売上(収入)金額を調べることで、その最も規模の大きい事業所のセンシティブな属性情報を高い精度で推定することができる可能性がある。

先行研究ではキー変数による層化を行わず単変量から ARD を計算しているが[11]、本研究では上記のような具体的な攻撃を想定し、その対応策として、多変量を用いて層ごとの露見リスクを総合的に評価する層化平均 ARD(stratified average ARD)を提案する。層化平均 ARD は、

[11] Hang et al. (2021) では、乱数を変更して生成した複数の合成データセットに対して ARD が適用されている。また、2 種類のシナリオ(①攻撃者が合成データにのみアクセスできる場合、②攻撃者が、自身がデータベースで 2 番目に大きいレコードであることを知っている場合)ごとに ARD が評価されている。本研究の層化平均 ARD は、単一の合成データセットを前提に①を評価した。

キー変数の全組み合わせでそれぞれ原データと合成データの最大値の乖離を計算し、その平均から算出される。図 2 の例では、まず原データと合成データのテーブルから、層を作らずに原データの売上の最大値、合成データの売上の最大値から ARD を求める。次に、地域については、関東で層化した ARD、関西で層化した ARD、東北で層化した ARD をそれぞれ計算する。同様に、産業についても、機械と鉄鋼でそれぞれ ARD を算出する。さらに、地域と産業でクロスした場合でも ARD の計測を行う。これらの ARD について、層化に用いた属性の数を考慮しつつ平均を取ることによって、層化平均 ARD が算出される。なお、計算に細かい層を使用する場合や合成するレコード数が限られる場合は、必ずしも原データと合成データの層がすべて一致するとは限らない(図 2 の例では、東北の層、東北および機械の層が該当する)。その場合には、全レコードの中央値(325)を代用することで計算を行った。この例では、例示を簡潔にするために売上のみの計算結果を示しているが、実験では 4 つの経理項目をすべて使用して計算を行った。

層化平均 ARD が大きいことは、層ごとの最大値が全体的に原データと合成データで異なることを意味する。その場合、属性情報の推定が発生するリスクは相対的に小さいと考えることができる。

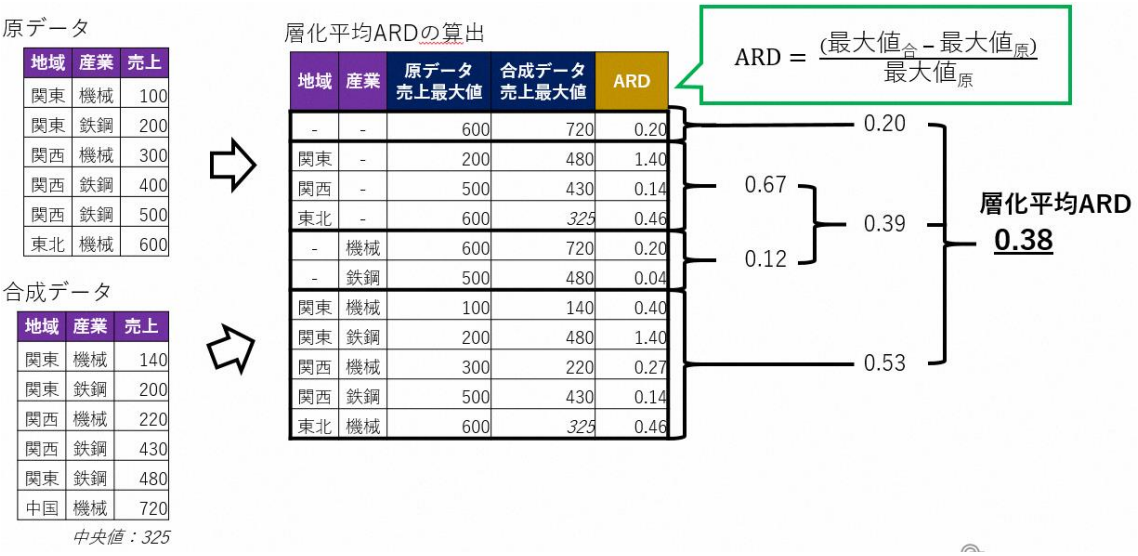

図 2　層化平均 ARD のイメージ

4.2.　有用性の評価指標

匿名化ミクロデータにおける有用性の定量的な評価方法としては、主として(1)記述統計量やクロス表における分布特性の比較や、(2)情報量損失(information loss)に関する指標の評価がある。前者においては、平均、分散等の記述統計量やクロス表における分布特性を比較したり、属性値の差、分散共分散行列や相関係数行列を比較したりすることが含まれる。属性値間の距離を定義し、その距離の近似度を計測する手法もこれに該当する。後者は、情報量損失に関する指標を定義した上で、原データから匿名化ミクロデータを作成した場合の情報量の低減の程度を定量的に評価する手法と言える(伊藤(2019))。

合成データにおける有用性の評価指標については、広義の尺度と狭義の尺度が存在する(Drechsler & Reiter(2009))。広義の尺度は原データと合成データとの距離を定量的に測った指標を含んでおり、Kullback-Leibler 情報量、Hellinger 距離、後述の傾向スコアに基づく pMSE といった指標が用いられる。狭義の尺度は原データと合成データの特定のモデルの差異に着

目するもので、ミクロデータを用いた各種の分析結果や信頼区間の差異が用いられる(Taub & Sakshaug(2020))。上記の中でも、傾向スコア(propensity score)(Woo et al. (2009))は合成データにおける代表的な有用性の評価指標である。傾向スコアとは、ある共変量が与えられた時、その個体がある群にあてはまる確率のことを示しており、医療統計等の分野で用いられることが少なくない。合成データの作成においては、そのレコードがどの程度合成データらしいか、その確率を傾向スコアとして算出することができる。傾向スコアを用いた評価は、米国センサス局でも採用されている(Drechsler & Reiter(2009))。

本研究では、傾向スコアを用いる評価指標の中でも、傾向スコア平均二乗誤差(propensity score Mean Square Error = pMSE)(Snoke et al. (2016))を採用した[12]。pMSE は以下の式で表される。

$$pMSE = \frac{1}{N}\sum(\hat{p}_i - c)^2 \tag{2}$$

Nは原データのレコード数と合成データのレコード数の和を、$\hat{p}_i$は各レコードの傾向スコアを、cはNに占める合成データのレコード数の割合を表す。

pMSE を算出するには、まず原データと合成データを縦に統合し、合成データか否かを示す属性を新たに追加する(合成データなら 0、そうでなければ 1)。この「合成データか否か」を目的変数に、地域や産業、売上(収入)金額といった属性群を説明変数に取り、ロジスティック回帰や CART といった手法を用いてモデリングすることで、各々のレコードがどの程度合成データらしいかを確率的に表す傾向スコアp_iを算出することができる。このp_iから、レコード全体に占める合成データの割合であるcとの乖離の差の平均を取ることでpMSE を計算する。pMSE が大きいということは、原データと合成データが区別しやすいということであり、それだけ合成データが原データから乖離している、すなわち有用性が低下していると評価することができる。

上記の pMSE は、属性間の相関性を直接的に評価する指標ではない。そこで本研究ではそれを補うために、相関係数行列の差の平均絶対誤差(mean absolute error of the difference of the correlation coefficient matrices)(Domingo-Ferrer et al.(2001))も指標として採用した。相関係数行列の差の平均絶対誤差は以下の式で表される。

$$\frac{\sum_{j=1}^{k}\sum_{1\le i\le j}|r_{ij} - r'_{ij}|}{\frac{k(k-1)}{2}} \tag{3}$$

出所　伊藤他(2014) 表 1

kは属性の数、rは原データの相関係数、r'は合成データの相関係数を表す。この相関係数行列の差の平均絶対誤差が小さいほど、原データの相関関係を保存できていると判断できる。

5.　攪乱的手法の有効性に関する比較・検証

[12] Woo et al. (2009) は、攪乱した匿名化ミクロデータの有用性評価について経験分布推定、クラスター分析、傾向スコアに基づく手法を比較し、傾向スコアが最も適していると評価した。また、Snoke et al. (2016) は傾向スコアを合成データの有用性評価に対して拡張した。

5.1.　攪乱的手法の概要

本節では、合成データの生成手法と比較・検討を行うための対象として、従来の攪乱的手法を用いて作成した匿名化ミクロデータに関する定量的な評価を行う[13]。攪乱的手法には、後述するミクロアグリゲーションと、標準偏差をもとに発生させた平均 0 の正規乱数を属性ごとにノイズとして付与する加法ノイズを使用した[14]。加法ノイズでは、発生させる正規乱数の標準偏差を変更することで、攪乱の程度を調節することが可能である。

ミクロアグリゲーションとは、原データを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ことで秘匿性を高める攪乱的手法である(伊藤(2009))。ミクロアグリゲーションにはいくつか種類が存在するが、本実験ではその中でも代表的な、①Z スコア総計法、②個別ランキング法、③MDAV 法を採用した。

Z スコア総計法は、標準化された属性の総計値でソートを行い、グループ化を行うミクロアグリゲーションの一手法である。対象となる属性群から算出された総計値をもとにレコード単位でソートを行うため、グループ化の対象となる属性値のばらつきが大きければ、攪乱の程度はより大きくなる。その場合、秘匿性の強度は高くなるものの、有用性の程度は低下する。個別ランキング法は、属性ごとに個別にソートした上でグループ化を行う手法である。レコード単位ではなく、属性ごとに個別にグループ化された近似的な属性値群に対して処理を行うことによって、攪乱の程度は相対的に小さくなる。したがって、個別ランキング法の場合、Z スコア総計法と比較して有用性は高くなるが、秘匿性はより小さくなることが知られている(伊藤(2009))。

MDAV(=Maximum Distance to Average Vector)法は、レコード間およびレコードとデータセットの全レコードの平均との距離を考慮し、均質なレコード群を形成するミクロアグリゲーションの一手法である(Domingo-Ferrer & Mateo-Sanz (2002), Hundepool et al. (2003))。ミクロアグリゲーションの中では近年研究事例が多く、匿名化ミクロデータ作成用 R パッケージ sdcMicro(Templ(2015))においてもミクロアグリゲーションに関するデフォルトの手法となっている。

図 3 は、MDAV 法の概要を示したものである。最初に、平均ベクトルを求めた上で、その平均ベクトルから最も距離が大きくなるレコードと、そのレコードからの距離が最大となるレコードを求める。つぎに、これらのレコードを中心にして、グルーピングを行う際の閾値である近傍 k 個のレコードの値の平均を取ることで、原データの有用性をある程度保持したまま、レコードの特定を困難にするような形で秘匿性を高めることができる。攪乱済みのレコードを取り除いた後、再帰的に処理を繰り返すことですべてのレコードが攪乱される。このように、レコード間の距離を考慮する点が、前述の Z スコア総計法や個別ランキング法との大きく異なる点であると言える。

なお、集約するレコード数の閾値 k の値を大きくするほど、攪乱の程度は大きくなるため、秘匿性の強度は相対的に高くなり、有用性の程度は逆に小さくなる。k の値を小さくするほど、原データの性質が残ることから秘匿性の強度は弱まるが、有用性はより大きくなる。したがって、設定する k の値によって攪乱の程度を調整することが可能である。この性質は、前述の Z スコア総計法や個別ランキング法にも共通するものである。

5.2.　攪乱的手法の有効性に関する定量的な評価

---

[13] 横溝・伊藤(2022)は、事業所・企業系の匿名化ミクロデータにおける各種の匿名化手法の評価や作成可能性について詳細に議論している。

[14] ノイズには属性値に直接係数を乗じる乗法ノイズも存在する。ドイツの匿名化ミクロデータ作成には、匿名化手法として乗法ノイズが採用されている(Brandt et al.(2008))。

図 4-1 は、ミクロアグリゲーションにおける Z スコア総計、個別ランキング法、MDAV 法と、加法ノイズのそれぞれを比較した R-U マップ(Risk Utility map) (Duncun et al. (2001))である。横軸は層化平均 ARD で秘匿性を、縦軸は pMSE で有用性をそれぞれ表している。

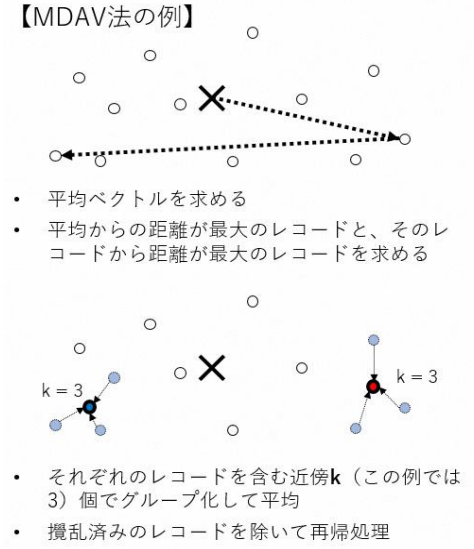

図 3　MDAV 法のイメージ

ミクロアグリゲーションについては集約するレコード数 k を、ノイズは攪乱率[%]を 3、5、10、30、50、100 と変化させてプロットした。k や攪乱率が大きくなるにつれて、秘匿性は大きく、有用性は小さくなる傾向が確認できる。本 R-U マップでは、秘匿性が高く、有用性も大きい理想的な手法は右下のエリアにプロットされるが、実際には秘匿性と有用性に関するトレードオフを考慮した上で、適切な手法が選択される。手法ごとに見ていくと、まず MDAV 法は、全体として秘匿性が小さいエリアにプロットされており、k の値を大きくすると有用性が大きく損なわれる傾向にある。次に個別ランキング法は、k の値にかかわらず、非常に秘匿性が小さいエリアにプロットされている。また、Z スコア総計法については、秘匿性は大きいものの、全体的に有用性が損なわれている。最後に加法ノイズは、秘匿性は攪乱率によって極端に変化するが、有用性はいずれの攪乱率でも損失が大きい傾向が見られた。

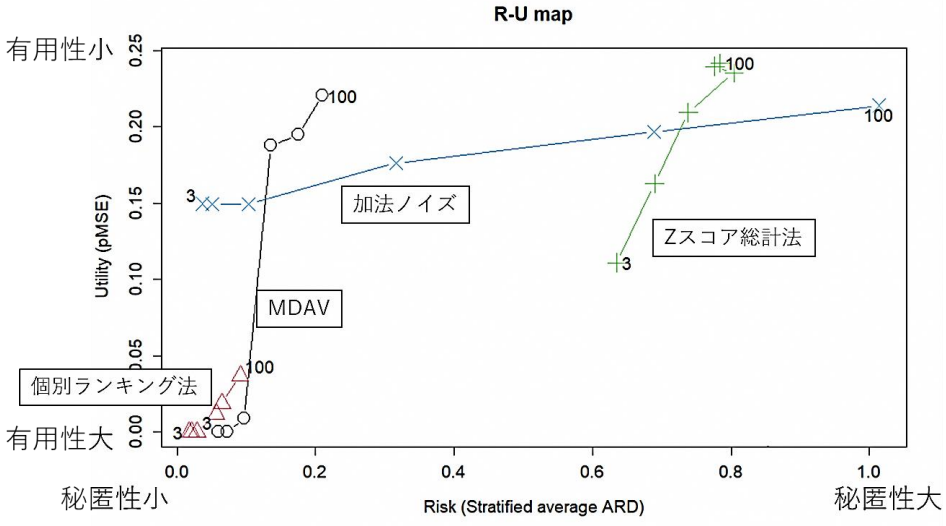

図 4-1　攪乱的手法に関する秘匿性(ARD)と有用性(pMSE)の評価結果

図 4-2 は、横軸については秘匿性として層化平均 ARD で、縦軸に関しては有用性として相関係数行列の差の絶対誤差で作成された R-U マップを示している。個別ランキング法、MDAV 法、加法ノイズは同一の曲線上に位置しており、トレードオフの関係性が明瞭に表れている。このことから、攪乱の程度によっては、秘匿性と有用性のバランスを取ることが可能である。それに対して、Z スコア総計法の場合、前述の手法と比べて有用性の著しい低下が顕著に示される結果となった。

以上の結果より、攪乱的手法については、個別ランキング法や MDAV 法を選択すれば、秘匿性の強度が相対的に小さくなるのに対して、Z スコア総計法や加法ノイズを選択すれば、有用性に課題が残ることが確認された。

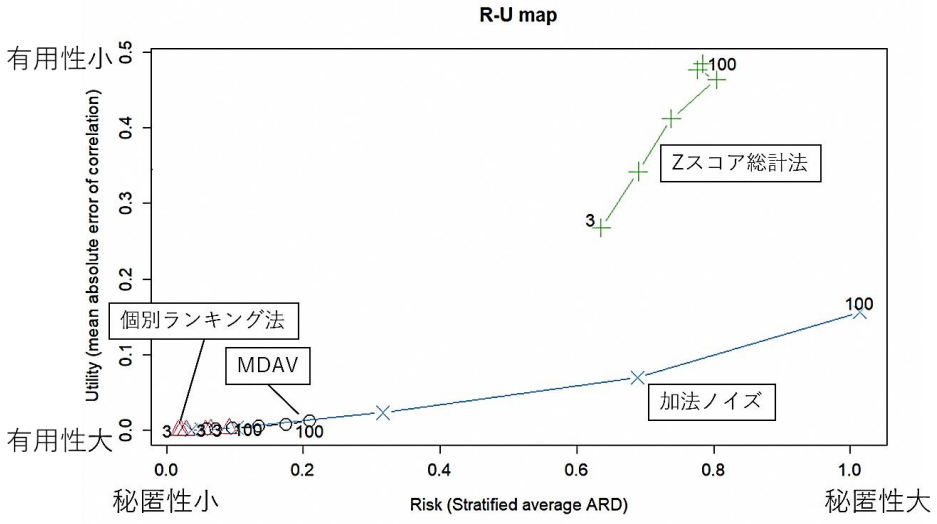

図 4-2　攪乱的手法に関する秘匿性(ARD)と
有用性(相関係数行列の差の平均絶対誤差)の評価結果

6.　合成データ生成手法の有効性の検証

前節では、匿名化手法を用いて匿名化ミクロデータを作成した上で、有用性と秘匿性の定量的な評価を行うことによって、それぞれの特徴を実証的に確認した。本節では、前節でその有効性を評価したミクロアグリゲーション技法の１つである MDAV を比較対象とし、各種の合成データを作成することでその差異を確認する。また、秘匿性や有用性の観点から、合成データ生成手法の有効性の検証を試みる。

6.1.　合成データ生成手法の概要

本研究では、合成データの生成については、合成データ生成用 R パッケージである synthpop を使用した(Nowok et al.(2016))。synthpop には、後述する CART、バギング、ランダムフォレストといった機械学習に基づくノンパラメトリックな手法や、パラメトリックな回帰分析を用いた手法があらかじめ実装されている。合成にあたっては、欠測値がなくキー変数としても重要な属性である地域の値を最初に生成し、産業、従業者規模、資本金階級、売上(収入)金額、付加価値額、給与総額と減価償却費の順に属性値の生成を行った。最初の属性である地域の数値は原データからランダムサンプリングで生成される。次の産業については、合成済

みの地域の属性値をもとにして確率的に生成が行われる。さらに、従業者規模については、すでに合成済みの地域と産業を用いて確率的に属性値が生成される。このように合成済みの属性を考慮しながら、属性値の生成を段階的に行うことによって、属性間の関係性を考慮した矛盾の少ない合成データを得ることができる[15]。

本実験で採用した合成データの生成手法は、①CART、②バギング、③ランダムフォレスト、④ランダムサンプリング、⑤回帰の 5 種類である。

CART (Classification And Regression Tree) (Breiman et al. (1984)) は、観測済みの属性から目的の属性を再帰的にグルーピングする、ノンパラメトリックな決定木分析手法である。質的属性と量的属性のいずれの合成にも適用可能であり、非線形の関係性を捉えやすいという利点がある。作成する木の深さや葉に振り分けるレコード数の制約(最小リーフサイズ)等を変化させることで、原データの性質をどこまで反映させるか調節することが可能である。合成データの作成手法として研究例が多く、synthpop ではデフォルトの合成手法にもなっている。

図 5 は、CART を用いて、地域と産業から売上を合成する際のイメージ図を示したものである。売上に関してできるだけ同質性が高いレコード同士のグループ化を可能にするために、地域と産業の質問を繰り返すことで決定木を作成する。この際、一般にジニ係数等の基準を用いて、どの属性のどの分類区分を採用するか、葉をさらに分割するか、あるいはその段階で分割を止めるかについての判断を行う。このようにして作成した決定木を用いることで、地域と産業から売上の値を合成することが可能となる。例えば、図 5 の例では、合成したレコードの中で地域が関東、産業が化学に該当する場合、あらかじめ作成しておいた決定木をたどることで、500 または 777 のいずれかの値が売上としてランダムに生成される[16]。

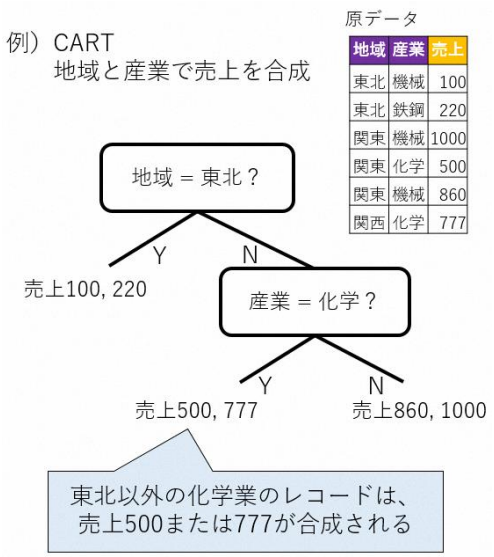

図 5　CART で地域と産業から売上を合成する際のイメージ

[15] 本実験では合成済みの属性を以降の属性の生成にすべて使用したが、属性ごとに使用する属性の有無を設定することも可能である。合成する順序や生成に使用する属性によって、作成される合成データの特性が変わるため、求められる水準の合成データが得られない場合は、属性の順序や使用する属性を探索的に入れ替えることや、分析上重要な属性で層化した上で合成を行うことが推奨される(Raab et al.(2021))。

[16] 777 のような特異な値が合成データに現れた場合、それが偶発的な属性情報の漏洩につながる可能性が存在する。synthpop ではこれを防ぐために、合成データ生成時にガウシアンカーネル密度推定(gaussian kernel density estimator)を行い、リーフ内の有限個の離散値を連続値に平滑化(smoothing)することで秘匿性を高めるオプションが用意されている(Nowok et al.(2016))。

バギング(bagging)は、CART を発展させた合成データの生成手法であり、ブートストラップ法を用いて決定木を複数作成し、それらを総合するアンサンブル学習を行う技法である。バギングは、機械学習の分野では一般に、CART に比べて過学習を抑えることができ、汎化性能に優れているとされる。ランダムフォレスト(random forest)はバギングの一種であり、ブートストラップ法に加え、使用する属性の組も変化させて決定木を複数作成し、アンサンブル学習を行う手法である。ランダムフォレストは、バギングよりもさらに汎化性能に優れているため、機械学習の分野ではより実践的な手法として知られている。

ランダムサンプリング(random sampling)は、復元抽出による無作為抽出を行う。属性ごとに独立にサンプリングを行うため、属性間の関係性のほとんどは失われることになる。最後に回帰(regression)は、回帰分析を用いて合成データを生成する手法である。本実験では、質的属性には多項ロジスティック回帰が、量的属性には線形回帰が用いられた[17]。

6.2.　合成データの生成手法の有効性に関する定量的な評価―攪乱的手法との比較

図 6-1 と図 6-2 はそれぞれ、合成データ生成手法である CART、バギング、ランダムフォレスト、ランダムサンプリング、回帰、また比較対象として攪乱的手法であるミクロアグリゲーション技法の 1 つである MDAV 法のそれぞれについて、R-U マップを用いて比較・検証を行ったものである。図 6-1 では、攪乱的手法の時と同様に、横軸は層化平均 ARD で秘匿性を、縦軸は pMSE で有用性を表している。決定木手法である CART、バギング、ランダムフォレストについては、最小リーフサイズを 3、5、10、30、50、100 と変化させてプロットした。最小リーフサイズとは決定木の末端である葉の大きさの制約であり、これが 3 であれば細かく枝の分割が行われるため原データの分布特性により近似的となる。決定木手法やランダムサンプリングには乱数発生に伴うばらつきが生じるため、手法ごとに 10 回ずつ合成した上でその平均値をプロットした。ミクロアグリゲーションについては前実験と同様に、レコード数 k を 3 から 100 まで段階的に変化させてプロットした。

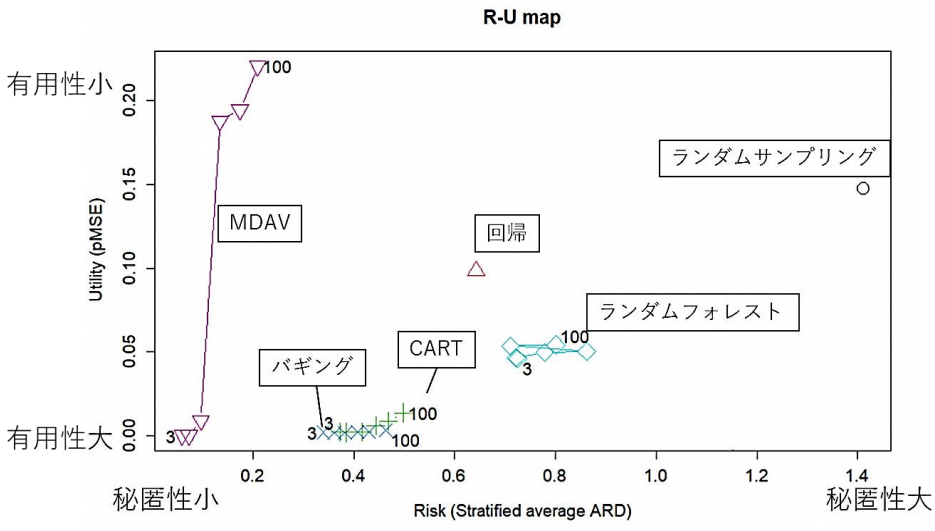

図 6-1 攪乱的手法と合成データ生成手法に関する秘匿性(ARD)と有用性(pMSE)の評価結果

[17] synthpop では、属性の性質ごとに回帰分析手法を使い分けることができる。デフォルト設定では、量的属性には線形回帰、質的属性(名義尺度、2 値)にはロジスティック回帰、質的属性(名義尺度、3 値以上)には多項ロジスティック回帰、質的属性(順序尺度)には順序多項ロジスティック回帰が適用される(Nowok et al. (2016))。

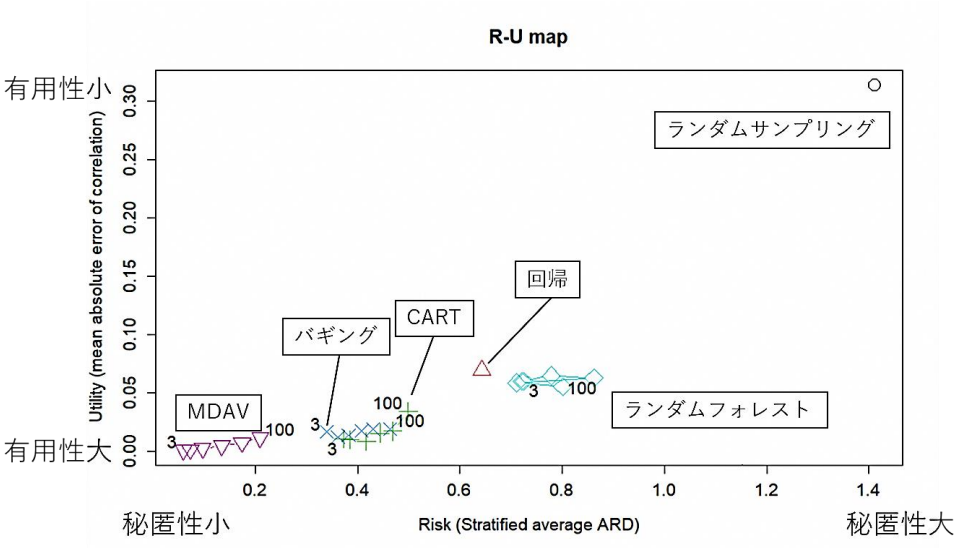

図 6-2 攪乱的手法と合成データ生成手法に関する秘匿性(ARD)と有用性(相関係数行列の差の平均絶対誤差)の評価結果

前実験と同様に、合成データにおいても秘匿性と有用性のトレードオフの関係性が表れている。まず、比較対象として用意した攪乱的手法である MDAV 法は、合成データと比較してもやはり秘匿性が小さいエリアに位置している。それに対して、ランダムサンプリングで合成した場合は最も秘匿性が大きく、有用性の低下も相対的に大きいことがわかった。回帰分析を用いた場合は、その中間的な位置にプロットされている。決定木手法である CART とバギングについては、ほぼ同様の結果が得られている。これは、ブートストラップ法を用いているかどうかの違いのみに起因しており、基本的には類似した決定木が生成されやすいからだと考えられる。いずれも有用性の低下は相対的に小さいが、秘匿性の数値は、k の値が小さい場合(3, 5, 10)の MDAV 法に比べて改善している[18]。決定木手法のうちランダムフォレストについては、CART やバギングと比較して、秘匿性は大きいが有用性は小さいという結果が得られた。また最小リーフサイズの変更に対して秘匿性や有用性の値が相対的にやや不安定になるという特徴が見られた。これについては、CART やバギングと違い、ランダムフォレストは決定木の作成に使用する属性の種類に変化が表れることが考えられる。使用する属性の数や組を変化させ、それらを総合して最終的な決定木を作成することから、定量的な評価の結果にもばらつきが生じやすい。ランダムフォレストは一般に汎化性能に優れているとされるが、合成データ作成においては未知のデータに対する予測力を高めることが目的ではないことを留意する必要がある。

つぎに、図 6-2 では、図 6-1 と同様に、横軸の秘匿性については層化平均 ARD を、縦軸に関しては有用性として相関係数行列の絶対誤差でそれぞれ表している。層化平均 ARD を用いた上図に比べて、トレードオフの関係性が顕著に表れている。MDAV 法、CART およびバギング、回帰、ランダムフォレスト、ランダムサンプリングの順に秘匿性が大きくなってい

[18] 層化平均 ARD は原データと合成データにおける層ごとの最大値の乖離を測る指標であるため、最大値以外のレコードリンケージに伴うリスクまでは考慮していない。合成データにおいてはこのリスクは非常に低いが、匿名化ミクロデータにおいてはレコードが特定されるリスクは相対的に大きくなる。この点を踏まえると、層化平均 ARD のみを見ている本実験の結果以上に、合成データの秘匿性の強度は高く、原データに含まれる個体を推定することが困難であると判断できる。

る。有用性については、MDAV 法、CART およびバギング、ランダムフォレスト、回帰、ランダムサンプリングの順になっている。注目すべき点としては、①層化平均 ARD の R-U マップと同様に、CART やバギングであれば有用性を保ったまま秘匿性を高められる可能性があること、②回帰に比べてランダムフォレストが秘匿性も有用性も大きくなっている点が指摘される。

以上の結果から、攪乱的手法、さらには回帰分析を用いた合成手法と比較して、決定木に基づく合成データの生成手法の有効性が示唆された。

7. CART によって作成された合成データの分布特性

本節では、前節における定量的な評価に関する実験にもとに、有望と判断できた合成データ生成手法の 1 つである CART(最小リーフサイズは 10 に設定)を用いて、合成データを生成した。その要約統計量および量的属性間の相関係数をそれぞれ表 4、表 5 に示す[19]。要約統計量は概ね再現されているが、mean(平均値)や at_99%(99%点)は属性によっては 10%近くの乖離が見られるものもあった。分布の歪みが大きいことから、平均値が合成されたデータにおける右裾の分布特性に大きく依存することが考えられる。逆に、そういった影響が相対的に小さいmedian(中央値)やat_1%(1%点)については、比較的原データに近い結果が得られた。相関係数についてはどの属性も概ね再現できており、差異はわずかであった。

表 4　要約統計量(上：原データ、下：CART で合成したデータ)

変数名	n	mean	sd	median	at_1%	at_99%
売上（収入）金額	10,000	53,768	427,885	3,500	0	859,421
付加価値額	10,000	10,640	55,503	1,459	-823	154,092
給与総額	7,618	2,430	6,034	600	0	24,622
減価償却費	7,618	371	1,607	34	0	5,847

変数名	n	mean	sd	median	at_1%	at_99%
売上（収入）金額	10,000	60,209	541,673	3,500	0	880,803
付加価値額	10,000	11,324	66,158	1,441	-738	170,776
給与総額	7,536	2,427	6,320	577	0	28,649
減価償却費	7,622	389	1,699	36	0	6,023

表 5　相関係数(上：原データ、下：CART で合成したデータ)

	売上（収入）金額	付加価値額	給与総額	減価償却費
売上（収入）金額	1.00	0.60	0.81	0.62
付加価値額	0.60	1.00	0.52	0.33
給与総額	0.81	0.52	1.00	0.50
減価償却費	0.62	0.33	0.50	1.00

	売上（収入）金額	付加価値額	給与総額	減価償却費
売上（収入）金額	1.00	0.59	0.81	0.63
付加価値額	0.59	1.00	0.53	0.35
給与総額	0.81	0.53	1.00	0.50
減価償却費	0.63	0.35	0.50	1.00

[19] 要約統計量は平均や標準偏差の値が直感的に理解しやすいように neg-log 変換前の値に戻して表示した。相関係数は neg-log 変換後の値を用いて求めている。

さらに、図 7 は、質的属性(リコーディング済の従業者規模、資本金額を含む)、量的属性の原データと CART で生成した合成データの構成比を表示している。図中の濃い棒線が原データを、薄い棒線が合成データをそれぞれ示している。いずれの属性においても構成比の違いはほとんどなく、原データの分布を再現できていることがわかる[20]。特筆すべき点としては、0 や未記入も概ね原データと同じような分布特性を示していることである。給与総額や減価償却費には原データに 0 や未記入が多く含まれるが、生成された合成データはこの特徴も含めて原データの分布を十分に再現できている。回帰を用いて合成データを生成する場合、0 や未記入の取り扱いが論点になりうるが、そういった点を考慮することなく合成データが生成可能であることは、CART の利点であると言える[21]。

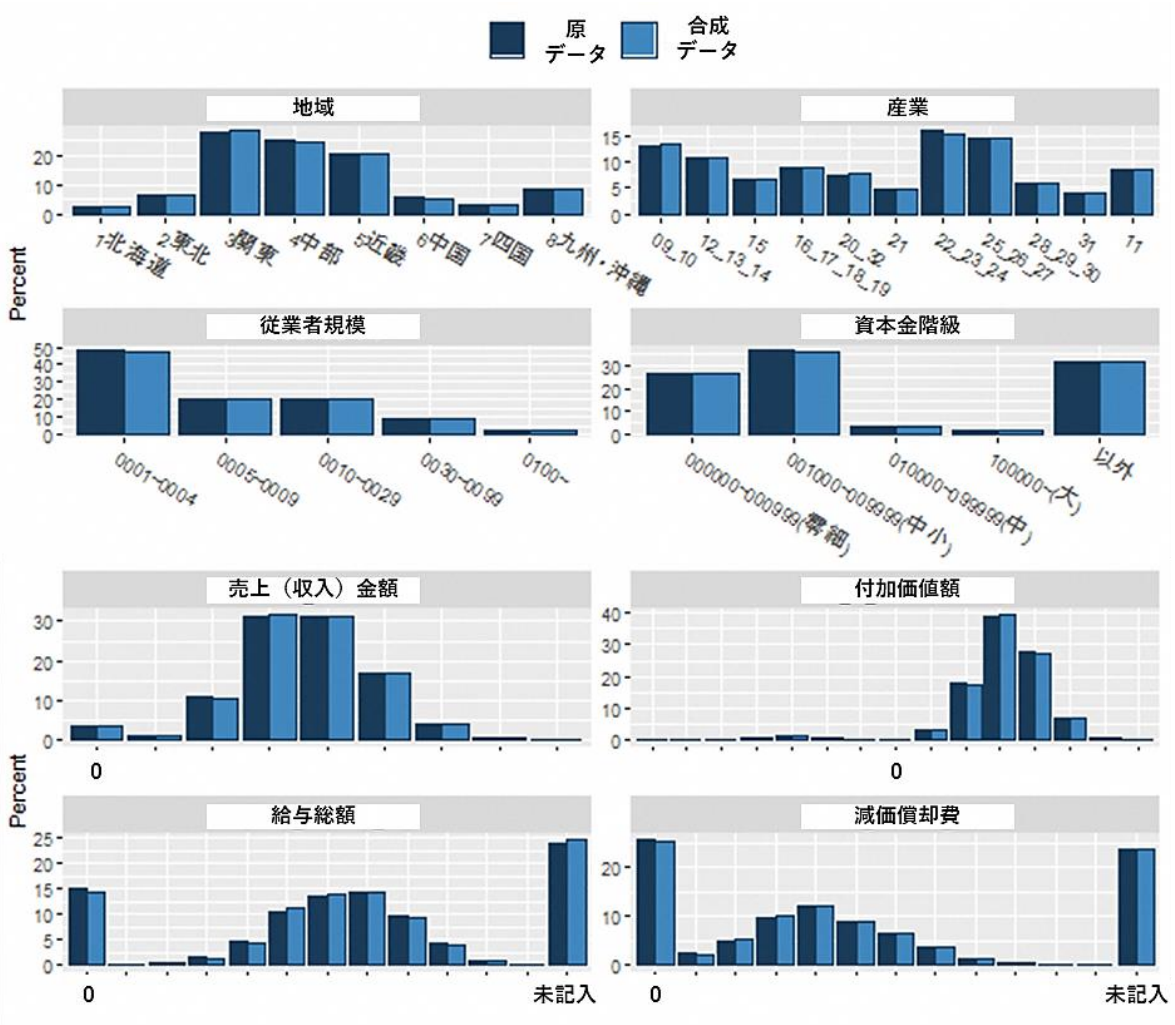

図 7　属性ごとの構成比の比較(濃：原データ、薄：CART で合成したデータ)

[20] Raab et al.(2021)は、合成データの有用性評価指標として、pMSE を標準化した比率 (standardized pMSE ratio = SpMSE) を用いた比較も推奨している。予備的に行った実験では、いずれの属性においても SpMSE は目安となる閾値 10 以下であり、有用性の損失の許容範囲であるという結果になった。

[21] 実務への適用を考える場合、そもそも合成データに未記入となっている空欄を残すべきかどうかという論点が存在する。仮に未記入の空欄を 0 や補完値に置き換える必要がある場合でも、置換を行った後に CART を適用することは容易である。

8. むすびにかえて

本稿では、海外とわが国における合成データの生成手法をめぐる議論を概括した上で、経済センサス活動調査の個票データをもとに事業所・企業系の合成データの生成手法の有効性を評価した。本稿においては、最初に、経済センサスのデータ特性を確認した上で、攪乱的手法および合成データ生成手法を共通の指標で定量的に評価することを目指し、秘匿性については層化平均 ARD を、有用性については傾向スコアに基づく pMSE や相関係数行列の差の絶対誤差の概要や特徴を明らかにした。次に、各種の攪乱的手法が適用された匿名化ミクロデータの定量的な評価を行い、秘匿性と有用性の観点から攪乱的手法の比較・検証を行った。さらに、回帰や決定木に基づく各種の合成データ生成手法を用いて、攪乱的手法の場合と同様の秘匿性と有用性の評価指標を用いて検証を行った。その結果から、代表的な決定木手法である CART やバギングであれば、攪乱的手法であるミクロアグリゲーションの MDAV 法と比較して、有用性を維持したまま秘匿性を高められる可能性があることを確認した。最後に、CART を用いて合成データを生成した上で、合成データでも要約統計量や相関係数、属性ごとの構成比を再現可能であることを実証的に明らかにした。

本研究の結果は、わが国の公的統計に対する各種の合成データの生成手法の可能性を示したものと言えるが、合成データの実務への適用可能性を追究しようとすれば、統計法を踏まえた形での作成方法の追究が今後の課題になると考えられる。本研究では、個票データ(調査票情報)をもとにノンパラメトリックな生成手法や機械学習の技法を用いて合成データの生成を行ったが、個票データに対して直接合成データの方法論を適用した場合、現行の統計法の下では、作成されたデータは個票データに対して「加工」が施された匿名データ(法 2 条第 12 項)に該当するため、その作成や提供については、統計法制度において匿名データとしての制約が課せられるだけでなく、適用された合成データの生成手法も、匿名化技法の 1 つとみなされる(伊藤(2022))。

こうしたことから、合成データの作成方法を工夫し、個票データに直接的に合成データの方法論を適用することに替えて、個票データとは異なるように作成されたミクロアグリゲートデータ(個票データに対してミクロアグリゲーションが適用されたデータ)(伊藤(2009))に対して合成データを生成することが考えられる。そのような方法に基づく合成データの生成が可能になれば、例えば、わが国において匿名データが作成されていない事業所・企業系の統計調査について、個票データの分析で使用するプログラムコード作成用のテストデータへの展開も期待できる。これについては、ミクロアグリゲーションのさらなる可能性を追究する観点から、別稿の課題となる。また、CART、バギングやランダムフォレストのような手法を適用しようとすれば、わが国の統計法の下では、合成データの作成・提供可能性の観点から、データ生成に関するモデル(決定木)を事前に公開することが求められる可能性もある。このとき、事前に公開すべき決定木の数が多くなる場合における統計実務レベルでの対処方法も、さらなる検討課題となるだろう。さらに、本研究は限られたレコードや属性を含むテストデータを用いて検証を行っているが、レコード数や属性の数が増大した場合に、CART 等の合成データ生成手法が同様の形で適用可能かどうかについての検討も、統計実務の観点から必要になると考える。これについても今後の研究課題としたい。

謝辞

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

たします。

参考文献

[1] 伊藤伸介 (2009)「匿名化技法としてのミクロアグリゲーションについて」熊本学園大学『経済論集』第 15 巻第 3・4 号合併号, pp.197-232.
[2] 伊藤伸介, 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.
[3] 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.
[4] 伊藤伸介 (2019)「公的統計データにおける秘匿性と有用性の評価のあり方に関する一考察―スワッピングを中心に―」, 坂田幸繁編『公的統計情報―その利活用と展望』中央大学出版部, pp.39-62.
[5] 伊藤伸介 (2022)「ミクロデータの匿名化と統計情報の秘匿可能性について」『経済学論纂(中央大学)』第 63 巻 1・2 合併号, pp.1-23.
[6] 伊藤伸介, 寺田雅之 (2020)「詳細な地域データにおける秘匿処理の適用可能性について」『日本統計学会誌』, 第 50 巻第 1 号, pp. 139-166.
[7] 高部 勲 (2017)「状態空間モデルに基づく季節調整法における改良方法の提案：一般化 neg-log 変換の活用に基づくゼロ・負の値を含む時系列データの安定化と季節調整値の推定精度向上」『統計研究彙報』第 74 号, pp.29-56.
[8] 高部 勲 (2022)「合成データの考え方に基づく公的統計疑似ミクロデータの作成方法の検討」『統計研究彙報』第 79 号, pp.111-130.
[9] 千田 浩司, 南 和宏, 寺田 雅之, 伊藤 伸介 (2022)「プライバシー保護型合成データの実用動向と今後の展望」『統計』2022 年 8 月号, pp.35-42.
[10] 南 和宏 (2022) 「プライバシー技法の動向と公的統計制度に求められる対応」『統計』2022 年 8 月号, pp.11-16.
[11] 山口幸三, 伊藤伸介, 秋山裕美 (2013)「教育用擬似ミクロデータの作成―平成 16 年全国消費実態調査を例として―」,『統計学』104 号, pp.1-15.
[12] 横溝 秀始, 伊藤 伸介 (2022)「事業所・企業系のミクロデータにおける匿名化措置の有効性の評価　－経済センサス - 活動調査を例として－」『統計研究彙報』第 79 号, pp.151-170.
[13] Bates, A. G., Špakulová, I., Dove, I., Mealor, A. (2019) “ONS methodology working paper series number 16 - Synthetic data pilot” https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber16syntheticdatapilot
[14] 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.
[15] Breiman, L., J. H. Friedman, R. A. Olshen, & C. J. Stone (1984). Classification and Regression Trees. Belmont, CA: Wadsworth.
[16] Chien, Chien-Hung, Alan Hepburn Welsh, & John D Moore. (2021). “Synthetic Business Microdata: An Australian Example”. Journal of Privacy and Confidentiality 10 (2).
[17] Domingo-Ferrer, J. & Torra, V. (2001). “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.

[18] 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.

[19] Drechsler, J. & Reiter, J.P. (2009). “Disclosure risk and data utility for partially synthetic data: an empirical study using the German IAB Establishment Survey”. Journal of Official Statistics, 25(4), 589-603.

[20] Drechsler, J. (2011). Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation, Springer.

[21] 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.

[22] Hang J. Kim, Jörg Drechsler, & Katherine J. Thompson. (2021). "Synthetic microdata for establishment surveys under informative sampling," Journal of the Royal Statistical Society Series A, Royal Statistical Society, vol. 184(1). pages 255-281, January.

[23] 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.

[24] Little, R. J. A. (1993). “Statistical Analysis of Masked Data”, Journal of Official Statistics Vol. 9, pp.407-426.

[25] Nowok, B., Raab, G. M., & Dibben, C. (2016). synthpop: Bespoke Creation of Synthetic Data in R. Journal of Statistical Software, 74(11), 1–26. https://doi.org/10.18637/jss.v074.i11

[26] Raab, G. M., Nowok, B., & Dibben, C. (2021). Assessing, visualizing and improving the utility of synthetic data. arXiv preprint arXiv:2109.12717.

[27] Reiter, J. P. (2005). “Using CART to Generate Partially Synthetic, Public Use Microdata”, *Journal of Official Statistics* Vol.21, pp.441-462.

[28] Rubin, D. B. (1993). “Discussion: Statistical Disclosure Limitation”, *Journal of Official Statistics*, Vol. 9, pp.462-468.

[29] Snoke, J., Raab, G.M., Nowok, B., Dibben, C., & Slavkovic, A.B. (2016). General and specific utility measures for synthetic data. Journal of the Royal Statistical Society: Series A (Statistics in Society). 181.

[30] Taub, J., Elliot, M., Pampaka, & M., Smith, D. (2018). Differential Correct Attribution Probability for Synthetic Data: An Exploration. In: Domingo-Ferrer, J., Montes, F. (eds). Privacy in Statistical Databases. PSD 2018. Lecture Notes in Computer Science. (), vol 11126. Springer, Cham.

[31] Taub, J., Elliot, M., & Sakshaug J. W. (2020). The impact of synthetic data generation on data utility with application to the 1991 UK samples of anonymised records. Transactions on Data Privacy, 13(1):1–23, 2020. ISSN 20131631.

[32] Templ, M. (2017). Statistical Disclosure Control for Microdata: Methods and Applications in R, Springer International Publishing.

[33] 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.

[34] Woo, M.-J., J. P. Reiter, A. Oganian, & A. F. Karr (2009). Global measures of data utility for microdata masked for disclosure limitation. Journal of Privacy and Confidentiality 1, 111–124.