統計数理(2024)
第 72 巻 第 2 号 217–231

特集「公的統計―社会の重要な情報基盤―」
[原著論文]

事業所・企業系の公的統計を対象にした合成データの生成技法に関する検討

―経済センサスを例として―

伊藤 伸介[1]・横溝 秀始[2]

(受付 2023 年 12 月 31 日;改訂 2024 年 6 月 10 日;採択 7 月 2 日)

要　　旨

わが国の事業所・企業系の統計調査においては,事業所や企業を対象にした匿名データが現在作成されていないだけでなく,事業所・企業系の統計調査の場合,一般公開型ミクロデータの作成が困難である.そのため,合成データが作成可能であれば,テストデータ等へのニーズに応えることが可能になる.そこで,本稿では,経済センサス活動調査の個票データを用いて,各種の合成データの生成技法について定量的な評価を行った.

本研究では,攪乱的手法であるミクロアグリゲーションの MDAV(=Maximum Distance to Average Vector)法,CART(=Classification And Regression Tree)等,さらには深層学習モデルの 1 つである CTGAN(=Conditional Tabular GAN)も用いて生成された合成データの有用性および秘匿性について定量的な評価を行った.本研究において CART を用いて合成データを生成した場合,要約統計量や相関係数といった分布特性が再現可能であることが確認された.また,CART は MDAV 法と比較して,有用性を保ったまま秘匿性の強度が高まる可能性がある.さらに,CTGAN については,CART と比較した場合,秘匿性の程度がより高くなっていることがわかったが,有用性の低下も相対的に大きいことが確認された.

キーワード:合成データ,ミクロアグリゲーション,CART,CTGAN,経済センサス.

1. はじめに―事業所・企業系のミクロデータの現状

わが国の公的統計においては,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.それに対して,事業所・企業系の統計調査に関しては,常用労働者を対象に匿名データが作成・提供されている賃金構造基本統計調査の事例はあるものの,調査客体である事業所や企業に匿名化技法を適用することによって作成された匿名データは,わが国では存在しない.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外となっている.このように,わが国では,事業所・企業系の統計調査においては,個票データ(調査票情報)のみの利用が可能になっている.

それに対して,教育目的のためのミクロデータの利用に対するニーズが高まっていることや,

[1] 中央大学 経済学部:〒 192–0393 東京都八王子市東中野 742–1
[2] 総務省 統計研究研修所:〒 162–8668 東京都新宿区若松町 19–1

オンサイト施設やリモートアクセスのようなセキュアな環境における個票データの利用に関して，プログラムコードのチェックを行うためのテストデータへのニーズが少なくないこともあり，海外では近年合成データ(synthetic data)に対する社会的関心が高まっている．なお，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属性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（Templ, 2017, p.157)と定義される．

事業所・企業系の統計調査の場合，一般公開型ミクロデータ(public use microdata)の作成が困難であることから，合成データを作成することができれば，テストデータ等へのニーズに応えることが可能になるように思われる．そこで，本稿は，経済センサスの個票データを用いて，事業所・企業系の統計調査の合成データの生成技法を定量的に評価する．なお，本研究においては，伝統的な匿名化技法としての攪乱的手法を用いて作成された匿名化ミクロデータと機械学習の方法論等を用いて生成した各種の合成データを比較・検証する．そのために，ミクロデータに対する秘匿性や有用性に関する評価指標に基づいて，合成データの定量的な評価を行う．

2. 公的統計における合成データ作成の現状

本節では，最初に海外の公的統計に関する合成データ生成の研究事例を紹介するだけでなく，公的統計における合成データの生成技法の比較検証に関する研究についても概括する．

2.1 公的統計における合成データ生成に関する海外の現状

海外における合成データの作成状況あるいは合成データの作成に関する研究事例については，例えば以下の事例を指摘することができる．第 1 は，Eurostat の事例である．Eurostat では，EU-SILC(=European Union Statistics on Income and Living Conditions)に対して，統計的なモデルを用いたシミュレーションによる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)が作成されている（伊藤, 2018)．

第 2 は，エディンバラ大学による事例である．エディンバラ大学では，スコットランド縦断調査(Scottish Longitudinal Study=SLS)を対象に，合成データの作成が行われてきた．R の合成データ作成用のパッケージである synthpop（Nowok et al., 2016)を用いることによって，人口センサスの縦断的な合成データが生成されている．

第 3 は，イギリス国家統計局(Office for National Statistics=ONS)による事例である．ONS では，労働力調査(Labour Force Survey)のミクロデータをもとに，synthpop を含む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（Bates et al., 2019)．

第 4 は，オーストラリアによる事業所・企業用の合成データの研究の事例である（Chien et al., 2021)．産業によって市場が寡占あるいは複占という特徴を有するため，事業所・企業系の統計調査においては，情報の削減や攪乱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がある．そこで，攪乱的手法との比較の結果，秘匿性と有用性を勘案した場合に，合成データの方法論が事業所・企業系のミクロデータに適用可能な手法として議論されている．

UNECE（2022)では，海外の統計作成部局における合成データの作成に関する最近の研究の事例が紹介されている．例えば，ニュージーランド統計局は，数理モデルによって生成された合成データを Synthetic Unit Record Files(SURFs)として展開している．また，2007 年所得調査に基づく SURFs および，2019 年の世帯貯蓄調査と人口センサスに基づき生成された ‘Census for School’ SURF が提供されている．つぎに，カナダ統計局は，センサスベースの合成

表 1．合成データのユースケースごとの推奨技法．UNECE（2022），表 7 をもとに作成．

手法の分類	手法	例	PUF / 分析テスト用	教育用	プログラムテスト用	備考
Sequential modelling	Fully Conditional Specification (FCS)	CART, Bagging, Random forest	○推奨	△可（作成コスト大）	△可（用途に対して高度すぎる）	元データ属性間の関係性が保持される
	Information Preserving Statistical Obfuscation (IPSO)	回帰	△全属性が線形回帰で表現できるなら推奨、以外は非推奨	△全属性が線形回帰で表現できるなら推奨、以外は非推奨	△可（用途に対して高度すぎる）	特に線形回帰に関連する結果と統計量が保持される
Simulated Data	Dummy files	-	×非推奨	△（分析を目的としないなら）	○推奨	分析には適さないが簡単かつ迅速
	Analytically advanced simulated data	一般用ミクロデータ	△想定する分析手法に対しては推奨、以外は非推奨	△想定する分析手法に対しては推奨、以外は非推奨	△可（用途に対して高度すぎる）	限られた統計量のみ保持される
	Pseudo Likelihood	-	△元の有限母集団から推定したい場合は推奨、合成ウェイトを期待する場合は非推奨	△可（作成コスト大）	△可（用途に対して高度すぎる）	元データ属性間の関係性が保持される
Deep Learning	Generative Adversarial Network (GAN)	CTGAN, Table GAN	△テキストまたは非構造化データが存在する場合は推奨	△可（作成コスト大）	△可（用途に対して高度すぎる）	非構造化データとテキストデータを処理する唯一の手法

データの生成に関する研究を行っている．具体的には，カナダの退職所得システム（Canadian retirement and income system）に関する動態的なミクロシミュレーションモデルの初期母集団のための研究が展開されている．オーストラリア統計局では，機械学習モデルを検証するためにミクロシミュレーションモデルに基づく合成データの生成が紹介されている．さらに，イギリス国家統計局では，2017 年に ONS 内部に設立された ONS Data Science Campus において，敵対的生成ネットワーク（Generative Adversarial Network=GAN）（Goodfellow et al., 2014）を用いた合成データを生成することが追究されている．

2.2　公的統計における合成データの生成技法の比較検証に関する研究

本節では，公的統計を対象にした合成データの生成技法の比較検証の事例を紹介したい．

UNECE（2022）では，合成データ生成技法やユースケースに応じた推奨技法が，体系的に整理されている（表 1）．合成データ生成技法は，逐次的なモデル化（Sequential modelling），シミュレーションによるデータ生成（Simulated Data），深層学習（Deep Learning）の 3 種類に大別される．逐次的なモデル化は，同時確率分布を設定した上で，逐次的に属性を合成する手法全般を表しており，属性間の関係性を保持しやすいという特徴がある．本研究でも実験に用いるノンパラメトリックな決定木手法である CART（= Classification And Regression Tree）はこれに含まれる．シミュレーションによるデータ生成は，対象となる元データの分布特性に関する情報が未知の場合に，合成データをシミュレーションで生成する手法全般を包含している．わが国の一般用ミクロデータは，その中の Analytically advanced simulated data に該当すると思われる．さらに，深層学習は GAN 等の手法を用いた生成技法であり，近年注目されている手法である．本研究で実験に用いる条件付き表形式 GAN（Conditional Tabular GAN=CTGAN）は，非構造化データやテキストデータを処理する唯一の方法であり，深層学習の一種である．

いずれの技法にもそれぞれ一長一短があり，原データの性質や用途に適した技法を選択することが重要である．

Taub et al.（2019）では，“The Synthetic Data Challenge” という形で，複数の研究チームが 1901 年スコットランドの歴史的な人口センサスを対象にした合成データの試行的な生成が追究されている．具体的には，①地域で層化を行った上での CART の適用，②分位点回帰やロジスティック回帰の適用，③ランダムサンプリング，④χ^2 統計量に基づく有向グラフ等，様々な生成方法を用いて作成された合成データを対象に，有用性と露見リスクの両面から検証が行われており，全体的に，有用性が高い合成データであるほど，露見リスクも相対的に高まるという結果が得られている．

また，Little et al.（2021）は，1991 年イギリス人口センサスの匿名化標本データ（Samples of Anonymised Records）を用いて，synthpop を用いて行った CART，DataSynthesizer を用いたベイズアプローチ，深層学習モデルと位置付けられる 2 つのタイプの GAN（CTGAN および TableGAN）を比較・検討を行っている．その結果，synthpop を用いた CART は，テストした 4 つの手法の中で最も実用性が高いが，個体が特定されるリスクも高いことが確認できた．それに対して，Table GAN における個体特定リスクは最も低いが，有用性が低いことが確認されている．

その一方で，横溝・伊藤（2023）においては，経済センサス活動調査の個票データを用いて，匿名化手法と合成データ生成技法の比較・検討を行っている．本研究では，ミクロアグリゲーションといった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・検証を行っただけでなく，回帰や決定木に基づく各種の合成データ生成技法を用いて，秘匿性と有用性の評価指標による検証を行った．本研究の結果から，CART については，ミクロアグリゲーション技法と比較して，有用性は保持された状態で，相対的に秘匿性が高いことが実証的に確認されている．さらに，伊藤・横溝（2024）では現行の統計法の下における合成データの展開可能性を指向しており，個票データに直接 CART を適用した結果とミクロアグリゲートされたデータに CART を追加的に適用した結果を比較・検証した上で，有用性と秘匿性の両面において大きな差が生じないことを確認している．なお，本研究は，横溝・伊藤（2023）に基づいて，合成データの生成技法としての CTGAN の適用可能性を追究している．

3. 使用するデータ

本研究で使用するデータは，「平成 28 年経済センサス-活動調査」（以下「経済センサス」と略称）の個票データである．テストデータにおいては，横溝・伊藤（2023）と同様に，製造業を対象とした従業者規模 1 人以上 1,000 人未満である 10,000 事業所が含まれており，このテストデータから 10,000 レコードを合成する．本研究で分析のために用いる属性群については，質的属性が，地域（8 区分），産業（11 区分），従業者規模（5 区分）と資本金階級（5 区分）であり[1)]，量的変数が，売上（収入）金額，付加価値額，給与総額と減価償却費である．

ところで，経済センサスのような事業所・企業系の統計調査においては，従業者規模や資本金が大きい企業および事業所が調査対象として常に含まれることが少なくない．こうした企業や事業所の場合には属性の多くが外れ値を有することから，相対的に分布の歪みが大きくなるだけでなく，入手可能な外部情報とのマッチングによって，個体情報が露見されるリスクが高くなる．また，量的変数間の相関性が高いことも事業所・企業系の統計調査における特徴と言える．

表 2 と表 3 はそれぞれ，使用する量的属性の基本統計量と相関係数行列を示したものであ

表 2. 本分析で使用する量的変数の基本統計量．出所　横溝・伊藤（2023）．

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
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
付加価値額	10,000	10,640	55,503	1,459	-823	154,092

表 3. 本分析で使用する量的変数の相関係数．出所　横溝・伊藤（2023）．

	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

る．量的変数においては，平均値と中位数の数値が大きく異なるため，分布の山のピークが中心の左側に位置し，右側の裾が大きくなるだけでなく，給与総額や減価償却費では 0 の値が存在しており，付加価値額では負の値も含まれることがわかる．そこで，本研究では，量的変数に関しては，neg-log 変換によって 0 や負の値も含めて対数化を行った[2)]．それによって，分布の歪み，さらには 0 や負値の問題を解決することが可能になる．例えば売上(収入)金額について neg-log 変換を行うと，概ね対数正規分布に近い分布に変換される．また，付加価値額の場合，正の山と負の山がそれぞれ現れることが確認された．表 3 の相関係数行列についても売上(収入)金額と給与総額では，0.8 という高い値を示している．

4. 合成データにおける秘匿性と有用性の評価方法について

先述のように経済センサスにおいては，外れ値(特異値)のような形でレコードが存在する可能性があるため，対象となるレコード群の中の個体が特定されるリスクが存在するだけでなく，合成データにおいて個体の属性情報が推定される可能性が指摘される．そこで，秘匿性に関する評価指標をもとに，合成データにおける露見リスクを評価することが求められる．また，合成データの利用可能性を追究しようとすれば，原データに対する合成データの有用性の定量的な評価を検討する必要がある．

本研究では，横溝・伊藤（2023）と同様に，絶対相対差分(Absolute Relative Difference=ARD)と呼ばれる属性漏洩(attribute disclosure)に関する評価指標（Kim et al., 2021）を用いた検証を行う．ARD は原データと合成データに含まれる属性値の最大値同士の乖離を評価する指標であり，以下の(4.1)式で表される．

$$\mathrm{ARD} = \frac{|\hat{L} - L|}{L} \tag{4.1}$$

L：原データに含まれる属性値の最大値

$\hat{L}$：合成データに含まれる属性値の最大値

合成データに対して侵入者が取りうる攻撃手段の 1 つは，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．例えば，地域と産業というキーとなりうる属性のクロスで考えた場合，攻撃者はその地域と産業の層に含まれるレコードの売上の最大値を調べることで，その地域において該当する産業の大規模な事業所に関してセンシティブな属性(売上)を推測できる可能性がある．そこで本研究では，横溝・伊藤 (2023) や伊藤・横溝 (2024) と同様に，キー変数の全ての組み合わせのそれぞれについて原データと合成データの最大値の乖離を計算し，その平均値を用いて算出する，層化平均 ARD(stratified average ARD)を秘匿性の評価指標として使用した．

つぎに本研究では，合成データにおける有用性の評価指標については，傾向スコア(propensity score) (Woo et al., 2009)を用いた評価指標の 1 つである，傾向スコア平均二乗誤差(propensity score Mean Square Error=pMSE) (Snoke et al., 2016)を有用性に関する定量的な指標と設定した上で実証分析を行った．なお，pMSE は以下の(4.2)式で表される．

$$\text{pMSE} = \frac{1}{N}\sum(\widehat{p_i} - c)^2 \tag{4.2}$$

N：原データのレコード数と合成データのレコード数の和

$\widehat{p_i}$：各レコードの傾向スコア

c：N に占める合成データのレコード数の割合

本研究でも，横溝・伊藤 (2023)および伊藤・横溝 (2024)と同様に，最初に原データと合成データに含まれるレコード群を統合し，つぎに「合成データか否か」を被説明変数として設定した上で，ロジット回帰モデル等を用いて，各レコードについて合成データである可能性がどの程度高いかを確率的に表す傾向スコア p_i を算出する．そして，傾向スコア p_i とレコード全体に占める合成データの比率を表す c との乖離の差の平均値として，pMSE を計測する．

また本研究では，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 and Torra, 2001; 伊藤 他, 2014)も有用性に関する指標として使用した．なお，相関係数行列の差の平均絶対誤差は，以下の(4.3)式で表される (伊藤 他, 2014)．

$$\frac{\sum_{j=1}^{k}\sum_{1\le i\le j}|r_{ij} - r'_{ij}|}{\frac{k(k-1)}{2}} \tag{4.3}$$

k：属性の数

r：原データの相関係数

r'：合成データの相関係数

5. 合成データの生成技法の有効性に関する比較研究

本研究は，経済センサスの個票データを例に，匿名化技法として用いられる攪乱的手法だけでなく，各種の合成データの生成技法を用いて，その有効性についての比較実験を行う．本実験における攪乱的手法については，ミクロアグリゲーションの 1 つである MDAV 法を用いる．また，本研究で用いる合成データの生成技法は，CART，バギング(Bagging)，ランダムフォレスト(Random Forest)，および深層学習モデルの 1 つである CTGAN である．

ミクロアグリゲーションは，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攪乱的手法 (Defays and Nanopoulos, 1993; 伊藤, 2009)である．その 1 つである MDAV(=Maximum Distance to Average Vector)法 (Domingo-Ferrer and Mateo-Sanz, 2002; Hundepool et al., 2003)は，対象レコードにおける平均値からの距離を考慮

した上でレコード間の近似性を最大にするように，平均ベクトルからの距離が最大となるレコードから優先的にグループ化を行う技法である．MDAV 法は，以下のような手順で適用される（横溝・伊藤, 2023）．それは，①対象となるレコード群について属性値の各々に関する平均値のベクトルを算出し，②平均からの距離が最大の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探索し，③しきい値 k を設定した上で，対象となるレコードを含む近傍の k 個のレコードをグループ化して平均値に置換し，④ミクロアグリゲート済（攪乱済み）のレコードを除去した後に，②と③の処理を繰り返すというものである．

つぎに，CART（Classification And Regression Tree）（Breiman et al., 1984）は，観測済みの属性値から目的変数となる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の手法である．CART の特徴として，質的属性と量的属性の両方においての適用可能な技法であることが指摘できる．

CART のメリットとして，決定木の作成における枝の分岐や決定木の末端である各々の葉の分布が可視化できることが指摘される．また，他の手法と比較して，適用された結果の解釈が容易であることから，実務への適用の観点から，幅広い分野での利用が可能である．さらに．葉のサイズに対して制約を与えることによって，枝の分割の細かさを制御できることから，CART によって作成された合成データの有用性と秘匿性の程度を調整することもできる．

図 1 は，CART を用いて，地域に基づいて産業を合成した場合の決定木の例を示している．「CHIKI8」は 8 区分にリコーディングした地域の分類区分を表しており，「09_10」は，産業中分類 “09”（食料品製造業）と “10”（飲料・たばこ・飼料製造業）をリコーディングした分類区分に対応している．本図においては，最初に原データを対象に，地域に基づいてそれぞれの葉が異なる分布を持つような産業の決定木を作成する．その場合の枝の分割の基準としては，ジニ係数といった指標が一般に用いられる．次に，合成の対象となるレコードの地域を用いて決定木を辿り，その先にある分布を用いて産業をランダムに合成する．図 1 では，例えば左端の Node2 は，クラスタリングの結果 2069 レコードの産業で葉が構成されている（横軸はリコーディング済の産業分類，縦軸は構成比）．枝の分割において北海道や東北といった地域を持つレコードの産業が振り分けられており，09_10 の食品・飲料等製造業等が他の地域に比べて相対的に多く合成されるように葉が形成されている[3)]．

また，表 4 は，CART によって作成された合成データにおける分布特性を示している．表 2 における原データと比較すると，概ね分布特性は再現されている．一方で経済センサスにおいては，世帯・人口系の統計調査と比較して分布の歪みが大きいことから，合成されたレコードにおいて特に右裾の部分に関しては，原データとは異なる分布特性になっていることが示唆される．これについては原データに分布特性を近づけるための合成の方法に関して改善の余地がある．なお，相関係数においては，原データをもとに相関係数が算出された表 3 と比較しても，相関係数は概ね再現されていることが確認される．

図 2 は，原データと CART による合成データについて，質的属性，量的属性の構成比を比較したものである．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，原データの分布を概ね再現できている[4)]．

バギング（bagging）は，CART を発展させた合成データの生成技法であって，ブートストラップ法を用いて決定木を複数作成する技法である．また，ランダムフォレストは，ブートストラップ法だけでなく，使用する属性の組も変化させて決定木を複数作成する手法である．これらの技法についても，本研究では，合成データの生成技法の有効性を検証するために用いた．

本研究の大きな特徴は，わが国の経済センサスの個票データを用いて，合成データの生成技法の 1 つである GAN の有効性について，MDAV 法や CART 等との比較・検証を行っている

レコード	地域（合成済）	産業（新規に合成）	…
1	北海道	食品・飲料等製造業	
2	北海道	繊維工業	
3	北海道	木材・家具等製造業	
…	…	…	
5001	関東	電子・電気・情報通信等製造業	
5002	関東	印刷関連業	
…	…	…	

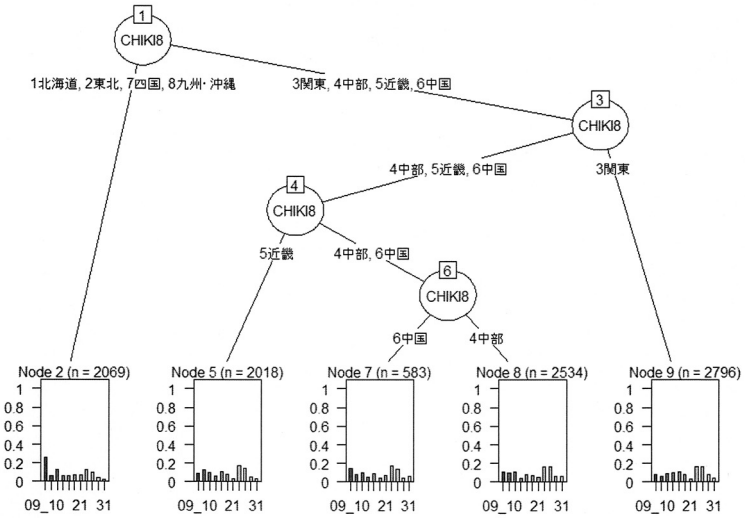

図 1. CART で作成した決定木の例.

ことである．敵対的生成ネットワーク GAN（Generative Adversarial Networks）は，教師なし学習で使用される人工知能アルゴリズムの一種であり，生成ネットワーク（generator）と識別ネットワーク（discriminator）を競わせることによって，学習データを模倣した精巧な出力を得ることができる技術である．図 3 で示されるように，生成される合成データの品質を向上させるために，①ノイズから合成データを生成し，②実データと合成データを識別させた上で，③識別の正否をフィードバックして学習させるという一連の手順が繰り返される．

本研究では，GAN の一手法である条件付き表形式 GAN（conditional tabular GAN=CTGAN）（Xu et al., 2019）を用いて合成データの生成を行った[5]．CTGAN は，表形式データの分布をモデル化し，その分布から合成データを生成する．GAN における合成データ生成の特徴としては，①離散値（質的属性）と連続値（量的属性）が混在するデータが合成の対象になっていること，②非ガウス分布や多峰性分布を対象にした合成データの生成が求められること，③スパースな質的属性に基づいて学習がなされる場合があること等が指摘されている．こうした分布の歪みやスパースな分類区分が存在する場合，合成データにおいて分布特性を精密に再現することは困難になることが想定される．そこで，CTGAN では，離散値を変分ガウス混合モデルで近似し，それぞれの分布の最頻値（mode）を用いて正規化することや，質的属性におけるすべて

表 4．CART によって作成された合成データにおける分布特性．

(1) 基本統計量

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

(2) 相関係数行列

	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

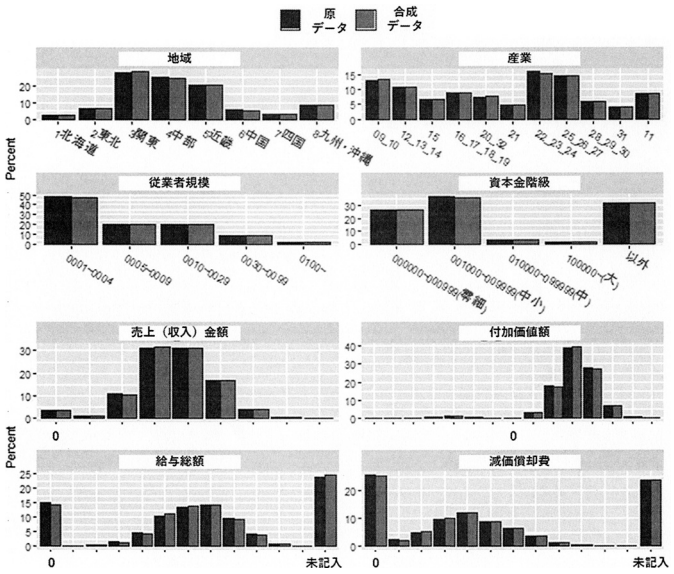

図 2．CART による合成データの属性別構成比．出所　横溝・伊藤（2023）．

の分類区分ができるだけ均等に学習されるように，条件付き(conditional)でオーバーサンプリングすることによって，これらの課題の改善が図られている．

図 4 は，予備実験として行った生成ネットワーク G と識別ネットワーク D の損失関数を用いた実験の結果を示したものである[6]．損失関数は深層学習全般に存在する概念であり，正解と予想の誤差(損失)をもとに，学習の成否や学習回数の判断するために用いられる．学習がうまくいっていない場合には，D が振動しながら発散する状態が見られるが，本実験においては D の損失関数は安定していることから，学習は概ね成功していると考えられる．

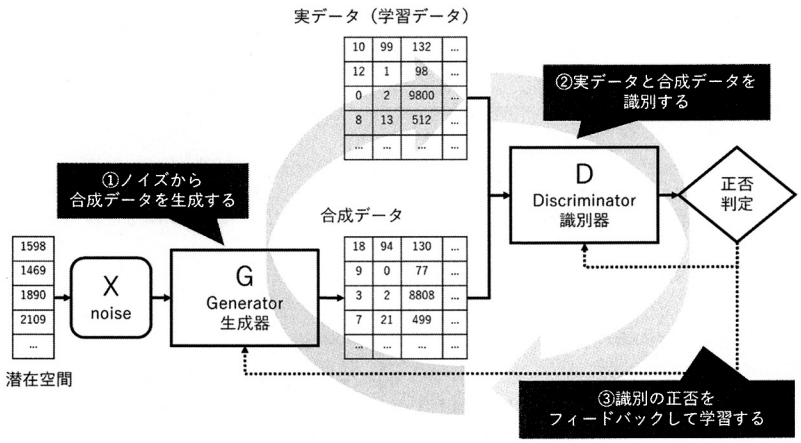

図 3. GAN のイメージ図.

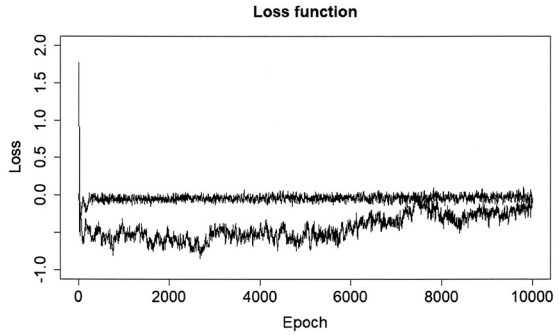

図 4. CTGAN の損失関数.

本実験では，MDAV 法，回帰による生成，サンプリングによる生成，CART，バギング，ランダムフォレストおよび CTGAN については，秘匿性と有用性の比較・検証を行っている．有用性の指標としては，前節で述べたように傾向スコアの指標である pMSE，および相関係数の平均絶対誤差(mean absolute error of correlation)，秘匿性の指標に関しては ARD をそれぞれ用いて検証を行った．比較の対象となる MDAV 法については，グループ化の対象となるレコード数 k を 3，5，10，30，50，100 と変化させた場合の，CART，バギング，ランダムフォレストに関しては，決定木における葉の大きさの制約である最小リーフサイズを 3，5，10，30，50，100 と変化させた場合の有効性の評価をそれぞれ行った．いずれも数字が大きくなるほど粗くクラスリングされるため，秘匿性は高くなる一方，有用性は低くなることが想定される．CTGAN については，epoch 数(学習の回数の単位)を 100，300，1000，3000，5000，10000 と変化させた上で，有効性の比較を行っている．こちらは学習回数が多いほど精巧な合成データに近付くため，有用性は高くなる一方，秘匿性は低くなることが想定される．その他のハイパーパラメータについては，各ミクロデータ生成ツールのデフォルトのパラメータが設定されている．

CART のような決定木による合成データの生成技法には乱数発生に伴うばらつきが生じることから，合成データを 10 回生成した上で，その平均値が本研究で用いられている．なお，決

定木手法だけでなく，CTGAN もこのような乱数の影響を受けることに留意されたい．

6. 実証実験の結果

本節では，経済センサスを用いた合成データの生成技法の有効性に関する実証実験の結果について述べる．図 5 と図 6 はそれぞれ，実証実験の対象となった各種の合成データの生成技法を比較した R-U マップ（Risk Utility map）（Duncan et al., 2001）を図示したものである．本分析で検証の対象となった MDAV 法，回帰による生成，サンプリングによる生成，CART，バギング，ランダムフォレストおよび CTGAN は，図 5 と図 6 ではそれぞれ，mdav，parametric，sample，cart，bag，rf および ctgan と表示されている．横軸は秘匿性の指標である層化平均 ARD を表している．また縦軸の有用性の指標としては，図 5 では pMSE，図 6 では相関係数の平均絶対誤差を用いた結果をそれぞれ表しており，秘匿性の強度と有用性の程度も大きいほど，右下のエリアにプロットされる．

図 5 を見ると，全体的に k が大きくなるにしたがって，秘匿性の程度は大きく，有用性は相対的に小さくなる傾向が見て取ることができる．ミクロアグリゲーションである MDAV 法における実証結果は，相対的に秘匿性の強度が小さなエリアにプロットされる傾向にある．すなわち攪乱的手法である MDAV を適用した場合，本実験で生成された各種の合成データと比較して相対的に秘匿性が小さくなる傾向にあることがわかる．

つぎに，CART やバギングに着目すると，最小リーフサイズを大きくしても有用性が大きく低減しないだけでなく，MDAV 法と比較して秘匿性の強度が大きいことが確認できる．ラン

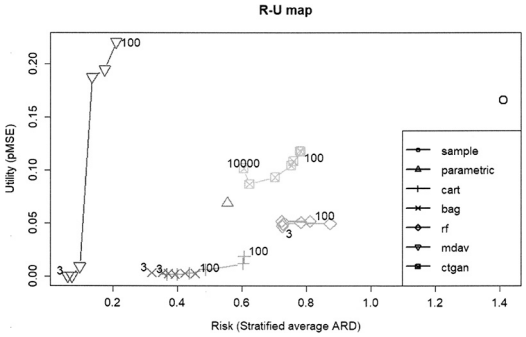

図 5. 秘匿性と有用性の評価結果：pMSE を用いた場合.

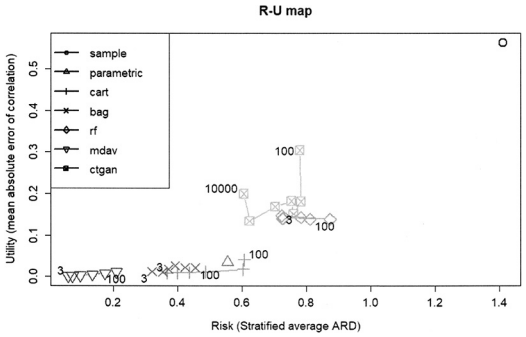

図 6. 秘匿性と有用性の評価結果：相関係数の平均絶対誤差を用いた場合.

ダムフォレストの場合も同様に，最小リーフサイズにかかわらず，有用性の程度は大きく変化しないが，MDAV 法との比較においては，グループ化の対象となるレコード数 k が 3 の場合には，MDAV 法においては，ランダムフォレストに対してより有用性が高くなっている．そして，CTGAN によって得られた実証結果については，CART，バギングとランダムフォレストを適用して生成した合成データにおける結果と比較すると，その秘匿性の程度は大きくなる傾向にあるが，有用性の低下も大きくなっていることが興味深い．

図 6 では，横軸の秘匿性については先ほどと同様に層化平均 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表している．有用性の指標として pMSE を有用性の指標として用いた図 5 と比較して，MDAV 法，CART・バギング・ランダムフォレスト，および CTGAN の位置関係がより顕著に表れている．すなわち，MDAV 法，CART，CTGAN の順に秘匿性の強度が大きくなっているが，MDAV 法と CART については，有用性の大きな差は見られない．このように，相関係数を有用性の指標とした場合でも，R-U マップにおいて類似した傾向が現れていることがわかる．その一方で，CTGAN においては，有用性に関しては epoch 数次第で不安定になる傾向が確認された[7)]．

7. おわりに

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・企業系のミクロデータに対して各種の合成データの生成技法の有効性を評価した．本研究においては，攪乱的手法であるミクロアグリゲーション（MDAV），合成データの生成技法である CART や回帰，および CTGAN を対象に，秘匿性や有用性に関する定量的な評価を行った．本研究でも，CART で合成データを生成した結果，分布特性として要約統計量や相関係数，属性ごとの構成比が再現可能なことがわかった．また，CART は MDAV と比較して，有用性を保ったまま秘匿性を高められる可能性があり，またその程度を調節することも可能なことが確認できた．さらに本稿では，公的統計のミクロデータに対する CTGAN の合成データとしての適用可能性を検証した．CTGAN の場合，CART と比較して，秘匿性の強度が高くなっていることが明らかになったが，有用性の低下も大きいことがわかった．

なお，合成データの生成技法の有効性に関する比較・検証については，統計実務の観点に立った場合には，合成データの利用可能性を検討するために，さらなる実証研究を進める必要があるように思われる．これについては，将来的な研究課題としたい．

注．

1) 本研究で使用した質的属性については，リコーディング済みの変数が用いられている．

2) neg-log 変換は，つぎの式で表現される．

$$Y_n = sgn(X_n) \times \ln(|X_n| + 1) \tag{3.1}$$

neg-log 変換がなされているのは量的属性のすべてではなく，売上等の経理項目のみであることに留意されたい．例えば，従業者数や資本金については，リコーディングのみが適用されている．なお，neg-log 変換に関する説明については，高部 (2017) も参照されたい．

3) 図の例では質的属性に対する CART であったが，量的属性に対しても CART を適用することができる．その際，原データの持つ特異な量的属性値（例えば 777 といった目立ちやすい売上など）であっても合成データとしてそのまま生成されるため，平滑化（smoothing）を行うことが考えられる．具体的には，ガウスカーネル密度推定（Gaussian

kernel density estimator)が用いられる．なお，カーネル密度推定は，synthpop にオプションとしてあらかじめ実装されている．

[4] CART の場合，特別の工夫を加えずとも 0 や未記入まで含めて属性ごとの分布を再現できる点は注目に値する．

[5] 米国の datacebo 社の提供する，合成データ生成ツール The Synthetic Data Vault(SDV)の一機能として，Python や R 用パッケージ “CTGAN” が公開されている．

[6] CTGAN の損失関数は以下で計算される．

$$\mathcal{L}_D = \frac{1}{m}\sum_{1}^{m}[D(x'^{(i)}) - D(x^{(i)})] \tag{5.1}$$

$$\mathcal{L}_G = \frac{1}{m}\sum_{1}^{m}[D(x'^{(i)})] + H \tag{5.2}$$

x：元データ
x'：合成データ
m：レコード数
H：交差エントロピー

[7] 深層学習では一般に，学習の初期の段階では epoch 数を増やすほど性能が向上することが多いが，ある点を越えると過学習(overfitting)を起こすことがある．本実験においては epoch 数 10000 では過学習の可能性が疑われる．

参 考 文 献

Bates, A. G., Špakulová, I., Dove, I. and Mealor, A. (2019). ONS methodology working paper series number 16 synthetic data pilot, https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber16syntheticdatapilot（最終アクセス日 2024 年 7 月 4 日）.

Breiman, L., Friedman, J., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*, 1st ed., Chapman and Hall/CRC, New York.

Chien, C.-H., Welsh, A. H. and Moore, J. D. (2021). Synthetic business microdata: An Australian example, *Journal of Privacy and Confidentiality*, **10**(2), https://doi.org/10.29012/jpc.733.

Defays, D. and Nanopoulos, P. (1993). Panels of enterprises and confidentiality: The small aggregates method, *Proceedings of 92 Symposium on Design and Analysis of Longitudinal Surveys*, 195–204, Statistics Canada, Ottawa.

Domingo-Ferrer, J. and 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. and Torra, V. (2001). Disclosure control methods and information loss for microdata, *Confidentiality, Disclosure and Data Access*: *Theory and Practical Applications for Statistical Agencies* (eds. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz), 91–110, Elsevier Science, Amsterdam.

Duncan, G., Keller-McNulty, S. A. and Stokes, S. L. (2001). Disclosure risk vs. data utility: The R-U confidentiality map, Technical Report, No.121, National Institute of Statistical Sciences, Washington, D.C.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014). Generative adversarial nets, *Advances in Neural Information Processing*

Systems, **27**, 2672–2680.

Hundepool, A., de Wetering, A. V., Ramaswamy, R., Franconi, L., Capobianchi, A., DeWolf, P.-P., DomingoFerrer, J., Torra, V., Brand, R. and Giessing, S. (2003). *μ-ARGUS Version 3.2 Software and User's Manual*, Statistics Netherlands, Voorburg NL.

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, 経済論集, **15**(3・4 号合併号), 197–232.

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, **47**(2), 77–101.

伊藤伸介, 横溝秀始 (2024). わが国の公的統計における合成データの展開可能性に関する一考察—事業所・企業系の統計調査を例に—, 経済学論纂, **64**(3・4 合併号), 147–164.

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, **71**, 83–124.

Kim, H. J., Drechsler, J. and Thompson, K. J. (2021). Synthetic microdata for establishment surveys under informative sampling, *Journal of the Royal Statistical Society Series A*: *Royal Statistical Society*, **184**(1), 255–281.

Little, C., Elliot, M., Allmendinger, R. and Samani, S. S. (2021). Generative adversarial networks for synthetic data generation: A comparative study, https://doi.org/10.48550/arXiv.2112.01925.

Nowok, B., Raab, G. M. and Dibben, C. (2016). synthpop: Bespoke creation of synthetic data in R, *Journal of Statistical Software*, **74**(11), 1–26.

Snoke, J., Raab, G. M., Nowok, B., Dibben, C. and Slavkovic, A. B. (2016). General and specific utility measures for synthetic data, *Journal of the Royal Statistical Society Series A*: *Statistics in Society*, **181**(3), 663–688.

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, **74**, 29–56.

Taub, J. and Elliot, M. (2019). The synthetic data challenge, Paper presented at Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, the Hague, the Netherlands, https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019_S3_UK_Synthethic_Data_Challenge_Elliot_AD.pdf（最終アクセス日 2024 年 7 月 4 日）.

Templ, M. (2017). *Statistical Disclosure Control for Microdata*: *Methods and Applications in R*, 99–132, Springer International, Cham, Switzerland, https://doi.org/10.1007/978-3-319-50272-4.

UNECE (2022). Synthetic data for official statistics — A starter guide, https://unece.org/sites/default/files/2022-11/ECECESSTAT20226.pdf（最終アクセス日 2024 年 7 月 4 日）.

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

Xu, L., Skoularidou, M., Cuesta-Infante, A. and Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN, *Advances in Neural Information Processing Systems*, **32**, 7333–7343.

横溝秀始, 伊藤伸介 (2023). 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—, 統計研究彙報, **80**, 97–116.

Study on Synthetic Data Generation Techniques for Official Statistics on Establishments and Enterprises: The Economic Census as an Example

Shinsuke Ito[1] and Shuji Yokomizo[2]

[1]Faculty of Economics, Chuo University
[2]Statistical Research and Training Institute, Ministry of Internal Affairs and Communications

Not only are anonymous data currently not available for statistical surveys of business establishments and enterprises in Japan, but producing publicly available microdata for statistical surveys of business establishments and enterprises is also difficult. Therefore, the development of a method to produce synthetic data would meet the need for test data. In this paper, we quantitatively evaluate various techniques for generating synthetic data using individual data from the Economic Census of Activity.

In this study, we quantitatively evaluated the usefulness and confidentiality of synthetic data generated using the maximum distance to average vector (MDAV) method of microaggregation, which is a disturbing method, the classification and regression tree (CART), and the conditional tabular GAN (CTGAN), a deep learning model. The results confirmed that the distributional properties such as summary statistics and correlation coefficients were reproducible when synthetic data were generated using CART. In addition, compared with the MDAV method, CART can potentially increase confidentiality while maintaining usefulness. Furthermore, for CTGAN, the degree of confidentiality was found to be higher compared with that for CART; however, the decrease in usefulness was also confirmed to be relatively greater.

Key words: Synthetic data, microaggregation, CART, CTGAN, Economic Census.