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――

伊　藤　伸　介*
横　溝　秀　始**

１．はじめに

海外では，近年合成データ（synthetic data）の作成とその実用化に向けた研究に対する関心が高まっている[1)]．合成データは，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と位置付けられる（Templ（2017, p.157））．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・レコード群に対して，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．合成データの作成に関しては，統計モデルに基づくパラメトリックな手法の適用だけでなく，CART（Classification And Regression Tree）（Breiman et al.（1984））等によるノンパラメトリックな手法も適用されてきた（Reiter（2005）等）．その一方で，合成データについては，対象となるすべてのレコードを対象に欠測値を含む属性群に対して擬似的に値が生成されたデータと，一部のレコード群に含まれるセンシ

1）　海外における合成データの生成に関する研究については，近年では，国際連合欧州経済委員会（United Nations Economic Commission for Europe＝UNECE）欧州統計家会議（Conference of European Statistician）の主催による「統計データの機密保護に関する専門家会議（Expert Meeting on Statistical Data Confidentiality）」および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「統計データベースのプライバシー保護会議 Privacy in Statistical Databases」において，多くの研究論文が発表されている．

ティブな属性にのみ擬似的な値によって補完したデータが存在する．前者は完全合成データ（fully synthetic dataset）（Rubin（1993））, 後者は部分合成データ（partially synthetic dataset）（Little（1993））と呼称される（Drechsler（2011）, 高部（2022））.

海外では，公的統計を対象にした合成データの方法論に関する様々な適用事例が存在する．欧州統計局（Eurostat）では，一般公開型ミクロデータ（public use microdata）を作成するために，合成データの方法論が適用されている．具体的には，Eurostat は，EU-SILC（＝European Union Statistics on Income and Living Conditions）において，統計的なモデルに基づくシミュレーションによる合成データの生成技法を用いて一般公開型ファイル（Public Use File＝PUF）を作成・公開している（伊藤（2018））．また，エディンバラ大学では，スコットランドの人口センサスの個票データを縦断的に連結するだけでなく，医療データもリンケージされた縦断的なデータ構造を有する[2)]．スコットランド縦断調査（Scottish Longitudinal Study）を対象に，合成データの作成が行われてきた．これについては，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．また，イギリス国家統計局においても，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，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））.

わが国においても，合成データの作成の可能性について議論が展開されている．例えば，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）に基づく合成データの作成（南（2022））や，民間のデータを主な対象にしたプライバシー保護型合成データ（Privacy Preserving Synthetic Data）の作成（千田他（2022））等が指摘される.

わが国の公的統計を対象にした合成データの作成方法に関する研究については，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（高部（2022））, 事業所・企業系の統計調査を対象にしたノンパラメトリックな合成データの作成方法の追究（横溝・伊藤（2023））等についての実証研究が存在する．しかしながら，海外の事例と比較すると，わが国では公的統計に基づく合成データの生成技法についての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

2） イギリス国家統計局（Office for National Statistics＝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.

と考える．

ところで，わが国において2018年に改正された統計法（以下「2018年統計法」と呼称），および2018年統計法に関連する施行規則やガイドラインでは，公的統計を対象にした合成データについての明確な定義がなされているとは言えない．したがって，合成データの作成可能性を追究するにあたっては，わが国における合成データの位置付けを明確にした上で，合成データの方法論を公的統計に対してどのように適用していくかについての検討が求められる．「合成データの方法論をわが国で検討することは，それを適用して作成されたデータをわが国の一般用ミクロデータのような匿名データの外に位置付けることができるか，あるいは2018年統計法ないしは統計法施行規則の下で，匿名データに適用される匿名化技法の１つとみなすことが可能かを議論する意味で，統計法制度上の課題にもなる」（伊藤（2022, 13頁））．その意味では，わが国の統計法制度の下で，公的統計ミクロデータにおける合成データを位置付けた上で，公的統計に対する合成データの作成方法の可能性を追究することが求められよう．

本稿では最初に，わが国の統計法制度の枠内において，公的統計の合成データの作成可能性を検討する．つぎに，事業所・企業系の統計調査である「平成28年経済センサス - 活動調査（以下，「経済センサス活動調査」と呼称）」を例に，合成データの生成技法の適用可能性に関する実証研究を行う．

2．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

わが国では，現在全国消費実態調査と就業構造基本調査の一般用ミクロデータが公開されている．一般用ミクロデータは，公表された統計表に含まれるセルの数値からミクロレベルの値を生成させることによって作成されることから，web 上での取得が可能になっている．また，一般用ミクロデータは，利用目的の制限がなく誰でも入手可能であることから，それは，わが国における一般公開型ミクロデータに該当すると言うことができる．

全国消費実態調査の一般用ミクロデータについては，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，一般用ミクロデータの作成に関する方法的な基礎となっている．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)）．これについては，全国消費実態調査における高次元の集計表を元データとし，その分布特性に近似するように属性値を新たに生成させることを指向していることから，合成データの方法論の適用事例の１つとみなされうる．

その一方で，教育用擬似ミクロデータは，匿名化技法の１つであるミクロアグリゲーション（mi-

croaggregation）（伊藤（2009））が方法的に展開されたものと位置付けられうる．ミクロアグリゲーションは，ミクロデータ（個票データ）をｋ個のレコードを有する同質的なレコード群にグループ化した上で，そのレコードにおける個々の属性値を平均値等の代表値に置き換える技法である（Domingo-Ferrer and Mateo-Sanz（2002, p.190），伊藤（2009, 201頁））．

教育用擬似ミクロデータは，「超高次元クロス集計表」[3]に基づき作成された集計表のセルに含まれる数値群を，質的属性値と量的属性値を含む個票データに準じたレコード群に変換したものから派生的に捉えることも可能である（伊藤（2009，211-212頁））．これらのレコード群は，ミクロアグリゲーションが適用されたミクロに準じたデータ（以下「ミクロアグリゲートデータ」と呼称）として位置付けられうる．具体的なミクロアグリゲートデータの作成手順は以下の通りとなる．第１に，超高次元クロス集計表の集計事項となる属性群から，属性の組み合わせを適当に選択することによって，超高次元クロス集計表に含まれるすべてのセルが０かｋ（ｋは閾値）以上の数値になるようにクロス集計表を作成する．第２にこの集計表から同一の属性値の組み合わせを有するレコード群を再編成することによって，ミクロアグリゲートデータを作成することが可能になる．

ところで，2018年統計法の下では，調査票情報（個票データ）に対して合成データの方法論を適用して作成されたデータは，個票データに対して「加工」（法２条第12項）が施された匿名データと考えられている．このことは，適用された合成データの作成手法は，匿名化技法の１つとみなされることを意味している（伊藤（2022））．したがって，わが国の現行法においては，例えば統計モデルに基づくパラメトリックな手法を用いて，個票データから変数群を生成した場合，その生成されたデータは合成データとは言えず，匿名データに位置付けられる．

それに対して，ミクロアグリゲートデータは，設定された閾値ｋに対して，グループ内にｋ以上のレコードが存在するようにアグリゲートされる．したがって，ｋの設定の仕方によっては，個票データとの対応付けが困難であるという観点から，ミクロアグリゲートデータは個票データとは異なるデータとなりうる．そのようなミクロアグリゲートデータに合成データの方法論を追加的に適用して作成したデータは，個票データに対して「加工」が施された匿名データとは異なる．こうしたことからアグリゲートされた公的統計のデータあるいは集計量に合成データの生成技法が適用されたデータは，公的統計において合成データの一形態になると考えられる[4]．

元データになるミクロアグリゲートデータが公表可能なデータ（集計表）となりうるかは，そこから作成された合成データが，誰でも入手可能な一般公開型ミクロデータに該当するかどうかに

３） 超高次元クロス集計表とは，「個別データが有するすべての属性群を集計事項の対象とした上で作成されるｎ次元の多重クロス集計表」であって（伊藤（2009, 211頁）），ミクロアグリゲーションの一形態としても位置付けられる．

４） ミクロアグリゲーションを適用する前に，特異値（外れ値）の削除といった処理を行うことも考えられる．

も関わってくる．統計表の結果数値や回帰分析のパラメータを含む統計量が公表可能な場合，そこから擬似的に生成された合成データは，一般公開型ミクロデータとしても位置付けることが可能である．わが国の一般用ミクロデータもこれに該当すると言える．このことは，合成データの中に一般用ミクロデータが包含されることを意味する．それに対して，合成データの生成の元となる統計表（あるいは集計量）やパラメータ等の統計量が，個体識別や属性漏洩のリスク等の懸念から公表可能ではないと判断されたとしても[5]，その元データが個票データとは異なるものとみなされるのであれば，そこから生成されたデータは匿名データとは異なる合成データに含まれうる．ただし，それが一般公開型ミクロデータとして公開可能となるためには，合成データに含まれる個体情報の漏洩リスクに関する検証が求められるだろう．

事業所・企業系の統計調査の個票データに対して直接合成データの生成技法を適用すれば，事業所・企業系の匿名データが作成される．そうした方向性も考えられるが，現状では事業所・企業系の匿名データがわが国で作成・提供されていないだけでなく，海外でも作成事例が少ないことを踏まえると，匿名化技法としての合成データの方法論を事業所・企業系の個票データに適用した場合の取り扱いに関しては慎重な議論が求められよう．一方で，統計調査の個票データから編成されたミクロアグリゲートデータに合成データの方法論を用いることによって，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としての合成データの展開も期待できる．その意味では，個票データにミクロアグリゲーションを適用することによって編成されたミクロアグリゲートデータ，あるいはアグリゲートされた集計量から合成データを生成することが，現行の統計法制度の下で事業所・企業系における擬似的なミクロデータの作成についての展開を図るための１つの可能性として期待できる．そこで次節では，ミクロアグリゲーションの特徴について述べることにしたい．

3．ミクロアグリゲーションの特徴について

先述の通り，ミクロアグリゲーションは，ミクロデータを同質的なレコード群にグループ化し，グループ内のレコードにおける属性値を代表値に置換する技法である．ミクロアグリゲーション技法については，①属性の性質による区分と②レコードをグループ化する場合の基準とな

5）　合成データの生成の元となる統計表（集計量やパラメータ等の統計量が表章された統計表も含む）が，集計計画に含まれないのであれば，統計作成部局においては，これらは，一般に公表可能な統計表ではないと判断されるであろう．しかしながら，統計法制度上，調査票情報の中に包含される中間生成物を除けば，公表可能でない統計表に基づいて，合成データを生成することは統計技術的には可能だと言える．したがって，合成データの生成の在り方については，わが国の統計法制度の下での合成データの位置付けの観点からも，さらなる議論が求められる．

るレコード数（閾値）の設定方法の観点から類型化することができる．本節では，伊藤（2009）に基づいて，ミクロアグリゲーションの特徴を概説する．

ミクロアグリゲーションについては，属性の性質によって，（１）量的属性と（２）質的（カテゴリカルな）属性に分類することができる．Domingo-Ferrer and Torra（2001）を参考にすれば，ミクロアグリゲーション技法は，主として量的属性を対象とした匿名化技法として位置付けられる[6]．さらに，量的属性に関するミクロアグリゲーションについては，レコード数（閾値）の設定方法の観点から，（１）閾値を固定した上でのグループ化と（２）閾値に基づく探索的なグループの設定に類型化することができる．前者に含まれるミクロアグリゲーション技法としては，①単一軸法（single axis method），②Ｚスコア総計法（sum of Z-scores method），③個別ランキング法（individual ranking method）（Anwar（1993）），および④MDAV 法（＝multivariate microaggregation based on Maximum Distance to Average Vector）（Domingo-Ferrer & Mateo-Sanz（2002），Hundepool et al.（2003））がある．

単一軸法は，ソートキーとなる特定の量的属性に着目し，ソートされたレコード群を一定のレコード数ごとにグループ化を行う技法であるが，Ｚスコア総計法は，各レコードにおける属性値群を対象に標準化された値の総計値（Ｚスコア総計値）に基づいて，レコードのグループ化を行う技法である（伊藤（2009））．それに対して，個別ランキング法は，量的属性の各々について個別にソート化とグループ化を行う技法として位置付けられる（Anwar（1993），伊藤（2009））．

さらに MDAV 法は，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手法等が存在する．MDAV 法の手順は，以下の通りである（横溝・伊藤（2023））．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する．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のレコードを探索する．③閾値ｋを設定した上で（例えばｋは３），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き換える．④ミクロアグリゲート済（攪乱済み）のレコードを除いた上で，②と③の処理を繰り返す．

後者の閾値に基づく探索的なグループの設定については，例えば，階層区分法がある（hierarchical clustering method, Domingo-Ferrer and Mateo-Sanz（2002））．階層区分法の場合，閾値ｋに基づいて，グループ内平方和（within-group sum of squares）を最小にするためのアルゴリズムを用いて探索的なグループ化が行われる．それは，Ward の階層区分法の一種であって，探索的な

6） 質的属性については，対象となる属性群各々において同一の属性値を有するレコードをグループ化することによって編成された「広義の」ミクロアグリゲーションとして位置付けることは可能である（伊藤（2009））．

（heuristic）ミクロアグリゲーション（k 分割（k-partition））と位置付けられる．

4．合成データに対する秘匿性と有用性の評価指標について

4-1　匿名化ミクロデータにおける秘匿性と有用性の評価方法

ミクロデータに対して適用される匿名化措置については，個票データに含まれる個体情報に関する秘密保護とミクロデータの有用性（利用可能性）の両面から，匿名化措置の適用可能性が追究される．そして，匿名化されたミクロデータについては，秘匿性と有用性の両面から，定量的な評価方法に関する数多くの研究がなされてきた．

伊藤（2019）によれば，匿名化ミクロデータにおける秘匿性の定量的な評価方法については，以下の 5 つの方法に類別することが可能である．第 1 は，外部情報とミクロデータのマッチングであって，個体情報の特定化を行うために用いられると想定される外部情報とミクロデータのマッチングを行うことによって，個体識別リスクを検証する．第 2 は，母集団一意に関する指標の計測であり，ミクロデータの中で母集団一意（population unique）に該当するレコード数を計測する．第 3 は，特殊な一意の分析（Special Uniques Analysis）であって，母集団一意に該当するレコードの中でも，特異な形で存在するレコードを「特殊な一意」とみなした上で，特殊な一意に該当するレコード数に関して検証を行う．第 4 は，レコードリンケージによるリスク評価であり，原データのレコードと匿名化ミクロデータのレコードとの間で対応付けが可能かどうかを判定することによってリスク評価を行う．そして，第 5 は，クロス集計表によるリスク評価であって，原データと匿名化ミクロデータのそれぞれに含まれる質的（カテゴリカルな）属性を用いてクロス集計表を作成し，クロス表の中で度数が 1 となるセルの総数をそれぞれ比較し，度数 1 となるセル数の変化の程度を比較する．

つぎに，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については，（1）記述統計量やクロス表における分布特性の比較と（2）情報量損失（information loss）に関する指標の評価を指摘することができる（伊藤（2019））．前者は，原データと匿名化ミクロデータの間で，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するだけでなく，原データと匿名化ミクロデータに含まれる属性値の差や，分散共分散行列や相関係数行列に見られる分布特性を比較・検証することが考えられる[7)]．それに対して後者は，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[8)]．

ミクロデータの秘匿性と有用性はトレードオフの関係にあると言える．そのため，秘匿性と有

7）　属性値間の距離を定義し，その距離の近さを測ることも想定される．

8）　傾向スコア（propensity score）の計測，クラスター分析による検証，経験分布関数における差異の評価等を用いて有用性を定量的に評価する方法も考案されている（Woo et al.（2009））．

用性の関係を定量的に明らかにし，それらのバランスを勘案しながら，匿名化措置の適用可能性を模索することが求められる．そこで，R-U マップ（R-U Confidentiality Map）（Duncan et al.（2001））のように，秘匿性と有用性の指標を用いて視覚的に図示した上で，匿名化措置の有効性を比較・検討することが考えられる[9]．

合成データの場合，ミクロレベルのデータが擬似的に生成されるが，事業所・企業系の統計調査に含まれる大規模な企業のデータのように，原データの中で外れ値（特異値）のような形で存在するレコードに関しては，合成データの作成方法を適用したとしても，個体が特定されるリスクが生じる可能性がある．また，事業所・企業系のデータであれば，合成データから個体情報についての属性が推定されるリスクも否定できない．そこで，合成データにおいても秘匿性に関する評価指標を検討することが考えられる．また，合成データの利用可能性を検討するためには，有用性についても定量的な評価が求められる．本研究では，各種の評価指標を検討した上で，合成データにおける秘匿性と有用性の定量的な比較・検証の方法を追究する．

4-2　合成データにおける秘匿性の評価指標について

先述の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，合成データに関するそれについての先行研究は多くないが，その中で，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）），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指摘することができる（Hang et al.（2021））．最後の 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って，合成データの秘匿性の評価指標として単純な定式化がなされているだけでなく，統計実務への適用が容易なことが特徴的である．ARD は，以下の（1）式で表される．

$$ARD = \frac{|\hat{L} - L|}{L} \qquad (1)$$

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

ARD は，原データと匿名化されたミクロデータにおいてレコード間の 1 対 1 の対応付けを必要としないことから，合成データだけでなく，攪乱的手法が適用された匿名化ミクロデータの場合

9）　わが国の公的統計ミクロデータを用いた定量的な研究事例としては，例えば伊藤・星野（2014）等を参照．

でも，秘匿性の定量的な評価を行うことが可能である．

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．具体的には，事業所・企業系の統計調査の場合，作成された合成データで把握される特定の地域や産業の中で最も大きな事業所の売上金額を調べることで，最大規模の事業所についてセンシティブな属性情報を高い精度で推定することが考えられる．そこで，本研究では，横溝・伊藤（2023）に基づいて，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）を評価指標として使用した．

層化平均 ARD は，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た上で，その平均値をもとに算出される．売上金額を例に挙げると，まず原データと合成データに含まれる売上金額の属性値からそれぞれの最大値を求めた上で，ARD を算出する．つぎに，データに含まれる変数別に ARD を計算するだけでなく，２つ以上の変数値で層化されたレコードにおいても ARD の計測を行う．算定された ARD の各々からの平均値を計測することによって，層化平均 ARD が算出される．層化平均 ARD が大きいことは，対象となる層化されたレコードから導出される合成データにおける最大値の原データのそれからの乖離が平均的に大きく異なることを意味することから，属性情報が推定されるリスクは相対的に小さくなるとみなすことができる．

4-3　合成データにおける有用性の評価指標について

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））．広義の尺度については，原データと合成データにおける距離が計測される指標が含まれており，Kullback-Leibler 情報量，Hellinger 距離，傾向スコア平均二乗誤差（propensity score Mean Square Error＝pMSE）といった指標が用いられる．それに対して，狭義の尺度については原データと合成データにおける特定のモデルに関する差異が注目され，例えばミクロデータを用いて推定される回帰分析のパラメータにおける信頼区間の差異が用いられる（Taub & Sakshaug（2020））．

これらの中で，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の１つと言うことができる．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である．合成データの作成にあたっては，当該レコードが合成データとして生成されている可能性について，その確率を計算するために，傾向スコアが用いられる．この傾向スコアを用いた有用性の評価は，アメリカセンサス局でも採用されていることが知られている（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[10]．なお，pMSE は以下の（２）式で表される．

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

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

$\hat{p}_i$：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

pMSE を算出するにあたっては，最初に原データと合成データに含まれるレコードをマージした上で，合成データか否かを示す属性値を設定する．この場合，合成データなら０，そうでなければ１という値が付与される．この「合成データか否か」を被説明変数に，地域や産業，売上（収入）金額といった属性群を説明変数として設定し，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 が算出される．この 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 が求められる．pMSE の数値が大きいことは，合成データが原データから乖離する傾向を示していることから，合成データの有用性が相対的に低下することを意味している．

pMSE は，属性間の相関性を直接的に評価する指標として定式化されていない．そこで，本研究では，横溝・伊藤（2023）と同様に，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 and Torra（2001））も有用性に関する指標として使用した．相関係数行列の差の平均絶対誤差は，以下の（３）式で表される（伊藤他（2014））．

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

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

この相関係数行列の差の平均絶対誤差が小さいほど，相対的に原データにおける相関関係が合成データにおいても保存されているとみなすことができる．

10） Woo et al.（2009）においては，攪乱的手法が適用され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・検証した上で，傾向スコアが有用性の指標として最も適していることが論じられている．また，Snoke et al.（2016）は，傾向スコアを合成データの有用性の評価指標に応用している．

5．合成データの作成方法の有効性の検証――経済センサス活動調査を用いて

本節では，事業所・企業系の統計調査である経済センサス活動調査を例に，ミクロアグリゲートデータに対して適用された合成データの生成技法の有効性に関する実証研究を行う．本研究においては，経済センサス活動調査を用いて，ミクロアグリゲーションとノンパラメトリックな手法の１つである CART を対象に，合成データの作成方法の有効性に関する比較・検証を行う．なお，本研究においては，合成データの作成・評価に関する R のパッケージである synthpop を用いる．

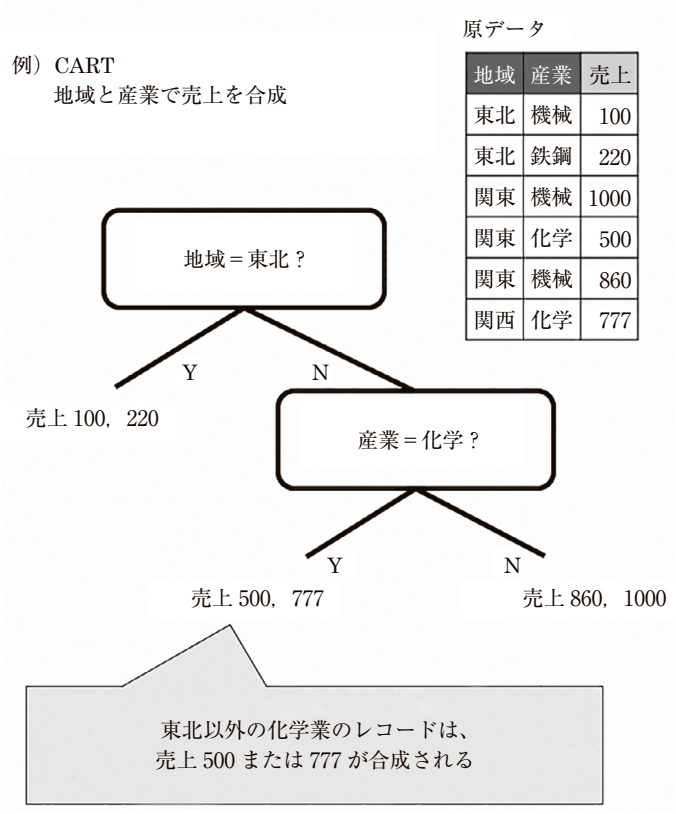
本研究で用いる合成データの生成技法である CART は，観測済みの属性値から目的変数となる属性を再帰的にグルーピングする，ノンパラメトリックな決定木分析手法である．CART の特徴として，質的属性と量的属性の両方における適用可能な技法であることが指摘できる．また，CART の場合，作成する木の深さや葉に振り分けるレコード数の制約条件（最小リーフサイズ）を変更することによって，原データにおける変数値の分布特性が合成データにどの程度保持されるかを調節することが可能である．なお，最小リーフサイズとは，決定木の末端である葉の大きさに関する制約である．最小リーフサイズが小さいほど，細かく枝の分割が行われることから，原データの分布特性に対してより近似的なデータが生成される．

図１は，CART を用いて，地域と産業に基づいて売上金額を合成した場合のイメージ図である．ジニ係数等の基準を用いて，どの属性のどういった分類区分が採用されるか，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がなされる．図１においては，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，地域と産業における分類区分がカテゴライズされた上で，決定木が生成される．このような形で生成された決定木を用いて，地域と産業における数値の組み合わせにしたがって，売上の値が合成される．図１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辿ることによって，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ことを例示している[11]．

本研究で使用するデータは，経済センサス活動調査で把握される従業者規模１人以上1,000人未満の製造業の事業所を対象に，テストデータ作成のために抽出された10,000レコードの事業所であ

11）図１において，例えば777のような特異な値が合成データに出現した場合，それによって偶発的な形で属性情報の推定につながる可能性がある．こうしたリスクを回避するために，synthpop では，合成データの生成において，ガウシアンカーネル密度推定（gaussian kernel density estimator）が行われ，リーフ内の有限個の離散値を連続値に平滑化（smoothing）することによって秘匿性を高めるオプションが存在する（Nowok et al.（2016））．

図 1　CART で地域と産業から売上を合成する際のイメージ



出所）横溝・伊藤（2023），図 5

る．この原データから同じ数の10,000レコードを持つ合成データが生成される．CART を適用する上での質的属性については，地域では 8 区分，産業に関しては11区分，従業者規模においては 5 区分，資本金階級については 5 区分にそれぞれ区分統合が施されている．また，量的属性については，売上（収入）金額，付加価値額，給与総額と減価償却費を使用する．

本研究では，（1）ミクロアグリゲーション，（2）CART および（3）ミクロアグリゲーションと CART を併用した上で作成された各種のデータについて，R-U マップによって秘匿性と有用性の比較を行った．ミクロアグリゲーションについては，①個別ランキング法（onedims），② Z スコア総計法（zscore）と③ MDAV（mdav）を行った．また，ミクロアグリゲーションと CART の併用については，MDAV を行った上で，CART が実施されている．なお，ミクロアグリゲーションの適用にあたっては，R のパッケージである sdcMicro が用いられている[12)]．

12）　経済センサスを用いたミクロアグリゲーションを含む匿名化技法の有効性の検証に関する実証研究については，伊藤・横溝（2021），横溝・伊藤（2022）を参照．

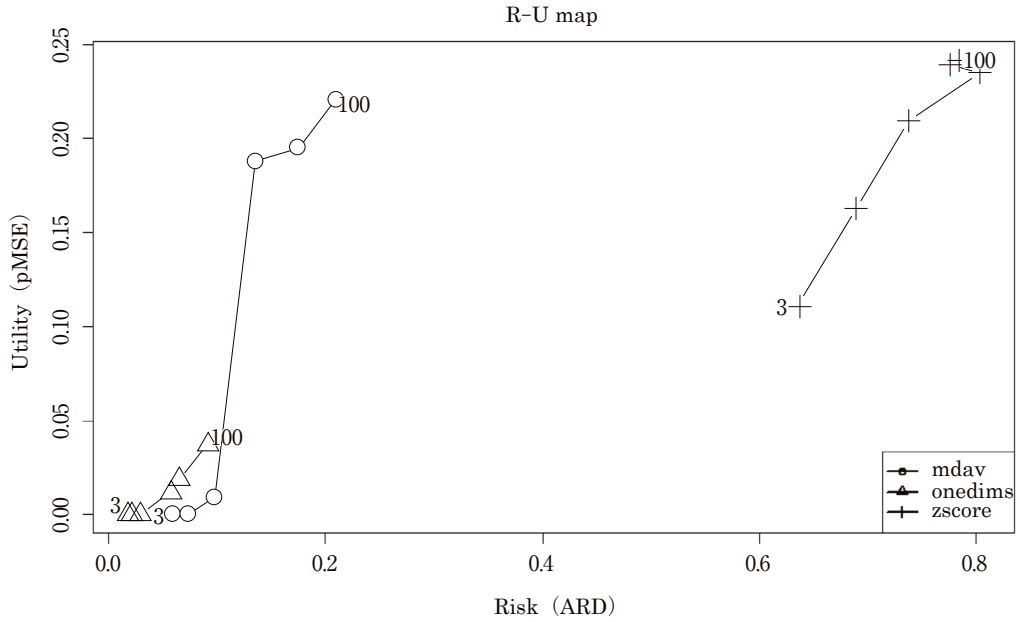
ミクロアグリゲーションについては，グループ化の対象となるレコード数 k を 3，5，10, 30, 50, 100と変化させた場合の有効性の評価，CART に関しては，最小リーフサイズを 3，5，10, 30, 50, 100と変化させた場合の定量的な評価をそれぞれ行った．なお，決定木手法には乱数発生に伴うばらつきが生じることから，本研究では，CART については10回ずつ合成した上でその平均値のプロットを行っている．

さらに，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．さらに合成データを10回作成した上で，その平均値が本研究で用いられている．

最初に，経済センサスの個票データに対してミクロアグリゲーションを適用した場合の有用性と秘匿性の検証結果について確認しておきたい．図2-1と図2-2はそれぞれ，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 法のそれぞれを比較した R-U マップを図示したものである．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2-1では pMSE，図2-2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ことが確認できる．

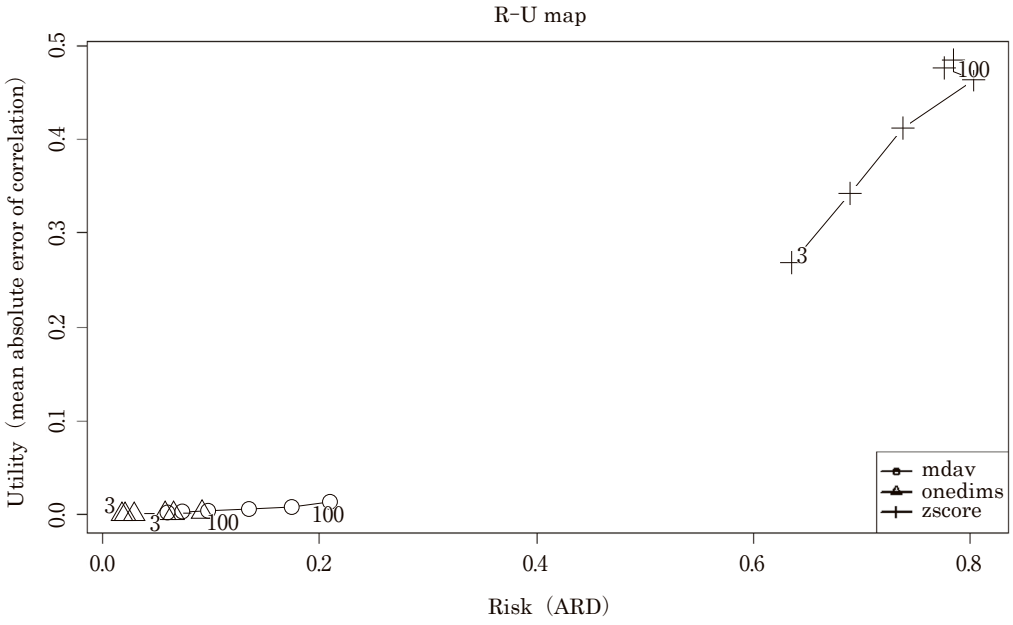
実証結果を見ると，図2-1と図2-2のいずれについても，全体的に k が大きくなるにしたがっ

図2-1　ミクロアグリゲーションにおける秘匿性と有用性の検証結果：pMSE を用いた場合



出所）横溝・伊藤（2023），図4-1を一部修正

図2-2　ミクロアグリゲーションにおける秘匿性と有用性の検証結果：相関係数の平均絶対誤差を用いた場合



出所）横溝・伊藤（2023），図4-2を一部修正

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れる．ミクロアグリゲーション技法ごとに見ていくと，最初に 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．また，k の値を大きくすると有用性が大きく低減する傾向にある．つぎに個別ランキング法における実証結果は，k の値にかかわらず，秘匿性が非常に小さいエリアにプロットされる．さらに，Z スコア総計法での実証結果においては，秘匿性の強度は相対的に大きいが，全体的に有用性の程度が他の 2 つの技法と比較して，著しく低減することがわかる．合成データの元データとしてのミクロアグリゲートデータを作成する場合に，有用性を重視するのであれば，MDAV 法が Z スコア統計法よりも望ましいことが確認できる．

図3-1と図3-2はそれぞれ，ミクロアグリゲーションと CART の比較・検証を行った実証結果を示したものである．さらに，ミクロアグリゲーションを適用した上で CART を適用した実証結果も示している．ミクロアグリゲーションの技法としては，有用性の観点から MDAV 法が選定されている．個票データに直接 CART を適用した場合における実証結果は，MDAV 法におけるそれと比較して，秘匿性の強度が大きなエリアにプロットされていることが確認できる．また，図3-1を見ると，k の値が小さい場合（3，5，10）には，有用性については CART と MDAV 法に大きな違いは見られないが，k が10を超えると，CART のほうが MDAV 法と比較して，相対的に良い実証結果が得られている．また，CART に着目すると，最小リーフサイズを大きくした場合，有用性

図3-1　ミクロアグリゲーションと CART の比較・検証の結果：pMSE を用いた場合

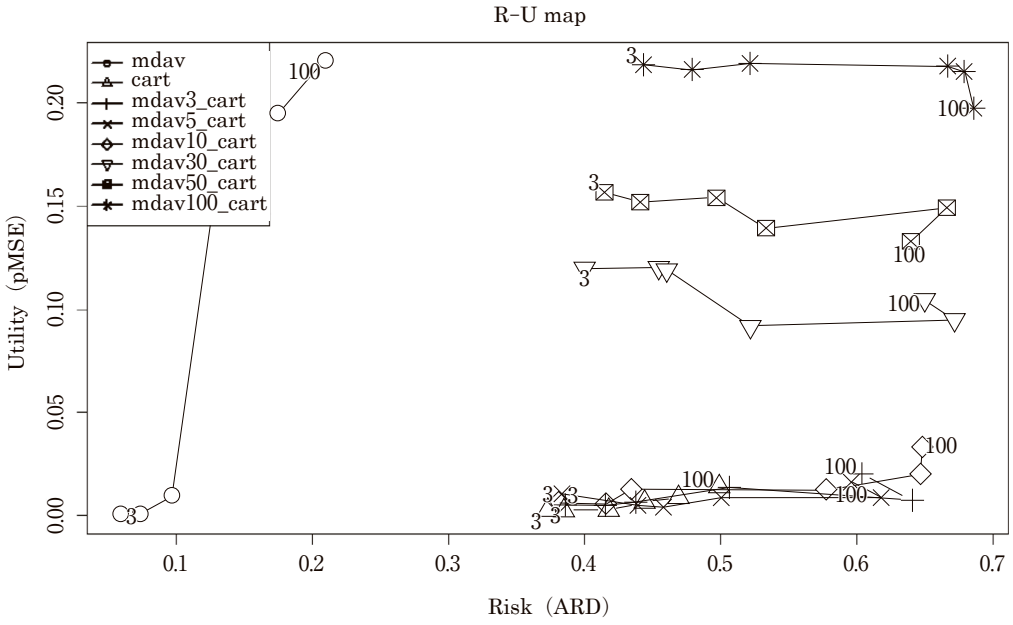
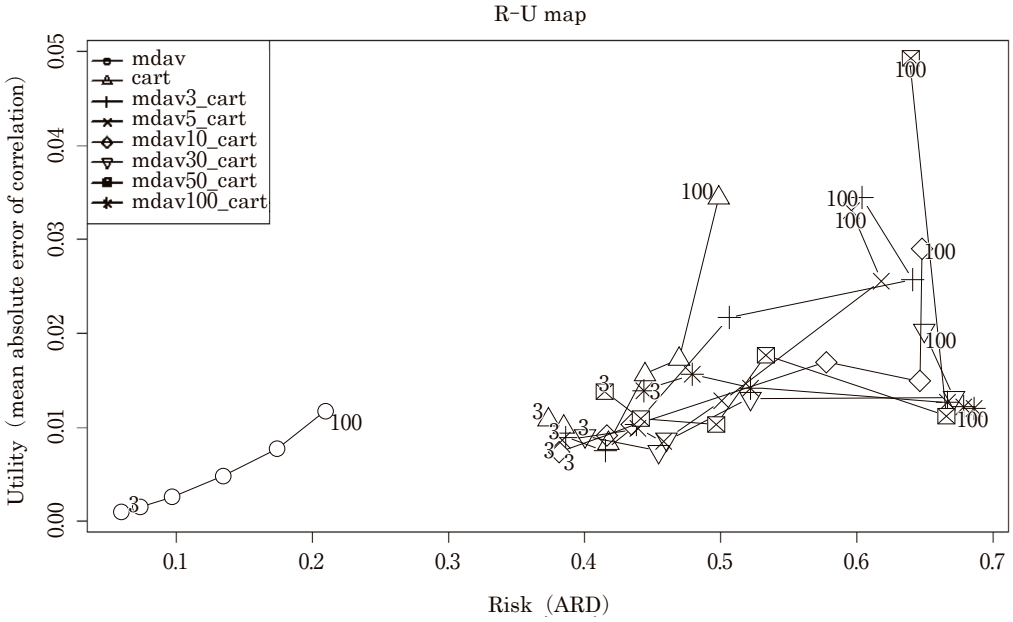


図3-2　ミクロアグリゲーションと CART の比較・検証の結果：相関係数の平均絶対誤差を用いた場合



については大きく低減することなく，秘匿性の強度が高まっていることも確認できる．

MDAV 法で作成したミクロアグリゲートデータに対して CART を適用して作成した合成データにおける実証結果を見ると，秘匿性と有用性の指標のいずれについても，CART のみを適用した結果と比較してそれほど実証結果が変わっていないことが確認できる．MDAV 法における k の値が大きくなるにつれて，秘匿性の程度は低減する傾向にあるが，k の値が小さい場合（3，5，10）については，CART のみの実証結果とほぼ近似的な結果が得られていることが興味深い．

図3-2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3-1と比較して，MDAV 法と CART の位置関係がより顕著に表れている．すなわち MDAV 法，CART の順に秘匿性の強度が大きくなっているだけでなく，MDAV 法，CART の順に有用性が低下していることが確認できる．図3-2において，注目すべき点としては，①層化平均 ARD の R-U マップにおける比較・検証結果と同様に，CART を用いた場合に，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 MDAV 法に CART を追加的に適用した合成データにおける実証結果は，最小リーフサイズが10以下であれば，CART のみを適用した場合の結果と比較して，有用性と秘匿性のいずれも大きく変わらないことが指摘される．

6．むすびにかえて

本稿では，わが国における合成データの作成可能性を追究するための１つの試みとして，ミクロアグリゲーションが適用されたデータに合成データの方法論を追加的に援用されたデータを試行的に作成した上で，その秘匿性と有用性について定量的な検証を行った．

個票データに直接 CART を適用した結果と，個票データから MDAV によって作成したミクロアグリゲートデータに対して CART を追加的に適用した場合の結果を比較すると，設定された閾値 k にもよるが，本実験では有用性と秘匿性の両面についてそれほど違いがないことが確認された．本研究成果を踏まえると，ミクロアグリゲートデータに CART を追加的に適用して合成データを生成することについては，さらなる展開可能性を図っていくことが期待できる．その一方で，ミクロアグリゲートデータから生成した合成データの実用性を追究するにあたっては，実験の対象となる質的属性をさらに増やした場合，どのような技法によってミクロアグリゲートデータを作成するかについてさらなる検討が求められる．これについては今後の検討課題としたい．

参 考 文 献

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，197-232頁

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（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，39-62頁

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，1-23頁

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，83-124頁

伊藤伸介・星野なおみ（2014）「国勢調査ミクロデータを用いたスワッピングの有効性の検証」，『統計学』第107号，1-16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

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，111-130頁

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，35-42頁

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，11-16頁

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，1-15頁

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，151-170頁

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』第80号，97-116頁

Anwar M. N.（1993）“Microaggregation: The Small Aggregates Method”, Eurostat Internal Report.

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

Breiman, L., Friedman, J. H. Olshen, R. A. and Stone C. J.（1984）*Classification and Regression Trees*. Belmont, CA: Wadsworth.

Chien, Chien-Hung, Alan Hepburn Welsh and Moore, John D.（2021）“Synthetic Business Microdata: An Australian Example”, *Journal of Privacy and Confidentiality* 10(2). https://doi.org/10.29012/jpc.733.

Domingo-Ferrer, J. and Torra, V.（2001）“Disclosure Control Methods and Information Loss for Microdata”, Doyle et al.（eds.）（2001）*Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies*, Elsevier Science, Amsterdam, pp.91-110.

Domingo-Ferrer, J. and Mateo-Sanz, J. M.（2002）“Practical Data-oriented Microaggregation for Statistical Disclosure Control”, *IEEE Transactions on Knowledge and Data Engineering*, Vol.14, No.1, pp.189-201.

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

Drechsler, J. and 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.

Duncan, G., Keller–McNulty, S. A. and Stokes, S. L. (2001) Disclosure Risk vs. Data Utility: The R–U Confidentiality Map. Carnegie Mellon University. Journal contribution.

Hang J. Kim, Drechsler, Jörg and Thompson, Katherine J. (2021) "Synthetic microdata for establishment surveys under informative sampling," *Journal of the Royal Statistical Society Series A*, Royal Statistical Society, Vol. 184(1). pp. 255–281.

Hundepool, A., de Wetering, A. V., Ramaswamy, R., Franconi, L., Capobianchi, A., De Wolf, P.–P., Domingo–Ferrer, J., Torra, V., Brand, R. and Giessing, S. (2003) μ–ARGUS version 3.2 Software and User's Manual, Statistics Netherlands, Voorburg NL.

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*, 184(1), pp. 255–281.

Little, R. J. A. (1993) "Statistical Analysis of Masked Data", *Journal of Official Statistics,* Vol. 9, pp.407–426.

Nowok, B., Raab, G. M. and Dibben, C. (2016) "Synthpop: Bespoke Creation of Synthetic Data in R", *Journal of Statistical Software*, 74(11), pp.1–26.

Reiter, J. P. (2005) "Using CART to Generate Partially Synthetic, Public Use Microdata", *Journal of Official Statistics*, Vol. 21, pp. 441–462.

Rubin, D. B. (1993) "Discussion: Statistical Disclosure Limitation", *Journal of Official Statistics*, Vol. 9, pp. 462–468.

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.

Taub, J., Elliot, M., Pampaka and 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.

Taub, J., Elliot, M. and 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.

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

Woo, M., Reiter, J. P., Oganian, A. and Karr, A. F. (2009) "Global Measures of Data Utility for Microdata Masked for Disclosure Limitation", *The Journal of Privacy and Confidentiality*, Vol.1, No.1, pp.111–124.

(*中央大学経済学部教授　博士（経済学））
(**総務省統計研究研修所研究部統計技術向上支援課)