統計研究彙報　第79号　2022年３月　（111-130）

# 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

高部　勲†

# A study of creating synthetic data that contributes to promoting the utilization of official statistics microdata

TAKABE Isao

公的統計ミクロデータの更なる利用促進のために、教育・訓練用の疑似的なミクロデータの必要性が指摘されている。このような目的で現状、提供されている一般用ミクロデータは、元のデータの持つ変数間の相関などの構造を可能な限り保持するという点で課題がある。そこで、諸外国で研究が行われている合成データに関するモデルベースの手法を基に、現行の法令・制度上の制約を満たしつつ元のミクロデータの構造を可能な限り保持した疑似的なミクロデータを作成する方法について、分析・検討を行った。その結果、元のデータの秘匿性に配慮した上で、中間的な集計結果や回帰分析などの結果を事前に公開することにより、現行の制度に沿った形で、元のデータの持つ構造を保持した疑似的なデータを作成・提供できる可能性があることが示された。

キーワード　公的統計、ミクロデータ、合成データ

The need for pseudo microdata for education and training has been pointed out to promote the use of official statistical microdata. Currently provided pseudo microdata has a problem in that it does not retain the structure well such as the correlation between variables of the original data. In this paper, we analyzed and examined the method of creating pseudo microdata based on synthetic data method that retains the structure of the original microdata as much as possible while satisfying the current legal and institutional restrictions. As a result, by disclosing the results of intermediate aggregation and regression analysis, it was shown that there is a possibility that the retained pseudo microdata can be created and provided.

Key Words　Official statistics, Microdata, Synthetic data

† 立正大学データサイエンス学部　Email：takabe.isao@ris.ac.jp

## 1. はじめに

行政機関が実施する公的統計調査の結果を基に作成される集計前のレコード単位のデータ（調査票情報）や、それらを特定の個人又は法人その他の団体の識別（他の情報との照合による識別を含む。）ができないように加工したもの（匿名データ）は、公的統計ミクロデータと呼ばれている。このようなレコード単位のデータを活用することにより、集計されたデータからは得られない、母集団に関する様々な構造（変数間の相関関係や回帰モデルなどにより表現される様々な関係、細かな地域・区分におけるデータの傾向など）についての計量分析などが可能となる。欧米ではEBPM（Evidence Based Policy Making：根拠に基づく政策立案）の観点から、公的統計のミクロデータを活用した多くの実証分析が行われている。我が国においても、2007年及び2018年の２度の統計法の改正により、公的統計の二次的利用に関する制度が拡充され、公的統計ミクロデータの利用要件が緩和されるなど、その利用に関する環境の整備が図られてきている。また、公的統計ミクロデータ研究コンソーシアムが設立されるなど、官学連携の取組も進んできている（高部・徳富（2020）、高部（2020））。

このように、公的統計ミクロデータの研究・教育利用の拡充が図られているものの、その制度等に対する利用者の理解が十分でないことや、利用方法がイメージしにくいことなどから、ミクロデータの利用が必ずしも進んでいるとは言えない状況にあり、公的統計ミクロデータに親しんでもらうための教育用、あるいはミクロデータ利用を想定したプログラムテスト用の疑似的なミクロデータの必要性についても指摘されている（山口（2019））。こうした教育用・プログラムテスト用のレコード単位のデータとして、諸外国では、合成データ（Synthetic Data）と呼ばれる疑似的なデータの作成方法に関する研究が行われている（伊藤（2018））。なお、本稿において考察する合成データは、疑似的なデータの作成方法の一部をカバーするものであり、乱数などによる完全にランダムなデータも含め、ほかにも様々な疑似的なデータの作成方法がある点に留意する必要がある。

ただし、我が国では、公的統計ミクロデータから直接的にレコード単位のデータを作成・提供するという方式は、現行の制度上、認められていない。本稿では、こうした課題を踏まえつつ、諸外国において研究や作成事例の蓄積がある合成データの作成方法（モデルに基づく疑似的なデータの作成方法）を基に、中間的な集計表や回帰モデルなどの推定結果などを、データの秘匿性に配慮した上で事前に公開することにより、そこから疑似的なミクロデータを作成する方法に関して分析・検討を行う。

## 2. 合成データの考え方に基づく疑似的なミクロデータの作成方法の検討

### 2.1. 合成データに関する先行研究

諸外国では、合成データと呼ばれる疑似的なデータ作成方法に関する研究が進んでいる。合成データとは、様々な計量分析に利用されることを想定して、元のデータの持つ様々な構造をできる限り保持した形で作成された人工的・疑似的なレコード単位のデータであり、一般への公開・利用を想定したものである。合成データの作成・提供により、元のデータの持つ秘匿性を確保した上で、世帯・個人・企業等のレコード単位のレベルでデータを提供することが可能となり、ミクロデータの利用を希望する者は、合成データを用いて、プログラムの開発・テストや、実際のミクロデータの利用を想定した分析方法の検討などを行うことができるようになる（伊藤（2018）、高部・徳富（2020）、谷道（2019））。以下では、Drechsler (2011) 及び Alfons et al.(2011)に基づいて、合成データに関する先行研究とその作成・提供事例について、説明する。

諸外国では、欧米を中心に、機密性・安全性に配慮した形で公的統計ミクロデータを作成・提供するための方法として、合成データに関する研究が行われてきている。Rubin（1993）は、観測値を欠測データとして扱い、欠測値の補完に用いられている多重代入法（multiple imputation）の考え方に基づいて合成データを作成することを提案した。この方法により作成された疑似データは、Fully synthetic dataset と呼ばれる。Little（1993）は、全ての変数ではなく、機密性の高い一部の変数の情報を複数の代入値に置き換える手法を提案した。この方法により作成された疑似データは、Partial synthetic dataset と呼ばれる。こうした多重代入法による方法は、その後も継続して研究され、様々な方法に拡張されている（Raghunathan et al（2003）、Drechsler et al（2008）、Reiter（2009））。モデルを明示的に仮定するのではなく、ノンパラメトリックな手法により合成データを作成するための方法についても研究が行われており、CART やランダムフォレストの手法に基づくノンパラメトリックな方法が提案されている（Reiter (2005)、Caiola and Reiter (2010)）。企業のデータに関しては、売上高を含め、非常に歪んだ分布を持つ変数が含まれている場合が多く、こうした状況に対応するための手法についても研究が行われている。Woodcock and Benedetto（2009）では、非常に歪んだ、または多峰性の分布を持つ変数に関して、カーネル密度推定に基づく方法を提案している。

カテゴリー変数を含む大きなサンプルサイズの世帯・個人を対象とした標本調査に対応する合成データの作成手法も提案されている（Munnich et al（2003）、Munnich and Schurle（2003））。この方法では、標本の各層に対し、乗率を考慮した形で年齢・性別などの変数の割合が元のデータと同じになるように世帯を抽出し、これらの変数を基に、元のデータから推定したモデルを用いて他の変数を生成する。この手法は、欧州連合の所得と生活条件に関する統計（EU-SILC: European Union Statistics on Income and Living Conditions）などに適用されている。上記の EU-SILC の方法では、パレート分布を利用することにより歪んだ分布を持つ変数の生成にも対応しており、また、ロジットモデル及び重回帰モデルに基づく２段階の推定方法により、0 などの特定の値に集中している変数の生成にも対応している（Alfons et al. (2011), Templ and Alfons (2010)）。

諸外国では、合成データの考え方に基づいて作成された疑似的なミクロデータの提供が行われており（Alfons et al.（2011））、実際のミクロデータから合成データを作成するための R のパッケージも開発・提供されている（Templ et al.(2017)、Nowok et al. (2016)）。ただし、これらのパッケージは、ミクロデータから直接的にレコード単位のデータを作成する仕様になっており、我が国の現行の制度上、こうしたパッケージ・ソフトウェアによって疑似データを作成・提供することは認められていない。

### 2.2. 我が国における公的統計に関する疑似データの作成・提供に関する課題と対応

我が国においても、疑似的なミクロデータの作成・提供に関する研究が行われている。独立行政法人統計センターでは、元のデータの持つ秘匿性に配慮した上で、中間的な集計表を事前に公開し、その集計表を基に多変数正規乱数を付加することにより、疑似的なミクロデータを作成する方法について研究を行っており、この方法で作成された疑似的なデータが提供され、多くのユーザに利用されている（山口・伊藤・秋山（2013））。この方法は、量的変数の値が多変量（対数）正規分布にしたがうことを仮定し、公的統計ミクロデータから作成した高次元の集計表を用いて、多変量（対数）正規乱数を生成するものであり、一部の質的変数の分布と、量的変数間の相関関係のみに着目する非常にシンプルなモデルを仮定している。この点で、本稿で考察するような、着目する変数とそれ以外の変数との関係を考慮したモデルを構築する合成データの方法とは異なるもので

ある。この方法では、質的変数については種類をかなり限定しており、量的変数については多変量正規分布に従うという強い仮定を置いており、変数間の構造を十分に反映できていないなど、その方法論に関しては課題が残る。

こうした状況に対応するために、前節で紹介した合成データの方法を適用することが考えられる。ただし、多重代入法に基づく手法などを適用して、公的統計ミクロデータから直接に、疑似的なレコード単位のデータを作成することは、現行の法令上は認められていない。

そこで本稿では、統計センターの一般用ミクロデータのように、中間的な集計表を作成する方法と合成データの方法を組み合わせる形で、事前に集計した統計表と回帰モデルの結果を基に、モデルを用いて疑似データを作成する方法について分析・検討を行う。具体的には、合成データの様々な作成方法の中でも、モデルを用いたパラメトリックな手法であり、中間的な集計表やモデルの推定結果を事前に公表することで、我が国の法令に沿った形で適用が可能である、Nowok et al. (2016), Alfons et al. (2011), Templ and Alfons (2010) で取り上げられている EU-SILK（欧州で作成・提供されている疑似的なデータ）の作成方法を応用した疑似データの作成・提供について研究を行う。

上記の方法による合成データの作成方法は、一部のレコード・変数を人工的に欠測させ、事前に構築した重回帰モデルやロジットモデルを用いて疑似データを発生させる方法であり（Templ (2017)）、変数間の関係を保持したデータの作成が可能となる。合成データのイメージを示したものが、以下の図１である。

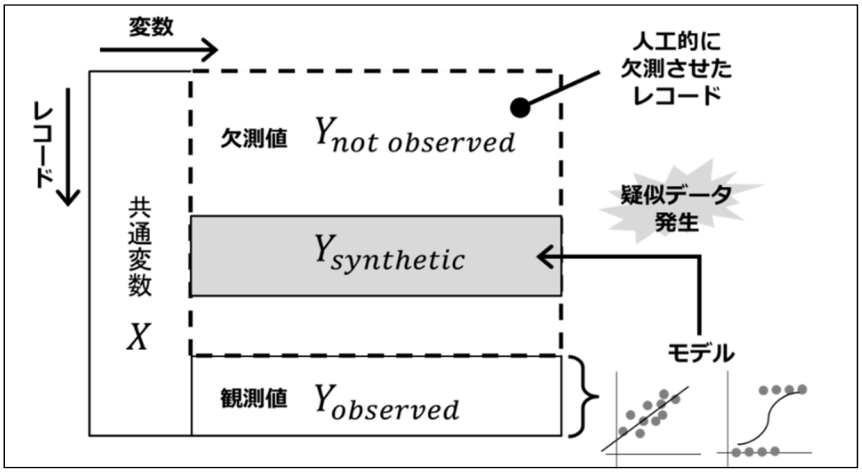


図１　合成データの概要

前述のパッケージでは、内部的に作成した統計モデルから疑似データを作成しているが、統計センターの作成する一般用ミクロデータと類似の方式で、このようなモデルの推定結果を、秘匿性・安全性に配慮した上で、事前に統計表の形で公開しておくことにより、現行制度に沿った形で合成データの考え方に基づく疑似データを作成することが可能となる（ただし，集計表だけではなく，重回帰モデルなどの推定結果や，その残差に関する情報についても公開する点が，統計センターの作成する一般用ミクロデータの方法と異なる）。

## 3. 実データに基づく合成データの作成

### 3.1. 分析に用いたデータの概要

本稿では、商用データを用いて、合成データの作成を試みた。使用したデータは、帝国データバンクの「COSMOS2」企業概要ファイルである。分析に使用したデータの概要については以下のとおり。なお、秘匿性を考慮し、地域などの詳細に関する情報は記載していない。

○データのサイズ：7,558 レコード
○データの時点：平成 24 年 2 月時点
○分析対象：資本金 300 万円以上 5,000 万円未満の、株式会社又は有限会社
○使用した変数：
(1) 地域（３地域：地域１，地域２，地域３）《離散変数》
(2) 産業（４区分：建設業、製造業、小売業、それ以外）《離散変数》
(3) 経営組織（２区分：株式会社、有限会社）《離散変数》
(4) 開設年（３区分：～1984 年，1985 年～1994 年，1995 年～）《離散変数》
(5) 資本金（４区分：300 万円以上 500 万円未満，500 万円以上 1000 万円未満，1000 万円以上 2000 万円未満，2000 万円以上）《離散変数》
(6) 従業員数（1 を足してから対数変換）《連続変数》
(7) 売上高（1 を足してから対数変換）《連続変数》

今回用いたデータでは、各変数について、欠測値は存在しない。データに欠測値が含まれる場合には、それらを単純に削除してしまうと、変数を発生させるためのモデルの結果や最終的な疑似データの精度に影響を及ぼす可能性があるので注意が必要である。なお、欠測値がある場合には、元データに近い疑似データを作成するという観点から、変数ごとの欠測の状況を考慮した上で、元データの状況に近い形で欠測値を発生させることが望ましいとされている（Templ, M. (2017)）。なお，資本金については，各種の制度などの関係で特定の値に集中しているという特殊性から、連続変数として扱うことが困難であったため、今回の分析では資本金額の階級で分類した離散変数として扱っている。

### 3.2. 各変数の記述統計量

今回使用したデータに関する記述統計量（カテゴリ別企業数・構成割合、要約統計量）を示したものが，以下の表１～表６である。なお，連続変数の要約統計量については、秘匿性の観点から、最大値及び最小値を表示していない。

表１　記述統計量：地域（カテゴリ別企業数、構成割合）

| | 地域１ | 地域２ | 地域３ | 合計 |
|---|---|---|---|---|
| レコード数 | 2970 | 3101 | 1487 | 7558 |
| 構成割合 | 0.39 | 0.41 | 0.20 | 1.00 |

表２　記述統計量：産業（カテゴリ別企業数、構成割合）

| | 建設業 | 製造業 | 小売業 | その他 | 合計 |
|---|---|---|---|---|---|
| レコード数 | 2389 | 1613 | 1942 | 1614 | 7558 |
| 構成割合 | 0.32 | 0.21 | 0.26 | 0.21 | 1.00 |

表３　記述統計量：産業（カテゴリ別企業数、構成割合）

| | 株式会社 | 有限会社 | 合計 |
|---|---|---|---|
| レコード数 | 4381 | 3177 | 7558 |
| 構成割合 | 0.58 | 0.42 | 1.00 |

表４　記述統計量：開設年（カテゴリ別企業数、構成割合）

| | ～1984年 | 1985年～1994年 | 1995年～ | 合計 |
|---|---|---|---|---|
| レコード数 | 3018 | 2656 | 1884 | 7558 |
| 構成割合 | 0.40 | 0.35 | 0.25 | 1.00 |

表５　記述統計量：資本金（カテゴリ別企業数、構成割合）

| | 300万円～500万円 | 500万円～1000万円 | 1000万円～2000万円 | 2000万円～ | 合計 |
|---|---|---|---|---|---|
| レコード数 | 1896 | 1170 | 2864 | 1628 | 7558 |
| 構成割合 | 0.25 | 0.15 | 0.38 | 0.22 | 1.00 |

表６　記述統計量：従業員数、売上高（要約統計量）

| 変数名 | 平均値 | 標準偏差 | 第１四分位数 | 中央値 | 第３四分位数 |
|---|---|---|---|---|---|
| 従業員数 | 12.32 | 40.10 | 2.00 | 5.00 | 12.00 |
| 売上高 | 365.87 | 2180.09 | 42.00 | 100.00 | 265.00 |

## 4. 実データに基づく合成データの作成

### 4.1. 疑似的なデータ作成の手順の概要

本稿では、企業に関する商用データを活用した合成データの作成を試みる。具体的な方針としては、秘匿性・安全性に配慮した上で中間的な統計表・モデルの推定結果をいったん公開し、その情報を基にモデルからデータを発生させる方法について検討した。その際に、単純なクロス集計表のほか、重回帰モデルやロジットモデルの推定結果や、残差の分布に関する要約統計量についても作成・公開する方法について検討した。

合成データの作成方法としては、Nowok et al. (2016), Alfons et al. (2011), Templ and Alfons (2010) で取り上げられている EU-SILK（欧州で作成・提供されている疑似的なデータ）の作成方法を参考にした。具体的には、以下の手順により、疑似的なミクロデータを作成した（Templ (2017)、Alfons et al. (2011)）。

【手順１】: 地域×産業の集計表の作成
【手順２】: 上記の集計表に合うような（地域・産業に属する）7,558 レコードを生成
【手順３】: 元データから事前に推定したモデルで変数を逐次的に推測

以下では、これらの手順に沿った形で、データの作成方法について説明する。作成したデータの分析結果については、次節で解説する。なお、秘匿性の観点からの安全性に関する統計表のチェックについては、独立行政法人統計センターによる「オンサイト利用における分析結果等の提供に関する標準的なチェック内容の解説と例」（独立行政法人統計センター（2019)）を参考にした。

### 4.1.1.【手順１】地域×産業の集計表の作成

地域及び産業でクロス集計を行った結果が、以下の表７である。各セルの度数には、各セルが 1 以上 10 未満の極端に小さな値は含まれておらず、また、行計又は列計の 90%超を占めるセルがないことから、「オンサイト利用における分析結果等の提供に関する標準的なチェック内容の解説と例」（独立行政法人統計センター（2019））における統計表のチェック内容を考慮すると，秘匿性・安全性は保たれており、この程度の粒度の集計表については、公開しても問題はないと考えられる。

表７　産業及び地域に関する周辺分布の表

| | 建設業 | 製造業 | 卸・小売業 | 左記以外 | 合計 |
|---|---|---|---|---|---|
| 地域１ | 1065 | 748 | 645 | 512 | 2970 |
| 地域２ | 935 | 508 | 831 | 827 | 3101 |
| 地域３ | 389 | 357 | 466 | 275 | 1487 |
| 合計 | 2389 | 1613 | 1942 | 1614 | 7558 |

### 4.1.2.【手順２】周辺分布に合うレコードの生成

上記の周辺分布に合うような（地域・産業に属する）レコードを生成する。具体的には、2 変数（地域及び産業）からなる 7,558 レコードを作成し、地域及び産業の組合せが上記のクロス集計表と一致するように、（例えば地域１と建設業の組合せのレコード数が 1,065 となるように）地域及び産業の属性を割り振る。このようにして作成した、２つの変数のみを含むデータからは、個別の企業を特定することはできないことから、秘匿性・安全性は確保されている。

### 4.1.3.【手順３】モデルに基づく変数の逐次的な推測

元データから事前に推定したモデルにより、変数を逐次的に推測する。今回の分析では、モデルにより変数の値を推測する際に、最初に離散変数の値を、カテゴリー数の少ない方から順に推測し、次に連続変数の値を推測している。具体的には、以下のようなモデル（重回帰モデル、多項ロジットモデル、順序ロジットモデル）を逐次的に用いて、経営組織、開設年、従業者数、売上高の順に、疑似的な変数の値を順次、推測・発生させる。

(1)「経営組織」を推測する２項ロジットモデルの構築
- 手順２で作成したデータを基に、「地域」及び「産業」から「経営組織」を推測する２項ロジットモデルを構築
- 構築したモデルにより、各レコードについて「経営組織」を推測・発生

(2)「開設年」を推測する順序ロジットモデルの構築
- 「地域」、「産業」及び「経営組織」から「開設年」を推測する順序ロジットモデルを構築
- 構築したモデルにより、各レコードについて「経営組織」を推測・発生

(3)「資本金」を推測する順序ロジットモデルの構築
- 「地域」、「産業」、「経営組織」及び「開設年」から「資本金」を推測する順序ロジッ

トモデルを構築
・構築したモデルにより、各レコードについて「資本金」を推測・発生

(4)「従業員数」を推測する重回帰モデルの構築
・「地域」、「産業」、「経営組織」、「開設年」及び「資本金」から「従業員数」の対数を推測する重回帰モデルを構築
・構築したモデルにより、各レコードについて「従業者数」の対数の予測値を算出
・重回帰モデルの残差から中央値（$MD$)、中央絶対偏差（$MAD$）を算出し、正規分布（$N(MD, MAD^2)$）に従う正規乱数を「従業者数」の予測値に加える（ただし 0 を下回る場合には 0 を代入する（折返し処理））

(5)「売上高」を推測する重回帰モデルの構築
・「地域」、「産業」、「経営組織」、「開設年」、「資本金」及び「従業員数」から「売上高」の対数を推測する重回帰モデルを構築
・重回帰モデルの残差から中央値（$MD$)、中央絶対偏差（$MAD$）を算出し、正規分布（$N(MD, MAD^2)$）に従う正規乱数を「売上高」の予測値に加える（ただし 0 を下回る場合には 0 を代入する（折返し処理））

ここで、各モデルの構築に当たっては、以下の R の関数を使用した。
○重回帰モデル：「lm」関数
○２項ロジットモデル：「glm」関数
○順序ロジットモデル：パッケージ「MASS」に含まれる「polr」関数

## 4.2. 元データに基づくモデルの推定結果

前節で示した各モデルの推定結果及び残差の情報（中央値（$MD$)、中央絶対偏差（$MAD$））について示したものが、以下の表８～表 12 である。なお、ここで示した推定結果は、いずれも自由度が 10 以上であり、残差に関する詳細な情報は示されておらず（残差の中央値及び中央絶対偏差のみを欄外に表示)、これらの結果には、個別の企業を特定できるような情報は含まれていないことから、秘匿性・安全性は確保されている。このような回帰モデルの推定結果を、秘匿性に十分配慮した上で公開することにより、「地域」及び「産業」の変数から順次、疑似的なデータを生成することが可能となる。

表８　「経営組織」を推測するための２項ロジットモデルの推定結果

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| 定数項 | -0.01634 | 0.04895 | -0.334 | 0.73852 |
| 地域：地域１【ベースライン】 | | | | |
| 地域２ | -0.20972 | 0.05284 | -3.969 | 7.22E-05 |
| 地域３ | -0.12303 | 0.06501 | -1.892 | 0.05843 |
| 産業：建設業【ベースライン】 | | | | |
| 製造業 | -0.57565 | 0.06702 | -8.59 | < 2e-16 |
| 小売業 | -0.23169 | 0.06207 | -3.732 | 0.00019 |
| その他 | -0.07967 | 0.06513 | -1.223 | 0.22123 |

表９　「開設年」を推測するための順序ロジットモデルの推定結果

| | | Value | Std. Error | t value |
|---|---|---|---|---|
| 地域： | 地域１【ベースライン】 | | | |
| | 地域２ | 0.02545 | 0.04886 | 0.5209 |
| | 地域３ | -0.21102 | 0.06031 | -3.4991 |
| 産業： | 建設業【ベースライン】 | | | |
| | 製造業 | -0.5192 | 0.06083 | -8.535 |
| | 小売業 | -0.50666 | 0.05838 | -8.6791 |
| | その他 | 0.10961 | 0.06043 | 1.814 |
| 経営組織： | 株式会社【ベースライン】 | | | |
| | 有限会社 | 1.03227 | 0.04475 | 23.0668 |
| 定数項： | ～1984年\|1985年～1994年 | -0.2457 | 0.051 | -4.8213 |
| | 1985年～1994年\|1995年～ | 1.3979 | 0.0538 | 26.0045 |

表 10　「資本金」を推測するための順序ロジットモデルの推定結果

| | | Value | Std. Error | t value |
|---|---|---|---|---|
| 地域： | 地域１【ベースライン】 | | | |
| | 地域２ | -0.14648 | 0.05117 | -2.8627 |
| | 地域３ | -0.04378 | 0.06243 | -0.7012 |
| 産業： | 建設業【ベースライン】 | | | |
| | 製造業 | -0.03121 | 0.06348 | -0.4917 |
| | 小売業 | -0.23733 | 0.06098 | -3.8917 |
| | その他 | -0.40506 | 0.06502 | -6.2302 |
| 経営組織： | 株式会社【ベースライン】 | | | |
| | 有限会社 | -3.59306 | 0.06609 | -54.3622 |
| 開設年： | ～1984年【ベースライン】 | | | |
| | 1985年～1994年 | -1.27705 | 0.04775 | -26.7425 |
| | 1995年～ | -0.28779 | 0.03997 | -7.1992 |
| 定数項： | 300万円以上500万円未満\|500万円以上1000万円未満 | -3.6484 | 0.0759 | -48.0615 |
| | 500万円以上1000万円未満\|1000万円以上2000万円未満 | -2.2158 | 0.0666 | -33.2767 |
| | 1000万円以上2000万円未満\|2000万円以上 | 0.775 | 0.0557 | 13.9034 |

表 11　「従業員数」を推測するための重回帰モデルの推定結果

| | | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|---|
| 定数項 | | 1.99846 | 0.02894 | 69.048 | < 2e-16 |
| 地域： | 地域１【ベースライン】 | | | | |
| | 地域２ | 0.01456 | 0.02448 | 0.595 | 0.552 |
| | 地域３ | -0.04671 | 0.03006 | -1.554 | 0.12 |
| 産業： | 建設業【ベースライン】 | | | | |
| | 製造業 | 0.38238 | 0.03072 | 12.448 | < 2e-16 |
| | 小売業 | -0.15804 | 0.0293 | -5.395 | 7.08E-08 |
| | その他 | -0.17644 | 0.03063 | -5.761 | 8.69E-09 |
| 経営組織： | 株式会社【ベースライン】 | | | | |
| | 有限会社 | -0.39709 | 0.03424 | -11.596 | < 2e-16 |
| 開設年： | ～1984年【ベースライン】 | | | | |
| | 1985年～1994年 | -0.12031 | 0.02185 | -5.507 | 3.78E-08 |
| | 1995年～ | 0.0924 | 0.01918 | 4.818 | 1.48E-06 |
| 資本金： | 300万円以上500万円未満【ベースライン】 | | | | |
| | 500万円以上1000万円未満 | 0.75028 | 0.03542 | 21.183 | < 2e-16 |
| | 1000万円以上2000万円未満 | 0.28027 | 0.02304 | 12.163 | < 2e-16 |
| | 2000万円以上 | 0.1226 | 0.02488 | 4.927 | 8.52E-07 |

中央値(MD)：0.003095962
中央絶対偏差(MAD)：0.8777895

表 12　「売上高」を推測するための重回帰モデルの推定結果

| | | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|---|
| 定数項 | | 2.948865 | 0.029566 | 99.737 | < 2e-16 |
| 地域： | 地域１【ベースライン】 | | | | |
| | 地域２ | 0.105774 | 0.019574 | 5.404 | 6.73E-08 |
| | 地域３ | 0.007828 | 0.024043 | 0.326 | 0.745 |
| 産業： | 建設業【ベースライン】 | | | | |
| | 製造業 | -0.103277 | 0.024815 | -4.162 | 3.19E-05 |
| | 小売業 | 0.683146 | 0.023473 | 29.104 | < 2e-16 |
| | その他 | -0.147661 | 0.024546 | -6.016 | 1.87E-09 |
| 経営組織： | 株式会社【ベースライン】 | | | | |
| | 有限会社 | -0.152575 | 0.027627 | -5.523 | 3.45E-08 |
| 開設年： | ～1984年【ベースライン】 | | | | |
| | 1985年～1994年 | 0.085169 | 0.017506 | 4.865 | 1.17E-06 |
| | 1995年～ | 0.01336 | 0.015361 | 0.87 | 0.384 |
| 資本金： | 300万円以上500万円未満【ベースライン】 | | | | |
| | 500万円以上1000万円未満 | 0.350539 | 0.029154 | 12.024 | < 2e-16 |
| | 1000万円以上2000万円未満 | 0.14328 | 0.018607 | 7.7 | 1.53E-14 |
| | 2000万円以上 | 0.083222 | 0.01993 | 4.176 | 3.00E-05 |
| 従業員数（対数） | | 0.912558 | 0.009206 | 99.129 | < 2e-16 |

### 4.3. 疑似的なデータの作成結果

前節の方法に基づいて、モデルから推定・発生させた疑似的なデータについて、元のデータとの分布の比較などを行う。

#### 4.2.1.離散変数の度数分布・構成割合の比較

「経営組織」及び「開設年」について集計した結果を比較したものが、以下の表13～表15である。これらの表を見ると、いずれの表においても、元データと疑似データとで、内訳の度数・構成比がほぼ等しくなっていることがわかる。

表 13　経営組織の比較

| 【レコード数】 | 株式会社 | 有限会社 | 合計 |
|---|---|---|---|
| 元データ | 4381 | 3177 | 7558 |
| 疑似データ | 4416 | 3142 | 7558 |

| 【構成比】 | 株式会社 | 有限会社 | 合計 |
|---|---|---|---|
| 元データ | 0.58 | 0.42 | 1.00 |
| 疑似データ | 0.58 | 0.42 | 1.00 |

表 14　開設年の比較

| 【レコード数】 | ～1984年 | 1985年～1994年 | 1995年～ | 合計 |
|---|---|---|---|---|
| 元データ | 3018 | 2656 | 1884 | 7558 |
| 疑似データ | 3092 | 2563 | 1903 | 7558 |

| 【構成比】 | ～1984年 | 1985年～1994年 | 1995年～ | 合計 |
|---|---|---|---|---|
| 元データ | 0.40 | 0.35 | 0.25 | 1.00 |
| 疑似データ | 0.41 | 0.34 | 0.25 | 1.00 |

表 15　資本金の比較

| 【レコード数】 | 300万円～500万円 | 500万円～1000万円 | 1000万円～2000万円 | 2000万円～ | 合計 |
|---|---|---|---|---|---|
| 元データ | 1896 | 1170 | 2864 | 1628 | 7558 |
| 疑似データ | 1873 | 1169 | 2920 | 1596 | 7558 |

| 【構成比】 | 300万円～500万円 | 500万円～1000万円 | 1000万円～2000万円 | 2000万円～ | 合計 |
|---|---|---|---|---|---|
| 元データ | 0.25 | 0.15 | 0.38 | 0.22 | 1.00 |
| 疑似データ | 0.25 | 0.15 | 0.39 | 0.21 | 1.00 |

### 4.2.2. 連続変数の分布の比較

「従業員数」及び「売上高」についての分布を箱ひげ図により比較したものが、以下の図２である。ただし秘匿性の観点から、横軸の目盛りの数値は示していない。これらの表を見ると、中央値や四分位範囲に大きな違いはないものの、全体の範囲については、重回帰モデルで推定した結果に正規乱数を付与している疑似データの方が狭くなっている。

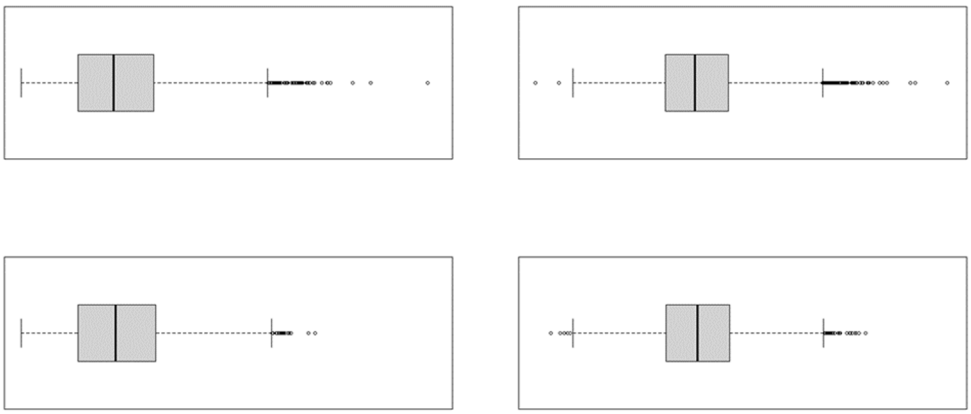

図２　「従業員数」及び「売上高」の分布の比較
（左：従業員数　右：売上高
上段：元データ　下段：疑似データ）

### 4.2.3. 売上高を予測する重回帰モデルの比較

元データと疑似データにおいて、売上高を、それ以外の変数で予測する重回帰モデルを推定した結果が表16である。元データと疑似データのそれぞれのデータで推定したモデルの結果を比較すると、ほとんどの変数で、回帰係数の符号は一致しており、係数の値もそれほど大きくは異なっておらず、ほぼ近い値となっていることがわかる。ただし、元データと疑似データとで、「地域」における「地域３」の回帰係数の符号が逆転している。この変数はp値が大きく（元データ：0.745、疑似データ：0.493）、推定値の誤差が大きいことが原因であると考えられる。

表 16　「売上高」を推測するための重回帰モデルの推定結果の比較

| | | 元データ | | | | 疑似データ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Estimate | Std. Error | z value | Pr(>\|z\|) | Estimate | Std. Error | z value | Pr(>\|z\|) |
| 定数項 | | 2.948865 | 0.029566 | 99.737 | < 2e-16 | 2.945073 | 0.027631 | 106.584 | < 2e-16 |
| 地域： | 地域１【ベースライン】 | | | | | | | | |
| | 地域２ | 0.105774 | 0.019574 | 5.404 | 6.73E-08 | 0.124272 | 0.017619 | 7.053 | 1.90E-12 |
| | 地域３ | 0.007828 | 0.024043 | 0.326 | 0.745 | -0.014865 | 0.021664 | -0.686 | 0.493 |
| 産業： | 建設業【ベースライン】 | | | | | | | | |
| | 製造業 | -0.103277 | 0.024815 | -4.162 | 3.19E-05 | -0.126383 | 0.022369 | -5.65 | 1.66E-08 |
| | 小売業 | 0.683146 | 0.023473 | 29.104 | < 2e-16 | 0.673735 | 0.02099 | 32.098 | < 2e-16 |
| | その他 | -0.147661 | 0.024546 | -6.016 | 1.87E-09 | -0.166643 | 0.022165 | -7.518 | 6.19E-14 |
| 経営組織： | 株式会社【ベースライン】 | | | | | | | | |
| | 有限会社 | -0.152575 | 0.027627 | -5.523 | 3.45E-08 | -0.140898 | 0.023638 | -5.961 | 2.63E-09 |
| 開設年： | ～1984年【ベースライン】 | | | | | | | | |
| | 1985年～1994年 | 0.085169 | 0.017506 | 4.865 | 1.17E-06 | 0.092696 | 0.015629 | 5.931 | 3.14E-09 |
| | 1995年～ | 0.01336 | 0.015361 | 0.87 | 0.384 | 0.008486 | 0.013555 | 0.626 | 0.531 |
| 資本金： | 300万円以上500万円未満【ベースライン】 | | | | | | | | |
| | 500万円以上1000万円未満 | 0.350539 | 0.029154 | 12.024 | < 2e-16 | 0.372735 | 0.02591 | 14.386 | < 2e-16 |
| | 1000万円以上2000万円未満 | 0.14328 | 0.018607 | 7.7 | 1.53E-14 | 0.161312 | 0.016809 | 9.597 | < 2e-16 |
| | 2000万円以上 | 0.083222 | 0.01993 | 4.176 | 3.00E-05 | 0.079515 | 0.016976 | 4.684 | 2.86E-06 |
| 従業員数（対数） | | 0.912558 | 0.009206 | 99.129 | < 2e-16 | 0.908968 | 0.0092 | 98.798 | < 2e-16 |

## 5. 変数の発生の順序を変更した場合の影響

### 5.1. モデルの推定・変数の発生の順序の変更

今回の分析では、モデルにより変数の値を推測する際に、最初に離散変数の値を、カテゴリー数の少ない方から順に推測し、次に連続変数の値を推測している。具体的には、「経営組織」、「開設年」、「従業員数」、「売上高」の順に、変数の値を推測している。今回の分析を行う上で参考にしたEU-SILK（Nowok et al. (2016), Alfons et al. (2011), Templ and Alfons (2010)）の作成においても、離散変数が先で、連続変数は後になるような順序で推定が行われている。

ただし、推測を行う際の変数の順序は、今回の方法に限られるものではなく、他の順序で発生させることも可能である。その際に、推測を行う変数の順序が、最終的な疑似データの内容や、当該データを用いた分析の結果に影響を与える可能性もある。そこで、発生させる変数の順序を変更した場合の影響を分析する。具体的には、以下のように、先に連続変数、後に離散変数を推測するモデルを元データから推定し、変数を発生させるように、順序を変更する。使用するモデル（重回帰、多項ロジットモデル、順序ロジットモデル）は前節と同様であり、連続変数における、残差の中央値等に基づく正規分布からの乱数の発生・付加の方法も同様である。

(1)「従業員数」を推測する重回帰モデルの構築
(2)「売上高」を推測する重回帰モデルの構築
(3)「経営組織」を推測する２項ロジットモデルの構築
(4)「開設年」を推測する順序ロジットモデルの構築
(5)「資本金」を推測する順序ロジットモデルの構築

### 5.2. モデルの推定結果

前節に置いて推定したモデルと、発生させる変数の順序を変更したモデルとの推定結果の比較を行ったものが、以下の表17～表21である。

表 17　「従業員数」を推測するための重回帰モデルの推定結果

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| 定数項 | 1.776213 | 0.026524 | 66.968 | < 2e-16 |
| 地域：地域１【ベースライン】 | | | | |
| 地域２ | 0.050134 | 0.028261 | 1.774 | 0.07611 |
| 地域３ | -0.004477 | 0.034775 | -0.129 | 0.897568 |
| 産業：建設業【ベースライン】 | | | | |
| 製造業 | 0.53635 | 0.035195 | 15.239 | < 2e-16 |
| 小売業 | -0.112421 | 0.033456 | -3.36 | 0.000783 |
| その他 | -0.225837 | 0.035293 | -6.399 | 1.66E-10 |

中央値(MD)：-0.05435458

中央絶対偏差(MAD)：1.002292

表 18　「売上高」を推測するための重回帰モデルの推定結果

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| 定数項 | 2.691199 | 0.023854 | 112.82 | < 2e-16 |
| 地域：地域１【ベースライン】 | | | | |
| 地域２ | 0.11347 | 0.020137 | 5.635 | 1.81E-08 |
| 地域３ | 0.016157 | 0.024773 | 0.652 | 0.514 |
| 産業：建設業【ベースライン】 | | | | |
| 製造業 | -0.119074 | 0.025454 | -4.678 | 2.95E-06 |
| 小売業 | 0.682748 | 0.023851 | 28.625 | < 2e-16 |
| その他 | -0.152061 | 0.02521 | -6.032 | 1.70E-09 |
| 従業員数（対数） | 1.008038 | 0.008197 | 122.97 | < 2e-16 |

中央値(MD)：-0.04752098

中央絶対偏差(MAD)：0.6881855

表 19　「経営組織」を推測するための２項ロジットモデルの推定結果

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| 定数項 | 3.07518 | 0.12655 | 24.3 | < 2e-16 |
| 地域：地域１【ベースライン】 | | | | |
| 地域２ | -0.18168 | 0.05883 | -3.088 | 2.01E-03 |
| 地域３ | -0.16766 | 0.0721 | -2.325 | 0.020048 |
| 産業：建設業【ベースライン】 | | | | |
| 製造業 | -0.28566 | 0.07517 | -3.8 | 1.44E-04 |
| 小売業 | -0.08003 | 0.0726 | -1.102 | 0.270277 |
| その他 | -0.41643 | 0.07322 | -5.687 | 1.29E-08 |
| 従業員数（対数） | -0.46103 | 0.04287 | -10.754 | < 2e-16 |
| 売上高（対数） | -0.50839 | 0.03515 | -14.464 | < 2e-16 |

表 20　「開設年」を推測するための順序ロジットモデルの推定結果

| | | Value | Std. Error | t value |
|---|---|---|---|---|
| 地域： | 地域１【ベースライン】 | | | |
| | 地域２ | 0.01965 | 0.04919 | 0.3996 |
| | 地域３ | -0.22795 | 0.06068 | -3.7565 |
| 産業： | 建設業【ベースライン】 | | | |
| | 製造業 | -0.40334 | 0.06175 | -6.5321 |
| | 小売業 | -0.59249 | 0.06201 | -9.5548 |
| | その他 | 0.04226 | 0.06112 | 0.6914 |
| 従業員数（対数） | | -0.3418 | 0.03552 | -9.6233 |
| 売上高（対数） | | 0.0401 | 0.02876 | 1.3945 |
| 経営組織：株式会社【ベースライン】 | | | | |
| | 有限会社 | 0.79752 | 0.04895 | 16.2919 |
| 定数項： | ～1984年\|1985年～1994年 | -0.8034 | 0.1106 | -7.2638 |
| | 1985年～1994年\|1995年～ | 0.8697 | 0.111 | 7.8378 |

表 21　「資本金」を推測するための順序ロジットモデルの推定結果

| | | Value | Std. Error | t value |
|---|---|---|---|---|
| 地域： | 地域１【ベースライン】 | | | |
| | 地域２ | -0.19688 | 0.05258 | -3.7445 |
| | 地域３ | -0.00991 | 0.06415 | -0.1545 |
| 産業： | 建設業【ベースライン】 | | | |
| | 製造業 | -0.19915 | 0.06605 | -3.015 |
| | 小売業 | -0.41801 | 0.06626 | -6.3087 |
| | その他 | -0.22419 | 0.06665 | -3.3635 |
| 従業員数（対数） | | 0.1733 | 0.03784 | 4.5803 |
| 売上高（対数） | | 0.43913 | 0.03135 | 14.0093 |
| 経営組織：株式会社【ベースライン】 | | | | |
| | 有限会社 | -3.26034 | 0.06767 | -48.1775 |
| 開設年： | ～1984年【ベースライン】 | | | |
| | 1985年～1994年 | -1.23411 | 0.04849 | -25.4486 |
| | 1995年～ | -0.35968 | 0.04092 | -8.7898 |
| 定数項： | 300万円以上500万円未満\|500万円以上1000万円未満 | -1.3502 | 0.1236 | -10.9204 |
| | 500万円以上1000万円未満\|1000万円以上2000万円未満 | 0.1424 | 0.1202 | 1.1854 |
| | 1000万円以上2000万円未満\|2000万円以上 | 3.4122 | 0.1263 | 27.0202 |

## 5.3. 疑似的なデータの作成結果

### 5.3.1.離散変数の度数分布・構成割合の比較

「経営組織」及び「開設年」について集計した結果を比較したものが、以下の表22～表24である。これらの表を見ると、いずれの表においても、元データと疑似データとで、内訳の度数・構成比がほぼ等しくなっていることがわかる。離散変数の疑似データの分布については、本稿の分析の範囲では、変数の発生の順序はそれほど大きな影響を与えていないことがわかる。

表 22　経営組織の比較

| 【レコード数】 | 株式会社 | 有限会社 | 合計 |
|---|---|---|---|
| 元データ | 4381 | 3177 | 7558 |
| 疑似データ | 4310 | 3248 | 7558 |

| 【構成比】 | 株式会社 | 有限会社 | 合計 |
|---|---|---|---|
| 元データ | 0.58 | 0.42 | 1.00 |
| 疑似データ | 0.57 | 0.43 | 1.00 |

表 23　開設年の比較

| 【レコード数】 | ～1984年 | 1985年～1994年 | 1995年～ | 合計 |
|---|---|---|---|---|
| 元データ | 3018 | 2656 | 1884 | 7558 |
| 疑似データ | 3026 | 2667 | 1865 | 7558 |

| 【構成比】 | ～1984年 | 1985年～1994年 | 1995年～ | 合計 |
|---|---|---|---|---|
| 元データ | 0.40 | 0.35 | 0.25 | 1.00 |
| 疑似データ | 0.40 | 0.35 | 0.25 | 1.00 |

表 24　資本金の比較

| 【レコード数】 | 300万円～500万円 | 500万円～1000万円 | 1000万円～2000万円 | 2000万円～ | 合計 |
|---|---|---|---|---|---|
| 元データ | 1896 | 1170 | 2864 | 1628 | 7558 |
| 疑似データ | 1923 | 1186 | 2946 | 1503 | 7558 |

| 【構成比】 | 300万円～500万円 | 500万円～1000万円 | 1000万円～2000万円 | 2000万円～ | 合計 |
|---|---|---|---|---|---|
| 元データ | 0.25 | 0.15 | 0.38 | 0.22 | 1.00 |
| 疑似データ | 0.25 | 0.16 | 0.39 | 0.20 | 1.00 |

### 5.3.2. 連続変数の分布の比較

「従業者数」及び「売上高」についての分布を箱ひげ図により比較したものが、以下の図３である。前掲と同様に、秘匿性の観点から、横軸の目盛りの数値は示していない。これらの表を見ると、中央値には大きな違いはないものの、前節の結果と比較して、第Ⅲ四分位の値がやや小さくなっており、四分位範囲が全体的に元データよりもやや狭くなっているように見える。地域、産業などの限られた離散変数から、連続変数である従業者数、売上高を推測することが難しいことが原因であると考えられる。

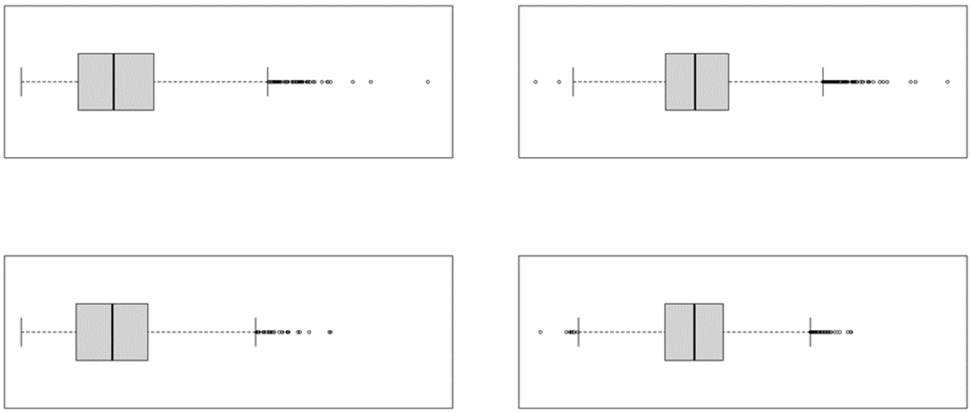

図３　「従業者数」の分布の比較
（左：従業者数　右：売上高
上段：元データ　下段：疑似データ）

### 5.3.3.売上高を予測する重回帰モデルの比較

元データと疑似データにおいて、売上高を、それ以外の変数で予測する重回帰モデルを推定した結果が表25である。元データと疑似データのそれぞれのデータで推定したモデルの結果を比較すると、全ての変数で、回帰係数の符号は一致しており、係数の値もそれほど大きくは異なっておらず、ほぼ近い値となっていることがわかる。ただし、元データと疑似データとで、「資本金」における「1000万円以上2000万円未満」の回帰係数が、元データのものよりも1/10程度になっている。前節の結果と同様に、この変数はp値が大きく（疑似データ：0.331）、推定値の誤差が大きいことが原因であると考えられる。売上高を予測する重回帰モデルについても、本稿の分析の範囲では、変数の発生の順序はそれほど大きな影響を与えていないことがわかる。

表 25　「売上高」を推測するための重回帰モデルの推定結果の比較

| | | 元データ | | | | 疑似データ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Estimate | Std. Error | z value | Pr(>\|z\|) | Estimate | Std. Error | z value | Pr(>\|z\|) |
| 定数項 | | 2.948865 | 0.029566 | 99.737 | < 2e-16 | 2.796904 | 0.02636 | 106.105 | < 2e-16 |
| 地域： | 地域１【ベースライン】 | | | | | | | | |
| | 地域２ | 0.105774 | 0.019574 | 5.404 | 6.73E-08 | 0.128458 | 0.017595 | 7.301 | 3.16E-13 |
| | 地域３ | 0.007828 | 0.024043 | 0.326 | 0.745 | 0.054134 | 0.021657 | 2.5 | 0.01245 |
| 産業： | 建設業【ベースライン】 | | | | | | | | |
| | 製造業 | -0.103277 | 0.024815 | -4.162 | 3.19E-05 | -0.106278 | 0.022441 | -4.736 | 2.22E-06 |
| | 小売業 | 0.683146 | 0.023473 | 29.104 | < 2e-16 | 0.666357 | 0.020981 | 31.76 | < 2e-16 |
| | その他 | -0.147661 | 0.024546 | -6.016 | 1.87E-09 | -0.1774 | 0.022085 | -8.033 | 1.10E-15 |
| 経営組織： | 株式会社【ベースライン】 | | | | | | | | |
| | 有限会社 | -0.152575 | 0.027627 | -5.523 | 3.45E-08 | -0.087761 | 0.022918 | -3.829 | 1.30E-04 |
| 開設年： | ～1984年【ベースライン】 | | | | | | | | |
| | 1985年～1994年 | 0.085169 | 0.017506 | 4.865 | 1.17E-06 | 0.067816 | 0.01574 | 4.309 | 1.66E-05 |
| | 1995年～ | 0.01336 | 0.015361 | 0.87 | 0.384 | 0.011121 | 0.013571 | 0.819 | 0.41255 |
| 資本金： | 300万円以上500万円未満【ベースライン】 | | | | | | | | |
| | 500万円以上1000万円未満 | 0.350539 | 0.029154 | 12.024 | < 2e-16 | 0.280496 | 0.02552 | 10.991 | < 2e-16 |
| | 1000万円以上2000万円未満 | 0.14328 | 0.018607 | 7.7 | 1.53E-14 | 0.016193 | 0.016657 | 0.972 | 0.33101 |
| | 2000万円以上 | 0.083222 | 0.01993 | 4.176 | 3.00E-05 | 0.033061 | 0.016873 | 1.959 | 0.0501 |
| 従業者数（対数） | | 0.912558 | 0.009206 | 99.129 | < 2e-16 | 0.945104 | 0.009171 | 103.051 | < 2e-16 |

## 6. まとめと今後の課題

本稿では、商用データを用いて、合成データの考え方に基づく疑似データを作成し、元データとの比較を行った。その結果、離散変数の集計値の構成比は、おおむね同じ値になっており、相関係数や、売上高を予測するための重回帰モデルの推定結果についても、元データと疑似データとで、ほぼ同様の傾向を示していることがわかった。秘匿性に配慮した上で、周辺分布の統計表や、回帰モデル（重回帰、２項ロジット、順序ロジット）の結果、残差の分布に関する情報（中央値、中央絶対偏差）を事前に作成し、公開することにより、それらの結果を用いて順次変数を生成することによって、元のデータの持つ構造をある程度保持した疑似データを作成できることが示された。

モデルの推定・変数の発生の順序を変更した場合においても、離散変数の分布、相関係数、売上高を予測する重回帰モデルについては、元データや順序の変更前の結果と比較して、それほど大きな影響は見られなかった。ただし、連続変数（従業者数、売上高）については、順序を変更する前と比較して、分布の範囲がやや狭くなっており、影響がみられた。

なお今回の分析では、元のデータに含まれるレコードの総数が7,558件と、それほど多くないこともあり、【手順１】における集計表を作成する際に、表中のセル内の度数が一定数確保されるような変数として、「地域×産業」を選定し、集計表を作成している。ただし、

この【手順１】における集計表の次元数や変数カテゴリー数が、最終的な疑似データの精度に影響する可能性があり、公的統計ミクロデータを用いた、このような影響の詳細な分析については、今後の課題である。

本稿では、資本金を連続変数としてではなく、離散変数として扱っている。これは、資本金額が特定の値（300万円、500万円、1000万円など）に集中しており、単純な回帰モデルの適用が困難であったためである。このような状況を考慮した、より精緻な推定を行うために、上記のような特定の値をとるかどうかについての多項（あるいは順序）ロジットモデルを使用し、それらの値を取らない場合に重回帰モデルを使用するという、２段階の推定を行うことが考えられる。

今回の分析では、回帰モデルなどの構築に使用した変数の数やカテゴリー数が少なく、多重共線性が問題となるような事象は生じていないが、実際に公的統計ミクロデータを用いて疑似的なデータを作成する場合には、変数の数やカテゴリー数なども多くなることから、多重共線性による問題が生じる可能性があり、こうした点を踏まえ、公的統計ミクロデータを用いた実証分析により、変数の選択などについて検討することも重要である。

今回の分析に用いたデータに関しては、変数間の論理的な矛盾が問題となることはなかったものの、例えば世帯のデータなどに関しては、変数間の論理的な関係を考慮せずにデータを発生させた場合、例えば親子関係や配偶関係と性別・年齢などについて、論理的な矛盾が生じる可能性があり、実際の公的統計ミクロデータを用いた分析に当たっては、こうした点を考慮してデータの適切な修正などを行っていく必要がある。

本稿では、作成した疑似データの有用性（元データと疑似データとの差異・相違）に着目して検討を行った。これに加えて、作成した疑似データの安全性（元データに含まれるレコードの情報が外部に流出しないようにすること）についても、評価を行っていく必要があると考える。具体的には、元データに偶然に一致する、あるいは非常に近いデータが生じることのないよう、提供の際には、例えば差分プライバシーの考え方を用いて適切なノイズを付与するという対応が考えられる（佐久間（2016），寺田（2019），Dwork, C. (2008)）。こうした点も、今後の課題である。

本稿における検討の結果を踏まえつつ、今後、実際の公的統計ミクロデータに対して、本稿で検討した方法を適用し、教育やプログラムのテストに耐えうる適切な疑似データが作成できるかということについて、検証・検討を重ねていく予定である。

謝辞

本稿について丁寧な査読をしていただき、多くの改善点の指摘及び有益なコメントをしていただいた匿名の２名の査読者に対し、深く感謝を申し上げたい。本研究は科研費（21K20133）の助成を受けている。

参考文献

[1] 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

[2] 佐久間淳（2016）．データ解析におけるプライバシー保護，講談社

[3] 高部勲（2020）．公的統計ミクロデータの利活用状況と課題：提供者及び利用者の観点から，統計，71(8)，4-9，日本統計協会

[4] 高部勲，徳冨智哉（2020）「公的統計ミクロデータ等に基づく Synthetic Data の作成及び分析の試み」、『ESTRELA』、321、19-24、統計情報研究開発センター

[5] 谷道正太郎（2019）. 公的統計における synthetic data(人工データ)の作成について, ESTRELA, 308，32-35，統計情報研究開発センター
[6] 寺田雅之（2019）．差分プライバシーとは何か，システム/制御/情報，63(2)，58-63
[7] 独立行政法人統計センター（2019）．オンサイト利用における分析結果等の提供に関する標準的なチェック内容の解説と例、
URL; https://www.e-stat.go.jp/microdata/sites/default/files/share/data-use/onsite_check.pdf
[8] 山口幸三（2019）. 改正された統計法と二次的利用の現状と課題, 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
[9] 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. 2013, No. 104, 1-15
[10] Alfons, A., Kraft, S., Templ, M., Filzmoser, P. (2011). Simulation of close-to-reality population data for household surveys with application to EU-SILC. *Statistical Methods & Applications*, 20(3), 383-407.
[11] Caiola, G. and Reiter, J. P. (2010). Random forests for generating partially synthetic, categorical data. Transactions on Data Privacy 3, 27–42.
[12] Drechsler J, Bender S, Rassler S (2008) Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB Establishment Panel. Transactions on Data Privacy 1, 105–130
[13] Dwork, C. (2008). Differential privacy: A survey of results, International conference on theory and applications of models of computation, 1-19, Springer, Berlin.
[14] Little, R. J. A. (1993). Statistical analysis of masked data. Journal of Official Statistics 9, 407–426.
[15] Munnich, R. and Schurle, J. (2003) On the simulation of complex universes in the case of applying the German Microcensus. DACSEIS research paper series No. 4, University of Tubingen.
[16] Munnich R, Schurle J, Bihler W, Boonstra HJ, Knotterus P, Nieuwenbroek N, Haslinger A, Laaksonen S, Eckmair D, Quatember A, Wagner H, Renfer JP, Oetliker U, Wiegert R (2003) Monte Carlo simulation study of European surveys. DACSEIS Deliverables D3.1 and D3.2, University of Tubingen.
[17] Nowok, B., Raab, G. M., & Dibben, C. (2016). synthpop: Bespoke creation of synthetic data in R. J Stat Softw, 74(11), 1-26.
[18] Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. Journal of Official Statistics 19, 1–16.
[19] Reiter, J. P. (2005). Using CART to generate partially synthetic, public use microdata, Journal of Official Statistics 21, 441–462.
[20] Reiter, J. P. (2009). Using multiple imputation to integrate and disseminate confidential microdata. International Statistical Review, 77(2), 179-195.
[21] Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. Journal of Official Statistics, 9, 462–468.
[22] Templ, M., Meindl, B., Kowarik, A., Dupriez, O. (2017). Simulation of synthetic complex data: The R package simPop. *Journal of Statistical Software*, 79(10), 1-38.
[23] Templ, M. (2017). *Statistical disclosure control for microdata*, Springer International Publishing.
[24] Woodcock, S. D. and Benedetto, G. (2009). Distribution-preserving statistical disclosure

limitation. Computational Statistics and Data Analysis 53, 4228–4242.