統計研究彙報　第83号　2026年3月(157～166)

変分オートエンコーダによる合成データの有効性評価

佐野　夏樹†
南　和宏*

Evaluation of the Effectiveness of Synthetic Data Using Variational Autoencoders

SANO Natsuki
MINAMI Kazuhiro

統計的開示性制御において、原データのレコードが再識別されないための匿名化手法に、グローバルリコーディングやノイズ付加による攪乱手法があるが、実際のデータを模倣して人工的に生成されたデータとして、合成データがある。一方、近年、人工知能分野で画像生成等に応用されている手法に変分オートエンコーダがある。変分オートエンコーダは、潜在変数が正規分布に従うとする確率モデルであるため、新規にデータを生成することが可能である。本研究では、教育用標準データセット SSDSE に対して変分オートエンコーダによる合成データの生成を行い、原データとの間で値や統計量の差異にもとづき、有用性評価を行った。またロジスティック回帰による合成データと原データの判別結果にもとづくリスク評価を行い、回帰分析の適用事例を比較することにより、変分オートエンコーダによるデータ生成の有効性について検討を行った。

キーワード：ミクロデータの匿名化、統計的開示制御、公的統計の二次利用、合成データ生成、
変分オートエンコーダ

In statistical disclosure control, techniques such as global recoding and noise addition are used to prevent the re-identification of records in the original data. In contrast, synthetic data refers to artificially generated data that emulate the statistical properties of real data. Recently, Variational Autoencoders (VAEs) have been applied in the field of artificial intelligence, particularly in image generation. VAEs are probabilistic models in which latent variables follow a normal distribution, enabling the generation of new realistic data samples. In this study, synthetic data was generated using a Variational Autoencoder based on the educational standard dataset SSDSE. The utility of the synthetic data was evaluated by comparing differences in values and statistical measures with those of the original data. Additionally, a risk assessment was conducted based on the ability of a logistic regression model to distinguish between synthetic and original records. Through comparative case studies involving regression analyses, the effectiveness of data generation using Variational Autoencoders was examined.

Keywords: Anonymization of microdata, Statistical Disclosure Control, Secondary analysis of official statistics,
Generation of Synthetic data, Variational Autoencoder

† 東京情報大学　総合情報学部　Email：ns207374@rsch.tuis.ac.jp
* 統計数理研究所　学際統計数理研究系

1. はじめに

公的統計におけるミクロデータやその集計結果、消費者の購買履歴や個人属性に関するデモグラフィックデータは、学術研究や企業のマーケティング活動において価値あるデータであり、第３者利用のためのデータの提供は実社会から強く望まれている。

そこで、2017 年施行の改正個人情報保護法からは、特定の個人を識別できないよう適切に加工された匿名加工情報であれば、本人の同意を得ずに、収集時の利用目的外であっても第三者提供が可能となった。ただし、匿名加工情報では、氏名などの直接識別子の削除だけでなく、第三者が保有するデータと照合されても個人が特定されないように、連結可能な符号の削除や特異な記述の削除など、追加的な加工が求められる。

公的統計の分野では、この様なデータを第３者に提供した際に個人が特定される再識別（re-identfication）リスクの他に、個人が特定されない状態でも、個人の属性が開示される属性開示（attribute disclosure）リスクや個人の属性が推測される推測開示（inferential disclosure）リスクが知られ、これらのリスクを総称して開示リスク（disclosure risk）と呼ぶ(Hundepool et al. (2012))。また開示リスクを制御する技術を、統計的開示制御（statistical disclosure control）と呼び、官庁におけるミクロデータ、企業におけるパーソナルデータを問わず、データの利活用を進める上で、重要な技術である。

統計的開示制御におけるデータのマスク方法として、直接識別子等の個人や組織の特定に結びつく変数の削除だけでは再識別リスクを完全には除去できないため、グローバルリコーディング[1](Sano and Hattori(2019), 佐野・服部(2020))やローカルサプレッション[2]、ノイズ付加[3]、PRAM[4](Gouweleeuw et al.(1997)), マイクロアグリゲーション[5](Defays and Nanopoulos (1993)), シャッフリング[6](Muralidhar and Sarathy (2006))が適用される。グローバルリコーディングやローカルサプレッションは、カテゴリー統合による抽象化やデータの秘匿は行うが、事実と異なるデータは公開されないため、非攪乱的手法と呼ばれる。一方で、ノイズ付加、ローカルサプレッション、PRAM、マイクロアグリゲーション、シャッフリングは、事実と異なるデータとなるため、攪乱的手法と呼ばれる。

またデータの利用者からすれば、攪乱的手法や非攪乱的手法によってマスクされたデータは, 元データと異なるほどデータ利用者が求めるものと異なる分析結果が得られやすく、データの有用性が損なわれる。この様にデータの開示リスクと有用性の間には一般的にトレードオフの関係があり、データの開示リスクと有用性を R-U マップ等により視覚化(横溝・伊藤 (2022))しながら、適用するマスク手法の特徴を把握することは重要である。

技術的に第３者に提供可能なデータの作成法として, 原データの特徴を保持する擬似データとして合成データが挙げられる。合成データは、完全合成データ、部分合成データ、ハイブリッドデータに分類できる。完全合成データは、提供するデータセットの中に、原データを含まず、合

[1] 特定の変数に対して連続値の離散化やカテゴリ統合を全サンプルに適用することで、変数の範囲をより抽象度の高いカテゴリへ再階層化し、再識別リスクを軽減する方法。

[2] 変数の組み合わせによる再識別リスクが高いと判断されたサンプルに対し、特定の変数の値を欠損値に置き換え、非公開とすることで、再識別リスクを軽減する方法。

[3] 原データが量的変数である場合に、元の値に対して特定の確率分布に従う乱数を付加した値に置き換える摂動手法。

[4] 原データが質的変数である場合に、遷移確率にもとづき、元のカテゴリを異なるカテゴリに置き換える摂動手法。

[5] クラスタリング適用後の各クラスタ内のサンプル値をクラスタ平均で置き換えることで、同じ変数の値の組み合わせになるサンプルを増やし、再識別リスクを軽減する方法。

[6]統計モデルを利用してデータを並べ替えることで、元の変数の値を平均、分散、相関等の統計的性質を保つ異なる値に置き換える方法。

成データのみで構成される。一方、部分合成データは、ローカルサプレッションと同様に、リスクの高い値を合成データで置き換えるため、部分的に原データを含む。ハイブリッドデータは、データの生成過程で原データと合成データの加算や積算を行い、生成されたデータである。

部分合成データの様に、部分的に原データを含んだり、ハイブリッドデータの様に特定のサンプルから作成されたデータは、攪乱的手法や非攪乱的手法によりマスクされた匿名化データと同様に、それぞれのサンプル（レコード）に対応したデータを生成するため、原データと匿名データの間で対応関係が存在する。一方で、データ生成法が、モデルの学習時には、原データを用いて生成メカニズムを構成し、データ生成時には、乱数を元にして生成する場合、原データに対して対応関係の無い合成データを生成できる。したがって乱数から生成した完全合成データは、再識別リスクの無い開示リスクの低いデータであると言える。

完全合成データの作成には、多重代入(Rubin (1993))、主成分分析（PCA）や人工ニューラルネットワークを用いた非線形 PCA を利用できる(Sano(2020)、Sano(2022))。 部分合成データでは、最も機密性の高いレコードまたは変数のみが合成データに置き換えられ、残りのレコードまたは変数は元のデータを保持される。ハイブリッドデータは、原データと合成データを組み合わせて作成され、元のデータの代わりに公開される。組み合わせ方法によっては、ハイブリッドデータは元のデータまたは合成データに似る可能性があり、ハイブリッドデータの作成方法として、十分性ベースの方法(Muralidhar and Sarathy(2008))がある。

上記の様に合成データは技術的には、今後さらなる展開が期待できるが、公的統計ミクロデータから直接、合成データを作成することは、法制度的には認められていないため、高部(2022)は、中間的な集計結果や回帰分析の結果を利用して合成データを生成する方法を提案している。一方で、将来において法制度が変更された場合や民間のパーソナルデータの第 3 者提供に備えて、合成データの生成方法の評価を行うことも重要である。横溝・伊藤(2023)は、CART やバギングにより合成データを生成し、MDAV 法と比較・評価を行っている。

機械学習の分野で Kingma and Welling (2014)により提案された変分オートエンコーダ（VAE）は、データ生成のためのニューラルネットワークの一つであり、顔や文字の画像生成、異常検知の分野に適用されている。本研究では、VAE を用いて、SSDSE（教育用標準データセット）に対して、合成データを生成し、有用性とリスクの評価を行う。 2 節において、変分オートエンコーダについて概観した後、 3 節で、変分オートエンコーダを用いたデータ生成法および、本研究で用いる有用性評価とリスク評価の指標について説明する。 4 節では、生成データに対する有用性とリスク評価の結果および回帰分析事例を通じて、現実的な側面から VAE を用いた合成データ生成法の有効性について検討を行う。

2. 変分オートエンコーダ

変分オートエンコーダ（Variational AutoEncoder、VAE）は、線形 PCA やニューラルネットワークを用いた非線形 PCA（オートエンコーダ）と同様に教師なし学習手法であり、入力データ $\boldsymbol{x}$ をエンコーダによって低次元の潜在変数 $\boldsymbol{z}$ に圧縮した後、デコーダによって $\boldsymbol{x}'$として再構成する。ただし、PCA やオートエンコーダは、確率モデルでは無いのに対し、VAE は確率モデルであり、潜在変数 z が平均μと分散σ^2の正規分布に従う確率変数であると仮定し、μやσ^2はエンコーダによって学習される。図 1 に VAE の概略図を示す。

一般に周辺対数尤度

$\log p_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(1)},\cdots,\boldsymbol{x}^{(n)}\right)=\sum_{i=1}^{n}\log p_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right)$,

における個々のサンプルの対数尤度は次のように書き換えられる。

$$\log p_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right)=D_{KL}\left(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}^{(i)})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}^{(i)})\right)+\mathcal{L}\left(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x}^{(i)}\right) \quad (1)$$

ここで、右辺第 1 項は近似事後分布$q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}^{(i)})$と真の事後分布$p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}^{(i)})$の間の KL ダイバージェンスであり、右辺第 2 項$\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x}^{(i)})$は i 番目のサンプルの周辺尤度に対する変分下限を表す。VAE は近似事後分布$q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}^{(i)})$をエンコーダとして構築し、対数尤度を直接最大化する代わりに、下限、$\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x}^{(i)})$の最大化を行う。下限は、さらに、次のように書き直せる。

$$\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x}^{(i)}) = -D_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}^{(i)})||p_{\boldsymbol{\theta}}(\boldsymbol{z})) + \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}^{(i)})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z})], \quad (2)$$

ここで、右辺第 1 項は正則化項であり、学習後のエンコーダ$q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}^{(i)})$が事前分布$p_{\boldsymbol{\theta}}(\boldsymbol{z})$に近いことを要求する。右辺第 2 項は、デコーダとして構築された$p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z})$の再構成誤差に相当する。事前分布$p_{\boldsymbol{\theta}}(\boldsymbol{z})$には、標準正規分布$N(\boldsymbol{0},\boldsymbol{I})$が用いられる。

潜在変数 z は、サンプリング層において、次の様にして正規分布からサンプリングされる。

$$z = \mu + \varepsilon\sigma, \quad (3)$$

ここでμとσはそれぞれ、潜在変数の平均と標準偏差であり、εは標準正規分布からサンプリングされた確率変数である。通常サンプリング操作は微分出来ないが、(3)式の様にサンプリングすることにより、誤差逆伝播する際のμとσの微分情報が伝播し，エンコーダによってμとσが学習される。

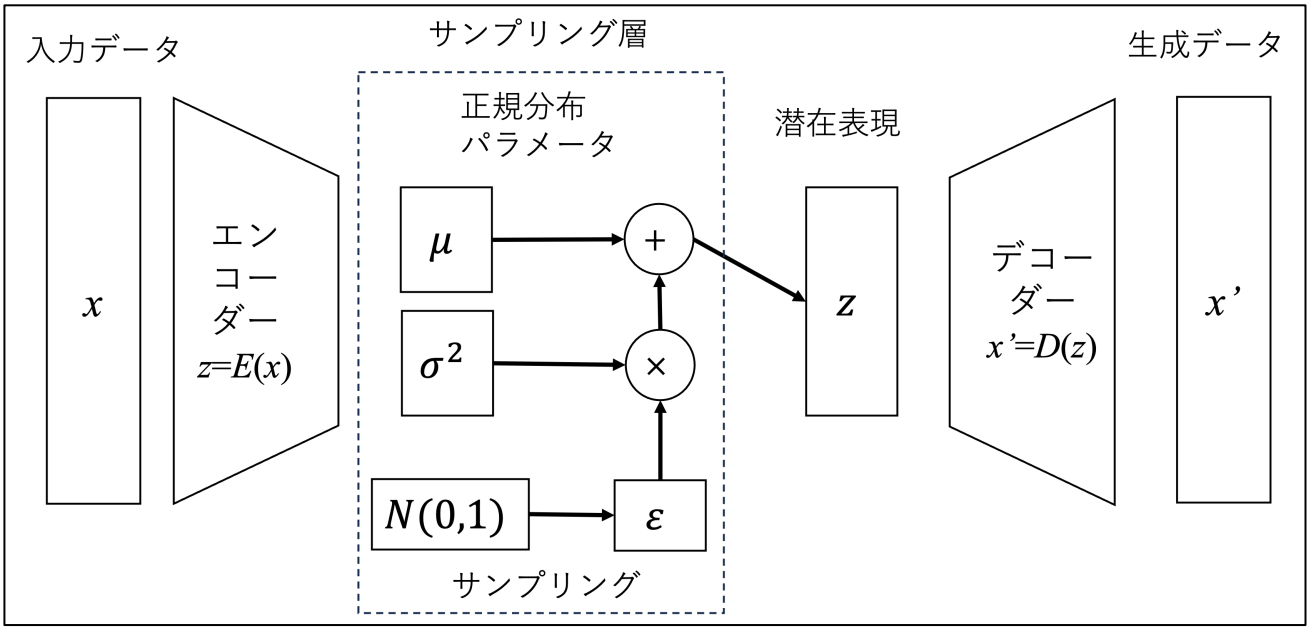

図 1. VAE の概略図

3. VAE を用いた合成データと有用性とリスクの評価指標

3.1 VAE を用いたデータ生成

本研究では、あらかじめ、元のデータに対して Min-Max スケーリングによる前処理を行い、図 2 にしめす構成の VAE により、2 種類の合成データ

(a) 再構成データ（Reconstructed Data）

(b) 生成データ（Generated Data）

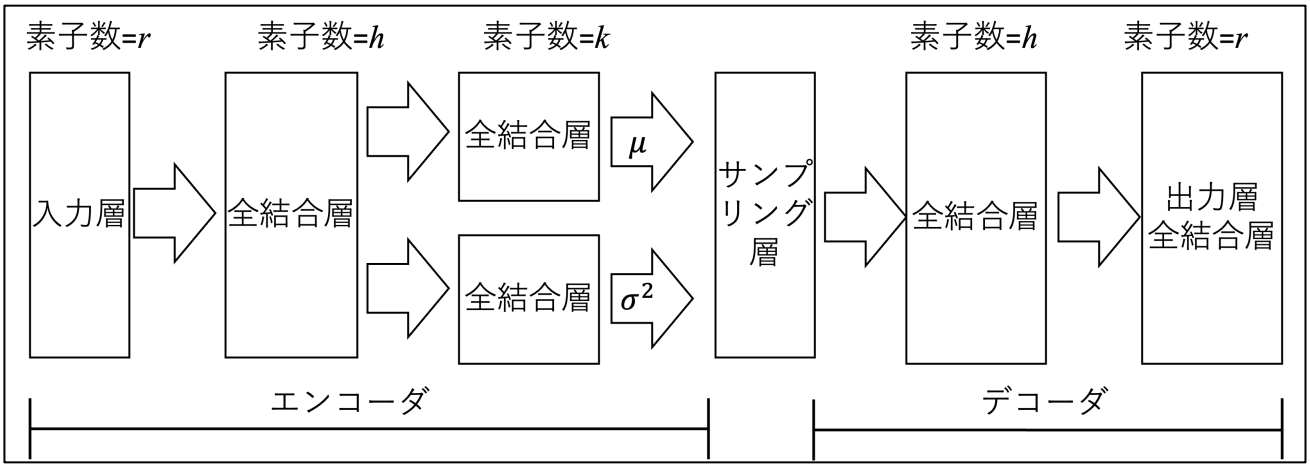

図 2. VAE の構成図

を生成する。ただし、r は入力変数の次元、h はエンコーダとデコーダの隠れ層の素子数、k は潜在変数の次元を表す。活性化関数には ReLU 関数を用いるが、出力層では、シグモイド関数による変換を行い、クロスエントロピー関数による損失の評価を行う。
(a)の再構成データは、学習フェーズ後に出力層から得られる出力値であり、出力されるサンプルは、原データのサンプルと対応関係がある。(b)の生成データは、学習フェーズ後にサンプリング層で新たに生成された潜在変数 z に対してデコーダを適用して生成されたデータであり、原データのサンプルとの間には対応関係は無い。(a)と(b)のデータ生成において、デコーダで出力されたデータは、Min-Max 前処理の逆変換し、合成データとする。

3.2 有用性評価指標

提案手法によって生成された合成データの有用性を情報損失として、以下の 3 つの平均変動指標(Domingo-Ferrer et.al. (2001))によって評価する。最初の指標は、観測値に対する平均絶対誤差率である。

$$MAEO = \frac{1}{nr}\sum_{i=1}^{n}\sum_{j=1}^{r}\left|\frac{x_{ij} - x'_{ij}}{x_{ij}}\right|, \tag{4}$$

ここで、x_{ij}およびx'_{ij}は、それぞれ j 番目の変数に対する i 番目の観測値の原データの値と生成データの値を表す。
2 つ目の指標は、各平均値に対する平均絶対誤差率である。

$$MAEM = \frac{1}{r}\sum_{i=1}^{r}\left|\frac{m_i - m'_i}{m_i}\right|, \tag{5}$$

ここで、m_iおよびm'_iは、それぞれ原データと生成データにおける i 番目の変数の平均値を表す。
3 つ目の指標は、相関係数に対する平均絶対誤差である。

$$MAEC = \frac{1}{r(r+1)/2}\sum_{i=1}^{r}\sum_{j>i}^{r}\left|c_{ij} - c'_{ij}\right|, \tag{6}$$

ここで、c_{ij}およびc'_{ij}は、それぞれ、原データと生成データにおける、i番目とj番目の変数間の相関係数を表す。より頑健な評価を行うために、これら3つの指標について平均の代わりに中央値の算出も行う。

3.3 リスク評価

原データと生成した合成データに対して、ロジスティック回帰分析による判別を行い、判別の難易度をAccuracy、Recall、Precision、F値により評価し、合成データのリスク評価とする。同様の評価指標として、pMSE (Woo et.al(2009))が挙げられる。pMSEは、ロジスティック回帰分析を用いて、原データか秘匿データかの判別を行い、秘匿データの予測確率と秘匿データ割合の間の平均二乗誤差を有用性評価の指標としている。pMSEが小さいことは、原データと秘匿データの区別がつかないほど、原データの特徴を保持し、有用性が高いことを意味する。同時に、pMSEをリスク評価指標とみると、pMSEの値が小さいことは、原データと秘匿データの区別がつかず、リスクが小さいことを意味する。各指標は、表1のコンフュージョン行列から次の様に計算される。

表1 コンフュージョン行列

予測＼実際	原データ	合成データ
原データ	a	b
合成データ	c	d

$$Accuracy = \frac{a+d}{a+b+c+d}, Recall = \frac{a}{a+c}, Precision = \frac{a}{a+b}, F = 2 \times \frac{Recall \times Precision}{Recall + Precision} \quad (7)$$

4. SSDSE-A データの合成データに対する有用性とリスクの評価

4.1 評価指標による合成データの有用性とリスクの検証

教育用標準データセットSSDSE-Aに対して合成データを生成し、3節の指標による有用性とリスクの評価を行った結果を表2と表3に示す。SSDSEは、都道府県や市町村を集計単位としたマクロデータであり、本研究が本来対象とするミクロデータではないが、集計単位が個人か市町村かの違いは、VAEによる合成データの定量的な評価を行う上で支障にならないと考え、また容易に利用可能なデータであることから、ここではSSDSEを用いた。

SSDSE-Aは、1741の市町村に対する人口・世帯、自然環境、経済基盤、行政基盤、教育、労働、文化・スポーツ、居住、健康・医療、福祉・社会保障に関する125変数からなるデータである。3章の有用性評価指標は、平均指標であるが、頑健な評価を行うために中央値による評価も行なった。エンコーダとデコーダの隠れ層の素子数はh=64、VAEの学習におけるバッチサイズを20、エポック数を200とした。

表2の有用性評価の結果からMAEOの値は、再構成データに関しては、7.068から8.546の値となっており、生成データに関しては、14.017から18.186の値となっており、再構成データの方が、生成データよりも小さな値であることがわかる。また中央値で計算した場合は、最大でも1を超えていないことから、平均値によるMAEOの値は、極端に大きな外れ値が発生していることを示唆している。

表２　再構成データと生成データの有用性評価　（括弧内の数値は中央値を表す）

	(a) MAEO		(b) MAEM		(c) MAEC	
k	再構成	生成	再構成	生成	再構成	生成
3	7.518 (0.623)	14.193(0.895)	0.053(0.057)	0.152(0.161)	0.219(0.094)	0.226(0.092)
5	7.585 (0.596)	14.017(0.895)	0.132(0.141)	0.183(0.197)	0.223(0.094)	0.223(0.089)
10	8.546(0.609)	15.508(0.905)	0.033(0.029)	0.136(0.145)	0.228(0.093)	0.227(0.091)
15	7.068(0.593)	14.161(0.886)	0.076(0.081)	0.170(0.181)	0.230(0.092)	0.228(0.089)
20	7.287(0.590)	18.186(0.904)	0.031(0.023)	0.117(0.112)	0.230(0.092)	0.234(0.089)
32	8.185(0.632)	14.732(0.892)	0.008(0.005)	0.125(0.127)	0.232(0.092)	0.230(0.091)

MAEM の値は、再構成データでは、0.008 から 0.132 の値をとるのに対して、生成データでは、0.117 から 0.183 の値を取っており、やはり再構成データの方が、生成データよりも小さな値である。

MAEC に関しては、再構成データと生成データの間で、それほど、大きくは変わらない結果であり、0.219 から 0.234 の値を取っているが、中央値は、最大でも 0.1 を超えていないことから、特定の変数間に原データと大きく異なる相関係数が発生していると考えられる。

以上のことから、再構成データの方が、有用性評価の側面からは、生成データよりも優れていると言える。また、潜在変数の数 k の増減に対しては、MAEO と MAEM は、ばらつきはあるが、特別な傾向性は無いことがわかる。MAEC に関しては、k の増減に対して、あまり変わらない結果であることがわかる。

表 3　再構成データと生成データのリスク評価

	(a) Accuracy		(b) Recall		(c) Precision		(d) F 値	
k	再構成	生成	再構成	生成	再構成	生成	再構成	生成
3	0.628	0.589	0.633	0.625	0.627	0.583	0.630	0.603
5	0.620	0.583	0.544	0.579	0.641	0.584	0.589	0.581
10	0.576	0.560	0.604	0.605	0.572	0.555	0.588	0.579
15	0.619	0.530	0.574	0.558	0.631	0.528	0.601	0.543
20	0.617	0.572	0.536	0.561	0.639	0.574	0.583	0.567
32	0.559	0.531	0.569	0.563	0.558	0.529	0.564	0.546

表 3 のリスク評価の結果からは、Accuracy と Precision は、全ての k において、生成データの方が、再構成データよりも小さな値をしめしている。Recall の値に関しては、k=3、15、32 においては、生成データの方が小さな値をしめしているが、それ以外では、生成データの方が大きな値をしめしている。Recall と Precision の調和平均である F 値においても、全ての k において、生成データの方が小さな値をしめしている。以上のことから、概ね再構成データよりも生成データの方が原データとの分類が困難であり、生成データの方が、リスク評価の側面からは、優れていると言える。また潜在変数の数 k の増減に対しては、有用性評価の指標と同様に、特別な傾向性は持っていないことが示唆される。

したがって、再構成データは、原データとの間に対応関係を持つものの、相対的に有用性評価の点で優れており、リスク評価の観点からは、生成データが優れていると言える。

4.2 回帰分析事例による有用性の検証

3 節における有用性評価指標は、基本的には、相対変化率にもとづく原データの情報損失量であるが、実際の分析における影響度を検証するため、回帰分析事例を通じて、偏回帰係数や寄与率の大小を原データに適用した結果と比較する。回帰分析に用いる変数は次の通りである。

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon, \varepsilon \sim N(0, \sigma^2), \quad (8)$$

y：地方税（千円）
x_1：15 歳～64 歳人口（人）
x_2：民営事業所数（事業所）
x_3：第 1 次産業就業者数（人）
x_4：第 2 次産業就業者数（人）
x_5：第 3 次産業就業者数（人）

原データ、再構成データ、生成データに対する(8)式の回帰分析を行った結果を表 4 にしめす。原データに対する全変数の偏回帰係数の符号が一致しているデータは無く、最も一致しているのは、k=3,10,15,32 の生成データが、切片$\hat{\beta}_0$の符号を除いて、符号が一致している。ただし、寄与率R^2は、全てのデータにおいて 0.99 程度の値をしめしている。回帰分析事例の比較結果から、合成データを用いて回帰分析を行った場合、原データに対する回帰分析と異なる結果が得られる可能性をしめしている。この理由は、回帰分析における回帰係数の推定において、平均と共分散行列が重要な役割を持つが、4 節表 2 の MAEM と MAEC の値は、合成データに対する回帰分析の結果が原データに対する回帰分析結果が同程度の結果を保証するには、情報損失が大きいこと示唆している。

5. まとめ

本研究では、VAE にもとづいた合成データ生成法として、再構成データと生成データを提案し、実際に SSDSE-A データに対して合成データを生成し、それらの有効性とリスクの評価を行った。その結果として、再構成データは、生成データに対して相対的に有用性の点で有効であり、リスク評価の観点から生成データの方が有効であることが示唆された。また回帰分析事例による合成データの有用性を評価した結果、潜在変数の次元 k の値によって、偏回帰係数の値にばらつきがあり、原データの分析結果とも符号が必ずしも一致していないことがわかった。表 2 の MAEO や MAEC の評価結果から、外れ値等の存在により、共分散情報の損失が起きていると考えられる。共分散情報を利用するその他の統計手法を適用した際にも、影響が起きると予想され、VAE による合成データを生成する際の今後の課題である。

謝辞
本稿について、丁寧な査読をしていただき、貴重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る。本研究は、JSPS 科研費 JP22K01427 の助成を受けたものである

表 4 回帰分析による合成データの有用性評価（有意水準 *** 0.1% **1% *5%）

(a) 原データに対する回帰分析の結果

$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	R^2
-1815303.13***	229.51***	1328.62***	-1727.02***	380.19***	-43.66	0.96

(b) 再構成データに対する回帰分析の結果

k	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	R^2
3	2692629.36***	132.43***	3060.56***	-4446.00***	-325.10***	85.51***	1.00
5	3532535.60***	151.46***	3072.81***	-6221.73***	-368.35***	83.30***	0.99
10	4111233.31***	-24.05*	3477.26***	-5865.88***	-530.26***	378.06***	0.99
15	1731194.05***	201.07***	4721.19***	-2709.03***	-851.94***	-130.07***	0.99
20	3064502.91***	238.36***	805.04***	-5372.46***	-461.95***	218.30***	0.99
32	300162.31	35.70***	2219.15***	-1040.29**	-693.83***	414.91***	1.00

(c) 生成データに対する回帰分析の結果

k	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	R^2
3	1885491.03***	148.15***	1467.96***	-116.06*	164.33***	-508165.97***	0.99
5	2491005.49***	109.78***	2057.60***	227.91***	84.96***	-676911.35***	0.99
10	1806247.61***	181.17***	1920.73***	-691.78***	175.70***	-348200.57***	0.99
15	772404.75***	169.16***	3119.21***	-523.74***	11.90	-268705.69***	0.99
20	1544136.51***	174.31***	543.07***	214.24***	154.09***	-559798.05***	0.99
32	750052.42***	81.76***	1805.39***	-549.90***	357.91***	-263507.39***	0.99

参考文献

[1] 佐野 夏樹, 服部 雄太 (2020), 「モデルの判別精度によるグローバルリコーディングの有用性評価」, 『統計研究彙報』, 第 77 号, pp.1-pp.14.

[2] 高部勲（2022）, 「合成データの考え方に基づく公的統計擬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-pp.130.

[3] 横溝秀始, 伊藤伸介 (2023), 「合成データ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, pp.97-pp.116.

[4] 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－」, 第 79 号, pp.151-pp.170.

[5] Defays D. and Nanopoulos P. (1993), “Panels of enterprises and confidentiality: the small aggregates method.” In: Proc. of 92 Symposium on Design and Analysis of Longitudinal Surveys, 195–204

[6] Domingo-Ferrer, J., Mateo-Sanz J.M., and Torra V. (2001), “Comparing SDC methods for microdata on the basis of information loss and disclosure risk”, Pre-proceedings of ETK-NTTS 2001 (Exchange of Technology and Know-how and New Techniques and Technologies for Statistics), 2, 493-526.

[7] Gouweleeuw J., Kooiman P., Willenborg L. and de Wolf, P. (1997), “Post randomization for statistical disclosure control: Theory and implementation”, Technical report, Statistics Netherlands.

[8] Hundepool A., Domingo-Ferrer J., Franconi L., Giessing S., Nordholt E., Spicer K. and de Wolf P. (2012), Statistical Disclosure Control. Wiley, Chichester, UK.
[9] Kingma D., Welling M. (2014), “Auto-encoding variational bayes”, In: Proc. of the 2nd International Conference on Learning Representation.
[10] Muralidhar K. and Sarathy, R. (2006), “Data shuffling: a new masking approach for numerical data”, Management Science, 52(5), 658–670.
[11] Muralidhar K. and Sarathy R. (2008), “Generating sufficiency-based non-synthetic perturbed data”, Transactions on Data Privacy, 1(1), 17–33.
[12] National Statistics Center of Japan: SSDSE (2025), https://www.nstac.go.jp/use/literacy/ssdse/ [Access：2025/5/5]
[13] Rubin D.B.(1993), “Discussion of statistical disclosure limitation”, Journal of Official Statistics 9(2), 461–468.
[14] Sano N.(2020), “Synthetic data by principal component analysis”, In: Proc. of 20th IEEE International Conference on Data Mining Workshops 2020, 101–105.
[15] Sano N. (2022), “Utility and risk evaluation of synthetic data by orthogonal transformation”, The Review of Socionetwork Strategies 16(1), 71–79.
[16] Sano N. and Hattori Y. (2019), “Utility evaluation measures for categorical data by classification performance”. In: Proc. of 19th IEEE International Conference on Data Mining Workshops, 356–361.
[17] Woo M. J., Reiter J. P., Oganian A. and Karr A. F. (2009), “Global measures of data utility for microdata masking”, Journal of Privacy and Confidentiality”, 1(1), 111-124.