パーソナルデータの合成データに関する研究

パーソナルデータは個人に関するデータの総称であり，様々な分野で，その利活用が期待されている．しかしながら，原データを公開した場合，個人が特定される可能性があるため，秘匿化処理は重要であり，経済統計の分野では，統計的開示制御の問題として古くから知られている．統計的開示制御では，原データ自体に対して，個人が特定されない様に加工した上で公開する方法の他に，原データの特性を保持した上で，原データとは異なるデータを生成する合成データの生成も重要な研究課題である．

佐野ゼミでは，以下の主成分分析によって合成したデータ（OT）のリスク評価と有用性評価を行っている．有用性評価は，加工データや合成データの原データに対する情報損失であり，個人の特定リスクとはトレード・オフの関係にある．

1. $p \times p$分散共分散行列$\boldsymbol{C}$を$p \times p$直交行列$\boldsymbol{U}$と$p \times p$対角行列に分解する．

$$\boldsymbol{C} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\boldsymbol{T}} = \sum_{i=1}^{p} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T$$

λ_i, $\boldsymbol{u}_i$は，それぞれ，$\boldsymbol{C}$の固有値・固有ベクトルを表す．

2. 平均ベクトル$\bar{\boldsymbol{y}}$を引き，中心化した個体ベクトル$\boldsymbol{z}$に対する合成データ

$\hat{\boldsymbol{y}}^{(q)} = \bar{\boldsymbol{y}} + \boldsymbol{U}_{(q)}\boldsymbol{U}_{(q)}^{T}\boldsymbol{z}$

を生成する．

$\boldsymbol{U}_{(q)}$は直交行列$\boldsymbol{U}$の1列目からq列目までの行列であり，$\boldsymbol{U}_{(q)}^{T}$はその転置行列である．

OTによる合成データは，qの値を指定することによって，次の様に有用性評価を行うことができる．

$$ILOT = \frac{\sum_{i=q+1}^{p} \lambda_i^R}{p} \qquad (1)$$

ここでλ_i^Rは相関係数行列に対する固有値である．

平成16年の全国消費実態調査の匿名データに対してOT, DRI（確定的回帰補間）, SRI（確率的回帰補間）, SRMI（逐次回帰多重補間）によって生成した合成データの有用性，および，リスク評価を比較した結果を表1にしめす．有用性は，絶対平均誤差，リスク評価は，レコードリンケージを用いて評価を行っている．詳細については, Sano N. (2022)[1]を参照されたい. OTの弧内の値は(1)式のILOTの値をしめし，提案手法の合成データは, ILOTの値に対応する最も近いqの値によって生成されている．表1の結果から，提案手法は，その他の手法に比べて有用性（情報損失）が小さく，リスクが大きいことがわかる．提案手法には，ねらいのILOT有用性評価値をもつ合成データを生成可能であるという特性がある．

表１ 主成分分析による合成データ生成法の有用性とリスクの評価

(a) 有用性評価

Method	DRI	SRI	SRMI	OT (0.01)	OT (0.05)	OT (0.1)	OT (0.2)
MAE	585.114	904.207	946.621	64.791	182.304	274.295	382.404

(b) リスク評価

Method	DRI	SRI	SRMI	OT (0.01)	OT (0.05)	OT (0.1)	OT (0.2)
k=1	0.000	0.000	0.000	0.996	0.930	0.694	0.271
k=3	0.001	0.000	0.001	0.997	0.949	0.775	0.377
k=5	0.002	0.000	0.001	0.997	0.955	0.801	0.425

OT の括弧内の値は, ILOT の値をしめす

参考文献

［1］ Sano N. (2022) : Utility and Risk Evaluation of Synthetic Data by Orthogonal Transformation, The Review of Socionetwork Strategies, Vol. 16, No.1, pp. 71-79.