ノイズが加えられた個票データのリスク評価

岡山商科大学　佐井 至道

個票データを公開することによってプライバシー侵害の可能性が高いと判断された場合，何らかの秘匿措置が施されることが多いが，国内ではカテゴリの併合のような非攪乱的な秘匿措置を行うことが多い．そのような個票データに対するリスク評価としては寸法指標を用いるのが一般的であり，データが標本調査で得られた場合には，推定された母集団寸法指標を用いることが多い．

これに対して，海外ではノイズの挿入やスワッピングのような攪乱的な秘匿措置が多用されており，近年，国内においても用いられることが多くなってきた．

特に，個体を特定するために用いられるキー変数に対してノイズの挿入を行うと，値が変化してしまうため，寸法指標によるリスク評価は意味を持たなくなる．そのため，秘匿後の個体が元の個体にリンクされる確率やそのような個体数などをリスク評価の指標とすることが多いが，これまでの議論では，個票データの中でリンクされるか否かについて検討するものものがほとんどである．

本報告では，標本から得られた個票データのキー変数に対してノイズを挿入した場合のリスク評価について，母集団も念頭に入れた検討と基礎的な実験の報告を行う．

個票データが標本調査で得られたものとし，母集団の大きさを N，標本の大きさを n，キー変数の個数を K とする．キー変数としては量的変数（連続型，離散型），カテゴリに順序のない質的変数の場合を考える．例えばすべてのキー変数が連続型の量的変数の場合，標本の i 番目の個体の k 個のキー変数のベクトルを $\mathbf{x}_i$，挿入するノイズのベクトルを $\mathbf{e}_i$，母集団の j 番目の個体のキー変数のベクトルを $\mathbf{a}_j$，2 つの個体のキー変数の組み合わせの距離を $\mathrm{d}(\cdot,\cdot)$ とするとき，$\mathrm{d}(\mathbf{x}_i+\mathbf{e}_i,\mathbf{a}_j) \leq \mathrm{d}(\mathbf{x}_i+\mathbf{e}_i,\mathbf{x}_i)$ となる $\mathbf{a}_j$ が少なくとも 1 つ存在すれば，個票データのノイズを加えられた個体について，母集団内で間違ったリンクが発生し，そうでない場合には真のリンクが発生したと考える．標本内での議論も同様である．

リンクの確率はキー変数が従う分布にも依存するが，シミュレーション実験においては正規分布，一様分布などを用い，挿入するノイズについても同様の分布を用いる．$x_{i,k}$ を標本で i 番目の個体の k 番目のキー変数の値などとするとき，$x_{i,k}$，$a_{j,k} \sim \mathrm{N}(0,1)$，$e_{i,k} \sim \mathrm{N}(0,\mathrm{c}^2)$ のように，キー変数の値，ノイズとも i，j，k について独立に正規分布に従う場合のシミュレーション結果を下の表に示す．数値は，個々の個体が真のリンクとなる確率である．例えば $k=1$，$c=10^{-2}$，$n=100$，$N=10000$ のとき，標本では $100 \cdot 0.68482 = 68.482$ 個が元の個体にリンクされるが，母集団を考慮に入れると $100 \cdot 0.02509 = 2.509$ 個がリンクされるだけとなる．

検討の詳細については当日報告する．

表: 個体が真のリンクとなる確率（上：$k=1$，下：$k=4$）

$c \setminus N$	10	100	1000	10000	100000
1.0	0.21973	0.02990	0.00358	0.00041	0.00005
1.0^{-1}	0.70040	0.17931	0.02509	0.00305	0.00035
1.0^{-2}	0.96099	0.68482	0.17854	0.02502	0.00304
1.0^{-3}	0.99601	0.95713	0.68344	0.17849	0.02501
1.0^{-4}	0.99959	0.99557	0.95680	0.68330	0.17848

$c \setminus N$	10	100	1000	10000	100000
1.0	0.53583	0.16207	0.03442	0.00602	0.00093
1.0^{-1}	0.99932	0.99297	0.93895	0.69507	0.31761
1.0^{-2}	1.00000	1.00000	0.99999	0.99993	0.99926
1.0^{-3}	1.00000	1.00000	1.00000	1.00000	1.00000
1.0^{-4}	1.00000	1.00000	1.00000	1.00000	1.00000