匿名データの個票開示リスク

金沢大・経　星野伸明

匿名データ作成では、匿名化されたデータについて個体識別できないことの根拠は統計委員会の見識に求められている。従って個体識別できないことの根拠について、専門家が説明責任を負う。しかし匿名化の程度の決定が透明でないと、説明することができない。従って個体識別行為をモデルで明確に表現することが望ましい。またこれにより、意思決定過程の改善点を焦点を絞って議論することが出来る。

個体識別可能性の判断に万全を期すには、根拠として観測情報も用いる方がよい。公開されたデータについて個体識別が起きたか否かの観測結果は、データが個体識別可能か否かについて（僅かながら）情報を持っており、確率モデルを用いることでこのような情報も拾い上げることが可能となる。

ただし観測が限られているので、母数の多い複雑なモデルの同定は望み薄である。故に星野 (2016) は単純な個体識別の一母数確率モデルを提案した。このモデルでは個体識別の要因を三つ挙げた。すなわち

(a) 攻撃用ファイルと公開ファイルに、誤記・誤分類や属性の経時変化がない（同個体なら両ファイルで変数の値が同じという意味）。

(b) 公開ファイルに個体が含まれている。

(c) 個体が母集団一意。

詳細な議論は省くが、このモデルでは個体識別が不可能な状況は $\Pr(a,b,c) \leq \beta$ と同値になる。ただし β は未知母数であり、観測から推定される。過去に個体識別が認知されていない事例の中で $\Pr(a,b,c)$ の最も高い評価値を γ と書く。同モデルでは β は γ 以上（かつ個体識別発生が認知されている事例の評価値未満）と最尤推定される。

このモデルでは、データについて個体識別の容易度 $\Pr(a,b,c)$ を評価し、閾値 β 以下ならそのデータは公開可能と明確に判断される。このような単純なモデルは個体識別と関係する多くの要因を捨象しているが、観測自体が統制されていれば、弊害は押さえられる。匿名データは利用者の条件がそろっているため、まさにそのような状況となる。

本報告では匿名データにおける β の最尤推定結果などを紹介する。ただし過去のデータ公開事例としては、制度開始から提供されている四調査（就業構造基本調査 (ESS)、住宅・土地統計調査 (HLS)、社会生活基本調査 (STULA)、全国消費実態調査 (NSFIE)）の当初から提供されている年次のデータ (ESS:92,97,02, HLS:93,98,03, STULA: 91,96,01,06, NSFIE: 89,94,99,04) だけを用いた。その他の匿名データファイルは、残念ながら利用件数が比較的限られている。

データでは HLS '98 が最も個体識別が容易な事例となった。HLS は公開されるレコード数が多いので、全体的に $\Pr(a,b,c)$ が高めである。STULA はレコード数が少ないため、全体的に $\Pr(a,b,c)$ が低めになっている。このような結果は、NSFIE,STULA,ESS について HLS 並に匿名化を緩和できることを示唆している。これら 3 調査は HLS と異なり都道府県コードを「三大都市圏か否か」に再符号化しているため、都道府県コードまで公開することは検討に値すると思われる。

References

[1] 星野伸明 (2016)「エビデンスに基づいた匿名化」, 日本統計学会誌, 46 巻, 1-42.