エビデンスに基づいた匿名化の実証

星野　伸明
金沢大・経

2015 年 11 月 27 日

背景の説明

- 匿名データは個体が識別できないようにミクロデータを加工してある。
  - 参）匿名データの定義（統計法第２条第１２項）:「一般の利用に供することを目的として調査票情報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」
- そのような加工を「匿名化」と呼ぶ。
- 最適な匿名化水準？
  - 識別リスクと有用性（分析価値）のトレードオフ有り。
  - 法的には個体識別が不可能な範囲で有用性を最大化するしかない。
  - 識別リスクの許容範囲を客観的に定めたい。
- ⇒ 公開された匿名データについて個体が識別されていないという観測から、識別リスクの許容範囲を統計的に推定する。

概要

1. 個体識別ができない状態の定式化
   (a) 個体識別の判別モデルと観測モデル
   (b) 母数推定
2. 識別リスクの計測
   (a) キー変数の選択方法
   (b) 住宅・土地統計調査 (H15) 匿名データの例

Marsh らの個体識別モデル

$$\Pr(\text{識別が実際に起きる}) = \Pr(\text{識別成功} \mid \text{識別を試みる}) \Pr(\text{識別を試みる}) \quad (1)$$

$$\Pr(\text{識別成功} \mid \text{識別を試みる}) = \Pr(a) \Pr(b|a) \Pr(c|a,b) \Pr(d|a,b,c) \quad (2)$$

ただし

(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) 個体が母集団一意である。
- 一意に照合される個体が母集団でも一意ということ。

(d) 個体が母集団一意と確証出来る。
- 既存情報で一意数は評価される。
- 本報告では追加情報による攻撃を確証の一種として考慮。

個体識別が不可能ということ

- 匿名データの定義における個体識別が不可能という状態の解釈：

$$\Pr(\text{識別成功} \mid \text{識別を試みる}) = 0 \tag{3}$$

- ここで (2) より

$$\Pr(\text{識別成功} \mid \text{識別を試みる}) = \Pr(a,b,c)\Pr(d|a,b,c) \tag{4}$$

  - $\Pr(a,b,c)$ か $\Pr(d|a,b,c)$ のいずれかが 0 なら個体識別は不可能。

母集団一意の確証

- 通常は $\Pr(a,b,c) \neq 0$ なので、個体識別が可能か不可能かは $\Pr(d|a,b,c)$ が 0 か否かの問題になる。
- $\Pr(d|a,b,c) = 0$ とは母集団一意の確証が不可能ということ。
- 母集団一意の確証方法：
  - 一意たらしめているキー変数の組み合わせ（指紋）について全数情報を集める。
  - 全数 $\subseteq$ 母集団。部分集団で一意なら母集団でも一意。
    - 例）日本の弁護士集団で一意なら、日本人でも一意。
  - 全数名簿が存在したり作りやすい場合は匿名化で対策する。
    - 例）「弁護士」を「弁護士か司法書士」と再符号化。

個体識別可能性の判別モデル

- 匿名化等によって母集団一意の確証要因はコントロールする。
  - それでも残る不確実性を母数 (β) に集約。
- 統計モデル化：適当な非負の β について

$$\Pr(a,b,c) \leq \beta \Leftrightarrow \Pr(d|a,b,c) = 0 \tag{5}$$

  - データが情報豊富なら、$\Pr(a,b,c)$ が高い。
  - 確証可能性は、データ情報度の単調関数と思われる。
- $\Pr(a,b,c)$ は母集団一意確証の「容易度」。

個体識別の観測モデル

- モデル (5) の母数 β を統計的に推定するには観測が必要。
- 個体識別が可能か否かは直接観測できないので、識別成功の社会的認知の有 ($X=1$) 無 ($X=0$) を観測：

$$\begin{aligned}\Pr(X=1) \quad = \quad & \Pr(\text{識別の社会的認知} \mid \text{個体識別が実際に起きる}) \\ & \times \Pr(\text{個体識別が実際に起きる})\end{aligned}$$

- $\Rightarrow \Pr(a,b,c)$ の評価値を γ で表せば

$$\Pr(X=1) = \begin{cases} p(\gamma) & \gamma > \beta \text{ の場合} \\ 0 & \gamma \leq \beta \text{ の場合} \end{cases} \tag{6}$$

- 適当な条件の下で $p(\gamma) > 0$.

閾値の最尤推定量 $\hat{\beta}$

- 過去の（匿名化した）データ公開事例 $(i = 1, 2, \ldots, n)$ をモデル (6) からの独立標本とみなす。i 番目の事例について $\Pr(a, b, c)$ の評価値 γ_i と個体識別発生認知の有無 x_i は観測できる。
- 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 $\bar{\gamma}$ と書けば、β は $\bar{\gamma}$ 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。
- 過大推定（$\hat{\beta} > \beta$）の確率は、$p(\cdot)$ が 0 に近いほど高い。それから真の β より γ が高い事例が少ないほど高い。
  - 新規に公開するデータの $\Pr(a, b, c)$ を $\bar{\gamma}$ と等しくすれば、真の β の位置によらず、過大推定の確率は単調非増加 $\Rightarrow$ 事例が安全性のエビデンスとなる。

キー変数の選択方法について

- Elliot et al. (2011)：個体情報を広範に調査した上でキー変数を選択。
  - 攻撃用情報の見当がついたとして、いかにキーを選ぶか？
- Fung et al. (2010)：“open problem”.
- 本報告：匿名化水準の管理にとって最適に選ぶ。
  - 既存研究は使い方を定めないので選べない。
  - k 変数からキーを選ぶ方法は 2^k 通りで、そのうちどれを採用するかと考える。

キー変数の選択に係る一意数の変化

- 2^k 個の母集団一意数の順序データ: $u_{(1)} \leq u_{(2)} \leq \cdots \leq u_{(2^k)}$
- 例）住宅・土地統計調査匿名データの部分評価 $(2^{11} = 2048)$
  - “11vars”：都道府県、住宅以外の建物の種類、住宅以外の建物の所有関係、建物の構造、建物の階数（うち一戸建て・長屋、うち共同住宅)、むねの建築時期、建築面積、敷地面積、エレベータの有無、高齢者対応か
  - “ex region”：都道府県削除
  - “ex date”：むねの建築時期削除

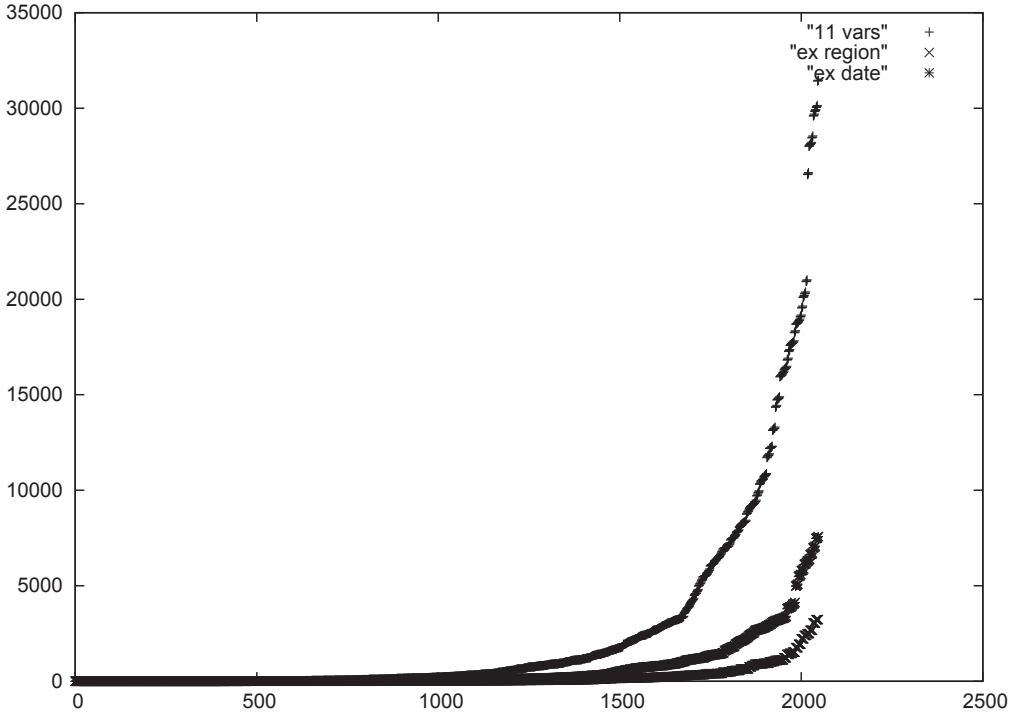

Figure 1: 標本一意数（縦軸）と順位（横軸）の関係

理論的なキー変数選択方針

- キー変数の選択とは、どの順位 r の $u_{(r)}$ を採用するか、という問題に他ならない。
- 選択した順位での一意数より小さい一意数を与えるキー変数しか使えない攻撃者は、（要因 (a,b) が一定なら）管理される。
- 順位 2^k を選べば全ての攻撃者を管理できる。しかし攻撃者がいない順位で評価したリスクは、実効リスクと違うので、識別不可能性の根拠にならない。
- ⇒攻撃者が存在する最大の順位 で一意数を評価する。
- 順位 $(1, 2, \ldots, 2^k)$ 上の攻撃者の分布で、最大値を推定したい。
  - 攻撃者は能力の範囲内で最も一意数を多く得られる順位に存在すると考える。

実際的なキー変数選択方針

- しかし分布の最大値の推定は困難で、分位点推定の方が現実的。
  - 資本規制でも 99%分位点を管理 (VaR)。
- 実際にはデータがないので、攻撃者分布の分位点は定量的に推定できない。
- 考察の主旨を活かせば、「大半」の攻撃者を管理するという方針でキー変数を選ぶのが実際的。つまり「公知」の変数をキーとする。
- 「大半」の外の攻撃者は、匿名化では管理できない。
  - 匿名化以外の手法が有効。例えば攻撃者分布の右裾に位置するような主体（名簿業者、個人情報収集組織等）にデータを渡さなければよい。
  - 識別事故が起きたときにうまく対応すれば当局への信認は上がると MacKey (2009) は議論。

例）住宅・土地統計調査 (H15) 匿名データ

- 公表サンプルサイズ： 31 万 266（世帯）；居住世帯ありのレコードのみ（つまり空き家は除く）。
- 母集団サイズ：4726 万（世帯）
- 標本抽出率=$\Pr(b|a)$：0.66 パーセント；単純無作為抽出とみなす。
- 攪乱は使われていないので $\Pr(a) = 1$ とみなす。

住宅・土地統計調査のキー変数

- Case 1：都道府県、世帯の種類、同居世帯の有無、夫婦の組数、家族類型、世帯の型、65 歳以上の世帯員の有無、75 歳以上の世帯員の有無、65 歳以上の世帯員のみか、75 歳以上の世帯員のみか、高齢夫婦の有無、世帯内の最高年齢
- Case 2：Case 1-都道府県
- Case 3：都道府県、世帯員各員について性別・年齢（15 歳未満は各歳）・配偶者の有無・続柄
- Case 4：Case 3+世帯主情報（性別、年齢、従業上の地位）
- Case 5：Case 4+現在の居住形態、所有の形態
- Case 6：Case 5+建物に関する事項、むねに関する事項、住宅の種類、所有関係、民営借家の所有区分、住宅の建て方、建築の時期
- Case 7：Case 6+地下室有無、自動車所有の有無、駐車スペースが敷地内、敷地外、住宅の購入・新築・建て替え等の別、H11 年以降の増改築有無
- Case 8：Case 7+台所、トイレ、浴室の設備状況

	S_1	$\Pr(c\|a,b)$	$\Pr(a,b,c)$
Case 1	4 918 819	.104	.00068
Case 2	1 683 983	.036	.00023
Case 3	5 038 968	.107	.00070
Case 4	6 871 365	.145	.00096
Case 5	9 374 185	.198	.00130
Case 6	29 082 561	.615	.00404
Case 7	35 610 454	.753	.00495
Case 8	42 962 590	.909	.00597

Table 1: 個体識別の容易度評価

まとめ

- 匿名データは、個体識別行為のモデルについて直接の実証対象である。
- 他の匿名データについても個体識別の容易度評価を行う予定。
- 本研究で使用した匿名データは統計法に基づいて（独行）統計センターから提供を受けた。