統計研究彙報　第77号　2020年３月　（1-14）

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

佐野夏樹†

服部雄太††

Utility Evaluation of Global Recoding by Classification Performance of Discrimination Model

SANO Natsuki
HATTORI Yuta

統計的開示性制御において、グローバルリコーディングは、代表的なカテゴリカルデータに対する秘匿方法である。一方で、カテゴリカルデータに対する有用性評価指標として、いくつかの情報損失評価指標が提案されているが、グローバルリコーディングは、複数のカテゴリを統合し、新たなカテゴリを生成するため、既存の情報損失評価指標が適用しにくいという問題がある。本研究では、グローバルリコーディングによって秘匿されたデータに対しても評価可能な、モデルの判別精度にもとづいた複数の情報損失評価指標を提案し、2010 年国勢調査データにもとづいて検証を行った。判別精度の尺度として、複数の尺度を比較検証した結果、Recall にもとづく評価指標が、情報損失評価指標として、妥当であることが明らかとなった。またその他の判別精度の尺度が適切でない理由として、日本における日本国籍の様に、一つのカテゴリが支配的な変数を予測する場合に、グローバルリコーディングの適用によってモデルの判別精度が向上している場合があることが確認された。

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、 プライバシー保護データマイニング、 ビッグデータ

In statistical disclosure control, global recoding is a typical masking method for categorical variable. Though some indices of utility evaluation of categorical data are proposed, they are actually difficult to be calculated for global recording data, because global recording merge plural categories and make a new category. In this study, we propose various indices of utility evaluation for categorical data which can evaluate the masked data by global recoding based on performance of discrimination model and verify them through numerical experiments for 2010 census data. As results, we obtain findings that the index based on recall is an appropriate measure. The reason why another indices are not appropriate is that they cannot evaluate utility appropriately when model predicts a variable including dominant category.

Keyword: Anonymization of micro data, Statistical disclosure control, Open data strategy of official statistics, Privacy preserving data mining, Big data

---

† 東京情報大学　総合情報学部　Email：ns207374@rsch.tuis.ac.jp
†† 名古屋大学大学院　医学系研究科　Email：u-ta.h@med.nagoya-u.ac.jp

1. はじめに

ミクロデータは、社会研究や経済研究など、社会科学分野において、しばしば利用されるデータであるが、原データを公開した場合、個人の特定等のリスクが存在する。通常は、この様なリスクを低下させるために、秘匿化処理が施される。秘匿化処理を施すことによって、リスクは低下するが、データ利用者が分析した結果が、本来得られた結果と異なる結果となり、利用者の有用性も低下する。このように、リスクと有用性の間には、トレードオフの関係があり、強力な秘匿化手法により、リスクを低減させた場合、有用性も大きく損なわれる。原データの秘匿化技術や、リスクや有用性の評価方法を総称して、統計的開示制御と呼ぶ。

秘匿化処理には、いくつかの技術が存在するが、まず、レコードの値そのものを変更するわけではない非攪乱手法として、グローバルリコーディングや局所秘匿（Hundepool et.al. 2012）があげられる。２番目に、レコードの値そのものを変更する攪乱手法としては、原データへのノイズ付加（Duncan and Pearson, 1991）、PRAM (Gouweleeuw et.al., 1998)、 マイクロアグリゲーション (Defays and Nanopoulos, 1993) 、シャッフリング (Muralidhar and Sarathy, 2006）があげられる。３番目に、原データの特定の性質を保存した合成データを生成する方法（Muralidhar and Sarathy, 2008）が挙げられる。

上記の方法によって秘匿化処理が施されたデータの有用性は、通常、秘匿データの原データに対する情報損失として評価される。原データが連続変数である場合は、情報損失評価は、比較的容易に実行できる。例えば、原データと秘匿データの間でレコード間の対応関係が明らかであれば、対応するレコード間の距離として計算できる。また対応関係が明らかでない場合は、原データと秘匿データの変数に対して、それぞれ、基本統計量を計算し、それらの基本統計量の乖離として、評価できる。より変数間の相関関係を考慮した情報損失を評価するならば、原データと秘匿データに対して、それぞれ、因子分析等の多変量解析手法を適用し、因子負荷量等の解析結果の乖離の程度として評価することも可能である。

しかしながら、カテゴリカルデータ、特にグローバルリコーディングによって秘匿化処理を行った場合は、カテゴリ統合によって、新たなカテゴリが生成されるため、原データと秘匿データの乖離の程度を把握することは、容易ではない。

本研究では、グローバルリコーディングを適用前の変数と適用後の変数を入力した２つの判別モデルを構築し、それらのモデル評価指標の差として、秘匿化手法がグローバルリコーディングであっても、評価可能な情報損失指標を提案する。また情報損失指標として、複数の判別精度基準を用いた指標を提案し、その中から、どの指標が妥当な指標であるか、2010 年の国勢調査データを用いた数値実験を通して、検討を行う。グローバルリコーディングを適用する際に、最小頻度の割合をどの程度にするかは、重要な問題であるが、本数値実験は、判別モデルを適用した際のモデル評価値として、秘匿化の影響の程度の示唆を与える。

2. グローバルリコーディング

本研究は、秘匿化処理方法として、もっとも基本的なグローバルリコーディングを対象とする。グローバルリコーディングは、対象とする変数に頻度の小さなカテゴリがある場合に、他のカテゴリと統合して、新たなカテゴリを生成する。例えば、表１(a)は原データ、表１(b)は、グローバルリコーディングを適用したデータの例である。原データでは、「農業」と「林業」は、異なるカテゴリとなっているが、グローバルリコーディングデータでは、「農業」と「林業」が統合され、「農業・林業」という新たなカテゴリが生成されていることがわかる。

頻度集計を行った結果として、頻度（頻度割合）の小さいカテゴリに属する個人や世帯は、特定されるリスクが高いため、頻度の小さいカテゴリは、その他のカテゴリと統合される。例えば、表２の頻度表を見ると、原データにおいて、「林業」カテゴリの頻度（割合）は、2（0.02）と、最も小さく、「農業」カテゴリと統合され、新たな「農業・林業」カテゴリを生成している。その際に、最小頻度割合 p を設定して、

p よりも頻度割合の小さいカテゴリを、統合の対象とする方法が、しばしば行われるが、最小頻度割合 p の設定によって、グローバルリコーディングの結果が異なるため、最小頻度割合 p の設定は、重要な課題である。

表１　原データとグローバルリコーディングデータの例

(a)　原データ

行番号	産業大分類
001	農業
002	林業
003	サービス業
004	不動産業
005	製造業

(b)　グローバルリコーディングデータ

行番号	産業大分類
001	農業・林業
002	農業・林業
003	サービス業
004	不動産業
005	製造業

表２　原データとグローバルリコーディングデータの頻度表の例

(a)原データの頻度表

	農業	林業	サービス業	不動産業	製造業	総数
頻度（割合）	5 (0.05)	2(0.02)	30(0.3)	13(0.13)	50(0.5)	100 (1.0)

(b) グローバルリコーディングデータの頻度表

	農業・林業	サービス業	不動産業	製造業	総数
頻度（割合）	7 (0.07)	30(0.3)	13(0.13)	50(0.5)	100 (1.0)

3.　カテゴリカルデータに対する情報損失指標の先行研究と課題

Domingo-Ferrer and Torra (2001) は、カテゴリカルデータに対する情報損失指標として、次の３つの情報損失指標（a）カテゴリ値の直接比較（b）分割表にもとづく指標（c）エントロピーにもとづく指標を紹介している。

3.1 カテゴリ値の直接比較

変数 V が名義尺度の場合に、元のカテゴリ c と秘匿後のカテゴリ c' の間の距離を、値を直接比較することによって、次の様に決める。

$$d_V(c,c') = \begin{cases} 0 \text{ if } c = c' \\ 1 \text{ if } c \neq c' \end{cases}, \qquad (1)$$

例えば変数Vは、性別であり、男もしくは女のカテゴリ値を取るとする。秘匿前のカテゴリはc =男、秘匿後のカテゴリは、c' =女とすると、(1)式による距離は、1 である。

変数Vが順序尺度の場合には、カテゴリ値 c と c' の順序によって最大値と最小値を決め、それらの間にあるカテゴリc''の数を変数Vの範囲で割ることによって次の様に決める。

$$d_V(c,c') = \frac{|c'':\min(c,c') \leq c'' < \max(c,c')|}{|D(V)|}, \qquad (2)$$

ここで $\leq (<)$ はVの範囲内における全順序（狭義）演算子を表し、$|D(V)|$ は、V の範囲の濃度を表す。

例えば変数Vは、年代を表す順序尺度のカテゴリ変数であり、10 歳未満、10 代、20 代、…、80 代、90 代超の 10 のカテゴリのいずれかを取るものとする。秘匿前のカテゴリは、$c = 30$ 代、秘匿後のカテゴリは、$c' = 50$ 代とすると、$|c''| = 2$、$|D(V)| = 10$であり、(2)式による距離は、0.2 である。

3.2 頻度表にもとづく尺度

原データセット X 、秘匿データセットX'、考慮変数集合Wが、最大次元Kが与えられた時に、頻度表（2 次元以上ならば分割表）にもとづく 情報損失指標は、次の様に計算される。

$$CTBIL(X,X';W,K) = \sum_{\substack{\{V_{j_1}\cdots V_{j_t}\}\subseteq W \\ |\{V_{j_1}\cdots V_{j_t}\}|\leq K}} \sum_{i_1\cdots i_t} \left| x_{i_1\cdots i_t} - x'_{i_1\cdots i_t} \right|, \quad (3)$$

頻度表にもとづく情報損失指標は、Wの部分集合の t 変数に対して、原データと秘匿データのそれぞれに、t 次元頻度表を集計し、その差を合算した指標である。ここで$x_{i_1\cdots i_t}$ と $x'_{i_1\cdots i_t}$ はそれぞれ、原データと秘匿データに対する頻度表の$i_1 \cdots i_t$要素を表す。頻度表にもとづく情報損失指標は、考慮集合の大きさ$|W|$、各変数のカテゴリ数、最大次元 K に依存するので、(3)式を頻度表のセル数で割り、正規化した指標も用いられる。

3.3 エントロピーベースの指標

PRAM は、推移確率行列$P(V'|V)$を用いた攪乱手法であるが、グローバルリコーディングも推移確率行列を用いて表現することができる。ここで、VとV'はそれぞれ、原変数と秘匿処理を施した変数を表す。De Wall と Willenborg（1999）は、推移確率行列とシャノンエントロピーを用いて、情報損失の評価を行った。推移確率 $P(V'|V)$ と事前確率$P(V)$ 、ベイズの定理を用いると、事後確率 $P(V|V')$ が計算できる。事前確率$P(V)$は、原データのカテゴリ頻度によって推定される。また事後確率は、秘匿後のカテゴリが得られている時の秘匿前カテゴリの推定確率を表す。特定のレコードの秘匿値が$V' = j$の値を取る時のエントロピーは、次の様になる。

$$H(V|V' = j) = -\sum_{i=1}^{n} P(V = i|V' = j)\log P(V = i|V' = j). \quad (4)$$

エントロピーベースの情報損失指標は、(4)式を全レコードに計算し、集計した次の指標である。

$$EBIL(P_V, X, X') = \sum_{r\in X'} H(V|V' = j_r). \quad (5)$$

3.4 グローバルリコーディングデータに対する情報損失評価の課題

グローバルリコーディングによって、原データの秘匿を行った場合、表１にしめされる様に、原データにおける複数のカテゴリーが統合され、新たなカテゴリが形成される。したがって、統合後のカテゴリと統合前のカテゴリを（1）式によって比較した場合、異なるカテゴリとみなされるため、適切に距離を計算することが、難しい。

頻度表による情報損失評価は、表２からわかる様に、原データとグローバルリコーディングデータから集計された頻度表の大きさが異なることから、(3) 式におけるセルの対応関係が明らかではなく、実際に計算することは難しい。

(4) 式によるエントロピーベースの情報損失指標は、秘匿手法が、グローバルリコーディングの場合においても、計算可能な評価尺度ではあるが、(1)式や(3)式による情報損失指標が、原データと秘匿データの乖離を評価しており、情報損失の尺度であることが容易に理解できることに対して、エントロピーを評価していることから、情報量の尺度ではあるが、損失の尺度とみなせるかどうかは、明らかではない。

4.　モデルベースの情報損失指標

4.1 情報損失指標の要求特性

本研究では、秘匿データに対して、何らかの統計モデルが適用される場面を想定し、原データに対するモデルの判別精度とグローバルリコーディングを適用したデータに対するモデル判別精度の差として、情報損失を定義し、有用性評価を行う。情報損失が大きいことは、匿名データの有用性が小さいことを意味し、逆に情報損失が小さいことは、匿名データの有用性が大きいことを意味する。本研究では、まず、モデルを用いて情報損失評価を行う際に、モデルベースの情報損失評価指標が満たすべき、要求特性として以下の特性を考える。

要求特性：モデルへの入力変数に攪乱を入れるほど、情報損失は増加する

上記の要求特性を分析の対象となる変数がカテゴリ、データの秘匿手法がグローバルリコーディングの場合に対して、より具体的に、以下の様に展開する。

(a) 情報損失は正の値をとる
(b) 最小頻度割合が増加するほど、情報損失は増加する
(c) モデルに入力する変数の数が増加するほど、情報損失は増加する
(d) 原データに対して、判別精度の高いモデルほど、情報損失が大きい

上記の要求特性(a)，(b)，(c)，(d)を実際に、情報損失評価指標が満たしているかどうかを検討する。

4.2 検討する情報損失評価指標

モデルベースの情報損失評価は、原データの変数を用いて原データの変数を予測するモデル（原データモデル）とグローバルリコーディングを適用した変数を用いて、原データの変数を予測するモデル（グローバルリコーディングモデル）の２つのモデルを構築し、これらのモデルの判別精度の差として、情報損失の評価を行う。

x_i, x_i^Gを原データ、グローバルリコーディングデータ、それぞれの i 番目の変数、i 番目の変数を除いた変数を、それぞれ、$\boldsymbol{X}^{-i}=\{x_1,x_2,\cdots,x_{i-1},x_{i+1},\cdots,x_m\}$、$\boldsymbol{X}_G^{-i}=\{x_1^G,x_2^G,\cdots,x_{i-1}^G,x_{i+1}^G,\cdots,x_m^G\}$とすると、原データモデルは、$\boldsymbol{X}_{sub}^{-i}\subseteq\boldsymbol{X}^{-i}$を入力変数として、$x_i$を予測するモデル、$f\left(\boldsymbol{X}_{sub}^{-i}\right)$である。一方で、グローバルリコーディングモデルは、$\boldsymbol{X}_{sub,G}^{-i}\subseteq\boldsymbol{X}_G^{-i}$を入力変数として、$x_i$を予測するモデル、$f^G\left(\boldsymbol{X}_{sub,G}^{-i}\right)$である。

以上の表記のもとで、次のモデルベースの情報損失指標

$$IL\left(M,x_i,\boldsymbol{X}_{sub}^{-i},\boldsymbol{X}_{sub,G}^{-i}\right)=M\left(x_i,f\left(\boldsymbol{X}_{sub}^{-i}\right)\right)-M\left(x_i,f^G\left(\boldsymbol{X}_{sub,\mathrm{G}}^{-i}\right)\right),\qquad(6)$$

を検討する。予測モデル$f(\cdot)$として、ここでは、多項ロジスティックモデルを用いる。M はモデルの判別精度を評価する尺度であり、本研究では、情報検索分野でしばしば、利用される 4 つの判別精度の尺度、Precision（適合率）、Recall（再現率）、F-value（F−値）、Accuracy（正確度）（Powers, David M W, 2011）を用いた場合の情報損失評価指標 IL を比較検討する。

4.3 多項ロジスティックモデル

多項ロジスティックモデルは、入力変数$\boldsymbol{y}=(y_1,y_2,\cdots,y_m)'$、所属クラスをしめす出力変数$z$がある時に、対象サンプルの所属クラスが j である確率をソフトマックス関数によって

$$p(z=j)=\frac{\exp\left(\alpha_j+\boldsymbol{\beta}_j'\boldsymbol{y}\right)}{\sum_{k=1}^{K-1}\exp(\alpha_k+\boldsymbol{\beta}_k'\boldsymbol{y})+1}$$

とするモデルである。ここで$\alpha_k,\boldsymbol{\beta}_k,k=1,\cdots,K-1$は、多項ロジスティックモデルのパラメータであり、

通常は最尤法によって、サンプルから計算される。多項ロジスティックモデルによって、予測を行う場合は、入力変数**y**に対して、

$$f(\boldsymbol{y}) = \arg\max_j p(z = j)$$

とする。本研究では、R パッケージ nnet 中の多項ロジスティックモデルの関数 multinom（Venables and Ripley (2002)）を用いた。

4.4 判別精度の尺度

表 3　混合行列

実際のクラス＼予測クラス	カテゴリ a	カテゴリ b	カテゴリ c	行合計
カテゴリ a	g_{aa}	g_{ab}	g_{ac}	$g_{a\cdot}$
カテゴリ b	g_{ba}	g_{bb}	g_{bc}	$g_{b\cdot}$
カテゴリ c	g_{ca}	g_{cb}	g_{cc}	$g_{c\cdot}$
列合計	$g_{\cdot a}$	$g_{\cdot b}$	$g_{\cdot c}$	全合計 $g_{..}$

あるサンプルのカテゴリを判別モデルによって予測したカテゴリと実際のカテゴリを集計した分割表を混合行列（表 3）と呼ぶ。混合行列の各要素には、度数が配置され、添字の左側は、実際のカテゴリ、右側は、予測したカテゴリを表す。カテゴリ a に着目した場合、g_{ba}、 g_{ca}は、カテゴリ a の第１種の誤りの度数に相当し、g_{ab}、 g_{ac}は、カテゴリ a の第２種の誤りの度数に相当する。通常、第１種の誤り度数と第２種の誤り度数は、トレードオフの関係を持つ。例えば、すべてのサンプルをカテゴリ a と予測した場合、g_{ab}、g_{ac}は 0 となるが、実際に全てのサンプルがカテゴリ a で無い限り、g_{ba}、g_{ca}は、相当に大きな値となる。混合行列をもとに、多クラス（カテゴリ）における Precision（適合率）、Recall（再現率）、F-value（F−値）、Accuracy（正確度）を判別精度の尺度として、以下の様に計算する。

$$p_a = \frac{g_{\mathrm{aa}}}{g_{\cdot \mathrm{a}}}, p_b = \frac{g_{\mathrm{bb}}}{g_{\cdot \mathrm{b}}}, p_c = \frac{g_{\mathrm{cc}}}{g_{\cdot \mathrm{c}}}, \quad \mathrm{Precision} = \mathrm{Mean}(p_a, p_b, p_c)$$

$$r_a = \frac{g_{\mathrm{aa}}}{g_{\mathrm{a}\cdot}}, r_b = \frac{g_{\mathrm{bb}}}{g_{\mathrm{b}\cdot}}, r_c = \frac{g_{\mathrm{cc}}}{g_{\mathrm{c}\cdot}}, \quad \mathrm{Recall} = \mathrm{Mean}(r_a, r_b, r_c)$$

$$f_a = \frac{2 \times p_a \times r_a}{p_a + r_a}, \quad f_b = \frac{2 \times p_b \times r_b}{p_b + r_b}, \quad f_c = \frac{2 \times p_c \times r_c}{p_c + r_c}, \quad \text{F-value} = \mathrm{Mean}(f_a, f_b, f_c)$$

$$\mathrm{Accuracy} = \frac{g_{\mathrm{aa}} + g_{\mathrm{bb}} + g_{\mathrm{cc}}}{g_{..}}$$

$g_{\cdot \mathrm{a}}, g_{\cdot \mathrm{b}}, g_{\cdot \mathrm{c}}$ は、それぞれ、カテゴリ a, b, c に予測されたサンプル数であり、$g_{\mathrm{a}\cdot}$, $g_{\mathrm{b}\cdot}, g_{\mathrm{c}\cdot}$ は、実際のクラスが、それぞれ、カテゴリ a, b, c のサンプル数である。p_a, p_b, p_cは、各カテゴリの Precision、r_a, r_b, r_cは各カテゴリのリコール、f_a, f_b, f_cは、各カテゴリの F-value である。また、$g_{..}$は、全サンプル数、Mean(·) は、括弧内の平均を表す。カテゴリ a の Precision, p_aは, カテゴリ a と予測したサンプルのうち、実際にカテゴリ a であった割合を表し，カテゴリ a の Recall, r_aは，実際のカテゴリが a であるサンプルのうち, 正しくカテゴリ a と予測できた割合を表し、統計学における感度（Sensitivity）に相当する。第１種の誤りと第２種の誤りがトレードオフの関係を持つことより、Recall と Precision もトレードオフの関係を持つため、調和平均により、平均を評価したのが、F-value である。

5.　数値実験

5.1　利用データ

2010 年国勢調査データから抽出された 128, 280　レコードを対象にし、使用する変数は、産業大分類、職業大分類、家族類型、国籍、子供の数と年齢による分類（子供の数と最年少・最年長の子供の年齢による類型）の５変数を対象に数値実験を行った。

5.2 評価対象モデル

4.2節において、m=5、x_1 = 産業大分類, x_2 = 職業大分類, x_3 = 家族類型, x_4 = 国籍, x_5 = 子供の数と年齢であり、原データモデルの予測対象の変数x_iの選び方は、5 つの変数から 1 変数を選ぶ組合せの数なので ${}_5C_1$=5 である。入力変数の数を I とすると原データモデルの入力変数$\boldsymbol{X}_{sub}^{-i}$の組合わせ数は、5 つの変数から予測対象の変数を除いた 4 変数から I 個の変数を選ぶ組合せの数 ${}_4C_I$ なので、I=1, 2, 3, 4 に対して、それぞれ、${}_4C_I$ は 4, 6, 4, 1 である。したがって、予測対象の変数x_iと入力変数$\boldsymbol{X}_{sub}^{-i}$の組合せは、$I$=1, 2, 3, 4 に対して、それぞれ、$5 \times {}_4C_I$ =20、30、20、5 となり、これらの数の原データモデルを構築する。グローバルリコーディングモデルにおいても同じ数のモデルを構築し、(6)式によって、情報損失の評価を行う。

5.3 最小頻度によるグローバルリコーディング手続き

最小頻度割合 p を 0.01,　0.03,　0.05 として、図１の手続きにしたがい、原データにグローバルリコーディングを適用し、グローバルリコーディングデータを作成する。

1. 入力 n: レコード数, p: 最小頻度割合
2. 閾値の計算 $th = n \times p$,
3. 各変数に対して、集計を行い、分割表を作成し、カテゴリ頻度によって、ソートする。

カテゴリ	C_1	C_2	…	C_i	…	C_s
頻度	h_1	h_2	…	h_i	…	h_s

4. 最小頻度カテゴリと二番目に頻度の小さなカテゴリを統合する。
   もし C_1 が唯一の最小頻度カテゴリなら、二番目に頻度の小さなカテゴリC_2 と統合する。
   そうでなく、複数の最小頻度カテゴリが存在するならば $(h_1 = h_2 \cdots)$、それらのカテゴリを統合する。
5. ステップ 3 とステップ 4 を最小頻度が閾値 th, を超えるまで再帰的に繰り返す。

図１ 最小頻度割合によるグローバルリコーディング手続き

表 4 は、原データおよびグローバルリコーディングデータの各変数のカテゴリ数をあらわす。表 4 から、p が増加する度に、カテゴリ数が減少し、特に、国籍は、p=0. 03,　0. 05 の時に、カテゴリ数が１となっていることがわかる。

5.4 予備解析：原データにおける変数間の相関関係

変数間の相関関係を把握するために、次のクラメールの連関係数（Cramér,H.1946）

$$V = \sqrt{\frac{\chi_0^2}{n\{(\min(q,r)-1\}}} \qquad (7),$$

を計算した結果を表５にしめす。ここで $\chi_0^2 = \sum_{i=1}^{q}\sum_{j=1}^{r}\frac{(n_{ij}-m_{ij})^2}{m_{ij}}$ はカイ二乗統計量であり、 n_{ij} と m_{ij} は、それぞれ、分割表の (i,j) セルの観測度数と期待度数を表す。n は総度数、q, r はそれぞれ、行カテ

ゴリ、列カテゴリのサイズを表す。クラメールの連関係数Vは、 $0 \leq V \leq 1$ の値を取り、2 つのカテゴリカル変数の相関関係を表す。
表 5 を見ると、産業大分類と職業大分類の間で、クラメール連関係数が 0.663 と最も、大きな値を取っていることがわかる。ついで、家族類型と子供の数と年齢による分類の間で 0.271 と比較的大きな値を取っていることがわかる。

表 4　グローバルリコーディング後の各変数のカテゴリ数

	原データ	$p = 0.01$	$p = 0.03$	$p = 0.05$
産業大分類	22	17	10	7
職業大分類	13	12	9	8
家族類型	27	16	10	7
国籍	52	2	1	1
子供の数と年齢による分類	97	29	12	8

表 5 クラメール連関係数
太字: 0.25 よりも大きな値

	産業大分類	職業大分類	家族類型	国籍	子供の数と年齢による分類
産業大分類	—	**0.663**	0.050	0.024	0.055
職業大分類	—	—	0.065	0.027	0.072
家族類型	—	—	—	0.023	**0.271**
国籍	—	—	—	—	0.020
子供の数と年齢による分類	—	—	—	—	—

5.5 要求特性の検証
5.5.1 要求特性(a)(b)(c)の検証
要求特性を検証するために、モデルの判別精度の評価尺度 M に(a)Precision(b)Recall(c)F-value(d)Accuracy を取り、各入力変数、各最小頻度割合 p におけるモデルの各情報損失指標を (6) 式にもとづいて計算し、その平均値を表 6 にしめす。

表 6 情報損失指標の平均値

(a) 判別精度尺度を Precision とした場合

	$p = 0.01$	$p = 0.03$	$p = 0.05$
入力変数の数= 1	−0. 003	0. 010	0. 014
入力変数の数= 2	−0. 080	−0. 102	−0. 071
入力変数の数= 3	−0. 105	−0. 087	−0. 121
入力変数の数= 4	−0. 085	−0. 040	−0. 052

(b) 判別精度尺度を Recall とした場合

	$p = 0.01$	$p = 0.03$	$p = 0.05$
入力変数の数= 1	0. 001	0. 004	0. 019
入力変数の数= 2	0. 002	0. 012	0. 035
入力変数の数= 3	0. 003	0. 021	0. 047
入力変数の数= 4	0. 005	0. 033	0. 060

(c) 判別精度尺度を F-value とした場合

	$p = 0.01$	$p = 0.03$	$p = 0.05$
入力変数の数= 1	−0. 108	−0. 137	−0. 134
入力変数の数= 2	−0. 133	−0. 212	−0. 203
入力変数の数= 3	−0. 124	−0. 176	−0. 210
入力変数の数= 4	−0. 133	−0. 225	−0. 211

(d) 判別精度尺度を Accuracy とした場合

	$p = 0.01$	$p = 0.03$	$p = 0.05$
入力変数の数= 1	0. 000	0. 007	0. 010
入力変数の数= 2	0. 001	0. 012	0. 019
入力変数の数= 3	0. 001	0. 018	0. 029
入力変数の数= 4	0. 000	0. 024	0. 041

表 6(*a*)の判別精度尺度を Precision にした場合の情報損失を見ると、負の値を取る場合が多く、要求特性 (*a*)、要求特性 (*b*)、要求特性 (*c*)を全て満たしていないことがわかる。一方で、表 6(*b*)の判別精度尺度を Recall にした場合の情報損失を見ると、要求特性 (*a*)、要求特性 (*b*)、要求特性 (*c*)を全て満たしていることがわかる。表 6(*c*)の判別精度尺度を F-value にした場合の情報損失を見ると、Precision と Recall の調和平均であるため、Recall の場合と同様に、要求特性 (*a*)、要求特性 (*b*)、要求特性 (*c*)を満たしていないことがわかる。表 6(*d*)の判別精度尺度を Accuracy にした場合の情報損失を見ると、要求特性 (*a*)、要求特性 (*b*)を満たしているが、入力変数の数 4、*p*=0. 01 の場合に、正確度が 0 であり、要求特性(*c*)を満たせていないことがわかる。

以上の結果から、情報損失評価の指標としては、判別精度尺度を Recall にした場合の情報損失尺度が、要求特性(*a*), (*b*), (*c*)を満たす評価指標であることがわかる。

5.5.2　要求特性 (*d*)の検証

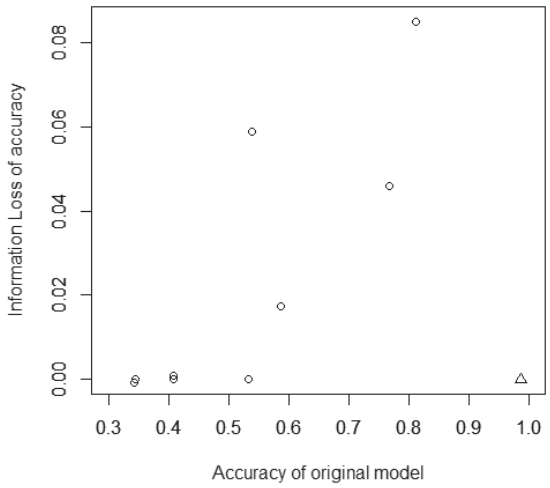
要求特性(d)は、以下の理由から、要求特性の一つとして、検証の対象としている。

4.1 節では、情報損失評価指標に要求される特性を「モデルへの入力変数に攪乱を入れるほど、情報損失は増加する」としたが、モデルへの入力変数の中でも、モデルの判別精度へ寄与しない変数と寄与する変数が存在する。モデルの判別精度に寄与する変数に対して、グローバルリコーディングを適用した場合、グローバルリコーディングモデルの判別精度は、原データモデルから大きく低下すると考えられる。一方で、モデルの判別精度に寄与しない変数に対して、グローバルリコーディングを適用しても、原データモデルとグローバルリコーディングモデルの判別精度は、それほど、変わらないと考えられる。結果として、高い判別精度をしめす原データモデルは、情報損失が大きく、判別精度の低い原データモデルは、情報損失が小さいため、原データモデルの判別精度と情報損失の間には、高い相関関係があると期待される。

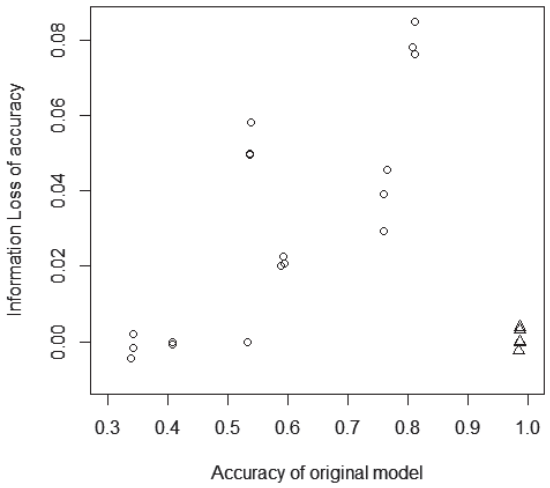
しかしながら、実際に数値実験によって、原データモデルの判別精度と情報損失の相関関係を検証したところ、支配的なカテゴリが存在する変数を予測するモデルにおいて、その他のモデルと異なる傾向が見られた。例えば、グローバルリコーディングにおける条件を $p=0.05$ にした場合の原データモデルの Accuracy と判別精度尺度に Accuracy を用いた場合の情報損失の間の散布図（図 2）を見ると、原データモデルの Accuracy と判別精度尺度に Accuracy を用いた場合の情報損失の間には、正の相関がある様に見える。一方で右下に、複数の外れ値（△のプロット）が存在する。これらのプロットは、国籍を予測するモデルであり、日本における国籍は、ほとんどが、日本国籍であるため、グローバルリコーディングモデルが、全員を日本国籍と予測した場合でも、ほとんど正しい予測であるため、高い Accuracy 値、低い情報損失となっている。

原データモデルの判別精度と情報損失との相関係数また、国籍の予測モデルを除いた場合の相関係数を表 7 にしめす。国籍の予測モデルを除いた相関係数は、全モデルを用いた相関係数に比べて、判別精度基準を Recall とした場合を除き、相関係数が大きく増加していることがわかる。例えば、Accuracy に対する全モデルを用いた相関係数は、入力変数の数が 1 から 4 までに対して、それぞれ（a）0.209（b）0.191（c）0.017（d）−0.402 と小さな値であるが、国籍予測モデルを除くと(a) 0.757 (b)0.780 (c) 0.726 (d) 0.679 まで増加していることがわかる。

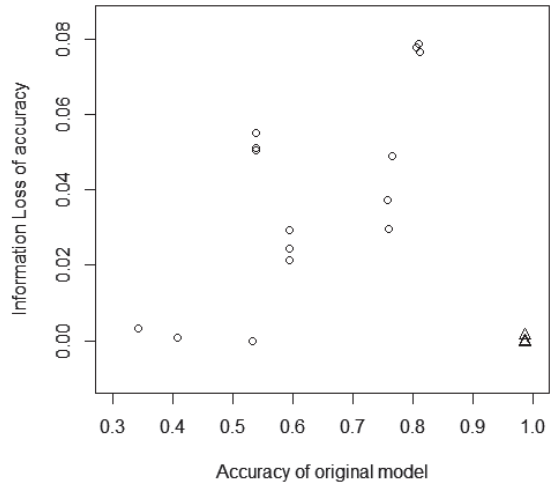
結論として、国籍の様な支配的なカテゴリが存在する変数の予測モデルを除くと、原データモデルの判別精度と情報損失の間には、高い相関関係が存在する。ただし、判別精度の尺度として Recall を用いた場合、支配的なカテゴリの有無に関わらず、高い相関係数の値がしめされた。以上のことから、情報損失評価を行う際の判別精度の尺度として、Recall を適用することが望ましい。



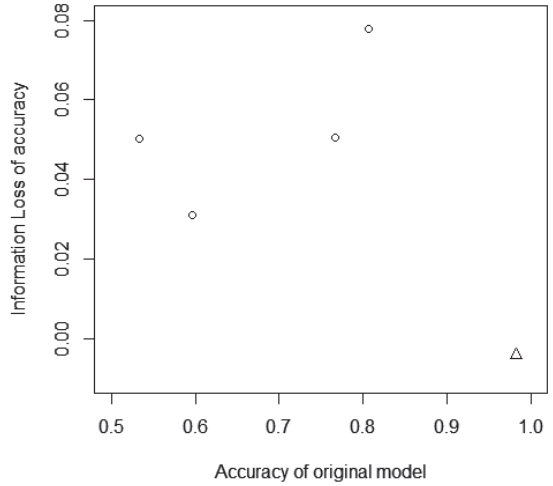
(a)　入力変数の数＝1



(b)　入力変数の数＝2



(c)　入力変数の数＝3



(d)　入力変数の数＝4

△：国籍の予測モデル　○：その他の予測モデル

図２　原データモデルの Accuracy と判別尺度を Accuracy とした場合の情報損失の散布図（p=0.05）

表 7 原データモデルの判別精度と情報損失との相関係数（$p = 0.05$）

(a) 入力変数の数 = 1

	Precision	Recall	F-value	Accuracy
全モデル（20 モデル）	0.277	0.985	0.850	0.209
国籍予測モデルを除いたモデル（16 モデル）	0.774	0.987	0.954	0.757

(b) 入力変数の数 = 2

	Precision	Recall	F-value	Accuracy
全モデル（30 モデル）	0.637	0.980	0.776	0.191
国籍予測モデルを除いたモデル（24 モデル）	0.860	0.979	0.897	0.780

(c) 入力変数の数 = 3

	Precision	Recall	F-value	Accuracy
全モデル（20 モデル）	0.820	0.983	0.616	0.017
国籍予測モデルを除いたモデル（16 モデル）	0.931	0.981	0.896	0.726

(d) 入力変数の数 = 4

	Precision	Recall	F-value	Accuracy
全モデル（5 モデル）	0.990	0.991	0.786	−0.402
国籍予測モデルを除いたモデル（4 モデル）	0.992	0.991	0.992	0.679

5.6 情報損失が大きいモデルの検証

どの変数を入力変数として、どの変数を予測するモデルの情報損失が大きいのかを検証するために、p = 0.05 の場合における情報損失値が大きな上位 5 モデルを表 8 にしめす。～の左に出力変数、右に入力変数を配置しているが、5.4 節の予備解析において、クラメール連関係数が大きな値を示した変数対（I：産業大分類、O：職業大分類）、（F：家族類型、C：子供の数と年齢による分類）を太字でしめしている。情報損失の値が大きなモデルは、クラメール連関係数の大きな変数対の片方を出力変数、もう片方を入力変数としていることがわかる。つまり、5.5.2 節で述べたように、出力変数に対して、予測に寄与する変数が入力変数となっている場合に、情報損失の値が大きくなっていることがわかる。

表８　情報損失（IL）の大きな上位５モデル（$p = 0.05$）

(a)　入力変数の数=1

Model	IL precision	Model	IL recall	Model	IL f	Model	IL accuracy
O~I	0.176	**O~I**	0.237	**O~I**	0.158	**O~I**	0.085
I~O	0.159	**I~O**	0.103	**I~O**	0.056	**C~F**	0.059
I~N	0.148	**F~C**	0.038	I~F	0.000	**I~O**	0.046
C~N	0.049	**C~F**	0.004	I~C	0.000	**F~C**	0.017
C~F	0.047	F~I	0.002	O~F	0.000	F~N	0.001

(b) 入力変数の数=2

Model	IL precision	model	IL recall	Model	IL f	Model	IL accuracy
N~IC	0.494	**O~I**N	0.238	N~OC	0.495	**O~I**N	0.085
O~IF	0.174	**O~I**F	0.196	**O~I**C	0.112	**O~I**C	0.078
I~ON	0.148	**O~I**C	0.182	**O~I**F	0.085	**O~I**F	0.076
O~IC	0.132	**I~O**C	0.133	**O~I**N	0.04	**C~F**N	0.058
O~IN	0.132	**I~O**N	0.111	**I~O**F	0.023	**C~I**F	0.050

(c) 入力変数の数=3

Model	IL precision	Model	IL recall	Model	IL f	Model	IL accuracy
C~ION	0.244	**O~I**FN	0.197	**O~I**FC	0.214	**O~I**NC	0.079
O~IFN	0.198	**O~I**NC	0.179	**O~I**NC	0.064	**O~I**FC	0.078
O~INC	0.162	**O~I**FC	0.179	**O~I**FN	0.04	**O~I**FN	0.076
O~IFC	0.144	**I~O**NC	0.131	N~IOF	0.001	**C~IOF**	0.055
I~OFN	0.127	**I~O**FC	0.098	N~IOC	0.000	**C~OF**N	0.051

(d) 入力変数の数=4

Model	IL precision	Model	IL recall	Model	IL f	Model	IL accuracy
O~IFNC	0.141	**O~I**FNC	0.190	**O~I**FNC	0.217	**O~I**FNC	0.078
F~IONC	0.126	**I~O**FNC	0.102	**I~O**FNC	0.020	**I~O**FNC	0.051
I~OFNC	0.048	**F~IONC**	0.004	**F~IONC**	-0.241	**C~IOF**N	0.050
C~IOFN	-0.212	**C~IOF**N	0.004	**C~IOF**N	-0.315	**F~IONC**	0.031
N~IOFC	-0.366	N~IOFC	0.001	N~IOFC	-0.733	N~IOFC	-0.004

I：産業大分類 O：職業大分類 F：家族類型 N：国籍 C：子供の数と年齢による分類

太字：クラメール連関係数が 0.25 以上の変数対

～の左が出力変数、右が入力変数をあらわす

6　結論

本研究では、グローバルリコーディングが適用された秘匿データに対して、情報損失の評価が可能なモデルベースの情報損失評価指標を複数個提案し、これらの中から、情報損失評価指標の要求特性を満たす指標を数値実験により検討した。結論として、Recall によるモデルベース情報損失評価指標が、要求特性を満たす指標であることが明らかとなった。またその他の情報損失評価指標が、要求特性を満たせない要因として、支配的なカテゴリが存在する変数の予測モデルを適切に評価することが難しいことがわかった。

また本数値実験の結果は、秘匿データ作成者からすれば、グローバルリコーディングを適用する際に、所望の情報損失評価値を得るためには、どの程度の最小頻度割合に設定すれば、良いかを示唆するものである。

今後の課題として、全ての変数に対して、提案手法によるモデルベースの情報損失評価を行う場合、考慮する変数の組合せが膨大になるため、事前に対応分析のようなカテゴリカル変数に対する情報要約手法を適用し、主要な変数グループ内において、提案手法を適用するような、より実用的な枠組みの開発が挙げられる。

謝辞

本論文の執筆にあたって、２名の査読者から大変、丁寧でかつ有益な助言を頂いた。心より感謝の意を表す。本研究は、文部科学省科学研究費補助金基盤研究（A)、関西大学ソシオネットワーク戦略研究機構公募研究による研究助成を受けて行なわれた。

References

1. Cramér, H. 1946 Mathematical Methods of Statistics. Princeton: Princeton University Press, pp. 282
2. Defays D. and Nanopoulos P. 1993 Panels of enterprises and confidentiality: the small aggregates method. Proc. of 92 Symposium on Design and Analysis of Longitudinal Surveys, pp. 195-204. Statistics Canada, Ottawa.
3. De Waal A.G. and Willenborg L.C.R.J. 1999 Information Loss Through Global recoding and Local Suppression, Netherlands Official Statistics (special issure on SDC), 14, pp.17-20.
4. Domingo-Ferrer J. and Torra V. 2001 Disclosure protection methods and information loss for microdata. In Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, pp. 111-134, North-Holland, Amsterdam.
5. Duncan G.T. and Pearson R.W. 1991 Enhancing access to microdata while protecting confidentiality: prospects for the future. Statistical Science 6, 219-239.
6. Gouweleeuw J.M., Kooiman P., Willenborg L.C.R.J. and de Wolf P.P. 1997 Post randomisation for statistical disclosure control: Theory and implementation. Technical report, Statistics Netherlands. Research paper no. 9731.
7. Hundepool A., Domingo-Ferrer J., Franconi L., Giessing S., Nordholt E.S., Spicer K., and de Wolf P.P. 2012, Statistical Disclosure Control. Wiley, Chichester, UK.
8. Muralidhar K. and Sarathy R. 2006 Data shuffling: a new masking approach for numerical data. Management Science 52(5), 658-670.
9. Muralidhar K. and Sarathy R. 2008 Generating sufficiency-based non-synthetic perturbed data. Transactions on Data Privacy 1(1), 17-33.
10. Powers D. M. W, 2011, Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies. 2 (1) 37–63.
11. Venables,W. N. and Ripley, B. D. 2002 Modern Applied Statistics with S. Fourth edition. Springer.
12. Statistics Bureau of Japan 2010 Summary Report of the 2010 Population Census (http://www.stat.go.jp/english/data/kokusei/2010/final_en/final_en.html, 2018 June 14 accessed)