西イングランド大学

Ritchie, F. (2011) 研究環境における統計的開示抑制。研究結果報告書。WISERD Data Resources.

公開されたバージョンを引用すること。
出版社の URL は以下のとおりである。
http://www.wiserd.ac.uk/files/7913/6543/6668/WISERD_WDR_006.pdf

紹介者：なし

(注記なし)

免責事項

UWE は、預託された資料の権原について、また、かかる資料の預託権利について全ての預託者から保証を得ている。

UWE は、預託された資料のいずれについても、商業的有用性、権原又は特定目的への適合性又は他の保証を明示的にも黙示的にも表明又は保証しない。

UWE は、資料の利用が特許、著作権、商標又は他の財産権或いは所有権を侵害しないことを一切表明しない。

UWE は、預託された資料の知的財産権の侵害について一切責任を負わないが、かかる侵害の申立てが発生した場合は、調査が終了するまで当該資料を公開しない意向である。

ウェールズ社会経済研究・データ・手法研究所
Sefydliad Ymchwil Gymdeithasol ac
Economaidd, Data a Dulliau Cymru

研究環境における
統計的開示検出及び抑制

WISERD DATA RESOURCES

WISERD/WDR/006

Felix Ritchie

2011 年 12 月

著者
Felix Ritchie

連絡窓口
国家統計局ミクロデータ分析・ユーザーサポート局
Cardiff Road Newport South Wales NP10 8XG

電子メール：felix.ritchie@ons.gsi.gov.uk 又は felix.ritchie@virgin.net

WISERD 本部連絡先

Cardiff University 46 Park Place Cardiff
CF10 3BB

電話：02920879338

電子メール：wiserd@cardiff.ac.uk

1. 摘要

研究環境における統計的開示抑制(SDC)は特殊な問題を引き起こす。SDC 研究の多くは、有限個の集計表の集まりが開示から安全であることを保証すること又は、ミクロデータセットを十分に匿名化することに関連する。研究環境はその性質上、機密データをほとんど制限されずに分析のために利用し得る環境である。この環境に特定して考案されていない SDC ルールを課すと、過度に複雑で、それにもかかわらず柔軟性及び有効性の目標を達成できないルールになる可能性がある。

本論文で主張するように、研究環境は、必要とされる「曖昧さ」を解釈に組み込んだ単純なルールを少数集めたものを基本とする、従来の SDC と異なるアプローチを要求する。このアプローチは、(a) SDC の原則及び目的全般に対する明確な同意、(b) 安全な成果物及び危険な成果物のクラスの実証及び(c) 研究者の積極的な関与を要求する。ただし、これは実務上の問題を複数引き起こす。

キーワード
機密保持、成果物保護

謝辞
本文のレビュー及び追加セクションの実用性に対する提言をいただいた Rhys Davies に感謝の意を表する。ONS の同僚並びに 2006 年の CAED 会議及び 2007 年のデータアクセスに関するワークショップの出席者の方々にも心より謝意を表する。瑕疵及び不作為は全て著者に帰する。本書に示した意見は著者の意見であり、必ずしも HM 政府の見解を表明するものではない。

本論文は当初、回覧書で、Ritchie F (2007)*研究環境における統計的検出及び開示抑制*、謄写版、国家統計局、6 月と称された。

1. はじめに

国家統計機関(NSI)の役割は古くから、個人及び企業のあらゆる側面について大量の情報を収集することであった。NSI の中核的役割は、上記の情報源から作成した集計表の公表であるが、ここ数年を見ると、基盤になるミクロデータの利用による研究可能性が次第に重要になりつつある。

NSI の多くは、ミクロデータに対する何らかのアクセスを提供しているが、この範囲は国によってもデータの種類によっても著しく異なる。例えば、社会的データは情報の内容を著しく損傷させることなく有効に匿名化できるため、これに対するアクセスは広範囲に及んでいる。これに対し、企業データの利用は通常、社会的データよりも制限が厳しく、仮にある場合でも、攪乱又は匿名化はほとんど行われない。

アクセス方法については、機密データの発行はたいてい、(オーストラリア及びニュージーランドのように) 特別なライセンスの使用又はリモートジョブ投入モデルに限定される。開示ミクロデータを研究するための専用のラボ施設を提供している NSI も複数ある。これは (米国、カナダ、イタリア又はドイツのように) 物理的に管理された場所に置かれることもあれば、(デンマーク、スウェーデン及びオランダのように)「バーチャルラボ」を通じて提供されることもある。

最新技術、特に、ユーザーに優しいシンクライアントシステムの開発は、ラボ施設の提供を次第に魅力的にしつつある。[1] この結果、データアクセスの向上に向けた NSI への要求はイノベーティブなラボソリューションによって次第に達成されつつある。このため、(データ操作も統計モデルの選択もほぼ制限されない)「研究環境」の提供は、柔軟なリモートジョブ投入システムと共に、着実に拡大しつつある。

---

[1]「シンクライアント」システムは、リモートコンピュータ上で処理が実行される場所である。つまり、このクライアントコンピューターはデータと完全に相互作用するように見えるが、単に指示を送るだけで、操作の成果物を確認する。「ファットクライアント」システムは、クライアントマシン上で処理が実行される場所である。「リモートジョブ投入」はプログラムが実行に向けてリモートコンピュータに送られると、プログラムの結果が戻される。つまり、データとの直接相互作用は発生しない。
シンクライアントシステムの主な利点は、データ管理の簡易化、セキュリティの向上及びロケーションとアクセスの分断である。NSI にとって 2 つ目と 3 つ目は特に魅力的であるが、シンクライアントプロセシングは古くから大型コンピュータシステムに対するデフォルトオペレーティングモードであるため、マイクロソフトウィンドウズ ™ システムに対するシンクライアントソリューションがサプライヤ及びユーザーに実行可能になったのはわずか 10 年前である。
従って、ラボ環境の提供に著しい成長が見られたのはここ数年である。

研究環境の利用に見られるこの拡大は、統計的開示抑制(SDC)に対する問題になっている。[2] SDC の主眼はこれまで、集計データの秘匿性又はここ数年では、研究用途向けの（たいてい、「一般公開型」ファイルと呼ばれる）非開示性データセットの生成に置かれるのが一般的であった。集計表及び一般公開型ファイルに対する SDC については大規模な文献がある。

しかし、研究環境におけるミクロデータの開示に関する SDC には異なるアプローチが必要である。主な問題は成果物の予測可能性である。この問題は、例えば、一般公開型ファイルの安全評価に用いられるシナリオベースモデリングを有効的に用いにくくする。

通常の SDC 研究と異なり、これに関する文献はほとんどない。 開示制限に関する *Journal of Official Statistics* の特集版(Feinberg and Willenborg, 1998)では、13 の論文のいずれにおいても研究環境が考察されなかった。最近の国際会議は、安全な環境の物理的側面又は配布用の安全なファイルの作成に主眼を置いている（例えば、UN （2004、2006、2008)、Domingo-Ferrer and Torra(2004)、Domingo-Ferrer and Franconi （2006)を参照)。欧州統計システムプログラムには、2008 年から ESSNet プロジェクトに初めて成果物開示の研究が組み込まれている。いずれも分析的成果物の公開を考察している Reznek (2004)、Corscadden et al (2006) Steel and Reznek (2006)及び Ritchie (2006a, 2006b)を別とすると、研究者がデータを制限なく扱える場合に生じる一般的問題を何らかの形で扱った分析はほとんどないようである。

これは NSI のリサーチセンターの設置を反映している面もある。このセンターはたいてい、NSI の一部で、比較的独立して活動し、関連する研究の実務経験を有する専門家で構成される。SDC の知識はリサーチセンターのスタッフの中で具現化されている。

しかしながら、今は、研究環境における有効な SDC の構成要素について考察する必要がある。これには 5 つの促進要因がある。まず、国際的な（特に EU 域内の）データ共有の増大に伴って、国際的なデータ共有の発生可能性を低減する SDC 基準に対する合意の欠如が懸念されている。第 2 に、研究作業の実施量の増大によって、研究に対する注目度は高まったが、考察が不十分なために、集計成果物及び匿名化用に考案された SDC ルールを採用して研究成果物に適用しようとする姿勢が見られるようになった。これは効果がなく、不適切な上、不必要に官僚的になる可能性があり、最悪の場合は、不適切なルールの盲目的適用によって、研究にとって致命的になる場合もある。第 3 に、研究環境で実施される分析の範囲は、SDC ルールの考案に用いられた従来のモデルをはるかに超えている。第 4 に、潜在的開示性データの現場外施設での公表を求める要求の高まりに伴って、例えば、大学

[2] 開示検出及び抑制は異なる概念であるが、本書では便宜上、SDC を用いて両方を網羅し、必要に応じて区別する。

の安全なリサーチセンターに管理職が異動される際に、NSI で安全に用いられるデータがその機密性を保持するように SDC 手続きを透明にする必要が生じている。最後に、集計及び匿名化に対する SDC は定期的に試験及び策定されているが、研究成果物に対するルールについての論議は行われていない、つまり、リサーチセンターの責任者が策定した内部ルールの独立した精査はほとんど行われていない上、「最良の事例」もあまり共有されていない。

本論文は、上記の問題、特に最後の問題に取り組むことを目的とする。本論文で主張するように、研究環境における SDC は、禁止的なルールベース手法と根本的に異なるアプローチー「原則事例」アプローチを要求する。これは、正確なルールを規定しようとする方針の限界を明らかに認めており、ルールが少しでも柔軟に連結されるような原則の理解に主眼を置く。これは、研究者の訓練及び自動システムの利用の両方に影響を及ぼす。

次のセクションでは研究環境について論評する。これに続いて、開示抑制に厳格なルールを適用する問題に目を向け、研究環境の性質が、ルールの規定を基本的に困難にしていることについて論じる。次のセクションでは、複雑な適用以外は極めて単純なルールを基本とする 1 つのアプローチを提言する。このアプローチには、研究者及び NSI の両方の教育が必要であり、成果物の承認規準は必然的に複雑になる。本書の最後では、英国の事例及び情報共有に関する論評を提示する。

2. 研究環境の特徴

SDC の多くは、集計表の安全化又はミクロデータの有効な匿名化に関係する。たいていは、有限個の集計表の集まり又は「侵入者」シナリオが特定可能で、その結果生じる「安全な」データはこれらの目標に照らして測定可能であるという理由で、これは実際的な目的である。

研究環境が他と顕著に違う点は成果物の予測不能性である。研究者は集計表を作成するが、この集計表は同じデータから作成される集計表と全く異なるものになる可能性がある。データは望ましくない方法で伸長されたり、歪曲されたり、併合されたりすることがある。研究者は、欠測している又は範囲外の変数に極めて個人的な処理を適用するかもしれないし、データの望ましくないサブサンプルを使用することかもしれない。データは様々な情報源から組み合わされる可能性もある。

線形集約から逸脱すると、研究成果物の範囲は、線形及び非線形推計、シミュレーション、確率論的モデリング、ベイズ分析、因子分析、動的モデリング、変遷データ等に著しく拡

大する。ミクロデータアクセスを提供する理由は、結局のところ、研究者が単純な線形集約からは不可能な又は、自動プロセスでは容易に定義付けが行えない一連の分析を探求できるようにすることである。

研究者側に基本的な統計能力が期待される。NSI はいずれも、研究者の背景及び資格に対する一定レベルの確認を行っている。これには、データに実施される作業が科学的に有効であるよう保証し、さらに、NSI に対する要求を低減する目的もある。NSI は研究者のデータ関連の疑問を支援するが、通常は、統計に関するメンタリング（指導的助言）は行わないようにしている。

上記をまとめ、本書では、研究環境は、専門家である研究者が開示データのアクセスをほぼ制限されずに成果物の予測不能な集合を作成できる環境であり、使用されるべき事前モデリング手法又はデータ変換の完全な規定が望ましくなく実用的でない環境と定義付ける。

ここでは、本論文の解釈上、ラボの研究者は意図的にデータを不正利用しないと信頼でき且つ、ラボの技術的セキュリティは容認可能であると想定する。上記は重要であるが個別の懸念事項である。考察については、例えば、 Desai (2004) or Ritchie (2006b)を参照。

3. 研究環境におけるルールベース手法に伴う問題点

SDC は全て、一定レベルの開示保護の保証を意図するルールに基づく。このルールは、明確で独立した且つ検証可能な基準一式になるように考案され、非開示性集計表又は匿名化されたデータセットの作成に不可欠である。

本書の目的は、ルールそのものが不適切であると主張することではない。ここで主張したいのは、研究環境の性質は、利用可能な変換の範囲を全面的に考慮しないルールに概ね基づいて SDC 戦略を定義付けようとする努力は失敗に向かって突き進むようなものだということである。これは、成果物の予測不能性が必然的に、一般的なルールを複雑な特殊ケースに変えてしまうためである。

本書では、単純な一次開示（つまり、他のセルと無関係なセルの開示リスク）を検討する方法でこれを実証する。*典型的な閾値ルールは*以下のようになり得る。

公開用の集計表には、表示されるセルの基礎になる最低 5 つの測定値の度数を組み込まなければならない。

これはデータ集計、例えば、産業別総取引高に適用される種類のルールである。セルの制限は、NSI が合意の可能性とみなすものに基づく可能性がある。この場合には、5 つの制限には、第 4 の当事者の暗示された価値の決定に結託するのは多くても 3 つの回答であるという NSI の考えが暗に含まれている。この仮定（及び二次的開示の可能性及び占有的数値の当面の無視）に基づくと、このルールはミクロデータの機密保持を保証する。

データそのものが開示的であり、データドナーで簡単に識別できる場合にはこれは適切になり得るが、この状態を維持できない場合は、これは過度に制限的である。

まず、データの開示性を考えてみよう。変換はこのルールを無効にする可能性がある。例えば、従業員当たりの生産性が表示される場合は、少数であることは懸念材料になり得ない。つまり、比率では、個々の調査回答が選別できないようにすることはできない。

閾値ルールはこの時点で以下のように改正できる。

データが変換されていない限り…

ただし、この情報は依然として有用である可能性がある。従業員当たりの取引高の*伸び*が表示されるとしよう。最初の調査情報はこのような複雑な変数からは決定できないが、企業の生産性がどう変化するかについての情報は企業秘密になる可能性がある。NSI はこの機密保持不履行を十分に検討する可能性があるため、ルールは再度以下のように改正する必要がある。

…そして、その結果生じる情報は機密保持不履行にならない。

ただし、この情報は既に公知である可能性もある。従業員当たりの生産性の伸びは従業員当たりの総収益の伸びから概算できる。会社が法人化される場合は、この情報は公開された会社勘定から入手できる見込みがある。研究情報から収集される情報は公開文書から入手できるものと定性的に同一であるため、

公的情報源から入手できない情報を提供することによって…

機密保持基準が違反されることはない。

ただし、情報が容易に入手できない場合は、NSI は企業秘密情報を提供しない義務を課されることがある。従って、以下が追加される。

…容易に…

さらに、類似する情報を公的情報源から容易に入手できる場合でも、NSI は（例えば、不確かな公開情報を裏付け得る）調査回答から推論が引き出せることを、機密保持規定の不履行になるとやはり感じるかもしれない。法的制限も存在し得る －秘密裡に提供される情報は、公知によって批准される場合でも、NSI によって公開されない場合がある。

次に識別の問題を考えると、これは少なくとも単純なルールには適しているようである。これはある程度当てはまるが、やはり隠れた問題が存在する。まず、直接的識別子（名前、住所、業種、立地）の範囲はデータセットによって異なる。当該識別子の文脈も重要である。
例えば、

- 英国では、郵便番号は中規模企業を特定するには十分であるが、個人又は世帯を特定できるのは極めて例外的状況のみである。
- 5 桁の標準産業分類(SIC)では、1 つの業種に数百社の企業が組み込まれることがあるが、国営事業等の別の業種では 1 社だけしか含まれない。
- 保健統計においては、その稀有性により強力な識別子になる特定の事象(珍しい癌等)もあれば、他のデータセットにおける遍在性により強力な識別子になる事象もある(出生等)。
- 地理情報そのものにはほとんど開示性はないが、他の変数と組み合わさると、重要な識別子の 1 つになる（例えば、Elliot (2004)を参照)。

さらに困難なことに、基礎になるデータが相対識別子レベルで収集されないこともある。英国の New Earnings Survey のデータの例を考えてみよう。これは従業員の 1%標本であるが、企業から収集される。集計表の各セルには 5 個より多い測定値が記載される可能性もあるが、これは従業員の人数だけを計上する。1 個のセル内の従業員が全て、1 つの企業に由来する可能性がある(例えば、集計表が国有産業の専門的職業を示す場合)。NSI の開示ルールが企業利益の識別を基準とする場合は、データの度数が高いセルは、NSI ルールを侵害する可能性がやはりある。

（企業レベルと異なり）プラントレベルデータ又は個人の特徴が家族単位を通じて識別される可能性がある個人データについても、同様の事例が引き出される可能性がある。セルの値は無効になることがある。重要なのは識別の単位の度数である。

識別が行われない状態で、データの公開が開示性になることはあり得ないが、複数因子を組み合わせると識別可能になることがある。これは、文脈によって極めて異なる。

つまり、単純なルールはこの時点で以下のようになる。

> *当該データが変換されない限り、公開用の集計表には、表示されるセルの基礎になる関連する開示抑制客体の最低 **5** つの測定値の度数を組み込まなければならず、そして、その結果生じる情報は、一般情報源から容易に入手できない情報の提供によって機密保持不履行にならない。*

これは、はるかに複雑なもので、上記の問題の一部に対応している。残念ながら、開示検出モデルと同様に、これも操作が難しい。「容易に入手できない」等の文言は、ルールの不可欠な部分であるが、一般的なケースでは特定することが不可能である。表現はあらゆるケースを網羅するように意図的に曖昧になるが、結果的にはどのケースも明示的に網羅されない。

この定義にはトートロジーも組み込まれている。つまり、データはそれが変換された時点で非開示性になり、非開示性である場合は、データは変換されている。「これは変換されたデータである」と記述する独立した一節はない。

このルールでは、識別をセル最小値で暗黙的にしか言及していない。これは、有意の一般的ルールでは特定が難しいためである。

最後に、このルールは関連する開示抑制客体も集計表に含まれない可能性があることを明示的に認識している。

つまり、この「ルール」は解釈される必要があるガイドラインになったのである。

線形集約の開示抑制は、差分抽出による開示可能性により、当然ながら極めて難しい。本論文の目的は、架空の論議を設定することではない。閾値ルールは識別の中核的存在であるが、本論文で主張するように、閾値ルールはそれ自体を最終目的とみなしてはならず、SDC の原則のカプセル化とみなすべきであり、従って、文脈の中で評価される必要がある。

4. ルールの派生：動物園の研究

以前より複雑なルールの策定に伴う問題には、ルールが決定される様式が含まれる。前述した単純な閾値ルール等の基本的ルールは、第一原則から派生し得るため、より複雑な派生物には、一連の「こうしたらどうなるか」のシナリオが要求された。

このアプローチは通常、一般公開型データセットの開示性を試験する際に用いられる。安全が確認されたデータセットは、複数の「攻撃」シナリオを適用する方法で試験すればよい。潜在的に開示リスクがあるセル又は測定値が特定される場合は、抑制メカニズムを調整して適用し直せばよい。分析の結果が安全な集計を決定するルールにつながることもある。

シナリオを用いる鍵は、検討中のデータが「非公開の」システムに近いものを形成することである。集計表のデータの場合は、成果物の形態は既知である。一般公開型データセットの場合は、データから派生する成果物の最終形態はわからないが、各測定値の不確かさのレベルは評価することができる。再識別の確率の推定値を引き出し（例えば、Elliott and Manning 2004 を参照）、適切な再符号化又はルールを定義付ければよい。シナリオ試験では、全ての可能性を明確に網羅することは不可能なため、起こり得る攻撃の有限集合を定義付ければよい。

開示データを伴う研究環境は「開放的」システムである。ルールの導出を担当する個人は、結果の安全だけを試験するのではなく、その結果がどのような形態を取るかも予測しなければならない。これはさらに困難な提議である。

開示抑制は、動物に動物を安全且つ活動的に保つための動物に対する囲いを提供するようなものと想像することができる。集計表のデータでも一般公開型データセットでも、開示検出の目的は、柵や壁等の強度を精査し、収容される動物が良好な状態にあるようにすることである。研究環境に伴う問題は、SDC 職員が、居住者が鳥であるか魚であるか昆虫であるかを事前にわからない状態でこれを行わなければならないことである…　全てのルールが非常事態ルールになるのはこのためである。研究環境における SDC は動物園の設計であり、檻を評価することではない。

5. 研究環境における SDC に向けた原則・事例アプローチ

上記の考察は必然的に、SDC の展開を過度に単純化することになるが、その含意は明確である。つまり、研究環境の偶然性を全て網羅するような困難なルールを導出する努力はほぼ失敗する運命にある。

ただし、はじめにの中で論じたように、研究環境で適用可能な何らかの「基準」は必要である。この丸をどうやって四角にすればいいのか。

この解は SDC と異なるアプローチにある。これは、4 つの重要な問題、つまり、原則に対する理解、少数で単純だが柔軟性のあるルール、可能な場合は常にデータではなく関数形の明示的なモデリング及び、研究者の教育が基本になる。最初の 2 つはある程度 SDC に既に組み込まれているが、研究環境が必要とするアプローチの違いが重要になるのは残りの 2 つである。

5.1 原則の理解

研究環境における SDC は自動的に実行できるようなものではない。この SDC には、チェックされる成果物、潜在的開示リスク及び容認可能な開示リスクのレベルの把握が不可欠である。このため、重要な問題は SDC のねらい及び目標に対する合意があることである。これは、合意ルールと同じではない。原則は NSI 全体で共通する可能性もあるが、ルールを実施する方法は分野によって異なることがある。

例えば、英国の国家統計局(ONS)の実務指針(ONS 2002a) は、「機密性保護」に向けた指針書を定め、関連する*データアクセス・機密保持に関する議定書*の中で、この指針を実際にどのように解釈すればよいかを次のように提言している。

> *…統計的開示抑制方法は、データ又は統計データの設計又は両方の組合せを変える可能性がある。第三者が他の情報源又はこれまでに公開された国家統計局の成果物のいずれかから入手できそうな情報を踏まえて、以下の基準に照らして機密性の保証が維持できる場合は、この方法は十分であると判断される。*
>
> > 侵入者が統計データの客体を他から識別する又は、まだ公知になっていない客体に関する情報を公表するのに、過度に膨大な時間、作業及び専門知識を必要とする。
>
> ONS (2004) pp7-4

この意図は、結果が公開されてはならない理由について概ね包括的な見方を示すことである。この基準は、SDC の絶対的な基準を規定せず、「起こり得る」及び「過度に膨大な」という表現を使って判断が必要であることを示唆している点に留意すること。特定のルール又はパラメータはなく、優先される抑制尺度に関するものは何もないが、これは、SDC 活動全ての尺度にならなければならない最終的な基準である。

5.2　柔軟なルール

ルールの適用における柔軟性は、研究環境における有効なSDCにとって極めて重要である。SDC審査者の自由裁量で「免責される」可能性がある厳格なルールを設ける方法や、本質的に柔軟性を有するルールを設ける方法などがある。重要な点は、成果物の不確かさをルールに組み込むことである。例えば、閾値ルール(セクション3)を以下と差し替えることが可能である。

> *集計表セルは通常、客体の度数が5以上である場合は非機密性であるとみなす。度数がこれを下回る場合は、機密保持の原則(…を参照)が破られないことが実証できることを条件に公開が許される。データの多様性が不十分である場合又はデータが少数の統計単位で特定できる場合には、これより高い度数が要求される可能性がある。*

これは、「可能性がある」、「不十分である」、「通常、」「少数の」を使って、経験則を展開しており、「グレー領域」がどこで発生するかを明示的に説明している。これは、下限値を主張する責任を研究者に課しているのに対し、上限値を主張する責任は黙示的にSDCの管理者に課している。

最も重要な点は、一部の機密保持原則に対する直接的言及と、セクション3で展開された機構的な閾値ルールと対照をなすことである。認識されるように、このルールは独立して存在することはできず、どの成果物も適切な文脈でとらえられることが必要である。

従って、独立した基準があることは望ましいと思われるが、ルールはそれだけでは成立しなくてもよい。この場合の問題は、この「ファジーな」ルールの遵守を確認するために精査の必要があるデータが全て公開されるのをどのように防ぐかである。これは、研究者及びSDCの従事者の教育が重要になる問題である。

5.3　モデルベースの理由付け

研究環境では、成果物の大部分は（データの非線形な集約と定義付けられる）何らかの「分析的成果物」の形態で出現する。問題になるのは、研究者がデータを無限に操作できるようになる可能性である。上記で述べたように、このアプローチはわからない動物の小屋を建てる方法を考える努力とほぼ同じである。無限個のルールを持つ集合を策定する必要はあるのだろうか。

本書のアプローチの目的は、比較的少数の「動物種」が実際に存在することを認識することである。動物をクラス（飛ぶもの、土を掘るもの、登るもの、人肉を食べるもの）に分類できる場合は、適切な手続きの策定はかなり単純化される。実際の動物の集合は無限に存在するが、1 匹の動物の特徴は概ね把握されている。

研究環境にとって不可欠な SDC の重要な変更はこれ、つまり、成果物そのものではなく、成果物の作成プロセスに目を向けることである。本書では、データの開示性ではなく、データが利用される方法に着目する。本書では、これを「モデルベース」の理由付けと呼び、「データベース」の理由付けと区別する。

単純な線形回帰分析を考えてみよう。従来のアプローチの目的は、外れ値、影響点、カテゴリカル変数、適合度、一般公開型データの用途等に関するルールを引き出すことである。これは、急速に極めて複雑化している。例えば、影響点を回帰分析に組み込むと、品質の悪い統計データになる可能性があるが、回帰分析の結果は必ずしも開示性になるとは限らない。

これに対し、関数形の線形回帰分析は、回帰分析結果がほぼ常に非開示性であること、問題ケースが少数識別集合であること、問題の多くは平均値及び合計値の同時公表に起因すること、データの開示性の簡単なチェック方法が存在すること及び、それゆえに、使用されるデータに関係なく回帰分析結果を非開示性にする単純な補正措置が存在することが明らかになっている（詳細については、Ritchie （2006a）を参照）。この分析は直接的であり、結論は明確である。「回帰分析」型の具体的な成果物の例は比較的簡単にチェックできる。

この場合の問題は見えるほどには悪くない。発生し得る成果物群を全てシャッフルして、特性を調査できる比較的少数のクラスにすることは可能である。これは、開示抑制が簡単になるということではない。例えば、パーセンタイルは集計表として処理できるが、カテゴリーの並び替えは異なる識別問題を発生させる。これはむしろ、SDC に用いられるルールは小規模で管理可能であり且つわかりやすい状態に維持できるという意味である。

モデルベースの理由付けは、極めて「開示リスクが高い」成果物に直接注意を払うのに役立ち得る。モデル分析は性質上、一部の成果物が内在的に開示問題に寄与することを実証する。この場合には、特殊な例を詳しく見直すことが必要である。「開示リスクが相対的に低い」成果物の場合は、限定されたチェックリストでも十分に機密保持の達成を証明できる。動物園の例えに戻ると、ライオンも羊もそれぞれに異なる。情報源が乏しい動物園の

飼育員は、適切な方法で観察できないことによる潜在的危険性がはるかに大きいため、ライオンを個々に理解することに集中する。

最後に、ルールは関数形が独立して成立できることを基本とする。すなわち、

> *少なくとも 1 つの係数が有効に秘匿される場合は、線形回帰分析は非開示性になる。つまり、線形回帰分析は公表された情報から合理的に決定することはできない (Ritchie 2006b)。*

この文脈では、集計表の分析をそのあるべき場所に据えることができる。線形集約は差分抽出による開示可能性があるため、本質的に安全ではない。集計表の作成方法は承認できないため、成果物が以下の適切な基準を達成しなければならない。

> *集計表も他の線形集約も、機密保持指針を達成していることが証明できない限り、公開されてはならない。*

これを見てはっきりわかるように、集計表そのものは容認できず、その公開に対する根拠を示す必要がある。これには、機密保持指針の直接的適用などがある。これは、成果物の公開を阻むものではなく、立証責任を単に移行するだけである点に留意すること。一定の規準を満たすことを条件に、集計表が公開されることを記述するルールはもうない。現在は、規準が満たされていることを実証できない限り、集計表は公開されない。これは微妙な違いだが、強調される重要な変化である。

5.4　教育

適用すべき SDC のルールが判断の要素を組み込む場合は、研究者は開示検出及び抑制に関する十分な情報を得ることが極めて重要である。この教育には、原則、ルール及びそれがどのように原則から派生したか及びルールがどのように適用及び解釈されているかを組み込む必要がある。解釈に関する指針がない場合は、整合性の達成は困難になる可能性があり、研究者は明らかに恣意的な決定により当惑又は混乱する可能性がある。これに対し、教育された研究者は容認可能な成果物を予測する能力の点で上回ることになり、成果物が承認されない理由を理解するはずであり、大量の容認不能な成果物で成果物の検査者に負担を与えることはないはずである。

これは明らかに、SDC をこれまでよりはるかに、研究者と SDC チーム間の協力的な作業にする。これは意図されたものである。つまり、両方の当事者に、効率的な非開示性デー

タの公開という、同じ目標を共有させるという意図がある。研究者は結果が速やかに許可されることを希望する。SDC チームは結果が正確な方法で許可されることを希望する。上記の目標は、成果物が遵守しなければならない原則及び基準を理解しこれに同意していれば矛盾することはない。

利点は他にもある。まず最初に、新しい状況(例えば、SDC チームがルールも事例も持っていない新しい関数形)が発生した場合に、SDC チームと研究者が協力して適切な指針を策定できる。第 2 に、研究者を枠組みの策定に参加させることにより、この枠組みは SDC 方法の妥当性に対する即時的フィードバックになる。最後に、研究環境は、経験豊富な卓越した分析者に直接アクセスの機会を与える。継続的なピアレビュー源を無視するのは不名誉であるようである。

研究者を SDC の枠組みに組み込むことには危険性もある。最も重要な点は、SDC チームがその立場を防御する自信と能力を備える必要があることである。準備不足の SDC チームが次第に低い基準を容認する脅威に晒されるにつれて、抑制緩和を強化する方向に流されてしまうこともある。解決策の 1 つは、開示リスク及び原則の解釈の問題において、NSI が最終決定権を持つように、最終決定責任が SDC チームにあることを明確にしておくことである。2 つ目は、NSI のルールが定期的且つ個別に見直されるようにすることである。

これは、統計データの相違を討議すればよいということではなく、ルールに異議を唱えることを望む研究者は、より良い方法の存在を証明することは研究者の責任であることを認識する必要があるという意味である。これは、NSI に変更の提言を無視することを認めるということでもない。SDC チームの技術的知見が不十分な場合は、理解の格差を埋める合理的な努力を施す必要があり、これを行わない場合は、両当事者間の信頼感は失われる。

ゆえに、教育の重要な役割は NSI と研究者間に良好な関係を築くようにすることである。Desai (2004) は次にように述べている。

> 「セキュリティの最良の形態は、ユーザーとの良好な関係である。ユーザーが助けを求める立場にあるというよりも利益を享受していると感じている場合は、ユーザーは少しでも自分の行動に責任を持つようになる可能性がある。」
> Desai (2004, p5)

これまで主張されてきたように、研究者の SDC への関与には危険性もある。つまり、SDC に関与させることによって、システムを破壊する方法について有用な情報が与えてしまうということである。これは無益な主張だと考える。まず、悪意のある研究者は、調査結果

が開示抑制を通過するようにするよりもラボからデータを排除する方が簡単な方法であることに気付くはずである。[3]

第 2 に、全ての成果物が SDC の対象になる場合でも、悪意のある研究者であれば、結果がどのようなルールの下でも容認可能であるように見える程度まで成果物を編集できてしまう。困難に陥る SDC チームは、意図的に不正操作された成果物に気付くだろうか。

第 3 に、そしてこれが最も重要なのだが、本書での考察は研究者が開示性のある結果の検出に関与すること及び、実施可能な要素において研究者を教育することについてである。NSI の多くは、検出及び抑制方法について何らかの情報を提供しており、特定の成果物に適用される特殊な抑制の詳細を討議しないようにしている。本書でも同じことが当てはまる。研究者に検出を指導する目的は、優れた成果物にまで抑制が必要になることを回避することでもある。

5.5　実用性

このアプロ―チの実施は成果物の量及び NSI の技能に関連する 3 つの典型的な懸念を引き起こす。

まず第 1 に、この戦略は、線形集約についても、自動的な SDC 方法の範囲が小さくなることを含意する。これは、実施される調査の量に比例して SDC の作業量が増加することを暗に意味する。有効な SDC 手続きを設けるねらいは研究を促進することであるため、これは NSI にとって非生産的になる可能性があり得る。

第 2 に、手作業による SDC チェックには検査者側に一定レベルの統計的専門知識が求められる。統計的基礎知識を有する個人の場合でも、これには養成にかなりの時間を要する。公開可能な成果物に対する高度な問合せ（ハーフィンダール指数は安全か。ジニ係数は安全か）の処理に備えた、統計的知識の養成も同様に必要になる。十分に養成された統計的知識を有する個人でも、他の利用者の研究に対して SDC を特に興味深い又は意欲を沸かせると思う見込みはあまりない。複数の職種を充当し資金を提供するのは困難な場合がある。

第 3 に、本書で示した SDC モデルの成功は、NSI と研究者間の関係に依存するところが大きい。新しい方法の開発、未解決問題をめぐる意見の対立の回避、提出された成果物の容認可能性は全て、良好な職場関係によって助長される。この依存性に起因して、研究者の

[3] この主張はリモートジョブ投入には当てはまらないことがある。

作業経緯又は NSI の活動基盤である制限に対する研究者の知識不足に NSI が気付かず、行き詰まる可能性もある。両方の当事者がこの関係に力を注ぐことが必要である。

6. 事例：英国における企業調査データの研究

最後に、英国の国家統計局(ONS)におけるバーチャル・ミクロデータ・ラボラトリー（VML)の一例を挙げる。このシンクライアントラボ施設は、機密データ、主に企業のミクロデータアクセスを ONS、政府及び大学の研究者に提供する。研究は主に、分析的経済学及び計量経済学である。成果物は全て開示チェックに向けて VML チームに送られる。

VML の重要な特徴の 1 つは、それが現役の研究者によって設計及び管理されており、それゆえに、チームによる SDC 指針の展開は標準的な成果物に対する実務経験によって伝えられることである。これによって、VML は、表面上はありえない程に厳格な体制を運営しているにもかかわらず、活発な研究基盤を確保した。この特徴は、VML チームと研究者間の関係を育て、チームの主張における権限を強化する上でも役立っている。

研究者は全て、短期的な訓練セッションを受けている。この多くは SDC によって採用されており、原則、占有ルールと閾値ルール及び、企業データの文脈での解釈が組み込まれている。参加型の事例は、研究者及び VML スタッフの両方にとって主要な教育手法であり、多くは観察された問題から派生した。VML の開示抑制原則は、ONS と同じであるが、研究者にとって少しでも適切になるような様式で言い直されている。

研究者は、成果物が「安全な」カテゴリーと「安全でない」カテゴリーに分類されることを示され、「安全な」成果物が拒否されるかどうか又は、「安全でない」成果物が承認されるかどうかにどの因子が影響を及ぼすかについてのガイダンスも提示されている。

この訓練の結果、BDL の研究者は、自身の成果物の開示性評価に他よりも有能である。セクション 5.2 で述べたように、柔軟なルールの利用は、全ての成果物が長期的精査を受ける結果につながるように見える。実際には、これは当てはまらない。これは、SDC の枠組みを速やかに認識する研究者は、安全な成果物のメッセージを学習し、早い段階で取扱い許可を得る結果を生産するためである。ただし、これ自体も問題を引き起こしている。

まず、このアプローチの必須部分は、SDC に用いられるパラメータ値（例えば、占有度／不確かさの限度）について研究者に情報を提供することである。この情報が ONS の成果物を攻撃する他の文脈の中で利用される可能性について懸念が示された。VML は既に、通常の ONS 成果物よりも高い閾値限度を使って、差分抽出による開示から防護している。これ

は、全てのパラメータ値に拡充された。VML の訓練では現在、VML 固有の値についてのみ、研究者と討議されている。研究者は、VML の SDC ルールは ONS のルールよりも厳格であり、成果物に対する要求は VML ルールのみに照らして判断されると教えられている。

第 2 に、成果物は結果のチェックに必要なデータ（基本になる度数なしの集計表等）を伴わずに提示されることがある。このような成果物は、情報の追加要請と共に研究者に戻され、研究者は経時的に、必要な情報の提供を学習する。

第 3 に､成果物の量は次第に増えてきている｡BDL に提示される成果物ファイルの中には、規模が大きすぎてファイルをチェックする時間が膨大になるものもある。安全な結果を作成する研究者の能力を BDL がいかに信頼しているとしても、BDL には、開示性成果物が ONS に留まるようにする法的責任を負っており、これによって、開示性ではなく量を根拠に成果物を拒絶したことがある。成果物の数は(差分抽出による開示可能性に起因して)結果の公開を拒絶する妥当な根拠であるが、これはあまり満足できる結果ではないため、BDL はこれまで、その訓練プログラムを調整して、最小限の成果物集合の強調を強化せざるを得なかった。

この協力的な SDC アプローチは総じて、スタッフへの少なからぬ投資及び VML のメッセージを周知させるための相応の努力を必要とした。しかし、長期的に見ると、このアプローチは、データセキュリティが高度で回答時間が容認可能な、低コストで、スケーラブル且つ透明な SDC システムを調達した。

新任研究者にはトライアンドエラー期間が概ね存在する。これはたいてい、フラストレーションを引き起こす時間であり、慎重に管理することが必要である。それでも、教育された研究グループを備えた結果､目標とする研究結果の公開時間が 2003 年の 2 週間から 2004 年には 2 日間に大幅に短縮されるという全体的影響が見られた。実際のところ、2007 年には結果は 90%のケースで 1 営業日に短縮された。拒絶率は 1 ヵ月当たりおよそ 1 論文であり、拒絶の主な理由は単に成果物の量である。[4]

7. 結論

状況の範囲を拡大した SDC 基準を策定する必要性は明らかである。研究環境のケースは、成果物が予測不能であるため特に困難である。このため、極めて限られた数の場合を除き自動的ツールを用いることを理由に、多くの状況で支持されない 1 つの絶対的基準に依存

[4] 研究者 1 人がクリアするために要請するラインログファイルは 300,000 個にも及んだ。しばらく検討された後、これは却下された。

せざるを得ない。しかし、成果物の構造に集中する方法で、結果を秘匿性の異なる複数のクラス、つまり、集計表は本質的に開示リスクがあり個別の評価が必要である、パネルデータ推計値は本質的に安全である等に分類することができる。重要な点は、研究者はこの規準を認識した上で、適用する能力を必要とするということである。

絶対的ルールの規定は困難になる可能性があるのに対し、原則は決定及び合意するのがこれよりはるかに容易である。原則は、特殊なケースを評価できる対照になる包括的な枠組みを形成する。原則は、組織間及び組織内の整合性を提供するものにもなり得る。内在的柔軟性は、ルールベースのシステムに比べて原則ベースのシステムを不透明にするが、手続きを試験する対照になる共通の判断基準は維持される。

原則を定義付け、発生し得る成果物の数学的構造をモデル化するこのアプローチには、知識の豊富な SDC チーム －データだけでなく関数形及び新しい状況の評価方法にも通じたチームが暗に含まれる。チームは、SDC の範囲が経時的に拡大することを明確に認識し、新しい動向を組み込むためのシステムを設置する必要がある。このため、事例を介した学習が、SDC スタッフの訓練プログラムの重要な部分になる。SDC スタッフは統計学全般に精通している必要があり、関係する研究者が用いる共通性の高い関数形に特に卓越している必要がある。

最後に、研究者は SDC プロセスにも関与していることが極めて重要である。まず、研究者及び SDC チームは、成果物が速やかに且つ安全に且つ容易に取扱い許可を得ることに関心を抱いており、これは、全ての当事者が枠組み及びルールに精通している時に最良の形で達成される。研究者は SDC に費やされた時間に対するみかえりを確認することができる。第 2 に、原則ベースシステムの柔軟性が高まるほど、研究者の関与によってシステムの透明性もそれだけ高まり、従って、混乱及び意見の不一致が発生する範囲が縮小される。第 3 に、研究者は、新しいルール及び手続きの策定に協力することに意欲的であり、適切なソリューションを提示せずに新規の成果物を要求する可能性は低い。第 4 に、SDC の訓練を利用すれば､研究者と SDC スタッフ間の信頼というコミュニティを構築することができる。

要するに、研究環境における SDC は他の状況の場合ほど手際よく抑制できないため、共通の基準に照らして比較できる透明でアクセシブルな手続きを策定する範囲が膨大になる。

参考文献

Corscadden, L, Enright, J., Khoo, J., Krsinich, F., McDonald, S., and Zeng, I. (2006). Disclosure assessment of analytical outputs, mimeo, Statistics New Zealand, Wellington

Desai, T. (2004) “Providing remote access to data: the academic perspective” in UN(2004)

Domingo-Ferrer, J. and Torra, V. (2004) Privacy in Statistical Databases: CASC Project International Workshop Proceedings, Springer-Verlag, Berlin

Domingo-Ferrer, J. and Franconi, L. (2006). Privacy in Statistical Databases: CENEX-SDC Project International Conference Proceedings, Springer-Verlag, Berlin

Elliot, M.E. and Manning, A. (2004). The methodology used for the 2001 SARS Special Uniques Analysis. Mimeo. University of Manchester.

Feinberg S.E., and Willenborg, L.C.R.J. (1998). Introduction to the Special Issue: Disclosure Limitation Methods for Protecting the Confidentiality of Statistical Data. Journal of Official Statistics, 14:4, 337-345

Reznek, A (2004). Disclosure risks in cross-section regression models. Mimeo, Center forEconomic Studies, US Bureau of the Census, Washington

Ritchie, F.J. (2006a). Disclosure control for regression outputs. Mimeo, Office for National Statistics, Newport

Ritchie, F.J. (2006b). Access to business data: dealing with the irreducible risks in UN(2006)

Steel, P and Reznek, A. (2006) Issues in designing a confidentiality-preserving model server, in UN (2006)

UN (2004) Monographs in Official Statistics: Work session on Statistical Data Confidentiality Luxembourg 2003, Eurostat, Luxembourg

UN (2006) Monographs in Official Statistics: Work session on Statistical Data Confidentiality Geneva 2005, Eurostat, Luxembourg

UN (2008) Monographs in Official Statistics: Work session on Statistical Data Confidentiality Manchester 2007, Eurostat, Luxembourg