S.ヴェイル(Steven Vale : 国連欧州経済委員会)編著『政府統計のための行政的なデータソースと第2次的なデータソースの利用 : 原則と実務にかんするハンドブック』第2部(完)

浜砂, 敬郎
九州大学 : 名誉教授

西村, 善博
大分大学経済学部 : 教授

https://doi.org/10.15017/1462164

出版情報：經濟學研究. 81 (1), pp.45-82, 2014-06-30. 九州大学経済学会
バージョン：
権利関係：

（翻訳資料）

Ｓ．ヴェイル（Steven Vale：国連欧州経済委員会）編著
『政府統計のための行政的なデータソースと第２次的なデータソースの利用
―原則と実務にかんするハンドブック―』第２部（完）

浜砂　敬郎　　西村　善博　分担・共訳

内容目次

訳者序文
序文
目次
注釈
第１章　行政的なデータソースと第２次的なデータソースとは何か？
第２章　行政的なデータソースを利用する利点
第３章　行政的なデータソースを利用するための体制
第４章　一般的な問題と解決策（以上、第80巻５・６合併号）
第５章　精度と行政的なデータ（以下、本号）
第６章　データの連結と照合
第７章　統計登録簿における行政的なデータの利用
第８章　統計調査を補完するための行政的なデータの利用
第９章　登録簿にもとづく統計システムに向けて

第５章　精度と行政的なデータ

5.1　はじめに

第４章において述べたように、行政的なデータの精度にたいする懸念はしばしば、統計目的のために、その利用を拡大することにたいする主な障害のひとつとなっている。この懸念は、正当化され、あるいは正当化されないかもしれないが、しばしば適時性のような精度の特定な側面にだけもとづいている。このような懸念に対応するためには、客観的な精度管理の要綱が必要である。それは精度に関連するすべての側面を考慮する要綱であって、判定が事前に通知されることを許容する要綱である。

多くの統計機関はすでに、伝統的な調査方法によって収集されたデータについては、精度にかんするなんらかの要綱を定めているが、その方法手続きを行政的なソースからのデータを包摂するように

拡張している統計機関は、比較的に僅かである[21)]。

5.2 精度の定義

そのような要綱を設定する基点は、精度そのものの定義である。ここでもまた、多くの調査研究が、この分野において、国々の統計機関と国際的な統計機関によって行われてきており、そのほとんどは、国際標準 ISO9000/2005[22)] にもとづいている。それは、精度をつぎのように定義している。

「一組の基本的な指標が要請を満足する程度。」

不幸なことに、この定義は、理解することがとくに容易ではなく、さらにいくらかの説明が必要である。それは、解釈を助けるために、つぎのような部分に分けられる。

1）「要請」

それは、特定の財やサービスにたいする利用者の要求を意味するためにもちいられているが、また生産者の、さらには社会全体の要求をも考慮に入れるべきであるという議論もなされるであろう。例えば、大きなエンジンを搭載する高速車は、ある個人の要求に適うかもしれないが、公害や安全走行にかんする社会の要求には対応しないかもしれない。しかしながら、政府統計機関の最終生産物、すなわち統計そのものは通常、「公共財」として、公共部門内部において生産され、この場合、いろいろな階層の要求が重なり合っていることが多い。

もし、われわれが行政的なデータセットを、性格上たとえ生産物と見なしたとしても、生産者（例えばある行政省庁）と利用者（統計機関）の要求には相当な相違がある。さらに、「その取引」はしばしば、市場条件ではなされないので、生産者にとって利用者の要求を考慮する誘因は、ほとんど働かない。それは精度にかんする緊張感をもたらし、第３章において記述したように、しっかりした組織的な体制にたいする要求を強めることになる。

2）「基本的な一組の指標」

いかなる財やサービスの利用者も、その財やサービスのいろいろな属性に関連する一組の規準に照らして、その品質を判断する。それはしばしば、レストランにおける食事のように、そう判然とは行われない。ある人は、食事を、食物が料理され、提供される方法、食事の量、レストランのサービス、装飾や雰囲気によって、さらにはいくつかの他の規準によって、その食事の質を判断するだろう（さしあたり、経費は含まれない、本章の後半で触れたい）。同じように、統計の精度は、基本的な一組の指標に照らして判断される。

いくつかの統計機関は統計データの精度を評価する一連の規準を開発しているが、主要な国際機関は、つぎのような規準の一覧表にかんしては、一致している。

21）例としては、オランダ統計局によって開発された方法（http://isi2011.congressplanner.eu/pdfs/950481.pdf）とスウェーデン統計局によって開発された方法（http://www.scb.se/statistik/_publikationer/OV9999_2011A01_BR_X103BR1102.pdf）がある。

22）http://www.iso.org/iso/catalogue_detail?csnumber=42180　を参照。

- **重要性（Relevance）**―統計が現実のあるいは潜在的な利用者の要求に応ずる程度。したがって、重要性は、必要な統計が作成されているかどうか、また作成される統計が必要とされているかどうかにかかわっている。それはまた、使用される概念（定義や分類など）が利用者の要求を反映している程度を包含する。
- **正確性**―統計的推定値の真値にたいする近さ。
- **適時性**―それは、利用できるデータと、それが記述する事象ないは現象の時間差の長さを反映する。
- **即時性**―データが実際に公表された日付と（しばしば事前に案内される）目標となる公表日との時間差。
- **取得可能性**―利用者がデータを得ることができる物理的な条件、すなわち取得箇所、請求方法、送付時間、決済方法、売買条件の便宜（著作権等）、ミクロまたはマクロデータの利用可能性、（データ媒体）の種類（印刷物、電子ファイル、CD-ROM、インターネット…）など。
- **明確性・説明可能性**―データに十分で適切なメタデータが添えられているかどうか、グラフや地図のような説明が提供されるデータの価値を高めているかどうか、データの精度にかんする情報が有効かどうか。
- **首尾一貫性・整合性**―ソースが異なるデータ、とくに性格や周期が異なる統計調査のデータが、調査方法、分類や方法論が異なるために完全に整合していないかもしれない。したがって、完全に首尾一貫した説明が利用者に伝えられないかもしれず、例えば、同じ変数にかんして、二つの異なった尺度が異なった数値とともに公表されるならば、利用者は混乱するだろう。
- **比較可能性**―統計間の相違が、統計指標の真値の相違や方法論的な食い違いに帰因する程度。比較可能性は、つぎのような規準を含む。
  - ○時間的な比較可能性―異なる時点のデータが比較できる程度。
  - ○場所的な比較可能性―異なる国や地方のデータが比較できる程度。
  - ○分野間の比較可能性―異なる統計分野のデータが比較できる程度。

この規準の一覧表は、行政的なデータに関連して二つの方法において利用されることができる。第一に、それは、結果として生成する統計の精度を評価するために、および行政的なデータソースにもとづくデータを、調査によるデータと比較するために利用される。第二に、規準の一覧表は、いろいろな行政的なデータソースそのものの精度を評価するために役立てられる[23)]。例えば、もし、統計家が、幸運にも、複数の行政的なデータソースを選択することができる状況にあるならば、より高い精度をそなえるデータソースを決定するために役立てられる。

しかし、もし規準の一覧表が、これから行政的なデータソースの精度を評価するために利用されるのならば、つぎのことが指摘されるべきである。それは、母集団とデータの収集過程にかんする十分な支援情報がないならば、絶対的な正確性を決定することは困難であるということである。この場合、二つの要因、すなわちデータソースの信頼性とデータの妥当性が考慮されるべきである。すなわちデー

23）この方法の応用と拡張については、オランダ統計局の論文を参照「行政データベースの精度を評価するための検査表」http://www.cbs.nl/NR/rdonlyres/0DBC2574-CDAE-4A6D-A68A-88458CF05FB2/0/200942x10pub.pdf

タソースが信頼されるかどうか、そしてデータが他のデータソースと比較して、また統計家が期待する数値と比較して、データが適正とおもわれるかどうかである。より客観的な方法としては、ある種の精度調査が一定の変数の正確な価値を決定するために、必要であろう。

行政的な単位と変数が、統計目的のために要求される単位と変数に近いことが、行政的なデータソースの精度を決定する重要な要因である。必要となる変換が少ないほど、誤差や偏りの危険性も低いであろう。その性質は、首尾一貫性の規準の一部と見なされる。

5.3　経費の制約

経費は、より制約的と考えられるから、統計にかんする多くの精度規準表から、意図的には除かれている。精度が決定されると、経費は、経費効率性にかんする実際的な決定がなされる方程式に組み込まれる。

しかし、経費は行政的なデータソースの場合にはとくに重要である。それは、行政的なデータソースが調査データよりも、絶対的に低い水準の精度をもたらすときでも、行政的なデータソースに、なお十分な経費上の長所があって、最も費用効率的な方法であるからである。多少とも経費が節約されると、それを精度の改善に向けることが可能であって、それが精度の差を縮小したり、ないしは打ち消す。

5.4　精度測定の実際

行政的なデータソースの精度と、その統計の精度にたいする影響を十分に理解するためには、つぎの3つの要素を考える必要がある。

1）新しく入力されるデータの精度

行政的なデータソースからであれ、調査データのソースからであれ、新しく入力されるデータは、先に表示したような一組の規準に照らして評価することができる。

最も重要な規準は、適時性と重要性であって、後者はデータソースの捕捉範囲と概念が要求に対応する程度によって測られる。他のデータソースとの比較可能性も、また重要であって、ソースが異なるデータを調整するなんらかの種類の操作が、明確な精度像を得るために、時々、必要かもしれない。精度の点検調査が、ときにはこの目的のために利用される。

心に留める重要性がある一つの要点は、データの（作成…訳者注）主体がデータの精度に関心を寄せる程度である。データの提供に投入される努力と入念さの大きさは、データ収集にたいして認められる価値または重要性にしたがって異なる。したがってデータの主体が、いくつかの事例では統計目的のためよりも、行政目的のために、より良い精度のデータを提供するかもしれない。

2）データ処理の精度

新しく入力されるデータが完全であるときでさえ、その精度はなお、それが統計として利用される前に通過する処理の相違によって、影響を受ける。理想的には、精度が処理によって高められるべき

であるが、不幸にも、それはいつものことではない。データの処理がどのように精度に影響するかについては、つぎのような例がある。

・データの照合と連結－過剰な量の偽の一致がデータに誤差をもたらす。過剰な量の偽の不一致が重複をもたらし、関心を寄せている母集団の大きさを過大にし、偏りを持ち込むかもしれない。
・はずれ値の検出と処理－誤差を検出するために、はずれ値の探査方法をもちいることは、データの精度を高めることに役立つかもしれないが、一般に、はずれ値が極端であるほど、それは誤りである可能性も大きい。しかしながら、はずれ値の過度な処理は、原データの価値を変化させ、データが欠損する重大な傾向をもたらす。
・データ編成の精度－はずれ値の検出および処理と同様に、データの編成は精度を高めるにちがいないが、慎重に行わなければ、それは誤差と偏りを招く[24)]。
・補記（imputation）の精度－もし、補記が欠けている数値や記録を補充するためにもちいられるならば、それは捕捉範囲を拡げるかもしれないが、ここでも、利用される方法は、偏りが差し込まれることを避けるために、慎重な精査を受ける必要がある。

常にしたがうべき一つのたいへん重要な原則は、とくに行政的なソースのデータが加工されるときには、原データ（それに添えられているいかなるメタデータも）の複写を、必要ならば、遡及するために、保管することである。処理前後のデータを比較することは、処理の精度を評価し、特定のいかなる問題を識別するために役立つ。

３）統計の精度

ISO の精度にかんする定義にたいする統計機関の一般的な説明によると、精度は、利用者の要求に対応することが、すべてである。したがって、統計の精度は、このような事情において決定される。それは、この要求を確定することが必要であって、利用者と議論し、例えば利用者の満足度調査によって、定期的に、その反応を把握する。

調査から行政的なデータソースへの移行は、明らかに統計（output）の精度に影響するであろう。影響は一般的に、いくつかの精度規準については肯定的であって、その他の規準については、否定的である。すべての場合に、影響を全体的に吟味することが必要であって、利用者が最も重要と考える規準に、大きな重みが置かれる。例えば、利用者は、とくに短期的な経済データについては、適時性の改善が正確性の低下を相殺して余りあると感じているかもしれない。別の考え方では、時系列データへの影響が考慮されるべきであって、変動を追跡する十分な期間について整合的な系列を作成できるかどうかが考慮されるべきである。

とくに、少なくとも、統計家の意識にたいしてと同じくらい利用者の見解に重みをおくことが重要であって、いくつかの場合に、統計家の意識は、正確性にかんする慣習的な観念に焦点をあてすぎている。全体として、統計への影響にかんするいかなる判断も、想定ではなく、客観的な証拠にもとづ

24）いろいろなデータの編成問題にかんする包括的な論文集については、国際連合欧州委員会が組織した統計データの編集にかんする部会つぎの論文（working papers）を参照。http://www1.unece.org/stat/platform/display/kbase/UNECE+Work+Sessions+on+Statistical+Data+Editing

くことが肝要であって、それが、第4章において述べたように、変化にたいする抵抗力に対処するただ一つの方法であるからである。精度報告の利用を確保する一つの方法は、データソースを変える影響を記録化し、伝えるために標準的な定型書式（templates）[25] に従うことである。

5.5 メタデータの役割

メタデータ[26] は、作成者と利用者双方に、データの精度にかんする情報を与えるために重要である。それは、前節に述べた3つのすべての段階について提供される。新しく入力されるデータは、それを完全に理解するために、十分なメタデータが添えられるべきであって、それが、数値に関係する変数に正確に割り付けられることを保証すべきである。利用されたデータの収集方法や加工方法についてだけでなくて、データソースの概念、定義や目的にかんする詳しい資料も、また重要である。それは、考えられる精度問題にかんする理解をより良くし、データの処理行程においてデータを編成する規則の基礎となるべきである。

データ処理においては、施された処置、およびどの記録と数値に、いかなる処理が施されなされたかを記録することが重要である。それは、処理の精度を評価するために大切な情報を与えるだけでなくて、処理において考えられる問題を精査し、いかなる誤りも取り除くための技法を提供する。

統計には、十分なメタデータが添えられるべきであって、利用者が統計を修復し、正確に解釈し、その精度を判定できるようにする。統計を規則的にかつ大量に使用する利用者にとって、なるべく標準的な書式にしたがう3つの行程にかんする十分な参考資料が、利用者がデータから正しい結論を析出することができるように、必要な情報を用意するであろう。精度にかんする情報の伝達を正しく行うことは、難しい。それは、ある利用者が集計度がたいへん高い指標で満足するのにたいして、別の利用者は、十分な詳細さを求めるからである。利用者に異なる水準の情報を理解させることができる一つのメタデータのモデルは、集計表から出発するが、詳細な事項をも理解することできる方法であって、おそらく最も適切なモデルであろう。

5.6 要約

行政的なデータソースの精度を評価する最良の方法は、データソースを完全に認識することであって、それはデータソースの基本的な目的、およびデータが収集され、処理される方法を含んでいる。データソースの完全な理解は、データの長所と短所にかんするより正確な評価を可能にする。

異なるデータソースを利用する影響を評価するために、データソース、およびそれを統計に変換するために利用される処理にかんする知識を、その結果である統計の利用者がもつ見解と結びつけることが必要である。それが、統計調査のデータと比較することによって、行政的なデータを利用する影響にかんする客観的かつ全体的な評価を可能にするであろう。

25）例えば、欧州連合統計局の提案がある。
http://epp.eurostat.ec.europa.eu/portal/page/portal/ver-1/quality/documents/ESQR_FINAL.pdf を参照。

26）その他のデータを定義し、記述するデータ（出所：ISO/IEC FDIS 11179-1「情報技術 ── メタデータ登録簿 ── 第1部：枠組み作業」、March 2004）

第６章　データの連結と照合

6.1　はじめに

統計作成過程における行政的なデータの利用には、おもに、つぎのような２つの方法がある。

・統計の直接的なデータソースとして

・他のデータソースと組合せた間接的なデータソースとして

もし、いくつかの行政的なソースのデータを、調査データを補うために、あるいは統計登録簿に追加するために利用するならば、国家の統計機関はそれらのデータを連結する何らかの方法を備えることが必要であろう。それは一般的に照合という処理方式をとるであろう。照合とは、それらのデータソースに存在する共通の特徴にもとづいて、ソースが異なるデータを連結することであると定義できる。

6.2　共通の識別子とは何か

もし、データソースに共通の特徴として、ある種の共通の参照番号あるいは識別番号（これ以降は共通の識別子と呼称）があれば、その処理は完全照合と呼ぶことができるし、比較的に容易である。完全照合では、２つの結果が考えられる。すなわち、ソースが異なる２つのレコードが、もちいられた共通の識別子にもとづき正確に連結するか、しないかのいずれかである。換言すると、識別子123456のレコードは、（データソースが同じ単位を捕捉していると仮定して）ソースが異なる同じ識別子のレコードと連結できるのにたいして、識別子123457の単位とは連結できない。

完全照合は、それぞれのデータソースで使われる照合変数の精度に大きく依存する。もし、少なくとも１つのデータソースにおける共通の識別子に誤りがあれば、誤った単位を連結するか、または連結すべき単位を連結しそこなう。このために、共通の識別子が照合されるすべてのファイルに存在するときでさえ、完全照合だけに依存することは十分でないことがある。

ときどき、識別子は審査用の数字を含むことがある。すなわち、識別子における、その他の数字にもとづき、標準アルゴリズムに従って生成される１つ以上の文字を含むことがある。もし審査用の数字が存在すれば、それはほとんどの入力ミスや読み取りの誤りを除去することで、ある水準の精度を保証するために役立つはずである。

6.3　照合用キー変数と識別力の概念

共通の識別子が存在しない場合、あるいは求められる照合の正確度を与えるために十分な精度が存在しない場合、関連データソースに共通する他の変数の利用を考慮することが必要である。選択される変数をしばしば「照合用キー変数」と呼ぶ。注記：その変数を導出することができるいくつかの事例のように（たとえば、囲み資料4.4の１人当り売上高比率に関する議論を参照せよ）、それが２つのデータソースに存在することは必ずしも必要ではない。共通の識別子以外の変数を利用するとき、照合の定型的な処理手続きでは、どのレコードが連結するかを決定するために確率に依存する傾向がある。

この種の確率的な照合にほぼ共通に利用される変数は、氏名、住所、生年月日、職業コードないし経済活動コードである。照合に使われる変数の選択は、各変数の「識別力」も考慮に入れるべきである。識別力は、照合用キー変数の数値の一意性に関係する。つぎのような変数は、他の変数よりも高い識別力をもつ。

・高い識別力：参照番号、完全な氏名、完全な住所
・低い識別力：性、年齢、都市、国籍

「完全な氏名」のような変数のなかで、いくつかの値が他の値よりも高い識別力をもつことがある。唯一の氏名は最高の識別力をもつであろう。これに対して、ありふれた氏名（たとえば、英語圏の多くの国における John Smith）は、とても低い識別力をもつであろう。

識別力はまた詳細度にも依存することがある。すなわち、

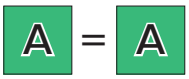
・“Born 1960, Paris（1960年生まれ、パリ）” は、識別力が低い。
・“Born 23 June 1960, rue de l’Eglise, Montmartre, Paris（1960年６月23日生まれ、教会通り、モンマルトル、パリ）” は、識別力が高い。

したがって、識別力の概念を考慮に入れて照合用キー変数を慎重に選択することは、照合の操作を首尾よく行うことに大きな影響を与えることができる。

6.4　いくつかの基本的な照合の用語

２つのレコードを比較するとき、それを「組」と呼ぶことができる。以下のシナリオは、一組のレコードに照合の技法を適用することから見込まれる主な結果を例示している。

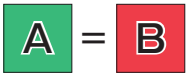
１）一致－実際に同一の実在物を示す一組



２）不一致－実際に２つの異なった実在物を示す一組

３）一致の可能性－一致かそれとも不一致かを決定するために十分な情報がない一組

４）誤った一致－照合処理で一致と誤って指定された一組（偽陽性）



５）誤った不一致－実際には一致しているが、照合処理で不一致と指定された一組（偽陰性）

照合の概念と実際の問題をよりよく理解するためには、本章末の照合の演習問題をご覧ください。それは架空ではあるが、現実的なデータを使用している。これは、共通の識別子を欠いた照合がいかに一定の判断力を必要とするか、あるいはしばしば、照合が精密な科学というよりはむしろ芸術的であるかを例示するためである。いかなる種類の確率的な照合も、一定割合の誤った一致や不一致をもたらすおそれがあるし、一致の可能性にかんして、いっそうの探索の必要性をもたらすであろう。

6.5　照合の技法

照合の技法は、２つの基本的な種類に分けることができる。

１）職員による照合－これは定義上、かなりの人的な資源を必要とするから、つぎのようになりそうである。

・経費がかかる
・一貫性がない
・遅い
・しかし、判断力がある

２）自動的な照合 －ひと度、運用可能になると（すなわち、１回限りの設置費を度外視するとして）、この方法は人的な介入を最小化するので、つぎのようになりそうである。

・安価である
・一貫性がある
・速い
・しかし、判断力は限られる

したがって最良の解決策は、明白な一致や不一致を見つけるために、自動的な照合の応用プログラムを利用することであり、専門的な事務職員に一致の可能性の処理を任せることである。費用効率性を高めるために、自動的な照合の割合を最大化する一方で、職員の介入を最小化することでなければならない。本章の残りでは、自動的な照合の主な特徴と、それが実際にいかに利用・改善できるかを考察する。

6.6　自動的な照合の進め方

自動的な照合の応用プログラムは通常、連続した類似的な段階をたどるが、特別な応用によっては、除外される段階や追加される段階もある。もっとも一般的な段階は以下のとおりである。

1）*標準化*

この段階は、主としてテキスト変数、すなわち、特定の書式に従うべき変数のために使われる。標準化処理の例はつぎのとおりである。

- 略語と通称の用語を標準文字列に置き換える。たとえば、文字列 “ltd” を “limited（有限責任会社）” に、“mfg” を “manufacturing（製造）” にそれぞれ変換することができよう。
- 通称名を標準化する。たとえば、都市名の別称がある（ベルギーの “Brussel〔ブリュッセル〕” ないし “Bruxelles”、北アイルランドの “Derry〔デリー〕” ないし “Londonderry〔ロンドンデリー〕”）。同じような処理が人名にたいしても必要であるのは、同名の異なる綴り（“Jane” ないし “Jayne”）や、ありふれた名前の短縮形（“William” にたいして “Bill”）が使われるときである。これは都市名と同様な処理であり、略語の標準化と併用できる。
- 「不要な（noise）」語を除去する。一般的に、それは、非常に低い識別力をそなえた語や句である。たとえば、住所における “road” ないし “street” をあげることができよう。
- 郵便番号、生年月日などには、共通の形式を与える。たとえば、“3 January 1985” を “030185” に変換することができよう。

標準化処理は言語に大きく依存している。また、照合されるレコードの種類に応じても変化することがある。それゆえに上述した例は、その処理を説明するにすぎない。照合を実際に行う場合にはそれぞれ、どの標準化規則を適用すべきかを決定するために、通常は、データの精査にもとづく前処理（prior work）が必要となるであろう。

標準化はまた、一種のデータの整理（data cleaning）と見なすことができる。それゆえに、つぎのような危険性をともなう。すなわち、標準化はデータの精度を歪め、または低下させることがある。極端な場合では、正しい一致を見つける可能性を低下させることもあろう。そのような危険性は、たいてい非常に小さく、通常、標準化される文字列の曖昧さに原因がある。英語の例に略語 “St.” がある。これは、“street” あるいは “saint” のいずれかを指している。また、“Chris” という名前がある。これは “Christopher”（男性）あるいは “Christine”（女性）の短縮形であろう。

照合処理の最初の段階としてときどき利用される、もう１つの種類の標準化は通常、政府郵政局のもっとも信頼のおける住所録と照らし合わせて住所を審査することである。これは、郵便番号と町ないし都市ないし地域との組合せの有効性の審査から、全住所の完全な審査に及ぶことがある。そのような審査の成功は明らかに、利用される住所参照ファイルの精度に大きく依存するであろう。

審査の結果が「整理済み」住所を使用することであれば、原データのコピーを保存しておくことも適切な実務である。いくつかの事例（イギリスの企業データの照合を含む）において、整理済み住所を利用することは、いくつかのレコードを連結する可能性を高めるが、他のレコードについては、その可能性を低下させることが分かっている。２つの並行的な照合の処理結果を組合せることは、すなわち、一方で、整理済み住所を利用し、他方で、整理前の住所を利用することは、しばしば最良の結果を与えることができる。

整理済み住所を利用することによって見込まれる２つの別の結果を、たとえその住所が厳密には照合に関係していないとしても、留意すべきである。第１に、いくつかの国における郵政局は、使われた住所が一定の基準に一致する大量の郵便を値引きすることから、整理済み住所を利用することは、データ整理の処理費用と照合の処理費用を相殺するのに役立つことがある。他方、回答者によって与えられた住所を整理済み住所に代えることは、いくつかの事例では、回答者を不快にすることがある。もし整理済み住所を統計調査票の郵送に使用するならば、回答率に影響を与えるであろう。これらのことは、できる限り整理済みデータと原データの双方を保持するための追加的な論拠である。

２）*構文解析（parsing）*

構文解析はある程度、標準化の拡張と見ることができる。この段階では、文字列を人が容易に認識できる形式から、むしろコンピュータ処理の論理にかなった形式に変換する。したがって、正確に一致する可能性が高い。その結果、もたらされる文字列をしばしば照合用キー変数と呼ぶ。英語における構文解析の初期の方法はしばしば、1918年に最初の特許を得た「Soundex 関数のアルゴリズム」を利用した。このアルゴリズム、またはそれから派生したアルゴリズムが、照合にかんする多くの応用プログラムの基礎を成している。しかしながら、構文解析の規則は言語間でかなり変化するので、関係データにたいして最良の結果を与えるように調整すべきである。

構文解析の規則にかんする例には、つぎのようなことがある。

- 類似の音声をもつ文字ないし文字群を共通の文字列に変換する。たとえば、“f”、“v” および “ph” を “ f ” に変換する。
- 無声音の文字を除去する。たとえば、“Thomas” という名前の “h” を除去する。
- すべての文字を大文字か、または小文字に変換する。
- 母音字を１文字に変換する。
- 名称または単語の最後の母音字を除く。
- ２重の文字を１文字に置き換える。たとえば、“Ann” は “An” となる。

たとえば、上記のすべての規則を利用する構文解析を通じて、文字列 “Steven Thomas Vale” を “stafan tamas fal” に変換できよう。文字列 “Stephen Tomos Vael” も同じ結果を与えるであろう。これは、構文解析が氏名の異なる綴り方の影響や綴りの誤りの影響を減らすことによって、一致率を高めるために、いかに役立つことができるかを示している。構文解析の規則を適用する順序の変更が、結果に影響を与えることもまた注意すべきである。

しかしながら標準化と同様に、構文解析が照合されるデータに十分うまく適合しないならば、それは役に立つというよりも、むしろ害を与える危険性がある。少なくとも、初めの段階では、構文解析の定型的な処理手続きの影響を慎重に分析すべきであって、いかなる場合でも、原データのコピーを比較のために保持すべきである。

3）*区画化*

もし、照合されるファイルがとても大きいならば、処理時間を節約するために、より小さな「区画」に分けることが必要であろう。それを行うために、いくつかの方法がある。たとえば、照合されるレコードに町の住所があれば、そのレコードを、全国のすべてのレコードとではなくむしろ、その町の他のレコードを含む区画と照合しさえすれば十分であろう。

区画化は慎重に利用されなければならず、しばしば全体の一致率の低下をもたらすであろう。しかしながら、その低下がごくわずかであって、より迅速な処理から得られる時間的利益が大きいならば、区画化は照合処理の費用効率性を高めることができる。いくつかの事例では、さまざまな区画化基準をもちいて、２つ以上の照合を試みることが適切でさえある。たとえば、データセット全体にたいして、比較的に限定的な照合基準を適用した後に、さほど限定的でない区画化基準を引き続きもちいて、当初は一致しないレコードの部分集合を再処理することが有益になるであろう。

区画化は明らかに、人口センサスの個人レコードのようなとても大きなデータセットにもっとも適切であるが、それは、照合の技法として、コンピュータの処理能力と処理速度の増大につれ、有用性を低下させるおそれがある。

4）*点数化*

自動的な照合の定型的な処理手続きでは、たいていの場合、２つのレコードが一致する可能性を評価するために、ある種の点数化を利用する。点数は、照合変数がどの程度緊密に一致するかをもとに割り当てる。割り当てた点数は、一組のレコードが明確な一致か、それとも一致の可能性か、それとも不一致かを決定するために利用できる。図6.1（原文では図5.1…訳者注）は、100点満点で示された、閾値の点数ｘとｙをもとに、３種類の照合結果を、いかに割り当てることができるかを示している。

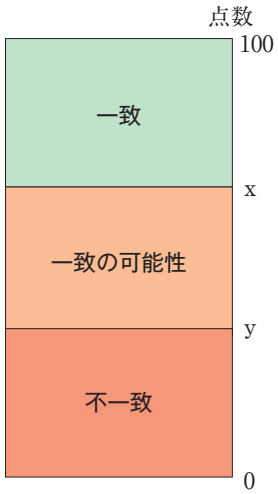


図6.1　閾値の点数を利用した照合結果の種類の決定

つぎの論理的な問題は、ｘとｙの値を決定する方法である。１つの選択肢は、Fellegi and Sunter[27]が提案したように、モデルにもとづく方法を利用することである。とはいえ実際には、それと同時に試行錯誤的な方法をもちいることになろう。

繰り返される照合の操作にたいして、データの精度は時間の経過とともに変化するから、ｘとｙの値を定期的に再評価することが必要となる。同様に、照合されるデータに対する要求の変化、または照合に利用できる資源の変化は、閾値の修正を招く可能性がある。閾値はまた、さまざまなデータセット間でかなり変化することもある。

ｘとｙの値を設定するときに、さまざまな種類の照合の判定ミスを考慮することも、また必要である。もし誤った一致によって、あ

27）Ivan P. Fellegi and Alan B. Sunter ,“A Theory for Record Linkage” を参照せよ。http://www.jstor.org/stable/view/2286061

る単位の統計的な情報が別の単位に露見するおそれがあるならば、そのとき、x の値は、誤った一致の危険性を許容できるほどに低くするために、十分に高く設定すべきである。しかしながら、もし、露見する危険性がなく、一定割合の誤った一致が結果に重大な影響を与えそうにない研究に、照合の結果を利用するのであれば、x の値をむしろ低くすることができる。

一致の可能性を精査するために職員を利用できると、しばしば実際には、x と y の距離を制約することがある。そのようなすべての場合に、職員による介入を優先すべきである。これは点数をもとに行うことから、最高点をもつ一致の可能性を最初に審査する。これが最大の利益を与えると想定できるからである。代わりに、関連単位のいくつかの他の特性（例、企業の雇用者数）では、見込まれる重複の影響を最小化できるように、職員による作業に優先順位を付けることがある。

6.7　照合の応用プログラムの実際

人々はしばしば、上述した方法とはかなり異なる方法で作業するが、多くの人々がもっとも良く知っている照合の応用プログラムは、インターネットの検索エンジンである。それは（利用者が入力した）文字列を受取り、それに関連するウェブページを検索する。しばしば、検索の結果を点数化し、認知される関連度の順序で結果を返す。検索の結果において、または綴りの代替候補が提案されるので、ある種の構文解析が存在することも明らかであろう。

インターネットの検索エンジンはまた、識別力概念にかんする良い実例を与える。たとえば、本書の執筆時点では、文字列 “matching” の www.google.com（グーグル）による検索は約７億件の結果を返したのにたいして、“statistical matching” は約3,000万件の結果、“parsing techniques in statistical matching（統計的な照合における構文解析法）” は約1,600万件の結果を返した。文字列を詳しくすると明らかに、検索の焦点を絞るのに役立つ。

政府統計の世界では、データ照合の応用プログラムを開発するために、２つの主な方法がある。

- 「在庫があってすぐに購入可能な」市販ソフトウェアを利用すること。たとえば、Informatica Identity Resolution（インフォマティカ社製の検索・照合ソフトウェア…訳者注）（SSAName3〔IBM 社製で検索・照合の中核的ソフトウェア…訳者注〕を組み込む）[28] を利用することが、それである。しかしながら、どの市販パッケージもその能力を完全に利用できるまでに、ある種の仕様変更が必要であろう。
- 組織内で照合の処理プログラムを開発すること。たとえば、米国センサス局[29]、カナダ統計局[30] およびイタリア国立統計研究所 ISTAT（イタリア統計局[31]）によって開発されたソフトウェアがある。

28）http://www.informatica.com/products_services/identity_resolution/Pages/index.aspx
29）http://www.census.gov/srd/papers/pdf/rr2001-03.pdf
30）http://www1.unece.org/stat/platform/display/msis/G-Link
31）http://forge.osor.eu/projects/relais/

代わりの照合方法は「３文字区分」法である。これは、文字列を３文字のグループに分け、２つの文字列間における同じグループの割合を計算する。

たとえば、文字列 “Steven Vale”：Ste/tev/eve/ven/en /n V/ Va/Val/ale と、文字列 “Stephen Vale”：**Ste**/tep/eph/phe/hen/**en /n V**/ **Va**/**Val**/**ale** を照合する。

結果は、双方の文字列について、全体で13の３文字グループのうち、一致した３文字グループ（太字体）は６つあるので、点数は６/13、すなわち0.46となる。文字列の構文解析は、点数を高めるのに役立つことがあるが、上述したように、誤りを招くおそれもある[32)]。

囲み資料6.1　── 事例研究 ── スティーヴン　ヴェイルとマイク　ヴィラーズによる「共通の識別子がないレコードの照合 ── イギリスの経験」からの抜粋

この資料は、下記 URL で検出することができる縮約されていない論文からの引用である。

http://www1.unece.org/stat/platform/download/attachments/56230020/matching+paper.pdf?version=1

イギリスの企業統計登録簿は、いくつかの行政的なソースと統計的なソースのデータを利用している。それらのデータソースのなかで最も重要なものは、付加価値税レコードと源泉課税方式の所得税レコードである。この２つのデータソースの捕捉範囲にはかなりの重複があるから、重複を最小化するために、各データソースの新たな単位が本物であり、他のデータソースからまだ追加されていないことを審査することが不可欠である。各データソースには、それ自体に単位の識別子システムがある。これは、名称と住所（ないしは所在地…訳者注）にもとづく照合が最良の解決策であることを意味する。

入力されたファイルは、つぎの４つの段階で処理される。

・整理―この定型的な処理手続きでは、名称の文字列を編集する。すなわち、特殊文字を除き、小文字を大文字に置き換える。
・書式設定―この定型的な処理手続きでは、名称の文字列を別の語に編集する。すなわち、「ストップワード」を除去し、選択された語を取り換え、接頭語をつなぎ合わせる。
・標準化―この定型的な処理手続きでは、名称を「標準化する」。たとえば、２重文字を除去する。
・照合用キー変数の生成―これは入力された文字列にもとづいたコードを生成する。たとえば、その文字列が “Steven Vale” であれば、作成されるキー変数はつぎのとおりである。
STEVEN → STAFAN → XJXM$$$　および　VALE → VAL → YLVO$$$$

YLVO$$$$ は、氏名のなかで氏にたいするキー変数であり、主要なキー変数としてもちいられる。それを一致の潜在的可能性を見つけるために、登録簿に保持された各レコードの氏名から生

32）その方法にかんして、SAS コードとしてプログラム化された実用的な応用プログラムが、企業動態にかんする統計を開発するための欧州連合統計局プロジェクトのなかで、フィンランド統計局によって明らかにされた。

成された氏名キー変数の一覧表と照合する。入力された氏名、住所および郵便番号を、一致の潜在的可能性がある各レコードの氏名、住所および郵便番号と比較し、100点満点の点数を与える。もし、点数が79点よりも高いならば、その組は明確な一致とみなし、点数が60点と79点の間であれば、それは一致の可能性に入る。59点よりも低い点数であれば、それは不一致とみなす。

明確な一致の一覧表におけるレコードの重複を除去する。そして、明確な一致の一覧表にあるレコードで、一致の可能性の一覧表にあるレコードも除去する。それから、明確な一致一覧表のレコードが、自動的に登録簿の対応する単位と連結される。一致の可能性一覧表にあるレコードと、より多数の不一致レコードが、職員による審査のために報告される。典型的な更新のために、約37％のレコードが明確な一致であり、約35％が一致の可能性に入る（その約80％を職員によって一致させることができる）。

直面した問題は、たとえば “Mike Villars T/A Mike’s Coffee Bar（マイク　ヴィラーズはマイクのカフェバーとして商売を営む）” のような企業名における “Trading as” あるいは “T/A” の使用であった。この場合、“Bar” を主要なキー変数として利用したいが、イギリスには多数のバーがあるので、その識別力は低い。解決策は “T/A” の直前の語、すなわち Villars が主要なキー変数であるように、企業名を分割することであった。

第６章の付録 ── 照合の演習問題

この演習問題は、新しいレコードをひとそろいの既存レコードと自動的に照合する５つの例を含んでいる。いかなる明確な一致も見つけられないが、一致の可能性に区分される最高の点数がついた５つのレコードを職員による審査に提出する。これらのデータは現実的であるが、実際には事実の記録ではない。新しいレコードに最も良く一致する既存レコードを選択して下さい。あるいは、一致の可能性に区分されるどのレコードも、新しいレコードとほぼ同じと思われないならば、不一致であると決定することができます。解答は例５の後にあります。

例１

新しいレコード	一致の可能性	
企 業 名：Bob the Butcher 所 在 地：16 ”Lawrence Street Southfleet Gravesend 郵便番号：DA11 7ZP	1	Bob Daley Butchers 17 Barwick Green Sidcup Kent DA15 8HP
	2	Brian Dunn Brians Family Butchers 16 Pembroke Close Pembroke Street Dover Kent DA6 1FB
	3	Mr B Dunn and Mrs V Dunn Brian’s Family Butcher Pembroke Street Gravesend Kent DA6 1AA
	4	B & B Butchers Mr B Jones 3 Clive Road Dartford Kent DA1 5RH
	5	B Washbrook Bob the Butcher 16 Lawrence Drive Castle Lane Southfleet Gravesend Kent DA11 7ZF

例２

新しいレコード	一致の可能性
企 業 名：Cars of Southfleet 所 在 地：3-5 Old Hill Southfleet Dartford 郵便番号：DA1 9KT	1 Fleet Motors 31-35 Old Dover Road Dartford Kent DA15 7JF
	2 Southwold Cars 1A Southwold Close Greenhithe Kent DA23 9BC
	3 Mr D Crane T/A Southeast Cars 12A Old South Road Greenhithe Gravesend Kent DA2 9BN
	4 Mr C James & Mr G Smith Fleet Motors 29-35 Old Dover Road Fleet Kent DA15 9XX
	5 Southfleet Cars 33 Old Hill Southfleet Dartford Kent DA1 9XT

例3

新しいレコード	一致の可能性	
企 業 名：Retail Co-operative Limited 所 在 地：35, Station Parade Station Road Dartford 郵便番号：DA1 7ED	1	Mr A Cooper Paintcraft Unit 132 Greenway Estate Lower Station Lane Welling Kent DA18 6GT
	2	Retail Co-op Ltd 030001 35 Station Street Dartford DA1 7DH
	3	Co-operative Funeral Services 362 Longfield Street Dartford DA1 1HD
	4	Co-operative Funeral Services Ltd, CFS（No14）Ltd & CFS Pension Fund 29 Station Street Bexleyheath Kent DA32 4RH
	5	Arts Co-operative 62 Highfield Street Dartford DA21 8JD

例４

新しいレコード	一致の可能性	
氏　　名：Dr James Johnson 住　　所：Griffons Penny Lane Eynsford Dartford 郵便番号：DA46 8FF	1	Mr James John Cunningham 35 Griffin Drive Darenth Dartford Kent DA4 6FF
	2	Mr John Jameson 56 Whinfell Road Gravesend Kent DA21 8GF
	3	Mr James Johnson 123 Penny Lane Aynsford Kent DA46 3JF
	4	John James 23 Perry Lane Dartford Kent DA28 3PF
	5	Mr James John Smith 18 Cornfield Lane Eynsford Dartford Kent DA46 8FF

例5

新しいレコード	一致の可能性	
企 業 名：Redipure Ltd 所 在 地：26A Queens Rd Welling 郵便番号：DA13 8RS	1	Redipure Limited Perseverence House 36A Cross Road Howley Dartford Kent DA27 8RR
	2	Eradicure Ltd Perseverence House Cross Rd Howley Dartford Kent DA27 8RT
	3	Redpull Ltd 152 Lower Wickham Lane Wellington Kent DA13 8ED
	4	Redpull Ltd 12 Lower Wickham Welling Kent DA13 3ED
	5	Redipure Holdings Ltd Crossroads Howley Dartford DA12 3LF

解答

この演習問題は、照合において100％の確実性が稀であることを示している。以下の解答では、事務的な照合の専門家にしたがって、新しいレコードがどの既存レコードと一致の可能性が最高となるかを考察する。

例１－一致の可能性がもっとも高いのは既存レコードの５番である。この既存レコードの商号は、わたしたちの新しいレコードの企業名と一致し、所在地はかなり似ている。郵便番号の違いは１文字である。すなわち、“F” の代わりに “P” であって、それらのレコードの１つに転記ミスがあることは容易に考えられる。

例２－またもや、一致の可能性がもっとも高いのは既存レコードの５番である。企業名と所在地は十分に似ており、例１と同様に、郵便番号の違いは１文字にすぎない。この場合はまた、既存レコードの１番と４番の組も一致であろうという興味深い問題を際立たせている。それは既存レコード間の重複を意味しており、そのような重複の危険性を減らすために、１つのデータセットをそれ自体と定期的に照合する有用性を示している。

例３－もっとも緊密に一致する既存レコードは２番である。主な相違は、既存レコードの企業名における略語（Ltd は limited〔有限責任会社〕、Co-op は Co-operative〔協同組合の〕の略語）の使用に関係している。これは、自動的な照合の定型的な処理手続きが、略語をその完全版に関係づけるときに、十分に有効ではないことを示唆している。そのような略語はしばしば、言語あるいはデータセットにさえ特有であって、照合されるデータの種類に応じて、自動的に照合を行う応用プログラムの仕様を変更できることが有用であることを示している。

例４－もっとも緊密に一致する既存レコードは３番である。既存レコードの５番は、郵便番号だけでなく、住所の大部分についても完全に一致しており、自動的な照合において高い点数の可能性があろう。これは、確かな一致に低すぎる閾値を設定する危険性を例示している。

例５－ここではレコードの証拠だけにもとづいて、一致が存在するとは思えない。しかしながら、この場合は、職員による照合において追加的な情報を利用する有用性を例示している。新しいレコードの企業名における略語 “Ltd” は、有限責任会社であることを示している。多くの国において、有限責任会社は、法律によって、他に存在しない名称をもつ必要がある。これは、もし照合が同じ企業の単位を連結することを目的とするならば、新しいレコードは既存レコードの１番と連結すべきであることを示唆している。異なった所在地は単純に、その会社が操業している異なった場所（場所的単位あるいは事業所）を指しているかもしれない。法人企業を認知するために自動的な照合の定型的な処理手続きを改善することは、企業名をいっそう重視するならば、自動的な一致率を高めるのに役立つことができよう。この戦略は、英国の企業統計登録簿に利用される照合の定型的な処理手続きに首尾よく採用された。

第７章　統計登録簿における行政的なデータの利用

7.1　はじめに

前章では、行政的なデータを取得すること、およびデータを統計目的のために利用にする適合性を確保することにかかわるいろいろな問題を考察した。これらの問題の多くは、日々行われている統計登録簿の管理にとって重要であるが、ここでは繰り返さない。その代わりに、本章では、行政的なデータが、統計登録簿に統合されることによって、統計の作成過程に投入される方法を考察する。最初に、統計登録簿を定義し、それから、行政的なデータを統合するためにもちいられる、いろいろなモデルを考察する。

7.2　統計登録簿の定義

しばしば共通の課題についてではあるが、いろいろな登録簿の定義がある。広範にもちいられている定義は、つぎのとおりである。

「*登録簿は、特定の組の対象にかんする事項とその細目を、定期的に記入する書式による全数の記録である*[33]。」

典型的には、登録簿は、各単位について、いくらかの属性を含み、かつなんらかの定期的に更新する技術をそなえる単位の構造的な一覧表である。このような様式において、多くの行政的なデータファイルを登録簿と見なすことができるが、一回だけデータを収集した結果ではない。

統計が単一の行政的なデータソースから直接に作成されるところでは、そのソースは、登録簿とは見なされるべきではないし、同様にして、調査、またはセンサス結果でさえ、通常、登録簿とは見なされないという主張がなされる。この主張は、行政的なデータが個別的な単位次元のデータでなく、集計値の形式でもちいられるときには、より強くなされる。

統計登録簿は、統計的な概念と定義にしたがって、統計家が管理することによって、統計目的のために策定され、維持される登録簿である。したがって、行政的登録簿は統計登録簿のソースとして利用されるが、逆の方向は、通常、データの「一方的な流れ」の原則に矛盾すると考えられている[34]。

統計登録簿は、典型的には統計データソースと行政的なデータソース双方のいくつかのデータソースを統合するために、データを調整する手段の役割を果たす。それは、共通の識別子をもちいて記録を連結することによって、または第６章において記述したような照合技法を利用することによって策定される。それは、時には、単一のデータソースのデータを使用することが容易であるが、そのような場合には、しばしばデータソースの正確性を点検することは困難である。いくつかのデータソースが一つの統計登録簿の内部において利用され、統合されるときには、データの正確性をよりよく調査することができる。不幸なことは、その否定的な側面は、ソースが異なり、相反するデータを処理す

33）「統計的なメタデータにかんする用語」国連欧州経済委員会欧州統計家会議：統計標準と研究、No. 53、ジュネーブ、2000、http://www.unece.org/stats/publications/53metadaterminology.pdf.

34）政府統計の基本原則（原則６）を参照、http://www.unece.org/stats/archive/docs.fp.e.htm

るための戦略が必要になることである。しかし、もし、統計登録簿の変数がデータソースの符号と日付を付して入力されているならば、自動化された操作プログラムが、データソースの優先度を決めて、多くのデータの齟齬を解消することができる。

ソースが異なるデータを統合するだけでなくて、統計登録簿は、また新しい変数を導く可能性を与えるかもしれない。その一つの例は、いくつかの国[35]が、企業統計登録簿における法律的な形態、経済活動分類と国外所有にかんするデータをもちいて、国民経済計算のために利用される制度部門[36]を導いていることである。

伝統的には、統計登録簿は調査の標本抽出の枠組みとして利用されてきたが、それは、それ自身の性格によって、いよいよ統計データのソースと見なされてきている。母集団の地理的な小地域や小さな階層については、とくにそうである。統計登録簿は、ソースが異なるデータを時間的に連結する基礎を提供することによって、縦断面分析を許容する。この方法は、いくつかの国においてもちいられてきており、国民または企業のコーホート研究を許容している。

7.3　行政的なデータを利用する統計登録簿を策定し、維持するためのモデル

これまで述べてきたように、統計登録簿は、ソースが異なるデータを調整する重要な役割を担う。このようなデータソースが統計の標本抽出の枠組や統計を作成するために、利用されたり、結合されたりする方法は、多く存在する。本節では、いろいろな国において、しかも統計のいろいろな分野のために利用されているいくつかの方法を考察する。

使用できるデータソースが国々によって相当に異なるから、一つのモデルを輸出しようとすること、または国際的な標準モデルを定義しようとすると、しばしば困難に遭遇する。したがって、下記のいろいろなモデルは、すべての国において実施すべき勧告と見なされるべきではなくて、他の国が統計登録簿において行政的なデータを利用している方法を紹介する事例と見なされるべきである。既成の解決方法ではなく、むしろ特殊な国家的な環境に応じて採用できる考え方を提供することが、その目的である。

１）多数のデータソースの結合

下記の図7.1は、イギリスの企業統計登録簿を維持するために利用されているデータソースを単純化したモデルである。それは、意図的に統計登録簿を中心において、いろいろなソースのデータを結合し、調整する手段として紹介している。それは、後で本章において詳細に議論される補助登録簿（satellite registers）の概念と、データソースが行政的なデータと統計データの混合体であるという考え方を導入している。この事例では、地理情報システム（GIS）が、ある程度、統計的にモデル化された行

35）オーストリアの例については、N．ライナー、K．シュバルツ、R．シャウマンとT．カーナーによる「オーストリア統計局の企業登録簿における部門分類の導入にかんする報告」がある。本論文は、英語の要約を含み、欧州連合統計局によって取得が制限されている BR-Net' サイトを通して、インターネットで利用可能。

36）『国民経済計算体系2008』第４章参照、── http://unstats.un.org/unsd/nationalaccount/docs/SNA2008.pdf

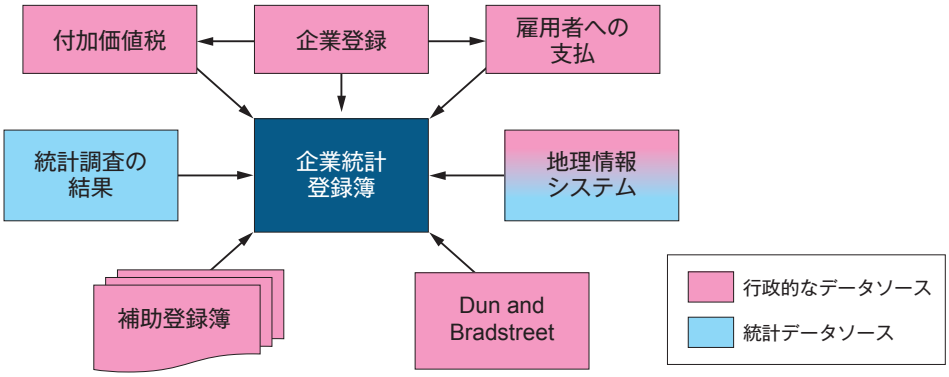


図7.1　イギリスにおける企業統計登録簿の単純なモデル

政的なデータ（おもに郵便業務データ）の混合体を含み、より統計的に等質的な地域を創出する人口センサスのデータを利用している。

２）集中化された行政登録簿の利用

集中化された行政登録簿は、しばしば政府内の効率性を高めるために策定されており、多くの事例において、重複を少なくし、したがって行政手続きを執る負担を軽減するように、登録簿の主体がいろいろな政府省庁と相互に連動することができる単一の接続装置（interface）を提供する。例えば、そのような登録簿が存在すると、個人や企業がその住所を変更するときには、その新しい事項を一度だしか提供する必要がなく、したがって、この事項は多くの関連する機関に共有される。

このような行政登録簿は、それがソースが異なるデータを照合し、調整する負担を、少なくともある程度まで除去するから、統計目的にとって大きな利益になることができる。しかしながら、この利益を最大化するためには、統計機関が、行政登録簿の開発と管理に、ある程度発言権をもつことによって、行政登録簿が単位、分類、定義および手続きにかんして、できるだけ統計的な要請に対応することを確保することが重要である。

この方法が実際に機能している良い事例の一つが、オーストラリア統計局による（行政的な）オーストラリア企業登録簿（ABR）の利用である[37)]。ABR は、いろいろな企業税を所管するオーストラリア税務庁によって開発されたが、経済活動分類のような特殊な分野において、入力操作の情報と専門知識を提供するオーストラリア統計局との密接な協力によって維持されている。

その結果、ABR が、最も大きな複雑な企業を除くすべての企業について、企業統計登録簿の適切な基礎となっている。事実、企業統計登録簿は、明確に２段構えの方法をとっている[38)]。多くの記録は、

37）http://www.abr.gov.au を参照
38）詳細は、つぎの文献を参照。
http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/8165.0Explanatory%20Notes1Jun%202007%20to%20Jun%202009?OpenDocument

ABR から直接に複写され、このデータソースによるだけで維持され、最も大きな複雑な企業の構造を把握することに、統計的な資源が解放され、集中されている。

３）データ共有センターの策定

単一の集中化された行政登録簿をめぐる話題の変種のひとつは、データ共有センターの概念である。このモデルでは、中心の組織体は、完全に独り立ちした登録簿ではなく、むしろ、いろいろな機関によって保有されているデータを検出し、照合する手段である。それは、非常に基本的な識別データを含んでいるが、その主要な目的は、いろいろな機関からのデータを政府省庁の内部において、共有することができる連絡路を提供することである。

図7.2は、イギリスにおけるそのような方法の実効性にかんする研究から取られている[39)]。この方法は実現されていないが、このモデルは、行政的なデータを共有するための有効な選択枝に入っている。薄ねずみ色の円（原文では青い円…訳者注）は、異なる政府機関を示し、それぞれが多くのデータ保有体（黒い円筒）をそなえている。それぞれのデータ保有体は、譲渡することができるデータと相手先を厳密に管理するポータルに結びついている。他方では、これらのポータルは連結されているデータ保有体の検索と照合を許容する十分なメタデータを含む中央のセンターに結びつけられている。このようにして、参加している機関の利用者は、中央のセンターを通して照会することができて、職員が取得権をもつ他の機関にある、関連するすべてのデータ保有体からデータを受け取ることができる。

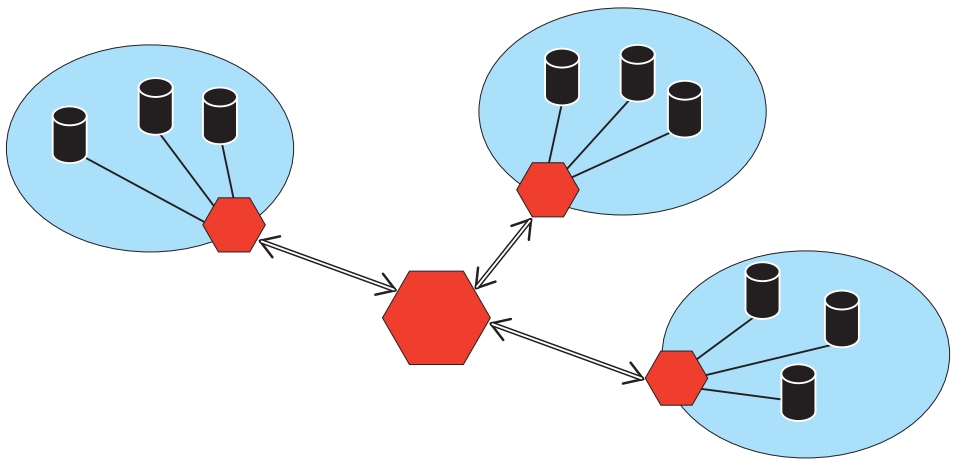


図7.2　データ共有センター

39）詳細は、つぎの文献を参照。
http://www.unece.org/stats/documents/ces/sem.46/5.e.pdf

4）補助登録簿による行政的なデータの利用

実際に行政的なデータを利用するやや異なったモデルは、行政的なデータを統計登録簿に結びつけるデータソースに特化した登録簿を組織することである。このようなデータソースに特化した統計簿が一定の基準に適っているならば、それは「補助登録簿[40]」と呼ぶことができる。補助登録簿は、国家統計システムにとって利用できる登録簿と定義することができ、単位と変数にかんする情報を含み、つぎのような条件を満たしている。

- それは、統計登録簿の統合された部分ではないが、それに連結することができる。
- それは、統計登録簿より範囲が限定的であるが、その範囲内において、統計登録簿の単位および変数の捕捉範囲を拡張することができる。
- それは、統計登録簿に検出されない単一または複数の変数を含む。そのような変数は、一般的に層別化のために、利用することができる。
- 通常、統計調査の結果が記録されるデータベースは、補助登録簿ではない。

したがって、補助登録簿は、統計登録簿に、ただ単位の部分集合にとって重要な行政的なデータを組み込むための手段である。それは、追加的な単位、または変数あるいは両方であるかもしれない。それは、行政的なデータソース、統計調査、または両者の結合体からの情報を利用することによって策定される。いくつかの事例では、それは、変数を追加するか、結合するか、さもなければ変形することがあるかもしれないが、他の事例では、それは特定のデータソースと、多かれ少なかれ同一であるかもしれない。補助登録簿が統計登録簿と十分整合することを保証するために、例えば、共通の単位識別子、共通の定義や分類のような追加的な規準を考慮に入れることが有効であるかもしれない。整合性が高いとそれだけ、補助登録簿は有効であろう。

図7.3 は補助登録簿が統計登録簿にどのように関係するかを示している。この図は、捕捉される単位と含まれる変数の両方において、解釈することができる。両方において、ある程度重なり合っているが、補助登録簿は、それは、追加的な単位についてであれ、または既存の単位の部分集合にかんする追加的な変数についてであれ、追加的な情報をもたらす。

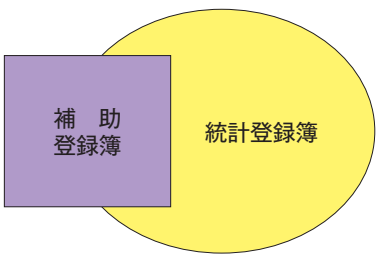


図7.3　補助登録簿と
統計登録簿の関係

補助登録簿にかんする最も新しい例は、企業データに関連しており、補助登録簿の範囲は、つぎのように決定される。

- 経済活動：補助登録簿は、例えば小売業、ホテル、道路運送のような特殊な活動を行う企業を含む。
- 規模：補助登録簿は、例えば「大企業」の部分集合のような一定数の雇用者をもつか、一定水準の取引高をそなえる単位を含む。
- 特性：補助登録簿は、例えば外国貿易に従事するような共通の特性をそなえる単位を含むかもしれない。

40）ときには連携登録簿（associated register）と呼ばれる。

補助登録簿に含まれる部分集団の単位に特化する変数の例は、ホテルの「種別」や「寝床数」、または小売業の「売り場面積」を含めることができていた。

補助登録簿は、層別や分析目的のために使用できる変数の範囲を拡げることによって、統計登録簿の重要性を増すことができ、層別変数の精度を高めることによって、標本抽出の効率性を改善することができる。それはまた、目的集団の捕捉範囲を大きくし、いくつかの場合には、統計調査によって収集することが必要な情報の量を減少させることによって、回答者の負担を軽減する。

５）登録簿にもとづく統計システム

登録簿にもとづく統計システムは、さらに第９章において議論されるが、ここでは、それが統計登録簿のために行政的なデータを利用するモデルを提供するかぎりにおいて述べる。先述したモデルと比較して、主な相違は、いくつかの連結された統計登録簿が広い範囲の行政的なデータを利用することによって策定されることである。このモデルは、おもにスカンディナビア諸国において、開発されてきており、３つか４つの中心となる統計登録簿をもちいている。図7.4は、スウェーデンにおいて採用されたモデルの簡略版を表示している。

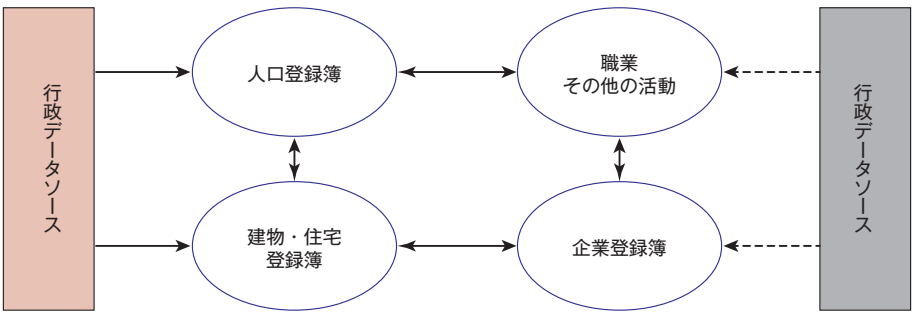


図7.4　スカンディナビア諸国の登録簿にもとづく統計システム

人口統計登録簿は、財産、または実物資産登録簿および企業統計登録簿に連結されていて、国民、財産と企業にたいする一義的な識別子のシステムを利用している。スウェーデンでは、第４の登録簿が導入されて、職業その他の活動にかんする項目が把握されている。この登録簿は、国民を賃金、年金と国家の社会保障給付を含む所得にかんするデータソースに連結しており、それによって国民と労働市場との関係を示している。

第７章の付録 ── 演習問題：企業統計登録簿の策定

あなたたちの政府が、企業家、および彼らが成功するか否かを規定する要因にかんするデータがもっと必要であると決定する。あなたたちの統計局は、この情報を提供する新しいデータの系列を作成することを決定する。あなたたちは、標本の抽出枠として利用する企業統計登録簿を、行政的なデータソースにもとづいて策定することを要請される。

あなたたちは、１万6,000ユーロの年次予算を得る。あなたたちが使用する各データソースを加工するために、2,000ユーロを支出する。それに加えて、データソースごとに異なるデータを購入する経費がある。

つぎのような行政的なデータソースを、あなたたちは使用することができる。

１．自営業の所得を申告する人々の税務記録

・内容：個人識別番号、氏名、住所、性別、申告所得額、企業名、企業の種類（国際産業分類〔ISIC〕２桁次元に対応する分類）。

・利用可能性：もし、あなたが、データの摘出と発送に要する税務局の経費を補填するために、年当たり2,500ユーロの料金を支払うならば、税務局は、毎年このデータを提供する。税務局は、CD-ROMによって、データを発送する。

・精度：データは、50％の正確さしかない「企業の種類」を除いて、95％の正確さである。あなたがデータを得る時点までに、データは、６ヶ月から18ヶ月の遅れがある。捕捉範囲は、合法的に企業を操業している国民の100％である。企業のほぼ20％が、（すなわち、所得を申告しない国民によって）不法に操業されていると推定される。

２．雇用者がいる企業の税務記録

・内容：企業識別番号、企業の名称と住所、企業の種類（ISIC の４桁次元に対応する分類）、企業が雇い主として最初に登録された年。

・利用可能性：もし、あなたが、データの摘出と発送に要する税務局の経費を補填するために、3,000ユーロの年間料金を支払うならば、税務局は、毎月、CD-ROMs によって、このデータを発送する。

・精度：データは、90％の正確さであって、一般的に２ヶ月から３ヶ月の遅れがある。税務局は、月ごとに、CD-ROMs によって、データを発送する。それは、合法的に国民を雇用している企業を100％捕捉している。企業の50％が雇用者を有し、その95％が合法的に操業されていると推定される。

３．住民行政登録簿

・内容：個人識別番号、氏名と住所、年齢、性別、教育水準、職業、国籍、出生国

・利用可能性：このデータは、すでに、3,000ユーロの年間経費によって、統計局によって利用されている。もし、あなたが、それを利用するならば、この経費の半額を支払うことになるだろう。データは、年次ごとに利用できて、それを、人口部門のあなたの同僚から、電子ファイルによって受け取ることができる。

・精度：データは、95％の正確さであるが、１年から２年の遅れがある。それは、合法的な国民の99％を捕捉しているが、住民全体のほぼ５％が、不法移民であって、捕捉されていないと推定される。

４．企業の電話帳（「黄書」）

・内容：企業の名称と住所、電話番号、企業の種類（独自の300種類の種目一覧表にしたがった分類）

・利用可能性：このデータは、民間部門の企業によって、商業的に販売されている。それは、毎月、CD-ROM によって利用することができる。年間予約は、通常7,000ユーロであるが、電話帳支給者は、統計局に、15％の割引で、提供する用意がある。

・精度：データは、電話帳支給者によって99％の正確さであると表明されている。支給者は、情報の正しさを保証することが、企業のためであると述べている。データは、一般的に１ヶ月から２ヶ月の遅れがある。税務局は、月ごとに、CD-ROMs によって、データを発送する。それは、すべての企業のほぼ95％を捕捉している（合法および非合法併せて）。

５．起業助成金にたいする応募者の一覧表

・内容：個人識別番号、氏名と住所、企業の識別番号、名称と住所、事業の種類（国際産業分類〔ISIC〕２桁次元に対応する分類）。

・利用可能性：ｅメールによって毎年３月に発送される。前年に助成金に応募した者を捕捉する記録シート１枚につき500ユーロ。

・精度：データは、少なくとも95％の正確さであるが、いくつかの住所には、時期遅れがある。起業を開始する者の約40％が起業助成金に応募しているが、彼らは、一般的には、最も成功する企業家である。それは、いかなる年でも、企業の母集団全体の６％である。

6.「全国企業協会」会員表

・内容：個人の氏名と住所、企業の名称、住所と電話番号、協会加入の日付

・利用可能性：毎年発行される企業家人名録（印刷物）一冊につき100ユーロ

・精度：少なくとも、90％の正確さであるが、いくつかの住所には、時期遅れがあるかもしれない。会費が相当に高く、企業家のほぼ10％しか会員になっていない。それはほとんど、少なくとも５年間は操業かつ成功している企業をもつ人々である。

問題：

１．限りある予算（１万6,000ユーロ）で、あなたたちは、どのデータソースを選択したいですか？

２．なぜ、あなたたちは、そのデータソースを選択したいとおもいましたか？

３．どのようにして、あなたたちは、異なるデータソースのデータを照合したいとおもいますか？

４．どのような種類の調査を、あなたたちは勧めたいとおもいますか？　個人面接調査ですか、電話調査ですか、それとも郵送調査ですか？

５．あなたたちは、どの変数を、調査標本を層別するために利用したいですか？

解答：

この演習問題には、実際にどのような正解も、誤った解答もないが、考慮すべき要因は、つぎの条件を含む。

・１～３のデータソースは、一般的に合法的な登録された単位についてだけではあるが、捕捉範囲が大きい公共部門の行政的なデータソースである。

・４のデータソースは、多くの国においていよいよ多くの統計目的のために考慮に入れられている民間部門の行政的なデータソースの典型的な例である。価格にかんして交渉する可能性に留意せよ。さらに低下する余地があるかも知れず、商業的な契約にかんする交渉の経験が有効であろう。

・５と６のデータソースは、典型的な補助登録簿と見なすことができ、捕捉範囲に制約があって、母集団全体とは属性が異なる特殊な部分集団に焦点をおいている。
・捕捉範囲、適時性、正確性、および付加価値が、各データソースについて費用効果分析の一部として考慮されるべきである。
・結果である統計にたいする利用者の要求にかんして、より多くの情報をもつことが有効であろう。それはデータソースの選択に影響するからである。経験ある統計家はしばしば、要求が少なくとも初めは、漠然としていて、利用者との進んだ対話が有効であることを認識している。明確化する課題は、つぎのことを含んでいる。
　○雇用を創出する企業に焦点をおくか、企業家数におくか。
　○要求される適時性と正確性の調整。
　○非公式の経済において操業する企業家の推定を試みることに利用者の関心があるか？　もし、そうであるならば、データソース４が、おそらくデータソース１との組み合わせにおいて必要とされるかもしれない。
・問題４と５は、ある程度トリックが仕掛けられている。それは、最初の応答において、調査が実際に必要であるか、または要求されているデータが選定されたデータソースの組み合わせによって策定される統計登録簿から直接に作成できるかどうか、明らかにされるべきであるからである。

第８章　統計調査を補完するための行政的なデータの利用

8.1　はじめに

本章は、統計調査において収集されたデータを補完するために、行政的なデータを利用するいろいろなモデルを概説する。それは、少ない経費、またはより良い精度で、さらには双方で統計を作成するためにもちいられる混合データソース法（a mixed-source approach）を紹介する。

統計データと行政的なデータの利用と連結に関連する問題の多くは、すでに第４章と第６章において取り上げたから、ここでは繰り返さない。その代わりに、本章では、統計を作成するために行政的なデータソースと統計的なデータソースの混合体のデータを利用するいろいろなモデルに焦点をあてる。

8.2　混合データソースモデル

１）母集団分割法（Split Population Approach）

本モデルでは、データを収集する目的のために、統計的な母集団が、二つ以上の部分に分割される。この方法は、第７章第３節において述べたように、オーストラリアの企業統計登録簿を保全するために利用される方法とたいへん類似している。行政的なソースのデータが、十分な精度を備えている単位について利用され、統計的なデータソースが残りの単位について利用される。

企業調査の典型的なシナリオは、調査が基本単位（大規模な企業、および複雑な構造をもつ企業）のデータを収集するために利用されるのにたいして、構造が単純な、比較的に小規模な企業にかんするデータが、税務申告から得られることである。税務データが利用される母集団の部分については、統計単位と行政単位が同一であるかもしれないし、非常に類似しており、統計的な概念や分類と、それに行政的に対応する規定が相違する影響は、最小か、少なくとも容易にモデル化することができるかもしれない。

残余の企業は代表的であって、統計の精度に個別的に大きな影響を与え、したがって正確なデータを得ることが最も重要な企業である。これらの単位は、また最も複雑な構造をそなえる単位であって、データが求められる統計単位を正確に定義するために、（第４章第５節において述べたように）しばしば一面化する（profile）することを要求する。これらの統計単位は、しばしば行政的な単位の結合体、またそれらの部分であって、雇用のようないくつかの変数が正確な総額を与えるように、単純に集計することができるのにたいして、販売や一定のその他の財務変数が与えることができないような別の変数は、そうすることができない。それらは、単純な集計では過剰計算をもたらすような一定額の単位内部の取引を含むからである。

企業調査における母集団分割法の実際的な例は、カナダ統計局が実施している企業統合調査（Unified Enterprise Survey）である。これは、いくつかの以前の調査を結合する年次企業データにたいする要求によっている。行政的なデータが、統計調査票によるデータの収集に代わって、調査対象の半数以上を占める構造が単純な企業について利用され、統計的な回答の負担を、ほとんど40％減少させている[41)]。

統計的な母集団が、国民または世帯であるところでは、学生、移民労働者または複数の居住地をも

つような特別な集団について、調査が必要な場合もあるであろう。それはすべて、行政的なデータが、とくに所在地に関連して、十分に更新できない、または正確でない単位が考えられる例である。

先行する章において幾度か述べたように、不法移民や非公式の経済において操業している企業のような、行政登録簿によって捕捉できない単位について考慮されなければならない。統計調査がそのような集団に限って利用されるかもしれないし、ある構成部分については推定が必要であるかもしれない。したがって、要求される統計を作成するために、第３のデータソースが導入される。このモデルは、下記の図8.1に例示されている。

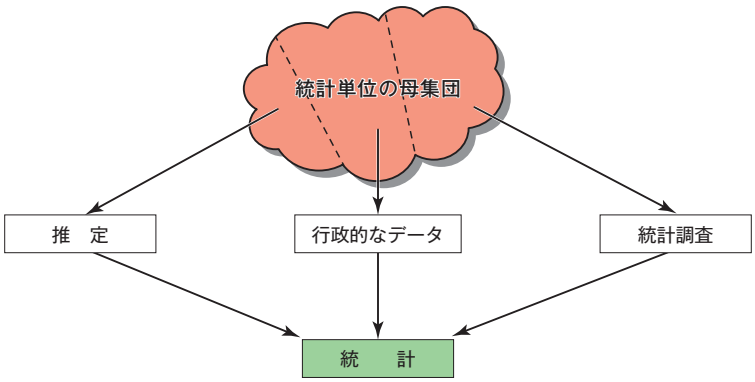


図8.1　母集団分割モデル

２）データ分割法

この方法では、統計単位の母集団とデータにたいする要求が同一である、例えば、母集団は特定の国に生活しているすべての人であって、データにたいする要求は、人口センサスに要求される通常の一組の変数である。上述の母集団分割法におけるように、母集団の一部分について、すべての変数を提供する代わりに、データ分割法では、行政的なデータソースが母集団全体について、いくつかの変数を提供するために利用される（行政的なデータソースが母集団のある部分について、いくつかの変数を提供する第３の方法も考えられる）。

したがって、データ分割法はデータを収集するために必要とされる調査票や面接の数を減少させないが、それぞれの調査票や面接において収集されるデータの量を少なくする。それは、通常、多くの変数が要求される大規模で、複雑なデータ収集に、したがって例えば、人口センサスに最も適している。行政的なデータと調査データは、統計のために利用されるデータセットを作成するために、個別の単位ごとにまとめられる必要がある。

41）詳細についてはマリー　ブロデュー（カナダ統計局）の「企業統合調査（UES）における税務データの利用」参照。http://unstats.un.org/unsd/economic_stat/Moscow_workshop/Canada%20-%20Use%20of%20tax%20data%20in%20the%20UES-E.pdf

データ分割法は、次章において述べられる登録簿にもとづく統計システムに移行する期間に、しばしばもちいられている。典型的には、統計的なデータ収集による変数は、多くの調査期間を経て、対応する行政的なデータソースからの変数によって置き換えられる。下記の表8.1は、この過程を、フィンランドにおける人口・住宅センサスのデータソースを紹介することによって表示している。

表8.1　フィンランド人口・住宅センサス（1960～2000）におけるデータ分割法

	1960	1970	1980	1990	2000
人口学的データ	Q	Q/R	R/Q	R	R
経済データ	Q	Q/R	Q/R	R/Q	R/Q
教育データ	Q	Q	R	R	R
世帯・家族データ	Q	Q	R	R	R
住宅データ	Q	Q	Q	R	R
営業用建物データ	Q	Q	R	R	n/c
建物データ	Q	Q	Q	R	R
夏季用別荘データ	Q	Q/R	Q/R	R	R

注記：記号説明
Q：統計調査票
Q／R：行政登録簿に補完された統計調査票
R／Q：統計調査票に補完された行政登録簿
R：行政登録簿　　n/c：収集せず
（出所）本表は、論文「統計目的のための登録簿と行政データベースの利用 ── フィンランド統計局の最良の活動 ──」付録２の要約版。
http://unstats.un.org/unsd/EconStatKB/KnowledgebaseArticle10169.aspx

３）事前記入済み調査票

この方法は、実際には、統計調査票が、統計単位にかんするデータを収集するために、なお利用されているデータ分割法の特殊な事例であるが、その調査票はできるかぎり行政的なデータを利用して事前に記入され、回答者は必要ならば、このデータを点検し、訂正することだけが求められる。図8.2は事前記入済み調査票の抜粋であって、イギリスの企業統計調査からの例である。

事前記入済み調査票は、主要な３つの利点をもつ。

・それは、データ提供者の時間を節約することによって、その負担を軽減するために役立つ（点検と訂正がデータの検出と記入より早いと仮定して）
・それは、行政的なデータの精度を点検する。
・それは、とくに行政的なデータの捕捉範囲が他のモデルに要求されるほど完全でないところでは、たいへん弾力的である。

主な弱点は、回答者が事前記入のデータを点検しないでそのまま受け入れる、または誤りを訂正するために、時間をかけないことによって、偏りが引き起こされる危険性である。

図8.2　事前記入調査票による調査からの抜粋

2．　あなたの勤務地の事業活動（説明の注5を見て下さい）
われわれの記録では、あなたの事業活動はつぎのとおり

木製家具の製造

2a.　もし、あなたの現在の事業活動が異なるならば、それを下記の欄に完全に記入して下さい（単語の間に一つの空欄をおいてマスに1字記入）

4）非回答者にたいする行政的なデータの利用

この方法は、データソース分割モデルとデータ分割モデルの双方の変形である。本法では、統計調査が、なおデータを収集する基本的な方法である。しかし、統計調査は、いろいろな程度の非回答を被り、それは、標本調査の処理の効率や調査結果の精度に影響する。非回答には、典型的には二つの形態があって、どちらかの一つをとる。単位にかんするデータが得られない「単位の非回答」であるか、部分的な回答は与えられるが、いくつかの事項についてはデータが空白である「事項にかんする非回答」であるかである。

非回答を処理することは、統計機関にとって、たいへん経費がかかる。それは、一般的には、欠けているデータを回収するために、郵便や電話による接触を繰り返し行うからである。この処理は、通常、「回答の追い駆け（response chasing）」として知られ、たいへん資源が費やされる傾向がある。

より廉価な代替法として、データが一定の期限までに提供されないならば、とくに調査結果にとって重要でないときには、単位（例えば、企業調査における小企業）について、データが行政的なデータソースから得られるか、導かれることが決定されるかもしれない。それは最も重要と考えられる単位に焦点を絞って、回答の追い駆けのために資源が投下されることを許容し、調査データによるより、行政的なデータを利用することによって、偏よりが最小化されることを意味する。それは、また調査結果の適時性を改善することに役立つ。精度に関連するいかなる問題についても、そうであるように、経費と、次元が異なる精度との間の妥協は避けられない。

また、行政的なデータは、ときには連結されるデータファイルのために、欠けている調査データを補記するための基礎として利用される[42)]。

42）例えば刊行書『所得参加計画調査の再構築』の第3章におけるアメリカセンサス局法を参照。 http://www.nap.edu/catalog/12715.html

５）推定のための行政的なデータの利用

標本調査が統計データを収集するために利用されるときには、しばしば推定技術を利用することが必要であって、（比率ではなく）母集団の全体値が要求されるときには、とくにそうである。したがって、母集団から標本抽出がなされなかった部分の数値を推定するためには、いくつかの基礎的な枠組みが必要である。この処理のために、ときどき標本を抽出するために利用される調査の枠組みから変数を得ることができるが、いくつかの場合には、推定過程における補助変数のように、行政的なソースのデータを利用することによって、正確さを高めることができる[43)]。実際には、この方法の多くの例が、小地域の推定値を改善するために行政的なデータを利用することにかかわっている[44)]。

8.3　考察を進めて

複数のデータソースを利用する、いかなる複雑な統計処理システムにおいても、メタデータの役割、とくに特定の項目のデータソースにかかわるメタデータのそれを考慮することは重要である。それは、データ項目を得た方法にしたがって、いろいろな過程（予期しない将来の過程を含めて）を経て、いろいろ方法で処理されることを許容する。データソースにかんする情報はまた、しばしば有効な精度指標であって、統計結果の精度水準にかんする決定を行うために役立てられる。

統計データと行政的なデータを混合して利用することは、とくに行政的なデータの捕捉範囲や精度が、統計的なデータの収集がまったく中止されるほどには、十分高くないと思われるときには、混合利用それ自身が最終点と考えられるかもしれない。それはまた、表8.1に説明したように、登録簿にもとづく統計システムへの漸次的な移行過程における一段階と見ることもできよう。

いずれにせよ、それは少なくとも、外部のデータ供給者に全面的に依存することや一般国民との接触が失われること等の不利益を避けると同時に、行政的なデータを利用する（経費節約を含む）いくつかの利点が実現される。それは、統計データと行政的なデータの精度を比較する可能性を与え、統計家を行政的なデータの利用に親しませ、処理精度を高める新しい技術を開発することを許容する。

このような理由から、データソースの混合利用法は現在、純粋な登録簿にもとづく統計システムよりもっと一般的であるが、時間とともに行政的なデータにたいする信頼度が高まり、その利用が拡大し、さらに利点が実現されて行くであろう。重心がさらに行政的なデータの方向に振れて行くときには、次章において記述するような登録簿にもとづくモデルに転換するか否かを考慮することが、ついには必要になるであろう。

43）例えば、『リトアニアの所得年次データをかくするための行政的なデータースの利用』参照。http://home.lu.lv/~pm90015/workshop2006/papers/Workshop2006_22_Slickute_Sestokiene.pdf

44）例えば、『アメリカ地域社会調査における小地域推定のための行政記録の利用』参照。http://www.fcsm.gov/99papers/mcf.html

第９章　登録簿にもとづく統計システムに向けて

9.1　はじめに

第８章の終わりに述べたように、行政的なデータが、統計登録簿を開発し、維持し、そしてまた、統計調査を補完するためにもちいられるならば、論理的につぎの段階は、そのような登録簿と統計調査を連結する方法を考察し、登録簿にもとづく統計システムに向かって進展することである。このような方途は、おもにスカンディナビア諸国の統計機関によって、しばしば登録簿にもとづく人口センサスの実現を初動的な焦点として、発展させられている。

純粋に登録簿にもとづく統計システムは、(特別な領域、または一組の領域にかんする）すべての統計が、おもに、二つ以上の連結した統計登録簿に結合されている行政なデータソースから作成されるシステムとして定義できるであろう。実際に、そのような純粋に登録簿にもとづく統計システムは、比較的に稀であって、小規模の統計調査がしばしば、精度を評価するために、また特別な変数や母集団の部分にかんする捕捉問題を克服するために必要である。したがって、より実際的な方途は、「登録簿にもとづく統計システム」という用語を、主として、連結した統計登録簿のなかに編制された行政的なデータにもとづくシステムを表現するために、もちいることである。

本章は、登録簿にもとづく統計システムに移行するいくつかの問題を簡潔に考察する。それは、国際連合欧州経済委員会刊行の『スカンディナビア諸国における登録簿にもとづく統計システム』[45] に含まれている本課題にかんするもっと詳しい研究と重複するというより、それを補うことを目的としている。本書は、人口・社会統計に焦点を絞った最良の活動を精査し、いくつかのスカンディナビア諸国の専門家によって準備されており、本課題にかんする信頼すべき研究と見なされるべきである。

9.2　実効性

登録簿にもとづく統計システムは、少なくとも短期的には、すべての国にとって、またはすべての領域の統計にとって実効的ではない。それは、このシステムを開発し、実現する実効性が、政策や社会経済的な基盤にかんする多くの前提条件に依存しているからであって、そのいくつかは、これまでの章において、いろいろ文脈のなかで述べられた。良好な登録簿にもとづく統計システムが成り立つ基本的な前提条件は、つぎのとおりである。

- 適切な行政的なデータソースの存在：目的母集団にかんする包括的な行政登録簿が必須である。例えば、不法移民や非公式の経済において活動する企業のような未登録の単位が多く存在すると、それは登録簿にもとづいて有効な統計を作成することを非常に困難にする。
- (データ) 取得の容易さ：行政的なデータソースが第３章に述べたいろいろな体制のもとで、統計家にとって容易に利用できなければならい。それは、統計家にとって、データ譲渡を容易にする様式において保管されていることを要求する。

45）http://www.unece.org/fileadmin/DAM/stats/publications/Register_based_statistics_in_Nordic_countries.pdf を参照

・共通の識別子の存在：第６章では、複数のデータソースに存在する単位に共通の識別番号が、絶対的に不可欠ではないことが示されたが、それは、データソース間の連結を相当に促進し、したがって、登録簿にもとづく統計を作成する効率を非常に高める。

・国民の受容度：第４章第２節において議論したように、政府省庁の内部におけるデータ連結とデータ共有にたいする一般国民の態度は、行政的なデータを統計目的のために利用することができる程度を決定する基本的な要因である。データ共有の効率性と個別単位に関連するデータ保護にたいする懸念を調整することは、国々の文化と伝統によって結果は異なるが、しばしば激しい論争を引き起こす原因である。いくつかの国においては、登録簿にもとづく統計システムの構想が現在、国民の大部分にとって受け入れられていないようにおもわれる。

これらの前提条件が整っていないならば、登録簿にもとづく統計システムを短期的な方途として考えることは、明らかに実効性がない。しかしながら、このモデルは、長期的な目標として有効であって、必要な前提条件を確立していく漸進的な変更計画を続けていくことによって、達成することができよう。スカンディナビア諸国の経験は、これらの国の登録簿にもとづく人口センサスの実現がほぼ20年を要したことが典型的であるように、長期的な計画の重要性を強調している。

9.3　包括的なモデル

第７章第３節は、行政データを統計登録簿において利用するモデルを提供するかぎりにおいて、登録簿にもとづく統計システムにかんする議論を含んでいた。この章の図7.4は、登録簿にもとづく統計システムの包括的なモデルを表示していたが、それは行政的な（データ）の投入だけに焦点を当てていた。下記の図9.1は、統計的な投入と産出を含むモデルに改めている。それには、つぎのように、二つの基本的な特徴がある。

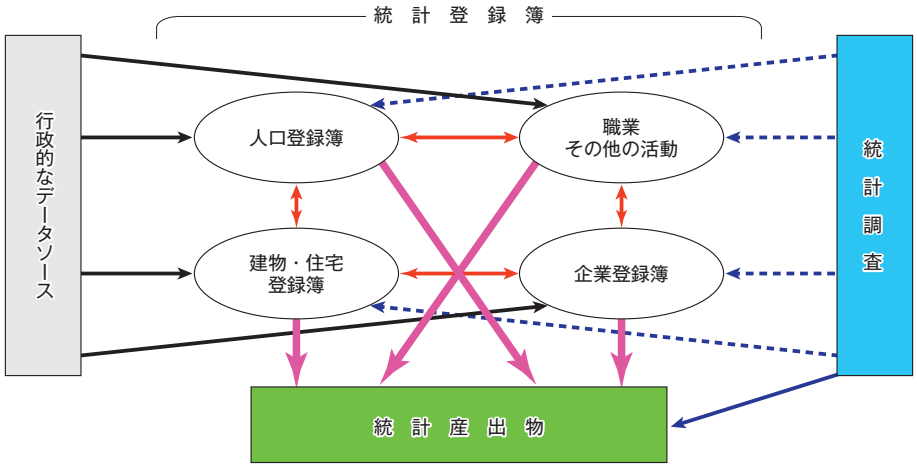


図9.1　登録簿にもとづく統計システム── 包括的なモデル

注記　職業・その他の活動の統計登録簿はこの国家版では常に存在するとは限らない。

・基本的な統計登録簿が連結されていること。また別の、もっと特殊な統計登録簿があるかもしれないが、明確さを保つために、図に示されていない。
・行政的なデータソースと統計的な産出物である調査データとが調整されていること。調整を決定する明確な規準はないが、行政的なデータソースが主要な情報源であると見通しておくことが合理的である。

9.4 要約

統計目的のために、行政的なデータを大量に利用することが考慮できるときには、登録簿にもとづく統計システムが究極の目標であることは、明白である。多くの国においては、それはたいへん遠い、おそらくずっと達成できないような目標であるようにおもえる。しかし、必要な前提条件を創出するために、漸進的に進展する戦略的な計画を採用することによって、この目標に徐々に近づいて行くことはできよう。

浜砂　敬郎〔九州大学名誉教授〕
西村　善博〔大分大学経済学部　教授〕