

ISSN 0385-2148

研究所報

No.45

フィンランドのビジネスレジスター

2015年3月

法政大学

日本統計研究所

はじめに

法政大学日本統計研究所では、現在、研究所助成による研究プロジェクトとして、「ビジネス・レジスターの国際比較研究」を遂行している。また、この間本研究所員ならびに客員研究員をコアメンバーとして科学研究費補助金による研究事業「政府統計データのアーカイビングシステムの構造と機能に関する国際比較研究」（基盤研究 B：代表者森博美）、「ビジネスレジスターによる企業動態統計の展開」（基盤研究 C：代表者菅幹雄）等を遂行してきた。本書は、このうちの酢が研究プロジェクトの研究成果を取りまとめたものである。

本研究所の所員である森と菅は、2011 年 9 月にフィンランド統計局を訪問し、同局経済動向部が、当時ビジネス・レジスターの構築を計画していたモンゴル統計局職員を対象に開催したセミナーに参加させてもらう機会を得た。その際に、レジスターベースの統計システムとしてわが国にも広く知られている北欧諸国におけるビジネス・レジスターに関して初めて体系的な情報提供を受けることができた。

その後本研究所では、ビジネス・レジスターをテーマにした国際ワークショップを開催することになり、これまでに 7 回のセミナーを開催している。このうち 2012 年には第 1 回目のワークショップとして「ビジネス・レジスターの新たな挑戦－北欧諸国の事例」（1 月 31 日）を、また 2014 年には第 6 回として「ビジネス・レジスターデータの応用例：フィンランド統計局のビジネス・レジスターデータを用いた分析」（11 月 19 日）と二度にわたりフィンランド統計局職員を招聘しわが国における事業所母集団データベース（日本版ビジネス・レジスター）の整備担当者と理論面、実務面の討議を行っている。本書の森論文は、現地調査とこの間のワークショップによって得た同国のビジネス・レジスターに関する情報に基づき、その概要を解説したものである。

また、菅を代表者とする研究プロジェクトではフィンランド統計局から 2002 年のビジネス・レジスターのマイクロデータの提供を受け、それに基づき研究を進めてきている。本書に収録されている菅論文、宮内論文、宮川論文は、いずれも同局から提供を受けたマイクロデータによる分析結果である。

ところで、わが国では、2009 年 4 月に施行された改正統計法（平成 19 年法律第 53 号）に事業所母集団データベースの整備条項（第 27 条第 1 項）が盛り込まれたのを契機にその整備に向けての取り組みがようやく開始された。4 年近い準備期間を経て 2013 年 1 月にはシステムの開発が完了し、同年 6 月から政府の統計作成のための年次フレームの提供が開始されている。

フィンランドでは欧州統計局（Eurostat）頭とも連携しつつ、LKAU と呼ばれる場所と事業内容をベースにした統計単位を基礎としたレジスター、さらには企業、企業グループレジスターの構築を既に完了しており、2013 年からは YTY と呼ばれるビジネス・レジスターの次世代型展開としてビジネス統計のデータ統合（data integration）に向けた取り組みを開始しており、この分野での欧州各国での取り組みを牽引する役割を果たしている。ちなみに、ここでいうビジネス統計データの統合とは、レジスターを骨格情報（コアス

テム)として公的統計をマイクロベースで相互にリンクするものである。その他にも同国のビジネス・レジスターでは、経緯度情報がすでに標準装備され、建築着工申請フォームにその情報の記載が義務づけられ、ビジネス・レジスターに自動的に搭載される仕組みになっている。さらにはビジネス・レジスター収録情報に基づくビジネス統計データの補完に関しても意欲的な取り組みが行われている。

このような意味でフィンランドのビジネス・レジスターに関して本書で取り上げられている諸話題は、わが国における経済時計の今後の展開方向を考える上で多くの示唆的な内容を持っていると考えられる。

本書がわが国の公的統計のこれからの展開にとっていくらかでも寄与できれば幸いである。

2015 年 3 月

法政大学日本統計研究所

フィンランドのビジネスレジスター

目 次

フィンランドのビジネス・レジスター	森 博美	1
フィンランド統計局のビジネスレジスターを用いたプロファイリング対象の選定の実験	菅 幹雄	27
フィンランド2002年ビジネスレジスターの補完実験	宮内 環	39
ビジネスレジスターの観光統計への応用	宮川幸三 Ossi Nurmi	63

フィンランドのビジネス・レジスター

森 博美(法政大学経済学部)

はじめに

1980年代から2000年代初頭にかけて、海外の政府統計機関ではビジネス・レジスターが、センサスに代わって、企業や事業所といった経済単位を調査対象とする統計調査のための母集団情報を提供する調査基盤情報として整備される。現在ではビジネス・レジスターは、多くの途上国も含めて、標本抽出のためのフレームとして本格稼働している。一方、わが国では、2007年の改正統計法(法律第53号)の第27条に「事業所母集団データベースの整備」が盛り込まれたことを契機に、ビジネス・レジスターの整備に向けての取り組みがようやく開始されたところである。

筆者がビジネス・レジスターの研究に取り組むようになったのは、イギリスの国家統計局(ONS)がその整備に取り組んできたIDBR(Inter-departmental Business Register)の存在を2000年代初頭に知ったことがその契機となっている。なお、IDBRの構造や機能については、法政大学日本統計研究所の『統計研究参考資料』No.86、2004年10月で詳しく紹介した。イギリスのIDBRを皮切りに、フランスのSIRENE(Les sources statistiques sur les entreprises)、オランダのビジネス・レジスターなどいくつかの国について、ビジネス・レジスターの整備状況、その構造や機能に関する文献研究や各国政府統計機関での担当者から直接聞き取りによる調査を実施してきた。

その過程でまず明らかになったことは、各国のビジネス・レジスターが共通して、(a)行政情報を主たる情報源としてビジネス・レジスターが整備されており、(b)それから得られる企業や事業所といった統計単位について、独自調査あるいは電話照会、ウェブや新聞等のメディア閲覧という手段によってその現状確認(これはプロファイリング(profiling)と呼ばれている)が行われ、(c)このような作業を経て母集団情報が経常的に更新され、(d)構築されたデータベースが標本調査のフレーム情報として活用されているという事実であった。これらはいずれも、統計調査に基づく統計作成といったこれまでの政府統計の在り方から捉えた場合、従来は見られなかった全く新たなタイプの統計業務である。

その一方で、各国のビジネス・レジスターの多様性もまた明らかになった。それは、各国の統計をとりまく実情の相違の反映でもある。それには、集中型・分散型といった国の統計機構の在り方だけでなく、その国の経済規模の大小、政府統計の作成にあたっての統計調査への依存度の違いなど様々な要因が関係している。好むと好まざるとにかかわらず、統計制度の存在形態がその国のいわば統計的風土によって制約される。その意味では、これからわが国で構築されることになるビジネス・レジスターも、自ずと日本型ビジネス・レジスターとならざるを得ないであろう。

わが国において持続可能な形でのビジネス・レジスターのシステム設計を行う場合、利用可能な行政情報とその情報特性、調査環境、秘密保護に対する国民の意識など、それはわが国固有の統計を取り巻く諸条件を踏まえたものでなければならない。そこでは、海外の事例はあくまでも参考情報に過ぎない。とはいえ、各国でどのような背景の下に特徴的なシステム設計が行われているかといった点に関する多様な知見は、これから具体的にシステム設計に取り組む際の有益な参考情報としての意義を持っている。

筆者は、2011 年 9 月にフィンランド統計局において、同国のビジネス・レジスターに関するインタビュー調査を実施した。周知のように、フィンランドは、レジスター・ベースの統計システムを持つ北欧諸国に属する。その点では、今日なお統計調査が政府統計作成における主要な情報源となっているわが国とはいわば対極に位置する。また、人口約 540 万人、法人数も約 32 万と経済規模もわが国とは大きく異なる。

今回、フィンランドを調査の対象国として選定したのは、次のような事情からである。これまで筆者自ら行った調査から、イギリス、フランス、それにオランダのビジネス・レジスターの相互の位置関係がおぼろげながら浮かび上がってきた。そこでは、イギリスと比較すればフランスが、さらにこれら 3 か国のビジネス・レジスターの中ではオランダのそれが、行政情報の統計への活用という点で最も環境が法制度的に整備されているように思われる。オランダ人がバイキングを祖先に持つこととこのようなビジネス・レジスターの特性とを直結させるのはいかにも短絡であろう。しかし、多様な各国のビジネス・レジスターの全体的位置づけを与えるためには、オランダのビジネス・レジスターの構造や機能の中に垣間見ることのできるレジスターを基礎にした統計システムにおけるその特徴を、一方の典型としておさえておくことが、その全体像を描く際に重要であると思われる。

北欧諸国は、今日、レジスター・ベースの統計システムを持つ国として、統計の世界の中で独特な位置を占めている。そのような国でビジネス・レジスターがどのように整備され、維持更新され、利用されているかは、他の諸国におけるビジネス・レジスターの特徴を読み取るうえで、有益な比較材料、すなわち、統計調査への依存度が相対的に大きい諸国におけるビジネス・レジスターとの対比的な意味を持つ。今回、フィンランドのビジネス・レジスターを取り上げたのは、このような問題意識からである。もちろん、レジスター・ベースの統計システムとして一括されている北欧諸国においても、それぞれの国情を反映して政府統計は多様な展開をしているものと推察される。この点の検討は今後の課題としたい。

以下本稿では、まず同国におけるビジネス・レジスター整備に至る経緯を紹介し、その作成機構、ビジネス・レジスターの構成と機能、所収の統計単位、維持更新のための情報源とその特徴、収録変数、ビジネス・レジスターの利活用等について述べてみたい。

1. ビジネス・レジスターの整備

本節ではまず、フィンランドにおけるビジネス・レジスターの今日に至る歩みを概観しておく。

フィンランドでは、1952 年に最初の経済センサスが実施された。しかし、同国における経済センサスは、1964 年を最後にその後実施されていない。また、1960 年代には、税務当局が取引税登録 (Turnover Tax Register: TTR) の整備に着手し、1968 年にはすでにビジネス・レジスターが作成されている。なお、当初のビジネス・レジスターの更新は、2 年周期とされていた。なお、このビジネス・レジスターでは、取引税の徴収単位としての企業あるいは事業所に近い概念とされる地域活動単位 (Local Kind of Activity Units: LKAU) が収録の統計単位として用いられていた。

1970 年代には、新たに税務当局が提供する月次の賃金登録が、当初、年 2 回の頻度でビジネス・レジスターの更新に利用できるようになった。さらに、税務当局では、1980 年代には、企業の損益勘定の登録を開始した。1984 年からは全ての雇用企業が、また 1989 年からは、経済活動を営む全企業が、ビジネス・レジスターに収録され、更新も年次更新へと改められた。なお、1980 年代後半にはビジネス・レジスターの対象範囲が公的部門にも拡張され、1985 年からは中央政府の諸

機関が、また 1990 年からは地方政府の各組織が新たに単位として収録されることになった。また、1995 年には新たに企業集団 (Enterprise Group :EG) がビジネス・レジスターの収録単位として追加された。なお 1991 年には、税務当局によるビジネス ID 法が成立し、それ以降、フィンランド統計局がこの ID 発行の責任を負うこととなった。

フィンランドでは 1995 年に税制が改正され、それまでの取引税 (Turnover Tax) が廃止され、付加価値税 (VAT) が新たに導入された。この税制改正はビジネス・レジスターの作成にも影響を及ぼすことになった。すなわち、最終消費だけでなく企業間で行われる個々の経済取引がインボイスによって把握されることになり、新たに VAT データに基づき月次での取引額を統計的に把握することが可能となった。

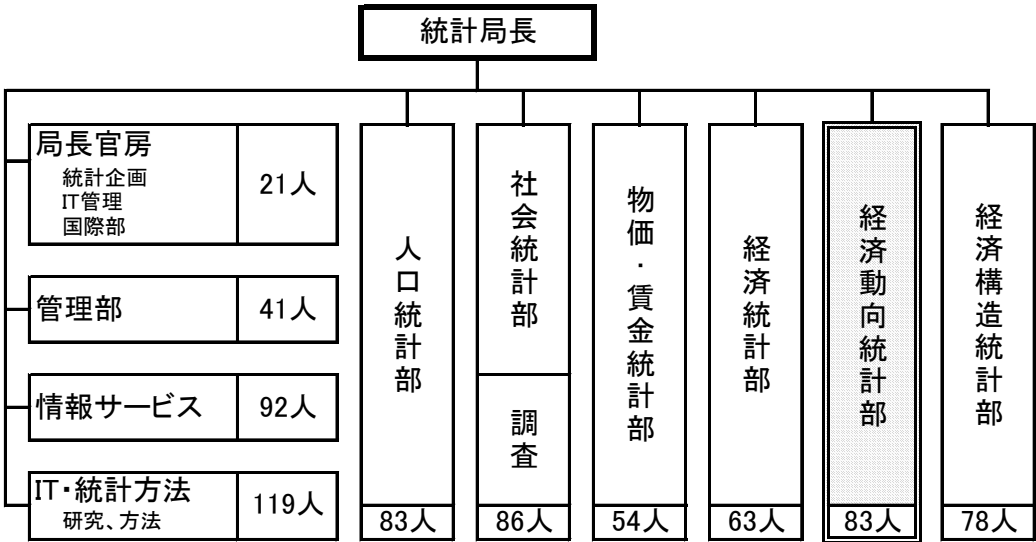
2001 年に税務当局のビジネス ID と特許登記庁 (National Board of Patents and Registration: NBPR) は、企業情報システム (Business Information System: BIS) と呼ばれる共同管理のデータベースを立ち上げ、それまで別々に付与していた ID を統合することによって統一ビジネス番号が導入された。

2. ビジネス・レジスターの作成機構

フィンランドでは現在 18 の政府機関や組織が政府統計を作成している。政府統計作成の中心機関が、フィンランド統計局 (Statistics Finland) である。フィンランド統計局そのものは常勤職員数が 840 人 (調査員を除く) と、わが国の政府統計組織規模と較べれば比較的小規模である。しかし、同国の人口が 500 万人である点を考慮すれば、その相対的な規模は大きいといえる。

図1に示したように、フィンランド統計局は、6つの分野別統計部門と情報サービス、研究部門、それに管理部門等からなっている。

図1 フィンランド統計局の組織構成



〔出所〕Tuula Viitaharju (2011b) p.7

このうち、ビジネス・レジスターは、製造業、建設業、貿易、サービス、運輸業、観光業関係の統計を所管する経済動向部に属するビジネス・レジスター課 (business register unit: BRU) が、その整備や維持更新の任務を遂行している。(図2参照)

図2 各統計部門が所管する統計

人口統計部	社会統計部	物価・賃金統計部	経済統計部	経済動向部	経済構造統計部
・人口 ・ジェンダー平等 ・人口センサス ・死因 ・雇用 ・司法・犯罪 ・選挙 ・教育 ・建物、住居、居住条件 ・地理情報	・労働市場 ・所得、消費 ・労働条件 ・文化、生活時間 ・成人教育 ・マスメディア、ICT利用	・住宅物価 ・消費者物価 ・国際物価比較 ・賃金・報酬、労働費用 ・生産者価格、費用	・国民経済計算 ・金融統計 ・地方経済統計 ・農林企業財務 ・消費者調査	・製造業 ・建設業 ・貿易 ・サービス ・運輸 ・観光 ビジネス・レジスター課 (30人)	・企業財務諸表 ・産業・地域別統計 ・品目別統計 ・サービス貿易 ・科学技術 ・環境 ・エネルギー ・情報社会 ・温室ガス

〔出所〕Tuula Viitaharju (2011b) p.8

ビジネス・レジスター課には現在 30 人の常勤職員が配置されている。そのうちの約半数にあたる 10-20 人は、ビジネス・レジスターのために提供あるいは独自に収集したデータのチェック、更新（調査・行政情報のミス等のチェック）業務に従事している。また、5 人の職員が大企業からの情報収集を担当している。企業内の構造把握などビジネス・レジスターに収録される統計単位の現状確認を行うプロファイラーは、現在 3 人に過ぎない。この他にビジネス・レジスター課には、3 人の情報サービス担当官が配置されている。

フィンランド統計局では、ビジネス・レジスターに提供される行政情報の品質は高いという認識がこれまで支配的であったことから、行政情報を含めたプロファイリングの必要性が自覚されるようになったのは比較的最近のことである。このため、数年前から試験的にプロファイリングが行われており、今後その体制が強化される見通しとのことである。なお、各業務への配置人員数は固定的ではなく、業務の閑繁を勘案して弾力的に運用されている。

3. ビジネス・レジスターの概念図

フィンランドのビジネス・レジスターでは、行政機関から提供される行政情報、統計調査を実施することによって得られる調査結果データ、その他民間データプロバイダー事業者等から購入するデータを用いることによって、企業や事業所等の新規開業や廃業、移転などを把握し、その結果がビジネス・レジスターに格納される。このようにして経常的に維持更新されるデータベースは、ビジネス・レジスター作成データベース (Business Register Production Database、以下 PRoDB と略称) と

呼ばれている。

PRoDB には、法人や事業所などビジネス・レジスターに収録されている統計単位の ID や所在地情報だけでなく、従業者や売上高規模など種々の変数が格納されている。このため、日常的に維持、更新されるビジネス・レジスターは、様々に利用されている。その詳細については後述するが、その機能は大きく次の二種類に区分される。(a)年次フレームとして統計調査実施のための基盤情報の提供を行う機能、(b)ビジネス・レジスター統計作成機能、がそれである。

このうち(a)調査実施のための基盤情報の提供に関してビジネス・レジスターは、サンプルフレームとしての機能を持つ。このためにフィンランド統計局では、PRoDB から月次でデータを移管することによって、ビジネス・レジスター・サービス・データベース(Business Register Service Database、以下 SeDB と略称)を作成している。

調査環境の悪化に伴う統計調査による客体把握度の低下は、長期にわたって母集団情報の提供機能を担ってきたセンサスの意義を次第に浸食することになる。このようなセンサスに代わって調査客体となる統計単位をプロファイリングという存否確認によって経常的にその把握を行うビジネス・レジスターは、資本金や従業員さらには売上高規模という統計単位の基本的な属性情報を持つことによって、標本抽出にとって必要な情報を提供することができる。

SeDB は、標本抽出に際しての標本の層別や抽出標本のローテーションといった標本の抽出方法や調査結果の母集団への復元のための乗率といった標本調査に関係する種々の情報だけでなく、絶対尺度としての母集団情報を保有することによって、欠損値や欠損レコードそのものの補定(補完)をも可能にする。またビジネス・レジスターに調査履歴情報を持たせることによって、調査が特定の客体に過度に集中しないように、いわゆる survey holiday 制度の導入によって回答負担を平準化することもできる。

ビジネス・レジスターは、種々の調査データ、行政データを個体(マイクロ)ベースでデータ統合したデータベースでもある。それは、企業や事業所といった経済単位の活動を捉える構造統計や速報統計作成のための情報源としても有効である。かつてのような周期的に実施されるセンサス結果に基づく母集団名簿の整備と異なり、ビジネス・レジスターは、日常的に維持更新される。さらにそれはいわゆる縦断型の母集団情報を提供することから、企業の新規開業や企業分割・統合、あるいは休廃業といったビジネスの動態面の把握に有効である。

ビジネス・レジスターが保有する情報に基づいて作成される統計は一般にビジネス・レジスター統計と呼ばれるが、ビジネス・レジスター統計の作成のためにフィンランド統計局では、PRoDB から年次ファイルを作成している。なお、財務諸表情報の提供に時間を要することから、ビジネス・レジスターの年次更新は 10 月末現在で行われている。

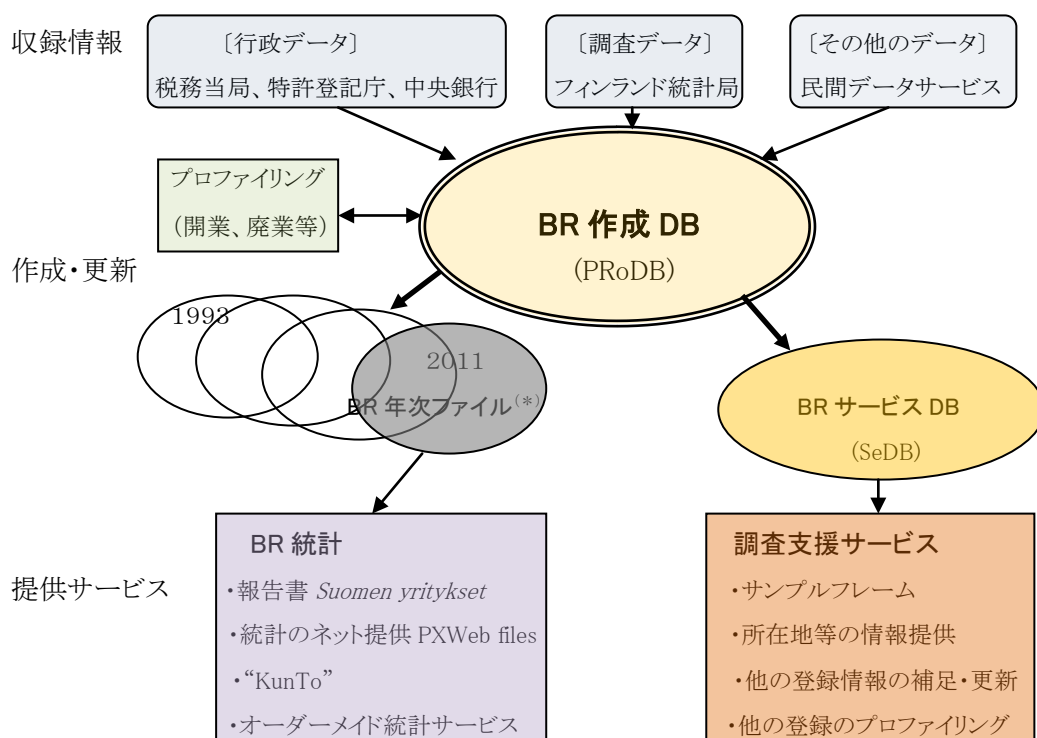
ビジネス・レジスターの年次ファイルは『統計年鑑(Suomen yritykset)』の編纂に用いられている。また、この他にも、ビジネス・レジスターが保有するデータに基づいて、四半期、さらには月次の速報統計が作成され、提供されている。

近年、EU 域内の厳しい雇用情勢の中で、雇用創出という実践的政策課題を受けて、OECD を中心に“Entrepreneurship Indicator Programme”という国際プロジェクトが進行中である。このようなプロジェクトを含め、企業の動態面に着目した企業動態分析(business demography)は、縦断面のデータベース機能をも持つビジネス・レジスターの整備、更新をその情報的基盤とすることによってはじめて成立するものである。

図3は、フィンランドにおけるビジネス・レジスター作成の情報的根拠とその機能について、

PRoDB、SeDB、それに年次ファイルの関係を含めて要約的に示したものである。

図3 フィンランドのビジネス・レジスターの情報源と機能



(*) 対象事業所・企業: 6 か月以上活動し 1.5 人以上の雇用又は売上高 9636EUR/年以上

〔出所〕Tuula Viitaharju (2011b) p.13, Jukka Pakola (2011b) p.7 (一部、加筆等の修正)

4. ビジネス・レジスターの収録単位

欧州連合の規則 (REGULATION (EC) No 177/2008) は、その付録資料に法人 (legal unit: LeU)、企業 (enterprise: ENT)、地域単位 (local unit: LU)、企業集団 (enterprise group: EG) をビジネス・レジスターが持つべき統計単位として、各単位が持つべき変数名とともに掲げている¹。なお、フィンランドのビジネス・レジスターでは、この他に地域活動単位 (the local kind-of-activity unit: LKAU) が、統計単位として追加されている²。

¹ EU 規則第 [C] 177 号: “REGULATION (EC) No 177/2008 OF THE EUROPEAN PARLAMENT AND OF THE COUNCIL of 20 February 2008” は、統計目的のためのビジネス・レジスターの共通枠組みを規定している。同規則の付属文書には、ビジネス・レジスターに収録される統計単位とそれが持つべき変数一覧を掲げている。〔本稿巻末の付属資料を参照〕

² 「共同体の生産システムの観察と分析のための統計単位に関する 1993 年 3 月 15 日付の EU 理事会規則 (“Council Regulation (EEC) No 696/93 of 15 March 1993 on the statistical units for the observation and analysis of the production system in the Community”) には、これ以外にも、制度単位 (the institutional unit)、活動単位 (the kind-of-activity unit:

(1) ビジネス・レジスターが保有する統計単位

(i) 法人(LeU)

フィンランドのビジネス・レジスターにおいて中心的統計単位となっているのは法人(LeU)である。同国で経済活動を行っている原則として全ての事業体は付加価値税登録(VAT Register)が義務付けられており、付加価値税登録番号(VAT ID)を持っている³。単独事業所で事業活動を行う企業の場合、通例、法人と企業は一致するが、法人と企業とで登録の場所が異なるケースもある。また法人には公的組織やNPOなども含まれる。このため、法人は企業とは、統計単位としては、必ずしも同一ではない。

フィンランドでは毎年 3 万件前後の新規登録が税務当局に対して行われているが、これは同国における登録法人数の約 0.5%に相当する。なお、税務当局は、月次で新規登録情報をフィンランド統計局に提供している。

(ii) 企業(ENT)

企業は最小の法人であり、資本、労働力、技術等の配置についての意思決定自主権を持ち、財貨・サービスの生産や取引を行う単位である。それは単一あるいは複数の場所で、単一あるいは複数の種類の事業活動を遂行する。多くの場合、法人と企業とは、組織単位としては同一である。他の多くのEU諸国と同様⁴に、フィンランドでは、企業は現在のところまだビジネス・レジスターにおける正式の統計単位とはなっておらず、目下その整備に向けての作業が進行中とのことである。

(iii) 地域単位(LU)

特定の場所(所在地)に存在し、企業ないしはその一部を構成する統計単位が、地域単位(LU)である。作業場、工場、店舗、倉庫、事務所、鉱山、貯蔵所といったものがそれに該当する。特定の場所あるいはそこを基点として経済活動が営まれ、パート就業者も含め、当該の場所において 1 人以上の者が就労することが、地域単位の条件となる。

法人の新規登録は、最低 1 つ以上の地域単位の創設を伴う。なお、一つの法人が事業活動を行う全ての場所において地域単位は所在する。

(iv) 企業集団(EG)

株式保有あるいは財務的結びつきによる企業の連合体は、企業集団(EG)と呼ばれる。企業集団は、複数の法人間の支配＝被支配関係構造を持つ。

企業集団の場合、生産、販売、収益に関する政策について、2 つ以上の意思決定センター(decision centers)を持つもの、財務管理や課税といった経営の特定の側面だけを集中するものなどその形態は多様である。企業集団は、グループ構成企業に関する選択権限を有する経済単位(economic entity)である。

KAU)、同種生産単位(the unit of homogeneous production: UHP) 地域活動単位(the local kind-of-activity unit: LKAU)、同種生産地域単位(the local unit of homogeneous production: local UHP) が掲げられていた。

³ 行政組織には、これとは別に固有の識別番号が付与されている。

⁴ EU 諸国のうち現時点で企業をビジネス・レジスターの統計単位としているのは、イギリスとオランダだけである。

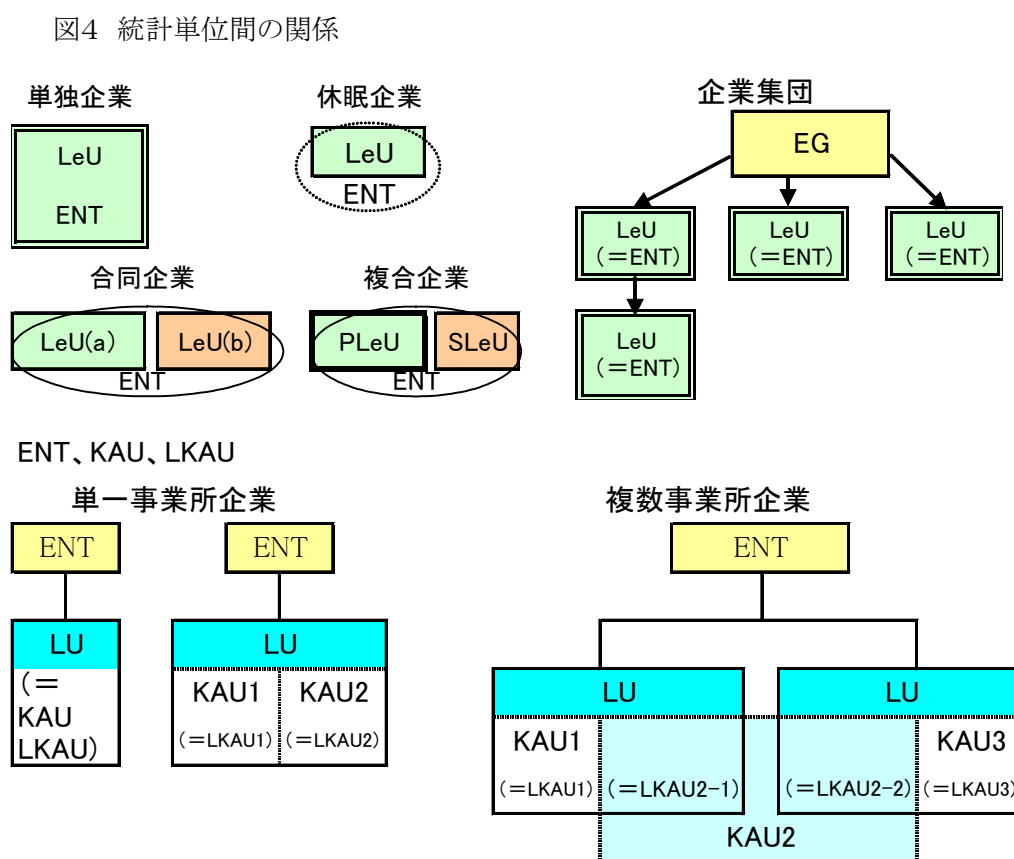
企業集団の構成にあたっては、海外の Dan & Bradstreet 社や Bureau Van Dijk 社の他、フィンランド国内の信用調査会社 Suomen Asiakastieto といった内外の民間データプロバイダーが保有する情報も活用している。しかし、こういった民間のデータプロバイダーが提供する企業集団に関する情報は一般に高価であり、しかも必ずしもお互いに整合的ではない。

(v) 地域活動単位(LKAU)

地域活動単位は、法人を産業活動の種類(NACE5 桁レベル)という観点から区分することによって構成される統計単位である。このうち、ある特定の場所で特定のカテゴリーに属する種類の活動を行うものが地域活動単位(LKAU)である。活動中の法人は、必ず一つ以上の地域活動単位を持つ。

ところで、LKAU と同様に、事業主体の活動の種類に注目した類似概念に活動単位(a kind of activity unit: KAU)がある。なお、フィンランド統計局では、複数の場所で同一の活動を行っている経済単位を、LKAUよりは統合度の高い活動分類区分によって活動単位(KAU)に編成し、統計単位化することを検討中である。

図4は、法人、企業、地域単位、企業集団、活動単位、地域活動単位の関係を図示したものである。



EU 加盟各国は、それぞれ歴史・文化的背景、取引商慣行、法制度、さらには大企業と中小企業の比重の違いといった社会経済的背景を異にする。このため、ビジネス・レジスターに収録すべき統計単位についても、EU 規則によって一応規定はされているものの、そのいずれを中心的に整備する統計単位とするかは、国によって一様ではない。ちなみに、加盟国の大半では法人がビジネ

ス・レジスターの中心的な統計単位とされている。近年、企業さらには域内外に本部を持つ多国籍企業を含む企業集団についてもビジネス・レジスターにおける統計単位として把握することが重要な課題となっている。なお、フィンランドも含め北欧各国のビジネス・レジスターに共通する特徴として、地域活動単位(LKAU)もビジネス・レジスターに収録すべき統計単位の一つとしている点を挙げることができる。

(2) 統計単位間の照合表

法人(LeU)と地域単位(LU)、地域単位(LU)と地域活動単位(LKAU)など、フィンランドのビジネス・レジスターで統計単位として採用されている単位については、単位相互間の照合表(relation table)が作成されている。照合表の記載は次のようなものである。例えばある法人が複数の地域単位を持つ場合、当該法人の識別番号(識別コード)に複数の地域単位番号(識別コード)が対応付けられており、各レコードは、法人と地域単位との関係の開始時点と解消時点情報を変数として持つ。なお、地域単位が当該法人との関係を維持していれば、解消時点記載カラムには NULL と表示されている。また、企業再編によってある地域単位が法人Aから分離され別の法人Bに移籍するケースも発生しうる。このような場合、法人と地域単位の照合表には、法人Aと当該地域単位のレコードには関係解消時点情報が記載され、また新たに法人Bと関連付けられた地域単位レコードが新設され、関係の開始時点情報と、関係の継続を示す NULL がそれぞれ記載されることになる。

現時点では照合表は二種類の統計単位間で作成、維持されている。この点についてフィンランド統計局では、企業集団(EG)、企業(ENT)、法人(LeU)、地域単位(LU)、地域活動単位(LKAU)にそれぞれつけた識別IDコード間の統一的な照合表の編成を今後の課題としている。

(3) 識別コード情報付与機関

上述したように、フィンランドのビジネス・レジスターは、法人、地域単位、企業集団、それに地域活動単位を統計単位に持つ。このうち法人については、企業情報機構(BIS)が共通識別番号を付与している。なお、この共通識別番号は税務情報に基づいていることから、法人だけでなく自営事業者に対しても発行される。

フィンランドでは、銀行口座の開設を初めとして様々な様式にこの共通識別番号の記載が求められる。事業活動展開の様々な局面で共通識別番号の提示が求められることから、そのカバレッジの精度は高いと考えられる。なお、法人の合併等については特許登記庁が管理する商業登記への報告が義務付けられており、BISは法人の動態情報も把握している。

地域単位、企業集団、それに地域活動単位については、フィンランド統計局が識別番号の付与機関となっている。

5. ビジネス・レジスターの収録変数

ビジネス・レジスターには、収録される統計単位別に、識別番号を含む名簿情報、活動区分情報、規模情報、企業等の動態情報、その他が収録されている。表1は企業、事業所(地域単位)、企業集団レコードに収録されている主な変数を例示したものである。

表1 ビジネス・レジスター収録変数例

企業レコード	(記入)	地域単位レコード	(記入)	企業集団レコード	(記入)
企業コード	*****	事業所コード	*****	企業コード	*****
名称	〇〇	名称	〇〇	名称	〇〇
所在地	〇〇	所在地	〇〇	企業集団コード	*****
郵便番号	*****	郵便番号	*****	企業集団名	〇〇
所在地の郵便局	〇〇	所在地の郵便局	〇〇	企業集団のタイプ	〇〇
市町村コード	***	電話局番	**等	企業集団の国籍	フィンランド等
市町村名	ヘルシンキ等	電話番号	*****	企業集団のタイプ2	〇〇
産業コード	*****	所在市町村コード	***	親子等の別	親/子/孫・
産業名	ホテル等	所在市町村名	ヘルシンキ等		
法的形態コード	**	産業コード	*****		
法的形態	有限会社等	産業名	ホテル等		
言語コード	*	従業員規模コード	*		
言語名	〇〇	従業員規模	階級区分表示		
電話番号	*****	当該市区町村内の従業員数	実数		
所有者タイプコード	*	位置座標情報	*****		
所有者タイプ名	外資等				
従業員規模コード	*				
従業員規模	階級区分表示				
売上高規模コード	*				
売上高規模	階級区分表示 (€)				
輸出・入者コード	*				
輸出・入業者の別	輸出者／輸入者				
従業員数	人数				
傘下事業所数	事業所数				

〔出所〕Päivi Kizywacki (2011) pp.6-7 より作成

これらの諸変数の他にもビジネス・レジスターには、統計単位によって、名簿情報としては、Fax、メールアドレスが、また規模情報としては、給与・報酬支払額が収録される。一方、動態情報としては、事業の開始/閉鎖年月日、事業の開始/閉鎖の形態、吸収/合併年月日、手放した企業/引受企業等の情報が収録される。なお、企業集団レコードには、海外子会社 (FAFs) も含め、親企業とのリンク情報、株式保有率 (%) も格納されている。またこれら以外にも、ビジネス・レジスターの各レコードは、ビジネス・レジスター調査の管理用変数、雇用主識別番号、VAT 登録番号、納税支払登録番号、輸出入業者番号、貿易統計申告コード等の情報を持つ。

6. ビジネス・レジスターの情報源

他の諸国と同様に、フィンランドのビジネス・レジスターにおいても、行政情報、調査情報、それに民間事業者が保有する情報が、その維持更新の情報源となっている。以下に、情報源のタイプ別に情報の主な特徴を紹介する。

(1) 行政情報

(i) 税務情報

フィンランドのビジネス・レジスターに収録される主要な行政情報は、税務当局から提供される税務記録情報である。すでに述べたように、税務当局は、ビジネス・レジスターのデータベース骨格情報である事業体そのものの把握⁵に関して、事業税登録、付加価値税 (VAT) 登録に係る新規事

⁵ 税務当局から提供される企業の名簿情報とは次のようなものである。

企業識別番号、所在地 (郵送住所)、産業部門コード、法人の開始/停止、法的形態、法人間の関係、名称・所在地・活動開始年月日

業体情報を月次⁶でフィンランド統計局に提供している。この他にも税務当局は、ビジネス・レジスターのデータベース上に格納される変数データについても、事業税、源泉徴収 (Pay-as-you earn: PAYE)、VAT、その他の徴税関連情報を提供している。

税務当局が管理する事業税登録は、企業の経理データを年単位で保有しており、これからは損益勘定、B/S、さらには企業間の株式所有関係といった一連の情報が得られる。また、賃金・報酬を支払う雇用主は、社会保障納付金の給与等からの天引き義務を持つ。これを制度的根拠とする源泉徴収登録は支払賃金・報酬額および雇用者数に関する包括的な所得申告データを保有しており、これらはフィンランド統計局に対して月次で提供されている。さらに、財貨・サービスを販売した事業体は、VATの支払いが義務づけられている⁷。VATの税額は取引のインボイス記載情報に従って算定される。このため、徴税業務のためのVAT登録は、事業体による取引額(売上高)情報を保有することになる。フィンランドのビジネス・レジスターは、税務当局が管理するVAT登録から月次で売上高情報の提供を受けている。

これらの他にも税務当局はフィンランド統計局に対して、年間賃金データや農林事業税データを提供している。このうち前者の情報ファイルからは個人の賃金報酬額や雇用主と雇用者の関係に関する情報が得られ、後者からは農林業における年間所得額に関する情報が得られる。

税務情報は、第一回目の提供で売上高と賃金・報酬額全体の約 9 割をカバーしている。しかし、税務情報には、情報の把握時点とフィンランド統計局にそれが提供されるまでの間に、少なからずタイムラグが存在する。例えば源泉徴収登録データは、最長で把握時点から 3 週間の、また VAT 登録データは、一般に 50 日前後、最長 75 日の遅れを伴う。

この他にも税務データには、納税下限額や免税に伴うカバレッジの問題も存在する。例えば、年間売上高€8500 未満の小企業は免税扱いとなっており、保健、社会保障、宝くじや出演料なども把握対象外である。

人的記入ミスや光学読取りミスなどもあるが、概して、税務当局から提供される情報の品質は良好であるといわれている。

(ii) その他の行政情報

税務情報以外の行政情報としては、中央銀行が外国資本による株式の保有状況を、税関が輸出入業者登録番号と貿易申告コード情報を、また特許登記庁は、企業の合併や休業、企業集団の連結財務諸表データをフィンランド統計局に提供している。なお中央と地方政府に関しては、財務省が中央政府とその職員情報を、また地方政府年金機関は、地方政府に関するそれらの情報を提供する。

フィンランドのビジネス・レジスターでは、地域単位 (LU) や地域活動単位 (LKAU) に関する各レコードは、該当する統計単位の所在地の緯度・経度情報を保有している。この情報は、建物・住宅登録 (Buildings and Dwellings Register) からビジネス・レジスターに供給される⁸。なお、地域単位の敷

⁶ それまで四半期で提供されていたが、2005 年からは月次で提供されるようになった。

⁷ 2010 年 7 月から標準税率が 23%に引き上げられた。なお、食料品については 13%、交通費については 9%の特例税率が適用されている。

⁸ フィンランドでは、事業用や居住用の建物の建設に際して提出が義務付けられている建築着工申請様式に所在地の緯度・経度情報の記載欄が設けられている。

地内に複数の建物が存在する場合には、位置情報は建物ごとにではなく、単一の緯度・経度情報がその地域単位に割り当てられている。

(iii) 行政情報に固有の問題

行政情報は行政上の業務の遂行のために収集あるいはその遂行過程で作成されるものである。従って、そのような行政情報を統計目的に使用する場合、行政情報に固有な様々な問題が存在する。

レジスター・ベースの統計システムを持つフィンランドでは、行政情報の統計への活用がすでに制度化されている。このため、行政機関側で行政情報の収集に際して、ビジネス・レジスターに配慮した報告様式の修正を行うケースもないわけではない。しかし、一般にはあくまでも行政利用を主目的に行政情報は収集・作成される。このようなことから、行政事由で把握される単位(行政上の単位)と統計単位とは必ずしも一致しない。

また、概念の定義も統計基準とは非整合的である場合もありうる。行政データの場合、行政上の制度変更が概念定義の変更に直結し、頻繁な制度改正はデータの連続性を大いに損なうことになる。更に、上にも指摘したように、行政情報の場合、把握時点と実際に統計局のビジネス・レジスター課に報告が寄せられるまでの間に、少なからずタイムラグがある。これらの事情もあり、行政情報を統計目的で使用するためには、データの調整が必要な場合が少なくない。なお、行政データはフィンランド統計局に対して有料で提供される。しかし、実際に提供機関側から請求されるのは提供ファイルの編成に関する提供機関側での作業量を反映した経費であることから、必ずしも高額ではない。

(2) ビジネス・レジスター調査情報

フィンランドは、レジスター・ベースの統計システムを持つ国のひとつとして知られている。しかしこのことは、この国の政府統計が全面的に行政情報に基づいて作成されていることを意味しない。[森(2012)]でもすでに紹介したように、同国では、大方の予想とは異なり、ビジネス・レジスターの更新のための調査を含め、多数の統計調査を実施している。

上にも述べたように、行政情報は一般に、各行政機関で計数を取りまとめ、その結果がフィンランド統計局に報告されるまでに多少の時間を必要とする。速報統計(short term statistics: STS)の作成にタイムリーな行政情報入手できない場合、統計調査を実施することによって直接データの収集を行う必要がある。

フィンランドには約 32 万の法人がある。フィンランド統計局では、このうち毎年 25000～30000 法人に対してビジネス・レジスターの情報更新のための調査(ビジネス・レジスター調査)を実施している。ビジネス・レジスター調査では、全国のトップ 2,000 社の企業内取引を、また賃金・報酬支払額についても、約 80 の活動種類別にそれを調査している。なお、フィンランドでは、50 の主要企業が、国内の全取引額の約 30%を占めている。

表2は、現在フィンランドで実施されているビジネス・レジスター調査の調査方法、調査対象(数)等を一覧的に示したものである。

表 2 ビジネス・レジスターのための主要な統計調査

調査名	調査方法	調査対象	対象数等
地域雇用統計・複数地域単位企業調査	ネット	複数事業所企業 (KAUs, 事業所)	- 年2回(1月、2月) - 年間約5000企業 - 年間約30 000事業所
複数地域単位企業調査 (BR 独自調査)	ネット	複数事業所企業	- 年1回(1月) - 年間約2 000企業 - 年間約5 000事業所
データファイル形式での複数地域単位企業調査	秘匿処理メール	複数事業所企業	- 約90企業 - 約8 500事業所
単独地域単位企業調査	ネット	全大企業および標本企業で所有者あるいは事業所数に変化のあった企業	- 年間3回 - 年間約 8 000 - 9 000企業
単独地域単位企業調査	ネット	- 層化標本(1 500企業): BRの母集団の品質指標を作成するために統計的指標に従って対象企業を選定(小規模単独事業所企業) - 所在地あるいは産業分類が変更になったと思われる小規模標本企業(1 500企業)	- 年間1回 - 年間約3 000企業
新設企業調査	ネット	- 規模による標本調査 - BRは税務当局から新設企業情報を入手	- 年間4回 - 年間約6 000企業
新設企業電話調査	電話	大規模新設企業	- 月次随時調査 - 年間約400企業
地域雇用統計・公共部門調査	ネット	中央政府機関	- 年間1回 - 400の中央政府機関 - 2 000の地方組織

[出所] Jarmo Ranki (2011) 報告資料

フィンランド統計局では、ウェブによる統計調査を 2006 年に初めて導入した。2011 年からはビジネス・レジスター情報の更新のための調査を全面的にウェブ⁹と電話による調査に切り替えた。なお、電子調査票には、企業等のIDコード、名称、郵送先住所、所在地、産業分類コード、前年の雇用者数と売上高がプレプリントされている。

ビジネス・レジスター調査の調査回答率は、全体として、複数の地域単位を持つ企業(わが国の複数事業所企業に相当)が 93%、単独の地域単位企業(同じく、単独事業所企業)のそれが 70%である。得られた回答結果の中の 10-40%はビジネス・レジスターのデータ・ベースの更新に直ちに用いられるが、残りの 60-90%については、点検のための事務的作業が必要である。点検により適切と判定されたデータについてはバッチによりデータ・ベースに収録される。なお、ビジネス・レジスター調査は、名称、所在地、売上高、従業員数など主にビジネス・レジスターの骨格 (backbone) 情報を更新するための情報収集を目的に実施されている。

ビジネス・レジスター調査は、行政情報よりも良質の情報を提供することもあるが、いくつかの課題も抱えている。

フィンランド統計局のビジネス・レジスター課では、企業側の報告担当者登録 (Enterprise Informant Register: EIR) が整備し、調査実施支援データベースとして常に更新している。しかし、法人や地域活動単位において内部の報告体制がどのように組織されているかについては、必ずしも十分な情報は持っていない。多数の地域単位を持つ企業の場合、各地域単位からの報告の取

⁹ ウェブ調査実施の際には、調査客体に対して、Logon に必要な ID 情報とパスワードが、事前に郵送される。

りまとめ作業が必要となる。また、調査によっては既存の経理データからは得られない計数の報告を求めるものもある。さらに、ERPソフト¹⁰、SAP社のビジネスアプリケーションソフトウェア、Oracleに未対応の企業の場合、調査報告に際して、企業側に多大な報告負担を求める調査もある。

企業、特に報告の提出要請が集中する大企業の報告負担を軽減するために、現在、フィンランドではいくつかの方策が講じられている。まずビジネス・レジスター側の対応としては、報告担当者登録が保有する調査履歴情報の活用がある。多くの国の政府統計機関と同様、フィンランド統計局でも、調査履歴情報を用いることによって、報告負担の均等化が図られている。

第2の対応は、報告徴集者である統計局側と報告者である企業側の人的連携の強化である。フィンランド統計局では報告担当者登録を基盤情報として双方の連携関係を日常的に更新しているのに加え、地域活動単位についても報告者側の担当者¹¹を調査別に定めている。

また、今後の課題としては、主要企業を直接訪問することによって、報告単位を確定するとともに、ERP、SAP、Oracleといったデータ管理方式を特定することで、現行のような調査ベースではなく、個々の変数ベースでデータ収集を再編成することによって企業ごとにデータの提供＝収集方式をカスタマイズすることが現在検討されている。フィンランドでは巨大複合企業は比較的限られることから、これらの企業からは直接データを収集することで、ビジネス・レジスターの主要な部分を把握するという発想がその根底にはある。

(3) その他のデータ

フィンランドのビジネス・レジスターでは、政府の行政情報や調査から得られる情報に加えて、民間のデータプロバイダーなどが保有する情報も活用されている。

フィンランド郵便会社(National Postal Company: NPC)は、郵便配達用に住所情報を維持、更新している。同社が持つ街路名、番地、郵便番号コードといった企業への郵便配達に用いる住所情報は、ビジネス・レジスターが保有する住所記載内容の点検に使用されている。同国では住所の記載がすでに全国的に標準化されており、郵便会社の住所情報は、法人や地域単位、地域活動単位といった事業体が所在する市町村名を特定するのにも使用されている。

ヨーロッパでは、Dun & Bradstreet 社や Bureau van Dijk 社が保有する企業情報が、企業集団レジスターの維持、更新に広く用いられている。フィンランドでは、両社に加え、国内の民間データハウスで格付け信用企業である Suomen Asiakastieto 社が保有するデータが、特許登記庁(NBPR)がビジネス・レジスターに提供する連結財務諸表情報とともに、同国における企業集団レジスター構築における主要情報として使用されている。

7. プロファイリングとデータの格納

(1) プロファイリング体制の整備

フィンランドのビジネス・レジスターでは、企業情報機構(BIS)が月次で提供する法人情報が、その骨格情報となっている。しかし、BISが月次で提供する法人情報には、当該法人による地域単位

¹⁰ Enterprise Resource Planning software の略称

¹¹ ここでの担当職員は、必ずしも報告の提出を行う直接の当事者とは限らない。

保有の有無、地域単位の数やそれらの所在地といった企業組織に関する情報は含まれていない。このためビジネス・レジスター課では、企業構造については、表2に示したようなビジネス・レジスター調査、ウェブや各種メディアの閲覧といった独自の方法で現状確認を行っている。

プロファイリングと呼ばれているこうした現状確認業務は、二つの側面からなる。その 1 は、地域単位や地域活動単位の存否確認であり、地域単位や地域活動単位の改廃、所在地の移転の確認といった業務がそれに含まれる。その 2 は、それらの活動内容の確認に関するものである。ここでは、地域単位や地域活動単位の産業分類(NACE)¹²あるいは制度部門(sector)による格付け、従業員数や売上高といった活動に関する属性情報の確認が行われる。なお、フィンランドのビジネス・レジスターでは、不動産業や事務管理支援といった事業所を対象とした補助的業務サービスに従事している法人は、一般にサービスを受受する顧客法人与同一の産業分類として格付けすることになっている。

フィンランド統計局での担当者からのヒアリング結果によれば、同国ではこれまで行政情報の精度に対する信頼が厚かったことから、プロファイリングが現実的課題として認識されるようになったのは比較的最近のことである。近年、フィンランドでも、規制緩和や行政予算の大幅な削減による行政取得情報の減少が、ビジネス・レジスターの整備や品質の確保という点で問題視されつつある。調査回収率の低下とともに、こういった行政情報の質の劣化は、ビジネス・レジスターの整備当事者に対して、プロファイリングの必要性を提起しているとのことである。このような中でフィンランド統計局では、経済動向部内に企業調整官(Business Coordinator)を配置し、特に企業組織が複雑な大規模複合企業の現状確認を、電話、Eメール、さらには直接訪問によって実施するという体制を整備しつつある。

(2) 収集データの点検と格納

ビジネス・レジスターに提供される情報の中には欠損値、他の統計との整合性を欠くデータ、さらには税込と税別のデータの混在といったようなケースも少なからず含まれる。このため、ビジネス・レジスターの更新にあたっては、その点検作業が所管部門の主要な業務となる。フィンランドのビジネス・レジスターでは、従業者数が欠損値となっている場合、賃金・報酬支払いデータを用いて回帰推定することで補定が行われている。

データの点検によるビジネス・レジスターの更新については、更新される情報源の種類、変数、統計単位に従って数百の手順が策定されており、その多くの作業が独自に開発したアプリケーションプログラムによってすでに自動化されている。しかしながら、実際には論理エラーをはじめ、種々の原因でこのアプリが適切に処理しきれないケースが発生しうる。このような場合にデータはエラーバッファファイルに一時的に格納される。当該企業への照会も含め、規模の大きい単位から順次、手作業によるプロファイリングを経てビジネス・レジスターに格納処理される。

¹² NACE(Nomenclature statistique des activités économiques dans la Communauté européenne)「欧州共同体における経済活動の統計的分類」の略称

8. ビジネス・レジスターの利用

(1) 調査インフラとしての利用

ビジネス・レジスターの作成データベース(Production Register DB: PRoDB)から作られるサービスデータベース(Service DB: SeDB)は、サンプルフレーム(標本抽出枠)として、フィンランド統計局が実施する統計調査の母集団情報を提供する(図3参照)。フレームとしての機能を果たすことができるように PRoDB は、調査対象となる統計単位ならびに産業分類(NACE)、従業員や売上高規模といった層化情報を保有している。

ビジネス・レジスターは、個体レコードから構成されるデータベースである。しかもその中には営業上の秘密事項に属する変数(sensitive variables)も含んでいる。PRoDB に係る秘匿性の担保ならびに整合的な更新業務の遂行のために、PRoDB へのアクセスは、経済動向部のビジネス・レジスター課職員だけにしか認められていない。一方、PRoDB から作成される SeDB については、所定の手続により許可を受ければ、フィンランド統計局の職員であれば誰でもオンラインないしバッチでアクセスすることができる。なお、このデータベースにはメタデータ情報サービスが付帯されており、利用者はデータの品質、更新スケジュールに関する情報さらには変数それ自体や処理方式についての説明情報を入手できる。

経済動向部ビジネス・レジスター課では、同部自らが実施しているビジネス・レジスター調査のために SeDB をフレーム情報として対象標本の抽出を行っている。一方、標本調査の実施を計画しているフィンランド統計局の他の部門に対しては、現在のところ同課は同様のサービスを提供してはいない。調査実施部門は、調査実施に関する申請、承認手続きを経た上で同課から SeDB そのものの提供を受け、自ら対象標本を抽出する仕組みになっている。なお、ビジネス・レジスター課から提供される SeDB は有料である。しかし、この場合も同課は申請部門に対して、ファイルの作成費相当額を経費として請求するだけである。また、SeDB の提供を受けた調査実施機関では、調査業務終了後、ファイルを削除することが義務付けられている。

フィンランドの統計システムは完全な一極集中型ではなく、統計局以外にも政府統計の作成を行う機関が複数存在〔森(2011)〕する。統計局外の調査実施者についてはフレームとして SeDB を使用できる体制には現在のところまだなっていないが、フィンランド統計局では将来は外部機関へのフレーム情報の提供を構想している。

(2) ビジネス・レジスターに基づく統計作成

ビジネス・レジスターは、標本調査のためのフレームとして他の統計調査実施の支援機能を持つと共に、それ自体が保有するデータに基づいた統計作成機能も保有している。このようにして作成される統計は一般にビジネス・レジスター統計と呼ばれる。ビジネス・レジスター統計には、他の統計調査と同様に、構造統計と速報統計とがある。

(i) 構造統計

構造統計は PRoDB から年 1 回の頻度で定期的に更新される年次ファイルに基づいて作成される。年次ファイルは更新作業完了後 1 か月以内にフィンランド銀行に提供されているほか、それから作成される統計は、『企業・事業所年報』(*Suomen yritykset*)として公刊され、ウェブ提供も行われている。なお、年次ファイルに収録されているのは、6 ヶ月以上活動し 1.5 人以上の雇用者または年間売上高 96.36€以上の企業・事業所のみである。

(ii) 速報統計

売上高のうち商品取引データは、速報が 45 日後、確報が 75 日後、といったスケジュールで公表されている。一方、サービス取引データについては、確報が 75 日後に公表される。フィンランド統計局以外の省庁が作成している製造業と建設業に関する速報データは、75 日後の公表となっている。なお、このような統計の公表とは別に、欧州標本 (European sample) に指定されている品目¹³については、商品取引速報データが 30 日後、サービス取引速報データが 45 日後に Eurostat に提供されている。一国全体の支払い賃金・報酬額データは 45 日後に公表される。

企業の開廃業状況に関する集計結果については、四半期ごとにビジネス・レジスターのデータに基づいて作成、公表され、それは企業動態統計として Eurostat にも定期的に報告されている。この他にも、フィンランド銀行と税務当局に対しては、月次で速報データ (short term statistics: STS) が提供されている。なお、OECD が Entrepreneurship Indicator Programme (EIP) との関連で提唱しているガゼル (Gazell) と呼ばれる急成長企業に関する統計も、現在フィンランド統計局において策定に取り組んでいるとのことである。

(3) 対外部利用者サービス

(i) オーダーメイドデータ処理サービス

前掲の図 3 にも示されているように、ビジネス・レジスターの PRODB から作られる年次ファイルは、オーダーメイド統計サービス (tailored service) という方式によって、フィンランドにおけるビジネスの動向等の分析のために、局外の利用者、すなわち、他省庁、地方自治体、各種業界団体、企業や研究者等にも広く提供されている。

フィンランド統計局がこの方式によるビジネス・レジスターデータの加工処理を受託する変数としては、売上高、輸出入額、賃金・報酬額、従業員数といったものがある。オーダーメイドデータ処理サービスの対象となるこれらの変数については、月次、四半期、そして年次での更新データを持つ。なお、データの加工処理にあたっては、例えば集計表の作成の場合、地域区分は全国 6 区分 (州レベル) と 20 区分 (県レベル)、また産業分類については NACE による 700 区分による作表が可能である。この他にもフィンランド統計局では、年次ファイルによる企業集団別の集計依頼も受託している。

(ii) 企業等に関する基礎情報の公開

フィンランドでは、ビジネス・レジスターが保有し恒常的に更新している情報の一部、特に企業等の識別番号、所在地、従業員数といった基礎的的属性情報¹⁴については、企業等に関するデータの

¹³ EU とユーロ圏では、域内の加盟国について必要な統計を時宜的に整備する目的で、欧州標本企画 (European sample scheme: ESS) が作られている。このうち速報統計に関する ESS は、2005 年 7 月 6 日付の規則第 1158 号で、その施行細則は、2007 年 6 月 14 日の規則第 657 号に規定されている。 ([http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Glossary:European_sample_scheme_\(ESS\)](http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Glossary:European_sample_scheme_(ESS))) 参照)

¹⁴ 公表が義務付けられた情報

(i) 雇用主、自営業者、企業、機関

- ①事業識別番号、その有効期間、法的形態、名称、産業、言語コード、現在地及び郵便住所、公開の連絡先情報、②所有者のタイプ、③所在地と活動の場所、④売上高の階級区分、⑤総

取り扱いを規定した 2004 年統計法第 18 条を根拠規定として公開され、同国における社会共通の情報インフラとして一般の利用に供されている。

(iii) リサーチラボ

(ii) で紹介したように、フィンランドでは企業等に関する基礎的属性情報の一部の変数は公表されている。とはいえ、フィンランドのビジネス・レジスターが保有する企業や事業所に関する情報の中には、それぞれの事業活動と深く結びついた事業者にとって営業上の秘密に属する情報 (sensitive data) も当然ながら含まれている。周知のように、個人や世帯統計については、カテゴリーの統合、トップ (ボトム) コーディング、スワッピング、丸めや誤差の付加といった様々な匿名化措置を施した匿名標本データ (いわゆるマイクロデータ) が作成され、研究あるいは民間での営利目的での使用に供されている。こういった個人や世帯データの匿名化に有効な種々の技法も、企業データの匿名化には必ずしも有効ではない。

そこで、ビジネス・レジスターが保有する個人情報について、利用ニーズへの対応と秘匿性の確保との両立を図る制度的仕組みとして、フィンランド統計局では局内にリサーチラボ (research laboratory: RL) と呼ばれる施設を開設している。所定の手続きを経て利用が承認された利用者は、統計局内 (in house) で、しかも限定された変数についてのみその使用が認められている。

9. 今後に向けての課題

フィンランドでは、統計調査データだけでなく行政情報についても共通識別番号が付与されている。そのため、調査結果あるいは行政機関から新たに提供される個体レコードは、ビジネス・レジスター内の既存のレコードと容易にしかも高い確度で紐づけることができる。統計局内でも、例えば産業分類を所管する NACE 部門と制度部門担当グループとは密接な連携関係を維持している。

ビジネス・レジスター課では、特に大規模複合企業を中心に、報告側の担当者 (contact person) との照会業務に従事する局側の担当者 (business coordinators) を決め、日常の業務にあたっているほか、調査実施のたびに企業報告者登録 (Enterprise Informant Register: EIR) を常に維持更新している。この ERR には、報告側の担当者名のほか、電話、Fax、Eメールといった調査や照会に必要な様々な連絡情報が格納されている。

フィンランドでは 2014 年からの供用開始を目指して、ビジネス・レジスターを基盤データベースに持つ政府統計のアーカイビングシステムの構築に向けてのプロジェクトが目下進行中である。ビジネス・レジスター・データウェアハウス (BR data Warehouse) 計画がそれである。データウェアハウスでは、各統計単位に付与されたユニークな識別番号をリンクキー変数として、統計調査や行政機関から提供される個体レコードが相互に個体ベースで統合されることになる。ビジネス・レジスターを識別番号をリンクキーとするリレーショナルなデータベースとしてデータウェアハウスを設計する

従業者数と市町村別の従業者数、⑥貿易への従事の有無、⑦VAT支払い義務、雇用主の活動、源泉所得税登録、⑧(企業集団の場合には) 集団との関係

(ii) 地域単位 (事業所)

①事業所識別番号、②活動の期間、③名称、④産業分類、⑤所在地、⑥公開の連絡先情報

ことで、フィンランドにける政府統計データのアーカイビングシステムは、新たなデータ統合 (data integration) の段階へと歩みを進めることになる。

大半の場合、統計単位としての法人と企業とは事実上一致する。しかし、所有に基づく法人概念と支配関係に基づく企業概念とは、本来異質なものである。フィンランドではこれまで法人がビジネス・レジスターの中心的な統計単位として扱われてきており、企業という統計単位は、いまだ構築途上にある。この点についても、データウェアハウスが稼働開始予定年である 2014 年までに、企業を正式な統計単位として採用できるための作業が現在、統計局内で進行中とのことである。

ビジネス・レジスターについては、他にもいくつか課題が残されているが、その一つが、市町村といった地方政府組織に関する組織的な情報収集体制の構築である。また、企業集団との関連では、多国籍企業のプロファイリングによる欧州企業集団レジスター (European enterprise group register: EGR) の構築や国境によって分断された企業 (Truncated enterprise) をビジネス・レジスターにどう取り込んでいくかといった課題が残されている。

むすび

レジスター・ベースの統計システムを持つフィンランドでは、税務当局と特許登記庁が共同管理する企業情報システム (BIS) が提供する法人情報を中心的統計単位として、ビジネス・レジスターが構築され、維持更新されている。

BIS が月次でビジネス・レジスター課に提供する法人登録情報は、VAT ならびに源泉徴収所得税情報に基づいて作成されることから、制度的に適用対象外となる事業体を除き、自営事業者も含めて包括的にカバーしている。そのため同国では、法人という統計単位の存否それ自体の確認に係る現状確認業務 (プロファイリング) は、基本的に不要である。

その一方で、BIS が提供する法人登録情報は法人内の組織に関する情報を保有していないという点で、ビジネス・レジスター構築への利用面で本質的な情報制約を持っている。特に複数の地域単位を持つ法人については、それぞれの所在や開廃等の確認が、ビジネス・レジスターの維持更新のためには不可欠の作業となる。ビジネス・レジスター課では、これらの確認作業に必要な情報を、ビジネス・レジスター調査、ウェブやメディアの閲覧、電話やメールによる照会によって収集している。世界各国の政府統計システムを鳥瞰した場合、フィンランドは一般に、レジスター・ベースの統計システムを持つ国の一つとして類別される。このようなフィンランドにおいて、行政情報を補完する目的で、ビジネス・レジスターの維持更新のために多くの調査が実施されていることが、今回の考察から明らかになった。

企業や地域単位 (事業所) といった経済主体が展開する事業は、NACE 等の統計分類のレベルにもよるが、必ずしも 1 種類とは限らない。大分類→中分類→小分類→細分類といった位階的分類が細分されるに従って、複数の異なる分類に格付けされる事業を同一の企業あるいは地域単位 (事業所) において営む主体が卓越することになる。

複数の産業部門にまたがる事業に従事している企業や地域単位 (事業所) について、これまで統計では、売上高 (取引額) 等が最も大きい部門に産業分類上は格付けされてきた。細分類から小分類、中分類、大分類へと産業分類のレベルが統合される過程で、各企業は相対的にシェアの大きな活動が属する産業部門にその全活動が帰属処理されることになる。このような積み上げ結果は、品目別集計と産業別集計との結果数字の乖離に見られるような、統計的反映結果を作り出し

てきた。

支配概念である企業といういわばトップダウン方式での把握ではなく、企業や事業所において展開されている活動の産業分類的な意味での種類(activities)の側面に注目し、いわばボトムアップ方式で統計を再構成するのが活動種類による統計単位の設定である。フィンランドのビジネス・レジスターでは、地域活動単位(LKAU)を原初的な統計単位として設定し、それからデータ処理によって事後的にある種同様な活動を行う単位として(a Kind of Activity Unit: KAU)を編成するとしている。上述したような産業格付けに係る統計の宿命的とも思われる問題を直視し、実態を統計技術的に最大限可能な範囲で反映した統計の作成を data integration によって実現するという発想が、地域活動単位(LKAU)、さらにはそれを踏まえて想定されている活動単位(KAU)その背景にはある。

統計技術的に既存の情報からKAUをどのように構成するかについては、今後検討すべき多くの課題が残されている。しかし将来、企業や地域単位(事業所)を一度地域活動単位(LKAU)に解体しその活動そのものの種類という観点から再構成することが仮にできるとすれば、このようにして作成される統計は、より精度の高いGDP統計の推計だけでなく、部門別の生産性の適正な比較評価にとっても、不可欠の要件と考えられる。その意味では、企業や事業所といった経済活動主体をどのようにビジネス・レジスターにおける統計単位として位置付け、今後それらの体系化、相互連関をどう図っていくかについて、フィンランドにおけるビジネス・レジスターの取り組みは、多様な統計編成を可能にするデータの置き方そのものについての重要な検討課題をわれわれに提起しているように思われる。

〔参考文献〕

森 博美(2012)「レジスター・ベースの統計システム下の統計調査」経済統計学会政府統計研究部会ニュースレター No.16

Antti Santaharju (2011), "The Importance of the Enterprise Group register as an Integral Part of the Business Register of Statistics Finland."

Jaakko Salmela (2011), "Linking administrative and survey data – employment variable for enterprise and establishments in Finnish Business Register."

Jarmo Ranki (2011), "Data Collection, Sources, Channels and Methodologies, Data Checks, Data Control and Data Validation Methodologies of Business Register in Statistics Finland."

Jukka Pakola (2011a), "Handling of Mergers."

Jukka Pakola (2011b), "The Use of BR: Statistics and Sampling Frame."

Jukka Pakola (2011c), "Design and Creation of Statistical Units in Business Registers vs. EU Regulations."

Jukka Pakola (2012), "Business Register in Finland."

Marko Tuomiaro (2011), "Development of Questionnaire Sets to Big Corporations, responding to Numerous Inquiries."

Marko Tuomiaro (2012), "The Euro Groups Register."

Päivi Kizywicki (2011), "The Use of Business Register for Providing Chargeable

Services.”

Rina Tammisto (2011),”Geographic Information System (GIS), Use of Spatial Data in Statistics Finland.”

Sami Saarikivi (2011),”Revision Project of the Business Register (BR) and Business Statistics in 2009-2014.”

Tarja Hatakka (2011),#Profiling and Work with Large Enterprises.”

Tuula Hausmann (2011),”Statistics Finland and Co-ordination and Co-operation with External Stakeholders.”

Tuula Viitahaju (2011a),”New trends in Developing Business Registers.”

Tuula Viitaharju (2011b),”Business Register in Finland.”

Tuula Viitaharju (2011c),”Co-operation Between Statistics Finland and Administrative Data Providers – case: Tax Authority.”

〔謝辞〕

本稿は、2011 年 9 月 5 日～7 日にフィンランド統計局（経済動向部）において実施した調査の席上配布された各資料（参考文献参照）と当日の各説明担当官との質疑、さらには 2012 年 1 月 31 日に日本統計研究所が開催した第 1 回国際ワークショップ「ビジネス・レジスターの新たな挑戦-北欧諸国の事例」における報告者との質疑に基づいている。フィンランド統計局経済動向部ビジネス・レジスター課の上席統計官諸氏、とりわけ Jukka T. Pakola と Marko T. Tuomiaro 両氏にこの場でお礼を申し述べておきたい。

本論文は、平成 23 年度日本学術振興会科学研究費補助金基盤研究 B「政府統計データのアーカイビングシステムの構造と機能に関する国際比較研究」（課題番号 22330070）ならびに挑戦的萌芽研究「GPS 情報の活用による公的統計の新たな展開可能性に関する多角的研究」（課題番号 40105854）による研究成果の一部である。

〔初出〕本稿は、『オケーショナルペーパー』（法政大学日本統計研究所）No.31 2012 年 3 月を再掲したものである。

【付属資料】

“REGULATION (EC) No 177/2008 OF THE EUROPEAN PARLAMENT AND OF THE COUNCIL of 20 February 2008”に掲げられた統計単位別の変数一覧

([O] は任意変数、[C] は条件付き変数であることを示す)

1. 法人

〔識別情報〕

- 1.1. 法人識別番号
- 1.2a. 名称
- 1.2b. 最も詳細な所在地（郵便コードを含む）
- 1.2c. [O] 電話・Fax 番号、e-メールアドレス、データの電子収集許可情報
- 1.3. 付加価値税登録番号又はそれがいない場合は他の行政識別番号

〔動態特性情報〕

- 1.4. 法人設立年月日又は事業活動を営む自然人としての公式承認年月日
- 1.5. 法人が 3.3 に記載された企業の一部であることをやめた年月日

〔経済的/制度的情報〕

- 1.6. 法的形態

〔他登録とのリンク情報〕

法人が記載され統計目的のために使用できる情報を持つ関係登録情報

- 1.7a. 欧州共同体規則 No 638/2004 (1)に基づく域内事業者登録情報、関税ファイル又は域外事業者登録情報
- 1.7b. [O] 貸借対照表情報（公表義務事業者）及び貿易収支ないし対外直接投資登録情報及び農業事業者登録

＜企業集団に属する法人に追加的に求められる変数＞

〔企業集団とのリンク情報〕

- 1.8. その単位が属する全居住企業集団/境域で分断された企業集団の識別番号
- 1.9. 全居住企業集団/境域で分断された企業集団への加入日
- 1.10. 全居住企業集団/境域で分断された企業集団からの離脱日

〔単位管理情報〕

居住法人の支配関係はトップダウン方式(1.11a)又はボトムアップ方式 (1.11b)で記載。直接又は間接支配という第 1 次関係だけが各単位について記載（これらを組み合わせることで全ての支配関係が得られる）

- 1.11a. その法人によって支配されている居住法人の識別番号（複数）
- 1.11b. その法人を支配している居住法人の識別番号
- 1.12a. その法人が支配している非居住法人の登録国、識別番号、名称および所在地
- 1.12b. [C] その法人が支配している非居住法人の VAT 登録番号
- 1.13a. その法人が支配している非居住法人の登録国、識別番号、名称および所在地
- 1.13b. [C] その法人を支配している非居住法人の VAT 登録番号

〔単位所有関係情報〕

[C] 居住法人の所有はトップダウン方式(1.14a)又はボトムアップ方式(1.14b)で記載情報の記載と株保有の閾値は、行政情報の利用可能性に依存。推奨されている閾値

は、直接所有の 10%以上。

1.14a. [C] その法人が所有する居住法人の(a)識別番号および(b)株式保有(%)

1.14b. [C] その法人が所有する居住法人の(a)識別番号および(b)株式保有(%)

1.15. [C] その法人が所有する非居住法人の(a)登録国、(b)識別番号又は名称、所在地、および VAT 登録番号および(c)株式保有(%)

1.16. [C] その法人を所有する非居住法人の(a)登録国、(b)識別番号又は名称、所在地、および VAT 登録番号および(c)株式保有(%)

2.地域単位

〔識別情報〕

2.1. 地域単位識別番号

2.2a. 名称

2.2b. 最も詳細な所在地（郵便コードを含む）

2.2c. [O] 電話・Fax 番号、e-メールアドレス、データの電子収集許可情報

2.3. その地域単位が所属する企業(3.1)の識別番号

〔動態特性情報〕

2.4. 活動の開始日

2.5. 活動の最終的停止日

〔経済的/制度的情報〕

2.6. 主活動コード（NACE 4 桁レベル）

2.7. [C] （もしある場合には）二次的活動（NACE 4 桁レベル）これは調査対象となった地域単位のみ

2.8. [O] それが属する企業の副次的活動として地域単位が行う活動の有無

2.9. 雇用者数

2.10a. 雇用数

2.10b. [O] フルタイム換算雇用数

2.11. 地理的位置（緯度経度情報）

〔他登録とのリンク情報〕

2.12. [C] （関係する登録が存在する場合）地域単位が記載され統計目的のために使用できる情報を持つ関係登録情報

3. 企業

〔識別情報〕

3.1. 企業識別番号

3.2a. 名称

3.2b. [O] 郵便住所、e-メールアドレス、URL

3.3. その企業を構成する法人の識別番号

〔動態特性情報〕

3.4. 活動開始日

3.5. 活動の最終的停止日

〔経済的/制度的情報〕

3.6. 主活動コード（NACE 4 桁レベル）

3.7. **[C]** (もしある場合には) 二次的活動 (NACE 4 桁レベル) これは調査対象とな
った企業のみ

3.8. 雇用者数

3.9a. 雇用数

3.9b. **[O]** フルタイム換算雇用数

3.10a. 3.10b に規定された以外の売上高

3.10b. **[O]** 農林水産狩猟業、公務、防衛、義務的社会保障、雇用者を持つ家計および
域外組織の売上高

3.11. 欧州勘定体系による制度部門とその副次部門

<企業集団に属する企業に関する追加収集情報>

〔企業集団とのリンク情報〕

3.12. 企業が所属する全居住企業集団/境域で分断された企業集団(4.1)の識別番号

4.企業集団

〔識別情報〕

4.1. 全居住企業集団/境域で分断された企業集団の識別番号

4.2a. 全居住企業集団/境域で分断された企業集団の名称

4.2b. **[O]** 居住企業集団/境域で分断された本部の郵便住所、e-メールアドレス、URL

4.3. 全居住企業集団/境域で分断された企業集団本部の条件付識別番号 (居住企業集
団本部である法人の識別番号と同じ)

支配単位が事業活動に従事していない自然人である場合には条件付き (**[C]**) (記載は
利用可能な行政情報に依存)

4.4. 企業集団のタイプ

1. 全て居住企業

2. 居住企業が支配する境域で分断された企業集団

3. 非居住企業が支配する境域で分断された企業集団

〔動態特性情報〕

4.5. 全居住企業集団/境域で分断された企業集団結成日

4.6. 全居住企業集団/境域で分断された企業集団の解散日

〔経済的/制度的情報〕

4.7. 全居住企業集団/境域で分断された企業集団主活動コード (NACE 2 桁レベル)

4.8. **[O]** 全居住企業集団/境域で分断された企業集団の二次的活動 (NACE 2 桁レベ
ル)

4.9. 全居住企業集団/境域で分断された企業集団の雇用者数

4.10. **[O]** 連結売上高

<多国籍企業集団についての追加収集情報 (4.4 のタイプ 2 および 3)>

第 11 条に規定された多国籍企業集団の情報が報告されるまでの間は、4.11 および 4.12a
の記載は任意 (**[O]**)

〔識別情報〕

4.11. 世界企業集団の識別番号

4.12a. 世界の企業集団名称

4.12b. **[O]** 世界の企業集団本部の登録国、郵便住所、e-メールアドレス、URL

4.13a. 企業集団本部が居住者の場合、世界の本部識別番号(企業集団本部である法人の識別番号と同じ)。企業集団本部が非居住者の場合はその登録国。

4.13b. **[O]** (非居住企業集団の場合) 世界の集団組織の識別番号、名称、所在地

〔経済的/制度的情報〕

4.14. **[O]** 世界の雇用者数

4.15. **[O]** 世界の総売高

4.16. **[O]** 世界の意思決定センターの所在国

4.17. **[O]** 企業と地域単位が所在する国 (複数)

フィンランド統計局のビジネスレジスターを用いた プロファイリング対象の選定の実験¹

菅 幹雄（法政大学経済学部）

1. はじめに

筆者が代表者を務める研究グループ（平成 24～25 年度科学研究費補助金基盤研究(C)、研究課題名「ビジネスレジスターによる企業動態統計の開発」）ではフィンランド統計局より、2002 年についてのビジネスレジスターのマイクロデータの提供を受けた。

ビジネスレジスターの作成・メンテナンスには、税務記録情報など行政記録情報が使用されているため、その情報の秘匿は法律で厳しく規定されており、ビジネスレジスターのマイクロデータを一般に提供している国はほとんどない²。その中であってフィンランドはビジネスレジスターのマイクロデータを一般に有償で提供している希有な事例である。

本稿はフィンランド統計局が提供しているビジネスレジスターのマイクロデータの構造と分布を観察した上で、このマイクロデータを用いたプロビット・モデルによるプロファイリング対象の選定の実験について報告する。

2. フィンランドのビジネスレジスターのマイクロデータの構造と分布

ビジネスレジスターのマイクロデータは企業データ（Data on enterprises）と事業所データ（Data on establishments）の二つのファイルから構成されている。企業データに収録されているのは表 1 の一番左の列、事業所データに収録されているのは表 1 の真ん中の列に示された項目である。

事業者識別番号が企業データと事業所データの両方にあるので、二つのファイルはマッチング可能である。このほかに両方のデータにある項目としては、市町村コード、産業分類、従業者数の規模階級、開設年月日がある。企業データにしかない項目には、従業者数、売上高の規模階級、法的形態（経営組織）、所有者の区分、雇用主／非雇用主の別、付加価値税納税者／非納税者の別、輸入／輸出の有無である。一方、事業所データにしかない項目には、事業所コード、郵便番号、市町村の従業者数、経度・緯度がある。

雇用主／非雇用主の別、付加価値税納税者／非納税者の別、輸入／輸出の有無、法的形態（経営組織）が企業データにしかないのは、企業が法的単位（legal unit）であるからである。すなわち、雇用者の税金・社会保険等の源泉徴収、付加価値税の納税、EU域内取引の輸入に関する付加価値税の仕入税額控除は法的単位で申告されるからである。

¹ 本稿の内容については、フィンランド統計局のユッカ・パコラ（Jukka Pakola）氏、慶應義塾大学の宮内環氏、総務省統計局基本構造統計課の永島勝利氏、高部勲氏からコメントをいただいた。ここに記して感謝したい。

² 米国では研究者からビジネスレジスターのマイクロデータの提供の要望が強くあり、合成データ（synthetic data、擬似マイクロデータのこと）による提供を模索している Kinney, Reiter[2013]。

売上高の規模階級が企業データにしかないのは、フィンランド統計局のビジネスレジスターでは、売上高を付加価値税の情報に基づいて把握しているが、付加価値税は企業単位で申告されるのであり、事業所単位ではないからである。このように行政記録の単位によって、存在する情報の種類が異なる。

表 1 フィンランド統計局のビジネスレジスターのマイクロデータに収録された項目

企業データ Data on enterprises	事業所データ Data on establishments	内容
Business ID	Business ID	事業者識別番号(8 桁)
	Establishment code	事業所コード(9 桁)
	Postal code	郵便番号(5 桁)
Municipality	Municipality	市町村コード(3 桁)
Industry, TOL 2002	Industry, TOL 2002	産業分類(5 桁)
Personnel		従業者数
Size category of personnel	Size category of personnel	従業者数規模階級
Size category of turnover		売上高規模階級
Start of operation	Start of operation	開業・開設年月日(企業は年月まで)
Legal form		法的形態(経営組織)
Type of owner		所有者の区分
Employer activity		雇用主／非雇用主の別
VAT activity		付加価値税納税者／非納税者の別
Importer/Exporter		輸入／輸出の有無
	Number of personnel in the municipality	市町村の従業者数
	Co-ordinates	経度・緯度

従業者数規模階級は企業と事業所でやや異なっており、事業所では 200 人以上がまとめられている(表 2)。なお企業については従業者数の情報があるので、これとは異なる規模階級を設定することが可能である。

表 2 従業者数規模階級

企業 (単位: 人)	事業所 (単位: 人)
1 = 0-4	1 = 0-4
2 = 5-9	2 = 5-9
3 = 10-19	3 = 10-19
4 = 20-49	4 = 20-49
5 = 50-99	5 = 50-99
6 = 100-199	6 = 100-199
7 = 200-499	7 = 200-
8 = 500-999	
9 = 1000-	

売上高規模階級情報は事業所データにはなく、企業データだけにある。表 3 は企業デー

タの売上高規模階級を示している。売上高の一番小さい規模階級 1 は 0～999 ユーロであり、1 ユーロ約 120 円で換算すると 0～約 12 万円しかない。表 4 の法的形態別売上高規模階級別企業数を見るとその多くは有限責任会社 (Limited company)、自然人 (Natural person)、住宅営団 (Housing corporation) である。自然人は我が国の個人企業に相当する。このように売上高が少ない企業は、おそらく有限責任会社の場合は活動実態がほとんどなく、自然人 (個人企業) の場合は趣味であろう。その次の規模階級 2 (1,000～39,999 ユーロ) は約 12～約 480 万円、規模階級 3 (40,000～99,999 ユーロ) は約 480～1200 万円である。表 4 を見るとその多くは自然人 (個人企業)、有限責任会社、合名会社であり、零細な自営業者を多く含んでいると考えられる。ちなみにフィンランド統計局のビジネスレジスターにおいて「企業」 (Enterprise) と呼ぶ場合は、このように会社企業以外も含んでいる。

表 3 売上高規模階級

企業 (単位: ユーロ)
1 = 0 - 999
2 = 1 000 - 39 999
3 = 40 000 - 99 999
4 = 100 000 - 399 999
5 = 400 000 - 1 999 999
6 = 2 000 000 - 9 999 999
7 = 10 000 000 - 39 999 999
8 = 40 000 000 - 199 999 999
9 = 200 000 000 -

表 4 法的形態別売上高規模階級別企業数

法的形態	売上高規模階級 (単位: ユーロ)									合計
	0-999	1,000-39,999	40,000-99,999	100,000-399,999	400,000-1,999,999	2,000,000-9,999,999	10,000,000-39,999,999	40,000,000-199,999,999	200,000,000-	
11 自然人	1579	43888	30281	13130	1400	476	2			90756
12 遺産	39	71	33	16	8	4				171
13 納税義務がある法人	119	277	151	136	34	3	1	1		722
14 合名会社	78	1602	2275	3150	729	88	4			7926
15 破産財団	9	3	8	7	5	1				33
21 合資会社	181	4649	6643	9298	2905	302	8	1		23987
22 共同所有輸送会社						3				3
31 有限責任会社	3770	14589	17797	31499	21360	6868	1706	610	144	98343
32 相互保険会社	19									19
33 貯蓄銀行	40									40
34 年金財団	35									35
35 住宅営団	1165									1165
41 協同組合	336	428	202	209	87	54	31	24	8	1379
51 財団	40									40
52 任意団体	48									48
53 相互損害賠償保険組合	115									115
54 経済団体	708	65	33	37	15	6	1		1	866
62 国営企業	1					1		1	4	7
63 公共団体、公社、公団	1									1
72 その他の宗教団体	1									1
90 その他の法的形態	138	37	46	134	199	142	53	18	2	769
合計	8422	65609	57469	57616	26742	7948	1806	655	159	226426

注) 濃い網掛けは 1 万社以上、薄い網掛けは 1 千社以上。

表 5 は事業所数規模階級別従業者数規模階級別企業数を示している。わが国では統計に

おける事業所の要件に「財又はサービスの生産と供給が、人及び設備を有して、継続的に行われていること」があるので、必ず事業所数≤従業者数の関係が成り立つはずである。フィンランドでも、一部例外があるようであるが、ほぼ成り立っているようである。相対的に高い規模階級を見ると、同じ従業者数規模階級でも事業所数がかなりばらついている。カテゴリーデータである2項目間の関連性の強さを表わすクラメールの連関係数³が0.244と低いことから、従業者数規模が大きい企業が事業所数規模も大きいとは限らないことがわかる。

表5 事業所数規模階級別、従業者数規模階級別企業数

		従業者数規模階級									合計
		0-4	5-9	10-19	20-49	50-99	100-199	200-499	500-999	1000-	
事業 所 数 規 模 階 級	0	29	6	3	2	1					41
	1-9	193835	16448	8397	4800	1390	616	334	89	33	225942
	10-29			3	28	52	56	70	19	13	241
	30-39			1	2	4	14	36	12	12	81
	40-49					2	4	14	6	12	38
	50-59						1	6	8	3	18
	60-69					1		3	3	4	11
	70-79						3	3	8	2	16
	80-89							1	1	2	4
	90-99								1	3	4
	100-199					1			2	2	5
	200-299								8	8	16
	300-399						1		1		2
	400-499									3	3
	500-599									1	1
	600-699									1	1
	700-799									0	0
800-899									2	2	
合計		193864	16454	8404	4832	1451	695	467	158	101	226426

表6は事業所数規模階級別、売上高規模階級別企業数を示している。売上高規模階級が一番低い階級（0～999ユーロ）に事業所数が多い企業がある。参考までに売上高規模階級1（0～999ユーロ）について事業所数規模階級別法的階級別企業数を見ると、事業所数が多い企業の法的形態は主に有限責任会社である。同階級について事業所数規模階級別産業分類5桁別企業数を見ると、事業所数が多い5桁産業は「貯蓄銀行による金融仲介業」（65120 Monetary intermediation by deposit banks）、「非生命保険会社」（66031 Non-life insurance companies）である。したがって、売上高が非常に小さいにもかかわらず、事業所数が多い企業は金融・保険業であること、この産業の売上高にはアウトプットの一部しか計上されていないと推察される。

表6に戻ると、売上高規模階級2（1,000～39,999ユーロ）から5（400,000～1,999,999

³ クラメールの連関係数は $\sqrt{\frac{\chi^2}{n(\min[k,l]-1)}}$ で計算した。 χ^2 はカイ自乗値。 n はデータの個数、 k,l はクロス集計表の列数と行数である。

ユーロ) までは傘下事業所数はほとんどの企業で 9 カ所以下である。それより高い規模階級を見ると、同じ売上高規模階級でも事業所数はかなりばらついている。クラメールの連関係数は表 5 よりさらに低く 0.182 である。このことから売上高規模と事業所数規模の関係性は弱いことがわかる。

表 6 事業所数規模階級別、売上高規模階級別企業数

		売上高規模階級									合計	
		0-999 39,999	1,000- 39,999	40,000- 99,999	100,000- 399,999	400,000- 1,999,999	2,000,000- 9,999,999	10,000,000- 39,999,999	40,000,000- 199,999,999	200,000,000-		
事業所数規模階級	0	2	13	6	7	5	8				41	
	1-9	8377	65596	57463	57609	26732	7878	1712	503	72	225942	
	10-29	19				5	53	59	76	29	241	
	30-39	15					6	14	33	13	81	
	40-49	1					1	10	15	11	38	
	50-59	1						3	10	4	18	
	60-69						1		1	4	5	11
	70-79	2						5	4	5	16	
	80-89	1								2	1	4
	90-99										4	4
	100-199	1					1				3	5
	200-299	2						1	7	6	16	
	300-399								1		1	2
	400-499	1								1	1	3
	500-599										1	1
	600-699										1	1
700-799										0	0	
800-899										2	2	
合計		8422	65609	57469	57616	26742	7948	1806	655	159	226426	

3. 規模別開設率

ビジネスレジスターの精度向上にはプロファイリングが欠かせない。プロファイリングとは行政記録情報等から入手した開業・開設、廃業・閉鎖、規模拡大・縮小などの企業組織構造の変化に関する情報を確認することである。その確認方法は、米国センサス局のように紙媒体の調査票によるものや、カナダ統計局のように電話やインターネットによるものや、英国統計局のように両方を実施するなどさまざまである。共通しているのは、全部の事業者について毎年確認することは不可能であるので、対象を絞って行っていることである。そして現在の我が国で問題になっているのは、その対象をいかに絞るかである。

そこでフィンランド統計局のマイクロデータを用いて、対象企業を選定する方法を検討するための実験を行う。フィンランド統計局のマイクロデータには事業所の開設年月の情報があるので、開設事業所を特定することはできる。ただし、閉鎖した事業所のデータはない。そこで企業組織構造の変化の有無ではなく、開設事業所の有無について分析する。このとき開設事業所とは過去 2 年以内（2000 年 1 月以降）に開設した事業所とする。

まず規模別開設率を観察する。ここでは開設率とは、開設事業所が有った企業の割合と定義する。表 7 は事業所数規模階級別、従業者数規模階級別開設率である。事業所数規模階級で見ると規模が大きい方が開設率も高い傾向が観察される。従業者数規模階級別に見ると、一番小さい階級（0～4 人）を除けば、同様に規模が大きい方が開設率も高い傾向が

観察される観察される。表 8 は事業所数規模階級別、売上高規模階級別開設率である。売上高規模階級別に見ると、下から 5 番目の階級（400,000～1,999,999 ユーロ）以降は、規模が大きい方が開設率も高い傾向が観察されるが、それより下の階級ではむしろ逆に傾向が観察される。

表 7 事業所数規模階級別、従業者数規模階級別開設率

		従業者数規模階級									合計
		0-4	5-9	10-19	20-49	50-99	100-199	200-499	500-999	1000-	
事業所数規模階級	0	0%	0%	0%	0%	0%					0%
	1-9	18%	11%	12%	16%	20%	26%	28%	27%	27%	17%
	10-29			67%	75%	79%	57%	66%	63%	46%	66%
	30-39			100%	100%	100%	79%	78%	83%	75%	80%
	40-49					50%	100%	86%	100%	75%	84%
	50-59						100%	83%	100%	67%	89%
	60-69					100%		100%	100%	75%	91%
	70-79						100%	100%	88%	50%	88%
	80-89							100%	100%	100%	100%
	90-99								100%	100%	100%
	100-199					100%			100%	100%	100%
	200-299								100%	88%	94%
	300-399						100%		100%		100%
	400-499									100%	100%
	500-599									100%	100%
	600-699									100%	100%
	700-799										
800-899									100%	100%	
合計		18%	11%	12%	16%	23%	31%	41%	53%	59%	17%

表 8 事業所数規模階級別、売上高規模階級別開設率

		売上高規模階級								合計	
		0-999	1,000-39,999	40,000-99,999	100,000-399,999	400,000-1,999,999	2,000,000-9,999,999	10,000,000-39,999,999	40,000,000-199,999,999		200,000,000-
事業所数規模階級	0	0%	0%	0%	0%	0%	0%				0%
	1-9	21%	24%	17%	13%	11%	14%	21%	26%	26%	17%
	10-29	53%				60%	85%	68%	63%	48%	66%
	30-39	73%					100%	93%	76%	77%	80%
	40-49	100%					100%	80%	87%	82%	84%
	50-59	100%						100%	80%	100%	89%
	60-69						100%	100%	100%	80%	91%
	70-79	100%						80%	100%	80%	88%
	80-89	100%							100%	100%	100%
	90-99									100%	100%
	100-199	100%					100%			100%	100%
	200-299	100%						100%	100%	83%	94%
	300-399							100%		100%	100%
	400-499	100%							100%	100%	100%
	500-599									100%	100%
	600-699									100%	100%
700-799											
800-899									100%	100%	
合計		21%	24%	17%	13%	11%	15%	24%	37%	52%	17%

4. プロビット・モデルによるプロファイリング対象の選定

規模別開設率の観察からは、事業所数規模階級と開設率との関係性は明確であった。そこで、開設事業所の有無を被説明変数とし、事業所数規模階級と従業者規模階級を説明変数とする。また、従業者数規模階級については、一番下の階級（0～4 人）を除けば、開設率との関係性は明確であった。このように従業者数規模が非常に小さい企業のプロファイリングは費用の面で割り合わないので、行政記録から得られた企業組織構造の変化に関する情報をそのままビジネスレジスターに反映することが国際的になされている。そこで、従業者数規模階級が一番下の階級は分析の対象外とする。

また、欧米諸国で行われている実際のプロファイリングでは、従業者数あるいは売上高の規模の大きい企業を、毎年、プロファイリングするのが一般的である。そこで従業者数規模あるいは売上高規模において一番上の階級に区分される企業は、優先的にプロファイリングの対象となると想定する。ちなみにこれに該当する企業は 191 社である。

以上より、開設事業所の有無を被説明変数とし、事業所数規模階級と従業者規模階級を説明変数とし、従業者数規模階級の一番下の階級と一番上の階級を除き（10 人以上 999 人以下）かつ売上高規模階級の一番上の階級を除いた（1, 999, 999 ユーロ以下）の企業を分析対象とする。事業所数規模階級を説明変数とするのは、事業所数が多い企業は、開設に関するノウハウがあるので、開設しやすいと考えるからである。従業者数規模階級を説明変数とするのは、従業者数が多い企業は、開設に必要な人員に余裕があると考えられるからである。

開設事業所の変化の有無は二項選択であるので、プロビット・モデルを適用する。なおプロビット・モデルの適用については Colin & Triverdi[1998]を参照した。モデルは次の通りである。

$$Y_i^* = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$
$$Y_i = \begin{cases} 1 & Y_i^* > 0 \\ 0 & Y_i^* \leq 0 \end{cases}$$

ここで Y_i は開設事業所の変化の有無、 X_{2i} は従業者数規模階級、 X_{3i} は事業所数規模階級、 ε_i は誤差項である。このモデルに誤差項の分布として標準正規分布を仮定したのがプロビット・モデルである。表 7 はプロビット・モデルの推定結果である。サンプルの大きさが 32,374 社と大きいので、z 統計量の p 値はすべて 0.000% である。マクファーデンの決定係数 (McFadden R-squared) が 0.035677 とかなり低い。従業者数規模階級 (PERSONNEL)、事業所数規模階級 (ESTAB) の係数のいずれも正值であり、これは従業者数が多い、あるいは事業所数が多い企業の開設率が高いことを示しており、妥当である。

モデルの当てはまり(表 8)を見ると、正しく判定した割合は高い(0.6%+86.3%=86.9%)。開設事業所が有ると判定した企業 253 社のうち、74.7% (189 社) は実際に開設事業所があったことになり、これに従ってプロファイリングした場合の効率は良い。だが、実際には開設事業所があった企業 4,378 社のうち 4.3% (189 社) しか、「開設事業所有り」と判定されておらず、大半の企業が漏れてしまう。

表 7 プロビット・モデルの推定結果 (モデル 1)

Dependent Variable: NEW ESTAB				
Method: ML - Binary Probit (Quadratic hill climbing)				
Date: 06/02/14 Time: 15:36				
Sample: 1 32374				
Included observations: 32374				
Convergence achieved after 5 iterations				
Covariance matrix computed using second derivatives				
Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-3.104278	0.107489	-28.88002	0.0000
ESTAB	0.798036	0.054970	14.51773	0.0000
PERSONNEL	0.129989	0.007461	17.42184	0.0000
McFadden R-squared	0.035677	Mean dependent var	0.135232	
S.D. dependent var	0.341976	S.E. of regression	0.334677	
Akaike info criterion	0.764339	Sum squared resid	3625.823	
Schwarz criterion	0.765116	Log likelihood	-12369.36	
Hannan-Quinn criter.	0.764588	Deviance	24738.72	
Restr. deviance	25653.99	Restr. log likelihood	-12826.99	
LR statistic	915.2645	Avg. log likelihood	-0.382077	
Prob(LR statistic)	0.000000			
Obs with Dep=0	27996	Total obs	32374	
Obs with Dep=1	4378			

表 8 プロビット・モデルの当てはまり (モデル 1)

		判定		
		有り	無し	計
実際	有り	189	4189	4378
	無し	64	27932	27996
	計	253	32121	32374

タテ構成比		判定		
		有り	無し	計
実際	有り	74.7%	13.0%	13.5%
	無し	25.3%	87.0%	86.5%
	計	100.0%	100.0%	100.0%

ヨコ構成比		判定		
		有り	無し	計
実際	有り	4.3%	95.7%	100.0%
	無し	0.2%	99.8%	100.0%
	計	0.8%	99.2%	100.0%

総計に対する 構成比		判定		
		有り	無し	計
実際	有り	0.6%	12.9%	13.5%
	無し	0.2%	86.3%	86.5%
	計	0.8%	99.2%	100.0%

最初のモデルでは、実際のプロファイリングに適用するには不十分であるので、追加すべき説明変数を検討する。企業の開業年と、傘下の事業所の開設率を見ると、2000 年以降に開業した企業の開設率が 70%と他の年次と比較して非常に高い（表 9）。1999 年以前に開業した企業について見ると、新しい企業の開設率が高いとは必ずしも言えない。したがって、2000 年以降に開業した企業に着目して説明変数を追加することが考えられる。

表 9 企業の開業年と傘下の事業所の開設率

開業年	分析対象企業数	内開設事業所が 有った企業数	開設率
1900-1909	12	3	25%
1910-1919	33	6	18%
1920-1929	35	1	3%
1930-1939	33	5	15%
1940-1949	118	23	19%
1950-1959	215	26	12%
1960-1969	760	64	8%
1970-1979	3870	224	6%
1980-1989	9420	625	7%
1990-1999	15041	1405	9%
2000-	2837	1996	70%
計	32374	4378	14%

以上を踏まえて 2000 年以降に開業した企業を 1、1999 年以前に開業した企業を 0 とするダミー変数を説明変数として追加してモデルを修正した。

$$Y_i^* = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 D_i + \varepsilon_i$$

$$Y_i = \begin{cases} 1 & Y_i^* > 0 \\ 0 & Y_i^* \leq 0 \end{cases}$$

ただし、 D_i はダミー変数である。このモデルの推定結果（表 10）を見ると、上のモデルと同様にサンプルの大きさが 32,374 社と大きいため、z 統計量の p 値はすべて 0.000%である。マクファーデンの決定係数は 0.269020 とかなり改善している。従業者数規模階級（PERSONNEL）、事業所数規模階級（ESTAB）、ダミー変数（START_2000S）の係数のどれも正値であるが、これは従業者数が多い、あるいは事業所数が多い、あるいは最近開業した企業の開設率が高いことを示している。

モデルの当てはまり（表 11）を見ると、正しく判定した割合は高い（6.6%+83.8%=90.4%）。開設事業所があると判定した企業 3,005 社のうち、70.9%（2,130 社）は実際に開設事業所があったことになり、これに従ってプロファイリングした場合の効率は良い。また、実際には開設事業所があった企業 4,378 社のうち 48.7%（2,130 社）が、「開設事業所有り」と判定されており、約半分がカバーされている。

表 10 プロビット・モデルの推定結果（モデル 2）

Dependent Variable: NEW_ESTAB				
Method: ML - Binary Probit (Quadratic hill climbing)				
Date: 06/02/14 Time: 22:31				
Sample: 1 32374				
Included observations: 32374				
Convergence achieved after 5 iterations				
Covariance matrix computed using second derivatives				
Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-3.657472	0.110877	-32.98684	0.0000
ESTAB	0.820630	0.056479	14.52975	0.0000
PERSONNEL	0.189582	0.008119	23.35142	0.0000
START_2000S	2.017117	0.027350	73.75304	0.0000
McFadden R-squared	0.269020	Mean dependent var	0.135232	
S.D. dependent var	0.341976	S.E. of regression	0.285233	
Akaike info criterion	0.579494	Sum squared resid	2633.545	
Schwarz criterion	0.580530	Log likelihood	-9376.272	
Hannan-Quinn criter.	0.579825	Deviance	18752.54	
Restr. deviance	25653.99	Restr. log likelihood	-12826.99	
LR statistic	6901.444	Avg. log likelihood	-0.289624	
Prob(LR statistic)	0.000000			
Obs with Dep=0	27996	Total obs	32374	
Obs with Dep=1	4378			

表 11 プロビット・モデルの当てはまり（モデル 2）

		判定		
		有り	無し	計
実際	有り	2130	2248	4378
	無し	875	27121	27996
	計	3005	29369	32374

タテ構成比		判定		
		有り	無し	計
実際	有り	70.9%	7.7%	13.5%
	無し	29.1%	92.3%	86.5%
	計	100.0%	100.0%	100.0%

ヨコ構成比		判定		
		有り	無し	計
実際	有り	48.7%	51.3%	100.0%
	無し	3.1%	96.9%	100.0%
	計	9.3%	90.7%	100.0%

総計に対する構成比		判定		
		有り	無し	計
実際	有り	6.6%	6.9%	13.5%
	無し	2.7%	83.8%	86.5%
	計	9.3%	90.7%	100.0%

プロビット・モデルでは、開設確率を

$$P = F(Y^*) = F(\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 D_i)$$

と推定することができる。F は正規分布の累積分布関数である。表 12 はプロビット・モデルによって推定した開設確率である。開設確率が高い企業から順番にプロファイリングすることにより、少ない費用で企業組織構造の変化を把握できると考えられる。

表 12 プロビット・モデルによる開設確率

企業の開業年が1999年以前の場合		従業者数規模階級						
		5-9	10-19	20-49	50-99	100-199	200-499	500-999
事業 所 数 規 模 階 級	0	1%	1%	2%	3%	4%	7%	9%
	1-9	5%	7%	10%	14%	19%	25%	31%
	10-29	21%	27%	33%	40%	48%	55%	63%
	30-39	50%	58%	65%	72%	78%	83%	87%
	40-49	80%	84%	89%	92%	94%	96%	98%
	50-59	95%	97%	98%	99%	99%	100%	100%
	60-69	99%	100%	100%	100%	100%	100%	100%
	70-79	100%	100%	100%	100%	100%	100%	100%
	80-89	100%	100%	100%	100%	100%	100%	100%
	90-99	100%	100%	100%	100%	100%	100%	100%
	100-199	100%	100%	100%	100%	100%	100%	100%
	200-299	100%	100%	100%	100%	100%	100%	100%

企業の開業年が2000年以降の場合		従業者数規模階級						
		5-9	10-19	20-49	50-99	100-199	200-499	500-999
事業 所 数 規 模 階 級	0	33%	40%	48%	55%	62%	69%	76%
	1-9	65%	72%	78%	83%	87%	91%	94%
	10-29	89%	92%	94%	96%	97%	98%	99%
	30-39	98%	99%	99%	100%	100%	100%	100%
	40-49	100%	100%	100%	100%	100%	100%	100%
	50-59	100%	100%	100%	100%	100%	100%	100%
	60-69	100%	100%	100%	100%	100%	100%	100%
	70-79	100%	100%	100%	100%	100%	100%	100%
	80-89	100%	100%	100%	100%	100%	100%	100%
	90-99	100%	100%	100%	100%	100%	100%	100%
	100-199	100%	100%	100%	100%	100%	100%	100%
	200-299	100%	100%	100%	100%	100%	100%	100%

5. おわりに

ビジネスレジスターのマイクロデータを用いた分析を応用することにより、企業組織構造に変化があった企業を効率的に見つけることができる可能性がある。実施のプロファイリングでは、新聞記事等の情報も活用できるわけであり、もっと正確に組織構造に変化があった企業を見つけることができるであろう。

参考文献

- Cameron, A. Colin, Pravin K. Trivedi[1998] *Regression Analysis of Count Data*, Econometric Society Monograph No.30, Cambridge University Press
- Kinney, Satkartar K., Jerome P. Reiter[2013] *SynLBD: Providing firm characteristics on synthetic establishment data*, Proceedings of the 59th World Statistics Congress of the International Statistical Institute, 2013

フィンランド 2002 年ビジネスレジスターの補完実験*

宮内 環 (慶應義塾大学経済学部)†

概要

本稿では、1990 年代後半よりマイクロデータにおける欠損値の補完に広く適用されるようになった NIM (Nearest-neighbor Imputation Method) による補完の精度の評価を、フィンランド国家統計局の 2002 年ビジネスレジスターを用いて行い、その結果について報告する。

当該ビジネスレジスターの特徴として次の二点が挙げられる。第一点目は、ビジネスレジスターを構成する事業所ファイルの事業所レコードには、各事業所の緯度・経度の情報が含まれており、任意の二つの事業所間の地理的距離が計算可能である、という点。第二点目は、このビジネスレジスターは事業所ファイル、企業ファイルの二種類により構成され、事業所レコードと企業レコードには各々識別番号が付与されており、事業所レコードと当該事業所が属する企業の企業レコードとを接続することが可能である、という点である。第一点目の事業所の緯度・経度情報は、事業所レコードのある一定の調査項目について生じる欠損値を補完するにあたり、NIM を適用するために不可欠である。さらに二点目により、事業所レコードのある一定の調査項目について生じる欠損値を NIM により補完する場合、当該事業所と同一の属性を持ち、当該の調査項目に欠損を生じていないドナーの候補としての事業所の範囲を特定するにあたり、事業所レコードに含まれる産業分類の情報だけを用いる場合、さらに欠損値が生じている当該事業所が属する企業の売上高階層の情報を併せて用いる場合、といった、情報量の差による補完の精度の比較が可能となる。

本稿では、フィンランド国家統計局のこうした特長を活かし、シミュレーションの手法により事業所の従業者数に擬似的に欠損値を生じさせ、これを NIM により補完を試み、その補完により得た値と、欠損を生じさせる前の値との比較により、補完の精度を評価した。

本稿における最も重要な結論は、つぎの二点である。第一点目は、ドナーの事業所の範囲を特定するにあたり、欠損値が生じた事業所に関する情報だけでなく、当該事業所が属する企業に関する情報も併せて用いると、補完の精度がより向上することが示される、という点。第二点目は、事業所の従業者数の欠損値を補完するにあたり、当該事業所の従業者階層に関する情報を併用すると、補完の精度が飛躍的に向上する、という点である。

キーワード: ビジネスレジスター、欠損値、補完、NIM (Nearest-neighbor Imputation Method)、シミュレーション
JEL 分類コード: C81

*この研究成果は科学研究費助成事業 (学術研究助成基金助成金 (基礎研究 (C)(一般))) 「ビジネスレジスターによる企業動態統計の開発」 (補助事業期間 平成 24 年度～平成 26 年度) の助成を得ることにより具体化されたものである。この研究に用いたビジネスレジスターはフィンランド国家統計局より有償で譲り受けたもので、この取得のための費用は前述の助成金の一部により賄われた。さらに同科学研究費助成事業の助成により組織された研究会の研究代表者である菅幹雄 (法政大学経済学部)、同研究会の研究協力者である森博美 (法政大学経済学部)、宮川幸三 (慶應義塾大学産業研究所)、Jukka Pakola (Statistics Finland)、Ossi Nurmi (Statistics Finland) の各氏からは、研究会において当該研究成果に関する研究上の多大なる助言と示唆を得た。ここに記して感謝の意を表する。なお、本稿におけるすべての誤謬は筆者の責に帰するものである。

†email: miyauchi@econ.keio.ac.jp

1 はじめに

統計調査においては、たとえそれがセンサス調査であっても、ある一定の調査項目について欠損値が生じやすい、という傾向はしばしば経験されるところである。ある調査項目に欠損値が生じた場合、これを集計する前の個票の段階で補完を試みる方法が近年において盛んに議論され、その方法の有力なものとして Nearest-neighbor Imputation Method (以下では NIM とよぶ) が近年注目されている。

本稿では、この NIM における補完の精度、およびその補完における追加的な情報が補完の精度にいかなる効果を及ぼすか、の二点について客観的に評価するために、フィンランド国家統計局の 2002 年ビジネスレジスターのデータセットを用いて、シミュレーションの手法による補完に関する数値的な実験を行うことにした。本稿における実験では、主に次の二点について確認を行った。まず、欠損確率が一樣である場合と個体の属性により変わる場合の各々において、補完の精度はどのように変化するか数値的に調べた。つぎに、補完に際して追加的に利用可能な情報が、補完の精度にいかなる与える影響を与えるかを数値的に調べた。

NIM による補完の精度を確認するにあたり、本稿の実験で用いるフィンランドの 2002 年ビジネスレジスターのデータセットの特徴として次の二点を指摘しておくべきであろう。

第一点目として、大別して次に述べる二種類のファイルより構成されていることである。その一つ目は、フィンランド国内の事業所ファイル、二つ目はフィンランド国内の企業ファイルである。前者の事業所データセットに含まれる各事業所レコードには、後者の企業データセットの各企業レコードと接続可能な識別番号が付与されており、これによって各事業所レコードを当該事業所が属する企業の企業レコードと接続することが可能となっている。

第二点目として、前者の事業所ファイルの事業所レコードには、事業所の従業者数や操業開始年月などの属性の他に、当該事業所の緯度・経度の情報が含まれているので、任意の二つの事業所間の地理的な距離を知ることができる。NIM による補完にはこの地理的な位置情報が欠かせない。

本稿における実験においては、シミュレーションの手法により事業所レコードに含まれる従業者数の数値項目に擬似的に欠損値を生じさせ、NIM により補完した結果と、当該数値項目に欠損値を生じさせる前の値との比較により、補完の精度に関して客観的な評価が可能となる。本稿の実験のデザインは主に次の二点である。

第一点目は、擬似的な欠損を与える確率を、事業所の属性とはかかわりなく一樣とする場合、事業所の属性により変化する場合、の二通りを設定したことである。本稿の実験では、こうした欠損確率の分布の違いにより補完の精度がどのような影響を受けるのかを評価した。

第二点目は、欠損値が生じた事業所において観察される属性 (欠損値が生じている数値項目以外の属性) と類似の属性を持つ別の事業所を探すにあたり、観察される属性を、事業所レコードから得られる情報に限定した場合、事業所レコードだけでなく当該事業所が属する企業の企業レコードから得られる情報も併せて用いる場合、の二通りを設定したことである。すなわち、実験を行うにあたり、事業所レコードと当該事業所が属する企業の企業レコードを接続データセット、言わば事業所と企業の「名寄せ済みファイル」を後者の場合のために、事業所ファイルとは別に準備した。本稿における実験では、欠損値を NIM により補完する試みにおいて、ある調査項目に欠損値が生じている事業所と同一の産業分類に属し、当該項目に欠損値が生じていない他の事業所をドナーの候補として特定してい

る。このとき、欠損値が生じている事業所についてその産業分類以外の属性として、当該事業所が属している企業の売上高階層がわかれば、この情報を用いてドナーの候補をさらに限定することができよう。本稿における実験のデザインの第二点目は、事業所レコードより得られる産業分類の情報のみによりドナーの候補を特定した場合、「名寄せ済みファイル」により産業分類の他に、当該事業所が属する企業の売上高階層の情報も併せてドナーの候補を特定した場合、の各々において補完を行い、これら両者の場合の補完の精度を比較することを行った。

本稿の構成は次のとおりである。第2節では、先行研究の概略を述べ、第3節では、本稿の実験で用いたフィンランド国家統計局のビジネスレジスターのデータの概略について述べる。第4節では、本稿の実験方法について述べ、第5節では、実験の結果についてその概略を述べる。第6節では結語を述べる。

2 先行研究

欠損値の補完は、かねてより統計調査の課題であった。当初は集計表における欠損セルを補完する方法が主であったが、1970年代ころからマイクロデータが統計の活用において主流となり始めたとともに、マイクロデータの各レコードに含まれる欠損値の補完が主要な問題として議論されるようになった。Fellegi and Holt (1976) がマイクロデータの調査項目の観測値について、整合性の検討 (Editing) と異常値や欠損値を補完 (Imputing) をコンピュータを用いて行う方法を提示し、以後 “Fellegi-Holt Method” として知られている。さらに同時期に Rubin (1976) も欠損値についての推測の議論を行っている。Little and Rubin (1987;2002) も含め、これらは主に統計的方法を背景としている。その後、Winkler and Chen (2001) では、“Fellegi-Holt Method” の展開が示されている。その他に、Rubin (1987;1996) は単一の欠損値だけでなく、複数の欠損値のセットを同時に補完する方法について提示している。すなわち欠損値の個々についてその周辺分布に基づいて補完をするのではなく、複数の欠損値のセットの背後にある同時分布を想定して補完を行うという考え方である。

一方、Bankier, et al. (1997) の提案による NIM (Nearest-neighbor Imputation Method) をカナダの国勢調査へ適用し、その後 Statistics Canada (1999;2002) ではその展開を行い、以後マイクロデータにおける補完にこの NIM が適用される事例が増えている。Andridge and Little (2010) は、以上の補完に関する歴史的展開を踏まえ、補完の方法論に関する包括的な議論を行っている。

わが国では森・菅 (2012) が事業所企業統計をビジネスレジスターと位置づけ、その個票データを用いて補定の精度をシミュレーションにより検証している。そのシミュレーション手法の概略をおおよそ次の通りである。まず事業所企業統計調査の本調査である平成 18(西暦 2006) 年のデータセットにおいて、従業者数の変数についてランダムに擬似的欠損値を作り出す。つぎに当該欠損値が生じた事業所が (1) 既存事業所であるか、あるいは (2) 新規事業所であるかにより、次の二通りの異なる補完方法を採用している。第 1 に当該欠損値が生じた事業所が既存事業所である場合には、過去に行われた平成 16(西暦 2004) 年あるいは平成 13(西暦 2001) 年調査における同一事業所の観測値を用いて補完する。第 2 に当該欠損値が生じた事業所が新規事業所である場合には、同じ平成 18 年の新規事業所のうち、欠損値を生じていない事業所における従業者数の地域別・産業別平均値により補完している。以上の方法により補完された従業者数と、実際 (擬似的に欠損値とされる前) の従業者

数の誤差 (および誤差率) の分布により、補完の精度を確認している。なお、欠損確率は、地域区分・産業区分にかかわらず一様の場合、都道府県や市町村といった地域区分および産業区分によって変わる場合を試みている。シミュレーションの結果として、筆者は次の二点を挙げている。第一に、既存事業所の欠損値を過去のデータにより補完する方法は時系列的なトレンドが急な変化でなければ良好な結果を与えるであろう。第二に、新規事業所の欠損値を欠損値を生じていない事業所の観測値で補完する方法は、事業所の特性に関する情報 (たとえば地域区分、産業区分、資本金などの規模区分) の情報を活用できれば良好な結果を与えるであろう、としている。

わが国では他に、高橋・伊藤 (2013) は売上高の補完についての検討を行っている。

3 フィンランド国家統計局のビジネスレジスター

我々はフィンランド国家統計局から 2002 年のビジネスレジスター (以下では “FBR2002” とよぶ) を有償で譲り受け、これを使う許可を得た。ここでは、FBR2002 の概略と本稿で述べる NIM (Nearest-neighbor Imputation Method) による補完の実験のために準備されたデータセットの概略を述べる。

3.1 事業所ファイルと企業ファイル

FBR2002 は、事業所ファイルと企業ファイルの二つのファイルより構成されている。前者の事業所レコードと後者の企業レコードにはユニークな企業 ID が付与され、この企業 ID によってある企業を構成する事業所のレコードを名寄せすることが可能となっている。表 1 と表 2 には各々、事業所レコード、企業レコードに含まれる変数を掲載した。

表 1: FBR2002 の事業所ファイルの事業所 (レコード) 数と変数

事業所数	255,127	
変数	桁数	備考
企業 ID	11	FBR2002 の企業レコードとのリンケージに利用可能
事業所コード	9	
郵便番号	5	
行政区番号	3	
産業分類記号	5	
事業所の従業者数階層	1	0-4, 5-9, 10-19, 20-49, 50-99, 100-199, 200- の 7 階層
操業開始年月日	8	西暦 4 桁、月 2 桁、日 2 桁
事業所の従業者数	6	補完の実験ではこの変数を擬似的に欠損値とした。
事業所の位置座標	14	前 7 桁が緯度、後 7 桁が経度

なお、以上の事業所レコードおよび企業レコードの「従業者数」には欠損値のコードは見当たらない¹。従って、第 4 節で述べる補完の実験において、擬似的に「従業者数」に欠損値が生じたとしてこれを補完した場合、その補完の精度を、欠損を生じさせた当該レ

¹ 数値の零 (ゼロ) は頻出するが、筆者はこれを「従業者数」の欠損値とはみなしておらず、当該変数の観測値として 0 の値が得られたと解釈している。これに対し、「事業所の位置座標」には座標の値が代入されていないレコードが多数あり、この場合は明らかに欠損値であると判断できる。この点を考慮すると、「従業者数」に現れる 0 の値は欠損値ではなく、0 という観測値が得られたと理解することが妥当であろう。

表 2: FBR2002 の企業ファイルの企業 (レコード) 数と変数

企業数	226,426	
変数	桁数	備考
企業 ID	11	FBR2002 の事業所レコードとのリンケージに利用可能
行政区番号	3	
産業分類記号	5	
企業の従業者数	6	
企業の従業者数階層	1	0-4, 5-9, 10-19, 20-49, 50-99, 100-199, 200-499, 500-999, 1000- の 9 階層
売上高階層	1	9 階層
事業開始年月日	8	西暦 4 桁、月 2 桁、日 2 桁
法律上の組織形態区分	2	コードの詳細は不明
所有形態区分	1	私有国内、国有、地方自治体、オーランド自治州、外国人、その他、の 6 区分
雇用する者の活動状況区分	1	コードの詳細は不明
付加価値税の状況区分	1	コードの詳細は不明
輸入・輸出の状況区分	1	輸入・輸出の有無の別

コードの「従業者数」の値と、補完によって得た値との乖離によって評価することが可能である。

本稿の第 4 節で述べる補完の実験では、上述の事業所ファイルのみを用いた実験の他に、FBR2002 の事業所レコードを、FBR2002 の企業ファイルの企業レコードに名寄せをしたレコード (以下では「名寄せ済みレコード」とよび、この名寄せ済みレコードから構成されるファイルを「名寄せ済みファイル」とよぶ) も用いた実験を行なっている。次項ではこの名寄せ済みレコード作成と名寄せ結果の概略について述べる。

3.2 事業所レコードの企業レコードへの名寄せ

「名寄せ済みレコード」の作製には、表 1 に示した事業所レコードの企業 ID と、表 2 に示した企業レコードの企業 ID とを比較し、両者が完全に一致した場合に、事業所レコードの企業レコードへの名寄せを行なった。名寄せの結果は非常に良好で、その状況を表 3 および表 4 に掲示した。

表 3: FBR2002 の事業所レコードの企業レコードへの名寄せ状況

場合	事業所レコード数
企業レコードに名寄せ (接続) でできなかった事業所レコード数	760
企業レコードに名寄せ (接続) できた事業所レコード数	254,367
合計	255,127

表 4: FBR2002 の企業レコードの事業所レコードへの名寄せ状況

一つの企業レコードに名寄せ (接続) できた事業所レコード数	企業 レコード数
0	41
1	219,006
2	4,445
3	1,244
4	503
5	276
6	208
7	124
8	62
9	74
10	52
11	37
12	40
13	21
14	14
15	22
16	14
17	16
18	16
19	9
20	14
21- 50	125
51-100	38
101-	25
合計	226,426

4 補完の実験の概略

本稿で述べる補完の実験は、ある事業所レコードの従業者数について擬似的に欠損値が発生したとして、その欠損値を NIM により補完し、この補完によって得られた従業者数の値、および当該レコードに記録された従業者数を真の値とし、両者を比較して補完の精度を評価している。なお、欠損値の発生確率は、事業所の属性にかかわらず一様に与えた場合、事業所の産業大分類および従業者階層別に異なる値を与えた場合を設定している。前者の場合は、一様に 15%, 30%, 45% の 3 ケースとし、これらを各々「ケース 1」、「ケース 2」、「ケース 3」とよぶ。さらに後者の場合は、事業所の産業大分類および従業者数規模別の欠損確率を表 5 に示し、これを「ケース 4」とよぶ。なお、産業大分類は大文字のアルファベット一文字で表され、その意味を表 6 に示した。

4.1 事業所ファイル、および名寄せ済みファイルによる 2 種類の補完の実験

本稿で述べる補完の実験は、用いたファイルによって 2 種類の実験に大別することができる。第 1 番目は、事業所ファイルを用いる実験であり、第 2 番目は名寄せ済みファイルを用いる実験である。補完のための NIM を用いる場合、欠損値が生じた事業所と類似の属性を持つ (欠損値が生じていない) 事業所を「ドナー」として探す必要があるが、このと

表 5: 産業大分類別・事業所従業員規模別に設定した欠損確率：産業・規模により異なる場合

産業分類	事業所従業員規模						
	0～4	5～9	10～19	20～49	50～99	100～199	200～
A	0.30	0.25	0.25	0.15	0.15	0.05	0.05
B	0.30	0.25	0.25	0.15	0.15	0.05	0.05
C	0.30	0.25	0.25	0.15	0.15	0.05	0.05
D	0.15	0.15	0.15	0.10	0.10	0.05	0.05
E	0.15	0.15	0.15	0.10	0.10	0.05	0.05
F	0.30	0.25	0.20	0.10	0.10	0.05	0.05
G	0.30	0.25	0.20	0.10	0.10	0.05	0.05
H	0.30	0.25	0.20	0.10	0.10	0.05	0.05
I	0.30	0.25	0.20	0.10	0.10	0.05	0.05
J	0.15	0.15	0.15	0.10	0.10	0.05	0.05
K	0.30	0.25	0.20	0.10	0.10	0.05	0.05
L	0.15	0.15	0.15	0.10	0.10	0.05	0.05
M	0.30	0.25	0.20	0.10	0.10	0.05	0.05
N	0.30	0.25	0.25	0.15	0.15	0.05	0.05
O	0.30	0.25	0.25	0.15	0.15	0.05	0.05
X	0.15	0.15	0.15	0.10	0.10	0.05	0.05

注) 産業分類 "P", "Q" の事業所は件数ゼロなので表示を省略。

表 6: 産業大分類コードと大分類産業部門名

産業分類	大分類の産業部門名
A	"Agriculture, hunting and forestry"
B	Fishing
C	Mining and quarrying
D	Manufacturing
E	"Electricity, gas and water supply"
F	Construction
G	"Wholesale and retail trade; repair of motor vehicles, motorcycles and personal and household goods"
H	Hotels and restaurants
I	"Transport, storage and communication"
J	Financial intermediation
K	"Real estate, renting and business activities"
L	Public administration and defence; compulsory social security
M	Education
N	Health and social work
O	"Other community, social and personal service activities"
P	Private households employing domestic staff and undifferentiated production activities of households for own use
Q	Extra-territorial organizations and bodies
X	Industry unknown

き事業所の属性に関する情報量が、第1番目の事業所ファイルを用いる実験(この実験を以後は「実験1」とよぶ)と、第2番目の名寄せ済みファイルを用いる実験(これを以後は「実験2」とよぶ)では異なり、一般に実験2のほうが前者に比べ情報量が多くなる。具体的には、前者の事業所ファイルを用いる実験では、事業所の産業分類や事業所の位置座標といった特定の事業所に固有の情報のみが得られるに過ぎないが、名寄せ済みファイルを用いる実験2では、事業所に固有の情報に加えて、当該事業所が属する企業全体の売上高階層などの情報も利用可能となる。これら二つの実験の目的は、事業所に関する属性の情報の追加による補完の精度の向上を確認することであり、事業所ファイルによる実験1におけるよりも、より豊富な情報を利用可能な名寄せ済みファイルを用いる実験2においていっそう高い補完の精度が期待される。

補完の精度は、本稿のこの節(第4.1節)に以下に述べる実験手続きの詳細の第8項から第11項に示されるように、補完によって得られた従業者数の値、および従業者数の真の値である当該レコードに記録された従業者数の誤差および誤差率によっている。

以上に述べた補完の実験の手続きの概略は次の通りである。実験1、実験2のいずれにおいても、まず事業所レコードの一部に従業者数に擬似的な欠損値を生じさせ、従業者数に欠損値が生じていない一定の他の事業所レコードの集合からNIMにより選び取られた事業所レコードを「ドナー」とし、当該ドナーの従業者数により欠損値を補完する。実験1と実験2の違いは、NIMを適用する「従業者数に欠損値が生じていない他の事業所レコードの集合」(この集合を以後は「ドナーの候補集合」とよぶ)の違いである。ただし、いずれの実験においても、ドナーの候補集合に含まれる事業所レコードは、欠損値が生じている事業所レコードと同一の産業大分類に属する事業所のそれ(の全部あるいは一部)に限られ、この集合から、表1中の「事業所の位置座標」を用いて直線距離で最も近い事業所を「ドナー」として選び取る。以下に実験の手続きをより詳細に述べる。

1. 事業所ファイル、あるいは名寄せ済みファイルにおける、事業所の「従業者数」に欠損値を生じさせる確率を定める。欠損確率の定め方は、大別して二通りとし、まず事業所の属性にかかわらず一様に15%(ケース1)、30%(ケース2)、45%(ケース3)とした場合と、つぎに表5に従って事業所の産業大分類別と従業者階層別に欠損確率が変化する場合(ケース4)を設定する。

ただし、(ケース4)における全産業および全従業者階層の平均的欠損比率は、(ケース2)の30%にほぼ近い水準の、おおよそ27%となるように設定してある。この点は表7および表8の左上にある産業計・従業者数規模計の「欠損比率平均」の欄を参照されたい。なお、前者は実験1の(ケース4)における欠損比率の実績、後者は実験2のそれを示す。

以下ではこの(ケース1)から(ケース4)の種類の別を、添え字 k ($k = 1, \dots, 4$)で示す²。

2. 以下の3から9までの手続きを、1回の試行として1,000回繰り返す。各試行には i ($i = 1, \dots, 1000$)の番号³を付与する。
3. あらかじめ与えた上述の欠損確率に従い、試行 i において事業所ファイル、あるいは名寄せ済みファイルにおける先頭レコードから末尾レコードまでの間で、事業所の「従業者数」に、擬似的に欠損値を生じさせる。この擬似的な欠損値が生じた事業所(以

² ケースの種類の別を示す k の添え字は、Kind of a caseの先頭の文字より採用した。

³ 試行の番号 i はIterationの先頭の文字より採用した。

後は便宜的に「欠損事業所」とよぶ⁴⁾に先頭から番号 j ($j = 1, \dots, J_{ik}$) を付与する。ただし、 J_{ik} の値は、試行 i ごとにも異なる場合があるだけでなく、一般に上記 1 の (ケース 1) から (ケース 4) によっても変化する。

4. 事業所の従業者数が欠損となった欠損事業所 j に対応し、ドナーの候補集合として、従業者数が欠損している事業所と類似の属性を持ち、かつ欠損値が発生していない事業所 (レコード) の範囲を特定する。ただし、本稿における実験では、ドナーの候補集合は、事業所ファイル、名寄せ済みファイルのいずれを用いるかにより、次に述べる通りに二通りのものを設定した。

事業所ファイルを用いた実験 1 における「ドナーの候補集合 $D_{jik}^{(1)}$ 」: 試行 i のケース k において、従業者数について欠損値が発生している第 j 番目の欠損事業所と同一の産業大分類に属する事業所のうち、従業者数について欠損を生じていないもの。以下では実験 1 におけるドナーの候補集合を「 $D_{jik}^{(1)}$ 」と記す。

名寄せ済みファイルを用いた実験 2 における「ドナーの候補集合 $D_{jik}^{(2)}$ 」: 試行 i のケース k において、従業者数について欠損値が発生している第 j 番目の欠損事業所と同一の産業大分類に属する事業所で、かつ当該事業所の企業の売上高階層と同一の売上高階層にある企業に名寄せされた事業所のうち、従業者数について欠損を生じていないもの。以下では実験 2 におけるドナーの候補集合を「 $D_{jik}^{(2)}$ 」と記す。

5. ドナーの候補集合から、擬似的に欠損値を生じた事業所の位置座標から直線距離にして最も近い事業所のレコードをドナーとして選び出し、その事業所の従業員数によって、欠損値を補完する。併せてドナーとして選び出された事業所について、ドナーとなった回数を記録する。
6. 上の手続きによってすでにドナーとなった回数が 5 回に達している事業所があれば、その事業所はドナーの候補集合から除外し、上記 5 の手続きによりドナーを選び出す。
7. ドナーとして選ばれた事業所の従業者数を、従業者数の欠損値の補完に用いる。
8. 上の 4 において従業者数が擬似的に欠損値とされた事業所で報告されている本来の従業者数と、上の 7 で得られた補完値との誤差および誤差率を計算する。
9. 事業所ファイル、あるいは名寄せ済みファイルにおいて、擬似的に従業者数が欠損値とされたすべての事業所について上の 4 から 8 の手続きを終えたら、これを 1 回の試行として、この 1 回の試行において欠損値が生じたすべての事業所について計算した補完値の誤差および誤差率の平均値や標準偏差などの基本的統計量を記録する。
10. 上の 3 に戻り、新たに事業所ファイル、あるいは名寄せ済みファイルにおける先頭レコードから末尾レコードまでの間で、事業所の「従業者数」に、新たに擬似的に欠損値を生じさせ、以上の試行を 1,000 回繰り返す。ただし、毎回の試行 i ごとに、ドナーの候補集合 $D_{jik}^{(1)}$ および $D_{jik}^{(2)}$ に含まれるすべての事業所について、ドナーとなった回数をすべてゼロに戻してから毎回の試行 i を開始する。
11. 1,000 回の試行をすべて終えたら、上の 9 で計算された各試行における補完の誤差および誤差率の平均値や標準偏差などの基本的統計量について、1,000 回の試行全体にわたる平均値を計算し、実験 1 および実験 2 の各々におけるケース 1 からケース 4 について補完の精度を相互に比較する。

⁴⁾ 「欠損事業所」とは、当該事業所のレコードは事業所ファイル中に存在するが、当該事業所のレコードにおける従業者数が欠損値となっている場合を示すのであって、当該事業所の捕捉それ自体には問題がない点に注意されたい。

以上を要約すると、2種の実験はつぎのように示されよう。

実験1：ドナーを事業所ファイルに含まれる、同一産業のレコードから構成される「ドナーの候補集合 $D_{jik}^{(1)}$ 」(試行 i のケース k において、 j 番目の欠損事業所と同一の産業大分類に属する事業所のうち、従業者数について欠損を生じていないもの) から採用して補完する 1,000 回の試行

実験2：ドナーを名寄せ済みファイルに含まれる、同一産業かつ同一売上高階層のレコードから構成される「ドナーの候補集合 $D_{ji4}^{(2)}$ 」(試行 i のケース k において、第 j 番目の欠損事業所と同一の産業大分類に属する事業所で、かつ当該事業所の企業の売上高階層と同一の売上高階層にある企業に名寄せされた事業所のうち、従業者数について欠損を生じていないもの) から採用して補完する 1,000 回の試行

なお、次節5では、これら実験1、実験2の「拡張」として、上記の、事業所の産業大分類別と従業者階層別に欠損確率が変化する(ケース4)の場合についてのみ、別途実験を行った結果も併せて報告する。実験1、実験2の拡張では、(ケース4)の各々のドナーの候補集合 $D_{ji4}^{(1)}$, $D_{ji4}^{(2)}$ の部分集合⁵として、ドナーの候補集合に含まれる事業所について、試行 i における欠損事業所 j の従業者数階層と同一の従業者数階層に属するものに限定したドナーの候補集合 $D_{ji4}^{(1E)}$, $D_{ji4}^{(2E)}$ よりドナーを選ぶ実験を行った。これらの「拡張」として行った実験を以後は、各々「実験1E」、「実験2E」⁶とよぶ。いま、試行 i における欠損事業所 j について、この事業所の従業者数規模と同一の事業所の集合を、実験1Eで用いる事業所ファイルにおいては $C_{ji4}^{(1)}$ 、実験2Eで用いる名寄せ済みファイルにおいては $C_{ji4}^{(2)}$ としよう⁷。これらの集合を用いれば、実験1E、実験2Eは、つぎのように示されよう。

実験1E：ドナーの候補集合を $D_{ji4}^{(1E)} \equiv D_{ji4}^{(1)} \cap C_{ji4}^{(1)}$ として補完する 1,000 回の試行

実験2E：ドナーの候補集合を $D_{ji4}^{(2E)} \equiv D_{ji4}^{(2)} \cap C_{ji4}^{(2)}$ として補完する 1,000 回の試行

5 実験結果

この節では、実験結果を大別して次の二つの視点から比較検討する。

まず前節で述べた第 i 試行における、欠損値を生じている事業所 j の補完のためのドナーの候補集合を、 $D_{jik}^{(1)}$ とした場合(実験1)と、 $D_{jik}^{(2)}$ とした場合(実験2)の各々の実験結果を、欠損確率が産業分類・従業者数規模にかかわらず一様であるが、欠損確率の水準が 15%, 30%, 45% と変化する場合である(ケース1)から(ケース3)の場合、さらに欠損確率が産業大分類別および従業者数規模別に異なる(ケース4)の場合について示す。すなわち、ドナーの候補集合の設定の方法は(ケース1)から(ケース4)共通であるが、欠損確率の違いが補完の精度に与える影響をこれらの結果の比較により示すことがこれら実験の趣旨である。

つぎに、(ケース4)の場合に限定して、これら実験の拡張として行った「実験1E」と「実験2E」の結果を、「実験1」と「実験2」の(ケース4)における結果と比較する。この比較により、従業者数規模の情報の有無が補完の精度に与える影響を示すことができよう。

⁵ 右下添え字の4は、欠損確率が(ケース4)の場合に限定されている、すなわち $k=4$ であることを示す。

⁶ これらの実験の名前に付した“E”の文字は、従業者数階層である Employee Class, あるいは実験の「拡張」を意味する Extension の最初の文字より採用した。

⁷ この集合の文字 C は、Class of Employee の最初の文字より採用した。

なお、「実験 1E」および「実験 2E」の第 i 番目の各々の試行において、擬似的に欠損値が発生している事業所の集合 (以後はこの集合を便宜的に「欠損事業所集合」とよぶ) を各々 $M_{i4}^{(1E)}$, $M_{i4}^{(2E)}$ とし⁸、さらに (ケース 4) の下で「実験 1」および「実験 2」の第 i 番目の各々の試行における「欠損事業所集合」を各々 $M_{i4}^{(1)}$, $M_{i4}^{(2)}$ とする。このとき

$$\begin{aligned} M_{i4}^{(1)} &= M_{i4}^{(1E)}, & (i = 1, \dots, 1000) \\ M_{i4}^{(2)} &= M_{i4}^{(2E)}, & (i = 1, \dots, 1000) \end{aligned}$$

となるように実験が統御されているので、実験 1 と実験 1E の各々の補完の精度の間の差異は、純粋に「ドナーの候補集合」 $D_{ji4}^{(1)}$, $D_{ji4}^{(1E)}$ (一般に $D_{ji4}^{(1)} \supseteq D_{ji4}^{(1E)}$ であるが、ほぼ 1 の確率で $D_{ji4}^{(1)} \supset D_{ji4}^{(1E)}$ が成り立つ) の差異に帰することができ、同様に実験 2 と実験 2E の各々の補完の精度の間の差異はいずれも、純粋に「ドナーの候補集合」 $D_{ji4}^{(2)}$, $D_{ji4}^{(2E)}$ (一般に $D_{ji4}^{(2)} \supseteq D_{ji4}^{(2E)}$ で、ほぼ 1 の確率で $D_{ji4}^{(2)} \supset D_{ji4}^{(2E)}$ である) の差異に帰することができる。さらに、ほぼ確率 1 で

$$\begin{aligned} M_{i4}^{(1)} &\neq M_{\ell 4}^{(1E)}, & (i \neq \ell, i, \ell = 1, \dots, 1000) \\ M_{i4}^{(2)} &\neq M_{\ell 4}^{(2E)}, & (i \neq \ell, i, \ell = 1, \dots, 1000) \end{aligned}$$

が成立する。

なお、(ケース 4) の下での実験 1、実験 1E の欠損比率の実績は、表 7 に、同じく (ケース 4) の下での実験 2、実験 2E の欠損比率の実績は、表 8 に示した。

5.1 実験 1(E) と実験 2(E) の結果

実験 1 と実験 2、および実験 1E と実験 2E の結果の概略を表 9 に示した。ここでは主にこの表から読み取れる実験結果について考察する。

実験 1 および実験 2 のケース 1 からケース 4 の比較 : 注目すべき結果は次の二点であろう。

まず第一点目として、欠損確率が事業所の産業分類や従業者数規模にかかわらず一様であれば、ケース 1 からケース 3 にかけての欠損確率の水準の変化は、誤差および誤差率に対して大きな影響を与えない、という点である。ケース 1 からケース 3 は欠損確率が事業所の産業大分類や従業者数規模にかかわらず一定であるが、これらケース 1 からケース 3 にかけては欠損確率が 3 倍に増加している。このようなケース 1 からケース 3 の結果を比較すると、興味深いことに、各試行における誤差あるいは誤差率の平均値ならびに標準偏差の 1,000 回の試行全体にわたる平均および標準偏差は大きく変化していないことがわかる。

第二点目として、欠損確率が事業所の産業分類や従業者数規模に応じて変化するケース 4 の場合、表 7 や表 8 に示すように、産業計および従業者規模計の平均的な欠損率が 0.268 と、ケース 2 のそれよりもやや小さくても、誤差の絶対値が、欠損率が一様である場合に比べて有意に大きくなる、という点である。ケース 1 からケース 3 の

⁸ 「欠損事業所集合」を示すこれらの右下添え字の 1 番目は、試行 i における欠損事業所集合であることを示す。同様に右下添え字の 2 番目は、欠損値の発生確率が (ケース 4) である条件の下における欠損事業所集合であることを示す。この 2 番目の添え字が 4 であるのは、実験 1E および実験 2E が (ケース 4) の場合においてのみ行われているからである。

表 7: 事業所ファイルを用いる「実験 1 の (ケース 4)」における産業大分類別・事業所従業員規模別の欠損比率の実績

産業分類	統計量	事業所従業員規模							
		規模計	0～4	5～9	10～19	20～49	50～99	100～199	200～
産業計	標本サイズ	208259	172615	18196	9323	5359	1598	728	440
	欠損比率平均	0.268	0.284	0.234	0.193	0.103	0.103	0.050	0.050
A	標本サイズ	5858	4999	534	241	59	20	4	1
	欠損比率平均	0.292	0.300	0.250	0.251	0.150	0.150	0.046	0.047
	同標準偏差	—	0.006	0.019	0.028	0.048	0.080	0.107	0.212
B	標本サイズ	405	387	16	1	1	0	0	0
	欠損比率平均	0.298	0.300	0.248	0.258	0.170	N/A	N/A	N/A
	同標準偏差	—	0.023	0.106	0.438	0.376	N/A	N/A	N/A
C	標本サイズ	809	688	71	24	19	3	3	1
	欠損比率平均	0.288	0.299	0.252	0.247	0.150	0.143	0.048	0.049
	同標準偏差	—	0.017	0.052	0.085	0.081	0.207	0.123	0.216
D	標本サイズ	21450	15292	2230	1600	1271	563	274	220
	欠損比率平均	0.143	0.150	0.150	0.150	0.100	0.100	0.050	0.050
	同標準偏差	—	0.003	0.007	0.009	0.008	0.012	0.013	0.014
E	標本サイズ	762	551	61	68	47	26	5	4
	欠損比率平均	0.144	0.150	0.150	0.151	0.099	0.102	0.047	0.049
	同標準偏差	—	0.015	0.045	0.043	0.042	0.059	0.095	0.109
F	標本サイズ	26213	21697	2385	1244	654	146	66	21
	欠損比率平均	0.284	0.300	0.250	0.200	0.100	0.100	0.049	0.049
	同標準偏差	—	0.003	0.009	0.012	0.012	0.025	0.026	0.046
G	標本サイズ	45907	36859	5185	2347	1142	259	84	31
	欠損比率平均	0.282	0.300	0.250	0.200	0.100	0.099	0.050	0.051
	同標準偏差	—	0.002	0.006	0.008	0.009	0.019	0.024	0.040
H	標本サイズ	9878	7667	1368	567	227	35	12	2
	欠損比率平均	0.282	0.300	0.250	0.200	0.100	0.103	0.048	0.044
	同標準偏差	—	0.005	0.012	0.017	0.020	0.052	0.064	0.149
I	標本サイズ	21061	17756	1658	834	504	151	87	71
	欠損比率平均	0.284	0.300	0.250	0.200	0.100	0.102	0.050	0.049
	同標準偏差	—	0.003	0.010	0.014	0.014	0.024	0.023	0.027
J	標本サイズ	4128	2792	617	401	204	60	37	17
	欠損比率平均	0.146	0.150	0.150	0.150	0.101	0.099	0.051	0.054
	同標準偏差	—	0.006	0.014	0.017	0.021	0.039	0.036	0.056
K	標本サイズ	41085	35450	2749	1481	958	263	126	58
	欠損比率平均	0.286	0.300	0.250	0.200	0.100	0.101	0.050	0.049
	同標準偏差	—	0.002	0.008	0.010	0.010	0.018	0.019	0.027
L	標本サイズ	36	19	4	3	3	0	1	6
	欠損比率平均	0.127	0.150	0.146	0.162	0.093	N/A	0.045	0.052
	同標準偏差	—	0.080	0.181	0.210	0.166	N/A	0.207	0.090
M	標本サイズ	1664	1448	103	39	44	25	5	0
	欠損比率平均	0.285	0.300	0.248	0.196	0.097	0.098	0.049	N/A
	同標準偏差	—	0.012	0.044	0.062	0.045	0.060	0.097	N/A
N	標本サイズ	13194	12117	714	246	92	19	5	1
	欠損比率平均	0.295	0.300	0.250	0.250	0.149	0.147	0.050	0.040
	同標準偏差	—	0.004	0.016	0.027	0.038	0.081	0.096	0.196
O	標本サイズ	15795	14879	501	227	134	28	19	7
	欠損比率平均	0.296	0.300	0.251	0.248	0.149	0.149	0.049	0.051
	同標準偏差	—	0.004	0.019	0.031	0.032	0.070	0.048	0.083
X	標本サイズ	14	14	0	0	0	0	0	0
	欠損比率平均	0.152	0.152	N/A	N/A	N/A	N/A	N/A	N/A
	同標準偏差	—	0.097	N/A	N/A	N/A	N/A	N/A	N/A

注 1) 産業分類 "P", "Q" の事業所は件数ゼロなので表示を省略。

注 2) 産業計・規模計においては欠損比率の標準偏差の表示を省略。

注 3) 表中の "N/A" は標本サイズがゼロのため計算できないことを示す。

注 4) 「実験 1」では事業所の緯度・経度が欠損しているレコードを除外してあるので、産業計・規模計の標本サイズ 208, 259 は表 3 のレコード数合計 255, 127 よりも除外した分だけ小さい。

表 8: 名寄せ済みファイルを用いる「実験 2 の (ケース 4)」における産業大分類別・事業所従業員規模別の欠損比率の実績

産業分類	統計量	事業所従業員規模							
		規模計	0～4	5～9	10～19	20～49	50～99	100～199	200～
産業計	標本サイズ	207718	172195	18124	9291	5346	1595	728	439
	欠損比率平均	0.268	0.284	0.234	0.193	0.103	0.102	0.050	0.050
A	標本サイズ	5837	4981	533	239	59	20	4	1
	欠損比率平均	0.291	0.300	0.250	0.250	0.150	0.148	0.046	0.048
	同標準偏差	—	0.006	0.019	0.029	0.044	0.078	0.103	0.214
B	標本サイズ	405	387	16	1	1	0	0	0
	欠損比率平均	0.297	0.300	0.248	0.255	0.149	N/A	N/A	N/A
	同標準偏差	—	0.023	0.104	0.436	0.356	N/A	N/A	N/A
C	標本サイズ	809	688	71	24	19	3	3	1
	欠損比率平均	0.288	0.300	0.250	0.247	0.148	0.161	0.050	0.049
	同標準偏差	—	0.017	0.051	0.088	0.081	0.208	0.124	0.216
D	標本サイズ	21381	15251	2213	1592	1268	563	274	220
	欠損比率平均	0.143	0.150	0.150	0.150	0.100	0.100	0.050	0.050
	同標準偏差	—	0.003	0.008	0.009	0.008	0.013	0.013	0.015
E	標本サイズ	762	551	61	68	47	26	5	4
	欠損比率平均	0.144	0.150	0.149	0.150	0.102	0.100	0.049	0.046
	同標準偏差	—	0.015	0.046	0.043	0.044	0.059	0.098	0.105
F	標本サイズ	26207	21692	2384	1244	654	146	66	21
	欠損比率平均	0.284	0.300	0.250	0.200	0.100	0.101	0.050	0.049
	同標準偏差	—	0.003	0.009	0.011	0.012	0.026	0.027	0.047
G	標本サイズ	45767	36732	5182	2339	1140	259	84	31
	欠損比率平均	0.282	0.300	0.250	0.200	0.100	0.100	0.050	0.049
	同標準偏差	—	0.002	0.006	0.008	0.009	0.018	0.025	0.038
H	標本サイズ	9695	7538	1334	556	221	32	12	2
	欠損比率平均	0.282	0.300	0.251	0.200	0.100	0.102	0.053	0.045
	同標準偏差	—	0.005	0.013	0.017	0.021	0.054	0.065	0.149
I	標本サイズ	21057	17753	1658	833	504	151	87	71
	欠損比率平均	0.284	0.300	0.251	0.200	0.100	0.099	0.050	0.051
	同標準偏差	—	0.003	0.011	0.014	0.013	0.024	0.023	0.026
J	標本サイズ	4128	2792	617	401	204	60	37	17
	欠損比率平均	0.146	0.150	0.150	0.151	0.100	0.099	0.050	0.053
	同標準偏差	—	0.007	0.014	0.018	0.021	0.037	0.034	0.055
K	標本サイズ	41022	35393	2746	1480	956	263	126	58
	欠損比率平均	0.286	0.300	0.250	0.200	0.100	0.101	0.051	0.049
	同標準偏差	—	0.002	0.008	0.010	0.009	0.019	0.020	0.029
L	標本サイズ	36	19	4	3	3	0	1	6
	欠損比率平均	0.125	0.147	0.155	0.147	0.096	N/A	0.053	0.052
	同標準偏差	—	0.080	0.181	0.207	0.171	N/A	0.224	0.094
M	標本サイズ	1662	1446	103	39	44	25	5	0
	欠損比率平均	0.285	0.300	0.252	0.199	0.101	0.102	0.051	N/A
	同標準偏差	—	0.012	0.042	0.065	0.044	0.059	0.098	N/A
N	標本サイズ	13191	12115	714	245	92	19	5	1
	欠損比率平均	0.295	0.300	0.250	0.250	0.151	0.146	0.053	0.045
	同標準偏差	—	0.004	0.015	0.027	0.037	0.080	0.102	0.207
O	標本サイズ	15745	14843	488	227	134	28	19	6
	欠損比率平均	0.296	0.300	0.250	0.251	0.150	0.149	0.050	0.048
	同標準偏差	—	0.004	0.020	0.029	0.031	0.067	0.049	0.089
X	標本サイズ	14	14	0	0	0	0	0	0
	欠損比率平均	0.150	0.150	N/A	N/A	N/A	N/A	N/A	N/A
	同標準偏差	—	0.097	N/A	N/A	N/A	N/A	N/A	N/A

注 1) 産業分類 "P","Q" の事業所は件数ゼロなので表示を省略。

注 2) 産業計・規模計においては欠損比率の標準偏差の表示を省略。

注 3) 表中の "N/A" は標本サイズがゼロのため計算できないことを示す。

注 4) 「実験 2」では事業所の緯度・経度が欠損しているレコードを除外してあるので、産業計・規模計の標本サイズ 207,718 は表 4 のレコード数合計 226,426 よりも除外した分だけ小さい。

表 9: 実験 1(E) と実験 2(E) の比較：地理的近接の同一産業大分類事業所による補完誤差 (率) の記述統計

ケース	各試行 における 誤差 (率) の 統計量	誤差				誤差率			
		実験 1		実験 2		実験 1		実験 2	
		1,000 回の試行全体の 平均	標準偏差	1,000 回の試行全体の 平均	標準偏差	1,000 回の試行全体の 平均	標準偏差	1,000 回の試行全体の 平均	標準偏差
ケース 1 (15%)	標本サイズ	31238.8	17.5	31156.8	18.9	30554.3	29.8	30480.6	28.8
	平均値	0.262	0.804	-0.094	0.639	-9.609	1.176	-1.661	0.527
	標準偏差	134.797	7.408	106.793	8.488	202.925	86.043	74.315	67.382
ケース 2 (30%)	標本サイズ	62477.6	29.8	62312.9	31.3	61109.1	41.7	60961.4	40.7
	平均値	0.243	0.640	-0.192	0.485	-9.978	0.879	-1.757	0.434
	標準偏差	137.439	5.028	109.370	5.648	215.860	64.137	91.072	59.956
ケース 3 (45%)	標本サイズ	93716.3	37.7	93467.6	38.6	91664.3	47.5	91438.1	47.9
	平均値	0.161	0.569	-0.266	0.472	-10.566	0.761	-1.886	0.388
	標準偏差	140.143	4.192	111.994	4.674	224.234	50.904	106.284	52.380
ケース 4	標本サイズ	55811.2	75.1	55664.5	73.1	54643.7	78.8	54513.3	76.2
	平均値	-2.561	0.533	-1.092	0.345	-10.241	0.838	-1.414	0.296
	標準偏差	112.811	4.695	76.771	5.813	192.802	58.406	56.908	49.800
ケース 4		実験 1E		実験 2E		実験 1E		実験 2E	
	標本サイズ	55810.6	75.1	55648.1	73.1	54643.0	78.9	54496.9	76.2
	平均値	0.292	0.505	-0.133	0.335	-3.882	0.437	-0.679	0.118
	標準偏差	107.732	4.743	75.415	5.438	93.121	34.230	16.041	23.123

注 1) 表側の「ケース 1」から「ケース 4」は擬似的に発生させる欠損確率を示す。「ケース 1」から「ケース 3」では産業大分類や従業員数規模にかかわらず一定の欠損確率とし、これを括弧内に示す。「ケース 4」では産業大分類や従業者数規模により、表 5 に示すとおりに欠損確率が変化するとした。

注 2) 表側の「標本サイズ」は各試行における欠損事業所数を示す。

注 3) 表側の「平均値」は各試行における補完の誤差 (率) の平均値を示す。たとえば実験 1 のケース 4 では、各試行における誤差の平均値を、1,000 回の試行全体を通して平均すると -2.561、その標準偏差は 0.533、同じく各試行における誤差率の平均値を、1,000 回の試行全体を通して平均すると -10.241、その標準偏差は 0.838 となることを示す。

注 4) 表側の「標準偏差」は各試行における補完の誤差 (率) の標準偏差を示す。たとえば実験 1 のケース 4 では、各試行における誤差の標準偏差を、1,000 回の試行全体を通して平均すると 112.811、その標準偏差は 4.695、同じく各試行における誤差率の標準偏差を、1,000 回の試行全体を通して平均すると 192.802、その標準偏差は 58.406 となることを示す。

場合、誤差の平均値の 1,000 回にわたる試行全体の平均値は、実験 1 では 0.161 から 0.262 の範囲、実験 2 では -0.266 から -0.094 の範囲であるのに対し、ケース 4 の場合、実験 1 では -2.561、実験 2 では -1.0992 と、その絶対値は有意に大きくなっている。

一方で、誤差率は、ケース 1 からケース 3 の場合と、ケース 4 の場合を比較しても有意な差は見られない。

実験 1 と実験 2 の比較：注目すべきは、実験 2 において各試行における誤差率の平均値の 1,000 にわたる平均値と標準偏差について、平均値の絶対値は劇的に減少し、さらに標準偏差も減少している点である。

一方で、誤差については、こうした若干の傾向は見られるものの、その変化の傾向は誤差率のそれに比べ必ずしも明確ではない。

実験 1 と実験 1E、実験 2 と実験 2E の比較：誤差、および誤差率のいずれにおいても、各試行における平均値の、1,000 回の試行全体にわたる平均値の絶対値は有意に小さくなっており、補完の精度が明らかに向上していることが読み取れる。

一方で、各試行における平均値の、1,000 回の試行全体にわたる標準偏差には明確な傾向は現れていない。

誤差と誤差率の比較：注目すべきは次の二点であろう。

まず、誤差率は実験 1(E) および実験 2(E) のいずれのケースにおいても、有意に負値となっているが、誤差の絶対値が有意に正値となるのは実験 2 のケース 4 の場合に限られる、という点である。

つぎに、誤差あるいは誤差率の絶対値が有意に正値となる場合には、すべての場合において、誤差あるいは誤差率が有意に負値となっている点である。これは事業所の従業員数規模の分布が、多数の小規模事業所と、ごく少数の大規模事業所という、事業所規模の形状の特殊性によるものと考えられる。

実験結果のまとめ：以上の結果をまとめると、注目すべきは以下の三点であろう。

第一点目は、実験 1 と実験 1E、あるいは実験 2 と実験 2E を各々比較することにより、従業員規模に関する情報の追加が補完の精度に大きく貢献する点が明らかとなる。これは、調査項目として従業員数における欠損値が多発する場合であっても、その階層に関する情報があれば、補完の精度を向上させることが可能であることを示唆している。この点は、誤差および誤差率のいずれについても当てはまる。

第二点目は、実験 1 と実験 2、あるいは実験 1E と実験 2E を各々比較することにより、事業所に関する情報だけでなく、当該事業所が属する企業の売上高階層などの情報の追加が、やはり補完の制度の向上に貢献することが明らかとなる。この点は、とくに誤差率について当てはまる。

第三点目は、ケース 1 からケース 3 までの結果とケース 4 の結果を比較することにより、欠損値の発生確率が事業所の産業や規模に依存して変化する場合には、それが一様である場合に比べ、補完の精度が悪くなる、という点が明らかとなる。この点は、とくに誤差に当てはまる。欠損値の発生状況が、調査対象の個体属性により変化することは一般的に広く経験されるのは周知の通りである。従って、「ドナーの候補集合」

を限られた情報により設定し補完を行った場合に十分な精度が確保できない恐れがある。こうした場合、上記の第一点目や第二点目の結果は、ある特定の調査項目の数値に欠損が生じている場合であっても、当該項目に関する階層の情報を追加したり、個体が属するグループに関する追加的情報により「ドナーの候補集合」を設定することが、補完の精度を向上させるよい決め手となることを示していると言えよう。

なお、参考までに表 10、表 12 には各々、実験 1E、実験 2E における補完の精度に関する結果を産業大分類別に示した。さらに、表 11、表 13 には各々、実験 1E、実験 2E における補完のドナーとして選ばれた事業所までの地理的距離の状況を産業大分類別に示してある。

表 10、表 12 の補完の精度には、産業大分類の違いによる大きな差異は見られないが、表 11、表 13 の補完のドナーとして選ばれた事業所までの地理的距離の状況には、産業大分類の違いによる差異が明らかに見られる。しかし、この違いは、クラスターを形成するといった産業立地の特性だけでなく、標本サイズの違いにも依存しているようにも見えることから、さらなる分析が必要である。

6 おわりに

本稿ではフィンランドの 2002 年ビジネスレジスターのデータセットを用い、NIM による補完の実験を行った。NIM による補完の基本は、基本的にある特定の項目について欠損値が生じている個体と共通の属性を持つ別の個体集合「ドナーの候補集合」を得て、その集合の中から、欠損値が生じている当該個体と地理的に最近接している個体を選び、その最近接している個体の値を用いるという点である。本稿では、この「ドナーの候補集合」の設定において、補完の精度が追加的な情報により向上することを示すことができた。ここで言う、追加的な情報とは大別して次の二点である。

第一点目は、欠損値が生じている項目の数値情報に替えて、当該数値が属する階層に関する情報である。本稿で確認できたように、従業者数に欠損値があっても、これに替えて従業者数階層の情報が得られるのであれば、補完の精度が飛躍的に向上することが明らかにされた。この点は、調査実施についても次に述べる重要な示唆を与えている。すなわち、数値を記入してもらう調査項目だけでなく、当該数値が属する階層を選択して答えてもらう形式の調査項目を併用することにより、前者のについて欠損が生じて、後者の情報が得られない場合に比べ、より高い精度の補完が可能となる点である。

第二点目、欠損が生じた個体が属するグループに関する情報である。本稿では、事業所の従業者数について欠損が生じた場合、当該の欠損事業所の産業分類に加え、当該の欠損事業所が属する企業の売上高階層の情報を付加して「ドナーの候補集合」を設定することにより、この売上高階層の情報を用いない場合に比べ、補完の精度がより向上することが確認できた。

本稿では、ビジネスレジスターの事業所に関する調査項目である従業者数についての欠損値の補完の精度を検証してきたが、同様の議論は、事業所に関する調査項目に限らず、たとえばある企業に雇用される従業者に関する属性の欠損値の補完についても同様に当てはまるであろう。

以上の点を実際に施すには、調査全体の設計を考慮することが必要である。経験が示すところによれば、欠損値が生じ易い数値項目にはある一定の傾向があることから、そうし

た傾向を持つ数値項目を調査票に盛り込む場合、当該数値項目の補完の精度を確保するに足りるであろう、追加的な情報も得られるような調査の設計が要請されることを、本稿の実験結果は示していると言えよう。

表 10: 実験 1E: 地理的近接の同一産業大分類事業所による補完誤差 (率) の記述統計

同一産業に属す、地理的距離に近い一事業所による補完					
産業	統計量	誤差		誤差率	
		平均	標準偏差	平均	標準偏差
Total	標本サイズ	55810.6	75.1	54643.0	78.9
	平均値	0.292	0.505	-3.882	0.437
	標準偏差	107.732	4.743	93.121	34.230
A	標本サイズ	1707.7	34.5	1688.7	34.3
	平均値	-0.360	2.052	-2.170	1.189
	標準偏差	72.355	24.883	24.362	27.419
B	標本サイズ	120.1	9.3	118.6	9.2
	平均値	-0.038	0.195	-1.393	0.497
	標準偏差	1.809	0.274	4.504	2.168
C	標本サイズ	233.1	12.8	228.7	12.6
	平均値	-0.181	0.843	-2.246	1.352
	標準偏差	9.615	6.928	11.135	16.943
D	標本サイズ	3075.5	48.4	3050.8	48.3
	平均値	0.312	2.050	-1.811	0.592
	標準偏差	106.033	30.533	19.573	23.038
E	標本サイズ	109.9	9.8	90.7	9.1
	平均値	1.543	17.025	-5.807	4.006
	標準偏差	138.960	106.767	30.651	23.176
F	標本サイズ	7436.2	69.0	7423.8	69.0
	平均値	-0.079	0.269	-1.305	0.088
	標準偏差	21.258	9.935	5.698	3.037
G	標本サイズ	12967.7	83.2	12570.1	82.0
	平均値	0.195	0.936	-5.568	1.069
	標準偏差	96.908	8.955	98.364	60.415
H	標本サイズ	2784.9	43.7	2769.2	43.6
	平均値	2.661	5.155	-18.225	4.326
	標準偏差	245.410	10.405	209.978	78.367
I	標本サイズ	5982.6	60.1	5952.1	60.1
	平均値	0.113	2.358	-5.065	2.037
	標準偏差	166.543	14.305	131.036	88.416
J	標本サイズ	600.8	21.5	445.1	19.2
	平均値	4.695	17.509	-11.492	15.616
	標準偏差	404.535	61.937	144.422	300.804
K	標本サイズ	11747.4	80.7	11314.3	80.2
	平均値	0.140	0.402	-1.710	0.279
	標準偏差	40.206	7.849	18.659	21.828
L	標本サイズ	4.5	1.9	3.6	1.7
	平均値	-0.782	43.947	-1.417	3.883
	標準偏差	37.969	73.209	3.076	5.607
M	標本サイズ	474.8	18.3	463.8	18.1
	平均値	0.272	4.072	-2.513	1.323
	標準偏差	71.753	42.287	16.606	17.669
N	標本サイズ	3890.1	51.0	3880.5	50.9
	平均値	0.052	0.342	-1.939	0.374
	標準偏差	19.823	1.966	21.572	9.651
O	標本サイズ	4673.1	57.7	4641.0	57.6
	平均値	0.209	0.761	-1.778	0.813
	標準偏差	46.945	10.094	42.553	37.479
X	標本サイズ	2.1	1.4	2.1	1.4
	平均値	0.028	0.328	-0.308	0.845
	標準偏差	0.268	0.262	0.626	0.843

注) 産業分類 "P","Q" の事業所は件数ゼロなので表示を省略。

表 11: 実験 1E: 地理的近接の同一産業大分類ドナー事業所までの「距離」の記述統計

同一産業に属す、地理的距離が近い一事業所による補完					
産業分類	統計量	平均	標準偏差	最小値	最大値
Total	標本サイズ	55811.2	75.1	55591	56020
	平均値	973.63	20.34	920.13	1036.58
	標準偏差	4403.54	271.16	3610.10	5280.83
A	標本サイズ	1707.8	34.5	1610	1827
	平均値	2551.27	85.29	2331.45	2888.98
	標準偏差	3399.10	314.93	2807.97	4609.51
B	標本サイズ	120.5	9.3	95	151
	平均値	10374.14	1313.80	6789.26	15080.96
	標準偏差	14229.35	3023.86	7900.53	26772.37
C	標本サイズ	233.2	12.8	190	282
	平均値	7563.69	625.60	5917.85	10004.53
	標準偏差	9328.91	2199.36	5934.01	17998.18
D	標本サイズ	3075.5	48.4	2929	3239
	平均値	956.11	62.75	752.60	1199.25
	標準偏差	3516.72	774.79	2089.76	6307.33
E	標本サイズ	109.9	9.8	79	140
	平均値	5522.31	891.87	2831.94	9046.00
	標準偏差	8509.41	2144.77	4395.75	21224.56
F	標本サイズ	7436.2	69.0	7126	7651
	平均値	1029.81	46.98	892.64	1171.90
	標準偏差	3818.06	651.19	2195.04	5667.21
G	標本サイズ	12967.7	83.2	12684	13202
	平均値	677.17	38.50	560.35	804.46
	標準偏差	4073.71	574.18	2856.80	6643.43
H	標本サイズ	2784.9	43.7	2658	2920
	平均値	1238.15	97.05	988.08	1632.27
	標準偏差	4895.15	1121.71	2725.13	9444.45
I	標本サイズ	5982.6	60.1	5780	6238
	平均値	1063.28	50.18	917.40	1285.43
	標準偏差	3625.43	654.74	2340.99	5919.61
J	標本サイズ	600.8	21.5	532	668
	平均値	1426.17	200.96	911.50	2272.39
	標準偏差	4711.75	1271.21	2455.47	11223.86
K	標本サイズ	11747.4	80.7	11488	12032
	平均値	643.27	45.14	534.83	822.12
	標準偏差	4619.63	650.11	2814.78	7526.84
L	標本サイズ	4.6	1.9	0	12
	平均値	19955.39	34315.07	0.00	304270.35
	標準偏差	35897.66	53731.06	0.00	200891.67
M	標本サイズ	474.8	18.3	421	539
	平均値	2874.41	386.87	1849.94	4367.43
	標準偏差	7958.97	1605.37	4353.59	14278.70
N	標本サイズ	3890.1	51.0	3735	4064
	平均値	848.80	61.22	688.01	1126.97
	標準偏差	3915.74	884.25	2430.18	7137.77
O	標本サイズ	4673.1	57.7	4525	4866
	平均値	824.80	50.96	684.68	1033.00
	標準偏差	3403.34	695.04	2199.57	6355.84
X	標本サイズ	2.1	1.4	0	8
	平均値	54465.19	37191.18	0.00	226313.11
	標準偏差	25432.91	26395.76	0.00	98930.23

注) 産業分類 "P", "Q" の事業所は件数ゼロなので表示を省略。

表 12: 実験 2E: 地理的近接の同一産業大分類事業所による補完誤差 (率) の記述統計

同一産業・企業の同一売上階層に属す、地理的距離に近い一事業所による補完					
産業	統計量	誤差		誤差率	
		平均	標準偏差	平均	標準偏差
Total	標本サイズ	55648.1	73.1	54496.9	76.2
	平均値	-0.133	0.335	-0.679	0.118
	標準偏差	75.415	5.438	16.041	23.123
A	標本サイズ	1697.8	34.4	1680.3	34.2
	平均値	0.154	1.682	-0.725	0.319
	標準偏差	62.265	23.627	10.445	8.486
B	標本サイズ	119.7	9.0	118.2	9.0
	平均値	-0.034	0.129	-0.341	0.123
	標準偏差	1.193	0.274	1.257	0.215
C	標本サイズ	231.5	13.0	227.0	12.9
	平均値	-0.241	0.640	-0.552	0.244
	標準偏差	8.047	2.944	2.684	1.748
D	標本サイズ	3065.6	49.9	3041.5	49.6
	平均値	-0.475	1.996	-0.640	0.355
	標準偏差	103.469	30.254	11.377	17.251
E	標本サイズ	108.5	9.4	89.5	8.7
	平均値	-1.363	17.263	-2.632	2.790
	標準偏差	122.340	108.356	17.791	21.159
F	標本サイズ	7437.2	68.8	7425.3	68.7
	平均値	-0.129	0.232	-0.332	0.038
	標準偏差	17.290	8.293	1.951	1.688
G	標本サイズ	12924.9	85.3	12531.3	84.1
	平均値	-0.236	0.739	-0.745	0.039
	標準偏差	77.655	9.851	4.258	0.773
H	標本サイズ	2732.2	45.3	2720.3	45.3
	平均値	-1.071	1.224	-0.594	0.099
	標準偏差	57.688	8.521	4.944	1.855
I	標本サイズ	5980.9	60.5	5950.9	60.3
	平均値	0.323	1.408	-0.551	0.117
	標準偏差	101.283	16.777	7.929	4.615
J	標本サイズ	601.1	22.8	445.4	19.7
	平均値	5.151	17.373	-12.233	14.277
	標準偏差	405.781	60.447	147.326	263.532
K	標本サイズ	11728.9	81.5	11299.6	79.9
	平均値	-0.165	0.309	-0.618	0.036
	標準偏差	28.233	7.640	3.523	1.327
L	標本サイズ	4.4	2.0	3.5	1.8
	平均値	3.434	45.341	-1.549	4.124
	標準偏差	38.056	73.810	3.264	5.715
M	標本サイズ	472.1	18.7	461.2	18.4
	平均値	-0.651	1.320	-0.767	0.487
	標準偏差	24.182	9.309	7.016	6.667
N	標本サイズ	3886.4	49.1	3876.8	49.0
	平均値	-0.154	0.246	-0.478	0.053
	標準偏差	13.109	2.413	3.215	0.724
O	標本サイズ	4655.4	57.9	4624.6	57.6
	平均値	-0.272	0.567	-0.456	0.044
	標準偏差	32.781	10.970	2.712	0.838
X	標本サイズ	1.6	1.2	1.6	1.2
	平均値	0.005	0.292	-0.153	0.552
	標準偏差	0.174	0.216	0.285	0.458

注) 産業分類 "P","Q" の事業所は件数ゼロなので表示を省略。

表 13: 実験 2E: 地理的近接の同一産業大分類ドナー事業所までの「距離」の記述統計

同一産業・企業の同一売上階層に属す、地理的距離が近い一事業所による補完					
産業分類	統計量	平均	標準偏差	最小値	最大値
Total	標本サイズ	55664.5	73.1	55441	55933
	平均値	2244.36	32.54	2136.26	2359.86
	標準偏差	7576.30	305.87	6637.34	9272.91
A	標本サイズ	1700.4	34.4	1592	1835
	平均値	7662.89	275.22	6896.61	8695.54
	標準偏差	10433.10	1263.24	7606.35	18527.13
B	標本サイズ	120.4	9.1	93	153
	平均値	26802.34	3971.41	18949.24	46349.35
	標準偏差	36197.59	9740.40	19770.04	90361.82
C	標本サイズ	233.3	13.1	189	276
	平均値	22544.89	1876.01	17597.31	28894.54
	標準偏差	27934.46	5375.73	14862.16	47991.77
D	標本サイズ	3066.2	49.9	2908	3216
	平均値	2139.26	100.89	1823.15	2609.10
	標準偏差	5576.98	928.01	3808.25	9308.46
E	標本サイズ	109.6	9.5	76	141
	平均値	23199.82	2984.22	14767.07	36816.41
	標準偏差	29309.67	6886.43	16073.20	74337.06
F	標本サイズ	7438.2	68.8	7221	7670
	平均値	2106.71	64.71	1910.01	2314.00
	標準偏差	5413.31	474.27	4148.39	7276.66
G	標本サイズ	12926.0	85.2	12676	13204
	平均値	1303.92	42.10	1167.87	1422.28
	標準偏差	4790.57	383.26	3799.41	6065.33
H	標本サイズ	2733.1	45.3	2560	2894
	平均値	3612.06	186.50	3109.38	4331.58
	標準偏差	9377.53	876.57	7320.66	15584.92
I	標本サイズ	5982.1	60.5	5773	6149
	平均値	2461.26	88.39	2231.38	2853.04
	標準偏差	6559.49	808.76	4806.25	12888.81
J	標本サイズ	601.1	22.8	537	673
	平均値	1805.89	430.84	935.09	3771.55
	標準偏差	9317.78	4824.20	2757.50	29044.39
K	標本サイズ	11729.2	81.5	11454	11994
	平均値	1131.99	48.21	1009.37	1378.51
	標準偏差	5265.21	523.66	3958.80	6901.34
L	標本サイズ	4.5	2.0	0	11
	平均値	20610.31	36627.43	0.00	304270.35
	標準偏差	36166.44	54452.67	0.00	200925.23
M	標本サイズ	474.3	18.8	424	528
	平均値	9798.29	940.08	6467.50	13034.18
	標準偏差	21020.71	2837.35	13480.94	33265.90
N	標本サイズ	3887.6	49.1	3729	4065
	平均値	2178.33	123.12	1811.39	2675.58
	標準偏差	7221.01	722.83	5503.51	9889.72
O	標本サイズ	4657.1	58.0	4474	4835
	平均値	2031.74	108.26	1758.70	2424.38
	標準偏差	7380.71	712.05	5653.65	12009.95
X	標本サイズ	1.6	1.2	0	6
	平均値	67310.09	47207.93	0.00	270019.21
	標準偏差	20160.33	26306.51	0.00	145650.04

注) 産業分類 "I", "Q" の事業所は件数ゼロなので表示を省略。

参考文献

- [1] Andridge, R. and R. Little (2010); “A Review of Hot Deck Imputaion for Survey Non-response,” *International Statistical Review*, **78**(1), pp.40-64.
- [2] Bankier, M., A.-M. Houle and M. Luc (1997); “1996 Canadian Census Demographic Variables Imputation,” *Proceedings of the Survey Methods Section, SSC Annual Meeting, June 1997*.
- [3] Bankier, M., P. Poirier and M. Lachance (2001); “Efficient Methodology within the Candadian Census Edit and Imputation System (CANCEIS),” *Proceedings of the Annual Meeting of the American Statistical Association, August 5-9, 2001*.
- [4] Chen, J. and J. Shao (2000); “Nearest Neighbor Imputation for Survey Data,” *Journal of Official Statistics*, **16**(2), pp.113-131.
- [5] Fellegi, I. and D. Holt (1976); “A Systematic Approach to Automatic Edit and Imputation,” *Journal of the American Statistical Association*, **71**(353), pp.17-35.
- [6] Little, R. and D. Rubin (1987); *Statistical Analysis with Missing Data*, John Wiley and Sons, Inc.
- [7] Little, R. and D. Rubin (2002); *Statistical Analysis with Missing Data*, Second Edition, New Jersey: John Wiley and Sons, Inc.
- [8] Rubin, D. (1976); “Inference and Missing Data,” *Biometrika*, **63**(3), pp.581-592.
- [9] Rubin, D. (1987); *Multiple Imputation for Nonresponse in Surveys*, John Wiley and Sons, Inc.
- [10] Rubin, D. (1996); “Multiple Imputation after 18+ Years”, *Journal of the Americal Statistical Association*, **91**(434), pp.473-489.
- [11] Statistics Canada (prepared by M. Bankier), (1999); “Experience with the New Imputation Methodology Used in the 1996 Canadadian Census with Extensions for Future Censuses,” *Statistical Commission and Economic Commission for Europe Working Paper*, No.24.
- [12] Statistics Canada (prepared by M. Bankier, P. Mason and P. Poirier), (2002); “Imputation of Demographic Variables from the 2001 Canadian Census of Population,” *Statistical Commission and Economic Commission for Europe Working Paper*, No.25.
- [13] Winkler, W. and B. Chen (2001); “Extending the Fellegi-Holt Model of Statistical Data Editing,” *Proceedings of the Annual Meeting of the American Statistical Association, August 5-9, 2001*.
- [14] 高橋将宜・伊藤孝之 (2013); 「経済調査における売上高の欠測値補定方法について」『統計研究彙報』第70号, pp.19-86.

- [15] 森博美・菅幹雄 (2012); 「事業所母集団データベースの更新情報等を活用したレジスター統計に関する研究」『平成 23 年度統計研修所共同研究報告書』総務省統計局統計研修所

ビジネスレジスターの観光統計への応用¹

宮川幸三（慶應義塾大学産業研究所）

Ossi Nurmi（フィンランド統計局）

1. 本研究の目的と概要

2012 年度より日本でもビジネスレジスター（事業所母集団データベース）の運用が開始された。ビジネスレジスターには、全ての事業所をカバーできるという大きな利点がある反面、収集される各企業や事業所の項目は基礎的なものに限定されている。そのため日本におけるビジネスレジスターは、主に、標本調査を実施する際の母集団名簿や標本データから母集団推計を行う際の情報として使用され、『経済統計を正確に作成するための名簿情報の提供及び管理のための重要なインフラ』²とされている。

このような、標本調査を実施する際の補助的な役割に加えて、諸外国では、ビジネスレジスターの情報それ自体に基づいた統計の作成や分析が行われているケースも多い。そのうちの 1 つは、事業所や企業の動態に関する統計の作成である。ここで動態と呼んでいるのは、企業や事業所の参入（誕生）（Birth）や退出（死亡）（Death）、生き残り（Survival）や成長（Growth）等、個別企業や事業所の変化を意味しており、これらの企業・事業所の動態を表した統計は、**Business Demography Statistics**と呼ばれている³。ビジネスレジスターは、常に個別企業や事業所の最新の情報を保持していることから、ビジネスレジスターのみを用いて精度の高い**Business Demography Statistics**を作成することが可能となる。

もう 1 つの方向性として、事業所の立地に関する情報を用いた分析をあげることができる。前述のように、一般的なビジネスレジスターには、各企業や事業所の基礎的な情報のみが収録されているケースが多い。これは例えば、名称、所在地、資本金、従業員規模、売上高、開業年次、産業分類、といった項目である。このうち所在地情報に関しては、地理情報システム（**Geographic Information Systems : GIS**）を合わせて使用することによって、企業や事業所が立地する地域の詳細な情報を入手することが可能となり、それら立地にまつわる様々な変数を分析に利用することができる。例えば **Jarmin and Miranda (2006)**では、アメリカのビジネスレジスターより得られる事業所の所在地情報を用いて、ハリケーンが地域経済に及ぼした影響について分析を行っている。また **Miyagawa et al.**

¹ 本稿における研究の一部は、平成 26 年度文部科学省科学研究費補助金（基盤研究（C）、研究課題：地理情報システムを利用した地域の観光 GDP 推計手法の開発、研究代表者：宮川幸三、研究課題番号：25380270）の助成を受けている。

² 総務省統計局サイト（<http://www.stat.go.jp/data/jsdb/gaiyou.htm>）より引用。同サイトには、ビジネスレジスターの主な機能として、・各種統計調査へ母集団情報を提供、・母集団情報提供の際、調査対象者の負担を平準化（重複是正）、・各種統計調査実施の支援等、の 3 点があげられている。

³ 例えば、各国で比較可能な **Business Demography Statistics** の作成に向けたマニュアルとして、**Eurostat-OECD (2007) "Eurostat-OECD Manual on Business Demography Statistics"**があり、そこでは Birth や Death といった各動態の定義や計算方法の詳細が示されている。

(2014)では、ビジネスレジスターの利用方法の1つとして、事業所の所在地情報を用いて地域の観光客向けの売上額を推定する方法を示し、事業所の所在地情報およびGISの利用によって推定精度が改善されることを明らかにしている。これらの事例は、ビジネスレジスターが、存在する企業や事業所のすべてをカバーしているからこそ可能になる分析であり、企業や事業所の所在地情報を利用した分析は、ビジネスレジスターの重要な役割の1つであると言える。

この他に、ビジネスレジスターを利用して小規模な企業や事業所のデータを推定する試みも行われている。一般的に、小規模な企業や事業所の数は多く、従って統計調査にかかるコストは大きい、生産額や従業者数等のボリュームはそれほど大きくないため、費用対効果の観点から言えば小規模な企業や事業所の統計調査を行うことは大規模のケースに比較して非効率である。また調査に対する回答率に関しても、多くの統計調査では小規模な企業や事業所の回答率が大規模に比較して低い。以上の観点から言えば、小規模な企業や事業所のデータを統計調査を実施することなく収集できれば、費用面や統計の精度面からそのメリットは大きいものになるであろう。Nurmi (2014)では、フィンランドのビジネスレジスターを用いて、小規模な宿泊事業所の延べ宿泊者数を推定する方法が示されており、その推定結果はフィンランドの実際の宿泊統計として公表されている。

このようにビジネスレジスターは、単に標本調査を実施する際の母集団名簿や標本データから母集団推計を行う際の情報としての利用のみならず、多岐に渡って活用されている。ここであげたいいくつかの活用事例のうち、Miyagawa et al. (2014)およびNurmi (2014)は、いずれも観光に関連してビジネスレジスターを利用した推定を行うものである。一般的に、観光客は常に流動的であるため、観光客を対象として調査を行う際には多くの制約がある。一方で観光に関連する事業所や企業は小規模なものも多く、事業所や企業を対象として観光に関連する調査を実施することも容易ではない。特に一国よりも小さな地域を対象として観光の規模を把握することは多くの困難を伴い、従って地域の観光統計調査では、通常の家計や企業・事業所を対象とした調査とは大きく異なる方法によって調査が行われている。しかし、これまでに行われている調査手法が、精度の面で問題を抱えていることも否定できない⁴。このような状況の中で、観光統計に関連してビジネスレジスターを活用した推定手法を導入することは、より精度の高い地域観光規模の把握を達成するための1つの有用な方法になりうるものである。そこで本稿では、上述のNurmi (2014)およびMiyagawa et al. (2014)の詳細を紹介し、ビジネスレジスターの観光統計への活用に関する展望を明らかにする。

以下では、次節および第3節において、それぞれNurmi (2014)およびMiyagawa et al. (2014)の研究内容を紹介した上で、第4節では、ビジネスレジスターの活用に関する今後の展望について簡単に述べている。

⁴ 具体的には、①季節変動を把握することが困難、②詳細な費目について調査を行うことができない、③母集団名簿の欠如、といった問題がある。日本の観光統計の状況と課題や問題点については、宮川 (2009) に詳しい説明がある。

2. ビジネスレジスターの観光統計への活用事例 I

2.1. フィンランドの観光統計と EU レギュレーション

Nurmi (2014)は、フィンランドにおける小規模な宿泊事業所の延べ宿泊者数の推定にビジネスレジスターを活用したものである。内容を紹介するのに先立って、まずはフィンランドにおける観光統計の現状とそれを取り巻く EU の統計レギュレーションを明らかにする。

表 1 は、フィンランドの観光統計の現状を表している。

表 1. フィンランドの観光統計			
	アウトバウンド	国内観光	インバウンド
日帰り	フィンランド 旅行調査	フィンランド 旅行調査 (2018 年～)	国境インタビュー (2012 年)
宿泊 (支払いを伴う)	フィンランド 旅行調査	宿泊統計	宿泊統計
宿泊 (支払いを伴わない)	フィンランド 旅行調査	フィンランド 旅行調査	国境インタビュー (2012 年)

出典：Nurmi (2014)より筆者翻訳。

フィンランドでは、観光関連の統計調査としてフィンランド旅行調査 (Finnish travel survey) が行われている。この調査はサンプル調査であり、15 歳から 84 歳のフィンランド国内居住者を対象として毎月電話調査によって実施されている。サンプルサイズは 2,350 人/月であり、主にアウトバウンドと国内観光に関する日帰り旅行、宿泊旅行の目的地や移動手段、宿泊の形態などを明らかにしている。ただし国内の日帰り旅行者については、2018 年から調査を行う予定となっている。

一方でインバウンド観光に関しては、国境インタビュー (Border Interview Survey) が行われている。この調査は、空港や港、国境検問所などにおけるインタビュー調査として実施されており、インバウンド旅行者の国籍や居住国、訪問理由、目的地、訪問期間、宿泊手段などが調査されている。またいくつかの地域では、観光消費等も尋ねている。

以上は、観光旅行者に対する需要サイドの統計調査であるが、国内観光およびインバウンド観光の支払いを伴う宿泊客に関しては、宿泊統計 (Accommodation statistics) においてホテル等の施設を対象に月次の調査を行っている。調査項目は、部屋数やベッド数、客室稼働率、宿泊の目的、宿泊の価格等であり、国内宿泊客とインバウンド宿泊旅行を分けて調査している。調査対象となる宿泊施設数は、およそ 1,500 施設である。

これら 3 種類の調査に基づいて、フィンランドにおけるアウトバウンド観光、国内観光、

インバウンド観光の実態が網羅的に把握されており、これらの結果を用いて TSA (Tourism Satellite Account) 等の加工統計が推計されることになる。

一方でEU諸国については、観光統計に関するレギュレーション (EU 692/2011) が定められており⁵、その中では以下のように観光統計における宿泊施設の取り扱い方法が定められている。

1. 10 ベッド以上の宿泊施設を統計の対象とする。
2. 当該国の 1 年間の延べ宿泊者数 (人×泊) が EU 全体の 1% よりも小さい場合は、20 ベッド以上の宿泊施設を対象としても良い。
3. 2 番目の基準を適用した国は、調査対象規模以下の毎年の延べ宿泊者数を推定し、提供しなければならない。

フィンランドは、このうち 2 番目の基準が適用される国であるため、前述の宿泊統計では、20 ベッド以上の施設のみを対象として調査が行われている。従って、上述の 3 番目の基準にあるように、20 ベッド未満の宿泊施設に関しては、毎年の延べ宿泊者数を何らかの方法で推定しなければならない。そこでフィンランドでは、この小規模宿泊施設分の推定にビジネスレジスターを利用した方法を用いている。次項以降では、その方法の詳細を明らかにする。

2.2. 小規模宿泊施設の選定

小規模宿泊施設の推定を行うためには、まず当該事業所の規模を把握し、宿泊調査の調査対象であるのか、あるいは小規模施設推定の対象となるのかを選定する必要がある。宿泊施設の規模把握に使用されるソースは以下の 3 つである。

1. 宿泊施設名簿
2. ビジネスレジスター
3. 農業企業の納税データ

1 は、宿泊調査の母集団名簿としても使用されるものであり、常に情報のアップデートを行いながら 20 ベッド以上の宿泊施設約 1,500 施設が含まれている。施設の単位は、地域活動単位⁶ (Local Kind of Activity Unit) であり、各地域活動単位のキャパシティに関する情報とともに連絡先情報が保持されている。また地域活動単位には、共通のビジネスID

⁵ 観光統計に関するレギュレーション (EU 692/2011) については、Eurostat(2014)において詳細な説明がなされている。

⁶ 地域活動単位は、日本の産業統計における「事業所」の概念に類似したものである。ただし地域活動単位の定義が、①単一の主体に所有されており、②単一の場所に立地しており、③主に単一の生産活動を行っていることであるのに対し、日本では複数の生産活動を行っていたとしても、①および②の条件に従っているものは事業所とみなされることになる。

および地域活動単位IDが付されているため、宿泊施設名簿に掲載される宿泊施設については、ビジネスレジスター等の他の行政データに容易にリンクすることが可能である。

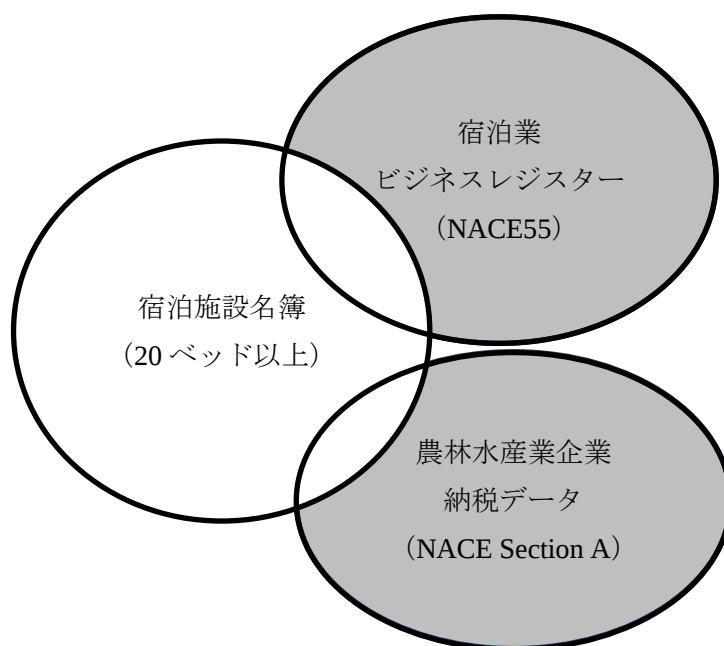
2 のビジネスレジスターは、フィンランドで活動する全ての企業、地域活動単位および企業グループの毎年のデータを収録している。フィンランドのビジネスレジスターでは、NACE⁷ Rev.2 と互換性のあるTOL2008 と呼ばれる標準産業分類に従って地域活動単位が格付けられており、また全ての地域活動単位について売上額の情報も収録されている。情報は毎年 11 月に更新される。

宿泊業は、NACE で言えば、「55. 宿泊業 (Accommodation)」に該当するものであり、1 の宿泊名簿より把握される 20 ベッド以上の宿泊施設と、2 のビジネスレジスターにおける NACE55 に格付けられる地域活動単位を照合することによって、20 ベッド未満の宿泊施設を特定化することができる。このようにして把握される 20 ベッド未満の宿泊施設は、2012 年時点で約 3,500 施設である。

NACE55 に格付けられる宿泊業を主業とする地域活動単位に加えて、フィンランドでは、農林水産業を主業とする企業が宿泊サービスを提供している、いわゆるアグリツーリズムあるいはグリーンツーリズムと呼ばれるようなケースが数多く存在する。このような農林水産業企業の提供する宿泊サービスを把握するために、3 の納税データが使用される。このデータは、NACE で言えば「Section A. 農林水産業 (Agriculture, Forestry and Fishing)」に格付けられる企業の納税データを収集したものであり、その中でこれらの企業が提供した宿泊サービスより得られる収入を分離して申告している。この農林水産業に関する納税データに関しても、1 のベッド数 20 以上の宿泊施設名簿と照合することによって、20 ベッド未満の施設を特定化することが可能となる。言うまでもなく、ここで述べたようなデータソース間の照合作業は、全てのデータに共通のビジネス ID および地域活動単位 ID が存在するために容易に行うことができるものである。

図 1 は、ここで述べた 3 種類のデータの関係を表している。図 1 に描かれる 3 つのデータを照合することによって、灰色部分に該当する 20 ベッド未満の小規模宿泊施設を把握することが、推定のための第 1 段階となる。2011 年の結果としては、ビジネスレジスターの宿泊業 (NACE55) に含まれる小規模施設数が 1,952 (図 1 右上円内の灰色部分に該当)、農業企業納税データに含まれる小規模施設数が 2,475 (図 1 右下円内の灰色部分に該当)であった。

⁷ NACE は、ヨーロッパの標準産業分類 (Statistical Classification of Economic Activities in the European Community) である。詳細については、Eurostat (2008) を参照のこと。



出典：Nurmi (2014)より筆者翻訳。

図 1. 宿泊規模識別のためのデータソース

次の段階としては、把握された 4,427 (=1,952+2,475) の地域活動単位のうち、推計の対象にならない地域活動単位を取り除く作業が行われる。例えば、宿泊業を主業とする企業の地域活動単位であったとしても、本社や営業所等のように、地域活動単位自体が宿泊サービスを提供していないケースについては、これを推計対象から取り除くことになる。また、ビジネスレジスターでは把握されているが宿泊施設名簿では把握されておらず、かつ 20 ベッド以上の規模である施設が存在することもあり、このような地域活動単位も小規模施設推定の対象から除く必要がある。これらの選別作業は手作業で行われ、結果として 3,522 の地域活動単位が小規模推計の対象となる。

2.3. 延べ宿泊者数の推定

延べ宿泊者数の推定に際しては、「同一産業に格付けられる施設である場合、施設自体の規模（ベッド数）に関わらず 1 泊当たりの売上額は一定である」という仮定を置いている⁸。20 ベッド以上の規模を持つ宿泊施設に関しては、3.1 項で述べた宿泊統計より 1 泊当たりの売上額を求めることができる。また全ての宿泊施設は、宿泊業に関する 3 つの NACE クラス（「55.1 ホテルおよび類似の宿泊施設（Hotels and similar accommodation）」、「55.2 休暇施設およびその他の短期滞在宿泊施設（Holiday and other short-stay

⁸ 通常の売上額には、レストラン等の宿泊サービス以外の売上が含まれているため、本来は「1 泊あたり売上額」ではなく「1 泊あたり宿泊料金」を一定として推定を行うべきである。しかし後に示すように、ビジネスレジスターでは各地域活動単位の売上額のためのデータが収録されているため、「1 泊あたり売上額」を一定とするような仮定が用いられている。

accommodation)」、「55.3 キャンプ場, RVパークおよびトレーラーパーク (Camping grounds, recreational vehicle parks and trailer parks)」) のいずれかに格付けられており、このそれぞれの部門について1泊当たり売上額が一定であるという仮定を置くことになる。

表 2 は、推定の詳細を表したものである。

表 2. 規模以下宿泊施設 延べ宿泊者数推定結果

NACE クラス		(1) 地域活動 単位数	(2) 売上額 (百万€)	(3) 1泊 あたり 売上額 (€)	(4) 延べ 宿泊者数 (千人泊)	(5) 非居住者 シェア	(6) 延べ 宿泊者数 居住者 (千人泊)	(7) 延べ 宿泊者数 非居住者 (千人泊)
55.1	ホテルおよび類似 の宿泊施設	237	11.7	84.56	139	23%	107	32
55.2	休暇施設およびそ の他の短期滞在宿 泊施設	3,234	46.7	25.49	1,834	21%	1,451	384
55.3	キャンプ場, RV パ ークおよびトレイ ラーパーク	51	1.1	45.58	25	16%	21	4
合計 (2011 年推定値)		3,522	59.6	29.84	1,998	21%	1,579	420

出典：Nurmi (2014)より筆者翻訳。

表 2 の(1)列は、前項において選定された、推定の対象となる 20 ベッド以下の宿泊施設（地域活動単位）の数を表している。3,522 施設の大半が、NACE クラスで言えば 55.2 に格付けられていることがわかる。(2)列は、ビジネスレジスターおよび農林水産業企業納税データより得られる売上額を、部門ごとに合計したものである。

一方(3)列は、宿泊統計より計算される 20 ベッド以上の施設に関する「1泊当たり売上額」（施設の売上額／延べ宿泊者数）である。ただし NACE クラス 55.2 については、そもそも小規模施設が多く、20 ベッド以上の施設を対象とした宿泊統計ではほとんどその実態を把握できていないため、別途、休暇施設およびその他の短期滞在宿泊施設のための集中予約サービス（centralized reservation services）より得られる情報に基づいて、(3)列の 1泊当たり売上額を計算している。結果として、NACE クラス 55.2 の 1泊当たり売上額がもっとも小さく、55.1 がもっとも大きくなっている。

(4)列は、(2)列の売上額を(3)列の 1泊当たり売上額で割ることによって計算した延べ宿泊者数である。更に延べ宿泊者数は、居住者と非居住者に分割して求められなければならない。そこで(5)列には、宿泊統計より入手される「延べ宿泊者数に占める非居住者のシェア」を掲載している。このシェアを用いて(4)列の値を案分することによって、(6)列および

(7)列において居住者、非居住者別の延べ宿泊者数が推定されることになる。

2.4. 推定手法の評価とまとめ

前項までに説明した推定手法を適用することによって、20 ベッド未満の小規模宿泊施設に関する延べ宿泊者数を、新たに統計調査を実施することなく把握することができる。20 ベッド以上で宿泊統計の対象となる施設の数が約 1,500 であるのに対し、推定の対象となる 20 ベッド未満の小規模施設数は約 3,500 に上っており、ここで紹介した推定手法を適用することによって、調査のコストや記入者負担が大きく削減されることは明らかであろう。

また一般的には、小規模主体による調査への回答率は低いものであるため、ビジネスレジスターを用いた今回の推定手法を適用することによって全ての小規模施設をカバーできることを考えれば、精度面からもメリットがあるだろう。しかし一方で、脚注 8 でも述べたように、推定時に使用される施設ごとの「売上額」には宿泊サービス以外の売上が含まれているため、そのことが推定結果にバイアスをもたらす可能性を否定することはできない。

産業のカバレッジという観点から言えば、今回の手法では、宿泊業を主業とする地域活動単位だけでなく、農林水産業でありながら宿泊サービスを提供するケースも含んで推定を行っている。この点はこの推定手法のメリットの 1 つであるが、一方で、他の産業部門において宿泊サービスを提供しているケースが存在する可能性もあり、これを考慮していない点は本手法の欠点でもある。産業のカバレッジを拡張することは、今後の課題の 1 つとなるであろう。

特に観光統計の場合には、データ周期の問題も重要な要素となる。これは、観光関連の支出に関しては季節変動が激しいためであり、理想的には月次周期でのデータ収集が望まれる。フィンランドにおいても 20 ベッド以上の施設を対象とした宿泊統計では、月次データが収集されている。しかしビジネスレジスターにおいて収録されている売上額データは、月次で更新されるものではない。従って小規模施設の推定結果は、基本的には年に 1 回のペースで改定されることになる。小規模施設に対する統計調査を行わずにこの問題を解決することは困難であり、ここであげたような周期の問題は、ビジネスレジスターを用いる手法の限界であろう。

3. ビジネスレジスターの観光統計への活用事例Ⅱ

3.1. 分析の目的と概要

本節では、ビジネスレジスターの観光統計への活用事例の 1 つとして、Miyagawa et al. (2014)の内容を明らかにする。この研究の目的は、ビジネスレジスターを利用して一国より小さな地域の観光規模を推定する手法を開発することである。

一般的に、地域の観光規模を把握する際には、観光地において観光客を対象として消費額に関する調査を実施するケースが多い。これは言うてみれば、需要サイドからのアプロ

一チである。日本でも、地域観光規模把握を目的とした観光統計として、各都道府県の手によって「観光入込客統計」が実施されている。この統計は、観光庁において定められた「観光入込客統計に関する共通基準」⁹に従って共通の方法によって作成されるものである。共通基準では、観光地点における観光客を対象としたパラメータ調査を通じて観光客の訪問地点数や消費額を調査し、最終的に地域の入込客数や観光消費額を推定する手法が細かく決められている。しかし、費用面の問題からパラメータ調査は四半期に一度行われるのみであり、従って観光消費の季節変動を厳密に把握することはできず、また観光客の回答者負担の観点から詳細な調査を行うことができない、といった問題があることも事実である。

そこで観光庁では、上述のような需要サイドアプローチの問題点を考慮し、供給サイドから観光規模を把握する試みとして 2012 年に「観光地域経済調査」を実施した。「観光地域経済調査」は、宿泊業や飲食業、小売業等、観光に関連する産業部門に格付けられる事業所を対象とした調査である。調査対象事業所は約 9 万事業所にのぼり、国内の 904 の地域を対象として調査を実施している。調査事項は、事業所の従業者数や経営組織、売上額などの一般的な項目に加えて、売上額に占める観光の割合や、観光に関連する事業所の特性（観光協会に加盟しているかどうか、観光ガイドブックに掲載されているかどうか、大型バス用の駐車場があるかどうか、など）等多岐にわたっている。

この調査はサンプル調査であることから、調査の結果を用いて地域内の全事業所の観光向け売上総額を求める、といったような場合には、事業所の母集団名簿、つまりビジネスレジスターを用いて母集団推定を行わなければならない。地域の観光向け売上額は、当該地域の政府や観光関連事業者にとって最低限把握すべき重要な情報であることから、できる限り精度の高い手法によって母集団推定を行うことが望まれる。しかしながら、一般的なビジネスレジスターでは、事業所の産業分類や従業者数、場合によっては生産額といった基礎的な情報が収録されているのみで、観光に関連する事業所の情報を保有しているケースはほとんどない。そのため、母集団推定を行うとしても、産業ごとに分割して推定を行う、従業者数や生産額のデータを用いて規模別の推定や比推定を行う、といった程度の手法を採らざるを得ない。観光向け売上額は、当該事業所の立地によって大きく異なるものであることから、ここで述べたような一般的な手法によって精度の高い観光向け売上額を推定することは困難であろう。

そこで Miyagawa et al. (2014)では、事業所の立地に着目し、ビジネスレジスターより得られる住所情報から割り出される緯度・経度に基づき、GIS を用いて事業所の周辺情報を利用した新たな母集団推定手法を提案している。次項以下では、その詳細を明らかにする。

3.2.ビジネスレジスターを用いた観光売上額推定

⁹ <http://www.mlit.go.jp/kankochou/siryou/toukei/irikomi.html>

図 2 は、一般的に想定される、観光地域経済調査とビジネスレジスターを用いた地域観光売上総額推定のフローを表したものである。

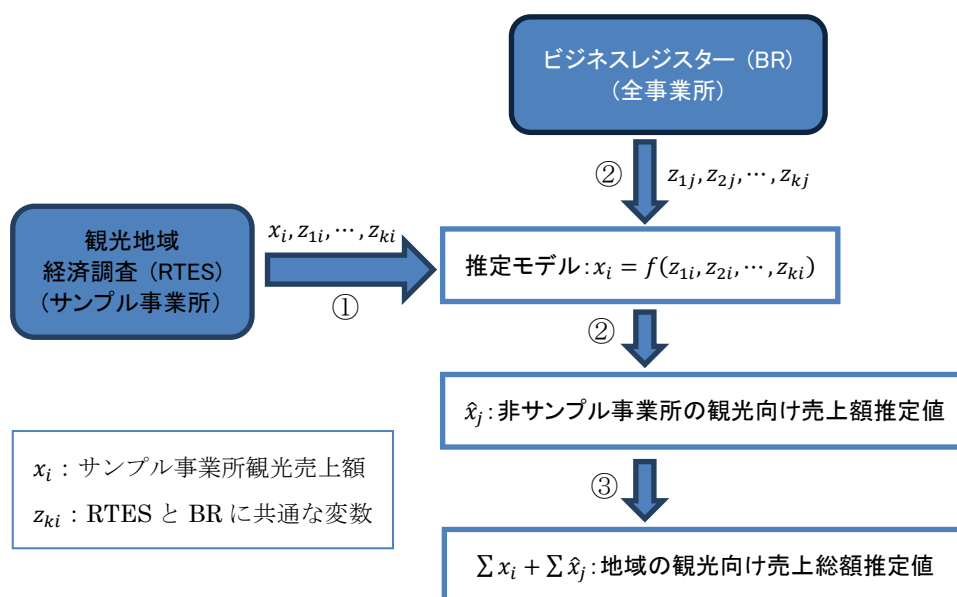


図 2. ビジネスレジスターを用いた地域観光売上推定のフロー

推定は以下のようなプロセスで行われる。

- ① 最初に、観光地域経済調査より得られる個別事業所の観光売上額 x_i と各事業所にまつわる変数 $z_{1i}, \dots, z_{ki}$ を用いて、事業所の観光売上額を推定するモデル $x_i = f(z_{1i}, z_{2i}, \dots, z_{ki})$ のパラメータを決定する。ここで $z_{1i}, \dots, z_{ki}$ は、観光地域経済調査とビジネスレジスターに共通な変数であり、一般的には産業格付けや従業者数、売上額等を想定している。
- ② $z_{1i}, \dots, z_{ki}$ はビジネスレジスターにも収録されている変数であるため、観光地域調査の対象となっていない非サンプル事業所の $z_{1i}, \dots, z_{ki}$ についても、ビジネスレジスターより入手することができる。これを①で構築した推定モデルに代入することによって、非サンプル事業所の観光向け売上額 $\hat{x}_j$ を推定することができる。
- ③ 観光地域経済調査より得られる全てのサンプル事業所の観光売上総額 $\sum x_i$ と、②で推定した非サンプル事業所の観光向け売上総額 $\sum \hat{x}_j$ の合計が、当該地域の観光向け売上総額推定値となる。

ここで重要な問題は、非サンプル事業所の観光向け売上額 $\hat{x}_j$ の推定精度であり、それは $z_{1i}, \dots, z_{ki}$ としてどのような変数を用いるかによって大きく左右されることになる。Miyagawa et al. (2014)では、観光地域経済調査を利用して行う推定シミュレーションを行

い、 $z_{1i}, \dots, z_{ki}$ として事業所周辺の立地特性変数を用いることによって、他の方法に比較して推定精度が飛躍的に高まることを示している。ここで立地特性変数と呼んでいるのは、各事業所の経度・緯度情報に基づき、GIS を用いて得られる事業所周辺の状況を表す変数であり、具体的には、事業所から観光地点や鉄道の駅までの距離等である。小売業や飲食店における観光向け売上を考えれば、観光地点や鉄道までの距離が大きく影響することは明らかであろう。このような立地特性変数は、事業所の住所情報のみから作成することができるため、住所情報を保持するビジネスレジスターより、全ての事業所の観光売上額を推定することが可能となる。次項では、推定シミュレーションの詳細を明らかにする。

3.3.観光売上額推定シミュレーション

図3は、Miyagawa et al. (2014)における推定シミュレーションの手法を示したものである。

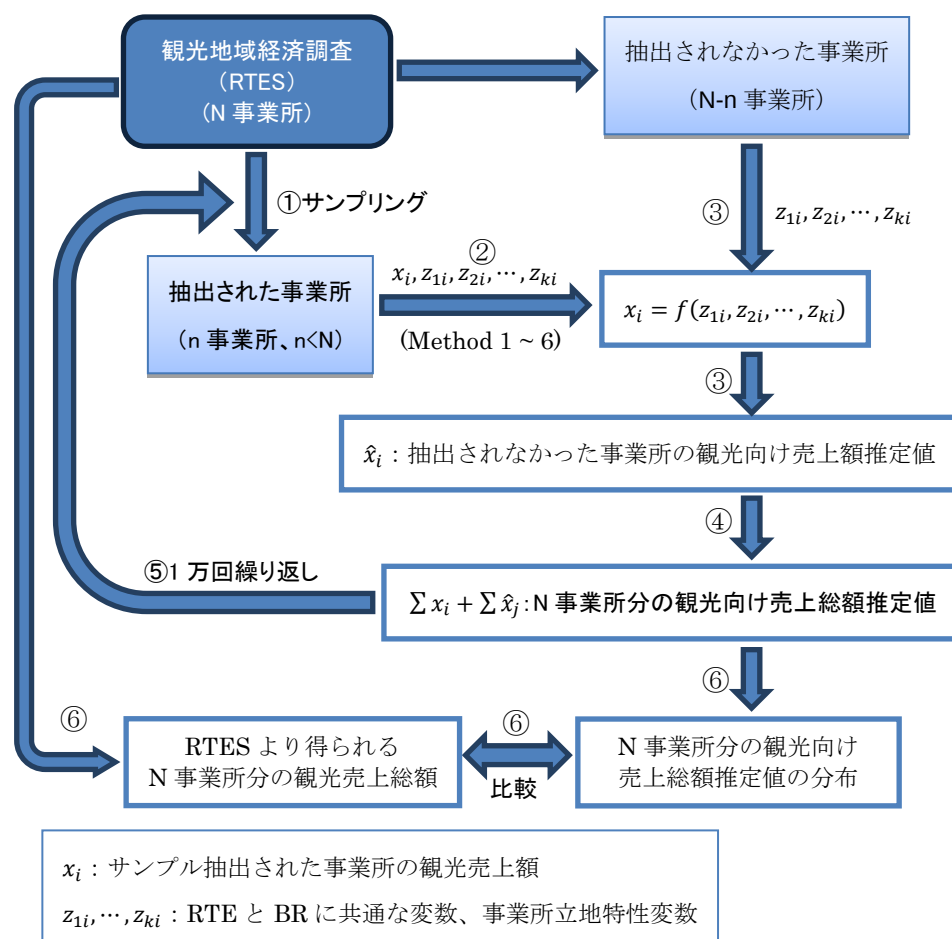


図 3. 推定シミュレーションのフロー

シミュレーションは以下のような方法で行われる。

- ① まず、観光地域経済調査の結果データ（N 事業所分）から無作為抽出を行い、n 事業所分を抽出する。実際の計算では、抽出率は 50%となっている。（ $N=2n$ である。）
- ② 抽出された事業所の観光売上額 x_i と各種変数 $z_{1i}, z_{2i}, \dots, z_{ki}$ を用いて、観光売上額推定モデル $x_i = f(z_{1i}, z_{2i}, \dots, z_{ki})$ のパラメータを決定する。この際には、Method 1～6の異なる 6 つのモデルを用いる。詳細は後に説明するが、Method 1～4 は立地特性変数を用いない手法であり、5・6 は立地特性変数を用いた手法である。論文の目的は、5・6 の手法による推定精度が、1～4 の手法に比較して高いことを示すことである。
- ③ ①で行ったサンプリングにおいて抽出されなかった事業所（N-n 事業所分）について、観光地域経済調査より得られる各種変数 $z_{1i}, z_{2i}, \dots, z_{ki}$ を②で構築した観光売上額推定モデルに代入することによって、抽出されなかった事業所の観光向け売上額推定値 $\hat{x}_j$ を計算する。
- ④ ①で抽出された事業所の実際の観光売上総額 $\sum x_i$ （これは観光地域経済調査のデータそのものである）と、③で推定した抽出されなかった事業所の観光向け売上総額推定値 $\sum \hat{x}_j$ の合計を、観光地域経済調査の対象となった N 事業所分の観光向け売上総額推定値とする。
- ⑤ ①～④のプロセスを、6 つの手法それぞれについて 1 万回ずつ繰り返し行う。これは、推定結果の分布を描くためである。
- ⑥ ⑤の繰り返し計算の結果に基づき、各手法ごとに観光向け売上総額推定値（1 万回分）の分布を描き、観光地域経済調査から得られる N 事業所分の実際の（調査結果としての）観光向け売上総額と比較する。もしも推定結果が観光地域経済調査から得られる売上総額の近くに集中していれば、その手法の推定精度は高いものであると言える。

以上のような方法によって、立地特性変数を用いた推定手法と用いない推定手法の精度を比較することができる。実際の分析では、静岡県の小売業と飲食サービス業のそれぞれについてこのシミュレーションを行い、結果を示している。ここで述べた事業所別観光向け売上額推定の手法（Method 1～6）の詳細は、以下のようなものである。

Method 1

Method 1 は、図 3 の①で抽出された事業所についての 1 事業所あたり平均観光向け売上額を、抽出されなかった事業所の観光向け売上額として使用するものである。

Method 2

Method 2 は、図 3 の①で抽出された事業所についての 1 事業所あたり平均観光向け売

上割合を、抽出されなかった事業所の観光向け売上割合として使用し、観光向け売上額を推定するものである。

Method 3

Method 3 では、被説明変数を観光向け売上割合、説明変数を各事業所の売上額（観光向け以外も含む）としたロジスティック回帰によって観光向け売上額を推定している。¹⁰

Method 4

Method 4 も、ロジスティック回帰の手法を用いたものである。Method 3 と同様に被説明変数は観光向け売上割合であり、説明変数としては事業所の従業者数を用いている。

最終的にビジネスレジスターを用いた母集団推定を行うことを考えた場合、ここで説明変数として使用できる変数はビジネスレジスターに収録されている変数でなければならない。ビジネスレジスターには、通常、売上額や従業者数等の基本的な情報のみが収録されているため、Method 3 や 4 のようなモデルによって観光売上額の推定を試みている。

Method 5

Method 1～4 が立地特性変数を使用しない手法であったのに対し、Method 5 は、立地特性変数を説明変数としてロジスティック回帰を行うものである。立地特性変数とは、具体的には以下のようなものである。

表 3. 事業所立地特性変数リスト

変数名	内容
観光地点	事業所から観光地点までの距離を表す変数。1 つの事業所が複数の異なる規模の観光地点の近隣に立地しているケースもあるため、観光地点リストで把握されている全ての観光地点について、各観光地点の入込客数をその観光地点から当該事業所までの距離で割った値を求め、その合計を変数として使用した。これによって、事業所の距離が遠い観光地点は、この変数の中でほぼ 0 に近い評価をされ、距離が近く入込客数の大きな観光地点ほど大きく評価されることになる。近傍に大規模な観光地点があれば、当該事業所の観光向け売上割合も大きくなると考えられるため、期待されるパラメータの符号は(+)である。
新幹線駅	事業所から最も近い新幹線の駅までの距離を表す変数。新幹線を利用する観光客が駅の近くで消費する可能性があるため、期待されるパラメータの符号は(-)である。
鉄道駅	事業所から最も近い新幹線以外の JR の駅からの距離を表す変数。JR 線を利用する観光客が駅の近くで消費する可能性があるため、期待されるパラメータの符号は(-)である。
高速道路出口	事業所から最も近い高速道路の出口からの距離を表す変数。高速道路を利用する観光客が出口付

¹⁰ ロジスティック回帰については、Dobson and Barnett (2008)などに詳しい説明がある。

	近で消費する可能性があるため、期待されるパラメータの符号は(-)である。
宿泊施設	事業所から宿泊施設までの距離を表す変数。観光地点のケースと同様、1つの事業所が複数の異なる規模の宿泊施設の近隣に立地しているケースもあるため、経済センサスで把握されている全ての宿泊施設について、各宿泊施設の売上額をその宿泊施設から当該事業所までの距離で割った値を求め、その合計を変数として使用した。これによって、事業所の距離が遠い宿泊施設は、この変数の中でほぼ0に近い評価をされ、距離が近く売上額の大きな宿泊施設ほど大きく評価されることになる。近傍に大きな宿泊施設があれば、当該事業所の観光向け売上割合も大きくなると考えられるため、期待されるパラメータの符号は(+)である。

これらの事業所立地特性変数は、事業所の住所から割り出される緯度・経度情報と、一般的に使用されている地図や経済センサス等より得られる事業所周辺の状況のみによって求められるものであるため、ビジネスレジスターを含め、住所情報を保有する全てのデータにおいて上述のような変数を用いた分析を行うことができる。

Method 6

Method 5 は、立地特性変数に加えて事業所の観光に関連する特性変数（以下では観光特性変数と呼ぶ）を説明変数としてロジスティック回帰を行うものである。観光特性変数は、具体的には以下のようなものである。

表 4. 事業所観光特性変数リスト

変数名	内容
観光協会	当該事業所が観光協会に加盟している場合に1をとるダミー変数。
地域の観光パンフレット	当該事業所が地域の観光パンフレットに掲載されている場合に1をとるダミー変数。
観光ガイドブック	当該事業所が観光ガイドブックに掲載されている場合に1をとるダミー変数。
クレジットカード	当該事業所においてクレジットカードを使用できる場合に1をとるダミー変数。
大型バス用駐車場	当該事業所に大型バス用の駐車場がある場合に1をとるダミー変数。
自社ホームページ	当該事業所が自社ホームページを持っている場合に1をとるダミー変数。
観光協会ホームページ	当該事業所が地域の観光協会のホームページに掲載されている場合に1をとるダミー変数。
旅行会社ホームページ	当該事業所が旅行会社のホームページに掲載されている場合に1をとるダミー変数。

これらの観光特性変数に関して、期待されるパラメータの符号はすべて(+)である。

3.4. 分析結果

推定シミュレーションの結果を明らかにする前に、観光地域経済調査で収集された全事業所のデータを用いて、Method 5 および 6 で用いたロジスティック回帰を行った際の推定結果を示す。分析対象は、静岡県内の事業所であり、小売業および飲食サービス業のそれぞれについて被説明変数を観光向け売上割合とした推定を行っている。表 5 が推定結果である。

表 5. 静岡県内全事業所推定結果

部門	飲食サービス業			小売業		
Method	Method 5		Method 6	Method 5		Method 6
観光地点	0.145	(0.000) **	0.063 (0.000) **	0.019 (0.000) **	0.022 (0.000) **	
新幹線駅	0.036	(0.000) **	0.054 (0.000) **	0.063 (0.000) **	0.061 (0.000) **	
鉄道駅	0.112	(0.001) **	0.068 (0.001) **	0.078 (0.000) **	0.045 (0.000) **	
高速道路出口	-0.069	(0.001) **	-0.017 (0.001) **	-0.208 (0.001) **	-0.141 (0.001) **	
宿泊施設	0.130	(0.002) **	0.167 (0.002) **	0.696 (0.002) **	0.719 (0.002) **	
観光協会			0.144 (0.007) **		0.976 (0.004) **	
観光パンフレット			0.290 (0.007) **		0.411 (0.006) **	
観光ガイドブック			0.145 (0.006) **		2.045 (0.005) **	
クレジットカード			0.248 (0.005) **		0.681 (0.004) **	
大型バス用駐車場			1.704 (0.006) **			
自社 HP			0.182 (0.006) **			
観光協会 HP			0.440 (0.007) **			
旅行会社 HP			0.648 (0.007) **			
自由度	668		660	743		739
尤度比インデックス	0.385		0.605	0.436		0.577

1. カッコ内は標準誤差を表す。

2. * および** は、それぞれ有意水準 5% および 1% で有意であることを示している。

3. 定数項を含めて推定を行っているが、表中には結果を表示していない。

この結果を見れば、Method 5 では、飲食サービス部門と小売業の両部門について全ての立地特性変数について有意な結果が得られている。ただし「新幹線駅」と「鉄道駅」に関する予想されるパラメータの符号はマイナスであったのに対し、推定結果は両部門についてプラスとなった。これは、駅から近い事業所ほど観光向け売上比率が低いことを示唆するものである。駅の利用者は観光客だけでなく、むしろ地域の居住者の多くが駅およびその付近の飲食店や小売店を利用しているため、このような結果が得られたと解釈することもできる。一方で Method 6 については、全ての立地特性変数について有意な結果が得られたものの、小売業では大型バス用駐車場やホームページに関する観光特性変数が有意ではなかったため、それらを除いた推定結果を掲載している。しかし、有意であった観光特性変数のパラメータの符号については、全て予想と同じ結果となっている。以上の結果から、ここで用いている立地特性変数や観光特性変数が、事業所の観光向け売上に影響することは明らかとなった。実際に行ったシミュレーションでは、観光地域調査から無作為抽出

した事業所のデータを用いて、表 5 のモデルを推定している。

図 4 および図 5 は、図 3. 推定シミュレーションのフロー図 3 で示した推定シミュレーションを各手法について 1 万回行ったうえで、その分布を描いたものである。

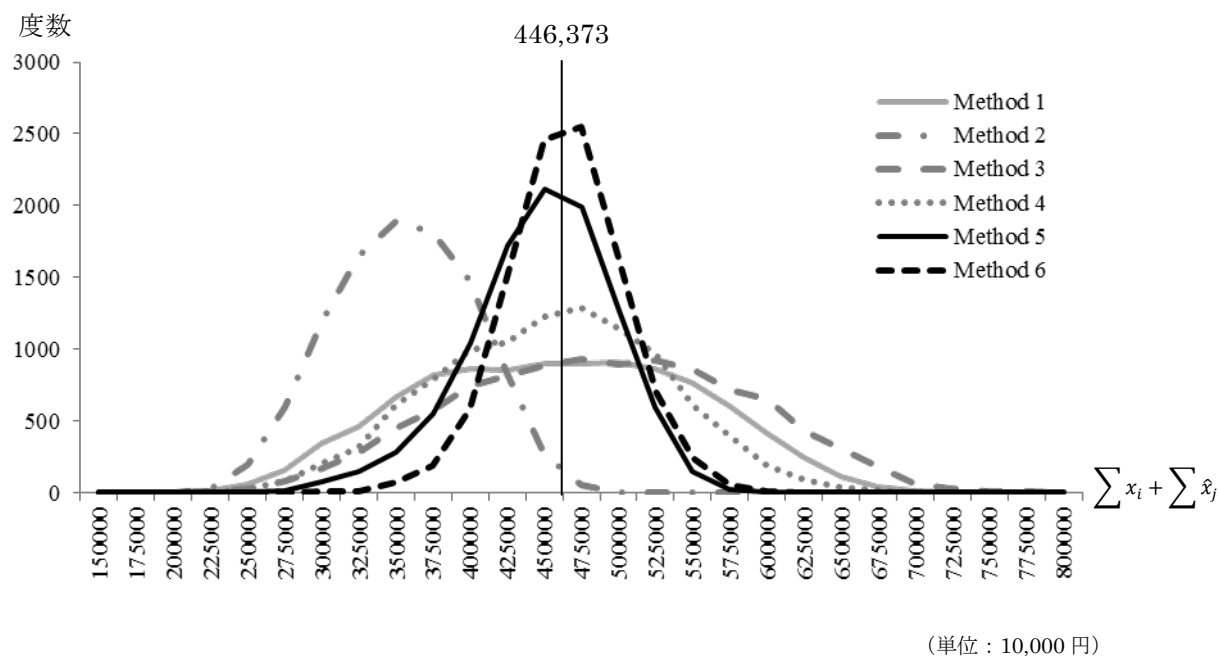


図 4. 飲食サービス業シミュレーション結果

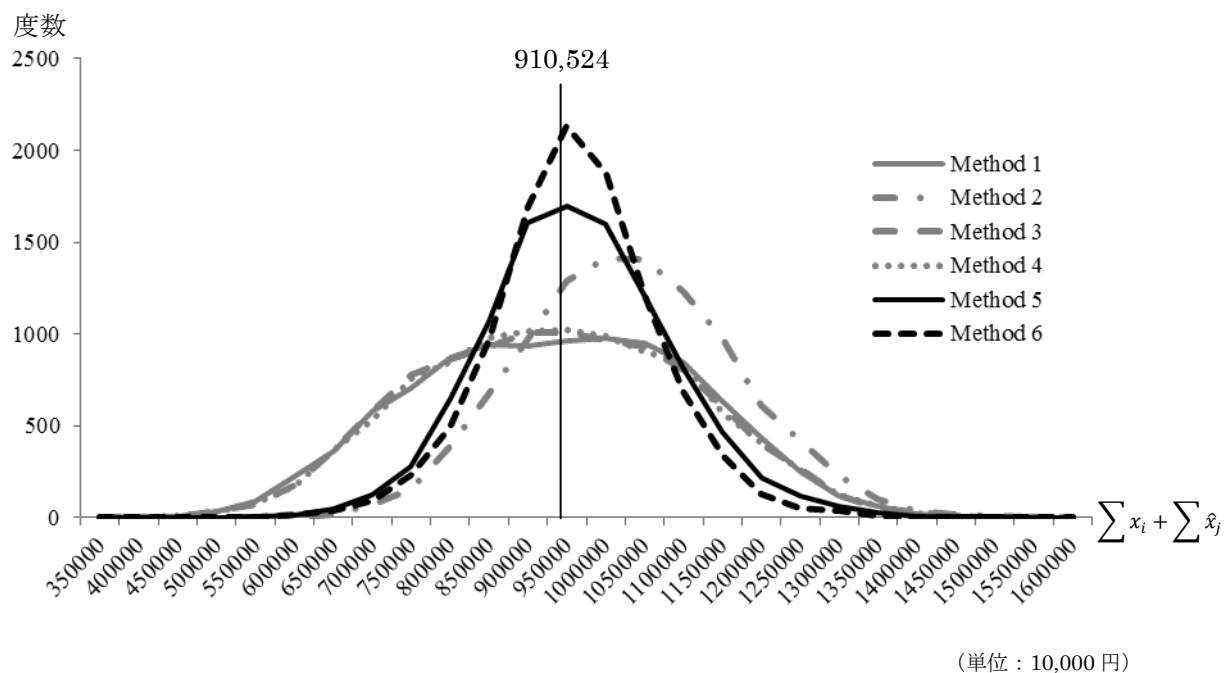


図 5. 小売業シミュレーション結果

図 4 および図 5 では、横軸が推定された観光売上総額を、縦軸が度数を表している。各手法について 1 万回の繰り返しシミュレーションを行っているため、手法ごとの度数の合計はいずれも 1 万である。図 4 の 446,373 および図 5 の 910,524 は、それぞれ飲食サービス業と小売業の観光地域経済調査における全事業所の観光向け売上総額であり、これが推定のターゲットとなる金額である。推定シミュレーションの結果がこの金額に近ければ近いほど、その推定結果は望ましいものであるということになる。

図 4 についていえば、Method 1、3、4 の結果は、類似した分布の形状を描いている。理論的な観点からいえば、Method 1 による推定値は不偏性を持ち、Method 3、4 の推定値は漸近的不偏性を持つ。図 4 においても、Method 1、3、4 の分布の中心は、ほぼターゲットとなる金額（446,373）の近くにあることがわかる。Method 2 は、分布の中心が左方向に大きく偏った形状となっている。Method 2 では、1 事業所あたり平均観光向け売上割合を用いた推定を行っているが、各事業所の観光向け売上割合と売上総額の間に相関がある場合には、このような偏りが発生することになる¹¹。Method 1～4 に比較して、Method 5 および 6 の分布の分散は明らかに小さい。理論的には Method 5 および 6 の推定値も漸近的不偏性を持ち、分布の中心もほぼターゲット金額近傍にあることがわかる。この結果は、事業所の立地特性変数を用いて推定を行った Method 5 および 6 の手法が、それらを用いない Method 1～4 に比較して精度の高い推定法であることを示唆している。

図 5 の小売業の結果も図 4 とほぼ同様であり、Method 1～4 に比較して、Method 5 および 6 が明らかに望ましい推定手法であることがわかる。

表 6 は、各手法の推定値とターゲットとなる観光売上総額（飲食サービス業 446,373、小売業 910,524）の差の絶対値を計算し、その 1 万回分の平均値を計算したものである。

表 6. 推定値と現実値の差の絶対値平均

	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6
飲食サービス業	76,816	104,255	81,852	61,160	38,266	30,569
小売業	145,964	131,396	143,380	141,946	96,116	82,174

値が小さいほど、平均的な推定の誤差が小さいため、より精度の高い推定手法であると判断することができる。この結果からも、Method 1～4 に比較して Method 5 および 6 がより望ましい手法であることがわかる。また、飲食サービス業と小売業の両部門ともに、Method 5 よりも Method 6 の値が小さいことから、立地特性変数だけでなく、観光特性変数も合わせて使用することによって、推定の精度がより一層高まることも明らかにされて

¹¹ Method 2 では、全ての非サンプル事業所について単一の観光向け売上割合を仮定することになるため、例えば、実際には観光以外の売上も含む売上額の大きな事業所ほど観光売上げ割合も高いようなケースでは、図 4 のように左方向に偏った過小推定となる。実際に飲食サービス業について売上額と観光売上げ割合の相関係数を求めたところ、その値は 0.181 であった。

いる。この点は、観光地域経済調査のような観光統計において、観光特性を調査することの重要性を示唆しているといえる。

以上の結果より、住所情報のみから求めることができる立地特性変数を使用することによって、地域の観光向け売上額の推定精度は大きく高まることが明らかとなった。これは、ビジネスレジスターの1つの有効な活用方法を示す事例であるといえよう。

4. 本稿のまとめとビジネスレジスターの活用に関する今後の展望

本稿では、ビジネスレジスターの観光統計への活用事例として、Nurmi (2014)およびMiyagawa et al. (2014)を紹介した。

Nurmi (2014)では、フィンランドのビジネスレジスターを利用して、宿泊統計における小規模な宿泊施設の延べ宿泊者数を推定する手法が示された。このような小規模事業所の推定は、観光統計に限るものではなく、経済センサス等の事業所を対象とした様々な統計調査にも応用できる可能性がある。小規模事業所については、一般的に事業所数が多く、統計調査への回答率も大規模事業所に比較して低いことから、調査コストや調査の精度という面からいえば、ビジネスレジスターの活用によって小規模事業所を対象とした統計調査を省略することができれば、そのメリットは大きいものであろう。

ただしフィンランドの事例において、小規模宿泊施設の推定を行うことができたこと背景には、2つのポイントがある。1つは、フィンランドのビジネスレジスターにおいて、各事業所の売上総額に関する情報が収集されていたことであり、もう1つは事業所の詳細な産業分類が把握されていたことである。Nurmi (2014)においては、ビジネスレジスターより得られる事業所の売上額を、NACEのクラス別の「1泊あたり売上額」で割ることによって延べ宿泊者数を求めていた。これは、「1泊あたり売上額」のレベルが、NACEで言えば同じ「55. 宿泊業」であっても、更に詳細なNACEクラスごとに大きく異なっていたためである。これらの結果は、ビジネスレジスターの活用の際には、事業所の売上額と産業格付けの情報が重要や役割を果たすことを物語るものであった。

日本においても、「平成26年経済センサス・基礎調査」では事業所の年間総売上（収入）金額に関する調査が行われ、その結果は「事業所母集団データベース」に収録されることになる。今後は、「事業所母集団データベース」を利用した小規模事業所の推定に関しても、その可能性を探ってゆく必要がある。

一方でMiyagawa et al. (2014)では、ビジネスレジスターに収録される事業所の住所情報に基づいて、事業所の立地特性に関する変数を導出し、それを利用することによって観光売上額の推定精度が高まることが示された。近年のGISの発展によって、地図から得られる情報は膨大なものであり、それを有効に活用することができれば、観光に限らず様々な分野の統計調査に関してより精度の高い推定を行うことが可能になるであろう。

言うまでもなく、このような事業所立地特性を利用した推定を行うためには、各事業所

の緯度・経度情報がビジネスレジスター上で正確に把握されていることが必要不可欠である。今後は、日本の「事業所母集団データベース」においても、全ての事業所について正確な緯度・経度情報を収集することが望まれる。

参考文献

Dobson, Annette J. and Adrian Barnett (2008) *An Introduction to Generalized Linear Models*, Third Edition, Chapman&Hall.

Eurostat (2008) *NACE Rev. 2 Statistical classification of economic activities in the European Community*, Eurostat Methodologies and Working papers.

Eurostat (2014) *Methodological manual for tourism statistics version 3.1, 2014 edition*, Eurostat manuals and guidelines.

Eurostat-OECD (2007) *Eurostat-OECD Manual on Business Demography Statistics*, European Communities/OECD.

Jarmin, Ron and Javier Miranda (2006) “The Impact of Hurricanes Katrina, Rita, and Wilma on Business Establishments: A GIS Approach”, Center for Economic Studies Discussion Paper CES 06-23.

Miyagawa, Kozo, Hiroyuki Kamiyama, Ryuta Shimamura, Fumikado Yamamoto (2014) “Estimating establishment-level tourism sales using the Regional Tourism Economic Survey and Geographical Information System”, presented in the 12th edition of the Global Forum on Tourism Statistics, <http://tsf2014prague.cz/programme/15-may>.

Nurmi, Ossi (2014) “Estimating Nights Spent in Finnish Accommodation Establishments Based on Business Registers” presented in the 12th edition of the Global Forum on Tourism Statistics, <http://tsf2014prague.cz/programme/15-may>.

宮川幸三（2009）「我が国の観光統計をめぐる現状と課題－地域観光統計体系の整備に向けて－」『産業連関』Vol.17, No.1,2.

研究所報(最近刊行分)

号数	タイトル	刊行年月日
23	民間統計	1997. 01. 31
24	統計環境実態調査	1998. 01. 31
25	ミクロ統計データの現状と展望	1999. 01. 31
26	The2000-01 World Population Census and the Related Issues	2000. 01. 31
27	統計と人権および開発－IAOS 2000 をめぐって	2001. 03. 15
28	第4回日本・中国経済統計学国際会議	2002. 03. 15
29	職安求職者にみる失業の実態	2002. 12. 20
30	国連ミレニアム開発目標と統計	2003. 10. 20
31	Workshops on "the Population Censuses" and "the Use of Census Micro Data"	2003. 12. 20
32	ミクロデータとその利用	2004. 04. 20
33	International Symposia on Population Census and Micro Data Archives	2005. 01. 10
34	政府統計の二次的利用	2005. 04. 20
35	ジェンダー（男女共同参画）統計	2007. 02. 20
36	人口センサスの現状と新展開	2007. 04. 01
37	統計における官学連携	2007. 04. 20
38	ジェンダー（男女共同参画）統計 II	2009. 02. 10
39	社会生活基本調査とその利用	2010. 01. 15
40	地方統計の現状と課題	2010. 09. 15
41	Exploring Potential of Individual Statistical Records	2011. 11. 05
42	観光統計	2013. 02. 05
43	国民経済計算関連統計の新たな展開	2014. 01. 30
44	タウンページデータによる事業所立地分析	2014. 02. 15

研 究 所 報 No. 45

2015 年 3 月 20 日

発行所 法政大学 日本統計研究所

〒194-0298 東京都町田市相原 4342

Tel 042-783-2325,6

Fax 042-783-2332

jsri@adm.hosei.ac.jp

発行人 森 博美

BULLETIN
OF
JAPAN STATISTICS RESEARCH INSTITUTE

No.45

March 2015

Finnish Business Register

CONTENTS

Preface

Finnish Business Register

Hiromi MORI

An Experiment to Select Units for Profiling by Using Business Register Data of
Statsitcs Finland

Mikio SUGA

On an Experiment for Assessing a Method for Imputing Missing Values Using
2002 Finnish Business Register

Tamaki MIYAUCHI

Application of Business Registers for Tourism Statistics

Kozo MIYAGAWA
Ossi NURMI

Edited by
JAPAN STATISTICS RESEARCH INSTITUTE
HOSEI UNIVERSITY
TOKYO, JAPAN