統計数理 第 72 巻 第 2 号

（通巻 140 号）

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

目　次

特集「公的統計―社会の重要な情報基盤―」

「特集 公的統計―社会の重要な情報基盤―」について
　千野 雅人 …… 149
公的統計における総合的品質管理の取組［研究資料］
　川崎 茂・上田 聖 …… 153
我が国における統計ビジネスレジスターの整備
　―社会の重要な情報基盤としての「事業所母集団データベース」―［総合報告］
　髙橋 雅夫 …… 175
労働時間を考慮した日本の消費構造に関する計量分析
　―全国家計構造調査を用いて―［原著論文］
　伊藤 伸介・出島 敬久・村田 磨理子 …… 195
事業所・企業系の公的統計を対象にした合成データの生成技法に関する検討
　―経済センサスを例として―［原著論文］
　伊藤 伸介・横溝 秀始 …… 217
多項ロジットモデルに基づく統計的マッチングの欠測値補完への応用［研究ノート］
　高部 勲 …… 233
公的統計における外れ値への対処とソフトウェア
　―R の MSD 法パッケージの実装について―［統計ソフトウェア］
　和田 かず美 …… 245
公的統計を利用した教育の実践例［統計ソフトウェア］
　佐藤 正昭 …… 261

高次元統計解析で探る銀河の分子ガスの物理状態と天文学への展望［研究詳解］
　竹内 努・矢田 和善・江頭 健斗・青嶋 誠・吉川 耕司・石井 晶・加納 龍生・施 文・
　曹 愛奈・馬 海霞・松井 瀬奈・中西 康一郎・クレ スチェータ・河野 孝太郎 …… 273

2024 年 12 月

大学共同利用機関法人 情報・システム研究機構 統計数理研究所

〒 190-8562 東京都立川市緑町 10-3　　電話 050-5533-8500（代）

本号の内容はすべて https://www.ism.ac.jp/editsec/toukei/ からダウンロードできます

ISSN 0912-6112

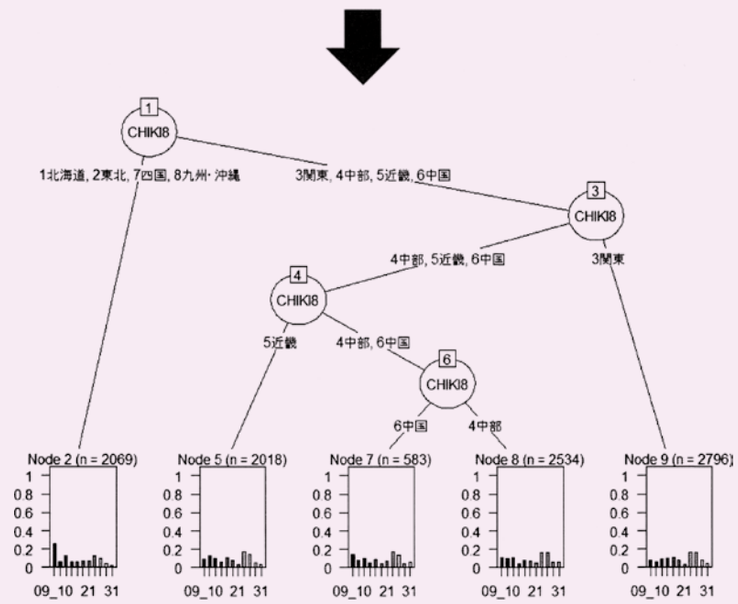
統計数理　PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

第 72 巻 第 2 号 2024

統計数理 Vol. 72, No. 2

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

レコード	地域（合成済）	産業（新規に合成）	…
1	北海道	食品・飲料等製造業	
2	北海道	繊維工業	
3	北海道	木材・家具等製造業	
…	…	…	
5001	関東	電子・電気・情報通信等製造業	
5002	関東	印刷関連業	
…	…	…	



統計数理研究所

統計数理

（年 2 回発行）

編集委員長　朴　堯星
編集委員　小山　慎介
林　慶浩
逸見　昌之
村上　大輔
村上　隆夫
特集担当編集委員　千野　雅人
南　和宏

編集室
池田　広樹　　川合　純華　　長嶋　昭子

「統計数理」は，統計数理研究所における研究成果を掲載する統計数理研究所「彙報」として 1953 年に歴史を始め，1985 年に誌名を変更し今の形となりました．現在は，統計数理研究所の研究活動に限らず，広く統計科学に関する投稿論文を掲載し，統計科学の深化と発展，そして統計科学を通じた社会への貢献を目指しています．

投稿を受け付けるのは，次の 6 種です．

a. 原著論文　　b. 総合報告　　c. 研究ノート
d. 研究詳解　　e. 統計ソフトウェア　　f. 研究資料

投稿された原稿は，編集委員会が選定・依頼した査読者の審査を経て，掲載の可否を決定します．投稿規程，執筆要項は，本誌最終頁をご参照ください．

また，上記以外にも統計科学に関して編集委員会が重要と認める内容について，編集委員会が原稿作成を依頼することがあります．

その他，「統計数理」に関するお問い合わせは，各編集委員にお願いします．

All communications relating to this publication should be addressed to associate editors of the Proceedings.

大学共同利用機関法人　情報・システム研究機構
統計数理研究所
〒190-8562 東京都立川市緑町 10-3　　電話 050-5533-8500（代）
https://www.ism.ac.jp/
© The Institute of Statistical Mathematics 2024
印刷：笹氣出版印刷株式会社

表紙の図は本誌 224 ページを参照

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

Vol. 72, No. 2

Contents

Special Topic : Official Statistics — Critical Information Infrastructure for Society

On the Special Topic “Official Statistics —Critical Information Infrastructure for Society”
Masato Chino ····· 149
Total Quality Management in Official Statistics
Shigeru Kawasaki and Sei Ueda ····· 153
Development of the Statistical Business Register in Japan
—“Establishment Frame Database” as an Important Information Base for Society —
Masao Takahashi ····· 175
Econometric Analysis of Consumption Patterns and Working Hours in Japan: Based on National Survey of Family Income Consumption and Wealth
Shinsuke Ito, Takahisa Dejima and Mariko Murata ····· 195
Study on Synthetic Data Generation Techniques for Official Statistics on Establishments and Enterprises : The Economic Census as an Example
Shinsuke Ito and Shuji Yokomizo ····· 217
Missing Value Imputation Using a Statistical Matching Method Based on Multinomial Logit Model
Isao Takabe ····· 233
Dealing with Outliers in Official Statistics —R Packages Implementing the MSD Estimators —
Kazumi Wada ····· 245
Practical Examples of Education Using Official Statistics
Masaaki Sato ····· 261

Research Review

Exploration of the Physical Properties of Molecular Gas in a Galaxy with High-dimensional Statistical Analysis and Future Prospects to Astronomy
Tsutomu T. Takeuchi, Kazuyoshi Yata, Kento Egashira, Makoto Aoshima, Kohji Yoshikawa, Aki Ishii, Ryusei R. Kano, Wen E. Shi, Aina May So, Hai-Xia Ma, Sena A. Matsui, Koichiro Nakanishi, Sucheta Cooray and Kotaro Kohno ····· 273

December, 2024
Research Organization of Information and Systems
The Institute of Statistical Mathematics
10–3 Midori-cho, Tachikawa, Tokyo 190–8562, JAPAN

統計数理(2024)
第 72 巻 第 2 号 149–151
©2024 統計数理研究所

「特集 公的統計―社会の重要な情報基盤―」について

千野 雅人[†](オーガナイザー)

「公的統計」とは,「行政機関,地方公共団体又は独立行政法人等が作成する統計」のことである(統計法第 2 条第 3 項).政府が作成する統計は,以前は,「官庁統計」や「政府統計」などと称していたが,2007 年に旧統計法を全面改正して制定された統計法により,「公的統計」として明確に定義された.ここで,「行政機関」とは国の府省庁等のことであり,「独立行政法人等」には,独立行政法人のほか,国立大学法人,大学共同利用機関法人,日本銀行,日本年金機構などの公的な法人が含まれる.統計法上は,統計数理研究所や国立大学が作成する統計は公的統計であるが,私立大学が作成する統計は公的統計ではない.この線引きには異論もあるかもしれないが,法令では,用語の定義が明確で紛れないことが重要であり,これにより,時々の為政者や行政担当者による恣意や裁量の余地がなくなるのである.

なお,国の府省庁や都道府県には,統計調査の実施について総務大臣への承認申請や届け出などの厳格な規律が規定されている一方で,統計数理研究所や国立大学には,基本理念に沿って統計を作成する責務は生じるものの,届け出手続きなどの規律は規定されていない.

この特集では,公的統計を「社会の重要な情報基盤」と位置付けたが,これも,統計法に基づくものである.終戦後の 1947 年 3 月に公布された旧統計法では,戦前の公的統計の真実性がおろそかになっていたのではないかという疑念を踏まえ,「統計の真実性を確保し,統計調査の重複を除き,統計の体系を整備し,及び統計制度の改善発達を図る」ことを法律の目的としていた.このため,公的統計と社会・国民との関係や位置付けは必ずしも明確ではなく,公的統計は個別の行政に付随して行われる副次的な活動であるという認識を持つ者もいたと思われる.

一方,データを取り扱う情報技術の進化や国民への浸透とともに,公的統計は,行政のみならず,産業界,学界,国民など広く社会の構成員に利用されるべき情報であり,社会を支える重要な情報インフラである,という認識が広まってきた.このような中で,統計法第 1 条「法律の目的」では,「公的統計が国民にとって合理的な意思決定を行うための基盤となる重要な情報である」という基本的認識を明確に規定した.これは,一見すると当然のことのように見えるが,旧統計法からの流れを見れば,「行政のための統計」から「社会の情報基盤としての統計」への転換を目指した画期的な認識を示すものであると言えるだろう.

また,2014 年 1 月に国連総会で決議された「公的統計の基本原則」(United Nations Fundamental Principles of Official Statistics)でも,「公的統計は,経済・人口・社会・環境の状態についてのデータを政府,産業界及び国民に提供することにより,民主主義社会の情報システムの中で不可欠な要素を構成している.」(基本原則 1)と規定している.このように,国際的に見ても,公的統計は社会に不可欠な情報基盤であることが,共通認識とされていることがわかる.

では,実際には,公的統計はどのように重要なのだろうか?

外国の例を挙げると,米国の国勢調査は,合衆国憲法の規定によって実施が義務付けられて

[†]統計数理研究所:〒 190–8562 東京都立川市緑町 10–3

いる．統計調査の実施根拠が，法律や政令や規則ではなく，憲法なのである．合衆国憲法第 1 章第 2 条第 3 項には，「議員議席及び直接税は，各州の人口に応じて，これを各州に配分する．人口の実地調査は，第 1 回合衆国議会の開会から 3 年以内に，以後は 10 年ごとに，法律の定める方法により行う」旨が規定されている．米国では，この規定に従って，合衆国憲法制定の 3 年後に当たる 1790 年に第 1 回国勢調査を実施し，以後，例外なく 10 年ごとに国勢調査を実施している．また，この憲法の規定から，国勢調査の目的が，国会議員の議席と国家徴収の税金を各州に公正に配分するという，国家が国家として存立するための根源にあることがわかる．実際，米国では，10 年ごとの国勢調査の実施後に，その人口に従って合衆国下院の 435 議席を 50 州に配分し直している．

このような状況は日本でも同様であるが，日本の場合その根拠は法律である．国勢調査の実施は，統計法の規定により政府（総務大臣）の義務となっており，また，衆議院小選挙区の区割りは，選挙関係法律の規定により国勢調査の人口に従って改定される．実際，2020 年国勢調査の人口が 2021 年に公表された後，2022 年に公職選挙法が改正され，衆議院小選挙区の区割りが改定された．その結果，2023 年以降の衆議院選挙では，小選挙区数は，5 都県で 10 増加する一方，10 県で 10 減少することとなった．小選挙区数は議員議席数であるため，とくに，小選挙区数が減少する県では，政党の公認候補者の調整に難航するなど，混乱する様子も見られる事態となった．また，国が徴収した税金を地方に配分する地方交付税の額は，地方交付税法により，国勢調査の人口を始めとする各種統計を基に算出されている．

このほか，消費者物価指数は，日本銀行の物価安定目標や公的年金額の毎年の改定に利用され，全国家計構造調査（かつての全国消費実態調査）は，生活保護基準額の 5 年ごとの改定や人事院による給与勧告に利用されるなど，公的統計の重要な活用事例を挙げれば，枚挙にいとまがない．公的統計がないと，民主主義も国家も企業も私たちの生活も成り立たなくなってしまうと言っても過言ではなく，公的統計は，まさに社会に不可欠な情報基盤となっている．

このような重要な役割を十分に果たすためには，公的統計は，社会の様々な主体から信頼される有用なものでなくてはならない．このためには，急速に進化するデジタル経済，サプライチェーン拡大などの経済のグローバル化，脱炭素社会に向けた活動，観光の振興や課題，働き方の多様化，加速する少子高齢化など，変化する社会経済の姿を的確に把握することができるよう，公的統計を整備していくことが求められる．また，公的統計がその有用性を十分に発揮するためには，アクセスしやすい統計情報の提供，学術研究のための調査票情報オンサイト利用の更なる利便性向上など，利用者の視点に立った統計データの利活用を促進していくことも重要である．そして，そのような取組を確実に進めるためには，的確なマネジメントとプロセスによって質の高い公的統計を作成・提供する体制を確立するなど，総合的品質管理を深化させることが必要である．

一方，この 5 年間には，毎月勤労統計調査と建設工事受注動態統計調査における不適切事案が相次いで発覚し，公的統計への信頼を揺るがす深刻な事態となった．これらは，調査設計や推計方法の変更とそれに伴い必要となる処理や手続きが，適正に行われなかった事案であった．総務省統計委員会（委員長：椿広計統計数理研究所長）は，このような重大事案の再発防止に向けて問題の点検と課題の抽出を行い，その結果を「建議」として取りまとめた．現在，この建議を基に，公的統計の総合的品質の向上に向けた各種の取組が本格化したところである．

これらの状況は，川崎・上田論文において，公的統計の品質の概念や，国際機関と日本における総合的品質管理に向けた取組・動向の一環として，概説されている．また，正確で信頼される公的統計を作成するためには，社会の重要な情報基盤である公的統計の重要な情報基盤として，事業所・企業を対象とする各種統計調査に母集団情報を提供する「統計ビジネスレジスター」の的確な整備が必要である．高橋論文では，日本の統計ビジネスレジスターである「事業

所母集団データベース」の構築の経緯や概要，今後の発展の方向や国際連携への期待などを論じている．

公的統計への信頼を高めるためには，統計利用者からみた有用性の向上も重要である．2007年に全面改正された統計法により，公的統計のミクロデータの一つである「匿名データ」の提供が開始された．さらに，2018年に一部改正された統計法により，国公私立大学等に所属する教員は，オンサイト施設において「調査票情報」（個票データ）を用いた学術研究分析を行うことが可能となった．伊藤・出島・村田論文では，「全国家計構造調査」の調査票情報を用いた独自分析により，夫婦の働き方の違いや配偶者の労働時間が，宿泊費などの費目の消費支出額に有意な影響を与えることを確認している．また，事業所・企業系の公的統計では，匿名データも一般公開型ミクロデータも作成が困難とされているが，伊藤・横溝論文では，「経済センサス活動調査」の調査票情報を用いて，これらを代替する役割を果たすミクロレベル疑似的データである「合成データ」について，各種の作成技法を有用性と秘匿性の両面から定量評価している．このほか，匿名データは，教育の発展に資する目的で利用することができる．佐藤論文では，大学の学部レベルの演習において，匿名データを用いて集計や分析を行うことができるRのパッケージについて解説している．

近年の公的統計の課題の一つが，調査環境の悪化による無回答や無記入・誤記入の増加への対応である．実地調査の現場や周知・広報の段階において，公的統計に対する国民の十分な理解の獲得に向けた努力は行われているが，そのような取組にも限界があり，集計の段階においても，調査回答における欠測値の補完や外れ値への対処を効果的に行うことが重要となる．高部論文では，多項ロジットモデルに基づく統計的マッチングの手法を紹介し，マッチング確率を欠測値の補完に活用する方法について分析を行っている．また，和田論文では，単変量外れ値と多変量外れ値の違いを解説し，MSD（Modified Stahel-Donoho）推定量による多変量外れ値検出法を実装したRのパッケージを紹介している．

最後になるが，本特集の編集に当たり，寄稿していただいた著者の皆さま，査読者の方々，編集委員の南和宏教授（統計数理研究所），編集室の皆さまに，お礼を申し上げたい．これを機に，公的統計に対する理解が進み，その発展に少しでも貢献することができれば，幸いである．

参 考 文 献

内閣 (2023). 公的統計の整備に関する基本的な計画（2023年3月28日 閣議決定）, https://www.soumu.go.jp/main_content/000871085.pdf.

United Nations General Assembly (2014). Fundamental Principles of Official Statistics, https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx.

United States Census Bureau (2021). Census in the Constitution, https://www.census.gov/programs-surveys/decennial-census/about/census-constitution.html.

統計数理(2024)
第 72 巻 第 2 号 153–173
©2024 統計数理研究所
特集「公的統計―社会の重要な情報基盤―」
[研究資料]

公的統計における総合的品質管理の取組

川崎 茂[1,2]・上田 聖[3]

(受付 2023 年 12 月 22 日;改訂 2024 年 2 月 7 日;採択 2 月 13 日)

要　　旨

本稿では,国際機関及び日本における公的統計の総合的品質管理に関する取組について概説する.日本及び世界においてこの分野の取組が体系的に行われ始めたのは 1990 年代以降のことであり,その内容は今日まで段階的に発達してきた.国際的な取組の中では,国連統計委員会の制定した「公的統計の基本原則」及び国際通貨基金(IMF)の「特別データ公開基準」などが世界の国々における公的統計の品質向上に大きな影響を与えた.また,日本では,2007 年に全面改正された統計法が公的統計の品質管理の発展に大きな役割を果たしている.最近,政府の 2 つの基幹統計調査において不適切な処理が行われていた問題が発生し,これを契機として総合的な品質管理の取組が拡充されている.結びに,公的統計の総合的品質管理を推進するために重要な要素について考察する.

キーワード:公的統計,公的統計の基本原則,統計法,総合的品質管理,TQM,国連統計委員会.

1. はじめに

政府が作成する統計は「公的統計」と呼ばれる.公的統計の作成プロセスは,統計という製品を作成するという意味において,工業生産のプロセスと類似している面がある.しかし,公的統計には工業製品とは異なるいくつかの特徴があり,公的統計の品質管理においてはその点に留意する必要がある.以下,その主なものを 3 点挙げる.

(1)品質の要件が定義しにくいこと

品質は,一般に「対象に本来備わっている特性の集まりが,要求事項を満たす程度」(ISO, 2015)と定義される.しかし,公的統計において要求事項を明確に定めることは必ずしも容易ではない.

公的統計の品質における重要な要素の一つは精度である.精度を高くすることは望ましいが,それにはコストがかかるため,一定の制約が生じる.公的統計は「国民にとって合理的な意思決定を行うための基盤となる重要な情報」(統計法第 1 条)であり,多くの利用者が様々な用途で利用することを想定する必要がある.要求される精度は利用目的によって異なり,精度向上の費用対効果を十分考慮する必要がある.統計の精度と他の品質要素(速報性など)の間に

[1] 一般財団法人日本統計協会:〒 169–0073 東京都新宿区百人町 2–4–6 メイト新宿ビル 6 階
[2] 滋賀大学 データサイエンス AI イノベーション研究推進センター:〒 522–8522 滋賀県彦根市馬場 1 丁目 1–1
[3] 総務省 統計局:〒 162–8668 東京都新宿区若松町 19 番 1 号

はトレードオフの関係があるため，どの品質要素をどのような利用に合わせて設定するのがよいか，そのバランスの判断は難しい．また，統計の精度という概念は必ずしも一般の人に理解しやすいものではなく，統計作成機関からの分かりやすい情報提供が必要となる．

さらに，公的統計は，複数の統計を組み合わせて利用されることが多いことから，関連する統計が整備されていれば価値が高まる．ただし，様々な統計が存在しても，それらの間に比較可能性や整合性が欠けていれば，利用者の混乱を招いたり，誤解が生じたりするなど問題が生じる．すなわち，品質は，個々の統計だけでなく，多種多様な統計全体の総合的な体系にも依存すると考えられる．

(2)公的統計は政府が法令に基づいて作成するものであること

行政機関が統計の作成・提供などの業務を行う上では，統計法を始め，様々な法令や行政手続き等に従う必要がある．統計作成に必要なリソースの配分は，各府省内の調整や政府内での査定を通じて決定されるので，統計の品質向上に必要な予算の確保は統計部局だけの判断で決定することはできない．政府の予算は国会の議決により決定され，財政事情によっては統計予算に縮減の圧力がかかりやすい．

また，公的統計は各府省の様々な部門で所管の行政に応じて作成されており，すべての部門で統一的な品質向上の取組を進めるのは容易なことでない．特に統計が専門でない機関は十分な体制が確保できない場合もある．仮に一部の統計において信頼を失う大きな問題が発生すれば，その影響は公的統計全体に波及するリスクがある．

(3)統計作成には国民の協力が必要であること

公的統計の作成における主な「原材料」は，世帯や企業等を対象とする統計調査などを通じて入手される情報である．公的統計の品質向上には，原材料の調達段階から品質の確保が必要となるが，それには国民が漏れなく正確な回答を提出してくれることが前提となる．近年，プライバシー意識の高まりなどから，統計調査への協力度は低下しており，このことは公的統計の品質に大きな影響を与えている．このため，統計調査に関し国民の理解・協力を得ることは品質向上のための重要な要素となる．

また，統計利用のニーズに対応して統計の充実・改善を行うには，統計調査の新設や拡充が必要となることが多い．その場合，回答者の負担が増加するなど，情報収集上の制約もある．国民は統計の利用者であると同時に情報の提供者であるので，このようなメリットと負担のバランスも考慮する必要がある．

このように，公的統計には独特の性質があることから，その品質向上の取組においてはその特徴を考慮に入れる必要がある．品質を広義で捉えるならば，公的統計全般が社会において信頼される有用な情報として提供されるか，という総合的な観点に立つ必要がある．他方，狭義で捉えるならば，個別の統計がその分野のユーザーのニーズを満たす正確で使いやすい情報となっているか，という個別の観点に立つ必要がある．「神は細部に宿る」という言葉があるように，公的統計では，全般的な信頼性，有用性といった広義の品質も，また，個々の統計の正確性，使いやすさといった狭義の品質も必須の要件である．両者は排他的なものではなく，それらが重なり合いながら公的統計の品質の概念を構成している．

初期の品質管理では，主として狭義の品質に焦点が当てられていた．戦後，世界的に標本調査の活用が拡大した時期には，統計数値の誤差の管理が品質に関する重要課題とされた．1980 年代以降の国際的な議論では，統計全般の信頼性，中立性，利用価値などに注目が集まり，徐々に広義の品質に視点が移っていった．1990 年代には，公的統計全体への信頼確保の視点も含められ，品質の高い統計の作成に必要な組織，制度等の基盤整備も視野に入れられるようになった．このような長年の積み重ねを通じて，多面的な品質の概念について共通認識が形成され，

公的統計において総合的品質管理が導入されるようになった．

以下では，まず，国際的な取組として，国連統計委員会及び国際通貨基金(IMF)による国際的な品質向上の取組について紹介する．次に，日本に関して，2007 年の統計法改正を契機とする品質管理の導入の動きを紹介する．さらに，2019 年及び 2021 年に発生した 2 つの基幹統計調査における不適切事案と，それに対応した品質向上の取組について紹介する．最後に，公的統計の品質管理において重要な要素について考察する．

2. 公的統計の品質に関する国際動向

2.1 国連の公的統計基本原則(FPOS)

世界の統計整備に関しては，1946 年に設立された国連統計委員会(United Nations Statistical Commission = UNSC)の場を中心に，世界各国の代表者が統計整備の枠組を検討し，基準や方針等を整備している．UNSC が 1990 年代以前に定めた基準の大半は特定分野の統計整備や統計手法等に関するものであり，統計全体の品質の概念に明示的に触れたものは見られなかった．

1994 年，UNSC は，公的統計の最も基本的な共通規範として，「公的統計の基本原則(Fundamental Principles of Official Statistics = FPOS)」を制定した（United Nations, 1994）．FPOS は，公的統計の品質確保の基礎を与えるものとして大きな役割を果たすものとなった．FPOS の制定の背景には，当時の国際情勢の大きな変化があった．1990 年代前半，計画経済の国々が市場経済に移行し，これに対応して西欧諸国が中心となって FPOS の原型を制定した．計画経済諸国でも公的統計は作成されていたが，その役割は国の経済計画を支えることが主目的であり，国民共通の情報基盤とは見なされていなかった．他方，市場経済下では国民，企業など様々な主体がオープンな情報に基づいて判断し，行動することが大前提となるため，計画経済が市場経済に移行するには，公的統計の役割，要件等の規範を新たに定義する必要が生じた．しかし，当時は市場経済諸国にも明文化された規範が存在しなかったことから，まず 1990 年頃，国連欧州経済委員会で新たな基本原則が制定された．この原則は世界全体においても有意義であると認識され，1994 年，国連統計委員会では欧州とほぼ同じ内容のものが，各国の公的統計が目指すべき規範として採択された．

FPOS は表 1 のとおり 10 項目から成る．その最も基本的な理念は原則 1 に要約されている．そこでは，公的統計は民主社会に不可欠な情報基盤であること，そのため，有用な統計が公正な方法・手続きにより作成され，社会での利用に供されるべきとしている．原則 2 以降もそれを深掘りする形で，異なる視点から公的統計の満たすべき要件を定めている．

FPOS は公的統計の品質に直接言及していないが，第 1 章で述べたように，公的統計の満たすべき要件が明示しにくい中で，このような根幹となる原則が国際的に合意されたことは，世界各国の公的統計の品質管理を推進するための基礎を与えた．

FPOS は，当初，統計専門の機関である UNSC の決議により採択されたが，2014 年には国連総会においてもこれを承認する決議が行われた（United Nations, 2014）．これにより，FPOS

表 1．公的統計の基本原則(項目名)．

原則 1　有用性，公正性，公平性	原則 6　個別データの秘密保護
原則 2　専門性，科学の原理，倫理	原則 7　統計関連法令の制定・公開
原則 3　情報源，作成方法等の公開	原則 8　国内の統計機関の間の調整
原則 4　統計の誤用への対応	原則 9　国際統計基準の使用
原則 5　統計作成上の情報源の選択	原則 10 統計に関する国際協力

(注) ここでは，国連において通常用いられている原則の簡略な名称を意訳して示した(United Nations, 1994).

は世界の政策担当者をも含めた，より広範な関係者における合意事項と位置付けられた．

2.2 国連の公的統計品質保証枠組（NQAF）

FPOS は公的統計の基本理念や原則などを端的に表現しており，各国の統計に関する指針として重要な役割を果たしたが，表現が抽象的であるため，それを実践する方法は必ずしも明確でなかった．FPOS の実践方法の具体化に貢献したのが，UNSC が 2012 年に採択した「公的統計の品質保証枠組（National Quality Assurance Framework = NQAF）」（United Nations, 2012）である．

NQAF の策定に当たっては，この分野で先駆的な取組を行ってきたいくつかの欧米の国が主導的役割を果たした．Lyberg（2012, p.111）によると，1980 年代後半から 1990 年代にかけて，欧米のいくつかの国で公的統計に関する品質マネジメントの導入を検討する動きが始まった．その背景には，厳しい予算制約の中でユーザーニーズにどのように対応すればよいかとの問題意識があった．また，この時期，1987 年に国際標準化機構（ISO）が品質管理の国際規格 ISO 9000 を制定し，総合的品質管理（Total Quality Management = TQM）が標準化されて注目を集めた．このような背景から，公的統計においても TQM への関心が高まり，いくつかの国が TQM の導入に着手した．

1990 年代の後半には，国際的な経済危機の発生に伴い，公的統計の信頼確保の必要性が注目されるようになり，公的統計の品質に関する国際会議が開催されるなど，世界的に統計の品質に関する関心が高まった．さらに 2000 年代に入ると，UNSC において FPOS を踏まえた品質管理の必要性が認識されるようになった．2010 年には UNSC の要請によりカナダ統計局がこの問題に関する国際的なレビューを行い，その結果を UNSC に報告した（United Nations Statistical Commission, 2009）．カナダなどいくつかの国で行っている品質保証の枠組は他の国々にも参考になると考えられ，同様な取組を国際的に進めていくこととされた．

UNSC では，それを踏まえて専門家グループを設置し，2012 年に各国共通の品質保証枠組のひな型を示した NQAF のガイドライン（United Nations, 2012）を作成した．各国は，そのひな型を参考に，それぞれの事情に合わせて修正して独自の枠組を作成することとされた．

NQAF の目次項目は表 2 のとおりである．ここでは，統計の作成・提供の活動全体を A〜D の 4 つのレベルに分けて 19 項目の原則を示し，それぞれについて，「保証すべき要素」を質問文の形で示している．併せて，保証のために採りうる方法を例示している．また，これら 19 項目が FPOS のどの原則に関係しているかも明示されている．

この NQAF は，各国でそれぞれの事情に合わせるよう検討された後，その検討状況を踏まえてマニュアル（United Nations, 2019）が作成された．このマニュアルでは，NQAF の「保証すべき要素（Elements to be assured）」は「要求事項（Requirements）」とされ，品質保証に沿った用

表 2．国連「公的統計品質保証枠組（NQAF）」の項目．

A 統計制度の管理	B 組織運営の環境整備	C 統計プロセスの管理	D 統計結果の管理
1 国の統計制度の調整 2 利用者・データ提供者・その他の関係者との調整 3 統計基準の管理・運用	4 専門的独立性の確保 5 中立性・客観性の確保 6 透明性の確保 7 秘密保護とデータ保全 8 品質保証 9 所要のリソースの確保	10 健全な手法の採用 11 費用対効果 12 適切な統計手続き 13 回答者負担	14 有用性の確保 15 正確性・信頼性の確保 16 遅延のない結果公表 17 アクセスしやすさ，明快さ 18 統計間の整合性 19 メタデータの管理

（出典）United Nations (2014)から筆者が作成.

語に改訂された．また，このマニュアルでは，2015 年の国連サミットで採択された「持続可能な開発目標（Sustainable Development Goals = SDGs）」の実施状況をモニターするために設定された SDG 指標のデータも対象とすることとされた．SDG 指標には多様なものが含まれており，既存の公的統計のほか，自然観測から得られるデータなど様々なものが用いられる．そのようなデータは公的統計以上に品質保証が難しいことから，マニュアルはその点も視野に入れて記述されている．UNSC は，今後，NQAF の未導入国においてもこのマニュアルに沿って統計の品質向上策を効果的に講じることができるよう，支援することを構想している．

2.3 IMF の特別データ公表基準（SDDS）

IMF は国際金融政策のために加盟国から金融・経済分野の様々な統計データを収集し，データベース上に公開している．IMF は，その事業の一環として，加盟各国における統計品質向上の自主的な取組を促すことを目的として，1996 年に各国の統計データの品質に関する情報公開の枠組である「特別データ公開基準（Special Data Dissemination Standard = SDDS）」を公表した．SDDS の初版は現在参照することができないが，その後改訂が行われ，現在は 2013 年版（IMF, 2013）が最新となっている．今日，多くの国が SDDS に従って経済，金融等の統計を公表しており，この枠組は世界の統計の品質向上に寄与してきた．

IMF が SDDS を導入したきっかけの一つには，1994〜95 年のメキシコ経済危機がある（Carson, 2001, p.6）．当時，IMF 理事会では，メキシコ経済危機を誘発した一因は政府統計の作成・公表の不備であったと考え，各国が信頼性の高い統計をタイムリーに公表することにより経済政策の透明性を高めることが必要と判断し，SDDS を設けることとした．

SDDS は，各国が経済，金融等の統計データを公表するに当たって，①データの対象範囲，公表周期，速報性，②アクセスのしやすさ，③データの公正さ，④データの品質の観点から満たすべき基準を示している．ここで「データの品質」とは，個々の統計に関する精度，正確さ，国際基準への準拠等を指すものであり，言わば狭義の品質である．

SDDS に参加するか否かは各国の自主判断に委ねられている．参加を決定した国は，公表する統計が SDDS の基準にすべて準拠していることを約束しなければならない．準拠しているか否かについては，IMF が確認して問題がなければ，その旨が公表される．基準に到達していない場合，必要な改善を行った上で再度 IMF の確認を受け，基準に達すれば SDDS 準拠として公表される．

SDDS に参加することは，その国の統計が一定水準に達していることを示す証しとなり，その国の統計に対する信頼度が高まると期待される．先進国の多くは SDDS 準拠が比較的容易であるが，途上国では改善を要する場合も多く，その場合，SDDS への準拠が目標とされ，統計改善のインセンティブとなっている．

SDDS に未到達の国に対しては，1997 年，基準を若干緩和した「一般データ公表体系（General Data Dissemination System = GDDS）」が用意された．これに参加する各国は，それを出発点として改善を進め，SDDS 準拠を目指すこととされた．

SDDS の求める品質の主要項目は表 3 のとおりである．ここでは，品質に関する要件を 4 つの分野に区分して記述している．

「①データの対象範囲等」では，提供されるべきデータが 4 部門に分けて具体的に列記されており，それぞれについて満たすべき公表周期，公表時期が示されている．「②データへのアクセス等」では，公表時期に関する恣意性をなくすとともに，すべての利用者を公平に取扱うことを求めている．「③データの公正さ」では，統計を作成する機関の中立性が法的に確保されるなど，FPOS を遵守する環境が整備されているほか，統計データの扱いに関する透明性を求めている．「④データの品質」では，統計の作成方法等の情報公開，データの整合性確認の可能性，

表 3. IMF「特別データ公表基準(SDDS)」の主要項目.

①データの対象範囲等	②データへのアクセス等	③データの公正さ	④データの品質
実体経済部門 財政部門 金融部門 国外部門	統計データの公表予定の事前公表 全ての利用者への同時公表	統計機関が FPOS を順守できる環境の整備 政府内の事前情報提供範囲の明確化 改訂，手法変更に関する情報の事前周知	作成方法，情報源等に関する情報源の公開 統計内・統計間での整合性チェックの可能性 国際基準等からの乖離に関する情報提供

(出典) IMF (2013)から筆者が作成.

国際基準との整合性に関する情報の公表などを求めている．参加国は①〜④のすべてを満たすことが望ましいが，国の状況に応じ，必要のない項目や重要でない項目である場合には，理由を付して適用除外としてもよいとされている．

IMF では，SDDS の参加国における遵守状況に関し，基準・規則遵守状況報告(Report on the Observance of Standards and Codes = ROSC)という制度に基づき，第三者による評価を行っている．ROSC は SDDS に限定されたものではなく，IMF が加盟国に遵守を求めている国際金融等に関する 12 分野の基準・規則等について評価を行うものである．ROSC の実施に際しては，調査対象国に対して IMF が事前に書面調査を行い，それを踏まえて専門職員が当該国を訪問してヒアリングを行い，その結果を報告書としてまとめている．日本の公的統計に関する ROSC は，2006 年にマクロ経済統計に焦点を当てて実施された．その報告書 (IMF, 2006)によれば，総論として，日本ではデータの収集，処理，公開が客観的な方法により行われ，専門性と透明性をもって対応が行われていると評価された．ただし，詳細な記述の中では，SNA 統計を担当する職員数が不足していること，統計部局における職員の短期の人事ローテーションが専門性の蓄積の障害となる懸念があることなどが課題として指摘された．これらの課題については，その後，関係機関において改善の努力がなされた．

SDDS 及び GDDS は，その後見直し・改善が行われている．SDDS については，2008 年のリーマンショックによる経済危機の経験を踏まえ，2012 年，内容を拡充した SDDS Plus が上位の基準として設定された．また，GDDS は 2015 年に内容が拡充(enhance)され，現在は e-GDDS と呼ばれている (IMF, 2015)．SDDS Plus は，SDDS と同様の形で基準を定めているが，特に①，②について内容が拡充されている．e-GDDS は SDDS よりも低い要求水準となっているが，参加国は，統計の中・長期改善計画を IMF に提出し，それを公表することが条件とされている．

各国の準拠の状況をみると，2000 年には加盟 183 カ国のうち，SDDS が 47，GDDS が 22，未準拠が 114 であったが，2023 年には加盟 190 カ国のうち，SDDS Plus が 31，SDDS が 47，e-GDDS が 107，未準拠が 5 となっている．IMF による取組は開始から約 30 年が経過し，これにより，各国においてより高いレベルの品質を目指す努力が継続的に進められている．

3. 公的統計の品質に関する日本の動向

日本の公的統計における品質向上の取組は，国際的な動向とよく似た順序をたどって進展した．日本では，1947 年に制定された統計法(以下「旧統計法」)の下で統計の整備・充実が着実に進められたが，品質向上の取組が体系的に行われるようになったのは 2007 年の統計法の全面改正以後のことである．この改正統計法(以下「新統計法」)により，公的統計の満たすべき要件が基本原則として明文化されるとともに，PDCA サイクルを実施するための制度的な基盤が整備された．これを契機として，品質管理の枠組が具体化され，その取組が徐々に進められた．

2010 年代後半には，公的統計の精度や信頼性をめぐる疑義や批判が政府内で生じたほか，不適切な統計処理の事案が発生したことから，品質管理の必要性に関する認識が高まった結果，その取組が本格化した．

この章では，統計法の改正とそれを起点として進められてきた品質向上の取組の展開を紹介する．

3.1 2007 年の統計法改正

旧統計法では，同法の目的を「統計の真実性を確保し，統計調査の重複を除き，統計の体系を整備し，及び統計制度の改善発達を図ること」（第 1 条）と定めていた．当時は「統計の品質」の概念が未確立であり，また，法令上の「整備」や「改善発達」などの文言が具体的に意味するところも明らかでなかったため，同法の下では品質向上を意識した体系的な取組はほとんど見られなかった．

21 世紀に入ると，公的統計が新しい環境により適切に対応したものとなるよう，2007 年に統計法の全面改正が行われた．この改正では，公的統計の満たすべき基本理念が明確に記述され，品質管理を進める上で重要な指針を与えた．以下，新統計法の中で，公的統計の品質に関係の深い 3 つの条項を挙げる．

第 1 条では，この法律の目的を「公的統計が国民にとって合理的な意思決定を行うための基盤となる重要な情報であることにかんがみ，公的統計の作成及び提供に関し基本となる事項を定めることにより，公的統計の体系的かつ効率的な整備及びその有用性の確保を図り，もって国民経済の健全な発展及び国民生活の向上に寄与すること」としている．（下線は筆者による．）これにより，社会における公的統計の位置付け，公的統計の整備に関する要件及び達成すべき目標が，抽象的な記述ながらもより明確にされた．

第 3 条では，公的統計に関する基本理念として次の 4 点を定めている．

①行政機関等の相互協力・役割分担の下で体系的に整備されるべき
②適切かつ合理的な方法で，中立性及び信頼性が確保されるよう作成されるべき
③国民が容易に入手し，効果的に利用できるものとして提供されるべき
④統計作成に用いられた個人又は法人等に関する秘密は保護されるべき

これと同様の考え方は旧統計法にも内在していたと考えることもできるが，新統計法ではこれが明記されたことにより，公的統計の満たすべき要件が明確化された．この内容は FPOS の趣旨にも合致している．

第 4 条では「政府は，公的統計の整備に関する施策の総合的かつ計画的な推進を図るため，公的統計の整備に関する基本的な計画を定めなければならない」と定めている．「公的統計の整備に関する基本的な計画」（以下「基本計画」）の実施状況については新統計法に基づき毎年度フォローアップし，必要に応じて対策が講じられる．そして，政府は公的統計に関して実施した施策の効果を評価した上で，基本計画をおおむね 5 年ごとに見直すこととされている．すなわち，この規定は公的統計の品質向上に関する PDCA サイクルを導入する根拠を与えており，基本計画は公的統計全体に関わる品質管理の重要な要素となっている．基本計画は閣議により決定されるので，政府自身がそれにコミットすることとなる．また，基本計画を策定する時は，関係行政機関と協議するほか，統計委員会の意見を聴き，パブリックコメントを行うこととされている．これにより，政府内だけでなく，統計利用者などを含む多方面の意見が反映される．この手続きは，公的統計が「国民にとって合理的な意思決定を行うための基盤」となるために必要な仕組みと考えられる．

3.2 第 1 期基本計画―品質要素の定義と品質保証の導入

新統計法の下での最初の基本計画(以下「第 1 期基本計画」)は，対象期間を 2009～2013 年度として 2009 年 3 月に閣議決定された．基本計画は，公的統計全体に通じる改善の目標や個々の統計の充実・改善などの方向性を示しており，それ自体が広義の品質管理の中核としての役割を担っている．その中では，公的統計の品質の確保・向上の方策が明記された．具体的には，総務省が中心となって「IMF データ品質評価フレームワーク等を基に，『統計の品質表示のための共通様式』を含めた統計の品質に関する自己評価のためのガイドラインを策定する．この際，作成過程の一層の透明化や，公表期日前の統計情報を共有する範囲・手続等について規定」し，各府省は，所管する統計について，それに基づき「自己評価を計画的に実施」することとされた(閣議決定, 2009, p.59)．これに基づき，2010 年 3 月，「公的統計の品質保証に関するガイドライン」(以下，「品質保証ガイドライン」)が統計企画会議の申合せとして決定された(統計企画会議, 2019)．

このガイドラインでは，公的統計の品質を「正確性のみならず，利用者ニーズの適合性，公表の適時性，統計データの解釈可能性などを含む概念」と規定し，「品質保証」を「利用者ニーズに対応した公的統計の作成・提供，その品質の表示・評価・改善を通じ，公的統計の有用性及び信頼性の確保・向上を目指す」体系的な活動を行うこととしている(p.1)．また，「品質保証」という概念は関係者になじみがないものであると想定されたことから，品質保証とは「製品に瑕疵があった場合に補修する，損害を補償するといった，製品保証のような概念ではない」旨も明記されている(p.1)．

品質保証ガイドラインに示された公的統計の品質要素は表 4 のとおりである．各府省は，これに基づいて品質の自己評価を実施し，それをもとに計画的に品質の改善に努めることとされた．このため，各品質要素に関して，評価の観点及びチェックポイントが詳細に示されている．

また，各府省は，所管する公的統計に関する品質表示の充実を図ることとされ，品質表示を行うべき事項は表 5 のとおりとされた．当面，基幹統計を優先して取り組み，それ以外の統計

表 4．公的統計の品質要素及び定義．

要素		定義
主要要素	ニーズ適合性	社会の様々な主体に広く有効に活用され得る情報基盤として，利用者のニーズを可能な限り満たした統計が作成されていること． (注) 利用者とは，国，地方公共団体，研究者，エコノミスト等に加え，広く一般利用者を想定
	正確性	社会の様々な主体に広く有効に活用され得る情報基盤として，作成された統計が社会経済の実態を可能な限り正しく表していること．
	適時性	作成された統計が利用者のニーズ・作成目的に応じて適時に公表（提供）されていること．
	解釈可能性・明確性	利用者が統計情報を適切に理解し，有効に活用するため，必要な情報が容易に入手・利用できるように提供されていること，及び統計の作成方法（統計データの収集，処理，蓄積，公表の方法・手続）等に関する情報が公表されていること．
補足的要素	信頼性	統計作成過程及び統計作成機関が利用者から信頼されるよう，統計の作成方法が，専門的な見地から決定され，公表されること，及び適切な秘密保護措置が講じられること．
	整合性・比較可能性	関連する複数の統計を用いて分析，地域間比較，時系列比較等を行うことが可能となるように，統計に用いられる概念，定義，分類等の整合が図られていること．
	アクセス可能性	基本的な情報を含め，作成された統計が，利用者のニーズに応じた形で容易に入手・利用できるように提供されていること．
	効率性	費用，報告者負担等の観点から，最も適切な情報源・作成方法によって作成されていること．

(資料) 統計企画会議申合わせ「公的統計の品質保証に関するガイドライン」(平成 22 年 3 月 31 日決定，最終改定平成 31 年 4 月 26 日)

表 5. 公的統計の品質表示事項(調査統計に関するもの).

1 調査の概要 (1)調査の目的 (2)調査の沿革 (3)調査の根拠法令 (4)調査の対象 (5)標本の抽出方法 (6)調査事項 (7)調査票 (8)調査の時期 (9)調査の方法 (10)その他(関係する研究会の審議概要等)	2 調査の結果 (1)用語の解説 (2)結果の概要 (3)集計・推計方法 (4)利用上の注意(統計基準，季節調整法，結果精度，速報・確報間の相違等) (5)正誤情報 (6)統計表一覧 (7)利活用事例 (8)その他(結果の国際比較，過去の結果との比較等)	3 公表予定 4 Q&A 5 問い合わせ先 6 過去の提供情報 7 その他(標語，研究論文，掲載媒体，外国語による情報，調査票情報の保管方法等)

（注）原資料には，調査統計以外の統計についても品質表示事項が掲載されているが，ここでは省略した.
（資料）統計企画会議申合わせ「公的統計の品質保証に関するガイドライン」（平成 22 年 3 月 31 日決定，最終改定平成 31 年 4 月 26 日）

についても可能な範囲で対応することとされた.

このガイドラインについては，以後，各府省の取組，関連する学会の研究成果，国際動向などを参考に，不断の見直しを行うこととされた．この方針に従い，品質保証ガイドラインはその後 2011 年，2016 年，2019 年に改定が行われて現在に至っている.

3.3 第 2 期基本計画―プロセス保証の導入と未諮問基幹統計調査の確認

第 1 期基本計画は，決定の約 4 年後から見直しが行われ，種々の検討・調整を経て 2014 年 3 月に第 2 期基本計画(対象期間：2014～2018 年度)（閣議決定, 2014)が決定された．新たな基本計画では，新たに「品質保証活動の推進」と題する節が設けられ，品質管理について 2 つの新しい事項が盛り込まれた.

1 点目は，品質保証活動における「プロセス保証」の導入である．プロセス保証とは，「決められた手順・やり方どおりに行えば，プロセスの最終アウトプットが目的・基準通りになるための一連の活動」（日本品質管理学会, 2015, p.6)とされる．公的統計については，結果数値や統計表を丁寧に検査・確認しても，十分な精度を満たしているか，各種のニーズを満たしているかなどを判断することは極めて難しい．このため，統計が適正なプロセスに従って作成されることを通じて統計の品質を保証するものである.

2 点目は，「未諮問基幹統計調査」に関する実態の確認及び課題等の検討を行うことである．未諮問基幹統計調査とは，長年にわたり統計委員会に諮問されていない基幹統計調査のことである．新統計法では，基幹統計調査を実施する行政機関がその調査を変更しようとする場合には，予め総務大臣の承認を得なければならないと定めている．その際，総務大臣は，原則として統計委員会に諮問してその意見を聴くこととされている．この手続きにより，統計調査の変更を専門家がチェックし，必要に応じて改善の提言が行われる．この手続きは，旧統計法における指定統計調査に関する手続きが継承されたものであり，第三者によるチェックを通じて，旧指定統計調査を引き継ぐ基幹統計調査の品質向上を促すものである．しかし，長年にわたり変更がなかった基幹統計調査はこの手続きを経ていないため，そのような統計調査に対しても一定のチェックを行う必要があると考えられたため，このことが基本計画に盛り込まれた.

これにより，2015 年度には，家計調査，人口動態調査など 5 本の基幹統計調査について，また 2016 年度には，賃金構造基本統計調査及び建築着工統計調査について統計委員会によるレビューが行われた．その結果，各統計調査の課題や改善方策等が明らかにされ，それに沿って改善が進められた.

3.4 第 3 期基本計画―経済統計に関する疑義・批判への対応と品質評価の導入

第 2 期基本計画が施行されてから間もない 2015 年頃から，政府内で GDP を中心とする経済統計の精度等に関し疑義が示され，批判などが行われるようになった．これに対応して，基本計画は当初の予定から 1 年前倒しで改定され，2018 年 3 月には，対象期間を 2018〜2022 年度とする第 3 期基本計画（閣議決定, 2018）が決定された．従前は統計の品質について主に自己評価に委ねられていたところ，新たな基本計画には第三者評価が導入され，政府全体として品質保証活動を推進する取組が盛り込まれた．

以下，その経緯と内容について順を追って解説する．

2015 年 10 月 16 日に開催された経済財政諮問会議における麻生太郎副総理（兼財務大臣）の発言が発端となり，政府内において経済統計の精度等に関し議論が活性化した．麻生副総理は，家計調査と商業動態統計の動向の乖離や毎月勤労統計の標本入れ替え時に生じるギャップなどについて疑問を呈し，統計の改善の必要性を訴えた（経済財政諮問会議, 2015, pp.9–10）．これに対し，高市早苗総務大臣は，麻生副総理の疑問に答え，GDP の精度向上のために基礎統計の充実と，GDP 作成におけるそれら基礎統計の活用方法の開発が必要であることを指摘した（経済財政諮問会議, 2015, pp.10–11）．統計の精度をめぐる議論がハイレベルな政治の舞台で行われたことにより，GDP 及びその基礎統計を始めとする公的統計の品質向上が政府の重要課題の一つとして位置付けられた．

これを契機として，政府内において公的統計の精度向上に対する関心が高まり，新たな検討の場が設けられ議論が行われた．2016 年 12 月には経済財政諮問会議において「統計改革の基本方針」（経済財政諮問会議, 2016）が決定され，それに基づいて設置された統計改革推進会議は 2017 年 5 月に提言を取りまとめた（統計改革推進会議, 2017）．その中では，政府が EBPM の推進と統計改革を一体として進めていくという方針を提示し，主要な柱として，①EBPM 推進体制の構築，②GDP 統計を軸にした経済統計の改善，③ユーザー視点に立った統計システムの再構築と利活用推進，④報告者負担の軽減と統計業務・統計行政体制，業務効率化，基盤強化が示された．その一環として，個々の統計の評価を行うために，「統計棚卸し」を実施することとされた．その具体的な内容は明示されていないが，統計の品質の確保・向上の観点から既存の統計全般についての事後的なモニタリングを行うこととされた．その趣旨や方向性は品質保証ガイドラインの推進を後押しするものとなった．

このような背景の下，2018 年 3 月には，2018〜2022 年度を対象期間とし，「統計改革の基本方針」を反映した第 3 期基本計画（閣議決定, 2018）が決定された．また，同年 6 月には，基本計画に基づく公的統計に関する PDCA サイクルの強化を図るため，統計法の改正が行われ，統計委員会の機能強化（①基本計画の進捗に対する統計委員会の勧告権の付与，②統計委員会の建議権の付与，③統計委員会をサポートし各府省において所管統計の品質面も含め総括する統計幹事を統計委員会に新設）が措置された．

第 3 期基本計画に基づいて進められた主な品質管理活動には，次のものがある．

- 品質保証ガイドラインの運用を通じた統計作成過程の更なる透明化
- 各統計について，精度向上，業務効率化，報告者負担軽減等の観点から第三者によるモニタリングの実施（「統計棚卸し」）
- 各統計の精度に関する情報の公表状況の統一的な検査（「見える化状況検査」）
- 政府全体を通じて統計の品質面，技術面等の改革を支援するチームの設置
- 統計委員会に「統計業務プロセス部会」と「評価分科会」を新設

「統計棚卸し」は，2018 年 11 月に統計委員会に新設された統計業務プロセス部会が中心となって行うこととなった．その取組内容は，3〜5 年の間にすべての統計について，統計の企

画，実査，審査・集計，提供・利活用の各段階に共通の視点を設けて統計の業務プロセスを点検するものであった．また，統計精度に関する標準的な検査である「見える化状況検査」もこれと一体的に行うこととされた．このような取組を通じて，品質に関する第三者評価が徐々に進められた．

3.5 毎月勤労統計調査の不適切事案と第 3 期基本計画の変更

第 3 期基本計画が実施される中，重要な基幹統計調査の 1 つである毎月勤労統計調査（以下「毎勤統計」）において不適切な処理が行われていた事案が発生した．この問題は，最初，2018 年末に新聞で報道され，2019 年 1 月，厚生労働省がこの事案について正式に公表した．（以下，「毎勤統計事案」．）

毎勤統計の結果となる賃金動向は，当時，政治的に高い注目を集めていた．また，毎勤統計の数値は雇用保険の給付額の算定に用いられていることから，この事案は大きな政治問題となった．このため，統計委員会が中心となって再発防止及び統計の品質向上のための方策を建議として取りまとめた．その方策は，総合的品質管理（TQM）の考え方に基づくものであり，政府において実行に移されるとともに，それを第 3 期基本計画に反映するために同計画の部分的な変更が行われた．

以下，その経緯について解説する．

毎勤統計は，厚生労働省が毎月実施する基幹統計調査である．常時 5 人以上を雇用する約 33,000 の事業所を対象として，そこに勤務する労働者の雇用状況，給与及び労働時間等を調査している．毎勤統計事案の問題点を要約すると，次のとおりである（「統計」編集部, 2019）．

⑴ 従業者規模 500 人以上の層は悉皆調査を行うものとして設計されていたが，東京都の当該層において回収困難な事業所が増加したため，厚生労働省は，2004 年からこの部分を抽出調査に変更した．その際，母集団復元のために抽出率の逆数を推計乗率として適用すべきところ，そのための集計システムの修正を怠った．これにより，それ以降の結果数値に偏りが生じることとなった．

⑵ このような標本設計の変更については，統計法に基づき，調査計画の変更として総務省に承認の手続きを行うべきであった．しかし，厚生労働省はそれを行っておらず，また，このような変更を行っていた事実をホームページ等で公表していなかった．

⑶ この調査では，数年ごとに標本の抽出替えが行われ，その際に断層が生じるという問題があった．その改善のために，厚生労働省は 2018 年に「ローテーションサンプリング」の手法を導入した．その際，⑴の推計方法の問題を是正した．この変更により，時系列に段差が生じたが，厚生労働省は推計方法の変更について対外的に公表していなかった．

⑷ 不適切な集計が行われていた時期に集計に用いられた基礎データが保存されていなかったため，その時期の数値を遡及して再計算することが困難となった．

この調査結果は雇用保険，労災保険などの支給額の決定に用いられており，この不適切処理によって 2004 年以降の保険受給者に対し，過去にさかのぼって追加の給付を行うことが必要となった．追加給付に係る経費（事務経費などを含む．）として約 795 億円が特別会計の予算に計上された（日本経済新聞電子版, 2019）．

なお，毎勤統計事案は，統計委員会において第 3 期基本計画に基づき統計改善の取組について確認・評価を行う過程で明らかになったものである．その意味では，基本計画に基づくチェック機能がこの問題の発見を促したものと見ることができる．

厚生労働省は，この問題の実態を解明するため，統計の専門家，弁護士等の外部有識者で構成される「毎月勤労統計調査等に関する特別監察委員会」を設置して調査を行った．その結果，

表 6. 2019 年統計委員会建議における基本的な視点.

① 品質はプロセスで作り込む 事後的な検査，外部からの監察・評価には限界がある．プロセスの中での品質保証に注力することが王道であり，また，最も効果的である．
② 透明性を確保する 統計の仕様・品質に関する情報の開示は，適切な統計利用及び利用者からの信頼確保に不可欠である．
③ 継続的に PDCA サイクルを回す 統計作成プロセスの中で PDCA（Plan-Do-Check-Act）のサイクルを回すことにより，不断の品質改善に取り組む必要がある．
④ 業務記録の保存を徹底する 業務の遂行を適切に管理する上でも，そして，PDCA サイクルを的確に回すためにも，業務記録を通じた再検証が不可欠であり，記録の保全はその前提条件である．
⑤ 必要な業務体制を整備する 統計の品質の確保・改善は，精神論だけでは実現できない．高い専門知識を有する人材・組織体制，その適切な運営・管理が必要である．
⑥ 府省間でノウハウ，リソースを有効活用する 府省間の比較で判明したグッド・プラクティスの共有に努めるとともに，優れたノウハウ，リソースを有する機関の協力を得て改善に取り組む．
⑦ ガバナンスを確立する 上記の確実な実行を保証するため，トップマネジメントが責任を持って取り組むこととし，その取組を可能な限り可視化する．

（資料）統計委員会（2019）p.2

2019 年 1 月から 2 月にかけて 2 つの報告書が公表された（厚生労働省 毎月勤労統計調査等に関する特別監察委員会, 2019a, 2019b）．

また，2019 年 1 月，毎勤統計事案を受けて，他の統計調査においても同様の問題が発生していないか政府全体で確認することとされ，総務省が中心となって各府省による一斉自己点検が行われた．その結果，毎勤統計事案と同程度の重大な不適切処理は発見されなかったが，影響度は小さいながら多数のエラー等が発見された．

これを踏まえ，統計委員会は 2019 年 2 月以降，全府省の統計調査及び業務体制について点検検証を進め，2019 年 6 月，毎勤統計を含むすべての統計調査に共通する総合的な対策として「公的統計の総合的品質管理を目指した取組について」（統計委員会, 2019）を建議として取りまとめ，総務大臣に提出した．この建議では，「総合的品質管理（TQM）」の考え方に沿って統計作成プロセスの継続的な改善を進め，万が一のエラー発生時にはその影響を最小化することを目指すこととされた．この建議における基本的な視点は表 6 のとおりである．

建議で提案された主な取組は次のとおりである．

(1) PDCA サイクルによるガバナンスの確立を図る観点から，各統計調査について，調査計画の履行状況等を点検・評価を行い，必要に応じて業務マニュアルの修正，調査計画の改定，調査の見直し等の措置を講じる．なお，この提案に沿って，2020 年 7 月，「PDCA サイクル確立に向けた点検・評価ガイドライン」（統計行政推進会議, 2020）が全府省の統計幹事による申合せとして決定され，同年 10 月から運用が開始された．

(2) エラーの発生防止及び早期発見の観点から，統計の結果を調査担当から独立した立場で分析的審査を行うために，「統計分析審査官」を新設する．なお，この提案に沿って，緊急措置

として，2019 年 7 月から 31 人の統計分析審査官が各府省に配置され，その任に当たった．

⑶誤りが発見された場合に重大な事態に至ることを防ぐとともに，再発防止等の適切な対策が講じられるよう，組織として明確な対応ルールを定める．なお，この提案に基づき，2020 年 6 月，府省共通の対応ルールのひな型が示され，それを基に各府省は順次対応ルールを定めて運用を開始した．

⑷統計作成者は，結果数値等の誤りが判明した場合に備えて，政策部局等による統計の利活用状況を予め把握し，万一の場合に備え，利用者への連絡ルール等を定める．

⑸建議に示された品質管理の取組を着実に進めるため，必要な人員，組織体制，職員の専門性確保を図る．なお，これを踏まえて 2021 年度から人員，組織体制等の強化が進められた．

また，この建議の公表後の 2019 年 9 月には，統計部門だけでなく政策部門とも連携した総合的な対策を検討するため，内閣官房長官を議長とする「統計改革推進会議」の下に各方面の学識者等からなる「統計行政新生部会」が設置され，さらなる検討が行われた．その結果は，2019 年 12 月に取りまとめられた（統計改革推進会議 統計行政新生部会, 2019）．この報告では，統計委員会による建議を踏襲しつつ，公的統計に対する国民の信頼確保のためには，各府省の統計部局だけでなく，政策部局等においても信頼確保の取組を進めるとともに，官房などでリソースの確保や人材育成などに携わる幹部職員等にも十分に配慮すべきことを指摘している．報告に含まれる主な項目は次のとおりである．

⑴公的統計の統計作成プロセスが当然満たすべきと判断される「要求事項」を定め，それに基づき統計作成プロセスの第三者による診断に着手し，作成プロセスの水準を段階的に向上させる．これについては，2020 年 11 月，統計委員会に統計作成プロセス部会が設置され，同部会が中心となって「要求事項」が取りまとめられた．それに基づいて「統計作成プロセス診断」が行われるようになった．

⑵統計作成の要所要所で，専門知識を有する職員が適切な判断を下すことができるよう，政府内に統計専門資格「統計データアナリスト」等を新設し，各府省においてその人材を育成し，必要な部署に有資格者を配置する．

統計委員会の建議及び統計行政新生部会の報告の内容については，その取組が確実に実施されるよう，2020 年 6 月に第 3 期基本計画の一部を変更する閣議決定が行われた（閣議決定, 2020）．基本計画は，対象期間を通常 5 年程度としているところ，決定から約 2 年という短い時間で変更したことは，公的統計の PDCA サイクルの根幹となる基本計画の枠組が柔軟に運用されていることを示している．

3.6 2021 年の不適切事案と統計委員会建議

前述の取組が進められていた途上の 2021 年 12 月，建設工事受注動態統計調査（以下，「受注統計」）における不適切な処理の事案（以下，「受注統計事案」）が発覚した．この受注統計は国土交通省が毎月実施する基幹統計調査であり，建設事業者等に対し建設工事の受注高等の報告を求めるものである．この事案は，毎勤統計事案からわずか 2 年後に発覚したものであることから，再び社会的に大きな注目を集めた．

この問題についても統計委員会が中心となって再発防止策を検討し，建議を取りまとめた．建議の検討においては，毎勤統計事案において講じられた方策の有効性等が点検された．その結果，毎勤統計事案における各種の対策は引き続き有効であることが確認され，受注統計事案を通じて新たに発見された課題を踏まえて追加の対策が立案された．

以下，建議の検討経緯や概要について解説する．

受注統計事案を要約すると，次のとおりである．

⑴受注統計が開始された 2000 年当時から，一部の調査対象が，当月の調査票の提出とともに，過去に未提出だった調査票を期限に遅れて提出する事例があった．その際，国土交通省は，年間合計値の精度を確保する観点から，遅れて提出された月の値を当月の値に合算し，その値を当月の調査票に修正記入して集計していた．

⑵2003 年から，回収率の低下に対応するため，受注高の合計値に回収率の逆数を乗じる「欠測値補完」を導入した．これにより，前記の合算処理と欠測値補完が併存することとなった．過月分の受注高が算入されているのに加えて，欠測値補完が行われたことにより，遅延提出分の数値が合計値に二重に計上される形となった．これは，通常「二重計上」の問題と呼ばれている．

⑶2019 年，国土交通省は二重計上が生じていることを認識したにもかかわらず，誤りの公表，訂正等の適切な対応を行わなかった．

⑷合算処理を行った際，調査票の数値を書き換え，元の数値を残していなかったため，二重計上の誤りを遡及して訂正することが困難となった．

受注統計事案の発覚を受け，同月，国土交通省では，弁護士や学識者から成る第三者検証委員会を設置し，事案の検証・分析が進められた．また，これと並行して，総務省及び統計委員会によるこの問題への対応について検証を行うため，統計委員会にタスクフォースが設置された．検証委員会及びタスクフォースは，いずれも 2022 年 1 月に報告書を取りまとめた（国土交通省 建設工事受注動態統計調査の不適切処理に係る検証委員会, 2022; 統計委員会, 2022a）．

これらの検証結果を踏まえ，2022 年 1 月，統計委員会は「公的統計品質向上のための特別検討チーム」（以下「特別検討チーム」）を設置して政府全体の再発防止策の検討を行った．この検討では，調査票の遅延提出，欠測値補完処理等は他の統計調査でも生じうることから，他の統計調査においても同様の問題が生じていないか，すべての基幹統計調査について点検・確認が行われた．その結果，基幹統計調査において新たな二重計上や書き換えの問題は確認されなかった．

特別検討チームでは，最初に，毎勤統計事案を踏まえた対策が行われていたのに，なぜ受注統計事案の発生が防止できなかったのか，対策に不十分な点があったのではないか，との観点から検討が行われた．その結果，受注統計事案の原因となる二重計上の問題は 2003 年に発生しており，第 3 次基本計画（変更後）の対策が過去にさかのぼって効果を持つのは困難であること，他方，仮にそのような対策が問題発生あるいは問題発見の当時に講じられていたとすれば，相当程度の効果が期待できたことなどが確認された．この結果を踏まえ，これまでの「総合的品質管理」の方針を維持しつつ，新たに発見された課題にも的確に対応できるよう，可能な限り対策を改善することとされた．

特別検討チームの検討結果は，2022 年 8 月に統計委員会の建議として総務大臣に提出された（統計委員会, 2022b）．この建議には 2019 年の統計委員会の建議からさらに踏み込んだ重要な指摘がいくつか含まれている．特に重要なのは，エラーの発生に対するトップマネジメントの基本姿勢のあり方である．組織においてエラーの発生を抑止し，エラーが重大事象に至るのを防ぐためには，トップが業務の特性を理解し，業務の内容や量に応じた体制を確保する必要があると指摘した．その上で，「エラーの発生自体を悪とする」のではなく，「エラーに対して社会や統計ユーザーを第一に考えた対応がなされないことを悪とする」という意識を府省内に浸透させ，品質優先で風通しのよい組織風土の醸成を強く提唱している（p.3）．この記述は，行政が無謬性に固執する姿勢を戒めるものでもあり，政府に対するこれまでの提言にはあまりみられない着目すべき点である．

表 7．2021 年統計委員会建議の 10 の取組．

＜総合的品質管理の推進＞
1　PDCA サイクルの確立と業務マニュアルの整備・共有の改善
2　業務マニュアルに記載のない事態が生じた場合の対応
3　変更管理の取組の導入
4　遅延調査票の取扱いの明確化
＜ガバナンスのための組織内外のコミュニケーションの確保＞
5　誤り発見・発生時の適切対処の徹底，備えと品質優先の組織風土の定着
6　地方公共団体や民間事業者との目的意識の共有と十分な意思疎通の確保
＜デジタル化による人間系ミスの低減と業務プロセスの改善＞
7　デジタル化の推進
＜品質優先の組織風土のための基盤の整備・強化＞
8　品質優先の組織風土の定着に向けたマネジメント能力の向上と職員の人材育成
9　各府省の体制強化
10　中央統計機構の相談対応の充実と体制強化

（資料）統計委員会 (2022b)の目次から抜粋

そのほか，具体的な取組として，表 7 に示す 10 項目が提言されている．

これらの項目の多くは，2019 年の建議における項目とも類似しているが，いずれも従来以上に実効性を高めるよう，具体的な提言が盛り込まれている．

また，2019 年の建議には無かった新たなものとして，2〜4 の項目が追加された．

2 は，受注統計において遅延提出の調査票の取扱がエラーの発端となったことに関連する事項である．業務マニュアルに記載されていない事態が発生した場合にはエラーが発生するリスクが増大するので，まずは状況に応じて例外的な処理を行って問題の解消に注力した上で，そのような状況を今後の改善に向けた見直しのヒントと認識し，統計作成プロセスや業務マニュアルの見直しに活かすことを推奨している．

3 は，受注統計事案が欠測値補完を新たに導入したのに伴って発生したことからの教訓である．誤りは「3H（変化，初めて，久しぶり）」のタイミングで発生しやすいことを指摘し，変更時には変化や変更が及ぼす影響を確認してエラーの未然防止に努めるよう促している．

4 は，遅延調査票は受注統計以外の統計調査でも発生すると考えられることから，その適切な取り扱い方を示している．

この建議では，過去の建議の内容を改めて評価した上で，それ以降に発見された課題にも適切に対応するために，推奨される取組の内容を深化させている．

3.7　第 4 期基本計画―政府における最近の総合的品質管理の取組

2022 年 8 月の建議は，第 3 期基本計画の終了する 2022 年度末まで余すところ約半年の時期にとりまとめられたため，建議の検討と基本計画の改定が並行して進められた．新たな基本計画には建議の提言内容がすべて盛り込まれ，2023 年 3 月に対象期間を 2023〜2027 年度として第 4 期基本計画（閣議決定, 2023）が決定された．これにより，不適切処理事案を踏まえた総合的品質管理の取組は，新たな基本計画の中で一体的に整理され，着実に進められる枠組が整った．

第 4 期基本計画では，統計の品質に焦点を当てた取組のあり方が，「第 I 部施策展開に当たっての基本的な方針」，「第 3 章第 IV 期基本計画における施策展開の基本的な視点」の「(4) 品質の高い統計の作成のための基盤整備」に記されている．以下にその主な部分を引用する．

公的統計が社会の情報基盤としての役割を果たす上で，統計ユーザー等にとって，信頼できる，有用性の高い統計が継続的に提供されていることが不可欠であり，品質の高い統計を確実に作成するために必要な基盤を整備する．

特に，各府省は，幹部職員の下，社会や統計ユーザー等を第一に考えて，主体的に統計の総合的品質管理（Total Quality Management．以下「TQM」という．）や統計作成プロセスの標準化やメタデータを含む情報提供の質の向上に取り組む．また，統計は，多くの者が関与する「総合プロジェクト」であることを認識し，統計の品質上の問題は，常に発生し得るものとして十分な備えを行い，仮に発生した場合には，迅速に対応する．さらに，品質管理体制を含め，必要なリソースの確保や専門性の高い人材の育成に計画的に取り組む．総務省は，中央統計機構として，各府省に対して，人的支援や相談対応を含む技術的な支援，各府省が共同で利用することができるシステムの提供を拡充するとともに，統計調査の最前線である地方公共団体や統計調査員による実査や審査，調査環境の改善を支援する．（p.7）

このような方針に基づき，総合的品質管理の実践に向けて様々な具体的な取組が行われている．以下，主な取組を紹介する．

（1）PDCA サイクルに関するガイドラインの整備と実践

各統計調査における総合的品質管理を実践するために，従前は「公的統計の品質保証に関するガイドライン」及び「PDCA サイクル確立に向けた点検・評価ガイドライン」が並立していたが，2023 年 7 月，両者を一体的かつ整合的に実施できるよう，新たなガイドラインとして「PDCA サイクルによる公的統計の品質確保・向上のためのガイドライン」（統計行政推進会議, 2023）（以下「PDCA ガイドライン」）に集約された．

このガイドラインは，各府省が統計作成プロセスについて自律的に PDCA サイクルを回すための指針を与えるものであり，それを実践するには，統計作成に関する業務マニュアルの整備などが必要となる．このため，ガイドラインの整備と並行して，各統計調査の業務マニュアルを整備するための参考として，「統計作成ガイドブック」が作成され，各府省に提供された．また，PDCA の「C（Check）」の実効性を高めるための重要な取組の一つである「統計作成プロセス診断」については，統計作成プロセスの品質保証を行うための方針等が定められた．この中には業務の各段階において品質の確保・向上に必要とされる「要求事項」が明記されており，統計作成者はそれを満たすように業務設計と業務遂行の水準を向上させていく取組が PDCA ガイドラインにおいても求められることとなった．

PDCA ガイドラインの適用実績としては，前身のガイドラインのものも含め，2020 年 10 月の運用開始から 2022 年度末までに 239 の統計調査に対して各府省自らが行う点検・評価が実施されている．これが適用された統計調査においては，①オンライン調査の導入・拡充，②速報値公表の追加，③業務委託・民間事業者の活用，④調査項目・集計項目の見直し，⑤業務マニュアル内容の充実・共有化，⑥標本設計・推定方法の見直し，⑦調査実施期間・調査票提出期限の見直し，などの多くの改善事例が報告されており，PDCA サイクルが統計の品質向上に寄与していることが確認されている．

（2）統計作成プロセス診断の実施・充実

統計作成プロセス診断は，各統計の作成プロセスの適正化及び改善を図り，統計の品質向上に寄与することを目的として，専門家が第三者の立場から診断を行うものである．この診断は，第 3 期基本計画（変更後）に基づき，2021 年に一部の統計調査に対し試行的に実施され，2022 年度には，不適切事案が生じた受注統計と，それと同じ部署で作成されていた建築着工統計の 2 つの基幹統計に対して先行的な取組が実施された．

これらの実績を踏まえ，2023 年 7 月に「統計作成プロセス診断の方針（フレームワーク）」が決定され，以降，これに基づいて計画的に実施されている．この方針には，統計調査が満たすべき「要求事項」が明示されており，これによって本格的な統計作成プロセス診断が可能となった．統計作成プロセス診断は，統計作成府省による継続的な品質管理の取組の充実及び実効性の確保を意図して行われるものであり，これにより，統計作成を担当する各府省の自主的な取組を支援するものであることが「統計作成プロセス診断の方針（フレームワーク）」に明示されている．また，これを通じて把握された改善などの好事例は，総務省を通じて府省横断的に横展開することとされている．第 4 期基本計画では「統計作成プロセス診断を全ての基幹統計調査を対象に計画的に行う」（閣議決定, 2023, p.68）という目標が掲げられ，現在，その計画に沿った対応が着実に進められている．

（3）体制整備及び人材育成

政府全体として総合的品質管理を充実・定着させていくためには，その業務を担う体制が必要である．このため，2023 年度から 40 名の統計品質管理官が総務省に定員として措置され，各府省の統計作成部署に派遣されている．統計品質管理官は，統計の総合的品質管理の中核として PDCA ガイドラインに基づく各府省内での自主・自律的な統計作成プロセスの不断の見直しの推進を図る役割を担う．また，各府省における総合的品質管理の支援のため，2023 年 4 月，総務省は，品質管理等に関する相談対応を担う外部有識者として統計品質アドバイザー 8 名，統計技術アドバイザー 4 名を任命している．

「品質はプロセスで作り込む」との考え方を実践する上では，担当職員には業務に必要な専門性が求められる．毎勤統計事案及び受注統計事案では，問題発生の早期発見及び未然防止ができなかったことの一因として，担当職員に統計の専門知識が不足していたことが挙げられている．このため，政府は，統計の専門資格として統計データアナリスト等を新たに設け，職員にその資格の取得を督励している．この資格は，所定の研修を修了することに加え，一定の業務経験を有することが要件とされている．2023 年 3 月末までに，政府全体で統計データアナリスト 27 人，統計データアナリスト補 187 人の認定が行われている．

統計専門人材の育成は，政府全体の人事方針にも取り入れられている．「令和 5 年度における人事管理運営方針について」（内閣人事局, 2023）には，「「公的統計の整備に関する基本的な計画」（令和 5 年 3 月 28 日閣議決定）等を踏まえ，統計作成に携わる職員について，各府省等が定める育成目標に従い，研修受講や統計データアナリスト等の資格取得の促進，能力向上と適切な処遇配置，品質管理や誤り防止の取組等への的確な評価等を行い，EBPM 及び統計人材の確保と育成を着実に行う」（p.13）ことが明記されている．これにより，計画的な人材育成をより強力に進めることが可能となる．

以上，最近の取組の一端を紹介した．公的統計の総合的な品質管理の取組は，第 4 期基本計画に沿って確実に進められており，基本計画を要とする公的統計全体の PDCA サイクルを回すことを通じて，その取組は新たな課題に対応して進化を続けている．

4. 結び

以上説明したとおり，海外においても国内においても，公的統計の品質管理に関する取組が段階的により高度なものに発展している．このような動向を通じて，公的統計の品質管理を成功させるために重要な要素を 3 点指摘しておきたい．

1 点目は，公的統計の品質管理を進めるための概念及び枠組を明確に定め，それを現実の課題に適用して改善を進めることである．

冒頭で述べたとおり，公的統計の品質の概念はとらえにくいものであるため，基本的な概念

を明確に定義することは品質管理を進める上で不可欠な第一歩である．その上で，品質向上のための仕組みを構築し，実践する必要があるが，枠組が抽象的に存在するだけでは品質管理を進めにくい．他方，現実には，統計に対する疑義，批判，さらには品質に影響する問題などが発生する．そのような場合にこそ，具体的な課題に即して品質管理の枠組を効果的に実践することが可能となる．新統計法は，このような枠組の根拠を与えるものであり，新統計法は公的統計の品質管理の仕組みを支える重要な要素と言える．

公的統計の利用ニーズは，社会・経済の環境とともに常に変化し，品質の目標自体も変化することから，品質向上の取組には終わりはない．品質管理の枠組を活用して継続的な品質向上の取組を行いつつ，枠組自体も必要に応じて改良していくことが必要である．

統計の品質向上とは，ややもすると統計のエラーをなくすことと見なされがちだが，より重要なのは利用価値を高めることである．品質管理が形骸化することのないよう，その本来の意義を統計組織内に浸透させることが必要である．

2 点目は，品質管理に関する各統計組織内での自律的な対応と政府全体を通じた連携である．日本の統計制度は分散型であり，政府内の様々な機関が統計を作成している．公的統計の品質向上のためには集中型に移行すべきとの議論もあるが，大規模な組織再編は容易ではなく，また，組織の形態を変えることで直ちに品質が向上するほど簡単な問題ではない．多種多様な統計の品質を 1 か所で集中的に管理することは実務上困難であるので，一義的には各統計組織が責任を持って自律的に品質管理を行うことが不可欠となる．このことは，統計組織を集中型に変更したとしても同様である．現状は，すべての統計組織を通じた共通の方針・枠組が明確に定められ，各組織が品質について責任を持って対応することが求められている．

しかし，府省によっては専門的なノウハウがない場合もあることから，中央統計機構である総務省の統計機関（統計局，政策統括官，統計研究研修所）等が各府省に対し必要に応じてアドバイスや支援を行うことが重要となる．また，重大事象を誘発するリスク要因に関する情報や，各統計機関において良好な取組事例を横展開することも品質の向上に効果的である．

このような意味で，政府全体を通じた連携が非常に重要であり，その中で中央統計機関に期待される役割は大きい．

3 点目は，統計の品質向上には，公的統計の意義，仕組み等に関し国民の理解を得ることである．

品質管理でしばしば用いられる表現に「後工程はお客様」というものがある．この言葉を統計作成のプロセスに当てはめると，国民は統計作成の「前工程」と「後工程」の両方を担っている．統計作成の前工程とは，統計調査などを通じて統計作成の原材料となる情報を国民から収集するプロセスに当たる．この工程においてしばしば発生する統計調査への非回答や回答誤りは，しばしば統計の品質低下の原因となるので，正確かつ漏れのない回答が得られることが重要となる．その意味では，国民が，国の統計組織を「後工程はお客様」と考えてくれれば，統計の品質向上に大きく貢献するだろうが，そのようなことはにわかには期待しがたいであろう．

他方，統計作成の後工程とは，作成した統計を国民の利用に供することに当たり，これについては統計組織が「お客様」に対して配慮しなければならないことは言うまでもない．本稿で説明している公的統計の品質管理の取組はまさにこの部分に該当する．

このように，国民は統計作成の前工程と後工程の両方に関わっているので，前工程である統計調査等において漏れなく正確な回答を提出することは，最終工程の「お客様」である国民自身のメリットにはね返ってくる．統計機関が国民に対してこのような両面の役割を十分説明し，理解を得るよう努力することは，統計の品質向上に寄与することとなる．その意味で，統計組織は国民との良好なコミュニケーションの確立に努める必要がある．

統計の品質の維持・向上は，一義的には統計組織自らが努力すべきことであるが，それには

国民の理解・協力が必要である．それには，統計組織はどのような場合でも誠実で分かりやすい情報発信を行う必要がある．統計組織が努力を重ねても，不測のエラーが発生することもあるが，そのような場合にも，その情報を誠実に公開し，速やかに是正措置を講じて信頼回復に努めなければならない．

このことと関連して，国民の統計リテラシーの向上に努めることは，公的統計の品質向上にとって重要である．利用者が高い統計知識を持てば，公表された統計から新たな知見が引き出されるなど，統計の利用価値が高まると期待される．そのためにも，統計学，データサイエンス，情報技術等の知識が国民に広まるよう，政府，学界，産業界等が連携して取り組むことが大切である．

高い品質の公的統計が社会に提供され，有効に活用され続けるためには，社会全体として，公的統計のこのような特質に関する理解が高まるよう継続的な努力が必要である．

謝　辞

本稿で解説した総合的品質管理の進展は，統計委員会委員，審議協力者，総務省の担当者等の皆様による多大なご尽力に負うものであり，この場を借りて深甚な敬意と感謝を申し上げる．また，原稿を注意深くお読みいただき，適切な助言をいただいたことに対し，匿名査読者及び編集委員に感謝を申し上げる．

参 考 文 献

Carson, C. (2001). Toward a framework for assessing data quality, IMF Working Papers, https://www.imf.org/-/media/Websites/IMF/imported-full-text-pdf/external/pubs/ft/wp/2001/_wp0125.ashx（最終アクセス日 2024 年 2 月 3 日．他の URL もすべて同じ）．

IMF (2006). Japan: Report on observance of standards and codes — Data module, response by the authorities, and detailed assessments using the Data Quality Assessment Framework (DQAF), https://www.imf.org/external/pubs/ft/scr/2006/cr06115.pdf.

IMF (2013). The special data dissemination standard: Guide for subscribers and users, https://dsbb.imf.org/content/pdfs/sddsguide13.pdf.

IMF (2015). Guide for users: Enhanced General Data Dissemination System, https://dsbb.imf.org/content/pdfs/eGDDS_Guide_for_Participants_and_Users.pdf.

ISO (2015). ISO 9000:2015 Quality management systems, https://www.iso.org/obp/ui/en/#iso:std:iso:9000:ed-4:v1:en:term:3.5.3.

閣議決定 (2009). 公的統計の整備に関する基本的な計画（平成 21 年 3 月 13 日）, https://www.soumu.go.jp/main_content/000283571.pdf.

閣議決定 (2014). 公的統計の整備に関する基本的な計画（平成 26 年 3 月 25 日）, https://www.soumu.go.jp/main_content/000536498.pdf.

閣議決定 (2018). 公的統計の整備に関する基本的な計画（平成 30 年 3 月 6 日）, https://www.soumu.go.jp/main_content/000536501.pdf.

閣議決定 (2020). 公的統計の整備に関する基本的な計画（令和 2 年 6 月 2 日）, https://www.soumu.go.jp/main_content/000536501.pdf.

閣議決定 (2023). 公的統計の整備に関する基本的な計画（令和 5 年 3 月 28 日）, https://www.soumu.go.jp/main_content/000871085.pdf.

経済財政諮問会議 (2015). 議事録（平成 27 年第 16 回）, https://www5.cao.go.jp/keizai-shimon/kaigi/minutes/2015/1016/gijiroku.pdf.

経済財政諮問会議 (2016). 統計改革の基本方針, https://www.kantei.go.jp/jp/singi/toukeikaikaku/dai1/sankou2.pdf.

国土交通省 建設工事受注動態統計調査の不適切処理に係る検証委員会 (2022). 建設工事受注動態統計調査の不適切処理に係る調査報告書, https://www.mlit.go.jp/report/press/kanbo09_hh_000047.html.

厚生労働省 毎月勤労統計調査等に関する特別監察委員会 (2019a). 毎月勤労統計調査を巡る不適切な取扱いに係る事実関係とその評価等に関する報告書, https://www.mhlw.go.jp/content/10108000/000472506.pdf.

厚生労働省 毎月勤労統計調査等に関する特別監察委員会 (2019b). 毎月勤労統計調査を巡る不適切な取扱いに係る事実関係とその評価等に関する追加報告書, https://www.mhlw.go.jp/content/10108000/000483640.pdf.

Lyberg, L. (2012). Survey quality, *Survey Methodology*, **38**(2), 107–130.

内閣人事局 (2023). 令和 5 年度における人事管理運営方針について(令和 5 年 3 月 28 日 内閣総理大臣決定), https://www.cas.go.jp/jp/gaiyou/jimu/jinjikyoku/files/housin_r05.pdf.

日本品質管理学会 (2015). 日本品質管理学会規格「プロセス保証の指針」(JSQC-Std 21-001:2015).

日本経済新聞電子版 (2019).「来年度予算案、101 兆 4571 億円に修正 統計不適切問題受け」2019 年 1 月 17 日.

統計行政推進会議 (2020). PDCA サイクル確立に向けた点検・評価ガイドライン, https://www.soumu.go.jp/main_content/000709829.pdf.

統計行政推進会議 (2023). PDCA サイクルによる公的統計の品質確保・向上のためのガイドライン, https://www.soumu.go.jp/main_content/000903680.pdf.

「統計」編集部 (2019). 毎月勤労統計の不適切処理をめぐる問題の概要, 統計, **70**(5), 2–5.

統計委員会 (2019). 公的統計の総合的品質管理を目指した取組について(建議), https://www.soumu.go.jp/main_content/000647066.pdf.

統計委員会 (2022a). 統計委員会タスクフォース精査結果報告書―建設工事受注動態統計調査を巡る事案への総務省政策統括官室の対応―, https://www.soumu.go.jp/main_content/000844625.pdf.

統計委員会 (2022b). 公的統計の総合的な品質向上に向けて(建議), https://www.soumu.go.jp/main_content/000829746.pdf.

統計改革推進会議 (2017). 統計改革推進会議最終取りまとめ, https://www.kantei.go.jp/jp/singi/toukeikaikaku/pdf/saishu_honbun.pdf.

統計改革推進会議 統計行政新生部会 (2019). 統計改革推進統計行政の新生に向けて～将来にわたって高い品質の統計を提供するために～, https://www.kantei.go.jp/jp/singi/toukeikaikaku/pdf/20191224_shinsei_honbun.pdf.

統計企画会議 (2019). 公的統計の品質保証に関するガイドライン(平成 22 年 3 月 31 日，改定 平成 23 年 4 月 8 日，平成 28 年 2 月 23 日，平成 31 年 4 月 26 日), https://www.stat.go.jp/data/guide/pdf/guideline.pdf

United Nations (1994). Fundamental Principles of Official Statistics, https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx.

United Nations (2012). Guidelines for the template for a generic national quality assurance framework, https://unstats.un.org/UNSDWebsite/statcom/session_43/documents/BG-NQAF-E.pdf.

United Nations (2014). Resolution adopted by the General Assembly on 29 January 2014 (A/68/L.36 and Add.1) 68/261, Fundamental Principles of Official Statistics, https://unstats.un.org/unsd/dnss/gp/fp-new-e.pdf.

United Nations (2019). United Nations National Quality Assurance Frameworks Manual for Official Statistics, https://unstats.un.org/unsd/methodology/dataquality/un-nqaf-manual/.

United Nations Statistical Commission (2009). Statistics Canada: national quality assurance frameworks, https://unstats.un.org/unsd/statcom/doc10/2010-2-nqaf-e.pdf.

Proceedings of the Institute of Statistical Mathematics Vol. 72, No. 2, 153–173 (2024) 173

Total Quality Management in Official Statistics

Shigeru Kawasaki[1,2] and Sei Ueda[3]

[1]Japan Statistical Association, Inc.
[2]Data Science and AI Innovation Research Center, Shiga University
[3]Statistics Bureau, Ministry of Internal Affairs and Communications

This paper outlines the efforts in total quality management of official statistics by international organizations and the Government of Japan. Quality management in this field has been developed stepwise since the 1990s. Among international efforts, the Fundamental Principles of Official Statistics established by the United Nations Statistical Commission and the Special Data Dissemination Standard of the International Monetary Fund have led to substantial improvements in the quality of official statistics worldwide. In Japan, the Statistics Law, which was fully revised in 2007, has played a key role in the broad application of total quality management in official statistics in Japan. Recently, the government encountered two major cases of inappropriate processing in two major statistical surveys. As a result, total quality management is now extensively applied to all official statistics. We conclude with a discussion of the necessary elements for promoting total quality management of official statistics.

Key words: Official statistics, Fundamental Principles of Official Statistics, Statistics Law, total quality management, United Nations Statistical Commission.

統計数理(2024)
第 72 巻 第 2 号 175–193
©2024 統計数理研究所

我が国における統計ビジネスレジスターの整備

―社会の重要な情報基盤としての「事業所母集団データベース」―

髙橋 雅夫†

(受付 2023 年 12 月 14 日；改訂 2024 年 2 月 22 日；採択 3 月 21 日)

要　　旨

本稿は，公的統計を社会の情報基盤と位置付けた現行の統計法において，その整備を進めることが定められた事業所母集団データベースの開発・整備の経緯と当該データベースの概要について紹介するとともに，今後の方向性について論じるものである．我が国の事業所母集団データベースは，国際的には統計ビジネスレジスターと呼ばれるものに相当する．統計ビジネスレジスターとは，統計目的で利用される事業所・企業等の経済単位のデータベースで，通常各国の国家統計局によって維持・管理されるものである．事業所母集団データベースは，経済センサスや各種統計調査及び行政記録等をその情報源とし，事業所や企業に関する各種統計調査のための母集団情報を提供するとともに事業所に関する統計を作成することを目的として整備されている．本稿では，事業所母集団データベース構築の経緯を紹介した後，データベースの概要について述べる．さらに，今後期待される発展の方向性として，データベースのカバレッジの更なる改善や国際連携などについても論じる．

キーワード：経済センサス，行政記録情報，サンプリング・フレーム，レジスター統計，ビジネスデモグラフィー．

1. はじめに

現行の統計法は，1947 年に制定された統計法が 2007 年に全面的に改正されたものである．改正の際の基本的な視点は「行政のための統計」から「社会の情報基盤としての統計」へというものであった（平田, 2007）．この改正においては，いくつかの重要な取組を新たに実施することが定められた．例えば，公的統計の整備に関する基本的な計画を策定することや調査票情報の 2 次利用の推進などであるが，その中に「事業所母集団データベースの整備」も含まれていた．「事業所母集団データベース」とは，統計法によれば，「事業所に関する情報の集合物であって，それらの情報を電子計算機を用いて検索することができるように体系的に構成したもの」と定義されている．

本稿においては，上記のように統計法において位置付けられることになった，社会の重要な情報基盤の一つと考えられる「事業所母集団データベース」の開発・整備の経緯を国連等における国際的な動向を踏まえて紹介した後，現在の事業所母集団データベースの概要について述べる．さらに，事業所母集団データベースの今後の展開として，データベースのカバレッジ，国際連携，同データベースに格納されたデータをある時点で集計して得られる事業所・企業に関

† 長野大学 企業情報学部：〒 386–1298 長野県上田市下之郷 658–1

する統計(レジスター統計)の作成及び事業所・企業の動態統計(ビジネスデモグラフィー統計)の作成などについて論じる.

なお,本稿における事業所母集団データーベースの概要については,主に同データベースの情報源や利用についてまとめたものであり,データベース本体の技術的仕様等のシステムに関する事項についての紹介は別の機会に譲ることとする.

2. 統計ビジネスレジスターとしての事業所母集団データベース

我が国の「事業所母集団データベース」は,国際的には「統計ビジネスレジスター」(Statistical Business Register)というものに相当する.統計ビジネスレジスターとは,統計目的で整備された事業所や企業などの経済単位のデータベースのことで,「統計ビジネスレジスターに関する国連ガイドライン」(United Nations, 2020)によれば,次のように定義されている.すなわち,統計ビジネスレジスターとは,「国家統計局によって維持管理される,ある地域範囲についての定期的に更新される経済単位の構造化データベースで,統計目的に利用されるもの」(筆者訳)である.

なお,「統計ビジネスレジスター」という用語に付されている「統計」という言葉は,行政目的で利用される事業所・企業の登録簿である「ビジネスレジスター」と区別するためのものと考えられるが,特に支障のない場合には単に「ビジネスレジスター」と呼ばれている.

事業所母集団データベースは,統計法第 27 条に基づき総務大臣が整備することとされている.整備の目的は,統計法によれば,「行政機関等による正確かつ効率的な統計の作成及び統計調査その他の統計を作成するための調査における被調査者の負担の軽減に資すること」とされている.また,その利用の目的として統計法では 2 つ挙げられている.1 つは,事業所に関する統計を作成するための調査の対象の抽出のためであり,もう 1 つは,事業所に関する統計の作成のためである.

データベース整備に当たっての具体的な企画は総務省統計局が担い,整備に関する様々な取組は,各府省や独立行政法人統計センター(以下「統計センター」という)の協力の下,進められている.

3. 統計ビジネスレジスターに関する国際的取組

海外の政府統計機関では,欧州の統計機関を中心に早くはノルウェーやフィンランドなどで 1960 年代半ばからビジネスレジスターの整備が始められたが,デンマーク,スイス,カナダなど多くの国々では 1980 年代以降にその構築にとりかかっている(Blancas, 2022).そのような動きの中で,1986 年にビジネスレジスターに関する経験を交換するための国際的な専門家グループによる「ビジネスフレームに関する国際円卓会議」(International Roundtable on Business Survey Frames)が組織され,第 1 回の会合がオタワ(カナダ)で開催された.この会議は,国連の統計委員会にその活動が承認されたいわゆるシティグループの一つに位置付けられ,2007 年には他のシティグループの名づけ方に合わせて「ビジネスレジスターに関するヴィースバーデングループ」(Wiesbaden Group on Business Registers; 以下,単に「ヴィースバーデングループ」という)と改名された(United Nations Statistics Division, 2024a).このヴィースバーデングループの会合には欧米の統計部局からの統計専門家が中心となって参加しているが,我が国の統計部局からも毎回のように参加し,論文を発表するなど情報交換を行ってきている.

なお,国連のシティグループは,統計手法の開発と実装における特定のテーマに関する課題について議論し,対処するための非公式協議グループとして各国統計局等の専門家を参加メンバーとして結成されてきたもので,扱われるテーマは例えば,ビジネスレジスターや物価指数,

サービス統計など，グループごとに定められている．各グループの名称は，その多くが最初の会議が開催された場所にちなんだものとなっているため，総称してシティグループと呼ばれている．シティグループでの議論の結果は国連統計委員会に報告され，国際統計基準の開発・改善などに役立てられている．

ヴィースバーデングループの会合は，2008 年までは毎年開催されてきたが，それ以後は隔年で開催されるようになり，これが開催されない年には国連欧州経済委員会(UNECE)が中心となってビジネスレジスターに関する専門家会合(Meeting of the Group of Experts on Business Registers)が開催されることとなったため，結果的にビジネスレジスターに関する専門家会合が毎年のように開催されることは継続されている．

このような流れの中で，ビジネスレジスターに関する国際的なガイドラインが作成されるようになった．2015 年には国連欧州経済委員会(UNECE)によって「統計ビジネスレジスターに関するガイドライン」(United Nations Economic Commission for Europe, 2015)が作成された．このガイドラインは，2003 年に欧州委員会統計局(Eurostat)により作成され 2010 年に第 2 版として改訂された「ビジネスレジスター推奨マニュアル」(Eurostat, 2010)を元に作られており，統計ビジネスレジスターを構築・維持していくための実用的ガイダンスと推奨事項を提供するものとなっている．

この UNECE のガイドラインを元に国連によって，統計制度があまり発達していない国々にも適用できるように「統計ビジネスレジスターに関する国連ガイドライン」が作成されており，現在のところ 2020 年版の最終稿が国連ウェブサイトで公開されている（United Nations, 2020）．

4. 事業所母集団データベース開発の経緯

我が国では，上記の国際的動向と連動する形で事業所・企業に関するデータベースの構築が進められてきた．本節では，同データベースの開発の経緯を高橋（2007a, 2007b）に基づき紹介する．

4.1 事業所に関する初期のデータベース

前述のように現在の我が国の「事業所母集団データベース」は 2007 年に統計法が全部改正された際にその整備が定められたものであるが，それ以前にも統計ビジネスレジスターに相当する，事業所に関するデータベースの整備が総務省統計局により進められていた．

事業所に関する最も初期のデータベースは，1998 年 1 月に運用を開始した「事業所・企業名簿情報データベース」である．このデータベースは，事業所や企業に関する様々な標本調査の標本抽出枠(サンプリング・フレーム)となる事業所や企業の名簿を各府省に提供することを目的に構築されていた．データベースの情報源は，平成 8 年事業所・企業統計調査の調査結果のみであり，また，データベースの各府省での利用はオフラインのみとなっていた．

次に開発されたデータベースは「事業所・企業データベース」と呼ばれ，2002 年 4 月からその運用が開始された．事業所・企業データベースは，統計調査の実施のための事業所・企業の母集団情報の提供及び各事業所・企業が各種の統計調査の標本として選定されたか否かに関する履歴情報の提供を目的としていた．この履歴情報は，特定の事業所・企業が一定の回数を超えて標本として選定されることを回避するために利用され，報告者負担の軽減に寄与するものであった．事業所・企業データベースの主要な情報源は，事業所・企業統計調査，商業統計調査，工業統計調査，及び法人企業統計調査の最新の調査結果であり，また，各府省は，当時の中央省庁の通信ネットワークであった霞が関 WAN を通じてこのデータベースへアクセスすること

が可能となっていた.

4.2 新しいデータベースの必要性

事業所・企業データベースは，その目的に沿った一定の役割を果たしていたが，以下に述べるような改善すべき点がいくつか明らかとなっていた.

まず，第 1 の点は，データベースのデータ更新があまり頻繁でないことであった．事業所・企業データベースの主要な情報源であった事業所・企業統計調査等の統計調査のデータは，調査の実施間隔に従って得られるため，データベースのデータが全体的に更新されるのに 2 年から 3 年の期間が必要となり，他の統計調査のための母集団情報の提供というデータベースの役割を踏まえると，更新頻度の更なる改善が望まれていた.

第 2 の点は，データ更新に相当程度のタイムラグが生じていたことであった．前述のようにデータベースの情報源は主に大規模な統計調査の結果であるが，各々の統計調査の実施から結果の公表までは通常少なくとも数ヶ月を要する．その上，データベースの更新に利用されるデータが提供されるのは，さらにその一定期間後となっている．例えば，2004 年に実施された事業所・企業統計調査(2004 年 6 月 1 日現在で実施)の結果に基づきデータベースのデータが更新されたのは調査日から 20 ヶ月以上後のことであった.

第 3 の点としては，その利用方法であった．事業所・企業データベースは，当初の目的に従って事業所・企業の母集団情報を提供することに加え，各種統計調査の標本として選定された履歴情報を提供することにより，特定の事業所・企業が各種統計調査の標本として過度に選定されないように調整する役割を果たしていた．しかし，当時諸外国において進められていた，統計ビジネスレジスター自体から統計を作成する取組については，まだ事業所・企業データベースにおいては実現されていなかった.

上記の状況を改善するため，総務省統計局において事業所・企業に関する新たなデータベースを構築することとなった．以下では，新たなデータベースを構築するに当たり，当時進められていた新データベースを取り巻く大きな流れについていくつか述べる.

4.3 統計調査等業務の業務・システム最適化計画

2006 年 3 月 31 日に当時の各府省情報化統括責任者連絡会議において，「統計調査等業務の業務・システム最適化計画」(以下，「最適化計画」という)が決定された (各府省情報化統括責任者連絡会議, 2006)．この最適化計画に基づき，それまで各府省によってそれぞれ別々に開発・運用されてきた統計に関する情報システムを統合することにより，業務とシステムの簡素化と合理化が図られることとなった．統合後のシステムには「政府統計共同利用システム」という名称が与えられることとなり，2008 年 1 月から運用が開始され，その後改良が加えられつつ現在でも運用が行われているものである.

政府統計共同利用システムは，事業所母集団データベース(当初は，事業所・企業データベース)，オンライン調査システム，統計地理情報システム(統計 GIS)等の 13 のシステムから構成されており，そのシステムの構成は図 1 に示すとおりである．この最適化計画において，事業所母集団データベースで整備する母集団情報の基礎となる統計調査として経済センサスを用いることとされ，母集団情報の更新は，各種統計調査の結果や行政記録情報等を活用して随時行うものとされた.

なお，事業所母集団データベースは，後述する「公的統計の整備に関する基本的な計画」(第 I 期基本計画)に基づき必要となる機能が追加されるまでは引き続き「事業所・企業データベース」という名称を用いることとされた.

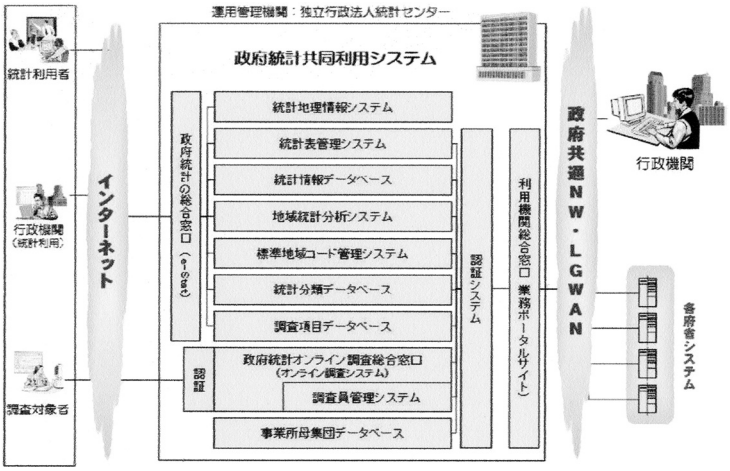


図 1. 政府統計共同利用システムの構成図．独立行政法人統計センター（2022）を元に筆者が加工．

4.4 統計法の全部改正

「経済財政運営と構造改革に関する基本方針 2004」（2004 年 6 月 4 日閣議決定）において既存統計の抜本的見直しや統計制度の充実が掲げられたことを受けて，2004 年 11 月に経済社会統計整備推進委員会（以下「推進委員会」という）が内閣府に設置され，経済社会統計の整備の推進を図るための議論が開始された．議論の結果は 2005 年 6 月 10 日に「政府統計の構造改革に向けて」と題する報告書としてまとめられた（内閣府経済社会統計整備推進委員会, 2005）．報告書では，経済センサスの創設，GDP 関連統計などの既存統計の抜本的見直しや統計制度の充実についての具体案が提言された．

推進委員会の報告書の公表直後には「経済財政運営と構造改革に関する基本方針 2005」（2005 年 6 月 21 日閣議決定）が公表され，統計法制度を抜本的に見直す方針が打ち出された．この方針に沿った形で 2005 年 9 月に推進委員会の後継組織として内閣府に統計制度改革検討委員会（以下「検討委員会」という）が設置された．この検討委員会において統計制度の抜本的見直しに関する議論が行われ，その検討結果として，2006 年 6 月 5 日に「統計制度改革検討委員会報告」が公表された（統計制度改革検討委員会, 2006）．この報告書は，「推進委員会」の報告書と併せて，我が国の統計の改革に向けた基本的な認識を示し，時代の変化に対応した新たな統計整備の仕組みについて提言するものとなっている．その提言の主要なポイントとして，統計に関する新たな法制度を整備することとと併せて，ビジネスフレーム（事業所・企業に関する共通の母集団情報）を整備することも挙げられていた．なお，上記の推進委員会及び検討委員会における議論の経過については，井上（2008）に詳しく記されているのでそちらも参照いただきたい．

上で述べた経緯を受けて「経済財政運営と構造改革に関する基本方針 2006」（2006 年 7 月 7 日閣議決定）において，統計法制度を抜本的に改革するための法律案を次期通常国会に提出することとされた．この統計法を全部改正する法案は 2007 年 2 月 13 日に国会に提出され，5 月 16 日に可決成立し，5 月 23 日に公布された．公的統計は，改正された統計法（以下，特に断りのない場合，「統計法」はこの改正された法律のことを指す）の第 1 条において「国民にとって合理的な意思決定を行うための基盤となる重要な情報」と位置付けられることとなった．

既に言及したように統計法第 27 条では，行政機関等による正確かつ効率的な統計の作成及び統計調査その他の統計を作成するための調査における被調査者の負担の軽減に資することを

目的として，基幹統計調査又は一般統計調査に係る調査票情報の利用，法人その他の団体に対する照会その他の方法により，総務大臣が「事業所母集団データベース」を整備することとされた．また，同法第 27 条第 2 項では，行政機関の長，地方公共団体の長その他の執行機関又は独立行政法人等は，以下の目的のため，総務大臣から事業所母集団データベースに記録されている情報の提供を受けることができるものとされた．

(1) 事業所に関する統計調査その他の事業所に関する統計を作成するための調査の対象の抽出
(2) 事業所に関する統計の作成

4.5 事業所母集団データベースの整備に関する計画

統計法第 4 条において，「政府は，公的統計の整備に関する施策の総合的かつ計画的な推進を図るため，公的統計の整備に関する基本的な計画（以下この条において「基本計画」という）を定めなければならない」とされている．この基本計画は，2009 年 3 月に最初の計画（以下「第 I 期基本計画」という）が策定され，おおむね 5 年ごとに統計をめぐる社会経済情勢の変化を勘案して変更するものとされている．

基本計画はその後これまでに，2014 年 3 月，2018 年 3 月（2020 年 6 月に一部変更）及び 2023 年 3 月と大きく 3 回にわたり変更されてきた（それぞれ，「第 II 期基本計画」，「第 III 期基本計画」，「第 IV 期基本計画」と呼ばれている）．基本計画の変更に当たっては，学識経験者から構成される統計委員会（統計法に基づき総務省に設置）において調査審議がなされた後に閣議決定が行われている．また，統計委員会は，統計法に基づき毎年度総務大臣から統計法の施行状況について報告を受けることとなっており，その際には総務大臣又は関係行政機関の長に対して意見を述べることができるとされている．すなわち，基本計画を含む統計法施行の PDCA サイクルの中で統計委員会が非常に重要な役割を担っていることになる．

これらそれぞれの基本計画では，事業所母集団データベースの整備についても具体的な計画が記されており，同データベースはこの計画に沿って整備が進められてきている．各期において事業所母集団データベース関連で記載された計画のうち，データベース整備において中心的な役割を果たしている総務省に関連する主なものは，以下のとおりである（詳しくは付録の表 A-1〜表 A-5 を参照）．

まず，第 I 期基本計画（2009 年 3 月閣議決定）においては，事業所母集団データベース整備のために，経済センサス-活動調査の中間年に経済センサス-基礎調査を実施すること，法人企業の母集団情報整備のために往復郵便等による業種名，従業者数，事業所数等の照会を定期的に実施すること，雇用保険適用事業所等に関する届出情報から事業所等の新設，廃業等を把握することなどについて定められた．

この第 I 期基本計画が策定されたことを受けて総務大臣が事業所母集団データベースを具体的に整備していくことになったわけであるが，まず初めに「事業所母集団データベースの整備方針」が 2011 年 3 月 25 日に総務大臣により決定された（総務省, 2011）．この整備方針では，事業所母集団データベースの整備に関して，その目的，整備サイクル，統計関係業務支援機能の整備，記録する統計調査，行政記録情報等の活用，及び整備スケジュールが定められた．この後，この方針を出発点として事業所母集団データベースの整備が進められることとなった．

2014 年 3 月に閣議決定された第 II 期基本計画では，新たな行政記録情報等の活用や企業への直接的な照会などにより企業組織構造の変化を継続的に確認する方法について検討することや，事業所母集団データベースを活用して，我が国の事業所・企業の実態を把握する統計に加え，事業所・企業の異動状況や産業の成長・衰退等に着目した統計を作成することなどが定め

られた.

また，第 III 期基本計画(2018 年 3 月閣議決定, 2020 年 6 月一部変更)においては，経済センサス-基礎調査をローリング調査や統計センターにおけるプロファイリング活動に移行し，事業所・企業や各種法人等に係る母集団情報の更なる整備促進を図ること，法人番号を事業所母集団データベースに登録すること，レジスター統計の作成内容について検討を行うことなどが計画された.

さらに，最新の第 IV 期基本計画(2023 年 3 月閣議決定)では，経済センサス-活動調査の中間年における事業所母集団データベースの更なる整備を図るため，事業所・企業の基本的事項の更新範囲や頻度を高めることについて，経済センサス-基礎調査の見直しを含めて検討を行うこととされた.

5. 事業所母集団データベースの概要

事業所母集団データベースの整備は，上述の「事業所母集団データベースの整備方針」に基づき開始され，その後は幾たびか変更された「基本計画」に沿って整備が進められており，本稿の執筆時点においても引き続き第 IV 期基本計画に従って進められているところである．以下では，事業所母集団データベースの概要について，総務省 (2011)，独立行政法人統計センター (2022)及び総務省統計局 (2024)に基づき述べる.

5.1 データベースの構成単位

事業所母集団データベースに収録されているデータの基本的な単位は，事業所である．事業所の定義は，後述する同データベースの主要な情報源である経済センサスにおける事業所の定義(日本標準産業分類における定義に準拠)に従っている．具体的には，令和 3 年経済センサス-活動調査の用語の解説 (総務省統計局, 2021b)によると，事業所の定義としては，以下のように述べられている.

> 事業所とは，経済活動が行われている場所ごとの単位で，原則として次の要件を備えているものをいう.
> - 一定の場所(1 区画)を占めて，単一の経営主体のもとで経済活動が行われていること.
> - 従業者と設備を有して，物の生産や販売，サービスの提供が継続的に行われていること.

すなわち，事業所とは，一般に工場，製作所，事務所，営業所，商店，飲食店，旅館，娯楽場，学校，病院，役所，駅，鉱業所，農家等と呼ばれるものである (総務省, 2013).

5.2 データベースの主な機能

事業所母集団データベースの整備の目的は，既に述べたように統計法第 27 条に規定されているとおり，(1)行政機関等による正確かつ効率的な統計の作成及び(2)統計調査その他の統計を作成するための調査における被調査者の負担の軽減に資することである．この目的を達成するためにこのデータベースは以下に示すようないくつかの機能を備えている.

- 事業所と企業に関する母集団情報を保持すること
- 各種統計調査を実施する府省等に母集団情報を提供すること
- 各種統計調査の標本として抽出された履歴を保持すること
- 母集団情報提供の際，調査対象者の負担を平準化すること(重複是正)

5.3 データベースの情報源

5.3.1 経済センサス

事業所母集団データベースの基盤情報となっているのは，経済センサスの調査結果である．経済センサスには，基礎調査と活動調査の 2 種類の調査がある．経済センサス-基礎調査は，すべての産業分野における事業所及び企業の基本的構造―例えば事業の内容や従業者数など―を把握するための調査で，これまで 2009 年，2014 年及び 2019 年の各年度に実施されてきた．一方，経済センサス-活動調査は，全ての産業分野における事業所及び企業の経済活動の実態―例えば売上(収入)など―を明らかにするための調査で，これまでに 2012 年，2016 年，2021 年に実施されてきた．

このうち 2019 年に実施された経済センサス-基礎調査は，基本計画に基づきローリング形式で行われており，2019 年 6 月 1 日から 2020 年 3 月 31 日までの期間に主に事業所の開業・廃業状況を把握する目的で実施された．なお，ローリング形式の調査とは，調査を 1 時点で全国的に実施するのではなく，地域ごとに順次日にちをずらして一定期間内に全国の調査を完了させる形態の調査方法である．

5.3.2 各種統計調査

経済センサス以外の統計調査についても，事業所母集団データベースに記録されている値の補完など，情報の更新のために利用されている．利用する統計調査は，以下の基準により選定されている．

- 特定の産業において，悉皆(又はおおむね悉皆)となっている統計調査
- 幅広い産業を対象とし，一定の悉皆層を有する統計調査
- 幅広い産業を対象とし，調査客体数が多い統計調査
- 行政記録情報等と連動することにより，新たな統計の作成が期待される統計調査
- その他一般統計調査のうち，特に事業所母集団データベースの整備に有効であると考えられる統計調査

上記の基準により選定された統計調査を所管府省ごとに整理したものを表 1 に示す．

5.3.3 行政記録情報等

統計調査結果に加え，行政記録情報も事業所母集団データベースの更新のための情報源として利用されている．利用されているのは，主に労働保険情報及び商業・法人登記簿情報である．

労働保険情報については，毎月，厚生労働省より雇用保険・労災保険の加入等に関する情報を入手し，また，商業・法人登記簿情報については，毎月，法務省より商業・法人登記簿に基

表 1. 事業所母集団データベースの更新に利用されている統計調査(経済センサス以外)．

府　省	統計調査名	府　省	統計調査名
総務省	経済構造実態調査	農林水産省	農林業センサス（法人組織経営体）
	サービス産業動向調査		漁業センサス（法人組織経営体）
	科学技術研究調査	経済産業省	経済産業省企業活動基本調査
	個人企業経済調査		特定サービス産業動態統計調査
財務省	法人企業統計調査		エネルギー消費統計調査
文部科学省	学校基本調査		中小企業実態基本調査
厚生労働省	毎月勤労統計調査		商業動態統計調査
	賃金構造基本統計調査	国土交通省	建設工事施工統計調査
	医療施設調査		

づく法人の設立，移転，商号変更，合併，解散等の情報を入手している．ただし，労働保険情報からは，雇用者のいる事業所のみの情報が得られ，また，商業・法人登記簿情報からは法人の情報のみが入手できていることに留意が必要である．

これらの行政記録に基づく情報は，事業所の新設や廃業を捉えることを主な利用目的としているが，行政記録に新たに登録されたからといって，必ずしもそれが新たに活動を開始した事業所，又は活動を廃止した事業所であるとは限らないため，総務省では，これらの行政記録情報に基づき，関係の事業所・企業に対して郵送や電話等による照会を行って，新設・廃業に関する正しい情報を事業所母集団データベースに格納するようにしている．さらに，2023 年からは，この事業所・企業に対する照会範囲を拡大し，存続する事業所・企業のうち複数事業所企業に対して売上高，従業員数等の基本的事項を郵便で照会する事業を開始することとしており，結果として売上高でみて 90% に相当する事業所に関する基本的事項の情報を毎年更新できるようになる予定である（Kurematsu et al., 2023）．

一方，必ずしも事業所母集団データベースへの直接の情報源ということではないが，事業所母集団データベースの精度向上に資する取組の一環として，基本計画に基づき統計センターにおいて企業調査支援事業（プロファイリング活動）が実施されている．

プロファイリング活動とは，企業ごとに専任の担当者（サポートスタッフ）が割り当てられ，企業との関係を密に構築しながら情報を収集する取組のことで，統計センターにおいては，2019 年以降このプロファイリング活動を導入し，約 5,000 企業を産業分類等で 5 つのグループに分け，1 グループあたり 6 名のサポートスタッフを配置した 5 つのチームを編成して対応に当たっている．サポートスタッフ 1 人当たり約 160 企業を担当し，統計調査への回答支援と正確な回答データの作成を実施している．回答支援の内容は，企業の回答者の把握，企業の合併・分割等の把握，調査対象名簿の整備，調査に関する情報提供，問い合わせへの対応，調査票の回収・督促，回答内容の審査・訂正，企業への照会などである．

このプロファイリング活動は，経済センサスや経済構造実態調査など企業を対象とする統計調査に適用され，調査に回答する企業を支援することを通じて調査の精度を確保し，ひいては事業所母集団データベースの精度向上に役立てられているところである．

5.4 格納・提供される母集団情報

上記のように様々な情報源から得られる情報が事業所の母集団情報としてデータベースに格納されるわけであるが，格納される項目は，経済センサス-活動調査における産業共通の調査事項が基本となっている（具体的な調査事項は，例えば 2021 年の経済センサス-活動調査の産業共通調査票の調査事項（総務省統計局, 2021a）を参照いただきたい）．また，調査事項をデータベースに格納する際には詳細な符号化が行われ，結果としてデータベース内の項目数は約 250 項目となっている．

データベースに収録されているデータのレコードは，主に各回の経済センサスの結果により更新される．例えば，2021 年の経済センサス-活動調査に関しては，活動中の事業所（前回調査から存続している事業所及び新設された事業所で，約 600 万事業所）のデータ更新及び追加並びに廃業事業所（約 150 万事業所）への廃業識別コードの付与が行われている．さらに，5 年ごとの経済センサス-活動調査の中間年には行政記録情報等を活用した情報収集により新設・廃業の事業所を捉えて，その情報をデータベースに付加していくため，データベースのレコード数は年々増加することとなる．なお，2024 年 2 月現在のデータベース内のレコード数（事業所数）は，約 1,400 万レコードである．

事業所母集団データベースに格納された母集団情報は，「年次フレーム」という名称で毎年更新され，統計法に基づき統計調査の対象の抽出又は統計の作成を行うことを目的として，国の

行政機関，都道府県，市区町村等に提供されている．

年次フレームで提供されている項目は前述の経済センサス-活動調査における産業共通の調査事項が基本となっているが，その主なものは以下のとおりである．

- 事業所に関する項目
  名称，所在地，電話番号，事業の開始日，単独・本所・支所の別，売上高，従業者数，常用雇用者数，産業分類　など
- 企業に関する項目
  経営組織，法人番号，資本金，売上高，従業者数，常用雇用者数，産業分類　など

なお，これまで述べてきた事業所母集団データベースの要点を模式図として図 2 に示す．

5.5 報告者負担の軽減措置

事業所母集団データベースの持つ機能の一つである，統計を作成するための調査での報告者の負担軽減を図るため，同データベースには統計調査の標本として選定された回数に関する情報を各事業所のデータに付与して保存している．この情報を活用して，各事業所・企業が統計調査の標本として過度に選定されないよう選定回数の上限を設定している．上限値は，会社等の資本金や従業者数によって定められており，具体的には表 2 のとおりである．

ただし，以下に示す事業所・企業は，選定回数の上限設定からは除かれている．

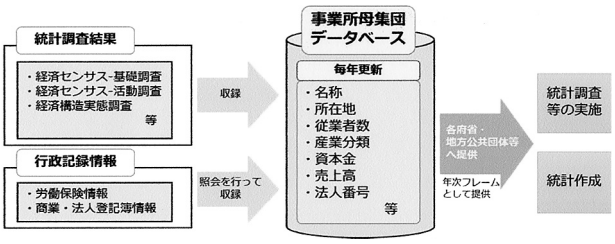


図 2. 事業所母集団データベースの模式図．独立行政法人統計センター（2022）を元に筆者が加工．

表 2. 統計調査の標本として選定される回数の上限．

区　　分	1 年間の上限（回）
1. 会社の本所又は単独事業所	
(1) 資本金 1 億円未満	20
(2) 資本金 1 億円以上 10 億円未満	34
(3) 資本金 10 億円以上	48
2. 会社の支所事業所	
(1) 従業者数 30 人未満	14
(2) 従業者数 30 人以上 100 人未満	28
(3) 従業者数 100 人以上	42
3. 会社以外の事業所	
(1) 従業者数 30 人未満	14
(2) 従業者数 30 人以上 100 人未満	28
(3) 従業者数 100 人以上	30

- 全数調査の対象となるもの
- 集落抽出法による標本調査で，集落内の全てを調査対象とする統計調査の対象となるもの
- 調査実施時に調査対象が決まる統計調査の対象となるもの
- 層別抽出法による標本調査で，一部悉皆部分に含まれるもの
- 国の機関に係るもの並びに地方自治法に規定する普通地方公共団体及び特別地方公共団体（特別区など）の機関に係るもの

6. おわりに

結びに代えて，今後の事業所母集団データベースの更なる発展が期待される事項についていくつか述べる．

6.1 データベースのカバレッジ

事業所母集団データベースは，事業所と企業に関する母集団のフレームとして機能することを意図したものと考えられるが，第 III 期基本計画において「法人企業統計の母集団名簿と事業所母集団データベースの企業数等のかい離」を改善するための方策を検討することとされていることからもわかるように，データベースのカバレッジが完全でない可能性を指摘されていた．実際，2016 年の時点では経済センサス-活動調査で捉えられた法人数が約 190 万法人であるのに対して，法人企業統計の母集団名簿による法人数が約 293 万法人となっており，経済センサスの方が 100 万法人程度少ない状況となっていた．

これを受けて総務省統計局により，2019 年に実施された経済センサス-基礎調査において，国税庁が提供している法人番号公表サイトから約 160 万法人の名簿（名称・所在地）を追加するとともに，2016 年以降に事業所・企業に照会して把握した新たな法人の名簿約 10 万法人を追加して現地調査を通じてそれらの法人の活動状態を確認することが行われた．その結果，活動状態を確認できた約 100 万法人が事業所母集団データベースに新たに登録されるに至った（総務省統計局, 2022）．

このように，事業所母集団データベースのカバレッジは法人企業については大きく改善されてきているが，個人企業に関してはまだ改善の余地が残されていると考えられる．小規模な個人企業については，経済センサスでその存在を把握することが困難な場合があるとともに，先に述べた経済センサス後に実施している事業所・企業に対する照会事業において利用されている労働保険情報及び商業・法人登記簿情報からは，雇用者のいない個人企業に関する情報が得られないため，それらの個人企業についての開業や廃業の情報の更新が現状では困難な状況にあると考えられる．

これに対して，政府部内では国税庁が管轄する個人事業主に関する e-Tax の申告情報を経済センサスにおいて活用できないか検討を行っており（内閣官房, 2021），今後このような検討が実を結び，データベースのカバレッジの更なる改善が進むことが期待される．

6.2 国際連携

ビジネスレジスターに関する国際的な取組については，本稿の第 3 節において紹介したところであるが，今後ますます国際的な連携が重要となってくると考えられるため，ここではこれに関する事項について述べる．

前述のように，ビジネスレジスターに関する国際的な専門家会合であるヴィースバーデングループや UNECE 等によるビジネスレジスターに関する専門家会合に我が国の統計部局からもたびたび参加をして情報交換を行ってきているところである．また，統計ビジネスレジスター

に関する国連ガイドラインの作成に際してはそれを扱う企業・貿易統計専門家委員会が 2018 年度に国連統計委員会の下に設置され（United Nations Statistics Division, 2024b），その会合に我が国統計部局からも参加するなど，我が国としてビジネスレジスターに関する国際的な連携に積極的に関わってきている．

この国連の企業・貿易統計専門家委員会は，企業関連統計の開発のための調整とガイダンスを提供することを目的としており，ビジネスレジスターもその専門家委員会での議論のテーマの一つとして位置付けられている（United Nations Statistics Division, 2024c）．この専門家委員会にはビジネスレジスター，企業の単一識別子，起業家精神などに関するいくつかのタスクチームが設置されている．このうちビジネスレジスターに関しては，ビジネスレジスターに地理空間情報を組み込むこと，多国籍企業のグローバルレジスター，ビジネスレジスターと貿易レジスターの統合，企業の単一識別子に関する世界的取組などが議論されている（United Nations Statistical Commission, 2024）．

近年，企業の多国籍化が進展しており，各国が整備しているビジネスレジスターもその国内だけでは留まらず，国際的な接続・連携が求められるようになってきている．また，ビジネスレジスターに地理空間情報を組み込むに際しては，国際的に標準化された体系に基づくものを利用していく必要がある．このように今後ますます国際的な連携が重要となってくるため，我が国としても引き続き積極的に国際的な活動に関与していくことが期待される．

6.3 レジスター統計

事業所母集団データベースの利用目的のひとつに，事業所に関する統計を作成することがある（統計法第 27 条第 2 項）．これに関連して前述のように，2018 年 3 月に閣議決定され，2020 年 6 月に一部変更された第 III 期基本計画において「レジスター統計」の作成内容について検討を行うこととされた．

総務省では，まず，2019 年の年次フレームを基にレジスター統計を集計することからこの課題への取組を開始した（Yagi et al., 2021）．ただし，事業所母集団データベースは，これまで 5 年に 2 度実施されてきた経済センサスの中間年においては，特定の産業や調査事項等の情報が更新されるのみであるので，レジスター統計を年次で作成していくために一定程度の推計を行っている．推計計算においては，経営組織や産業，事業所・企業が新設か存続かの別によって異なる推計モデルを作成し，付加価値額等の推計が行われている（総務省統計局, 2021c）．

レジスター統計は，これまでのところ，2019 年及び 2020 年の年次フレームを基に 2 カ年分についてその試算値が公表されている．作成されている試算値には，事業所数，企業数，従業者数，売上（収入）金額，付加価値額等の経理事項が含まれている．なお，この試算に用いられている集計方法や推計方法等については研究段階のものであり，今後さらに改善されていくことが期待される．

6.4 ビジネスデモグラフィー統計

事業所母集団データベースから作成される統計としてこれまでに公表されたレジスター統計の試算値は，毎年の年次フレーム作成時点のいわゆるスナップショット（静態）としての事業所・企業の状態を表す統計である．これに加えて原理的には，事業所母集団データベースから事業所・企業の動態を表す統計も作成することが可能である．これはビジネスデモグラフィー統計と呼ばれているものである．ビジネスデモグラフィー統計は，「統計ビジネスレジスターに関する国連ガイドライン」（United Nations, 2020）によると，「特定の期間における企業（や事業所）の開業数，廃業数のデータや，前の期間に開業して存続している企業（や事業所）の数に関するデータを提供するものである」（筆者訳）とされている．

我が国において実際に事業所母集団データベースからビジネスデモグラフィー統計が作成できるようになるためには，例えば税務情報などの更なる情報源に基づいてデータベースがより広範囲・頻繁に更新されるようになる必要があると考えられ，今後の進展が期待されるところである．

なお，複数の経済センサス（又は事業所母集団データベースから作成される年次フレーム）の結果を基にビジネスデモグラフィー統計を確率的に推計する方法も筆者を中心に研究・開発しているところである．一般に，経済センサス等 1 年以上の間隔で実施される統計調査の結果に基づくビジネスデモグラフィー統計の推計においては，いわゆる「短命の開業企業」（経済センサス等の調査実施後に開業し，次の調査の前に廃業してしまいセンサス等では捉えられない企業）が見落とされてしまうことが指摘されている（Ahmad, 2006）．筆者らの研究においては，それらの「短命の開業企業」も考慮に入れた推計モデルを開発し，ビジネスデモグラフィー統計の推計値の試算を行っている（Takahashi, 2022）ことを参考までに記しておきたい．

6.5 ユーザーの拡大

統計法によると，事業所母集団データベースに記録されている情報を利用できるのは，国や地方公共団体などの公的機関に限られている．したがって現状では研究者が研究目的で利用することはできず，研究者が事業所・企業の母集団情報を利用する際には，経済センサスの調査票情報をオンサイト施設などで利用することとなり，その場合は当然，センサスとセンサスの中間年の事業所・企業の新設・廃業等の情報―すなわち，ビジネスデモグラフィーに関する情報―を利用することはできない．

前述のように仮に将来，事業所母集団データベースが今まで以上に広範囲・頻繁に更新されるようになり，このデータベースから直接ビジネスデモグラフィー統計が作成できるような段階になった際には，研究者が研究目的でこのデータベースを利用できるようになることが望まれる．

謝　辞

本稿は，筆者が総務省統計局や独立行政法人統計センター等の公的統計機関に在籍中に携わった統計ビジネスレジスター関連の業務での経験を踏まえてとりまとめたものである．この場をお借りして本稿の執筆に関連してお世話になった関係の方々に深甚の謝意を表する．ただし，本稿における意見や見解は筆者のものであり，筆者の所属する又は所属した組織のそれとは全く関係のないことに留意いただきたい．

付 録

A. 「公的統計の整備に関する基本的な計画」における事業所母集団データベース整備関連の主要な計画

表 A-1. 第Ⅰ期基本計画における事業所母集団データベース整備関連の主要な計画.（「別表 今後5年間に講ずべき具体的施策」のうち,「第2 公的統計の整備に関し総合的かつ計画的に講ずべき施策」部分より）

項 目	具体的な措置，方策等	担当府省	実施時期
2 統計相互の整合性及び国際比較可能性の確保・向上に関する事項 (2) ビジネスレジスターの構築・利活用 ア 母集団情報の的確な整備	○ 経済センサス - 活動調査の中間年に当たる平成26年に，事業所に関する属性情報，企業の親子関係を的確にとらえ，本社と支社の組織的な連携関係を明らかにする経済センサス - 基礎調査を引き続き実施するため所要の準備を行う.	総務省	平成25年度までに所要の準備を実施する.
	○ 法人企業の母集団情報の整備を行うため，往復郵便等による業種名，従業者数，事業所数等の照会を定期的に実施する.	総務省	平成21年度から実施する.
	○ 厚生労働省の協力を得て，雇用保険適用事業所設置届及び労働保険保険関係成立届から事業所等の新設，廃止等を把握することについて検討する.	総務省	平成22年から検討する.
イ ビジネスレジスターの充実と拡張	○ 工業統計調査の出荷額等，全数調査の調査結果の他，一定規模以上の企業に関する法人企業統計調査の売上高等の主要な経理情報をビジネスレジスターの情報源として利用することについて，関係府省との検討を開始する.	総務省	平成21年度から検討する.
	○ ＥＤＩＮＥＴ情報をビジネスレジスターに収納することを検討する．併せて，ＥＤＩＮＥＴ情報とビジネスレジスターの情報を法人企業統計に活用する具体的方策を検討する.	総務省，財務省	平成21年度から検討する.
	○ 特許庁の協力を得て，産業財産権の企業出願人の名称及び所在地と企業の登記情報との照合作業を行い，ビジネスレジスターに両者の照合情報を収納する.	総務省	平成21年度から検討を開始し，速やかに実施する.
	○ 事業所・企業識別番号と「日本輸出入者標準コード（ＪＡＳＴＰＲＯコード）」（輸出入申告書，蔵入承認申請書，積戻し申告書等に記載されている輸出入者に対応したコード）の照合を行うに当たり，費用対効果を考慮しつつ，どのような有用性が得られるかについて検討を開始する.	総務省	平成21年度から検討する.

表 A-2. 第 II 期基本計画における事業所母集団データベース整備関連の主要な計画.（「別表 今後 5 年間に講ずる具体的施策」のうち，「第 3 公的統計の整備に必要な事項」部分より）

項　目	具体的な措置，方策等	担当府省	実施時期
1　統計作成の効率化及び報告者の負担軽減 (1) 事業所母集団データベースの整備・利活用	○　新たな行政記録情報等の活用や企業への直接的な照会などにより企業組織構造の変化を経常的に確認する方法について検討し，母集団情報の効果的かつ効率的な整備を推進する．また，これらの取組に当たり，専門性を持つ人材の育成や体制整備等に努める．	総務省	平成 26 年度から順次実施する．
	○　事業所母集団データベースを活用して，我が国の事業所・企業の実態を把握する統計に加え，事業所・企業の異動状況や産業の成長・衰退等に着目した統計を作成する．また，地理情報の活用等についても研究を推進する．	総務省	平成 26 年度から順次実施する．

表 A-3. 第 III 期基本計画における事業所母集団データベース整備関連の主要な計画（その 1）.（「別表 今後 5 年間に講ずる具体的施策」のうち，「第 2 公的統計の整備に関する事項」部分より）（注）表 A-3～表 A-5 の「具体的な措置，方策等」欄において，基幹統計に係る事項を「◎」とし，その他の公的統計に係る事項を「○」としている．

項　目	具体的な措置，方策等	担当府省	実施時期
1　国民経済計算を軸とした横断的・体系的な経済統計の整備推進 (2) 経済構造統計を中心とした経済統計の体系的整備の推進等	○　事業所母集団データベースに格納される企業対象の統計調査結果における収録事項の定義の統一・共通的に把握すべき調査事項を設定した上で，同データベースを活用した企業統計の提供を推進するとともに，大規模企業の活動実態を横断的に把握する統計整備を検討する．	総務省，関係府省	平成 34 年度 (2022 年度) までに一定の結論を得る．

表 A-4. 第 III 期基本計画における事業所母集団データベース整備関連の主要な計画(その 2).
(「別表 今後 5 年間に講ずる具体的施策」のうち,「第 2 公的統計の整備に関する事項」部分より)

項 目	具体的な措置,方策等	担当府省	実施時期
(3) 国民経済計算及び経済統計の改善に向けた基盤整備・連携強化	◎ 経済センサス - 基礎調査をローリング調査や独立行政法人統計センターにおけるプロファイリング活動に移行し,事業所・企業や各種法人等に係る母集団情報の更なる整備促進を図る.また,平成 33 年 (2021 年) 経済センサス-活動調査の円滑・効率的な実施に向け,行政記録情報等により新たに捕捉した事業所等も含めた確認作業を早期に完了するとともに,プロファイリング活動において把握する情報や把握方法等に関する具体的な検討を推進し,その内容を事前周知する.	総務省	平成 31 年度 (2019 年度) から実施(初回のローリング調査は 32 年 (2020 年) 年央までに実施)する.
	○ 事業所・企業や各種法人等に係る統計調査を実施するに当たり,引き続き法人番号の把握に努め,これを事業所母集団データベースに登録するとともに,法人番号を活用した欠測値の補完や集計の充実等を検討する.	各府省,総務省	平成 30 年度 (2018 年度) から実施する.
	○ 農林業センサスにおける一戸一法人・非法人の組織経営体や,建設業許可事業者名簿に掲載された企業情報を,事業所母集団データベースに登録する方向で検討する.	総務省,関係府省	平成 30 年度 (2018 年度) 末までに結論を得る.
	○ 事業所母集団データベースに格納する統計調査の範囲拡充,法人番号公表サイトの変更情報や決算等に関する企業の公表情報等の活用,行政記録情報や民間データの活用及びレジスター統計の作成内容について検討を行い,結論が得られた取組から順次実施する.	総務省,各府省	平成 30 年度 (2018 年度) から順次実施する.
	○ 法人企業統計の母集団名簿と事業所母集団データベースの企業数等のかい離について,産業や資本金階級別などの企業属性ごとにかい離の状況を明らかにするとともに,その要因を把握した上で,かい離を改善するための方策を検討する.	総務省,財務省	令和 3 年度 (2021 年度) 末までに結論を得る.
	○ 専従の役員・労働者等が存在しない法人等を含めた事業所母集団データベースに登録する法人・事業所等の情報について,事業所母集団データベースへの格納方法や,具体的な母集団情報としての提供を検討する.	総務省	平成 30 年度 (2018 年度) 末までに結論を得る.
	○ 月次・年次調査の標本抽出に資するため,事業所の活動状態を随時更新した母集団情報の提供を検討する.	総務省	平成 30 年度 (2018 年度) 末までに結論を得る.

表 A-5. 第 IV 期基本計画における事業所母集団データベース整備関連の主要な計画.（「別表 今後 5 年間に講ずる具体的施策」のうち，「第 2 公的統計の整備に関する事項」部分より）

項　目	No.	具体的な措置，方策等	担当府省	実施時期
2　経済統計の体系的整備の推進 (3) 経済計作成の改善に向けた取組	25	◎　経済センサス - 活動調査の中間年における事業所母集団データベースの更なる整備を図るため，報告者の負担や地方公共団体・統計調査員の事務負担にも配慮しつつ，事業所・企業の基本的事項の更新範囲や頻度を高めることについて，経済センサス - 基礎調査の見直しを含めて検討を行う.	総務省	令和 6 年経済センサス - 基礎調査の企画時期までに結論を得る.

参　考　文　献

Ahmad, N. (2006). A proposed framework for business demography statistics, OECD Statistics Working Papers, No.2006/03, OECD Publishing, Paris, https://doi.org/10.1787/145777872685.

Blancas, A. (2022). Country progress report on SBR 2021, Meeting of the Group of Experts on Business Registers, https://unece.org/sites/default/files/2022-09/S_5_Mexico%20CPR.pdf（最終アクセス日 2024 年 3 月 27 日）.

独立行政法人統計センター (2022). 事業所母集団データベースの整備について, https://www.nstac.go.jp/sys/files/2022_rengo_02.pdf（最終アクセス日 2024 年 3 月 27 日）.

Eurostat (2010). *Business Registers Recommendations Manual*, European Union, Luxembourg.

平田佳嗣 (2007). 国民の財産であり、社会の情報基盤としての統計〜統計法案〜, 立法と調査, **267**, 9–16.

井上達夫 (2008). 60 年ぶりの統計制度改革—新統計法と 2 つのキーワード—, 日本統計学会誌, **37**(2), 301–312.

各府省情報化統括責任者連絡会議 (2006). 統計調査等業務の業務・システム最適化計画, https://www.stat.go.jp/info/guide/public/keikaku/pdf/honbun.pdf（最終アクセス日 2024 年 3 月 27 日）.

Kurematsu, R. and Tanabe, T. (2023). Efforts for further development and enhancement of Statistical Business Register in Japan — Improvement of update frequency —, 28th Meeting of the Wiesbaden Group on Business Registers, https://cdn.aanmelderusercontent.nl/i/doc/cd126a3233eacc8014f6dfb78395396e（最終アクセス日 2024 年 3 月 27 日）.

内閣府経済社会統計整備推進委員会 (2005). 政府統計の構造改革に向けて, https://warp.da.ndl.go.jp/info:ndljp/pid/11670228/www5.cao.go.jp/keizai-shimon/special/statistics/promote/report.pdf（最終アクセス日 2024 年 3 月 27 日）.

内閣官房 (2021). 経済センサスへの税務情報の活用について, https://www.kantei.go.jp/jp/singi/toukeikaikaku/toukeigijutsu_data_source/dai4/siryou4.pdf（最終アクセス日 2024 年 3 月 27 日）.

総務省 (2011). 事業所母集団データベースの整備方針，https://www.stat.go.jp/info/kenkyu/jsdb/pdf/23_1-2-1.pdf（最終アクセス日 2024 年 3 月 27 日）.

総務省 (2013). 日本標準産業分類一般原則，https://www.soumu.go.jp/main_content/000286955.pdf（最終アクセス日 2024 年 3 月 27 日）.

総務省統計局 (2021a). 経済センサス-活動調査調査票（産業共通）, https://www.stat.go.jp/data/e-census/2021/pdf/chousahyo_01.pdf（最終アクセス日 2024 年 3 月 27 日）.

総務省統計局 (2021b). 令和 3 年経済センサス-活動調査用語の解説, https://www.stat.go.jp/data/e-census/2021/kekka/pdf/k_yougo.pdf（最終アクセス日 2024 年 3 月 27 日）.

総務省統計局 (2021c). レジスター統計(試算値)の作成に関する研究結果(中間報告), https://www.stat.go.jp/info/kenkyu/jsdb/register_index.html (最終アクセス日 2024 年 3 月 27 日).

総務省統計局 (2022). 事業所母集団データベースの整備・利活用について, https://www.soumu.go.jp/main_content/000832068.pdf (最終アクセス日 2024 年 3 月 27 日).

総務省統計局 (2024). 事業所母集団データベース(ビジネスレジスター), https://www.stat.go.jp/data/jsdb/index.html (最終アクセス日 2024 年 3 月 27 日).

髙橋雅夫 (2007a). 新しいビジネスフレームの開発—第 20 回ビジネスフレーム円卓会議発表論文—その 1 背景と経緯, 統計, 2007 年 11 月号, 49–54.

髙橋雅夫 (2007b). 新しいビジネスフレームの開発—第 20 回ビジネスフレーム円卓会議発表論文—その 2 概要と課題, 統計, 2007 年 12 月号, 33–38.

Takahashi, M. (2022). A method for estimating the number of short-lived births of businesses based on a stochastic model, *Procedia Computer Science*, **207**, 494–503.

統計制度改革検討委員会 (2006). 統計制度改革検討委員会報告, https://warp.da.ndl.go.jp/info:ndljp/pid/11670228/www5.cao.go.jp/keizai-shimon/special/statistics/reform/report.pdf (最終アクセス日 2024 年 3 月 27 日).

United Nations (2020). United Nations Guidelines on Statistical Business Registers (Final draft prior to official editing), New York, https://unstats.un.org/unsd/business-stat/SBR/Documents/UN_Guidelines_on_SBR.pdf (最終アクセス日 2024 年 3 月 27 日).

United Nations Economic Commission for Europe (2015). *Guidelines on Statistical Business Registers*, United Nations, New York and Geneva.

United Nations Statistical Commission (2024). Report of the Committee of Experts on Business and Trade Statistics, https://unstats.un.org/UNSDWebsite/statcom/session_55/documents/2024-18-Business&Trade-E.pdf (最終アクセス日 2024 年 3 月 27 日).

United Nations Statistics Division (2024a). Wiesbaden Group on Business Registers, https://unstats.un.org/wiesbadengroup/ (最終アクセス日 2024 年 3 月 27 日).

United Nations Statistics Division (2024b). Committee of Experts on Business and Trade Statistics, https://unstats.un.org/unsd/business-stat/UNCEBTS.cshtml (最終アクセス日 2024 年 3 月 27 日).

United Nations Statistics Division (2024c). Mandate and governance of the Committee of Experts on Business and Trade Statistics, https://unstats.un.org/unsd/business-stat/UNCEBTS/Documents/Mandate_and_governance.pdf (最終アクセス日 2024 年 3 月 27 日).

Yagi, M., Yoshida, T. and Tsuji, K. (2021). New Outputs of the Business Register in Japan (Register-based Statistics), 27th Meeting of the Wiesbaden Group on Business Registers, https://www.inegi.org.mx/eventos/2021/wiesbaden/doc/3_1_PAP_YAGI_YOSHIDA_TSUJI.pdf (最終アクセス日 2024 年 3 月 27 日).

Development of the Statistical Business Register in Japan —"Establishment Frame Database" as an Important Information Base for Society—

Masao Takahashi

Faculty of Business and Infomatics, Nagano University

This report provides an overview and possible future directions for developing the Establishment Frame Database, stipulated by the current Statistics Act, which positions official statistics as an information base for society. Databases such as the Establishment Frame Database in Japan are internationally referred to as statistical business registers, which are databases of economic units such as enterprises and establishments and are used for statistical purposes; they are typically maintained and managed by national statistical offices in various countries. The Establishment Frame Database is designed to utilize information sources such as the Economic Census, various statistical surveys, and administrative records to provide population information about establishments and enterprises for various statistical surveys and to create statistics related to establishments. In this report, after a brief history of the Establishment Frame Database is provided, an overview of the database is presented, followed by discussions of future possible directions, including further improvement of the coverage of the database and international cooperation.

Key words: Economic Census, administrative records information, sampling frame, register-based statistics, business demography.

統計数理(2024)
第 72 巻 第 2 号 195–215
©2024 統計数理研究所
特集「公的統計―社会の重要な情報基盤―」
[原著論文]

労働時間を考慮した日本の消費構造に関する計量分析

―全国家計構造調査を用いて―

伊藤 伸介[1]・出島 敬久[2]・村田 磨理子[3]

(受付 2023 年 12 月 31 日；改訂 2024 年 6 月 7 日；採択 7 月 2 日)

要　　旨

日本を含む先進国全般に，共働き世帯の割合が増加している．それに伴って，夫婦の所得の構成と時間の配分が変化しつつあることから，様々な経済政策の効果を検討するためには，世帯類型ごとに夫婦の所得の構成と労働時間が消費支出に与える影響を調べることが求められる．その関係について，本稿では，調査項目として新たに週間就業時間が追加された全国家計構造調査の個票データを用いて，説明変数に労働時間を導入することで明らかにした．

具体的には，先行研究で推定されてきた家計消費の詳細費目に関する消費関数を設定した上で，説明変数に労働時間を新たに取り入れた．さらに，消費に与える恒常所得の影響をより正確に考慮するために，賃金構造基本統計調査で推定される賃金関数をもとにした賃金の期待値と実現値の差も説明変数に導入した．

その結果，第 1 に先行研究と同様に，夫婦の働き方の違いが家事消耗品や宿泊費等の支出に影響を与えることが観察された．第 2 に，労働時間が消費の費目に一部ではあるが影響をもたらしていることが確認された．以上の点については，共働き世帯で配偶者の労働時間が長くなるほどに，時間節約型の消費傾向が強まることによって説明することができる．

キーワード：家計消費，消費関数，共働き世帯，労働時間．

1. はじめに―働き方の変化と家計の消費行動

他の先進国同様に日本でも，夫婦に関しては，片働き世帯の割合が減少し，共働き世帯の割合が増加している(図 1)．その共働き世帯においても，配偶者がパート・アルバイトなどの短時間労働だけでなく，正規就業の割合が増大している．また，配偶者が非正規就業でも，パート・アルバイトの他に，有期契約労働者，派遣労働者の様々な就業形態が増加しており，就業形態は多様化しているといえる．こうした夫婦の働き方の違いによって，夫婦の所得の構成は大きく異なるため，消費財の選好に関する両者の交渉力や時間制約の影響を通じて，消費支出を変化させる可能性がある．所得を変化させる様々な経済政策が消費に与える影響を評価する場合には，そのような働き方の違いに注意した世帯類型ごとに，消費の構造を捉えることが求

[1] 中央大学 経済学部：〒 192–0393 東京都八王子市東中野 742–1
[2] 上智大学 経済学部：〒 102–8554 東京都千代田区紀尾井町 7–1
[3] (公財)統計情報研究開発センター：〒 101–0051 東京都千代田区神田神保町 3–6 能楽書林ビル 5F

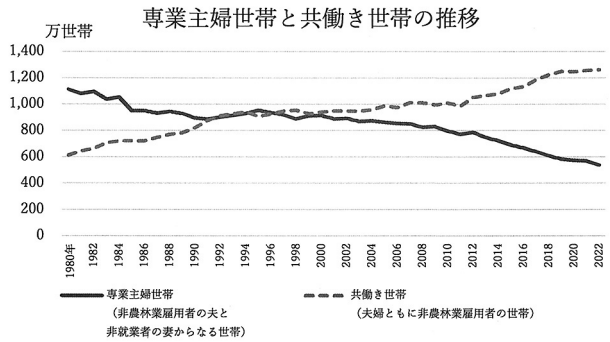


図 1．専業主婦世帯と共働き世帯の推移．出典：労働政策研究・研修機構「早わかり　グラフでみる長期労働統計」II 労働力，就業，雇用，図 12.（注：本図は，総務省統計局「労働力調査（一部は特別調査）」各年から作成されている．）

められる．

本稿では，公的統計の個票データを用いて，世帯類型ごとに，就業を所与とした場合に世帯主と配偶者の所得の構成が双方の選好に関わる費目の消費支出に与える影響について明らかにする．

具体的には，これまでの先行研究を踏まえ，全国家計構造調査の個票データを用いて，その詳細費目の消費支出に関して，夫婦の所得の影響を推定する．さらに，恒常所得の影響をより正確に制御するために，賃金構造基本統計調査の個票データも併用しながら，先行研究のモデルに加えて，労働時間に関する情報を有効に活用した上で，世帯類型や就業特性を考慮した消費構造との関係を実証的に明らかにする．

2. 近年の男女の雇用形態の変化

実証分析に先立ち，夫婦の労働時間を大きく決定づける雇用形態の変化を，男女別に比較しておこう．図 2-1 と図 2-2 は，日本の雇用形態別の雇用者数を男女別に示したものである．

まず，雇用者数全体としては，男性でほとんど伸びが見られない．その一方で，女性では持続的な増加が観察される．さらに，男性と女性で，正規雇用と非正規雇用（パート・アルバイト，派遣社員，契約社員など）の構成が大きく異なり，女性では男性に比べて，非正規雇用の構成割合が依然として大きいことがわかる．

ただし，女性については，雇用者数全体の伸びが続いている中で，雇用者数のうち正規の職員・従業員（正規雇用者）が 2015 年頃から増加していることも観察される．これについては，2012 年の労働契約法の改正によって，非正規雇用で標準的な有期労働契約がその後の更新により通算 5 年を超えた場合，労働者が申し込めば無期労働契約に転換できる権利（無期転換申込権）が設けられたことが指摘できる．改正労働契約法の施行は 2013 年 4 月であり，そこから通算 5 年となる有期労働契約が発生する 2018 年 4 月から，無期転換申込権の適用が開始されている．この転換申込権に対応できるように，企業など雇用主は，有期労働契約の見直しを事前に進めていたことが知られている．また，非正規雇用の中でも，派遣社員，契約社員等，労働時間が比較的長い雇用形態の増加も観察される．

以上のことは，まず夫婦での片働きから共働き世帯の増加とともに，共働き世帯の中でも，女性の正規雇用の増加，さらに非正規雇用の中でも比較的長時間の労働である雇用形態の増加を示している．このことは，夫婦の所得と労働時間の構成を変化させ，夫婦双方の消費に関す

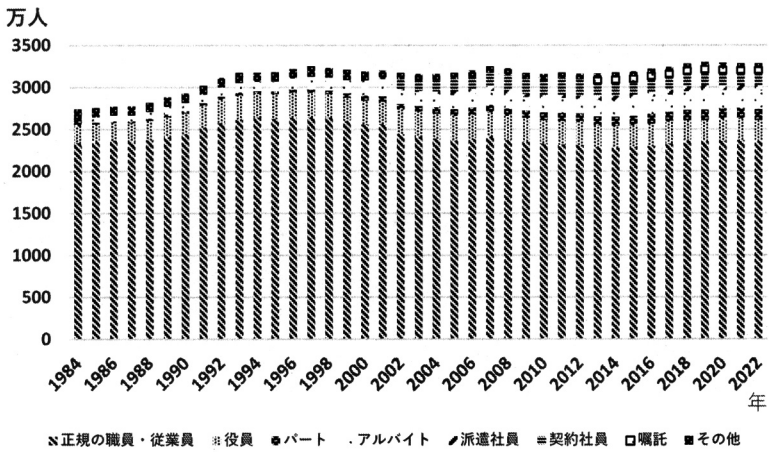


図 2-1. 雇用者の雇用形態の構成：男性．出典：労働政策研究・研修機構「早わかり　グラフでみる長期労働統計」II 労働力，就業，雇用の図 8-2.（注：本図は，総務省統計局「労働力調査（一部は特別調査）」各年から作成されている．ここでの派遣社員とは，労働者派遣事業所の派遣社員に限ったものである．2012 年以前は，契約社員に嘱託が含められている．）

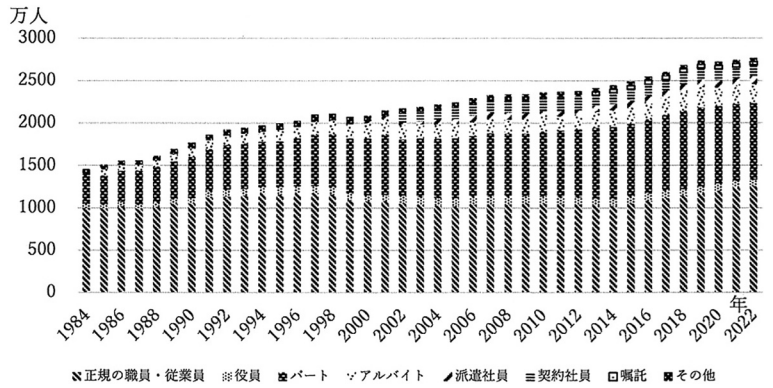


図 2-2. 雇用者の雇用形態の変動：女性．出典：図 2-1 に同じ．原典と注に関しても同様．

る交渉力や家計内生産の時間的制約に影響を与える可能性がある．この変化は，政府の財政政策や社会保障制度の効果を変えるために，以上の関係の実証が求められる．

3. 理論モデルと先行研究

3.1 家計消費の理論モデル

家計消費の理論的説明に関しては，家計の生涯にわたる効用最大化から得られる条件であるオイラー方程式が基本として知られている．また，その実証研究においては，恒常所得の影響とリスク回避度を勘案した消費の平準化についての検証がなされてきた．それに関しては，日本でも世帯単位のミクロデータを利用した計量分析の方向が追究されている（阿部, 2011）．

近年の日本を含む先進国での共働きが増加することによって，恒常所得の稼得主体が夫と妻になることが指摘できる．このことは，家計の選択が世帯主の選択と同一視できない部分を拡

げることから，効用を最大化する主体もまた，家計ではなく，夫と妻のそれぞれと位置付けられる余地が拡大することを示唆している（伊藤・出島, 2022）．

3.2 世帯における所得の構成と消費構造に関する先行研究

世帯構成員の所得の比率に応じて，消費にどのような影響を与えるかを検証した日本における先行研究としては，例えば，Hayashi（1995），小原（2008），澤田（2008），坂本（2008）等がある．

坂本（2008）によれば，家計経済研究所が 2006 年に実施した「世帯内分配と世代間移転に関する研究」の調査個票データを用いることによって，夫婦の働き方の組みあわせによって両者の交渉力が変わることが，消費や余暇の配分に影響をもたらすことを実証的に明らかにしている．具体的には，夫妻における自由裁量消費（被服・履き物・教養・娯楽費）の比率，余暇時間の配分に関する夫妻での比率の決定要因が定量的に追究されている．それによって，本分析では，夫と妻のそれぞれの賃金と一部では子どもの人数が，これらの比率に有意な影響を与えていることが検証されている．妻の賃金が高くなるにつれて，妻の自由裁量消費が傾向的に大きくなるだけでなく，余暇時間に関しても，妻の労働時間が長く，余暇時間の比率が小さくなる傾向にあることが確認されている．

これらの先行研究においては，世帯員の中で高所得である者の交渉力が相対的に高くなり，高所得である世帯員にとって限界効用の高い財の支出構成が大きくなる傾向が明らかにされている．その結果，配偶者の稼得所得の増大が，配偶者にとって限界効用が高い財の支出の増加が予想される．

配偶者の稼得所得の向上とともに労働時間が増加した場合，家事時間（家計内生産時間）を節約するインセンティブが大きくなるため，調理済み食品等，時間節約的な財の購入が拡大する影響も予想される．また，世帯主と配偶者の選好が異なる費目，例えば妻の衣服や子供の経費のような費目については，配偶者の所得が世帯の所得全体に占める比率が上昇すると，それらの支出割合が増大することが考えられる．

こうした労働時間と家計内生産を考慮した消費支出との関係を推定するには，世帯員の労働時間と消費の支出構成がすべて観察されるデータが求められる．あいにく日本には，それらのすべてが観察される公的統計がない．そこで，阿部 他（2021）は，余暇目的の時間と金銭投入に関して独自調査をすることで，家計内生産関数の形状に関する推計を行っている．その結果，労働時間と家計内生産に用いる財投入の間に強い代替性はなく，十分な時間を余暇に割くことができない人は，財投入も少なくなることが示されている．このことは，本稿との関係で述べれば，労働時間が長いほど，家計内生産に多く用いる財の支出が少なくなることを示唆している．

伊藤・出島（2022）や伊藤 他（2023）は，全国消費実態調査の個票データを用いて，配偶者の就業状況をより細分化した上で，世帯主と配偶者の働き方の差が消費の費目に与える影響を実証的に明らかにした．

伊藤・出島（2022）では，賃金構造基本統計調査を用いて推計された賃金関数から算出された，勤め先収入の実現値と期待値の差が，どの費目に対して消費の変化をもたらすかを追究した．本分析から，配偶者の働き方の差が食料費や教育費等の支出に影響を与えることが確認された．また，賃金の実現値と期待値の差分が消費支出に及ぼす影響については，賃金の実現値が期待値を上回った場合に，食料費や教養娯楽費といった費目で家計の消費支出が増大する傾向にあることがわかった．

伊藤 他（2023）における分析結果からは，配偶者の選好に関わる消費支出の詳細費目については，配偶者の所得の比率が上昇すると，消費の比率が高まる費目があることから，配偶者の

所得比率の増大は，家計内の配偶者の交渉力を相対的に拡大させる傾向にあることが明らかにされた．さらに時間節約的な財の購入量についても影響を受けている部分が観察された．

4. 使用するデータ

本研究で使用するデータは，2019 年の全国家計構造調査の個票データ，および 2019 年の賃金構造基本統計調査の個票データである．本研究では，世帯主が 20 歳〜59 歳である核家族世帯を対象とした上で，世帯主と配偶者の就業形態(正規就業・非正規・非就業)の組み合わせに基づいて，計量分析を行う．また，世帯類型を勘案するために，分析においては末子の年齢についても考慮する．さらに本研究は，勤労者世帯のみを分析の対象とする．

全国家計構造調査の特徴は以下のとおりである．

全国家計構造調査は，2019 年より実施されている，5 年に 1 回の調査でサンプル・サイズが約 90000 世帯である．家計の消費や所得さらには貯蓄の詳細な情報を把握するために，家計における年単位の所得と資産・負債情報，および調査月である 10 月と 11 月の 2 ヶ月間について家計簿を回答する調査である．また，10 大費目といった区分で消費支出を詳細に把握できることから，就業状況をコントロールした場合の所得の構成と消費支出の関係を精密に分析することができる．その意味では，勤労所得と勤め先の属性が把握できるだけでなく，家計資産および資産所得等の情報も捕捉する唯一の公的統計ということができる．

2014 年までの全国消費実態調査と 2019 年の全国家計構造調査では，標本設計は異なる．したがって，モデル分析を行った場合のパラメータの比較可能性については留意することが求められる．また，全国家計構造調査の場合，収支項目が把握されているのは，「ロングフォーム」の調査票に回答した世帯に限定される．このため，報告者負担の軽減から，標本の一部については家計調査の対象世帯が組み込まれている．

全国家計構造調査の留意点として，標本設計においては，家計収支に関する情報の一部については，家計調査で捕捉される収支項目に関する情報が援用されていることが指摘できる．したがって，全国家計構造調査の調査対象世帯と家計調査が援用された世帯におけるデータ特性が異なる可能性がある．また，全国家計構造調査の調査事項については，学歴や労働時間といった拡充された項目もあるが，産業，世帯主以外の構成員の職業のような削除された項目も存在する．

全国家計構造調査において，学歴や労働時間が新たに捕捉されたことによって，例えば学歴を含めた形で，労働時間あたりの賃金を被説明変数とする賃金関数の推定が可能になる．なお，労働時間については，世帯員ごとに，「ふだんの 1 週間の就業時間」(以下「週間就業時間」と呼称)を尋ねている．調査世帯に配布される記入のしかたでは，就業規則などで定められた就業時間(所定内労働時間)に関係なく，1 週間の実労働時間(いわゆる残業時間を含む)を回答することが求められている．以下では，この週間就業時間を労働時間として解釈する．

週間就業時間については，無回答である世帯が存在する．その理由の 1 つとして，事業場外みなし労働時間制，裁量労働制が該当する者が，みなし労働時間適用者に含まれることが考えられる．さらに，みなし労働時間もない場合でも，高度プロフェッショナル制度の対象者は，週間就業時間については無回答者に該当する場合がある．

表 1 は，厚生労働省の「就労条件総合調査」に基づくみなし労働時間制の適用労働者の割合を示したものである．表 1 を見ると，みなし労働時間が適用される労働者は，事業場外みなし労働時間制では 7.6%，裁量労働制では 1.2% に留まる．このことから，みなし労働時間制の適用対象となる労働者が，無回答者の一部に存在するものと考えられる．

表 1. みなし労働時間制の適用労働者の構成割合．出典：厚生労働省「就労条件総合調査」，2020 年 1 月 1 日現在.

労働者の割合(%)	みなし労働時間制の適用を受ける労働者	適用制度の種類			みなし労働時間制の適用を受けない労働者
		事業場外みなし労働時間制	専門業務型裁量労働制	企画業務型裁量労働制	
企業規模計	8.9	7.6	1.0	0.2	91.1
1,000 人以上	11.4	9.2	1.8	0.3	88.6
100〜999 人	8.2	7.3	0.7	0.2	91.8
30〜99 人	5.8	5.3	0.3	0.3	94.2

5. 世帯主と配偶者の所得の構成が双方の選好に関わる費目の消費支出に与える影響

本節では，Ito and Dejima (2020)，伊藤・出島 (2022)，伊藤 他 (2023)における先行研究を踏まえて，全国家計構造調査で初めて調べられた労働時間に関する情報を活用するように拡張する．以上の先行研究では全国消費実態調査で把握される 10 大費目だけでなく，中分類費目や小分類費目といった詳細な消費支出にも着目し，世帯主と配偶者の双方の選好が異なる費目まで細分化した上で，世帯主と配偶者の所得の構成がそれらの詳細な費目に及ぼす影響を事実発見的に明らかにしたが，本研究では，それらの情報に追加して労働時間に関する情報を説明変数に導入する．

最初に，伊藤・出島 (2022)や伊藤 他 (2023)との比較を可能にするために，これらの先行研究で導入したモデルに基づいて，本研究ではモデル 1a，モデル 1b，およびモデル 2 を設定した．このうちモデル 1a(式(5.1a))，モデル 1b(式(5.1b))は，消費支出に対する第 i 費目の比率が，世帯主の所得(勤め先からの年間収入)と配偶者の勤め先収入がそれに占める比率に影響を受けることを，他の世帯属性を制御した上で定量化するものである．

$$\frac{C_i}{C} = \beta_{a0i} + \beta_{aHi} \ln Y_H + \beta_{aWi} \frac{Y_W}{Y_H} + \beta_{aMi} \frac{Y_M}{Y_H} + \boldsymbol{X_1 \gamma_{ai}} + \varepsilon_{ai} \tag{5.1a}$$

$$\frac{C_i}{C} = \beta_{b0i} + \beta_{bHi} \ln Y_H + \beta_{bWi} \frac{Y_W}{Y_H} + \beta_{bMi} \frac{Y_M}{Y_H} + \alpha_{HHi} H_H + \alpha_{HWi} H_W + \alpha_{EHi} E_H + \alpha_{EWi} E_W + \boldsymbol{X_1 \gamma_{bi}} + \varepsilon_{bi} \tag{5.1b}$$

ここで，各変数とパラメータは以下を表している．

$\frac{C_i}{C}$：消費支出に対する第 i 費目の比率

Y_H, Y_W, Y_M：世帯主，配偶者，それ以外の世帯員(65 歳未満)それぞれの勤め先からの年間収入

H_H, H_W：世帯主と配偶者それぞれの週間就業時間ダミー

E_H, E_W：世帯主と配偶者それぞれの学歴ダミー

表 2. 本研究で用いたモデルで設定される説明変数の一覧.

	モデル1a	モデル1b	モデル2
被説明変数	消費支出に対する詳細費目の比率	消費支出に対する詳細費目の比率	消費支出に対する詳細費目の実額
説明変数	配偶者の正規就業ダミー	配偶者の正規就業ダミー	配偶者の正規就業ダミー
	配偶者のパート・アルバイト就業ダミー	配偶者のパート・アルバイト就業ダミー	配偶者のパート・アルバイト就業ダミー
	配偶者の派遣労働ダミー	配偶者の派遣労働ダミー	配偶者の派遣労働ダミー
	配偶者のその他就業ダミー	配偶者のその他就業ダミー	配偶者のその他就業ダミー
	持ち家の有無ダミー	持ち家の有無ダミー	持ち家の有無ダミー
	性別ダミー	性別ダミー	
	世帯主の年齢(各歳)	世帯主の年齢(各歳)	世帯主の年齢(各歳)
	世帯主の年齢の2乗	世帯主の年齢の2乗	世帯主の年齢の2乗
	勤め先からの年間収入(対数)	勤め先からの年間収入(対数)	勤め先からの年間収入(対数)
	配偶者の勤め先収入の年間収入に対する比率	配偶者の勤め先収入の年間収入に対する比率	配偶者の勤め先収入の年間収入に対する比率
	65歳未満の世帯員の勤め先収入の年間収入に対する比率	65歳未満の世帯員の勤め先収入の年間収入に対する比率	65歳未満の世帯員の勤め先収入の年間収入に対する比率
	財産収入(対数)	財産収入(対数)	財産収入(対数)
	貯蓄現在高 (対数)	貯蓄現在高 (対数)	貯蓄現在高 (対数)
	住宅・土地購入のための借入金残高(対数)	住宅・土地購入のための借入金残高(対数)	住宅・土地購入のための借入金残高(対数)
	住宅・宅地資産額(対数)	住宅・宅地資産額(対数)	住宅・宅地資産額(対数)
	大都市圏ダミー	大都市圏ダミー	大都市圏ダミー
	末子の年齢ダミー	末子の年齢ダミー	末子の年齢ダミー
		世帯主の週間就業時間ダミー	世帯主の週間就業時間ダミー
		配偶者の週間就業時間ダミー	配偶者の週間就業時間ダミー
		世帯主の学歴ダミー	世帯主の学歴ダミー
		配偶者の学歴ダミー	配偶者の学歴ダミー
			世帯主の勤め先からの時給の実現値-期待値
			世帯主の配偶者の勤め先からの時給の実現値-期待値

$\boldsymbol{X_1}$：その他に制御すべき家計の属性(表 2 を参照)

$\beta_{a0i}, \beta_{aHi}, \beta_{aWi}, \beta_{aMi}, \alpha_{HHi}, \alpha_{HWi}, \alpha_{EHi}, \alpha_{EWi}, \boldsymbol{\gamma_{ai}}, \boldsymbol{\gamma_{bi}}$：各説明変数の回帰係数

$\varepsilon_{ai}.\varepsilon_{bi}$：誤差項

なお，表 2 は，これらのモデルに含まれる被説明変数と説明変数の一覧を示している．伊藤・出島 (2022)では，世帯主と配偶者の勤め先からの年間収入の対数がモデルに設定されているが，本研究では，伊藤 他 (2023)でも導入されている勤め先からの年間収入(対数)と配偶者の勤め先収入の年間収入に対する比率をモデルの説明変数として用いている．

モデル 1a とモデル 1b では，被説明変数として，家計消費の 10 大費目の詳細費目である中分類費目あるいは小分類費目を対象に，消費支出に対する各比率を用いている．具体的には，食料費については，家事時間の節約を把握するために食料費の内訳である主食的調理食品や一般外食を分析の対象とした．さらに，子育てに家計内でかける時間を学習支援業のサービスの

購入で代替するために，教育費に関しては，内訳である補習教育の消費支出に対する比率を算出した．

本研究では，食料費と教育費以外については，世帯主と配偶者の選好に関わる中分類あるいは小分類の費目として，家賃，電気代，ガス代，上下水道料，家事用消耗品，家事サービス，被服，交通，自動車等維持，ガソリン，通信，宿泊料，月謝類，理美容サービス，こづかい，交際費，および仕送り金を選定した上で，これらのおのおのについて消費支出に対する各費目の比率を算出した．なお，2019 年の全国家計構造調査の場合には，男女別に被服費を算出することができないため，被服費全体で分析を行った．

本研究においては，夫婦が共働きか片働きかによって，さらに配偶者の就業形態の違いが，消費支出の詳細費目別比率にどのような影響を与えるかを明らかにするために，配偶者の就業形態別に重回帰分析を行った．

また，本研究では，伊藤・出島 (2022)や伊藤 他 (2023)で定義されたモデルを拡張して，モデル 1a における説明変数に世帯主の週間就業時間ダミー，配偶者の週間就業時間ダミー，世帯主の学歴ダミーと配偶者の学齢ダミーを追加的に設定したモデル 1b も設定した上で，計量分析を行った．

モデル 1b の特徴は，全国家計構造調査で把握可能な週間就業時間と学歴も説明変数として用いられていることである．週間就業時間は，調査事項「ふだんの 1 週間の就業時間」における分類区分，すなわち①15 時間未満，②15〜29 時間，③30〜34 時間，④35〜39 時間，⑤40〜48 時間，⑥49〜59 時間，⑦60 時間以上，⑧決まっていない，および⑨休業中を用いて設定された．なお，本分析では，労働時間が 0 時間の場合を無業とし，「⑨休業中」については無業に，「⑧決まっていない」については，「①15 時間未満」に含めることにした．

これらをもとに，世帯主の週間就業時間ダミー，および配偶者の週間就業時間ダミーのカテゴリーは以下のように設定された．

世帯主の週間就業時間ダミーについては，週間就業時間が 40 時間以上の場合は 1，それ以外は 0 とした．また，配偶者の週間就業時間ダミーについては，0 時間ダミー，15 時間未満ダミー，15 時間以上 40 時間未満ダミー，および 40 時間以上ダミーの 4 つの区分に応じてダミー変数が設定される．なお，本研究では，0 時間ダミーをリファレンス・グループとする．

世帯主と配偶者の学歴ダミーについては，高卒ダミー，短大卒ダミー，大学・大学院卒ダミーを設定した(中卒ダミーがリファレンス・グループ)．

モデル 1a とモデル 1b に性別ダミーが導入されているが，世帯主が妻であるケースが非常に少ないため，係数を評価することが困難であることから，後述するモデル 2 においては，世帯主が夫，配偶者が妻とした場合のモデルを設定している．

ところで，就業時間無回答者の対象範囲としては，(1)回答する世帯員が，就業者の労働時間を把握できない者，(2)労働時間が実際とは異なりうる者，(3)みなし労働時間がある者が考えられる．調査客体は，本来であれば，みなし労働時間が存在すれば，労働時間数を回答することが考えられる．しかし，回答者がみなし労働時間への回答について無知であれば，無回答のままかもしれない．家族においてもその把握は困難と推測される．

さらに，本分析では，モデル 1b の消費関数に加えて消費を決定づける所得の要因をより精密に定量化するために，賃金が労働者の人的資本で定まるとし，その決定要因を労働者と雇用主の属性で表すミンサー型賃金関数(5.3)と(5.4)を設定した上で，賃金率の実現値と期待値の差が，消費支出に及ぼす影響について実証研究を行った[1)]．賃金率を分析に用いるのは，過去の教育訓練などの人的資本で定まる単位労働時間当たりの稼得能力を評価するためであり，この水準が恒常所得を規定して消費に影響をもたらすと考えられるからである．この定式化がモデル 2 であり，以下の式(5.2)〜(5.4)で示される．

$$(5.2)\quad \frac{C_i}{C} = \beta_{20i} + \beta_{2Hi}\ln Y_H + \beta_{2Wi}\frac{Y_W}{Y_H} + \beta_{2Mi}\frac{Y_M}{Y_H} + \alpha_{2HHi}H_H + \alpha_{2HWi}H_W + \alpha_{2EHi}E_H + \alpha_{2EWi}E_W + \boldsymbol{X_2\gamma_{2i}} + \delta_{Hi}(W_H - E(W_H)) + \delta_{Wi}(W_W - E(W_W)) + \varepsilon_{2i}$$

$$(5.3)\quad E(W_H) = \boldsymbol{X_H}\hat{\boldsymbol{\theta}}_{\boldsymbol{H}}$$

$$(5.4)\quad E(W_W) = \boldsymbol{X_W}\hat{\boldsymbol{\theta}}_{\boldsymbol{W}}$$

ここで，各変数とパラメータは以下のものを表している．

$\frac{C_i}{C}$：消費支出に占める第 i 費目の比率
W_H, W_W：世帯主と配偶者それぞれの賃金率
$\boldsymbol{X_2}$：家計の属性(表 1 を参照)
$\boldsymbol{X_H}$：世帯主とその雇用主の属性(年齢，年齢の 2 乗，地域，企業規模等)
$\boldsymbol{X_W}$：配偶者とその雇用主の属性(年齢，年齢の 2 乗，地域，企業規模等)
$\beta_{20i}, \beta_{2Hi}, \beta_{2Wi}, \beta_{2Mi}, \alpha_{2HHi}, \alpha_{2HWi}, \alpha_{2EHi}, \alpha_{2EWi}, \boldsymbol{\gamma_{2i}}, \delta_{Hi}, \delta_{Wi}$：各説明変数の回帰係数
$\hat{\boldsymbol{\theta}}_{\boldsymbol{H}}, \hat{\boldsymbol{\theta}}_{\boldsymbol{W}}$：世帯主と配偶者それぞれの賃金関数の回帰係数
ε_{2i}：誤差項

特に δ_{Hi} および δ_{Wi} は，賃金率の実現値と期待値との差が対象となる詳細費目の消費支出に対する比率に及ぼす影響に関する回帰係数となっている．

式(5.1)〜式(5.4)をもとに，本研究でモデル 2 が設定されている．

式(5.3)と式(5.4)で設定される賃金関数は，賃金が人的資本で規定されるものと考えて，その人的資本に影響を与える要因を，雇用主(産業など)と労働者(年齢など)の属性の関数としてモデル化したものである．雇用主・労働者の属性に対するこの関数の予測値(期待値)が，その人的資本で得られる平均的な賃金といえる．現実に受け取っている賃金は，その期待値と同額ではないから，その差分が消費支出に影響を及ぼす可能性がある．この差分は，人的資本のうち，賃金関数の説明変数として観察された部分と，そうでない部分の差といえ，生涯の予算制約に影響を与えうる情報といえる．

モデル 2 を推定するための具体的な手順は以下のとおりである．

第 1 に，賃金率の期待値を求めるために，賃金構造基本統計調査の個票データを用いて，以下の式(5.5)を推定する．

$$(5.5)\quad \text{賃金率} = f\,(\text{年齢，年齢の 2 乗，学歴ダミー，地域，企業規模，企業区分(官公庁ダミー)})$$

なお，(5.5)式の推計は，賃金関数が大きく異なる，男女別かつ雇用形態(正規，非正規)別に行った[2)]．

第 2 に，全国家計構造調査の個票データをもとに，世帯主および配偶者の勤め先からの年間収入のそれぞれに基づいて賃金率を計測し，これを賃金率の実現値とする[3)]．

第 3 に，つぎの式(5.6)をもとに賃金率を算出した上で，対数変換を行う．

$$(5.6)\quad \text{賃金率} = (\text{きまって支給する現金給与額} \times 12 + \text{特別給与}) / (\text{所定内実労働時間数と超過実労働時間数} \times 12)$$

(5.6)式で賃金構造基本統計調査において算定された賃金率の対数を就業者のみが分析対象となっているミンサー型賃金関数をもとに年齢等で回帰する．

第 4 に，全国家計構造調査の個票データを用いて，式(5.5)と式(5.6)をもとに男女別，雇用形態別(正規 or 非正規)の賃金率の期待値の対数を求めた上で，実数値に変換を行う．この数

値を用いて，世帯主と配偶者の賃金率の実現値と期待値の差を計測する．

なお，妻が無業の場合，本分析上の作業仮説として，実現値と期待値の差の計算において，今働いていない人が働いているとしたらと仮定した場合の(妻の)年間収入の期待値を 0 と設定した[4]．

モデル 2 における仮説としては，実現値と期待値の差が 0 より大きい場合，集団から予想される賃金より今の自分の賃金がより高くなることを示している．それによって，予算制約が緩やかな状態になることから，消費支出を拡大するのが合理的であることが示唆される．

6. 推定結果と解釈

最初に，伊藤・出島 (2022)と伊藤 他 (2023)で定式化されたモデルを調査年次が異なる全国家計構造調査の個票データを用いて推計した結果が，表 3，表 4-1 と表 4-2 で示されている．この中で，表 3 は，モデル 1a を推定して得られた，年間収入の対数が詳細費目の比率に与える影響を示したものである．また，表 4-1 と表 4-2 はそれぞれ，モデル 1a において，年間収入の対数が詳細費目の比率に与える影響を世帯類型別に見たものである[5]．なお，モデル 1a において，分析対象となった全世帯と世帯類型ごとの説明変数に関する基本統計量はそれぞれ，付表 1-1 と付表 1-2 で示されている．

これらの表から明らかにされるのは，年間収入が配偶者の選好に関わる詳細費目に与える影響についてである．家事用消耗品の場合，勤め先収入の年間収入が与える影響については，配偶者が片働きと共働きのいずれにおいても，マイナスに有意になっていて，係数に差が認められる．つぎに，被服については，年間収入が高いほど，消費支出に占める被服の消費の割合が増加する傾向にあることがわかる．さらに，電気代，ガス代や上下水道料を見ると，配偶者の就業形態に関わらず，年間収入が消費支出に及ぼす影響については，マイナスに有意な結果が得られているだけでなく，電気代，ガス代，上下水道料のいずれにおいてもパラメータの絶対値の大きさは，配偶者が正規＞非正規＞無業の順になっていることが確認できる[6]．これについては，在宅時間の影響を表していることが考えられる．なお，2019 年の調査結果において

表 3．年間収入の対数が 10 大費目の詳細費目に与える影響，2019 年，全世帯．

費目	係数	費目	係数
主食的調理食品	-0.0041***	ガソリン	-0.0075***
一般外食	0.0139***	通信	-0.0198***
家賃	-0.0011	補習教育	0.0108***
電気代	-0.0076***	宿泊料	0.0016***
ガス代	-0.0058***	パック旅行	0.0012
上下水道料	-0.0051***	月謝類	0.0057***
家事用消耗品	-0.0019***	理美容サービス	-0.0017
家事サービス	0.0014	こづかい	-0.0048**
被服	0.0088***	交際費	0.0091***
交通	0.0062***	仕送り金	0.0147***
自動車等維持	-0.0135***		

注) ***： 1% 有意，**： 5% 有意，*：10% 有意，
〈 〉はリファレンス・グループである(以下同様)．

表 4-1．年間収入の対数が 10 大費目の詳細費目に与える影響，2019 年，世帯類型別．

	世帯主：正規 配偶者：正規	世帯主：正規 配偶者：非正規	世帯主：正規 配偶者：無業
主食的調理食品	-0.0004	-0.0063***	-0.0054***
一般外食	0.0203***	0.0146***	0.0132***
家賃	0.0078*	-0.0028	-0.0042
電気代	-0.0066***	-0.0100***	-0.0108***
ガス代	-0.0054***	-0.0057***	-0.0074***
上下水道料	-0.0048***	-0.0060***	-0.0071***
家事用消耗品	-0.0022***	-0.0016***	-0.0024***
家事サービス	0.0002	-0.0004	-0.0003
被服	0.0110***	0.0114***	0.0093***
交通	0.0117***	0.0076***	0.0069***

表 4-2．年間収入の対数が詳細費目に与える影響，2019 年，世帯類型別．

	世帯主：正規 配偶者：正規	世帯主：正規 配偶者：非正規	世帯主：正規 配偶者：無業
自動車等維持	-0.0142***	-0.0178***	-0.0140***
ガソリン	-0.0101***	-0.0092***	-0.0069***
通信	-0.0280***	-0.0246***	-0.0211***
補習教育	0.0086***	0.0141***	0.0110***
宿泊料	0.0023	0.0038***	0.0017
パック旅行	0.0012	0.0001	0.0004
月謝類	0.0040	0.0096***	0.0084***
理美容サービス	0.0020*	0.0326***	0.0030***
こづかい	-0.0106***	-0.0096***	-0.0024
交際費	0.0168***	0.0106***	0.0026
仕送り金	0.0140***	0.0178***	0.0197***

は，年間収入が増大した場合，主食的調理食品費が消費支出に占める比率については，配偶者が正規の場合を除けば，マイナスで有意な結果が出ている．このことから，配偶者が非正規と無業の場合には，エンゲルの法則が実証的に示されていることがわかる．なお，配偶者が正規の場合，所得が高くなっている傾向にあることから，在宅時間が短いことによるマイナスの影響が，配偶者の所得の増加に伴い，予算制約が比較的大きく拡張された影響によって相殺された可能性が指摘できる．

その他の消費支出について見ると，こづかいに関しては，配偶者が就業している場合に，年間収入が増加するにつれて，消費支出に対する比率が減少する傾向にあることが確認できる．また，仕送り額に関しては，年間収入が増えると全般的に増大する傾向にある．なお，補習教

表 5．配偶者の勤め先収入の年間収入に対する比率が詳細費目に与える影響，2019 年．

費目	係数	費目	係数
主食的調理食品	0.0030	ガソリン	0.0092***
一般外食	-0.0024	通信	0.0216***
家賃	-0.0015	補習教育	-0.0024
電気代	0.0009	宿泊料	-0.0040**
ガス代	-0.0000	パック旅行	-0.0005
上下水道料	-0.0031	月謝類	-0.0016
家事用消耗品	-0.0043***	理美容サービス	-0.0017
家事サービス	0.0014	こづかい	-0.0031
被服	0.0002	交際費	-0.0057
交通	-0.0099**	仕送り金	0.0023
自動車等維持	0.0209***		

表 6-1．配偶者の勤め先収入の年間収入に対する比率が詳細費目に与える影響，2019 年，世帯類型別．

	世帯主：正規 配偶者：正規	世帯主：正規 配偶者：非正規	世帯主：正規 配偶者：無業
主食的調理食品	0.0041	-0.0037	0.0067
一般外食	-0.0176**	0.0031	0.0195
家賃	-0.0048	0.0076	-0.0073
電気代	0.0061	-0.0093*	-0.0005
ガス代	0.0005	0.0010	-0.0029
上下水道料	-0.0019	-0.0042	-0.0131
家事用消耗品	-0.0025*	-0.0066***	-0.0053
家事サービス	-0.0002	0.0019	0.0050
被服	-0.0104	0.0052	0.0266*
交通	-0.0203***	0.0027**	- 0.0112

育費については，共働き世帯，片働き世帯のいずれについても，年間収入が高くなると，プラスに有意な傾向が見られる．

つぎに表 5 は，モデル 1a から推定された，配偶者の勤め先収入の年間収入に対する比率が詳細費目に与える影響を示したものである．また，表 6-1 と表 6-2 はそれぞれ，モデル 1a において，配偶者の勤め先収入の年間収入に対する比率が詳細費目に与える影響を世帯類型別に示したものである．これらから明らかになるのは，配偶者の年間収入の比率が配偶者の選好に関わる詳細費目に与える影響である．一般外食については，配偶者が正規である場合にのみ，パラメータがマイナスで有意となっている．これについては，主食的調理食品には弁当等が含ま

表 6-2. 配偶者の勤め先収入の年間収入に対する比率が詳細費目に与える影響，2019 年，世帯類型別．

	世帯主：正規 配偶者：正規	世帯主：正規 配偶者：非正規	世帯主：正規 配偶者：無業
自動車等維持	0.0158	0.0273**	0.0397
ガソリン	0.0108**	0.0063	0.0318***
通信	0.0350***	0.0090	0.0011
補習教育	-0.0010	-0.0045	0.0043
宿泊料	-0.0113***	0.0071**	-0.0120
パック旅行	-0.0028	0.0049	-0.0109
月謝類	0.0045	-0.0212***	-0.0190
理美容サービス	-0.0005	-0.0026	0.0025
こづかい	0.0058	-0.0262**	0.0270
交際費	0.0004	-0.0207**	0.0021
仕送り金	-0.0095*	0.0108	0.0449*

れており，正規雇用者が勤務日に時間を取ることができない場合，外食への支出を弁当等の調理済みの食品の購入で代替している可能性がある．すなわち，隠れた需要要因として勤務中の昼食が存在し，それが，一般外食から主食的調理食品に代替されていることが考えられる．なお，一般外食と主食的調理食品の区分について今後より精密な再検討が求められる．

家事用消耗品の場合，年間収入に占める配偶者の勤め先収入の比率をみると，配偶者が正規就業かパート・アルバイト就業において，家事用消耗品に対してマイナスに有意の結果が得られている．このことから，夫婦共働きの場合，片働き世帯の場合と比較して，家事時間(家計内生産時間)が短縮していることが見て取れ，それゆえに，家事用消耗品への購入が減少している可能性がある．さらに，被服については，配偶者が無業である場合，配偶者の勤め先収入の年間収入に対する比率に関しては，プラスに有意な結果がそれぞれ導かれている．

その他の消費支出については，こづかいに関しては，配偶者が非正規就業の場合には，年間収入に占める配偶者の勤め先収入の比率がマイナスに有意になっている．共働き世帯において各世帯員の効用最大化が図られていることから，伊藤・出島 (2022)で議論されている，世帯員による個別の意思決定が家計の交渉力によって反映される Collective Model と整合的な結果が得られていることを指摘することができる．仕送り額に関しては，年間収入に占める配偶者の勤め先収入の比率については，配偶者が正規就業者の場合マイナス，無業の場合プラスにそれぞれ有意な結果が得られている．仕送り額のような子供の経費に対する支出は，配偶者にとっての効用が相対的に高くなる費目であることを示唆している．

さらに，電気代，ガス代や上下水道料を見ると，年間収入に占める配偶者の勤め先収入の比率については，配偶者が非正規の場合に電気代でマイナスに有意な結果が得られている．共働き世帯において，年間収入に占める配偶者の勤め先収入の比率が高いほど，配偶者における在宅時間が短くなる傾向にあることが考えられ，その結果として，消費支出に占める電気代の割合が低下する傾向にあることがわかる．なお，このような結果が得られた背景としては，正規であると労働時間が一定であるだけでなく，産業や職業によって時間当たりの賃金率は多様になっていることが指摘される．そのため，在宅時間と配偶者の勤め先収入の間に相関はないことから，電気代のパラメータも有意ではないが，非正規の場合には労働時間が長いほど勤め先

表 7-1. 世帯主と配偶者の労働時間が詳細費目に与える影響，全世帯と世帯類型別.

費目	週間就業時間ダミー	全世帯	夫婦共働き	夫婦片働き
被服	世帯主が40時間以上	-0.0003	0.0010	-0.0025
	配偶者が１時間以上15時間未満	-0.0058*	-0.0076**	—
	配偶者が１5時間以上40時間未満	-0.0039	-0.0053*	—
	配偶者が40時間以上	-0.0054*	-0.0057*	—
通信	世帯主が40時間以上	-0.0036*	-0.0042*	-0.0034
	配偶者が１時間以上15時間未満	0.0003	-0.0005	—
	配偶者が１5時間以上40時間未満	0.0008	-0.0002	—
	配偶者が40時間以上	0.0088**	0.0065	—

表 7-2. 世帯主と配偶者の労働時間が詳細費目に与える影響，全世帯と世帯類型別.

費目	週間就業時間ダミー	全世帯	夫婦共働き	夫婦片働き
宿泊	世帯主が40時間以上	-0.0003	-0.0000	-0.0009
	配偶者が１時間以上15時間未満	-0.0033**	-0.0042***	—
	配偶者が１5時間以上40時間未満	-0.0026**	-0.0034***	—
	配偶者が40時間以上	-0.0029**	-0.0032**	—
交際費	世帯主が40時間以上	-0.0026	-0.0022	-0.0030
	配偶者が１時間以上15時間未満	-0.0063*	-0.0054	—
	配偶者が１5時間以上40時間未満	-0.0104***	-0.0094***	—
	配偶者が40時間以上	-0.0102***	-0.0094***	—

収入が大きく，さらにその場合は在宅時間が短くなるためにパラメータがマイナスで有意になっている．ガソリン代と通信費に関しては，配偶者が正規就業の場合にはそれぞれ，配偶者の勤め先収入の年間収入に対する比率はプラスに有意な結果が得られている．なお，配偶者が無業者である場合に回帰係数がマイナスで有意になっていることについては，配偶者が調査月以外において就業していて勤め先収入を得ていることが指摘できる．

表 7-1 と表 7-2 はそれぞれ，モデル 1b から推定された，全世帯と世帯類型別に世帯主と配偶者の労働時間ダミーが詳細費目に与える影響を示したものである．具体的には表 7-1 と表 7-2 は，被服，通信，宿泊と交際費に与える労働時間の影響を表示している．また，表 7-3 と表 7-4 はそれぞれ，被服と交際費に関する推定結果を示している．労働時間が把握されることで，こ

表 7-3. 被服に関する推定結果，全世帯と世帯類型別.

世帯類型	全世帯			夫婦共働き			夫婦片働き		
説明変数	係数	標準誤差	有意性	係数	標準誤差	有意性	係数	標準誤差	有意性
配偶者の正規就業ダミー	0.0049	0.0029	*						
配偶者のアルバイト・パート就業ダミー	0.0025	0.0029							
配偶者の派遣社員ダミー	0.0081	0.0046	*						
配偶者のその他就業ダミー	0.0011	0.0041							
持ち家の有無ダミー	0.0014	0.0028		0.0021	0.0035		-0.0017	0.0050	
性別ダミー	-0.0032	0.0046		-0.0010	0.0058		-0.0153	0.0132	
世帯主の年齢	-0.0010	0.0006		-0.0010	0.0008		-0.0010	0.0011	
世帯主の年齢の2乗	0.0000	0.0000		0.0000	0.0000		0.0000	0.0000	
年間収入(対数)	0.0088	0.0014	***	0.0118	0.0019	***	0.0094	0.0025	***
配偶者の勤め先収入の年間収入に対する比率	0.0039	0.0049		0.0023	0.0048		0.0283	0.0167	*
65歳未満の世帯員における年間収入の比率	0.0178	0.0293		-0.0108	0.0496		0.0093	0.0463	
財産収入(対数)	-0.0003	0.0003		0.0002	0.0004		-0.0010	0.0005	
貯蓄現在高(対数)	0.0000	0.0003		-0.0002	0.0003		0.0002	0.0005	
負債現在高(対数)	-0.0001	0.0002		-0.0001	0.0002		-0.0001	0.0003	
住宅・宅地資産額(対数)	0.0006	0.0004		0.0005	0.0005		0.0009	0.0007	
大都市圏ダミー	0.0016	0.0010	*	0.0012	0.0012		0.0025	0.0017	
末子の年齢<0～5歳>									
6～12歳	-0.0027	0.0013	**	-0.0036	0.0016	**	-0.0005	0.0023	
13～15歳	-0.0064	0.0018	***	-0.0067	0.0021	***	-0.0065	0.0036	*
16～18歳	-0.0074	0.0021	***	-0.0075	0.0025	***	-0.0092	0.0042	**
19歳以上	-0.0078	0.0028	***	-0.0091	0.0036	**	-0.0025	0.0050	
世帯主の週間就業時間ダミー	-0.0003	0.0015		0.0010	0.0019		-0.0025	0.0027	
配偶者の週間就業時間ダミー〈0時間ダミー〉									
1時間以上15時間未満ダミー	-0.0058	0.0032	*	-0.0076	0.0031	**			
15時間以上40時間未満ダミー	-0.0039	0.0028		-0.0053	0.0027	*			
40時間以上ダミー	-0.0054	0.0029	*	-0.0057	0.0030	*			
世帯主の学歴〈中卒ダミー〉									
高卒	-0.0009	0.0033		-0.0017	0.0040		0.0080	0.0073	
短大卒	0.0011	0.0035		-0.0017	0.0041		0.0151	0.0076	
大卒・大学院卒	0.0009	0.0034		0.0007	0.0040		0.0074	0.0074	
配偶者の学歴〈中卒ダミー〉									
高卒	-0.0007	0.0039		-0.0037	0.0057		0.0006	0.0060	
短大卒	-0.0011	0.0039		-0.0040	0.0057		-0.0006	0.0061	
大卒・大学院卒	-0.0008	0.0040		-0.0043	0.0058		0.0016	0.0062	
定数	0.0096	0.0157		-0.0031	0.0212		0.0105	0.0288	
Adj. R^2	0.032			0.033			0.035		
F	7.324			6.052			3.720		
N	5684			3832			1721		

のような推定が可能になっている.

本研究から明らかになった知見は，労働時間が消費支出に与える影響である．被服，宿泊と交際費については，配偶者が就業すると，消費支出に対する比率が有意に減少する傾向にある.

表 7-4. 交際費に関する推定結果，全世帯と世帯類型別．

世帯類型	全世帯			夫婦共働き			夫婦片働き		
説明変数	係数	標準誤差	有意性	係数	標準誤差	有意性	係数	標準誤差	有意性
配偶者の正規就業ダミー	0.0087	0.0034	**						
配偶者のアルバイト・パート就業ダミー	0.0080	0.0034	**						
配偶者の派遣社員ダミー	0.0210	0.0054	***						
配偶者のその他就業ダミー	0.0067	0.0049							
持ち家の有無ダミー	0.0046	0.0033		0.0017	0.0040		0.0116	0.0062	*
性別ダミー	0.0005	0.0054		0.0051	0.0066		0.0018	0.0165	
世帯主の年齢	-0.0021	0.0007	***	-0.0031	0.0010	***	-0.0004	0.0014	
世帯主の年齢の2乗	0.0000	0.0000	***	0.0000	0.0000	***	0.0000	0.0000	
年間収入(対数)	0.0082	0.0017	***	0.0143	0.0022	***	-0.0005	0.0032	
配偶者の勤め先収入の年間収入に対する比率	0.0048	0.0057		0.0023	0.0055		0.0160	0.0209	
65歳未満の世帯員における年間収入の比率	0.0260	0.0345		0.0143	0.0568		-0.0127	0.0580	
財産収入(対数)	0.0004	0.0004		0.0002	0.0005		0.0006	0.0007	
貯蓄現在高(対数)	0.0009	0.0003	***	0.0008	0.0004	**	0.0009	0.0006	
負債現在高(対数)	-0.0001	0.0002		-0.0001	0.0002		0.0000	0.0004	
住宅・宅地資産額(対数)	0.0006	0.0005		0.0008	0.0005		0.0000	0.0009	
大都市圏ダミー	0.0015	0.0011		0.0015	0.0014		0.0017	0.0021	
末子の年齢<0～5歳>									
6～12歳	-0.0031	0.0015	**	-0.0050	0.0018	***	0.0019	0.0029	
13～15歳	-0.0103	0.0021	***	-0.0110	0.0024	***	-0.0101	0.0046	**
16～18歳	-0.0088	0.0025	***	-0.0109	0.0029	***	-0.0045	0.0053	
19歳以上	-0.0036	0.0033		-0.0010	0.0041		-0.0065	0.0063	
世帯主の週間就業時間ダミー	-0.0026	0.0018		-0.0023	0.0022		-0.0030	0.0033	
配偶者の週間就業時間ダミー〈0時間ダミー〉									
1時間以上15時間未満ダミー	-0.0063	0.0038	*	-0.0054	0.0036				
15時間以上40時間未満ダミー	-0.0104	0.0033	***	-0.0094	0.0031	***			
40時間以上ダミー	-0.0102	0.0034	***	-0.0094	0.0034	***			
世帯主の学歴〈中卒ダミー〉									
高卒	0.0047	0.0039		0.0059	0.0046		0.0047	0.0039	
短大卒	0.0057	0.0041		0.0070	0.0047		0.0057	0.0041	
大卒・大学院卒	0.0079	0.0040	**	0.0078	0.0046	*	0.0079	0.0040	
配偶者の学歴〈中卒ダミー〉									
高卒	0.0030	0.0046		-0.0022	0.0065		0.0076	0.0075	
短大卒	0.0024	0.0046		-0.0025	0.0065		0.0059	0.0076	
大卒・大学院卒	0.0061	0.0047		-0.0002	0.0066		0.0124	0.0077	
定数	-0.0030	0.0185	**	-0.0089	0.0243		0.0079	0.0360	
Adj. R^2	0.039			0.046			0.025		
F	8.661			8.136			2.942		
N	5684			3832			1721		

ただし，消費支出が最も減少するのは，労働時間が短い部分にも認められる．配偶者の就業に伴って，買い物，旅行，友達との交際といった私的な行動に対する時間が制約されることが考えられる．一方，通信費については，上記と異なる傾向が観察される．配偶者の労働時間が増

表 8-1．世帯主の勤め先からの賃金率の実現値と期待値の差が詳細費目に与える影響，2019 年．

	係数		係数
主食的調理食品	0.0136	ガソリン	0.0150
一般外食	-0.1702	通信	0.1536
家賃	1.2098***	補習教育	0.0744
電気代	0.1989**	宿泊料	0.1418
ガス代	0.0489	パック旅行	-0.3829
上下水道料	0.0057	月謝額	-0.1420
家事用消耗品	0.0744*	理美容サービス	0.0194
家事サービス	0.1056**	こづかい	0.5171*
被服	0.4863**	交際費	-0.2536
交通	0.2099	仕送り金	0.7840
自動車等維持	-0.0095		

表 8-2．世帯主の配偶者の勤め先からの賃金率の実現値と期待値の差が詳細費目に与える影響，2019 年．

	係数		係数
主食的調理食品	-0.0838	ガソリン	-0.0766
一般外食	0.6095	通信	-1.0447**
家賃	-0.6429	補習教育	-0.9168
電気代	-0.0445	宿泊料	0.0081
ガス代	0.2082	パック旅行	-0.0322
上下水道料	-0.0395	月謝額	-0.3274
家事用消耗品	-0.1027	理美容サービス	0.3550*
家事サービス	0.1421	こづかい	-1.1045
被服	-0.0774	交際費	1.4803***
交通	-0.4120	仕送り金	-1.6923
自動車等維持	0.1664		

大した場合，夫婦の通信需要が増大することが想像される．また，仕事に関する通信費の自己負担分が含まれる可能性があることから，有意に増大する傾向にあることが確認できる．

表 8-1 は，モデル 2 から推定された，全世帯を対象に世帯主の勤め先における賃金率の実現値と期待値の差が 10 大費目の詳細費目に与える影響を全世帯で見たものである．また，表 8-2 は，モデル 2 において，世帯主の配偶者の勤め先における賃金率の実現値と期待値の差が 10 大費目の詳細費目に与える影響を示したものである．

本分析の結果から，表 8-1 と表 8-2 のいずれにおいても，プラスに有意な結果が出ている費目があることが確認できる．世帯主の場合，想定通り，実現値と期待値の差分が消費支出に対してプラスに有意になっている項目が存在する．表 8-1 では，家賃，電気代，家事消耗品・家

事サービス，被服，こづかいについてプラスに有意な値が導かれている．特に，家賃については，長期にわたって同じ賃貸住宅に住むことを想定しているのが通常であることから，恒常所得と最も正確に対応する消費となる．本分析結果から，自分の賃金が観察される属性よりも高くなっている場合，将来の予算制約がより拡張的であることから，消費支出を増加させることが合理的であるものと解釈される．

なお，本分析では以下のような留保条件を勘案する必要がある．それは，全国家計構造調査における標本設計の変化についてである．先述のとおり，2014 年までの全国消費実態調査と 2019 年の全国家計構造調査では，標本設計が異なることが指摘できる．そのため，全国家計構造調査では家計調査で援用された世帯の消費構造が，2014 年全国消費実態調査の調査対象世帯の消費構造と異なる可能性がある．これは，全国消費実態調査における調査客体と比較した場合，全国家計構造調査では相対的に少ない標本が本分析の対象となっていることが原因である．したがって，全国家計構造調査においては，家計調査の対象世帯における家計簿情報と全国家計構造調査の対象世帯における家計簿情報については，データ特性上の差異が生じる可能性があることに留意する必要がある．

7. おわりに

本稿では，全国家計構造調査と賃金構造基本統計調査の個票データを併用した上で，労働時間を考慮した場合に，世帯類型，さらには世帯主と配偶者の働き方の差に応じて，世帯主と配偶者の所得の構成が，双方の選好に関わる詳細費目の支出に及ぼす影響に関する実証分析を行った．

本分析の結果，第 1 に先行研究と同様に，夫婦の働き方の違いが家事消耗品や宿泊費等の詳細費目の支出に影響を与えることが観察された．第 2 に，被服費，宿泊費，交際費のような一部の詳細費目に対して，配偶者の労働時間が影響を与えていることが確認された．本研究の結果から，共働き世帯で配偶者の労働時間が長くなるにしたがって，配偶者の消費に関する交渉力の増加と時間節約型の消費傾向が強まっていることと解釈される．

注．

1) 日本における公的統計を用いたミンサー型賃金関数の推定については，川口 (2011)が網羅的な検討をしている．

2) 全国家計構造実態調査には学歴に関する変数はあるものの，産業や勤続年数に関する変数がないことから，本研究の賃金関数にはこれらの変数を説明変数に設定することができない．このことが，賃金率の期待値に関する留保条件となる．

3) 伊藤・出島 (2016)は，就業構造基本調査と賃金構造基本統計調査を併用した上で，賃金の期待値と実現値の差が，転職意識に及ぼす影響を計測している．

4) 就業履歴が利用可能であれば，本人の人的資本と就労した場合の賃金の期待値をより正確に算出することが可能になる．

5) 本稿の目的は，注目している回帰係数の t 検定による有意性を確認するものである．ただし，一部の費目に関しては，モデルの適合度についての F 検定が，有意でない結果も存在している．本稿では，F 検定が有意である部分について説明と解釈を行っている．

6) 配偶者が無業である世帯において，全国家計構造調査の調査対象ではない期間で勤め先収入が発生する場合には，配偶者の年間収入は 0 にならないことに留意されたい．

謝　辞

本稿は，令和 5 年度日本学術振興会科学研究費補助金「外国人労働力を考慮した雇用創出と労働移動の計量経済分析」（基盤研究(C)）（研究代表者：中央大学 伊藤伸介）における研究成果の一部を発表するものである．また，本研究において使用した「全国消費実態調査」および「賃金構造基本統計調査」の調査票情報は，統計法第 33 条に基づき提供を受けたものであり，本報告で作成した集計表等は提供を受けた調査票情報を独自集計したものである．記して関係各位に御礼申し上げたい．

付　録

付表 1-1．モデル 1a に関する基本統計量，全世帯．

説明変数	度数	平均	標準偏差
配偶者の正規就業ダミー	6342	0.2711	0.4445
配偶者のアルバイト・パート就業ダミー	6342	0.3926	0.4884
配偶者の派遣社員ダミー	6342	0.0143	0.1189
配偶者のその他就業ダミー	6342	0.0218	0.1459
持ち家の有無ダミー	6342	0.7617	0.4260
性別ダミー	6342	0.9890	0.1045
世帯主の年齢	6342	43	7
世帯主の年齢の2乗	6342	1865	621
勤め先からの年間収入(万円)	6342	735	317
配偶者の勤め先収入の年間収入に対する比率	6342	0.1725	0.1707
65歳未満の世帯員における年間収入の比率	6342	0.0029	0.0160
財産収入(円)	6342	320	4104
貯蓄現在高(万円)	6342	868	1128
負債現在高(万円)	6342	1094	1214
住宅・宅地資産額(万円)	6342	1678	1747
大都市圏ダミー	6342	0.4475	0.4973
6～12歳	6342	0.2884	0.4531
13～15歳	6342	0.1175	0.3220
16～18歳	6342	0.0829	0.2758
19歳以上	6342	0.0430	0.2030

付表 1-2．モデル 1a に関する基本統計量，世帯類型別．

	世帯主正規・配偶者正規			世帯主正規・配偶者非正規			世帯主正規・配偶者無業		
説明変数	度数	平均	標準偏差	度数	平均	標準偏差	度数	平均	標準偏差
持ち家の有無ダミー	1686	0.7924	0.4057	2656	0.8020	0.3986	1857	0.7017	0.4576
性別ダミー	1686	0.9781	0.1465	2656	0.9970	0.0548	1857	0.9952	0.0695
世帯主の年齢	1686	42	7	2656	43	7	1857	42	8
世帯主の年齢の2乗	1686	1809	619	2656	1927	584	1857	1826	655
勤め先からの年間収入(万円)	1686	907	339	2656	735	317	1857	663	301
配偶者の勤め先収入の年間収入に対する比率	1686	0.3673	0.1424	2656	0.1567	0.0967	1857	0.0128	0.0507
65歳未満の世帯員における年間収入の比率	1686	0.0018	0.0075	2656	0.0031	0.0143	1857	0.0029	0.0174
財産収入(円)	1686	267	2982	2656	270	3124	1857	460	5952
貯蓄現在高(万円)	1686	1023	1316	2656	753	950	1857	926	1171
負債現在高(万円)	1686	1255	1318	2656	1080	1135	1857	1024	1225
住宅・宅地資産額(万円)	1686	1765	1758	2656	1624	1602	1857	1750	1935
大都市圏ダミー	1686	0.3541	0.4784	2656	0.4492	0.4975	1857	0.5342	0.4990
6～12歳	1686	0.2604	0.4390	2656	0.3438	0.4750	1857	0.2396	0.4270
13～15歳	1686	0.1079	0.3104	2656	0.1468	0.3540	1857	0.0824	0.2750
16～18歳	1686	0.0801	0.2715	2656	0.1005	0.3008	1857	0.0587	0.2351
19歳以上	1686	0.0380	0.1912	2656	0.0414	0.1993	1857	0.0485	0.2148

参 考 文 献

阿部修人 (2011).『家計消費の経済分析』, 一橋大学経済研究叢書, Vol.59, 岩波書店, 東京.

阿部修人, 稲倉典子, 小原美紀 (2021). 家計内サービス生産関数及び時間制約に関する一考察, RCESR Discussion Paper Series, No.DP21-3, 1–33, 一橋大学経済研究所経済社会リスク研究機構, 東京.

Hayashi, F. (1995). Is the japanese extended family altruistically linked? A test based on engel curves, *Journal of Political Economy*, **103**(3), 661–674.

伊藤伸介, 出島敬久 (2016). 賃金の期待値と実現値の差が転職希望に与える影響—賃金構造基本統計調査と就業構造基本調査の個票データを併用した計量分析—, 上智経済論集, **61**(1・2 号合併号), 19–57.

Ito, S. and Dejima, T. (2020). The relationship between household type and consumption patterns in Japan — Evidence from Japan's National Survey of Family Income and Expenditure, Discussion Paper Series A, No.712, 1–32, 一橋大学経済研究所, 東京.

伊藤伸介, 出島敬久 (2022). 世帯類型の違いが消費構造に与える影響の計量分析—公的統計ミクロデータを用いて—, 経済分析, **205**, 27–52.

伊藤伸介, 出島敬久, 村田磨理子 (2023). 世帯主と配偶者の所得の構成が双方の選好に関わる費目の消費支出に与える影響, 経済学論纂, **63**(5・6 合併号), 245–273.

川口大司 (2011). ミンサー型賃金関数の日本の労働市場への適用,『現代経済学の潮流 2011』(阿部顕三, 大垣昌夫, 小川一夫, 田淵隆俊 編), 67–98, 東洋経済新報社, 東京.

小原美紀 (2008). 家計内交渉と家計の消費変動,『世帯内分配と世代間移転の経済分析』(チャールズ・ユウジ・ホリオカ, (財)家計経済研究所 編), 48–72, ミネルヴァ書房, 京都.

坂本和靖 (2008). 世帯内における消費・余暇配分の構造,『世帯内分配と世代間移転の経済分析』(チャールズ・ユウジ・ホリオカ, (財)家計経済研究所 編), 21–47, ミネルヴァ書房, 京都.

澤田康幸 (2008). 日本における世帯内リスクシェアリングの分析,『世帯内分配と世代間移転の経済分析』(チャールズ・ユウジ・ホリオカ, (財)家計経済研究所 編), 73–90, ミネルヴァ書房, 京都.

Econometric Analysis of Consumption Patterns and Working Hours in Japan: Based on National Survey of Family Income Consumption and Wealth

Shinsuke Ito[1], Takahisa Dejima[2] and Mariko Murata[3]

[1]Faculty of Economics, Chuo University
[2]Faculty of Economics, Sophia University
[3]Statistical Information Institute for Consulting and Analysis

The share of dual-earner households is increasing in developed countries, including Japan. With the accompanying changes in income composition and time allocation, examining their effects on consumption patterns and expenditures is worthwhile because doing so will enable the study of the effects of various economic policies. This paper examines the relationship between income composition and consumption patterns using individual data from the National Survey of Family Income Consumption and Wealth, to which weekly working hours have been newly added as a survey item.

Specifically, after setting up a consumption function for the detailed expense items of household consumption that have been estimated in previous studies, we newly introduced working hours as an explanatory variable. To more accurately account for the effect of permanent income on consumption, we also introduced as an explanatory variable the difference between expected and realized wages based on the wage function estimated by the Basic Survey on Wage Structure.

Consistent with previous studies, we observed that differences in the way couples work affect their spending on household consumables, lodging, and other items. In addition, we confirmed that working hours have an effect, albeit partial, on consumption expenditures. These results can be explained by the tendency among dual-earner households to increase time-saving consumption as the working hours of spouses of households increase.

Key words: Household consumption, consumption function, dual-earner households, working hour.

統計数理(2024)
第 72 巻 第 2 号 217–231
©2024 統計数理研究所
特集「公的統計―社会の重要な情報基盤―」
[原著論文]

事業所・企業系の公的統計を対象にした合成データの生成技法に関する検討

―経済センサスを例として―

伊藤 伸介[1]・横溝 秀始[2]

(受付 2023 年 12 月 31 日;改訂 2024 年 6 月 10 日;採択 7 月 2 日)

要　　旨

わが国の事業所・企業系の統計調査においては,事業所や企業を対象にした匿名データが現在作成されていないだけでなく,事業所・企業系の統計調査の場合,一般公開型ミクロデータの作成が困難である.そのため,合成データが作成可能であれば,テストデータ等へのニーズに応えることが可能になる.そこで,本稿では,経済センサス活動調査の個票データを用いて,各種の合成データの生成技法について定量的な評価を行った.

本研究では,攪乱的手法であるミクロアグリゲーションの MDAV(=Maximum Distance to Average Vector)法,CART(=Classification And Regression Tree)等,さらには深層学習モデルの 1 つである CTGAN(=Conditional Tabular GAN)も用いて生成された合成データの有用性および秘匿性について定量的な評価を行った.本研究において CART を用いて合成データを生成した場合,要約統計量や相関係数といった分布特性が再現可能であることが確認された.また,CART は MDAV 法と比較して,有用性を保ったまま秘匿性の強度が高まる可能性がある.さらに,CTGAN については,CART と比較した場合,秘匿性の程度がより高くなっていることがわかったが,有用性の低下も相対的に大きいことが確認された.

キーワード:合成データ,ミクロアグリゲーション,CART,CTGAN,経済センサス.

1. はじめに―事業所・企業系のミクロデータの現状

わが国の公的統計においては,現在 7 種類の世帯・人口系の統計調査が匿名データとして提供されている.それに対して,事業所・企業系の統計調査に関しては,常用労働者を対象に匿名データが作成・提供されている賃金構造基本統計調査の事例はあるものの,調査客体である事業所や企業に匿名化技法を適用することによって作成された匿名データは,わが国では存在しない.また,全国消費実態調査と就業構造基本調査については,一般用ミクロデータが公開されているが,事業所・企業系の統計調査は現時点では作成の対象外となっている.このように,わが国では,事業所・企業系の統計調査においては,個票データ(調査票情報)のみの利用が可能になっている.

それに対して,教育目的のためのミクロデータの利用に対するニーズが高まっていることや,

[1] 中央大学 経済学部:〒 192–0393 東京都八王子市東中野 742–1
[2] 総務省 統計研究研修所:〒 162–8668 東京都新宿区若松町 19–1

オンサイト施設やリモートアクセスのようなセキュアな環境における個票データの利用に関して，プログラムコードのチェックを行うためのテストデータへのニーズが少なくないこともあり，海外では近年合成データ(synthetic data)に対する社会的関心が高まっている．なお，合成データとは，元になるデータからその分布特性が近似するように属性値を新たに生成することによって作成され，個体情報の秘匿性が確保されたミクロレベルの擬似的なデータ（Templ, 2017, p.157)と定義される．

事業所・企業系の統計調査の場合，一般公開型ミクロデータ(public use microdata)の作成が困難であることから，合成データを作成することができれば，テストデータ等へのニーズに応えることが可能になるように思われる．そこで，本稿は，経済センサスの個票データを用いて，事業所・企業系の統計調査の合成データの生成技法を定量的に評価する．なお，本研究においては，伝統的な匿名化技法としての攪乱的手法を用いて作成された匿名化ミクロデータと機械学習の方法論等を用いて生成した各種の合成データを比較・検証する．そのために，ミクロデータに対する秘匿性や有用性に関する評価指標に基づいて，合成データの定量的な評価を行う．

2. 公的統計における合成データ作成の現状

本節では，最初に海外の公的統計に関する合成データ生成の研究事例を紹介するだけでなく，公的統計における合成データの生成技法の比較検証に関する研究についても概括する．

2.1 公的統計における合成データ生成に関する海外の現状

海外における合成データの作成状況あるいは合成データの作成に関する研究事例については，例えば以下の事例を指摘することができる．第 1 は，Eurostat の事例である．Eurostat では，EU-SILC(=European Union Statistics on Income and Living Conditions)に対して，統計的なモデルを用いたシミュレーションによる合成データの方法論を用いた一般公開型ファイル(Public Use File)が作成されている（伊藤, 2018)．

第 2 は，エディンバラ大学による事例である．エディンバラ大学では，スコットランド縦断調査(Scottish Longitudinal Study=SLS)を対象に，合成データの作成が行われてきた．R の合成データ作成用のパッケージである synthpop（Nowok et al., 2016)を用いることによって，人口センサスの縦断的な合成データが生成されている．

第 3 は，イギリス国家統計局(Office for National Statistics=ONS)による事例である．ONS では，労働力調査(Labour Force Survey)のミクロデータをもとに，synthpop を含む合成データ作成用の複数のパッケージを比較・検討した上で，統計実務の観点から合成データの作成可能性を追究している（Bates et al., 2019)．

第 4 は，オーストラリアによる事業所・企業用の合成データの研究の事例である（Chien et al., 2021)．産業によって市場が寡占あるいは複占という特徴を有するため，事業所・企業系の統計調査においては，情報の削減や攪乱といった従来の匿名化手法が世帯・人口系ほど有効でない可能性がある．そこで，攪乱的手法との比較の結果，秘匿性と有用性を勘案した場合に，合成データの方法論が事業所・企業系のミクロデータに適用可能な手法として議論されている．

UNECE（2022)では，海外の統計作成部局における合成データの作成に関する最近の研究の事例が紹介されている．例えば，ニュージーランド統計局は，数理モデルによって生成された合成データを Synthetic Unit Record Files(SURFs)として展開している．また，2007 年所得調査に基づく SURFs および，2019 年の世帯貯蓄調査と人口センサスに基づき生成された ‘Census for School’ SURF が提供されている．つぎに，カナダ統計局は，センサスベースの合成

表 1．合成データのユースケースごとの推奨技法．UNECE（2022），表 7 をもとに作成．

手法の分類	手法	例	PUF / 分析テスト用	教育用	プログラムテスト用	備考
Sequential modelling	Fully Conditional Specification (FCS)	CART, Bagging, Random forest	○推奨	△可（作成コスト大）	△可（用途に対して高度すぎる）	元データ属性間の関係性が保持される
	Information Preserving Statistical Obfuscation (IPSO)	回帰	△全属性が線形回帰で表現できるなら推奨、以外は非推奨	△全属性が線形回帰で表現できるなら推奨、以外は非推奨	△可（用途に対して高度すぎる）	特に線形回帰に関連する結果と統計量が保持される
Simulated Data	Dummy files	-	×非推奨	△（分析を目的としないなら）	○推奨	分析には適さないが簡単かつ迅速
	Analytically advanced simulated data	一般用ミクロデータ	△想定する分析手法に対しては推奨、以外は非推奨	△想定する分析手法に対しては推奨、以外は非推奨	△可（用途に対して高度すぎる）	限られた統計量のみ保持される
	Pseudo Likelihood	-	△元の有限母集団から推定したい場合は推奨、合成ウェイトを期待する場合は非推奨	△可（作成コスト大）	△可（用途に対して高度すぎる）	元データ属性間の関係性が保持される
Deep Learning	Generative Adversarial Network (GAN)	CTGAN, Table GAN	△テキストまたは非構造化データが存在する場合は推奨	△可（作成コスト大）	△可（用途に対して高度すぎる）	非構造化データとテキストデータを処理する唯一の手法

データの生成に関する研究を行っている．具体的には，カナダの退職所得システム（Canadian retirement and income system）に関する動態的なミクロシミュレーションモデルの初期母集団のための研究が展開されている．オーストラリア統計局では，機械学習モデルを検証するためにミクロシミュレーションモデルに基づく合成データの生成が紹介されている．さらに，イギリス国家統計局では，2017 年に ONS 内部に設立された ONS Data Science Campus において，敵対的生成ネットワーク（Generative Adversarial Network=GAN）（Goodfellow et al., 2014）を用いた合成データを生成することが追究されている．

2.2　公的統計における合成データの生成技法の比較検証に関する研究

本節では，公的統計を対象にした合成データの生成技法の比較検証の事例を紹介したい．

UNECE（2022）では，合成データ生成技法やユースケースに応じた推奨技法が，体系的に整理されている（表 1）．合成データ生成技法は，逐次的なモデル化（Sequential modelling），シミュレーションによるデータ生成（Simulated Data），深層学習（Deep Learning）の 3 種類に大別される．逐次的なモデル化は，同時確率分布を設定した上で，逐次的に属性を合成する手法全般を表しており，属性間の関係性を保持しやすいという特徴がある．本研究でも実験に用いるノンパラメトリックな決定木手法である CART（= Classification And Regression Tree）はこれに含まれる．シミュレーションによるデータ生成は，対象となる元データの分布特性に関する情報が未知の場合に，合成データをシミュレーションで生成する手法全般を包含している．わが国の一般用ミクロデータは，その中の Analytically advanced simulated data に該当すると思われる．さらに，深層学習は GAN 等の手法を用いた生成技法であり，近年注目されている手法である．本研究で実験に用いる条件付き表形式 GAN（Conditional Tabular GAN=CTGAN）は，非構造化データやテキストデータを処理する唯一の方法であり，深層学習の一種である．

いずれの技法にもそれぞれ一長一短があり，原データの性質や用途に適した技法を選択することが重要である．

Taub et al.（2019）では，“The Synthetic Data Challenge” という形で，複数の研究チームが 1901 年スコットランドの歴史的な人口センサスを対象にした合成データの試行的な生成が追究されている．具体的には，①地域で層化を行った上での CART の適用，②分位点回帰やロジスティック回帰の適用，③ランダムサンプリング，④χ^2 統計量に基づく有向グラフ等，様々な生成方法を用いて作成された合成データを対象に，有用性と露見リスクの両面から検証が行われており，全体的に，有用性が高い合成データであるほど，露見リスクも相対的に高まるという結果が得られている．

また，Little et al.（2021）は，1991 年イギリス人口センサスの匿名化標本データ（Samples of Anonymised Records）を用いて，synthpop を用いて行った CART，DataSynthesizer を用いたベイズアプローチ，深層学習モデルと位置付けられる 2 つのタイプの GAN（CTGAN および TableGAN）を比較・検討を行っている．その結果，synthpop を用いた CART は，テストした 4 つの手法の中で最も実用性が高いが，個体が特定されるリスクも高いことが確認できた．それに対して，Table GAN における個体特定リスクは最も低いが，有用性が低いことが確認されている．

その一方で，横溝・伊藤（2023）においては，経済センサス活動調査の個票データを用いて，匿名化手法と合成データ生成技法の比較・検討を行っている．本研究では，ミクロアグリゲーションといった攪乱的手法が適用された匿名化ミクロデータの定量的な評価を行い，秘匿性と有用性の観点から攪乱的手法の比較・検証を行っただけでなく，回帰や決定木に基づく各種の合成データ生成技法を用いて，秘匿性と有用性の評価指標による検証を行った．本研究の結果から，CART については，ミクロアグリゲーション技法と比較して，有用性は保持された状態で，相対的に秘匿性が高いことが実証的に確認されている．さらに，伊藤・横溝（2024）では現行の統計法の下における合成データの展開可能性を指向しており，個票データに直接 CART を適用した結果とミクロアグリゲートされたデータに CART を追加的に適用した結果を比較・検証した上で，有用性と秘匿性の両面において大きな差が生じないことを確認している．なお，本研究は，横溝・伊藤（2023）に基づいて，合成データの生成技法としての CTGAN の適用可能性を追究している．

3. 使用するデータ

本研究で使用するデータは，「平成 28 年経済センサス-活動調査」（以下「経済センサス」と略称）の個票データである．テストデータにおいては，横溝・伊藤（2023）と同様に，製造業を対象とした従業者規模 1 人以上 1,000 人未満である 10,000 事業所が含まれており，このテストデータから 10,000 レコードを合成する．本研究で分析のために用いる属性群については，質的属性が，地域（8 区分），産業（11 区分），従業者規模（5 区分）と資本金階級（5 区分）であり[1)]，量的変数が，売上（収入）金額，付加価値額，給与総額と減価償却費である．

ところで，経済センサスのような事業所・企業系の統計調査においては，従業者規模や資本金が大きい企業および事業所が調査対象として常に含まれることが少なくない．こうした企業や事業所の場合には属性の多くが外れ値を有することから，相対的に分布の歪みが大きくなるだけでなく，入手可能な外部情報とのマッチングによって，個体情報が露見されるリスクが高くなる．また，量的変数間の相関性が高いことも事業所・企業系の統計調査における特徴と言える．

表 2 と表 3 はそれぞれ，使用する量的属性の基本統計量と相関係数行列を示したものであ

表 2. 本分析で使用する量的変数の基本統計量．出所　横溝・伊藤（2023）．

変数名	n	mean	sd	median	at_1%	at_99%
従業者合計	10,000	17	47	5	1	226
資本金額	6,843	80,119	1,200,141	1,000	84	1,192,520
売上（収入）金額	10,000	53,768	427,885	3,500	0	859,421
給与総額	7,618	2,430	6,034	600	0	24,622
減価償却費	7,618	371	1,607	34	0	5,847
付加価値額	10,000	10,640	55,503	1,459	-823	154,092

表 3. 本分析で使用する量的変数の相関係数．出所　横溝・伊藤（2023）．

	売上（収入）金額	付加価値額	給与総額	減価償却費
売上（収入）金額	1.00	0.60	0.81	0.62
付加価値額	0.60	1.00	0.52	0.33
給与総額	0.81	0.52	1.00	0.50
減価償却費	0.62	0.33	0.50	1.00

る．量的変数においては，平均値と中位数の数値が大きく異なるため，分布の山のピークが中心の左側に位置し，右側の裾が大きくなるだけでなく，給与総額や減価償却費では 0 の値が存在しており，付加価値額では負の値も含まれることがわかる．そこで，本研究では，量的変数に関しては，neg-log 変換によって 0 や負の値も含めて対数化を行った[2)]．それによって，分布の歪み，さらには 0 や負値の問題を解決することが可能になる．例えば売上(収入)金額について neg-log 変換を行うと，概ね対数正規分布に近い分布に変換される．また，付加価値額の場合，正の山と負の山がそれぞれ現れることが確認された．表 3 の相関係数行列についても売上(収入)金額と給与総額では，0.8 という高い値を示している．

4. 合成データにおける秘匿性と有用性の評価方法について

先述のように経済センサスにおいては，外れ値(特異値)のような形でレコードが存在する可能性があるため，対象となるレコード群の中の個体が特定されるリスクが存在するだけでなく，合成データにおいて個体の属性情報が推定される可能性が指摘される．そこで，秘匿性に関する評価指標をもとに，合成データにおける露見リスクを評価することが求められる．また，合成データの利用可能性を追究しようとすれば，原データに対する合成データの有用性の定量的な評価を検討する必要がある．

本研究では，横溝・伊藤（2023）と同様に，絶対相対差分(Absolute Relative Difference=ARD)と呼ばれる属性漏洩(attribute disclosure)に関する評価指標（Kim et al., 2021）を用いた検証を行う．ARD は原データと合成データに含まれる属性値の最大値同士の乖離を評価する指標であり，以下の(4.1)式で表される．

$$\text{ARD} = \frac{|\hat{L} - L|}{L} \tag{4.1}$$

L：原データに含まれる属性値の最大値

$\hat{L}$：合成データに含まれる属性値の最大値

合成データに対して侵入者が取りうる攻撃手段の 1 つは，キーとなる属性の層ごとのセンシ

ティブな属性の最大値を推定することである．例えば，地域と産業というキーとなりうる属性のクロスで考えた場合，攻撃者はその地域と産業の層に含まれるレコードの売上の最大値を調べることで，その地域において該当する産業の大規模な事業所に関してセンシティブな属性(売上)を推測できる可能性がある．そこで本研究では，横溝・伊藤 (2023) や伊藤・横溝 (2024) と同様に，キー変数の全ての組み合わせのそれぞれについて原データと合成データの最大値の乖離を計算し，その平均値を用いて算出する，層化平均 ARD(stratified average ARD)を秘匿性の評価指標として使用した．

つぎに本研究では，合成データにおける有用性の評価指標については，傾向スコア(propensity score) (Woo et al., 2009)を用いた評価指標の 1 つである，傾向スコア平均二乗誤差(propensity score Mean Square Error=pMSE) (Snoke et al., 2016)を有用性に関する定量的な指標と設定した上で実証分析を行った．なお，pMSE は以下の(4.2)式で表される．

$$\text{pMSE} = \frac{1}{N}\sum(\widehat{p_i} - c)^2 \tag{4.2}$$

N：原データのレコード数と合成データのレコード数の和

$\widehat{p_i}$：各レコードの傾向スコア

c：N に占める合成データのレコード数の割合

本研究でも，横溝・伊藤 (2023)および伊藤・横溝 (2024)と同様に，最初に原データと合成データに含まれるレコード群を統合し，つぎに「合成データか否か」を被説明変数として設定した上で，ロジット回帰モデル等を用いて，各レコードについて合成データである可能性がどの程度高いかを確率的に表す傾向スコア p_i を算出する．そして，傾向スコア p_i とレコード全体に占める合成データの比率を表す c との乖離の差の平均値として，pMSE を計測する．

また本研究では，相関係数行列の差の平均絶対誤差(mean absolute error of the difference of the correlation coefficient matrices) (Domingo-Ferrer and Torra, 2001; 伊藤 他, 2014)も有用性に関する指標として使用した．なお，相関係数行列の差の平均絶対誤差は，以下の(4.3)式で表される (伊藤 他, 2014)．

$$\frac{\sum_{j=1}^{k}\sum_{1\le i\le j}|r_{ij} - r'_{ij}|}{\frac{k(k-1)}{2}} \tag{4.3}$$

k：属性の数

r：原データの相関係数

r'：合成データの相関係数

5. 合成データの生成技法の有効性に関する比較研究

本研究は，経済センサスの個票データを例に，匿名化技法として用いられる攪乱的手法だけでなく，各種の合成データの生成技法を用いて，その有効性についての比較実験を行う．本実験における攪乱的手法については，ミクロアグリゲーションの 1 つである MDAV 法を用いる．また，本研究で用いる合成データの生成技法は，CART，バギング(Bagging)，ランダムフォレスト(Random Forest)，および深層学習モデルの 1 つである CTGAN である．

ミクロアグリゲーションは，同質的なレコード群にグループ化した上で，個々の属性値を平均値等の代表値に置き換える攪乱的手法 (Defays and Nanopoulos, 1993; 伊藤, 2009)である．その 1 つである MDAV(=Maximum Distance to Average Vector)法 (Domingo-Ferrer and Mateo-Sanz, 2002; Hundepool et al., 2003)は，対象レコードにおける平均値からの距離を考慮

した上でレコード間の近似性を最大にするように，平均ベクトルからの距離が最大となるレコードから優先的にグループ化を行う技法である．MDAV 法は，以下のような手順で適用される（横溝・伊藤, 2023）．それは，①対象となるレコード群について属性値の各々に関する平均値のベクトルを算出し，②平均からの距離が最大のレコードとそのレコードからの距離が最大となるレコードを探索し，③しきい値 k を設定した上で，対象となるレコードを含む近傍の k 個のレコードをグループ化して平均値に置換し，④ミクロアグリゲート済（攪乱済み）のレコードを除去した後に，②と③の処理を繰り返すというものである．

つぎに，CART（Classification And Regression Tree）（Breiman et al., 1984）は，観測済みの属性値から目的変数となる属性を再帰的にグルーピングする，ノンパラメトリックな決定木分析の手法である．CART の特徴として，質的属性と量的属性の両方においての適用可能な技法であることが指摘できる．

CART のメリットとして，決定木の作成における枝の分岐や決定木の末端である各々の葉の分布が可視化できることが指摘される．また，他の手法と比較して，適用された結果の解釈が容易であることから，実務への適用の観点から，幅広い分野での利用が可能である．さらに．葉のサイズに対して制約を与えることによって，枝の分割の細かさを制御できることから，CART によって作成された合成データの有用性と秘匿性の程度を調整することもできる．

図 1 は，CART を用いて，地域に基づいて産業を合成した場合の決定木の例を示している．「CHIKI8」は 8 区分にリコーディングした地域の分類区分を表しており，「09_10」は，産業中分類 “09”（食料品製造業）と “10”（飲料・たばこ・飼料製造業）をリコーディングした分類区分に対応している．本図においては，最初に原データを対象に，地域に基づいてそれぞれの葉が異なる分布を持つような産業の決定木を作成する．その場合の枝の分割の基準としては，ジニ係数といった指標が一般に用いられる．次に，合成の対象となるレコードの地域を用いて決定木を辿り，その先にある分布を用いて産業をランダムに合成する．図 1 では，例えば左端の Node2 は，クラスタリングの結果 2069 レコードの産業で葉が構成されている（横軸はリコーディング済の産業分類，縦軸は構成比）．枝の分割において北海道や東北といった地域を持つレコードの産業が振り分けられており，09_10 の食品・飲料等製造業等が他の地域に比べて相対的に多く合成されるように葉が形成されている[3)]．

また，表 4 は，CART によって作成された合成データにおける分布特性を示している．表 2 における原データと比較すると，概ね分布特性は再現されている．一方で経済センサスにおいては，世帯・人口系の統計調査と比較して分布の歪みが大きいことから，合成されたレコードにおいて特に右裾の部分に関しては，原データとは異なる分布特性になっていることが示唆される．これについては原データに分布特性を近づけるための合成の方法に関して改善の余地がある．なお，相関係数においては，原データをもとに相関係数が算出された表 3 と比較しても，相関係数は概ね再現されていることが確認される．

図 2 は，原データと CART による合成データについて，質的属性，量的属性の構成比を比較したものである．図中の濃い棒線が原データを，薄い棒線が合成データをそれぞれ示している．いずれの属性においても構成比の違いはほとんどなく，原データの分布を概ね再現できている[4)]．

バギング（bagging）は，CART を発展させた合成データの生成技法であって，ブートストラップ法を用いて決定木を複数作成する技法である．また，ランダムフォレストは，ブートストラップ法だけでなく，使用する属性の組も変化させて決定木を複数作成する手法である．これらの技法についても，本研究では，合成データの生成技法の有効性を検証するために用いた．

本研究の大きな特徴は，わが国の経済センサスの個票データを用いて，合成データの生成技法の 1 つである GAN の有効性について，MDAV 法や CART 等との比較・検証を行っている

レコード	地域（合成済）	産業（新規に合成）	…
1	北海道	食品・飲料等製造業	
2	北海道	繊維工業	
3	北海道	木材・家具等製造業	
…	…	…	
5001	関東	電子・電気・情報通信等製造業	
5002	関東	印刷関連業	
…	…	…	

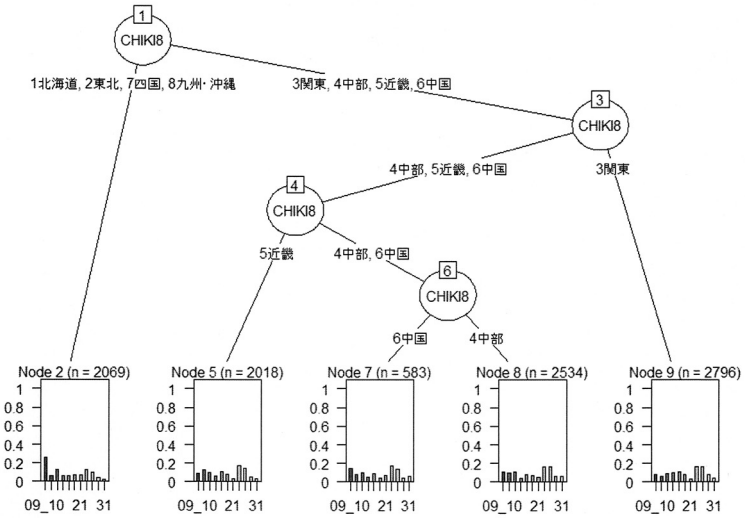


図 1. CART で作成した決定木の例.

ことである．敵対的生成ネットワーク GAN（Generative Adversarial Networks）は，教師なし学習で使用される人工知能アルゴリズムの一種であり，生成ネットワーク（generator）と識別ネットワーク（discriminator）を競わせることによって，学習データを模倣した精巧な出力を得ることができる技術である．図 3 で示されるように，生成される合成データの品質を向上させるために，①ノイズから合成データを生成し，②実データと合成データを識別させた上で，③識別の正否をフィードバックして学習させるという一連の手順が繰り返される．

本研究では，GAN の一手法である条件付き表形式 GAN（conditional tabular GAN=CTGAN）（Xu et al., 2019）を用いて合成データの生成を行った[5]．CTGAN は，表形式データの分布をモデル化し，その分布から合成データを生成する．GAN における合成データ生成の特徴としては，①離散値（質的属性）と連続値（量的属性）が混在するデータが合成の対象になっていること，②非ガウス分布や多峰性分布を対象にした合成データの生成が求められること，③スパースな質的属性に基づいて学習がなされる場合があること等が指摘されている．こうした分布の歪みやスパースな分類区分が存在する場合，合成データにおいて分布特性を精密に再現することは困難になることが想定される．そこで，CTGAN では，離散値を変分ガウス混合モデルで近似し，それぞれの分布の最頻値（mode）を用いて正規化することや，質的属性におけるすべて

表 4．CART によって作成された合成データにおける分布特性．

(1)基本統計量

変数名	n	mean	sd	median	at_1%	at_99%
売上（収入）金額	10,000	60,209	541,673	3,500	0	880,803
付加価値額	10,000	11,324	66,158	1,441	-738	170,776
給与総額	7,536	2,427	6,320	577	0	28,649
減価償却費	7,622	389	1,699	36	0	6,023

(2)相関係数行列

	売上（収入）金額	付加価値額	給与総額	減価償却費
売上（収入）金額	1.00	0.59	0.81	0.63
付加価値額	0.59	1.00	0.53	0.35
給与総額	0.81	0.53	1.00	0.50
減価償却費	0.63	0.35	0.50	1.00

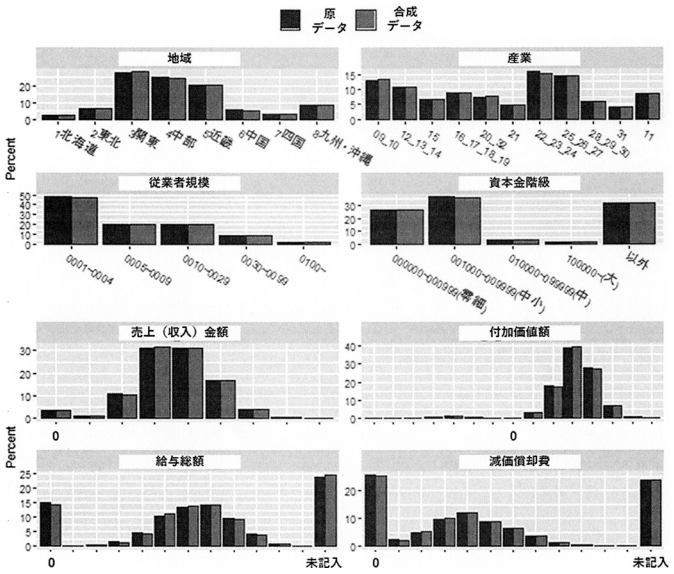


図 2．CART による合成データの属性別構成比．出所　横溝・伊藤（2023）．

の分類区分ができるだけ均等に学習されるように，条件付き(conditional)でオーバーサンプリングすることによって，これらの課題の改善が図られている．

図 4 は，予備実験として行った生成ネットワーク G と識別ネットワーク D の損失関数を用いた実験の結果を示したものである[6]．損失関数は深層学習全般に存在する概念であり，正解と予想の誤差(損失)をもとに，学習の成否や学習回数の判断するために用いられる．学習がうまくいっていない場合には，D が振動しながら発散する状態が見られるが，本実験においては D の損失関数は安定していることから，学習は概ね成功していると考えられる．

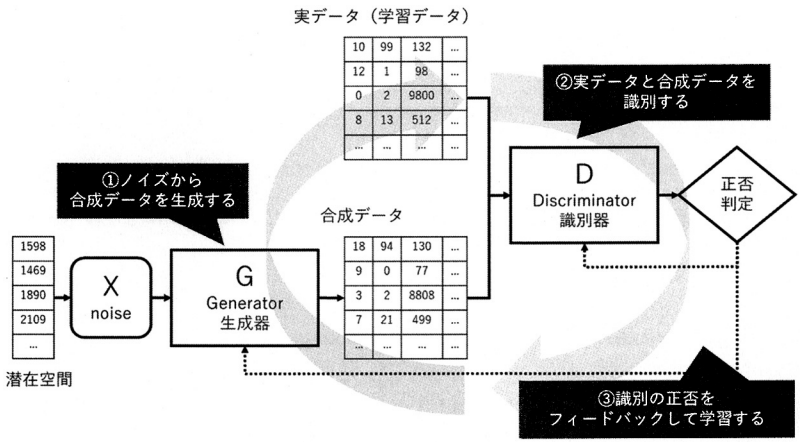


図 3. GAN のイメージ図.

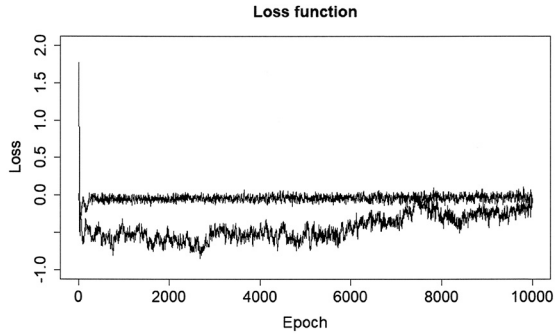


図 4. CTGAN の損失関数.

本実験では，MDAV 法，回帰による生成，サンプリングによる生成，CART，バギング，ランダムフォレストおよび CTGAN については，秘匿性と有用性の比較・検証を行っている．有用性の指標としては，前節で述べたように傾向スコアの指標である pMSE，および相関係数の平均絶対誤差(mean absolute error of correlation)，秘匿性の指標に関しては ARD をそれぞれ用いて検証を行った．比較の対象となる MDAV 法については，グループ化の対象となるレコード数 k を 3，5，10，30，50，100 と変化させた場合の，CART，バギング，ランダムフォレストに関しては，決定木における葉の大きさの制約である最小リーフサイズを 3，5，10，30，50，100 と変化させた場合の有効性の評価をそれぞれ行った．いずれも数字が大きくなるほど粗くクラスリングされるため，秘匿性は高くなる一方，有用性は低くなることが想定される．CTGAN については，epoch 数(学習の回数の単位)を 100，300，1000，3000，5000，10000 と変化させた上で，有効性の比較を行っている．こちらは学習回数が多いほど精巧な合成データに近付くため，有用性は高くなる一方，秘匿性は低くなることが想定される．その他のハイパーパラメータについては，各ミクロデータ生成ツールのデフォルトのパラメータが設定されている．

CART のような決定木による合成データの生成技法には乱数発生に伴うばらつきが生じることから，合成データを 10 回生成した上で，その平均値が本研究で用いられている．なお，決

定木手法だけでなく，CTGAN もこのような乱数の影響を受けることに留意されたい．

6. 実証実験の結果

本節では，経済センサスを用いた合成データの生成技法の有効性に関する実証実験の結果について述べる．図 5 と図 6 はそれぞれ，実証実験の対象となった各種の合成データの生成技法を比較した R-U マップ（Risk Utility map）（Duncan et al., 2001）を図示したものである．本分析で検証の対象となった MDAV 法，回帰による生成，サンプリングによる生成，CART，バギング，ランダムフォレストおよび CTGAN は，図 5 と図 6 ではそれぞれ，mdav，parametric，sample，cart，bag，rf および ctgan と表示されている．横軸は秘匿性の指標である層化平均 ARD を表している．また縦軸の有用性の指標としては，図 5 では pMSE，図 6 では相関係数の平均絶対誤差を用いた結果をそれぞれ表しており，秘匿性の強度と有用性の程度も大きいほど，右下のエリアにプロットされる．

図 5 を見ると，全体的に k が大きくなるにしたがって，秘匿性の程度は大きく，有用性は相対的に小さくなる傾向が見て取ることができる．ミクロアグリゲーションである MDAV 法における実証結果は，相対的に秘匿性の強度が小さなエリアにプロットされる傾向にある．すなわち攪乱的手法である MDAV を適用した場合，本実験で生成された各種の合成データと比較して相対的に秘匿性が小さくなる傾向にあることがわかる．

つぎに，CART やバギングに着目すると，最小リーフサイズを大きくしても有用性が大きく低減しないだけでなく，MDAV 法と比較して秘匿性の強度が大きいことが確認できる．ラン

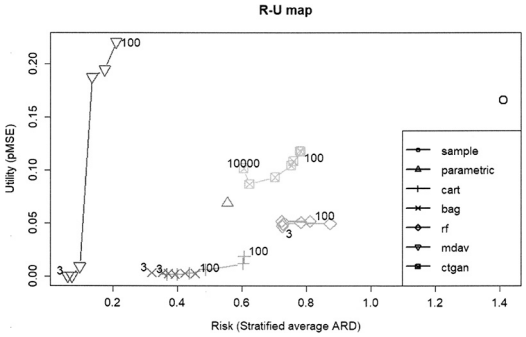


図 5．秘匿性と有用性の評価結果：pMSE を用いた場合．

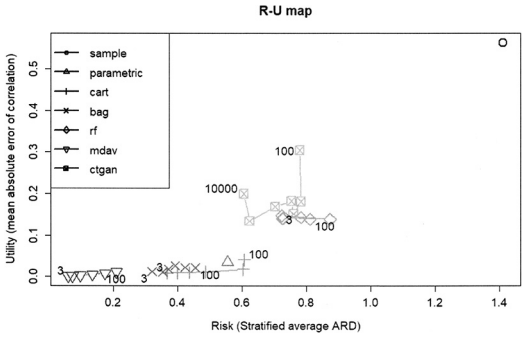


図 6．秘匿性と有用性の評価結果：相関係数の平均絶対誤差を用いた場合．

ダムフォレストの場合も同様に，最小リーフサイズにかかわらず，有用性の程度は大きく変化しないが，MDAV 法との比較においては，グループ化の対象となるレコード数 k が 3 の場合には，MDAV 法においては，ランダムフォレストに対してより有用性が高くなっている．そして，CTGAN によって得られた実証結果については，CART，バギングとランダムフォレストを適用して生成した合成データにおける結果と比較すると，その秘匿性の程度は大きくなる傾向にあるが，有用性の低下も大きくなっていることが興味深い．

図 6 では，横軸の秘匿性については先ほどと同様に層化平均 ARD を，縦軸に関しては有用性として相関係数行列の絶対誤差で表している．有用性の指標として pMSE を有用性の指標として用いた図 5 と比較して，MDAV 法，CART・バギング・ランダムフォレスト，および CTGAN の位置関係がより顕著に表れている．すなわち，MDAV 法，CART，CTGAN の順に秘匿性の強度が大きくなっているが，MDAV 法と CART については，有用性の大きな差は見られない．このように，相関係数を有用性の指標とした場合でも，R-U マップにおいて類似した傾向が現れていることがわかる．その一方で，CTGAN においては，有用性に関しては epoch 数次第で不安定になる傾向が確認された[7)]．

7. おわりに

本稿では，経済センサス活動調査の個票データを用いて，事業所・企業系のミクロデータに対して各種の合成データの生成技法の有効性を評価した．本研究においては，攪乱的手法であるミクロアグリゲーション（MDAV），合成データの生成技法である CART や回帰，および CTGAN を対象に，秘匿性や有用性に関する定量的な評価を行った．本研究でも，CART で合成データを生成した結果，分布特性として要約統計量や相関係数，属性ごとの構成比が再現可能なことがわかった．また，CART は MDAV と比較して，有用性を保ったまま秘匿性を高められる可能性があり，またその程度を調節することも可能なことが確認できた．さらに本稿では，公的統計のミクロデータに対する CTGAN の合成データとしての適用可能性を検証した．CTGAN の場合，CART と比較して，秘匿性の強度が高くなっていることが明らかになったが，有用性の低下も大きいことがわかった．

なお，合成データの生成技法の有効性に関する比較・検証については，統計実務の観点に立った場合には，合成データの利用可能性を検討するために，さらなる実証研究を進める必要があるように思われる．これについては，将来的な研究課題としたい．

注．

1) 本研究で使用した質的属性については，リコーディング済みの変数が用いられている．

2) neg-log 変換は，つぎの式で表現される．

$$Y_n = sgn(X_n) \times \ln(|X_n| + 1) \tag{3.1}$$

neg-log 変換がなされているのは量的属性のすべてではなく，売上等の経理項目のみであることに留意されたい．例えば，従業者数や資本金については，リコーディングのみが適用されている．なお，neg-log 変換に関する説明については，高部 (2017) も参照されたい．

3) 図の例では質的属性に対する CART であったが，量的属性に対しても CART を適用することができる．その際，原データの持つ特異な量的属性値（例えば 777 といった目立ちやすい売上など）であっても合成データとしてそのまま生成されるため，平滑化（smoothing）を行うことが考えられる．具体的には，ガウスカーネル密度推定（Gaussian

kernel density estimator)が用いられる．なお，カーネル密度推定は，synthpop にオプションとしてあらかじめ実装されている．

[4] CART の場合，特別の工夫を加えずとも 0 や未記入まで含めて属性ごとの分布を再現できる点は注目に値する．

[5] 米国の datacebo 社の提供する，合成データ生成ツール The Synthetic Data Vault(SDV)の一機能として，Python や R 用パッケージ “CTGAN” が公開されている．

[6] CTGAN の損失関数は以下で計算される．

$$\mathcal{L}_D = \frac{1}{m}\sum_{1}^{m}[D(x'^{(i)}) - D(x^{(i)})] \tag{5.1}$$

$$\mathcal{L}_G = \frac{1}{m}\sum_{1}^{m}[D(x'^{(i)})] + H \tag{5.2}$$

x：元データ
x'：合成データ
m：レコード数
H：交差エントロピー

[7] 深層学習では一般に，学習の初期の段階では epoch 数を増やすほど性能が向上することが多いが，ある点を越えると過学習(overfitting)を起こすことがある．本実験においては epoch 数 10000 では過学習の可能性が疑われる．

参 考 文 献

Bates, A. G., Špakulová, I., Dove, I. and Mealor, A. (2019). ONS methodology working paper series number 16 synthetic data pilot, https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber16syntheticdatapilot（最終アクセス日 2024 年 7 月 4 日）.

Breiman, L., Friedman, J., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*, 1st ed., Chapman and Hall/CRC, New York.

Chien, C.-H., Welsh, A. H. and Moore, J. D. (2021). Synthetic business microdata: An Australian example, *Journal of Privacy and Confidentiality*, **10**(2), https://doi.org/10.29012/jpc.733.

Defays, D. and Nanopoulos, P. (1993). Panels of enterprises and confidentiality: The small aggregates method, *Proceedings of 92 Symposium on Design and Analysis of Longitudinal Surveys*, 195–204, Statistics Canada, Ottawa.

Domingo-Ferrer, J. and Mateo-Sanz, J. M. (2002). Practical data-oriented microaggregation for statistical disclosure control, *IEEE Transactions on Knowledge and Data Engineering*, **14**(1), 189–201.

Domingo-Ferrer, J. and Torra, V. (2001). Disclosure control methods and information loss for microdata, *Confidentiality, Disclosure and Data Access*: *Theory and Practical Applications for Statistical Agencies* (eds. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz), 91–110, Elsevier Science, Amsterdam.

Duncan, G., Keller-McNulty, S. A. and Stokes, S. L. (2001). Disclosure risk vs. data utility: The R-U confidentiality map, Technical Report, No.121, National Institute of Statistical Sciences, Washington, D.C.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014). Generative adversarial nets, *Advances in Neural Information Processing*

Systems, **27**, 2672–2680.

Hundepool, A., de Wetering, A. V., Ramaswamy, R., Franconi, L., Capobianchi, A., DeWolf, P.-P., DomingoFerrer, J., Torra, V., Brand, R. and Giessing, S. (2003). *μ-ARGUS Version 3.2 Software and User's Manual*, Statistics Netherlands, Voorburg NL.

伊藤伸介 (2009). 匿名化技法としてのミクロアグリゲーションについて, 経済論集, **15**(3・4 号合併号), 197–232.

伊藤伸介 (2018). 公的統計ミクロデータの利活用における匿名化措置のあり方について, 日本統計学会誌, **47**(2), 77–101.

伊藤伸介, 横溝秀始 (2024). わが国の公的統計における合成データの展開可能性に関する一考察—事業所・企業系の統計調査を例に—, 経済学論纂, **64**(3・4 合併号), 147–164.

伊藤伸介, 村田磨理子, 高野正博 (2014). ミクロデータにおける匿名化技法の適用可能性の検証—全国消費実態調査と家計調査を用いて—, 統計研究彙報, **71**, 83–124.

Kim, H. J., Drechsler, J. and Thompson, K. J. (2021). Synthetic microdata for establishment surveys under informative sampling, *Journal of the Royal Statistical Society Series A*: *Royal Statistical Society*, **184**(1), 255–281.

Little, C., Elliot, M., Allmendinger, R. and Samani, S. S. (2021). Generative adversarial networks for synthetic data generation: A comparative study, https://doi.org/10.48550/arXiv.2112.01925.

Nowok, B., Raab, G. M. and Dibben, C. (2016). synthpop: Bespoke creation of synthetic data in R, *Journal of Statistical Software*, **74**(11), 1–26.

Snoke, J., Raab, G. M., Nowok, B., Dibben, C. and Slavkovic, A. B. (2016). General and specific utility measures for synthetic data, *Journal of the Royal Statistical Society Series A*: *Statistics in Society*, **181**(3), 663–688.

高部勲 (2017). 状態空間モデルに基づく季節調整法における改良方法の提案：一般化 neg-log 変換の活用に基づくゼロ・負の値を含む時系列データの安定化と季節調整値の推定精度向上, 統計研究彙報, **74**, 29–56.

Taub, J. and Elliot, M. (2019). The synthetic data challenge, Paper presented at Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, the Hague, the Netherlands, https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019_S3_UK_Synthethic_Data_Challenge_Elliot_AD.pdf（最終アクセス日 2024 年 7 月 4 日）.

Templ, M. (2017). *Statistical Disclosure Control for Microdata*: *Methods and Applications in R*, 99–132, Springer International, Cham, Switzerland, https://doi.org/10.1007/978-3-319-50272-4.

UNECE (2022). Synthetic data for official statistics — A starter guide, https://unece.org/sites/default/files/2022-11/ECECESSTAT20226.pdf（最終アクセス日 2024 年 7 月 4 日）.

Woo, M.-J., Reiter, J. P., Oganian, A. and Karr, A. F. (2009). Global measures of data utility for microdata masked for disclosure limitation, *Journal of Privacy and Confidentiality*, **1**, 111–124.

Xu, L., Skoularidou, M., Cuesta-Infante, A. and Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN, *Advances in Neural Information Processing Systems*, **32**, 7333–7343.

横溝秀始, 伊藤伸介 (2023). 合成データの生成手法の有効性に関する定量的な評価—事業所・企業系のミクロデータを用いて—, 統計研究彙報, **80**, 97–116.

Study on Synthetic Data Generation Techniques for Official Statistics on Establishments and Enterprises: The Economic Census as an Example

Shinsuke Ito[1] and Shuji Yokomizo[2]

[1]Faculty of Economics, Chuo University
[2]Statistical Research and Training Institute, Ministry of Internal Affairs and Communications

Not only are anonymous data currently not available for statistical surveys of business establishments and enterprises in Japan, but producing publicly available microdata for statistical surveys of business establishments and enterprises is also difficult. Therefore, the development of a method to produce synthetic data would meet the need for test data. In this paper, we quantitatively evaluate various techniques for generating synthetic data using individual data from the Economic Census of Activity.

In this study, we quantitatively evaluated the usefulness and confidentiality of synthetic data generated using the maximum distance to average vector (MDAV) method of microaggregation, which is a disturbing method, the classification and regression tree (CART), and the conditional tabular GAN (CTGAN), a deep learning model. The results confirmed that the distributional properties such as summary statistics and correlation coefficients were reproducible when synthetic data were generated using CART. In addition, compared with the MDAV method, CART can potentially increase confidentiality while maintaining usefulness. Furthermore, for CTGAN, the degree of confidentiality was found to be higher compared with that for CART; however, the decrease in usefulness was also confirmed to be relatively greater.

Key words: Synthetic data, microaggregation, CART, CTGAN, Economic Census.

統計数理(2024)
第 72 巻 第 2 号 233–244
©2024 統計数理研究所
特集「公的統計―社会の重要な情報基盤―」
［研究ノート］

多項ロジットモデルに基づく統計的マッチングの欠測値補完への応用

高部 勲†

(受付 2024 年 2 月 16 日；改訂 7 月 22 日；採択 7 月 25 日)

要　　旨

統計的マッチングは，異なるデータを組み合わせて有用なデータを構築するための手法である．統計的マッチングにより，追加の調査やデータの収集を行うことなく，有益なデータを作成することが可能となり，近年，様々な分野で利用が進んでいる．本研究では，高部・山下(2021)，Takabe and Yamashita (2020)及び高部・山下 (2018)で提案された，多項ロジットモデルに基づく統計的マッチングの手法について紹介するとともに，その副産物として得られるマッチング確率を欠値補完に活用する方法について検討し，実データを用いて，その試算・分析を行った．

キーワード：統計的マッチング，多項ロジットモデル，ウエイト付き距離関数，欠測値補完．

1. はじめに

1.1 公的統計データの欠測値補完に関する概要

社会学におけるアンケート調査，企業による市場調査，行政機関が実施する公的統計調査など，様々な調査においては，無回答・無記入により，調査対象や調査項目の一部についての情報が得られない場合がある．このような欠落した値を欠測値という（日本統計学会, 2022, 2023)．欠測値の要因である無回答には，白紙や未回収など，構成要素単位で発生する全項目無回答(unit nonresponse)と，いくつかの質問に回答がないなど，変数単位で発生する一部項目無回答(item nonresponse)がある（高橋・渡辺, 2017)．欠測値が発生した場合，そのままでは統計的計算処理が不可能となることから，調査対象への再調査等により，可能な限り正しいと考えられる結果を得ることが望ましい．しかし，そのような対応が困難な場合には，欠測値を何らかの方法により補完(imputation)する場合がある（日本統計学会, 2022, 2023)．

単一の統計表の公表を念頭に置いた欠測値の補完においては，1 つの欠測値に 1 つの値(平均値，モデルによる予測値など)を代入する単一代入法(single imputation)が用いられる場合が多い．そのような場合に，得られたデータを用いて関心のあるパラメータの推定を行った場合，例えば分散の過小推定が起こるなど，パラメータの推定値にバイアスが生じる可能性があることが問題点として挙げられる．こうした問題を回避するために，ベイズ統計の考え方を基に，欠測値を含むデータの分布から独立かつ無作為に抽出された複数のシミュレーション値によって欠測値を置き換える多重代入法(multiple imputation)が用いられる場合がある（高橋・渡辺,

† 立正大学 データサイエンス学部：〒 360–0194 埼玉県熊谷市万吉 1700

2017; 高井 他, 2016)．欠測値を含むデータの分析においては，完全にランダムな欠測(MCAR: missing completely at random)，ランダムな欠測(MAR: missing at random)などの欠測のメカニズムを勘案しつつ，推定対象の特性も考慮しながら推定を行う必要があり，そのための様々な手法が提案されている（高橋, 2022; 高橋・渡辺, 2017; 高井 他, 2016; 星野, 2009; 岩崎, 2002)．

ところで，国勢調査のような公的統計調査のデータに欠測値が生じた場合の補完については，我が国を含む多くの国において，公表される多様な集計表の元となる単一の集計用ミクロデータを作成することを念頭に，整合性のあるデータを当該欠測値に代入して補完を行う単一代入法が用いられており，特に同じ調査データの中から欠測した値に類似していると考えられるレコードを当該データから検索して代入するホットデック法(hot deck imputation)や，行政記録などの外部データから検索して代入するコールドデック法(cold deck imputation)が多く利用されている（北原・寺垣内, 2023; 坂下, 2018; 統計委員会担当室, 2013)．

ホットデック法及びコールドデック法では，欠測値を伴うレコードをレシピエント(recipient)と呼び，代入値を提供するレコードをドナー(donor)と呼ぶ．そして，それらのレコードに共通に含まれる変数から何らかの形で距離関数を定義して，距離が最小となるレコードをドナーとする最近隣法(nearest neighbor imputation)が用いることが多い．また，単一のレコードの値を代入するのではなく，複数のレコードの値を何らかのウエイトに基づいて加重平均した値を代入する fractional imputation と呼ばれる手法も考案されている（De Waal et al., 2011)．ホットデック法及びコールドデック法は，2 つのデータの間で類似するレコードを探索するという観点から，統計的マッチングと本質的には同じ考え方であることが指摘されている（高橋・渡辺, 2017)．

本稿では，こうした考え方に基づき，国内外の動向も勘案しつつ，公的統計データを念頭に置いた，統計的マッチング(statistical matching)の手法の欠測値補完への応用について検討する．その準備として次節では，統計的マッチングの概要について説明する．

1.2 統計的マッチングの概要

近年，情報処理技術の発展やネットワーク環境の向上により，様々なデータが利用可能になっており，これらのデータを何らかの形で結合することができれば，新たに統計調査やデータの収集等を行うことなく，情報量の多い有用なデータを構築することができる．

こうした中で，複数のデータを結合するためのデータリンケージの手法が，様々な分野で注目を集めている（Herzog et al., 2007; Christen, 2012; Harron et al., 2015)．ところで，データリンケージを行う際に，各レコードを識別できる照合キー(共通一連番号，名称，所在地など)が存在する場合には，それらを利用した完全照合(exact matching)が可能となる（村田・伊藤, 2016; 山口, 2014)．しかし，このような照合キーが利用できない場合には，各データの共通変数を基に算出した距離が近いレコード同士を結合する方法が用いられる．これを統計的マッチング(statistical matching)という（美添, 2005)．統計的マッチングのイメージを示したものが以下の図 1 である．

統計的マッチングに関する研究は 1960 年代から行われてきており，初期には，名称・所在地などを基に，異なるレコードを同一と判定する確率と，同一の対象を表すレコードが正しく同一であると判定される確率の比率を基にマッチングの適否を判定する方法（Newcombe et al., 1959; Fellegi and Sunter, 1969)が開発されてきている．その後，共通変数以外の変数を欠測値とみなして重回帰モデルやベイズ統計学の枠組みに基づいて推測を行う方法（D'Orazio et al., 2006; Rassler, 2002; 栗原, 2015)，各レコードがどちらのデータに属するかという確率(傾向スコア(Propensity Score))の値が近いレコード同士をマッチングする方法（Rubin, 1986; Stuart, 2010)などの様々な手法が研究・開発されている．

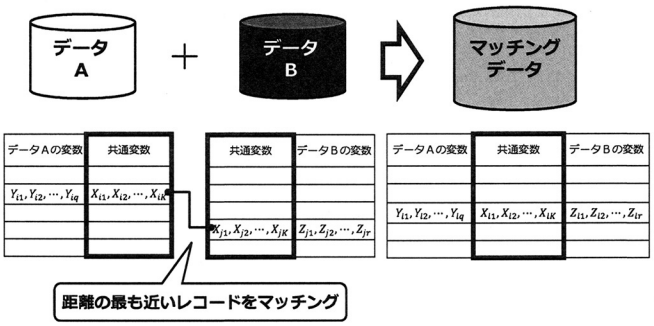


図 1. 統計的マッチングのイメージ.

距離に基づく統計的マッチング(各データに共通の変数を用いてレコード間の距離を計算し，最も近いレコード同士のマッチングを行う手法，D'Orazio et al., 2006)は，比較的初期の段階から研究が行われてきている．ただし，この方法では，各変数の重要度やスケール調整の方法をどのように決定するかについて一般的な基準が無く，各変数のウエイトの決定方法が恣意的になるおそれがある．この問題に対応するため，高部・山下 (2021)，Takabe and Yamashita (2020)及び高部・山下 (2018)では，多項ロジットモデルを用いた統計的マッチングの手法を提案している．この手法は，前述の先行研究における課題を克服している点が特長である．具体的には，距離の最適なウエイトをデータから統計的に(最尤法により)推定することが可能であり，名称・所在地などの詳細な文字情報が利用できない場合でも適用可能な方法であり，さらに，連続変数とカテゴリ変数が混在する場合でも適用可能な方法となっている．また，通常の多項ロジットモデルの推定の際に得られる t 値・p 値などの統計量により，ウエイトの推定精度を分析することが可能であり，レコードの一致確率(マッチング確率)を推定し，マッチングの精度を定量的に評価することが可能である．

1.3 本稿で提案する欠測値補完への応用

本稿では，高部・山下 (2021)，Takabe and Yamashita (2020)及び高部・山下 (2018)における多項ロジットモデルを用いた統計的マッチングの手法について，その概要を説明するとともに，モデルの推定の過程で副産物として得られるマッチング確率の欠測値補完への効果的な利活用方法について提案・検討する．具体的には，データから推定した統計的マッチングに関する多項ロジットモデルを用いて，外部のデータからマッチング確率が最大となるレコードを検索し，そこから値を代入するコールドデック法について検討する．また，本来はホットデック法を念頭に提案されている fractional imputation の考え方を準用し，マッチング確率をウエイトとして用いて，単一のレコードではなく複数のレコードの加重平均により欠測値を補完するコールドデック法の改善方法についても検討する．そして，提案手法を実際のデータ(公的統計データ及び商用データ)を基に，人工的に欠測させた値に対して提案手法を適用し，その効果について検証する．

次節では，提案手法のベースとなる，多項ロジットモデルを用いた統計的マッチングの手法について，高部・山下 (2021)，Takabe and Yamashita (2020)及び高部・山下 (2018)を基に，その手法や推定方法などの概要を説明する．

2. 多項ロジットモデルに基づく統計的マッチング

2.1 手法の概要

以下の 2 種類のデータ(データ A 及びデータ B)の統計的マッチングを行う場合を想定する.

- データ A(マッチング元,ドナー(Donner)):レコード数 M
- データ B(マッチング先,レシピエント(Recipient)):レコード数 N

ここで,データ A の i 番目のレコードと,データ B の j 番目のレコードが同一の対象である確率 P_{ij} を考える(以下,これをマッチング確率という).ここで P_{ij} は,レコード間の距離 D_{ij} を用いて次のように表現できるものとする.

$$P_{ij} = \frac{\exp(-D_{ij})}{\sum_{j=1}^{N} \exp(-D_{ij})} \tag{2.1}$$

距離 D_{ij} の形式については様々なものが考えられるが,以下の絶対値距離(Manhattan 距離)が用いられることが多い.

$$D_{ij} = \sum_{k=1}^{p} \beta_k |X_{ik} - X_{jk}| \tag{2.2}$$

カテゴリ変数(離散変数)に対しては,以下の距離が用いられる.

$$D_{ij} = \sum_{k=1}^{p} \beta_k I(X_{ik} = X_{jk}) \tag{2.3}$$

ここで $I(X_{ik} = X_{jk})$ は,以下のように定義される関数である.

$$I(X_{ik} = X_{jk}) = \begin{cases} 1 & (X_{ik} = X_{jk}) \\ 0 & (X_{ik} \neq X_{jk}) \end{cases} \tag{2.4}$$

距離のウエイト β_k については,連続変数の場合には,レンジ(最大値から最小値を減じたもの)の逆数やデータの標準偏差の逆数,質的変数の場合は 1 が用いられることが多い(Gower 距離,Gower, 1971).

全てのレコードの組合せに対して計算した距離 D_{ij} の値を基に,式(2.1)を用いてマッチング確率 P_{ij} を推定し,その値が最も大きいレコードと結合することにより,統計的マッチングを行うことができる.多項ロジットモデルに基づく統計的マッチングのイメージを示したものが,以下の図 2 である.

2.2 距離のウエイトの推定方法

次に,距離のウエイト β_k をデータから推定する方法について述べる.式(2.1)を基に,対数尤度関数 L を,以下の式(2.5)のように構成することができる.

$$l(\boldsymbol{\beta}) = \log \left(\prod_{i=1}^{M} \prod_{j=1}^{N} P_{ij}(\boldsymbol{\beta})^{\delta_{ij}} \right) = \sum_{i=1}^{M} \sum_{j=1}^{N} \delta_{ij} \log(P_{ij}(\boldsymbol{\beta})) \tag{2.5}$$

$\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ は,距離に含まれるウエイトを表している.δ_{ij} は,データ A(Donner)のレコード i と,データ B(Recipient)のレコード j が同一の場合に 1,それ以外の場合に 0 となる変数である.対数尤度関数 L を $\boldsymbol{\beta}$ に関して最大化することにより,ウエイトの最尤推定値 $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)$ が得られる.

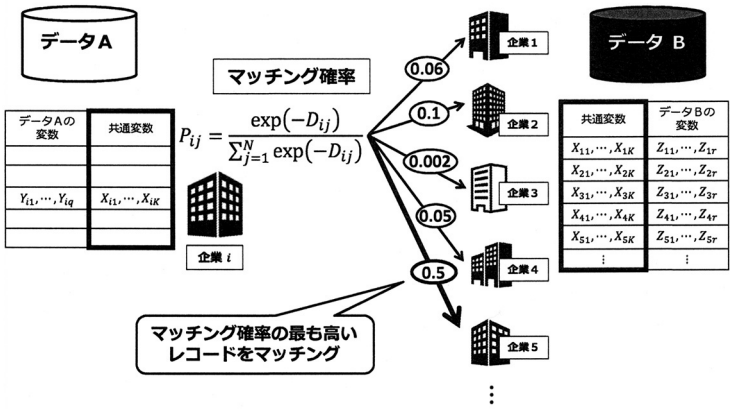


図 2. 多項ロジットモデルに基づく統計的マッチングのイメージ.

前節では，多項ロジットモデルの枠組みに基づき，最尤法により最適な距離のウエイトを推定する方法について述べた．ここでは，距離を用いた統計的マッチングにおいて，距離のウエイトを何らかの形で推定している先行研究とその課題，及びそれらと比較した場合の提案手法の利点について述べる．

2.3　実データを用いた多項ロジットモデルの推定結果

本稿では，平成 24 年経済センサス-活動調査のミクロデータ（経済センサスミクロデータ）及び帝国データバンクの企業データ（帝国データバンクデータ）を対象として分析を行う．推定の対象地域については，3 つの県のデータ（県 A～C）を対象に推定を行う．データの概要については以下のとおり．

経済センサスミクロデータ

- 分析には，平成 24 年調査の結果を使用（調査の期日は平成 24 年 2 月 1 日現在）．
- 一部の変数に関して欠測値が含まれていることから，MICE（Buuren, 2012）の手法に基づき，事前に欠測値を補完（R のパッケージ mice を使用）．

帝国データバンクデータ

- 「COSMOSII」企業概要ファイル・レイアウト C を使用（平成 24 年 2 月時点）．
- 資本金 300 万円以上 5,000 万円未満の企業を対象．
- 完全照合できなかったレコードについては，分析対象から除外．

なお，経済センサスミクロデータと帝国データバンクデータでは，変数の定義などに違いがあり，以下に示すように，各種の調整を行っている．

従業者数及び従業員数

- 帝国データバンクデータの従業員数には，パート・アルバイトを含む場合と含まない場合が混在していると想定されるデータが見受けられ，経済センサスミクロデータの従業者数については，どちらの情報も得られる．
- そこで，上記の 2 つの場合に関して距離を計算し，このうち小さい方を従業者数・従業員数に関する距離とする．

産業
- 帝国データバンクデータの産業分類を，日本標準産業分類を基に組み替えて使用．
- 「S：公務(他に分類されないものを除く)」及び「T：分類不能の産業」は対象外．

開設年
- 帝国データバンクデータでは，開設年が年単位で記録されているが，平成 24 年経済センサス-活動調査では，開設時期がカテゴリ変数となっている．
- そこで，開設年を(1) 1984 年以前，(2) 1985 年〜1994 年，(3) 1995 年〜2004 年，(4) 2005 年以降，の 4 つの時期に区分して，カテゴリ変数として使用．

上記のデータを名称，所在地により完全照合し，各データから 2/3 のレコードを無作為抽出して学習用データとし，残りの 1/3 のレコードを，後述する欠測値補完の検証用のテストデータとした．地域ごとの各データのレコード数については，以下の表 1 のとおり．

経済センサスミクロデータおよび帝国データバンクデータの両方のデータに共通に含まれる変数(共通変数)は，以下の表 2 に示した 6 種類である．

このほかに，売上高も変数として含まれているが，これは後述するように，欠測値補完の検証をする際の変数として用いることとしている．

データには連続変数とカテゴリ変数が含まれていることから，距離関数としては，式(2.2)及び式(2.3)を組み合わせた形のものを用いることとする．その際に，高部・山下 (2021)，Takabe and Yamashita (2020)及び高部・山下 (2018)の結果に基づき，モデルの推定精度を向上させるため，距離の対数変換値(以下の式(2.6))を用いた多項ロジットモデルについて推定を行う．

$$D_{ij} = \sum_{k=1}^{K} \beta_k \log(|X_{ik} - X_{jk}| + 1) \tag{2.6}$$

これらの距離に基づく多項ロジットモデルの推定結果を示したものが，表 3 である．推定結果を見ると，全ての地域で，ほぼ全ての変数について 0.1 パーセントの有意水準で有意となっている．

なお，推定した多項ロジットモデルに基づく統計的マッチングの精度についても重要な観点

表 1．各地域のデータのサイズ．

	地域 A	地域 B	地域 C
(1) 学習用データ	14,744	13,289	18,122
経済センサスミクロデータ	5,095	4,184	5,539
帝国データバンクデータ	9,649	9,105	12,583
(2) テストデータ	7,288	6,646	9,066
経済センサスミクロデータ	2,463	2094	2,774
帝国データバンクデータ	4,825	4,552	6,292
合計 ((1) + (2))	22,032	19,935	27,188

表 2．分析に使用する変数．

変数	単位・区分	種類
従業者数（従業員数）	連続変数	人
資本金額	連続変数	万円
産業分類（大分類）	カテゴリ変数	18 区分
開設年	カテゴリ変数	4 区分
地域	カテゴリ変数	市又は郡に応じた区分
経営組織	カテゴリ変数	株式会社，有限会社又は経営組織不明

表 3. 多項ロジットモデルの推定結果.

変数	地域 A		地域 B		地域 C	
従業員数	1.277	***	1.306	***	1.222	***
	(57.953)		(48.380)		(58.394)	
資本金額	0.816	***	0.859	***	0.775	***
	(76.089)		(58.764)		(69.941)	
産業	3.541	***	3.571	***	3.547	***
	(76.553)		(66.119)		(80.612)	
開設年	1.352	***	1.427	***	1.478	***
	(39.844)		(36.382)		(44.017)	
地域（市・郡）	9.709	***	9.148	***	8.217	***
	(16.787)		(22.354)		(36.774)	
経営組織（株式会社・有限会社）	3.845	***	4.493	***	4.019	***
	(29.731)		(21.094)		(27.863)	
対数尤度	-17221		-11517		-20559	
McFadden の疑似決定係数	0.632		0.698		0.606	

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, . $p < 0.01$
( ) は各回帰係数の t 値である.

ではあるものの，今回は，当該モデルを欠測値補完に応用した場合の効果に焦点を当てていることから，マッチングの精度の検証については省略する．様々な距離関数を用いた多項ロジットモデルに基づく統計的マッチングの精度の比較検証に関しては，高部・山下 (2021)，Takabe and Yamashita (2020) 及び高部・山下 (2018) を参照.

3. マッチング確率の欠測値補完への応用

本研究では，多項ロジットモデルに基づく統計的マッチングの手法を欠測値補完に応用する方法について検討する．具体的には，帝国データバンクデータの売上高が全て欠測している場合を想定し，その値を，経済センサス-活動調査データの類似したレコードにおける売上高の値で補完する場合を想定する．なお，本研究では，欠測値があるデータ (レシピエント) を帝国データバンクデータとし，欠測値補完に使用されるデータ (ドナー) を経済センサス-活動調査データとしている[1)].

表 3 における多項ロジットモデルの推定結果を用いて，地域ごとのテストデータにおける欠測値補完の精度を検証する．具体的には，地域ごとにマッチング確率を算出し，レシピエントにおいて最もマッチング確率が大きいレコードの売上高の値を，対応するドナーの売上高の代入値とする方法を検討する．その際に，マッチング確率を用いて，経済センサスの全てのレコードにおける売上高の値を加重平均した結果による欠測値補完ついても検討する．具体的には，ドナー側のデータにおける i 番目のレコードの売上高 (欠測値) を y_i $(1 \leq i \leq M)$ とし，レシピエント側のレコードの売上高の値 z_j $(1 \leq j \leq N)$ とした場合に，マッチング確率 P_{ij} による加重平均により，$y_i = \sum_{j=1}^{N} P_{ij} z_j$ として欠測値補完を行う方法を検討する.

また，これらの提案手法に対する比較手法として，従来の統計的マッチングの研究においてよく用いられている Gower 距離 (式 (2.2) において β_k を X_{ik}, X_{jk} のレンジ (最大値-最小値) の逆数とし，式 (2.3) において β_k を 1 としたもの，Gower, 1971) に基づく単純な最近隣法 (レシピエントにおいて最も Gower 距離が小さいレコードの売上高の値を代入値とする方法) による結果についても検討する.

欠測値補完の精度については，帝国データバンクデータの売上高の値を真値として，以下の平均絶対誤差率 (MAPE: Mean Absolute Percentage Error) を用いて評価する.

$$MAPE = \frac{100}{n} \sum_{i=1}^{M} \left| \frac{\hat{y}_i - y_i}{y_i} \right| \tag{3.1}$$

表 4. マッチング手法ごとの欠測値補完の精度(MAPE).

マッチング手法	地域 A	地域 B	地域 C
Gower 距離に基づく最近隣法	129.06	183.09	111.85
マッチング確率に基づく最近隣法	114.74	147.08	105.57
マッチング確率に基づく加重平均	107.43	133.44	104.78

ここで，y_i は真の売上高，$\hat{y}_i$ は，ドナーとして選択されたレコードの売上高を示している．各地域における欠測値補完の手法ごとに，平均絶対誤差率の結果を示したものが表 4 である．

上記の結果を見ると，いずれの地域においても，Gower 距離に基づく最近隣法よりも，マッチング確率に基づく手法の方が，平均絶対誤差率でみた場合の精度が向上している．また，マッチング確率による加重平均値を用いた手法が，最も精度がよい結果となっている．データから推定された最適な距離に基づくマッチング確率を用いることにより，欠測値補完の精度が向上することが示された．Gower 距離のようにウエイトをデータに応じて調整していない距離を用いるよりも，マッチング確率のようにデータから最適な距離のウエイトを算出している指標を用いる方が，欠測値補完の精度は向上すると考えられる．また，マッチング確率に基づく売上高の加重平均を用いることにより，はずれ値の影響が緩和されると考えられることから，マッチング確率の加重平均値を用いた方が，欠測値補完の精度は向上すると考えられる．こうした結果が，表 4 から確認できる．

なお，地域ごとに平均絶対誤差率の水準に違いがあることが確認される．これについて，平均をとる前の事業所ごとの絶対誤差率を地域ごとにプロットした結果が，図 3，図 4 及び図 5 である．これらの図をみると，特に地域 B において極端に大きな絶対誤差率を持つ事業所が多いことがわかる．このように，2 つのデータの時点などの違いにより，売上高の結果が大きく異なるにもかかわらず従業差数や資本金などの結果が近い値となっている事業所が多く存在する地域においては，他の地域よりも平均絶対誤差率の水準が全体的に大きくなることがわかる．このような場合，地域間の平均絶対誤差率の比較はできないものの，地域内における平均絶対誤差率を基にした各手法の比較・評価は可能である．

4. おわりに

本研究では，高部・山下 (2021)，Takabe and Yamashita (2020) 及び高部・山下 (2018) で提案された多項ロジットモデルに基づく統計的マッチングの手法を基に，マッチング確率を用いた加重平均により欠測値補完を行う方法について提案し，実データを基にその検証を行った．検証の結果，いずれの地域においても，Gower 距離に基づく最近隣法よりも，マッチング確率に基づく最近隣法の方が，平均絶対誤差率でみた場合の欠測値補完の精度が向上しており，データから最適なウエイトを計算した距離関数に基づくマッチング確率は，欠測値補完を行う上でも有用であることが示唆される結果となった．データの構造が大きく変わらないことを想定すれば，一度，距離関数(多項ロジットモデル)を推定しておけば，その結果を用いることで，外部データからのコールドデック法に基づく欠測値補完を効率的に行うことが可能となる．ただし，マッチング確率による加重平均値を用いた手法については，一部の地域で補完の精度が低下することが明らかとなった．売上高は一般的に外れ値を持つことから，加重平均によって補完値の精度が低下した可能性もあり，例えば加重平均を計算する範囲をマッチング確率が上位のレコードに限定するなど，さらに工夫や検証を重ねる必要があると考えられる．なお，本研究では，帝国データバンクデータの売上高が全て欠測している場合を想定しているが，本稿で提案した手法は売上高などの特定の属性値が全てのレコードにおいて欠測している場合に限定されるものではなく，売上高などの属性値が一部のレコードにおいて欠測している場合にも

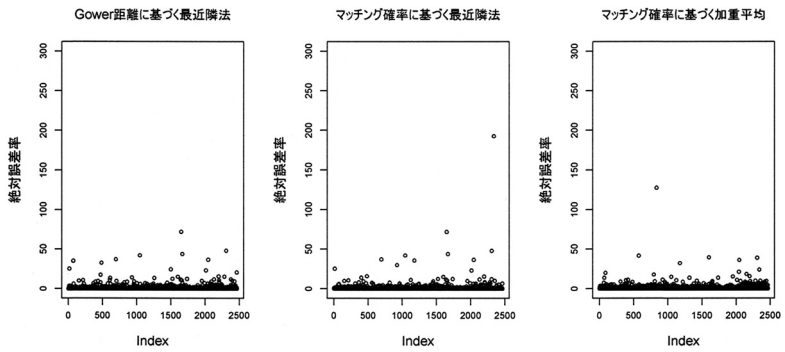


図 3. 地域 A の欠測値補完手法ごとの各事業所の絶対相対誤差.

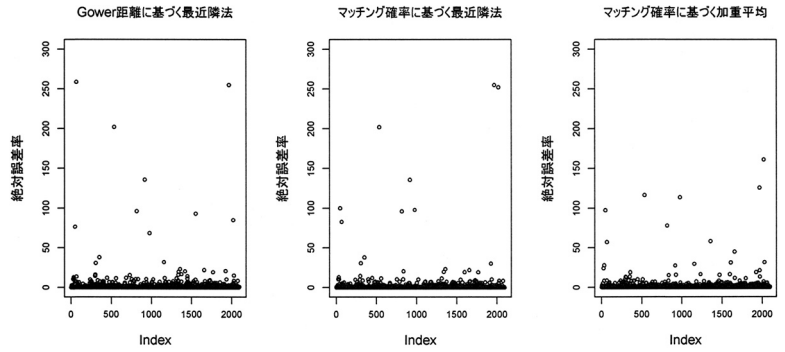


図 4. 地域 B の欠測値補完手法ごとの各事業所の絶対相対誤差.

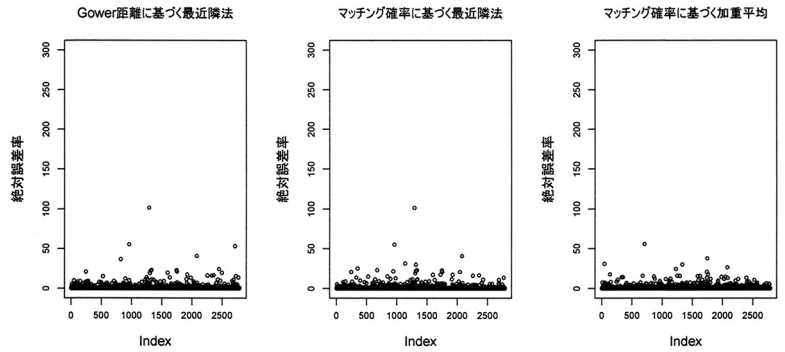


図 5. 地域 C の欠測値補完手法ごとの各事業所の絶対相対誤差.

適用可能である.

今後の課題として，今回のデータを用いて構築したモデルを全く別のデータに適用し，その精度を検証することが考えられる．本研究における手法は，企業データだけでなく，世帯・個人などが対象のデータにも適用することが可能であることから，企業データ以外の様々なデータに対しても本研究における手法を適用することにより，提案手法の有効性を確認していくことが必要であると考える．また，今回は売上高という単一の連続変数について検証を行ったが，複数の変数を対象とする場合の検討や，地域・産業などのカテゴリ変数の欠測値補完をど

のように考えるのかという点も，今後の課題である．

令和 5 年 3 月に決定された政府の第 Ⅳ 期公的統計基本計画において，公的統計ミクロデータの利活用に関しては，調査票情報の利活用に係る安全及び国民の安心確保を図りつつ，学術研究分野からの要請への対応を図っていくため，調査票情報のオンサイト利用を始めとする更なる利便性向上を図ることとされている．こうした状況を鑑み，公的統計のミクロデータや社会調査の個票データなど，レコード単位の様々な統計調査のミクロデータの利活用が進められていく中で，欠測値の補完は今後も一層，重要なテーマになっていくものと考えられ，本稿で提案した手法も含め，今後も，継続的な手法の開発・改善を続けていく必要があると考える．

注．

[1)] 本研究では，帝国データバンクデータをレシピエントとし，経済センサス-活動調査データをドナーとしているが，その役割を逆にした(ドナーを帝国データバンクデータとし，レシピエントを経済センサス-活動調査データとした)モデルを想定することも可能である．高部・山下 (2021)では，ドナーとレシピエントの役割を入れ替えたモデルも合わせて活用することにより，統計的マッチングの精度を向上させる方法について分析している．

参 考 文 献

Buuren, S. (2012). *Flexible Imputation of Missing Data*, CRC press, Boca Raton.

Christen, P. (2012). *Data Matching*: *Concepts and Techniques for Record Linkage*, *Entity Resolution*, *and Duplicate Detection*, Springer, Berlin.

D'Orazio, M., Di Zio, M. and Scanu, M. (2006). *Statistical Matching*: *Theory and Practice*, Wiley, Chichester.

De Waal, T., Pannekoek, J. and Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*, John Wiley & Sons, New Jersey.

Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage, *Journal of the American Statistical Association*, **64**, 1183–1210.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties, *Biometrics*, **27**, 623–637.

Harron, K., Goldstein, H. and Dibben, C. (2015). *Methodological Developments in Data Linkage*, Wiley, Chichester.

Herzog, T. N., Scheuren, F. J. and Winkler, W. E. (2007). *Data Quality and Record Linkage Techniques*, Springer, Berlin.

星野崇宏 (2009).『調査観察データの統計科学：因果推論・選択バイアス・データ融合』, 岩波書店, 東京.

岩崎学 (2002).『不完全データの統計解析』, エコノミスト社, 東京.

北原昌嗣, 寺垣内雅子 (2023). 諸外国の国勢調査におけるインピュテーション方法, 統計研究彙報, **80**, 137–162.

栗原由紀子 (2015). 統計的マッチングにおける推定精度とキー変数選択の効果：法人企業統計調査ミクロデータを対象として, 統計学, **108**, 1–15.

村田磨理子, 伊藤伸介 (2016). 事業所・企業系のミクロデータを用いたデータリンケージの可能性：賃金構造基本統計調査を例に, 統計学, **110**, 1–17.

Newcombe, H. B., Kennedy, J. M., Axford, S. J. and James, A. P. (1959). Automatic linkage of vital records, *Science*, **130**, 954–959.

日本統計学会 (2022).『経済統計の実際』, 共立出版, 東京.

日本統計学会 (2023).『調査の実施とデータの分析』, 共立出版, 東京.

Rässler, S. (2002), *Statistical Matching*, Springer, New York.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations, *Journal of Business and Economic Statistics*, **4**, 87–94.

坂下信之 (2018). 諸外国における統計調査の欠測値補完方法の動向と手法の体系について, 総務省統計研究研修所リサーチペーパー, **43**, 総務省統計研究研修所, 東京.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward, *Statistical Science*, **25**, 1–21.

高部勲, 山下智志 (2018). 多項ロジットモデルを用いた新たな統計的マッチング手法の提案, 統計学, **115**, 1–16.

高部勲, 山下智志 (2021). 企業データの統計的マッチング及びその精度改善, 統計研究彙報, **78**, 21–40.

Takabe, I. and Yamashita, S. (2020). New statistical matching methods using multinomial logistic regression model, *Advanced Studies in Classification and Data Science* (eds. T. Imaizumi, A. Okada, S. Miyamoto, F. Sakaori, Y. Yamamoto and M. Vichi), 265–274, Springer, Singapore.

高橋将宜 (2022).『統計的因果推論の理論と実装』, 共立出版, 東京.

高橋将宜, 渡辺美智子 (2017).『欠測データ処理』, 共立出版, 東京.

高井啓二, 星野崇宏, 野間久史 (2016).『欠測データの統計科学』, 岩波書店, 東京.

統計委員会担当室 (2013).『統計データの補完推計に関する調査報告書』, 野村総合研究所, 東京.

山口幸三 (2014).『失われし 20 年における世帯変動と就業異動：1991 年〜2010 年のミクロ統計データの静態・動態リンケージにもとづく分析』, 日本統計協会, 東京.

美添泰人 (2005). 統計的照合手法の基礎理論と最近の適用例, 青山経済論集, **56**, 43–71.

Missing Value Imputation Using a Statistical Matching Method Based on a Multinomial Logit Model

Isao Takabe

Faculty of Data Science, Rissho University

Statistical matching is a technique for combining different data to construct useful data. Statistical matching enables the creation of useful data without additional research or data collection and has recently been used in various fields. In this study, we introduce the method of statistical matching based on the multinomial logit model proposed in Takabe and Yamashita (2018, 2020, 2021) and also discuss the use of matching probabilities obtained as a byproduct for missing value imputation.

Key words: Statistical matching, multinomial logit model, weighted distance function, missing value imputation.

統計数理(2024)
第 72 巻 第 2 号 245–260
©2024 統計数理研究所
特集「公的統計―社会の重要な情報基盤―」
［統計ソフトウェア］

公的統計における外れ値への対処とソフトウェア
―R の MSD 法パッケージの実装について―

和田 かず美†

（受付 2024 年 1 月 9 日；改訂 2 月 10 日；採択 2 月 14 日）

要　　旨

公的統計の分野において，連続値の外れ値検出法で最も多く使用されているのはレンジチェックと呼ばれる単変量の方法であるが，調査統計の成果物が従来の統計表に加えて個別データの提供も進むにつれて，多変量な外れ値への対処の重要性も徐々に認識されつつある．本稿では，単変量の外れ値と多変量の外れ値の違いについて解説し，多変量の外れ値検出法のうち単峰で対称な楕円分布を前提とする手法の 1 つである MSD（Modified Stahel-Donoho）推定量による方法を実装した R のパッケージ RMSD 及び RMSDp とその利用方法について紹介する．

キーワード：データクリーニング，楕円分布．

1. はじめに

公的部門の統計調査データの集計処理においては，何らかのエラーや，間違いではないが他の大部分のデータとは傾向が違うデータの発生を避けることが難しい．そのような観測値は「外れ値」と呼ばれ，広義には分布の裾にあり他の大多数のデータとは異なる傾向を持つものと定義される（Eurostat, 2014）が，厳密には統計調査の各集計プロセスに応じて様々に定義される．

調査票情報を電子化し，記入ミスや誤りを検出するデータチェックのプロセスにおいては，検出すべき外れ値は誤りの可能性がある観測値が対象となる．また，例えば欠測を回帰モデルに基づき補完するプロセスにおいては，検出すべき外れ値は梃子比と Cook の距離が共に大きいデータが回帰推定及びそれに基づく補完値を歪める可能性がある観測値となるだろう．データのエラーを除外・修正し，必要に応じて欠測を補完した後の，集計値に乗率と呼ばれる標本抽出率の逆数を乗じて統計表の数値を作成する母集団推定プロセスでは，誤りではなくとも，特異な値が大きな乗率を持ち，集計結果に影響するようなデータが検出すべき外れ値となり，そのようなデータについては乗率の調整等の措置を検討することになる．統計作成プロセス毎の外れ値の処理については，Wada (2020) などに詳しい．

公的統計における伝統的な外れ値検出法は，何らかの方法で正常な値の範囲を決めるレンジチェックと呼ばれる単変量の方法である．この正常値範囲をデータ自体から導き出す方法については，野呂・和田 (2015) を参照されたいが，複数の相関のある数量項目を含む調査統計において，単変量の外れ値検出法を個々の項目に適用する場合，特定の変数で極端な値をとらない

† 総務省 統計研究研修所：〒 185–0024 東京都国分寺市泉町 2–11–16

が調査項目間の相関関係が他の大部分の観測値と異なるタイプの外れ値を検出することができない．本稿では，そのような外れ値を「関係性の外れ値」と呼ぶことにする．公的統計の集計プロセスでは，そのような外れ値はデータを層化したり，相関の高い変数との比率についてレンジチェックを適用することにより疑似的に対応することが多く，実務上多変量の外れ値検出法はあまり普及はしていない．この理由としては以下のようなことが考えられる．

- 単変量の方法と比較して技術的に難易度が上がる．
- 実データでの正解（真の外れ値）はわからない．
- 多くの検出法が提案されているが，ベストな方法は状況により変わりうる．
- 検出された外れ値の審査や対処が単変量のものよりも困難である．

本稿では，楕円分布モデルを前提とした多変量な外れ値検出法の一つである MSD（Modified Stahel-Donoho）推定量に基づく方法とその実装パッケージについて解説し，公的統計への適用事例について併せて紹介する．これは，先に述べた多変量の方法が普及しない理由のうち，技術的難易度の緩和に貢献するものである．なお，p 次元確率変数 $\boldsymbol{x}$ が楕円分布に従うとき，その密度関数 $f(\boldsymbol{x})$ は次のように表現される．

$$f(\boldsymbol{x}) \propto g\left[(\boldsymbol{x}-\mu)^\top \Sigma^{-1}(\boldsymbol{x}-\mu)\right]$$

ここで，μ は $\boldsymbol{x}$ が従う楕円分布の平均値ベクトル，Σ は分散共分散行列とする．また，g は非負値関数で，確率変数の定義域全体での積分が有界となる．公的統計分野において，家計における所得と消費や事業所・企業における従業者数と売上高などは，対数変換などの変数変換によって対称化することで楕円分布と見なすことができる．探索的データ解析に基づく変数変換等により多変量正規分布に近づくデータも，典型的な楕円分布モデルに基づく外れ値検出法の適用範囲となる．

まず，第 2 節では本稿で取り上げる MSD 法に関する先行研究について紹介し，第 3 節でその実装アルゴリズムと並列化の方法についても解説する．第 4 節ではこれらのソフトウェアのパッケージ化とその仕様を実例を交えて紹介する．

2. 先行研究

本稿で取り上げる MSD 法は，Patak（1990）が考案した，Stahel-Donoho（SD）推定量と射影追跡（PP: Projection Pursuit）法という二つの手法の組み合わせである．SD 推定量は Stahel（1981）と Donoho（1982）がそれぞれ考案したもので，観測値を様々な方向の基底ベクトルに射影し基底ベクトル上の中心からの乖離度に応じて各観測値に付与したウェイトを用いて，平均値ベクトルと分散共分散行列のロバスト推定を行う．SD 推定量は，変数の数によらず高い破局点を持つことが知られている（Maronna and Yohai, 1995）．続く PP 法のステップでは，SD 推定量である分散共分散行列から固有値分解による主成分分析（PCA: Principal Component Analysis）を行う．PCA は，相関のある変数を含むデータセットを観測値の持つ多様性を極力保持しつつ主成分得点に変換するが，主成分同士は相関を持たないために前節で述べた「関係性の外れ値」を想定する必要がない．このため，個々の主成分に対しては単変量の外れ値検出法の考え方を適用することができる．

周知のように，第一主成分は最大の固有値を持つ固有ベクトルに観測値を射影した結果であり，分散が最も大きくなるよう設定される．第二主成分は，データから第一主成分に直交する空間への射影方向の中で，同様に分散を最大化する方向を示す．第三主成分はデータから第一・第二主成分に直交するという制約の下で，同様に分散を最大化する方向を示すベクトルが

選択される．第一主成分は多変量データの共通性を代表する尺度，通常サイズ因子と呼ばれているが，第二主成分以降は共通性よりも徐々に観測値の特異性が集約されていくため，外れ値検出に関してはむしろ有用性が高い．

このように，Patak (1990)が考案したMSD法は，SD推定量として得られたロバストな分散共分散行列にPP法を適用してさらに良い平均値ベクトルと分散共分散行列を推定し，これに基づき各観測値のマハラノビス平方距離を算出し，その数値が大きなものがF検定統計量を目安に外れ値として検出される．このMSD法は，大標本での破局点が約0.5と高く，直交変換不変(orthogonally equivariant)である (Franklin and Brodeur, 1997)．

伝統的な外れ値への対応策としては，母集団が正規分布などの特定の分布に従うことを前提に仮説検定を行う方法論や，回帰など異なるモデルを前提とするものも存在するが，ここでは観測値が平均値ベクトルと分散共分散行列で決まる楕円分布族に従うことを前提とする検出法に焦点を当てる．統計調査の集計実務への適用可能性を重視し，国連欧州経済委員会(UNECE)やEU統計局(Eurostat)などが行っている公的統計分野の取り組みであるUNSC/UNECE刊行のStatistical Data Editingシリーズや，2001〜2003年にEU域内の政府統計部局・大学及び企業の統計専門家が参加したEUREDITプロジェクト[https://www.cs.york.ac.uk/euredit/euredit-main.html]の報告書などを調査し，多変量外れ値検出法の先駆的な実用化事例であるカナダ統計局の年次卸売・小売業調査(AWRTS)に使用されたMSD法に着目した．

カナダ統計局での適用方法とその手法の詳細については，Franklin and Brodeur (1997)に詳しい．EUREDITプロジェクト報告書の一部であるBéguin and Hulliger (2003)においてこれが紹介され，改善点についていくつか指摘がなされている．和田 (2010)は，カナダ統計局の方法とEUREDITでの指摘に基づいた改良手法を統計ソフトRで実装し，実際にBéguin and Hulliger (2003)による改良が外れ値検出の性能を向上させることを確認している．

3. 実装

3.1 Modified Stahel-Donoho 推定量のアルゴリズム

大きさnでp次元のデータセット$\boldsymbol{X}$について，そのi番目の観測値を$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$とすると，Béguin and Hulliger (2003)に基づくMSD法のアルゴリズムは以下のとおり．

(1)直交基底の作成

ランダムなb個の直交基底$\boldsymbol{\nu}^{(k)}$ $(1 \leq k \leq b)$を作成するために，$b \times p \times p$個の一様乱数を生成してb個のp行p列の行列を作成，これにGram-Schmidtの直交化法を適用する．Béguin and Hulliger (2003)は，pが大きくなっても検出性能を維持するために必要な基底数bについて，pの増加に伴い指数関数的に増加すると述べたMaronna and Yohai (1995)を根拠として1変数当たり$\exp(2.1328 + 0.8023p)/p$としているので，$b = \exp(2.1328 + 0.8023p)$となる．

(2)射影と残差の標準化

直交基底行列$\boldsymbol{\nu}$のj番目の基底ベクトル$\boldsymbol{\nu}_j$ $(1 \leq j \leq p)$が張る直線上に観測値$\boldsymbol{x}_i$を射影した座標$\boldsymbol{\nu}_j^\top \boldsymbol{x}_i$を中央値(median)と中央絶対偏差(MAD: median absolute deviation)を用いて

$$r_{ij} = \frac{|\boldsymbol{\nu}_j^\top \boldsymbol{x}_i - \mathrm{med}(\boldsymbol{\nu}_j^\top \boldsymbol{x})|}{\mathrm{mad}(\boldsymbol{\nu}_j^\top \boldsymbol{x})/0.674} \tag{3.1}$$

によりロバストに標準化し，標準化残差r_{ij}を求める．ここで，medは中央値，madはMADを示す．データが標準正規分布に従うと仮定すると，MADは0.674になるため，MADを0.674で割ることにより標準偏差と基準を揃える．

(3)残差に基づく次元別ウェイトの算出

i 番目の観測値 $\boldsymbol{x}_i$ を j 番目の基底ベクトル $\boldsymbol{\nu}_j$ に射影して得られる標準化残差 r_{ij} から，

$$\tilde{r}_{ij} = \begin{cases} r_{ij} & \text{if} \quad 0 \le r_{ij} \le c, \\ c^2/r_{ij} & \text{if} \quad c \le r_{ij}, \end{cases} \qquad c = \sqrt{\chi^2_{p,0.95}} \tag{3.2}$$

により刈り込み残差 $\tilde{r}_{ij}$ を算出し，

$$w_{ij} = \tilde{r}_{ij}/r_{ij} \tag{3.3}$$

により次元別のウェイト w_{ij} を得る．

Maronna and Yohai (1995)は，定数 c について，$\boldsymbol{X}$ が p 次元の正規分布に従う場合に射影残差の二乗が漸近的に自由度 p のカイ二乗分布に従うことを根拠としている．

(4)一次ウェイトの算出

各観測値に対し，直交基底 1 組毎に得られる p 個のウェイトについて，まず

$$w_i = \prod_{j=1}^{p} w_{ij} \tag{3.4}$$

により次元別ウェイトの積和をとると，各観測値毎に基底数と同じ b 個のウェイト $w_i^{(k)}$ が得られるので，$\breve{w}_i = \min\limits_{1\le k\le b} w_i^{(k)}$ として，観測値別に b 個の中から最小のウェイトを選択し，これを一次ウェイト $\breve{w}_i$ とする．

基底ベクトルへの射影後，単変量の位置と尺度の推定量に中央値と MAD を使用し，残差 $r_i = (r_{i1}, \ldots, r_{ip})$ としてウェイト $w_i = w(r_i)$ が正で連続かつ w_i と $r_i^2 w_i$ が $r_i \le 0$ で有界であれば，Donoho (1982)に基づき，SD 推定量は有限標本での破局点の最大値である $\varepsilon_0^* = [(n-p+1)/2]/2$ を達成する (Tyler, 1994)が，式(3.3)のウェイト関数はこの条件を満たす．

(5)重み付き主成分分析

一次ウェイト $\breve{w}_i$ を用いて，

$$\begin{aligned} \hat{\boldsymbol{u}} &= \frac{\sum_{i=1}^n \breve{w}_i \boldsymbol{x}_i}{\sum_{i=1}^n \breve{w}_i} \\ \hat{\boldsymbol{V}} &= \frac{\sum_{i=1}^n (\boldsymbol{x}_i - \hat{\boldsymbol{u}})(\boldsymbol{x}_i - \hat{\boldsymbol{u}})^\top \breve{w}_i^2}{\sum_{i=1}^n \breve{w}_i^2} \end{aligned} \tag{3.5}$$

により平均値ベクトル $\boldsymbol{u}$ と分散共分散行列 $\boldsymbol{V}$ の推定値を求める．

(6)主成分射影と二次ウェイト算出

主成分分析により，p 次元の固有ベクトルが p セット得られるが，これも一組の直交基底とみなすことができる．一次ウェイト算出時と同じようにこれらの固有ベクトルに観測値を射影し，式(3.1)，(3.2)，(3.3)及び(3.4)により算出したウェイトの中から観測値別に最小の値を持つものを選び，これを二次ウェイトと呼ぶ．

(7)最終ウェイトの決定と平均値ベクトル・分散共分散行列の推定

各観測値別にさらに一次ウェイトと二次ウェイトを比較し，値が小さい方を最終ウェイトとする．この最終ウェイトを用いて再度式(3.5)により平均値ベクトルと分散共分散行列を推定する．

(8)マハラノビス距離の算出と外れ値の特定

ステップ(6)で推定した平均値ベクトルと分散共分散行列を用いて，

$$D^2(\boldsymbol{x}_i) = (\boldsymbol{x}_i - \hat{\boldsymbol{u}})^\top \hat{\boldsymbol{V}}^{-1} (\boldsymbol{x}_i - \hat{\boldsymbol{u}})$$

によりマハラノビス平方距離 $D^2(\boldsymbol{x}_i)$ を算出する．$D^2(\boldsymbol{x}_i)$ の検定統計量 F_i は，自由度 p 及び $(n-p)$ の F 分布に従い，

$$F_i = \frac{(n-p)n}{(n^2-1)p} D^2(\boldsymbol{x}_i)$$

により得られる．外れ値と判定する基準は，Franklin and Brodeur（1997）に準じて F 検定統計量の 99.9% 点をデフォルトとする．

3.2 シングルコア版の性能評価

Maronna and Yohai（1995）及び Peña and Prieto（2001）は，外れ値検出法の評価に次のようなモデルに従う乱数データを使用している．

$$(1-\alpha)N_p(\boldsymbol{0}, \boldsymbol{I}) + \alpha N_p(\delta e_1, \lambda \boldsymbol{I}), \quad \boldsymbol{I} = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$$

ここで，α は外れ値の割合，p は変数の数，λ は外れ値の分散，δ が第 1 変数上の正常値からの外れ値の距離を示す．外れ値ではない観測値の量はデータセットの大きさ $\times(1-\alpha)$，原点中心で分散共分散行列が I の p 次元標準正規分布乱数で，外れ値の量はデータセットの大きさ $\times\alpha$，平均は第一変数以外 0，分散 λ の正規分布に従う．この評価用データセットは外れ値もそれ以外の観測値もそれぞれ変数間に相関のない正規分布に従うため，関係性の外れ値は発生しない．このため，和田（2010）において実装された MSD 関数の評価には，上述のモデルの外れ値ではない観測値には相関関係を導入し，

$$(1-\alpha)N_p(\boldsymbol{0}, \boldsymbol{R}) + \alpha N_p(\delta e_1, \lambda \boldsymbol{I}), \quad \boldsymbol{R} = \begin{pmatrix} 1 & & r \\ & \ddots & \\ r & & 1 \end{pmatrix} \tag{3.6}$$

に従う乱数データセットを使用した．加えて，実データはより裾が重く歪みのある分布である場合も多いため，外れ値ではない観測値に Skew-T 分布，複合ポワソン分布及び対数正規分布に従う乱数を採用した．また，Howkins-Bradu-Kass（Hawkins et al., 1984），スイスのレストラン業（Béguin and Hulliger, 2003），Herzsprung-Russell（Rousseeuw and Leroy, 2003），Bushfire（Campbell, 1989），Stackloss（Rousseeuw and Leroy, 2003），Modified Wood Specific Gravity（Rousseeuw and Leroy, 2003）といった評価用データにより，EUREDIT 版がカナダ版より外れ値検出性能が高いことが確認された．

Wada et al.（2020）では，調査実務への適用を目的として，和田（2010）で実装された関数の EUREDIT 版と，BACON 法（Billor et al., 2000），Fast-MCD 法（Rousseeuw and Van Driessen, 1999）及び NNVE 法（Wang and Raftery, 2002）を比較し，上述の乱数データに加え，四乗根変換等を施した企業の経理項目データで性能比較を行い，裾が重く歪んだデータについて MSD 法の性能が高いことを確認している．比較に使用されたソフトウェアは表 1 のとおりで，オプション等は全てデフォルト設定を採用している．

MSD 法は，特に変数間の相関が高い場合に検出力が高いが，一方で計算負荷が上述の他の手法と比較しても高く，特に変数の数 p の増加に伴い計算負荷は指数関数的に増大する．和田（2010）で実装された MSD 関数について，大きさ 100 のデータセットについて，4GB メモリの 32bit 機では 11 変数を越えるとメモリ破綻する（Wada and Tsubaki, 2013）．

表 1．使用ソフトウェア．

方法	説明
BACON	Béguin and Hulliger (2003)で公開されている S-plus のコードを R に移植（詳細は https://github.com/kazwd2008/BEM）
Fast-MCD	R の rrcov パッケージ収録の covMcd 関数
NNVE	R の covRobust パッケージ収録の cov.nnve 関数

3.3 並列化

和田 (2010)による MSD 関数について，より高次元のデータセットへの適用を可能とするため，Wada and Tsubaki (2013)は和田 (2010)による MSD 関数の並列化実装を行った．本節以降では，和田 (2010)で実装された MSD 関数をシングルコア版と呼ぶ．並列化版は，for や apply 等の関数による繰り返し処理を並列化するための CRAN パッケージ foreach 及び doParallel を使用し，3.1 節のアルゴリズムのうち(1)から(4)までのステップの処理を小分けにして複数のスレッドに振り分けることによりメモリ破綻を防ぐ．

具体的には，シングルコア版では計算速度を重視し，b 個の基底を構成する一様乱数の直交化や $b \times p \times p$ 個の基底ベクトルへの射影と残差・ウェイトの計算をそれぞれ apply 関数を用いて一括で計算処理するが，並列化版はこれらをより小さなチャンクに分けて複数の子プロセスで実行し，1 セットの一次ウェイトに情報を集約した後に親プロセスに戻す．このため，同じデータセットにシングルコア版と並列化版関数を実行する場合，コア数が少なければ計算速度はシングルコア版の方が速い．

並列化版の性能評価は，シングルコア版と同様にモデル(3.6)に基づく正規分布に従う乱数データに適用し，大きさ 100，20 変数のデータセットにおいて，シングルコア版による少ない変数のデータセットでの結果と比較して外れ値検出性能が落ちないことが確認された．加えて，独立行政法人統計センターが公開していた 2004 年全国消費実態調査を模した合成データである疑似ミクロデータ(現在は疑似ミクロデータの公開は終了し，一般用ミクロデータが公開されている)への適用事例も紹介された．テストに使用されたのは EPSON Pro4700 で，CPU は Core™ i5(3.33GHz)，メモリ 4GB で 32bit 版の Windows 7 Pro である．約 2 万レコードの疑似ミクロデータについて，並列化版の処理は 8 変数では 55 秒，10 変数で 150 秒を要した．

4. パッケージ化

シングルコア版も並列化版も，平均値ベクトルと分散共分散行列をロバスト推定する関数コードの形でそれぞれ和田 (2010)及び Wada and Tsubaki (2013)で公開されているが，これに外れ値判定プロセスを追加してパッケージ化し，2023 年 11 月に CRAN の審査を経て公開された．本節ではその仕様と利用方法について解説する．

4.1 シングルコア版

4.1.1 仕様

パッケージ RMSD に収録された RMSD 関数は，Béguin and Hulliger (2003)に基づく MSD 法による多変量外れ値検出の関数で，並列化版の RMSDp::RMSDp() と比較すると高速だが，高次元データへの適用はメモリ破綻を起こしやすい．引数と戻り値をそれぞれ表 2 及び表 3 に示す．

引数は，inp に行が各観測値で列が変数の行列形式で外れ値を検出する欠測のないデータを指定する．nb は基底数で，$b = \exp(2.1328 + 0.8023p)$ を変更したい場合に指定する．変数の数

表 2. RMSD 関数の引数.

引数	説明
inp	入力データ（行列形式）
nb	射影を行う基底の数（デフォルト値推奨）
sd	再現性が必要な場合に指定する乱数シード (任意の整数値)
pt	外れ値検出の閾値（デフォルト 99.9% 値）

表 3. RMSD 関数の戻り値.

変数名	説明
u	平均値ベクトル
V	分散共分散行列
wt	最終ウェイト
mah	マハラノビス平方距離
FF	F 検定統計量
cf	外れ値判定の閾値（pt の値に対応するマハラノビス平方距離）
ot	外れ値判定結果（1: 外れ値ではない 2: 外れ値）

が多くメモリ破綻を起こす場合は，デフォルトよりも小さな数を指定することで破綻を防げる可能性がある一方，外れ値検出力が低下する．この関数は直交基底の作成に一様乱数を使用しているために，同じデータについて複数回実行した場合，結果が動く可能性がある．sd の指定は，基底を構成する乱数のシードに使用されるため，関数の実行結果を固定したい場合に使用する．pt は外れ値検出の閾値でデフォルトは 99.9% 値が設定されているがこれはあくまでも目安であり，正規分布よりも裾の重いデータセットの場合は外れ値判定するマハラノビス二乗距離の閾値 cf をデータに応じて調整し，その調整値よりも大きなマハラノビス平方距離 mah のものを外れ値として検出し直す．また，企業の経理項目や世帯収入等の歪みの大きなデータは，Wada et al. (2020)での事例のように事前に適切な変換を施す．

戻り値は u がロバスト推定された観測値の平均値ベクトル，V が分散共分散行列である．この二つの推定値により楕円分布の形が決まる．wt は各観測値の最終ウェイトで，この値が 0 に近いほど観測値はデータ中心から外れている．mah は各観測値のデータ中心からのマハラノビス二乗距離で，これが F 検定統計量に対応する閾値 cf の値よりも大きい場合に外れ値として ot の値に 2 がセットされる．

4.1.2 利用例

ここでは，robustbase パッケージに収録された Herzsprung-Russell データセット(パッケージ内での名前は starsCYG)について RMSD 関数を適用し，結果を外れ値判定の閾値とともに散布図に図示するコード例を紹介する．

```
install.packages("RMSD")
library(RMSD)
data(starsCYG, package="robustbase")
o1.msd <- RMSD(starsCYG)

plot(starsCYG, pch=21, col="gray", bg=c("gray80", "black")[o1.msd$ot], cex=1.5)

# 確率楕円描画
n <- nrow(starsCYG);        d <- ncol(starsCYG)
eg <- eigen(o1.msd$V)                          # 共分散行列を固有値分解
P   <- eg$vectors                              # 固有ベクトル
D  <- matrix(rep(0, d*d), ncol=d)              # 対角成分が固有値、それ以外は 0
diag(D) <- eg$values
```

```
PP <- solve(P)                                          # 固有ベクトルの逆行列
cf <- o1.msd$cf*(n^2-1)*d / ((n-d)*n)                     # 検定統計量から距離の閾値を逆算
     # o1.msd$cf は、外れ値判定に使用したカットオフ値(デフォルトで 99.9% 点)
ax1 <- sqrt(cf * eg$values)                             # 楕円の長軸・短軸の計算

nb <- 0:200                                              # データの刻み
dw <- 2 * pi / 200;  w <- dw * nb
XA1 <- ax1[1] * cos(w);    XA2 <- ax1[2] * sin(w)
XX1 <- t(t(PP) %*% t(cbind(XA1, XA2)) + o1.msd$u)    # 有向楕円に
lines(XX1, col="gray60", lwd=3, lty=2)
points(o1.msd$u[1], o1.msd$u[2], pch=18, cex=2)  # データ中心
```

図 1 の菱形の点は推定されたデータ中心(平均値ベクトル)，外れ値の目安としている F 検定統計量の 99.9% 点を灰色の点線による楕円で示している．

4.2 並列化版

パッケージ RMSDp に収録された RMSDp 関数は，シングルコア版である RMSD::RMSD() を並列化したもので，シングルコア版よりも変数の多いデータに適用できるが，シングルコア版がメモリ破綻せず動く場合はシングルコア版の方が処理時間が短い．

関数 RNSDp の入力パラメータと出力はそれぞれ表 4 及び表 5 のとおり．RMSD::RMSD() の引数に cores と dv が追加されている．core は，関数が利用するコアあるいはスレッドの数を指定するもので，デフォルトでは関数を実行する PC で利用可能な全てのコアを使用する．dv は並列化のための処理のチャンクの大きさを指定するもので，デフォルト値の 10000 は一括で計算処理を行う行列の要素数を指定する．このデフォルト値は 4GB メモリの 32bit 機を基準に設定されているため，より大きなメモリのある PC の場合は値を増やすと処理速度が上がるが，大きくしすぎるとメモリ破綻する．戻り値はシングルコア版に準拠する．

4.2.1 利用例

ここでは，(独)統計センターが公開している一般用ミクロデータから平成 21 年全国消費実態調査の詳細品目データを使用して，RMSDp パッケージを利用した外れ値検出を行う．一般用ミクロデータは，集計表から調査の個別データを模して乱数により疑似的に作成されたもので，全国消費実態調査の詳細品目データは二人以上の世帯約 5 万世帯についての家計の収入・支出などの 430305 項目を収録し，調査の概要や符号表とともに公開されている．

データチェックを行う場合も，Wada (2020)で紹介されている個人企業経済調査のように補完を行う場合でも，全てのレコードについて全数量項目をまとめて外れ値検出するわけではな

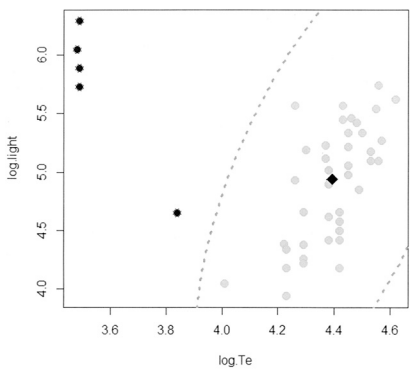


図 1. Herzsprung-Russell データセット．

表 4. RMSDp 関数の引数.

引数	説明
inp	入力データ（行列形式で 1 行が 1 つの観測値）
cores	利用するコア（スレッド）数
nb	射影を行う基底の数（デフォルト値推奨）
sd	再現性が必要な場合に指定する乱数シード (任意の整数値)
pt	外れ値検出の閾値（デフォルト 99.9% 値）
dv	チャンクの大きさ（デフォルトは最大要素数 10000）

表 5. RMSDp 関数の戻り値.

変数名	説明
u	平均値ベクトル
V	分散共分散行列
wt	最終ウェイト
mah	各観測地のマハラノビス平方距離
cf	外れ値判定の閾値（pt の値に対応するマハラノビス平方距離）
ot	外れ値判定結果（1: 外れ値ではない 2: 外れ値）

表 6. 飲料項目と変数名.

項目名	変数名	項目名	変数名
緑茶	E080	紅茶	E081
他の茶葉	E082	茶飲料	E083
コーヒー	E085	コーヒー飲料	E086
ココア・ココア飲料	E087	果実・野菜ジュース	E089
炭酸飲料	E090	乳酸菌飲料	E091
乳飲料	E092	ミネラルウォーター	E093
他の飲料のその他	E094	清酒	E096
焼ちゅう	E097	ビール	E098
ウイスキー	E099	ワイン	E100
発泡酒・ビール風アルコール飲料	E101	他の酒	E102

い．世帯調査であれば世帯の属性によりできるだけグループ内の類似性が高くグループ間の相違が大きくなるようにグループ分けを行う．このデータセットの例では，調査の対象となる世帯の収入や支出に影響があると思われる属性でグループ化し，世帯主が就業者で年齢階級が 30 歳未満，かつ世帯人員が 2 人という属性を持つ世帯に着目する．また項目について，原則として相関関係があるものを選択するが，ここでは表 6 に示す飲料の 20 項目を対象とする．

これらの項目は全て金額なので，適切な変換を施して楕円分布に近づける必要がある．どの程度の変換とするかは，散布図行列の目視結果や Box-Cox 変換（Box and Cox, 1964）のパラメータ推定結果，検出外れ値数などを勘案して決定する．Box-Cox 変換のパラメータ推定値は，0 であれば対数，0.25 で四乗根，0.5 で平方根変換ということになるが，外れ値の影響を受けやすいパラメトリックな方法であるため，極端なデータが混入している場合により小さな推定値になる傾向がある．

Box-Cox 変換のパラメータ推定結果を表 7，RMSDp パッケージを用いた MSD 法による外れ値検出結果を表 8，QQ プロットを図 2 に示す．外れ値検出結果から対数変換が最も検出数が少ないが，Box-Cox 変換のパラメータ推定結果からは少なくとも対数変換は適切ではないことがわかる．次に，QQ プロットを見ると，変換なしあるいは平方根が適切に見えるが，変換な

表 7. Box-Cox 変換のパラメータ推定結果.

E080	E081	E082	E083	E085
0.12532751	0.07193657	0.11722085	0.20877778	0.13916805
E086	E087	E089	E090	E091
0.12729565	0.06312193	-0.04562682	0.04230969	-0.00431776
E092	E093	E094	E096	E097
0.12543596	0.15176793	0.16776192	0.08195253	0.15086917
E098	E099	E100	E101	E102
-0.06567368	0.14506019	0.02271221	-0.08031569	0.11825778

表 8. 検出された外れ値の数.

データ変換	正常値	外れ値
変換なし	247	58
二乗根変換	269	36
四乗根変換	280	25
対数変換	282	23

しの場合外れ値がかなり多く検出されることと，散布図行列を比較すると平方根変換を施した方がより楕円分布に近いため，平方根変換を採用する．ただし，裾の重い分布なので外れ値と判定するための閾値を広げるか，あるいは QQ プロットで検出したい個数を決めてマハラノビス距離の大きい順に外れ値判定する．

この事例についてのサンプルコードを以下に示す．コードを流す前に，(独)統計センターのサイト[https://www.nstac.go.jp/use/archives/ippan-microdata]からデータファイル ippan_2009zensho_z_dataset.csv と符号表 ippan_2009zensho_s.xls をダウンロードし，符号表を加工する．冒頭 7 行を削除し，行番号・項目名・階層・項目番号・変数名以外の列を削除，さらに変数名が空欄の行を削除して csv 形式で ippan_items.csv という名前で保存する．

全国消費実態調査の品目項目はレベル 5 までの階層構造で，内訳を持たない品目だけを選ぶために加工した符号表の情報を使用する．具体的には，次の行の品目レベルが同じであれば内訳を持たないと判定する．

この 305 レコード 20 変数の外れ値検出には，8 コア 32GB メモリの PC(VAIO SX12，モデル VJS123C12N，Intel Core™ i7-1065G7) を使用し，計算時間は約 3 時間半であった．

```
install.packages("RMSDp")
library(RMSDp)
#
# データファイルの読込み
dat <- read.table("ippan_2009zensho_s_dataset.csv",
                header=TRUE, sep="," ,        skip=8, colClasses=c(rep("factor",7),
                rep("numeric",423)))
attach(dat)
#
# 符号表情報の読込み
items <- read.csv("ippan_items.csv", header=TRUE)
# 内訳を持たない末端の品目特定
f.lv <- rep(0, length(items$Level))   # 末端フラグ
for (i in 1:(length(items$Level)-1)) if (items$Level[i] >= items$Level[i+1]) f.lv[i] <- 1
#
# 飲料関係の 20 項目一覧
(s.itm <- items[which(f.lv==1),][79:98,])
```

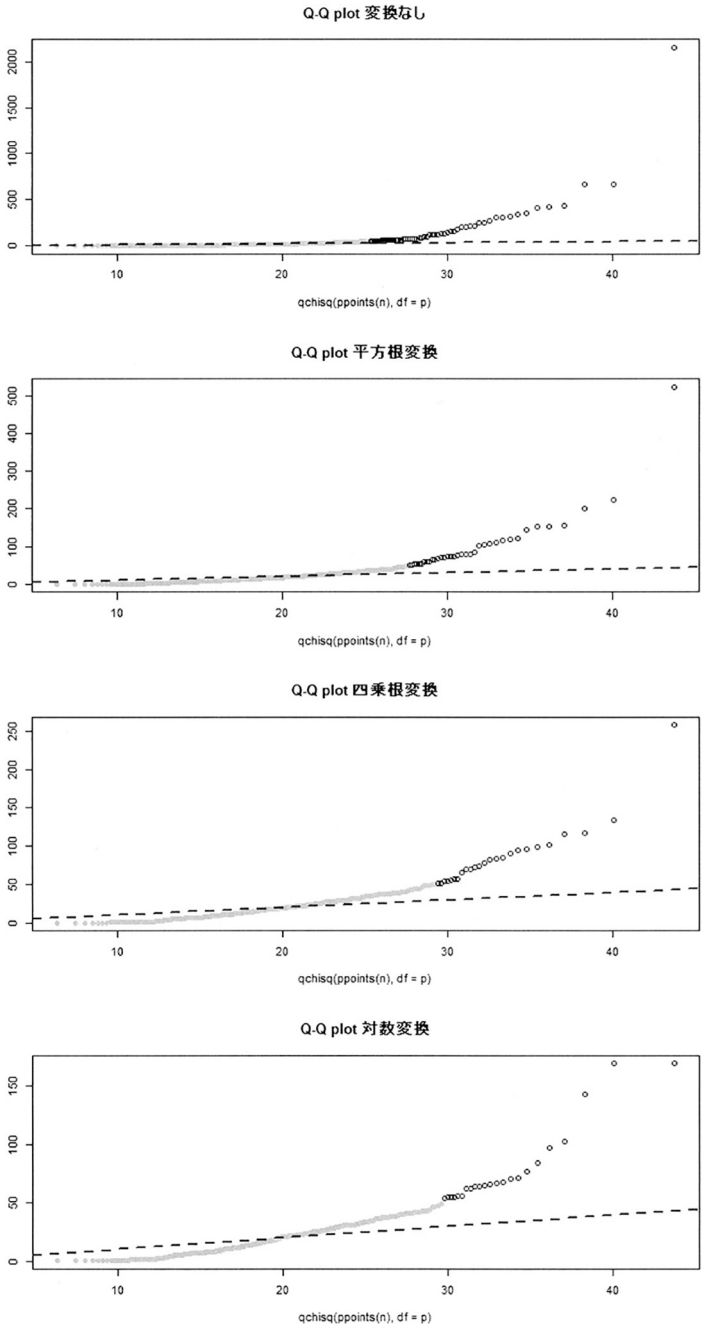


図 2. 一般用ミクロデータ(全国消費実態調査・飲料)QQ プロット.

```
#
# 世帯主が就業者で 30 歳未満 (T_Age_5s が 1) かつ世帯人員 2 人 (T_SeJinin が 2) の世帯
require(dplyr)
s1.dat <- dat |> filter(T_Age_5s == 1, T_SeJinin == 2) # レコード選択
d1 <- s1.dat[,which(f.lv==1)][,79:98]   # 項目選択
dim(d1)                # [1] 305  20
#
# 外れ値検出
(s1_time <- Sys.time())
ot1 <- RMSDp(d1)                  # 変換なし
(e1_time <- Sys.time())
#
(s2_time <- Sys.time())
ot2 <- RMSDp(sqrt(d1))           # 平方根変換
(e2_time <- Sys.time())
#
(s3_time <- Sys.time())
ot3 <- RMSDp(d1^(1/4))           # 四乗根変換
(e3_time <- Sys.time())
#
(s4_time <- Sys.time())
ot4 <- RMSDp(log10(d1))           # 常用対数変換
(e4_time <- Sys.time())
#
# [参考] 外れ値検出の進捗状況表示方法
# require(progress)  # プログレスバー表示
# pp <- 100      # 100% 刻み
# pb <- progress_bar$new(total = n)
# for (i in 1:pp) {
#    pb$tick()
#    ot1 <- RMSDp(d1)
#    Sys.sleep(1/pp)
# }
#
# 外れ値検出結果 (1:正常値; 2:外れ値)
table(ot1$ot)
table(ot2$ot)
table(ot3$ot)
table(ot4$ot)
#
# 並行座標プロット
require(MASS)
par(mfrow=c(4,1))
parcoord(d1, col=c(rgb(0,0,0, alpha=0.2), rgb(1,0,0, alpha=0.5))[ot1$ot])
parcoord(sqrt(d1), col=c(rgb(0,0,0, alpha=0.2), rgb(1,0,0, alpha=0.5))[ot2$ot])
parcoord(d1^(1/4), col=c(rgb(0,0,0, alpha=0.2), rgb(1,0,0, alpha=0.5))[ot3$ot])
parcoord(log10(d1), col=c(rgb(0,0,0, alpha=0.2), rgb(1,0,0, alpha=0.5))[ot4$ot])
#
# 散布図行列
pairs(d1, cex=0.1, pch=20, col=c(rgb(0,0,0, alpha=0.2),
                rgb(1,0,0, alpha=0.3))[ot1$ot])
pairs(sqrt(d1), cex=0.1, pch=20, col=c(rgb(0,0,0, alpha=0.2),
                rgb(1,0,0, alpha=0.3))[ot2$ot])
pairs(d1^(1/4), cex=0.1, pch=20, col=c(rgb(0,0,0, alpha=0.2),
                rgb(1,0,0, alpha=0.3))[ot3$ot])
pairs(log10(d1), cex=0.1, pch=20, col=c(rgb(0,0,0, alpha=0.2),
                rgb(1,0,0, alpha=0.3))[ot4$ot])
```

```
#
# QQ プロット
n <- nrow(d1);  p <- ncol(d1)
par(mfrow=c(2,2))
qqplot(qchisq(ppoints(n), df=p), ot1$mah, pch=19, col=ot1$ot[order(ot1$mah)],
        main="Q-Q plot 変換なし")
        abline(0, 1, col="green")
qqplot(qchisq(ppoints(n), df=p), ot2$mah, pch=19, col=ot2$ot[order(ot2$mah)],
   main="Q-Q plot 平方根変換")
        abline(0, 1, col="green")
qqplot(qchisq(ppoints(n), df=p), ot3$mah, pch=19, col=ot3$ot[order(ot3$mah)],
   main="Q-Q plot 四乗根変換")
        abline(0, 1, col="green")
qqplot(qchisq(ppoints(n), df=p), ot4$mah, pch=19, col=ot4$ot[order(ot4$mah)],
   main="Q-Q plot 対数変換")
        abline(0, 1, col="green")
#
# Box-Cox 変換のパラメータ推定
require(car)
(lambda1 <- powerTransform(d1))
#
# 外れ値判定の閾値を広げる
# F 検定統計量
FF <- ot2$mah * (n-p) * n /((n^2-1) * p)
ot2r <- rep(1, n)  # 外れ値フラグ

ot2r[which(FF > (ot2$fs*1.5))] <- 2  # 元の閾値を 1.5 倍に
table(ot2r)
```

4.3 処理時間や処理の限界について

シングルコア版は，すべての基底への射影計算を一度にメモリに展開するアルゴリズムのため，特に変数の数が増えればメモリの破綻を起こして処理不能となる．4GB メモリ搭載の 32bit 機の場合，大きさ 100 のデータセットによるシミュレーションで，11 変数が限界であった（Wada and Tsubaki, 2013），並列化版はデフォルト設定では同様の環境において処理変数を増やすために開発され，メモリ破綻を防止するために基底への射影計算を分割し，複数コアに割り振っている．

データ分割の最小単位はデータセットのサイズになるため，並列化版については理論上はメモリ展開不能なサイズのデータでない限り，時間はかかるが処理は可能である．大容量のメモリが利用可能な場合は，関数 RMSDp の引数 dv を大きくすることで処理速度の向上を図ることができる．

処理速度は，コア数が 1 桁台の汎用 PC であればシングルコア版が速いが，コア数の多いワークステーションあるいはスパコン等では並列化版がより高速となる．

コア数 4 で 16GB のメモリを搭載した 64bit のノート PC において，シングルコア版の RMSD パッケージの試算を行った結果は表 9 のとおり．併せて並列化版のテストも行い，4 コア全てを使用して 16 変数テストを行ったところ，7 時間半で正常終了した．17 変数のテストについては，フル充電の上電源も接続していたが，演算終了前にバッテリー切れで電源が落ちて計算不能となった．ただしこれはテスト機(表 10)のハードウェアに起因するものである．

5. おわりに

本稿では，破局点の高い楕円分布モデルに基づく外れ値検出法の一つである MSD 法のパッ

表 9．シングルコア版の限界（コア数 4・16GB メモリ機の場合）．

変数の数	処理時間	実行可能性
12	54 分	○
13	2 時間	○
14	4 時間半	○
15	10 時間半	○
16	30 時間	○
17	メモリ破綻	×

表 10．テスト機について．

シリーズ・モデル	Vaio Z, VJZ13A
プロセッサ	Intel Core^{TM} i7-5557U プロセッサ（TDP 28W）
メインメモリ	16GB
プラットフォーム	x86_64-w64-mingw32/x64 (64-bit)
R バージョン	4.3.2 (2023-10-31 ucrt)

ケージを紹介した．現在のところ，外れ値を判定する閾値の調整は F 検定統計量のパーセンタイル値のみであるが，今後より裾の重い分布に対応できる引数の追加など，より使いやすいものになるよう改善していきたい．

ここで紹介した MSD 法のシングルコア版は，2019 年個人企業経済調査から，相関のある経理項目の補完プロセスについて，最近隣ホットデック法の補完値候補となるドナーデータセットから他の大部分と傾向の違う特異な観測値を除外するために使用されている．補完プロセスはデータチェック後に行われ，外れ値検出対象となるデータは誤り等を含まないクリーンデータであるため，検出された観測値は，審査や修正を行うことなくそのまま集計に使用されるが，欠測の補完値として使用されない．この実用化例では，冒頭で紹介した多変量外れ値検出法の利用が進まない理由の三番目への対応として，複数の候補からベストな方法をシミュレーションにより選択している．さらに四番目の理由に関連して，クリーンデータへの適用であるため検出された外れ値について審査をする必要がない．

謝　辞

本研究について，統計数理研究所の椿広計所長には前職の筑波大学ビジネス科学研究科から長年に亘るご指導をいただき，心より感謝の意を表します．

参 考 文 献

Béguin, C. and Hulliger, B. (2003). Robust multivariate outlier detection and imputation with incomplete survey data, https://www.cs.york.ac.uk/euredit/results/Results/Robust/Part%20C.zip（最終アクセス日 2024 年 3 月 5 日）.

Billor, N., Hadi, A. S. and Velleman, P. F. (2000). BACON: Blocked adaptive computationally efficient outlier nominators, *Computational Statistics & Data Analysis*, **34**(3), 279–298.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations, *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **26**(2), 211–243.

Campbell, N. (1989). Bushfire mapping using noaa avhrr data, Technical Report, Commonwealth Scientific and Industrial Research Organisations (CSIRO).

Donoho, D. L. (1982). Breakdown properties of multivariate location estimators, Ph.D. Qualifying Paper, Department of Statistics, Harvard University, Boston.

Eurostat (2014). Statistics Explained [Glossary:Outlier], http://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Outlier（最終アクセス日 2024 年 3 月 5 日）.

Franklin, S. and Brodeur, M. (1997). A practical application of a robust multivariate outlier detection method, *Proceedings of the Survey Research Methods Section*, *American Statistical Association*, 186–191.

Hawkins, D. M., Bradu, D. and Kass, G. V. (1984). Location of several outliers in multiple-regression data using elemental sets, *Technometrics*, **26**, 197–208.

Maronna, R. A. and Yohai, V. J. (1995). The behavior of the Stahel-Donoho robust multivariate estimator, *Journal of the American Statistical Association*, **90**(429), 330–341.

野呂竜夫, 和田かず美 (2015). 統計実務におけるレンジチェックのための外れ値検出方法, 統計研究彙報, **72**, 41–54.

Patak, Z. (1990). Robust principal component analysis via projection pursuit, Ph.D. Thesis, University of British Columbia.

Peña, D. and Prieto, F. J. (2001). Multivariate outlier detection and robust covariance matrix estimation, *Technometrics*, **43**(3), 286–300.

Rousseeuw, P. J. and Leroy, A. M. (2003). *Robust Regression and Outlier Detection*, John Wiley & sons, New Jersey, USA.

Rousseeuw, P. J. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator, *Technometrics*, **41**(3), 212–223.

Stahel, W. A. (1981). Breakdown of covariance estimators, Research Report, No.31, Fachgruppe für Statistik, Eidgenössische Technische Hochschule (ETH), Zürich.

Tyler, D. E. (1994). Finite sample breakdown points of projection based multivariate location and scatter statistics, *The Annals of Statistics*, **22**(2), 1024–1044.

和田かず美 (2010). 多変量外れ値の検出〜MSD 法とその改良手法について〜, 統計研究彙報, **67**, 89–157.

Wada, K. (2020). Outliers in official statistics, *Japanese Journal of Statistics and Data Science*, **3**(2), 669–691.

Wada, K. and Tsubaki, H. (2013). Parallel computation of modified Stahel-Donoho estimators for multivariate outlier detection, *2013 IEEE International Conference on Cloud Computing and Big Data*, 304–311.

Wada, K., Kawano, M. and Tsubaki, H. (2020). Comparison of multivariate outlier detection methods for nearly elliptical distributions, *Austrian Journal of Statistics*, **49**(2), 1–17.

Wang, N. and Raftery, A. E. (2002). Nearest-neighbor variance estimation (NNVE): Robust covariance estimation via nearest-neighbor cleaning, *Journal of the American Statistical Association*, **97**(460), 994–1019.

Dealing with Outliers in Official Statistics —R Packages Implementing the MSD Estimators—

Kazumi Wada

Statistical Research and Training Institute,
Ministry of Internal Affairs and Communications (MIC)

In the field of official statistics, the univariate method known as range checking is still the predominant method for detecting outliers in continuous values; however, the importance of dealing with multivariate outliers is gradually being recognized because the products of statistical surveys increasingly include individual data, in addition to traditional statistical tables. This paper explains the difference between univariate and multivariate outliers. It also introduces the R packages RMSD and RMSDp, which implement the modified Stahel-Donoho (MSD) estimators as a multivariate outlier detection method that assumes a unimodal symmetric elliptic distribution.

Key words: Data cleaning, elliptical distribution.

統計数理(2024)
第 72 巻 第 2 号 261–271
©2024 統計数理研究所
特集「公的統計―社会の重要な情報基盤―」
［統計ソフトウェア］

公的統計を利用した教育の実践例

佐藤 正昭†

（受付 2024 年 2 月 2 日；改訂 3 月 22 日；採択 3 月 27 日）

要　　旨

公的統計を利用した統計教育の一端を紹介するとともに，その際必要となる R survey パッケージの分析利用の方法を解説する．

キーワード：ミクロデータ，公的統計調査匿名データ，統計教育，survey パッケージ．

1. はじめに

公的統計は，社会経済について知ることができる国民の情報基盤であることから，公的統計の教育利用が有意義なものとなることは論を待たないところであり，様々な形で教育利用が行われている．大別すると，e-Stat から入手できる既製品とも言える集計表の利用，ミクロデータのひとつである公的統計調査匿名データの利用の 2 つの方法がある．それらの実践例について，筆者の経験の一端を紹介し，さらなる教育利用の広がりを期待するものである．ここで，e-Stat と公的統計調査匿名データという用語について説明しておく．日本の統計が閲覧できる政府統計ポータルサイトとして，「e-Stat」と名付けられた政府統計の総合窓口 (2024)がインターネット上で運営されている. 統計法第 2 条に定義されているように，公的統計(同条第 3 項)を作成するために実施される統計調査(同条第 5 項)の調査票情報(同条第 11 項)から，匿名データ(同条第 12 項)が作成され提供されている．この匿名データについて，「公的統計調査匿名データ」と言うこととする．

2. e-Stat により提供されるデータを利用した統計教育

図 1 で示す e-Stat を通じて，各府省が作成している様々な分野の統計データ(集計表)や関連する情報を，分野，府省名，及びキーワードにより検索して，ダウンロードできるようになっている. さらには，地方公共団体などが提供している統計データの所在源情報も「統計関係リンク集」として掲載されている. e-Stat により提供されている様々な分野の統計データから主要なものを教育用標準データセットとしてパッケージ化した SSDSE というものが用意されており，また，上記で示したように様々な分野の統計調査や加工統計の集計表が利用可能で，それらを利用した統計データ分析教育の実践例が積み重ねられているので，筆者の実践例は省略するが，データを入手する手法として，web スクレイピングの技術も教育の中で取り入れることが重要ではないかと考えている．例えば，メッシュ統計は，第一次メッシュ区画別に境界データと統計データが e-Stat 上で提供されているので，広範囲にわたるデータを入手しようとすると，web スクレイピングによる効率化が求められる．web スクレイピングを実現す

† 滋賀大学 データサイエンス学部：〒 552–8522 滋賀県彦根市馬場 1–1–1

図 1．e-Stat.

る手法としては，様々なものがあるが，筆者は，Windows Subsystem for Linux（WSL），Docker Google-chrome, RSelenium パッケージの組合せで，Windows PC 上で実践している．メッシュ統計だけでなく，他のデータの web スクレイピングにも応用可能であろう．

3. 大学院における公的統計調査個票データを利用した研究指導

教育の実践例からは外れるが，関連することとして紹介する．滋賀大学においては，地方公共団体の職員が，社会人派遣枠で，このところ数名，大学院博士前期過程に入学している．筆者もこれまで 2 人の研究指導を担当した．そのうちの 1 人は，統計法第 33 条第 1 項の規定に基づき，調査票情報のオンサイト利用を活用して，研究を行い論文を執筆することができた．統計法第 33 条第 1 項の規定に基づいた調査票情報の利用は，大学院生が自治体職員であったため可能となった特殊な例ではあるが，今後は，政府が推進している，根拠に基づいた政策立案と訳される Evidence-Based Policy Making（以下，EBPM という. 行政改革推進本部, 2024 も参照）に関連する分析を含めてデータ分析を実践できる公務分野の職員の養成がさらに求められると予想されるところであり，このような形態の研究指導の需要も増加するのではないかと考えている．e-Stat からもたどることができる miripo と名付けられたミクロデータ利用ポータルサイト（2024）の情報や関連する研究会の情報に接していると，調査票情報の研究利用が推進されていると理解できるので，そのような方針とも合致すると考えている．（ただし，筆者の理解では，修士論文を執筆するためという理由だけでは，利用は承認されないので注意を.）

4. 公的統計調査匿名データを利用した統計教育，今後の課題

学部での演習といった，公的統計調査匿名データを利用した教育の意義やメリット，並びに演習環境の例については，佐藤（2023）に記したので参考にしていただきたい．また，全国でどのような教育実践例があるかは，上記で紹介した miripo（ミクロデータ利用ポータルサイト, 2024）において検索可能である．教育利用においては，筆者が利用承認を受けた経験では，総務省令に基づいた利用環境の整備が求められる．具体的には，分析に利用する PC がインターネットに接続されていないことが必要，データを USB に保存して教員が管理する必要，さらには，教員が管理する演習室や研究室を用意することが必要である．したがって，学生が自習しようとしても，できないのが現状である．教員が，分析例を示して，学生が再現実験するだけならばよいが，探究的に学生が試行錯誤しながら分析をすすめることをしようとすると，自習環境の整備も必要と考えている．具体的には，学内 LAN 上に，Linux サーバーを設置し，学生は閉じた演習室の環境において，USB を利用したシンクライアントを立ち上げ，バーチャル

デスクトップ環境で分析する自習環境が構築できないかと考えており，最近テストを開始したところである．

5. R survey パッケージを利用した匿名データの分析方法

5.1 R survey package 利用の意義

公的統計調査匿名データを集計するためには，一般には，抽出率を考慮した集計が必要となる．大学学部レベルの演習において，そのような集計を簡単に行うことができるソフトウエアはいろいろと考えられるが，経費を抑えるという観点からは，例えば，R を使うことも一案であり，R のパッケージ survey を利用することができる．もともと，survey パッケージは，標本抽出やそのためのシミュレーションなどを行うためのパッケージであることから，標本抽出における抽出率を考慮した分析機能（クロス集計や多変量解析，検定など）も実装されている．詳細は Lumley et al.（2024）を参照．ところが，既存の検索システムや，国立情報学研究所（CiNii research）において，「R survey package パッケージ」と検索してもたところ，survey パッケージを活用した分析方法一般を解説した記事は見当たらなかったので，ここに当該パッケージの分析機能の利用方法の概要を解説し，参考にしてもらうこととしたい．参考に，他のソフトウエアの例を記しておくと，SAS ソフトウエアにおいては，weight statement を指定することが可能なプロシージャーでは，抽出率を考慮した集計が可能である．

5.2 データの準備

申請手続きをすることなく入手できる「一般用ミクロデータ」を用いて解説することとする．そのために，独立行政法人統計センター（2024）から，利用規約に同意し，簡単なアンケートに答えて，平成 21 年全国消費実態調査（詳細品目）ippan_2009zensho_s.zip をダウンロードし，zip ファイルを解凍する．その中には，コードブックとも呼ばれる符号表（ippan_2009zensho_s.xls）も含まれており，符号表は 2 つのシートから構成されている．全国消費実態調査についての情報は，総務省統計局（2024）を参照されたい．以下，ippan_2009zensho_s_dataset.csv を用いて，survey パッケージの利用方法を解説していく．

5.3 データの読み込み

```
#R version 4.3.2 (2023-10-31 ucrt) -- "Eye Holes"にて動作確認
#RStudio 2023.12.0 Build 369
#survey_4.2-1.zip
setwd("c:/data/pubstat") #各自の環境に応じてフォルダ(ディレクトリ)を指定
library(tidyverse)
library(survey)

#符号表を参照して，明示的に，f 指定，d 指定を，430 の変数にそれぞれ割り当てる
num_of_factor <- "f"
for( i in seq(6)){
  num_of_factor<-paste(num_of_factor,"f" ,sep = "")
}
str_count(num_of_factor,"f")

num_of_numeric <- "d"
for( i in seq(422)){
```

```
  num_of_numeric<-
    paste(num_of_numeric,"d" ,sep = "")
}
str_count(num_of_numeric,"d")

col_format<-
  paste(num_of_factor,num_of_numeric,sep= "")
col_format
nchar(col_format)
```

出力結果例 1

```
> nchar(col_format)
[1] 430
```

430 の変数に対応して，フォーマット指定の文字列の文字数が 430 となっていることが確認できる．

```
#emEditor などを使って，一般ミクロデータのヘッダーを確認の上，読み込み
zensho_s2009 <-
  read_csv("ippan_2009zensho_s_dataset.csv",
           col_types =col_format,
           skip=8
  )
dim(zensho_s2009)
str(zensho_s2009)
#R studio おいて，430 個の変数について，意図どおり型が指定されているか確認
View(zensho_s2009)
#出力結果は，図 2 のとおりである．
```

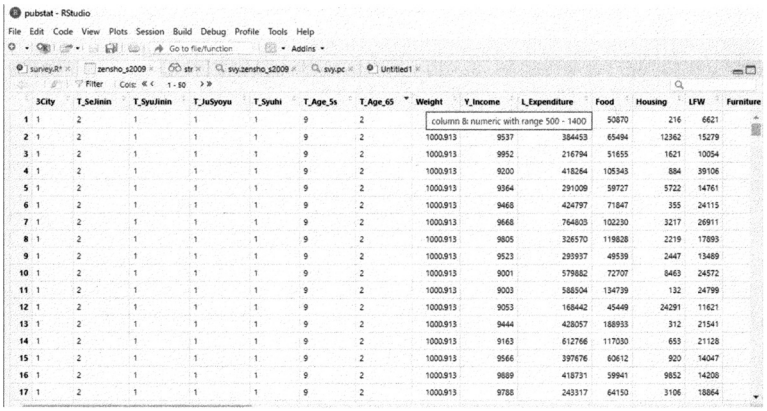
出力結果例 2

```
> dim(zensho_s2009)
[1] 45811   430
```

上記で示した符号表の 1 枚目のシート「符号表」に変数の数が 430 と記してあり，2 枚目のシート「基本数」というシートに，オブザベーション数が 45,811 と記してあるので，R の出力結果と照合し確認するとよい．なお，変数の名称は，csv ファイルのヘッダーにも記されており，符号表にも記されているので，R の実行例では，その名称をそのまま用いている．

5.4 surveydesign オブジェクトの作成

survey パッケージを利用して，分析機能を利用するためには，まず，surveydesign 型のオブジェクトを作成する必要がある．surveydesign() 関数を用いてオブジェクトを作成することになり，その際の引数としては，標本抽出を行うわけではないので，層が一つしかないと考えて ids=~1 と指定し，加えて抽出率の逆数(重み，weight)が格納されている変数の名称を指定する



	3City	T_SeJinin	T_SyuJinin	T_JuSyoyu	T_Syuhi	T_Age_5s	T_Age_65	Weight	Y_Income	L_Expenditure	Food	Housing	LFW	Furniture
1	1	2	1	1	1	9	2	1000.913	9537	384453	65494	12362	15279	
2	1	2	1	1	1	9	2	1000.913	9952	216794	51655	1621	10054	
3	1	2	1	1	1	9	2	1000.913	9200	418264	105343	884	39106	
4	1	2	1	1	1	9	2	1000.913	9364	291009	59727	5722	14761	
5	1	2	1	1	1	9	2	1000.913	9468	424797	71847	355	24115	
6	1	2	1	1	1	9	2	1000.913	9668	764803	102230	3217	26911	
7	1	2	1	1	1	9	2	1000.913	9805	326570	119828	2219	17893	
8	1	2	1	1	1	9	2	1000.913	9523	293937	49539	2447	13489	
9	1	2	1	1	1	9	2	1000.913	9001	579882	72707	8463	24572	
10	1	2	1	1	1	9	2	1000.913	9003	588504	134739	132	24799	
11	1	2	1	1	1	9	2	1000.913	9053	168442	45449	24291	11621	
12	1	2	1	1	1	9	2	1000.913	9444	428057	188933	312	21541	
13	1	2	1	1	1	9	2	1000.913	9163	612766	117030	653	21128	
14	1	2	1	1	1	9	2	1000.913	9566	397676	60612	920	14047	
15	1	2	1	1	1	9	2	1000.913	9889	418731	59941	9852	14208	
16	1	2	1	1	1	9	2	1000.913	9788	243317	64150	3106	18864	
17	1	2	1	1	1	9	2	1000.913						

図 2. 出力結果例 3. Rstudio において，ポップアップ機能で変数の型を確認.

ことがポイントとなる．下記の例では，svy.zensho_s2009 が surveydesign 型オブジェクトであり，このオブジェクトを survey パッケージに含まれる分析用の各種関数に渡して分析を実行していくこととなる．各関数を利用する場合は，重みとなる変数を明示的に指定する必要がなくなるメリットがある．

```
svy.zensho_s2009 <-
  svydesign(ids=~1,data=zensho_s2009,weights=zensho_s2009$Weight)
class(svy.zensho_s2009)#型を確認
```

出力結果例 4

```
> class(svy.zensho_s2009)
[1] "survey.design2" "survey.design"
```

surveydesign 型オブジェクトであることが示される．

5.5 属性別の分布確認，基本数の確認

符号表ファイルの基本数シートの数値と一致するか確認するために svytable() 関数を利用する．以下の実行例では，引数として，年齢 5 歳階級を示す属性変数「T_Age_5s」の名称が「~」（チルダー）をつけて渡され，次に上記で作成した surveydesign 型のオブジェクト名が渡されている．svytable() 関数では，属性変数を複数指定して，クロス集計を行うこともできる．ここでは，基本数シートに掲載されている重み考慮ありの総世帯数 31,761,998 に一致することが確認できる．

```
svytable(~T_Age_5s,design=svy.zensho_s2009)
sum(svytable(~T_Age_5s,design=svy.zensho_s2009))
#クロス集計の例
svytable(~T_Age_5s +T_Syuhi,design=svy.zensho_s2009)
```

出力結果例 5　2 次元クロス集計表

```
> svytable(~T_Age_5s +T_Syuhi,design=svy.zensho_s2009)
        T_Syuhi
T_Age_5s       1       2
       1  758995       0
       2 1859482       0
       3 2779608       0
       4 2923673       0
       5 2878926       0
       6 3039765       0
       7 3424996       0
       8 2936320       0
       9 3101564       0
       0       0 8058669
```

基本数シートに掲載されている重み(ウエイト)考慮ありの場合の属性別世帯数に一致していることが確認できる．

クロス集計を行う場合，上記の例では，集計値として世帯数をカウントしていたが，集計値を平均値としたい場合もある．その場合は，svyby() 関数を利用するのも一案である．集計値を平均値としたい場合は，下記の例のように引数として「svymean」を指定する．引数として，集計対象となる変数名や属性変数名を渡す場合の記述方法は，これまでの実行例から類推できる．下記の実行例で出力されたクロス集計表の値は，基本数のシートにおいて示されている数値と一致する．

```
#年齢 5 歳階級別収入平均値
svyby(~Y_Income,
      ~T_Age_5s,
      svy.zensho_s2009,
      svymean
)
#年齢 5 歳階級別(収入平均値，消費支出平均値)
svyby(~Y_Income+L_Expenditure,
      ~T_Age_5s,
      svy.zensho_s2009,
      svymean
)
#年齢 5 歳階級別就業状態別(収入平均値，消費支出平均値)
svyby(~Y_Income+L_Expenditure,
      ~T_Age_5s+T_Syuhi,
      svy.zensho_s2009,
      svymean
)
```

5.6 ヒストグラムの作成

最初に，重みの分布を通常の hist() を使ってみて以下の実行例でみてみると，ばらつきがあることがわかる．

```
hist(zensho_s2009$Weight,plot=FALSE)
```

次に，年間収入を表している数値変数「Y_Income」についてヒストグラムを図示する実行例は以下のとおりとなる．この関数でも，変数名の前に「~」（チルダー）をつけることがポイントとなる．Lumley et al.（2024）を参照するとわかるとおり，残念ながら，「plot=FALSE」オプションはサポートしていない．

```
#ヒストグラムを図示する場合の例
svyhist(~Y_Income, svy.zensho_s2009, main="Survey weighted",
freq=TRUE,col="purple",xlim=c(0,5e4),ylim=c(0,3.0e7))#明示的に freq 指定が必要
```

5.7 箱ひげ図の作成

箱ひげ図を描く関数は svyboxplot() であるが，数値変数を指定した上で，属性別に一挙に箱ひげ図を描くことができる仕様となっているので，数値変数「Y_Income」に「~」（チルダー）をつけて属性変数を指定する．属性別の作成が必要でない場合は，「1」を指定すればよい．

```
#箱ひげ図
svyboxplot(Y_Income~1,svy.zensho_s2009,all.outliers=TRUE)
#属性別箱ひげ図
svyboxplot(Y_Income~T_Age_5s,svy.zensho_s2009,all.outliers=TRUE)
```

5.8 四分位数の算出

四分位数の算出は，以下の関数を利用する．引数の渡し方は，今までの例から簡単に類推できるかと思う．

```
#四分位階級
svyquantile(~Y_Income,svy.zensho_s2009, c(.25,.5,.75))
```

出力結果例 6

```
> #四分位階級
> svyquantile(~Y_Income,svy.zensho_s2009, c(.25,.5,.75))
$Y_Income
     quantile ci.2.5 ci.97.5      se
0.25     3800   3772    3829 14.5407
0.5      5508   5472    5544 18.3672
0.75     8003   7946    8071 31.8875

> income_cdf<-svycdf(~Y_Income,svy.zensho_s2009)
#5508 の位置を確認すると中位数であることがわかる
> income_cdf[[1]](5508)
[1] 0.5000038
```

この関数は面白い使い方ができる．例えば，年間収入が 6000 千円の場合，何パーセント点にあたるのか，教えてくれる．実行例は以下のとおりである．

```
income_cdf<-svycdf(~Y_Income,svy.zensho_s2009)
income_cdf
income_cdf[[1]](6000)
```

5.9 回帰分析

回帰分析を行うためには，svyglm() 関数を利用するが，通常の glm() 関数と同様に引数を渡せばよく，引数「family=""」も有効である．疑似的にオブザベーションが作成されている一般ミクロデータを利用しているため，下記の例では，回帰分析の結果はあくまで実行例である．

```
reg_income<-
  svyglm(Y_Income~Food, design=svy.zensho_s2009)
summary(reg_income)
library(performance)
performance(reg_income)#評価指標を出力
```

5.10 主成分分析

主成分分析を行うためには，svyprcomp() 関数を利用するが，通常の prcomp() 関数と同様に引数を渡せばよい．関数より作成されるオブジェクトも通常の prcomp() の場合と同様に扱えばよい．class() 関数を使えば，そのことがわかる．なお，疑似的にオブザベーションが作成されている一般ミクロデータを利用しているため，下記の例では，統計的な意味はあまりなく，結果はあくまで実行例である．

```
svy.pc<- svyprcomp(
  ~Y_Income+L_Expenditure+Food+Housing+LFW+Furniture+Clothes+Health
  +Transport+Education+Recreation+OL_Expenditure,
  design=svy.zensho_s2009,scale=TRUE,scores=TRUE)
class(svy.pc)
svy.pc$rotation#主成分負荷量を表示
svy.pc$x#主成分得点を表示
dim(svy.pc$x)
svy.pc$rotation[,1]#PC1 主成分負荷量を表示
round(sum(svy.pc$rotation[,1]*svy.pc$rotation[,2]),1)#内積により
直交していることの確認

#累積寄与率計算など
vars <- svy.pc$sdev^2
pov <-vars/sum(vars)
cpov <- cumsum(pov)
cpov#累積寄与率の表示
screeplot(x=svy.pc,type="lines",main="Scree Plot",npcs=12)
#npcs を明示しないと 10 成分分しか表示されない
```

出力結果例 7
```
#主成分分析の結果の一部
> dim(svy.pc$x)
[1] 45811    12

> round(sum(svy.pc$rotation[,1]*svy.pc$rotation[,2]),1)
#内積により直交していることの確認
[1] 0

> #累積寄与率計算など
> vars <- svy.pc$sdev^2
> pov <-vars/sum(vars)
> cpov <- cumsum(pov)
> cpov
 [1] 0.2792098 0.3799106 0.4724950 0.5539764 0.6318810 0.7077415
 0.7811314  0.8526595  0.9166016 0.9625722 1.0000000 1.0000000
#当然に 1 になる.
```

```
screeplot(x=svy.pc,type="lines",main="Scree Plot",npcs=12)
#出力結果は，図 3 のとおりである.
```

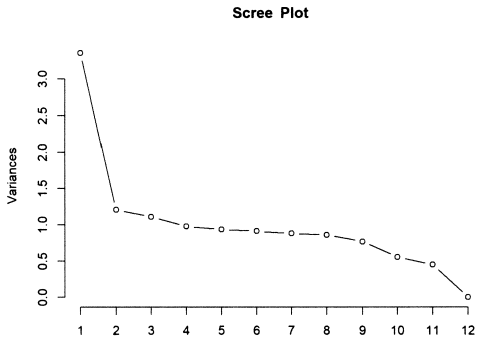


図 3. 出力結果例 8. Rstudio において Screeplot 図を確認.

参 考 文 献

独立行政法人統計センター (2024). 一般用ミクロデータの利用, https://www.nstac.go.jp/use/archives/ippan-microdata/（最終アクセス日 2024 年 4 月 30 日）.

行政改革推進本部 (2024). 政府の行政改革／ EBPM の推進, https://www.gyoukaku.go.jp/ebpm/index.html（最終アクセス日 2024 年 4 月 30 日）.

Lumley, T., Gao, P. and Schneider, B. (2024). Package 'survey' Analysis of Complex Survey Samples, https://cran.r-project.org/web/packages/survey/survey.pdf（最終アクセス日 2024 年 4 月 30 日）.

ミクロデータ利用ポータルサイト (2024). miripo, https://www.e-stat.go.jp/microdata/（最終アクセス日 2024 年 4 月 30 日）.

佐藤正昭 (2023). 公的統計調査匿名データを活用した教育活動の可能性について, 月刊「統計」, 2023 年 10 月号, 54–55.

政府統計の総合窓口 (2024). e-Stat, https://www.e-stat.go.jp/（最終アクセス日 2024 年 4 月 30 日）.

総務省統計局 (2024). 平成 21 年(2019 年)全国消費実態調査, https://www.stat.go.jp/data/zensho/2009/index.html（最終アクセス日 2024 年 4 月 30 日）.

Practical Examples of Education Using Official Statistics

Masaaki Sato

Faculty of Data Science, Shiga University

This paper introduces a component of statistical education using official statistics and explains the method of analysis using the R survey package, which is necessary for that purpose.

Key words: Microdata, anonymous data from official statistical surveys, statistical education, survey package.

統計数理(2024)
第 72 巻 第 2 号 273–303
©2024 統計数理研究所
［研究詳解］

高次元統計解析で探る銀河の分子ガスの物理状態と天文学への展望

竹内 努[1,2]・矢田 和善[3]・江頭 健斗[4]・青嶋 誠[3]・吉川 耕司[5]・石井 晶[4]・加納 龍生[1]・施 文[1]・曹 愛奈[1,8]・馬 海霞[1]・松井 瀬奈[1]・中西 康一郎[6,9]・クレ スチェータ[6,7]・河野 孝太郎[10]

(受付 2024 年 1 月 30 日；改訂 7 月 18 日；採択 7 月 19 日)

要　　旨

現代科学のデータ解析では，次元が d，標本数が n であるデータにおいて，$n \ll d$ となる場合が往々にして見られる．天文学では従来，このような状況は不適切と見なされ，データの持つ次元のほとんどの情報を破棄して $d < n$ にする以外に選択肢はないと考えられていた．$n \ll d$ となるデータは，高次元小標本(high-dimensional low sample size: HDLSS)と呼ばれる．HDLSS を含む高次元データの解析には，高次元データ特有の新しい理論と方法論が必要となる．青嶋らの研究グループは，それらを高次元統計解析と名付け，さまざまな統計理論とその方法論を構築した．本論文ではまず高次元統計解析を紹介し，その代表的な手法であるノイズ掃出し主成分分析(principal component analysis by noise-reduction method: NRPCA)と自動スパース主成分分析(automatic sparse PCA: A-SPCA)を導入する．これらの方法の実際のデータへの応用例として，アタカマ大型ミリ波/サブミリ波干渉計(Atacama Large Millimeter/Submillimeter Array: ALMA)が撮影した近傍宇宙の典型的な爆発的星形成(スターバースト)銀河 NGC 253 の分光マップに適用する．ALMA の分光マップは典型的な HDLSS データである．NGC 253 の中心部は全体として回転しており，放射される電磁波は回転速度に対応したドップラー効果によって波長が偏移している．元のデータはドップラー効果がそのまま表れており，予備解析としてまずこの元データを解析した．その結果，高次元 PCA はドップラー効果を正確に抽出し，物理的なモデルを介在させることなく回転の空間構造を正確に記述できた．次に，同じ方法を回転のドップラーシフトを補正したデータに適用し，より詳細なスペクトルの特徴を解析

[1] 名古屋大学 素粒子宇宙物理学専攻：〒 4664–8602 名古屋市千種区不老町
[2] 統計数理研究所 統計的機械学習研究センター 客員：〒 190–8562 東京都立川市緑町 10–3
[3] 筑波大学 数理物質系数学域：〒 305–8571 つくば市天王台 1–1–1
[4] 東京理科大学 創域理工学部：〒 278–8510 千葉県野田市山崎 2641
[5] 筑波大学 計算科学研究センター：〒 305–8571 つくば市天王台 1–1–1
[6] 国立天文台：〒 181–8588 東京都三鷹市大沢 2–21–1
[7] 日本学術振興会 特別研究員(PD)
[8] 学習院大学 理学部：〒 171–8588 東京都豊島区目白 1–5–1
[9] 総合研究大学院大学 物理科学研究科：〒 181–8588 東京都三鷹市大沢 2–21–1
[10] 東京大学 大学院理学系研究科附属天文学教育研究センター：〒 181–0015 東京都三鷹市大沢 2–21–1

した．これらの解析により，NRPCA と A-SPCA によって ALMA スペクトルマップの非常に複雑な特性を定量化できることが示された．特に，これらの方法はモデルを仮定することなく NGC 253 の中心からの大規模な質量流の情報を抽出でき，高次元統計学の強力さを示した．この方法は，分光サーベイデータだけでなく，HDLSS であるあらゆるタイプのデータに適用可能である．

キーワード：高次元統計解析，高次元主成分分析，星間物質，分子輝線，爆発的星形成，銀河進化．

1. はじめに

最近 10 年間で「ビッグデータ」という用語は広く認識され，天文学を含む科学研究のほぼすべての分野でみられるようになってきている．現代の天文観測装置は，従来に比べ圧倒的に大量のデータを取得でき，それらの物理状態についての定量的情報を提供してくれる．天文学で最も典型的なビッグデータとして，Sloan Digital Sky Survey（SDSS）[1]，Dark Energy Spectroscopic Instrument（DESI）Survey[2]，Gaia[3]，PanSTARRS（Panoramic Survey Telescope And Rapid Response System）[4] などが挙げられる．そのデータサイズは数十万から数百万に達し，将来計画では数億のオーダーになると期待されている．

しかし実は，天文学にはこれとは異なった種類のビッグデータが存在する．天文学では往々にして，詳細な観測には非常に時間がかかる．たとえばアタカマ大型ミリ波/サブミリ波干渉計（Atacama Large Millimeter/submillimeter Array: 以下 ALMA）によって対象天体をマッピング観測する場合，天体上の多くの点で電波強度を測定することは一般に非常に時間がかかり，容易ではない．このような状況は，より一般的に面分光[5] で頻繁に見られる（例：Very Large Telescope（VLT）での UVES[6]，William Herschel Telescope の SAURON[7]，SDSS MaNGA data[8] など）．この種の別の典型的な例としては，電波干渉計（例えば ALMA）や X 線，紫外線，または中/遠赤外線波長の衛星機器による分光マッピング調査が挙げられる．波長（または振動数）の次元を d で表し，標本数を n で表すと，分光マッピング観測の場合往々にして $n \ll d$ となる．

従来の統計解析では，標本数 n がデータの次元 d ($n \gg d$) よりもはるかに大きいことを前提としている．すでに述べたように，様々な自然科学研究においてこの前提は成り立っていないことが多い．天体物理学では，伝統的にこのような状況は不良設定問題と見なされ，上述の分光マッピングの場合では波長分散方向の情報を大幅に捨てることによって $d < n$ としていた．2010 年までこの状況は他の研究分野でもほぼ同様で，多くの研究者は単純にその方向での更なる解析を諦めていたか，ビッグデータ時代の到来以前は，そもそもそのような問題に対処する必要さえ見出していなかった．例えば，有名なデータセット集である Hand et al. (1994) には，当時の統計学で扱われた 500 を超えるデータが掲載されているが，そのほとんどは $d < n$ と従来型の前提を満たしている．しかし 1990 年代後半，情報科学の発展に伴い，突如として $d \gg n$ なる高次元データが脚光を浴びるようになった．有名な例として，Golub et al. (1999) では，白血病患者の遺伝子発現データが扱われ，次元数 d は 7129 であるのに対し，標本数 n は 72 であった．1990 年代の統計学は $d < n$ を大前提にして構成されていたため，このような場合の統計的推測に精度を保証することができなかった．

標本数がデータの次元数よりも低い ($n \ll d$) データには，以下に代表されるようなさまざまな問題が知られている．

(1) 標本共分散行列の逆行列が不安定もしくは存在しないので，古典的な統計学では扱えない．

(2) さまざまなタイプの「次元の呪い」が生じる．
(3) 往々にして計算コストが膨大になる．

これらの問題により，古典的な統計学の方法では高次元のデータ空間の特性を捉えることができず，データに内在する豊富な情報を活用できない．よって $n \ll d$ のデータを解析するには，高次元データ特有の統計理論と方法論を構築する必要があった．

このような状況は 2000 年代に劇的に変化した．まず，確率論とランダム行列理論に基づく理論物理学の分野から重要な結果がもたらされた．d と n が同じオーダーで増大する（すなわち $n/d \longrightarrow c > 0$ (c：定数)）場合に，母集団がガウス分布という仮定の下，標本共分散行列の固有値の漸近的挙動が議論された（たとえば Baik et al., 2005; Baik and Silverstein, 2006; Johnstone, 2001; Paul, 2007）．ただし，HDLSS データにおいては，$n/d \longrightarrow c > 0$，および母集団がガウス分布という仮定は現実的ではない．

Hall らは論文 Hall et al. (2005) において，高次元小標本 (high-dimensional low sample size: HDLSS) という新しい枠組みでの漸近理論を導入した．彼らの研究では，標本数 n を固定し，データ次元 d を $d \longrightarrow \infty$ と極限をとっている．これは，d を固定して $n \longrightarrow \infty$ とする従来の漸近理論からの革新的な脱却であった．

Hall et al. (2005) および Ahn et al. (2007) は HDLSS データの重要な幾何学的な表現について説明したものの，それでもガウス分布は仮定されていた．対照的に，Yata and Aoshima (2012) はこの仮定を取り払った新しい理論を展開し，ガウス分布の仮定が満たされない場合について全く異なるタイプの幾何学的表現も発見した (詳細は Yata and Aoshima, 2012 や 青嶋・矢田, 2019 を参照のこと)．これらの研究は，高次元のデータ解析に関する数理統計学の研究の基礎をなし，現在流行のビッグデータサイエンスの重要な部分として発展した．これらの新理論は，ゲノム解析，医学，神経科学，画像および形状解析などのさまざまな研究分野に新たな光をもたらしている．そして青嶋 誠，矢田 和善，石井 晶と共同研究者らは，この種の問題を扱う統計的方法論の枠組みをさらに発展させた（たとえば Aoshima and Yata, 2011, 2014, 2015, 2019; Yata and Aoshima, 2012, 2013; Ishii et al., 2016; およびその参考文献）．高次元統計解析自体の数学的背景に関心のある読者は，いくつかの総説論文を参照されたい（たとえば Aoshima and Yata, 2017; Aoshima et al., 2018）．

ところが，他の分野での大きな進歩にもかかわらず，この新しい統計的方法論は天文学においては全く知られていなかった．本論文では，天文学における HDLSS データの高次元統計解析の最初の応用例を紹介する．天文データに対する高次元統計解析手法のパフォーマンスを示すため，近傍銀河 NGC 253 の中心領域の分光マップに応用した結果を示す．NGC 253 は，近くにある典型的なスターバースト銀河[9]であり，さまざまな波長で観測されている（Rieke et al., 1980）．本研究では高次元主成分分析 (PCA) を ALMA 分光マップ（Ando et al., 2017）に適用した結果を示す．Ando et al. (2017) は，ALMA Band 7 に基づく NGC 253 の中心 $\sim$ 200 pc (パーセク) の領域における 8 pc $\times$ 5 pc の分解能での分光マップを取得した[10)11)]．以下で詳細に紹介するように，ALMA マップは通常 HDLSS データであるとみなせる．NGC 253 のスペクトルには，非常に多数の分子輝線が存在する（たとえば Ando et al., 2017; Martín et al., 2021）．本研究のもう 1 つの焦点は，輝線やバンド輝線[12]が非常に多数存在する ALMA のスペクトルを PCA によって客観的に分類することである．銀河の進化は，主として星間物質 (interstellar medium: ISM) から星への遷移である星形成によって駆動される．星形成進化には ISM のさまざまな相[13]が関連しており，ISM の進化は銀河の進化の完全な理解に向けた鍵となる．分光観測は，銀河の物質の情報を抽出して解釈するために非常に重要な手段である（たとえば Martín et al., 2021）．よって，高次元 PCA による輝線スペクトル分類が有効であるならば，これは銀

河進化研究のための強力なツールとなり得る．

本論文の構成は以下のようになっている．2 章では高次元 PCA に重点を置き，HDLSS データと高次元統計解析の概念を紹介する．ここでは詳細な数学的証明なしに，HDLSS データの驚くべき性質を特徴付けるいくつかの重要な定理を提示する．次に 3 章では天文データの実際の解析を示す．ここでは本解析に使用する NGC 253 の ALMA マップの性質について説明する．本論文のデータ解析は 2 段階構成になっている．まず 4 章では高次元統計解析が分光マップデータの解析に本当に有効かどうかを検証する．その優れたパフォーマンスを確認し，5 章では NGC 253 分子線のより詳細な解析に進む．結論は 6 章で示す．付録 A では星間物質からの分子輝線の基礎，付録 B では放射のドップラー効果とデータへの補正法を解説している．本論文は，主として Takeuchi et al. (2024)の内容をもとに著者らの研究を解説したものである．

2. 方法：高次元統計解析

2.1 高次元統計解析

本節では，高次元統計解析のいくつかの重要な統計理論とその方法論を紹介する．より詳細な解説は例えば 青嶋・矢田 (2013, 2019)を参照のこと．

2.1.1 HDLSS データにおける強不一致性

まず，強不一致性と言われる高次元データ特有の現象を 1 つ例示する．標本 $\vec{x}_i$ $(i = 1, \ldots, n)$ が平均 $\vec{\mu}$ をもつ d 次元母集団から抽出されたとし，標本平均 $\overline{\vec{x}}$

$$\overline{\vec{x}} \equiv \frac{1}{n}\sum_{i=1}^{n} \vec{x}_i \tag{2.1}$$

を考える．

従来の大標本の枠組みである $d/n \longrightarrow 0$ といくつかの条件の下では

$$\left\|\overline{\vec{x}} - \vec{\mu}\right\| \xrightarrow{\mathrm{P}} 0 \tag{2.2}$$

なる一致性をもつ．ここで，$\xrightarrow{\mathrm{P}}$ は確率収束を表す[14]．一方，高次元統計解析の枠組みである $d/n \longrightarrow \infty$ では，

$$\left\|\overline{\vec{x}} - \vec{\mu}\right\| \xrightarrow{\mathrm{P}} \infty \tag{2.3}$$

となる．この性質は，強不一致性と呼ばれる．式(2.3)は，次元 d の増加に伴って観測データのノイズが急激に増大することを意味し，データが HDLSS である場合は従来の統計手法を適用できないことが分かる．

このような現象をより具体的に捉えるために，この巨大なノイズの振る舞いを理論的に検証する．以下でその背景となる数学的概念を紹介する．

2.1.2 標本共分散行列の双対表現

簡単のため，平均ベクトル $\vec{\mu} = \vec{0}$ とし，正定値行列である共分散行列 $\tilde{\Sigma}$ をもつ母集団を考える．共分散行列 $\tilde{\Sigma}$ の固有値を

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \ (> 0) \tag{2.4}$$

とすれば，

$$\tilde{\Sigma} = \tilde{H}\tilde{\Lambda}\tilde{H}^{\top} \tag{2.5}$$

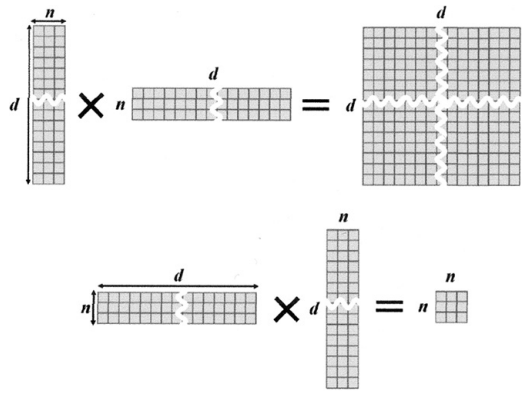


図 1. 標本共分散行列 $\tilde{S}$ とその双対行列 $\tilde{S}_{\rm D}$ の間の関係の概念図．上のパネルは標本共分散行列を，下のパネルはその双対行列を表す．

$$\tilde{\Lambda} \equiv \mathrm{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_d) \tag{2.6}$$

と固有値分解することができる．ただし，$\top$ は転置を表す．ここで $\tilde{H} = [\vec{h}_1, \ldots, \vec{h}_d]$ はそれら固有値に対応する固有ベクトルで構成された直交行列である．

ここで，この母集団から n 個の標本 $(d > n)$，$\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n$ を抽出し，$(d \times n)$ のデータ行列

$$\tilde{X} = (\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n) \tag{2.7}$$

で表現する．ただし，$\vec{x}_1, \ldots, \vec{x}_n$ は互いに独立に同分布に従う (i.i.d.)．標本共分散行列 $(d \times d)$ は

$$\tilde{S} = \frac{1}{n}\tilde{X}\tilde{X}^{\top} \tag{2.8}$$

で与えられる．これに対応する双対標本共分散行列 $(n \times n)$ は

$$\tilde{S}_{\rm D} \equiv \frac{1}{n}\tilde{X}^{\top}\tilde{X} \tag{2.9}$$

として定義される (図 1)．この行列 $\tilde{S}_{\rm D}$ の n 個の標本固有値

$$\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_n \geq 0 \tag{2.10}$$

は ($d > n$ のもと) $\tilde{S}$ の標本固有値のうち最初の n 個と同一であることに注意する．HDLSS データにおいて $\tilde{S}_{\rm D}$ には以下のような利点がある．

(1) 推測における計算コストの大幅な削減．
(2) 双対空間上での高次元現象の可視化．
(3) 双対空間上での漸近理論の展開．

2.1.3 高次元データの幾何学的表現

高次元データの幾何学的表現を最初に扱ったのは，Hall et al. (2005) である．彼らはその幾何学的表現を分類問題に応用した．さらに，Aoshima and Yata らの一連の研究により，HDLSS データにおける特徴的な幾何学的表現が導出され，高次元のさまざまな統計的推測に応用された[15]．本節では，高次元 PCA に関連する HDLSS の幾何学的表現を 1 つ紹介する．

共分散行列の固有値分解 $\tilde{\Sigma} = \tilde{H}\tilde{\Lambda}\tilde{H}^{\top}$ に基づき，

$$\tilde{X} \equiv \tilde{H}\tilde{\Lambda}^{\frac{1}{2}}\tilde{Z} \tag{2.11}$$

とおく．ただし，$\tilde{Z} = [\vec{z}_1, \ldots, \vec{z}_d]^\top$, $\vec{z}_i = (z_{i1}, \ldots, z_{in})^\top$ $(i = 1, \ldots, d)$ とする．ここで，$\mathbb{E}(z_{ij}z_{i'j}) = 0$ $(i \neq i')$ および $\mathbb{V}(\vec{z}_i) = \tilde{I}_n$ である（$\mathbb{E}$ および $\mathbb{V}$ は期待値および分散，$\tilde{I}_n$ は n 次元単位行列とする）．$\tilde{Z}$ のそれぞれの変数の 4 次モーメントは一様有界であると仮定する．$\tilde{X}$ がガウス分布に従うとすれば，すべての z_{ij} は i.i.d. の標準正規分布に従う確率変数となる．また，$\tilde{S}_{\mathrm{D}}$ の固有値分解を

$$\tilde{S}_{\mathrm{D}} = \sum_{i=1}^{n} \hat{\lambda}_i \vec{u}_i \vec{u}_i^\top \tag{2.12}$$

で定義する．ここで $\vec{u}_i$ は $\hat{\lambda}_i$ に対応する（長さ 1 の）固有ベクトルである．$\hat{\vec{h}}_i$ を $\tilde{S}$ の i 番目の（長さ 1 の）固有ベクトルとすると，$\hat{\vec{h}}_i$ は

$$\hat{\vec{h}}_i = (n\hat{\lambda}_i)^{-1/2}\tilde{X}\vec{u}_i \tag{2.13}$$

と計算できる．ここで以下で与えられる条件

$$\frac{\mathrm{tr}\left(\tilde{\Sigma}^2\right)}{\left(\mathrm{tr}\,\tilde{\Sigma}\right)^2} = \frac{\displaystyle\sum_{i=1}^{d}\lambda_i^2}{\left(\displaystyle\sum_{i=1}^{d}\lambda_i\right)^2} \longrightarrow 0 \quad \text{as} \quad d \longrightarrow \infty \tag{2.14}$$

を仮定する．ここで，$\mathrm{tr}(\tilde{\Sigma}^2)/(\mathrm{tr}\,\tilde{\Sigma})^2 \in [1/d, 1)$ となり，$\Sigma = c\tilde{I}_d$ $(c > 0)$ の場合にその最小値 $1/d$ を与える．すなわち，共分散行列が単位行列のような球形に近い場合はその比が小さくなる．なお，(2.14) は球形条件とも呼ばれる．$\tilde{X}$ がガウス分布に従う場合，条件 (2.14) の下で，

$$n\left(\sum_{i=1}^{d}\lambda_i\right)^{-1}\tilde{S}_{\mathrm{D}} \xrightarrow{\mathrm{P}} \tilde{I}_n \quad \text{as} \quad d \longrightarrow \infty \tag{2.15}$$

が得られる（Ahn et al., 2007; Jung and Marron, 2009; Yata and Aoshima, 2012）．ここで

$$\vec{w}_i = n\left(\sum_{i=1}^{d}\lambda_i\right)^{-1}\tilde{S}_{\mathrm{D}}\vec{u}_i = n\left(\sum_{i=1}^{d}\lambda_i\right)^{-1}\hat{\lambda}_i\vec{u}_i \tag{2.16}$$

とおく．Yata and Aoshima (2012) はガウス分布を緩めた適当な条件で，n 次元ベクトル $\vec{w}_i$ について

$$\vec{w}_i \xrightarrow{\mathrm{P}} 1 \quad \text{as} \quad d \longrightarrow \infty \ \ (i = 1, \ldots, n) \tag{2.17}$$

を示した．すなわち，データの双対空間における固有地・固有ベクトルが単位 n 次元球表面に集中する「球面集中現象」である．式 (2.17) で示される幾何学的表現により，標本固有値が一定に定まることが見て取れる．これは，$\tilde{S}_{\mathrm{D}}$（あるいは $\tilde{S}$）を用いた従来型の推定量では固有値の推定が困難であることを意味する．一方，Yata and Aoshima (2012) は，データが非ガウス分布のもとで，まったく異なる幾何学的表現（座標軸集中現象）を与えている．詳細は，Yata and Aoshima (2012) や 青嶋・矢田 (2019) を参照のこと．

これで双対空間におけるノイズの幾何学的挙動を調べる準備が整った．平均 $\vec{\mu} = \vec{0}$ と共分散行列 $\tilde{\Sigma}_d = \tilde{I}_d$ の d 次元ガウス分布から抽出された標本を考える．この分布から $n = 3$ の標本による $\vec{w}_i$ をいくつか（独立に）生成すると，図 2 に示される幾何学的表現が得られる．次元 d が

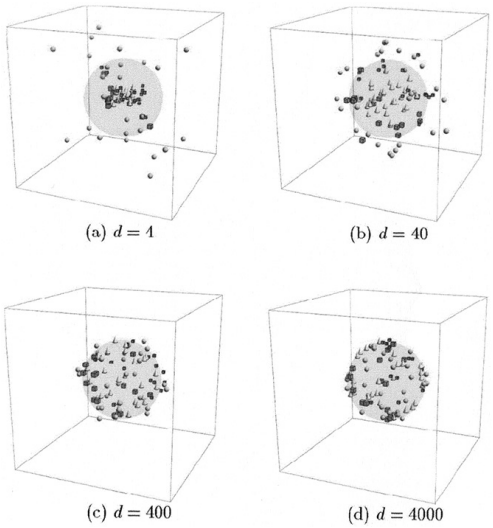


図 2. 式(2.17)で示される幾何学的表現.

低い場合，直感からそれほど外れた形状をしていない，つまり $\vec{w}_i$ は原点に集まり，ある程度散らばっている．一方より高い次元 d の場合，$\vec{w}_i$ は半径 $(d/n)^{\frac{1}{2}}$ の球に集中する．これは，高次元データになるにつれ，データの本質的な特徴が巨大なノイズ球に集まることを明示している．この幾何学的表現が，高次元データ特有の重要な性質である．

2.2 高次元主成分分析

前節と同様に，平均 $\vec{\mu} = \vec{0}$, 共分散行列 $\tilde{\Sigma}_d$ を持つ母集団を考える．$\tilde{\Sigma}_d$ の固有値について，Yata and Aoshima (2009)で導入された「一般化スパイクモデル」と呼ばれる次のモデルを仮定する[16]．

$$\lambda_i = \begin{cases} a_i d^{\alpha_i} & \text{for} \quad i = 1, \ldots, m \\ c_i & \text{for} \quad i = m+1, \ldots, d \end{cases} \tag{2.18}$$

ここで $a_i(>0)$, $c_i(>0)$, α_i $(\alpha_1 \geq \cdots \geq \alpha_m > 0)$ は未知の定数，m は未知の正の整数であり，$\lambda_1 \geq \cdots \geq \lambda_d$ を満たすとする[17]．固有値 λ_i $(i \leq m)$ は潜在的な情報を表し，λ_i $(i \geq m+1)$ はノイズを表している．ここで，

$$\frac{\sum_{i=m+1}^{d} \lambda_i^2}{\left(\sum_{i=m+1}^{d} \lambda_i\right)^2} \longrightarrow 0 \quad \text{as} \quad d \longrightarrow \infty \tag{2.19}$$

が成立する．つまりノイズの空間は球形条件 [式(2.14)] を満たす．

2.2.1 従来型 PCA の限界

次元 d に依存する標本数を $n(d) = d^{\gamma}$ $(\gamma > 0)$ の形で定義しておく．Yata and Aoshima (2009)は，HDLSS データに適用した場合における従来型 PCA を漸近的に評価し，従来型 PCA による推定が一致性を持つための条件を求め，従来型 PCA の限界を示した[18]．

彼らは，以下の条件

$$z_{ij} \quad (i = 1, \ldots, d,\ j = 1, \ldots, n) \text{ は独立} \tag{2.20}$$

を満たす場合と満たさない場合について，一般化スパイクモデル(2.18)のもと固有値に関する次の 2 つの定理を得た[19)]．

定理 1. [Yata and Aoshima (2009) I] $\tilde{Z}$ の成分が条件(2.20)を満たすならば，$i \leq m$ について
(1)$\alpha_i > 1$ を満たす i について $d \longrightarrow \infty$ かつ $n \longrightarrow \infty$
(2)$\alpha_i \in (0, 1]$ を満たす i について $d \longrightarrow \infty$ かつ $d^{1-\alpha_i}/n(d) \longrightarrow 0$
の条件のもとで

$$\frac{\hat{\lambda}_i}{\lambda_i} \xrightarrow{\mathrm{P}} 1 \tag{2.21}$$

が成り立つ．

定理 2. [Yata and Aoshima (2009) II] $\tilde{Z}$ の成分が条件(2.20)を満たさないならば，$i \leq m$ について
(1)$\alpha_i > 1$ を満たす i について $d \longrightarrow \infty$ かつ $n \longrightarrow \infty$
(2)$\alpha_i \in (0, 1]$ を満たす i について $d \longrightarrow \infty$ かつ $d^{2-2\alpha_i}/n(d) \longrightarrow 0$
の条件のもとで式(2.21)が成り立つ．

定理 1 と 2 は，従来型 PCA による固有値推定が一致性を持つための条件を与える．これらの定理において，(1)の $d \longrightarrow \infty$ と $n \longrightarrow \infty$ で記述された条件は，n は d に依存していない．すなわち，その n が d よりもはるかに小さい HDLSS の場合も含む．しかし(2)では，特に $\tilde{Z}$ の成分が条件(2.20)を満たさない場合，n は d に強く依存している．したがって，従来型 PCA を HDLSS データに適用する場合は細心の注意を払う必要がある．この従来型 PCA の高次元における問題を克服するため，Yata and Aoshima (2010, 2012)は HDLSS データのための 2 つの新しい PCA を提案した．本論文ではそのうちの 1 つであるノイズ掃出し法 (Yata and Aoshima, 2012)に焦点を当てる[20)]．

2.2.2 ノイズ掃出し PCA(NRPCA)：HDLSS データのための新しい方法

ここで高次元 PCA を導入する．図 3 は高次元 PCA のイメージを説明している．HDLSS データは，図 2 に見られるような球面に集中する巨大なノイズを持つため，データが持つ潜在的に重要な情報(潜在空間)はノイズ球に完全に埋もれてしまい，隠されてしまっている．しかし，実はこの状況下でもデータの特性を推定することは不可能ではない．これは，ノイズの挙動が球面近傍で特定され，それを差し引くことができるという重要な事実による．その手法が，Yata and Aoshima (2012)によって提案されたノイズ掃出し(NR)法である．本論文ではノイズ掃出し PCA(noise-reduction PCA: NRPCA)と呼ぶ．

以下でもう少し厳密な説明を試みる．Yata and Aoshima (2012)は，幾何学的表現 [式(2.15)] から派生した NR 法を提案した．まず $\tilde{S}_{\mathrm{D}}$ を次のように分解する．

$$n\tilde{S}_{\mathrm{D}} = \sum_{i=1}^{m} \lambda_i \vec{z}_i \vec{z}_i^{\top} + \sum_{i=m+1}^{d} \lambda_i \vec{z}_i \vec{z}_i^{\top} \ . \tag{2.22}$$

第 2 項はノイズを表現しており，条件(2.20)のもとで式(2.15) と同様に，次のようなノイズの幾何学的表現が得られる．

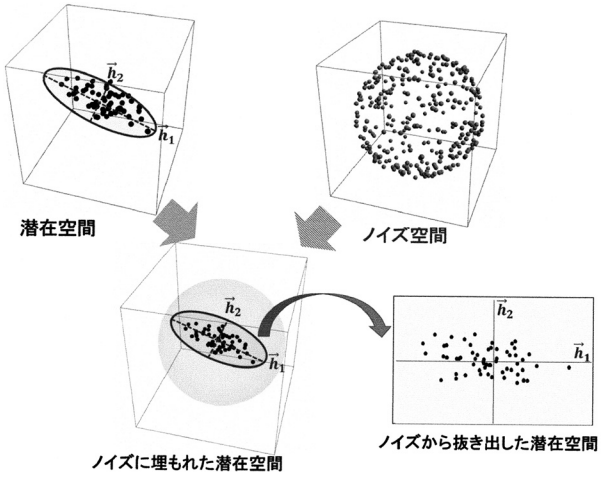


図 3. 高次元主成分分析(PCA)の概略．HDLSS データでは，巨大なノイズをもつため，潜在空間がノイズに埋もれてしまう．高次元 PCA を用いることで，その潜在空間の抽出が可能となる．

$$\frac{\sum_{i=m+1}^{d} \lambda_j \vec{z}_i \vec{z}_i^{\top}}{\sum_{i=m+1}^{d} \lambda_j} \xrightarrow{\mathrm{P}} \tilde{I}_n \quad \text{as} \quad d \longrightarrow \infty \,. \tag{2.23}$$

つまり，一般化スパイクモデルにおけるノイズ空間は，球面集中現象をもつことがわかる．この結果を用いてノイズ部分を除去した固有値の推定量

$$\check{\lambda}_i = \hat{\lambda}_i - \frac{1}{n-i}\left(\mathrm{tr}\,\tilde{S}_{\mathrm{D}} - \sum_{j=1}^{i} \hat{\lambda}_i\right) \qquad (i = 1, \ldots, n-1) \tag{2.24}$$

が得られる．ここで，以下の定理が得られている．

定理 3. [Yata and Aoshima (2012) (I)]　$\tilde{Z}$ の成分が条件(2.20)を満たすとき，$i \leq m$ について
(1)$\alpha_i > 1/2$ を満たす i に対し，$d \longrightarrow \infty$ かつ $n \longrightarrow \infty$
(2)$\alpha_i \in (0, 1/2]$ を満たす i に対し，$d \longrightarrow \infty$ かつ $d^{1-2\alpha_i}/n(d) \longrightarrow 0$
の条件のもとで

$$\frac{\check{\lambda}_i}{\lambda_i} \xrightarrow{\mathrm{P}} 1 \tag{2.25}$$

が成り立つ．

定理 3 と 1 を比較すると，(条件(2.20)の下で)$\check{\lambda}_i$ の一致性は $\hat{\lambda}$ よりも緩い条件で成立することが分かる．

次にノイズ掃出し法を主成分ベクトル推定に適用する．$\check{\vec{h}}_i \equiv (n\check{\lambda}_i)^{-1/2}\tilde{X}\vec{u}_i$ を定義し，その主成分ベクトル $\vec{h}_i$ の推定量としての性能を検証する．ここで $\check{\vec{h}}_i = (\hat{\lambda}_i/\check{\lambda}_i)^{-1/2}\hat{\vec{h}}_i$ であることに注意する．Yata and Aoshima (2012)は次の定理を示した．

定理 4. [Yata and Aoshima (2012) II] λ_i $(i \leq m)$ は単根とする．定理 3 の条件の下で，

$$\check{\vec{h}}_i^{\top} \vec{h}_i \xrightarrow{\mathrm{P}} 1 \tag{2.26}$$

が成り立つ．

また，主成分スコアについても評価する．x_i の i 番目主成分スコアは $\vec{h}_i^{\top} \vec{x}_j = \vec{z}_{ij} (\lambda_i)^{\frac{1}{2}} \equiv \vec{s}_{ij}$ で与えられる．$\vec{u}_i \equiv (\hat{u}_{i1}, \ldots, \hat{u}_{in})^{\top}$ $(i = 1, \ldots, d)$ とすると，主成分スコアは $\hat{\vec{u}}_{ij} (n \check{\lambda}_i)^{\frac{1}{2}} \equiv \check{\vec{s}}_{ij}$ と書ける．ここで，i 番目主成分スコアの平均二乗誤差(MSE)を

$$\mathrm{MSE}(\check{\vec{s}}_i) \equiv \frac{1}{n} \sum_{j=1}^{n} (\check{s}_{ij} - s_{ij})^2 \tag{2.27}$$

と定義すると，もう 1 つの重要な定理を得る．

定理 5. [Yata and Aoshima (2012) III] λ_i $(i \leq m)$ は単根とする．定理 3 の条件の下で，

$$\frac{\mathrm{MSE}(\check{s}_i)}{\lambda_i} \xrightarrow{\mathrm{P}} 0 \tag{2.28}$$

が成り立つ．

ここまでデータは平均 0 としてきたが，(中心化することにより)与えた定理はすべて平均が 0 でない場合にも適用できることに注意する．

このように，HDLSS データの固有値(寄与率)，主成分ベクトル，および主成分スコアの推定において，NRPCA が高いパフォーマンスを持つことが分かる．以降，固有値推定には NRPCA で得られた値を用いる．

2.2.3 オートマティック・スパース PCA(A-SPCA)

さらに，主成分ベクトルをより精度良く推定するために，Yata and Aoshima (2022)で提案されたオートマティック・スパース PCA (automatic sparse PCA: A-SPCA)を導入する．すべての i に対して $\hat{\vec{h}}_i = (\hat{h}_{i(1)}, \ldots, \hat{h}_{i(d)})^{\top}$ とする．閾値 $\zeta_i > 0$ が与えられたとき，

$$\hat{h}_{i*(s)} = \begin{cases} \hat{h}_{i(s)} & \text{if } |\hat{h}_{i(s)}| \geq \zeta_i, \\ 0 & \text{if } |\hat{h}_{i(s)}| < \zeta_i \end{cases} \quad \text{for } s = 1, \ldots, d \tag{2.29}$$

とおく．$\hat{\tilde{\vec{h}}}_{i*} = (\hat{h}_{i*(1)}, \ldots, \hat{h}_{i*(d)})^{\top}$ とすると，$\vec{h}_i$ の閾値に基づく推定量は次のように定義される．

$$\hat{\tilde{\vec{h}}}_{i**} = \hat{h}_{i*} / \|\hat{h}_{i*}\|. \tag{2.30}$$

しかし，この推定量は ζ_i に大きく依存している．

最近，青嶋と矢田は NR 法を拡張した A-SPCA を次のように提案した (Yata and Aoshima, 2022)．NR 法による主成分ベクトル $\check{\tilde{\vec{h}}}_i = (\check{h}_{i(1)}, \ldots, \tilde{h}_{i(d)})^{\top}$ について，簡単のため $|\check{h}_{i(1)}| \geq \cdots \geq |\check{h}_{i(d)}|$ と仮定する．与えられた定数 $\omega \in (0, 1]$ に対して，成分の累積寄与率が ω より大きくなる主成分ベクトルを考える．次を満たすある整数 $\check{k}_{i\omega} \in [1, d]$ が一意に存在する．

$$\sum_{s=1}^{\check{k}_{i\omega}-1} \check{h}_{i(s)}^2 < \omega \quad \text{および} \quad \sum_{s=1}^{\check{k}_{i\omega}} \check{h}_{i(s)}^2 \geq \omega. \tag{2.31}$$

さらに，

$$\check{h}_{i\omega(s)} = \begin{cases} \check{h}_{i(s)} & \text{if } |\check{h}_{i(s)}| \geq |\check{h}_{i(\check{k}_{i\omega})}|, \\ 0 & \text{otherwise} \end{cases} \quad \text{for } s = 1, \ldots, d.$$

を定義する．$\check{\tilde{h}}_{i\omega} = (\check{h}_{i\omega(1)}, \ldots, \check{h}_{i\omega(d)})^\top$ とする．Yata and Aoshima (2022)は HDLSS データで $\check{\tilde{h}}_{i\omega}$ のいくつかの一致性に関する性質を示し，実際のデータ解析でそのパフォーマンスを検証した．本論文では $\omega = 1/2$ とする[21]．A-SPCA は，NRPCA によって構築された主成分ベクトルの最も重要な成分を抽出でき，それらの成分が実質的に主成分の方向を決めているといえる．

3. データ

本研究の対象銀河 NGC 253 は，3.5 Mpc(1″ は 17 pc に対応)の距離にある近傍棒渦巻銀河であり（Rekola et al., 2005），形態は Sc とされている．この銀河の太陽中心速度[22]は 243 ± 2 km s^{-1} ($z = 0.000811$: Kamphuis et al. 2015) で，主に宇宙膨張による．この銀河は大小マゼラン銀河を除いて最も明るい遠赤外線放射源の 1 つであり，多くの波長で広く研究されてきた．NGC 253 は純粋なスターバーストとして知られており，中心の分子領域での星形成率 (SFR) は SFR $\sim 2\ M_\odot$ [yr^{-1}] と推定されている(Rieke et al., 1980; Keto et al., 1999; Bendo et al., 2015)[23]．中心には大質量，高密度で暖かい ISM の存在が確認されている（e.g. Sakamoto et al., 2011）．さらに，スーパースタークラスター[24]も可視光および近赤外線(NIR)で観測されている（Fernández-Ontiveros et al., 2009）．非常に強いアウトフローもこの領域で観測されている（Matsubayashi et al., 2009; Bolatto et al., 2013）．スーパースタークラスター，アウトフローともスターバースト活動に密接に関連している．

3.1 NGC 253 の ALMA 分光マップ

本研究では，Ando et al.（2017）によって取得された，近傍の典型的スターバースト銀河 NGC 253 の ALMA Band 7 での分光マップを用いる．最終的な画像の合成ビームサイズは $0.''45 \times 0.''3$ で，これは NGC 253 の距離において 8 pc × 5 pc に相当する．スペクトルにはさまざまな輝線が検出されている（Ando et al., 2017）．振動数方向のサンプリング数は 2248 である．

Ando et al.（2017）の元マップは，空間領域(赤経および赤緯)で 864×864 点，振動数領域で 2248 成分(単位 [Hz])で構成される．ただし，ピクセルスケールはマップの合成ビームサイズ $0.''45 \times 0.''3$ よりも小さいためオーバーサンプリングであり，各空間グリッド点は互いに独立になっていない．この問題を克服するため，$0.''45 \times 0.''3$ の楕円ビームに対応する円形ビームサイズは，$0.''37$ を使用して，イメージマップをリサンプリングした．得られた独立なグリッドからなるマップは 230×230 ピクセルで構成される．また信号の微弱な領域を除外する必要がある．このため振動数の全範囲にわたって輻射強度強度を積分し，強度 I [Jy beam^{-1}] > 2.0 の領域を選択した．図 4 は有意な信号が検出された明るい領域および「マスク」領域をあらわしている．信号のあるピクセルの数 n は 231 となった．

3.2 宇宙膨張による赤方偏移および系の回転によるドップラーシフト

NGC 253 の中心部は大きく傾いている ($i \sim 78°$)．これにより，ドップラーシフトによって分光マップ上にかなりコヒーレントな回転パターンが生じる．この効果を考慮するため，スペクトル上の 3 本の顕著な輝線 HCN(4–3)，HNC(4–3)，および CS(7–6) にガウス関数フィットを行い，輝線の中心波長を推定した．速度場のマップを図 5 に示す．これら 3 本の輝線から得られたドップラーシフトを平均することにより，シフトの大きさを決定した．次に，宇宙膨張に

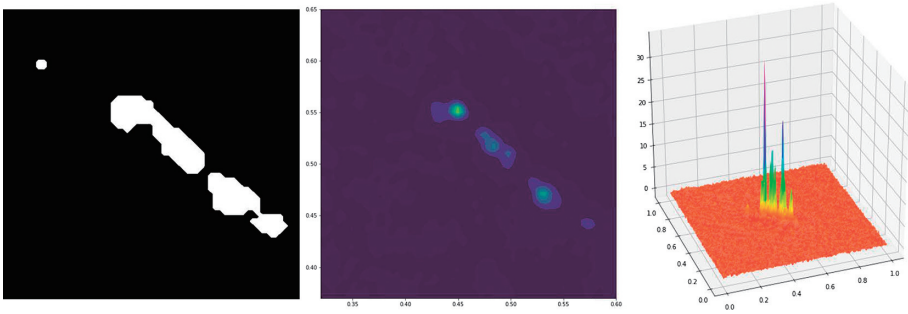


図 4. マスクで抽出した NGC 253 マップの明るい領域（Takeuchi et al., 2024）．左：マスク領域マップ．白抜きは有意な強度の信号がきている領域．中央：信号の強度マップ．右：信号の強度マップの鳥瞰図．

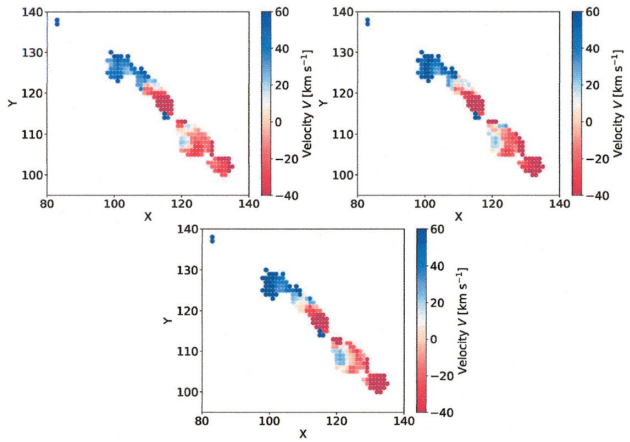


図 5. HCN(4–3)（左上），HNC(4–3)（右上），および CS(7–6)（下）輝線から推定した NGC 253 の速度マップ（Takeuchi et al., 2024）．この銀河の系全体の速度 243 [km s^{-1}] を差し引いた値を示す．

よって引き起こされる NGC 253 の系全体の速度[25]に関して振動数をシフトした．この補正によってコヒーレントな回転の系統的影響のないマップが得られ，サブミリ波輝線のより詳細な特性を解析できる．ドップラーシフトについては付録 B に詳しく述べている．

データキューブ[26]には振動数にギャップがある．速度に応じてスペクトルをシフトすると，スペクトルの境界もシフトするため，本研究では単純にすべてのピクセルに共通の振動数範囲を採用した．データの最終的なサイズは，空間位置に沿って 231, 補正後のスペクトル要素数は 1971 である．

3.3 HDLSS データとしての分光マップ

上記のように，NGC 253 のマップは 空間次元 231 × スペクトル次元 1971, つまり $n = 231$ および $d = 1971$ であり，明らかに $n \ll d$ の条件を満たしている．DNA マイクロアレイデータ（通常，$n \sim 10$–100 および $d \sim 10^4$–10^5）ほど極端ではないものの，この ALMA マップは典型的な HDLSS データである．天体物理学の観点から問題となるのは，スペクトルに非常に多く

の情報が含まれていること，そして n に比べて多種多様なスペクトル輝線が観測されることである．しかし，これは高次元統計解析においては特別な状況ではない．

ただし，分光データへの高次元統計の適用は全く単純というわけにはいかない．スペクトル線は，ISM の乱流運動や他の物理機構によって拡がったりシフトしたりする．その場合，各スペクトル分解能ユニットが持つ情報は独立にはならない[27]．この複雑な問題は，HDLSS データを扱う他の研究分野で調べられたことはない．しかし，天文学のあらゆる分光データに共通して見られる問題であるため，まず高次元統計解析が実際に輝線のシフトと広がりのある天体分光マップに対して機能するかどうかを本格的な解析の前に検証する．

4. 予備解析

すでに述べたように，ゲノム解析やその他の分野における HDLSS データの典型的な例とは異なり，天文分光データには特有の潜在的な問題が存在する．分光データをベクトルデータ $\vec{x} = (x_1, \ldots, x_d)^\top$ と考えたとき，根本的な問題は隣接する振動数(または波長)指標 i と $i+1$ $(i = 1, \ldots, d-1)$ における情報が独立ではないことである．つまり，銀河の ISM には内部運動があり，ドップラー効果によって振動数方向に広がる．また，天体の観測領域全体にわたってシステムの系統的な回転が見られる(3 章)．高次元統計解析がこのようなタイプの分光データに対して機能するかどうかを調べるため，高次元 PCA を用いて NGC 253 の元の分光マップデータの予備解析を行った．「元の」という言葉は，系全体の回転によるマップ上のドップラーシフトを補正せずにこの解析をしたことを意味する．すなわち，高次元 PCA をそのまま元データに適用し，どのような特徴が検出されるかを検証した．このため画像の 2 次元座標を赤経方向，次に赤緯方向に沿って並べ替え，データを 1 次元ベクトルと見なして指標を付けなおした．PCA を 2 次元画像に直接適用することも可能だが，本論文では解析を単純にするため 1 次元化する方針を採用した．

4.1 固有スペクトル(eigenspectra)

次の段階の解析では，主として A-SPCA の結果に注目する．ただし，NRPCA は全てのスペクトル特徴を対象として機能するので，NRPCA で求めた固有スペクトル (PCA によって観測されたスペクトルから構築された固有ベクトル)も示す．図 6 の上のパネルは NGC 253 の ALMA 分光マップから NRPCA によって構成された固有スペクトルを示している．この図では，すべての PC に対応する固有スペクトルにおいて，同じ振動数範囲で同一のスペクトルの特徴が現れていることが見て取れる．これは，一部のスペクトルの特徴が，他のスペクトルの領域よりもスペクトル全体を決定する重要な情報を含んでいることを示唆する．

オートマティック・スパース PCA(A-SPCA)は，これらの特定のスペクトルの特徴に対応する最も重要な要素を選び出し，局在化させる機能を持つ．図 6 の下のパネルは，A-SPCA によって得られた固有スペクトルを示す．図 6 からわかるように，A-SPCA はある特徴が全体のスペクトルの決定に重要となる場合のみ値を残し，他の成分を強制的に 0 にする．したがって，A-SPCA の固有スペクトルはすべての特徴を含むスペクトルの再構成には用いず，完全なスペクトルの再構成には NRPCA を用いる．A-SPCA のこの利点を利用し，次の解析でスペクトル全体を決める特徴を特定する．

4.2 PC の寄与率分布

まず NRPCA によって得られた寄与度の大きさ順にソートした PC の固有値分布を図 7 に示す．最初のいくつかの固有値は，残りの固有値よりも著しく大きい．この振舞いは，高次元統

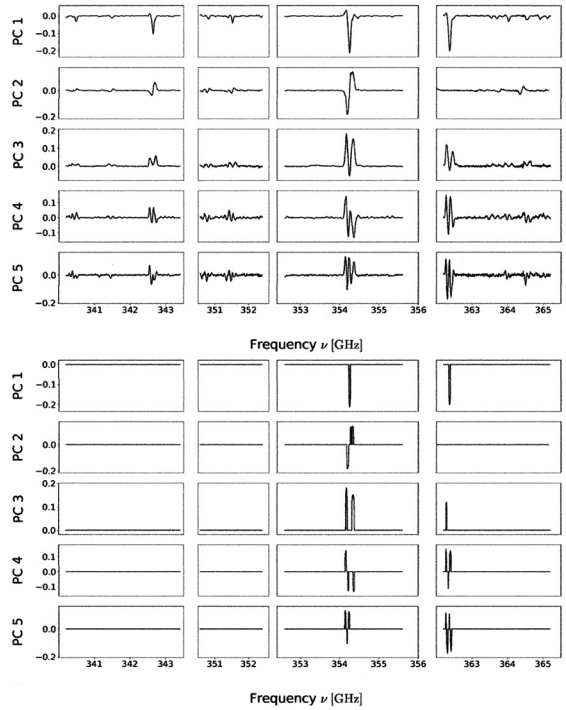


図 6. NGC 253 の ALMA 分光マップから作成された第 1 から第 5 主成分(PC)に対応する固有スペクトル (Takeuchi et al., 2024)．上段：ノイズ掃出し PCA(NRPCA)により得られた固有スペクトル．下段：オートマティック・スパース PCA(A-SPCA)によって得られた固有スペクトル．PC を決定付ける成分のみが非ゼロの値を持つ．NRPCA と通常の PCA の結果は，$\check{\vec{h}}_i = (\hat{\lambda}_i/\check{\lambda}_i)^{-1/2}\hat{\vec{h}}_i$ (p.11) より，形状は同じでスケールが違うだけの図となるためここでは示していない．

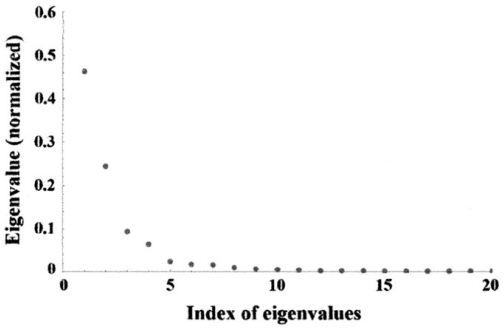


図 7. NRPCA で得られた NGC 253 の ALMA 分光マップの規格化した固有値(寄与率) (Takeuchi et al., 2024)．

計解析の文脈では「スパイク(spike)」と呼ばれる．PCA の固有値は各 PC の寄与を表しているので，PCA は共分散行列全体を各 PC からの寄与に分解することが分かる．得られた固有値分布がこのスパイク構造を示していることから，高次元 PCA が NGC 253 の ALMA 分光マップのいくつかの特徴を検出したことが理解できる．すなわち，本研究のデータが通常の HDLSS

データとみなせ，高次元 PCA の前提が成り立つことが保証されている．

天体物理学的観点からは，図 7 は分光マップデータに対する高次元統計解析の極めて有望な可能性を示している．Ando et al. (2017)は，マップ上で 30 を超える分子ないしラジカル輝線(Ando et al., 2017 の図 1 を参照)，およびいくつかの広がったスペクトル放射を特定した．たとえ興味の対象を輝線に限定したとしても，物理的直感だけで輝線を分類することは非現実的である．このような分光データから物理情報を抽出するために，伝統的には輝線比が使用されてきた．しかし，輝線の数が増えるにしたがって，輝線の組み合わせ数は爆発的に増加する(組み合わせ論的爆発)ため，力業で多数の輝線比を一度に診断することは全く現実的ではない．これと対照的に，図 7 に示されているように，複雑な分光マップの全情報は最初の数個の PC で表現できることがわかる．これは，30 本を超える輝線を含む複雑なスペクトルの特徴が，ごく少数の PC に対応するベクトル基底(固有スペクトル：eigenspectra)て決まっていることを意味する．これは PCA を含む，一般に次元削減として知られる方法を用いることのよく知られた利点である．(e.g., Galaz and de Lapparent, 1998; Ronen et al., 1999; Wang et al., 2011; Pace et al., 2019; Portillo et al., 2020)，しかし，これら先行研究はすべて，十分に大きな標本数 n についての 1 次元積分スペクトルへの応用にとどまり，HDLSS データに適用したものではなかった．本論文の解析は，これら先行研究の解析とは全く異なることを強調しておく．ここで示す方法は，第 2 章で述べたように，従来型 PCA では巨大なノイズ球の存在のため歯が立たなかった典型的 HDLSS データである分光マップに問題なく応用できる．

このように，高次元統計解析の方法が天体分光マップデータに対してうまく機能することを確認できた．これに基づき，高次元 PCA によるさらなる解析に問題なく進むことができる．

4.3 第 1 および第 2 主成分の物理的意味

得られた第 1 主成分(PC1)および第 2 主成分(PC2)が何を表すかを考察するのが次の段階である．図 8 に PC1 および PC2 の散布図を示す．最も顕著な特徴は，x 軸に関してほぼ対称な蝶の羽根のようなパターンである．これは，NGC 253 中心領域のコヒーレントな回転を連想させる．PC1 と 2 の 2 次元構造を再構築したマップを図 9 に示す．

明らかに，PC1 は輝線強度を表し，PC2 は NGC 253 中心の対称でコヒーレントなパターンを表している．実際，図 9 の右パネルの PC2 マップと，同じ領域で直接測定した視線方向の速度場マップ(図 5)は非常によく一致している．視線方向の速度場は，運動する物質から放射される輝線の振動数がドップラーシフトしていることから，振動数のシフト量を特殊相対論の式を用いて速度に換算することができる．ドップラーシフトについては付録 B で詳しく述べている．

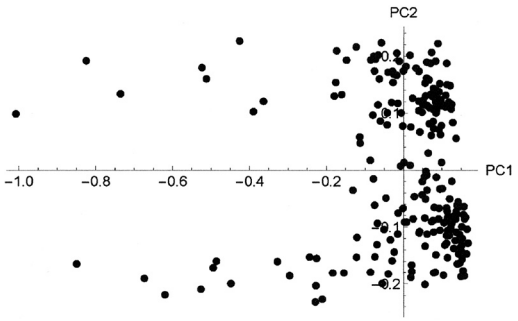


図 8．NGC 253 の ALMA 分光マップから A-SPCA によって得られた PC1 と PC2 の分布(Takeuchi et al., 2024)．

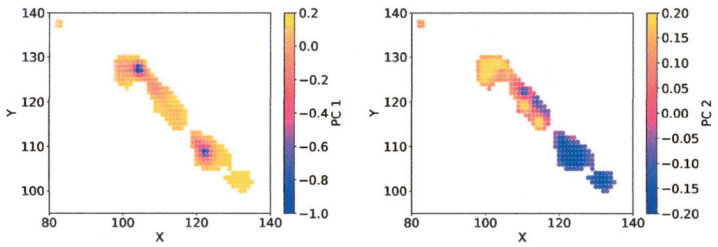


図 9. A-SPCA によって得られた PC1 と PC2 のマップ上での 2 次元分布．左パネルは第 1 主成分(PC1)，右パネルは第 2 主成分(PC2)のマップをそれぞれ示している．

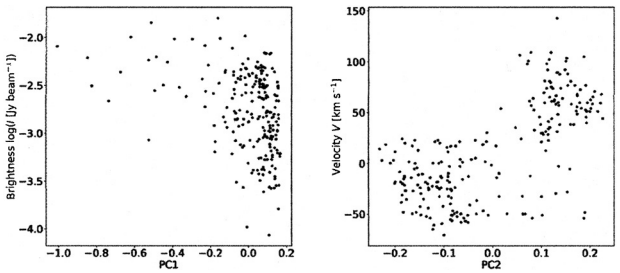


図 10. 主成分と観測された物理的特性の関係（Takeuchi et al., 2024）．左：NGC 253 の ALMA マップ上での PC1 と振動数全域での積分強度 I [Jy beam^{-1}] との相関．右：PC2 と視線速度の相関．

この関係をさらに詳しく調べるため，PC と観測された特性の相関を図 10 に示す．図 10 の左パネルは，PC1 と各ピクセルにおける振動数範囲全体にわたって積分した全強度との間の散布図である．これらの間にはよい相関が見られる．PCA はデータの線型な特徴のみを抽出するのに対し，PC と物理量との関係は必ずしも線型ではないため，この相関は必ずしもきれいな線型相関とはならない．図 10 の右パネルは固有速度と PC2 の関係を示す．すでに 2 次元マップで示したように，PC2 は図 10 で視線速度場を近似的に表している．しかし，PC2 と固有速度が一致しない特異領域が存在することに注目しよう．これらの領域は，強い質量流出など，何らかの局所的な現象の影響を受けている可能性がある．この問題については，後述の詳細な解析を用いて再検討する．

4.4 主成分に対応するスペクトルの特徴：元データの場合

各主成分に具体的にどのような特徴的スペクトルが関与しているかを検証するのは非常に興味深い解析であるが，従来の PCA はこのような目的には不向きであった．一方，A-SPCA は各主成分を決める特徴的スペクトルを特定できる．A-SPCA によって特定された，主成分を決める特徴的スペクトルを図 11 に示す．この図において，スペクトルは銀河の解析領域全域にわたって積分されている．星印は PC1 を決める特徴的スペクトル，三角形は PC2 を決める特徴的スペクトルを示している．図を見やすくするため，高次の主成分の寄与は示していない．PC1 の情報は HCN(4–3) および HNC(4–3) 分子の 2 本の輝線，特に輝線の中心部に集中している(分子輝線放射機構の詳細と表記法については付録 A を参照)．PC2 の情報は HCN(4–3) 輝線と関連するが，PC1 の場合とは対照的に輝線のウィング部に割り当てられている．この図のスペクトルは NGC 253 の解析領域全体にわたって積分されているため，各輝線のウィング部は実際にはドップラーシフトによって静止系波長からずれた輝線が重なることで形作られてい

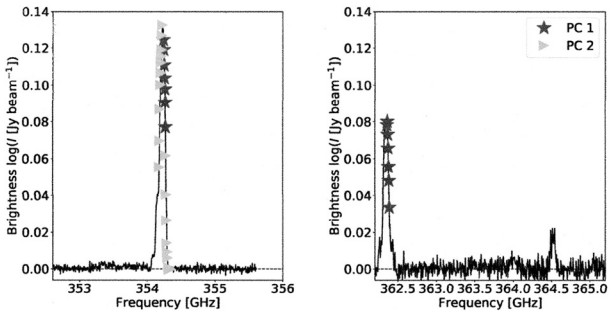


図 11. A-SPCA によって求めた NGC 253 の ALMA 分光マップの PC を特徴付ける特徴的スペクトル．星印は PC1 を決める特徴的スペクトル，三角形は PC2 を決める特徴的スペクトルである．

る．これは，PC2 のマップ(図 9)が NGC 253 のドップラー速度場マップ(図 5)とよく一致するということと整合的である．

このように，高次元 PCA は ALMA 分光マップ上の輝線の特徴抽出において高い能力を示すことが分かった．この予備解析により，HDLSS データとしての ALMA 分光マップに対する A-SPCA のパフォーマンスが保証されたので，次章では問題なく分子輝線の詳細な解析に進むことができる．

5. 本解析および議論

前章(4 章)で高次元統計解析の性能を確認した．ここでは更に詳細な物理的解析に進む．本章では，NGC 253 の分光マップの輝線の振動数から各点ごとに求めたドップラーシフトを補正したデータを用いる．これをドップラーシフト補正済み分光マップと呼ぶことにする．

5.1 NGC 253 のドップラーシフト補正済み分光マップへの高次元 PCA の応用

4 章と同様の方法で，NGC 253 のドップラーシフト補正済みマップに高次元 PCA を適用する．PC2 については 4 章と同じ表記を用いるが，ここでの PC2 はドップラーシフト補正済みマップから得られた値であり，4 章で示した値とは異なることに注意されたい．また，ドップラーシフト補正済みマップに NRPCA および A-SPCA を適用すると，PC のすべての寄与が補正の影響を受けることも注意する必要がある．

予備解析と同様に，まずドップラーシフト補正済み分光マップの固有スペクトルを図 12 に示す．図 6 と比較すると，各固有スペクトルの形状の意味がより明確になっている．スペクトル全体のドップラーシフトを既に差し引いているため，輝線の振動数はもはやコヒーレントな回転の速度を表していないことをここで強調しておく．したがって，速度構造は常に系全体の回転パターンからの 1 次のオーダーの変位であり，輝線プロファイルの広がりまたは非対称な歪みとして現れる．もう 1 つの重要な違いは，元データの解析では特定されなかった重要なスペクトル輝線が PC5 (図 12 の下のパネルの PC5 を参照)で識別できることである．これは，系の回転の圧倒的な寄与がマップから適切に取り除かれ，ドップラーシフト補正済みマップでより微妙なスペクトルの特徴が顕著になったことによる．新しい PC2 は，HCN をはじめいくつかの輝線のウィング[28]を表しており，ガスの非常に局所的な膨張を反映している．PC3 と 4 は，膨張速度よりも速いガスの動きによって生じる輝線の青方偏移を表している．PC5 はそれほど目立たないが，PC3 および 4 と同様に，輝線の赤方偏移を表している．

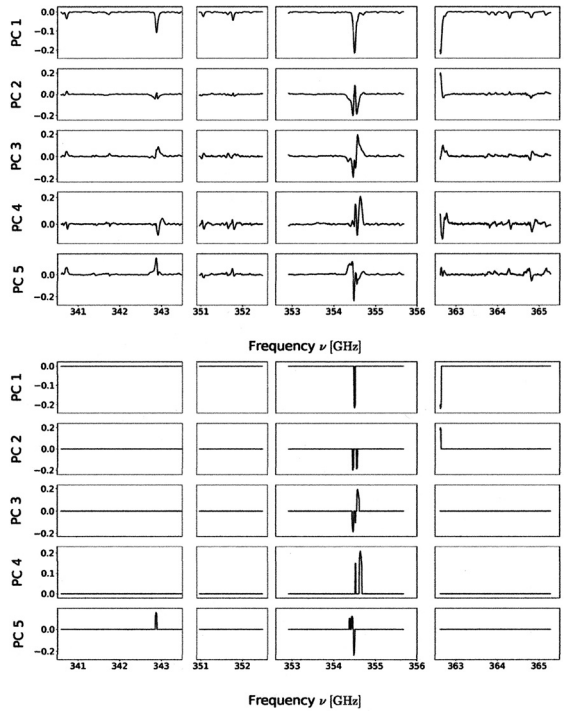


図 12. ドップラー補正済み ALMA 分光マップから高次元 PCA によって構築された第 1–第 5 主成分(PC)に対応する固有スペクトル．上 5 枚のパネルは NRPCA による固有スペクトル，下 5 枚のパネルは A-SPCA で得られた固有スペクトルを示している．

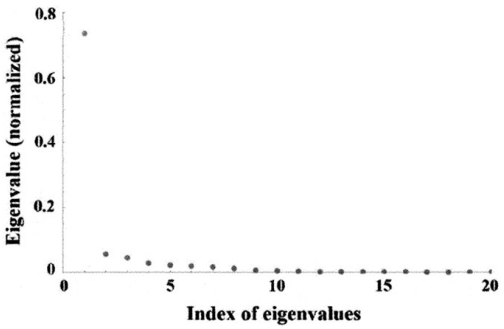


図 13. 系全体の回転によるドップラーシフト補正後の NGC 253 の ALMA 分光マップデータから NRPCA によって得られた固有値寄与率（Takeuchi et al., 2024）．

NRPCA によって得られた固有値の寄与率を図 13 に示す．ドップラーシフト補正済みマップから求められた新しい PC2 の寄与率は，図 7 の PC2 よりもはるかに小さくなっている．これは系統的回転の除去がうまくいっていることによる．そして，新しい PC2 の寄与率は小さいながらも無視できない大きさを持つ．つまり，新しい PC2 は NGC 253 の分光マップにおける，局所的で詳細な特徴を表現している．これは，固有スペクトルからの示唆とも一致している．

このことは，PC の散布図においてより顕著に表れる．各 PC 間の関係を図 14–17 に示す．

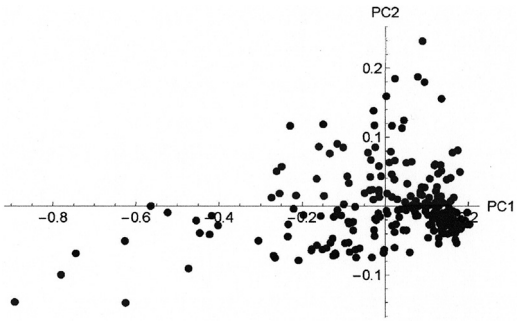


図 14. 系全体の回転によるドップラーシフト補正後の NGC 253 の分光マップから A-SPCA によって求めた PC1 と PC2 の 2 次元分布．この図における PC2 は図 9 の PC2 とは異なることに注意．

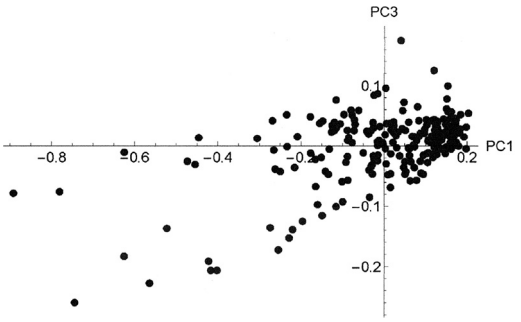


図 15. 図 14 と同様，NGC 253 のドップラーシフト補正済み分光マップから A-SPCA によって求めた PC1 と PC3 の 2 次元分布．

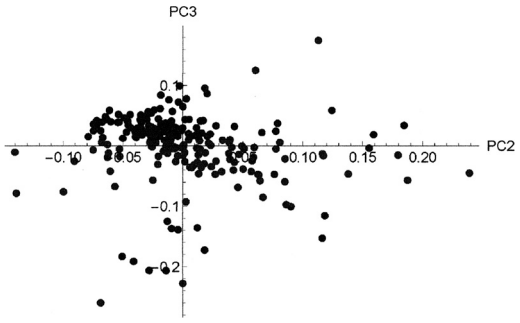


図 16. 図 14 と同様，NGC 253 のドップラーシフト補正済み分光マップから A-SPCA によって求めた PC2 と PC3 の 2 次元分布．

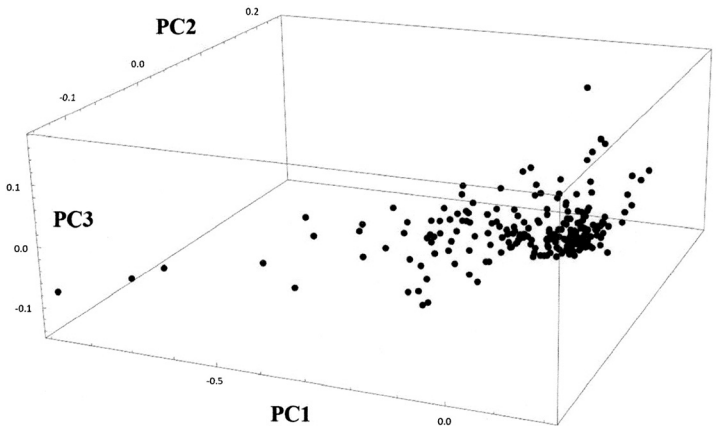


図 17. NGC 253 のドップラーシフト補正済み分光マップから求めた PC1, PC2, PC3 の分布の 3 次元構造.

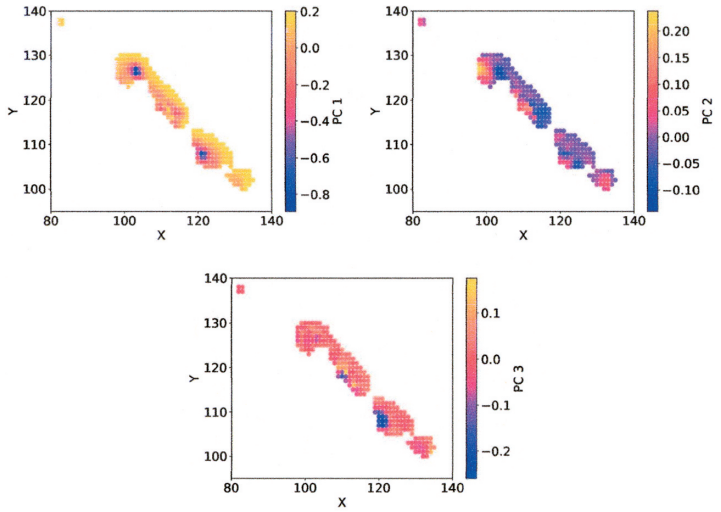


図 18. A-SPCA によって得られた NGC 253 の PC1, PC2, および PC3 の空間構造.

図 14 は，PC1 と PC2 の 2 次元分布を示している．新しい図においては，図 8 に見られる蝶の翅のようなパターンが完全に消失していることがわかる．図 15 および 16 はそれぞれ PC1 と PC3, PC2 と PC3 の散布図である．図 14 と同様，図 15 と 16 に対称なパターンはない．図 17 は PC1, 2, 3 の分布の鳥瞰図である．これらの図から，ドップラーシフト補正済みマップのスペクトルの特徴に明らかな対称パターンはなく，PC 間の散布図は比較的連続的な主要部と外れ値からなることがわかる．

5.2 NGC 253 のスペクトル主成分マップ

NGC 253 の分光マップ上での PC1, 2, 3 の特徴をさらに検証するため，図 18 で各 PC の空間構造[29]を再構成した．この図でも PC1 が輝線強度を表していることが再確認できる．一方ドップラーシフト補正済みマップの PC2 は，図 9 とは対照的に NGC 253 でより複雑で局所

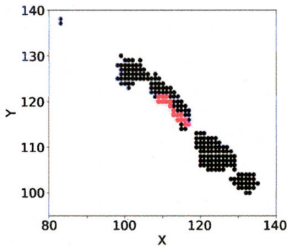


図 19．図 10 マップにおける速度異常がみられる領域．

的なパターンを表現している．PC3 の 2 次元マップには，小スケールの構造も示されている．PC2 が局所的な質量流出を示唆していることを思い出そう．Ando et al.（2017）によると，PC2 の値が最も大きい領域の空間スケール（直径）は ∼ 10 pc であり，確かに空間的に小さいスケールでの現象である．

特に興味深いのは，図 18 の下パネルに見られる PC3 が負の領域である．固有スペクトルは，PC3 および部分的に PC4 が，局所的ではあるがよりスケールの大きいアウトフロー（質量流出）を表しており，輝線の青方偏移[30]に現れていることを示している．Bolatto et al.（2013）は，電波連続波の中心[31]からの分子アウトフローを報告した．最近，Walter et al.（2017）および Krieger et al.（2019）は ALMA によるこの領域の詳細な観測を行った．彼らは詳細な解析により，この質量アウトフローを「SW ストリーマー」と名付けた．このアウトフローは私たちに向かって運動しており，SW ストリーマーの根元の位置は，図 18 に示された PC3 が負の領域と正確に一致している（Krieger et al., 2019, 図 1 を参照）．図 19 の速度異常領域も SW ストリーマーの根本とよく一致する．さらに分子輝線観測だけでなく，ファブリーペロー分光器[32]を用いて Hα[33]やその他の輝線でもアウトフローが観測されている（Matsubayashi et al., 2009）．Hα などの輝線は電離ガスのアウトフローをトレースするが，質量流の位置，形状，方向は分子のアウトフローとかなりよく一致している．これらすべての手がかりは，高次元 PCA が ALMA 分光マップからアウトフロー現象を（物理モデルに依存しないという意味で）純粋に客観的に抽出できたことを示している．

5.3 各 PC に対応するドップラー補正データの特徴的スペクトル

図 20 では，星と三角形は図 11 と同様に PC1 および PC2 に関係する特徴的スペクトルを表し，黒丸は PC3 に関係する特徴である．図 11 と比較すると，まず輝線の系統的なシフトが見られる．もちろん，これは銀河全体の宇宙論的赤方偏移の効果である．ALMA バンド間の境界は観測者の静止系に固定されているため，ドップラーシフト補正後はスペクトルの一部が境界からはみ出す．その結果，図 20 ではスペクトルのかなりの部分が失われている．ドップラー補正後の PC1 は，HCN(4–3) および HNC(4–3) 輝線の強度をより正確に表している．これは，スペクトルの関連する特徴を担う部分がこれらの輝線のピーク付近に集中しているという事実に反映されている．PC2 は HCN(4–3) の輝線中心から少し離れた振動数のプロファイルに割り当てられている．ただし，これは HNC(4–3) 輝線の低周波側に関連しているため，ドップラーシフト補正により低振動数側が切り取られてしまっている．PC3 は HCN(4–3) 輝線の高振動数側のさらに遠いプロファイルと関連するが，PC2 に関連する低振動数側のスペクトルプロファイルとも一部重複している．これらの特性は固有スペクトルから既に示唆されており（図 12），A-SPCA によって得られた詳細な情報で確かめられたといえる．したがって，スペクトルが非常に複雑に見える場合でも，NRPCA や A-SPCA など高次元 PCA を用いれば情報を解読

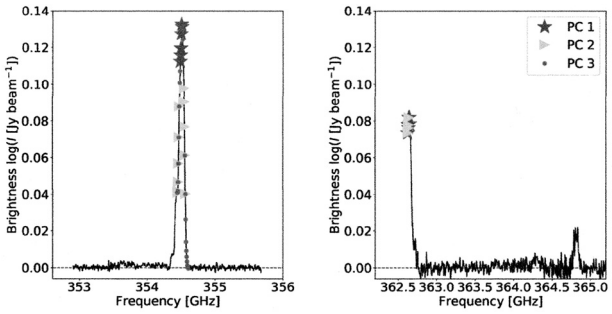


図 20. 系全体の回転によるドップラーシフトについて補正後の ALMA 分光マップから A-SPCA によって求められた，PC1, 2, 3 に対応する特徴的スペクトル．星と三角形は図 11 と同様，PC1 および PC2 に関係する特徴的スペクトルを表し，丸は PC3 に関係する．

し，分光マップを決める特徴的スペクトルを見つけ出すことができる．さらに驚くべきことに，HCN と HNC 輝線の詳細なプロファイルは，アウトフロー現象を含む NGC 253 の ALMA 分光マップのほぼすべてのスペクトル特性を支配している．NGC 253 の ALMA 分光マップによるさらに詳細な天体物理学的素過程のさらなる検証は，今後の研究で行う予定である．

6. まとめと結論

銀河進化は主に星形成によって駆動される．星形成とそれに続く元素合成によって星間物質が進化するので，これが銀河進化を理解する鍵となる．分光観測はこの過程を明らかにする重要な天文学的手法だが，詳細な分光マッピングは多大な時間がかかるため，多くの観測点で独立したデータを得るのは難しい．その結果，分光マップデータは高次元小標本(HDLSS)となる．HDLSS データには高次元特有の困難があり，その解析には古典的方法に代わる新しい統計手法が必要となる．この問題を解決するため，高次元統計解析が発展してきた．

HDLSS データの解析には「次元の呪い」に起因する多くの問題が存在し，古典的統計手法は往々にして正しくない結果を導く．高次元統計解析は双対行列とその幾何学的表現を用いてこの問題を克服し，HDLSS データの特異な振舞いを利用して解析を行う枠組みを提供する．我々は NGC 253 の ALMA 分光マップデータに対して高次元 PCA を適用し，スペクトルマップを特徴づけるスペクトルの特徴成分を特定した．高次元 PCA はモデルを仮定することなく，NGC 253 中心部の系全体の回転を抽出することに成功した．またこうして得られた回転のドップラー効果を補正したデータに高次元 PCA を用いることで，爆発的星形成に起因するガスのアウトフロー現象や速度場異常を示す固有スペクトルが得られた．

これらの結果は，物理的直観を頼りに手作業で特徴を選択するのではなく，高次元データの完全な情報を利用することによって得られたものである．このように，高次元統計解析は将来の宇宙望遠鏡の分光マップ解析に新たな方法論を提供する．また，たとえば極めて稀な天体の分光データを分類する目的においても有用である．別の例として，活動銀河核からの電波連続波を銀河間物質の水素原子が吸収することで生じる HI の森(HI forest)の SKA による観測が挙げられる．我々は希少天体のスペクトルや銀河間物質の吸収線系データに高次元統計解析を適用し，銀河形成の過程解明を目指している．このように，高次元の統計学は天文学の幅広い課題に対して全く新しい扉を開くことになるであろう．

注.

[1] https://www.sdss.org/surveys/.

[2] https://www.desi.lbl.gov/.

[3] https://www.cosmos.esa.int/web/gaia.

[4] https://panstarrs.ifa.hawaii.edu/pswww/.

[5] 天球上で広がった天体の電磁波スペクトルを，天体上の各点で取得する観測をこう呼ぶ.

[6] https://www.eso.org/sci/facilities/paranal/instruments/uves.html.

[7] https://www.ing.iac.es/PR/wht_info/whtsauron.html.

[8] https://www.sdss.org/surveys/manga/.

[9] 爆発的な勢いで星を形成している銀河.

[10] ALMA の電波分光計は，特定の振動数範囲に限ったミリ波電波を検出する．この範囲を band とよぶ.

[11] パーセクは天文学で使われる距離の単位で，1 [pc] は 3.08×10^{18} [cm] に相当する.

[12] 輝線とは，ある量子力学的エネルギー準位から別のエネルギー準位に遷移するときのエネルギー差に相当する波長あるいは振動数の放射のことで，エネルギー差が離散的な値を取ることに対応して狭い波長に集中した放射となる．バンド輝線とは，そのような輝線が比較的狭い波長域に多数存在し，お互いに重なり合って太い輝線となったものを指す.

[13] 電離ガス(プラズマ)相，原子ガス相，分子ガス相などを指す.

[14] 式(2.2)は，任意の $\epsilon > 0$ について $\lim_{d/n \to 0} \mathbb{P}(\|\overline{x} - \vec{\mu}\| > \epsilon) = 0$ を意味する．ここで $\|\cdot\|$ はユークリッドノルムを表し，$\mathbb{P}(A)$ は事象 A の生起確率である．式(2.3)は，任意の $\epsilon > 0$ について $\lim_{d/n \to \infty} \mathbb{P}(\|\overline{x} - \vec{\mu}\| < \epsilon) = 0$ を意味する.

[15] 青嶋・矢田 (2019)を参照のこと.

[16] 従来のスパイクモデルは Johnstone (2001)で与えられ，$\lambda_1, \ldots, \lambda_m$ は 1 よりも大きい定数，$\lambda_{m+1} = \cdots = \lambda_d = 1$ なる非常に限定的なモデルであった．それを高次元データの特徴に合わせて一般化したのがモデル(2.18)である．ただし，一般化スパイクモデル(2.18)は，高次元データ空間の次元数と，それに伴う固有値の大きさ，そして，それを推定するための標本数との関係をみるための簡便なモデルであり，Yata and Aoshima (2013)は，一般化スパイクモデルを拡張したパワースパイクモデルを与え，高次元 PCA の漸近理論を構築している.

[17] 4 節および 5 節で見られるように実高次元データの固有値の最初のいくつかは非常に飛びぬけており，一般化スパイクモデルはその性質を非常によく表している.

[18] この文脈での一致性とは，推定量 $t(n)$ (n: 標本数)が任意の $\epsilon > 0$ について，

$$\lim_{n \to \infty} \mathbb{P}\{|t_n - \theta| < \epsilon\} = 1\,.$$

が成り立つことを指す.

[19] 式(2.20)はガウス分布を緩めた条件である．その条件のもと，ノイズの空間は図 2 に示されるような幾何学的表現をもつ.

[20] もう一つは Yata and Aoshima (2010) で提案されたクロスデータ行列法である．これら A-SPCA のコードについては，https://github.com/Aoshima-Lab/HDLSS-Tools を参照のこと.

[21] Takeuchi et al. (2024)において，いくつかの ω の値で検証している.

[22] Heliocentric velocity: 太陽を基準として測定した視線速度．銀河系内の太陽系の運動速度ベクトルを差し引いていない.

[23] ここで $M_\odot$ は太陽質量，すなわち太陽の質量 1.99×10^{33} [g] を 1 [$M_\odot$] とする単位．
[24] 活発に星が形成される領域でみられる巨大な星の集団．
[25] 宇宙膨張による輝線のシフトは厳密には速度によるものではないが，近傍銀河については近似的に速度とみなすことができる．これをハッブル速度とよぶ．
[26] 分光マップのデータは画像の空間方向 2 次元＋振動数分散方向 1 次元の 3 次元構造を持つため，データキューブと呼ばれている．
[27] 連続放射（輝線ではなく，広い振動数域にわたって現れる放射）の存在は状況をさらに複雑にするが，本研究ではこの問題は扱わない．
[28] 分子の熱運動によるドップラー効果により輝線の振動数が変わり，静止系振動数の周りに輝線が広がる効果をドップラーブロードニングという．この効果で広がった輝線の周辺部をウィングとよぶ．
[29] ここでの空間構造とは，天球上の 2 次元マップ上での各量の位置依存性を意味する．
[30] 放射する物体が我々に向かって進むとき，ドップラー効果によって放射の波長が青い方（波長の短い側，振動数の高い側）にシフトすることを指す．
[31] 輝線ではなく，電波の波長（振動数）について連続的な放射の観測画像で最も明るい領域をこのように表現する．
[32] 電磁波の干渉を利用した面分光装置の一種．
[33] 水素が電離して正の電荷をもったイオン（この場合は 1 個の陽子）になっているとき，電子が結合すると余剰分のエネルギーを放射してより安定な状態に遷移する．水素原子内の電子のエネルギーは量子力学によって決まる離散的準位を取る．電子は高いエネルギーから低いエネルギーに遷移していき，エネルギー準位の差に応じた輝線を放射する．これが水素の再結合線で，準位 3 から 2 への遷移によって放射されるのが Hα 線と呼ばれる．Hα 線は可視光線の赤に対応する波長を持つ．

謝　辞

本論文の査読者には大変有益なコメントおよび議論をいただき，論文の可読性と議論の明快さを改善することができた．ここに心より感謝する．本論文は統計数理特集号『諸科学における統計数理モデリングの拡がり II』のために執筆したが，責任著者のスケジュール上の都合により投稿が〆切に間に合わなかったため，一般論文として投稿したものである．担当編集者の島谷健一郎氏の示唆に深く感謝する．この論文では ALMA データ（ADS/JAO.ALMA#2013.1.00099.S, ADS/JAO.ALMA#2013.1.00735.S）を使用している．ALMA はチリ共和国の協力のもと ESO, NSF，NINS，NRC，MOST および ASIAA，KASI によって維持される共同研究機関である．ALMA 天文台，ESO，AUI/NRAO，および国立天文台によって運営されている．本研究は日本学術振興会科学研究費補助金（21H01128，24H00247，19K03937，および JP17H06130）のサポートを受けている．本研究の一部は，住友財団平成 30 年度基礎科学研究費補助金（180923），統計数理研究所共同利用研究費「データサイエンスによる銀河進化研究の新展開」の支援も受けて行った．

河野海氏には本研究の初期の解析に大きな貢献をいただいた．また内容について有意義な議論をしていただいた中山優吾氏，池田思朗氏，原田ななせ氏に感謝する．

付　録

A. 電波領域における分子スペクトル輝線の放射機構

スペクトル輝線の放射機構は基本的に量子力学的効果である．原子の大きさ程度以下のミクロなスケールでは，素粒子の運動は古典力学ではなく量子力学で記述される．古典力学との顕著な違いは，エネルギーなど物理量が離散的な値を取ることである．これを「量子化されている」と表現する．その詳細は文献に譲るが（たとえば Wilson et al., 2013），離散化されたエネルギーの値をエネルギー準位と呼び，その間で量子力学的遷移が生じるとエネルギー差 ΔE に対応する波長

$$\lambda = \frac{hc}{\Delta E} \tag{A.1}$$

の輝線が放射される（h はプランク定数，c は光速）．

分子は原子に比べて複雑な構造をしており，分子の振動や回転にともなう複数の量子状態間の遷移によって電磁波を放射する．分子の振動準位間の遷移にともなう放射は多くの場合赤外線領域となるが，回転遷移では電波領域の放射も多数あり，電波天文学の重要な観測対象である．星間空間，特に低温の領域では，H_2，CO，OH，SiO などの 2 原子分子が存在する．古典力学的では，分子の重心周りの回転エネルギー $E_{\rm rot}$ は

$$E_{\rm rot} = \frac{\omega^2}{2}\left(m_1 r_1^2 + m_2 r_2^2\right) \tag{A.2}$$

と表される．ここで，m_1，m_2 は 2 つの原子の質量，r_1，r_2 は分子の重心からそれぞれの原子までの距離，ω は回転角速度である．一方，2 原子間の距離を r とすると

$$r_1 = \frac{m_2}{m_1+m_2}\, r,\ r_2 = \frac{m_1}{m_1+m_2}\, r \tag{A.3}$$

であることを用いると

$$E_{\rm rot} = \frac{\omega^2}{2}\left(\frac{m_1 m_2}{m_1+m_2}\right) r^2 \equiv \frac{m_{\rm red}}{2} r^2 \omega^2 \tag{A.4}$$

とできる．ここで $m_{\rm red}$ は換算質量と呼ばれる量

$$m_{\rm red} = \frac{m_1 m_2}{m_1+m_2} \tag{A.5}$$

である．さらに慣性モーメント I を

$$I \equiv m_{\rm red} r^2 \tag{A.6}$$

で定義すると，

$$E_{\rm rot} = \frac{I\omega^2}{2} \tag{A.7}$$

と表される．また回転の角運動量 $L_{\rm rot}$ は

$$L_{\rm rot} = I\omega \tag{A.8}$$

なので，

$$E_{\rm rot} = \frac{L_{\rm rot}^2}{2I} \tag{A.9}$$

が得られる．

一方，量子力学では $L^2_{\rm rot}$ を角運動量演算子

$$\hat{L}^2 \equiv -\hbar^2 \left\{ \frac{1}{\sin\theta} \left(\frac{\partial}{\partial\theta} \right) \left[\sin\theta \left(\frac{\partial}{\partial\theta} \right) \right] + \frac{1}{\sin^2\theta} \left(\frac{\partial^2}{\partial\varphi^2} \right) \right\} \tag{A.10}$$

で置き換えることにより，エネルギーの演算子であるハミルトニアン $\hat{H}$ を作り，シュレーディンガー方程式

$$E_{\rm rot}\psi = \hat{H}\psi = \frac{\hat{L}^2}{2I}\psi \tag{A.11}$$

を構成する ($\hbar \equiv h/2\pi$)．変数 (θ, φ) は 3 次元球座標系における角度変数であり，ψ は系の状態を記述する波動関数と呼ばれる関数である．

量子的な状態は確率で表現されるので，その力学的考察のためには式(A.11)を解いて ψ を求めればよい．演算子 $\hat{L}^2$ の固有値は $J = 0, 1, 2, \ldots$ として $\hbar^2 J(J+1)$ であるから，回転運動のエネルギー固有値 E_J は

$$E_J = \frac{\hbar^2}{2I} J(J+1) \tag{A.12}$$

となる．これが分子の回転のエネルギー準位の具体的表式である．本文で分子輝線の名前を HCN(4–3) のように表記しているが，これは分子が HCN で，回転準位 $J = 4 \longrightarrow 3$ の遷移にともなって放出された輝線であることを意味している．

B. ドップラーシフト(Doppler shift)

本章では物体の運動によって生じる光の波長の偏移，いわゆるドップラーシフト(Doppler shift)について説明する．この効果を起こす物理を説明するのは特殊相対性理論である．天体が観測者に対して速度 v(ここでは遠ざかる方向を正と定義する)で等速直線運動しているとし，θ は視線方向と運動方向のなす角を表すとする．

天体が観測者に対して静止しているとき，天体から観測者へ伝播する波長 λ の電磁波を考える．天体と観測者間の距離を d とすると，時間間隔 $t = d/c$ の間に $n = d/\lambda$ 個の波が放出される．一方，天体が観測者に対して速度 v で運動している場合，天体と観測者間の距離は時間間隔 t ののち

$$d' = d + (v\cos\theta)t \tag{B.1}$$

に増加している．ここで波の数 n が保存することに注意すれば，

$$\frac{\lambda'}{\lambda} = \frac{d + (v\cos\theta)t}{d} = 1 + \beta\cos\theta\,, \tag{B.2}$$

$$\beta \equiv \frac{v}{c} \tag{B.3}$$

が得られる．波長と振動数の間には $\lambda\nu = c$ の関係があるので，式(B.2)を振動数の関係式に換算すると

$$\frac{\nu'}{\nu} = \frac{1}{1 + \beta\cos\theta} \tag{B.4}$$

となる．これは，視線方向の運動によって電磁波の波長が引き伸ばされる効果である．

特殊相対論では，観測者から見ると運動している天体の時間は遅れる．つまり，時計の進み方は $\sqrt{1-\beta^2}$ 倍になり，振動数もこれに比例して小さくなる．すなわち，

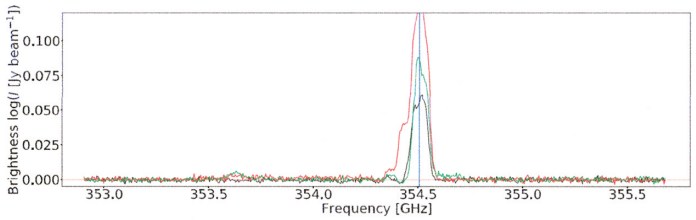


図 21. ドップラーシフト補正された分子輝線 HCN(4–3) の例．赤，緑，青は，線の中心がその静止系振動数と一致するように補正されたスペクトルである（Takeuchi et al., 2024）．

(B.5) $$\frac{\nu''}{\nu} = \sqrt{1-\beta^2}$$

である．この効果は天体の運動速度の大きさのみによるので，天体が視線に直交する方向に運動している場合 ($\theta = 2\pi$ [rad]) でも振動数が変化する．これを横ドップラー効果（transversal Doppler effect）とよぶ．

これら 2 つの効果により，最終的に特殊相対論的なドップラー効果による振動数変化は

(B.6) $$\nu_{\rm obs} = \frac{\sqrt{1-\beta^2}}{1+\beta\cos\theta}\,\nu_{\rm em}$$

と書ける．ここで，$\nu_{\rm em}$ は静止系で観測される振動数であり，$\nu_{\rm obs}$ は観測される振動数である．視線方向のみの効果を考えるならば（$\theta = 0$），

(B.7) $$\nu_{\rm obs} = \frac{\sqrt{1-\beta^2}}{1+\beta}\,\nu_{\rm em} = \sqrt{\frac{1-\beta}{1+\beta}}\,\nu_{\rm em}$$

となり，波長では

(B.8) $$\lambda_{\rm obs} = \sqrt{\frac{1+\beta}{1-\beta}}\,\lambda_{\rm em}$$

が求める関係である．下付き添え字 obs，em については式（B.7）と同様である．

本文 5 章で議論しているように，詳細な物理過程を考察するためには，特殊相対論的ドップラーシフトを補正して静止系での振動数に戻すことが必要である．これには式（B.7）を用いればよい．具体例として，HCN(4–3) 輝線のドップラー補正を図 21 に示す．赤，緑，青は NGC 253 のマップ上の異なる位置から補正されたスペクトルであり，輝線の中心がその静止系振動数と一致する．各輝線は複雑なプロファイルを持っているため，輝線の中心はガウス関数フィッティングによって決定している．青い縦の実線は，HCH(4–3) 輝線の静止系振動数 354.5 GHz を示している．

参 考 文 献

Ahn, J., Marron, J. S., Muller, K. M. and Chi, Y.-Y. (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions, *Biometrika*, **94**(3), 760–766.

Ando, R., Nakanishi, K., Kohno, K., Izumi, T., Martín, S., Harada, N., Takano, S., Kuno, N., Nakai, N., Sugai, H., Sorai, K., Tosaki, T., Matsubayashi, K., Nakajima, T., Nishimura, Y. and Tamura, Y. (2017). Diverse nuclear star-forming activities in the heart of NGC 253 resolved with 10-pc-scale

ALMA images, *The Astrophysical Journal*, **849**(2), 81, http://dx.doi.org/10.3847/1538-4357/aa8fd4.

Aoshima, M. and Yata, K. (2011). Two-stage procedures for high-dimensional data, *Sequential Analysis*, **30**(4), 356–399, http://dx.doi.org/10.1080/07474946.2011.619088.

青嶋 誠, 矢田和善 (2013). 論説：高次元小標本における統計的推測, 数学, **48**, 225–247.

Aoshima, M. and Yata, K. (2014). A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data, *Annals of the Institute of Statistical Mathematics*, **66**, 983–1010, http://dx.doi.org/10.1007/s10463-013-0435-8.

Aoshima, M. and Yata, K. (2015). Geometric classifier for multiclass, high-dimensional data, *Sequential Analysis*, **34**(3), 279–294, http://dx.doi.org/10.1080/07474946.2015.1063256.

Aoshima, M. and Yata, K. (2017). Statistical inference for high-dimension, low-sample-size data, *Sugaku Expositions*, **30**(2), 137–158, http://dx.doi.org/10.1090/suga/421.

Aoshima, M. and Yata, K. (2019). High-dimensional quadratic classifiers in non-sparse settings, *Methodology and Computing in Applied Probability*, **21**, 663–682, http://dx.doi.org/10.1007/s11009-018-9646-z.

青嶋 誠, 矢田和善 (2019).『高次元の統計学』, 共立出版, 東京.

Aoshima, M., Shen, D., Shen, H., Yata, K., Zhou, Y.-H. and Marron, J. S. (2018). A survey of high dimension low sample size asymptotics, *Australian and New Zealand Journal of Statistics*, **60**, 4–19, http://dx.doi.org/10.1111/anzs.12212.

Baik, J. and Silverstein, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models, *Journal of Multivariate Analysis*, **97**(6), 1382–1408, http://dx.doi.org/10.1016/j.jmva.2005.08.003.

Baik, J., Ben Arous, G. and Péché, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices, *Annals of Probability*, **33**(5), 1643–1697, http://dx.doi.org/10.1214/009117905000000233.

Bendo, G. J., Beswick, R. J., D'Cruze, M. J., Dickinson, C., Fuller, G. A. and Muxlow, T. W. B. (2015). ALMA observations of 99 GHz free-free and H40α line emission from star formation in the centre of NGC 253, *Monthly Notices of the Royal Astronomical Society*, **450**(1), L80–L84, http://dx.doi.org/10.1093/mnrasl/slv053.

Bolatto, A. D., Warren, S. R., Leroy, A. K., Walter, F., Veilleux, S., Ostriker, E. C., Ott, J., Zwaan, M., Fisher, D. B., Weiss, A., Rosolowsky, E. and Hodge, J. (2013). Suppression of star formation in the galaxy NGC 253 by a starburst-driven molecular wind, *Nature*, **499**(7459), 450–453, http://dx.doi.org/10.1038/nature12351.

Fernández-Ontiveros, J. A., Prieto, M. A. and Acosta-Pulido, J. A. (2009). The nucleus of NGC 253 and its massive stellar clusters at parsec scales, *Monthly Notices of the Royal Astronomical Society*, **392**(1), L16–L20, http://dx.doi.org/10.1111/j.1745-3933.2008.00575.x.

Galaz, G. and de Lapparent, V. (1998). The ESO-Sculptor Survey: Spectral classification of galaxies with Z $<$ 0.5, *Astronomy & Astrophysics*, **332**, 459–478.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science*, **286**(5439), 531–537, http://dx.doi.org/10.1126/science.286.5439.531.

Hall, P., Marron, J. S. and Neeman, A. (2005). Geometric representation of high dimension, low sample size data, *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **67**(3), 427–444, http://dx.doi.org/10.1111/j.1467-9868.2005.00510.x.

Hand, D. J., Daly, F., McConway, K., Lunn, D. and Ostrowski, E. (1994). *A Handbook of Small Data Sets*, 1st edition, Chapman and Hall, London.

Ishii, A., Yata, K. and Aoshima, M. (2016). Asymptotic properties of the first principal component and equality tests of covariance matrices in high-dimension, low-sample-size context, *Journal of*

Statistical Planning and Inference, **170**, 186–199, http://dx.doi.org/10.1016/j.jspi.2015.10.007.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis, *Annals of Statistics*, **29**(2), 295–327, http://dx.doi.org/10.1214/aos/1009210544.

Jung, S. and Marron, J. S. (2009). PCA consistency in high dimension, low sample size context, *Annals of Statistics*, **37**(6B), 4104–4130, http://dx.doi.org/10.1214/09-AOS709.

Kamphuis, P., Józsa, G. I. G., Oh, S. H., Spekkens, K., Urbancic, N., Serra, P., Koribalski, B. S. and Dettmar, R. J. (2015). Automated kinematic modelling of warped galaxy discs in large H I surveys: 3D tilted-ring fitting of H I emission cubes, *Monthly Notices of the Royal Astronomical Society*, **452**(3), 3139–3158, http://dx.doi.org/10.1093/mnras/stv1480.

Keto, E., Hora, J. L., Fazio, G. G., Hoffmann, W. and Deutsch, L. (1999). A super-star cluster in NGC 253: Mid-infrared properties, *The Astrophysical Journal*, **518**(1), 183–189, http://dx.doi.org/10.1086/307246.

Krieger, N., Bolatto, A. D., Walter, F., Leroy, A. K., Zschaechner, L. K., Meier, D. S., Ott, J., Weiss, A., Mills, E. A. C., Levy, R. C., Veilleux, S. and Gorski, M. (2019). The molecular outflow in NGC 253 at a resolution of two parsecs, *The Astrophysical Journal*, **881**(1), 43, http://dx.doi.org/10.3847/1538-4357/ab2d9c.

Martín, S., Mangum, J. G., Harada, N., Costagliola, F., Sakamoto, K., Muller, S., Aladro, R., Tanaka, K., Yoshimura, Y., Nakanishi, K., Herrero-Illana, R., Mühle, S., Aalto, S., Behrens, E., Colzi, L., Emig, K. L., Fuller, G. A., García-Burillo, S., Greve, T. R., Henkel, C., Holdship, J., Humire, P., Hunt, L., Izumi, T., Kohno, K., König, S., Meier, D. S., Nakajima, T., Nishimura, Y., Padovani, M., Rivilla, V. M., Takano, S., van der Werf, P. P., Viti, S. and Yan, Y. T. (2021). ALCHEMI, an ALMA comprehensive high-resolution extragalactic molecular inventory. Survey presentation and first results from the ACA array, *Astronomy & Astrophysics*, **656**, A46, http://dx.doi.org/10.1051/0004-6361/202141567.

Matsubayashi, K., Sugai, H., Hattori, T., Kawai, A., Ozaki, S., Kosugi, G., Ishigaki, T. and Shimono, A. (2009). Galactic wind in the nearby starburst galaxy NGC 253 observed with the Kyoto3DII Fabry-Perot mode, *The Astrophysical Journal*, **701**(2), 1636–1643, http://dx.doi.org/10.1088/0004-637X/701/2/1636.

Pace, Z. J., Tremonti, C., Chen, Y., Schaefer, A. L., Bershady, M. A., Westfall, K. B., Boquien, M., Rowlands, K., Andrews, B., Brownstein, J. R., Drory, N. and Wake, D. (2019). Resolved and integrated stellar masses in the SDSS-IV/MaNGA survey. I. PCA spectral fitting and stellar mass-to-light ratio estimates, *The Astrophysical Journal*, **883**(1), 82, http://dx.doi.org/10.3847/1538-4357/ab3723.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model, *Statistica Sinica*, **17**(4), 1617–1642.

Portillo, S. K. N., Parejko, J. K., Vergara, J. R. and Connolly, A. J. (2020). Dimensionality reduction of SDSS spectra with variational autoencoders, *The Astrophysical Journal*, **160**(1), 45, http://dx.doi.org/10.3847/1538-3881/ab9644.

Rekola, R., Richer, M. G., McCall, M. L., Valtonen, M. J., Kotilainen, J. K. and Flynn, C. (2005). Distance to NGC 253 based on the planetary nebula luminosity function, *Monthly Notices of the Royal Astronomical Society*, **361**(1), 330–336, http://dx.doi.org/10.1111/j.1365-2966.2005.09166.x.

Rieke, G. H., Lebofsky, M. J., Thompson, R. I., Low, F. J. and Tokunaga, A. T. (1980). The nature of the nuclear sources in M82 and NGC 253, *The Astrophysical Journal*, **238**, 24–40, http://dx.doi.org/10.1086/157954.

Ronen, S., Aragon-Salamanca, A. and Lahav, O. (1999). Principal component analysis of synthetic galaxy spectra, *Monthly Notices of the Royal Astronomical Society*, **303**(2), 284–296, http://dx.doi.org/10.1046/j.1365-8711.1999.02222.x.

Sakamoto, K., Mao, R.-Q., Matsushita, S., Peck, A. B., Sawada, T. and Wiedner, M. C. (2011). Star-

forming cloud complexes in the central molecular zone of NGC 253, *The Astrophysical Journal*, **735**(1), 19, http://dx.doi.org/10.1088/0004-637X/735/1/19.

Takeuchi, T. T., Yata, K., Egashira, K., Aoshima, M., Ishii, A., Cooray, S., Nakanishi, K., Kohno, K. and Kono, K. T. (2024). High-dimensional statistical analysis and its application to an ALMA map of NGC 253, *The Astrophysical Journal Supplement Series*, **271**(2), 44, http://dx.doi.org/10.3847/1538-4365/ad2517.

Walter, F., Bolatto, A. D., Leroy, A. K., Veilleux, S., Warren, S. R., Hodge, J., Levy, R. C., Meier, D. S., Ostriker, E. C., Ott, J., Rosolowsky, E., Scoville, N., Weiss, A., Zschaechner, L. and Zwaan, M. (2017). Dense molecular gas tracers in the outflow of the starburst galaxy NGC 253, *The Astrophysical Journal*, **835**(2), 265, http://dx.doi.org/10.3847/1538-4357/835/2/265.

Wang, L., Farrah, D., Connolly, B., Connolly, N., Lebouteiller, V., Oliver, S. and Spoon, H. (2011). Principal component analysis of the Spitzer IRS spectra of ultraluminous infrared galaxies, *Monthly Notices of the Royal Astronomical Society*, **411**(3), 1809–1818, http://dx.doi.org/10.1111/j.1365-2966.2010.17811.x.

Wilson, T. L., Rohlfs, K. and Hüttemeister, S. (2013). *Tools of Radio Astronomy*, Springer Verlag, Heidelberg, http://dx.doi.org/10.1007/978-3-642-39950-3.

Yata, K. and Aoshima, M. (2009). PCA consistency for non-Gaussian data in high dimension, low sample size context, *Communications in Statistics — Theory and Methods*, **38**(16-17), 2634–2652, http://dx.doi.org/10.1080/03610910902936083.

Yata, K. and Aoshima, M. (2010). Effective PCA for high-dimension, low-sample-size data with singular value decomposition of cross data matrix, *Journal of Multivariate Analysis*, **101**(9), 2060–2077, http://dx.doi.org/10.1016/j.jmva.2010.04.006.

Yata, K. and Aoshima, M. (2012). Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations, *Journal of Multivariate Analysis*, **105**(1), 193–215, http://dx.doi.org/https://doi.org/10.1016/j.jmva.2011.09.002.

Yata, K. and Aoshima, M. (2013). PCA consistency for the power spiked model in high-dimensional settings, *Journal of Multivariate Analysis*, **122**, 334–354, http://dx.doi.org/10.1016/j.jmva.2013.08.003.

Yata, K. and Aoshima, M. (2022). Automatic sparse PCA for high-dimensional data, *Statistica Sinica* (in press), http://dx.doi.org/10.5705/ss.202022.0319.

Proceedings of the Institute of Statistical Mathematics Vol. 72, No. 2, 273–303 (2024) 303

Exploration of the Physical Properties of Molecular Gas in a Galaxy with High-dimensional Statistical Analysis and Future Prospects to Astronomy

Tsutomu T. Takeuchi[1,2], Kazuyoshi Yata[3], Kento Egashira[4], Makoto Aoshima[3], Kohji Yoshikawa[5], Aki Ishii[4], Ryusei R. Kano[1], Wen E. Shi[1], Aina May So[1,8], Hai-Xia Ma[1], Sena A. Matsui[1], Koichiro Nakanishi[6,9], Sucheta Cooray[6,7] and Kotaro Kohno[10]

[1]Division of Particle and Astrophysical Science, Nagoya University
[2]Research Center for Statistical Machine Learning, The Institute of Statistical Mathematics
[3]Institute of Mathematics, University of Tsukuba
[4]Department of Information Sciences, Tokyo University of Science
[5]Center for Computational Sciences, University of Tsukuba
[6]National Astronomical Observatory of Japan
[7]Research Fellow of the JSPS (PD)
[8]Department of Physics, Gakushuin University
[9]Department of Astronomy, School of Science, Graduate University for Advanced Studies (SOKENDAI)
[10]Institute of Astronomy, Graduate School of Science, The University of Tokyo

If we denote the dimension of data as d and the number of samples as n, we often meet a case with $n \ll d$. Traditionally in astronomy, such a situation is regarded as ill-posed, and they thought that there was no choice but to throw away most of the information in data dimension to let $d < n$. The data with $n \ll d$ is referred to as high-dimensional low sample size (HDLSS). To deal with HDLSS problems, a method called high-dimensional statistics has been developed rapidly in the last decade. In this work, we first introduce the high-dimensional statistical analysis. We apply two representative methods in the high-dimensional statistical analysis methods, the noise-reduction principal component analysis (NRPCA) and automatic sparse principal component analysis (A-SPCA), to a spectroscopic map of a nearby archetype starburst galaxy NGC 253 taken by the Atacama Large Millimeter/Submillimeter Array (ALMA). The ALMA map is a typical HDLSS dataset. First we analyzed the original data including the Doppler shift due to the systemic rotation. The high-dimensional PCA could describe the spatial structure of the rotation precisely. We then applied to the Doppler-shift corrected data to analyze more subtle spectral features. The NRPCA and A-SPCA could quantify the very complicated characteristics of the ALMA spectra. Particularly, we could extract the information of the global outflow from the center of NGC 253. This method can also be applied not only to spectroscopic survey data, but also any type of HDLSS data.

Key words: High-dimensional statistical analysis, principal component analysis, interstellar medium, molecular emission line, starburst, galaxy evolution.

2024 年 11 月改訂

「統計数理」投稿規程

1. 「統計数理」は，統計科学の深化と発展，そして統計科学を通じた社会への貢献を目指すものである．投稿原稿は，統計科学に関連した内容を持つもので，和文の原稿に限る．
2. 投稿原稿は次の 6 種とする．
   a. 原著論文 (Paper)
      統計科学の発展に貢献すると考えられる研究結果．
   b. 総合報告 (Review Article)
      特定の主題に関する一連の研究およびその周辺領域の発展を著者の見解に従って総括的，かつ体系的に報告したもの．
   c. 研究ノート (Letter)
      研究速報，新しい発想，提言，問題提起，事例報告など研究上，記録にとどめておく価値があると認められるものや，既発表の論文等に対するコメントで，研究上，記録にとどめておく価値があると認められるもの．
   d. 研究詳解 (Research Review)
      特定の研究領域における理論的あるいは応用的成果を，最近の結果や知見を加えてわかりやすく説明したもの．
   e. 統計ソフトウェア (Statistical Software)
      有用な計算法や解析法に関する短いプログラムおよびサブルーチンのリスト，利用手引き，実行例など．
   f. 研究資料 (Research Archives)
      歴史的なデータ，入手困難なデータや統計的手法の比較検討のために有用なデータ，あるいは，歴史的文献の翻訳や解説など．
   いずれも原則として，未発表のものに限る．
3. 投稿された原稿は，編集委員会が選定・依頼した査読者の審査を経て，掲載の可否を決定する．
4. 投稿原稿は電子投稿査読システム https://www.editorialmanager.com/toukei/ より投稿するものとする．原稿は pdf ファイルとし，必要なフォントはすべて埋め込み，原稿全体を一つのファイルにまとめることとする．論文が採択になった場合，著者は最終稿のソースファイルとハードコピーを提出するものとする．
5. 著作権
   (1) 掲載される論文等の著作権はその採択をもって統計数理研究所に帰属するものとする．統計数理研究所は，紙媒体の「統計数理」のほか電子媒体などを通じて論文等を公表することができる．特別な事情がある場合は，著者と本編集委員会との間で協議の上措置する．
   (2) 投稿原稿の中で引用する文章や図表の著作権に関する問題は，著者の責任において処理する．
   (3) 著者が自分の論文等を複製，転載，翻訳，翻案等の形で利用するのは自由である．この場合，著者は掲載先に出典を明記する．
6. 原稿は次の執筆要項に従って作成する．

「統計数理」執筆要項

1. 原稿は A4 用紙に 1 行 36 字から 40 字で 1 行おき，1 頁あたり 22 行程度とする．原稿の長さは原則として表・図を含めて 30 頁相当以内とし，各ページにページ番号を付す．図表は別紙にまとめ，本文中には挿入箇所のみを指定する．LaTeX で原稿を作成する場合は，「統計数理」スタイルファイルの使用を推奨する．
   https://www.ism.ac.jp/editsec/toukei/
2. 文体は「である体」とし，句読点は「,」「.」を用いる．
3. 原稿は以下の順に書くものとする．
   ［第 1 頁］ 標題，著者名，所属名および所在地，メールアドレス，和文要旨 (500 字程度，文献の引用および数式は原則として避ける)，和文キーワード (6 語以内)．
   ［第 2 頁］ 英語による標題，著者名，所属名，Abstract (450 ワード程度)，Key words (6 words and phrases 以内)．Abstract は，問題の所在と得られた結果等がそれだけで理解できるようなものとする．

［第 3 頁以降］

① 本文：章，節の番号は，第 1 章にあたるものは，“1.”, 第 1 章 第 1 節にあたるものは，“1.1” というようにつける．また，式の番号は，章ごとに (2.1), (2.2) のようにし，式の左側に配置する．

② 数式：数式は簡明さを心がけ，添字にさらに添字をつけるのはなるべく避ける．

③ 参考文献：書き方は本要項 第 4 項を参照．

④ 表：一枚の用紙に一つの表を書く．表の番号は論文中に現れる順に従って，表 1，表 2,... または，Table 1, Table 2,... のようにする．

⑤ 図：一枚の用紙に一つの図を描く．図はそのまま写真製版できる鮮明なものを用意する．大きさは印刷出来上がりの 1〜2 倍とし，トレースが必要な場合は原則として著者が行うものとする．図の番号は論文中に現れる順に従って，図 1，図 2,... または，Fig. 1, Fig. 2,... のようにする．図は原則としてモノクロ印刷とするが，カラー印刷を必要とする場合は編集委員会に相談すること．

⑥ 注：本文中の注釈は極力避ける．やむを得ず注釈をつける場合は脚注とせず，論文末尾に後注とする．後注は，順番に “1, 2,...” の番号を付け，本文中では上付きで示す．

4. 本文中での参考文献の引用は，著者名 (出版年) とする．たとえば，Efron (1982), 清水・湯浅 (1984), Cox and Snell (1981), 坂元 他 (2004), Nakano et al. (2000).

5. 参考文献の書き方

［雑誌の場合］いずれの場合も雑誌名は省略しないものとする．

① ページ番号がある文献

著者名 (出版年)．標題，雑誌名，巻，ページ〔始–終〕.

【例】Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically, *Journal of the American Statistical Association*, **68**, 361–368.

【例】Bligh, E. G., Dyer, W. J. (1959). A rapid method of total lipid extraction and purification, *Canadian Journal of Biochemistry and Physiology*, **37**, 911–917, https://doi.org/10.1139/o59-099.

② ページ番号がない文献

著者名 (出版年)．標題，雑誌名，巻，ページ〔始–終〕．DOI があるものは DOI を入れ，DOI がない場合は URL と最終アクセス日を記載する．

【例】Lien, G. Y., Kalnay, E. and Miyoshi, T. (2013). Effective assimilation of global precipitation: Simulation experiments, *Tellus A,* **65***,* https://doi.org/10.1175/WAF-D-13-00032.1.

【例】Kiraly, F. J. and Oberhauser, H. (2019). Kernels for sequentially ordered data, *Journal of Machine Learning Research*, **20**(31), http://jmlr.org/papers/v20/16-314.html (最終アクセス日 2023 年 10 月 1 日)．

［叢書の中の一巻の場合］

著者名 (出版年)．書名 (編集者名)，叢書名，発行所名，発行地名．

【例】Sakamoto, Y., Ishiguro, M. and Kitagawa, G. (1983). *Akaike Information Criterion Statistics*, Mathematics and Its Applications, Reidel, Dordrecht.

［単行本等の場合］

著者名 (出版年)．書名，発行所名，発行地名．

【例】Cressie, Noel (1993). *Statistics for Spatial Data*, Wiley, New York.

［編集書の中の一部の場合］

著者名 (出版年)．標題，編集書名 (編集者名)，巻，ページ，発行所名，発行地名．

【例】Akaike, H. (1980). Likelihood and the Bayes procedure, *Bayesian Statistics* (eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), 143–166, University Press, Valencia, Spain.

なお，同じ著者によるものが同一年に複数個現れる場合には，(1980a), (1980b) などとして区別する．文献は，日本人も含め，著者名のアルファベット順に並べる．

6. 著者校正は原則として一回とする．その際，印刷上の誤り以外の字句や図版の訂正，挿入，削除等は原則として認めない．