製表技術参考資料 42

ISSN　1348－7345
ISSN－L 1348－7329

情報ソースの多様化に対応するための研究

―ビッグデータの利用を中心にして―

NSTAC

Working Paper No.42

令和３年６月

独立行政法人　統計センター

製表技術参考資料は、独立行政法人 統計センターの職員が、その業務に関連して行った製表技術に関する研究の結果を紹介するためのものである。

ただし、本資料に示された見解は、執筆者の個人的見解である。

目次

情報ソースの多様化に対応するための研究

―ビッグデータの利用を中心にして―

田村　義保　　武藤　杏里　松本　正博 *

要旨

本報告は、統計センターアクションプラン「情報ソースの多様化に対応するための研究」並びに令和２年度統計センター中期研究方針及び中期研究計画に基づき収集した情報の報告と分析の取り組みについてまとめたものである。

現在、政府全体のビッグデータや行政記録情報の公的統計への活用の検討が必ずしも進んでいない状況を踏まえ、当面は基礎的な研究を進めることとしており、ビッグデータ等を統計作成に活用する可能性について、内外の事例の研究、技術的な課題の研究を行うこととしていた。

そこで、本報告では、国内外におけるビッグデータ等を統計に活用した（あるいはその試みの）事例等の他、その事例を分析するための、ビッグデータの特徴及び統計に活用するための利点や課題についての研究事例、並びにビッグデータ等の所在に関する情報を取集している。

* 独立行政法人統計センター統計技術・提供部技術研究開発課

情報ソースの多様化に対応するための研究 †

―ビッグデータの利用を中心にして―

田村　義保　　武藤　杏里　　松本　正博

1 はじめに

本報告は、統計センターアクションプラン「情報ソースの多様化に対応するための研究」並びに令和２年度統計センター中期研究方針及び中期研究計画に基づき収集した情報の報告と分析の取り組みについてまとめたものである。令和２年度統計センター中期研究方針及び中期研究計画による研究内容、スケジュール等は下記のとおりである。

1.1 研究内容

公的統計は、これまでは主に統計調査により得られた調査票を基に作成されてきたところであるが、統計調査への回答者の負担軽減の観点や調査環境の悪化等による回答内容の品質低下の観点から、統計調査以外の方法で得られる情報も活用して統計を作成していく必要性が高まっている。加えて、近年の情報のデジタル化の流れと相まって、統計作成にビッグデータや行政記録情報を活用できる可能性も高まっているところである。

現在、政府全体の検討が必ずしも進んでいない状況を踏まえ、当面は基礎的な研究を進めることとする。統計センターとしては、「情報ソースの多様化に対応するための研究」 として、データエディティングへの活用なども視野に入れつつ、ビッグデータや行政記録情報を統計作成に活用する可能性について、内外の事例の研究、技術的な課題の研究を行うこととする。

1.2 具体的な研究の内容

今後、情報の入手方法・入手技術、情報の質等に関する検討を進めるために、国内外におけるビッグデータ、行政記録情報の活用状況を収集し整理する。

どのようなビッグデータ、行政記録情報を統計作成に活用できるか、活用するための課題の抽出と解決に向けた研究を行う。

まず、国内において統計作成にビッグデータを活用している事例を収集し、収集した情報を継続的に整理する。続いて、諸外国の事例を収集する。

1.3 スケジュール

［令和２年度］国内において統計作成にビッグデータを活用している事例の研究を行う。諸外国において統計作成にビッグデータを活用している事例の研究を行う。統計作成にビッグデータを活用している事例の研究の取りまとめ（報告書作成）を行う。国内において統計作成に行政記録情報を活用している事例の研究を行う。

---

† 本論文に示された意見・見解は執筆者の個人的見解であり、統計センターの見解を示すものではない。

［令和 3 年度］諸外国において統計作成に行政記録情報を活用している事例の研究を行う。統計作成に行政記録情報を活用している事例の研究のとりまとめ（報告書作成）を行う。
［令和 4、5 年度］ビッグデータ及び行政記録情報を統計編成業務に活用するための課題と解決方法について検討する。

2 ビッグデータの特徴と統計に活用するための視点

本章は、次章以降のビッグデータ等を活用した統計の事例について、分析するための視点の例を提供することを目的とする。

2.1 ビッグデータとは（ビッグデータの５Ⅴ）

田村(2014)[1]によると、ビッグデータは以下の 5 つの V という観点から語りうる。

1. 容量（Volume）
2. 種類（Variety）
3. 頻度・スピード（Velocity）
4. 正確さ（Veracity）
5. 価値（Value）

以下では、この 5 つの観点よりビッグデータについて語ることで、本稿で扱うビッグデータの性質を明確にする。

2.1.1 容量（Volume）

容量の観点からビッグデータを特徴づけると、以下のようになる。
「ビッグデータは、典型的なデータベースソフトウェアが把握し、蓄積し、運用し、分析できる能力を超えたサイズのデータを指す。この定義は、意図的に主観的な定義であり、ビッグデータとされるためにどの程度大きいデータベースである必要があるかについて流動的な定義に立脚している。（中略）ビッグデータは、多くの部門において、数十テラバイトから数ペタバイト（a few dozen terabytes to multiple petabytes）の範囲に及ぶだろう。」（平成 24 年版情報通信白書(2012)[2]）

---

[1] 田村義保,「ビッグデータとは何だろうか」, ESTRELA, No. 245, 2014 年 7 月.
[2] 平成 24 年版情報通信白書「第 1 部　特集　ICT が導く震災復興・日本再生の道筋」「第 2 章「スマート革命」が促す ICT 産業・社会の変革」「第 1 節「スマート革命」―ICT のパラダイム転換―」「4 知識情報基盤として新たな付加価値を創造する ICT とビッグデータの活用」「(1) ビッグデータとは何か」
https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h24/html/nc121410.html

2.1.2 種類（Variety）

種類の観点からビッグデータを特徴づけると、以下のようになる。

「ビッグデータとは、一般的には「3 つの V」で、その特徴を説明されることが多く、具体的には、「Volume（多量性）」、「Variety（多様性）」、「Velocity（流動性）」の特徴を持ったデータのことを指します。多量性はデータの総量、多様性はデータの種類、流動性はデータが生成されるスピードを示しています。近年、スマートフォンやタブレット、SNS（※ 1）などのソーシャル・メディア、M2M 通信（※ 2）の普及に伴い、世界中で生成・蓄積されているデータ量は急増しています。また、扱われるデータの種類は、従来からの販売や在庫などに関する数値や文字列のデータ（構造化データ）のみならず、ツイッターなどの代表されるテキストデータ、センサーやカメラから得られる位置情報やセンサーデータ、音声、動画、クリックストリーム（※ 3）などのデータ（非構造化データ）が増加しており、ビッグデータの 8 割を非構造化データ（※ 4）が占めていると言われています。」「（※ 1）Social Networking Service（ソーシャル・ネットワーキング・サービス）の略。（※ 2）Machine to Machine 通信の略。人間を介在せず、ネットワークを通じて機器同士が情報交換を行う通信方法。（※ 3）ウェブサイト内で訪問者がどのページを歩き回ったのかを示す履歴。（※ 4）定型フォーマットで記録される構造化データとは異なり、非構造化データは、音声や動画、テキストデータなど構成する要素が多様で、容易に分類化・体系化しづらいデータを指します。」（なるほど統計学園高等部 [3]）

なお、引用部の「構造化データ」についての説明は、広義である可能性がある。狭義の構造化データとは、リレーショナルデータベースに代表されるようなメタデータを含んでいるものを指す。構造化データの解析は標準的な統計的手法で十分である場合が多い一方、非構造化データの解析では、データの種別に適した手法を考えて行く必要がある。そのため、現時点では公的統計作成に用いることができるビッグデータは POS データのような構造化データに限られてくる。

2.1.3 頻度・スピード（Velocity）

頻度・スピードの観点からのビッグデータについて、丸山(2013）では以下のように述べられている。

「CPS が普及すると，センサーで発生する大量のデータを扱う必要がある。その際に必要とされる，新たなアーキテクチャはどんなものだろうか。（株）Preferred Infrastructure の岡野原大輔氏は，データの大部分がネットワークのエッジ部分で格納・処理される「エッジ・

[3] なるほど統計学園高等部 https://www.stat.go.jp/koukou/trivia/bigdata.html
（国立国会図書館インターネット資料収集保存事業で、上記 URL により参照［保存日：2018/07/01 - 2021/01/04］）

ヘビー・データ」の時代が来ると予測している 2)。なぜ，ビッグデータの時代には，ネットワークのエッジに多くのデータがたまるようになるのだろうか？

われわれは少なくともコスト，プライバシー，そしてレイテンシー（遅延時間）の 3 つの主要な要件があると考える。それらの要件を必要とする典型的な利用シナリオとして，(1) 監視カメラ，(2) 生体センサー，(3) パーソナルモビリティについて考える。」（丸山(2013)[4]）

ここで、CPS は Cyber-Physical System の省略形で、サイバーフィジカルシステムとは実空間（Physical）で集めたデータをサイバー空間である計算機で解析して、付加価値をつけて、実空間に戻すようなシステム（概念）のことを指す。

このように、ビッグデータとは通信ネットワークのエッジ部分において高頻度・高スピードで保持されるデータであるということができる。

2.1.4 正確さ（Veracity）

データが大量にあるといえども、あまりにノイズや誤りの多い、正確さを欠いたデータであれば分析に使うことができない。そのため、ビッグデータとは、データがある程度の正確さを持っているものを指す。

ある程度の正確さを担保する技術は、例えばデータクリーニング、データクレンジング[5]、そしてデジタルキュレーション[6]が挙げられる。

公的統計作成のために、ビッグデータを用いようとした場合は、企業がビッグデータ分析

---

[4] 丸山宏，「エッジ・ヘビー・データとそのアーキテクチャ―ビッグデータ時代の IT アーキテクチャ―」，情報処理，56 巻，5 号，269-275，2013.

[5] データベースの中から誤りや重複を洗い出し、異質なデータを取り除いて整理すること。データベースの精度を高めることにより、経営やマーケティングに有用な相関関係やパターンを探り出すデータマイニングなどに役立てることができる。（Goo 辞書：https://dictionary.goo.ne.jp/word/%E3%83%87%E3%83%BC%E3%82%BF%E3%82%AF%E3%83%AC%E3%83%B3%E3%82%B8%E3%83%B3%E3%82%B0/#:~:text=%E3%83%87%E3%83%BC%E3%82%BF%E2%80%90%E3%82%AF%E3%83%AC%E3%83%B3%E3%82%B8%E3%83%B3%E3%82%B0%E3%80%90data%20cleansing%E3%80%91&text=%EF%BC%BB%E5%90%8D%EF%BC%BD(%E3%82%B9%E3%83%AB)%E3%83%87%E3%83%BC%E3%82%BF%E3%83%99%E3%83%BC%E3%82%B9,%E3%83%87%E3%83%BC%E3%82%BF%E3%82%AF%E3%83%AA%E3%83%BC%E3%83%8B%E3%83%B3%E3%82%B0%E3%80%82）

[6] デジタル資産を選択、保存、維持管理、組織化そしてアーカイブする一連の行為である。インターネット上にあふれる膨大な情報を整理し、新たな意味づけを行い、多くの人と共有すること。（Wikipedia：https://ja.wikipedia.org/wiki/%E3%83%87%E3%82%B8%E3%82%BF%E3%83%AB%E3%83%BB%E3%82%AD%E3%83%A5%E3%83%AC%E3%83%BC%E3%82%B7%E3%83%A7%E3%83%B3）

をする時よりも、さらに綿密なキュレーションが必要になる。物価指数研究会（第 13 回）[7]では、ウェブスクレイピング技術を用いた宿泊料の価格取集及び指数作成方法の検討について紹介している[8]。

どのようなWebサイトから集めるかがキュレーションであり、集めたデータの重複を取り除き、誤りを修正するのがクレンジングである。

2.1.5 価値（Value）

データの価値は、上述の 4 V を活用して生み出されるものである。ビッグデータとは、なにより価値を生み出せなければならない。

2.2 ビッグデータを統計に活用するための視点

総務省では、「公的統計の整備に関する基本的な計画」（平成 30 年 3 月 6 日閣議決定）を踏まえ、各府省、地方公共団体、民間企業等におけるデータ等の相互利活用を推進することを目的とした「ビッグデータ等の利活用推進に関する産官学協議のための連携会議」[9]（以下、ビッグデータ連携会議）を開催している。

ビッグデータ連携会議の具体的な検討事項は大きく分けて次のとおりである。

1. 官民における、統計的分析や統計作成におけるビッグデータ等の利活用の先行事例及び先行研究の分析について
2. 統計的分析や統計作成における優先度の高いビッグデータ等の選定と応用可能性について
3. 関係者との情報共有及び優良事例の横展開の可能性について
4. データ保護及びデータ取得方法について

このビッグデータ連携会議の中で、第 1 回では、利活用上の論点・課題として、日本経済団体連合会が、2016 年に公表した「公的統計の改善に向けた提言」から、ビッグデータ活用のメリットとビッグデータ活用に向けた課題を 3 点ずつ挙げている。

（ビッグデータ活用のメリット）

1. 速報性（リアルタイムかそれに近いような速度でデータの入手が可能）
2. データ量の膨大さ（調査対象に限って見れば、全数もしくはそれに近い規模での把握が可能であり、詳細な内訳や属性別データの入手も可能）

---

[7] https://www.stat.go.jp/info/kenkyu/cpi/giji013.html

[8] 消費者物価指数 2020 年基準改定計画で、航空運賃、外国パック旅行費及び宿泊料についてウェブスクレイピングの技術を活用するとしている。
https://www.stat.go.jp/data/cpi/pdf/2104022.pdf

[9] 総務省ビッグデータ等の利活用推進に関する産官学協議のための連携会議
https://www.soumu.go.jp/main_sosiki/kenkyu/big_data/index.html

3. 低コスト（一旦データ収集の仕組みが整備されれば、データの収集に追加的なコストをほとんど要しない）

（ビッグデータ活用に向けた課題）

1. 母集団代表性の担保（特定の業態など、調査対象に偏りがある場合も多く、公的統計としての母集団代表性が担保されていない）
2. データ形式の統一（現状では、ビッグデータの作成主体によって商品コードが異なるなど、異なるビッグデータの間でデータを集計することが難しい）
3. データ提供のインセンティブ付け（民間企業にとっては、営利目的で作成したビッグデータを公的統計のデータソースとして提供するインセンティブがない）

また、第 2 回の資料と示された別所(2018)[10]では、ビッグデータ活用の本格化に向けて望ましいと思われる実務上の検討プロセスの雛形を提示している。特に、Struij (2016)[11]を引用し、以下のビッグデータ利活用方法及び潜在的効果を示している。

1. 新たな統計の作成への活用
2. 公的統計よりも詳細な情報提供への活用
3. 公的統計の早期化への活用
4. 公的統計のナウ・キャスティングへの活用
5. 公的統計の精度向上への活用
6. 公的統計の報告者軽減への活用
7. 公的統計の作成コストの削減と効率性向上への活用

他方、民間ビッグデータを活用した統計作成の課題として、「統計作成当局は、統計作成にあたって品質コントロールを左右できる立場であることが前提となる。しかし、民間が保有するビッグデータをデータ源として活用する場合には、データの調査主体（data producers）ではなく、むしろ消費者（data consumers）として向き合わざるを得ないところに難しさがある。ビッグデータを公的統計のデータ源として活用するためには、（中略）統計作成プロセスのあり方を根本から考える必要がある。」として、検討プロセスのひな型を提示している。提示されたひな形の検討プロセスを列挙すると以下のとおりである。

1. 目的をしっかり立てる
2. 利用するデータを特定する
3. 利活用は試行的に始める（データノイズの存在、データバイアス（母集団代表性の欠如）の存在、データ形式・メタデータの課題）
4. データホルダーとの協力関係の構築（協力の義務づけの妥当性、対価を払ったデータ

---

[10] 別所英実, (2018), 民間ビッグデータを統計として活用するためには、何が必要か：諸外国の取組事例の紹介と日本における課題の整理. 総務省統計委員会担当室ワーキングペーパー. 2018-WP01.

[11] Struijs, P (2016), “BIG DATA for official statistics,” Basque Statistics Office.

提供、データ取得の継続性を巡る課題等）

5. 対外コミュニケーション（対データ保有主体、対データ保有主体以外、データコミュニティの深化）
6. 個人情報や企業情報の保護
7. 技術・人材の確保
8. コーディネーション（民間統計の精度確認、政府認定）

次いで、経済産業省の「平成２６年度ビッグデータを活用した新たな経済指標・分析手法の動向に関する調査研究」[12]についてみる。まず、以下にその目的を示す。

「最近の ICT 技術の進展により活用が可能となって来た民間等保有のビッグデータを基にした、新たな速報性の高い「ナウキャスティング（Now-casting、足元予測）」・「フォーキャスティング（Forecasting、将来予測）」や膨大な相関分析といった分析手法（アナリティクス）による、マクロ・ミクロの新経済指標を研究開発する期待・実現性が急速に高まっている。ビッグデータには、速報性だけでなく、データによっては代表性があり費用効率的なものも少なくないため、公的統計の改善や新たな経済指標の作成に資する可能性があり、ひいては我が国の企業構造改革に資することも期待される。

一方で、公的統計の作成目的で政府が収集した調査票情報についても、民間等保有のビッグデータと同様に大容量のデータであるため、その更なる有効活用を図ることができる可能性がある。

そこで本調査研究では、民間企業等のビッグデータ保有状況や、国内外における研究・利活用事例等を整理することによって、公的統計や新たな経済指標づくりへのビッグデータの活用可能性を検討する。」

ビッグデータの利点と公的統計の課題の関係性については、以下の 4 点について紹介している。

1. 速報性―公表の遅さ
2. 網羅性―対象範囲の限定等
3. 補完―解釈・ニーズへの対応
4. 費用効率性―調査実施のコスト

ここで、1、2、4 については、ビッグデータは公的な統計調査に対する優位性があると述べ、3 については、ビッグデータの持つ偏りや公的統計との分類の整合性について論じ、そのまま用いることはできなくても公的統計の補完情報として用いることができる可能性があるとしている。

また、ビッグデータの公的統計の利用の視点についても次のようにまとめている。

---

[12] 三菱 UFJ リサーチ＆コンサルティング株式会社（経済産業省大臣官房調査統計グループ総合調整室委託）, (2015), 平成２６年度ビッグデータを活用した新たな経済指標・分析手法の動向に関する調査研究報告書.

図表１　ビッグデータ（BD）の公的統計への利用可能性の整理

BD 活用の視点		該当する公的統計と具体的な課題			該当するビッグデータの候補
BD の利点	統計の課題	統計	業種等	具体的な課題	
速報性	公表の遅さ	第３次産業活動指数	移動電気通信業	ウエイトが特に大きく、データが四半期ごとかつ公表が遅い	電気通信事業協会月次契約数（非ビッグデータ）
			一般貨物自動車運送業	ウエイトが特に大きいが、データの公表が遅い	VICS、ETC データ、プローブデータ
			生命保険業	ウエイトが比較的大きく、データの公表が遅い	―
			病院・一般診療所		レセプトデータ
			自動車整備業		国土交通省「自動車登録検査業務電子情報処理システム（MOTAS）」
			無店舗販売小売業	ウエイトが大きく、データの公表が遅い	―
			金融商品取引業		―
			損害保険業		―
			旅館		―
			職業紹介・労働者派遣業		―
			廃棄物処理業	ウエイトが大きく、代替的なデータに依拠せざるを得ない	―
網羅性	対象範囲の限定	特定サービス産業動態統計調査	インターネット付随サービス業	回収率が低い	―
			外国語会話教室		―
補完	解釈・ニーズへの対応	商業動態統計	小売販売額	速報性や予測系列に対する利用者ニーズが高い	POS データ、JAN コード、ポイントカード
			卸売販売額		VAN 運営会社の EDI データ
			家庭用電気器具販売額	データの時系列変動に対する解釈が難しい	POS データ

費用効率性・利用可能性		特定サービス産業動態統計調査	クレジットカード業	利用相手先別の「その他」の把握	クレジットカードデータ

3 ビッグデータ及び行政記録情報に関する情報（国内）

3.1 ビッグデータ等の活用事例

国内におけるビッグデータ活用例について、官公庁の報告書・資料等、および刊行物や学会・研究会等から得た情報を報告する。

なお、官公庁の公的資料としては、平成 26 年度から令和元年度までの情報通信白書[13]等の記載、またそれ以外でビッグデータ活用例の記載が見つかった官公庁の報告書・資料等を引用する。また、刊行物や学会・研究会等の学術的資料としては、府省設置のワーキンググループ（WG）の成果が論文等として発表された刊行物や学会や研究会等の予稿集や発表スライド等を引用する。

3.1.1 総務省「情報通信白書」

3.1.1.1 平成 26 年版情報通信白書

平成 26 年版の白書に挙げられた活用例にあるデータには、公的統計作成に活用できるような事例は見当たらなかった。第 1 部第 3 章第 1 節「様々な価値を生み出すビッグデータ」にある事例を以下に示す。

1. マツダ（株）：部品（約 1 万）の記録を活用した品質向上
2. 本川牧場：牛の個体情報やセンサー計測情報（200 から 300）をクラウドで分析
3. （株）グリーン＆ライフイノベーション：データに基づく漁場予想
4. （株）あきんどスシロー：皿の RFID の情報を分析
5. イーグルパス（株）：車載 GPS データ、センサーデータを活用したバスダイヤ最適化
6. （株）マイクロアド：Web 閲覧の ID、Cookie を利用した Demand Side Platform 事業

[13] 総務省 情報通信白書（平成 26 年版から令和元年版）
https://www.soumu.go.jp/johotsusintokei/whitepaper/h26.html, ～ h30.html,
https://www.soumu.go.jp/johotsusintokei/whitepaper/r01.html

3.1.1.2 平成 27 年版情報通信白書

平成 27 年版の白書の第 2 部第 5 章第 4 節「ICT 化の進展がもたらす経済構造の変化」には、リモートセンシングデータ等を活用した農業情報サービスの事例が挙げられており、田村（2014）は、農林水産省の「作物調査への活用が期待される」としている。

3.1.1.3 平成 28 年版情報通信白書

平成 28 年版の白書の第 1 部第 3 章第 3 節「公共分野における先端的 ICT 利活用事例」を以下に示す。

1. 医療ヘルスケア分野
   - ・MySOS、Join、Team
   - ・パーソナルデータの分散管理
   - ・Enlitic（人工知能（AI）による悪性腫瘍の検出）
   - ・機能繊維素材を活用した実証実験
2. 教育分野
   - ・スタディサプリ
   - ・プログラミング教育
   - ・Qubena（人工知能（AI）を用いた算数・数学のタブレット教材）
3. 交通分野
   - ・公共交通オープンデータ協議会
4. 防犯分野
   - ・警備における ICT 活用とボランティア連携
5. 防災・減災部門
   - ・アンケート結果（災害が身の回りで起こる場合、災害の情報を収集するのに最も利用するメディア）
   - ・平成 28 年（2016 年）熊本地震

3.1.1.4 平成 29 年版情報通信白書

平成 29 年版の白書の第 1 部第 2 章は「ビッグデータ利活用元年の到来」と題されている。その中第 1 節「広がるデータ流通・利活用」には、総務省「安心・安全なデータ流通・利活用に関する調査研究」（平成 29 年）を出典とした「NTT ドコモ「モバイル空間統計」」等の事例が挙げられている。

3.1.1.5 平成 30 年版情報通信白書

平成 30 年版の白書には、ビッグデータを表題とした章等は見当たらず、公的統計作成に直接つながると考えられる事例やビッグデータを中心とした記述はない。ただし、ビッグデータについては、第 1 部の「はじめに」には、「データ主導社会へ」「データの価値」という小節に「大量のデジタルデータ（Big Data：ビッグデータ）の生成、収集、蓄積が進みつつある。それらデータの AI（Artificial Intelligence：人工知能）による分析結果を、業務処理の効率化や予測精度の向上、最適なアドバイスの提供、効率的な機械の制御などに活用することで、現実世界において新たな価値創造につなげる」という記述はある。「はじめに」の小節は、次いで「デジタルトランスフォーメーション」、「Society5.0」、「人口減少時代の ICT による持続的成長」となっており、第 1 章以下のこれらの観点による記述中にビッグデータが散見される。

3.1.1.6 令和元年版情報通信白書

令和元年版の白書にも、ビッグデータを表題とした章等は見当たらず、公的統計作成に直接つながると考えられる事例の記述はない。なお、統計に関する記述として、第 1 部第 2 章第 1 節には、「データに価値をもたらす「4V」」のうち「Veracity」（正確性）について、「例えば統計では調査対象全体（母集団）から一部を選んで標本とすることが行われるが、ビッグデータでは、この標本を母集団により近づけることにより、母集団すなわち調査対象全体の性質をより正確に推計できるようになる。」との記述がある。

3.1.2 総務省「ビッグデータ等の利活用推進に関する産官学協議のための連携会議」

第 8 回連携会議で示された「これまでにビッグデータ連携会議で取り上げた事例」[14]は以下の表のとおりである。

[14] ビッグデータ等の利活用推進に関する産官学協議のための連携会議(第 8 回)（令和元年 11 月 13 日)「参考 1 これまでにビッグデータ連携会議で取り上げた事例」より https://www.soumu.go.jp/main_content/000654853.pdf

図表２　これまでにビッグデータ連携会議で取り上げた事例

（府省による利活用等の検討事例）

分野	統計（調査）名 ビッグデータ名	取組	メリット ◎：BD 利活用の目的（主なメリット） ○：副次効果（副次的なメリット）					主な課題	備考 （所管府省）
			Q	C	D	B	内容		
消費	消費動向指数（CTI） POS データ･クレジットカード等	データホルダーを含む産官学が連携した「消費動向指数研究協議会」において、CTI 作成への、民間企業データ活用を検討	◎		◎		・民間企業が保有する様々なデータを利用することによる精度向上及び早期に利用可能なデータによる公表早期化の可能性	・推計方法の検討（民間企業データのバイアス補正等） ・データの安定的な入手	第 5 回 BD 連携会議 総務省統計局消費統計課
交通	パーソントリップ（PT）調査 携帯基地局情報等	従来の統計調査データに携帯基地局情報等を組み合わせることにより、従来の PT 調査より詳細な OD データを作成（現在検証中）	◎	○		○	・総量を PT 調査で、BD で比率を算出することにより、目的別手段別小ゾーン間 OD の把握が可能	･PT 調査における携帯電話基 地局データの精度の検証 ・統計調査（アンケート調査）で把握する移動（トリップ）と、基地局情報等の移動との定義のずれの解消	第 5 回 BD 連携会議 国土交通省都市局都市計画調査室
商業	ビッグデータを活用した商業動態統計調査（試験調査：家電大型専門店分野） POS 等ビッグデータ（POS データの他、店舗マスター等）	商業動態統計調査（以下、「本体調査」）において、ＰＯＳ等ビッグデータを活用するといった新たな調査方法の採用とその調査事務について実地の検討を行い、「報告者負担の軽減化」、「統計業務の効率化」、「公表の早期化」の他、「景気動向把握の向上に資するための把握内容の詳細化」等の実現可能性などの精査に必要な基礎資料を得ることを目的として実施。	○	◎	◎	◎	・報告者負担の軽減、統計調査業務の効率化、公表の早期化 ・集計区分の詳細化（都道府県別、経済産業局別が可能）	･BD を活用した新たな調査スキームの検討（POS データプラットフォーマーの活用） ・調査スキーム維持に要するコスト ・本体調査との整合性確保 ・二次的利用の対象範囲	第 6 回 BD 連携会議 経済産業省大臣官房調査統計グループ調査分析支援室
物価	消費者物価指数（CPI） WEB 掲載データ	CPI を構成する費目の中で、外国パック旅行費・航空運賃・宿泊料に関して大手旅行業者・航空会社の WEB サイトをスクレイピングにより格情報を収集し、それらの情報から 2020 年基準 CPI を作成（予定）	◎	○		○	・ネット販売価格の取り込み ･WEB からの膨大なデータの使用による統計精度の向上 ・データの自動収集による報告者負担の軽減と業務効率化	･WEB サイト保有企業との継続的な信頼関係の構築（サイトアクセスの承諾、販売実態を踏まえた価格代表性の確保） ・適切なデータノイズ除去、システム障害発生時等のリスク管理及び事前の備え ・統計作成までの業務体制（リソース配分）	第 8 回 BD 連携会議 総務省統計局物価統計室

Q：統計の精度向上、詳細化　C：統計作成の効率化、コスト削減　D：統計公表の早期化　B：報告者負担の軽減

（民間ビッグデータの紹介事例）

ビッグデータ	内容	関連分野	備考 （講演した企業）
転職情報	人材紹介会社において、転職が決まったときの転職先企業での職種・業種・給与と、元の会社での職種・業種・給与から転職時の賃金変動状況が把握できる	労働（賃金変動）に関する分野	第 2 回 BD 連携会議 株式会社リクルートキャリア
流動人口	携帯端末アプリから取得したGPS位置データからメッシュ単位の流動人口（ある時点に、ある場所に存在した人口）が推計できる	人口の流動情報を活用する分野	第 3 回 BD 連携会議 株式会社 Agoop
地図情報	GIS を使用して、地物、人に関する情報、経済に関する情報等の様々なデータを地図上で可視化することにより、政府の施策立案や民間マーケティングを効果的に行える	人や物の地理的な配置情報を活用する分野	第 4 回 BD 連携会議 株式会社ゼンリンジオインテリジェンス
電力データ	スマートメーターからの設備情報と電力量情報から、世帯の居住等に関する状況や、住民の活動状況を位置情報と合わせて把握できる	住民の居住や活動情報を活用する分野	第 7 回 BD 連携会議 グリッドデータバンク・ラボ有限責任事業組合

3.1.3 菅氏、飯島氏、兵頭氏　他「メッシュ型流動人口検証」[15]

ビッグデータ連携会議の下に設置されたワーキンググループ（WG）における研究である。取り組みの概要は、以下の通りである。

「東京都において GPS 方式で収集したメッシュ型の流動人口データ（GPS データ）を国勢調査や基地局方式で収集したデータ（基地局データ）と比較し、統計的に分析した結果をまとめたものである。この結果から GPS データは、国勢調査や基地局データと相関があり、

[15] 菅 愛子, 飯島 信也, 兵頭 大史, 他. (2019). 東京都における流動人口データの有効性の検証. 総務省統計委員会担当室ワーキングペーパー 2019-WP03.
https://www.soumu.go.jp/main_content/000630006.pdf
なお、先行研究として、独立行政法人統計センター, 株式会社 NTT ドコモ. (2013). 「平成 24 年度 共同研究報告書 官庁統計とモバイル空間統計に基づく新たな統計の創出に関する共同研究」https://www.nstac.go.jp/services/pdf/sankousiryou2503-4.pdf がある。

一定レベルの信頼性が確保されているとともに、利活用にあたり国勢調査を補完しうるものとして有効であることを示すことができた。更に以下の GPS データの特徴に留意して使用することで、より効果的な活用が可能であると考えられる。」

なお、GPS データはソフトバンク株式会社の子会社の株式会社 Agoop から提供を受けている。メッシュ型 GPS データに関しては、以下のような利点が挙げられていた。

- 国勢調査 5 年間における差率（2010 年と 2015 年調査の差）の標準偏差を基準とした評価から、GPS データは流動人口 2,000 人以上のメッシュ（500m 四方）において一定レベルの信頼性が確保されており、そのエリアは東京都 23 区を中心に居住地域の面積の約半分に該当する。
- 時間帯別の GPS データは、国勢調査では把握対象外である通勤・通学以外の勤務中の移動や余暇・消費活動による人の動きを、ビジネス街や住宅地などの地域の特性と整合する形で捉えている。
- GPS データは解像度が高いため、大量の人の動線のハブとなる都心ターミナル駅や海岸沿い等、隣接するメッシュ間の人口差が大きいエリアにおいて、より強みを発揮できる。
- GPS データはアプリユーザーの属性に依存することから、標本の偏りが発生しやすいことを理解した上で、利活用を進めていくことが肝要である。

WG では、2015 年から 2017 年の位置情報、NTT ドコモ基地局情報により把握した東京都のメッシュデータ 500m メッシュ及び 1km メッシュで 2015 年の国勢調査のメッシュデータとの相関や差異を分析している。位置情報や NTT ドコモ基地局データが国勢調査の補完に役立つ可能性があることを示している。

3.1.4 経済産業省「【試験調査】ビッグデータを活用した商業動態統計調査」[16]

2019 年に経済産業省の経済産業省大臣官房調査統計グループが行った「ビッグデータを活用した商業動態統計調査（試験調査：家電大型専門店分野）」に関わる取り組みについて述べる。

この調査は、従来の商業動態統計調査の一部を「本体調査」とし、POS 等ビッグデータを活用した新たな方法による調査を「試験調査」とし、その比較をしたものである。

この調査の目的は、以下に引用する通りである。

「本試験調査は、本体調査において、ＰＯＳ等ビッグデータを活用するといった新たな調査方法の採用とその調査事務について実地の検討を行い、「報告者負担の軽減化」、「統計業務

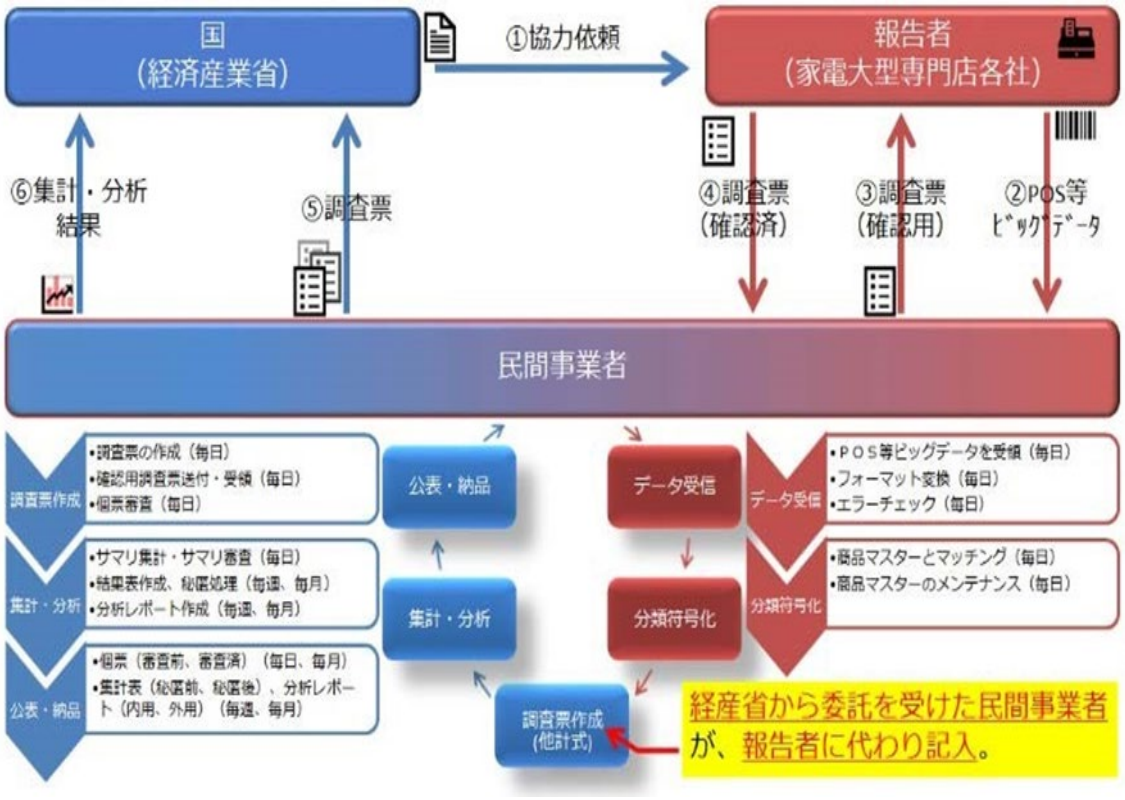
[16] 経済産業省，【試験調査】ビッグデータを活用した商業動態統計調査.
https://www.meti.go.jp/statistics/tyo/bigdata_syoudou/index.html
https://www.meti.go.jp/statistics/tyo/bigdata_syoudou/pdf/20190524_hikakukekka_honbun.pdf

の効率化」、「公表の早期化」の他、「景気動向把握の向上に資するための把握内容の詳細化」等の実現可能性などの精査に必要な基礎資料を得ることを目的としています。」

ビッグデータとして民間企業が集めたデータを利用しているのではなく、自らがビッグデータのソースからデータを収集し、ビッグデータを作成する形になっている。従来の調査票調査のオンライン化も兼ねている。

調査方法及び、その実施スキームは下記[17]の通りである。

図表３　ビッグデータを活用した商業動態統計調査の実施スキーム



[17] 経済産業省大臣官房調査統計グループサービス動態統計室/調査分析支援室. (2019). ビッグデータを活用した商業動態統計調査（試験調査：家電大型専門店分野）」の検証結果について. https://www.meti.go.jp/statistics/tyo/bigdata_syoudou/pdf/20190524_hikakukekka_honbun.pdf

図表４　本体調査と試験調査との実施計画上の違い

項目	本体調査	試験調査
種別	基幹統計調査	一般統計調査
調査の名称	商業動態統計調査	ビッグデータを活用した 商業動態統計調査
調査票	丁２調査：家電大型専門店	試験調査：家電大型専門店分野
調査対象の範囲	・地理的範囲全国 ・属性的範囲日本標準産業分類に掲げる「細分類５９３１－電気機械器具小売業（中古品を除く）」又は「細分類５９３２－電気事務機械器具小売業（中古品を除く）」に属する事業所（以下「家電専門店」という。）を有する企業であって、経済産業大臣が指定する条件を満たすもの。 ・経済産業大臣が指定する条件 売場面積が 500 ㎡以上の家電専門店を 10 店舗以上有する企業。	・地理的範囲全国 ・属性的範囲日本標準産業分類（平成２５年１０月改定）に掲げる「細分類５９３１－電気機械器具小売業（中古品を除く）」又は「細分類５９３２－電気事務機械器具小売業（中古品を除く）」に属する売場面積 500 ㎡以上の事業所（家電大型専門店）を 10 店舗以上有する企業
報告を求める事項	①企業名 ②商品販売額 ③店舗数 ④商品手持額	①企業名、法人企業番号 ②店舗番号 、都道府県番号 ③商品販売額 ④期末商品手持額（３月、６月、９月、１２月の各月末）
基準となる期日	商業動態統計調査は、毎月末日現在 によって行う。ただし、商品販売額は、月初めから月末までの１か月間、丁２調査の調査事項のうち商品手持額については、毎四半期末日現在によって行う。	平成 27 年１月１日～平成 30 年 12 月 31 日の間の 毎日 （原則として、毎日０時から２４時までの２４時間。）の実績なお、期末商品手持額については、毎四半期末日現在。
調査組織	経済産業省－民間事業者－報告者	経済産業省－民間事業者－報告者

以下に調査結果をまとめる。

試験調査の回収率については約 73.9%であり、対象企業数 23 企業中、回答数 17 企業であった（本体調査の回収率を 100%としている）。金額ベースでの回収率は 3 兆 1,877 億円／ 4 兆 3,192 億円の 72.6%であった。

暦年ごとの商品販売総額（全国／全分類）における差額は、平成２８年の１９６億円（本体調査を１００とした場合の試験調査との差額が占める割合：０.６％）が最大であり、最小は、平成２７年の▲８億円（同▲０.０％）であった。

また、本体調査と試験調査の商品販売総額の相関を時系列に見ると、相関係数は０．９９と非常に高い結果となっている。

3.1.5 経済産業省「平成２８年度　ＩｏＴ推進のための新産業モデル創出基盤整備事業（ビッグデータを活用した新指標開発事業）」[18]

この事業は、

1. ビッグデータを活用した新指標開発実証事業の実施
2. 有識者委員会の運営、及び新指標への活用に向けた包括的調査事業の実施
3. 成果の取りまとめ・普及活動

から構成されており、目的は、以下のとおりである。

「民間企業が保有するＰＯＳデータ、サイバースペース上に蓄積されているブログや Twitter を始めとしたソーシャルネットワーキングサービス（ＳＮＳ）等の書き込み、政府等行政機関が保有する統計情報や行政記録情報等のビッグデータについて、解析技術やＡＩ技術等を活用して分析を行うことで、既存の政府統計の補完、拡充、詳細化を実現し、従来の統計よりも速報性に優れた指標を開発して、政府においては迅速で正確な景気判断・政策決定を、民間においては迅速で的確な経営判断・意思決定を可能とすることを目的とする。」

この事業の委託事業者 PwC あらた有限責任監査法人から再委託された事業者とその実施内容は以下の通り。

- ジーエフケーマーケティングサービスジャパン株式会社（以下、「GfK」又は「GfK ジャパン」）：ＰＯＳ等のビッグデータの収集・加工・維持・提供、およびＰＯＳ等のビッグデータを活用した新指標開発
- 株式会社エヌ・ティ・ティ・データ（以下、「NTT データ」）：Twitter ビッグデータの収集・加工・維持・提供
- 株式会社ホットリンク（以下、「ホットリンク」）：ブログビッグデータの収集・加工・維持・提供

[18] PwC あらた有限責任監査法人（経済産業省大臣官房調査統計グループ調査分析支援室委託）.(2017). 平成２８年度ＩｏＴ推進のための新産業モデル創出基盤整備事業（ビッグデータを活用した新指標開発事業）報告書.
https://www.meti.go.jp/meti_lib/report/H28FY/000071.pdf

- 株式会社インテージリサーチ（再々委託先：株式会社インテージテクノスフィア）、株式会社インテージ（３社を合わせ以下、「インテージ」）：ＰＯＳ等のビッグデータを活用した新指標開発
- 野村證券株式会社（以下、「野村證券」）：ＳＮＳ等のビッグデータとその解析技術を用いた新指標開発・評価・検証、および新指標の算出・公開サービスの実装・提供
- 公益財団法人流通経済研究所（以下、「流通経済研究所」）（NTT データによる再々委託先）：ＰＯＳ、ＳＮＳ等のビッグデータ、及び政府統計等を組み合わせた新指標開発

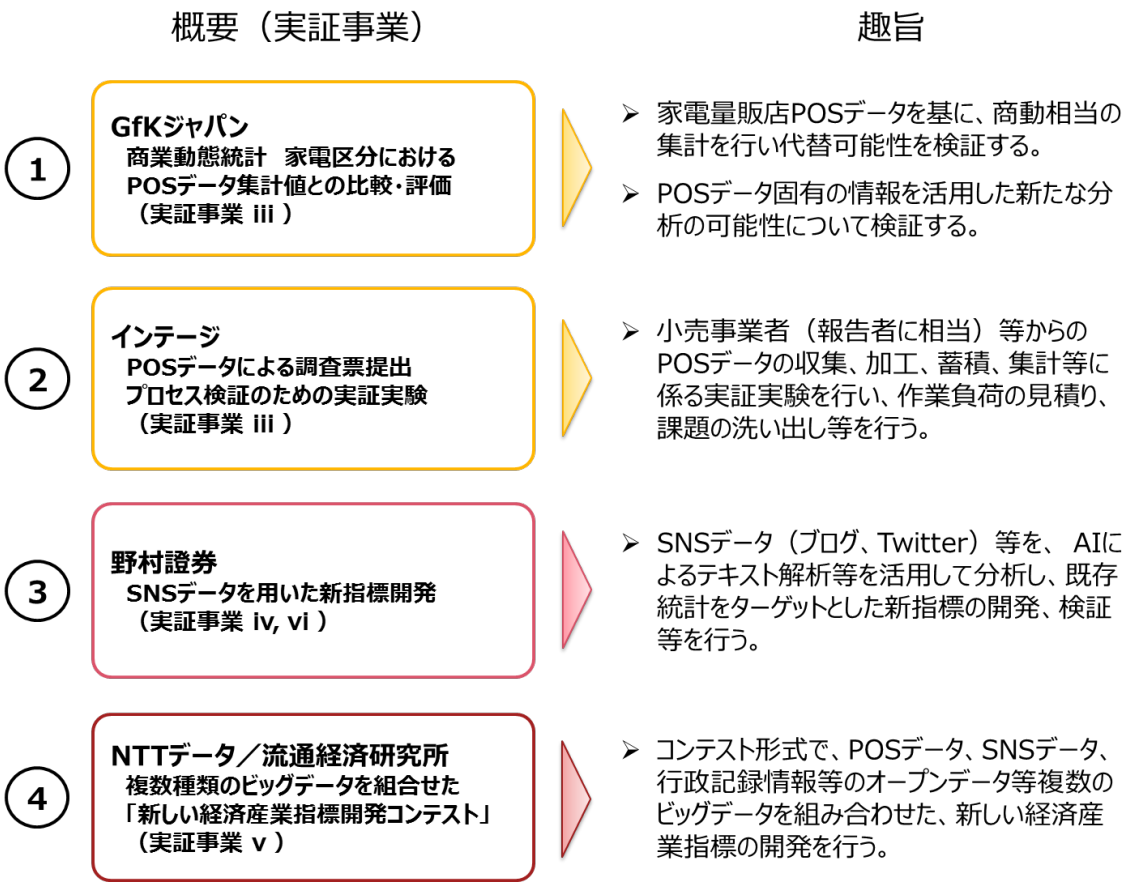
提供されたデータは以下の通り。

図表５　再委託事業者により取集・加工・提供されたデータ

GfK ジャパン	データの内容	家電大型専門店等から収集されたＰＯＳデータ
	対象期間	２０１４年１月～２０１６年１２月
	時系列区切り	月次、週次、日次
	対象地域	日本全国（都道府県別）
	商品分類単位	家電製品１０３分類
	データ項目	数量、金額、平均価格、店舗数
NTT データ	データの内容	Twitter データ
	対象期間	２００８年１月～２０１７年２月
	時系列区切り	日次
	対象範囲	プライベートアカウントのツイート･ダイレクトメッセージを除く全ての日本語ツイート
	加工条件	特定の期間内、特定のキーワードを含む（含まない）等、再委託事業者等の要望に応じた加工を実施
ホットリンク	データの内容	ブログデータ
	対象期間	２００７年１月～２０１７年２月
	時系列区切り	日次
	対象範囲	アメーバブログ、livedoor blog、Yahoo!ブログ、seesaa ブログ、楽天ブログ、はてな Diary、goo ブログ、エキサイトブログ、ココログ、ヤプログ、FC2 ブログ、ウェブリブログ、DTI ブログ、
		LOVELOG、ジュゲムを含む主要なブログポータル
	加工条件	特定の期間内、特定のキーワードを含む（含まない）等、再委託事業者等の要望に応じた加工を実施（加工条件に合致するブログの記事数を提供することも可能）

各企業による事業結果の概要は以下の通り。

図表 6　各再委託事業者による実施内容（実証事業）の概要とその趣旨



このうち、GfK ジャパンによる、経済産業省の商業動態統計丁 2 調査票「家電大型専門店販売」を代替する、POS データによって調査された新指標は特に精度の良い指標となっていた。

3.1.6 佐藤氏、牧本氏、齊藤氏「POS データを活用した商業動態統計の速報性向上に関する研究」

経済産業省平成 28 年度 IoT を活用した新ビジネス創出推進事業（ビッグデータを活用した新指標開発事業）の一環として、筑波大学の佐藤忠彦教授、牧本直樹氏、齊藤敬氏、武井明則氏により実施され、2018 年度統計関連学会連合大会で発表された研究である。

研究の目的としては、POS データを用いて、商業動態統計（以下、「商動」）の速報性を向上できないかを検証することである。具体的には、商動予測モデルの検証、すなわち以下の項目にまとめられる。

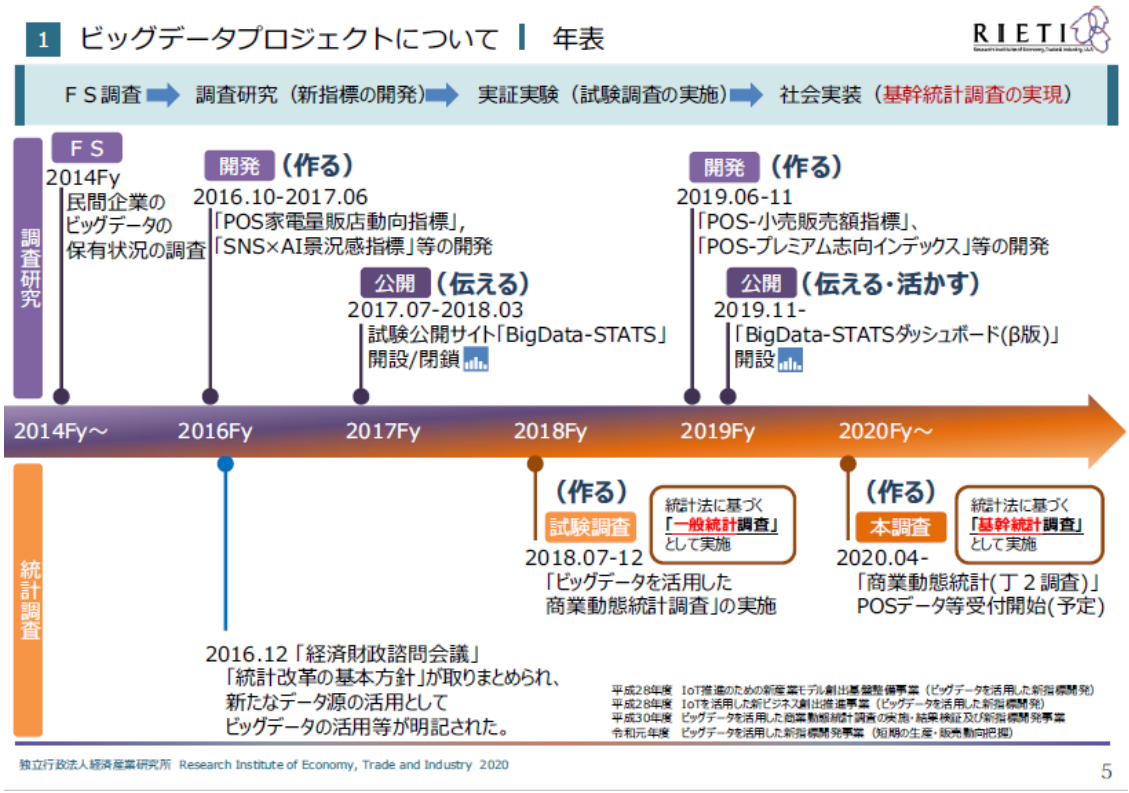
(a) 商動のみを用いてモデル化、推定する
(b) 商動データに POS データの情報を追加し、モデル化、推定する
(c) (a)と(b)の結果を比較し、POS データを用いることの意義を検討する

結論としては、商動のみを用いてもかなりの予測精度が担保できるが、多くの業態、県で POS を併用することで予測精度が向上できるというものである。

3.1.7 小西葉子氏　RIETI-BBL セミナー「ビッグデータと公的統計調査：「作る・伝える・活かす」工夫」[19]

2020 年 1 月 22 日に経済産業省主催の RIETI-BBL セミナーにて、小西葉子氏（RIETI 上席研究員、大阪大学経済学研究科　特任教授（常勤））平成 28 年度以降の経済産業省のビッグデータプロジェクトについて総括する年表を示していた。

図表 7　ビッグデータプロジェクトについて|年表



小西氏は講演にて POS データを用いた基幹統計を作成することが総務大臣に承認されていることに触れたうえで、ビッグデータを用い、AI 手法やデータサイエンス手法を活用し、報告者の負担を減らし、公表時期を早め、ユーザーが利用しやすい環境を提供する重要性について述べていた。

また、民間企業が保有するＰＯＳデータを活用し、各小売業態の商品別販売について、週

[19] RIETI-BBL セミナー　ビッグデータと公的統計調査：「作る・伝える・活かす」工夫 https://www.rieti.go.jp/jp/events/bbl/20012201.html

次、地域別など、より詳細な動向を簡便に把握することを可能とするための指標「METI POS―小売販売額指標（ミクロ）」を発表した。この指標のほかにも、ビッグデータをより活用しやすくするための様々な経済指標を公開している「BigData-STATS ダッシュボード（β 版）」[20]も紹介した。

3.1.8「統計技術・データソースの多様化等検討会」

効果的・効率的な統計作成に資する統計技術、統計作成に用いるデータソースの多様化等について調査・検討するため、統計改革推進会議―統計改革調査部会の下で「統計技術・データソースの多様化等検討会」（以下、検討会）が開催されている。

この令和２年８月に開催された第１回検討会[21]では、事業所母集団データベースにおいて、行政記録情報として、労働保険情報（月次）及び商業・法人登記簿情報（月次）について対象企業について照会を行い、結果をデータベースに収録するとの例が参考資料に掲載されている。

このほか、前述のビッグデータ連携会議で取り上げられた公的統計についての事例が整理され「ビッグデータの公的統計への利活用事例等」との資料として提出。これには、データの取集方法等についてまとめられていることから、その抜粋を以下に示す。

[20] BigData-STATS ダッシュボード（β 版）.
https://www.meti.go.jp/statistics/bigdata-statistics/bigdata_pj_2019/index.html

[21] 統計技術・データソースの多様化等検討会（第１回）配布資料.
https://www.kantei.go.jp/jp/singi/toukeikaikaku/toukeigijutsu_data_source/dai1/siryou.html

図表８　ビッグデータの公的統計への利活用事例等について（抜粋）

分野	物価	消費	商業	交通
公的統計	消費者物価指数(CPI)	消費動向指数(CTI)	商業動態統計(家電大型専門店分野)	パーソントリップ(PT)調査
ビッグデータ	WEB 掲載価格データ	POS データ・クレジットカード等	POS データ	携帯基地局情報
統計の目的	全国の世帯が購入する家計に係る財及びサービスの価格等を総合した物価の変動を時系列的に測定する。家計の消費構造を一定のものに固定し、これに要する費用が物価の変動によって、どう変化するかを指数値で示したもので、毎月作成。指数計算に採用している各品目のウエイトは家計調査の結果等に基づいている。毎月の品目価格は小売物価統計調査によって得られる。	家計調査の結果を補完し，消費全般の動向を捉える分析用のデータとして開発中の参考指標。家計消費指数を吸収するとともに，単身世帯を含む当月の世帯の平均的な消費(CTI ミクロ), 家計最終消費支出の総額の動向(CTI マクロ）を推計している。	全国の商業を営む事業所及び企業の販売活動などの動向を明らかにすることを目的として実施。全国の卸売・小売事業所を対象として、商品販売額、販売先別商品販売額、月末従業者数、期末商品手持額等を調査している。	都市における人の移動に着目した調査。世帯や個人属性に関する情報と１日の移動をセットで尋ねることで､「どのような人が、どのような目的で、どこからどこへ、どのような時間帯に、どのような交通手段で」移動しているかを把握することが可能で、都市交通の現況の把握、将来交通需要の予測、都市交通マスタープランの作成等、都市交通に関する計画等を策定する上での基礎資料とすることを目的としている。
データの収集方法	WEBにおける公表情報をスクレイピングにより取集。	消費動向指数研究協議会を通じて、POS データ保有企業、クレジット・カード会社等から提供。	POS データを収集・分析するマーケティングリサーチ（MR）会社における加工を経て、情報を収集。	携帯キャリア会社から提供。

概況（背景と現状）	【背景】統計改革の基本方針等において、インターネット販売価格の更なる捕捉に向けた検討を行うとの提言がなされたことを受け、CPIにおけるインターネット価格のより精緻な把握に向けた検討を開始。	【背景】消費全体の動向を、マクロ、ミクロの両面でとらえる、速報性を備えた包括的な消費関連指標の在り方について検討することを目的として、総務大臣主催の「速報性のある包括的な消費関連指標の在り方に関する研究会」を開催し、消費動向指数を開発。	【背景】BD を活用した新指標の開発に向けて調査研究を進める中で、公的統計への BD 活用が統計改革の基本方針、第Ⅲ期基本計画に明記されたことも踏まえ、BD を活用した商業動態統計の実施に向け検討を開始。	【背景】1970 年代以降、PT 調査の実施とこれに基づく総合都市交通計画の立案が継続的に行われて、科学的な分析結果に裏打ちされた都市交通施策の推進に大きく寄与してきた。一方で、①近年の自治体における都市交通上の課題として「短中期的・ミクロな交通政策」までニーズが拡大したこと、②全国あらゆる場所で、24 時間 365 日データが蓄積される交通系ビックデータが登場したこと、など状況が変化してきたことを受け、PT 調査へのビッグデータ活用の検討を開始。
	【現状】有識者を交えた「物価指数研究会」の検討、平成 30 年度統計法施行状況に関する審議を経て、旅行サービス 3 品目について、消費者物価指数 2020 年基準の 2020 年 1 月分からウェブスクレイピングの実運用を開始。現行の 2015 年基準は 2021 年 12 月分公表迄であり、現行と並行して進めていく予定。	【概況】消費動向指数研究協議会を設立後、消費動向指数の研究を継続的に実施し、既存統計をデータソースとした消費動向指数（CTI）を開発し、平成 30 年 1 月分から参考指標として公表開始。今後、研究分析・検証を得た後、ビッグデータを順次活用する予定。	【現状】2018 年に試験調査を実施し一定の目処が立ったため、2020 年より本格的に実施。BD を活用するかどうかは企業の意向次第（調査票記入と POS データの活用の選択制）。	【現状】PT 調査データとビックデータを組み合わせ、それぞれの強みを生かすことのできる総合交通調査体系の構築を目指す一環として、「総合都市交通体系調査におけるビッグデータ活用の手引き」を作成。H30 年の東京都市圏 PT 調査でも本手引きを活用し、調査設計・分析等を実施。今後、近畿圏 PT 調査・中京圏 PT 調査でも活用していく予定。

メリット（主目的）	WEBスクレイピング技術で大量のネット価格（旅行関係）を取得することによる、CPI の精度向上。	POSデータ等を活用した高精度で消費動向を捉えることのできる速報性の高い指標の構築。（当面は、CTI マクロにおける GDP 統計の構成要素である家計消費支出の予測精度向上の実現を目指して検討中。）	家電大型専門店に関する調査をPOSデータで代替すること（MR 会社による POSデータの調査票情報への変換）による、報告者負担の軽減。	トリップ総量を PT 調査、各ゾーンにおけるトリップ目的・トリップ手段に関する比率を携帯基地局データで算出することによる、目的別手段別小ゾーン間 OD の推計
データ選定と対象分野の絞り込み	品質のバラツキを少なくできるため、WEBスクレイピングによる価格取集に馴染む、航空運賃・宿泊費・外国パック旅行を対象とした。	CTI マクロの検討に当たって、ＰＯＳデータ、クレジットカードデータ等を候補として検討中。	商業動態統計の対象業種のうち、商品・店舗（企業のみならず事業所も含む）に関して網羅的なデータが既に把握されており、かつ当該データの活用がビジネスとして確立している、家電大型専門店分野を対象とした。	交通系ICカードから得られるビッグデータなどもあるが、交通手段やエリア的な網羅性が最も高い携帯基地局情報を選択した。調達先は安価で協力してくれることを重視して選定した。
ビジネスモデル	WEBスクレイピングは（サイト運営者のサーバーに負荷はかけるが）公表情報を収集するものであるため、データの取得そのものの購入費等は不要。	現在は、データホルダー（個別企業等からデータを収集している企業）から、研究のために不定期に無償でデータ提供を受けている。	MR 会社における家電量販店 POS データの収集・分析業務は、従来より家電メーカーなどを顧客としたビジネスモデルが確立されている。今回は MR 会社に対して組替え集計業務を委託するものである。	調査委託先での BD の調達コストが上昇する可能性が高く、今後も適切な予算を確保することが課題となる。

3.1.9 小寺氏、藤田氏　他「ＰＯＳ・テキストデータを用いた消費分析―機械学習を活用して―」[22]

POS・テキストデータを公的統計作成や Editing に用いた例ではないが、示唆に富む活用法やレビューであるとして記述する。

消費分析に対して新しい視点を提供することを目的として、ＰＯＳ・テキストデータをもとに機械学習の手法も活用しながら、３つの論点について検討を行う内容であった。

結果は、以下の通りである。

「（１）一般的な需要・供給曲線を想定し、ＰＯＳの価格・数量データを用いて、価格・数量変化が需要要因と供給要因のどちらに起因するのかの要因分解を行った。

（２）ＰＯＳデータ等の速報性の高いデータから、機械学習により小売業販売額全体の動きのナウキャストを行った。ＰＯＳデータのみでも、一定程度の精度を持つナウキャストを行うことは可能であったが、天候データも加えることで、ナウキャストの精度が向上するとの結果が得られた。

（３）ディープラーニングの手法を用いて、新聞記事の内容がどの程度ポジティブであるかを示す指数を紙面別に作成し、消費者マインド等との相関関係の強さを検証した。」

また、ＰＯＳデータを利用した関連研究として以下の研究がレビューされていた。

- 渡辺・渡辺 [23]:ＰＯＳデータを用いて、ほぼリアルタイムで物価動向を観察することができる「日経・東大日次物価指数」と呼ばれる指標を作成
- 丸山他 [24]:「カバレッジ、品質調整、計算方法等の点においてＣＰＩと異なっていることから、既存のマクロ統計と比較する際には留意が必要となる」と指摘
- Imai and Watanabe[25]：ベース:ＣＰＩと計算手法や銘柄選定基準を同様にした指数の作成も行われている
- 渡辺 [26]：ＰＯＳデータが利用可能な店舗毎に物価前年比と売上高前年比を計算するこ

---

[22] 小寺信也，藤田隼平，井上祐介，新田尭之.「POS・テキストデータを用いた消費分析－機械学習を活用して－」内閣府経済財政分析ディスカッション・ペーパーDP/18-1. 2018. https://www5.cao.go.jp/keizai3/discussion-paper/dp181.pdf

[23] 渡辺広太，渡辺努．「スキャナーデータを用いた日次物価指数の計測」．東京大学金融教育研究センターワーキングペーパーCARF-J-094，2013.

[24] 丸山歩，嶋北俊一，落合牧子，上田聖．「CPI と東大指数の乖離の分析について」．統計研究彙報 第 72 号. 55-78. 2015 年 3 月.

[25] S. Imai and T. Watanabe. “Replicating Japan’s CPI Using Scanner Data, Replicating Japan’s CPI Using Scanner Data”, JSPS Grants-in-Aid for Scientific Research, Working Paper Series No. 072. 2015.

[26] 渡辺努．「店舗別インフレ率から読み取れること」．ナウキャスト『マンスリーレポート 2016 年 2 月号』. 2016 年 2 月 17 日.

とで、価格の変化が需要曲線と供給曲線のどちらがシフトすることで起こっているのかを考察

- 上田他[27]：小売店における定価の改定頻度や特売の頻度と失業率や総労働時間等のマクロ経済指標との相関関係
- 藤田[28]：ＰＯＳデータから作成した価格指数と消費者マインド・景気循環との関係性を分析し、家計が小売価格の変化に敏感になった可能性や、定価の改定や特売と景気循環との間には有意な相関関係がみられると指摘。ただし、併せて、定価の改定や特売等の系列を束ねたヒストリカルＤＩを作成し、内閣府による景気基準日付との整合性を調べたところ、合致率は６割強にとどまったことも報告
- Watanabe et al[29]：ＰＯＳデータに対して機械学習を活用した分析。ＰＯＳデータと天候データの双方を利用して販売数量を予想するモデルを推計。通常の線形モデルに加え、ニューラルネットワークによる推計

以上に加えて、テキストデータを利用した関連研究として以下の研究がレビューされていた。

- Doms and Morin[30]：アメリカのデータを利用した実証分析により、ニュース媒体は３つのチャネルを通して消費者マインドに影響を与えていることを報告
- Hollanders and Vliegenthart[31]：オランダにおいてネガティブな報道記事が消費者マインドを低下させていることを実証分析
- Soroka[32]：イギリスのデータを利用して、ポジティブなニュースとネガティブなニュースとでは消費者マインドに与える影響が異なり、人々はよりネガティブなニュース

---

[27] 上田晃三，須藤直，渡辺広太.「ＰＯＳデータによる「特売」の分析」，慢性デフレ真因の解明，渡辺努編．日本経済新聞. 97-114. 2016.

[28] 藤田隼平．「ＰＯＳデータを用いた経済分析の試み－小売価格と景気動向との関係性の検証－」．経済財政分析ディスカッション・ペーパー・シリーズ DP/17-4. 2017.

[29] T. Watanabe, H. Muroi, M. Naruke, K. Yono, G. Kobayashi and M. Yamasaki. “Prediction of regional goods demand incorporating the effect of weather ”. In Big Data (Big Data). IEEE International Conference. 3785-3791. 2016.

[30] M. E. Doms and N. J. Morin. “Consumer sentiment, the economy, and the news media”. FRBSF Working paper. 2004.

[31] D. Hollanders and R. Vliegenthart. “The influence of negative newspaper coverage on consumer confidence: The Dutch case”. Journal of Economic Psychology, 32(3), 367-373. 2009.

[32] S. N. Soroka. “Good news and bad news: Asymmetric responses to economic information”. Journal of Politics. 68(2). 372-385. 2006.

に反応していることを報告

- 饗場・山本 [33]：従来型と機械学習（ディープラーニング）の双方の手法のどちらが、内閣府「景気ウォッチャー調査」のコメント正しく分類できるかを比較し、機械学習による改善を確認
- Heston and Sinha[34]：ニューラルネットワークを活用することで、90 万件の記事データから株式のリターンを予測できるかを研究し、1 週間分のニュースから作成したセンチメントが、１四半期先までの予測力があることを報告
- 山本・松尾 [35]：再帰型ニューラルネットワーク（Recurrent Neural Network: RNN）により内閣府「景気ウォッチャー調査」を学習させたモデルを用いて、内閣府「月例経済報告」や日本銀行「金融経済月報」の指数化を行い、日経平均とある程度の相関があることを確認
- 五島他 [36]：景気ウォッチャー調査を畳み込みニューラルネットワーク（Convolutional Neural Network: CNN）により学習させたモデルを用いて、150 万以上の記事のスコア化を行い、ニュース記事が株・為替・債券等にどのような影響を与えているかを分析
- 塩野 [37]：物価関連記事からＣＰＩを予想するモデル（ニューラルネット）を作成し、高いパフォーマンスでＣＰＩのナウキャスティングを行うことができたことを報告
- Shapiro et al.[38]：文章の内容が特定の感情分類に分類される確率を計算する機械学習モデルを用いて新聞記事の指数化を行い、同指数を予測モデルに加えることでインフレ率等の予測精度が向上したと報告
- 内閣府政策統括官 [39]：簡易なニューラルネットワークモデルにより景気ウォッチー調査を機械に学習させたモデルを用いて、インターネット上にある「景気」を含む記事

---

[33] 饗場行洋，山本裕樹.「データサイエンスと新しい金融工学」. 野村證券.2018.

[34] S. L. Heston and N. R. Sinha. “News vs. Sentiment: Predicting Stock Returns from News Stories”, Financial Analysts Journal. 73(3). 1-17, 2016.

[35] 山本裕樹，松尾豊.「景気ウォッチャー調査の深層学習を用いた金融レポートの指数化」. 第 30 回人工知能学会全国大会. 2016.

[36] 五島圭一，山田哲也，高橋大志. 「畳み込みニューラルネットワークを用いた日次景況感指数の構築と資産価格変動との関連性」. 日本ファイナンス学会. 2017.

[37] 塩野剛志. 「人工知能とテキスト・データを活用した数量分析」. 日本銀行金融研究所ディスカッション・ペーパー・シリーズ 2018-J-9，2018.

[38] A. H. Shapiro, M. Sudhof and D. Wilson, “Measuring news sentiment”, Federal Reserve Bank of San Francisco. 2018.

[39] 内閣府政策統括官（経済財政分析担当）.『日本経済 2017 － 2018 －成長力強化に向けた課題と展望－』. 2018.

がどの程度ポジティブであるかについて指数化を行い、消費者マインドとのある程度の正の相関がみられたことを報告

3.1.10 行政データの活用例 1[40]

前橋市の空き家率推定に関するビッグデータの活用例について、六信氏及び馬場氏の研究を報告する。

六信氏はビッグデータによる自治体の政策のあり方について、複雑化する地域課題に対し「行政があらゆる公共的サービスを提供するのは限界」があるとし、これからのまちづくりのキーワードは「地域経営」であると述べている。すなわち、市民、企業・団体、行政それぞれが他人ごとではなく、「自分こと」として課題解決に取り組む必要がある。行政の役割は取り組みを「促し、つなげ、支援する」ことであり、そのためにはエビデンス（根拠となるデータ）が不可欠であるとしている。

また、以下の各担当からなる産官学連携の取り組みについても触れた。

- 官担当：自治体保有データ調査、自治体データ活用手続き（close data の場合）
- 学担当：空き家推定値算出
- 産担当：ダッシュボード構築

特に、産担当のダッシュボード構築では、以下の自治体の行政データを活用し、建物ごとの空き家確率を推定するためのモデルを用いていた。

- 住民基本台帳（所在地、性別、年齢）
- 固定資産税台帳（所在地、建築年、建物用途、構造）
- 水道使用料（所在地、月別使用料）

実際にダッシュボードに表示された空き家率情報を見ながら実態調査を行っており、空き家確率が 0.5 以上の家屋で実際に空き家であったのは、76.9%、空き家確率 0.5 未満の場合で、居住していた確率は 69.0%であったという。

一方、馬場氏は上述でいう学担当の立場として、自治体保有データを用いた空き家の時空間的予測モデルの構築を行った。

報告されたモデルは二時点間のデータを活用した機械学習的なモデルである。対象地域は前橋市（空き家率 173/2111 ≈ 8.2%）で、空き家の選定基準は、住居表示があり、中心市街地活性化区域内または周辺に立地する建物であることとしている。

データについては、以下に示すものを、集約などの前処理を施して用いる。

- 前橋市の自治体データ（括弧内は変数名）

---

[40] 2019 年 12 月 12 日に統計数理研究所で開催された研究集会「公的データの利用とプライバシー保護の理論」で発表された「ビッグデータで超スマート自治体を形成し、政策のあり方を見直す」（六信孝則：株式会社帝国データバンク）と「自治体保有データを活用した将来空き家率の推定」（馬場弘樹：東京大学・大学院工学系研究科）

住民基本台帳（建物内最高年齢、建物内最小年齢、建物内人員数）
固定資産台帳（木造ダミー、鉄骨造ダミー、RC-SRC ダミー、築年数）
水道栓情報（2014 年平均水道使用量、2018 年平均水道使用量）
空き家実態調査結果（緯度経度、空き家判定ダミー）

- 地図データ（括弧内は変数名）
  ゼンリンの住宅地図（建物面積）
  前橋市の地番地図

モデルは「時間不変特徴量＋2014 年時間変化特徴量」からなり、学習は教師データ「2015 年空き家判別結果」によって XGBoost を用いて行った。

2018 年の時間変化特徴量を用いて 2019 年の空き家率を推定した結果、以下の表が得られた。

図表 9　空き家率推定結果

予測データ		2019 年予測空き家	
		Yes	No
2015 年現地調査	Yes	110	63
	No	42	1896

空き家の予測正答率に関しては今後の発展が期待されるが、時間に関して一般化された予測モデルを構築したという点が評価される。

3.1.11 行政データの活用例 2[41]

福岡市の、ヘルスケア・ビッグデータと ICT を活用した地域包括ケアシステムの取り組みについて述べる。

このシステムはデータ収集から分析、在宅連携支援、情報提供までを一貫して行い、特に特筆すべきはデータ収集と分析についてである。収集されたデータ個人情報の匿名化や暗号化がされ、毎月更新されている。これまで蓄積されたデータは令和元年 9 月末現在で約 230 種 31 億件にも及ぶ。データには以下のものがある。

- 住民情報（約 20 年分）
- 介護保険に関わる被保険者情報（約 20 年分）
- 介護レセプト（約 20 年分）

[41] 雑誌「統計」の 2019 年 12 月号に掲載された「ヘルスケア・ビッグデータと ICT 活用による地域包括ケアシステムの実現」（福岡県福岡市、地方公共団体における統計利活用表彰（2018 年）統計局長賞）

- 国民健康保険及び後期高齢者医療加入者の被保険者情報（約 9 年分）
- 医療レセプトについて（約 9 年分）
- 健診情報については（約 12 年分）

平成 30 年度からシステム利用の職員研修も始まり、今後の発展に期待される。他の自治体でも同様のデータ収集が進めば、地方の行政上の利用だけでなく、製表への利用も期待できる。

3.2 ビッグデータ等の所在に関する情報

3.2.1 行政記録情報

行政記録情報については、内閣官房が、行政が保有するデータの棚卸結果を取りまとめ、公開[42]している。最新の「行政保有データ(行政手続等関連)の棚卸結果概要（令和 2 年 3 月とりまとめ)」(内閣官房情報通信技術（IT）総合戦略室（令和 2 年 8 月)) の内容を紹介する。

まず、「行政保有データ（行政手続等関連）の棚卸調査の概要」を以下に示す。

- 調査対象機関：国の行政機関(22 府省)
- 調査対象手続：各府省が所管する法令において規定されている全手続（法令に基づく行政手続及び民-民手続。約 56,000 種類)
- 調査対象：調査対象手続において得られるデータ
- 調査時点：平成 31 年 3 月 31 日（年間手続種類数等は、年度の記載がない限り、原則、平成 30 年 4 月 1 日～平成 31 年 3 月 31 日)
- 主な調査項目：データの管理状況、データの活用状況、データの活用先、データの公開状況、オープンデータ化未対応・非公開の理由、公開データのファイル形式

なお、棚卸リストは政府 CIO ポータル及び各府省庁の Web サイトに掲載されている。

行政手続等関連データについての、担当府省庁別のデータ管理状況は、以下のとおりである。

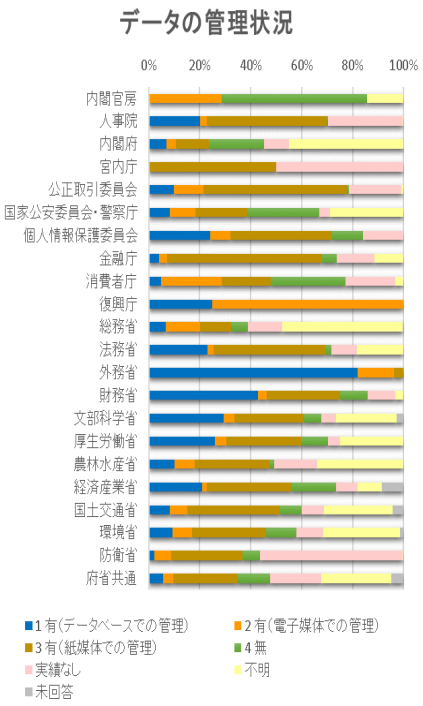
- 手続約 56,000 種類のうち、データの管理状況を把握しているのは約 35,000 種類(約 64%)。
- このうち、「電子媒体かつデータベース」又は「電子媒体」で管理されているデータのある手続の割合は約 34%にとどまっているが、前年（平成 30 年 3 月 31 日時点）の約 30%からは 4 ポイント上昇しており、増加傾向にある状況。
- また、省庁別に見ると、割合が 50%を超えている省庁もある一方、30%未満の府省もあり、バラツキがみられる。

---

[42] オープンデータ｜政府 CIO ポータル. 行政保有データの棚卸結果.
https://cio.go.jp/policy-opendata

図表 10　担当府省庁別のデータ管理状況

	総手続数（延べ数）	データの管理状況					実績なし	不明（※1）	未回答	管理されているデータのうちデータベースまたは電子媒体管理の割合
		計	1 有（データベースでの管理）	2 有（電子媒体での管理）	3 有（紙媒体での管理）	4 無				
内閣官房	7	6	0	2	0	4	0	1	0	33.3%
人事院	249	175	50	7	117	1	73	1	0	32.6%
内閣府	663	300	46	24	87	143	64	298	1	23.3%
宮内庁	2	1	0	0	1	0	1	0	0	0.0%
公正取引委員会	215	169	21	25	121	2	44	2	0	27.2%
国家公安委員会・警察庁	1,680	1,122	140	167	340	475	75	483	0	27.4%
個人情報保護委員会	25	21	6	2	10	3	4	0	0	38.1%
金融庁	3,826	2,823	150	127	2,324	222	567	436	0	9.8%
消費者庁	276	213	13	66	53	81	54	9	0	37.1%
復興庁	4	4	1	3	0	0	0	0	0	100.0%
総務省	4,034	1,565	271	547	483	264	547	1,910	12	52.3%
法務省	940	673	216	22	413	22	95	172	0	35.4%
外務省	133	133	109	19	5	0	0	0	0	96.2%
財務省	5,223	4,490	2,235	180	1,500	575	564	169	0	53.8%
文部科学省	768	520	225	34	209	52	44	183	21	49.8%
厚生労働省	9,209	6,480	2,381	423	2,695	981	425	2,304	0	43.3%
農林水産省	4,348	2,136	435	348	1,275	78	737	1,475	0	36.7%
経済産業省	6,228	4,577	1,298	123	2,075	1,081	533	583	535	31.0%
国土交通省	10,444	6,265	873	697	3,800	895	906	2,810	463	25.1%
環境省	2,919	1,688	273	224	844	347	308	882	41	29.4%
防衛省	675	295	15	45	188	47	378	2	0	20.3%
府省共通	3841	1827	220	146	964	497	770	1058	186	20.0%
総計	55,709	35,483	8,978	3,231	17,504	5,770	6,189	12,778	1,259	34.4%
総計に対する割合	100.0%	63.7%	–	–	–	–	11.1%	22.9%	2.3%	
①データ管理状況に対する割合	–	100.0%	25.3%	9.1%	49.3%	16.3%	–	–	–	
（参考）①の前年結果（割合）	–	100.0%	22.3%	7.7%	51.0%	19.0%	–	–	–	

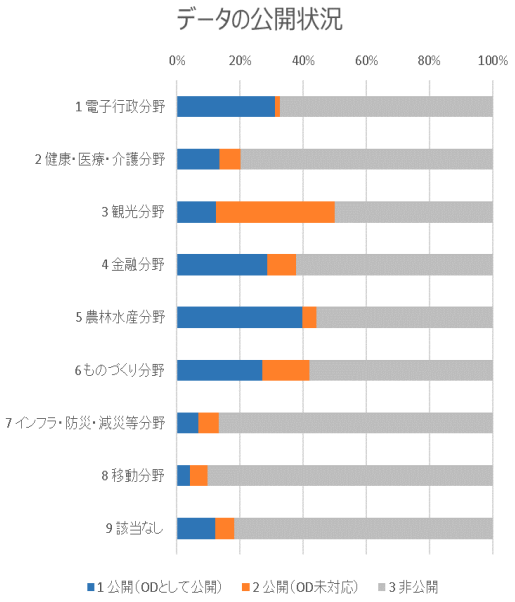


また、分野別のデータ公開状況としては、以下のとおりである。

- データベース又は電子媒体で管理されているデータ（延べ 16,400 種類）のうち、オープンデータとして公開しているデータが約 2,800 種類（約 17%）
- オープンデータ未対応で公開しているデータが約 900 種類（約 5 %）
- 非公開は 12,700 種類（約 77%）。

図表 11　各分野別のデータ公開状況

		データの公開状況			
		計	1 公開（ODとして公開）	2 公開（OD未対応）	3 非公開
該当する分野（※1）	1 電子行政分野	2,971	925	49	1,997
	2 健康・医療・介護分野	1,318	178	90	1,050
	3 観光分野	16	2	6	8
	4 金融分野	388	111	36	241
	5 農林水産分野	652	260	28	364
	6 ものづくり分野	188	51	28	109
	7 インフラ・防災・減災等分野	616	42	40	534
	8 移動分野	591	25	33	533
	9 該当なし	9,649	1,170	588	7,891
総計		16,389	2,764	898	12,727
割合		100.0%	16.9%	5.5%	77.7%
（参考）前年の割合			16.6%	6.2%	76.6%



3.2.2 民間のビッグデータ

民間のビッグデータの所在については、前出の経済産業省「平成２６年度ビッグデータを活用した新たな経済指標・分析手法の動向に関する調査研究報告」[43]で、8 企業・団体へのヒアリングを行うなどによる調査を行っており、POS データ、卸売りデータ、クレジットカードデータ、ポイントカードに分けてまとめている。その中から、データの所在等に関して参考となる部分を以下にまとめた。

1. POS データ

POS データについては、マーケティング会社を中心として整備が進んでいる。POS データでは、商品情報を JAN コード[44]レベルで管理している場合が大半ではある。POS データが有意になるかは商品マスターの整備状況次第であるが、それが必ずしも整備されていないこともあり、マーケティング会社では自社の POS データ用のマスターデータを丁寧に整備している。食品、飲料、日用品、化粧品、医薬品、家電等はすでにかなりの程度

---

43 9 頁脚注 12

44 JAN コードは「どの事業者の、どの商品か」を表す、世界共通の商品識別番号。
JAN コードは、商品のブランドを持つ事業者が、一般財団法人流通システム開発センターから貸与された GS1 事業者コードを用いて、商品ごとに設定される。通常、バーコードスキャナで読み取れるように、バーコードシンボルによって商品パッケージに表示される。
（一般財団法人流通開発センター. GS1 事業者コード･JAN コード(GTIN ジーティン)とは.（https://www.dsri.jp/jan/about_jan.html）より）

整備が進んでいる。

2. 卸売りデータ

ヒアリングされた会社は、メーカーと卸売業者との間の B to B 取引に関するデータを扱うサービスを行うもので、その会社がカバーしている業界取引全体に占めるシェアは金額ベースで 8 割を超えているものと推定されている。同社では、各レコードの中身をみることは一切ない。これは顧客との契約において定められた制限である。商品コードは基本的に JAN コードが用いられている。また、各レコードは、基本的にはユーザーが受信したことが確認できたら、ユーザーの再受信要請に応じる期間のみ保持し、システム上から消去してしまう。

3. クレジットカードデータ

クレジットカード会員の名前、住所、性別、職種等がデータ化されている。また、カード利用時の伝票情報もデータベース化されている。カード利用時の伝票情報としては、利用日時、加盟店名、業種などは分かるが、購入した商品・サービス名までは分からない。百貨店であれば、POS 情報によって購入した個別商品までわかる。

なお、クレジットカードデータを利用した公的統計づくりは、諸外国でも進展がみられる分野であるが、日本は、クレジットカードでの決済比率が低い国であるため、クレジットカードのみのデータを利用して消費動向全体を把握することは難しい可能性がある。

4. ポイントカードデータ

ポイントカードデータの多くも、POS データと消費者の属性を結びつけて把握することが可能である。

4 ビッグデータに関する情報（海外）

海外における、ビッグデータに関する情報、公的統計へのビッグデータの活用例についてまとめる。公的な報告書・資料、学会・研究会等での予稿集・レジュメに分けてまとめる。

4.1 公的資料等における海外のビッグデータ活用事例

4.1.1 平成２６年度ビッグデータを活用した新たな経済指標・分析手法の動向に関する調査研究[45]

1. 世界におけるビッグデータ活用の状況[46]

(a) ビッグデータプロジェクトに関するサーベイ調査とその概要

国連統計部（UNSD：United Nations Statistics Division）と国連欧州経済委員会（UNECE：

---

45 9 頁脚注 12 調査研究報告書. 42-49

46 9 頁脚注 12 調査研究報告書. 42 の脚注に「本節は Jansen, Ronald（2014）”UNSD/UNECE Survey and Project Templates” International Conference on Big

United Nations Economic Commission for Europe）は各国 65 のビッグデータプロジェクトの状況について、2014 年 10 月にサーベイ調査を実施した。サーベイではビッグデータを「ボリュームが大きく、即時性があり、バラエティの豊富なデータソースであり、効率的な方法によって測定、監督、管理、加工する必要があるもの」と定義している。調査対象者は、UNECE のビッグデータグループのメンバーと、国連のビッグデータに関するグローバルワーキンググループのメンバーである。33 の組織・戦略と、25 カ国 57 プロジェクトから回答を得た。

(b) サーベイ調査のまとめ

各国で実施しているプロジェクトの概要をみると、「予備的な調査（Exploratory/ research）」の割合が 49%と最も高くなっている。次いで、「パイロットプロジェクト（Pilot）」が 30%、「統計の作成（Production）」が 18%となっている。分野別にみると、多くのプロジェクトは、「経済・金融統計（Economic and financial statistics）」、「人口・社会統計（Demographic and social statistics）」、「物価統計（Price statistics）」に関連している。

外部機関とのパートナーシップの状況についてみると、約 3 分の 1 のプロジェクトでは「契約済み（Contract in place）」となっているが、45%のプロジェクトは「議論中（In discussion）」もしくは「プロトタイプ・テスト（Prototyping/ Testing）」となっている。パートナーに対する資金拠出は、約半数のプロジェクトで発生していない。

ビッグデータプロジェクトにおけるデータソースがあるかどうかをみると、「すでにデータソースを特定し、入手している（Yes, we have identified a new source and received the data）」がもっとも多く 29%、となっており、次いで「データソースがあり、データ提供者と議論をしている」が 26%、「データソースはあるが、データ提供者と議論は行っていない」が 20%となっている。

以上の調査結果から、Jansen（2014）は以下のように結論付けている。

- ビッグデータプロジェクトにおける大きな課題は、データに対するアクセス
- ビッグデータは多くの場合、民間でのみ所有されており、それらの多くはグローバル企業である。
- そのため、国際的な統計コミュニティが、データへのアクセスを得るために協調して交渉にあたる必要がある。
- いくつかの国からは、国際的なコミュニティが、国際的パートナーシップと、ビッグデータ活用に関する質と機密性確保のフレームワークの構築を行うことについて、提案があった。

Data for Official Statistics（2014 年 10 月）を参照している。」

2. 中国におけるネットワークデータの CPI への適用 [47]

近年、B2C[48]や C2C といった E コマースのウェブサイトが急速に発達しており、個人の消費行動に大きな影響を与えている。こうした中、中国国家統計局（市統計局）では、これらのネットワークデータを CPI 推計に適用する試みを報告している。

3. 中国におけるサーチエンジンデータを用いた住宅価格の予測 [49]

住宅市場は中国経済の成長の基点の一つであり、住宅価格は常に多くの人々の関心を集めている。だが、統計局が公表する住宅価格指数はリアルタイムでの公表ではないため、人々のニーズを満たすには至っていないという問題意識に基づき、サーチエンジンデータを用いた住宅価格の予測方法と結果を報告している。

4. 検索ログを用いたナウキャスティング

Choi and Varian（2012）[50]では、検索ログデータを用いて、さまざまな経済変数のナウキャスティング・予測を試みており、検索ログの利用によって予測力が高まるとしている。ラグ変数、外生変数に加えて、Google Trends の検索ログデータを用いることで、予測力を高めている。

イタリアの Istat（Italian National Institute of Statistics, 訳例：イタリア国家統計局）は、ビッグデータの潜在的な活用可能性を評価するため、Google Trends の検索データを補助変数として導入し、労働市場に関するモデルの試作を行っている。

国連統計委員会のビッグデータカンファレンス [51]では、既存の自己回帰モデルと、Google Trends の「job」カテゴリの検索結果や「job offers」という語句の検索結果を用いたいくつかの予測モデルとの比較を行っているプロジェクトを紹介している。その予備的な結論として、Google Trends をイタリアの労働市場に適用することで、予測や推定の精度を改善できる潜在的な可能性が示され、例えば失業率については「翌月の予測」や「従来の統計より小さな地域単位での推定（小地域推定）」に際しての有用性が期待されている。

---

47 9 頁脚注 12 調査研究報告書. 45 脚注に「本節は、Sun, YiBing（2014）”Application of network data in CPI”, Oct.29, 2014.を参照している。」

48 電子商取引の分野において、企業（business）と消費者（consumer）の取引のこと。

49 9 頁脚注 12 調査研究報告書. 47 脚注に「本節は Dong, Qian（2014） ”Housing Price Prediction Using Search Engine Query Data”, OCT.29, 2014 を参照している。」

50 Hyunyoung Choi and Hal Varian, “Predicting the Present with Google Trends”, Economic Record, 2012, June.

51 Emanuele Baldacci, “Web Scraping for Labour Statistics”, International Conference on Big Data for Official Statistics Organised by UNSD and NBS China Beijing, China, 28-30 October 2014.

5. OECD SWIFT Index

OECD の SWIFT Index は GDP を予測するために作られた指数であり、国際金融取引のデータに基づいて構築されている。SWIFT Index に基づく GDP 成長率の予測結果が下記の図である。GDP 成長率のナウキャスティングおよびフォーキャスティングを行うために、線形予測モデルを構築し、SWIFT Index を算出している。

6. クレジットカードを用いた観光消費の推計（ニュージーランド）

ニュージーランドの MBIE（Ministry of Business, Innovation & Employment, 訳例：ビジネス・イノベーション・雇用省）は、2012 年 12 月からカード式電子決済（electronic card transaction, 対象とするカードは debit card, credit card, charge card で、同国におけるコアな消費支出の 69%はこれらによって決済されていると見積もられている）のデータを用いた 2 種類の旅行関連統計を作成、公表している。そのうちの一つは、月次の RTI（The Regional Tourism Indicators, 遡及データは 200 年 1 月～）であり、もう 1 つは年次の RTE（The Regional Tourism Estimates, 遡及データは 2009 年 3 月期～）である。

RTI は 2008 年平均値を 100 とする支出金額に関する指数である。また、RTE は RTI のデータを用いて TSA（Tourism Satellite Account, SNA の旅行・観光サテライト勘定）と総額が一致し、MBIE の海外旅行者調査（International Visitor Survey）と整合するように推計された支出実額に関する統計である [52]。

7. mobile positioning data を用いた観光統計（エストニア）

Eurostat は、エストニアで「旅行関連統計に携帯電話の位置データを活用するためのフィージビリティスタディ」を実施（Eurostat Contract No.30501.2012.001-2012.452）し、その結果を 2014 年 6 月に報告書としてまとめている。プロジェクトは 2013 年 1 月から 2014 年 6 月にかけて行われ、その目的は携帯電話の位置情報データを使って旅行関連統計（tourism statistics）を作成することの可能性と限界を明らかにすることであった。

4.1.2 小西葉子氏　RIETI-BBL セミナー「ビッグデータと公的統計調査：「作る・伝える・活かす」工夫」による海外の動向 [53]

経済産業省のビッグデータプロジェクトの成果の報告のうち、「海外の動向（平成 30 年度海外調査の結果より）」として、以下の 3 か国の状況が紹介されている。

- イギリス　消費者物価指数（CPI）の一部品目で、スキャナーデータ（POS）を活用している事例はあるが、民間企業が保有するビッグデータを広範囲に収集し、公的な統計

[52] http://naratourismstatisticsweek.visitors.jp/global/index.html

[53] 22 頁脚注 19 参照　https://www.rieti.go.jp/jp/events/bbl/20012201_konishi.pdf

調査として実施した事例は存在しなかった。

- オランダ　消費者物価指数（CPI）の一部品目で、スキャナーデータ（POS）を活用している事例はあるが、民間企業が保有するビッグデータを広範囲に収集し、公的な統計調査として実施した事例は存在しなかった。
- シンガポール　民間企業が保有するビッグデータを広範囲に収集し、公的な統計調査として実施した事例は存在しなかった。

4.2 海外の学会・研究会等の報告書によるビッグデータ活用事例

府省設置のワーキンググループ（WG）の成果が論文等として発表された刊行物や学会や研究会等の予稿集や発表スライド等から収集した活用例等を報告する。

4.2.1 Big Data for 21st Century Economic Statistics[54]

2019 年 3 月にワシントン DC で開催された National Bureau of Economic Research 主催の研究会議の抄録を紹介する。

ここで紹介した他に、International Conference on Big Data for Official Statistics Organized by the United Nations Statistics Division（UNSD）and National Bureau of Statistics of China[55]などでもいくつかの Big Data を公的統計作成に利用する試みが発表されている。

1. The Scope and Impact of Open Source Software: A Framework for Analysis and Preliminary Cost Estimates (Carol Robbins, Jose Bayoan Santiago Calderon, Gizem Korkmaz, Daniel Chen, Sallie Keller, Aaron Schroeder, Stephanie S. Shipp, Claire Kelling)

オープンソースソフトウェアは、献身的なユーザーコミュニティによって育まれた特殊なアプリケーションとして、また毎日数百万人が使用するプラットフォームの基盤となるデジタルインフラストラクチャとして、いたるところに存在する。このタイプのソフトウェアは、企業、大学、政府研究機関、非営利団体、および個人としての貢献を通じて、民間部門内および民間部門外の両方で開発、保守、拡張されている。Robbins、Korkmaz、Calderon、Kelling、Keller 及び Shipp は、これらのセクターによって作成されたオープンソースソフトウェアの範囲と影響を文書化する方法を提案し、プロトタイプを作成した。これにより、公的資金による研究成果の既存の評価尺度を拡張した。研究者は、R、Python、Julia、および JavaScript のオープンソースソフトウェア言語のパッケージを開発するコストを見積るとともに、R パッケージの統計を再利用する。これらの再利用統計は、相対的な価値の尺度となる。研究者は、R、Python、Julia、および JavaScript を開発するためのリソースコストが 2017 年のコストに基づくと 30 億ドルを超えると推定している。

---

54 https://conference.nber.org/conferences/2019/CRIWs19/summary.html

55 https://unstats.un.org/unsd/trade/events/2014/Beijing/

2. Estimating the Benefits of New Products (W. Erwin Diewert, Robert C. Feenstra)

統計機関が直面する主な課題は、商品の入手可能性の変化に応じて価格と数量の指標を調整する問題である。この問題は、ある商品層の製品が小売店で表示されたり消えたりする時に、スキャナーデータのコンテキストで発生する。ヒックスは、指数理論への経済的アプローチの文脈でこの問題に対処するための予約価格の方法論を提案した。Feenstra と Hausman は、ヒックスのアプローチを実装するための特定の方法を提案した。Diewert と Feenstra はこれらのアプローチを検証し、Feenstra の論文と同じように計算された一定の弾性ゲインの半分を取ることを推奨している。このゲインは、弱い条件下では、ハウスマンが提案する需要曲線の線形近似から得られるゲインを上回るが、許容できる程度に近い値となっている。研究者は、CES 型と 2 次効用関数を使用して得られたゲインとを比較する。さまざまなアプローチが、オンラインで入手可能な冷凍ジュース製品に関するスキャナーデータを使用して実装されている。

3. Transforming Naturally Occurring Text Data into Economic Statistics: The Case of Online Job Vacancy Postings (David Copple, Bradley J. Speigner, and Arthur Turrell)

Copple、Speigner 及び Turrell は、地域と職業の両方による不均一性を含む英国の労働市場のより詳細な状況を得るために、公的データと自然発生データの両方を結びつけている。新しい自然発生データは、英国の大手求人サイトの 1 つに企業が投稿した 1500 万件の求人広告である。研究者は、この厄介なオンラインデータを、セクター、地域、職業の公式分類にマッピングする。人材派遣会社自身の非公式の職種分野は、手動で公式の部門分類にマッピングされ、部門ごとの公式の求人統計を利用して、バイアスを減らすために、データの重み付けを変更するために使用されている。研究者は各求人情報の緯度と経度を直接地域にマッピングする。標準的な職業分類（SOC）コードを用いた公的統計に一致させるために、研究者は、各職種に関連するテキストデータを取得して SOC コードにマッピングする教師なし機械学習アルゴリズムを開発している。このアルゴリズムは、ジョブの説明など、ジョブに関連付けられているすべてのテキストを使用し、テキストを公的分類にマッピングする必要がある他のさまざまな状況で使用できる。研究者は、アルゴリズムを GitHub 上で Python パッケージとして利用できるようにすることを予定している。これらのデータを公的統計と組み合わせて使用することで、危機後の期間で長く続く特徴である英国の弱い生産性と生産高の伸びを調べることができる。失業者と求人の間の労働市場の不一致は、以前は英国の生産性「謎」の 1 つの要因として関係していた（Patterson、Christina、et al.「間違った場所で一生懸命働く：英国の生産性の謎に対する不一致に基づく説明。」European Economic Review 84 (2016):42-56.）。完全にラベル付けされたデータセットを使用して、研究者たちは、職業的および地域的なミスマッチを解きほぐすことで、危機後の生産性と生産量の伸びがどの程度向上したかを調べている。結果に対する不一致の影響は、生産性、緊密性、およ

びマッチング効率の散らばりによってもたらされている(研究者はこれについて新しい推定値を提供している)。研究者は、これらのサブマーケット間での著しい散らばりの証拠を、重要な異質性を隠している集約データを用いて、示している。以前の研究とは反対に、研究者たちは、職業上のミスマッチを解きほぐすことが、危機後の成長に弱い影響を与えていたであろうことを発見している。ただし、地域のミスマッチが解消されると、現実の変遷に比べて生産性と生産性の伸びが大幅に向上し、危機前の傾向に戻ることになる。研究者は、自然発生データが公式統計を強力に補完する方法を実証している。

4. Nowcasting the Local Economy: Using Yelp Data to Measure Economic Activity (NBER Working Paper No. 24010) (Edward L. Glaeser, Hyunjin Kim and Michael Luca)

オンラインプラットフォームの新しいデータソースは、地域の経済活動の尺度付けに役立つだろうか？米国センサス局などの機関からの政府データセットは、地方レベルでの地方経済活動の標準的な尺度を提供している。ただし、これらの統計は通常、複数年の遅れの後にのみ表示され、一般向けのバージョンは郡または郵便番号レベルに集約される。対照的に、Yelp などのオンラインプラットフォームからクラウドソーシングされたデータは、多くの場合、同時的であり、公式の政府統計よりも地理的に細かい。Glaeser、Kim、Luca は、Yelp のデータが経済活動をリアルタイムに細かく、ほぼすべての地理的規模で測定することで政府の調査を補完できるという証拠を提示している。Yelp でレビューされた企業とレストランの数の変化は、郡のビジネスパターンにおける、施設とレストラン全体の数の変化の予測に使える。同時かつ時間差のある Yelp データを使用するアルゴリズムは、アルゴリズムの生成に使用されていないテストサンプルで、時間差のある CBP データを考慮した後の残差分散の 29.2%を説明できる。このアルゴリズムは、人口密度が高く、裕福で、教育水準の高い郵便番号地域に対してより正確である。

5. A Machine Learning Analysis of Seasonal and Cyclical Sales in Weekly Scanner Data (Rishab Guha, Serena Ng)

Guha と Ng は、2006 年から 2014 年までの郡レベルで 108 のグループについて収集された毎週のスキャナーデータを分析している。データには、正確な周期ではないが、相互に依存する多次元の週ごとの季節変動を示している。既存の単変量手順は不完全であり、集約時に強い季節性を示す調整済系列を生成する。研究者は、郡全体で情報をプールするパネルデータステップを用いることで単変量調整を拡張することを提案している。その後、機械学習ツールを使用して、年周期の季節変動を除去する。調整された予算シェアの需要分析では、3 つの要因が見つかる。1 つはトレンドであり、2 つは循環的なもので、消費者信頼感のレベルの変化とよく一致している。大不況の影響は、消費者が非必須商品から離れて家庭料理に代わることによって、場所や製品グループを変化させる。したがって、データは、季節の影響が除去された後のローカルおよび集約的な経済状況について有益である。2 段階の方法

論は、これらの変動が横断的に依存している限り、他の種類の攪乱変動を除去するように適合させることができる。

6. Re-Engineering Key National Economic Indicators（Gabriel Ehrlich, David Johnson, John C. Haltiwanger, Ron S. Jarmin, Matthew D. Shapiro）

企業や家庭からデータを収集する従来の方法は、ますます困難に直面している。これには、調査に対する回答率の低下、従来のデータ収集方法のコストの増加、経済の急速な変化に対応することの困難性が含まれている。実質的にすべての市場取引のデジタル化は、主要な国家経済指標を再構築する可能性を提供する。統計システムの課題は、このデータが豊富な環境での運用方法である。Ehrlich、Haltiwanger、Jarmin、Johnson 及び Shapiro は、データ源でアイテムレベルのデータを収集し、そのようなデータインフラストラクチャと整合性のある測定方法を使用して重要な指標を構築する機会に焦点を当てている。ユビキタスな取引のデジタル化により、価格と数量をデータ源で同時に収集または集約できる。経済統計に関するこの新しいアーキテクチャは、販売された品目の急速な変化から生じる課題を生み出している。研究者は、大規模な品目レベルのデータで価格と数量の指標を推定するために最近提案された幾つかの手法を探索している。これらの方法は非常に有望であるが、公的な経済統計の基礎となる準備が整うには、さらに多くの研究が必要である。最後に、研究者は、データ収集のトランザクションから公的統計を構築すること、および 21 世紀の統計機関の能力と組織に対する意味付けに取り組んでいる。

7. From Transactions Data to Economic Statistics: Constructing Real-Time, High-Frequency, Geographic Measures of Consumer Spending（Shifrah Aron-Dine, Aditya Aladangady, Wendy Dunn, Laura Feiveson, Paul Lengermann, and Claudia R. Sahm）

消費者支出に関するデータは、経済活動を追跡するため、経済政策立案者にリアルタイムで通知するために重要である。Aladangady、Aron-Dine、Dunn、Feiveson、Lengermann 及び Sahm は、消費者支出に関する新しいデータセットの構築について説明している。彼らは、大規模な支払い技術会社からの匿名化されたカード取引を、支出発生のたった数日後に利用可能な支出の毎日の地理的推定に変換している。センサス局の毎月の小売売上高の調査は、景気の循環的な状況を監視するための主要な情報源であるが、全国的な統計であり、局所的ショックまたは短期ショックの研究にはあまり適していない。さらに、調査のリリースの遅れとその後の（時には大規模な）改訂は、政策立案者にとっての有用性を低下させる可能性がある。公的調査を拡大してより詳細で迅速な公開を行うと、費用がかかり、回答者の負担が大幅に増加する。このアプローチは、消費者の支出に関するデータとして民間企業からのクレジットカードやデビットカード及びその他の電子決済データを使用することにより、これらの情報のギャップを埋めるのに役立つ。研究者の日次シリーズは 2010 年から現在まで利用可能であり、集計して公式のセンサス統計に類似した全国的な月間成長率を求めること

ができる。研究者のデータセットに含まれる新しい高頻度の地理情報のアプリケーションとして、Hurricanes Harvey と Irma の支出への影響をリアルタイムで定量化している。

（この研究は公的統計の結果と比較しているが、実際に新しい公的な統計指標とはなっていない。）

5 まとめ

アクションプラン「情報ソースの多様化に対応するための研究」に対する活動として令和元年度に収集した情報について述べた。

総務省「ビッグデータ等の利活用推進に関する産官学協議のための連携会議」、経済産業省「POS データを活用した商業動態統計の速報性向上に関する研究」が資料として充実していた。

これらから集められた公的統計の事例のうち、国内では、経済産業省「商業動態統計(家電大型専門店分野)」では、家電等の POS データによる指標作成がこれまでの調査に代わって公的指標となろうとしている。総務省においても、宿泊料金等の調査がウェブスクレイピングによるものに代わろうとしている [56]。

海外においては、この報告に挙げた事例の中では、ニュージーランドの旅行関連統計を除いて公的指標としているものはない。EU の一部の国では、公的統計調査の代わりにウェブスクレイピングを用いているところもある。今回調べた中には小西氏の報告の中にオランダの例がある。

これらの事例は、本報告のビッグデータを活用した視点において紹介した「母集団代表制の担保」等の課題がこれらの事例において問題になり難いものとも考えられるが、今回は、公的統計に対する課題やそれが事例に対してどのように解決されているか等について検討が及んでいない。今後は、個々の事例についてどのような課題があり、それがどのように解決されている等の体系的な調査・研究を進める必要がある。

---

[56] 統計委員会. 諮問第 142 号の答申 小売物価統計の指定の変更及び小売物価統計調査の変更について. 2020 年 9 月 9 日
https://www.soumu.go.jp/main_content/000706451.pdf

製 表 技 術 参 考 資 料 42

令和３年６月発行

編集・発行　　独立行政法人　統計センター

〒162－8668

東京都新宿区若松町 19－1

電　話　代表　03（ 5273 ）1200

掲載論文を引用する場合は、事前に下記まで連絡してください

情報技術センター技術研究開発課　　TEL ： 03－5273－1368

E-mail ： research-info@nstac.go.jp