統計的データ解析を取り巻く環境

田村　義保　　統計思考院　特任教授

統計学が「科学の文法」と呼ばれていることは良く知られている。2012年のHarvard Business Reviewの"Data Scientist: The Sexiest Job of the 21st Century"には、2009年8月のHal Varian氏（Googleチーフエコノミスト）の発言、"The sexy job in the next 10 years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?"が引用されている。今年は、2020年なので、最初の発言からは10年以上たってしまっている。

もう、「統計家」がもっとも、セクシーなジョブであることは終わってしまったのあろうか？残念ながら、日本においては、最初から、そのように思っていた人は無かったように思う。日本では、「データサイエンティスト」がこの10年で、最もセクシーなジョブと思われているように考える。私は、統計的手法は、データサイエンスの中核をなしていると考えている。職業名が「データサイエンティスト」であっても、「統計家」であっても良いが、データ解析に使われている手法の中心が「統計的データ解析」であることは認識して欲しいと考えている。

このような思いを持ちつつ、昨今、ちまたを騒がせている、「統計、統計学」に関係する「バズ」ワードについて述べつつ、データ利活用の現状について解説することが、本ポスターの目的である。

● ビッグデータ　● データサイエンティスト　● シンギュラリティ　● DX(デジタルトランスフォーメーション）
● アナリティクス　● データ中心科学　● AI（機械学習、Deep Learning)　● SDGs

今でもそうだが、ビッグデータが一時期、一番、聞いた用語であるように思われる。本ポスターもビッグデータに関連したことを中心にしている。当初は3Vと言われたビッグデータも今では5Vと言われることもある。

容量(Volume)　種類 (Variety)　頻度・スピード(Velocity)　価値（Value)　正確さ（Veracity）

頻度・スピードこそがビッグデータの本質だと考えている。スピードとも関係しているが、容量の増加についての参考情報をあげておく。

2014 年時点でのhttp://www-06.ibm.com/software/jp/data/bigdata/）によると、1 日あたり、2.5EB（エクサバイト）であった。また、https://www.otsuka-shokai.co.jp/media/byline/numbers/20160926.htmlによると2020 年の全世界のデジタルデータの生産量は44ZB（ゼッタバイト）（2013 年は4.4ZB）、https://dcross.impress.co.jp/docs/news/000202.html によると2025 年に163ZB になる。

京コンピュータ関連で使っていた磁気ディスク装置が100PB、統計科学スーパーコンピュータの磁気ディスク・ユーザー領域が3.6PBであることを考えるとその大きさが分かるであろう。ビッグデータは大きいだけでなく、その内容の豊富さ、重要さから、その利用法に、関心が集まっている。ビッグデータを活用して公的統計作成等に活かそうという試みや分析例の一部として次のようなものがある。

- 政府はビッグデータを活用した新たなマクロ経済指標の作成に乗り出す。実現すれば、世界で初めての試みになる。
（ロイター　2013年9月17日12:02 午後7年前更新）
現在の月次動向指数より、短い周期で公表可能な新指数作成を目指したようである。
- https://www5.cao.go.jp/j-j/wp/wp-je18/pdf/p01032.pdf　には、POSデータや新聞記事を用いた分析事例が出ている。(2018年内閣府白書）
- https://www.boj.or.jp/research/brp/ron_2015/data/ron150625a.pdf　には、景気ウォッチャー調査のテキスト分析の試みが出ている。(2015年BOJ Reports & Research Papers）

情報通信白書の平成25年版から27年版には民間企業のビッグデータ活用事例が出ていたが、なぜか、平成28年版以降には記述が無かった。民間には、消費者物価指数に関係する、JCB消費NOW、日経・東大物価指数（日経CPINow）などがある。公的統計作成に関するビッグデータ活用について「第10回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」(2020年9月30日）の資料を参考にしてまとめておく。

公的統計	消費物価指数 (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、クレジットカードデータ	POSデータ	携帯基地局情報
概況	宿泊費に関する情報をウエブスクレイピングにより収集し、CPIを作成。従来は、宿泊施設に調査員が出向いていた。	GDPの内の家計最終消費支出を予測するための指数として、CTIを作成している。CTIを予測するために、POSデータ等の情報を用いて、家計最終消費支出の早期予測を目指している。	商業動態統計の家電などの売り上げを従来は専門店から情報を集めていた。専門店のPOSデータを収集している機関からのデータ収集にしている。	PT調査データとビックデータを組み合わせた総合交通調査体系構築を目指している。

今年になって、非常に良く見聞きする用語に、デジタルトランスフォーメーション（DX)とSDGsがある。まずはデジタルトランスフォーメーションについて説明する。デジタル化が非常に遅れているのは、3年前に某地方自治体に漁獲量データの存在の調査に行った時に広辞苑くらいの資料集を見せてもらい、「必要ならお貸しするので、コピーをとってください。」と言われたことがある。デジタル化は直近の10数年ほどしかやっていないとのことであった。おそらく、すべての組織で、こんなものであると思う。

- Wikipediaによれば、『2004年にスウェーデンのウメオ大学のエリック・ストルターマン教授が提唱した。彼は「ITの浸透が、人々の生活をあらゆる面でより良い方向に変化させる」と定義し・・・』 とある。このページにはIDCによるDXの定義もある。
- https://dcross.impress.co.jp/docs/news/000202.html（DIGITAL X 編集部）には、『IDC は、企業がIoT やAI を使って業務のあり方を変革する「デジタルトランスフォーメーション（DX）」を成功させるためには、AI の分析対象となるデータの量を最大化しなければならないとする。そのためには、IoT 端末から得られるデータだけでなく、人が作り出すデータを組み合わせる必要があると指摘する。』 とある。

SDGsも今年になってから急にマスコミに登場するようになったと感じている。外務省のWebには『2015年9月の国連サミットで採択された「持続可能な開発のための2030アジェンダ」にて記載された2030年までに持続可能でよりよい世界を目指す国際目標です。17のゴール・169のターゲットから構成，・・・』とある。DXやSDGsのことを書いたのは、すべて、ビッグデータやそのデータ解析と関係していると思っているからである。文科省はデータサイエンティスト育成に力を入れている。JAXAはSDGsに関係した環境指数の作成を目指している。一昨年には、携帯端末位置情報と国勢調査人口との比較の研究も行われていた。統数研の研究者がこれらに貢献していくことを願って、結びとしたい。