第１部 論文集
原著

データに基づく施策形成におけるデータ分析スキル向上のための行政と大学の連携 ～R 共同学習環境の構築～

高崎 光浩[1] 佐々木 和美[2] 楠田 詞也[2] 川原 康義[3]
北島 健一[3] 古川 修一[3] 松田 智大[4]

佐賀大学 [1]
佐賀県統計分析課 [2]
佐賀県健康増進課 [3]
国立がん研究センター [4]

要 旨

【目的】 情報通信技術の社会生活への浸透によりデータサイエンスへの期待が高まっている。行政においてもデータに基づく政策形成（Evidence-Based Policy Making）という考え方として拡がりつつあるが、各分野の課題背景を理解した上で整理し解決する力を有し、統計解析やプログラミング技術を駆使して得られた知見の意味づけができ、使える形にして運用できる能力を備えた人材は稀である。
能力を備えた人材が協力しお互いの得意分野を活かすことが有効であると考え、行政と大学が連携した共同利用環境を構築しその効果について検討した。
【方法】 分析ツールとして広く用いられている統計解析用プログラミング言語 R を用いた。R は多機能であるが、日常的にデータ分析に携わっていない者が単独で使いこなすにはハードルが高い。パソコン内で単独利用する形態ではなく、統合開発環境 R Studio を介した web サービスとして利用し、行政の担当者が実際のデータを用いて分析するのを大学教職員がアドバイスする方式とした。
【結果】 R の分析では、ライブラリと呼ばれる分析用スクリプトを目的に応じて追加する必要があり、初心者が難しさを感じる一因となっていることがわかった。本研究では web サービスとしての利用なので全員が常に同一の環境で行うことができた。
R は Excel と同様に GUI 操作とスクリプト（簡易プログラム）を記述する２つの利用法がある。スクリプトによる方法は敬遠されがちであるが、コメント等を細かく入れることにより処理の理解が容易になり、分析の誤りも見つけやすくなることが実感できた。データを変えて同じ分析を行う場合は、GUI では全ての操作を手作業で繰り返すため操作ミス等が生じやすいが、スクリプトではデータファイル名の書き換えだけで確実に同じ分析ができ人的ミスが減らせることもわかった。
【結論】 web 上の R の共同利用環境を用いた行政と大学との連携はデータ分析スキルの向上に貢献できると示唆された。

１．はじめに

行政の役割は「社会的課題」の解決に向けた様々な対応である。限られた資源を効率的に活用しながら、社会の動向を多面的かつ体系的に把握、分析し、優先して対応すべき課題を見極め、解決の処方としての施策を決める。施策の決定に当たっては、その過程を体系的なエビデンスとして捉え、科学的合理性のある政策を形成することが求められる。行政においても Evidence-Based Policy Making として拡がってきている [1,2]。

情報通信技術の社会生活への浸透により、大量のデータが採取、蓄積されており、分析可能となってきている。

このような環境の中、データサイエンスへの期待が高まっている。大量のデータを蓄積しそこから新たな知見を得ることについては 1960 年代から取り組まれていたが、利用できるデータの質と量、コンピュータの処理能力等の制約により十分な成果が得られているとは言い難かった。ところが、近年の情報通信技術の飛躍的進歩によってデータ量と分析に利用できるコンピュータ資源は十分な実用レベルに達したといえる [3]。

一方で、各分野での課題背景を理解した上で整理し解決する力を有し、統計解析やプログラミング技術を使いこなし、得られた知見の意味づけを整理し、採取・蓄積したデータを使える形にして、さらに運用できる能力を備えたデータサイエンティスト 4,5) という人材が求められている。これらの要素を全て備えた人材を確保するのは容易ではないが、一つ一つの能力を持った複数の人材が協力することによって、Evidence-Based Policy Making の実現に近づくことが可能と考えられるため、大学と行政の連携によってそれを検証することとした。

行政における課題解決の場面において、データサイエンティストに求められる能力と比較してみると、行政担当者は、数学・統計学の知識やコンピュータスキルは一般に不足しているが、課題の認識と分析に利用できるデータは豊富に有している。大学教職員は、数学・統計学の知識やコンピュータスキルは有しているが、具体的データを持っておらず、解決すべき政策課題にも直面していないことが多い。

２．目的

行政と大学がそれぞれの能力、役割を発揮して連携することによって、データサイエンティストとして備えるべき能力が満たされ、科学的根拠に基づく政策形成に貢献できるのではないかと考え、その第一歩としてデータサイエンスにおける分析ツールの共同利用環境を整えその効果を検証した。

3．方法

科学的根拠に基づく政策形成にはデータ分析が不可欠であるため、大学と行政の担当者が共同で利用できる分析環境を構築した。本研究への協力が得られシステムを利用している者の背景は表１の通りである。

表１　本システムの利用者の背景

所属区分	職種	勤務形態	人数	備考
大学	教員・研究者	常勤	1	主たる研究者
大学	事務職員	常勤	1	がん登録実務者
研究所	研究者	常勤	1	統計分析の専門家
行政	事務職員	常勤	5	がん対策担当者
行政	事務職員	常勤	1	統計分析部門担当者

行政担当者も大学教職員も業務で使い慣れている Excel（Microsoft 社）で日常のデータ処理は行っているが、Excel でのデータ分析には限界があり、本格的な統計分析を行うには統計分析専用ソフトウェア（SAS、SPSS、JMP 等）が必要となるが、これらは一般的に高価であり、行政機関で保有することは難しい。

そこで本研究では、無料で利用でき、信頼性も高く多くの科学論文の統計分析で利用実績がある統計分析用プログラミング環境 R[6、7] を用いることとした。

また、R は Windows、Macintosh、Linux とコンピュータのオペーレーティングシステムの制約なく利用できるため、利用者のパソコンにインストールして R を単独で利用することも可能であるが、本研究のように複数で連携して分析を行うには、各利用者の PC で R の利用環境を同一にしておく必要があり手間がかかる。そこで、R 用の IDE（統合型開発環境）である RStudio をサーバ上で R とともに稼働させることにより、関係者全員が同じ環境で R を web サービスとして利用できるようにした。

R の分析環境を構築したサーバの仕様は表２の通りである。

表２　R の分析環境を構築したサーバの性能等

サーバ	HP Proliant DL160 Gen9
CPU	Intel(R) Xeon(R) CPU E5-2603 v3 @ 1.60GHz
メモリ	32GB
SSD	1TB
OS	Linux Kernel 3.10.x (CentOS 7)

分析環境は大学のサーバに構築した。デフォルト設定における R は、データや分析条件等が各ユーザーのホームディレクトリ（ユーザーフォルダ）に保存される。本研究では分析用データを共有したり、スクリプトを相互に編集したりすることもあるため、共有ディレクトリを作成してホームディレクトリにリンクを作成することにより、容易に共有フォルダと個人フォルダを使い分けられるように工夫した。

本システムで取り扱うデータについては、操作方法の例示には練習用のダミーデータをサンプルデータとして用いるが、利用者は実データを用いて操作を行うこととした。ただし、個人情報を含まないデータに限定した。

4. 結果

本研究で構築した共同分析環境を利用するには、インターネットに接続された端末からブラウザでアクセスする。RStudio のログイン画面（図１）が表示されるので、指定されたユーザーID とパスワードを入力し、[Sign In]ボタンをクリックすることで、R の分析環境が利用できるようになった。

近年、個人情報の不正取得等のため、ネットワーク資源への悪意を持った攻撃が増えており、その対応策として業務用端末からのインターネット利用に関して利用条件が厳しくなっている。それにより、システムの安全性は向上するが、これまで利用できていたサービスが利用できなくなるなどの問題も生じている。本システムもインターネット上の web サービスとして提供しているため、組織のセキュリティ設定の制約の影響を受ける可能性が心配されたが、本分析環境の利用に関して影響はなく、全ての機能を問題なく利用できた。

サンプルデータとそれを用いて基本的な統計分析を行うスクリプト（図２a）を大学教員が提供し、行政担当者が R にログインし、スクリプトを開き、Source ボタンをクリックすることにより、サンプルデータが読み込まれ、データの基本統計量が計算され、２つの群の平均値に差があるかどうかを確認する t 検定を行い、箱ひげ図をプロットするという一連の基本的分析操作を確認した。

実際に R を使用してみると、多機能で様々な分析に利用できるのは事実であるが、メニューやボタンが全て英語であるため、難しさが助長されていた。データのドラッグ選択、処理をボタンやメニューでの選択などマウス操作だけで分析を完結させることができず、手入力での分析が基本となっていることに対し、ほとんどの利用者が難しいと訴えていた。

図１　R Studio のログイン画面

(©RStudio® under AGPL v.3)

図２　R で分析を行うためのスクリプト例

(©RStudio® under AGPL v.3)

R で分析を行うには、ライブラリと呼ばれる分析用スクリプトを目的に応じて追加する必要があるが、本研究ではRの分析環境をwebサービスとして構築しているため、それらの環境設定はサーバー側で一度行うだけで全員が同じ環境で利用できた。

また、RStudio は R 専用の統合型開発環境であるため、初心者が苦手とする R のコード入力を補助するヘルプ機能や補完機能が充実しており、R を単独で用いるより R スクリプトのコード入力が容易に行えた。

5．考察

使い慣れた Excel を用いる分析と R による分析の比較

R はスクリプト（プログラム）を手入力して分析するのが基本であるため、日常的にデータ分析に携わっていない者が単独で使いこなすにはハードルが高い。しかし、Excel での分析では、離れた列に入力されているデータを用いて分析する場合、それらが隣り合うようにコピー&ペーストしてデータを移動させる必要がある。また、A 列に性別が入力されていて、B 列に血圧が入力されている場合、B 列の血圧は男性と女性の血圧が混在している。このような場合に男性と女性の血圧をグラフに示したくてもフィルタリングなどをして男性と女性のデータを別の列に配置し直す必要がある。さらに、そのような分析を別のエクセルデータで行うには、全く同じ操作を手作業で繰り返す必要がある。R の場合、一度スクリプトを記述すれば、データファイル名の部分だけ書き換えるだけで別のデータで同じ分析を行うことができる。１つのデータ列に男性と女性の血圧値が混在している場合でも、性別が入力されている A 列の値を用いてグラフを描き分けることも容易である。

スクリプトを記述して分析を行う利点

多くの分析は Excel でも可能であり、わざわざ難解なスクリプト記述方法を覚えることには抵抗が大きいと考えられる。「Excel でも同じ操作の繰り返しぐらいマウス操作が増えるだけだから問題ではない。」と感じるかもしれないが、実際の統計・分析を行う場面においてスクリプトを記述するメリットは大きく、Excel での操作においては、データを選択する際に異なるデータ範囲を選択することが多々あり、誤りが生じやすい。コピー&ペーストする際に間違うこともあり得る。メニューから操作を選ぶ際に誤ったコマンドを選ぶ可能性もある。いずれの場合でも操作の記録が残らないので、誤操作を確認する手段がなく、設定ミスや操作ミスが見つかれば、最初から全ての操作を繰り返さなければならない。その再操作の過程でも同じように誤操作が起こる可能性が残る。一方、スクリプトを記述する R では、全ての操作手順が文字として残っているので、分析実行前に間違いがないか確認でき、疑義が生じた場合は後からでもスクリプトを見直すことで確認ができる。共同研究等でデータ分析を複数の関係者で行う場合も、スクリプトファイルを共有するだけで確実に同じ分析を行えることが保証される。

R を web サービスとして共同利用するメリット

R は各自がパソコンにインストールして利用するのが一般的な利用法である。R は統計分析用の道具箱のようなもので、インストールしたばかりの R にはよく使う道具だけが入っている。行いたい分析によって、追加の道具が必要になり、それらを追加することにより便利さが増すが、使っているうちにそれぞれの R の機能に違いが生じることとなる。共同研究で R を利用する場合は、このような追加機能の環境まで統一させておかないとディスカッションが円滑に進まなくなる可能性がある。

web サービスとして共通の R 分析環境を用いれば常に同じ機能が利用できる。また、スクリプトを共有ファイルに入れておくことにより分析手法を全員で確認した上で、同一の手法で分析が確実に行えることになる。

大学と行政担当者で一緒に R を学ぶメリット

一般的な R の学習法は、web サイトの解説や解説本を用いる方法が主流である。この場合、解説する側は、R の操作法を知っており、プログラミングの知識はあるが、分析したい課題を有していないことが多い。そのため、R の使い方を教えるのが主目的となっている。一方、学ぶ側は、分析しなければならない課題に直面しており、R の使い方を覚えるのが目的ではない。このような関係性においては、教える側が、現実味のない例題を用いて、知っていることを列挙するだけとなり、学んでも達成感が得られず挫折してしまう可能性が高い。

当県では行政が保有するデータの利活用に積極的に取り組んでおり[8]、我々は以前から大学と行政で受託研究等を通じてデータ分析に関して議論を続けている。多くの場合、行政の担当者側が解決すべき課題に直面し、関連するデータを持っており、大学側が分析手法を知っているという関係であるが、単に行政側から大学が分析を請け負うという形式ではなく、図３に示すように、データを直接見ながら両者がディスカッションして分析を進めていくことが成果につながりやすいと思われる。

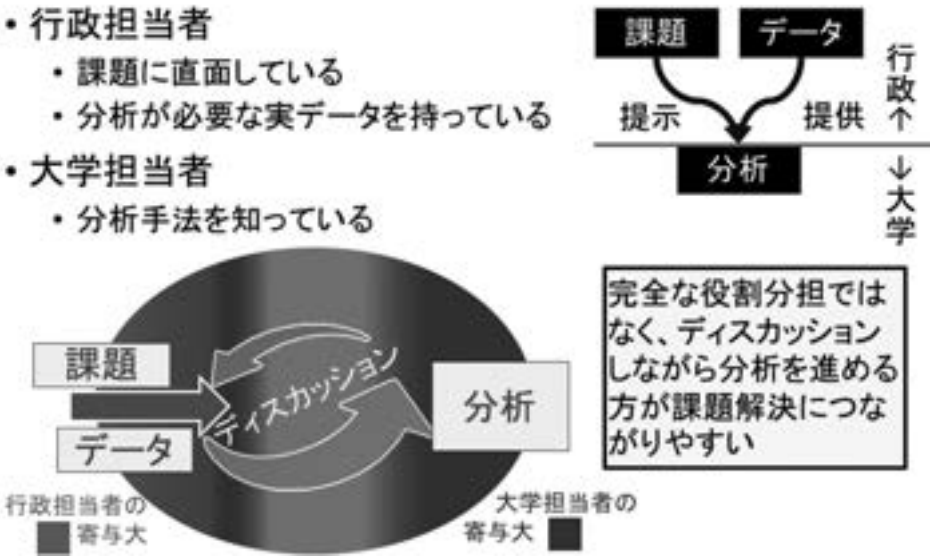

図３　データ分析における大学と行政との関係

本システムで取り扱うデータについて（個人情報を含むデータの取り扱いに関する展望も含めて）

行政が保有するデータには個人情報を含むものがあり、本システムでこれらのデータ分析を行うことも技術的には可能であるが、現段階では、データ分析のスキル向上を当面の目標と考えており、個人情報を含まないデータを用いて運用している。

今後、個人情報を含むデータ分析に適応範囲を広げていく必要性は生じると思われるが、安易に適応範囲を広げることは想定しておらず、県の情報セキュリティポリシー等に照らして、データ取扱いについての判断を行うことになると考えている。

また、個人情報を含むデータの共同分析を行う際にシステムの利用者がスクリプトに個人識別符号を記述する可能性がある。個人識別符号と上を入力させないという制限をシステムが事前にかけることは不可能であるため、本システムでスクリプトを実行する前に、個人情報が含まれていないかどうかをチ

ェックする仕組みの導入を検討する必要がある。さらに、並行して行う利用者向けの講義の中で、個人情報保護について反復して注意を喚起するとともに、どのような仕組みを使っても完全にその可能性を除外することはできないことも利用者に周知徹底すべきと考えている。

文献

1) 森川 正之：「エビデンスに基づく政策形成」に関するエビデンス:RIETI Policy Discussion Paper Series 17-P-008（2017 年３月）.

2) 佐藤　靖、松尾敬子、有本建男：科学的助言の概念の構造 JAIST 年次学術大会講演要旨集 31:466-469, 2016.

3) 岩崎　学：データマイニングと知識発見―統計学の視点から―　行動計量 26(1):46-58, 1999.

4) Data Scientist: The Sexiest Job of the 21at Century. Harvard Business Review, pp 70-76, October 2012.

5) The Data Science Education Dilemma. Technology Innovations in Statistics Education, 7(2):1-9, 2013.

6) 中澤　港：R による統計解析の基礎 (Computer in Education and Research). ピアソンエデュケーション, 2003.

7) R. A. Muenchen: The Popularity of Data Analysis Software. r4stats.com, 2013.（http://r4stats.com/articles/popularity/ 2018-09-25 確認）

8) 楠田　詞也：「佐賀県における「データ分析に基づく政策立案手法の導入」（データ利活用プロジェクト）の推進」、ESTRELA 特集・地方公共団体における統計利活用の取り組み：ESTRELA 2017 年 2 月（No.275）.