政策評価論における機械学習手法の応用

―潜在ディリクレ配分モデルを用いた行政事業レビューの組織間比較―

三上　真嗣

概　要

本稿では、機械学習手法の潜在ディリクレ配分モデルを用いて行政事業レビューにおける組織（府省庁）間の差異を量的に可視化する。その目的は、行動行政学の新たな潮流への対応と政策評価論研究の人的リソース不足への対応の 2 点を背景に、政策評価論における機械学習手法の応用例を示し、その有用性と限界を考える点にある。

ここで得られた結果は、本文中に図示した通り、府省庁によって異なるパターンで説明を行っていたというものであった。実務家や定性研究者には薄々気づかれていた事実が、「勘や経験」を量的に表現できた点が重要である。この意味で、政策評価論における機械学習手法はある程度で有用であると見通しがたつ。

機械学習は「規則や構造の抽出」、「パターンの認識」に長けており、因果推論とは異なるアプローチの道を示している。大量のデータからパターンが認識できれば、これまで処理しきれなかった情報を把握できる。これは、質的研究や歴史研究、比較研究とも相性がよいだろう。同様に行動研究や量的研究との接続を考える上でも有用となる。

ただし、利用可能な情報の種類には限界があるほか、利用者が動作原理をある程度理解しつつ、対象の性質や背景に専門性をもつ必要があるという限界もある。従来手法との上手な連携ができれば、長期にわたる丁寧な研究を主としつつ、補助として即応性と網羅性を重視したデータ分析を行う、そうした「二刀流」も可能となる。

1．はじめに

本稿では、機械学習（machine learning）手法を用いて行政事業レビュー[1]における組織（府省庁）間の差異を量的に可視化する。数ある手法のうち一例として、潜在ディリクレ配分モデル（Latent Dirichlet Allocation Model、以下 LDA）を用いる。ただし、データサイエンスの手法に接近するものの、その意図は帰無仮説検定による因果推論にはない（同様に法則性を示すことは主眼にない）。本稿の目的は、実務志向な政策評価論における機械学習手法の応用例を提示し、質的研究や歴史研究、比較研究を促進する際の有用性と限界を考える点にある。

実務志向な政策評価論においては、量的な研究はそれほどメジャーではない[2]。個別の評価が同質･均質ではない以上、評価の「制度運用」（南島 2020）など個別の具体的なメカニズムを知りたい場合には、「平均因果効果」があまり有用ではないのが 1 つの理由である。この事実に首肯しながらも、本稿があえて量的研究を試みる理由には次の 2 点がある。

第 1 に、近年の行政学や政策学に関する国

[1] 行政事業レビューの理論的な整理は、南島（2011）を参照。

[2] 政策評価論研究者にとっては、各種の実験手法や量的手法は慣れ親しんだものである。たとえば、Shadish, Cook, and Campbell（2002）は実験手法や準実験手法を議論する上での基礎であり、インパクト評価を理解する場合には計量経済学の手法や方法論にも触れる。ただし、政策評価の実務で計量経済学的な手法が用いられる事例は多くない。なお、計量経済学的な「効果検証」手法は、直近では安井（2020）や Cunningham（2021）を参照。

際的な動向に追随する必要がある。経済学や政治学と同様に、実験手法や統計的手法を応用した実証的な研究が広がりつつある。たとえば、心理学[3]の知見や手法を行政学に応用しようと試みる行動行政学（Behavioral Public Administration）[4]を志向する研究（日本在住の行政学者では野田遊や篠原舟吾らが一例、野田（2020）やJilke et. al.（2018）などを参照）や、経済学（計量経済学、実験経済学、行動経済学など）の理論や手法に接近する研究が進み出している。再現性ある「科学」を志向したり、ヒューリスティックな研究を深めるためには重要な取り組みである。

この研究動向のなか、政策評価論と密接である行政責任論やアカウンタビリティ[5]の研究領域にも変化が生じつつある。オランダ・イギリス・スウェーデン・デンマーク・スイス・オーストラリアの大学間では、実証的な研究との連携によって国際共同研究“Calibrating Accountability Project”が始まっているが、これはその一例である。ただし、彼らの意図は、因果関係の実証的な推論に限られず、プロジェクト名が示すように「アカウンタビリティの較正や調整」にあると考えられる。行動行政学の文脈では、実験手法の応用（Schillemans 2016）や社会心理学理論の応用（Aleksovska 2021）など一層実証性を強調した研究も増えつつある[6]。もっとも、アカウンタビリティ研究においては、心理学（あるいは、行動科学）に接近する議論は、Lerner and Tetlock（1999）のように特段目新しいものでもない[7]。

この国際共同研究における理論的な支柱の1つとして、当初は人的資源管理や心理学研究のもとで議論されていた‘Felt Accountability’[8]が再注目されている（Flink and Klimoski 1998；Hall, Frink, and Buckley 2017）。すなわち、アカウンタビリティを確保するための説明を受けた者（情報の「受け手」）がいかに「感じるか」を軸に議論を展開する。この「受け手」の反応（すなわち、行動）を心理尺度で計測する研究（Overman, Schillemans, and Grimmelikhuijsen 2021）も登場し、国際比較研究に発展している（たとえば、Overman et al.（2018））。こうした温故知新を図る議論は、比較政策評価論など新たな展開を考える手がかりになる。

第2に、日本国内における政策評価論研究者の人手不足に対処する必要がある。政府が公表する大量の報告書は、行政学や政策評価論に精通した研究者の（アクティブな）人的リソースの許容量を超えつつある。それを裏付けるように政府の報告書の多くが実務的にも学術的にも手つかずのままである。このため、日本における政策評価論（および、政策実施論や政策形成・決定論の評価が関連する領域）では研究対象の変化に十分に追いつけておらず、政策評価の比較研究である比較政策評価（Comparative Policy Evaluation、Furubo, Rist, and Sandahl（eds.）（2002）が国際的な代表例）の進展も遅れている。さらに政府のデジタル化や「アジャイル型政策形成・評価」の推進も相まって、今後は一層大量のデータがより高頻度に公表されるだろう。実態把握が追いつかない「空白領域」が広がれば、研究や実務改善にも悪影響が及び、政府活動が再びブラックボックスに閉ざされることとなる。透明性はあっても民主主義を損ねてしまう事態である。

政策評価論が直面するこれらの課題に取り組むうえで、「機械学習手法が政策評価論に有用ではないか」とここでは考える。この考えに筋道を付けるのが、行政学（および関連する政策学）の方法論に関する篠原らの議論である。篠原・小林・白取（2021：160）は、行政学の新たな方法論を論じるなかで、新たなテクノロ

[3] 政策を念頭においた心理学者による統計手法の解説には久保真人編（2016）がある。
[4] Jilke, Olsen, Tummers, en Grimmelikhuijsen（2016：10）は、行動行政学（gedraagsbestuurskunde）を「心理学と行政学（bestuurskunde）の知見に基づいた、個人やグループの行動（gedrag）や態度（houding）といった（ミクロ）視点からの行政学的テーマ（onderwerpen）の分析」と定義する（括弧内はオランダ語）。なお、英語論文のGrimmelikhuijsen, Jilke, Olsen, and Tummers（2017：46）がしばしば引用されるが、オランダ語論文が先行しており、英語版の定義と若干の差異がある。
[5] 政策評価との理論的関係は山谷（1997）を参照。アカウンタビリティの国際的な定義や議論は、Bovens（2007）、Bovens en Schillemans（2009：2）、Bovens, Goodin, and Schillemans（2014）に詳しい。
[6] 批判を加えるとすれば、Aleksovskaらの行政責任論研究は現実の制度運用や組織メカニズムの複雑さを理論的に単純化しすぎている。
[7] たとえば、Aleksovska, Schillemans, and Grimmelikhuijsen（2019）は、1970年代以降のアカウンタビリティに関する一連の研究をレビューしている。
[8] オランダ語圏では‘gevoelde verantwoording’として心理学の知見が使われている（Schillemans, Bokhorst, van Genugte, en Vrielink 2018：62）。

ジーの発展により質的研究と量的研究の境界がなくなり、融合が進む可能性を示唆している。国際的には機械学習の応用研究が始まりつつある（Anastasopoulos and Whitford 2019）が、日本の行政学や政策評価論ではほとんど取り組まれてこなかった。そこで本稿では、機械学習手法による政府報告書の実態把握の一例として、行政事業レビューにおける説明「行動」パターン（Reporting 活動）に迫ってみたい。

2. 方法論とモデル

2.1　生成モデル

社会科学の量的研究とくに因果推論の研究では、各種の統計検定や（重）回帰分析に基づいて帰無仮説の棄却をもとに「科学」的判定を行う場合が多い。だが、この狭義の「科学」に縛られない、より柔軟な現状把握に徹する道も考慮したい。

とくに、近年の統計学やこれに関係する領域では *p* 値に「過度に」依拠した手法と方法論が問題視されるようになっており、背景事情に注意を払う必要がある。たとえば、心理学においても「再現性の危機」問題が指摘されて論争と改革が生じている。『心理学評論』では「心理学の再現可能性」（2016 年、59 巻 1 号）、「統計革命」（2018 年、61 巻 1 号）、「心理学研究の新しいかたち」（2019 年、62 巻 3 号）と 3 回の特集が組まれてきた。友永・三浦・針生（2016）によれば、過去の心理学研究で再現性が確認されたものが 40% に満たないという結果が 2015 年に明らかになり、*Basic and Applied Social Psychology* 誌が帰無仮説検定の記載を廃止するなどの改革が進んだという。認知心理学者の大久保街亜は、社会心理学における研究を振り返り、帰無仮説検定への過度な依存が再現性を低下させる一因になっていると述べる（大久保 2016）。こうした騒動を受け、アメリカ統計学会は 2016 年に *p* 値についての声明を発表するに至った（Wasserstein and Lazar 2016）[9]。

心理学以外の統計学に携わる研究者たちからも *p*-hacking や HARKing（Hypothesizing After the Results are Known）といった問題が指摘されており、ネットワーク科学と生態学の研究者である阿部真人は、最先端の知見を総合して統計学を解説する際にこの問題を丁寧に扱っている（阿部 2022：232，254-60）。たとえば、*p* 値が有意水準を下回るまでサンプルサイズを追加したり、アンケート調査の回答を集めた後にすべての結果を採用せず、特定の結果のみを選択的に用い、*p* 値が有意になるまでモデルの統計検定を試し続けたりする研究手法は問題とされる。

こうした専門家たちの主張を考慮しつつ、政府報告書の実態把握をする道を考える必要がある。それは、データサイエンス的に述べれば、政府報告書のデータ生成過程[10] を明らかにするアプローチである。データの生成に関するモデルは、生成モデル（generative model）とよばれ（岩田 2015：19）、その作成には「すべての観測データに対する背後の生成過程を記述する」（須山 2017：33）ことが試みられる。モデルにデータを合わせるのではなく、現実データをもとにモデルを表現する。これがもう一つのアプローチとなる。

2.2　機械学習と階層ベイズモデル

政府が公表する膨大で複雑な実務データをもとに生成モデルを表現できれば、政策評価の実務と研究に示唆を与えることが可能になる。これは研究の目的に応じて柔軟にモデル観を変えることでもある。モデルについて、近年の統計学やデータサイエンスの視点では、①数理モデル・②統計モデル・③機械学習モデルの 3 分類[11] が

[9] 教育心理学者の岡田謙介は、この声明で提示された p 値の代替が、ベイズ統計のアプローチと一部重なると述べている（岡田 2017：90）。また、社会心理学者の清水裕士は、心理学におけるベイズ統計モデリングの有用性について詳しく説明し、帰無仮説検定に対する批判や従来では解けなかった問題に対する解決策として注目を集めつつあると述べる（清水 2018）。

[10] ただし、定性的な「過程」研究とは異なる意味合いである。馬場（2019）は、確率的な表現で作られたモデルを「確率モデル」とよび、この確率モデルにデータを適合したものを「統計モデル」と表現して区別する。さらに確率モデリングやベイズ推定によって、予測結果のあやふやさを織り込んでノイズや情報不足による曖昧さを確率の「雲」として表現できると考えられている（持橋・大羽 2019）。

[11] 細野（2021：39）は、公共政策を考える際のモデルを 3 つに区別している。すなわち、具象モデル（具体的な対象を縮尺した模型）、数理モデル、数値計算モデル（シミュレーションによるモデル分析）である。だが、技術発展によってこれらの境界線が曖昧になりつつある。

提示されており、次のように重視する目的が異なると考えられている（阿部 2021：329）。すなわち、既知のメカニズムから演繹的に導かれる①数理モデルでは、パラメータを変化させて知見を得ることができる。他方、帰納的に（少数の）データからつくる②統計モデルは、比較的単純な構造であり、解釈に向いている[12]。この 2 つに対して、大量のデータからつくる③機械学習モデルは、予測に特化している。ただし、複雑な構造のためにモデルの解釈は難しくなる。

この機械学習とは、「データに潜む規則や構造を抽出することにより，未知の現象に対する予測やそれに基づく判断を行うための計算技術の総称である」と考えられており（須山 2017：2）、計算機科学を起源とする機械学習は、工学を起源とするパターン認識（pattern recognition）と同じ分野を 2 つの側面から見たものであるという（Bishop 2006：vii=2015 上巻：iv）。ここから機械学習のタスクが「規則や構造の抽出」や「パターン認識」にあるとわかる[13]。

本稿が試みる機械学習手法は、Blei et al.（2013）が提唱したトピックモデル（topic model）の LDA である[14]。トピックモデルは、潜在的なトピックのパターンを抽出する③機械学習モデルである。テキスト・マイニングに用いられる場合も多いのだが、テキストだけを想定したモデルではない。Blei らの原論文に明示されているように LDA は 3 層の階層ベイズモデル（Hierarchical Bayesian Model）であり（Blei et al. 2013：997）、②統計モデルとしての側面もある。

本稿の読者は定量研究者だけではないので、この階層ベイズモデルの位置づけを生態学者の久保拓弥の議論に従って整理しておきたい。久保（2012）は、一般線形モデル、一般化線形モデル（Generalized Linear Model、以下 GLM）、一般化線形混合モデル（Generalized Linear Mixed Model、以下 GLMM）、階層ベイズモデルを紹介し、統計モデルの複雑さの段階を整理している。すなわち、第 1 段階の一般線形モデルを使う場合には等分散性や正規性といった前提を満たす必要があり[15]、厳密な実験や上手な研究デザインを採用する必要がある。残差[16]が最小になるように最小二乗法（Ordinary Least Square）を使ってモデル推定する。次に、残差が正規分布に従うと仮定できない場合には、第 2 段階の二項分布やポアソン分布などを採用した GLM が用いられる。サンプル・データでパラメータが得られる確率を最大化する最尤推定法（Maximum Likelihood Estimate Method）[17]がモデル推定に使われる。モデルに個体差・場所差といった変量効果（random effects）を想定する場合、第 3 段階の GLMM が採用される。ここに時間差や空間差も取り込めば一層複雑なモデルを作る必要が生じる。GLMM のパラメータをただ 1 つに定めるのではなく、確率分布（とその階層化）として扱うとき、第 4 段階の階層ベイズモデルが登場する。変量効果の数が多くなると、パラメータ間の組み合わせ数が非常に多くなったり、積分消去の回数が増えたりして、現実的な時間で解析が不可能となる。そのため近似（ベイズ推定やマルコフ連鎖モンテカルロ法（Markov Chain Monte Carlo method、以下 MCMC）などを利用）が行われる。以上が久保の整理である。

この近似にはベイズ統計の知見が応用され、ベイズ統計モデリングとも呼ばれる[18]。これは、機械学習モデルの一部とも無関係ではない。確率的なモデリングと確率推論を用いたアプローチをベイズ学習（Bayesian machine learning）といい、観測したデータをもとにモデルを更新す

[12] 政治学の因果推論では、②統計モデルに焦点を当てて議論することが多いだろう。
[13] 細野はパターン認識の難しさを次のように述べる。「人間はパターン化して物事を見た上で理解することにはたけているが，パターン化が難しい物事にはてこずる場合が多い。そして人間の認識能力の限界もある」（細野 2021：38）。
[14] なお、Transformer モデルが公表されて以来、大規模言語モデル（Large Language Models、以下 LLM）が相次いで登場している。LLM やニューラルネットワークを使った分析も考えられたが、ブラックボックスな作動原理が多い点、発展段階のため技術の変化が激しい点から本稿では採用を見送った。ただし、行政実務でも模索が進みつつあるので、深層学習技術等の応用については今後の検討課題となる。
[15] この条件を満たすために事前に正規性や等分散性の検定を行う場合がある。このプロセスがない行政学や政策学の研究もあるが、検定多重性の問題を回避するためのおそらく自覚的な試みである。
[16] モデルの予測値と観測データの差。誤差（error）ではない。
[17] 尤度（likelihood）とは、ある観測データにおいてモデルや分布のパラメータの尤もらしさ。対数尤度は対数をとった尤度。
[18] なお、近似計算には Stan（Carpenter et al. 2017）のような確率的プログラミング言語が便利である。GLM、GLMM、階層ベイズモデルの分析は、久保拓弥（2012）、松浦（2016）、馬場（2019）に詳しい。

る（須山 2017：29-30）。すなわち、①観測データ $\mathcal{D}$ と観測されていない未知の変数 X について同時分布[19] $p(\mathcal{D}, \mathrm{X})$ を計算し、②この同時分布をもとに、事後分布 $p(\mathrm{X}|\mathcal{D}) = \frac{p(\mathcal{D}, \mathrm{X})}{p(\mathcal{D})}$ を解析的または近似的に求める。このモデル更新においてベイズの定理[20] が使われている。

2.3　トピックモデル

文章に応用するトピックモデルは、文書中に現れる単語の多重集合[21]（Bag of Words、以下 BoW とよぶ）から特定の文書集合を生成する確率モデルである（岩田 2015：19）。当該領域に対するリスペクトを込めて、Blei et al.（2013）の議論でも扱われたユニグラムモデル、混合ユニグラムモデル、LDA と順に整理し、なるべく丁寧に理解を深めたい。ただし、厳密な数学的議論が本稿の目的ではない点に注意されたい。

確率変数間の関係を視覚的に表す際には、グラフィカルモデルが用いられる場合が多い（Bishop（2006：359-418）やWainwright and Jordan（2008）、渡辺（2008）が参考になる。）。図中の矢印「A → B」はチルダ「A ～ B」と同様に確率変数が確率分布に従うことを意味する。また、背景色がある円は観測された変数、白地の円は未知の変数、四角は繰り返しを表し、右下の数だけ繰り返す。3 つの生成過程をグラフィカルモデルで表現すれば、図 1 の通りである。

数式や記号については、岩田（2015）の表記法を軸として、佐藤（2015：5-7）、須山（2017：vii-viii）に従って表記する。まず、文書の総数を D とし、すべての文書を $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_D)$ と表す。この文書集合において添数が d である要素 $\mathbf{w}_d = (w_{d,1}, w_{d,2}, \ldots, w_{d,N_d})$ は、文書 d を BoW で表現した単語集合を表している。この単語集合における添数を n としたとき、$w_{d,n}$ は $\mathbf{w}_d$ に

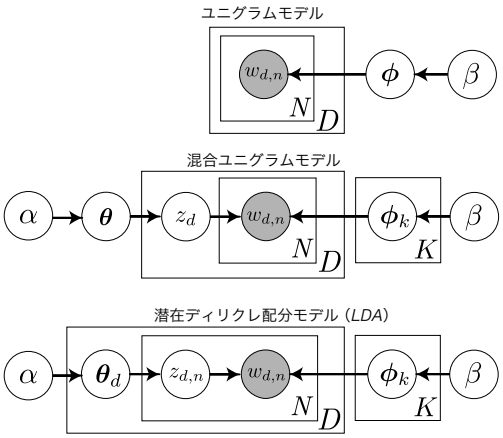

図 1　代表的なトピックモデル（グラフィカルモデル）

出典：Blei et al.（2013）と岩田（2015：58）を参考に三上作成。

おける n 番目の単語を表す。さらに総語彙数（単語の種類）を V として、添数 v を用いて特定の単語を表現する。文書全体の総単語数を N として、文書全体における単語 v の出現回数を N_v、文書 d における単語 v の出現回数を $N_{d,v}$、文書 d に含まれる単語数を N_d と表記する。

岩田（2015：19-20）によれば、第 1 のユニグラムモデル（Unigram Model）は、文書を構成するすべての単語がある単一のカテゴリ分布[22] に従って生成される。すなわち、次のように確率分布に従う[23]。

$$w_{d,n} \sim \mathrm{Categorical}(\boldsymbol{\phi}) \quad \text{for } n = 1, \ldots, N_d, \text{ for } d = 1, \ldots, D \text{（単語の生成）}$$

ここで、$\boldsymbol{\phi} = (\phi_1, \ldots, \phi_V)$ であり、ϕ_v は単語 v が生成される確率を表している。ϕ_v は確率であるので、$\phi_v \geq 0, \sum_{v=1}^{V} \phi_v = 1$ である。たとえば、

---

[19] 結合分布とも呼ばれる。

[20] ベイズの定理は、ある 2 つの確率 x, y について条件付き分布の式 $p(x|y) = \frac{p(x,y)}{p(y)}$ と同時分布である $p(x, y)$ の周辺化の式である $p(y) = \int p(x, y)\, dx$ から導くことができる（須山 2017：14-5）。

[21] 重複有りの集合。

[22] カテゴリ分布は、2 次元のベルヌーイ分布をより一般な K 次元に拡張した離散確率分布であり、次式で定義される（岩田 2015：136-7；須山 2017：52）。

$$\mathrm{Categorical}(\mathbf{s}|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{s_k}$$

ここで、s は K 次元のベクトルであり、各要素を 1 から k までの添数を用いて示すとき、s_k は 0 か 1 をとり、その総和が 1 である。π はそれぞれの発生確率を表す。

[23]「～」（チルダ記号）は確率変数が確率分布に従うことを意味する。なお、繰り返しを指して「for $k = 1, \ldots, K$」のように付す（須山 2017 に準じた記法）。2 つ以上連続する場合は、末尾に近いものから多重に繰り返す。

$\phi_{外交}$ = 0.1 であれば、文章を生成した際に 10%の確率で「外交」が出現する。

グラフィカルモデルでは一定の規則で数式を組み立てられるので、確率分布との関係を考慮すれば、パラメータ ϕ が与えられたとき、ユニグラムモデルにおける文書集合 W の生成確率はカテゴリ分布に従って独立に生成される単語 $w_{d,n}$ の総乗で表される。

しかし、すべての文書が同じトピックを有しているとは行政事業レビューのような現実のケースでは考えにくい。そこで第 2 の混合ユニグラムモデル（mixture of unigram models）では、それぞれの文書に 1 つずつトピックがあると仮定する（岩田 2015：33）。前述の記法に加え、総トピック数を K とし、あるトピックを添数 k で表す。そこでは、トピック k の単語分布は $\boldsymbol{\phi}_k = (\phi_{k,1}, \cdots, \phi_{k,V})$ と表現され、$\phi_{k,v} = p(v|\boldsymbol{\phi}_k)$ はトピック k において単語 v が生成される確率を表している（ただし、$\phi_{k,v} \geq 0, \sum_{v=1}^{V} \phi_{k,v} = 1$）。

また、それぞれの文書はただ 1 つのトピック $z_d \in \{1, \ldots, K\}$ を有すると考え、トピック z_d ごとの単語分布 $\boldsymbol{\phi}_{z_d}$ に従って文書内の単語が生成されると仮定する。これは、$w_{d,n} \sim$ Categorical $(\boldsymbol{\phi}_{z_d})$ と表現される。なお、すべての文書で各トピックがとる確率分布をトピック分布 $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$ で表す。$\boldsymbol{\theta}_k$ は文書にトピック k が割り当てられる確率であり、$\boldsymbol{\theta}_k = p(k \mid \boldsymbol{\theta})$ である（ただし、$\theta_k \geq 0, \sum_{k=1}^{K} \theta_k = 1$）。

混合ユニグラムモデルでは、まず超パラメータ[24] の α と β を用いて、文書全体のトピック分布 $\boldsymbol{\theta}$ と文書全体の単語分布 $\boldsymbol{\Phi} = (\phi_1, \ldots, \phi_k)$ を推定する必要がある。この $\boldsymbol{\theta}$ と $\boldsymbol{\Phi}$ の推定には、多項分布やカテゴリ分布の共役事前分布[25] であるディリクレ分布（Dirichlet distribution）[26] が利用される。すなわち、次の階層で表現される。

$$\boldsymbol{\theta} \sim \text{Dirichlet}(\alpha)\ （トピック分布\boldsymbol{\theta}を生成）$$
$$\boldsymbol{\phi}_k \sim \text{Dirichlet}(\beta)\ \text{for } k = 1, \ldots, K\ （単語分布\boldsymbol{\Phi}を生成）$$
$$z_d \sim \text{Categorical}(\boldsymbol{\theta}_d)\ \ \text{for } d = 1, \ldots, D\ （トピック集合\mathbf{Z}を生成）$$
$$w_{d,n} \sim \text{Categorical}(\boldsymbol{\phi}_{z_d})\ \ \text{for } n = 1, \ldots, N_d \text{ for } d = 1, \ldots, D\ （単語集合\mathbf{W}を生成）$$

第 3 の LDA においては、1 つの文書が複数のトピックをもち、文書ごとにトピック分布 $\boldsymbol{\theta}_d = (\theta_{d,1}, \ldots, \theta_{d,k})$ があると考える。$\theta_{d,k}$ は文書 d を構成する単語にトピック k が割り当てられる確率である（ただし、$\theta_{d,k} \geq 0, \sum_{k=1}^{K} \theta_{d,k} = 1$）。文書 d の単語 n に 1 つ 1 つトピック $z_{d,n}$ が割り振られる。このとき LDA のモデルは次の階層で表現され、たとえば、$\boldsymbol{\theta}_{分析対象の文書}$ によって複数のトピックがどの程度の確率で生じるかを把握できる。

$$\boldsymbol{\phi}_k \sim \text{Dirichlet}(\beta)\ \text{for } k = 1, \ldots, K\ （単語分布\boldsymbol{\Phi}を生成）$$
$$\boldsymbol{\theta}_d \sim \text{Dirichlet}(\alpha)\ \text{for } d = 1, \ldots, D\ （トピック分布\boldsymbol{\Theta}を生成）$$
$$z_{d,n} \sim \text{Categorical}(\boldsymbol{\theta}_d)\ \text{for}\ \ n = 1, \ldots, N_d\ \ \text{for } d = 1, \ldots, D\ （トピック集合\mathbf{Z}を生成）$$
$$w_{d,n} \sim \text{Categorical}\left(\boldsymbol{\phi}_{z_{d,n}}\right)\ \ \text{for } n = 1, \ldots, N_d \text{ for } d = 1, \ldots, D\ （単語集合\mathbf{W}を生成）$$

[24] ハイパーパラメタ。確率分布であるパラメータを定めるためのパラメータ。
[25] 共役事前分布は、事後分布の分布が事前分布と同じ分布になる分布である。ベイズ推定の計算負荷を減らすことができる。
[26] ディリクレ分布は、ベータ分布を K 次元に拡張した連続確率分布である。K 次元ベクトル π を生成する確率分布であり、π はそれぞれの要素が 0 から 1 の間で値を取り、総和が 1 となる。（岩田 2015：143-4；須山 2017：58-9）。

$$\text{Dirichlet}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_K\right)}{\prod_{k=1}^{K} \Gamma(\alpha_K)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

ここで、α の要素 α_k は正の実数値であり、Γ はガンマ関数。

なお、トピックの全体集合は $\mathbf{Z} = (z_{1,1}, \ldots, z_{1,N_1}, z_{2,1}, \ldots, z_{D,N_D})$、単語分布の全体集合は $\mathbf{\Phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K)$、トピック分布の全体集合は $\mathbf{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)$ である。この同時分布は次式に表現され、これを式変形しながら分析を進めることができる（岩田 2015：62；須山 2017：193-5；佐藤 2015：75）。

$$p(\mathbf{W}, \mathbf{Z}, \mathbf{\Theta}, \mathbf{\Phi} \mid \alpha, \beta) = p(\mathbf{W} \mid \mathbf{Z}, \mathbf{\Phi})\, p(\mathbf{Z} \mid \mathbf{\Theta})\, p(\mathbf{\Theta} \mid \alpha)\, p(\mathbf{\Phi} \mid \beta)$$
$$= \left[\prod_{d=1}^{D}\prod_{n=1}^{N} p(w_{d,n} \mid z_{d,n}, \mathbf{\Phi})\right]\left[\prod_{d=1}^{D}\prod_{n=1}^{N} p(z_{d,n} \mid \boldsymbol{\theta}_d)\right]\left[\prod_{d=1}^{D} p(\boldsymbol{\theta}_d \mid \alpha)\right]\left[\prod_{k=1}^{K} p(\boldsymbol{\phi}_k \mid \beta)\right]$$

3. 分析対象

3.1　データと手法

本稿で利用するデータは、2022 年度の行政事業レビュー（URL1）のすべての評価シートである。行政事業レビューを扱う理由は次の 3 点すなわち、府省横断的に統一的な評価基準で整理されている点、政策評価制度のなかで影響力を増している点、それにもかかわらず全体的な実態把握が十分に進んでいない点である。

このうち必要性に関する説明「事業所管部局による点検・改善－国費投入の必要性－事業の目的は国民や社会のニーズを的確に反映しているか。－評価に関する説明」の項を分析対象とした（*D*=5394）。必要性評価は、評価階層の理論において最も基礎的な部分であり、表現の違いは府省庁間の違いを示す重要なポイントになるからである。必要性評価のレーティング分布は大きく偏っており、“〇”が 5340 件（約 98.9%）、無記入が 21 件（約 0.4%）、省略表記が 33 件（約 0.6%）であった。

対象となる府省庁は、行政事業レビューの区別に準じ、内閣官房（85 件）、内閣府（231 件）、個人情報保護委員会（7 件）、カジノ管理委員会（2 件）、公正取引委員会（9 件）、警察庁（90 件）、金融庁（39 件）、消費者庁（38 件）、デジタル庁（39 件）、復興庁（147 件）、総務省（204 件）、法務省（73 件）、外務省（400 件）、財務省（65 件）、文部科学省（508 件）、厚生労働省（1180 件）、農林水産省（395 件）、経済産業省（458 件）、国土交通省（640 件）、環境省（335 件）、原子力規制委員会（52 件）、防衛省（357 件）である。

分析環境には R（ver. 4.3.0）および RStudio（ver. 2023.06.0+421）を利用した。テキストを分かち書きし、単語文書行列に変換する際には、R の Quanteda（ver.3.3.1）を用いた[27]。数字とアルファベットは正規表現で半角・全角を揃え、URL 表記、ひらがな、数字、記号のみで構成される語、1 文字の語、5 回未満の低頻度語を除外し、単語文書行列を得た（語彙数 *V*=5652、総単語数 *N*=104638）。

LDA には複数の実装が提案されているが、幅広い分野で利用されている R の topicmodels（ver.0.2-14）（Grüen and Hornik 2011）を採用した。なお、LDA は階層ベイズモデルとして Stan でも実装可能であり、モデルの事前確認に Stan を用いた[28]。LDA の推定手法には、原論文である Blei et al.（2013）の実装に準じ、変分ベイズ法[29] を採用した。変分ベイズ法（変分推論）では、推定される確率分布を真の確率分布に近づけるべくモデル更新を繰り返す。その

[27] 日本語文はスペースで単語を分けられないので、計量政治学などでしばしば利用される Quanteda を用いて分かち書きを行った。内部処理では ICU（International Components for Unicode）が使われている。形態素解析手段には、MeCab や Juman、Juman++ なども有名である。

[28] 本稿ではパッケージによる分析の前に Stan を用いた分析を試みた。しかし、MCMC（NUTS 法）の結果、収束条件 Gelman-Rubin 推定量 $\hat{R}$<1.1 を満たしたものの、筆者の有する計算資源では一部の確率変数で有効サンプルサイズ $\hat{n}_{eff}$ を十分に満たすことができなかった。この収束条件は Gelman et al.（2014：285-7）を参照。モデルを拡張したり、他の回帰分析等と組み合わせたりする場合に有用であり、今後の課題とする。実際に数式を実装して理解を深める際にも役立つ。

[29] パッケージと原著論文の実装では若干の違いがある。情報科学の研究者にとっては重要な差異であっても、政策評価論の議論にそれほど大きな影響を与えるわけではない。詳細については参考文献を比較参照されたい。

際、両者の「近さ」を最小化できればよい。大まかに説明すれば、この確率分布間の「近さ」は Kullback-Leibler Divergence[30]（以下 KL ダイバージェンス）によって把握できる（Kullback and Leibler 1951；佐藤 2015：41）。この値の最小化[31] を図り近似する。ここでは、1 文書あたり 500 回を上限とした近似計算を行った。

3.2　モデル選択

モデル選択において、トピック数 K の選択にはいくつかの指標[32] が用いられる。佐藤（2015：135-6）は、予測性能を示す Perplexity を用いたモデル選択を提示しており、値が低いものが好ましいとされる。Perplexity は次式で定義されており（Blei et al. 2013：1008）、算出には訓練済みのモデルに対してテストデータを投入する。

$$\mathrm{Perplexity}(D) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\right\}$$

さらにトピック品質を測るための指標 Coherence を採用する。Mimno et al.（2011：265）は、$D(v)$ を単語 v が出現する文書数、$D(v_1, v_2)$ を単語 v_1 と v_2 が共起する文書数としたとき、トピック k の Coherence を次式で定義している（複数の実装があるが、これは UMass 方式とよばれる。）。

$$\mathrm{Coherence}\left(V^{(k)}\right) = \sum_{m=2}^{M}\sum_{l=1}^{m-1}\log\frac{D\left(v_m^{(k)}, v_l^{(k)}\right)+1}{D\left(v_l^{(k)}\right)}$$

ここで $V^{(k)} = \left(v_i^{(k)}, \ldots, v_M^{(k)}\right)$ は、トピック k における生成確率で上位 M 個の単語である。本稿では、M=10 として全データを対象に計算を進めた。佐藤（2015：131）に従い、各トピックの Coherence の平均を計算してモデル全体の指標を得た。値は低いものが好ましい。

ここでは簡便に[33]、分析に用いる全データを 9：1 にランダム分割し、約 90% の訓練データ（$D_{\text{訓練データ}}$=4819）と約 10% のテストデータ（$D_{\text{テストデータ}}$=535）を得た。次に、訓練データを用いてトピック数を 2 から 75 まで 1 ず

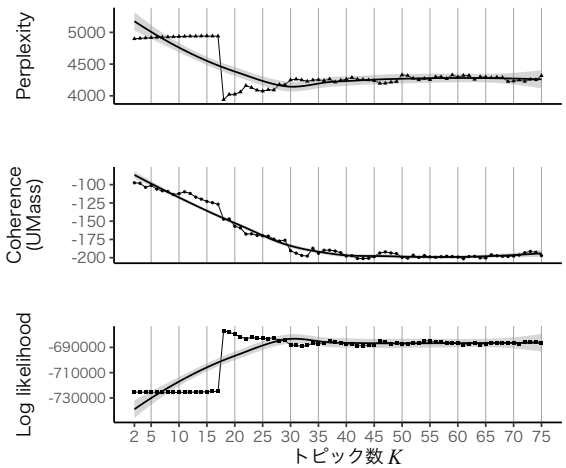

図 2　LDA のモデル選択指標

出典：三上作成。

[30] p と q が離散確率分布であるとき、KL ダイバージェンスは次式で定義される。

$$\mathrm{KL}(p,q) = \sum p(x)\log\frac{p(x)}{q(x)}$$

ここで、$p(x)$ と $q(x)$ はそれぞれ確率分布 p と q に従って選ばれた値が x である確率を指し、省略記法の Σ は全体の総和である。

[31] 凸関数の性質とイェンゼンの不等式によって変分下限の最大化問題に帰着する。ベイズ推定も含め、岩田（2015：61-6）に詳細な解説がある。

[32] これを「評価指標」というが、政策評価論や評価理論における概念と同一ではない。

[33] より正確を期する際には交差検証法（k-Fold Cross Validation、k=10 等）を用いる。

つ加算して LDA のモデルを作成した（K=2,3,…,75）。図 2 は、作成した 74 モデルをもとに、Perplexity と Coherence を平滑化曲線とともにプロットしたものである（図 2）。

図 2 によれば、Perplexity は K=18 で急落し、25 前後にかけて低い値を示す。平滑化した曲線をみれば K=30 の値を前後に収束する。他方、Coherence は K=2 が最大であり、トピック数の増加に対して減少を続けるため、判断基準から除外する。以上から K は 18 から 25 の範囲程度が妥当であると判断した。それぞれのモデルを実際のデータで作成し、最終的に結果の可読性をもとに K=25 を採用した。なお、モデルの統計学的な尤もらしさを示す対数尤度（log likelihood）は、K=18 以降は高い値を示す。

4. 結果と考察

行政事業レビューにおける必要性評価のパターンは図の通りである（図 3、図 4）。図 3 は全体の単語分布をトピックごとに示したものである（図 3）。生成確率上位 10 件の単語を表示した。この単語を参考に、「①ニーズの強調」から「㉕保険・共済」に至るラベルを質的な判断に基づいて付した。ラベル付けに際して、各トピックの該当確率上位 20 件ずつの行政事業レビューシートも確認した。

図 4 は府省庁間比較の結果である。トピックごとの生成確率を府省庁単位でプロットした（図 4）。ここでは、府省庁単位で抽出した各文書の生成確率について平均を計算した。府省庁の違いは変数として含まれておらず、記述の差異のみが反映されている。行政事業レビューの記述項目は評価シートの形式や組織間の申し合わせによって管理されているため、なるべく同じ視点で記述されているはずである。しかし、日頃から報告書に触れている研究者や実務家、あるいは外部有識者が薄々気づいていたとおり、組織間に差があったとわかる。この「勘や経験」を可視化したことになる。

トピックには個別の政策内容に関連したものも多い。たとえば、「②防衛・安全保障」のように防衛省にしかほとんど反応しないものもある。他方で内閣府、総務省、国土交通省のように全体的に同程度の比率でバランスよく分かれている組織もある。1 つのトピックが複数の組織に及ぶ場合もあり、たとえば「④外交・国際協力」は外務省のほか、財務省や公正取引委員会にも反応が見られる。財務省は有償資金協力を担っており、妥当な結果であろう。公正取引委員会は、発展途上国や他先進国を含めた国際的な競争政策を展開している。他方で、トピックには政策内容ではない行政管理・総務的な説明もあった（「⑩実施・評価」や「⑮計画の実施」、「⑳業務の遂行」、「㉒目的の遂行」）。これは施設や業務システムなどのインフラを整備、更新する場合に該当していた。また、ハンセン病対策事業や被爆者援護など法律や閣議決定、計画を理由として、その適切な実施を必要性の説明としていた。「①ニーズの強調」では「ニーズがあるため必要」とやや循環的な説明がなされ、ここには原子力規制委員会やカジノ管理委員会、消費者庁などが反応する。評価のパターンに応じて説明を試みるアカウンタビリティの種類が異なる以上、評価は十分に管理できていないことが示唆される。

なお、個別の文書ごとにも結果を見ることができる。たとえば、防衛省の事業がすべて「②防衛･安全保障」に依拠しているわけではなく、実際には事業特性に応じてばらつきがある。防衛省事業の一部をヒートマップで可視化した（図 5）ところ、装備品の開発、購入、修繕といった事業にはトピック②が強調されるが、「サイバー攻撃」関連事業には「⑬生活の安定」が強調されることがわかる。新たな政策を説明する際には、別のロジックが用いられていることを示唆する。

同様に、図 6 に外務省の分担金や拠出金の必要性評価の文書の一部を描出した（図 6）。微妙な違いを反映して異なる確率分布を示していると気づくだろう。たとえば、「④外交・国際協力」が強調される文書ばかりでもなく、「国際連合興業開発機関（UNIDO）分担金」のように「⑭科学技術・産業」が高い確率と推定されるものもある。また、同じ「経済協力開発機構・開発センター」でも拠出金と分担金で説明に差異が生じている点も興味深く、実務的には外務省や財務省が関心を寄せる点であろう。

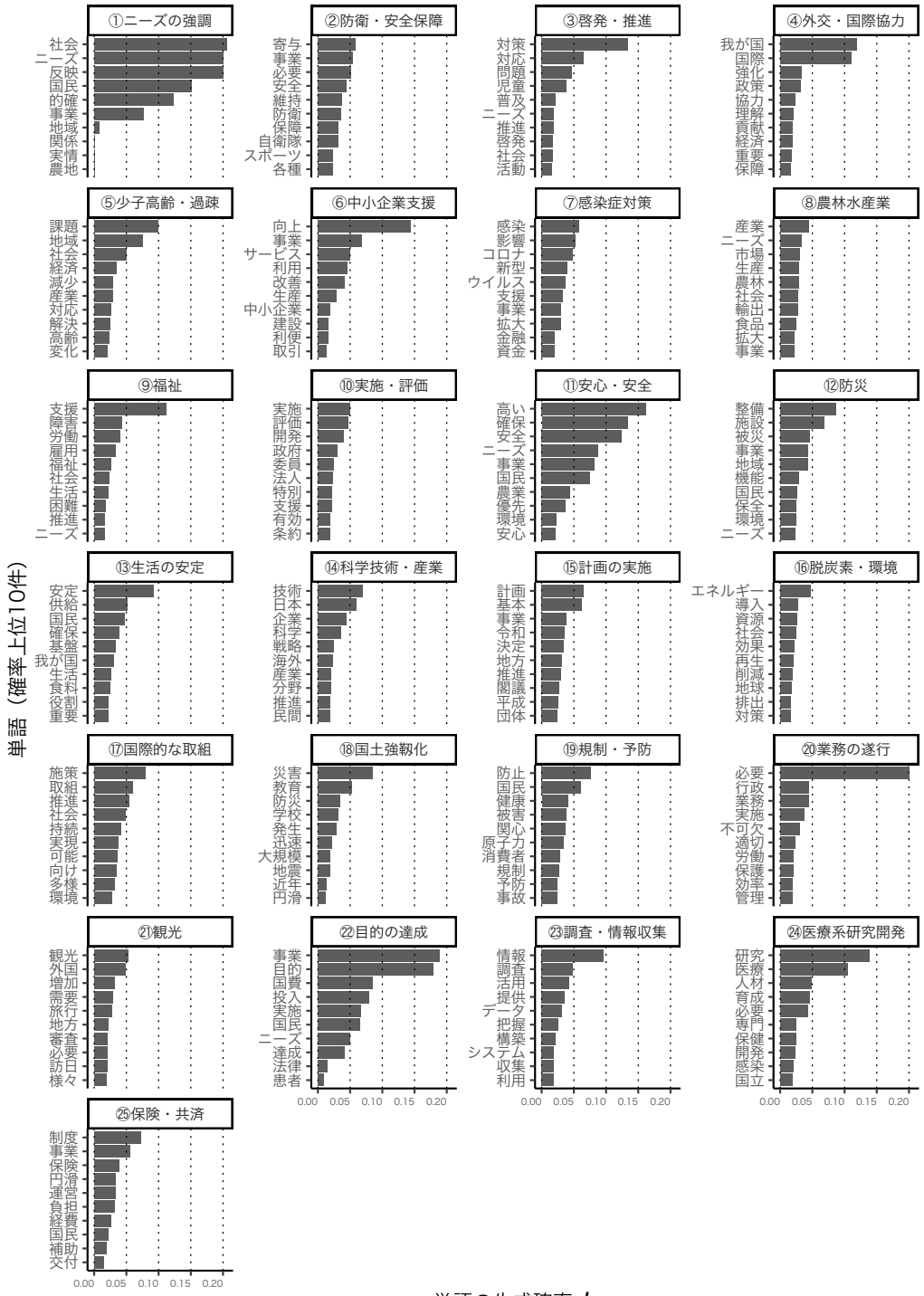

図３　全体の単語分布Φ（トピック間比較）

出典：分析結果をもとにトピック名称を付して三上作成。

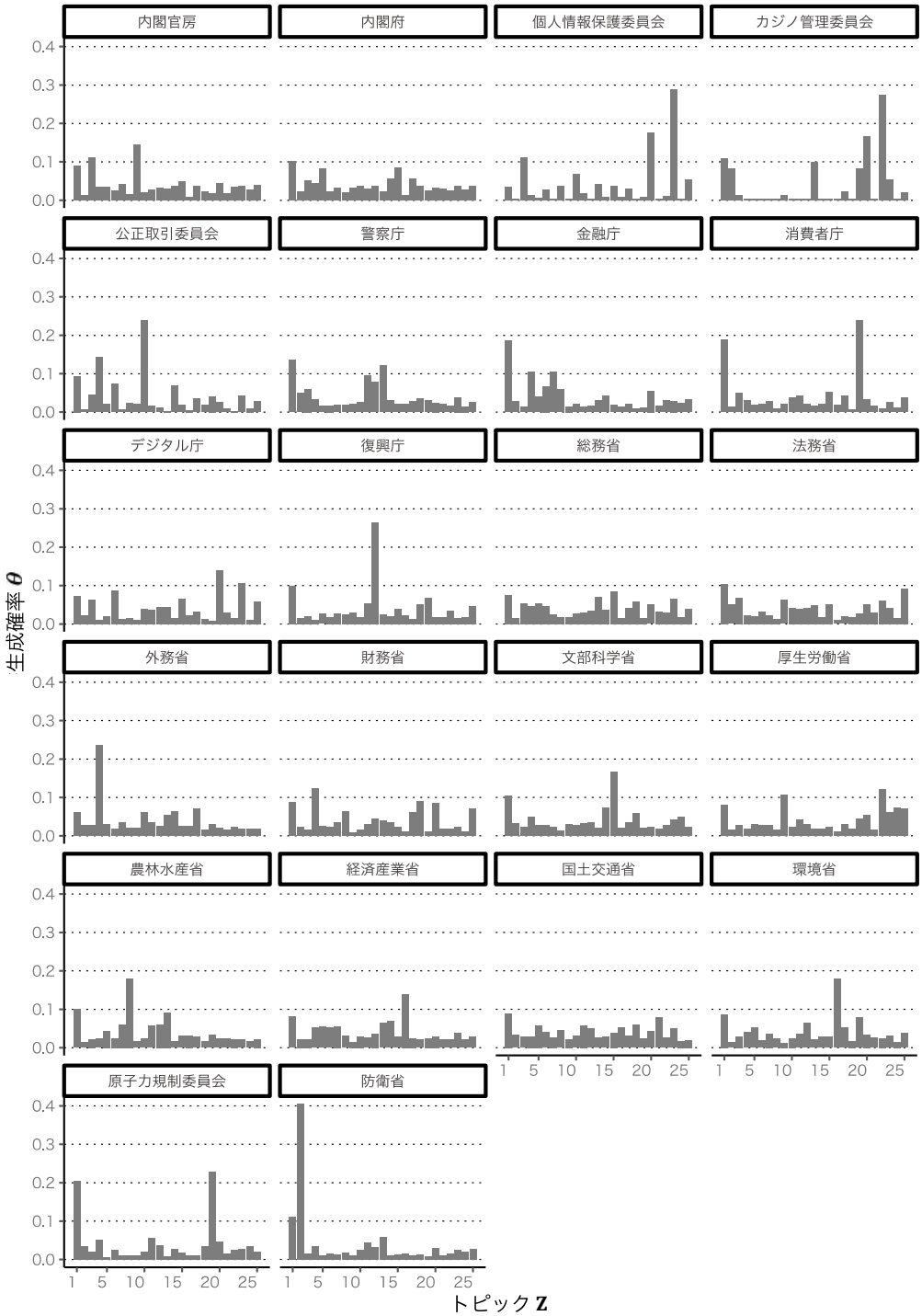

図 4　全体のトピック分布Θ（府省庁間比較）

出典：分析結果をもとに三上作成。トピック番号はすべて下端の横軸に対応し、それぞれ図 3 の丸付き文字と対応している。

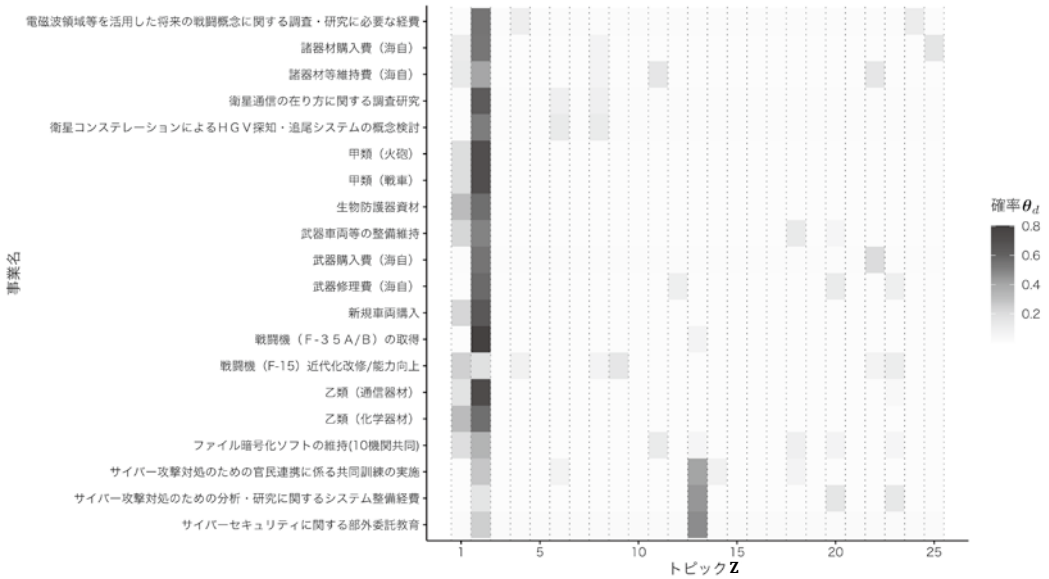

図５　防衛省事業（一部）のヒートマップ

出典：分析結果の一部をもとに三上作成。

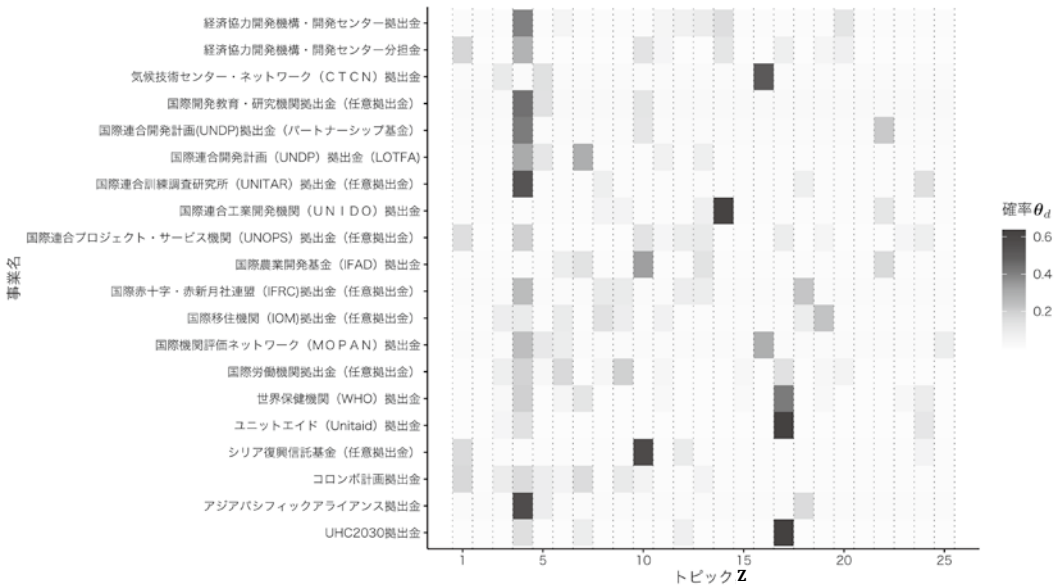

図６　外務省分担金・拠出金事業（一部）のヒートマップ

出典：分析結果の一部をもとに三上作成。

5. おわりに―限界と展望―

本稿では、機械学習手法を用いて行政事業レビューにおける組織（府省庁）間の差異を量的に可視化した。その結果は前節の通りである。ただし､行政事業レビューのうち「必要性評価」の文章情報のみに依拠した結論だという限界がある。

この分析の意図は、政策評価論における機械学習手法の応用に際し、その有用性と限界を考える点にあった。本稿が用いた手法はその一例にすぎないため更なる調査が必要ではあるものの、政策評価論の実務的課題を解決する上ではある程度の有用性が認められるのではないか。すなわち､「規則や構造の抽出」､「パターンの認識」に長けた機械学習は、これまでとは異なる研究アプローチの道を示してくれる。大量のデータから質的な情報を量的に把握できれば、これまで人手で処理しきれず埋もれていた情報を扱えるようになる。したがって、政策評価論における事例研究や歴史研究、あるいは実務における調査研究や予算査定を促進する際にも相性がよい。比較研究における分類軸としても利用できるほか、行動研究など他の量的研究との接続を考える上でも有用となる。

本稿の結果に基づいた今後の展望として、比較政策評価論の展開が考えられる。政府活動のパターンを把握するだけでなく、共時的・通時的に比較して法則性を明らかにすることもできる。他の評価制度、地方自治体、独立行政法人、あるいは国際比較にも展開できるだろう｡また、政府の活動パターンと市民や住民の納得との関係を行動行政学的に検討すれば、市民と住民の視点（松下 1975）から政策と行政の実態を監視し､「注意喚起情報（西尾勝 1990：114)」を定期的に生み出す実務的な仕組みを整えられるかもしれない。これは｢アジャイル型政策形成･評価」の実務動向にもリンクする点がある。

他方で、機械学習手法の限界も垣間見える。第 1 に、対象の性質や背景に関する専門性が必要である。本稿でも質的研究者の目線でラベル付けが必要であったように､「学習データの作成」や結果の解釈、最終的な判断には専門的な知見が必要となる。もちろん手法の動作原理をある程度理解することも必要である。第 2 に、機械学習手法が扱える情報の種類は、他の量的手法よりははるかに多いものの、政策評価論の研究者からすれば未だ満足できる水準にない。たとえば、制度やメカニズム、時間の流れ、歴史の文脈といった情報を総合して扱う手法はまだ発展途上であるし、特定のパターンが生じる理由にはやはり答えられない。

要するに、政策評価論においては機械学習の手法は魅力的ではあるが万能薬ではない。しかし、使い方次第では有用である。長期にわたる丁寧な研究を主としつつ、その補助として即応性と網羅性を重視したデータ分析を行う、そうした「二刀流」も可能となる。このメリットを最大限享受するためには、従来手法との上手な連携が大切である。

付記

本研究は、JSPS 科研費（JP23K18770）の助成を受けた研究成果の一部である。

参考文献

【日本語文献】
阿部真人（2021）『データ分析に必須の知識･考え方 統計学入門』ソシム。
岩田具治（2015）『トピックモデル』講談社。
大久保街亜（2016）「帰無仮説検定と再現可能性」『心理学評論』59（1）：57-67。
岡田謙介（2017）「ASA 声明とこれからの統計学の使われ方―最近の心理統計分野の動向から―」『社会と調査』（19）：88-93。
久保拓弥（2012）『データ解析のための統計モデリング入門―一般化線形モデル・階層ベイズモデル・MCMC―』岩波書店。
久保真人（編）（2016）『社会・政策の統計の見方と活用―データによる問題解決―』朝倉書店。
佐藤一誠（2015）『トピックモデルによる統計的潜在意味解析』コロナ社。
篠原舟吾・小林悠太・白取耕一郎（2021）「行政学における方法論の厳密化と多元的共存」『年報行政研究』56：145-64。
清水裕士（2018）「心理学におけるベイズ統計モデリング」『心理学評論』61（1）：22-41。
杉山聡（2022）『本質を捉えたデータ分析のための分析モデル入門』ソシム。
須山敦志（2017）『ベイズ推論による機械学習入門』講談社。
東京大学教養学部統計学教室（編）（1991）『統計学入門』東京大学出版会。
東京大学教養学部統計学教室（編）（1992）『自然科学の統計学』東京大学出版会。
友永雅己・三浦麻子・針生悦子（2016）「心理学の再現可能性―我々はどこから来たのか我々は何者か　我々はどこへ行くのか―（特集号の刊行に寄せて）」『心理学評論』59（1）：1-2。
南島和久（2011）「府省における政策評価と行政事業レビュー―政策管理・評価基準・評価階層―」『会計検査研究』43：57-71。
南島和久（2020）『政策評価の行政学』晃洋書房。
西尾勝（1990）『行政学の基礎概念』東京大学出版会。

西山慶彦・新谷元嗣・川口大司・奥井亮（2019）『計量経済学』有斐閣。
野田遊（2020）「大阪都構想の賛否の程度は情報提供で変化するか？」『同志社政策科学研究』21（2）：171-83。
馬場真哉（2019）『R と Stan ではじめる ベイズ統計モデリングによるデータ分析入門』講談社。
細野助博（2021）『公共政策のためのモデリングとシミュレーションの基礎』ミネルヴァ書房。
松浦健太郎（2016）『Stan と R でベイズ統計モデリング』共立出版。
松下圭一（1975）『市民自治の憲法理論』岩波書店。
蓑谷千凰彦（1988）『計量経済学（第 2 版）』東洋経済新報社。
持橋大地・大羽成征（2019）『ガウス過程と機械学習』講談社。
安井翔太（2020）『効果検証入門―正しい比較のための因果推論／計量経済学の基礎―』技術評論社。
山谷清志（1997）『政策評価の理論とその展開―政府のアカウンタビリティ―』晃洋書房。
渡辺有祐（2016）『グラフィカルモデル』講談社。

【英語文献】
Aleksovska, M. (2021) Accountable for What? The Effect of Accountability Standard Specification on Decision-Making Behavior in the Public Sector. *Public Performance & Management Review*, 44(4): 707-34.
Aleksovska, M., Schillemans, T., and Grimmelikhuijsen, S. (2019). Lessons from Five Decades of Experimental and Behavioral Research on Accountability: A Systematic Literature Review. *Journal of Behavioral Public Administration*, 2(2): 1-18.
Anastasopoulos, L. J., and Whitford, A. B. (2019) Machine Learning for Public Administration Research, With Application to Organizational Reputation. *Journal of Public Administration Research and Theory*, 491-510.
Bishop, C. M. (2006) *Pattern Recognition and Machine Learning*, Springer-Verlag.（=2012、元田浩・栗田多喜夫・樋口知之・松本裕治・村田昇監訳『パターン認識と機械学習―ベイズ理論による統計的予測―（上）（下）』丸善出版。）
Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003) Latent Dirichlet Allocation, *Journal of Machine Learning Research*, 3: 993-1022.
Bovens, M. (2007) Analysing and Assessing Accountability: A Conceptual Framework, *European Law Journal*, 13(4): 447-468.
Bovens, M., Goodin, R. E., and Schillemans, T. (eds.) (2014) *The Oxford Handbook of Public Accountability*, Oxford University Press.
Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017) Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, 76(1): 1-32.
Cunningham, S. (2021) *Causal Inference: The Mixtape*, Yale University Press.（= 加藤真・河中祥吾・白木紀之・冨田燿志・早川裕太・兵頭亮介・藤田光明・邊土名朝飛・森脇大輔・安井翔太訳（2023）『因果推論入門―ミックステープ：基礎から現代的アプローチまで―』技術評論社。）
Frink, D. D., and Klimoski, R. J. (1998) Toward a Theory of Accountability in Organizations and Human Resource Management. In *Research in Personnel and Human Resources Management*, 16: 1-51.
Furubo, J. E., Rist, R. C., and Sandahl, R. (eds.) (2002) *International Atlas of Evaluation*, Transaction Publishers.
Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014) *Bayesian Data Analysis, Third Edition*, CRC Press.
Grimmelikhuijsen, S., Jilke, S., Olsen, A.L. and Tummers, L. (2017) Behavioral Public Administration: Combining Insights from Public Administration and Psychology. *Public Administration Review*, 77: 45-56.
Grüen, B., and Hornik, K. (2011) Topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software*, 40(13): 1-30.
Hall, A. T., Frink, D. D., and Buckley, M. R. (2017) An Accountability Account: A Review and Synthesis of the Theoretical and Empirical Research on Felt Accountability. *Journal of Organizational Behavior*, 38: 204–24.
Jilke, S., Lu, J., Xu, C., and Shinohara, S. (2018) Using Large-Scale Social Media Experiments in Public Administration: Assessing Charitable Consequences of Government Funding of Nonprofits, *Journal of Public Administration Research and Theory*, 29(4): 627-39.
Kruschke, J. K. (2015) *Doing Bayesian Data Analysis, A Tutorial with R, JAGS and Stan 2nd ed*., Elsevier Inc.（=2017、前田和寛・小杉孝司監訳『ベイズ統計モデリング―R, JAGS, Stan によるチュートリアル―［原著第 2 版］』共立出版。）
Kullback, S., and Leibler, R. A. (1951) On Information and Sufficiency. *Annals of Mathematical Statistics*, 22: 79-86.
Lerner, J. S., and Tetlock, P. E. (1999) Accounting for the Effects of Accountability. *Psychological Bulletin*, 125: 255-75.
Mimno, D., Wallach, H., Talley, E., Leenders, M., and McCallum, A. (2011) Optimizing Semantic Coherence in Topic Models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 262-72.
Overman, S., Schillemans, T., and Grimmelikhuijsen, S. (2021) A Validated Measurement for Felt Relational Accountability in the Public Sector: Gauging the Account Holder's Legitimacy and Expertise. *Public Management Review*, 23(12): 1748-67.
Overman, S., Schillemans, T., Lægreid, P., Fawcett, P., Fredriksson, M., Maggetti, M., Papadopoulos, Y.G., Rubecksen, K., Rykkja, L.H., Salomonsen, H.H., Salomonsen, A., and Wood, M. (2018) *Comparing Governance, Agencies and Accountability in Seven Countries*, CPA Survey Report.
Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002) *Experimental and Quasi-experimental Designs for Generalized Causal Inference*, Houghton, Mifflin and Company.
Wainwright, M. J. and Jordan, M. I. (2008) Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 1(1-2): 1-305.
Wasserstein, R. L., and Lazar, R. L. (2016) The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, 70(2): 129-33.

【他外国語文献】
オランダ語文献
Bovens, M., en Schillemans, T. (2009) Publieke Verantwoording: Begrippen, Formen en Beoordelingskaders. In Bovens, M., en Schillemans, T. (red.), *Handboek Publieke Verantwoording*. Lemma, 19-34.
Jilke, S., Olsen, A. L., Tummers, L., en Gremmerikhuijsen, S. (2016) Gedragsbestuurskunde: Combineren van Inzichten uit de Bestuurskunde en de Psychologie. *Bestuurskunde*, 25(3): 9-16.
Schillemans, T., Bokhorst, M., van Genugten, M., en Vrielink, M. O. (2018) Voorlopige Contouren van Bestuursgericht Toezicht: Empirische Inzichten in Jonge Praktijken van Bestuursgericht Toezicht. *Bestuurskunde*, 27(4): 54-66.

【URL 等】
1. 内閣官房行政改革推進本部事務局（2023）「行政事業レビューシート 令和４年度データベース」政府の行政改革ウェブサイト（2023 年 6 月 1 日取得、https://www.gyoukaku.go.jp/review/database/database221209.xlsx）。