政府の予測力の現在地を探る④

～予測は当たるか、信頼できるか～

上席研究員　小林 郁雄

本稿は、政府予測の体系的な整理・分析を通じて、予測を活用しつつも惑わされないための実務的視点を提示するシリーズ最終報である。これまで、予測の価値の源泉、政府予測 319 件の全体像や政策分野別の特徴、政府予測の「情報インフラ」として役割などを明らかにしてきた。本稿では、それらを踏まえ、予測の当否や信頼性に焦点をあてて、政府予測の読み解き方を考察する。

予測の価値は利用者にとっての「有用度」で決まるが、その客観評価は容易でない。そこで、予測のプロセス・手法・結果から成る「予測の信頼度」という枠組みを提示する。「精度」のほかに、スキル、前提条件、透明性、改善履歴など、多様な評価の観点を位置付けたが、その背景には、約 6 割の政府予測で精度検証が困難という構造的要因がある。①答え合わせ自体が難しい超長期予測、②前提が現実と乖離しやすいシナリオ型予測、③大地震のような低頻度事象の確率予測、④政策誘導効果を伴う自己否定／自己実現型予測は、予測結果と実績の比較では当否が決まらない。すでに、地震の長期評価や気候変動予測、野生生物の個体数推計など、ベイズ統計的な政府予測は多数あり、今後 AI 型の予測が増加すれば「信頼度」を読む力の重要性が増す。

また、世界最高水準の精度を誇る気象予報の検証結果は、多くの示唆に富んでいる。降水の有無の予測には複数の評価指標があり、利用目的ごとに重視すべき指標は異なる。竜巻注意情報は適中率こそ 1 割にとどまるが、高い価値を持ち、当否のみで予測評価することの危うさを示している。3 日先の台風中心位置の予測は、最先端技術を用いてもなお「大阪－名古屋間」に匹敵する誤差を含み、予測性能には避けがたい限界がある。

こうした予測の限界と構造を踏まえ、「信頼度」を重層的に評価し、利用者自身の予測リテラシーを高めることが、不確実性の大きい時代における実務的な予測活用の鍵となる。

本稿でわかること

- 政府予測の多く（約 6 割）で「精度検証が困難」であることの構造的な理由
- 超長期予測・シナリオ型予測・低頻度事象に対する確率型予測・自己否定／自己実現型予測の特性
- 精度検証困難な予測を含めた品質評価の枠組み「予測の信頼度」（プロセス・手法・結果）の考え方
- 天気予報の精度検証から読み解く、「利用目的に合った評価属性・指標の使い分け」とその重要性
- 竜巻注意情報にみる、「適中率」が約 1 割でも高い価値が成立する理由
- 精度向上の最前線に位置する台風・線状降水帯・大雨予測の予測精度の現状と限界
- 予測の限界や不確実性を踏まえた利用者側の予測運用方針・設計の重要性

1. はじめに

本稿を含む 4 本のレポート（シリーズ①～④）では、政府予測を網羅的に調査して、その結果をもとに政府の予測力の一端を明らかにする。

シリーズ①[1]では、「予測とは何か、どれだけの予測が存在するか」と題して、予測と知能の本質的な関わり、予測の価値の源泉、今回の政府予測の定義、予測の類型化に資する視点など、基本的な事項を整理した。

「どのような予測があるか」と題したシリーズ②[2]では、政府予測 319 件の全体像を 8 つの視点で分析するとともに、経済・産業分野の具体的な予測リストを提示した。続くシリーズ③[3]では、経済・産業分野以外の 5 つの政策分野の予測リストを示し、具体的な事例を深掘りして政府予測の現状・課題を明らかにした。

シリーズ最終の本稿では、予測の当否とその検証可能性に焦点を当てる。前半では、当否と検証可能性を読み解くための知見や考え方を整理し、後半では、気象予報を題材に予測精度の現状や限界をみていく。

予測との向き合い方を再考する一助となるよう、可能な限り具体的な事例に基づく説明に努めた。しかし、「予測の信頼度」という枠組みや気象予報の検証指標に関して、説明が概念的でやや難解になってしまった部分もある。そのような箇所は適宜読み飛ばしていただき、事例から得られる示唆や気づきを中心に本稿を読み進めていただければ幸いである。

なお、本稿の「政府予測」とは、日本の府省庁が自らの予測として公表し、そのことを令和元年度以降の公表資料で確認できるものである。府省庁が独立行政法人などに要請・委託することによって実施された予測は、研究段階にとどまるものなどを除き「政府予測」に含まれる[4]。

また、本稿では、「予測」と「フォーサイト」を、重なる部分はあるが異なる概念として整理した[5]。このため、「フォーサイト」の多くは今回作成した予測リストには含まれていない。また、政府の「長期戦略等」は基本的に「政府予測」そのものではなく、多くの場合、政府予測の掲載資料という位置づけになる[6]。

2. 政府予測の当否や信頼度を読み解く

（1）政府予測の多くは当否が決まらない

シリーズ①では、「予測の価値は、当否そのものよりも、いかに意思決定に役立つかで決まる」と述べた。政府予測は、天気予報のように生活者の行動の拠り所となり、経済見通しのように企業の経営判断の材料として、さらには政府自身の計画策定や政策判断の基盤として、多様な主体の意思決定を支えている。このため、政府予測の価値は、本来的には利用者にとっての「有用度」で決まる（※）。

※シリーズ③では、政府予測が自らの政策形成や制度運用を支える「情報インフラ」として機能していること、すなわち政府自身にとっての有用度が大きいことを示した。

もっとも、このような「有用度」は利用者の立場や目的により異なるため、政府予測の価値を客観的に評価することは難しい。シリーズ①では、予測する側がその価値を高める取り組みとして、次の 2 点を挙げた。

・利用者のための利便性向上（情報量、わかりやすさなど）
・予測の確度向上（予測が当たる可能性を高めることなど）

---

[1] Insight Plus「政府の予測力の現在地を探る①～予測とは何か、どれだけの予測が存在するか～」
[2] Insight Plus「政府の予測力の現在地を探る②～どのような予測があるか 前編（全体像と経済・産業分野）～」
[3] Insight Plus「政府の予測力の現在地を探る③～どのような予測があるか 後編（各政策分野の詳細）～」
[4] 前脚注 2　図表 1
[5] 前脚注 2　図表 2
[6] 前脚注 2　図表 3

このうち、前者は利用者の主観に左右されるのに対して、後者の「確度」は一定の客観評価が可能であり、予測の利用価値を左右する重要な要素でもある。ところが、シリーズ②・③で示したとおり、政府予測の約6割は精度検証が困難な予測に分類される。なぜなら、予測が外れたように見えても、それを「外れ」と断定できない構造的な性質をもつためである（この性質の詳細は(3)で後述する）。

したがって、政府予測の価値を一定の客観性をもって読み解くためには、予測の「精度」に限らない総合的な品質評価の枠組み――本稿では「政府予測の信頼度」という――が必要になる。

（２）政府予測を読み解くための判断材料「予測の信頼度」

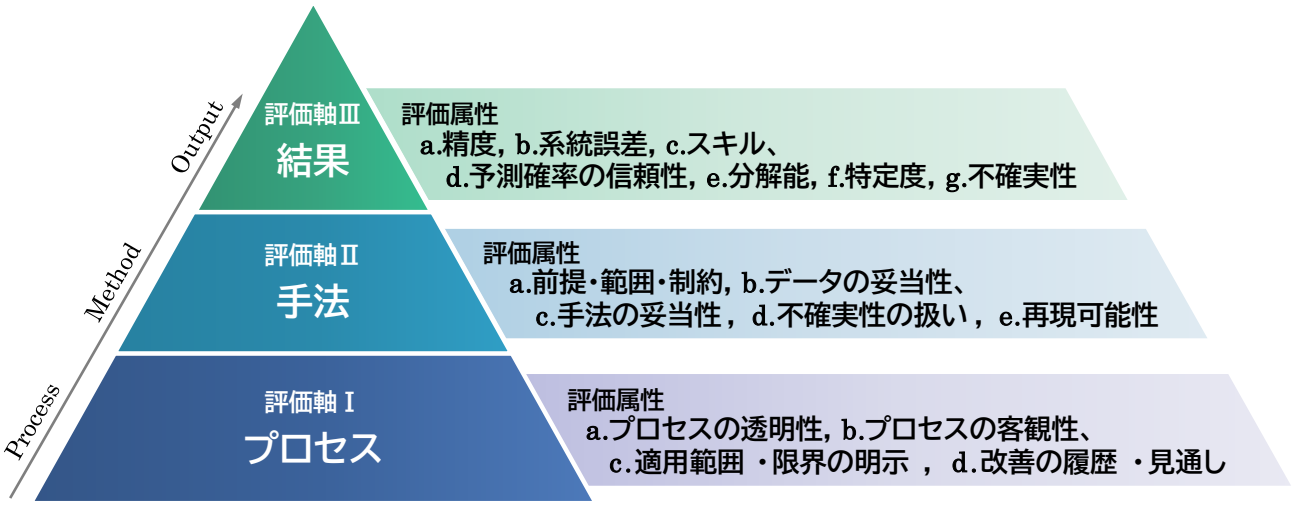
予測の品質評価の指標やガイドラインは分野ごとに複数存在する。例えば、気象・水文分野では、世界気象機関（WMO）により予測の検証に関するガイドラインが整備され[7]、主に予測結果の性能に関する科学的な検証方法が詳細に示されている。また、英国政府には、AQuA Book という分析業務の品質確保のための指針があり[8]、分析のプロセスや手法に求められる要件が詳細に規定されている。しかし、予測のプロセス・手法・結果を包括的に扱い、予測の品質を総合的に評価するための枠組みは必ずしも明確に定まっていない。

そこで本稿では、これら既存の枠組みを参考にしつつ、「予測の信頼度」をプロセス・手法・結果という3つの評価軸と、それらを構成する16の評価属性（評価の観点）から成る枠組みとして位置づける**≪図表1≫**。

このうち評価軸Ⅲは、精度検証が可能な予測だけに適用され、評価軸ⅠとⅡはすべての予測が対象となる。したがって、精度検証が困難な政府予測であっても、プロセスと手法という2つの評価軸、更には複数の評価属性（評価の観点）を用いて、「予測の信頼度」を重層的に、より深く読み解くことができる。

本節の冒頭で、政府予測の価値は利用者にとっての「有用度」で決まると述べたが、そこに「予測の信頼度」という客観性の高い視点を加えることによって、予測の価値をより的確かつ冷静に判断することが可能になる（評価属性の詳細については(4)で後述する）。

≪図表１≫ 予測の総合的な品質評価の枠組み――「予測の信頼度」



出典：当社作成

[7] 例えば、WMO-No.1364: Guidelines on the Verification of Hydrological Forecasts, 2025 < https://library.wmo.int/records/it_em/69478-guidelines-on-the-verification-of-hydrological-forecasts >
[8] GOV.UK, The AQuA Book, Published 30 July 2025 < https://www.gov.uk/guidance/the-aqua-book >

次項(3)では、精度検証困難な予測の特性について具体的な事例を交えて詳述する。続く(4)では、「予測の信頼度」の各評価軸を構成する評価属性の概要と留意点について詳述する。さらに(5)では、(3)で取り上げる低頻度事象の発生確率に関連して、確率の伝統的な解釈とベイズ的な解釈の違いについて説明し、その違いがもたらす予測実務への影響と政府予測との関係性に言及する。

（３）精度検証が難しい政府予測の代表例とその特性

予測の議論では、結果の当否、特に予測精度の高低に注目が集まりやすい。しかし、予測精度を科学的に検証するためには、予測が次のような前提条件を満たす必要がある。

・予測結果と対比できる実績が得られること（対比可能な観測結果が得られること）
・対比により得られた精度などの品質指標が、次回の予測の良し悪しを判断する材料となること（次回の予測に適用できるような、意味のある評価結果が得られること）
・実績が予測による影響を受けていないこと（観測結果が予測と独立であること）

これらの要件が満たされない予測、すなわち精度検証が難しい予測の代表例は 4 つある。

①超長期の予測
②シナリオ型の予測
③低頻度事象に対する確率型の予測
④自己否定型または自己実現型の予測

詳細は後述するが、遠い未来の予測である超長期の予測では、事後検証を行うこと自体が物理的に難しい。シナリオ型の予測は、前提とした社会・経済の動きや行動選択が現実と一致して推移することは稀であり、実績と予測を比較しても、そこから単純に“当たった・外れた”と判断することはできない。また、大地震のような稀にしか起きない低頻度事象に対する確率予測は、単一または少数の実績しか得られないため、予測された確率の当否を統計的に判定することに原理的な限界がある。さらに、自己否定性や自己実現性と呼ばれる性質をもつ予測では、予測の公表が未来に影響を及ぼすため、実績と予測を単純比較できない。

これらの予測は、“当否”という評価の仕方を適用できない性質をもつことから、外形的には“外れたように見える”予測であっても、科学的には外れたと断定することが難しい。

政府予測の約 6 割は精度検証が「困難」な予測に該当し、特に、防災・危機管理分野と食料・農林水産業分野を除くとその割合は約 8 割に上る。一部の政策分野を除くと、政府予測は“外れたと断定されるリスクが低い”という言い方もできる。以下では精度検証が難しい予測の代表例について、順に詳述する。

①超長期の予測

遠い未来を対象にした予測は、最終的な結果を事後的に検証できるようになるまでに数十年を要するため、主として物理的な制約の点から予測精度の検証が難しい。なお、超長期の予測であっても、不確実性の扱い方（特定型かシナリオ型かなど）によっては、予測実施から一定期間が経過した段階で、初期的または中間的な精度検証が可能な場合もある。

本シリーズでは、最長予測期間が 2045 年以降の予測（20 年以上先の未来を含む予測）を超長期の予測として整理し、その検証可能性を「困難」と分類した。50 年先や 100 年先という場合には、答え合わせ自体がほぼ不可能であることは容易に想像がつく。一方で、20 年程度の期間の場合は、検証が物理的に不可能とまでは言い切れないかもしれない。しかし、この場合でも、20 年以上前に設定された予測の前提条件（人口動態、経済動向、技術進歩、政策変更、国際情勢、自然災害など）などが現在に至るまで妥当性を維持している可能性は極めて低い。仮に実績値（観測値）を予測値と比較し、予測精度を算出できたとしても、その数値を評価することの意義はごく限定的となる。むしろ、20 年間における前提条件の変化や、実績値との乖離を生んだ要因を分析することの方が重要になる。

以上から、超長期予測については、物理的な制約に加えて、前提条件の維持に関する構造的な制約が重なることによって、予測精度の検証が困難になると解釈することができる。

今回の政府予測リスト 319 件には、最長予測期間が 2045 年以降となるような超長期予測が 108 件含まれる。さらに、50 年以上先の未来（2075 年以降）を対象としたものは 42 件あり、物理的な制約の面から精度検証が難しい予測と考えることができる（※）。超長期予測 108 件の分野別の内訳をみると、約 6 割は、年金財政、介護サービス、労働需給、人口、上下水道インフラ、食料需給、廃棄物処理など多岐にわたっており、約 3 割が気候変動やその影響に関する予測、約 1 割が地震・津波に関する予測となっている。

※地震・津波の予測については最長予測期間が明示されていないが、予測対象となる大～巨大地震の性質上、超長期予測に含めて整理した。

②シナリオ型の予測

本シリーズでは、シナリオ型の予測を、「将来の複数の仮定・前提条件（例：政策対応の有無、経済成長率など）に基づいて予測を行い、それぞれの結果を提示する予測、またはシナリオ間の差分や比較結果をもって予測値として示す予測」と定義した（詳細はシリーズ①）。このような予測の精度検証の難しさについては、すでにシリーズ②（3(3)②c. シナリオ型予測の精度検証の限界）で触れたが、改めて補足しておく。

シナリオ型の予測は、将来の出来事や状態を言い当てることを目的としたものではなく、将来に関する選択や意思決定を支えるために、複数の将来の姿を提示するものと位置付けられる。そのため、どのシナリオが実現するか、あるいは実現しやすいのかといった可能性は前もって特定しておらず、予測結果が“当たった・外れた”と一義的に判断できる構造になっていない。

その上、予測に用いた前提条件が当初描いたシナリオどおりに推移するとは限らない。特に社会・経済に関するシナリオでは、前提条件がシナリオどおり長期にわたって推移する可能性は極めて低い。実績値（観測値）と予測値を比較して、両者の乖離を分析すること自体は可能だが、乖離の原因がシナリオ自体の妥当性にある場合、予測の精度を形式上算定できたとしても、その値は意味をもたない。

すなわち、シナリオ型の予測は、予測の前提が複数のパターンに分かれるため、予測の当否を判断しにくい構造的特性を備えていることに加えて、実績値と予測値の乖離がシナリオの妥当性に強く影響されることも相まって、予測精度の検証が困難になると解釈することができる。

逆にいえば、前提条件がシナリオに完全に一致して推移した場合には、予測精度の検証は可能となる。そのような例として、シナリオが政策対応の有無（例：環境規制のあり・なし）であるケースや、2 つのシナリオのうち 1 つが非現実的なケース（例：すでに堤防が整備されている状況下での「堤防なし」シナリオ）が挙げられる。そのようなケースでは、前提条件がどちらかのシナリオどおりに推移する可能性が高いため、シナリオ型の予測であっても精度検証が可能な場合がある。

今回の政府予測リスト 319 件には、シナリオ型予測が 133 件含まれ、そのうち長期（5 年超 20 年以下）予測 55 件と超長期予測 68 件で、シナリオ型予測全体の 9 割を占める。

相対的に少数派となる短期的なシナリオ型予測には、シリーズ③の図表 21-1 の№20「富士山噴火後の首都圏での降灰の予測」や、№21「富士山噴火に伴う降灰被害の予測」が挙げられる（いずれも噴火後最大約 15 日間を対象とした予測）。また、シリーズ②の図表 7-7 に示したとおり、政策分野別では、経済・産業分野（27 件）と環境・エネルギー分野（48 件）でシナリオ型の予測の比率が高く、それらの分野で精度検証困難な予測が多くを占める要因となっている。

③低頻度事象に対する確率型の予測

確率型の予測の場合、結果の当否を判定するためには、予測値と実績値に関して十分な数のデータが必要になる。例えば、天気予報の 1 つである「降水確率予報」は、1 日当たり 8 つの予測値が生成される（※）。これに対して、実績値は予測時間帯における降雨の有無として得られるため、数か月程度の間に、統計的な評価に足る多数の予測値と実績値が得られることになる。

※この予測はシリーズ③の図表 21－2 の政府予測№34 に該当する。降水確率は 1 日に 2 回（5 時と 17 時）発表され、1 回の予測で 24 時間先までを 6 時間ごとに区分した 4 つの降水確率が 10%刻みで示される。

精度の検証方法は幾つかあるが、例えば「降水確率 80%」という予測を 100 回行い、そのうち 80 回で降雨があり、20 回で降雨が無かったと仮定する。この場合、実際に降雨があった頻度(frequency)は 80%（80 回／100 回）で、予測値の 80%と一致するため、高い予測精度であったことが客観的に評価できる（※）。このように、高頻度(high rate)で発生する現象の場合、確率型の予測であっても精度検証が「可能」となる。

※この検証方法は、確率予測の検証でよく知られた「信頼性(Reliability)」という評価属性に対応したものである（これについては(4)①の d で後述する）。

一方で、同じ確率型の予測でも、大きな地震のような稀にしか発生しない現象（低頻度事象：rare event）の予測では、予測値に対応した実績値が十分に得られない。例えば、「ある活断層で 30 年以内に地震が発生する確率は 80%」という予測に対して、その後 30 年間地震が発生しなかったとしても、予測が外れたとは断定できない。なぜならば、その期間内に地震が発生しない確率も 20%存在するからである。同じように、期間内に地震が発生したとしても、一度だけでは予測が当たったと結論付けることはできない。期間内に地震が発生しない確率は 20%あるという予測が外れたという見方もできるからである。

仮に同様の予測が 10 回繰り返され、そのうちの 8 回前後で期間内に地震が発生したことが確認できれば、

予測の精度を一定の客観性をもって検証することもできるが、活断層に伴う地震のように数千年に一度という極めて低頻度（low rate）（※1）の現象では、そうした検証は事実上不可能である（※2）。

※1　日本語の「頻度」には2通りの意味がある。地震の発生頻度のような頻度は、回／年という単位をもち、英語だと rate が用いられる。一方、前述の例のような降雨ありの頻度は、単位をもたず、割合や%で表され、英語では frequency が用いられる。

※2　ここでは古典統計的な確率解釈で表現したが、ベイズ統計の立場で論じても結論は変わらない。政府予測の中にはベイズ統計的な予測が数多くあり、その詳細については（5）で後述する。

とりわけ注意を要するのが、「30年以内の地震発生確率が0.1～3%」といった、低頻度で、なおかつ巨大な災害に関する予測結果の解釈である。この表現は、地震の発生確率が低いもののゼロではないことを示すものだが、一方で、30年以内にその地震が発生しない確率が97%以上あり、「発生しない確率が非常に高い」という解釈も科学的には誤りではない（※）。

※97%という数値だけをみると、IPCC（気候変動に関する政府間パネル）による可能性（Probability）の表現では「非常に高い」に相当する値である。気候変動に関する可能性の表現を地震発生確率に持ち込む意図はないが、「30 年以内に地震が発生する確率」に比べて「そうならない確率」の方が非常に高いことを示すことに変わりはない。統計用語では「発生する確率÷発生しない確率」はオッズと呼ばれ、上記の例では、オッズが0.1／99.9 ～ 3／97という非常に低い値となり、数値の上では、発生する見込みが非常に薄いという結論が導かれる。

このような低頻度巨大災害の予測を適切に解釈することは非常に難しい。1995年の阪神・淡路大震災の震源断層である「野島断層」を含む六甲・淡路島断層帯主部（淡路島西岸区間）は、震災後の調査によって平均活動間隔が1700～3500年と推定された。仮にこの知見を震災前に得ていたとすると、この断層帯の震災発生直前における 30 年以内発生確率は、0.02～8%という予測結果であったことが知られている[9]。そのほかにも、2016年の熊本地震の本震の震源断層とされる布田川断層帯では、地震発生直前における 30 年以内発生確率はほぼ0～0.9%と予測されていた[10]。したがって、活断層に伴う地震の長期評価結果を利用する際は、予測された発生確率がその程度の値であっても、地震が発生した実例が複数存在することを忘れてはならない。

六甲・淡路島断層帯主部の場合は、8%という発生確率が、他の活断層の長期評価結果と比較して相対的に高い値であることから、むしろ地震対応の必要性を示唆する予測結果であったという解釈が成り立つかもしれない。しかし、布田川断層帯の場合は、予測された地震発生確率からはそのような解釈が難しく、対策を最優先することを、むしろ躊躇させかねない予測結果であったという見方もできる（※）。

※布田川断層帯の場合は、隣接する日奈久断層帯で28時間前に発生したM6.5の大きな地震が引き金になった可能性が指摘されており、当該断層帯単独の活動を想定した長期評価の結果（0～0.9%という30年以内発生確率）はこのケースに適用すべきでないという見方もある。しかし現状では複数の断層帯の連動を想定した予測方法は確立されていないため、この事例は長期評価の限界を示すものともいえる。

9 地震調査研究推進本部「今までに公表した活断層及び海溝型地震の長期評価結果一覧」令和7年1月15日現在<https://www.jishin.go.jp/main/choukihyoka/ichiran.pdf>

10 同上

以上から、低頻度事象に対する確率型の予測の特徴として次の 2 点が挙げられる。

・一般に精度検証が困難である

・予測結果の解釈が難しく、特に巨大災害の場合は、発生確率の大小に惑わされないことが重要になる（人的被害の有無や社会・経済的な被害の規模を踏まえた「リスク」を評価するなど、利用者側の予測を読み解く力と予測との向き合い方が問われることになる）

今回の政府予測リスト 319 件には、確率型の予測が 30 件含まれている。このうち低頻度事象に対する確率型の予測には、地震の長期評価関連 5 件と気候変動予測 10 件を合わせた 15 件ほどが該当する。

④自己否定型または自己実現型の予測

政府予測が強いメッセージ性をもつことについてはシリーズ①で言及したが、それらの予測の中には、予測された未来の実現を回避させる効果をもつものがある。典型的な例が災害時の被害発生予測であり、予測結果を示すことによって、自治体や企業、生活者に対して、被害の回避・軽減のための行動を促す役割を担っている。予測の公表が、その後の対策や行動に影響を及ぼすことによって、結果として予測が当たりにくい状況を作り出すという点で、これらは自己否定型の予測と呼ばれる。このような予測は、防災・危機管理分野に限らず、環境・エネルギー分野などの政府予測にも一定程度存在する。

同様に、政府予測の中には、予測された未来の実現を誘導するような効果をもつものも多い。典型的な例として、政策の方向付けの根拠を担う予測が挙げられる。予測の公表が、それと整合的な政策措置の導入や、企業等の予見性の向上を通じて行動に影響を及ぼし、結果として予測が当たりやすい状況を作りだすという点で、これらは自己実現型の予測と呼ばれる。このような予測は、自己否定型の予測とは異なり、政府予測全般に広く存在するとみられる。

このような予測では、たとえ予測と実績に乖離が生じたとしても、それは予測の提示によって未来が変化した結果だと解釈できるため、単純に「予測が外れた」とは言い切れない。したがって、形式的に精度検証を行ったとしても、その結果を正しく解釈することは難しく、この点において精度検証が困難だといえる。

仮に、形式的に算定された精度が高い水準にあったとすると、期待された政策形成や行動変容が十分に実現されなかった可能性がある。特に自己否定型の予測の場合には、回避すべき未来というメッセージの訴求が不十分であった可能性が指摘され、低い評価につながる可能性すらある。

自己否定型の政府予測には、次のような例が挙げられる。

・シリーズ③の図表 3-2 の№53「災害リスク地域に居住する人口の予測」
　災害リスク地域内での居住の再検討を住民・自治体に促す効果⇒機能すれば、災害リスク地域における居住人口の抑制につながる。

・シリーズ③の図表 7-1 の№18「病虫害発生予報、予察警報、注意報」
　病虫害への備えを喚起して、その発生を未然に防止・抑制する効果⇒機能すれば、病虫害の発生回避・流行抑制につながる。

・シリーズ③の図表 16-1 のNo.1-10、14-21「気候変動関連の 4℃上昇等のシナリオ予測（※）」
　気候変動対策の実施や行動変容を促す効果⇒機能すれば、4℃上昇や深刻な影響の回避につながる。
※気候変動予測では複数のシナリオが設定されるが、4℃上昇などの温暖化シナリオは、将来の深刻なリスクを可視化することで、そのような未来を回避する行動を促す目的も含めて設定されている。
・シリーズ③の図表 16-2 のNo.43「今後 1 週間の電力需給見通し」
　需給の見通しは需要側の行動や対策を促す効果⇒機能すれば、電力需給ひっ迫の回避につながる。
・シリーズ③の図表 21-1 のNo.2「南海トラフ巨大地震発生に伴う被害の予測」
　事前防災対策の実施や行動変容を促す効果⇒機能すれば、地震被害の軽減（防災・減災）が実現する。
・シリーズ③の図表 21-1 のNo.21「富士山噴火に伴う降灰被害の予測」
　事前防災対策の実施や行動変容を促す効果⇒機能すれば、火山被害への防災・減災が実現する。

一方で、自己実現型の政府予測の具体例を挙げることは必ずしも容易でない。予測の目的から考えて、予測実施者が自己実現性を意識している可能性はあるものの、実際にそのような誘導効果がどの程度存在するのかという判定が難しいためである。ただし一般論として、次の 2 つの政府予測は、制度設計や政策運用との結びつきの強さからみて、自己実現型の予測とみなして差し支えないと考えられる。

・シリーズ②の図表 8-1 のNo.3 をはじめとした「中長期の経済財政に関する試算」
　潜在成長率の予測の場合、成長期待や投資行動の予見性を高め、その方向を企業等に示す効果⇒機能すれば、予測された潜在成長率に近い経済成長が現実の経済で実現に近づくことになる。
・シリーズ③の図表 16-1 のNo.24 をはじめとした「エネルギー需給・電源構成に関する長期的な見通し」
　電力需要・電源構成の見通しの場合であれば、原子力政策、送電網投資、企業の設備投資を方向付けする効果⇒機能すれば、予測された電源構成に近いエネルギー需給体制へと現実が変化する。

（4）「予測の信頼度」を構成する評価属性について

上記(2)では、予測の品質評価の枠組みとして「予測の信頼度」を掲げ、それを『プロセス』『手法』『結果』という 3 つの評価軸から体系化した。しかし、予測『プロセス』の品質を評価するといっても、どのような観点から読み解けばよいのかが明確でなければ、実務で活用することは難しい。

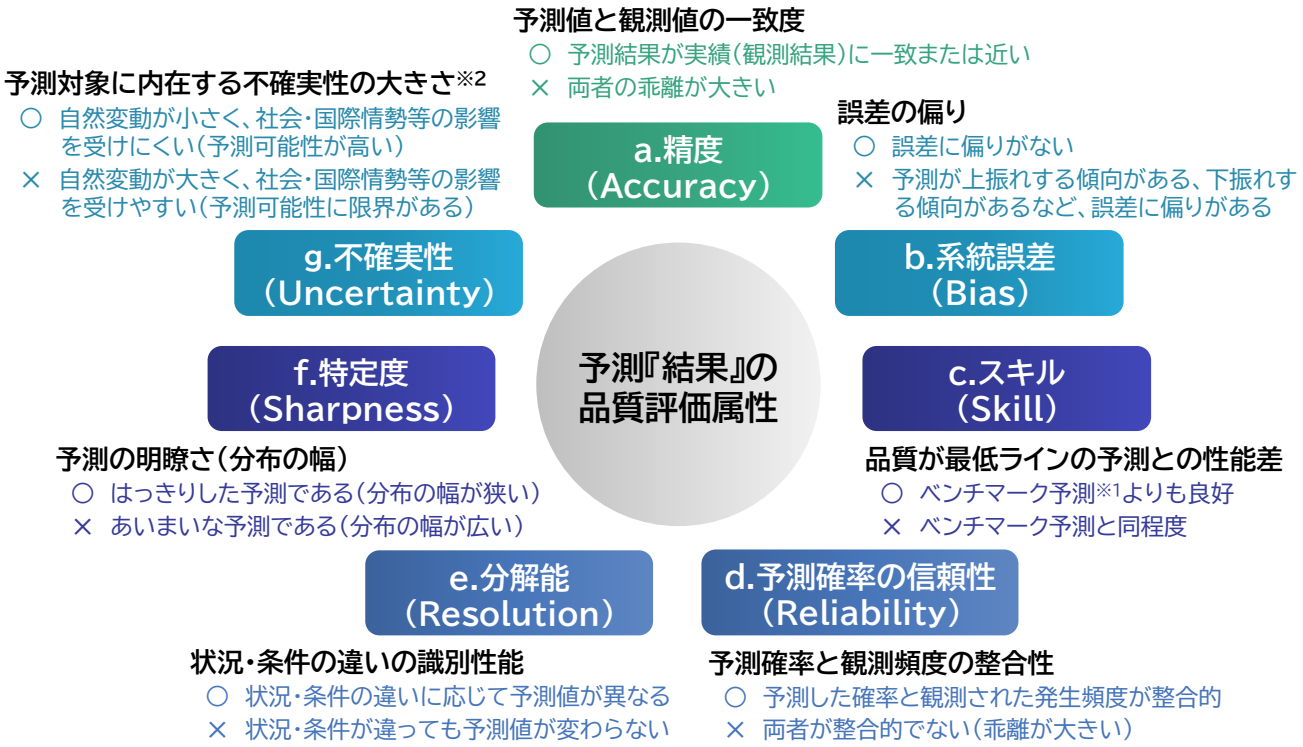
そこで本節では、各評価軸の構成要素である評価指標（評価の観点）の概要と留意点を整理する。なお、予測を実施する際の流れは、プロセス（予測の設計・手続き・統制）の検討→手法（使用データ・予測モデル）の検討→結果の取得という順序になるが、ここでは、予測品質の読み解きやすさという観点から、①結果、②手法、③プロセスの順で説明を進める。また、予測を読み解くための実務的視点の提示という本シリーズの目的に鑑み、ここでは、具体的な評価指標や計算式など技術的な詳細には立ち入らないこととする（巻末の**【参考１】**で、代表的な評価指標「ブライアスコア」「ブライアスキルスコア」について解説する）。

①　『結果』の品質を読み解く

気象・水文分野は、予測『結果』の検証に関して、国際的な標準化が進んでいる分野の 1 つである。世界気象機関（WMO）は、降雨・気温などの気象要素や季節変化などの気候要素、河川流況などの水文要素を対象に、予測の検証方法をまとめたガイドラインを整備している。

2025 年に公表された水文予測に関する最新のガイダンス[11]を参考にすると、予測『結果』の品質を読み解く上で汎用性が高いと考えられる評価属性を 7 つ抽出することができる。**≪図表 2≫**に、各評価属性の概要と高評価・低評価のイメージを整理した。

≪図表２≫ 予測『結果』の品質評価属性



※1:過去の平均値を予測値とする、または過去データを単純回帰するなど、最低限超えるべき予測のこと。
※2:予測対象の特性だが、予測結果の精度や検証可能性の制約要因であるため、品質属性に位置付けた。
※3:a～cおよびgは非確率型と確率型のいずれの予測にも適用できるが、d～fは主に確率型の予測に適用できる。

出典：WMO-No.1364: Guidelines on the Verification of Hydrological Forecasts, 2025 を一部参考にして当社作成

これらの評価属性は、予測『結果』に対する精度検証が「可能」な予測のみに適用できる。また、個々に留意すべき点もある。例えば、「精度」は概念として理解しやすく、意思決定にも活用しやすい一方で、予測期間の長短や予測対象に内在する不確実性の大きさに左右される。このため、比較対象となる他の予測が存在しない場合には、「精度」だけで品質の「良し悪し」を判定することはできない。

また、「スキル」には、ベンチマーク予測の設定次第で評価結果が大きく変わってしまう特性がある。「予測確率の信頼性」は、確率予測の評価に欠かせない基礎的な属性ではあるが、予測確率と観測頻度との差をどのように解釈し、意思決定に結びつけるかの判断は難しい。「特定度」についても、「予測の幅が狭いほど、より良い予測である」という誤解を生みやすい点に注意を要する。

このように、評価属性にはそれぞれの特徴や留意点があるため、予測『結果』の品質評価では、複数の評価属性を併用することによって、品質を多面的に読み解くことが望ましい。

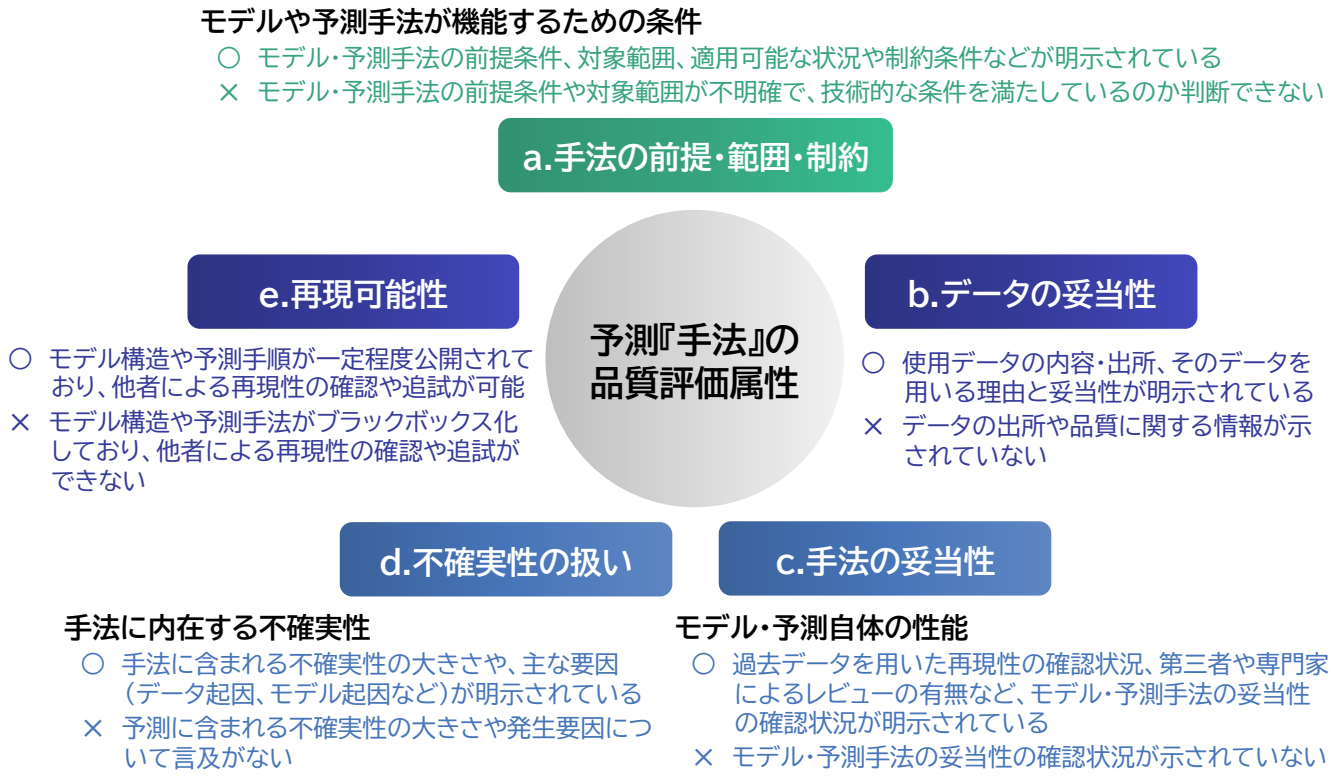
[11] 前脚注 7

② 『手法』の品質を読み解く

英国政府は、政府の意思決定を支える分析業務の品質保証を目的として、AQuA Book（the government Analytical Quality Assurance Book）という詳細な指針を整備している[12]。この指針は、予測を含む分析業務全般を対象としたものであり、品質確保における基本原則として、RIGOUR（再現性：Repeatable，独立性：Independent，現実性：Grounded in reality，客観性：Objective，不確実性の管理：Uncertainty-managed，頑健さ：Robust）を掲げている**≪BOX≫**。2025 年の改訂では、AI を活用したモデルなど、内部構造がブラックボックスになりがちな分析についても品質保証対象に含めるなど、内容の強化が図られた。

AQuA Book は、分析の発注者・分析者・保証担当者など、主として分析を行う側が満たすべき要件を規定した指針であり、分析結果を第三者が利用する際の留意点を整理したものではない。しかし、予測の利用者にとっても援用可能な有益な視点が多数含まれている。そこで、AQuA Book を参考にして、予測『手法』の品質を読み解く上で有用で汎用性が高いと考えられる評価属性を 5 つ抽出した。各評価属性の概要と高評価・低評価のイメージを**≪図表 3≫**に示す。

≪図表３≫ 予測『手法』の品質評価属性



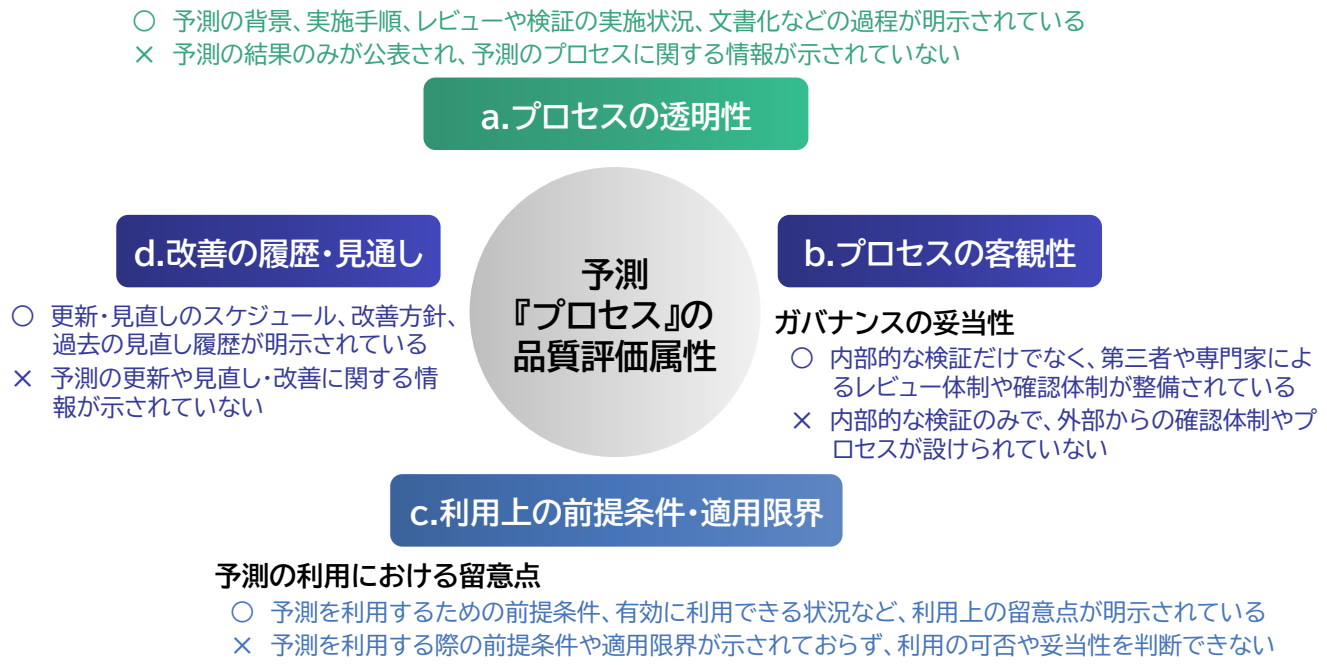
出典：GOV.UK, The AQuA Book, Published 30 July 2025 を一部参考にして当社作成

③ 『プロセス』の品質を読み解く

②と同様に、AQuA Book を参考にして、予測『プロセス』の品質を読み解く上で有用で汎用性が高いと考えられる評価属性を 4 つ抽出した。各評価属性の概要と高評価・低評価のイメージを**≪図表 4≫**に示す。

[12] 前脚注 8

≪図表４≫ 予測『プロセス』の品質評価属性



出典：GOV.UK, The AQuA Book, Published 30 July 2025 を一部参考にして当社作成

≪BOX≫ 英国政府 AQuA Book の"RIGOUR"（分析品質保証の基本原則）について

英国政府は、予測を含めた公的部門の分析全般を対象に、その信頼性と有用性を確保するための枠組みとして、AQuA Book (The government Analytical Quality Assurance Book) を定めている[13]。そこに示された、分析の各段階に共通する品質確保の基本原則が "RIGOUR" である。

・Repeatable（再現性）：同じ入力・前提・制約条件で分析すれば、同じ結果が得られること。
・Independent（独立性）：分析が偏見やバイアスによって歪められていないこと。
・Grounded in reality（現実との整合性）：
分析が現実世界の状況や結果と整合しており、現実とかけ離れた前提や想定に依存していないこと。
・Objective（客観性）：議論やレビューなどを通じて、主観や先入観が排除されていること。
・Uncertainty-managed（不確実性の管理）：
分析に含まれる不確実性が特定され、管理され、適切に説明されていること。
・Robust（頑健性）：
分析の不確実性や限界を踏まえても結果が大きく破綻せず、利用可能な品質が確保されていること。

本文では、「予測の信頼度」を構成する評価軸について、『手法』と『プロセス』を分けて評価属性を整理したが、それらは、より上位の評価概念である "RIGOUR" の 6 原則に基づいて体系化することもできる。その場合は、『手法』『プロセス』を、上記 6 つの観点を用いて統合的に評価することになる。

評価の抽象度はやや高まるが、簡潔さという利点もあることから、RIGOUR の 6 原則という大局的な評価軸を用いて、『手法』と『プロセス』を読み解くという選択肢も考えられる。

[13] 前脚注 8

（５）古典統計的な確率とベイズ的な確率の違い、それらと政府予測の関係について

人間は、確率という概念を直感的に扱うことを苦手としており、確率に基づく判断にはバイアスが入り込みやすいといわれる。伝統的な古典統計学において、確率とは「同じ条件で事象を繰り返したときの発生頻度」であり、一度しか起こらないような現象には確率という概念を適用できない。したがって、「A 社が 10 年以内に倒産する確率は 5%未満」といった表現は、古典的な確率の解釈では説明できない。同様に「火星に生命がいる確率は 1%程度」という状態についても、火星は一つしか存在せず、しかも生命がいるか否かの二択であってその中間は実在しないことから、古典統計学では扱うことができない。

しかし実際には、このような表現は日常的によく使われ、多くの人に違和感なく受け入れられている。こうした古典統計では扱いが難しい、いわば主観的な確率を扱えるのがベイズ統計である。ベイズ統計における確率は、「ある命題が真であることへの確信の度合い（主観確率）」と位置付けられ、一度しか起こらない現象でも確率の概念を適用できる。

両者の本質的な違いは、確率の捉え方にある。古典統計では、発生確率には唯一の真値が存在すると考える。したがって、データを蓄積すれば、観測された頻度はその真値に近づいていくと考える（これを大数の法則という）。一方、ベイズ統計では、発生確率そのものが不確実性をもつ量であると考え（※）、新たなデータが得られるたびに、その分布が更新されていくと考える（これをベイズ更新という）。

※発生確率を 1 つの固定した値と考えるのではなく、発生確率自体を、焦点の定まらない雲のような確率分布と捉えることになる。確率をさらに確率的に扱うという点で直感的にはわかりにくい。しかし、発生確率が十分に特定できていない段階では、その値をどこかに 1 つ仮定して決めてしまうことより、その値が存在する可能性のある範囲や傾向を分布として捉える方が、より理に適っていると考えることもできる。

以上の説明を宝探しに例えると、古典統計の場合、宝は最初から 1 か所に埋まっているので、穴を掘る回数を増やしていけばいずれ必ず見つかると考える。これに対して、ベイズ統計の考え方では、最初、宝の在りかは広く不鮮明に分布している。しかし、穴を掘るたびに情報（例：土の硬さや色など）が加わり、その分布が次第に狭まっていくので、宝の在りかが鮮明になっていくと考える。なお、一般的には、掘る穴を十分に増やしていけば、ベイズ的な探索で絞り込まれた宝の在りかと、古典統計的に探索した宝の在りかは、同じ場所に収束していくと期待される。古典統計ではデータの「量」に軸足を置き、ベイズ統計ではデータの「質（頻度以外の情報）の活用」に軸足を置く点に、両者の方向性の違いがある。

重要なのは、このような概念的な違いではなく、予測アプローチにおける実務的な相違の大きさにある。古典統計は、大量のデータを同じ条件下で繰り返し取得できる場合には強力な手段となるが、繰り返し観測することが難しい低頻度の現象には使いにくい。加えて、前者は発生頻度に基づく予測となるため、発生頻度以外の知見（地震であれば、地質構造やプレートの動きなど）を予測に組み込むことが構造的に難しい。これに対してベイズ統計では、専門家の判断や過去の研究で得られた知見を事前情報として予測に組み込みやすく、新たなデータが得られるたびに推定結果を更新するという構造的な特徴を備えている。

このため、現実社会の幅広い分野で、ベイズ統計的な確率予測が求められている（※）。

※一方で、頻度以外の知見を予測に組み込みやすいという利点は、ケースによっては留意点となる。例えば、事前情報に重大な誤りや偏りが含まれていた場合、少ない回数のベイズ更新では誤り等の影響を十分に排除できない。そのような場合、予測結果には重大なバイアスが含まれている可能性がある。

政府予測の中でも、地震の長期評価や気候変動予測のように、予測期間が超長期で繰り返し観測ができない現象では、古典統計の「頻度」に基づく確率解釈は適用しづらい。また、野生生物の個体数推計や漁業資源評価のように、モニタリングデータが追加されるたびに状況を見直す必要がある分野では、ベイズ更新を組み込んだ予測モデルの有効性が高い。このように、取得できるデータが限られる場合や、環境条件が変化を続ける場合、あるいは複数の情報源を有効に活用して予測を高度化したい場合は、ベイズ的手法による予測が実務的に適している。

ベイズ統計は、医療（疾病リスク予測など）や社会システム（電力需要予測など）、AI、ディープラーニングなど幅広い領域で社会実装が進んでいる。前述したように、人間は、ベイズ統計によって扱うことが可能になった「主観確率」に比較的違和感なく接している一方で、「30 年以内の地震発生確率が 0～1%」のようなベイズ的な予測結果を、意思決定にうまく活かすことには依然として苦労している。

このように、技術や社会がベイズ的な判断を前提に変化する中で、人間側の確率理解や判断能力がその変化に追いついていない状況だといえる。また、今回リストアップした政府予測には AI 学習モデル型の予測は見つからなかったが、今後は予測モデルの多様化が進み、政府予測におけるベイズ的な予測のウェイトが一層高まることが見込まれる。このため、政府予測の信頼度を正しく読み解き、意思決定に適切に活かす力が、これまで以上に求められることになる。

３.気象予報にみる予測精度の現状と限界

気象予報は、気象庁という行政部局の日常業務（現業）として実施されている点で、ユニークな政府予測である。しかも、その多くが精度検証可能な予測に分類され、予報の事後検証が継続的に実施・蓄積されている政府予測の代表例でもある。検証結果は、気象庁のホームページで広く公開されている。さらに、台風や大雨の予測といった、気象庁が戦略的に取り組んでいる重点分野の施策については、「気象庁業務レポート」という形で、予測の現状分析と課題を含めた詳細が毎年度公表されている。

ここでは、これらの公表資料に取り上げられた具体的な予測を題材として、予測『結果』の品質を中心に、検証結果の読み解きを進める。これを通じて、予測を読み解く際に役立つ知見や視点を整理しつつ、精度検証可能な政府予測の現在地を明らかにしていく。

以下、(1)では、一般に馴染みの深い天気予報の精度検証結果を取り上げ、予測『結果』を読み解く際の実務的な留意点を洗い出す。続く(2)では、予測の価値が当否だけで決まらないことを示す代表例として、竜巻注意情報を取り上げ、その精度検証結果を検討する。さらに(3)では、気象庁の予測の中でも精度向上の最前線に位置づけられる台風、線状降水帯、大雨に関する予測について、予測精度の現状と課題の整理を行う。

（１）天気予報の精度検証が示唆する「予測結果を読み解く視点」

日本の天気予報は世界でもトップクラスの精度を誇るとされるが、それでも、「明日は雨が降る」という前日の夕方 17 時時点の予測は、平均すると約 20%の確率で外れてしまう（＝空振りしてしまう）。

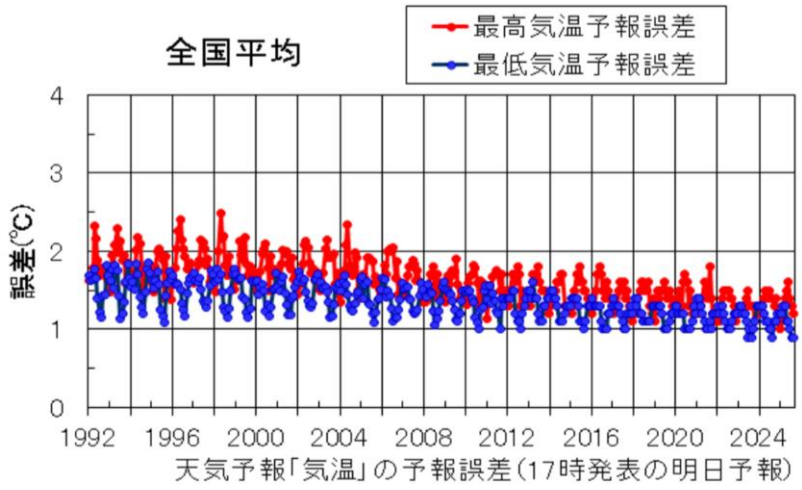
そのような検証結果は、気象庁のホームページで確認することができる。天気予報の場合は、①日中の最高気温・最低気温、②降水の有無、③降水確率予報について、『結果』の品質を中心とした評価が行われている。このうち、①と②では「精度（Accuracy）」に軸足を置いた検証が行われ、③では「予測確率の信頼性（Reliability）」の検証が行われるなど、同じ天気予報でも内容に応じた評価属性が適用されている。

①　「日中の最高気温・最低気温の予報」の精度

最高気温・最低気温の検証では、予測『結果』の評価属性である「精度」を測る指標として、RMSE（二乗平均平方根誤差）が採用されている。これらの予測は、結果を 1 つの数値として示す特定型の予測であるため、予測値と実績値（観測値）の差を「誤差」として定量化することができる。一般に、誤差が小さいほど、予測精度が高いことを意味する。

誤差による評価は、予測の精度を直感的に理解しやすいだけでなく、誤差の経年変化を示すことによって、予測性能の改善状況を視覚的に把握できる。**≪図表 5≫**は、毎日 17 時に発表される明日（翌日）の予報のうち、最高・最低気温の予報について、予測誤差（全国平均）の 1992 年以降の推移を示したものである。これをみると、これまでほぼ一貫して「精度」が改善されてきたことがわかる。

≪図表５≫ 最高・最低気温の予測誤差（全国平均）



出典：気象庁ホームページ「天気予報の精度検証結果」
〈 https://www.data.jma.go.jp/yoho/kensho/yohohyoka_top.html 〉

このような予測の改善状況や履歴の明示は、予測『プロセス』の品質評価属性である「改善の履歴・見通し」に対応している。前述したように、「精度」は、予測対象に内在する不確実性や予測期間の長短などに左右されるため、ある時点の指標値だけでは、その精度の良し悪しを比較・判断することが難しい。この事例では、予測『結果』の品質を「精度」の観点から評価し、『プロセス』の品質を「改善の履歴・見通し」の観点から評価できるため、「予測の信頼度」が継続的に向上している状況を知ることができる。

この事例は単純ながらも、「予測の信頼度」の評価では、1つの評価軸だけでなく複数の評価軸や評価属性を組み合わせることが、予測を読み解く上で重要になることを示している（※）。

※この事例の場合、コンピュータなどの計算機資源の飛躍的な性能向上や、新たな衛星観測を含めた観測体制の強化といった段階的な外的要因が、この予測の精度の向上に直接寄与した様子はうかがわれない。したがって、この予測の精度向上では、予測データの長期的な蓄積やモデル改良といった内部的な要因が大きなウェイトを占めることが示唆される。予測の『プロセス』という評価軸を加味することによって、例えば「精度の飛躍的な向上は見込みにくいが、今後も一定の改善が緩やかに続く可能性が高い」といった、意思決定に資する見通しを立てることも可能になる。

②「降水の有無に関する予報」の精度

a. 3種類の適中率の使い分け：利用目的に応じた指標選択の重要性

降水の有無に関する予報についても、予測『結果』の評価属性である「精度」の検証が行われている。なお、この場合の「降水あり」の予報とは、「雨」「曇り一時雨」「曇り時々雪」などの降水を含む予報を指し、「降水なし」の予報とは、「晴れ」「曇り」「晴れ時々曇り」などの予報を指す。一方、実績値については、1㎜以上の降水があれば「降水あり」、1㎜未満の降水であれば「降水なし」として集計される。

この予報は、降水の有無という二値分類の問題であるため、誤差ではなく「当否」を用いて予測の「精度」を検証できる。気象庁では、この検証に「適中率」という指標を用いており、適中率が100%であれば、外れのない予測であることを意味する。なお、この適中率には、次の3種類の指標が存在する。

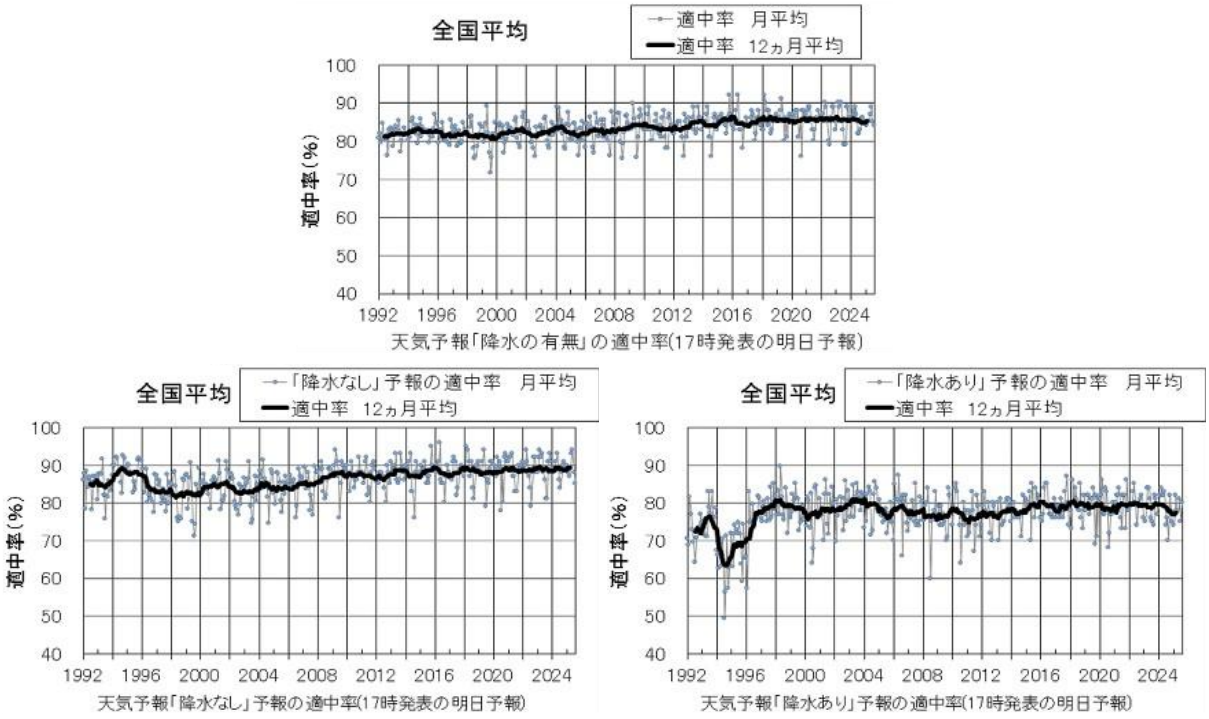
「降水の有無の適中率」…予測が当たった回数を全予測回数で割ったもので、一般的な的中率に相当。
「降水なしの適中率」　…降水なしの予測が当たった回数を降水なしの予測回数で割ったもの
「降水ありの適中率」　…降水ありの予測が当たった回数を降水ありの予測回数で割ったもの

≪図表6≫は、毎日17時に発表される明日（翌日）の予報のうち、降水の有無に関する予報について、上記3種類の適中率の1992年以降の推移を示したものである。これをみると、「降水の有無の適中率」**≪図表6上≫**は緩やかではあるが改善傾向にある。「降水なしの適中率」**≪図表6左下≫**についても、同様の傾向が読み取れる。これに対して、「降水ありの適中率」**≪図表6右下≫**は、1998年以降、ほとんど改善が認められない。また、適中率の水準をみると、「降水なしの適中率」が90%近い値となっているのに対して、「降水ありの適中率」はそれより約10ポイント低い約80%で推移している。

このように、同じ「適中率」であっても、「降水なしの適中率」と「降水ありの適中率」では、「精度」と「改善の履歴・見通し」の双方に明確な違いがみられる。

このような違いが生じる理由は、少なくとも2つ考えられる。1つめは、『結果』の品質評価属性である「不確実性（Uncertainty）」の違いに起因する。一般に、「降水なし」の予測は、広域的な気象場（例：天気図で示されるような大きなスケールの気圧配置）をもとに予測する場合が多いのに対して、「降水あり」の予測では、局地的かつ短時間の気象現象を予測する必要がある。そのため、後者の方が、予測対象に内在する不確実性が大きく、結果として「精度」が相対的に低くなりやすい。

≪図表６≫ 降水の有無の予報の精度検証における３つの「適中率」（全国平均）



出典：気象庁ホームページ「天気予報の精度検証結果」
〈 https://www.data.jma.go.jp/yoho/kensho/yohohyoka_top.html 〉

2 つめは、予測が外れることにより生じる社会的影響の非対称性である。一般に、生活者や多くの事業者（建設、農業、イベント運営、防災関連分野など）にとって、「降水なし」の予報が外れた場合、すなわち晴れの予報で雨が降ったケース（見逃し）の方が、「降水あり」の予報が外れて晴れたケース（空振り）に比べて、受ける被害・影響が大きい傾向がある。そのため、予報を作成する際の判定基準が、「降水なし」の予測の精度を相対的に高める方向に設定されている可能性がある。

したがって、この予測の利用者は、3 種類の適中率の精度の違いをあらかじめ理解して予測を活用することが重要になる。例えば、降水に対応して業務が発生する事業者（例：除雪事業者、雨水処理事業者）にとっては、「降水なしの適中率」よりも、むしろ「降水ありの適中率」の方が重要な指標となる。さらに、適中率は、地方別・月別の数値が公表されている。このため、例えば除雪事業者であれば、年間平均ではなく、冬季における「降水ありの適中率」を利用する方が、リスク評価や意思決定に適している。

この事例は、利用者によって重視すべき評価属性や評価指標が異なること、また「精度」と一口に言っても複数の指標があることを示すものであり、自らのニーズに適合した指標を選択することが、予測を適切に読み解く上で重要になることを示唆している。

b. ベンチマーク予測との比較：複数の品質属性を用いた評価の重要性

上述のとおり、「降水の有無の適中率」は地方別に算定され、それらの値には地域差がある。この場合、適中率の高い地域の予測の方が、信頼性が高い優れた予測といえるのだろうか。

このことを検討するために、降雨日数が少ないことで知られる岡山市と、その逆の秋田市を例にして、「降

水なしの適中率」を考えてみる。年間日数に占める降水なしの日の割合は、岡山市で 75%程度、秋田市では50%程度となっている。このため、仮に予測技術を使わないまま、毎日「降水なし」という予報をし続けたと仮定すると、「降水なしの適中率」は、岡山市では 75%程度、秋田市では 50%程度になる。これが、『結果』の品質評価属性の１つである「スキル」に対応した、ベンチマーク予測の一例である。

あくまで仮定だが、「降水なしの適中率」が岡山市で 90%程度、秋田市で 80%程度であったとすると、ベンチマーク予測との差は、岡山市で 15 ポイント、秋田市で 30 ポイントとなる。したがって、「精度」の観点では岡山市の方が高い評価であったとしても、「スキル」の観点からみると、秋田市の予測の方が高く評価できることになる。

この事例は、『結果』の品質評価において、「精度」だけでなく、「系統誤差」や「スキル」といった異なる評価属性を用いて多角的な評価を試みることが、予測を賢く読み解く上で重要であることを示している。

c.「適中率が高い＝良い予測」ではない理由：指標に惑わされないために

降水の有無に関する予報では、「適中率」だけでなく、「補足率」、「一致率」、「見逃し率」、「空振り率」（※）といった指標についても検証結果が公表されている。予測の「精度」に関して、このように複数の指標が設定されている理由について考えてみる。

上述のとおり、岡山市を例にすると、仮に毎日「降水なし」と予報し続けた場合でも、年間を通せば 75%程度という比較的高い「適中率」が得られる。ところが、実際に雨が降った日を「降水あり」と正しく予報できた割合、すなわち「補足率」を算定すると、その値は 0%となる。すなわち、見かけ上は高い「適中率」であっても、実質的には降水予測として機能していない場合もあることになる。このように、各指標には長所と留意点があり、評価の目的に適合しない指標を用いた場合、予測の信頼性を誤評価するおそれがある。

この事例は、予測『結果』の品質を評価する上で、評価属性や評価指標への理解を一定程度深めることが、予測に惑わされず、予測を的確に読み解く上で不可欠であることを示している。

※それぞれの指標を気象庁の定義に沿って表現すると次のとおり（同一の用語であっても、技術分野によって定義が異なる場合がある）。

- 補足率　：実際が「降水あり」⇒予報も「降水あり」となる頻度（降水ありを言い当てた頻度）
- 一致率　：予報が「降水あり」⇒実際も「降水あり」となる頻度（降水あり予報の適中率）
- 見逃し率：予報が「降水なし」⇒実際は「降水あり」となる頻度（降水なし予報が外れた頻度）
- 空振り率：予報が「降水あり」⇒実際は「降水なし」となる頻度（降水あり予報が外れた頻度）

③「降水確率の予報」の精度

降水確率の予報については、2(3)の③で取り上げたとおり、予測『結果』の評価属性である「予測確率の信頼性(Reliability)」に対応した検証結果が公表されている。**≪図表 7≫**は 24 時間先までの降水確率予報の精度検証の一例である。図の横軸は予測された降水確率（予測確率）であり、縦軸はその予報が出された回数のうち実際に降水が観測された頻度（観測確率）となっている。予測確率と観測確率が完全に一致する場合、プロットは赤の斜線上に整列する。

この図から、次のような傾向を読み取ることができる。

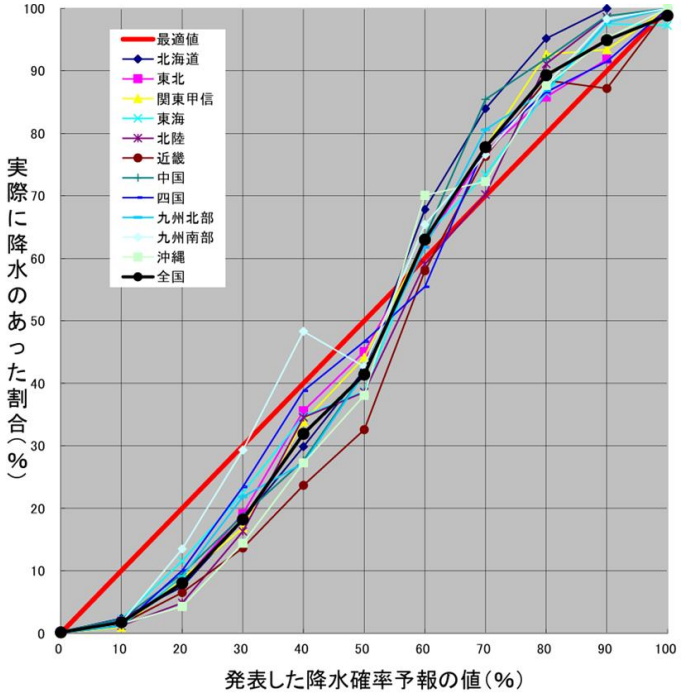
・予測確率が低い場合（10～40%）

⇒「予測確率＞観測確率」となる傾向があり、実際の降水確率は予測確率よりも低くなる傾向がある

・予測確率が高い場合（70%～90%）

⇒「予測確率＜観測確率」となる傾向があり、実際の降水確率は予測確率よりも高くなる傾向がある

≪図表７≫ 24 時間先までの降水確率予報の「予測確率の信頼性」（2025 年 3 月～2025 年 5 月）



出典：気象庁ホームページ「天気予報検証結果」

〈 https://www.data.jma.go.jp/yoho/kensho/HPdata/HPdata2505/fpop_25haru.pdf 〉

このような予測と実績の乖離にみられる体系的な偏りは、決定論的な予測であれば「系統誤差（Bias）」として扱われるが、この場合は、低確率域と高確率域で「予測確率の信頼性が低下している」という形で評価される。利用者は、上記の予測確率を次のように解釈することで、より合理的な意思決定が可能になる。

・予測確率が 10～40%という低い確率の場合

ほぼ降水がないと判断してよく、10%刻みの差を気にする必要性は小さい。

・予測確率が 80%～90%という高い確率の場合

ほぼ降水があると判断してよく、10%刻みの差を気にする必要性は小さい。

・予測確率が 50～70%といった中程度の場合

予報値の差（10%刻み）が相対的に重要な意味をもつため、慎重な解釈が望まれる。ただし、予報値の差の解釈よりも実務上重要なことは、「判断が割れる日だ」という理解であり、そのことが合理的な行動・判断につながると考えられる。

シリーズ①で述べたように、利用者側では、予測の使い方や受け止め方に工夫の余地がある。ここに挙げた例を参考にすると、予測の価値は、予測を実施する側の技術的な要素だけでなく、利用者側の予測の活用スキルにも左右されることがわかる。

この事例は、予測『結果』の品質評価において、評価属性や指標の検証結果の背後にある意味を理解することによって、自らのニーズに適した予測の付き合い方を見いだせる可能性を示唆している。

もっとも、この図に示された予測と実績の関係性は不変ではなく、年・季節・地域の違い、更には予測技術の改善によっても変化し得るため、上記のような解釈は例示したケースに限られる点に注意が必要である。

一般論として、人間は確率的な情報を直感的に二値化して決定論的に解釈する傾向がある。降水確率予報の場合も、50%を境に、それより低ければ「降らない（降水なし）」、高ければ「降る（降水あり）」といった二値的な解釈を無意識的に行う人が、ある程度の割合で存在すると考えられる。

図に示されたような予報と実績との関係性は、そのような人に対して、結果として空振り率（「降水あり」予報が外れる頻度）と見逃し率（「降水なし」予報が外れる頻度）の両方を低く抑えることに成功している（※）。こうした予測特性がどこまで意図的にコントロールされているかは定かでないが、予測を行う側が、利用者の心理的な特性や予測が活用される場面を念頭に置いて予測を設計している可能性がうかがえる。

予測の実施者によるこうした工夫を、利用者が理解して、意思決定に生かすことができれば、予測の有用度がさらに高まる可能性がある。

※単純な二値問題の予測であれば、見逃し率を低下させるために予測の判定基準を変更すると空振り率が上昇し、逆に空振り率を下げると見逃し率が上昇するという、トレードオフの問題が存在する。

（２）竜巻注意情報の精度検証が示唆する「予測の価値の構造」

気象庁では、竜巻等（ダウンバーストなどの激しい突風を含む）に対して注意を呼びかけるため、「竜巻注意情報」を発表している。この情報は、竜巻発生の確度が高まった場合や目撃情報が得られた場合などに発表され、情報の有効期間は発表から約 1 時間とされている。

「適中率」、「補足率」といった評価指標の平成 20 年以降の推移が、ホームページで公表されており、これによれば、竜巻注意情報の実に 9 割ほどが「空振り」という結果に終わっている。

① 評価指標の定義と検証結果

竜巻注意情報に関する主な評価指標は次の 3 つである（以下の指標名や説明は、気象庁の定義をもとにしつつも、わかりやすさという観点から一部改変したものであり、厳密には気象庁の定義を参照のこと）。

・突風適中率

竜巻注意情報の発表数のうち、有効期間内に竜巻等の突風が発生したものの割合（「竜巻等あり」という予報が当たった頻度）。(1)②の例でいうと「降水ありの適中率」に相当する。

・強風適中率

竜巻注意情報の発表数のうち、有効期間内に竜巻等の突風、または対象県内で最大瞬間風速 20m/s 以上が記録されたものの割合（「竜巻等あり」という予報が遠からず当たったとみなせる頻度）。

・突風補足率

実際に発生した突風回数のうち、竜巻注意情報が発表された割合（突風発生を言い当てた頻度）。

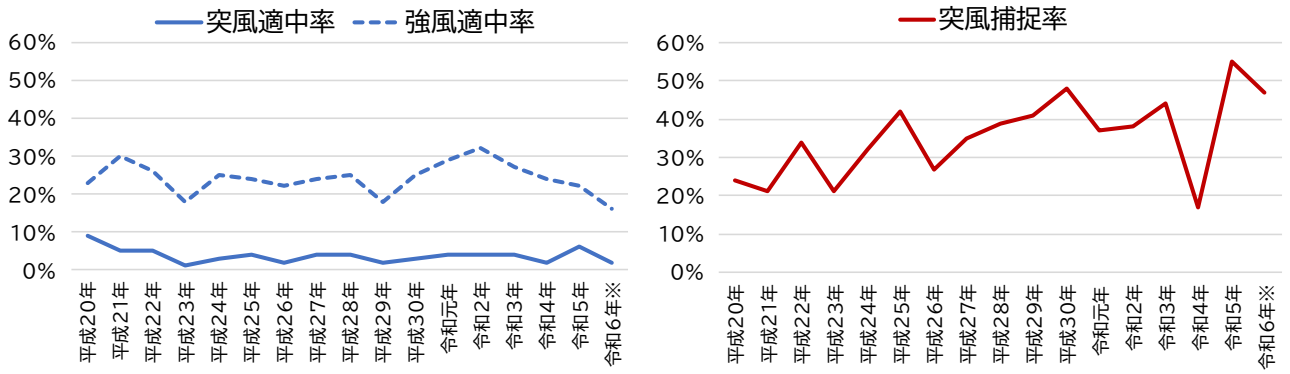
検証結果のうち、突風適中率、強風適中率、突風補足率の推移を**≪図表 8≫**に示す。主な特徴は 3 点ある。

・突風適中率は 1～9%、強風適中率は 16～32%であり、(1)の天気予報の場合に比べて適中率が総じて低い

・適中率の経年的な推移を見る限り、明確な改善傾向が認められない

・突風補足率は 17～55%と年変動は大きいが、緩やかな改善傾向が認められ、近年は 5 割近い水準である

≪図表８≫ 突風適中率、強風適中率、突風補足率の推移（平成 20 年 3 月～令和 6 年 12 月）



※令和 6 年の値は速報値

出典：気象庁ホームページ「竜巻注意情報の精度について　平成 20 年 3 月 26 日から令和 6 年 12 月 31 日までの精度の推移（令和 7 年 1 月 15 日更新）」の掲載データを用いて、当社にてグラフ化

〈 https://www.data.jma.go.jp/tatsumaki/tatsumaki_hyoka_top.html 〉

②空振り率と見逃し率のトレードオフ構造

この検証は二値問題に単純化されているため、ここでの「適中率」は(1)②の c で示した「一致率」と等しく、その裏返しが「空振り率」となる。したがって、突風の空振り率（竜巻注意情報を発表したが、突風は観測されなかった頻度）は、なんと 91～99% に達する。強風までを含めて適中とみなした場合でも、空振り率は 68～84%となる。

一方で、「突風捕捉率」の裏返しである、突風の見逃し率（実際に発生した突風の発生回数のうち、注意情報が発表されなかった頻度）は、近年は 5 割ほどであり、空振り率に比べれば低い水準に抑えられている。

繰り返しになるが、この事例のような単純な二値問題では、見逃し率を下げようとすると、必然的に空振り率が上昇するというトレードオフの関係がある。竜巻等の突風は、人命に関わる災害に直結しやすい現象であることから、突風の見逃し率をできるだけ低い水準に抑えることが重視され（※）、その代償として空振り率が 9 割を超える高水準となっている可能性が強く示唆される。

※(1)②の a では、予測が外れた場合に生じる社会的影響の非対称性に言及し、降水ありを見逃した場合の方が、降水あり予測が空振りした場合に比べて被害・影響が大きい傾向があるとした。突風の場合は、降水の場合に比べて、この非対称性が格段に顕著に現れることになる。

③適中率が低くても予測の価値が認められる要因

この予測は、予測対象の適中率が 1 割を下回る（すなわち 9 割は外れる）にもかかわらず、その意義や必要性を疑問視する声は必ずしも聞こえてこない。この事例は、予測の価値は単純な当否だけでは決まらないことを示している。この予測の価値を支える要因は、次の 5 点に整理できる。

「社会的な価値の大きさ」

竜巻等の突風災害を回避する行動につながる予測は、社会全体として死傷者を減らす効果をもち、社会的価値が大きい。

「利用者にとっての価値の大きさ」

たとえ外れが多くあったとしても、予測が適中した場合には利用者の人命を守る意思決定を支えるという点で、利用者にとっての価値が大きい。

「補足率の高さ（見逃し率の低さ）と改善傾向」

社会的な価値や利用者にとっての価値の源泉は、適中率の高さではなく、突風の発生を見逃さない確率（補足率）の高さにあり、その値は改善基調で推移し、おおむね 5 割まで上昇している。

「予測の空振りに対する社会的受容」

人命に係わる防災・危機管理情報については、多少の空振りは想定の上で運用されるべきという考え方が、社会的に一定程度共有されている。

「予測対象に内在する不確実性の大きさ」

竜巻等の突風は局地的かつ突発的に発生するため、予測の難易度は非常に高い。このため、「精度」の評価では、突風という現象のもつ不確実性の大きさを考慮する必要がある（これは、予測『結果』の品質に関する評価属性である「不確実性（Uncertainty）」に対応）。

④竜巻注意情報に残る課題と本節のまとめ

一方で、この予測の最大の課題は、次の三者のバランスの確保にある。

「9 割超という非常に高い水準の空振り率」

「それに対する社会的な許容度」

「自治体等の防災関係者における実務上の許容度」

空振り率が社会的な許容度を超えた場合には、いわゆるオオカミ少年効果が顕在化するおそれがある。また、防災関係者における許容度を超えた場合には、いわゆる警報慣れや警報疲れによる対応の遅れをまねき、予測本来の機能が発揮されなくなる可能性がある。なお、気象庁の検証結果の公表ページでは、補足率は明示されている一方で、「空振り率」という表現は登場しない。この点からも、社会的な受容とのバランスに一定の配慮が為されている可能性がうかがえる。

この事例は、予測の価値は単純な適中率の高さだけで決まるものではなく、見逃し率の低さ、社会的影響の大きさ、利用者にとっての便益、不確実性の大きさ、そして空振りに対する社会的な許容とのバランスなど、さまざまな要因によって定まることを示している。

（３）台風や豪雨の予測精度検証にみる「予測精度の現状と限界」

台風や豪雨のような大規模被害をもたらす気象現象の予測の精度向上は、気象庁が最も力を注いでいる取り組みの 1 つである。しかし、それでも、3 日先の台風の中心位置の予測には、「大阪－名古屋間」の距離に匹敵するほどの誤差が含まれているなど、課題も多く残されている。

気象庁は、4 つの戦略的方向性のもとに 10 の関連施策等を設定し、毎年、その進捗状況を公表している[14]。戦略的方向性の第 1 には「防災気象情報の的確な提供及び地域の気象防災への貢献」が位置付けられ、そのための関連施策として「台風・豪雨等に係る防災に資する情報の的確な提供」が掲げられている。

当該施策の進捗を測るための業務指標には次の 5 つがある。

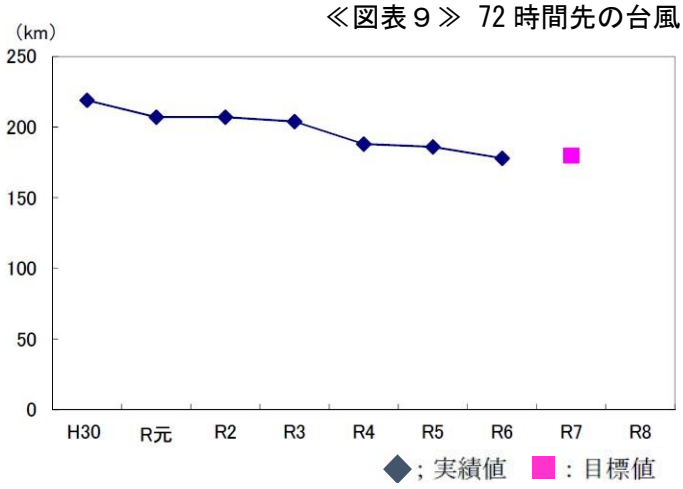
- 台風予報の精度の改善（台風中心位置の予報誤差）
- 線状降水帯情報の改善（①線状降水帯に関する防災気象情報の改善件数、②線状降水帯予測の捕捉率）
- 大雨の予測精度の改善（降水短時間予報の精度）
- 大雨に関する早期注意情報の予測精度の改善（①大雨に関する警報級の可能性[高]の適中率、②大雨に関する警報級の可能性[中]以上の捕捉率）
- 大雪の予測精度の改善（大雪の予測値と実測値の比）

これらの予測は、気象庁による政府予測の中でも、精度向上の最前線に位置していると考えられる。そこで、上記の業務指標の中から、台風中心位置の予測誤差、線状降水帯予測の補足率、大雨に関する警報級の可能性[高]の適中率の 3 つを取り上げ、予測『結果』の品質の現在地に関する情報を整理する。

①台風中心位置の予報誤差

台風中心位置の予報誤差の推移と目標値を**≪図表 9≫**に示す。ここでの予測誤差とは、72 時間先の台風中心位置についての予測位置と実績位置との間の距離であり、台風の発生件数や移動経路の年単位の変動を考慮して、当該年を含めた過去 5 年間分の平均値となっている。

≪図表９≫ 72 時間先の台風中心位置の予報誤差



出典：気象庁業務評価レポート（令和 7 年度版）資料 1
〈 https://www.jma.go.jp/jma/kishou/hyouka/hyouka-report/r07shiryo1.pdf 〉

これをみると、平成 30 年以降、予報誤差は緩やかに減少しており、令和 7 年目標値である 180km を達成しつつあることがわかる。とはいえ、この距離は、東京と静岡市、大阪市と名古屋市、熊本市と鹿児島市といった都市間の距離におおよそ相当する。3 日先の予測とはいえ、都道府県をまたがるほどの大きさの誤差が含まれていることを意味しており、防災担当者はこのことを頭に置いて業務を進める必要がある。また、設定されている目標値自体も、一般的な感覚に比べると控えめな水準にとどまっているようにみえる。

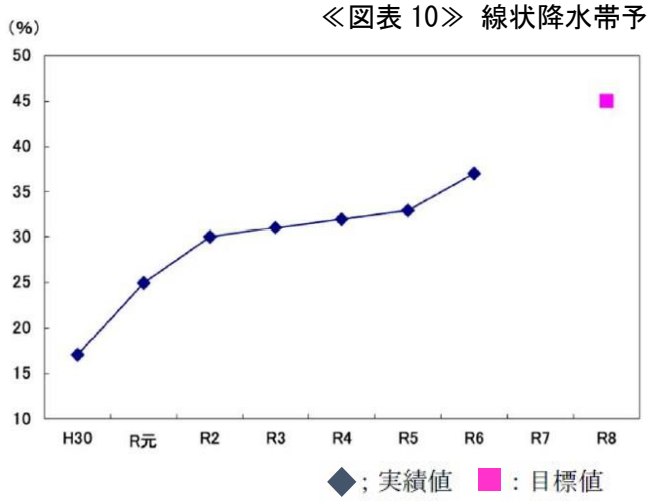
[14] 気象庁業務評価レポート（令和 7 年度版）本文< https://www.jma.go.jp/jma/kishou/hyouka/hyouka-report/r07honbun.pdf >

②線状降水帯予測の捕捉率

線状降水帯予測の捕捉率の推移と目標値を**≪図表 10≫**に示す。ここでの補足率とは、観測により線状降水帯が観測された事例のうち、発生の 15 時間前からの予測において、観測された場所から 100km 以内の範囲で線状降水帯が予測されていた事例の割合とされる。評価対象としては、線状降水帯の発生に年変動があることを考慮して、当該年を含めた過去 5 年間分が集計対象となっている。

これをみると、捕捉率は平成 30 年の 17%から令和 6 年の 37%へと大幅に向上している。しかし、この結果は、線状降水帯の発生という危険性の高い気象現象の 6 割以上を見逃してしまうという予測の現状を示している。前述した突風の見逃し率と比べても、予測性能の面で改善の余地が大きいといえるだろう。

≪図表 10≫ 線状降水帯予測の捕捉率の推移と目標値



出典：気象庁業務評価レポート（令和 7 年度版）資料 1
<https://www.jma.go.jp/jma/kishou/hyouka/hyouka-report/r07shiryo1.pdf >

③大雨に関する警報級の可能性[高]の適中率

「大雨に関する警報級の可能性[高]」とは、全国の気象台が 17 時に発表する早期注意情報のうち、翌日の 6 時から 24 時を対象期間として、大雨に関する警報級の可能性が高いことを示す情報である。このような早期注意情報は、「社会的に大きな影響を与える現象について、可能性が高くなくとも発生のおそれを積極的に伝える」という方針の下、平成 29(2017)年 5 月から提供が開始された[15]。

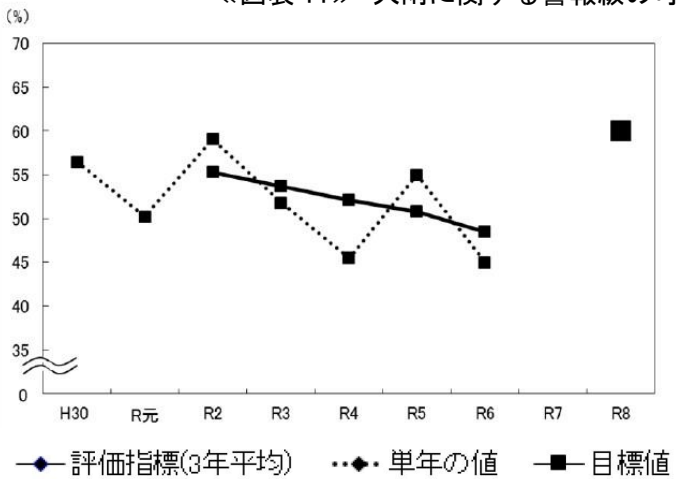
この早期注意情報の適中率の推移と目標値を**≪図表 11≫**に示す。ここでの適中率とは、大雨に関する警報級の可能性[高]を発表した事例のうち、実際に警報の基準に到達した割合を、全国予報区の前 3 年間について平均した値である。これをみると、3 年間の平均がとれるようになった令和 2 年の 55.3%から、令和 6 年の 48.8%にかけて、適中率が低下傾向にあることがわかる。加えて、目標値は適中率 60%以上と設定されており、一般的な感覚と比べて、かなり控えめな水準に設定されているようにみえる。

この予測の場合、実績が警報の基準に達しなかったことをもって、空振りと呼ぶべきかどうかは議論のあるところだが、実績が警報の基準に達しない可能性が 5 割近くあることになる。これを例えると、河川管理者が警報級の大雨に備えて夜間に人員を手配したものの、結果的にそのようなレベルの大雨が降らなかったというケースが、2 回に 1 回は発生することを意味する。

[15] 気象庁業務評価レポート（令和 7 年度版）資料 1< https://www.jma.go.jp/jma/kishou/hyouka/hyouka-report/r07shiryo1.pdf >

夕方の情報発表に伴い、河川管理者をはじめとした国や自治体の関係者が夜間に大掛かりな準備や対応を進めることも想定されるため、社会的な影響も大きい。竜巻等の突風の空振り率に比べれば低い水準とはいえ、警報慣れや警報疲れを回避しつつ予測の信頼性を維持するためにも、予測精度の向上が期待される。

≪図表 11≫ 大雨に関する警報級の可能性[高]の適中率の推移と目標値



出典：気象庁業務評価レポート（令和 7 年度版）資料 1
<https://www.jma.go.jp/jma/kishou/hyouka/hyouka-report/r07shiryo1.pdf >

④台風や豪雨に関する予測の現状と限界のまとめ

予測精度向上の最前線に位置づけられる 3 つの予測事例から、それらの構造的な課題を含めた予測の現在地を次のように整理することができる。

「台風・豪雨に関する予測性能の限界」（『結果』の現状）

台風中心位置の予測誤差は主要都市間距離に相当するスケールであり、線状降水帯の約 6 割は見逃され、大雨の警報級の可能性に関する予測の空振りが 5 割近くに及んでいる。予測結果には未だ大きな不確実性が含まれており、現在の先端技術をもってしても、台風や豪雨に関する予測『結果』に示された性能には明確な限界が存在している。

「台風・豪雨に関する予測の困難性」（『手法』の現状）

また、設定されている目標値そのものが、一般的な感覚に比べて控えめな水準にとどまっていることや、目に見えた改善が実現されていない指標も存在している。このことは、台風や線状降水帯といった現象が、物理的に極めて予測困難な対象であり、予測『手法』において限界が存在することを示している。

「緩やかな精度改善の推移」（『プロセス』の現状）

その一方で、線状降水帯の捕捉率にみられるように、一部の指標では着実な改善が認められている。現状の予測精度は、飛躍的な向上は認められていないが、着実に向上している段階にあると考えられる。

「不確実性を前提とした意思決定の重要性」（予測の信頼性を踏まえた利用のあり方）

したがって、台風や豪雨といった不確実性が大きい予測を意思決定に活用する場合には、予測を「当たるか外れるか」で単純に評価するのではなく、一定の見逃しや空振りを前提として予測を読み解くことが重要になる。その上で、どの程度の不確実性を許容し、どの段階でどのような判断を行うのかといった、利用者側の予測の運用方針・設計が、予測を意思決定に活かす上で欠かせない。

（4）気象予報の精度検証に関するまとめ

ここまで、精度検証可能な政府予測の代表例である気象予報を取り上げ、その検証結果から予測の現状等を整理してきた。

(1)では、気温、降水の有無、降水確率といった日常的な天気予報を題材として、「精度」「信頼性」「スキル」「補足率」「空振り率」など複数の評価指標が存在していること、また、同じ「適中率」であっても、利用目的や地域特性によって評価の意味合いが大きく異なることを示した。その結果、予測は単一の数値だけで良し悪しを判断するべきではなく、評価軸や指標の特性の違いを踏まえて多面的に読み解く必要があるという知見が得られた。

(2)では、竜巻注意情報を例に、適中率が1割未満という極めて低い数値であっても、見逃し率の低さ、社会的影響の大きさ、利用者にとっての便益、不確実性の大きさ、空振りに対する社会的許容といった要因の組合せによって、予測の価値が十分に確保されることを示した。また、空振り率・社会的許容度・防災実務上の許容度の三者のバランスが、予測の有効性を左右する課題として残されていることも明らかになった。

(3)では、気象庁が精度向上の最前線として位置づける台風、線状降水帯、警報級の大雨に関する予測を対象に、予測『結果』・『手法』・『プロセス』の現状を整理した。その結果、最先端技術をもってしても、依然として大きな不確実性が残されており、飛躍的な精度向上は容易には期待ではない一方で、線状降水帯の捕捉率に見られるように、精度の改善が着実に進捗している段階にあることも確認された。

以上をまとめると、気象予報は「当たるか外れるか」で単純に評価すべき予測ではないといえる。高い不確実性を構造的に内包する情報として、予測の『結果』・『手法』・『プロセス』の各段階に分けて評価し、さらには『利用』の段階までを含めて読み解くことが望ましい政府予測だといえる。

予測の実施者である気象庁には、予測精度の更なる向上に向けた技術開発の継続が望まれる。一方で、大きな不確実性を内包する予測の利用では、見逃しや空振りを前提に、不確実性をどこまで許容し、どの段階で判断を行うかといった、予測との向き合い方がその実質的な価値を左右する。気象予測に限らず多くの社会・経済予測が大きな不確実性を伴う以上、利用者には、予測が外れた場合も含めた判断・行動を想定した対応が重要となる。

4. おわりに～本シリーズの総括～

本シリーズの目的は、政府予測の網羅的・体系的な整理と分析を通じて、政府の予測力の一端を明らかにし、予測を活用しつつも惑わされないための実務的な視点を提示することにあった。

シリーズ①では、予測と人間の知的活動や知能との本質的な関わりに触れつつ、予測とは何か、成功例とは言い難い政府予測の具体例、予測の価値の源泉、予測を読み解くための8つの視点などの基本的な事項を整理した。特に、予測の真価が“未来を言い当てること”ではなく“意思決定に役立つこと”にあること、利用者目線では“正確な予測”より“役立つ予測”が重要であることを示した。

シリーズ②では、政府予測とフォーサイト、長期戦略等との関係を整理し、8つの視点を用いて政府予測

の全体像を分析した。その結果、政府予測319件の9割が継続的に実施されており、約半数が府省庁設置法以外の法的な根拠をもつことを明らかにした。また、政府予測の平均像が「社会・経済を対象にした、行政と企業向けの、中期～長期の予測」であり、「数理モデル型」が最も多く、不確実性の扱いでは「特定型」と「シナリオ型」が拮抗することを示した。さらに、「経済・産業分野」の政府予測44件の予測リストを提示し、その結果、「中期～長期」の「シナリオ型」予測が多くを占めることから、約3分の2 が精度検証困難に分類されることを示した。

シリーズ③では、「社会・医療・福祉」「食料・農林水産」「国土・インフラ」「環境・エネルギー」「防災・危機管理」の5分野275件の予測リストと特徴的な事例を取り上げ、予測がどのように作成され、政策に活用されているのかを分析した。その結果、政府予測の多くが「情報提供サービス」にとどまらず、政策形成や制度運用を支える基盤的な「情報インフラ」として機能していることを明らかにした。このことは、政府予測の質・量の改善や、予測と政策決定者との関係強化が、政策の質の向上に寄与することを意味する。さらに、COVID-19の数理モデル、SPEEDI、地震・津波の予測などを通じて、科学的予測を政策決定プロセスにどう位置付け、どのように社会に提示するべきかという観点からの課題を整理した。

最終報となる本稿では、予測が外れたように見えても「外れ」と断定できない構造的性質に触れ、政府予測の多くが当否や精度で客観評価できないことを指摘した。そこで、予測のプロセス・手法・結果から成る「予測の信頼度」という枠組みを用いた重層的な評価を提案し、スキル、前提条件、透明性、改善履歴といった「精度」以外の評価属性を示した。また、精度検証が困難な予測例として、超長期予測、シナリオ型予測、低頻度事象の確率型予測、自己否定／自己実現型予測の 4 類型を挙げ、それぞれの特性と事例を整理した。さらに、技術や社会がベイズ的判断を前提とする方向に変化する中で、それに人間の確率理解が追いついておらず、政府予測のベイズ的要素が高まるほど、政府予測を読み解く力の重要性も増すことを指摘した。

本稿の後半では、精度検証が可能な気象予報を題材に、予測の現状と限界を整理するとともに、予測を読み解く際に留意すべき点を整理した。具体的には、評価指標を理解した上で多面的に予測を読み解く必要があること、適中率が低くても予測の価値が確保されている実例が存在すること、さらに、予測には不確実性を排除しきれない本質的な限界があることを示した。

シリーズ全体の結論を一言で表現すると、「予測の価値は、予測の前提や不確実性を理解した上で、その結果をどのように解釈し、意思決定につなげるかという利用者側の対応に依存する」ということであろう。とりわけ、不確実性の大きい予測では、予測が外れた場合も含めた判断や行動が求められる。こうした結論は、シリーズ①で指摘した「予測との向き合い方や活用スキルの重要性」とも整合するものであり、予測を読み解くための「予測リテラシー」の向上が重要になる。

本シリーズが、予測との向き合い方を見直す契機となり、政策運営や企業経営の高度化・安定化に向けて、より良い判断と選択をもたらす一助となれば幸いである。

【参考１】ブライアスコアとブライアスキルスコア

① ブライアスコア

確率型の予測において、図表 2 の「a.精度」に関する最も基礎的かつ代表的な指標がブライアスコアである。ブライアスコアは、予測確率と観測結果との差を二乗して平均した値であり、値が 0 に近いほど予測と実績が一致しており、精度が高いことを示す。値が大きくなるほど、予測と実績の乖離が大きい。

例として、3 人の予測者が降水確率を 2 回予測し、結果が次のようになった場合を考える。

＜予測者 A＞
1 回目：降水確率 90%と予測し、雨が降った
2 回目：降水確率 10%と予測し、雨が降らなかった
この場合のブライアスコアは ⇒ $((0.90-1)^2+(0.10-0)^2)/2=0.01$

＜予測者 B＞
1 回目：降水確率 70%と予測し、雨が降った
2 回目：降水確率 30%と予測し、雨が降らなかった
この場合のブライアスコアは ⇒ $((0.70-1)^2+(0.30-0)^2)/2=0.09$

＜予測者 C＞
1 回目：降水確率 50%と予測し、雨が降った
2 回目：降水確率 50%と予測し、雨が降らなかった
この場合のブライアスコアは ⇒ $((0.50-1)^2+(0.50-0)^2)/2=0.25$

予測者 A と B はどちらも「当たった」ようにみえるが、利用者の直感と同様に、より確信度の高い予測を行った A の方が、ブライアスコアは 0 に近く、「精度」が高いと評価される。A の予測は相対的に強気の予測、B の予測は弱気または慎重な予測という受け止め方もできる。

一方、予測者 C は、いわゆるリスク回避的な予測である。「外れた」とは断定し難い反面、予測に期待される「意思決定に役立つ情報」とも言い難く、そのことはブライアスコアにも表れている。

ブライアスコアは、単純な計算式で定義される一方で、図表 2 の「d.予測確率の信頼性」「e.分解能」「g.不確実性」という 3 つの評価属性に分解して説明できるという興味深い特徴をもっている。このため、気象予測に限らず、社会科学分野における確率予測やリスク評価、機械学習でのモデル評価など、さまざまな分野で活用されている。

シリーズ①では、社会・経済の動向等に関する予測について、数百名の専門家による予測精度の平均が「あてずっぽうの予測と大差なかった」という結果を紹介したが、そこでの精度評価にもブライアスコアが使われている[16]。「年内に紛争が発生する／しない」という二値問題ではなく、「年内に紛争が発生する確率は○%である」という形で予測を行うことで、単純な的中率では捉えることができない「予測の精度」を、定量的に評価することができる。

② ブライアスキルスコア

ブライアスキルスコアとは、図表 2 の「c.スキル」に関する指標であり、ある確率予測がベンチマークとなる予測と比べてどの程度優れているかを評価するものである。計算式は次のとおり。

ブライアスキルスコア＝1－(評価する予測のブライアスコア)／(ベンチマーク予測のブライアスコア)

この値が 1 に近いほどベンチマーク予測よりも優れた予測であることを示し、0 であれば同等、マイナスの値の場合はベンチマーク予測よりも劣っていることを意味する。このため、ブライアスキルスコアは、確率予測が実務的に「使う価値のある予測かどうか」を判断する際に有用な指標となる。

[16] フィリップ・E・テトロック & ダン・ガードナー『超予測力 不確実な時間の先を読む 10 カ条』土方奈美訳, 早川書房, 2016 年, p.92

【参考２】政府予測の類型化に関して

シリーズ①では予測の類型化に資する 8 つの視点（分析軸）を提示し、シリーズ②・③では、それらを用いて政府予測全体と政策分野別の特徴・特性を整理した。当初は、この分析軸に基づいて政府予測の類型化に取り組むことを予定していたが、紙幅の制約もあり、別の機会に譲ることとした。ここでは、異なる分析軸同士の関係性を示す一例として、予測の時間軸をベースにした集計結果と、そこから読み取れることを、**≪参考図表≫**に示した。図中の右側に整理した内容自体は、予測に対する一般的なイメージと大きな違いはないようにも思われるが、政府予測 319 件に対する分析結果を通じて、そのことを改めて確認することができる。

≪参考図表≫ 政府予測件数の“時間軸×ターゲット層”ヒートマップ

表の見方
・図中の数値は該当する予測件数を表す
・配色は、0→紺 55→緑 110→黄色

		超短期	短期	中期	長期	超長期	時間軸の加重平均スコア
全政府予測 319件	行政向け	43	20	53	100	108	3.6
	企業向け	41	20	52	69	70	3.4
	生活者向け	27	7	4	2	6	2.0
予測対象 「社会・経済」の予測221件	行政向け	4	4	41	97	75	4.1
	企業向け	4	4	40	67	43	3.9
	生活者向け	0	3	4	2	5	3.6
予測対象 「技術・産業」の予測90件	行政向け	0	4	37	28	21	3.7
	企業向け	0	4	37	25	17	3.7
	生活者向け	0	2	0	1	0	2.7
予測対象 「自然現象」の予測113件	行政向け	42	13	12	5	42	2.9
	企業向け	40	13	12	1	30	2.7
	生活者向け	27	2	0	1	3	1.5
不確実性の扱い方 「特定型」の予測221件	行政向け	35	12	49	45	33	3.2
	企業向け	33	12	49	24	22	2.9
	生活者向け	24	5	4	0	0	1.4
不確実性の扱い方 「確率型」の予測90件	行政向け	5	4	2	2	17	3.7
	企業向け	5	4	2	1	16	3.7
	生活者向け	4	2	0	0	1	1.9
不確実性の扱い方 「シナリオ型」の予測113件	行政向け	6	2	2	55	68	4.3
	企業向け	6	2	1	45	42	4.2
	生活者向け	2	0	0	2	5	3.9

行政向けと企業向けの政府予測
⇒長期～超長期の予測が多い
生活者向けの政府予測
⇒超短期の予測が多い

「社会・経済」の政府予測
⇒長期の予測が多い
「技術・産業」の政府予測
⇒中期～長期の予測が多い
「自然現象」の政府予測
⇒超短期と超長期の予測が多い
時間軸の重心位置が最も
短期なのは「自然現象」の予測
長期なのは「社会・経済」の予測

「特定型」の政府予測
⇒中期の予測が多い
「確率型」の政府予測
⇒超長期の予測が多い
「シナリオ型」の政府予測
⇒長期～超長期の予測が多い
時間軸の重心位置が最も
短期なのは「特定型」の予測
長期なのは「シナリオ型」の予測

※予測対象の分類には重複があるため、合計は予測総数と一致しない。
※時間軸の加重平均スコアとは、超短期を1、短期を2、…、超長期を5と数値化し、件数と掛け合わせた場合の平均値。予測時間軸の重心の位置を表す。

出典：各種資料から当社作成