第３章　今後の分析業務に際しての課題と推進の方向性

第２章で詳述した今回分析は、昨年度の調査で課題とされた点（１．２章記載）を踏まえ、今回追加された 2014 年のデータも活用して実施した。ビッグデータによる経済・景気分析への活用という点では、昨年度、及び今年度の景気ウォッチャー調査による分析の結果は非常に有望、かつ将来性のある分析であると考える。

特に今年度は昨年度の分析における課題にある程度対処し、数々の改善策を導入した結果、景気ウォッチャー調査 DI 値の予測・推計精度の向上を図ることができた。しかしながら、今回の分析結果を検証するとさらなる改善が望まれる点もあれば、新たな課題も浮上している。

本章の以下各項においては、今後この分析業務をさらに発展、推進させていく際の課題や将来の方向性について記述する。

３．１　検証結果、検討委員会から得られた分析業務の課題

第２章から得られた知見や評価、さらには検討委員会における委員からの指摘を踏まえると、景気・経済に関する分析にビッグデータ[1]を活用することの課題は、大要以下のようにまとめることができよう。

３．１－１　検討委員会における主な指摘事項

第２章における分析結果を踏まえ、検討委員会において委員から頂いた主な指摘事項は大きく五つに分類される。以下、分類毎に指摘内容を記載する。

a.有意表現（特徴表現）の抽象化・グループ化の手法について

・有意表現を特徴表現から抽象化するにあたり、抽象化の括り方にレベル感の差があるように見受けられる。例えば、@食品@のように非常に多数の単語をひとつの上位概念にまとめたものもあれば、少ない単語数から構成されるものもある。グループ化に際しては概念の上下関係、階層構造等も考慮し、分析に使用するために最適な抽象化をすることがポイントと考える。

[1]一般に「ビッグデータ」というとデータそれ自体よりも「データと技術とビジネス（あるいは公共面での便益など）を統合した一連の価値創造活動、あるいはその手法」を示すことが多いが、本章では「世間一般に存在する大量のデータ」のことを指すこととする。

・また、地名の抽象化についても、アメリカのように単独の国が日本の景気に大きく影響を与えるものについては単独の語のまま残し、影響が小さいと考えられるマイナーな国については抽象化でまとめるという配慮も必要と考える。まとめる、まとめないの判断基準としては、例えば単語(地名)単位でカイ二乗を計算し、有意であればそのまま使用し、データ数が少なく有意でなければ、今回のようなまとめ方を行うなどの判別をしてはどうか。

・例えば「@需要@」という抽象化概念において、「消費税増税前なので駆け込み需要が発生した」は景気がよいことを意味するが、一方で「消費税増税前に比べて」のような表現があり、これは景気が悪い方を意味する。「@需要@」についてはこれら両方の意味が一緒になって抽象化されてしまった模様である。「@購買@」にも同様の傾向がありそうである。このように抽象化に限界があることを考慮した上で取り組むことが課題である。

・抽象化、階層化の方法については大きくは２つある。今回の場合は、人の知識に基づいて手でまとめ上げたということだが、もうひとつの方法は過去における共起表現[2]を使って自動的にまとめ上げる方法がある。そのデータソースとしてはウィキペディアを利用することが考えられる。これはどちらがよいかは難しい。景気指標を誰が見るかであるが、人の知識があった方が解釈性は高まり、より正確性が増すと思われる。人が分析する為のものならば、経済的な用語集を作成した方がよいと考える。なお、日本経済新聞社のシソーラスなどの情報を活用することが可能と考える。

b.分析対象データの採用・絞り込みの手法について

・景気ウォッチャーの個票は５値のデータがあるが、真ん中の「変わらない」のデータを抜いて２値にするというのは、せっかくの情報を捨てているのではという懸念が残る。

・また、「変わらない」のコメントの中に実はプラスやマイナスが有意に出てくる単語もあれば、「変わらない」という意味が有意に出てくる単語もあるのではないかと思われるので、その点をチェックしてみることも有用かもしれない。

---

[2] 任意の文書や文において、異なる２つの文字列が同時に出現することを共起と呼ぶ。
例えば、「消費税が増税された」という文において、消費税と増税は共起の関係にあり、
本報告では、このペアを共起表現としている。

・今回分析においては、景気の現状に対するコメントと先行きに対するコメントの両方を合算して特徴表現の作成を行っているが、特徴表現は現状に対するものと先行きに対するものとでは相違がみられると考える。従い、DI 値との相関を考慮する際には現状と先行きのそれぞれで特徴表現のセットの使い分けを行うことが必要と考える。

c.有意表現の分類・整理の手法について

・同じ単語でも、時期によってポジティブ／ネガティブが変化するものがあり（たとえば「増税前」)、これに応じて DI 値の変化との関係が確認できるといい。一方で、単語数が増加傾向にあるものはいつもポジティブになっているというような単語が確認できるとよく、このようなデータが多くあると望ましい。
前半の部分で学習して、後半の部分で当てにいくときには、「増税前」という単語は景気がよいという意味で効いてくるわけで、それで当てにいくのでどうしても外れることになってしまう。そういうことがないようにするには、時間的な依存性がないような単語を説明変数として使っていくということになる。

・可能であれば、テキストデータからポジ／ネガのインデックスを作成し、その原因も集約して出力できれば使いやすいと思われる。
例えばであるが、いくつかの経済ニュース配信社では、経済ニュースにポジ／ネガインデックスを付けて配信している。このように既に提供されているデータやデータの配信方法を調査し、参考にすることもポイントである。

・業種別の分類については、製造業と非製造業などの大まかな括り方で分類・分析できる可能性はあるかもしれない。

d.有意表現の分析結果の今後の活用法について

・現在の景気ウォッチャー調査では、毎月のデータを読んでコメントを EXCEL でカウントして人手で集計している。今回の調査研究の報告内容から、機械による解析・分析は活用することができると考える。一案としては、景気ウォッチャー調査結果はテキスト情報に答えが付いており、答え合わせができることから、まずテキスト情報をもとに景気の良し悪しを意味する有意表現のリストやモデルを構築し、これをベースとしてツイッター等のソーシャル・ネットワーキング・

サービス[3]（以下、ＳＮＳとする）のデータと突き合わせることでタイムリーな経済・景気分析に活用していくことを視野に入れるべきである。

・SNS 対応が不可能ならば、新聞の記事データを全部かけるとか１日のＴＶ放送をかけて、デイリーでどのような情報が出るかなどとの関連をみることには興味がある。

・今回の調査研究で作成した辞書等を内閣府が公開している景気個票辞書として公開し、有効活用を可能として頂きたい。将来的には、内閣府として景気指標に対応したシソーラスの公開が課題と考える。

e.説明変数（有意表現）やモデルの頑健性を確認するためのアイデア

・過学習やデータ不足の問題を解決するアイデアのひとつとしては、現在 DI 値のみを目的変数として使用しているが、GDP などの他の目的変数を増やして、それらの推計・予測分析の結果共通してできた説明変数は意味があると考える。また、DI 値の推計では説明変数として出てきたが、GDP の推計では出てこないというような特殊性を見つけるなど、説明変数の階層化をすることも検討ポイントであると考える。

・物価が目的変数に使えるのではないかと考える。マーケットは、DI よりも物価の方が結びつきは強い。

・推計式の構築に際し、異種混合学習で枝分かれする部分（門関数[4]）の条件を、過去の DI 値で制限することは可能と考えられる。そうすれば、過去の DI 値を説明変数として使用することによる弊害を回避できる他、推計式の納得性の向上にもつながると考えられる。

上記に指摘された内容を念頭に置いて今後の分析の枠組み、手法、精度の改善を図ることが望まれる。

[3]ソーシャル・ネットワーキング・サービス（英: Social Networking Service、SNS）とは、インターネット上の交流を通して社会的ネットワーク（ソーシャル・ネットワーク）を構築するサービス。

[4] 異種混合学習において、利用するモデルを切り替えるための条件式を門関数と呼ぶ。

３．２　今後の分析業務推進の方向性と将来像

検討委員会における上記の指摘事項を踏まえ、景気ウォッチャー調査を活用して景気・経済に関する分析を行うにあたり、着眼すべき課題は下記と考えられる。

（１） 分析精度向上に向けた取り組み

（２） 今後の分析業務を普及・拡充するための課題・取り組み

３．２－１　分析精度向上に向けた取り組み

景気ウォッチャー調査の個票データを利用して経済や景気の動向を把握するには、さらなる分析精度の向上が必要であり、今回の調査ではシソーラスの有効性が確認された。しかし今回分析においてはシソーラスの要改善点も指摘されており、整備については下記の点を考慮し経済や景気動向の分析目的に適応した内容にしていく必要がある。

① 経済分析に適した語彙体系の整備

経済を分析する上での必要十分な語彙体系を整備することが望ましい。

本調査研究における分析において、単語単位ではなく「増加傾向を示す表現」などフレーズ単位での抽象化が重要であることが示唆されている。そのため、一般的なシソーラスに加え、フレーズを含めた語彙体系を構築していくことが必要である。

② 新しく出現する語彙への対応

新しい単語は、日々生み出されており、それらが経済指標と相関を示すこともある。そのため、コメントに出現した語彙、シソーラス化を試みた語彙、新しく出現した語彙を一元的に管理し、新しい単語の抽象化が必要かどうか、必要の場合は語彙体系のどこに属するべきかを容易に判断できる仕組みが必要となる。

検討委員会において指摘された、整備後のシソーラスの公開についても今後検討することが必要と考える。その際、シソーラスの整備は、今回の調査研究で得られた知見のみをベースに整備するのではなく、公開されているシソーラスの活用・援用も踏まえた検討が必要である。また、２．３章に記載した通り、検討委員会で挙げられた以外の予測精度向上施策の可能性もある。

一方、分析精度の向上にはシソーラスの整備のみでなく、データの選別・取り込み方や推計式の作成手法の改善も重要なポイントである。今回は、景気ウォッチャーDI 値を二値で抽出した有意表現から推計した。しかし、他にも有意表現の中から良い答えと

悪い答えの比率を算出し、DI を推計する方法[5]などがある。これには各二値で抽出した言葉に対する重みの説明が複雑になるなどの課題もあるが、良い／悪いに有意表現を分けた上でマトリックスとして有意表現を整理するなどの方法はあるかもしれない。
以上のような点を考慮しつつ推計式の精度改善に向けた取り組みが求められる。

また、有意表現の選定に際しては、昨年度は LIBLINEAR[6]というオープンソースの判別器を用いて、景気が良いコメントと景気が悪いコメントを判別することによって得られる各単語やフレーズに対する回帰係数を用いた。しかしながら、この回帰係数の信頼性が測りづらいという難点があった。このため今年度は、カイ二乗分析が有意性の検定が容易であることから、カイ二乗分析を採用した。

しかしカイ二乗分析による検定を行う場合、今回分析においては景気が「良い」「悪い」の 2 通りでしか有意表現を分類していないため、仮に景気が非常に良くなったことを意味する有意表現と、景気が少しだけ良くなったことを意味する有意表現の 2 つがあった場合、その 2 つが同じカイ二乗値をとる可能性がある。このような課題についても検討が必要である。

３．２－２　今後の分析業務を普及・拡充するための課題・取り組み

現在、内閣府が公表している景気ウォッチャー調査の結果は、

① 調査結果（抜粋）
② 調査結果（全体版）
③ 景気判断理由（現状）
④ 景気判断理由（先行き）
⑤ 統計表

の５種類の情報が①～④は pdf 形式、⑤は Excel 形式で公開されているが、今後は③④についても CSV 形式等での公開により、政府が公表するデータを一般の産学関係者がより利用しやすい形で提供する、いわゆる「オープンデータ」化への動きが加速されることが望ましい。

---

[5] 例えば、ある月において「景気が良い」を意味する有意表現の登場数が 100 個、「景気が悪い」を意味する有意表現の登場数が 60 個あった場合、DI 値の推計値として 100/(100+60)=0.625 の計算式をもとに 62.5、と算出することを基本的な考え方とする方法。有意表現ごとに違った重みをかけて計算をする、より複雑な方法も考えられる。

[6] 台湾国立大学の Chih-Jen Lin 教授のチームが公開しているオープンソースの機械学習パッケージ。C++で書かれたライブラリと、その機能を使って機械学習と分類・回帰を行うコマンドラインユーティリティが含まれている。

検討委員会においてもオープンデータ化への取り組み強化についての指摘があったが、その端緒として景気ウォッチャー調査関連のデータが利用されることは、景気分析における各所からの知見をさらに得ることにつながることが期待される。３．２－１に記述した分析精度向上のための取り組みについても、本分析業務の他にも類似の研究・分析を行った事例があれば、その結果が精度改善に役立つ可能性がある。政府部門、学術分野、民間部門の全てが研究・分析できる土台を広げるためにも、データの利用拡大は重要である。

また、景気ウォッチャー調査のデータのみでなくツイッター等のＳＮＳのデータを活用することも検討する価値があると考える。ＳＮＳで発信される情報は世評や世論が反映されていると認識されており、このため景気や経済を分析するためのデータの一つとして価値があろう。今回分析業務等をきっかけとしてこの検討を深めることは今後重要度を増してくると考えられる。

ただし、ＳＮＳ、あるいはＷｅｂやメディアから公開されているデータを政府機関が利用することは、実績の不足、公開情報を利用した分析の妥当性の検証、公開情報の入手方法等の観点から更なる検討を要するものと考えられる。また、経済、景気の分析目的で同データを入手、利用する際の価格については、既に学術・研究目的での利用実績がある程度進んでいる[7]現状を踏まえ、利用目的、費用対効果の観点等に照らして妥当かつ適正な水準を確保するよう努めることが必要である。さらにはデータの取り扱いに関しては個人情報の漏えいの恐れをなくすためのルール構築など、一般の安心感を確保する枠組みを作っていくことが、分析主体による将来的なデータ利用の自由度、可用性を広げることにつながるものと考えられる。

（以上）

[7] 例えば、東日本大震災前後における Twitter のツイート数の変化から、Twitter ユーザー間のつながりの変化を分析した学術研究がある。（平成 25 年度 内閣府委託調査「東日本大震災後の日本経済の産業構造・景気循環分析」報告書 p.12-13 参照）