第１章　調査の概要

１．１　調査目的

東日本大震災は日本経済に大きなショックをもたらした。大震災を踏まえた適切な経済財政運営及び被災地復興に資するため、大震災の影響によって産業・雇用・人口構造に大きな変化のみられた地方経済に関する情報を的確・迅速に収集し、景気動向の調査・分析を行うことが求められている。

近年、インターネットのサイト検索語句の変化や被災地立地企業の取引先変化といったデータを使用してこうした構造変化を探るいわゆるビッグデータ業務が盛んに行われているが、経済構造や景気局面の変化に対する研究としてはまだ十分ではない。

こうした中、景気ウォッチャー調査については震災直後に実施された 2011 年 3 月調査においても東北地方の回答率が９割を超えるなど、ショックに対して安定的にデータを得られる調査であると評価されている。

このため、内閣府では 2013 年度に、2010 年から 2013 年の景気ウォッチャー調査の個票と各種経済指標の関係変化などを調査・分析する調査研究を実施した。調査によって、景気ウォッチャー調査の個票と各種経済指標の変化に関係があることや、東日本大震災によって景気ウォッチャー調査のコメントに含まれることばの変化があることがわかった。

今年度においては、2013 年度調査の結果の解釈を深化させるために、以下のテーマにつき、景気ウォッチャー調査の個票と各種経済指標の関係変化などを調査・分析する調査研究を実施する。

- 景気ウォッチャー調査のコメントによる各種経済指標の予測精度を改善する方法の検討と検証
- 景気ウォッチャー調査のコメントにおける各種経済指標に対する関係性や東日本大震災による変化について、地域ごとに差があるかの検証

また、本年度作業の実施に際して直近の状況を反映させるために 2014 年の景気ウォッチャーの個票データを追加する必要があるが、本年４月の消費税率引上げという大きな制度変更が実施されたことから、税率引上げが景気ウォッチャーのコメントに含まれることばに対する影響についても同時に検証する。

１．２　昨年度の分析業務における指摘事項と課題

2013 年度に実施された調査・分析業務（「東日本大震災後の日本経済の産業構造・景気循環分析」平成 25 年度内閣府委託調査）の結果に対し、検討委員会において委員から頂いた主な指摘事項は以下の通りであった。

・予測式（推計式）作成の際に、オーバーフィット[1]を回避するためにデータを地域別に分けるなどの工夫をしていたが、根本的には 400 という有意表現[2]の数が多すぎることがオーバーフィットの原因だと思う。

・例えば「タバコ」や「魚」など個別の単語を説明変数にしているが、単語のグループ化をして、例えば数十種類ぐらいの「生鮮食料品」という単語グループの回数の合計なり、回数を説明変数とするという、説明変数のほうを減らしてオーバーフィットを防ぐほうが、有望ではないかと思う。その方が、このビッグデータ景気動向指数のテキストの分析の良さが出ると思う。

・毎月の景気ウォッチャー調査の分析は、今は手作業でやっておられると思うが、例えば「ビッグデータ分析から見た今月の頻出語句」というような付録を付けるだけでも、世の中に「そういうことができるんだな」という、とっかかりを提供することによって、将来的にはこういうものに関する研究が増えていくような土台を作っておくということもいいと思う。

・例えば、「消費税」と「あるネガティブな文脈」、もしくは「ポジティブな文脈」で、一緒に出てきそうな単語は何かというのを示すだけで、例えば最初、「消費税」が例えば「駆け込み」だけと一緒に出てきたな、今度は「安売り」とかそういう方向へだんだんフォーカスがシフトしたというのが見える可能性がある。一緒に出てくる頻度を時間的に見せるだけでも、実は一般的な着目の変化を図式化する可能性はできると思う。

・キーワードの分析に関しては、もう少しハイブリッド的な分析ができるとよいという

---

[1] 学習データとして用いる入出力関係に過度に適合してしまい、学習に用いなかったデータに対する推定精度が悪くなる現象を指す。過学習と同意語。

[2] ある表現が、「景気が良いコメント」と「景気が悪いコメント」の集合において、一方にのみ存在することをカイ二乗値で数値化した場合、その値の上位のものを有意表現と呼ぶ。本報告では上位４００個を対象とした分析を行っている。

印象。単語の語数を拾うことはもちろん大事だが、どういう文脈で出てきているかについて、例えばある単語の後に否定的な言葉がついているとこれは関係無いということになるので、単語とその前後の文章を機械にまず書きださせて、それを人間が見て、こういう場合には使えない、こういう場合には使えるというようなことをチェックしながら高度な方法論が開発できる余地があるのではないかと思う。

・ビッグデータ分析においてデータ量を増やす場合には、いろいろな人から意見をとってきて単語の情報を増やすという空間的な増やし方と、時系列的な量を増やすという２通りの方法がある。今回の分析は大まかにみて全ての時期を同じように分析をかけているので、時系列な量のバイアスを完全に無視した方法ということになる。今回結果では、抽出単語の中に「中国」や「東日本」が出てきているが、「東日本」という単語が景気が悪いという意味を示す理由は時期によるものである。３年という短い期間だが、三つの期間くらいに分けて、時系列方向の統計的有意性も検証しないと的中精度はかなり落ちてしまう。

・キーワードの分析は積み上げが大事なので、それぞれの単語に統計的有意水準（1%とか 5%とか）を付けてもらったほうが後々積み上げていく分析のときに役立つと思う。

・たとえ予測がある程度難しいとしても、過去に 1 か月ごとに出ている DI 値の内挿というか、日次で補える、過去のデータの間を埋められるような指標を作れるだけでも非常に価値があると思う。

・既存の経済統計と景気ウォッチャー調査は、両方ともバイアスを持っていると思う。どっちにどういうバイアスがあるかということを比較、相対的に整理しておくということが、もうちょっとあってもよいと思う。物価指数については多少そういう分析があったが、物価指数以外、景気動向についても同じような議論があってもよいと思う。

今回の分析業務においては、これらの事項を念頭に置いて分析の枠組み、手法、精度の改善を図ることが最大の課題となった。

データや分析業務の作業期間など、さまざまな制約要因があるため上記全ての指摘に対し改善のための対策や分析項目の追加などの施策を実行することは今回分析業務ではできなかった。しかし、景気ウォッチャー調査のテキスト個票に登場する単語を上位概念（シソーラス）の作成によってグループ化を図るなどの施策を行い、それによって精度の向上を引き出す、あるいは分析結果の可視化の度合いを高める取り組みは実行した。

今回の分析業務にあたり、委員の主な指摘事項に対する対処、施策の状況を表１．２－１に示す。

表１．２－１　平成 25 年度検討委員会での主な委員指摘事項並びに
今回分析業務における対処状況

委員からの指摘事項	今回分析業務における対処	報告書中の記載箇所
説明変数の数（400）が多すぎる。単語のグループ化をして説明変数を減らし、オーバーフィットを防ぐことが有望	シソーラスの作成による単語のグループ化を実施	2．2　テキストデータの数値化と景気動向と相関の強いキーワードの抽出
調査個票における毎月の頻出語句などのデータを付録として付けるだけでも、ビッグデータの将来的研究・分析が増えていく土台を作ることができる／（単語の）頻度を時間的に見せるだけでも、一般的な着目の変化を図式化する可能性ができる	イベント（大震災、消費税率引き上げ）前後における頻出単語の変化と登場数の推移を追跡	2．4　東日本大震災前後の変化 2．5　消費税率引上げ前後の変化
（機械学習による分析と、人間の判断による知見を融合させた）ハイブリッド的な分析の実施が望まれる	シソーラス作成の際の概念整理など、一部分析においては研究者の判断・知見による設定が行われている	2．2　テキストデータの数値化と景気動向と相関の強いキーワードの抽出
時系列方向の統計的有意性も少し検証しないと、的中精度がかなり落ちてしまう	イベント前後における頻出単語の変化と登場数の推移を追跡することにより検証に役立てる	2．4　東日本大震災前後の変化 2．5　消費税率引上げ前後の変化
それぞれの単語に1%、5%等の統計的有意水準を付けた方が、後々の分析の際に役立つ	今回分析業務では主要単語において有意水準を付加	2．4　東日本大震災前後の変化 2．5　消費税率引上げ前後の変化

<以下は委員指摘事項のうち、今回分析業務のスコープ外となる内容>

過去の月次DI値の内挿（日次で補える、過去のデータの間を埋められるような指標を作る）だけでも非常に価値がある	今回分析業務の説明変数は景気ウオッチャー調査のテキスト（月次）をベースとしているので、次回以降の課題	なし
既存の経済統計や景気ウオッチャー調査の統計に内在するバイアス（どの統計にどのようなバイアスがあるか）を比較、相対的に整理しておくことが有用	景気ウオッチャー調査の統計が持つバイアスについては分析業務を通じ定性的な知見が得られつつある	3．1　検証結果、検討委員会から得られた分析業務の課題 3．2　今後の分析業務推進の方向性

１．３　調査・分析項目

今回の調査・分析においては、以下のような項目についての分析業務を実施した。（分析業務の詳細は第２章を参照）

①　「テキストマイニング分析業務」

景気ウォッチャー調査のコメントから、景気動向に関連する特徴的な表現（単語、フレーズ）を抽出・分類し、登場頻度等を数値化しデータベース化する

・分析対象の絞り込み

・コメントからの単語の抽出
・景気に関連する特徴表現の抽出
・同義語の抽出
・単語と特徴表現を抽象化（シソーラス作成）
・有意表現の抽出
・震災前後の有意表現の変化分析
・消費税率引上げ前後の有意表現の変化分析

② 「データマイニング分析業務」
① の結果から得られたデータベースと、景気ウォッチャー調査の現状判断／先行き判断 DI 値、及びその他の経済指標との相関関係を探り、検証する
・予測式の抽出
・予測式の評価

なお、分析する項目、手順、手法については、精度改善を目的とするため昨年度の分析業務から一部を変更、追加または改訂しており、また一部の項目では分析を実施していない。

昨年度と今回における分析業務の主な相違・変更点については表１．３－１に示す。

表１．３－１　今回の分析業務における調査・分析項目

（赤字が主な変更／追加点）

昨年度の分析項目	今回の分析項目	分析手法等の変更点
<テキストマイニング>		
	・分析対象の絞り込み （「変わらない」コメントを分析対象外）	追加
・コメントからの単語の抽出	・同左（自立語のみを抽出）	
・景気に関連する特徴表現の抽出	・同左	
・同義語の抽出	・同左	
	・単語と特徴表現を抽象化（シソーラス作成）	追加
・有意表現の抽出	・同左	有意性検証指標として カイ二乗値を使用
・コメントからの景気現状判断分類	今回は実施しない	
・コメントからのDI値の予測 （2値、3値、5値に分類→予測）	今回は実施しない	
・コメントからのDI以外の経済指標の 上昇/下降の予測	データマイニングの項で実施	
	・震災前後の有意表現の変化分析	追加
	・消費税引上げ前後の有意表現の変化分析	追加
<データマイニング>		
・テキストマイニングで得られたデータベースをもとに、景気ウオッチャー調査の現状/先行き判断DI値との相関関係を探り、検証		
－分析A:有意単語発生頻度 →現状判断DI（全国）を予測	今回は実施しない	
－分析B:有意単語発生頻度 →現状判断DI（全国+地域別）を予測	・同左	昨年度の手法に加え、前年差分、有意表現出現の割合化、シソーラスの追加等を組み合わせ「分析群A」「分析群B」を実施
－分析C:有意単語発生頻度 →先行き判断DI（全国+地域別）を予測	・同左	
－分析D:個人別の現状/先行き判断に 基づくDI値の予測	今回は実施しない	
・震災前後のテキスト特徴量の分析	テキストマイニングの項で実施	
・震災前後の特徴表現の出現回数分析 （月次推移）	今回は実施しない	
	・有意単語発生頻度 →DI以外の経済指標の予測	追加（「分析群C」を実施）

１．４　検証結果の検討委員会の設置

検証結果の作成方法、基礎データ、昨年度からの改善結果などを検討・評価し、さらなる改善策を検討すべく、検証のための委員会を設置し、有識者から意見を聴取した。

＜開催日・場所＞
日時：平成 27 年 3 月 16 日　14：31－16：30
場所：内閣府合同庁舎 8 号館　429 号会議室

＜委員（氏名は五十音順）＞
和泉　潔　氏　（東京大学 大学院工学系研究科 システム創成学専攻 准教授）
大守　隆　氏　（東京都市大学 環境学部 教授）
水野 貴之 氏　（国立情報学研究所 情報社会相関研究系 准教授）