ESRI Discussion Paper Series No.345

計量テキスト分析による景気判断

ーコーディングルールや主成分を使った時系列分析ー

山澤成康

March 2018

内閣府経済社会総合研究所
Economic and Social Research Institute
Cabinet Office
Tokyo, Japan

計量テキスト分析による景気判断

－コーディングルールや主成分を使った時系列分析－

山澤成康 [1]

要旨

景気ウォッチャー調査の文章情報を使い、計量テキスト分析で景気動向の把握や予測法を検討した。「景気判断理由集（現状）」に使われる約 19 万件の文章から単語を抽出し、その単語が各月の総文章数に対してどれくらい出現するか（出現率）を集計して時系列データとして使用した。分析は、（１）コーティングルールを使った分析（２）相関分析（３）主成分分析（４）GDP 予測への応用――に分かれている。

コーディングルールを使った分析では、分析者が作成した単語の組み合わせ（コーディングルール）に従って出現率を計算して、グラフ化した。政策効果などがわかりやすく示せることがわかった。

相関分析では、景気ウォッチャー調査の現状判断 DI と各単語の出現率の相関係数をとり、どのような単語の相関係数が高いかを調べた。景気に順相関あるいは逆相関する単語を選び、景気指標を作成した。

主成分分析では、頻出 150 語の出現率を時系列データとみなし、主成分を抽出した。ウエートの高い語やその語と同時に使用される語などを検討して、各主成分がどのような性質を持っているのかを検討した。第１主成分が、景気ウォッチャー調査の現状判断 DI や景気動向指数・一致指数との相関が高いことがわかった。

GDP 予測への応用では、近似ダイナミックファクターモデルなどを使用して、実質 GDP 成長率が予測できるかどうかを検討した。説明変数として、鉱工業生産指数のほか単語の出現率や主成分を用いると、予測精度が上がることがわかった。

ＪＥＬ　分類番号：Ｅ３２
キーワード：景気循環、テキスト分析、景気ウォッチャー調査

---

[1] 内閣府経済社会総合研究所上席主任研究官。

本稿の執筆に際し、ＥＳＲＩセミナーの討論者である松林洋一神戸大学大学院経済学研究科、また西崎文平所長をはじめ出席者の皆様より有益な指摘を頂いた。なお、本稿で示された内容や見解は筆者個人によるものであり、所属する機関のものではない。ありうべき誤りは筆者個人の責に帰するものである。

１．はじめに

計量テキスト分析とは

計量テキスト分析は大量の文章情報を数量化して処理する分析法である。分析は、文章から単語を抽出して、その出現回数を数えるところから始まる。文章を名詞、動詞などに分けることを形態素（言語の意味を持つ最小単位）解析と呼ぶ。

樋口（2014）はテキスト分析の手法を 2 つに分け、客観的なデータを作成して多変量解析を使って分析する Correlational アプローチと、主観的な基準（コーディングルール）で、言葉や文章を選択して分析する Dictionary-based アプローチに分けている。Correlational アプローチは分析者の理論や問題意識の影響を受けない分析手法であり、Dictionary-based アプローチは研究者が主観的に言葉を選び、理論仮説の検証や問題意識の追求を行うアプローチである。本論文では両方のアプローチを使って分析を試みた。ソフトウエアは、樋口（2014）が紹介している「KHcoder」（http://khc.sourceforge.net/ よりダウンロード）を利用した。

先行研究

テキスト分析の景気分析への応用の先行研究として、山澤（2009）は「月例経済報告」の言葉に注目し、その変化から景気指標を作成した。谷口ほか（2010）は、経済新聞の言葉が景気拡大を示すのか、後退を示すのかに関して分析した。新聞記事の内容を解釈することに主眼が置かれている。和泉ほか（2011）は日本銀行の「金融経済月報」から金融市場の動向を分析している。

岡崎・敦賀（2015）は、KHcoder を利用した分析を紹介している。景気ウォッチャー調査を使い、センチメント指標の作成と共起ネットワークを使った分析をしている。

内閣府の委託調査（日本電気株式会社 2014、日本電気株式会社 2015）では、景気ウォッチャーの現状判断 DI についてテキストデータを使って推計した。

山本・松尾（2016）は、景気ウォッチャー調査のテキストデータをもとに、深層学習を行い、月例経済報告や展望レポートを適用する指数を開発した。これは野村 AI 景況感指数として公表されている。

2017 年に入ってからシンクタンクなどで、テキスト分析を応用したものが相次いで発表された。2017 年 7 月 13 日に公表された大和総研の「大和地域 AI（地域愛）インデックス」は、景気ウォッチャー調査を使って言葉と景気の関係を学習し、日本銀行の「さくらレポート」を数値化し、地域別の景気動向指数を四半期ごとに作成している。個々の文章が消費、投資など、どの分野に言及しているかを判別し、地域別、分野別にインデックスを作成している。

経済産業省と野村證券は、SNS（ソーシャル・ネットワーキング・サービス）に登場する言葉を使って景気や経済指標を予測する「SNS×AI 景況感指数」、「SNS×AI 鉱工業生産予測指数」を開発し、2017 年 7 月 19 日に試作版を公表した。「SNS×AI 景況観指数」は、AI（人口知能）を使ってテキストデータから景気指標を作成している。ただ単にテキストに

出現した単語の数を数えた場合、「値下げ」、「減少」はいずれも景気に対してマイナスのイメージだが、「値下げする商品の減少」であれば、景気に対してプラスとなる。AIの学習機能により、「値下げする商品の減少」などもプラスと評価できるように工夫がされている。ジーエフケーマーケティングサービスジャパン（GfK Japan）が作成したＰＯＳデータを利用した指標とともに、経済産業省の「ビッグデータスタッツ（試作版）」というサイトに掲載されている。

RIETI（独立行政法人経済産業研究所）は「日本の政策不確実性指標」を作成し、6月2日から公表を始めた。主要新聞4紙（朝日新聞、日本経済新聞、毎日新聞、読売新聞）に、economy（経済）、uncertainty（不確実）に関連する語と、政策（財政、金融、通商、為替）に関連する語を含む文章を数えることで、不確実性の指標を作成している。economy については「経済」と「景気」、uncertainty については「不透明」「不安」「不確実」「不確定」を採用している。採用した言葉はArbatli et al. (2017)に掲載してある。

三井住友アセットマネジメントの渡邊誠シニアエコノミストは、統計ソフトRのパッケージ「RMeCab」を使い、景気ウォッチャー調査に不安、節約に関連する語がどの程度出現するかをグラフ化した(渡邉　2017)。

前田（2017）は、クリスマスやハロウィーンといったイベント系の言葉、忘年会、新年会といった会合系の言葉の出現数について分析している。

表1はテキストデータを利用した指標をまとめたものである。

表１　テキストデータを利用した指標例

名称	説明	作成機関	出典など
景気ウォッチャー調査DIのコメントからの再現	景気ウォッチャー調査のテキストデータから数値データを推計	内閣府	内閣府（2015)
野村AI景況感指数	景気ウォッチャー調査の言葉を深層学習し、「月例経済報告」や「展望レポート」に適用	野村證券	山本・松尾（2016)
大和地域AI（地域愛）インデックス	「さくらレポート」を使って地域ごとにインデックスを作成する。	大和総研	大和総研ホームページ
SNS×AI 景況感指数（ウオッチャーAI)	SNSのテキストデータを使って景気ウォッチャー調査を予測	経済産業省	経済産業省（2017)
日本の政策不確実性指数	新聞記事から不確実性に関する言葉などの出現比率を指数化	独立法人経済産業研究所	経済産業研究所（2017）、Arbatli et

			al. （2017）
「不安」､「節約」に関する指標	景気ウォッチャー調査のさまざまな言葉の出現頻度を時系列のグラフとして表す。	三井住友アセットマネジメント	渡邉(2017)

本論文の問題意識

本論文では、景気ウォッチャー調査で毎月発表される「景気判断理由集（現状)」のコメントを元に分析する。ある条件に当てはまる言葉がいくつ現れるかを毎月調べ、総単語数に対する比率（出現率）として時系列データとする。このデータを使って、グラフの描画、主成分分析の適用、景気指標との相関分析、GDP の予測などを試みた。

本論文の構成は以下の通りである。第 2 章ではコーディングルールを使った分析をした。グラフを描いて、様々な言葉の出現率の動きをみた。第 3 章では、頻出語の出現率と景気指標との相関係数を調べることで、どのような言葉が景気と相関、あるいは逆相関するのかを調べた。第 4 章では頻出語の出現率に対して主成分分析を使い、それぞれの主成分がどのような意味を持つかを検討した。第 5 章では、頻出語の出現率やそれらの主成分が GDP の予測に有用な情報を持っているかどうかを検定した。「おわりに」では、この論文のまとめと今後の課題を書いた。

２．コーディングルールを使った分析

データについて

本論文では、景気ウォッチャー調査のテキストデータを使った分析を試みる。使用データは、インターネット上で公開されている「景気判断理由集（現状)」とし、2010 年 1 月から 2017 年 9 月までを対象とした。総計 12 万 631 件のコメントがあり、これは文章（センテンス）の数では 19 万 5089 個になる。ここから抽出した単語を一つの時系列データとみなし、各月ごとに、総単語数に対する比率（出現率）を計算した。

文章から単語を取り出すには形態素解析ツールを使う。KH Coder では、ChaSen（茶筌）を利用することができる。ChaSen（茶筌）は、奈良先端科学技術大学院大学松本裕治研究室で開発された（ChaSen :http://chasen-legacy.osdn.jp/）。

実際に形態素解析をすると、不自然な言葉が登録されることがある。たとえば、「アクセサリー」が「アクセ」と「サリー」と別々の語として認識される場合がある。これを修正するには、単語を「アクセサリー」として強制抽出する方法がある。また、「1 月」、「2 月」といった単語は頻繁に使われるが、言葉の増減が景況感に影響を与えるわけではないので、「1 月」から「12 月」までの言葉は、分析対象に含めないこととした。強制抽出した語、分析対象に含めない語については付注 1 にまとめた。

コーディングルールとは

次に、コーディングルールを使った分析を試みる。コーディングルールとは、言葉の頻度を数える際のルールである。コーディングルールを使うと、「北陸新幹線」といった一つの言葉を取り出すこともできるし、複雑なルールの下での出現頻度も計算できる。一つの文章に「A」という言葉と「B」という言葉いずれかが入っている場合は、「AorB」とし、両方入っている場合は「AandB」というルールを作る。「良い or 好調 or 上がる」であれば、文章の中に、良い、好調、上がる、のいずれかが入っている文章の数を計算することになる。また、「near(震災-復興)[15]」とすれば、前後 15 語以内に「震災」と「復興」が含まれる回数を数えることができる。品詞を指定することもでき、「ない-->否定助動詞」とすると、「ない」という言葉のうち否定助動詞だけを取り出すことができる。そのほかさまざまなルールが作成できる。先行研究のうち、コーディングルールを使った分析について表２にまとめた。

表２　コーディングルールを使った分析

テーマ	コーディングルール	作成機関	出典など
景気	「感情極性対応表」を使い、上位 30 語の出現頻度を集計	日本銀行	岡崎・敦賀（2015）
不確実性	「経済」と「景気」が含まれ、かつ「不透明」「不安」「不確実」「不確定」が含まれる新聞記事を集計	独立法人経済産業研究所	Arbatli et al.(2017)
不安	心配、懸念、不透明、不安	三井住友アセットマネジメント	渡邉（2017）
節約	慎重、様子見、買い控える、買い控え、節約、財布のひも		
個別の言葉の出現頻度	為替、株価、中国、テロ、地震・震災、マイナス金利、日銀、オリンピック、消費税・増税　、台風　、英国・EU 離脱、不安		
イベント系	クリスマス、祭り、バレンタイン、花見、ハロウィーン、花火	大和総研	前田(2017)
会合系	宴会、忘年会、新年会、送別会、パーティー、歓迎会、女子会		

コーディングルールによる時系列データ

ある言葉が出現する文が全体の文に対してどの程度出現するか（出現率）を毎月調べることで、時系列データを作ることができる。たとえば、「プレミアム付き商品券」は、2015 年に多くコメントされ、2016 年にも話題が出たことがわかる。「セール」や「春物」「夏物」といった言葉は季節ごとに表れるが、毎年の比較をすることで年ごとの違いがわかる。

コーディングルールを用いれば、さまざまな言葉の出現率を組み合わせて時系列データが作成できる。これらの言葉の動きを見ることで、さまざまなイベントがどの程度景気に影響するかを把握することができる。

どのようにして、言葉を選ぶのかは検討の余地がある。言葉を選ぶ基準には①主観的な判断②シソーラスの利用③「感情極性表現（高村他 2006）」などほかの研究の利用④機械学習による変数の選択－などが考えられる。シソーラスは言語に上下関係をつけ、体系化したものだ。例えば日経シソーラスでは、「リラクゼーション」という言葉の下に、「アロマテラピー」、「タラソテラピー」、「ヨガ」がある。岡崎・敦賀（2015）は、感情極性表現を使って言葉と景気との関係をつなげた。

さまざまな考え方があるが、主観的に選んだコーディングルールを使って試みに作成したものが、図１である。こうしたデータと景気指標を合わせてみることで、景気変動の原因を探ったり、政策効果がどのように表れているのかを類推したりすることができる。

図１　コーディングルールによる出現率

（注）各月の総文章数に対する検索語の出現回数の比率。それぞれの言葉のコーディングルールは以下の通り。

「プレミアム付商品券」　プレミアム付商品券　or　プレミアム付き商品券

「エコカー」エコカー

「消費税」増税　or　消費税　or　駆け込み

「オリンピック」オリンピック　or　五輪　or　パラリンピック　or　オリパラ

「東日本大震災」東日本大震災　or　東北大震災

「復興」復興　or　ふっこう

「爆買い」爆買い　or　爆買

「人手不足」人手不足　「冷夏」冷夏

「不確実性」心配　or　懸念　or　不透明　or　不安

「節約」慎重　or　様子見　or　買い控える　or　買い控え　or　節約　or（財布のひも　and　固い）

「災害」台風　or　豪雨　or　地震　or　震災　or　洪水　or　津波

３．相関分析による景気指標の作成

景気版「感情極性対応表」の作成

コーディングルールの応用として、景気と関連の深い語を選び出し、それを使って景気指標の作成を試みた。

岡崎・敦賀（2015）では、「感情極性対応表[2]（高村他 2006）」を使って景気指標を作成している。「感情極性対応表」は言葉を肯定的・否定的という尺度で数値化したものだ。すべての単語はマイナス１から 1 までの値で表され、肯定的な言葉ほど数値が高い。たとえば「優れる」は１、「悪い」はマイナス１と評価される。岡崎、敦賀（2015）は、上位 30 語について出現頻度を感情極性でウエート付けしてセンチメント指標を作成し、景気ウォッチャー調査現状判断 DI（方向性）と似た動きをしていることが示された。

しかし、「感情極性対応表」は景気の評価のために作られたものではないため、実感とそぐわない面が多い。たとえば、「上昇」という言葉は、-0.189547 と、マイナスの評価になっている。また、上昇という言葉には、「賃金の上昇」という良い意味と、「失業率の上昇」という悪い意味が含まれるため、この言葉が景気に対してどういう意味を持つかは、実際の「景気ウォッチャー調査」のサンプルに当たらないとわからない。また、「高い」という言葉はこの対応表にはない。

そこで本節では、景気判断のための「感情極性対応表」を作成する。コメントから抽出された言葉と景気ウォッチャー調査の数値情報（現状判断 DI（方向性））との相関係数を計算し、景気に関して肯定的な言葉かどうかを決める。ある言葉が景気をどのように評価しているかを演繹的に探るのは難しい。そこで、帰納的に景況と言葉とを結びつけようとしたものだ。

まず、頻出語の上位 300 語について、各月の出現率を計算する。次に、出現率と景気ウォッチャー調査の現状判断ＤＩ（方向性）との相関係数をとり、相関が高いもの、逆相関のもの、無相関のもの（相関係数の絶対値が小さいもの）について並び替えた（表３）。

相関が高いのは、「良い」「増える」「好調」といった言葉で、逆相関のものは、「落ち込む」「減少」「減る」という言葉である。「影響」という言葉はマイナスの意味で使われることが多く、相関係数のマイナス値が最も高くなっている。無相関のものは、「売上」「関連」「安い」といった言葉である。

次に、相関の高い単語の出現率を組み合わせて景気インデックスを作成した。相関の高い言葉のうち解釈のしやすいものを選び、指標とした。順相関の高い言葉として「良い」、「増える」、「好調」、「上昇」を選び、逆相関の言葉として「影響」「落ち込む」「減少」「減る」を選んだ。逆相関の高い言葉に「東日本大震災」があるが、震災後の特定の期間に用いられるものとして用いなかった。これらの言葉の出現頻度について相関係数で加重平均して景気指標（Text Index）を作ると、景気の現状判断 DI（方向性）とほぼ同様の動きになった

[2] 「単語感情表現対応表」が（http://www.lr.pi.titech.ac.jp/~takamura/pndic_ja.html）よりダウンロードできる。

（図 2、相関係数は 0.874）。計算式は以下の通り。X_t^a は、単語 a の出現率の時系列データである。

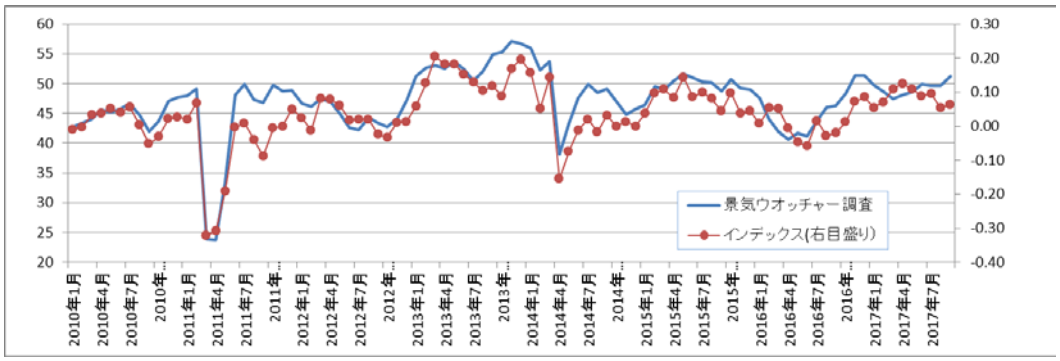
$$TextIndex_t = 0.742 * X_t^{良い} - 0.718 * X_t^{影響} + 0.685 * X_t^{増える} + 0.665 * X_t^{好調} + 0.631$$

$$* X_t^{上昇} - 0.596 * X_t^{落ち込む} - 0.572 * X_t^{減少} - 0.527 * X_t^{減る}$$

表３　景気の現状判断 DI（方向性）との相関ランキング

	順相関		逆相関		無相関	
	言葉	相関係数	言葉	相関係数	言葉	相関係数
1	良い	0.742	影響	−0.718	業界	0.001
2	増える	0.685	東日本大震災	−0.651	採用	0.001
3	好調	0.665	落ち込む	−0.596	売れる	0.003
4	上昇	0.631	以降	−0.578	飲食	0.005
5	変わる	0.619	減少	−0.572	サービス	0.006
6	増加	0.606	大きい	−0.570	安い	0.006
7	上がる	0.566	大幅	−0.551	個人	0.008
8	建設	0.520	生産	−0.546	様子	0.008
9	伸びる	0.512	減る	−0.527	店	0.010
10	高額	0.506	出る	−0.494	時期	0.010
11	高い	0.505	非常	−0.494	伴う	0.010
12	効果	0.492	落ちる	−0.450	派遣	0.011
13	上回る	0.490	低下	−0.433	利用	0.011
14	上向く	0.461	状態	−0.414	売上	0.013
15	推移	0.455	取引	−0.391	依然として	0.019
16	堅調	0.453	低迷	−0.385	意欲	0.021
17	景気	0.446	消費	−0.381	関連	0.035
18	期待	0.427	悪い	−0.375	店舗	0.039
19	変化	0.424	買う	−0.318	海外	0.041
20	感じる	0.418	必要	−0.317	買上	0.041
21	伸び	0.411	戻る	−0.307	思う	0.042
22	求人	0.402	厳しい	−0.290	新規	0.044
23	順調	0.392	見る	−0.286	食品	0.045
24	利益	0.388	広告	−0.267	企業	0.045
25	全体	0.359	状況	−0.265	確保	0.046
26	動き	0.352	自動車	−0.252	購入	0.057
27	前年	0.338	予約	−0.244	今年	0.061
28	比較	0.338	観光	−0.228	動く	0.064
29	求職	0.333	旅行	−0.211	回復	0.070
30	件数	0.323	客数	−0.210	特に	0.079

図２　現状判断 DI とテキストデータから作ったインデックス（Text Index）



否定語などに関する検証

言葉と景況感の対応を考える際には注意すべき点がある。言葉と景況感の関係が複雑な場合である（表４）。

一つ目は、同じ語が景気に正反対の評価をしている場合である。「上昇」という言葉は、主語が何になるかで景況に与える評価が変わっている。成長率が「高い」場合は景気にプラスになるが、失業率が「高い」場合は、景気にマイナスになる。この現象への対応策としては、「上昇」の主語を場合分けし、プラスに効く場合とマイナスに効く場合の効果を分けて考えることが必要だ。

景気ウォッチャー調査に関連して、こうした問題が起こるのは、「失業」に関する言葉だろう。この出現回数を調べてみると非常に少ない。毎月の文章数は平均 2098 個だが、「失業」が出現する回数は毎月平均 0.8回であり、出現しない月も多い。このため、今回の分析ではこの問題には特に対応しなかった。

表４　言葉と景況感の関係が複雑な場合

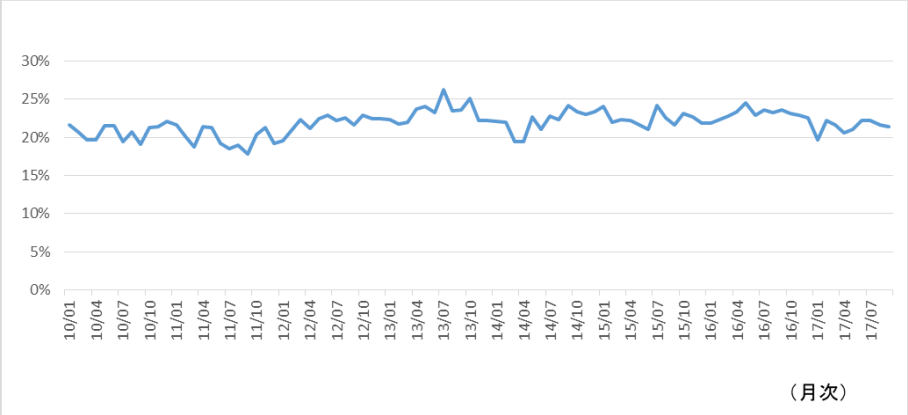
項目	例示
１正反対の意味	成長率が「高い」　⇔　失業率が「高い」
２文末否定	景気が良くなっているとは思えない　≒　景気が悪い 否定助動詞：好転するところまではいたっていない 形容詞：盛り上がりがない
３経済主体による違い	円高が進む　→輸入業者にとってはプラス、輸出業者にとってはマイナス

次に問題となるのは、ある言葉が否定形として使われる場合だ。「高い」という言葉が景気に肯定的な言葉だとしても、「高くない」というように、否定助動詞「ない」や形容詞「ない」が同時に使われる場合は正反対の意味になる。

ただ、否定助動詞「ない」の出現率が景気動向にかかわらず一定であるとすると、「低い」

を「高くない」といったり「多い」を「多くない」といったりする比率も変わらないと大まかには考えられる。景気ウォッチャー調査のサンプルでみると、否定助動詞「ない」が出現する比率は 20%から 25%の間でほぼ一定であることがわかる（図３）。

図３　否定助動詞「ない」が出現する比率



（注）各月の総文章数に対する「ない」という否定助動詞が含まれた文章数の比率。対象範囲は、景気ウォッチャー調査の 2010 年 1 月から 2017 年 9 月まで。

次に、「ない」の出現率が景気に関係がないことを統計的に検証した。前節で作成した景気インデックス（図２参照）を説明する回帰式において「ない」が出現する比率を説明変数に加えて有意に効くかどうかを調べた。推計式は以下の通りである。

$$Y_t = \alpha + \beta_1 X_t + \beta_2 Z_t + e_t$$

Y_t は現状判断 DI（方向）で、X_t はテキストデータから作成した景気インデックス、Z_t は、「ない」の各月の出現比率、e_t は誤差項である。

推計結果は以下の通り。

$$Y_t = 49.4 + 55.1X_t - 18.8Z_t + e_t$$
$$(12.5)\ (17.2) \qquad (\text{-}1.0)$$

自由度修正済み決定係数：0.773 ダービンワトソン比　0.580

この結果をみると、括弧内の数値が示す t 値からみて「ない」という言葉が景気ウォッチャー調査の現状判断 DI の動きを説明する上で、テキストデータから作成した景気インデッ

クスを補うような追加的な説明力をもっていはいない。

最後に、経済主体の違いで、言葉の意味が変わる場合である。円高という言葉は、輸入業者にとっては輸入価格が下がるのでプラスの意味に使われるが、輸出業者にとっては、ドル建て輸出物価の上昇を意味しておりマイナスの意味となる。景気ウォッチャー調査のコメント欄には、どのような業者が回答したかを答える項目があるので、それを使って分析をすることはできるが、本分析のデータセットにはそれを採り入れてないので、今後の課題とした。

４．テキストデータの主成分分析

前節では、それぞれの語と景気指標との相関係数をもとにインデックスを作成した。本節では、頻出語に含まれている主成分の中から景気に関連する主成分を抽出することを試みる。

頻出 150 語を抽出し、それぞれ語の出現率データを時系列変数としてみなし、その主成分をとりだして、景気指標との関係をみた。取り出した主成分の固有ベクトルをみてみよう（表５）。固有ベクトルは各主成分を作成する際のウエートを表すため、ウエートの高い言葉ほどその主成分に影響を与えていると考えられる。

また、各主成分がどのような言葉を中心に構成されているのかをみるために、共起ネットワークを作成した（付図１）。ここで作成した共起ネットワークは、各主成分の最もウエートの高い語（たとえば、第１主成分の場合は「良い」）を中心に、つながりの強い言葉同士を結んだものである。言葉のつながりの強さについては、Jaccard（ジャッカール）指数 [3] を使った。

和泉ほか（2011）では、まず、日銀金融経済月報から Jaccard 指数を用いて関係性の高い単語を抽出し、関係性の高い単語のみから主成分を構築している。本論文では、絞り込みは行っていない。ほかの単語との関係性が薄くても、独立して景気と相関している単語もあるためだ。

固有ベクトルや共起ネットワークから、それぞれの主成分がどのような性質を持っているかを類推してみよう。第１主成分は景気全般を表していると考えられる。「生産」や「景気」といった言葉のウエートが高い。第 2 主成分は共起ネットワークをみると最もウエーとが高い「共に」という言葉が、「客」「来客」「売上」などの言葉と使われているため、顧客の動向、第３主成分は、「単価」「価格」のウエートが上位にあることから、価格動向、第４主成分は「期待」、「景気」といった言葉のウエートが高いため景気の先行き、第５主成分は「消費」「観光」「旅行」といった言葉から、消費、観光関連の動向、第６主成分は、「感じる」「思う」「個人」といった言葉のほか「下回る」「落ちる」「好調（逆符号)」といった

[3] サンプル群間の類似度を比較するもの。テキスト分析では、ある語 w_1 が出現する回数を F_1、ある語 w_2 が出現する回数を F_2、w_1 と w_2 が同じ文に出現する回数を a として、$\frac{a}{F_1+F_2-a}$ で表される。大きいほどつながりが強い。

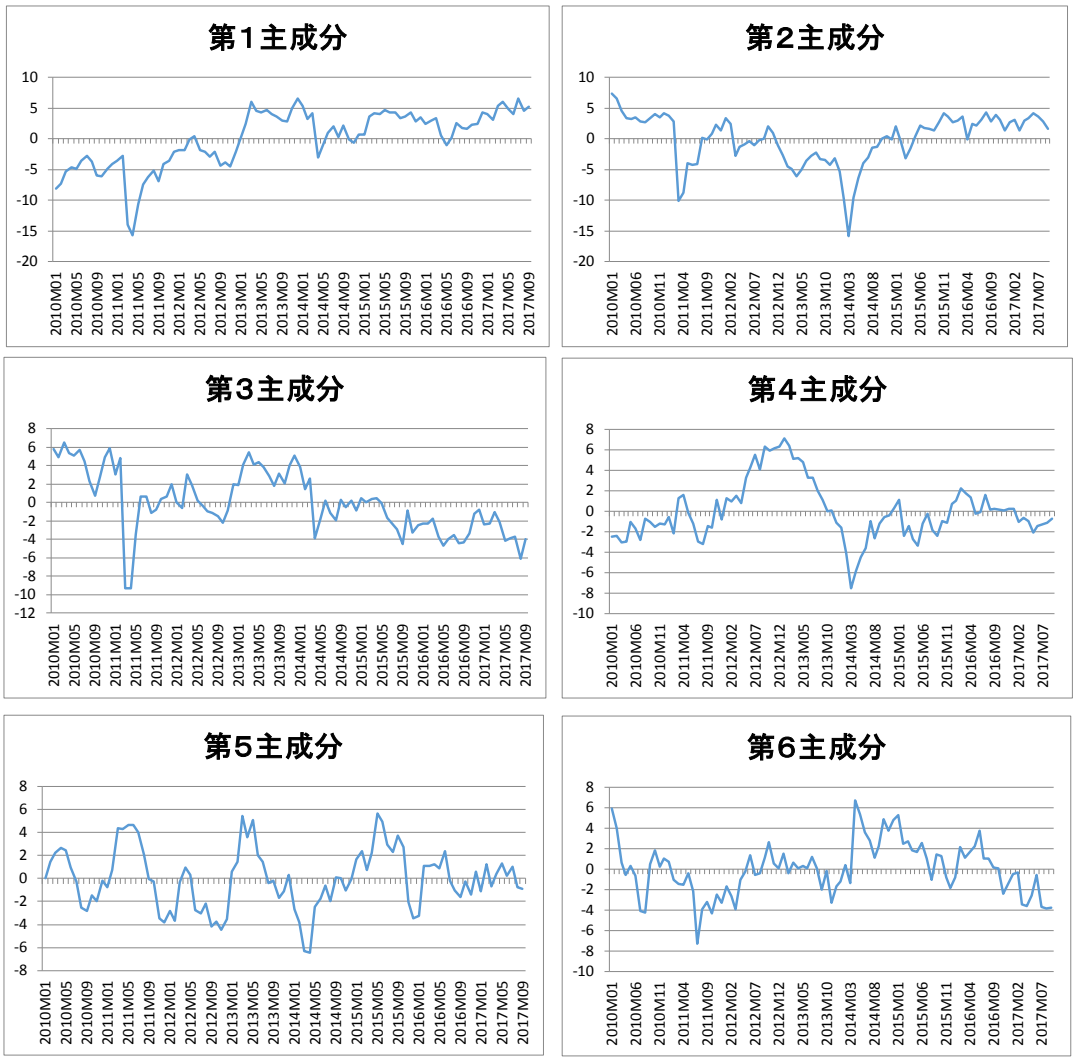
言葉のウエートが高く、消費マインド（の逆符号）を表していると考えられる。

これらの主成分が既存の景気指標とどの程度関係があるのかを調べたのが、表６である。第１主成分と景気指標との相関はかなり高い。第１主成分と景気動向指数 CI・一致系列との相関係数のほうが、第１主成分と景気ウォッチャー調査現状判断 DI との相関係数よりも高いという結果になった。図５は、第１主成分と景気ウォッチャー調査現状判断 DI のグラフである。第１主成分は DI の動きと似ているが相関係数は 0.743 で、前節で作成した景気インデックスの方が相対的に相関が高い。

消費者物価指数の伸び率と第２主成分の逆相関が比較的強い。使われている言葉のウエートから第２主成分は顧客動向を表していると述べたが、物価との相関が高いのは、駆け込み需要やその反動といった消費税増税前後の物価変動との相関が高くなっていると考えられる。

消費水準指数と第３主成分の相関が高い。また、第４主成分と鉱工業生産指数の逆相関が強いことがわかる。

図４　主成分分析の結果



（注）推計期間は、2010 年 1 月から 2017 年 9 月。

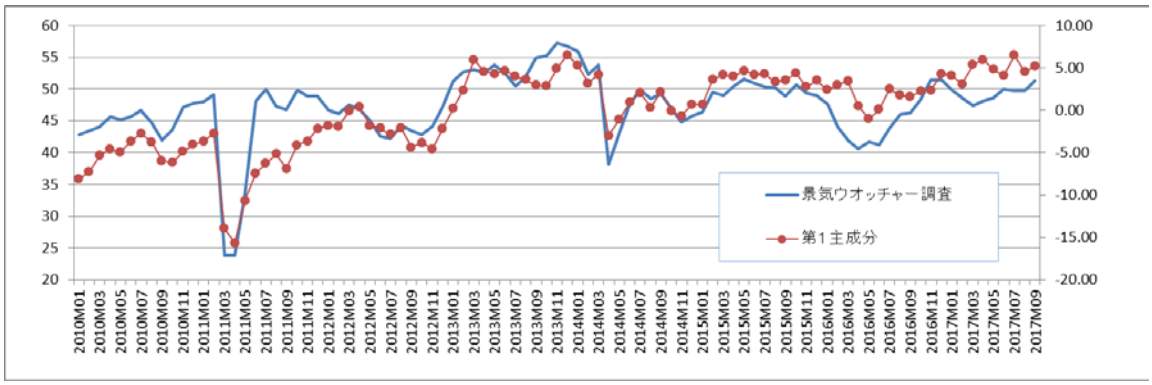
表５　各主成分の固有ベクトル（絶対値の大きい順）

	1		2		3		4		5		6	
1	良い	0.176	共に	0.180	単価	0.209	売上	−0.234	消費	0.199	感じる	0.201
2	変化	0.164	続く	0.177	観光	−0.183	期待	0.188	観光	0.198	中心	−0.178
3	非常	−0.154	前	−0.174	価格	0.173	増税	−0.163	企業	0.180	関連	−0.167
4	上昇	0.152	駆け込み	−0.163	影響	−0.170	上回る	−0.163	客	0.171	個人	0.164
5	生産	−0.152	消費税	−0.161	依然として	0.166	消費税	−0.158	駆け込み	−0.157	好調	−0.164
6	好調	0.151	需要	−0.159	増える	0.157	人	0.156	上向く	0.153	下回る	0.162
7	変わる	0.150	増税	−0.157	求人	0.148	景気	0.155	需要	−0.151	思う	0.156
8	景気	0.150	来客	0.156	上向く	0.147	駆け込み	−0.153	消費税	−0.150	増税	0.155
9	堅調	0.149	客	0.153	少し	0.147	競合	0.150	効果	0.149	気温	−0.151
10	大幅	−0.145	思う	−0.148	イベント	−0.143	今	0.149	増税	−0.149	落ちる	0.151

表６　主成分と経済指標の相関係数

	第1主成分	第2主成分	第3主成分	第4主成分	第5主成分	第6主成分
景気ウオッチャー調査・現状判断DI(方向性)	0.743	0.002	0.471	−0.071	−0.056	−0.208
景気動向指数・先行指数	0.671	−0.434	0.221	−0.183	0.085	−0.132
景気動向指数・一致指数	0.852	−0.107	−0.310	−0.227	−0.144	0.103
景気動向指数・遅行指数	0.725	0.030	−0.548	−0.204	−0.043	0.225
消費者物価指数（総合）（前年同月比）	0.336	−0.395	−0.124	−0.350	−0.239	0.373
消費者物価指数（生鮮食品を除く総合）（前年同月比）	0.321	−0.443	−0.152	−0.262	−0.221	0.378
消費者物価指数（食料（酒類を除く）及びエネルギーを除く総合）（前年同月比）	0.486	−0.238	−0.301	−0.273	−0.116	0.517
消費水準指数（世帯人員及び世帯主の年齢分布調整済）（二人以上の世帯）総合（前年同月比）	0.126	−0.010	0.359	0.166	0.000	−0.307
鉱工業生産指数 2010年基準（季節調整値）	0.303	0.319	0.364	−0.496	−0.269	−0.224

図５　　第１主成分と景気ウォッチャー調査現状判断 DI（方向性）



５．GDP の予測への応用

テキストデータを使って四半期別 GDP 速報（QE）の実質 GDP 前期比の伸び率の予測を試みた。QE の予測には、QE 発表以前に発表されたデータを使用することが一般的である。その候補はさまざまあるが、速報性を重視して対象四半期終了後 17 日前後で発表される鉱

工業生産指数の速報値[4]を使う。

この分析の目的は「テキストデータの情報を加えると予測精度は増すのか」という疑問に答えることである。分析の幅を広げるために、テキストデータのみで予測した場合や、景気ウォッチャー調査の現状判断 DI（方向性）を使った場合なども検討した。

景気ウォッチャー調査は、対象となる四半期が終わって 10 日程度で発表される。鉱工業生産指数の速報値は 17 日前後である。景気ウォッチャー調査だけを使った予測は、対象四半期終了後 10 日時点での予測、鉱工業生産指数速報値と景気ウォッチャー調査を使った予測は、対象四半期終了後 17 日後の予測ということになる。

予測対象として実用的なのは、第１次速報値だが、参考のため第２次速報や年次推計を被説明変数とした場合の値も検討した。使用する予測モデルは以下の通りである。

$$Y_t = \alpha + \beta_1 X_t + \Sigma \beta_i Z_{it} + e_t$$

Y_t は実質 GDP の前期比伸び率、X_t は鉱工業生産指数（速報値、以下同じ）の前期比伸び率、Z_{it} は景気ウォッチャー調査のテキストデータを使った時系列データ、e_t は推計残差である。 α 、β_i は係数である。月次データは３ヵ月分平均することで四半期データとした。鉱工業生産指数のデータを使わない場合は、X_t の項はない。

テキストデータを使った変数にはいくつかのパターンが考えられる。①出現頻度の高い単語を選んで当てはまりのよいものを探す②主成分分析を使う③テーマに沿ったコーディングルールを使う――という方法だ。

①は、景気ウォッチャー調査に使われている「減る」「増える」などの言葉の出現頻度を時系列変数として使うものだ。この推計結果をみることでどのような言葉が実質 GDP 前期比と相関が高いのかがわかる。ただ、複数の変数を使う場合は多重共線性の問題が発生する。

②は、対象とした頻度変数を主成分分析にかけて、新たに説明変数として使う方法である。近似ダイナミックフィルターと同様の手法である。

③は、さまざまなコーディングルールを作成して、それを説明変数として使うものだ。主観的な作業になるが、適切なコーディングルールを使って作成すれば、最も望ましい推計法といえる。試行錯誤が必要となり、今回の分析では行っていない。

説明変数にはさまざまな候補があるが、それぞれの式について当てはまりの高い順に並べた。当てはまりの基準としては決定係数を使った。同じ式の中での当てはまりの優劣をつける場合、当てはまりの基準を自由度修正済み決定係数にしても赤池情報量規準（ＡＩＣ）にしても結果は同じである。説明変数の数やサンプル数が同じだからである。

テキストをそのまま説明変数とする場合

被説明変数を実質 GDP 前期比伸び率（Y_t）とし、テキストデータを説明変数として推計

[4] 鉱工業生産指数の予測指数や第三次産業活動指数を使う予測も考えられ、さまざまな変数を使う余地はある。

した。全データから、頻度の高い順に単語を 150 語選び、それらの各月の出現頻度を説明変数 Z_{it} とした。実質 GDP 前期比伸び率には、第１次速報値、第２次速報値、年次推計値を使った。

$$Y_t = \alpha + \Sigma \beta_i Z_{it} + e_t \quad （式１）$$

$$Y_t = \alpha + \beta_1 X_t + \Sigma \beta_i Z_{it} + e_t \quad （式２）$$

式１は、テキストデータだけで回帰したもので、式２はテキストデータのほか鉱工業生産指数前期比伸び率（X_t）を説明変数としたものだ。

近似ダイナミックファクターモデル

テキストデータをそのまま使うと、多重共線性の問題が起こりうる。そこで、出現頻度の高い 150 語について、主成分分析を行い第 30 主成分までを使って推計した。

被説明変数を実質 GDP 前期比伸び率（Y_t）とし、説明変数をテキストデータの主成分（P_{it}）とした、近似ダイナミックファクターモデルである（近似ダイナミックファクターモデルについては、飯星 2009 参照）。

$$Y_t = \alpha + \Sigma \beta_i P_{it} + e_t \quad （式３）$$

$$Y_t = \alpha + \beta_1 X_t + \Sigma \beta_i P_{it} + e_t \quad （式４）$$

式３は、主成分だけで回帰したもの、式４は主成分のほか鉱工業生産指数前期比伸び率（X_t）を説明変数としたものだ。

説明変数とする主成分の数を１から５までそれぞれ推計した。30 個のあらゆる組み合わせのなかから最も決定係数の高い組み合わせを選んだ。

表７　推計結果（その１）

被説明変数：実質GDP前期比伸び率

（式1）説明変数：テキストデータのみ

	1次速報	R^2	2次速報	R^2	年次推計	R^2
1	落ちる	0.252	増える	0.258	増える	0.224
2	今年、落ちる	0.475	今年、落ちる	0.503	増える、上昇	0.433
3	今年、落ちる、業界	0.609	今年、落ちる、業界	0.623	好調、前月、依然として	0.610

（式2）説明変数：鉱工業生産指数速報前期比伸び率＋テキストデータ

	1次速報	R^2	2次速報	R^2	年次推計	R^2
1	今年	0.661	今年	0.656	点数	0.632
2	伸びる、高い	0.742	伸びる、高い	0.741	変わる、伸びる	0.711
3	伸びる、高い、派遣	0.800	価格、低い、利益	0.822	売上、変わる、伸びる	0.859

（式3）説明変数：主成分のみ

	1次速報	R^2	2次速報	R^2	年次推計	R^2
1	23	0.153	3	0.139	25	0.255
2	23,25	0.291	23,25	0.278	23,25	0.541
3	12,23,25	0.291	12,23,25	0.402	3,23,25	0.634
4	9,20,23,25	0.571	9,20,23,25	0.545	3,8,23,25	0.714
5	9,11,20,23,25	0.668	3,12,20,23,25	0.616	3,8,14,23,25	0.777

（式4）説明変数：鉱工業生産指数速報前期比伸び率＋主成分

	1次速報	R^2	2次速報	R^2	年次推計	R^2
1	19	0.641	8	0.635	6	0.543
2	6,19	0.702	8,26	0.717	6,27	0.729
3	6,19,23	0.755	8,16,19	0.752	6,10,27	0.778
4	6,16,19,23	0.802	8,16,19,24	0.747	6,10,23,27	0.729
5	6,8,16,19,23	0.832	6, 8,16,19, 24	0.830	6,10,24,27,29	0.813

（注）R^2は決定係数。推計期間：2010年第1四半期から2017年第2四半期（速報）、2010年第1四半期から2015年第4四半期（年次推計）。

推計結果

推計結果（表７）で、一つの単語だけで実質 GDP 前期比伸び率（1次速報値）を推計した場合（式１）をみると決定係数は 0.252 で、当てはまりのよい言葉はないことがわかる。3 つの言葉を組み合わせると、決定係数は 0.609 まで上がるが、それほど高くない。また、第 1 次速報値と第 2 次速報値は似た言葉を選択しているが、年次推計値は違う言葉の当てはまりがよい。

説明変数に鉱工業生産指数前期比伸び率（速報値）を加えると（式２）、当てはまりはよくなり、３変数を使うと、決定係数は 0.800 となる。

主成分のみを説明変数とした場合（式３）は、5 つの主成分を使っても、決定係数は 0.668 である。鉱工業生産指数前期比伸び率（速報値）を加えると（式４）、決定係数は 0.832 となり、４種類の定式化の中では最も決定係数が高くなる。

１～４式で、それぞれ最も決定係数が高いものについて、推計結果を載せたものが表８、推計値が図６である。説明変数に用いた主成分のグラフは付図２、付図３に載せた。左端はテキストデータを使わず、鉱工業生産指数のみを説明変数とした場合である。この場合の自由度修正済み決定係数は 0.566 だ。次に数値データとして景気ウォッチャー調査の現状判

断 DI を説明変数にしたものも推計した。数値データの係数は有意ではなく、実質 GDP 成長率の予測には寄与しないことがわかった。現状判断 DI の差分をとって説明変数とした場合も同様だった。

テキストデータのみを説明変数とした場合の（式１）は、鉱工業生産指数のみの場合よりも自由度修正済み決定係数は低い。（式３）はテキストデータの主成分のみを説明変数としたものだが、自由度修正済み決定係数は 0.599 と高くなっているものの、ＡＩＣやベイズ情報量規準（ＢＩＣ）をみると、数値が大きく（当てはまりは悪く）なっている。

一方、鉱工業生産指数に加えてテキストデータを説明変数とした式２でも式 4 でも自由度修正済み決定係数が上昇し、係数も有意である。この結果をみると、テキストデータを説明変数に使うことで予測精度が向上することがわかる。すなわち、テキストデータには鉱工業生産指数には含まれない有益な情報を含んでいるために、追加的な説明力をもつと解釈できる。

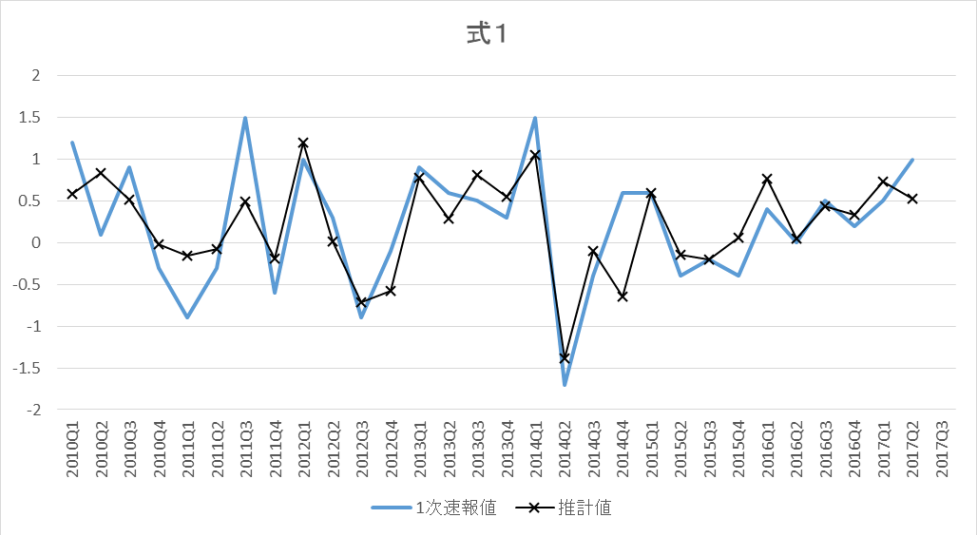
表８　推計結果（その２）

被説明変数：実質ＧＤＰ前期比伸び率（1 次速報値）

変数	テキストデータなし			景気ウオッチャー調査の現状判断DIを加えた場合			式1			式2			式3			式4		
α	定数項	0.183	(2.044)	定数項	0.174	(0.159)	定数項	−1.429	(−1.858)	定数項	0.681	(1.146)	定数項	0.220	(2.556)	定数項	0.185	(2.943)
β１	鉱工業生産指数	0.229	(6.235)	鉱工業生産指数	0.229	(5.446)				鉱工業生産指数	0.239	(9.333)				鉱工業生産指数	0.221	(7.835)
β２				現状判断DI	0.0002	(0.008)	「今年」	46.760	(3.864)	「伸びる」	47.468	(4.888)	第9主成分	−0.192	(−3.502)	第6主成分	−0.064	(−2.295)
β３							「落ちる」	−100.730	(−4.974)	「高い」	−51.585	(−4.553)	第11主成分	0.223	(2.657)	第8主成分	0.081	(2.040)
β４							「業界」	117.866	(2.985)	「派遣」	−106.171	(−3.522)	第20主成分	−0.430	(−3.755)	第16主成分	−0.162	(−2.559)
β５													第23主成分	0.606	(4.337)	第19主成分	0.221	(3.232)
β６													第25主成分	0.398	(3.775)	第23主成分	0.270	(2.721)
Adj. R2		0.566			0.550			0.564			0.800			0.599			0.789	
AIC		1.472			1.538			1.537			0.784			1.506			0.890	
BIC		1.565			1.679			1.724			1.018			1.786			1.217	
D.W.		2.540			2.540			2.410			2.365			1.309			2.524	

(注)推計期間は 2010 年１－3 月期から 2017 年４－６月期。カッコ内はｔ値。Adj.R2 は自由度修正済み決定係数、D.W.はダービンワトソン比。

図６（その１）　　実質ＧＤＰの予測値



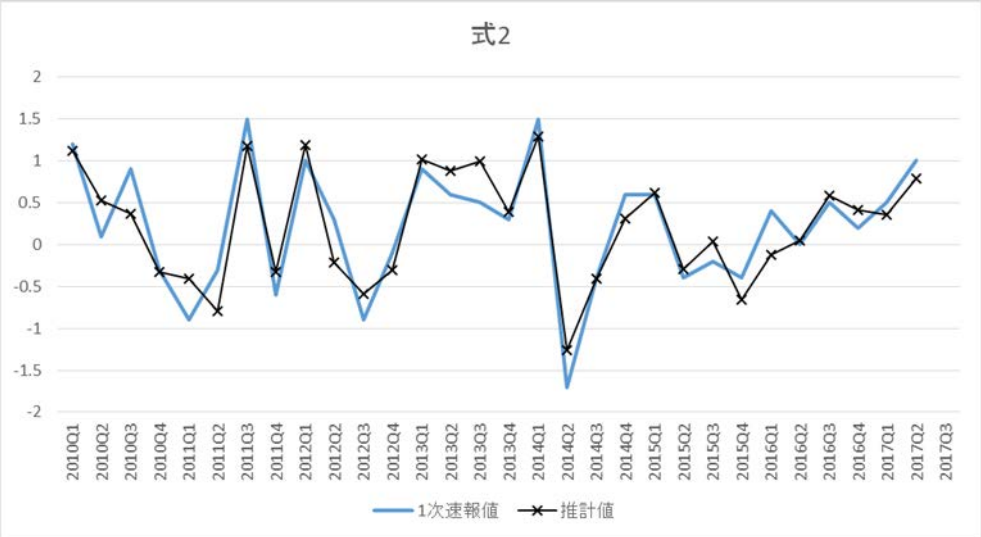
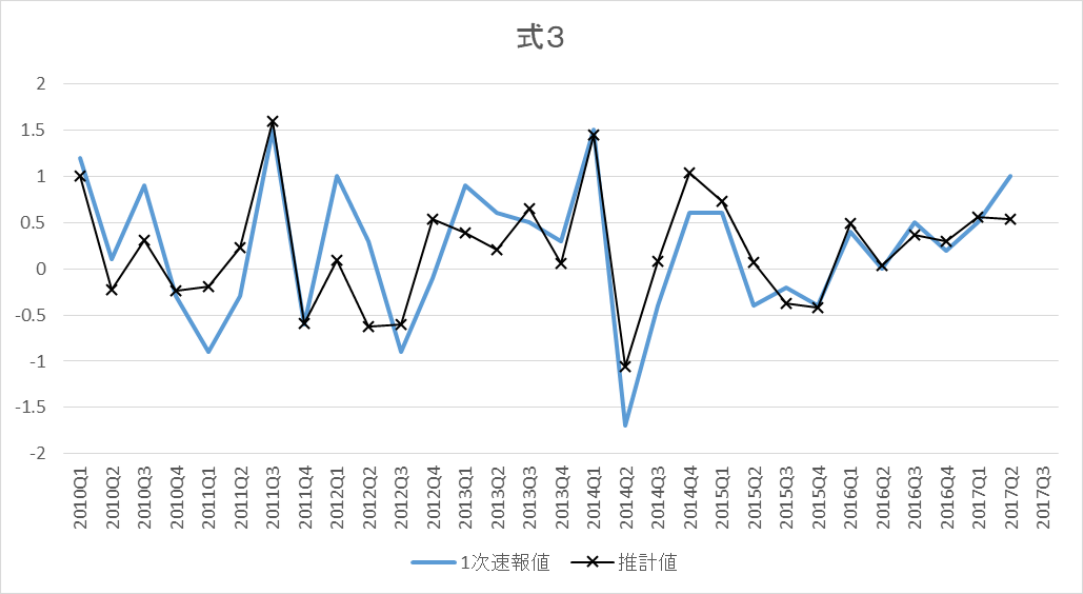
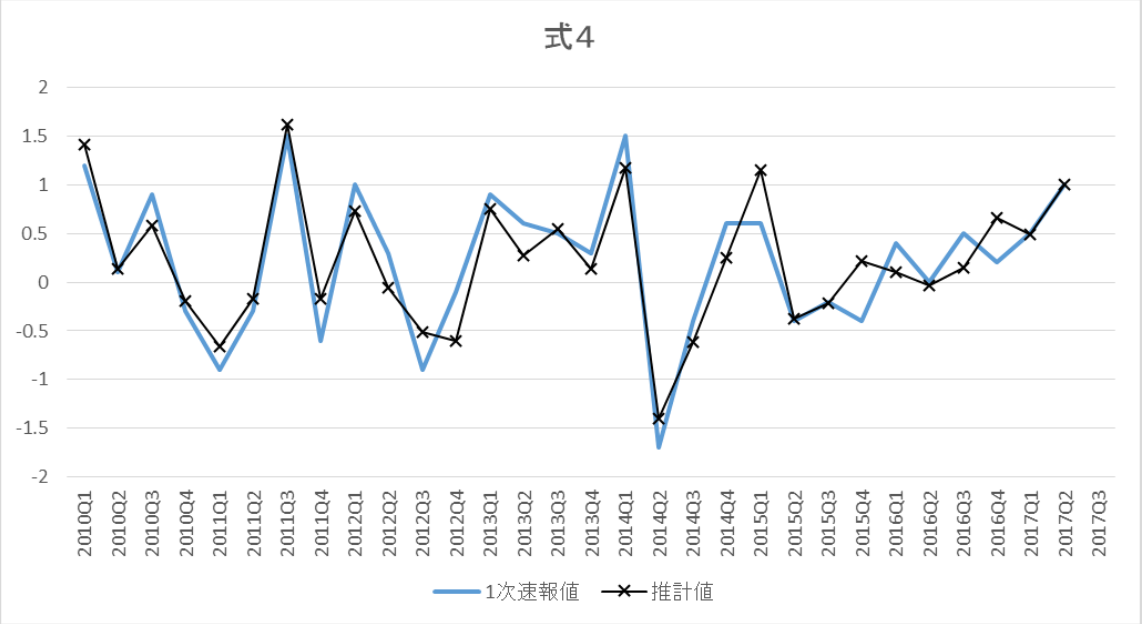


図６（その２） 実質 GDP の予測値(続き)





（注）予測に使用した推計式は、被説明変数を実質 GDP 成長率の１次速報値とした。式１は、３変数のテキストデータのみ、式２は鉱工業生産指数と３変数のテキストデータ、式３は、５主成分のみ、式４は鉱工業生産指数と５主成分で推計。

おわりに

この論文ではテキストデータの景気分析への応用を試みた。

分析者が作成した単語の組み合わせ（コーディングルール）を用いた分析では、「プレミ

アム付商品券」、「エコカー」、「消費税」といった単語の出現率を時系列に沿って描くことで、政策効果を視覚的に表現できることがわかった。

テキストデータと景気ウォッチャー調査の現状判断 DI(方向性)との相関係数を調べ、景気と相関の高い言葉を見つけた。「良い」「増える」「好調」などの言葉は景気と相関が高く、「影響」「東日本大震災」「落ち込む」などの言葉は景気と逆相関である。これらの語の出現率を使って、現状判断DI(方向性)とほぼ同じ動きをするインデックスを作ることができた。

頻出語を主成分分析にかけると、第１主成分と景気ウォチャー調査の相関が高く、第１主成分が景気動向を表していることが確認できた。

最後にテキストデータを実質 GDP 予測に使えるかどうかを検証した。鉱工業生産指数とともに「伸びる」「高い」「派遣」の出現率を説明変数に加えると予測精度が増し、テキストデータには鉱工業生産に含まれない有益な情報があることがわかった。

今後の課題として、景気ウォッチャー調査のテキストデータを 2001 年まで遡って収集することや、「月例経済報告」などほかのテキストデータを使って分析することが考えらえる。分析手法としては機械学習などを使うことが考えられる。

参考文献

飯星博邦(2009)「主成分分析によるマクロ経済パネルデータの共通ファクターの抽出とその利用」 ESRI Discussion Paper Series No.219

和泉潔、後藤卓、松井藤五郎（2011）「経済キスト情報を用いた長期的な市場動向推定」、『情報処理学会論文誌』Vol.52 No.12、2011 年、3309-3315 頁

岡崎陽介、敦賀智裕（2015）「ビッグデータを用いた経済・物価分析について－研究事例のサーベイと景気ウォッチャー調査のテキスト分析の試み」日本銀行

経済産業研究所（2017）「日本の政策不確実性指数」、経済産業研究所ホームページ、データ・統計、https://www.rieti.go.jp/jp/database/policyuncertainty/ (2017 年 12 月 7 日閲覧)

経済産業省（2017）「BigData-STATS」、経済産業省ホームページ、https://bigdata-statistics.meti.go.jp/ (2017 年 12 月 7 日閲覧）

高村大也・乾孝司・奥村学（2006）「スピンモデルによる単語の感情極性抽出」、『情報処理学会論文誌』Vol.47 No.2、pp.627-637

大和総研（2017）「大和地域 AI（地域愛）インデックス」、大和総研ホームページ、http://www.dir.co.jp/research/report/regionalindex/2017.html （2017 年 12 月 7 日閲覧）

谷口 将太、坂地 泰紀、酒井 浩之 他（2011）『経済新聞記事から抽出した景気動向を示す根拠表現への極性付与手法の提案』電子情報通信学会論文 D,Vol.J94-D,No.6,pp.1039-1043

「新聞記事のテキストマイニングによる長期市場動向の分析」, 人工知能学会論文誌 Vol. 28(2013) No. 3 , p.291-296 (2013)

内閣府 (2015)『地域の経済 2014』政策統括官（経済財政分析担当）

日本電気株式会社（2014）『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』（平成 25 年度内閣府委託調査）、2014 年 3 月

日本電気株式会社（2015）『東日本大震災及び消費税率引上げ後の日本経済の産業構造・景気循環分析業務報告書』（平成 26 年度内閣府委託調査）、2015 年 3 月

樋口耕一（2014）『社会調査のための計量テキスト分析－内容分析の継承と発展を目指して』

ナカニシヤ書店

前田和馬（2017）「パリピが景気動かす？」大和総研コラム、大和総研

松本裕治（2011）「ChaSen--形態素解析器」、奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座(松本研究室)、http://chasen-legacy.osdn.jp/ （2017 年 12 月 7 日閲覧）

渡邉誠（2017）「消費動向の分析②～サービス消費の持ち直しとその背景～」エコノミスト便り、三井住友アセットマネジメント

松本裕治（2011）「ChaSen--形態素解析器」、奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座(松本研究室)、http://chasen-legacy.osdn.jp/ （2017 年 12 月 7 日閲覧）

山澤成康（2009）「景気指標としての月例経済報告」JCER Discussion Paper 124、日本経済研究センター

山本裕樹・松尾豊(2016)「景気ウォッチャー調査の深層学習を用いた金融レポートの指数化」第 30 回人工知能学会全国大会

Arbatli ,C. E., S. J. Davis, A. Ito ,N. Miake and I. Saito(2017)" Policy Uncertainty in Japan" Working Paper No. 17/128,International Monetary Fund

付注１

＜強制抽出する語＞

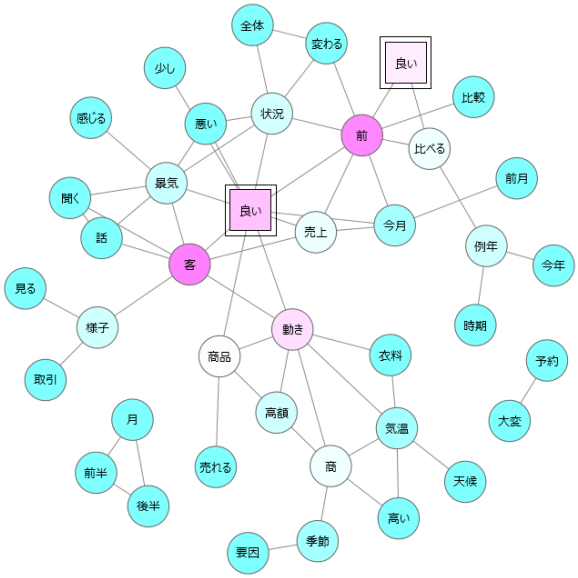
11月、12月、アウトレット、アウトレットモール、アクセサリー、インバウンド、エコカー、エコポイント、オン・デマンド、オンシーズン、オンリー、クリスマス、グロサリー、ゴーデンウイーク、コールセンター、サーチャージ、シークレット、ステテコ、スマートフォン、スマホ、テイクアウト、テイスト、デパ地下、テレマーケティング、トライアル、ナフサ、パート、パブリシティ、プレカット、プレクリアランス、プレセール、プレミアム商品券、プレミアム付商品券、ヘッドスパ、マイナンバー、ユニーク、リーマンショック、リクルーティング、リピーター、リピート、リモデル、ローコスト、ロープライス、ワイナリー、駅ナカ、牛タン、県央、元請、原子力発電、原子力発電所、原糸、原燃料、催行、少しづつ、消費税、水揚、政権交代、孫請、宅建業、東日本大震災、当県、道央、爆買い、北陸新幹線、満タン

＜排除する言葉＞

１月、２月、３月、４月、５月、６月、７月、８月、９月、１１月、１２月、11月、12月

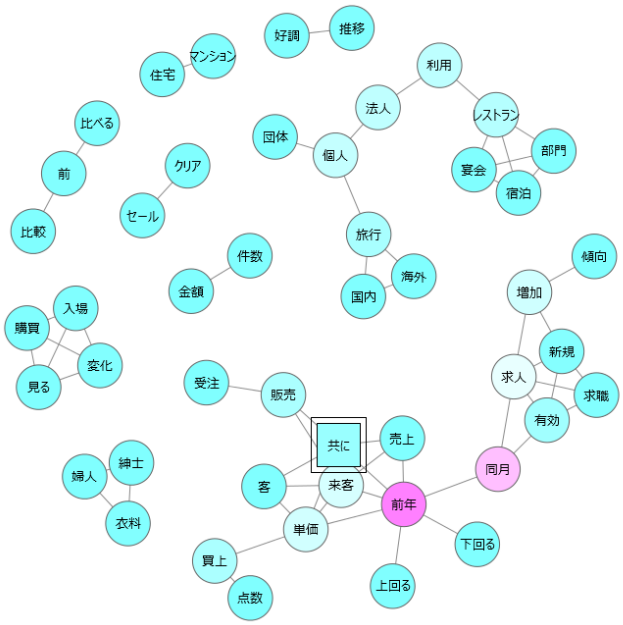
付図１　主成分ごとの共起ネットワーク

第１主成分

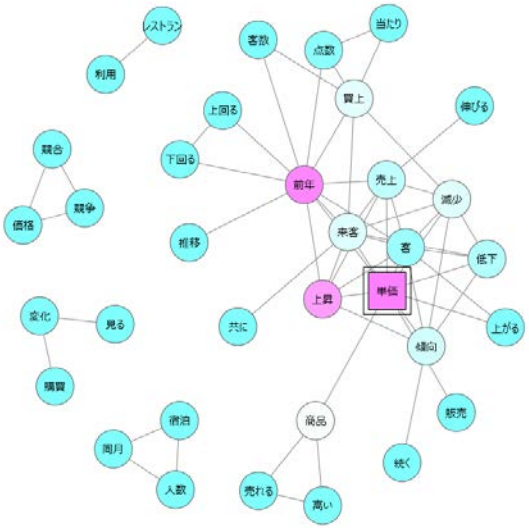


(注)2 重の正方形で囲まれた語は、第１主成分を構成する語のうち最もウエートの高い語。その語を中心に関係の深い語を結んでいる。色が濃いほど出現率が高い。第２主成分以下も同じ。

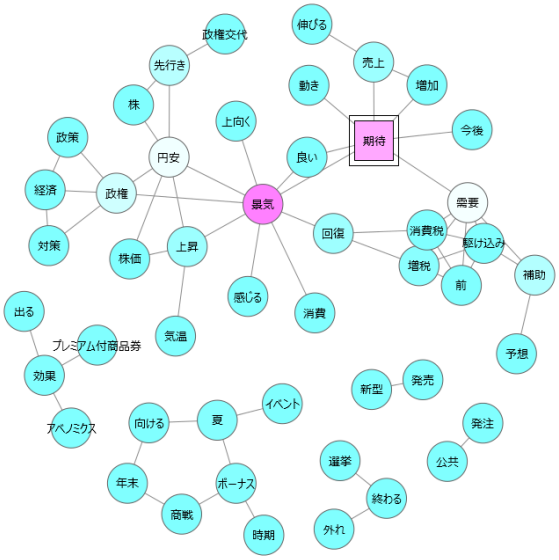
第２主成分



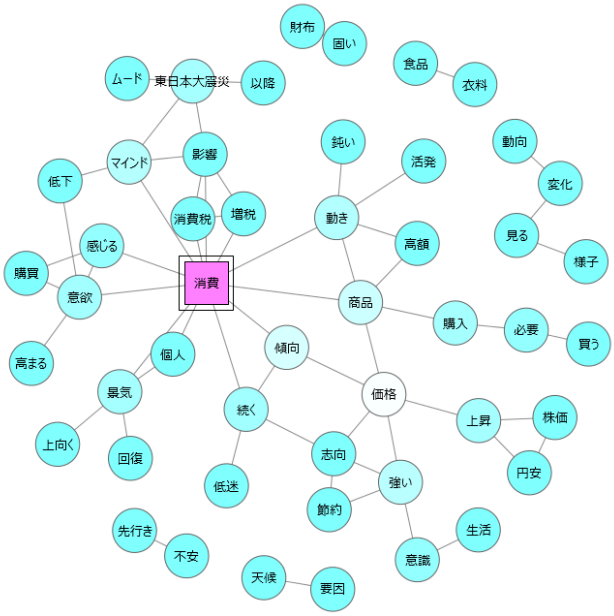
第 3 主成分


レストラン
利用
客数
点数
当たり
買上
伸びる
上回る
下回る
前年
売上
減少
競合
競争
価格
推移
来客
客
低下
単価
上昇
変化
見る
共に
上がる
傾向
購買
販売
滞泊
周月
人数
商品
売れる
高い
続く

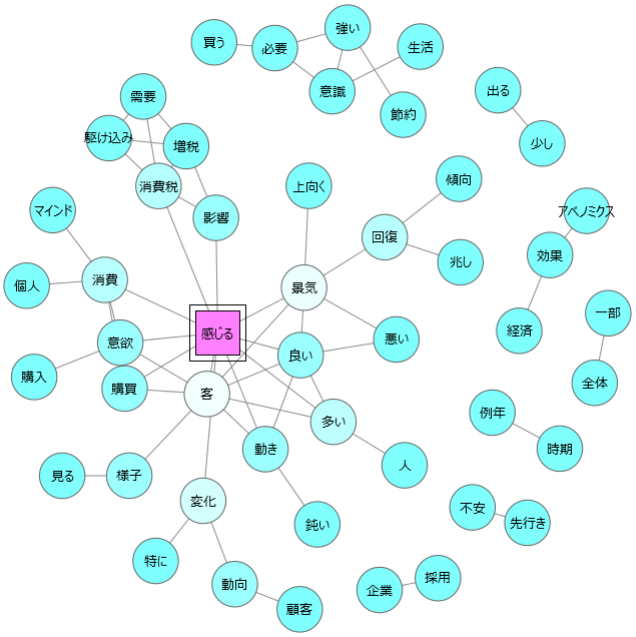
第 4 主成分


政権交代
伸びる
先行き
売上
株
動き
増加
上向く
期待
今後
政策
円安
良い
経済
政権
景気
需要
対策
株価
上昇
回復
消費税
駆け込み
増税
前
補助
感じる
消費
出る
気温
プレミアム付商品券
予想
効果
新型
発売
アベノミクス
向ける
夏
イベント
発注
選挙
公共
年末
ボーナス
終わる
商戦
外れ
時期

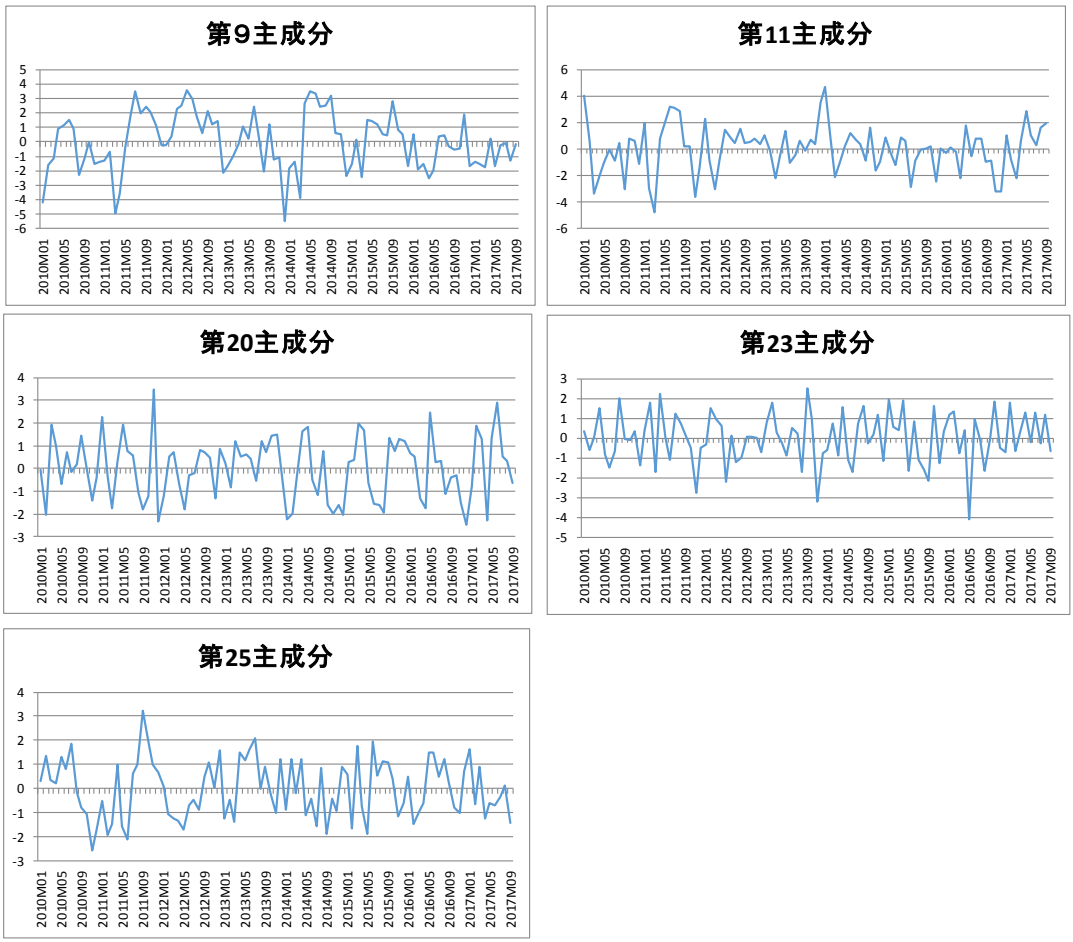
第 5 主成分


財布
固い
食品
衣料
ムード
東日本大震災
以降
マインド
影響
純い
活発
動向
変化
低下
消費税
増税
動き
高額
見る
様子
感じる
購買
消費
意欲
商品
購入
必要
買う
高まる
個人
傾向
景気
続く
価格
上昇
株価
上向く
回復
志向
円安
低迷
強い
節約
先行き
生活
不安
意識
天候
要因

第 6 主成分


買う
必要
強い
生活
意識
節約
出る
少し
需要
駆け込み
増税
消費税
上向く
傾向
アベノミクス
マインド
影響
回復
効果
個人
消費
景気
兆し
感じる
意欲
悪い
一部
経済
購入
良い
購買
客
全体
例年
多い
時期
動き
人
見る
様子
変化
不安
先行き
純い
特に
動向
企業
採用
顧客

付図２　主成分のグラフ（表７の式３）



付図３　主成分のグラフ（表７の式４）