



日本銀行ワーキングペーパーシリーズ

## 機械学習による景気分析 — 「景気ウォッチャー調査」のテキストマイニング —

大高一樹\*

kazuki.ootaka@boj.or.jp

菅和聖\*\*

kazutoshi.kan@boj.or.jp

No.18-J-8  
2018年9月

日本銀行  
〒103-8660 日本郵便（株）日本橋郵便局私書箱 30 号

\* 調査統計局

\*\* 調査統計局（現・国際局）

日本銀行ワーキングペーパーシリーズは、日本銀行員および外部研究者の研究成果をとりまとめたもので、内外の研究機関、研究者等の有識者から幅広くコメントを頂戴することを意図しています。ただし、論文の中で示された内容や意見は、日本銀行の公式見解を示すものではありません。

なお、ワーキングペーパーシリーズに対するご意見・ご質問や、掲載ファイルに関するお問い合わせは、執筆者までお寄せ下さい。

商用目的で転載・複製を行う場合は、予め日本銀行情報サービス局 (post.prd8@boj.or.jp) までご相談下さい。転載・複製を行う場合は、出所を明記して下さい。

# 機械学習による景気分析\*

## —「景気ウォッチャー調査」のテキストマイニング—

大高 一樹<sup>†</sup>・菅 和聖<sup>‡</sup>

2018年9月

### 【要旨】

本稿では、景気ウォッチャー調査のテキストデータに対し、統計的言語処理と機械学習の手法を用い、3つの分析を行った。第1に、過去対比で相対的に出現頻度が高い特徴語に対する共起ネットワーク図により、景気変動要因を効果的に可視化した。第2に、ナイーブベイズ分類器を用い、テキストデータから物価動向を表す指数を構築した。この指数は、インフレ率の先行指標として有用であることが分かった。第3に、予想インフレ率やインフレ率に対して説明力がある単語を、ランダムフォレストの変数重要度を基準に抽出した。これにより、物価変動の要因を理解するうえで有益な情報が得られることが分かった。

JEL 分類番号 : C55, C81, E31, E32

キーワード : テキスト分析, 機械学習, 共起ネットワーク図, ナイーブベイズ分類器, ランダムフォレスト, インフレ率, 予想インフレ率

---

\* 本稿の作成に当たり、日本銀行の一上響氏、宇野洋輔氏、加藤直也氏、城戸陽介氏、関根敏隆氏、近松京介氏、敦賀智裕氏、平形尚久氏、村越智文氏、吉羽要直氏から有益なコメントを頂いた。ただし、あり得べき誤りは筆者ら個人に属する。本稿の内容と意見は筆者ら個人に属するものであり、日本銀行の公式見解を示すものではない。

<sup>†</sup> 日本銀行調査統計局 (kazuki.ootaka@boj.or.jp)

<sup>‡</sup> 日本銀行調査統計局 (現・国際局, kazutoshi.kan@boj.or.jp)

## 1. はじめに

近年、景気動向や物価動向に関連する新たな情報が、情報技術やその基盤となるインフラの普及とともに増加している。例を挙げると、ニュース記事やインターネット上のテキスト情報、決済データ、気象データ、および交通・運輸のデータなどである。これらは情報量の多さからビッグデータとも言われ、情報の蓄積と利用可能性の向上が進んでいる。経済分析においても、これらの潜在的な有用性が高まっている [15, 16, 30, 34]。

本稿では、日本の経済と物価の情報を多く含んでいる景気ウォッチャー調査の景気判断理由集のテキストデータに対し、統計的言語処理と機械学習の手法を用い、3つの分析を行った。同調査は、景気の現状および先行きに対する5段階の評価だけでなく、その評価の理由もコメントとして収集し、景気判断理由集として毎月公表している。この景気判断理由集は、すべて景気変動に関連した記述となっていることからノイズが少なく、景気分析で利用し易い優れたデータセットとなっている。

景気ウォッチャー調査の景況感を表す指数は景気判断で参照されるが、景気判断理由集については、有益な情報が多く含まれている可能性があるにもかかわらず、月あたり約10万語に及ぶ巨大なデータであることもあり、十分活用されてこなかった。本稿では、景気判断理由集から景気分析に有用な情報をより多く抽出できる手法を提案する。

1点目は、共起ネットワーク分析手法の応用である。共起ネットワーク図は、テキストデータにおける単語の共起関係を可視化したものである。先行研究である岡崎・敦賀 [31] の手法では、各時点での情報のみを用いて、共起ネットワーク図を作成している。この方法は、景気変動要因を可視化する目的を一定程度果たしている。すなわち、テキストデータ中で出現頻度の高い語に関連する景気変動要因については、可視化することができる。もっとも、この方法には、2つの短所がある。第1に、季節的に出現頻度が高まる単語が毎年特定の月に同じように抽出されてしまう。第2に、単月で出現頻度が高いことのみを単語の抽出基準とするため、データセットの性質として常に出現頻度が高い一般的な単語が恒常的に抽出されてしまう。こうした短所により、景気判断において重要となる、季節的な要因を除いた景気の変化に関する情報が少なくなってしまう。本稿では、こうした短所を補うため、抽出する単語を、過去の同じ月と比較して相対的に出現頻度の高い単語（特徴語）に絞る手法を提案する。特徴語のみによる共起ネットワーク図を用いた結果、先行研究で提案されている手法と比較して、より効果的に

景気変動要因を可視化できることが分かった。もっとも、この方法にも、先行研究の手法とは異なる欠点がある。この方法では、景気変動を説明するうえで重要な頻出語であっても、出現頻度の変化が乏しい場合には、重要な単語として抽出されず、共起ネットワーク図に現れないことがある。こうした場合には、先行研究の手法も併せて用い、クロスチェックを行うことも、重要な要因の見落としを防ぐ意味で有用である。

2点目は、ナイーブベイズ分類器を用いた物価センチメント指数の開発である。景気ウォッチャー調査の景気判断理由集の各コメントを物価の上昇または下落を表すものに分類し、物価動向のDI (Diffusion Index)「物価センチメント指数」を構築した。この手法の特徴は、物価変動に対する評価とテキストの対応関係を集めた教師データに対して事前に機械学習を行うことで両者の関係を推定し、テキストデータから物価変動の評価を効率的に抽出できる点である。また、この手法は、物価変動に対する評価が明確でないコメントが含まれていても概ね頑健に動作する。実際に、物価センチメント指数は、インフレ率の先行指標として有用であることが確認された。

3点目は、ランダムフォレスト回帰を用いた予想インフレ率およびインフレ率の変動要因の分析である。テキストデータを用いた分析では、予想インフレ率の変動と関連が深い単語を抽出することで、特定の理論に依拠することなく様々な要因の分析が可能となる。実際の抽出結果からも、物価変動を理解するうえで有益な情報が得られることが分かった。

本稿の構成は、以下の通りである。2章では、先行研究との関係を述べる。3章で、データについて説明する。4章では、共起ネットワーク分析を利用して主要な景気変動の要因の可視化を行う。5章では、ナイーブベイズ分類器を利用して物価動向のDIを構築する。6章では、ランダムフォレストを利用して中長期の予想インフレ率やインフレ率に対する予測力をもつ単語の抽出を行う。7章が結びである。

## 2. 関連する文献

テキスト分析や機械学習を金融経済分野で応用する研究は、近年蓄積が進んでおり、関連する文献は膨大である。中央銀行においても、テキスト情報の政策分析への利用を模索する動きが広がっている [6, 7, 8, 11, 20]。ここでは、本稿の分析と関連が深いと考

えられるものを挙げる。

テキスト分析の金融経済分野への応用の先駆けは、Tetlock [28, 29] である。彼らは、ニュース記事からセンチメント指標を構築し、同指標による株価リターンの予測可能性を分析した。沖本・平澤 [32] は、日本の株式市場を対象に同様の分析を行っている。これらの研究は、金融市場への応用事例である。

テキストデータ的一种である検索データを経済変数の直近値の予測 (nowcast) に利用する研究も行われている。Choi and Varian [12, 13] は、Google の検索数を利用して失業保険申請件数や自動車販売、旅行者数を予測した。Scott and Varian [25, 26] は、同じく Google の検索データの膨大な指標から予測に有用な指標を選定する枠組みを構築し、失業給付と小売販売額の予測に応用した。Matsumoto *et al.* [23] は、検索データが日本の旅行取扱額に対する予測力を有することを示した。これらの研究では、検索データなどを用いることで、短期の予測パフォーマンスを高められることが報告されている。

テキスト分析を、物価やインフレ予想の分析に応用する研究もある。Kulkarni *et al.* [21] は、Google の検索数を利用して住宅価格を予測した。Guzmán [17] は、検索データからインフレ予想を表す指数を構築し、サーベイに基づく予想インフレ率に対する予測力を有することを示した。Seabold and Coppola [27] は、検索データを消費者物価指数の直近値の予測に用いた。塩野 [37] は、ニュース記事のテキストデータを用いて日本の消費者物価指数の直近値の予測を行った。

テキスト情報から不確実性等を表す新たな金融経済指標を構築する研究もある。Alexopoulos and Cohen [1], Arbatli *et al.* [2], Azzimonti [4] は経済政策の不確実性指標をニュース記事から構築した。Baker *et al.* [5] は、不確実性指標が投資や雇用などのマクロ経済指標に対する予測力を持つことを示した。Caldara and Iacoviello [10] は地政学リスク指標をニュース記事から構築した。これらは、テキストデータが持つ独自の情報を利用して、現状把握ないしは予測を行う試みである。

景気ウォッチャー調査を用いたテキスト分析を、金融経済分野に応用する研究もある [33, 35, 36, 40, 43, 44]。山本・松尾 [44] は、景気ウォッチャー調査を教師データとして、これら以外の文章のセンチメントを判別できる深層学習に基づく機械学習モデルを構築した。

以上のような先行研究は、テキスト分析や機械学習を金融市場、景気動向の分析に適

用した例であるが、本稿の特徴は、景気ウォッチャー調査の景気判断理由集というマイクロテキストデータに対して、テキスト分析および機械学習を用いることにより、わが国全体のマクロ的な景気循環および物価変動のダイナミクス<sup>1</sup>に関する多くの示唆を得られることを示した点である。

### 3. 使用するデータ

景気ウォッチャー調査は、地域ごとの景気動向を迅速に把握するために、内閣府が2000年1月から毎月実施している日本全国を対象としたサーベイ調査である。回答者は約2,000人に及び、小売業販売員やタクシー運転手など地域の景気動向を観察できる業種の就業者（景気ウォッチャー）から選ばれる。各回答者は、現状と先行きの景況感をそれぞれ5段階で評価し、それぞれの景気判断理由を短くコメントする。

各回答者の現状と先行きの景気判断は、それぞれ現状と先行きの景況感の指数（DI）に集約される。各回答者のコメントを集めた景気判断理由集は、テキストデータとして公表される。景気判断理由集の分量は、単月の調査分で約10万語に及び、エコノミストが毎月精読するには分量が多いという意味で「ビッグデータ」である。

景気判断理由集は、景気分析に利用するうえで望ましい性質を持っている。景気判断理由集のテキストは、すべて景気動向に関連したコメントであるため、経済関連以外のテキストも多く含まれている新聞記事やウェブスクレイピングデータよりもノイズが少ない。回答者による言い回しのばらつき（表記ゆれ）に関しても一部にこれを抑制する処理が施されているほか、分析ソフトウェアでの処理が容易なファイル形式（CSV形式）で提供されているなど、分析に適したデータセットとなっている。さらに、地域や業種といった回答者属性の偏りが小さくなるようサンプリングされているほか、調査から公表までの期間も約2週間と短く速報性もある、といった長所もある。

---

<sup>1</sup> テキスト情報を利用していないが、機械学習によって予想インフレ率の個票データを解析し、マクロ経済変数と企業のインフレ予想形成との関係を分析したものに、日本銀行調査統計局経済調査課経済分析グループ [41] がある。

## 4. 分析 1 : 特徴語の共起ネットワーク分析によるトピック抽出

本章では、まず、先行研究の共起ネットワーク分析の手法の概要とその問題点について議論する。次に、過去の同じ月と比較して相対的に出現頻度の高い単語である「特徴語」を用いた共起ネットワーク分析の手法とその分析結果について説明する。後述するように、同手法は、先行研究の手法よりも景気変動要因の分析に適していると考えられる。

### 4-1. 共起ネットワーク図の作成方法と問題点

はじめに、共起ネットワーク図の作成手法を説明する。共起ネットワーク図は、点を単語に、辺を共起関係に対応させた無向グラフである。共起ネットワーク図の作成は、大きく次の2つのステップに分けられる。第1のステップでは、重要度の高い単語を抽出する。第2のステップでは、抽出した単語を対象として、文中において同時に出現し易い（共起関係が強い）もの同士を辺で結ぶ。第2のステップにおいて、辺を結ぶか否かを判定する基準となる共起関係の強さは、集合の類似度を表す Jaccard 係数 (Jaccard coefficient [22]) で測る。単語A と単語B の Jaccard 係数は

$$J(X_A, X_B) = \frac{|X_A \cap X_B|}{|X_A \cup X_B|} = \frac{\text{単語 A と B を同時に含む文の数}}{\text{単語 A または B を含む文の数}}$$

と定義される。ここで、 $X_A$  と  $X_B$  は、それぞれ単語A と単語B を含む文の集合である。

これまで共起ネットワーク図を景気動向の分析に応用した研究としては、岡崎・敦賀 [31] がある。彼らは、本分析と同様のデータセットである景気判断理由集を利用し、単語の共起ネットワーク図による景気動向分析を提案した。具体的には、重要度の高い単語の抽出に当たり、毎月の調査回毎のデータを独立に分析し、各月ごとに出現頻度の高い単語を抽出することにより、共起ネットワーク図（以下、「岡崎・敦賀の共起ネットワーク図」と呼ぶ）を作成する手法を採用した。

この方法は、出現頻度の高い語に関連する景気変動要因については、可視化することができる。もっとも、季節要因を除いた景気の変動を分析する観点からは、次の2つの短所がある。第1に、分析対象とするデータセットにおいて、出現頻度が季節によって毎年増加する単語が存在する場合、景気分析で必要となる季節要因を除去した景気動向の特徴が捉えにくい。例えば、景気判断理由集では、「ゴールデンウィーク」「クリスマ

ス」などの季節性の強い単語が、前年からの変化が無いにも関わらず、毎年特定月に共起ネットワーク図に現れるため、季節要因を取り除いた各時点の特徴を捉えることが難しい。第2に、分析対象とするデータセットの性質により、出現頻度の常に高い単語が、各時点で共通して抽出されるため、景気動向の時系列的な変化を捉えにくい。

#### 4-2. 特徴語共起ネットワーク図の作成

先に述べた問題点は、重要な単語の抽出基準が、単月の調査時点のみにおける出現頻度に依拠しており、他の時点のデータを参照していないことに起因する。これを解消するため、本分析では、当月分に加えて、過去の同じ月のデータも利用する方法を提案する。具体的には、過去の同じ月と比べて出現頻度が相対的に高い単語である「特徴語」を抽出し、特徴語のみを対象に共起ネットワーク図を作成する（以下、「特徴語共起ネットワーク図」と呼ぶ）。この手法では、データセットの性質により恒常的に出現頻度の高い単語や、季節性により特定月のみに必ず出現頻度が高まる単語の抽出を抑制できる。抽出された単語は例年からの相対的な出現頻度の高まりを示すことになるため、景気動向の分析により適した共起ネットワーク図の作成が可能となる。

ここでの特徴語の基準である「特徴度」は、過去の同じ月対比の相対的な出現頻度の高さを、集合の類似度を表す Dice 係数 (Dice coefficient [22]) で測ったものとする。すなわち、過去の同じ月の文すべてのうち単語  $W$  を含む文の集合を  $X_W$ 、当月のすべての文の集合を  $Y$  として、単語  $W$  の特徴度を

$$\text{単語}W\text{の特徴度} = D(X_W, Y) = \frac{2|X_W \cap Y|}{|X_W| + |Y|}$$

と定義する<sup>2</sup>。

これは、当月の単語  $W$  を含む文の数を  $a$  個、過去の同じ月における  $W$  を含む文の数を  $b$  個、当月の  $W$  を含まない文の数を  $c$  個として、次のように簡単に計算できる：

---

<sup>2</sup> Dice 係数以外の特徴語の抽出基準として、ある単語の過去平均的な出現確率と当月の出現確率の比率（自己相互情報量）に基づくもの [14] が使用されることもある。確率比に基づく特徴度では、過去に一度も出現していない単語が、当月に一度でも出現すると特徴語の基準を満たしてしまう。このため、当月のテキストから景気変動にある程度以上の影響力のあるトピックを抽出する本分析の目的には適していない。



$$D(X_W, Y) = \frac{2a}{(a+b) + (a+c)} .$$

単語  $W$  が、共起ネットワーク図に登場するためには、この特徴度が高い必要がある。すなわち、当月の単語  $W$  を含む文の数  $a$  が過去の同じ月における  $W$  を含む文の数  $b$  と比較して相対的に高いことが必要となる。従って、データセットの性質から恒常的に出現頻度が高い単語や、季節的な特徴により当該月に出現頻度が高まる単語は、 $a$  対比で  $b$  が相対的に大きくなるため、重要な単語として抽出されにくい。また、ここで用いる特徴度は、過去のデータ対比での出現頻度だけでなく、当月における  $W$  を含まない文の数  $c$  対比での出現頻度の高さも加味される<sup>3</sup>。

ただし、特徴語共起ネットワーク図に特有の限界として、景気変動を説明するうえで重要な頻出語であっても、出現頻度の高い状態がある程度長期にわたるなど、出現頻度の変化が乏しい場合には、重要な単語として抽出されず、共起ネットワーク図に現れないことがあるという点が挙げられる。したがって、景気分析の実務では、特徴語共起ネットワーク図だけでなく、岡崎・敦賀の共起ネットワーク図も用いてクロスチェックを行うことが、重要な要因の見落としを防ぐ意味で有効である。

#### 4-3. 景気判断理由集の特徴語共起ネットワーク図

前節で定めた基準により抽出した特徴語を使って、特徴語共起ネットワーク図を作成する。共起関係の強さは、Jaccard 指数が 0.06 以上であることを基準とし、表示する単語の候補を約 100 語とした。これらの設定は、特徴語共起ネットワーク図において、各月の主要な景気変動の特徴（トピック）に対応する部分ネットワークが視認できる程度にネットワーク全体の疎性が維持されるよう選んだ。岡崎・敦賀 [31] も 0.06 と同程度の基準を採用している。表示する単語の候補数は、岡崎・敦賀 [31] の約 50 語よりも増やした<sup>4</sup>。これらの基準は、抽出したいトピックや用途に応じて調整することが可能である。

景気の現状判断に関するコメントを利用した特徴語共起ネットワーク図（図 1）は、

<sup>3</sup> 作成には KH Coder を用いた。KH Coder による作成方法の詳細は補論 A、B を参照。

<sup>4</sup> 表示する語を特徴語のみに絞ることにより、出現頻度が常に高く、他の語との共起関係も強い一般的な単語の出現が抑えられ、比較的ネットワークが疎となる。このため、視認性を保ったまま表示する単語を増やすことが可能となった。

岡崎・敦賀の共起ネットワーク図（図2）と大きく異なる結果となっており、当該月の景気変動の特徴をよりの確に捉えられていると考えられる。例えば、図1の③の部分ネットワークにはインバウンド需要のトピックが現れている。この部分ネットワークには、「インバウンド」「化粧（品）」「高額」「好調」といった単語が現れており、インバウンド需要により化粧品や高額商品の販売が好調であるといったトピックの具体的な内容が読み取れる。他方、同一のトピックに対応している図2の③の部分ネットワークは、「高額」と「商品」の2つの単語のみから成るため、トピックに関する情報量が少ない。「インバウンド」「化粧（品）」といった単語は、各月での出現回数が少ないため、図2の岡崎・敦賀の共起ネットワーク図には現れない。他方、これらの単語は、過去対比で相対的に出現回数が増加していることから、図1の特徴語共起ネットワーク図では重要な単語として捉えられていると考えられる。この他、同様の理由で、特徴語共起ネットワーク図では、④の部分ネットワークである株価上昇のトピックのような当該月を特徴づけるトピックが抽出されている。他方、岡崎・敦賀の共起ネットワーク図ではこうしたトピックは抽出されていない。景気の先行き判断に関するコメントについても、上記の現状判断の場合と同様に、より各月の景気変動を捉えるうえで重要な単語が抽出されていることが確認できる（図3，4）。

ただし、特徴語共起ネットワーク図と岡崎・敦賀の共起ネットワーク図に共通する限界として、最近のネットワーク図から読み取れる物価動向の情報が乏しい点が挙げられる。特徴語共起ネットワーク図では、物価動向の単語がほとんど現れていない（図1，3）。これは、前節で述べた特徴語共起ネットワーク図に特有の限界によるものである。すなわち、物価そのものを表す単語（「価格」「単価」「物価」）が、出現頻度の変化の乏しさから、特徴語の基準を満たさないため、特徴語共起ネットワーク図に現れないことに起因すると考えられる（図5，6）。岡崎・敦賀の共起ネットワーク図でも、物価に関しては、「価格」「単価」といった一般的な単語と「上昇」「下落」といった変動を表す単語との共起関係の有無程度の情報しか得られない（図2，4）。このように、共起ネットワーク分析のアプローチでは、物価動向をうまく捉えることができない。

そこで、次の分析では、機械学習を用いた新しいアプローチで物価情報の抽出を試みる。

## 5. 分析2：物価センチメント指数の構築

本章では、物価センチメント指数の作成方法と結果を説明する。物価センチメント指数は、ナイーブベイズ分類器を用い、各コメントを言及された物価動向に応じて分類することにより作成する。ナイーブベイズ分類器の学習には、過去のコメントの一部を手動で分類した教師データを用いる。本分析の特徴は、既存の物価指数を用いることなく、テキストデータのみから、物価に先行する指数を構築した点である。

### 5-1. 景気判断理由集からの物価動向の抽出

景気ウォッチャー調査には、景気動向のDIがある一方、物価動向については数値的な評価やDIが存在しない。もっとも、景気判断理由集では、景気動向に関してコメントする中で、物価動向についても言及されているものが多くみられる。こうした情報を集約し、景気ウォッチャー調査全体として物価上昇と下落のどちらが優勢であるかを把握することで、わが国の物価動向について示唆を得ることは可能だろうか。

こうした問いに答えるにあたり、まず、物価動向を定量化する単純なアプローチであるワードカウントでは物価動向をうまく抽出できないことを確認しておく。図7は「値上げ」「値下げ」の出現頻度を示しているが、振れが大きく、物価動向を読み取ることができない。他の単語についても同様の結果が得られる。この原因は、(1)物価動向の表現が多様であること、(2)回答者によって語彙選択にばらつきがあること、(3)否定形での言及や文脈によって同じ表現でも意味合いが変化することに起因する。

主に前者2つの表現の多様性に対応するため、機械学習手法を用いて、各コメントを言及された物価動向に応じて分類し、この分類に基づいて物価動向のDIを構築することを試みる。以下では、これを「物価センチメント指数 (Price Sentiment Index, PSI)」と呼ぶことにする。各コメントを分類するためのカテゴリは、物価への言及がない場合も考慮して、「a. 物価上昇」「b. 物価下落」「c. 物価に言及はあるが方向感なし (横ばい)」「d. 物価への言及なし」の4つとする。物価センチメント指数は、カテゴリ a のコメント数を  $\alpha$ 、カテゴリ b のコメント数を  $\beta$ 、カテゴリ c のコメント数を  $\gamma$  として、物価上昇のコメントの割合と物価下落のコメントの割合の差をもとに、

$$\text{物価センチメント指数} = \frac{\alpha - \beta}{\alpha + \beta + \gamma} \times 100$$

と定義する.

## 5-2. ナイーブベイズ分類器

物価変動に関するコメントの分類には、教師あり分類器の一つであるナイーブベイズ分類器 (Naïve Bayes classifier) を利用する [24, 39]. 景気判断理由集のデータの特徴は、コメント数がコメントに使用される単語の種類よりも少ないことである. 例えば、2017/10月の景気判断理由集 (現状判断) は、約 1,300 個のコメントに、約 3,000 種類の異なる単語を含む. ナイーブベイズ分類器は、こうしたデータのサンプルが特徴量の次元対比で少ない場合でもある程度頑健な推定結果を得られることから、今回のような分析に適している. また、ナイーブベイズ分類器は構造が単純であるほか、各コメントの分類の理由が分かり易い. こうした特徴は、分類の根拠を直接的に解釈することが本手法より難しいサポートベクターマシンやランダムフォレストといった他の教師あり分類器よりも優れた点である. 特に、景気分析では物価センチメント指数を計算することに加えて、それが示唆する物価動向を経済現象として解釈する必要があるため、コメントの分類の根拠が分かり易いことは重要である.

ナイーブベイズ分類器は、コメント  $d$  がカテゴリ  $c \in C$  に属する条件付き確率  $P(c|d)$  を最大化するカテゴリ  $c^* = \operatorname{argmax}_{c \in C} P(c|d)$  を返す. 条件付き確率  $P(c|d)$  を計算することは一般には難しいが、生成モデルに構造的制約を課すことで、コメント  $d$  がカテゴリ  $c$  に属する事後確率  $P(c|d)$  は、教師データにおけるカテゴリ別の各単語の出現回数のみから計算できる<sup>5</sup>.

## 5-3. 推定結果

ナイーブベイズ分類器の訓練に用いる教師データは、景気ウォッチャー調査 2000 年 1 月～2017 年 10 月分のコメント約 40 万個の中から、景気の実況判断と先行き判断についてそれぞれ約 1,500 個を抽出し、手動で分類して作成した. 分類のサンプルを図 8, 9 に示している. 例えば、「商品を値上げしたことで客単価が上がっている. 来客数も増えている.」といった物価上昇に言及しているコメントは、「a. 物価上昇」に分類する. 同様に、「b. 物価下落」「c. 物価もちあい」に言及しているコメントを、それぞれのカ

---

<sup>5</sup> 推定方法の詳細は、補論 A, C を参照.

カテゴリに分類する。一方で、物価の上昇、下落、もちあい、いずれにも言及していないコメントは、「d. その他」に分類する。この教師データから、各カテゴリの生起確率およびカテゴリ毎に各単語の条件付き生起確率を算出し、これらに基づいてすべてのコメントを分類する。直感的には、例えば、教師データにおいて「物価上昇」カテゴリでの出現確率が高かった単語を多く含むコメントほど、「物価上昇」カテゴリに属する確率が高くなる。

分類器の学習結果である各単語のカテゴリ毎の条件付き生起確率（スコア<sup>6</sup>）をみると、単語の意味から直感的に予想される大小関係と概ね整合的であることが確認できる（図10, 11）。物価上昇に関連する「高騰」「値上げ」や、エネルギー価格などコストに関連する「(価格) 転嫁」「燃料」「ガソリン」といった単語を含むコメントは、物価上昇を表すカテゴリ a に分類される確率が高い。逆に、「値下がり」「下落」「(競争) 激化」「デフレ」等を含むコメントは価格下落を表すカテゴリ b に分類される確率が高い。分類器の性能を評価するために、教師データを母集団とみなして5分割交差検証<sup>7</sup>を行った結果は、アウトオブサンプルでの平均正答率が現状判断で74.2%、先行き判断で76.1%となった<sup>8</sup>（図10, 11）。後述するように、物価センチメント指数と消費者物価指数が高い連動性を示すことを考慮すると、ナイーブベイズ分類器の性能は、目的に照らして十分であると評価できる。

算出された物価センチメント指数は、消費者物価指数と高い連動性を示している（図12）。両者の相関係数は約0.7と高く、短期的な変動の方向も概ね合致している。このことから、景気ウォッチャーの景気判断理由集は、わが国全体の物価動向についての情報を有していると考えられる。

---

<sup>6</sup> 図中のスコアは、条件付き生起確率の対数値を、最小値が0となるようスケールした値である。

<sup>7</sup> 交差検証では、教師データの一部をテストデータとみなし、それらを学習には用いず学習結果の汎化誤差（generalization error）の評価のみに用いる。5分割交差検証は、教師データを5分割し、その1つをテストデータとして用いた評価を5回繰り返す交差検証を意味する。

<sup>8</sup> 図中で示した Kappa 統計量  $\kappa$  は、正解の度合いを評価する指標であり、以下のように求められる：

$$\kappa = \frac{(\text{正答率}) - (\text{偶然正解する確率})}{1 - (\text{偶然正解する確率})}$$

すなわち、ランダムに分類しても正解したであろう程度の正答率の場合には0を、すべてのデータについて正解の場合に1の値をとる。

次に、物価センチメント指数と消費者物価指数の時差相関を調べると、いずれの消費者物価指数に対しても、物価センチメント指数が先行している（図13, 14）。グレンジャー因果検定でも12か月先の消費者物価に対して有意に説明力を持つ（図15）。従って、物価センチメント指数は、わが国の先行きの物価動向の情報を有していると考えられる。景気の現状判断のコメントが、物価の先行きの情報を有する理由として、回答者がまだ価格設定に反映していないコストの変動要因、価格設定スタンス、需給要因などがセンチメントに現れており、そうした要因がタイムラグを伴って徐々に将来の価格にも織り込まれていくためと考えられる。

なお、物価センチメント指数は、消費者物価指数（総合除く食料・エネルギー）よりも消費者物価指数（総合）との相関が高い。このことは、景気判断理由集がエネルギーや生鮮食品の価格変動に関する言及を含むことが影響していると考えられる。

本章では、景気判断理由集から示唆される物価動向の定量的な指標を作成した。テキスト分析の特色として、分析者が必要に応じて、物価センチメント指数の変動要因について、学習結果の各単語のスコアをもとに解釈できるほか、実際に分類された各コメントを見て考察できることが挙げられる。

この手法には制約と留意点もある。第1に、各カテゴリにおける単語の生起確率の時間的な変化を考慮していない。物価動向に新しい現象が生じて、新たな語彙が多く使われるようになった場合には、教師データにサンプルを追加すべきである。第2に、ナイーブベイズ分類器は、構造が単純であるため、否定形による言及や文脈情報を処理できない。テキストの語順を考慮でき、文脈情報のある程度捉えられる再帰型ニューラルネットワーク（Recurrent Neural Network）のような深層学習手法を利用すること等は今後の検討課題である。第3に、景気ウォッチャーのコメントは景気動向に対する評価であるため、物価動向に対する評価を読み取ることは人間でも難しい場合がある。従って、教師データの分類の正しさは、物価動向に関する記述の曖昧さに原理的に制約される点に留意すべきである。第4に、景気判断理由集では、データの性質上、景気動向に影響されにくい定例的な価格改定などの物価の情報が含まれない可能性がある。

## 6. 分析3：テキストデータを用いたインフレ率の変動要因の分析

### 6-1. 予想インフレ率・インフレ率に影響を与える語を抽出するモデル

ここでは、景気ウォッチャー調査の景気判断理由集から、予想インフレ率とインフレ率の変動に説明力のある単語を機械学習により自動で抽出する。前章でみたように、物価センチメント指数は物価との同時相関が高だけでなく、物価に先行する関係にもあることから、現状や先行きの物価変動の要因を理解するうえで有益な情報が含まれている可能性がある。

ここでは、景気判断理由集から予想インフレ率とインフレ率の決定要因を分析する枠組みとして、景気判断理由集の単語の出現頻度を入力変数としたランダムフォレストにより、予想インフレ率とインフレ率を予測するモデルをそれぞれ推定する。この推定結果から、それぞれの変数に対する説明力の高い単語を選別することができるため、そうした単語をもとに変動要因の分析が可能となる。

### 6-2. ランダムフォレスト

ランダムフォレスト [9, 19, 38] は、多数の決定木 (decision tree) による推定値を平均することで、入力変数 (特徴量) から目的変数を予測するモデルである。

決定木による回帰分析には、一般的な線形モデルなどによる回帰分析と比較して、いくつかの長所がある。まず、決定木は、二分木の性質と変数と閾値の選択方法から、入力変数の外れ値に対して頑健であるほか、入力変数の中に目的変数との関係が乏しいものが多く含まれる場合でも安定的に推定値を得られる。この点は、予想インフレ率やインフレ率との関係が乏しい多種類の単語を入力変数に含む本分析に適している。また、決定木は、その分岐構造自体が入力変数と目的変数の関係を表すため、解釈が容易である。もっとも、単一の決定木による推定精度は低い<sup>9</sup>ほか、決定木の分岐構造は再帰的であるため、入力変数値の僅かな違いで学習される分岐構造全体が大きく変化する可能性がある。

ランダムフォレストは、決定木の持つ長所を生かしつつ、これらの欠点を補うため、

---

<sup>9</sup> これは、決定木が、計算負荷を抑制する観点から、局所最適な選択を繰り返して全体の解を求めるアルゴリズムの考え方 (貪欲法, greedy algorithm) に基づいて構築されるためである。

復元抽出により生成したデータに対して決定木を複数生成し、その決定木を組み合わせることで、計算負荷を抑制しつつ精度の高い推定を可能にする手法である<sup>10</sup>。

### 6-3. 分析結果

本分析では、目的変数に Consensus Economics 「コンセンサス・フォーキャスト」の、6～10年先予想インフレ率と、インフレ率（消費者物価指数、総合除く生鮮食品・消費税）を用いた。予想インフレ率については、3か月ないしは6か月ごとの調査であるため、調査が行われなかった月のデータを線形補間して用いた。生成する決定木の数は、十分に多い  $B = 500$  とした。各決定木の入力変数の候補の個数は、概ね  $p = 200$  とした。これは、交差検証をもとにグリッドサーチで最適値として求めた。

モデル推定の頑健性を確認するために、予想インフレ率とインフレ率について、 $K$  分割交差検証を行った結果が図 16, 17 である。横軸はテストデータ期間（アウトオブサンプルの予測期間）の長さであり、横軸の値が大きいほど交差検証における教師データの分割数が少なく、期間が長くなる。ベンチマークとして AR(1) 過程を用いた予測結果と比較した。予想インフレ率とインフレ率に共通する結果として、短い期間の予測では、ランダムフォレストの予測精度は AR(1) に劣る。もっとも、長めの期間では、ランダムフォレストの方が高い推定精度となっており、景気判断理由集が予想インフレ率とインフレ率の予測精度の改善に寄与していることが確認できる。

次に、予想インフレ率とインフレ率の予測で寄与の大きい（変数重要度の高い）単語を抜き出す。予想インフレ率では、原材料価格や人手不足、為替などに関連する単語の重要度が高いとの結果となった（図 18）。インフレ率では、消費増税関連など物価の比較的短期の変動要因に関連する単語の重要度が相対的に高いとの結果となった（図 19）。

これらの単語の予想インフレ率およびインフレ率への寄与を定量的に測るために部分従属度をプロットする。例えば、「人手不足」は、予想インフレ率とインフレ率の両方の場合で右肩上がりの曲線になっている。これは、「人手不足」という単語の頻度が高まると予想インフレ率およびインフレ率が平均的に高まることを示している。こうした部分従属プロットと、各単語の出現頻度の時系列データ（図 20, 21）を利用すれ

---

<sup>10</sup> 推定方法の詳細は、補論 A, D を参照。



ば、予想インフレ率およびインフレ率の変動に寄与してきた語の変遷を知ることができる。

部分従属プロットの見方には、留意点がある。ある語の予想インフレ率に対する寄与の大きさと方向を判断する際には、ほかの語との関係（交互作用）を本来は考慮する必要がある。例えば、「物価」「上昇」という2つの単語が同時に増えた場合と、「物価」「下落」という2つの単語が増えた場合の「物価」の寄与は当然異なる。部分従属プロットは、こうした交互作用を平均することで周辺化したものであるため、部分従属プロットの変動の大きさと、変数重要度は必ずしも一致しない。

部分従属度から得られる予想インフレ率およびインフレ率への寄与を視覚的に表現するためにワードクラウドを利用する。ワードクラウドを用いて、各月において出現頻度が高く、予想インフレ率の推計に寄与の大きかった単語を1枚の図にまとめた（図2-2）。文字の大きさは、部分従属度を用いて算出した予想インフレ率およびインフレ率に対する影響度<sup>11</sup>の絶対値を表す。赤色はプラス方向、青色はマイナス方向の寄与であることを表す。ただし、部分従属プロットと同様、この色分けも、交互作用を周辺化したものである点には留意が必要である。

2018年1月調査の分析結果では、労働需給の逼迫に関する「人手不足」や、景況感全般や財・サービスの需要に関する「好調」「インバウンド」「観光」「春物」、原材料価格の上昇に関する「原料」「燃料」「高騰」「食料」などが幅広くプラス方向に寄与している一方、求職者の増加に関連する「求職」がマイナス方向に寄与している。2014年7月調査では、「消費税」「増税」「駆け込み」「転嫁」など消費税率の上昇要因がプラス方向に寄与している。2012年10月調査では、「デフレ」「円高」「冷え込む」など景況感の悪さや市況を示す単語のマイナス方向の寄与が目立つ。2007年10月調査では、「ガソリン」「原油」「高騰」といった原油価格上昇に関連する単語のプラス寄与がある一方で、「競合」「倒産」といった企業の競争環境に関連する単語がマイナス方向に寄与している。

次に、図2-3は、消費者物価指数の実績値データに適用した結果である。全体としては、予想インフレ率の場合と概ね同様のプラス方向の寄与の単語がみられる。異なる点を挙げると、2018年1月調査では求職者の増加に関連する「求職」のマイナス寄与が見

---

<sup>11</sup> 影響度とは、ある単語が過去平均並みに出現する場合の部分従属度と同単語の当月の部分従属度の差。

られないほか、2007年10月調査では「競合」「倒産」といった競争環境関連のマイナス寄与が相対的に小さい。

## 7. 結び

本稿では、景気ウォッチャー調査のテキストデータに対し、統計的言語処理と機械学習の手法を用い、3つの分析を行った。第1に、過去対比で相対的に出現頻度が高い特徴語に対する共起ネットワーク図により、景気変動要因の可視化を行った。第2に、ナイーブベイズ分類器でテキストを分類することにより構築した物価動向を表す指数を作成した。第3に、予想インフレ率や物価変動に対して説明力がある単語を、ランダムフォレストの変数重要度を基準に抽出することにより、物価変動に影響を与える要因を分析した。これらの分析結果から、統計的言語処理と機械学習によるテキスト分析は、従来の経済理論に基づいた分析手法を補完しうるものであり、景気動向やインフレ率の分析ツールとして実務的に有用であることが確認された。さらに、こうした手法は、従来のデータを用いた分析ではできなかったファクトファインディングを可能とすることで、経済理論の発展にも寄与し得るものと考えられる。

景気分析におけるテキスト分析やビッグデータの利用には、制約と限界もある。第1に、テキスト分析は、景気分析の軸である経済指標に対する定量的分析の代替手段とはならない。定量化したテキストデータと経済現象の規模との対応関係は一般に明らかではない。また、テキストデータは執筆者の主観にも影響される。従って、テキスト分析により得られた定量的な分析結果は、回帰分析等で有意に説明力をもっていたとしても過度な信頼はできない。第2に、英中銀チーフエコノミストの Haldane も指摘するように、ビッグデータと機械学習を利用したモデルは、理論的制約が弱く、既知のデータから得られる経験則に過度に依存してしまうことにより、未知のデータに対する予測パフォーマンスがむしろ悪化してしまうリスクがある [18]。

こうした議論を踏まえると、景気分析におけるテキスト分析の位置づけは、公的統計等を利用した従来の分析手法にはない新しい角度から有力な材料を提供することで、景気認識を容易にし、景気判断の精度を向上させるための補完的な役割を果たし得るものと考えられる。今後、経済理論による構造分析とビッグデータを用いた実証分析

の補完関係 [18] を意識しながら、機械学習手法を景気分析に活用していくことが重要である<sup>12</sup>.

---

<sup>12</sup> Haldane [18] は、より一般的に、機械学習が経済学に与える影響についても言及している。また、Athey [3] は、標準的な計量経済学的手法と機械学習の違いについて議論している。

## 参考文献

- [1] M. Alexopoulos, and J. Cohen (2015), "The power of print: Uncertainty shocks, markets, and the economy," *International Review of Economics and Finance*, Vol. 40, Issue C, pp. 8-28.
- [2] E. C. Arbatli, S. J. Davis, A. Ito, N. Miake, and I. Saito (2017), "Policy uncertainty in Japan," *NBER Working Paper*, No. 23411.
- [3] S. Athey (2018), "The impact of machine learning on economics." In A. K. Agrawal, J. Gans, and A. Goldfarb (Eds.), *The Economics of Artificial Intelligence: An Agenda*, University of Chicago Press.
- [4] M. Azzimonti (2018), "Partisan conflict and private investment," *Journal of Monetary Economics*, Vol. 93, Issue C, pp. 114-131.
- [5] S. R. Baker, N. Bloom, and S. J. Davis (2016), "Measuring economic policy uncertainty," *The Quarterly Journal of Economics*, Vol. 131, Issue 4, pp. 1593-1636.
- [6] Bank of England (2015), "One bank research agenda," February 2015.
- [7] D. Bholat, S. Hansen, P. Santos, and C. Schonhardt-Bailey (2015), "Text mining for central banks: Handbook," *Centre for Central Banking Studies*, No. 33, pp. 1-19.
- [8] B. Bok, D. Caratelli, D. Giannone, A. M. Sbordone, and A. Tambalotti (2018), "Macroeconomic nowcasting and forecasting with big data," *Annual Review of Economics*, Vol. 10, pp.615-643.
- [9] L. Breiman (2001), "Random forests," *Machine Learning*, Vol. 45, pp. 5-32.
- [10] D. Caldara, and M. Iacoviello (2018), "Measuring geopolitical risk," *International Finance Discussion Papers, Board of Governors of the Federal Reserve System*, No. 1222.
- [11] C. Chakraborty, and A. Joseph (2017), "Machine learning at central banks," *Bank of England Staff Working Paper*, No. 674.
- [12] H. Choi, and H. Varian (2009), "Predicting initial claims for unemployment benefits," *Technical Report, Google Inc.*
- [13] H. Choi, and H. Varian (2012), "Predicting the present with Google trends," *Economic*

*Record*, Vol. 88, Issue s1, pp. 2-9.

- [14] K. W. Church, and P. Hanks (1990), "Word association norms, mutual information, and lexicography," *Computational Linguistics*, Vol. 16, Issue 1, pp. 22-29.
- [15] L. Einav, and J. Levin (2014), "The data revolution and economic analysis." In J. Lerner, and S. Stern (Eds.), *Innovation Policy and the Economy*, Vol. 14, University of Chicago Press.
- [16] M. Gentzkow, B. T. Kelly, and M. Taddy (2017), "Text as data," *NBER Working Paper*, No. 23276.
- [17] G. Guzmán (2011), "Internet search behavior as an economic forecasting tool: The case of inflation expectations," *Journal of Economic and Social Measurement*, Vol. 36, Issue 3, pp. 119-167.
- [18] A. Haldane (2018), "Will big data keep its promise?," Speech at the Data Analytics for Finance and Macro Research Centre, King's Business School, 19 April 2018.
- [19] T. Hastie, R. Tibshirani, and J. Friedman (2009), *The elements of statistical learning - Data mining, inference, and prediction, Second Edition*, Springer.
- [20] Irving Fisher Committee on Central Bank Statistics (2015), "Central bank's use of and interest in 'big data'," *IFC Report*, October 2015.
- [21] R. Kulkarni, K. E. Haynes, R. R. Stough, and J. H. P. Paelinck (2009), "Forecasting housing prices with Google econometrics," *George Mason University School of Public Policy Research Paper*, No. 2009-10.
- [22] C. D. Manning, and H. Schütze (1999), *Foundations of statistical natural language processing*, MIT Press.
- [23] A. Matsumoto, K. Matsumura, and N. Shiraki (2013), "Potential of search data in assessment of current economic conditions," *Bank of Japan Research Papers*, No. 2013-04-18.
- [24] K. P. Murphy (2012), *Machine learning: A probabilistic perspective*, MIT Press.
- [25] S. L. Scott, and H. R. Varian (2014), "Predicting the present with Bayesian structural time series," *International Journal of Mathematical Modelling and Numerical Optimization*, Vol. 5, Issue 1-2, pp. 4-23.

- [26] S. L. Scott, and H. R. Varian (2015), "Bayesian variable selection for nowcasting economic times series." In A. Goldfarb, S. M. Greenstein, and C. E. Tucker (Eds.), *Economic analysis of the digital economy*, University of Chicago Press.
- [27] S. Seabold, and A. Coppola (2015), "Nowcasting prices using Google trends: An application to Central America," *Policy Research Working Paper*, No. 7398.
- [28] P. C. Tetlock, (2007), "Giving content to investor sentiment: The role of media in the stock market," *Journal of Finance*, Vol. 62, pp. 1139-1168.
- [29] P. C. Tetlock, M. Saar-Tsechansky, and S. Macskassy (2008), "More than words: Quantifying language to measure firms' fundamentals," *Journal of Finance*, Vol. 63, Issue 3, pp. 1437-1467.
- [30] H. R. Varian, (2014), "Big data: New tricks for econometrics," *Journal of Economic Perspectives*, Vol. 28, No. 2, pp. 3-28.
- [31] 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—」, 日本銀行調査論文.
- [32] 沖本竜義・平澤英司 (2014) 「ニュース指標による株式市場の予測可能性」, 証券アナリストジャーナル, 第 52 巻, 第 4 号.
- [33] 風戸正行・黒崎哲夫・五島圭一 (2018) 「金融政策のトーン分析：日本銀行の政策説明とメディア報道」, 日本銀行金融研究所ファイナンス・ワークショップにおける研究報告資料.
- [34] 株式会社 NTT データ経営研究所 (2015) 「公的統計におけるビッグ・データの活用に関する調査研究報告書」, 平成 26 年度内閣府大臣官房統計委員会担当室請負調査.
- [35] 五島圭一・山田哲也・高橋大志 (2017) 「畳み込みニューラルネットワークを用いた日次景況感指数の構築と資産価格変動との関連性」, 日本ファイナンス学会第 25 回大会予稿集.
- [36] 小寺信也・藤田隼平・井上祐介・新田堯之 (2018) 「POS・テキストデータを用いた消費分析—機械学習を活用して—」, 経済財政分析ディスカッション・ペーパー・

シリーズ, No. DP/18-1.

- [37] 塩野剛志 (2018) 「人工知能とテキスト・データを活用した数量分析」, 日本銀行金融研究所ディスカッション・ペーパー・シリーズ, No. 2018-J-9.
- [38] 杉山将 (2013) 「イラストで学ぶ機械学習: 最小二乗法による識別モデル学習を中心に」, 講談社.
- [39] 高村大也 (2010) 「言語処理のための機械学習入門」, 奥村学監修, コロナ社.
- [40] 敦賀智裕・岡崎陽介 (2017) 「テキストデータを用いた金融経済分析—サーベイと景気ウォッチャー調査を用いた分析」, 統計, 2017年4月号.
- [41] 日本銀行調査統計局経済調査課経済分析グループ (2017) 「企業のインフレ予想形成に関する新事実: Part II—機械学習アプローチ」, 日本銀行ワーキングペーパーシリーズ, No. 17-J-4.
- [42] 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」, ナカニシヤ出版.
- [43] 山澤成康 (2018) 「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」, ESRI Discussion Paper Series, No. 345.
- [44] 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」, 人工知能学会全国大会 (30回).

## 補論

以下では、本稿で行った分析の実装の詳細について解説する。前処理にあたる形態素解析や BoW (Bag of Words) ベクトルの出力、ナイーブベイズ分類器の学習、共起ネットワーク図の描画には、計量テキスト分析のためのフリーソフトウェアである「KH Coder (Ver. 2.00f)」<sup>13</sup>を用いた [42]。形態素解析システムには、「ChaSen」<sup>14</sup>を用いた。ランダムフォレストの学習には、統計処理ソフトウェア「R」<sup>15</sup>を用いた。

### A. データの前処理

使用するデータは、2000年1月以降の景気ウォッチャー調査の景気判断理由集の CSV 形式のファイルである。景気の現状判断については「追加説明及び具体的状況の説明」、先行き判断については「景気の先行きに対する判断理由」の列のテキストのみを切り出し、1つのテキストファイルに集約する。

次に前処理を行う。まず機種依存文字等を形態素解析が可能な文字に置換する。次に、これらに対して、調査回毎に見出し (KH Coder の「H1 タグ」で指定) を付与して、各コメントがどの調査回のものであるかを区別する属性情報を付加する。このような処理を調査回毎に行い、KH Coder への入力用として1つのテキストファイルにまとめる。1 コメントを1段落として、改行で分割し、以下のような形式に整形する。

<H1>調査年月：2000年1月</H1>  
価格が低下したことにより購買意欲が上向いており、売上点数が増加している。  
2～3か月前と比べると、売上が伸びている。12月に比べると、2倍の伸びを示している。  
3か月前と同様に、客は必要な物でも慎重な購買を行っている。今月スタートの冬物値下げで商品価格が下がっても、客は衝動買いをしていない。  
今月は雪が多いせいもあり、良いとはいえない。客単価も悪く、店の状態をみる限り、景気は良いとはいえない。

(以下略)

<sup>13</sup> <http://khc.sourceforge.net/>

<sup>14</sup> <http://chasen-legacy.sourceforge.jp/>

<sup>15</sup> <https://www.r-project.org/>



KH Coder に上記で作成したテキスト形式ファイルを読み込み、形態素解析を行う  
 <KH Coder での操作：「プロジェクト」→「新規」→「分析対象ファイル参照」→「開  
 く」>. これにより、分析対象ファイル内の文章から単語が切り出され、その結果から  
 データベースが作成される.

形態素解析を行う際、辞書に含まれていない主要な語（例：「インバウンド」「スマー  
 トフォン」等）については、別途リストを与え、強制的に単一の単語として抽出する。  
 また、景気判断理由集でどのような文にも含まれ得る些末な語（例：「今月」「思う」「感  
 じる」等）については、共起ネットワーク図の見やすさの観点から、分析に使用しない  
 語として指定する<「前処理」→「語の取捨選択」→「強制抽出する語の指定」/「使  
 用しない語の指定」,「前処理」→「前処理の実行」>. 形態素解析の結果を確認すると、  
 以下のようにになっている<「前処理」→「語の抽出結果を確認」>.

ID:2 価格 / が / 低下 / し / た / こと / に / より / 購買意欲 / が / 上向い / て / おり / , / 売上 / 点数 / が / 増加 / し / て / いる / .			
表層語	基本形	品詞	品詞 (ChaSen による出力)
価格		名詞	名詞-一般
が		その他	助詞-格助詞-一般
低下		サ変名詞	名詞-サ変接続
し	する	動詞 B	動詞-自立 連用形
た		その他	助動詞 基本形
こと		その他	名詞-非自立-一般
に		その他	助詞-格助詞-一般
より	よる	動詞 B	動詞-自立 連用形
購買意欲		タグ (強制抽出語)	
が		その他	助詞-格助詞-一般
上向い	上向く	動詞	動詞-自立 連用タ接続
て		その他	助詞-接続助詞
おり	おる	その他	動詞-非自立 連用形
,		その他	記号-読点
売上		名詞	名詞-一般
点数		名詞	名詞-一般
が		その他	助詞-格助詞-一般
増加		サ変名詞	名詞-サ変接続
し	する	動詞 B	動詞-自立 連用形
て		その他	助詞-接続助詞
いる		その他	動詞-非自立 基本形
.		その他	記号-句点
(以下略)			

景気判断理由集には、毎月新たな未知語が出現するため、必要に応じてリストを更新する。

## B. 特徴語の共起ネットワーク分析の詳細

KH Coder で特徴語抽出を行う。具体的には、分析対象月と過去の同じ月のデータからなる入力ファイルを前処理した後、データに付与した調査回毎の見出し属性に基づき、分析したい調査回を選択する<「ツール」→「外部変数と見出し」→「リスト」→（分析したい調査回を選択）→「特徴語」>。特徴語は、分析に利用しにくい助詞、助動詞、平仮名のみからなる語を除き、分析対象月を含む過去の同じ月対比での Dice 係数の上位 100 語程度を抽出する<「関連語検索」→「フィルタ設定」>。

特徴語のみを対象として共起ネットワーク図を描画する<「関連語検索」→「共起ネット」>。描画する共起関係の強さは、Jaccard 係数 0.06 以上とする。図を見やすくするため、出現数の多い語ほど文字とノードを大きく表示し、サブグラフ毎にノードを色分けする<「関連語・共起ネットワーク」→「カラー」→「サブグラフ検出 (modularity)」／「調整」>。

## C. 物価センチメント指数の構築の詳細

(ナイーブベイズ分類器によるコメント分類)

まず、ベイズの定理から、コメント  $d$  がカテゴリ  $c \in C$  に属する条件付き確率は、

$$P(c|d) = \frac{P(d|c) \times P(c)}{P(d)}$$

となる。右辺の分母は  $c$  に依存しないため、当該コメント  $d$  を、コメントから抽出した単語を出現した数だけ並べた単語ベクトル  $d = (w_{1,d}, w_{2,d}, \dots, w_{n,d})$  の形式で与えると、その属するカテゴリ  $c^*$  は、

$$\begin{aligned} c^* &= \operatorname{argmax}_{c \in C} P(d|c) \times P(c) = \operatorname{argmax}_{c \in C} \left( \prod_{1 \leq i \leq n} P(w_{i,d}|c) \right) \times P(c) \\ &= \operatorname{argmax}_{c \in C} \left( \prod_{w \in W} q_{w,c}^{n_{w,d}} \right) \times p_c \end{aligned}$$

と表せる。ナイーブベイズ分類器の特徴は、各単語の生起確率  $P(w_{i,d}|c)$  に条件付き独立を仮定する点である。上式で、 $n_{w,d}$  は単語ベクトル  $d$  に含まれる単語  $w$  の出現回数を表し、データから簡単に求まる。 $q_{w,c}$  ( $\forall w \in W, c \in C$ ) は、カテゴリ  $c \in C$  所与の下で各単語  $w$  が生起する確率である。 $p_c$  ( $c \in C$ ) は、ある単語ベクトルが特定のカテゴリ  $c \in C$  に属する確率である。これらのパラメータ  $q_{w,c}$  ,  $p_c$  は、教師データを用いて事前に推定しておく。

### (ナイーブベイズ分類器のパラメータの推定)

これらのパラメータの推定方法を説明する前に、生成モデルを詳述しておく。まず、 $\alpha \in \mathbb{R}$  をハイパーパラメータとするディリクレ分布  $\text{Dir}(\alpha)$  からカテゴリを決定する確率  $(p_c)_{c \in C}$  をドロウする。次に、同一のハイパーパラメータのディリクレ分布  $\text{Dir}(\alpha)$  から生成する単語を決定する確率  $(q_{w,c})_{w \in W}$  ( $\forall c \in C$ ) をドロウする。これらを用いて、次のような操作を反復して単語ベクトルを生成する。まずカテゴリ  $c$  を、 $(p_c)_{c \in C}$  をパラメータとする多項分布からドロウする。次に、単語ベクトルの長さ  $N_d$  をドロウする。この確率はカテゴリ  $c$  と独立であると仮定する。この場合、パラメータの推定結果には影響しない。最後に、 $N_d$  回だけ単語  $w_{i,d}$  ( $1 \leq i \leq N_d$ ) を、 $(q_{w,c})_{w \in W}$  をパラメータとする多項分布からドロウする。

こうした生成モデルに基づいて、パラメータの集合  $\theta = (\{q_{w,c}\}, \{p_c\})$  を、教師データをもとに MAP 推定 (Maximum A Posteriori estimation, 最大事後確率推定) する。推定値は、教師データ  $D = \{(c_k, d_k)\}_k$  に対して単語の出現回数を計算するだけで効率的に得られる。

MAP 推定値  $\theta^*$  は、

$$\theta^* = \operatorname{argmax}_{\theta \in \Theta} P(\theta|D) = \operatorname{argmax}_{\theta \in \Theta} P(\theta) \times P(D|\theta)$$

で定義される。まず、右辺第2項  $P(D|\theta)$  を求める。カテゴリ  $c$  所与の下で、各単語が多項分布に従って生成されることから、

$$P(D|\theta) = \prod_{(c,d) \in D} p(c, d|\theta) = \prod_{(c,d) \in D} p(c|\theta) \cdot p(d|c, \theta)$$

$$= \prod_{(c,d) \in D} p_c \times P(N_d) \times \frac{N_d!}{\prod_{w \in W} n_{w,d}!} \prod_{w \in W} q_{w,c}^{n_{w,d}}$$

と表せる。  $P(N_d)$  は、単語ベクトル  $d$  の長さが  $N_d$  となる確率であり、カテゴリ  $c$  とは独立であると仮定した。

次に、  $P(\theta)$  を求める。  $(p_c)_{c \in C}$  と  $(q_{w,c})_{w \in W}$  ( $\forall c \in C$ ) の事前分布がハイパーパラメータ  $\alpha$  のディリクレ分布に従うことから、

$$\begin{aligned} P(\theta) &= P((p_c)_{c \in C}) \times P((q_{w,c})_{w \in W}) \\ &= K_1(\alpha) \prod_{c \in C} p_c^{\alpha-1} \times K_2(\alpha) \prod_{w \in W} q_{w,c}^{\alpha-1}, \forall c \in C \end{aligned}$$

となる。ここで、  $K_1(\alpha)$  と  $K_2(\alpha)$  は規格化定数である。

以上より、  $\theta^*$  は、

$$\theta^* = \operatorname{argmax}_{\theta \in \Theta} \prod_{c \in C} p_c^{\alpha-1} \times \prod_{w \in W} q_{w,c}^{\alpha-1} \times \prod_{(c,d) \in D} p_c P(N_d) \frac{N_d!}{\prod_{w \in W} n_{w,d}!} \prod_{w \in W} q_{w,c}^{n_{w,d}}$$

となる。

これと、パラメータが確率として満たすべき制約  $\sum_{c \in C} p_c = 1$ ,  $\sum_{w \in W} q_{w,c} = 1$  ( $\forall c \in C$ ) を合わせてラグランジュの未定乗数法で解くと、任意のカテゴリ  $c \in C$  と単語  $w \in W$  に対して、

$$\begin{aligned} p_c &= \frac{N_c + (\alpha - 1)}{\sum_{c \in C} N_c + |C|(\alpha - 1)} \\ q_{w,c} &= \frac{n_{w,c} + (\alpha - 1)}{\sum_{w \in W} n_{w,c} + |W|(\alpha - 1)} \end{aligned}$$

を得る。ここで、  $N_c = \sum_{w \in W} n_{w,c}$  は、カテゴリ  $c$  に属する単語ベクトルの長さの総和を表す。  $n_{w,c}$  は、カテゴリ  $c$  に属する単語  $w$  の出現数を表す。すなわち、パラメータの MAP 推定値は、教師データについて、カテゴリ別、単語種別に出現回数を数えるだけで計算できる。

本分析では、すべてのカテゴリと単語に対して  $\alpha = 2$  として推定を行う。  $\alpha > 1$  とすることにより、出現回数  $n_{w,c}$  が 0 回の単語の出現確率も厳密には 0 とならず、過学習 (overfitting) を避けることができる。

$P(\theta|D)$  を最大化する MAP 推定と、 $P(D|\theta)$  を最大化する最尤推定法 (maximum likelihood estimation) との違いは、前者がパラメータの事前分布  $P(\theta)$  を考慮している点のみである。最尤推定法による推定値は、 $\alpha = 1$  を代入した場合の MAP 推定値と一致しているため、出現回数が 0 回の単語の出現確率の推定値は厳密に 0 となってしまう。これは、あるカテゴリにおいて出現回数が 0 回の単語を含む単語ベクトルは、そのカテゴリに属する確率が 0 となることを意味するため、分類上の不都合が生じる。MAP 推定では、出現回数が 0 回の単語でも、出現確率の推定値が厳密に 0 とならず、過学習を抑制できる点で優れている。 $\alpha = 2$  の場合には、すべての単語の出現回数を 1 回ずつ増やして算出した最尤推定値に相当する。

以上のように、ナイーブベイズ分類器は、データの生成モデルに一定の仮定を置くことによって、推定するパラメータ数を少なくし、少数のデータに対しても過学習を避けて汎化性能 (generalization performance) を維持している。もっとも、こうした仮定をデータが厳密に満たさずとも、実用的には十分な性能を得られることが多い [24]。

### (カテゴリ推定の数値例)

この点を直感的に理解するため、「販売価格を、原油価格の上昇により引き上げた」というコメントの分類を例にとって考える。このコメントおよび教師データ内で出現する単語は、「販売価格」「原油価格」「上昇」「引き上げる」の 4 種類とし、カテゴリは単純化のため、「物価上昇」「物価下落」の 2 つを考える。

教師データにおけるカテゴリごとの各単語の出現回数  $n_{w,c}$  は、以下の通りとする。

カテゴリ	$n_{w,c}$				$N_c$
	販売価格	原油価格	上昇	引き上げる	
物価上昇	1	2	2	3	8
物価下落	2	1	1	0	4

このとき、カテゴリ  $c$  所与の下で各単語  $w$  が生起する確率  $q_{w,c}$  および、あるコメントが特定のカテゴリ  $c$  に属する確率  $p_c$  は、以下のように推定できる<sup>16</sup>。

<sup>16</sup> この例では、カテゴリ「物価下落」における「引き上げる」の出現回数  $n_{w,c}$  が 0 回であるが、 $\alpha = 2 (> 1)$  としていることで、出現確率が 0 となることが回避されている。

カテゴリ	$q_{w,c}$				$p_c$
	販売価格	原油価格	上昇	引き上げる	
物価上昇	$\frac{1+1}{8+4 \times 1}$	$\frac{2+1}{8+4 \times 1}$	$\frac{2+1}{8+4 \times 1}$	$\frac{3+1}{8+4 \times 1}$	$\frac{8+1}{12+2 \times 1}$
	$= \frac{2}{12}$	$= \frac{3}{12}$	$= \frac{3}{12}$	$= \frac{4}{12}$	$= \frac{9}{14}$
物価下落	$\frac{2+1}{4+4 \times 1}$	$\frac{1+1}{4+4 \times 1}$	$\frac{1+1}{4+4 \times 1}$	$\frac{0+1}{4+4 \times 1}$	$\frac{4+1}{12+2 \times 1}$
	$= \frac{3}{8}$	$= \frac{2}{8}$	$= \frac{2}{8}$	$= \frac{1}{8}$	$= \frac{5}{14}$

これらの確率をもとに、「販売価格を、原油価格の上昇により引き上げた」というコメントの、各カテゴリにおける  $(\prod_{w \in W} q_{w,c}^{n_{w,d}}) \times p_c$  を計算する. すなわち、「物価上昇」「物価下落」のそれぞれのカテゴリについて、

$$\left( \prod_{w \in W} q_{w, \text{物価上昇}}^{n_{w,d}} \right) \times p_{\text{物価上昇}} = \frac{2}{12} \times \frac{3}{12} \times \frac{3}{12} \times \frac{4}{12} \times \frac{9}{14} = 0.00223$$

$$\left( \prod_{w \in W} q_{w, \text{物価下落}}^{n_{w,d}} \right) \times p_{\text{物価下落}} = \frac{3}{8} \times \frac{2}{8} \times \frac{2}{8} \times \frac{1}{8} \times \frac{5}{14} = 0.00105$$

となる. この場合、カテゴリ「物価上昇」の方が、この値が大きくなるため、「物価上昇」に分類される.

#### (データ作成方法と実装方法)

教師データは、2000年1月～2017年10月の景気判断理由集のコメントからランダムに抽出した1,500コメントを、5-1節で定めた4つのカテゴリに手動で分類して作成した. コメント毎のカテゴリを示すラベルをCSV形式で保存する.

KHCoderで形態素解析を行った後、上記のCSVファイルを読み込み「ツール」→「外部変数と見出し」→「読み込み」→「CSVファイル」>、教師データを用いたナイーブベイズ学習、および学習結果を用いた新規データの分類・集計を組み込みの機能で行う「ツール」→「文書」→「ベイズ学習による分類」→「外部変数から学習」/「学習結果を用いた自動分類」>.

## D. 予想インフレ率とインフレ率を説明する単語抽出の詳細

### (決定木による回帰)

決定木とは、木構造の条件分岐によって出力に対応する目的変数と入力変数を関係づける分類器である。

回帰に用いる決定木は、入力変数  $x$  (今回の分析ではすべての単語の出現頻度を並べたベクトル) の属する空間を、 $R_1, \dots, R_K$  の  $K$  個の互いに素な領域に再帰的に分割し、それぞれの領域に対して1個の推定値  $c_k$  ( $1 \leq k \leq K$ ) を定めたものである。 $x$  の属する領域  $R_k$  に対応する値  $c_k$  を、決定木による推定量  $T(x)$  とする：

$$T(x; \theta) = \sum_{1 \leq k \leq K} c_k \mathbb{I}(x \in R_k).$$

ここで、決定木のパラメータ  $\theta = \{c_k, R_k\}_{1 \leq k \leq K}$  は、観測データ  $\{(x_i, y_i)\}_{1 \leq i \leq N}$  から次のように推定する。

教師データに対して最適な木構造を求める問題は、効率的に解くことが難しいことが知られている<sup>17</sup>。このため、貪欲法で次のような最適化問題を繰り返し解くことで、自動的に変数  $j$  とその閾値  $s$  を選択する。

まず、1つの分岐を持つ決定木(決定株)を生成することを考える。このとき、次のような最適化問題を解いて、推定値の2乗誤差を最小化する変数  $j$  と閾値  $s$  を決定する：

$$\min_{j,s} \left[ \min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right].$$

2つ以上の分岐を持つ決定木を生成する場合には、分割された各領域を全体の領域とみなして、同様に分岐を増やすことを繰り返すことで、木を成長させて全体の分岐構造を推定していく。

木をどこまで成長させるかという停止条件には、分岐数や分岐を増やしたときの追加的な推定精度の改善幅を基準とした複数の設定方法がある。本分析では、簡単に各決定木の分岐の総数を直接指定する。各決定木の分岐数を適度に減らすことで過学習を防ぐ

---

<sup>17</sup> 厳密に言えば、教師データを完全に判別できる決定木の中で、条件分岐の判定の期待回数が最小となるものを求める問題は NP 困難である。

ことができる。

### (ランダムフォレストによる回帰)

決定木の短所は、データの僅かな変化に対して学習される木構造が大きく変化し、推定値のボラティリティが大きくなることである。ランダムフォレストは、この欠点を補うため、復元抽出により生成したデータに対して決定木を複数生成し、その決定木を組み合わせることで、計算負荷を抑制しつつ精度の高い推定を可能にする手法である。回帰の場合、推定結果は、以下のようなアンサンブル平均となる：

$$f(x) = \frac{1}{B} \sum_{1 \leq b \leq B} T(x; \theta_b).$$

ここで、 $f(x)$  はランダムフォレストによる推定量を表す。 $\theta_b$  は各決定木のパラメータである。

ランダムフォレストでは、2種類の乱択を組み合わせて決定木を生成する。1種類目は、復元抽出である。これによって、大数の法則からアンサンブル平均の推定精度が向上する。もっとも、決定木間の相関が高いため、推定精度は十分に改善できない。そこで、2種類目の乱択として、入力変数  $n$  個のうち一部の  $p$  ( $< n$ ) 個をランダムに選択してから決定木を生成し、決定木間の相関を低下させる。

### (変数重要度と部分従属プロット)

ランダムフォレストは、多数の決定木を用いることから、目的変数と入力変数の関係を解釈することが難しい。このため、各入力変数のうち目的変数の予測に重要な影響を及ぼすものを見出すために、各入力変数に対して変数重要度 (variable importance) を計る。さらに、見出した重要変数と目的変数との関係を捉えるため、これを部分従属プロット (partial dependence plot) によって可視化する。

変数重要度は、予測精度への寄与に基づいて計算する。まず、各決定木を生成する際に、サンプリングされなかったデータを OOB (Out-Of-Bag) サンプルと呼ぶ。この OOB サンプルのうち、注目する入力変数のみをランダムに並べ替え、新たなサンプルを生成する。この新たなサンプルを用いた場合の平均的な予測精度が、前者を用いた予測精度



と比較してどの程度悪化しているか (mean decrease in accuracy) を、注目する入力変数の変数重要度とする。

部分従属プロットは、注目する入力変数と目的変数との関係を低次元空間で捉えるものである。入力変数  $\{x_i\}_{1 \leq i \leq n}$  の添え字集合  $U = \{i : 1 \leq i \leq n\}$  を  $U = S \cup C$  となる互いに素な部分集合  $S$  と  $C$  に分割する。このとき、部分従属プロットは、

$$f_S(X_S) = E_{X_C}[f(X_S, X_C)]$$

で定義される [19]。これは、注目する入力変数  $X_S$  と目的変数との関係を、その他の入力変数  $X_C$  を周辺化 (average out) し、交差項を捨象することで求めることに相当する。もっとも、 $X_S$  と  $X_C$  の相関が低ければ、交差項を捨象することによる問題は大きくない。

これは、次のように推定することができる：

$$\hat{f}_S(X_S) = \frac{1}{N} \sum_{i=1}^N f(X_S, x_{i,C}).$$

ここで、 $\{x_{i,C}\}_{1 \leq i \leq N}$  は、観測データにおける  $X_C$  の値を表す。

#### (データ作成方法と実装方法)

KH Coder で過去の全データについて形態素解析を行い、調査回毎の BoW ベクトルを出力する < 「ツール」 → 「文書」 → 「『文書×抽出語』表の出力」 → 「CSV ファイル」 >。

ランダムフォレストの学習、各単語の変数重要度算出、およびワードクラウドの描画は、以下の R スクリプトにより行った。ここで、ハイパーパラメータである各決定木の分割を行う入力変数の候補の数は、10 分割交差検証によりグリッドサーチを行い、RMSE を最小とする値を用いた。

```

library(randomForest)
library(wordcloud)

#トレーニングデータの作成
# zpi: 目的変数 (予想インフレ率またはインフレ率) の時系列ベクトル
# bow: 各単語の出現頻度の時系列を束ねた行列
trains <- data.frame(zpi, bow)

#ランダムフォレストの学習
# p: グリッドサーチにより最適化した入力変数の選択数
rf_model <- randomForest(zpi~., trains, mtry=p, importance=TRUE)

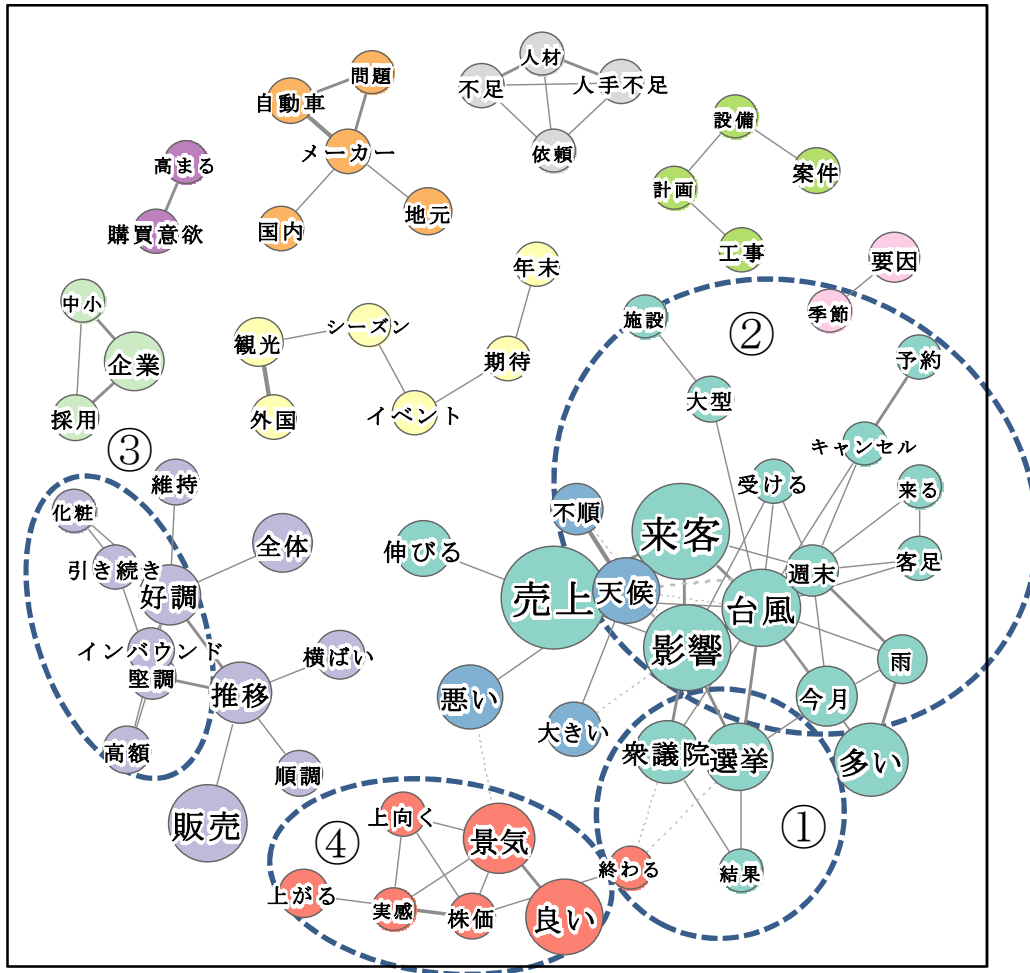
#変数重要度の取得
importance <- importance(rf_model, type=1)

#各単語の影響度算出
imp_words <- rownames(importance)[order(-importance)][1:150]
for(word in imp_words){
  pp <- partialPlot(rf_model, trains, as.vector(word))
  infl <- cbind(infl,
                approx(pp$x, pp$y, xout=bow[,word], rule=2)$y
                -approx(pp$x, pp$y, xout=mean(bow[,word]))$y)
}

#ワードクラウドの描画
# month: 分析対象月
freq_words <- bow[month, imp_words]>0
wordcloud(imp_words[freq_words], abs(infl[month,])[freq_words])

```

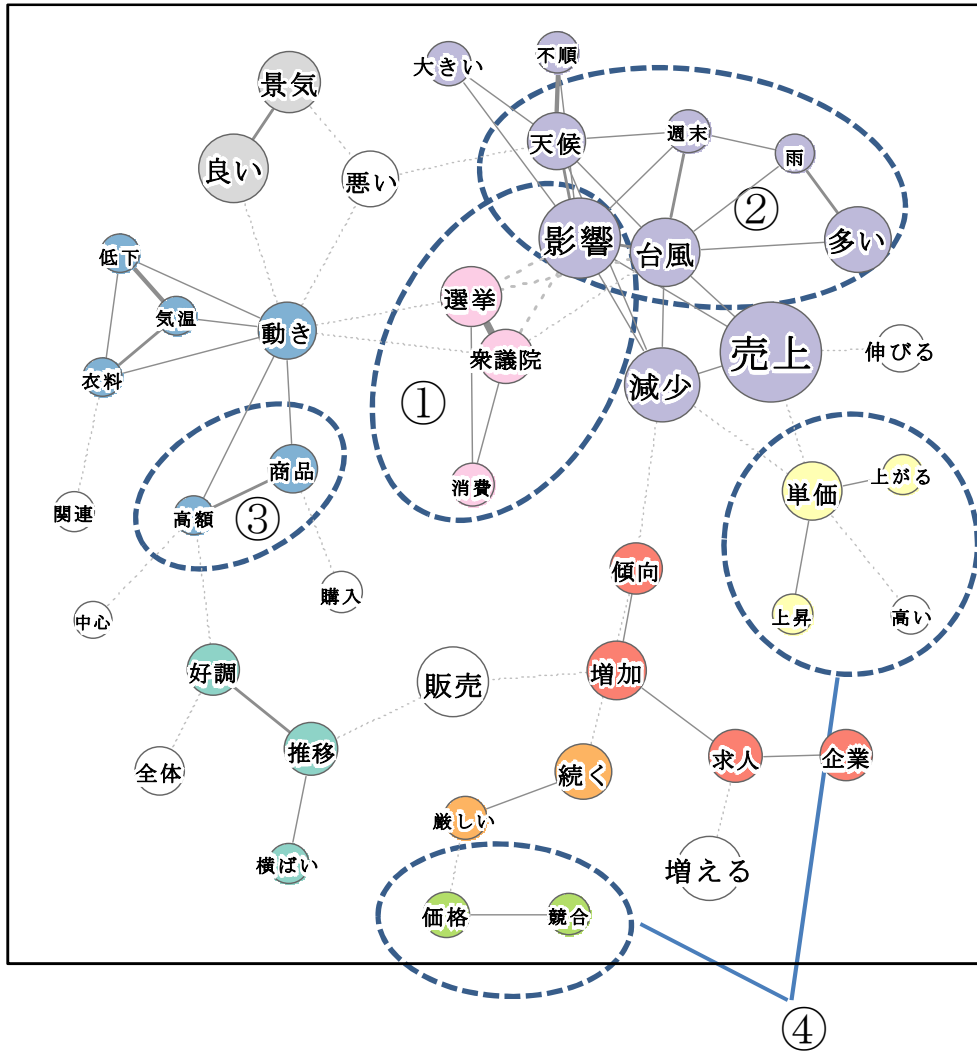
# 特徴語の共起ネットワーク図 (2017/10月・現状判断)



<各部分ネットワークとトピックの対応関係>

- ①衆議院選挙の影響
- ②台風，天候不順の影響
- ③高額品，インバウンド需要
- ④景気と株高

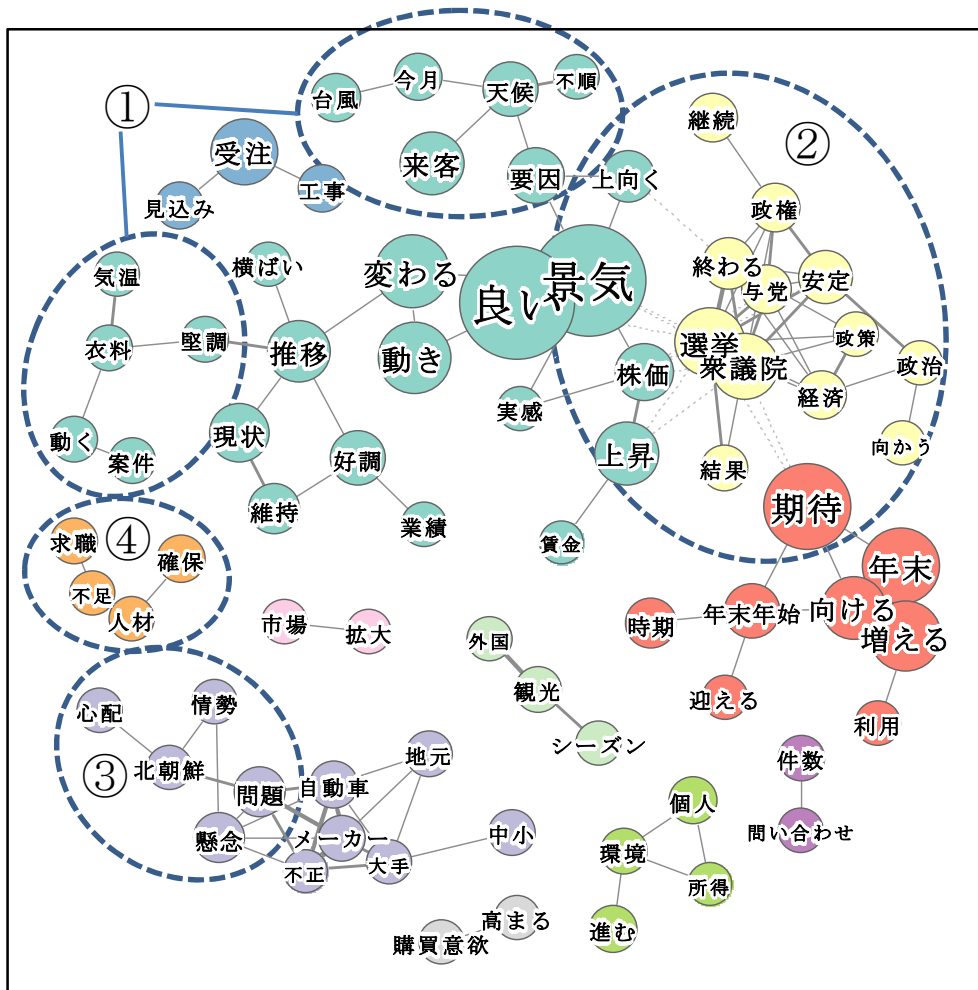
# 岡崎・敦賀の共起ネットワーク図 (2017/10月・現状判断)



<各部分ネットワークとトピックの対応関係>

- ①衆議院選挙の影響
- ②台風，天候不順の影響
- ③高額商品の動き
- ④価格の動き

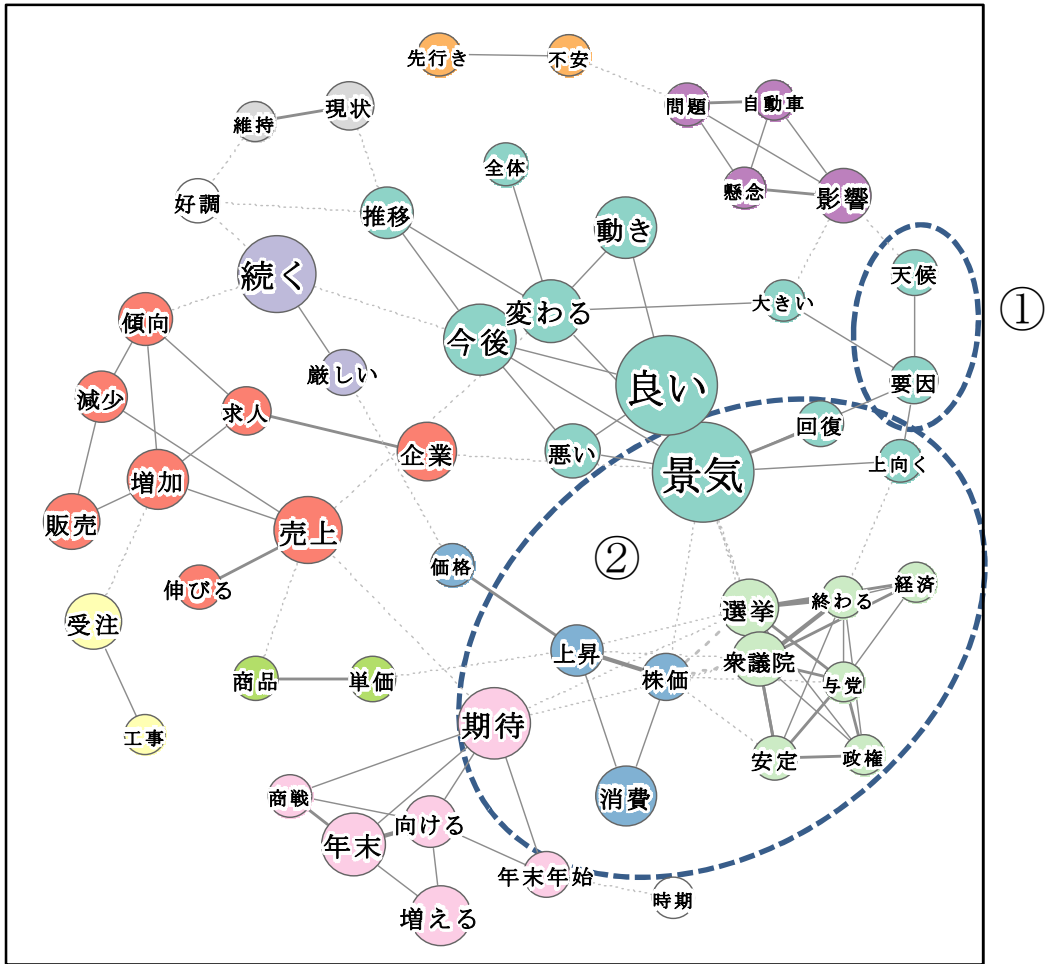
# 特徴語の共起ネットワーク図 (2017/10月・先行き判断)



## <各部分ネットワークとトピックの対応関係>

- ①台風，天候不順の影響
- ②衆議院選挙，株高の影響
- ③北朝鮮情勢
- ④人手不足の影響

岡崎・敦賀の共起ネットワーク図 (2017/10月・先行き判断)

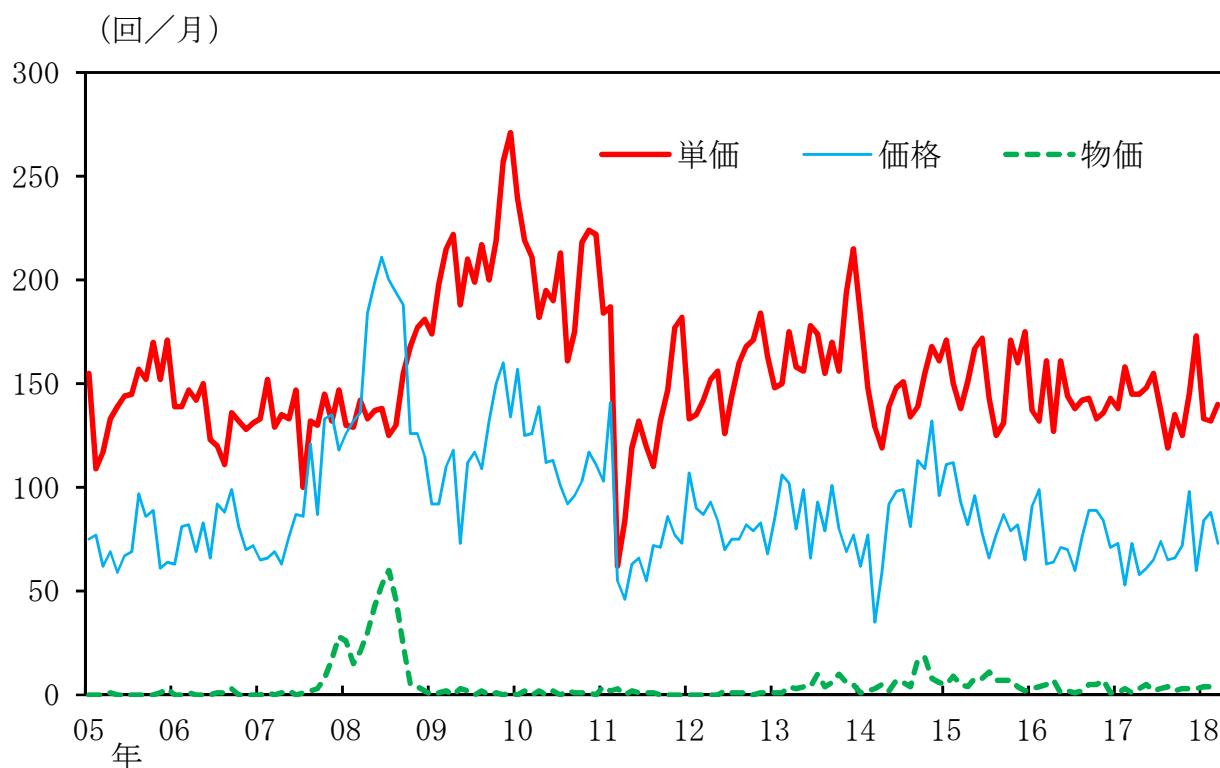


<各部分ネットワークとトピックの対応関係>

- ①天候不順の影響
- ②衆議院選挙，株高の影響

## 物価に関する単語の出現頻度と共起回数（現状判断）

## (1) 物価を表す単語の出現回数（現状）

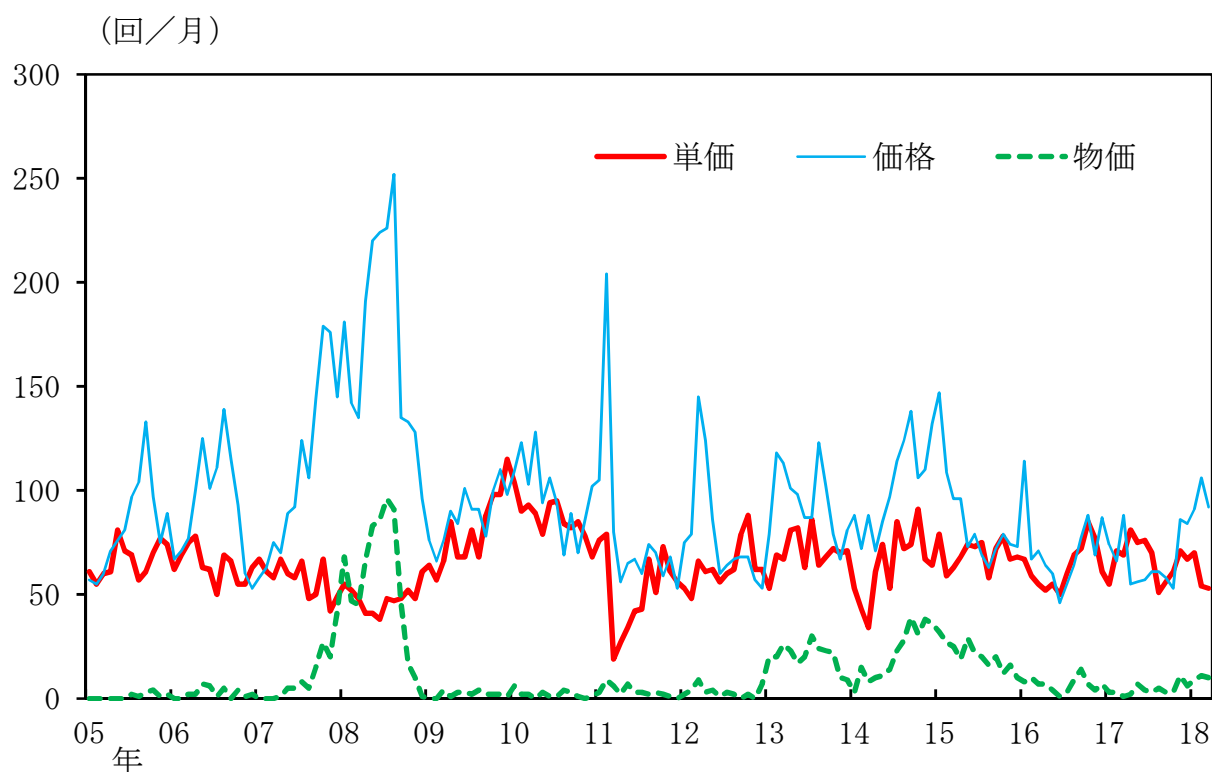


## (2) 物価に関する単語同士の共起回数（2018/3月，現状）

組み合わせ	共起回数	Jaccard係数
単価-来客	40	0.126
単価-売上	29	0.080
単価-上昇	27	0.147
単価-減少	21	0.086
単価-購入	19	0.104
単価-増加	18	0.067
価格-高騰	14	0.182
単価-傾向	14	0.067
単価-高い	13	0.075
価格-野菜	12	0.160
単価-上がる	12	0.075
価格-上昇	11	0.081
単価-点数	11	0.080
単価-買上	10	0.073
単価-低下	10	0.066
単価-上回る	10	0.063
単価-客数	9	0.059
単価-伸びる	9	0.045
価格-売上	9	0.028
単価-変化	8	0.043

## 物価に関連する単語の出現頻度と共起回数（先行き判断）

## (1) 物価を表す単語の出現回数（先行き）



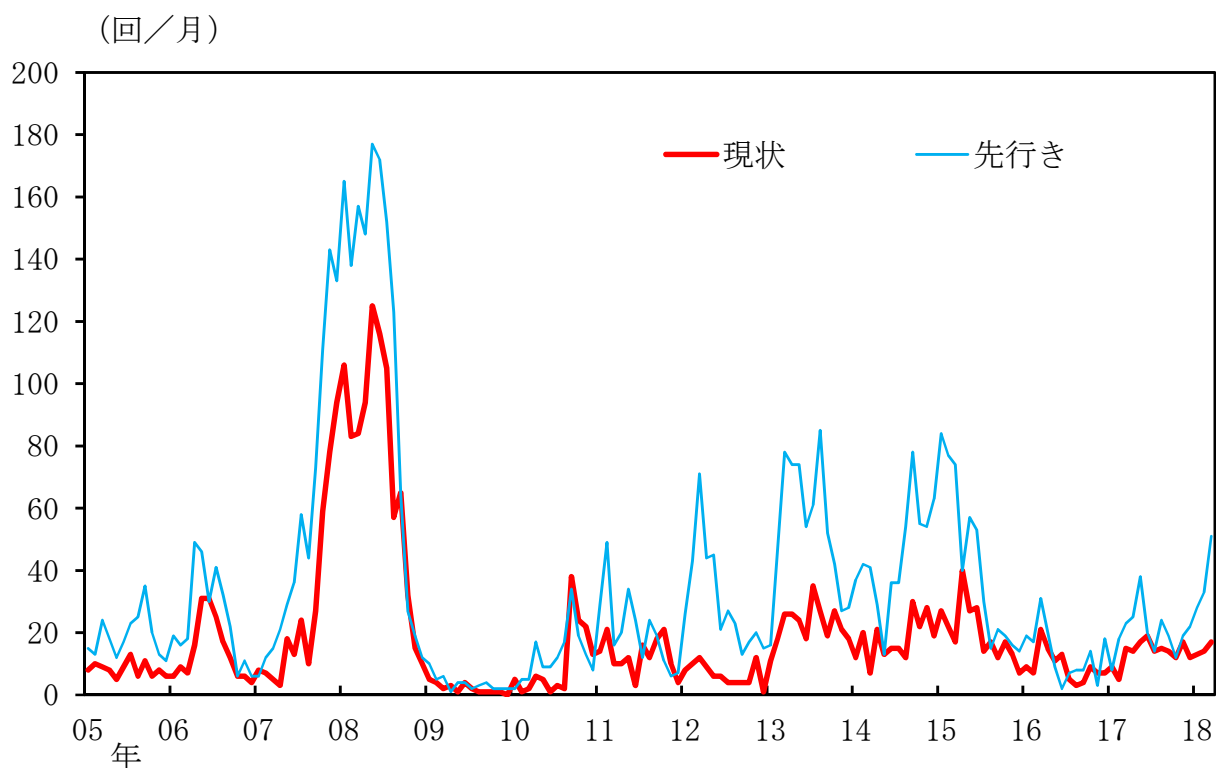
## (2) 物価に関連する単語同士の共起回数（2018/3月，先行き）

組み合わせ	共起回数	Jaccard係数
価格-上昇	19	0.128
価格-高騰	16	0.174
単価-来客	12	0.081
価格-原材料	11	0.122
価格-競合	11	0.107
単価-売上	11	0.060
価格-商品	10	0.069
価格-販売	10	0.055
単価-上がる	9	0.107
価格-転嫁	9	0.102
単価-良い	9	0.031
単価-上昇	8	0.064
価格-値上げ	8	0.064
価格-受注	8	0.045
価格-原料	7	0.081
価格-燃料	7	0.077
価格-人件	7	0.074
価格-野菜	7	0.072
価格-製品	7	0.069
価格-厳しい	7	0.044

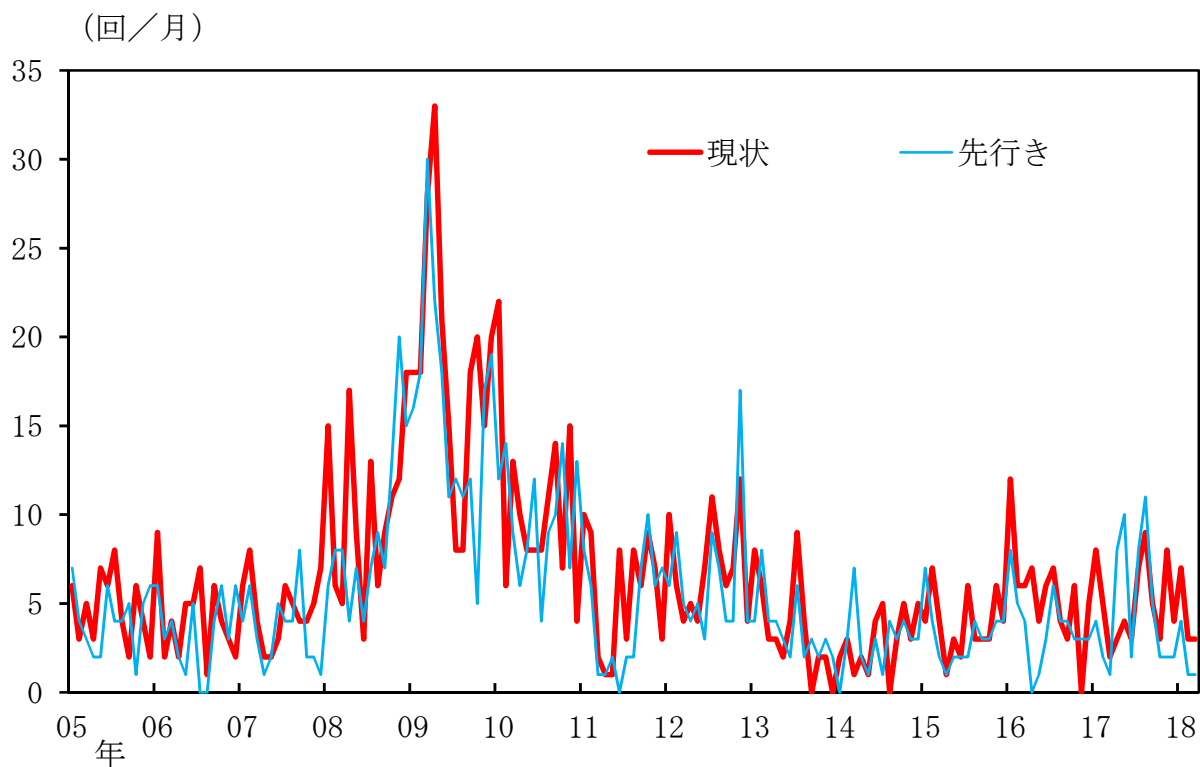


## 「値上げ」と「値下げ」の出現回数

### (1) 「値上げ」の出現回数



### (2) 「値下げ」の出現回数



## 教師データのサンプル（現状判断）

a. 物価上昇	b. 物価下落
<p>「商品を値上げしたことで客単価が上がっている。来客数も増えている。」</p> <p>「原油コストの上昇に伴い販売価格の値上げが実施されるため……。」</p> <p>「販売価格を上げたにもかかわらず、利用客数は今の段階で予想を上回っている。」</p>	<p>「価格低廉化の傾向は変わらず、薄利多売状況が続いている。」</p> <p>「価格の引下げがあり、利益が出ないのが現状である。」</p> <p>「衣料品関連の量販店で値下げの動きがあり……。」</p>
c. 物価もちあい等	d. その他
<p>「販売単価は前年を上回っていない。」</p> <p>「大口価格が下がる兆しがなく……。」</p> <p>「取引先等の受注価格，販売価格等の動きもあまりない……。」</p>	<p>「業界全体の景気が悪い。」</p> <p>「来客数，受注量，受注額の好調が続いている。」</p> <p>「業種によっては人手不足の状況が深刻化している。」</p>

## 教師データのサンプル (先行き判断)

a. 物価上昇	b. 物価下落
<p>「ガソリン価格の高騰による影響は、今後もしばらく続く。」</p> <p>「4月以降は様々な商品の値上げが予想され……。」</p> <p>「既に販売価格の値上げ交渉を行っており、徐々に進めば利益は確保できる……。」</p>	<p>「客から値下げの要望があり、販売単価は現行より低下する見込みである。」</p> <p>「荷主からは、値下げ圧力が続いている。」</p> <p>「日用品の値下げに踏み切るが、仕入価格は変わらないため、販売数が伸びても利益増にはつながらない。」</p>
c. 物価もちあい等	d. その他
<p>「原料価格は落ち着いており、当面は現状のまま推移する。」</p> <p>「仕事量は徐々に増加しているが、単価は一向に上昇してこない。」</p> <p>「1品単価、客単価は上がってきていない。」</p>	<p>「インバウンド消費は、今後も一層拡大すると見込んでいる。」</p> <p>「受注量等はあまり変わらないと予想している。」</p> <p>「個人消費が上向く材料が見当たらない。」</p>

## ナイーブベイズ分類器の学習結果（現状判断）

<各単語のスコア（カテゴリ別）>

	a	b	c	d
単価	5.40	6.00	5.34	2.08
高騰	4.31	1.19	3.08	0.69
求人	1.95	0.50	1.83	4.38
転嫁	3.24	0.50	2.75	0.00
物価	3.15	0.50	2.52	0.00
値上がり	3.68	0.50	2.52	0.69
下落	1.95	4.16	2.75	0.69
原材料	3.15	1.19	2.24	0.00
採用	0.85	1.19	1.14	3.71
円	2.23	3.06	1.14	0.00
激化	1.54	3.20	1.83	0.00
燃料	2.93	1.59	2.24	0.00
派遣	0.85	2.44	1.14	3.37
値上げ	3.94	1.59	2.24	1.39
上昇	4.74	2.29	3.62	2.40

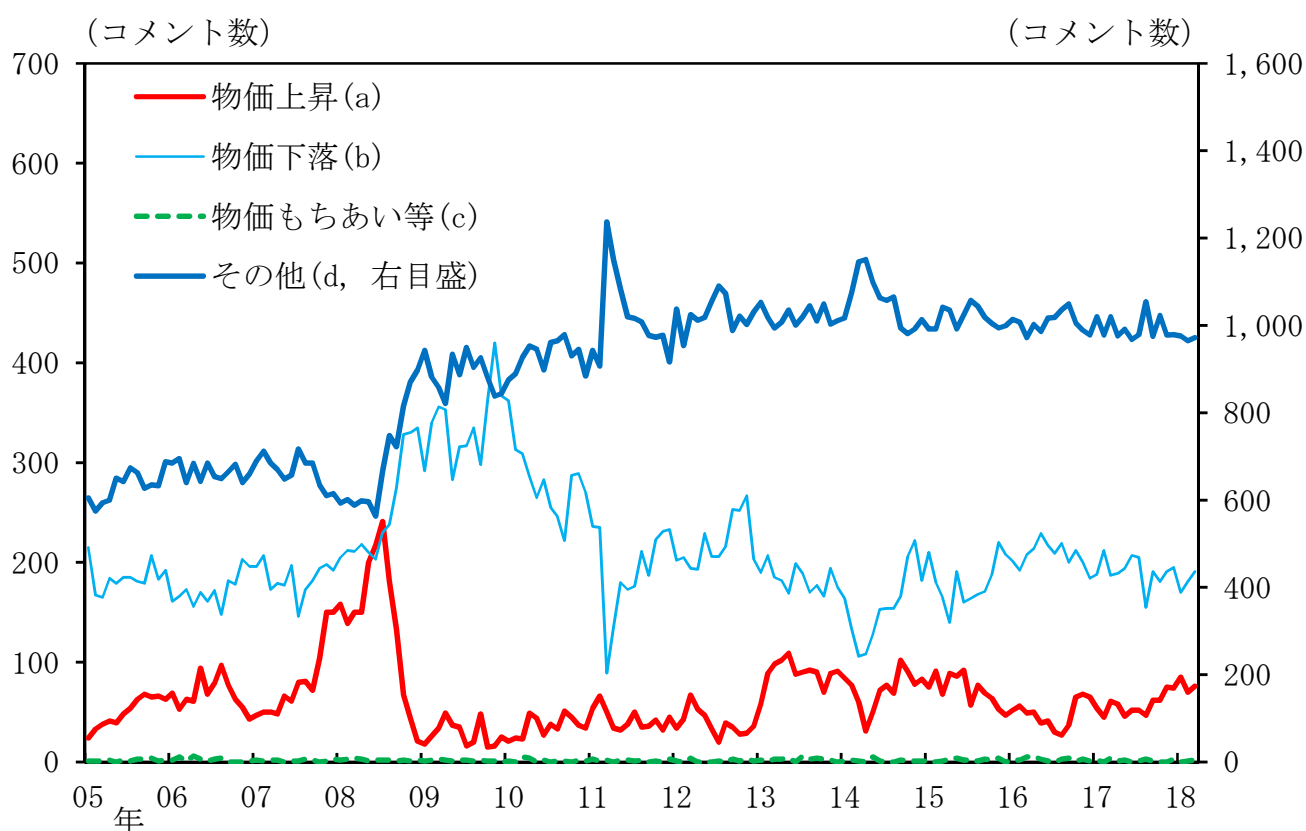
<5分割交差検証結果>

ベイズ学習による分類					
	a	b	c	d	
正解	a	76	61	1	53
	b	13	290	1	66
	c	12	56	0	30
	d	15	78	1	747

正解を得た数： 1,113 / 1,500 (74.2%)

Kappa 統計量： 0.552

<各カテゴリのコメント数>



## ナイーブベイズ分類器の学習結果（先行き判断）

<各単語のスコア（カテゴリ別）>

	a	b	c	d
値上げ	5.12	0.80	3.19	0.69
高騰	5.01	1.90	2.34	0.00
ガソリン	4.41	1.90	2.34	0.00
原材料	4.00	2.19	1.94	0.00
値上がり	4.17	0.80	2.34	0.69
単価	4.66	5.60	5.05	1.95
原油	4.52	3.11	2.63	0.69
デフレ	2.46	3.51	1.24	0.00
求人	2.75	1.50	1.24	4.32
転嫁	3.56	0.80	2.34	0.69
運賃	3.15	2.19	1.24	0.00
値下げ	1.77	3.64	2.85	0.69
改定	2.75	1.50	2.63	0.00
野菜	2.62	1.50	2.63	0.00
燃料	3.56	1.50	1.94	0.69

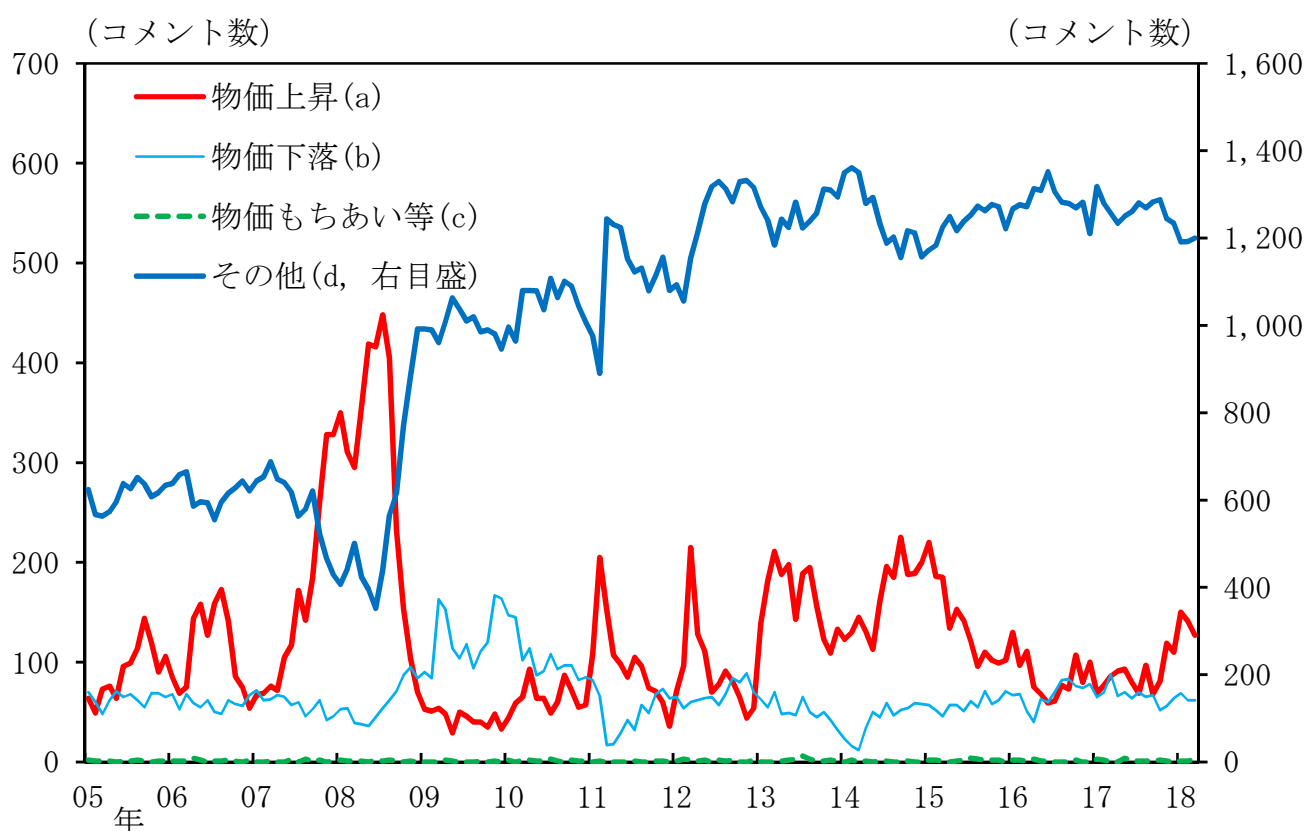
<5分割交差検証結果>

ベイズ学習による分類					
	a	b	c	d	
正解	a	195	21	0	66
	b	32	113	0	85
	c	20	19	0	37
	d	34	42	3	833

正解を得た数： 1,141 / 1,500 (76.1%)

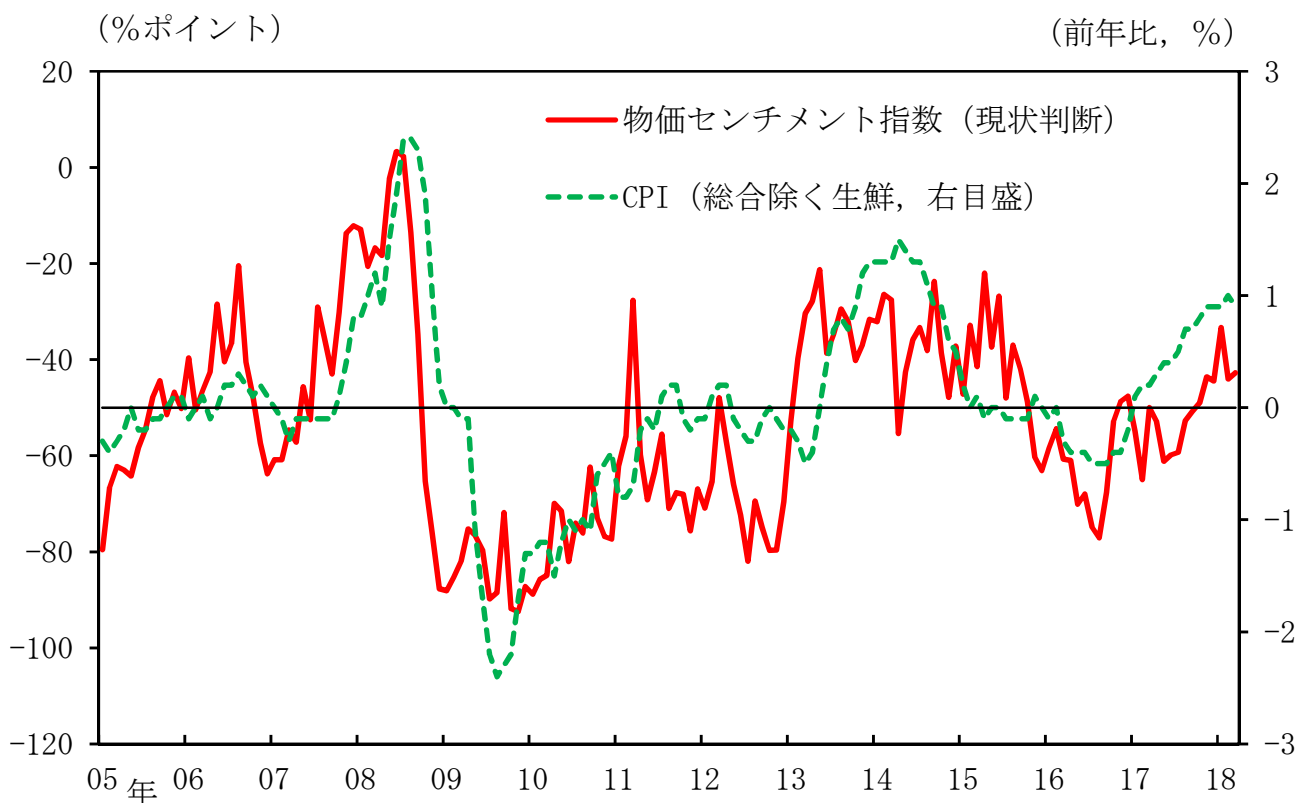
Kappa 統計量： 0.549

<各カテゴリのコメント数>

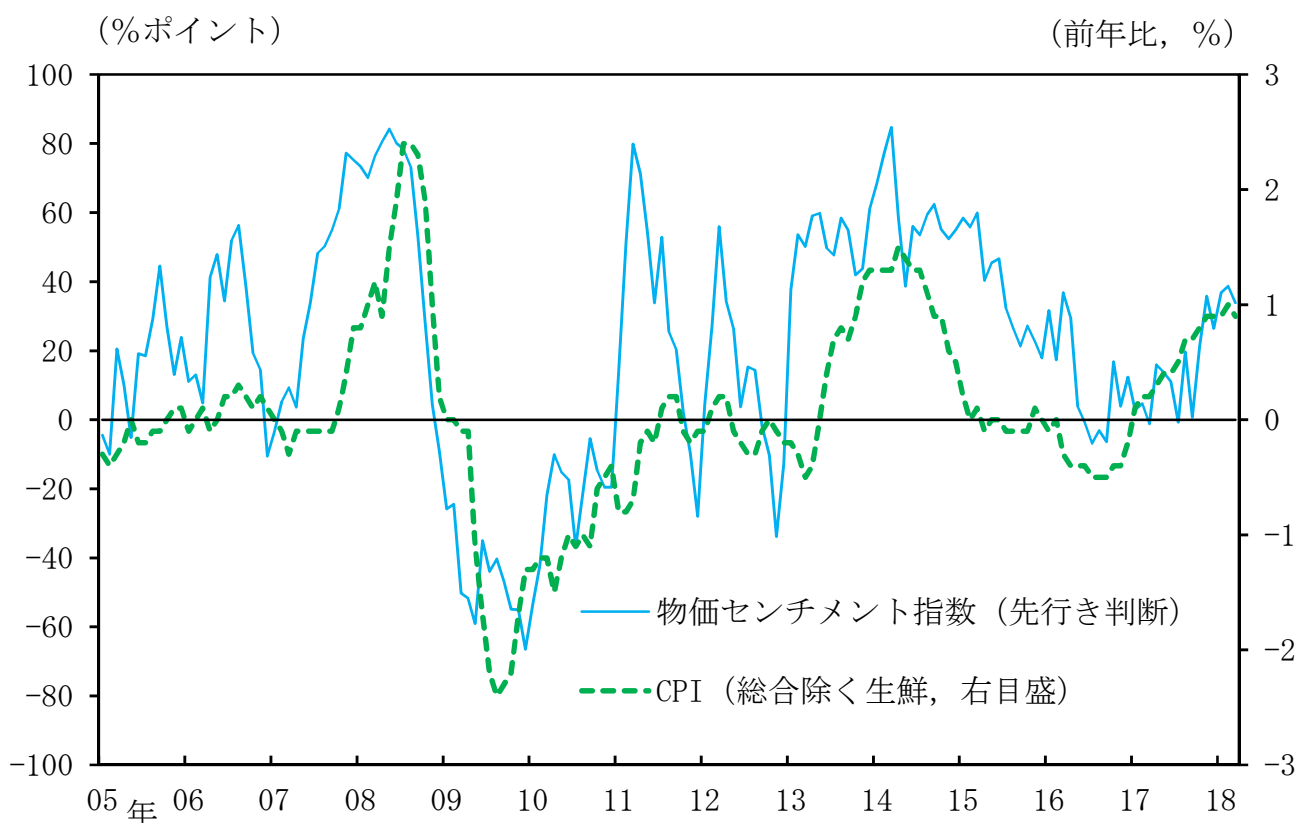


# 物価センチメント指数

## <現状判断>



## <先行き判断>

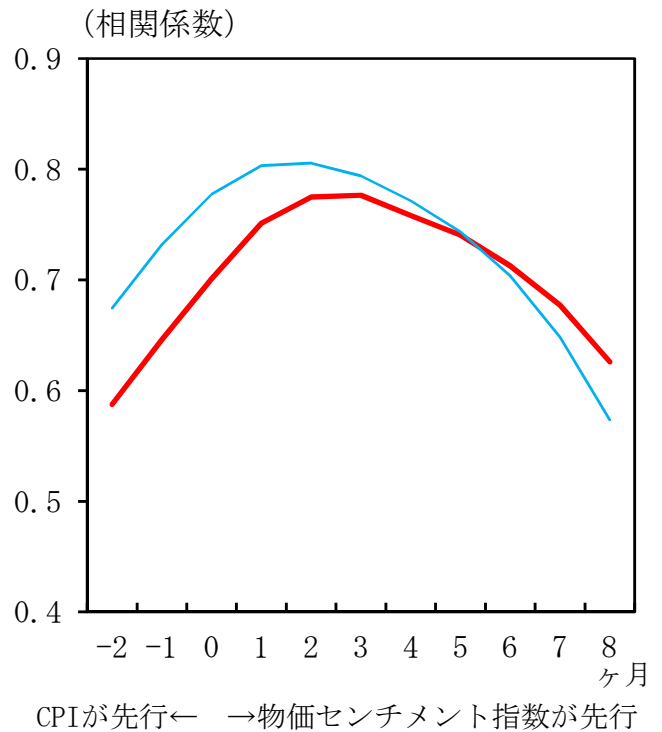
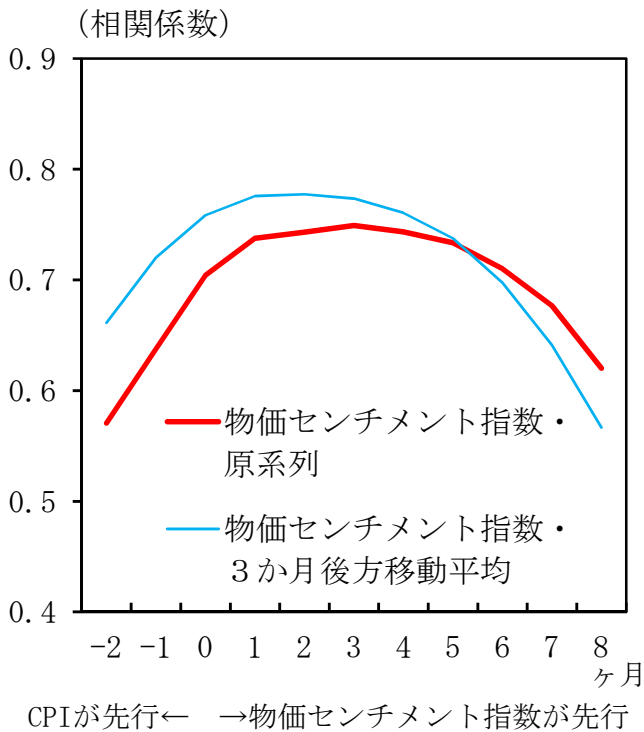


(注) CPIは、消費税調整済.

# 物価センチメント指数と消費者物価の時差相関（現状判断）

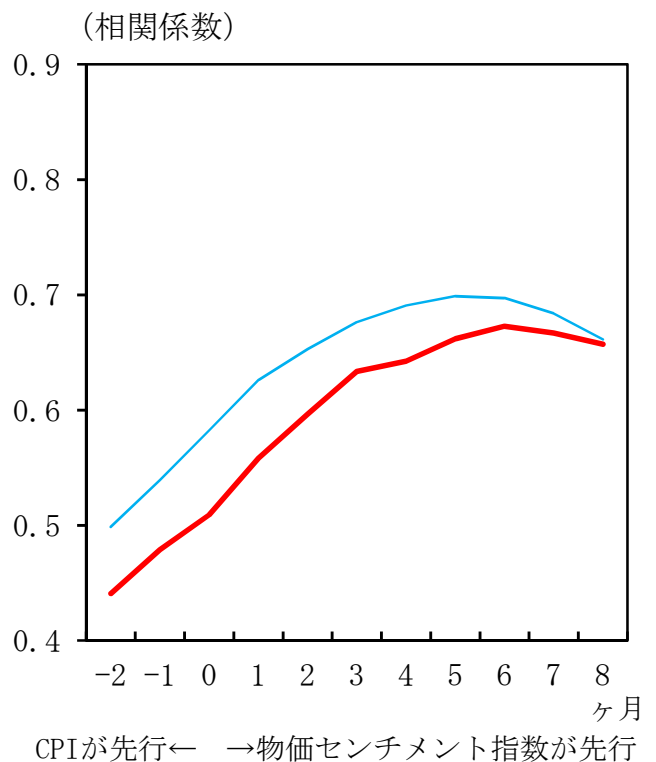
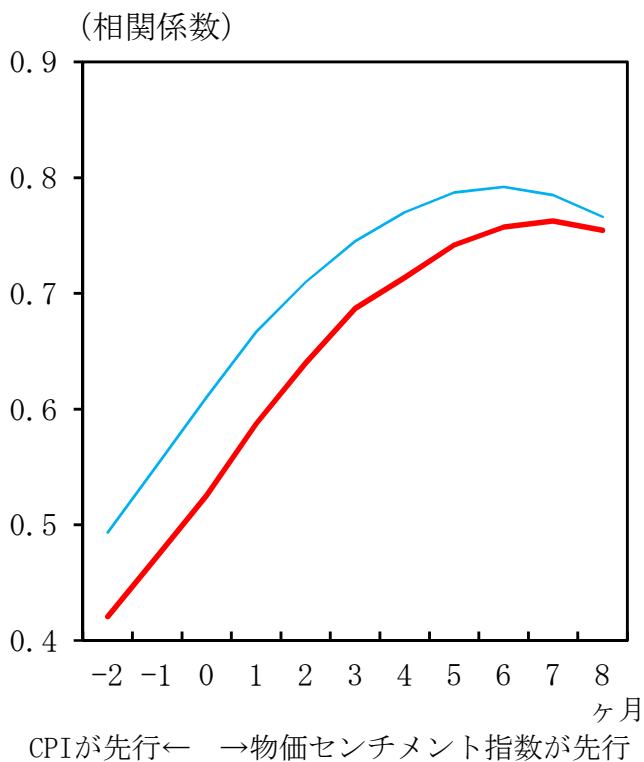
< C P I 総合 >

< 総合 除く生鮮食品 >



< 総合 除く生鮮・エネルギー >

< 総合 除く食料・エネルギー >

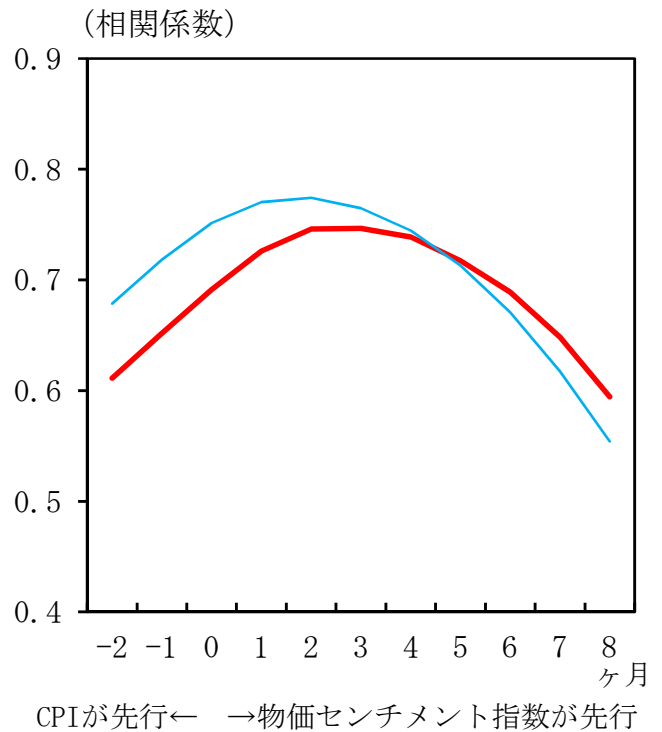
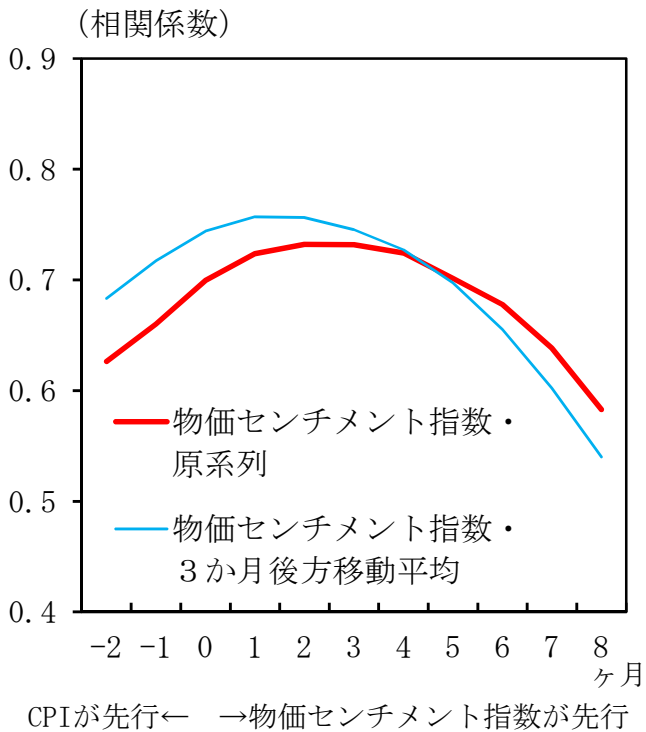


(注) C P I は、消費税調整済、前年比。

# 物価センチメント指数と消費者物価の時差相関（先行き判断）

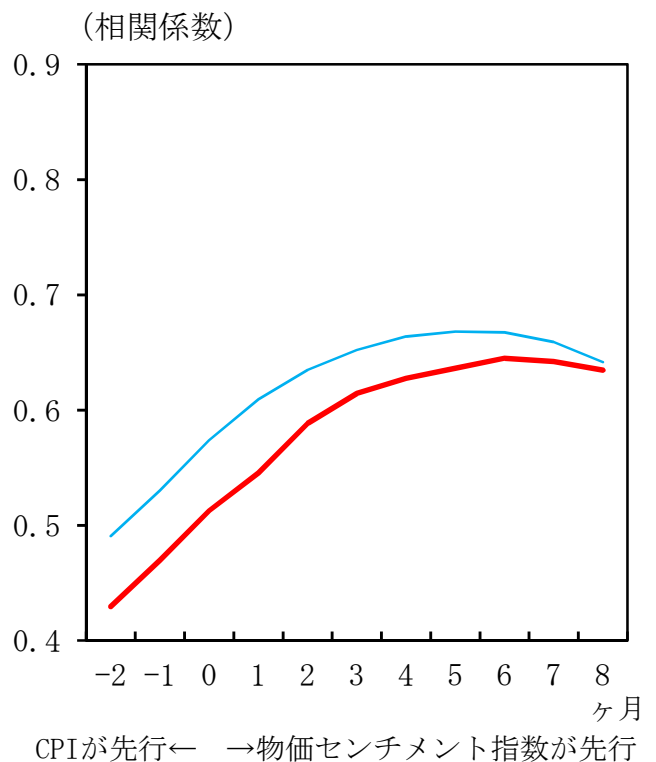
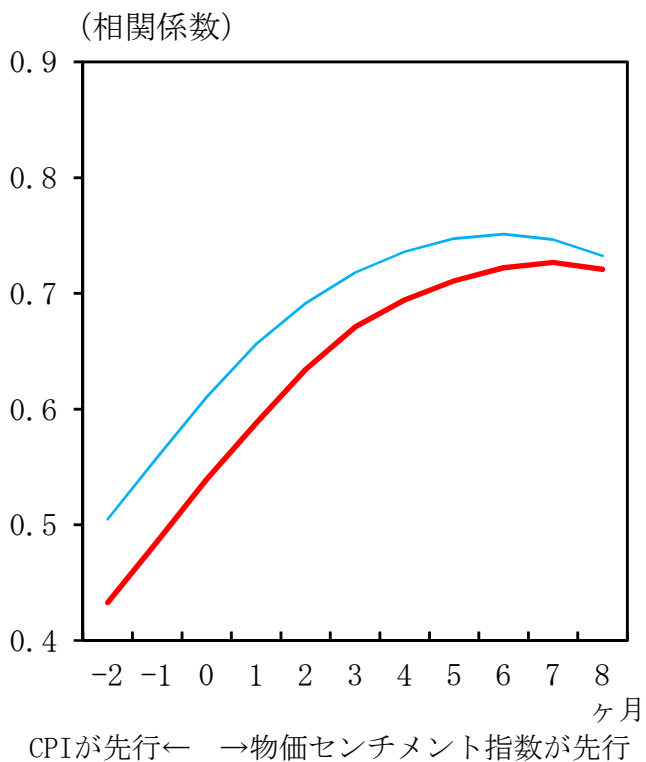
< C P I 総合 >

< 総合 除く生鮮食品 >



< 総合 除く生鮮・エネルギー >

< 総合 除く食料・エネルギー >



(注) C P I は、消費税調整済、前年比。



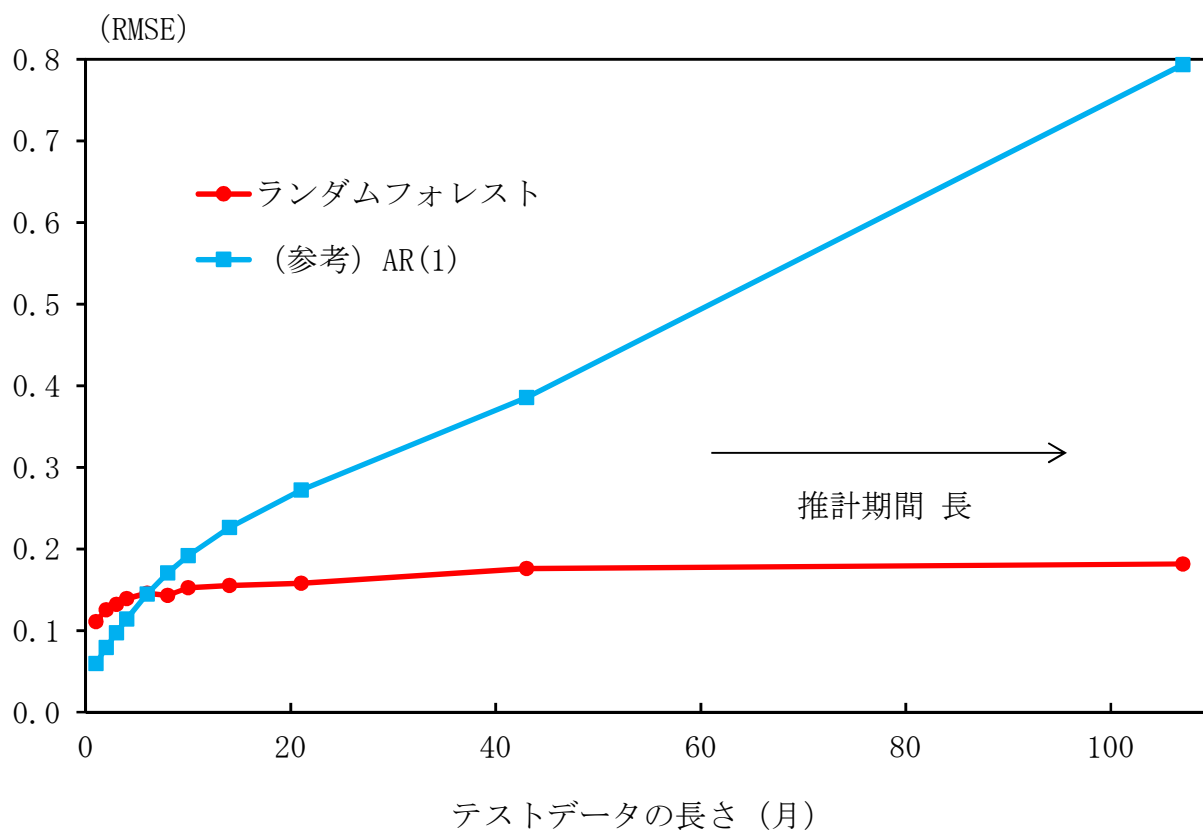
## グレンジャー因果性検定結果

帰無仮説：AはBに対してグレンジャーの意味での因果関係はない		F 値	p 値
A	B		
物価センチメント指数（現状判断）	→	C P I（総合）	5.575 *** 0.0000
C P I（総合）	→	物価センチメント指数（現状判断）	1.680 * 0.0744
物価センチメント指数（先行き判断）	→	C P I（総合）	3.514 *** 0.0001
C P I（総合）	→	物価センチメント指数（先行き判断）	0.793 0.6578
物価センチメント指数（現状判断）	→	C P I（総合除く生鮮）	6.079 *** 0.0000
C P I（総合除く生鮮）	→	物価センチメント指数（現状判断）	0.696 0.7545
物価センチメント指数（先行き判断）	→	C P I（総合除く生鮮）	3.390 *** 0.0002
C P I（総合除く生鮮）	→	物価センチメント指数（先行き判断）	0.900 0.5477

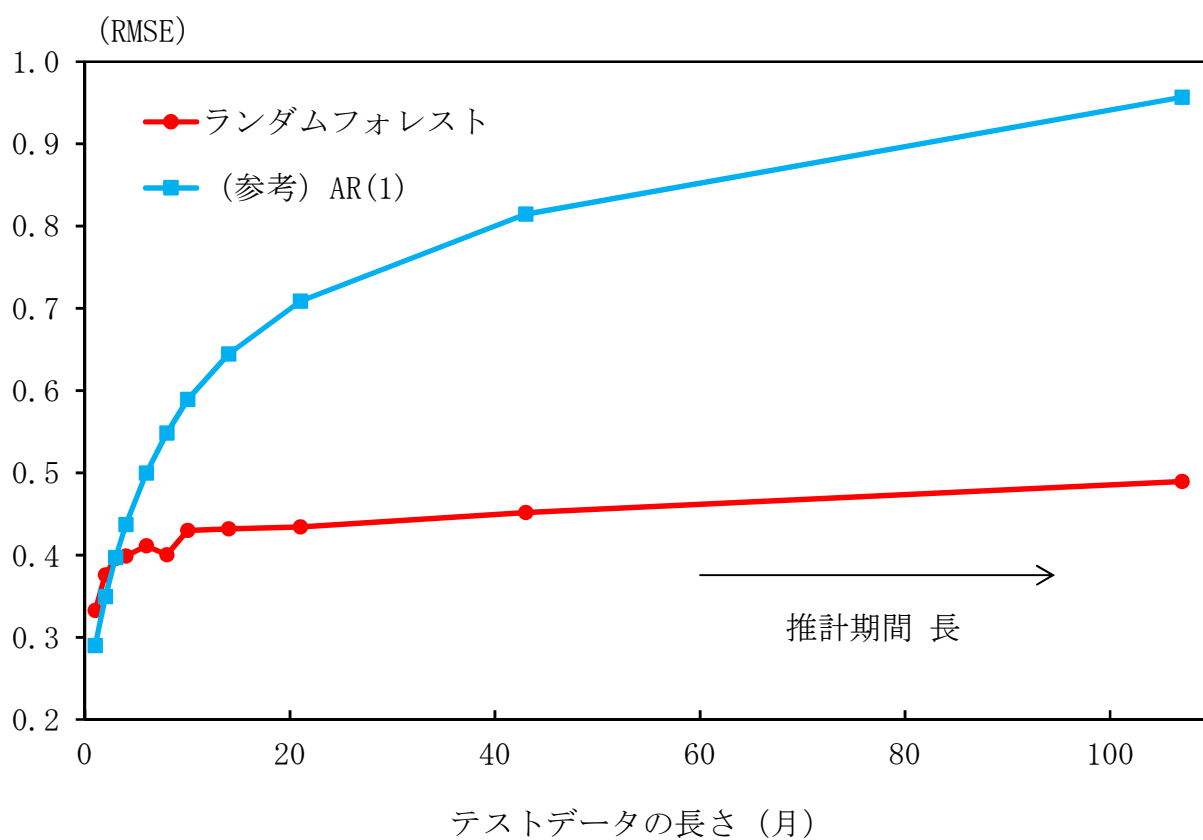
(注) 1. \*\*\*は1%水準, \*は10%水準でグレンジャーの意味での因果関係はないとする帰無仮説は棄却されることを示す.

2. ラグは12期を選択した.

### 交差検証 (目的変数：予想インフレ率)



### 交差検証（目的変数：C P I 総合除く生鮮）



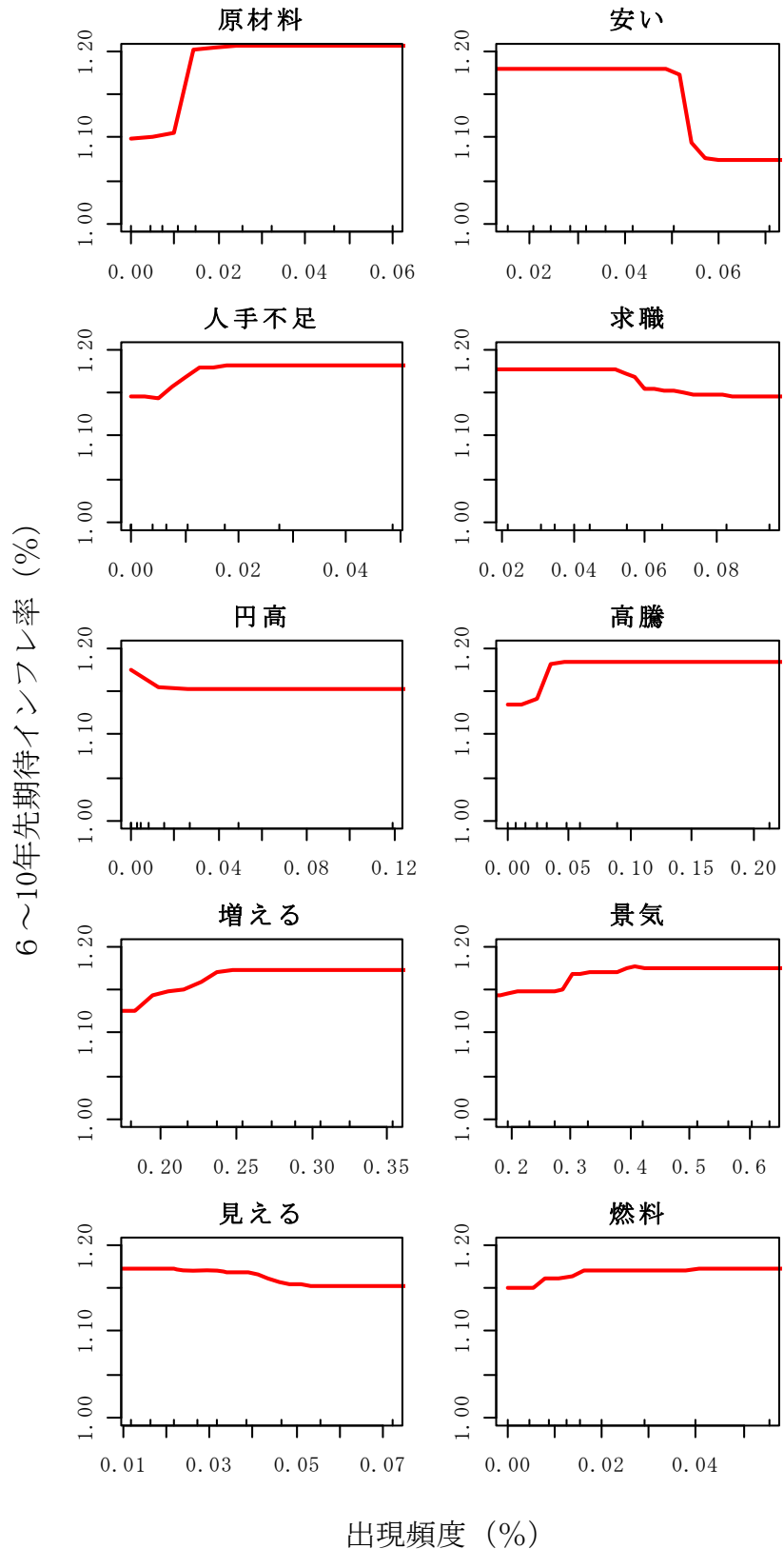
# 変数重要度と部分従属プロット (目的変数：予想インフレ率)

<変数重要度の高い単語>

<部分従属プロット>

(最大値=100)

原材料	100.0
安い	92.9
人手不足	74.1
求職	71.9
円高	69.2
高騰	67.3
増える	60.7
景気	58.7
見える	55.6
燃料	54.7
先行き	53.5
見る	47.4
減少	46.0
反動	41.9
人材	40.2
値上げ	39.2
消費	39.2
デフレ	37.8
明るい	37.2
国内	37.1



# 変数重要度と部分従属プロット (目的変数：C P I 総合除く生鮮)

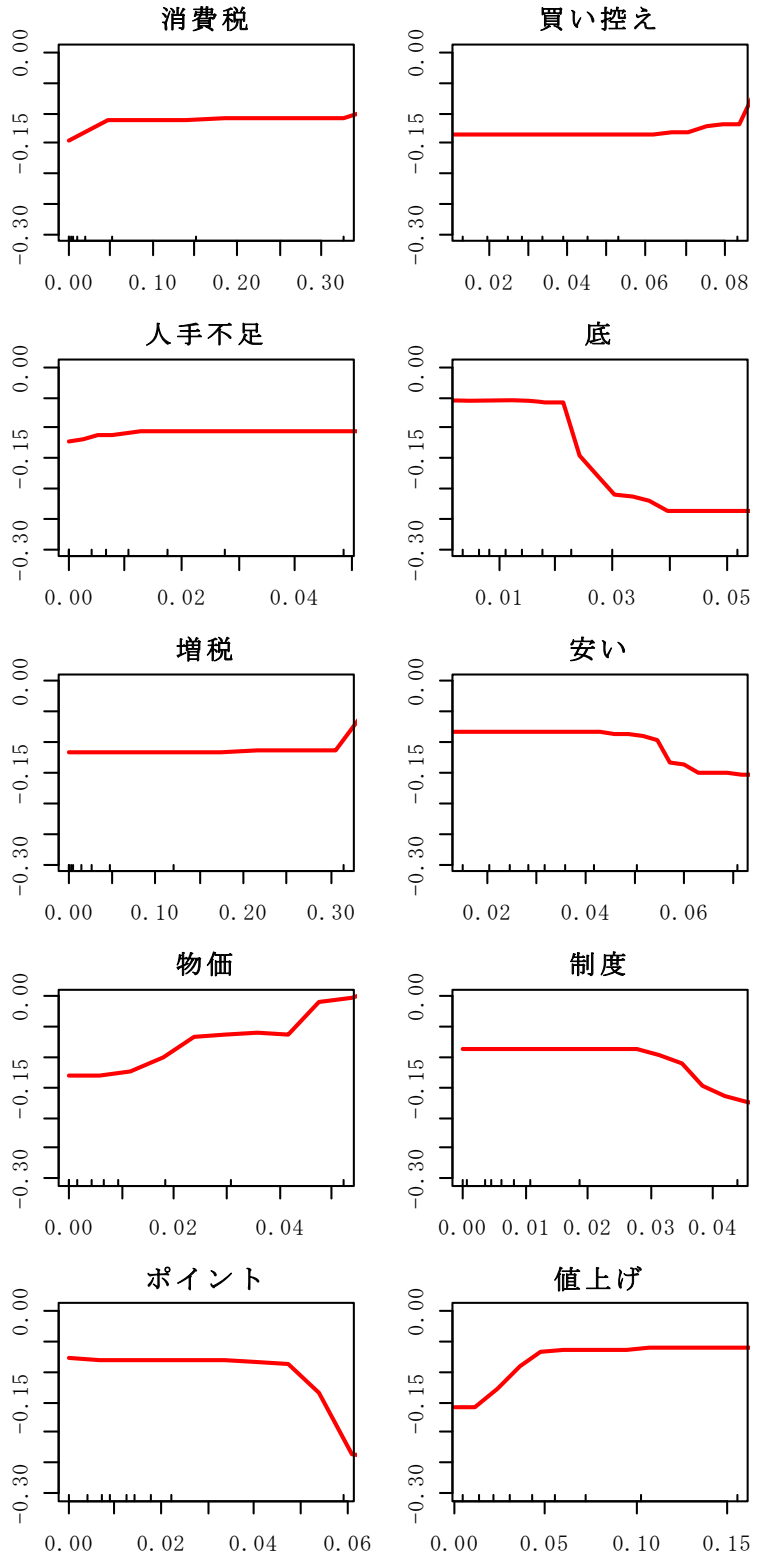
<変数重要度の高い単語>

<部分従属プロット>

(最大値=100)

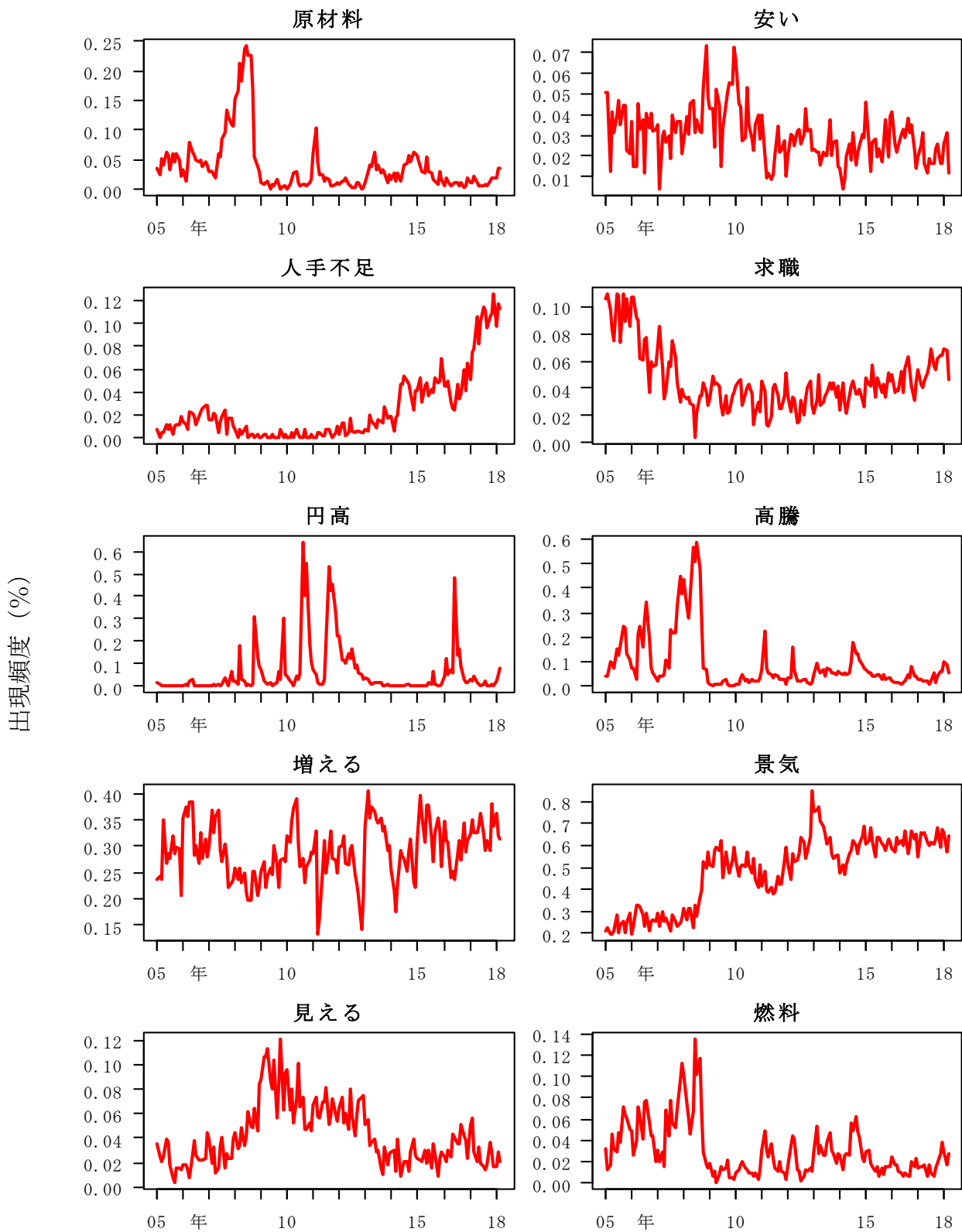
消費税	100.0
買い控え	98.6
人手不足	84.7
底	79.9
増税	72.9
安い	71.7
物価	70.0
制度	64.0
ポイント	61.8
値上げ	60.4
購入	54.8
補助	43.0
雇用	42.8
上がる	42.1
実感	39.7
景気	38.1
来客	37.8
デフレ	35.8
低下	35.2
エコ	35.1

C P I 総合除く生鮮 (前年比, %)

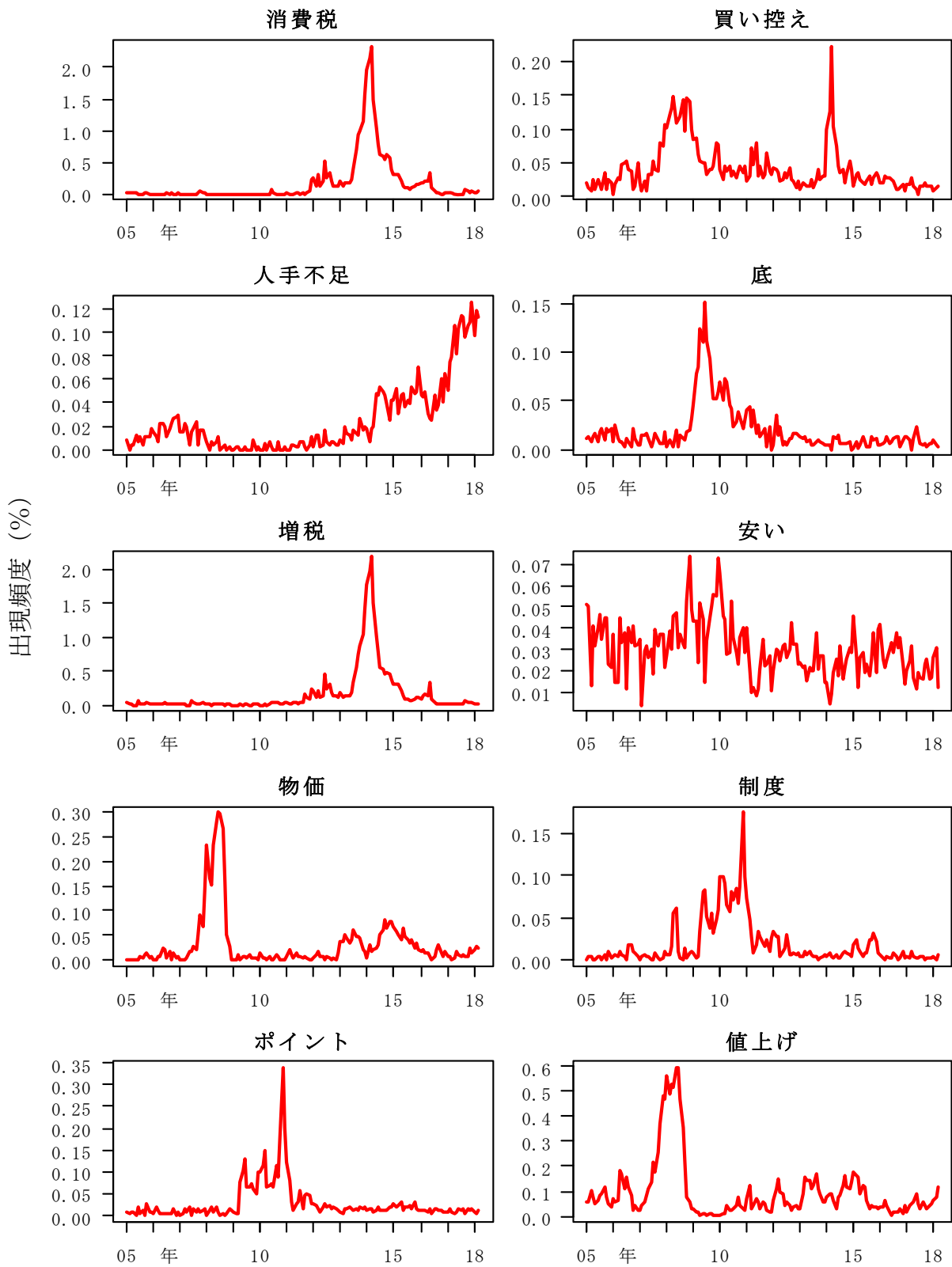


出現頻度 (%)

### 予想インフレ率に対する説明力が高い単語の出現頻度

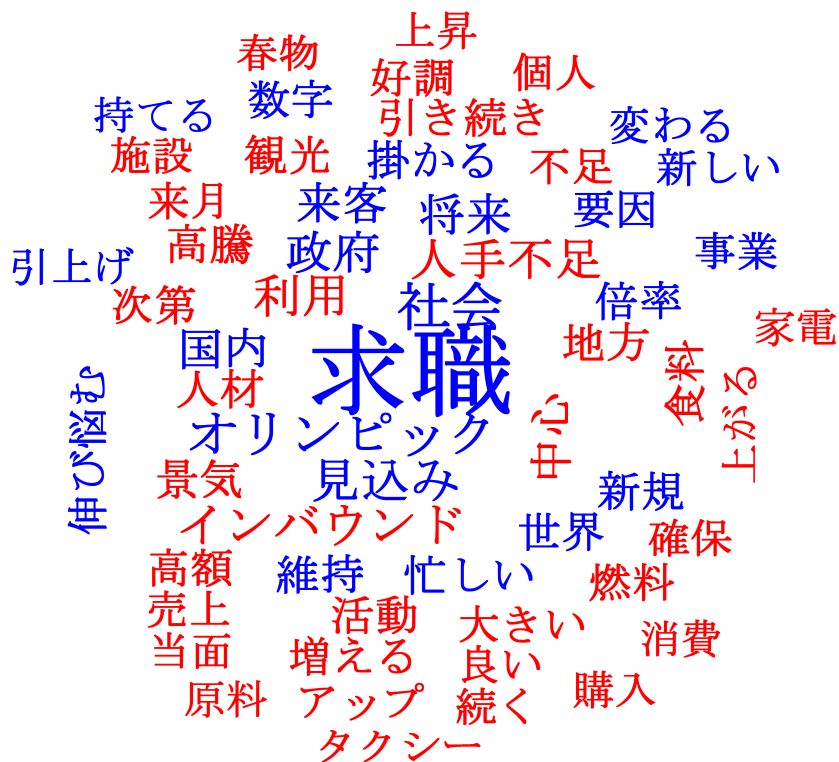


# CPI総合除く生鮮に対する説明力が高い単語の出現頻度



# ワードクラウド (予想インフレ率の変動要因分解)

<2018/1月>



<2014/7月>

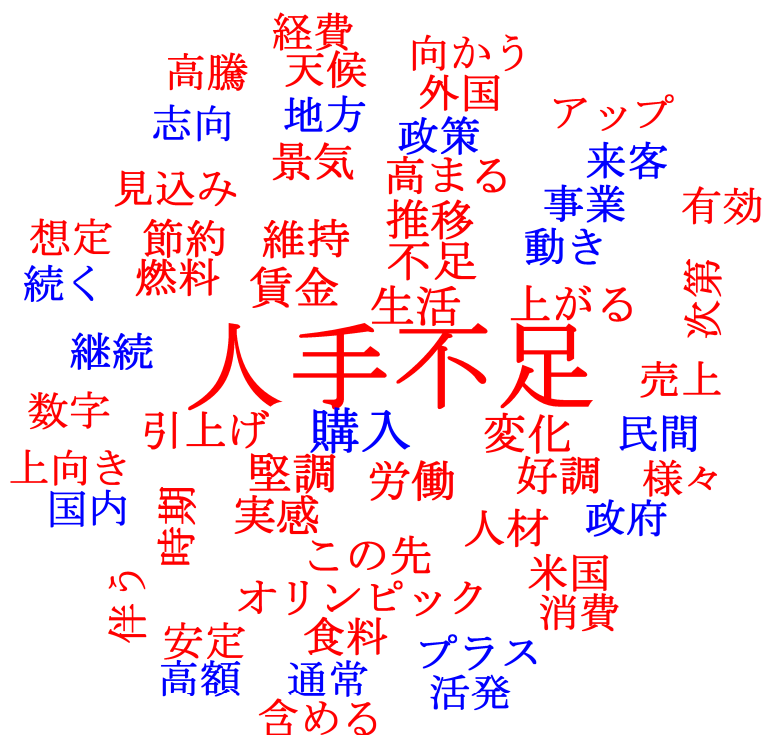




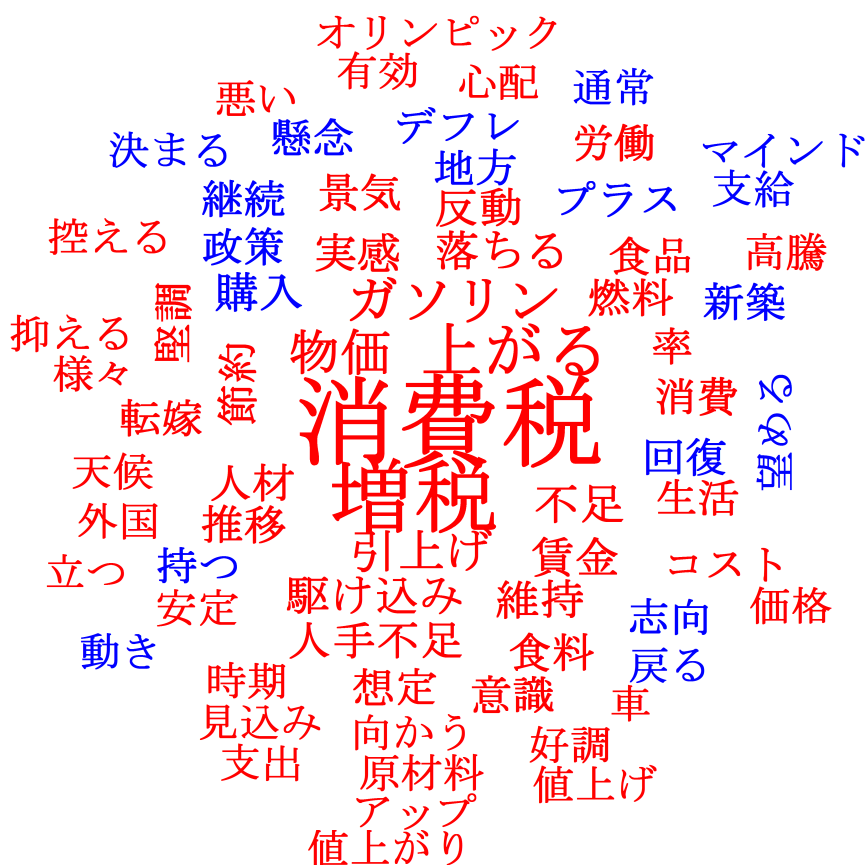


# ワードクラウド (C P I 実績の変動要因分解)

<2018/1月>

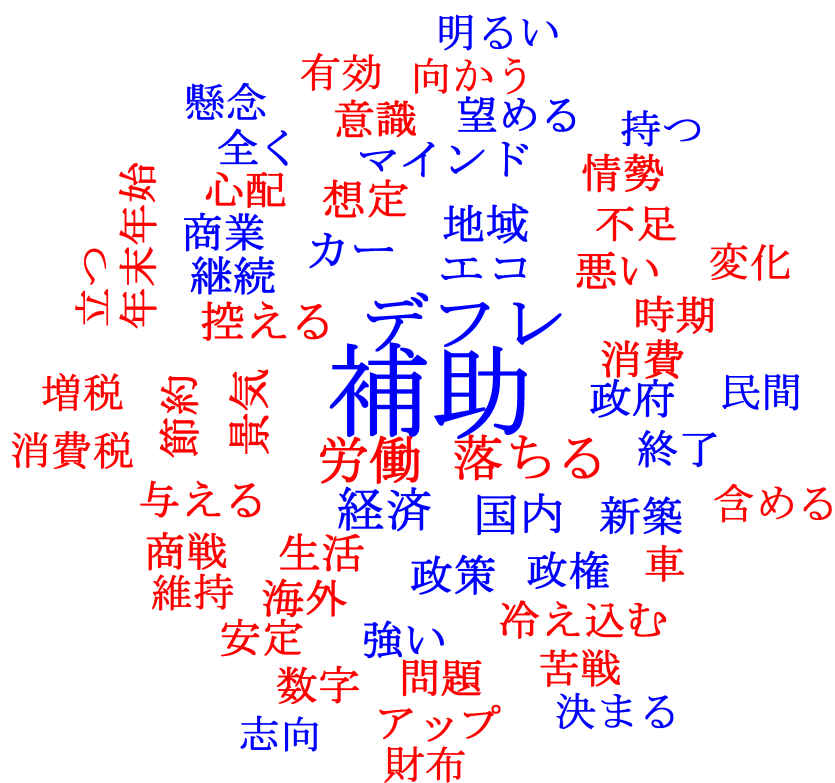


<2014/7月>



# ワードクラウド (C P I 実績の変動要因分解)

<2012/10月>



<2007/10月>

