

# 国会議事録を用いた経済指標のナウキャストイング

高杉 亮介<sup>†</sup> 山名 早人<sup>††</sup>

<sup>†</sup> 早稲田大学大学院 基幹理工学研究科 〒169-8555 東京都新宿区大久保 3-4-1

<sup>††</sup> 早稲田大学理工学術院 〒169-8555 東京都新宿区大久保 3-4-1

国立情報学研究所 〒101-8430 東京都千代田区一ツ橋 2-1-2

E-mail: †{takasugi,yamana}@yama.info.waseda.ac.jp

あらまし 政府による政策の効果を評価する際に様々な経済指標が使われている。しかし、既存の経済指標では統計調査に数ヶ月を要するため、政府による政策の効果を即座に評価することはできない。このような状況下において、近年ビッグデータを活用して足下の未公表データを予測し、景気動向を把握することが求められている。そこで本稿では、政府が将来の政策について議論した内容である国会議事録を分析することで、経済指標を予測する。具体的には、まず、政治家による発言内容から経済指標と相関の強い n-gram に着目し、政策運営状況を表す特徴量を得る。そして、得られた特徴量を用いて回帰分析と時系列解析との組み合わせにより経済指標を予測する。複数の経済指標をもとに本手法を実験した結果、従来の時系列解析による予測より小さい誤差で予測可能であることが示せた。

キーワード 経済指標予測, ナウキャストイング, 国会議事録

## 1. はじめに

政府が国策を実行・運営する上で、経済状況を定量的に把握しておくことは重要である。政府は、現状の経済状況を把握するため、あるいは将来予測を行うために、様々な経済指標を作成し使用している。経済指標がどの程度上下したか定量的に分析することで、政府が実行・運営する政策にどれほどの効果があったのかを評価することができる。例えば、政策の効果が目標より小さい場合には、より大きな影響を与えるよう政策を強化する。逆に、政策の効果が目標より大きい場合には、効果を抑えるよう政策を変更する。しかし、政府が統計をとり経済指標を作成するまでに数週間から数ヶ月を要するため、直近の経済指標でさえ公表に遅れが生じている [1] [2] [3]。したがって、政府による政策の効果を即座に評価することはできない。

従来から様々な機関がより直近の経済状況を把握するために、まだ公表されていない経済指標を予測するナウキャストイングと呼ばれる取り組みを行っている [4] [5]。ナウキャストイングを行うにあたって、これまで過去の統計や経済状況を踏まえて予測を行う経済学・ファイナンス分野における手法が試されてきた [6]。例えば、原ら [6] は、GDP (国内総生産)<sup>(注1)</sup> との関連が強い経済指標をもとにして実質 GDP 成長率を予測している。

このような状況下において、近年ビッグデータを用いて速報性のある経済指標を作成することや、足下の未公表データをナウキャストイングする取り組みが求められている [7] [8]。蔵本ら [9] は、新聞記事を分析して株価の騰落予測を行っている。蔵本らの研究では、投資家からのニーズに応えるために騰落正

答率で評価を行っており、株価指数そのものを予測誤差で定量的に評価するには至っていない。一方、POS データ (Point of Sales Data) を用いて消費者物価を指数化する取り組みがある [10] [11]。しかし、POS データが対象とする商品の範囲が小さく、消費者物価指数に対する説明能力が不十分である点が指摘されている [12]。Yahoo!JAPAN [13] は、検索クエリを用いて景気動向指数 (一致指数)<sup>(注2)</sup> の予測を行っている。しかし、景気動向に関わる事象は消費や雇用といった個人に関わる経済部門だけではなく、投融资や在庫管理といった企業に関わる経済部門も含んでいる [14]。したがって、インターネットユーザの興味を示すとどまる検索クエリデータでは、景気動向指数を説明する能力が不十分であると考えられる。

そこで本稿では、国会議事録<sup>(注3)</sup> のテキストデータを分析することで経済指標を予測することを試みる。国会で答弁される内容は国の政策に関わることであり、経済状況に大きな影響を与えるものである。また、国会では政策や法律が施行される前の時点における議論が行われている。よって、国会議事録は、未公表の経済指標を予測することや新たな先行指数を作成することに有用であると考えられる。具体的な手法としては、まず、国会で答弁された発言内容から経済指標との相関が強い形態素 n-gram に着目し、形態素 n-gram 出現回数と主成分分析により特徴量を得る。次に、得られた特徴量を用いて回帰分析により経済指標の増減予測を行う。さらに、回帰分析から得られる増減と時系列解析から得られる増減との組み合わせにより経済指標を予測する。本手法を用いて、経済状況を示す代表的な経

(注1) : 内閣府 : 国民経済計算 (GDP 統計), <http://www.esri.cao.go.jp/jp/sna/menu.html> (2014/12/23 アクセス)

(注2) : 内閣府 : 景気動向指数, <http://www.esri.cao.go.jp/jp/stat/di/menu.di.html> (2014/12/13 アクセス)

(注3) : 国立国会図書館 : 国会会議録検索システム, <http://kokkai.ndl.go.jp/> (2014/12/29 アクセス)

済指標である実質国内総生産（季節調整値）と消費者物価指数の予測を行う。

本稿は以下の構成をとる。まず、2. 節で関連研究について述べ、3. 節で提案手法について説明する。そして、4. 節で、実験・評価を行い、最後に、5. 節でまとめる。

## 2. 関連研究

従来より、経済指標予測に関する研究では、経済統計上の数値情報を用いて経済指標を推計する手法が発展してきた。近年では、データマイニング技術の発達に伴い、ビッグデータを用いて経済分析を行う取り組みが見られる。特に、各国中央銀行や調査機関が直近の未公表指標を予測するナウキャストイングと呼ばれる取り組みを行っている [4] [5] [6] [15] [16]。

以下では、経済指標予測に関連する研究として、経済統計情報を用いた研究とビッグデータを用いた研究についてまとめる。

### 2.1 経済統計情報を用いた経済指標予測

#### 2.1.1 GDP のナウキャストイングのための月次推計手法

従来は、GDP のナウキャストイング手法では、用いるデータを厳選してナウキャストイングを行うブリッジ・モデルが主流であった [15]。しかし、近年では、多数の情報を集約してナウキャストイングを行うファクター・モデルへと主流が移行している [16]。

こうした手法変化を踏まえた上で、ブリッジ・モデルとファクター・モデルの両方を組み合わせることで、扱うデータをバランスさせる手法を 2013 年に日本銀行の原ら [6] が提案した。原らは、速報性のある経済指標を用いて月次 GDP を推計し、実質 GDP 成長率をナウキャストイングした。具体的には、以下の 2 段階の構成をとっている。

#### ブリッジ・モデル過程

GDP 成長率を考える上で重要な鉱工業生産指数<sup>(注4)</sup> および第 3 次産業活動指数<sup>(注5)</sup> を用いて、GDP 変動の大部分を説明

#### ファクター・モデル過程

GDP 基礎統計から抽出した主成分を用いて、GDP 変動の残りを説明

また、GDP 基礎統計を主成分に集約する際、各基礎統計が持つ情報の種類に違いがあることに着目し、各基礎統計をグループ分けした。グループ分けは表 1 の通りである。

表 1 GDP 基礎統計のグループ分け

GDP 基礎統計	需要側統計	消費
		投資
		貿易
		その他
	供給側統計	

上記グループ分けにより予測モデルへの当てはまりが改善し

(注4)：経済産業省：鉱工業生産指数, <http://www.meti.go.jp/statistics/tyo/iip/> (2014/12/27 アクセス)

(注5)：経済産業省：第 3 次産業活動指数,

<http://www.meti.go.jp/statistics/tyo/sanji/> (2014/12/27 アクセス)

たという結果を得た。実質 GDP 成長率を予測した結果を平方平均二乗誤差により評価したところ、原らの手法では、日本経済研究センター [4] が発表する予測値と比較して平均的に高いパフォーマンスが得られた。

原らの手法は、用いるデータの選択やグループ分けが GDP に特化しており、ナウキャストイング対象を GDP 以外の経済指標に汎化することはできない。また、一般に、過去の経済統計情報から将来予測を行う手法は、用いる経済統計情報に依存する。したがって、原らのように過去に経験がない景気動向を考慮する場合は、経済統計情報に現れない情報も考慮する必要がある。

### 2.2 ビッグデータを用いた経済指標予測

#### 2.2.1 新聞記事を用いた株価指数の予測

2012 年、蔵本ら [9] は、新聞記事を分析することで株価指数を予測する研究を行った。蔵本らは、経済統計から得られる数値情報の解析だけで投資判断を下すことは困難であると論じている。その理由として、重要人物による発言や社会制度の変更、ニュースといった数値外の情報が欠落している点を挙げている。そこで、蔵本らは、数値情報には含まれない情報を持つテキストデータを分析することで、投資活動に役立つ市場予測を行った。

具体的には、日本経済新聞の記事を用いて、共起解析、主成分分析、重回帰分析により TOPIX（東証株価指数）<sup>(注6)</sup> や日経平均株価<sup>(注7)</sup>、19 業種の株価指数<sup>(注8)</sup> を予測した。

#### 共起解析

(1) 地方面以外の新聞記事データに対して形態素解析を行った。

(2) 同一文中から生成される動詞・名詞・形容詞のみからなる形態素 2-gram のうち、経済用語を含むものを抽出した。

(3) 抽出した形態素 2-gram の出現回数が 1 ヶ月間に 30 回以上であれば 1, 30 回未満であれば 0 として形態素 n-gram を数値化した特徴量を取得した。

(4) 特徴量の取得を 36 ヶ月分行った。

(5) 出現回数が 36 ヶ月中すべて 30 回未満であった形態素 2-gram を除去した。

#### 主成分分析

主成分分析により共起解析で得られた特徴量を 15 次元に次元削減し、15 主成分 × 36 ヶ月の行列を作成した。

#### 重回帰分析

主成分分析により 15 次元で表された特徴量を説明変数として 36 ヶ月分学習し、求める株価指数の増減を予測した。

経済状況を把握することで政府の政策運営に活かせるように

(注6)：株式会社東京証券取引所：東証株価指数（TOPIX）, <http://www.tse.or.jp/market/topix/topix.html> (2014/12/13 アクセス)

(注7)：日本経済新聞社：指数情報一日経平均プロフィール,

<http://indexes.nikkei.co.jp/nkave/index/profile?idx=nk225> (2014/12/13 アクセス)

(注8)：野村證券金融工学研究センター：NOMURA400,

<http://qr.nomura.co.jp/jp/n40/index.html> (2014/12/27 アクセス)

することを目的とする経済指標予測では定量性を考慮した予測・分析が必要である。例えば、経済状況が政府の目標とする景気回復量に達していなければ政策を強化しなければならない。逆に、経済状況が政府の目標とする景気回復量を超えていれば政策を緩和しなければならない。しかし、蔵本らは株価指数の騰落を正答率で評価しており、株価指数自体に対する定量的な予測性能分析には至っていない。

### 2.2.2 POS データを用いた消費者物価指数の動向分析

2007 年に一橋大学の阿部ら [10] が、2013 年には東京大学の渡辺ら [11] が、POS データを用いて日次物価指数を推計した。

阿部らと渡辺らは、日本経済新聞デジタルメディア社が提供する日次の POS データを用いた。この POS データはスーパーマーケットを中心とした日本全国に及ぶ小売店 280 店舗における販売情報を保持している。販売情報は販売商品や販売数、販売価格に関する情報を含んでいる。阿部らと渡辺らは、それぞれ独自に POS 物価指数を算出しているが、基本的には販売価格を販売数で加重平均したものを連鎖指数にするという考え方に則っている。

阿部らと渡辺らはそれぞれ、算出した POS 物価指数と総務省統計局が作成する公式の消費者物価指数とを比較し、マクロ経済分析に応用した。しかし、POS 物価指数と公式の消費者物価指数では用いているデータが違うため、必ずしも比較分析が適切とは言えないと総務省統計局は指摘している [12]。具体的には、POS 物価指数では、

- 扱うデータ対象が、加工食品と家事用品に限定されており家計消費の 2 割程度にすぎない
  - 商品の容量変化を考慮した価格調整が行われていない
  - 情報源が一部のスーパーマーケットに限られ、代表的な店舗で調査が行われているとは言えない
- といった注意点を挙げている。

### 2.2.3 検索クエリを用いた景気動向指数の予測

2013 年、Yahoo!JAPAN [13] は、足もとの景気を把握するために検索クエリを用いて回帰分析により景気動向指数（一致指数）を予測した。具体的には、

(1) 恒常的に検索される検索クエリから、検索数と景気動向指数との相関係数が 0.8 以上か 0.8 以下である検索クエリを抽出

(2) 抽出した強い相関を持つ検索クエリと類似する検索数推移をたどる検索クエリ群を足し合わせ

(3) 過学習を防ぐため回帰分析による学習と予測の誤差が近くなるような検索クエリを厳選

(4) 厳選した検索クエリの検索数を用いて重回帰分析により景気動向指数を算出  
という流れで予測を試みた。

しかし、以下のような理由から、検索クエリには景気動向指数を説明する能力は不十分であると考えられる。Yahoo!JAPAN が扱った検索クエリは、Yahoo!JAPAN が提供する検索エンジンを利用したインターネットユーザが持った興味・関心を示すにすぎない。したがって、検索クエリは個人消費や労働供給状況といった一般消費者による経済活動を説明する上ではある程

度適していると言える。一方、景気動向指数は幅広い経済部門を踏まえた指数となっている [14]。具体的には、a) 生産、b) 在庫、c) 投資、d) 雇用、e) 消費、f) 企業経営、g) 金融、h) 物価、i) サービスの各経済部門を指数化したものを合成している。よって、検索クエリには、企業の投融資や在庫管理といった企業活動も含めて指数化した景気動向指数を説明する能力は不十分である。

## 3. 国会議事録を用いた経済指標のナウキャストイング手法

2. 節で取り上げた既存研究では、

- 経済統計に反映されない情報を予測モデルに組み込めていない。
- 予測モデルに汎用性がなく特定の経済指標しか予測できない。
- 扱うデータに予測対象とする経済指標を説明する能力がない。

という問題があった。上記問題に対処するため、本手法では、社会・経済に大きな影響を与える国会での発言内容をデータとして用いる。そして、予測対象とする経済指標と関連のある特徴量を抽出して回帰分析を行うことで、様々な経済指標をナウキャストイングできるよう対処する。以下、本手法では、特徴量抽出・次元削減・回帰分析・予測値補正という 4 つの過程を経る。

### 3.1 国会議事録データ

本研究では、国会議事録<sup>(注9)</sup>をコーパスとして用いている。国会での発言は公式に記録され発言責任を伴うものであり、政策や法案を立案することに繋がる議論である。国会議事録はおおよそ 2, 3 週間のタイムラグで公開されている。よって、国会での発言は政治家が政治運営を行う上で重要であり、将来の日本社会に大きな影響を与える。以上を踏まえると、国会での発言を記録したものである国会議事録は、将来における日本社会や経済状況を予測する上で有用であると考えられることができる。

本研究で用いる国会議事録は、

- 衆議院
- 参議院
- 両院協議会・合同審査会等

の全議会を含んでいる。各議会で発言された内容が社会に影響を与える度合いは同程度と考える。よって、国会で開催された全議事を区別なくひとまとめにしたデータを扱う。この国会議事録に含まれる情報を表 2 に示す。上記国会議事録のうち、発言内容をテキストコーパスとして扱う。

国会議事録テキストコーパスから時系列で特徴量を得るため、下記の通り国会議事録トレーニングデータを定める。

(1) 発言日情報をもとに発言内容を 1 ヶ月ごとに分割する。ここでは年月  $t_m$  を変数  $m$  により

(注9)：村上 聡史：国会会議録データセット（対話コーパス）、[lod.sfc.keio.ac.jp/challenge2012/show\\_status.php?id=d036](http://lod.sfc.keio.ac.jp/challenge2012/show_status.php?id=d036), Linked Open Data Challenge Japan 2012 データセット部門 d036, (2013/10/15 アクセス)

表2 国会議事録のデータ構成

パラメータ	データ量
発言年月日	1947年～2012年のうち11,501日
会議上の発言者名	68,313通り
名寄せした発言者名	35,898名
名寄せした発言者の役職名	1,304種類
発言内容	48,400,884文
国会議事録検索システム URL	(全発言に URL が割り当て)

$$t_m = 100(2000 + \frac{m - \text{MOD}(m, 12)}{12}) + \text{MOD}(m, 12) + 1 \quad (1)$$

と定め、年月  $t_m$  の総発言内容を  $S_{t_m}$  とおく。ただし、 $\text{MOD}(m, 12)$  は  $m$  を 12 で割った余りを意味し、 $\text{MOD}(-1, 12) = 11, \text{MOD}(-2, 12) = 10, \dots$  とする。また、2000年1月を年月  $t_0 = 200001$ 、2000年2月を年月  $t_1 = 200002, \dots$  と表記する。以後、 $m = 0, 1, \dots, 130$  の範囲で変数  $m$  を考えることとする。

(2) 連続する6ヶ月分の発言内容を1つの国会議事録トレーニングデータとしてまとめる。つまり、1つの国会議事録トレーニングデータ  $T_{t_m}$  は

$$T_{t_m} = \sum_{j=m-5}^m S_{t_j} \quad (2)$$

で表される。

(3) 本手法では学習期間を24ヶ月とする。よって、24ヶ月分の国会議事録トレーニングデータ

$$T_{t_m}, T_{t_{m+1}}, \dots, T_{t_{m+23}} \quad (3)$$

を用意する。ここで、 $T_{t_m}$  と  $T_{t_{m+1}}$  の間には5ヶ月分の発言内容に重なりがある点に注意されたい。

以上で定めた国会議事録トレーニングデータから特徴量を抽出する。

### 3.2 形態素 n-gram の出現回数抽出

本項では、国会議事録トレーニングデータである国会議事録の発言内容から特徴量を抽出する手法について詳述する。特徴量抽出については、

(1) 国会議事録トレーニングデータにおける年月情報と経済指標における年月情報との時間間隔を設定

(2) 国会議事録トレーニングデータを1文ごとに分割

(3) 特定の品詞にあたる単語を抽出

(4) 抽出した単語から1国会議事録トレーニングデータごと

に形態素 n-gram を生成

(5) 24の国会議事録トレーニングデータから時系列 n-gram を用意

(6) 予測性能に優れた形態素 n-gram を厳選  
という手順をとる。

#### 3.2.1 国会議事録と経済指標とのタイムラグ設定

国会での発言によっては、経済状況に及ぼす影響が即座に現れるものと遅れて生じるものがあると考えられる。本手法では、国会での発言が経済状況に影響するまでのタイムラグをパラメータ  $l$  として取り入れる。年月  $t_m$  における経済指標を  $y_{t_m}$  とお

くとき、経済指標  $y_{t_m}$  と対応して学習に用いる国会議事録トレーニングデータは、 $T_{t_{m-l}}$  となる。最適なタイムラグ  $l$  を探索するため、 $l = 1, 2, \dots, 12$  と月数で変化させることとする。

#### 3.2.2 特定の品詞にあたる単語の抽出

3.2.1 で述べた国会議事録トレーニングデータに対して形態素解析エンジン MeCab<sup>(注10)</sup> を用いて形態素解析を行う。分解した形態素のうち、政策運営に関連する単語が多く含まれる品詞として動詞、名詞、形容詞、副詞のみを考え抽出する。ただし、ストップワードを取り除くため、非自立語、代名詞、形容詞接尾語、動詞接尾語、1文字の単語は除去する。

#### 3.2.3 形態素 n-gram の生成

国会で意見・現況報告・質疑と論理が展開するような、1文ごとに趣旨が区切れる場合を配慮し、3.2.2 項で抽出した単語をもとに1文単位で形態素 n-gram を生成する。本手法では、予測誤差を縮小すべく、形態素 n-gram における  $n$  の値を予測対象とする経済指標により探索的に調整する。形態素 n-gram の出現回数は、国会議事録トレーニングデータ1つごとに総和をとる。国会議事録トレーニングデータを24ヶ月分用いることから、24ヶ月に渡る形態素 n-gram を時系列データ  $C$  として考えることができる。ここで、24ヶ月分の国会議事録トレーニングデータ  $T_{t_{m-l}}, T_{t_{m-l+1}}, \dots, T_{t_{m-l+23}}$  から生成された形態素 n-gram が  $I$  種類あるとする。このとき、国会議事録トレーニングデータ  $T_{t_{m-l}}$  における  $I$  種類の形態素 n-gram 出現回数を  $c_{t_{m-l},0}, c_{t_{m-l},1}, \dots, c_{t_{m-l},I-1}$  とおく。ただし、1つの国会議事録トレーニングデータにおいて生成されなかった形態素 n-gram 出現回数は0回とする。以上をもとに時系列 n-gram  $C$  を出現回数を要素とする行列で

$$C = \begin{pmatrix} C_{t_{m-l}} \\ \vdots \\ C_{t_{m-l+23}} \end{pmatrix} = \begin{pmatrix} c_{t_{m-l},0} & \dots & c_{t_{m-l},I-1} \\ \vdots & \ddots & \vdots \\ c_{t_{m-l+23},0} & \dots & c_{t_{m-l+23},I-1} \end{pmatrix} \quad (4)$$

と表現する。

#### 3.2.4 閾値以上出現する形態素 n-gram の選出

国会議事録から生成される頻度の少ない形態素 n-gram で示される特徴は、国会で大きな議論となっていないものであると考えられる。そのため、頻度の少ない形態素 n-gram が社会・経済へ与える影響は小さく、特徴量として選択しない方がよいと判断した。そこで、3.2.3 項で生成した24ヶ月分の時系列 n-gram において、一度も一定回数以上出現しなかった形態素 n-gram を除去する。ここで、閾値を  $\theta_c$  とおくと、形態素 n-gram の出現回数が

$$c_{t_{m-l+j},i} < \theta_c \quad (j = 0, 1, \dots, 23) \quad (5)$$

を満たす形態素 n-gram  $i$  は時系列 n-gram  $C$  から削除する。なお、本手法では特徴量取得にて十分な次元を確保できるよう、閾値を  $\theta_c = 20$  に設定する。

(注10) : MeCab:

[http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html\(2015/1/5アクセス\)](http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html(2015/1/5アクセス))

### 3.2.5 経済指標と相関が強い時系列 n-gram の厳選

国会議事録トレーニングデータから生成した時系列 n-gram  $C$  のうち、予測対象とする経済指標と相関があるものを厳選する。これにより、予測対象とする経済指標の説明に適した特徴量を作成する。具体的には、まず、形態素 n-gram  $i$  の時系列出現回数  $c_{t_{m-l+j},i}$  ( $j = 0, 1, \dots, 23$ ) と、トレーニングデータとする時系列経済指標  $y_{t_{m+j}}$  ( $j = 0, 1, \dots, 23$ ) から、ピアソンの積率相関係数  $Cor_i$  を計算する。ピアソンの積率相関係数  $Cor_i$  は

$$Cor_i = \frac{\sum_{j=0}^{23} (c_{t_{m-l+j},i} - \bar{c}_i)(y_{t_{m+j}} - \bar{y})}{\sqrt{\sum_{j=0}^{23} (c_{t_{m-l+j},i} - \bar{c}_i)^2} \sqrt{\sum_{j=0}^{23} (y_{t_{m+j}} - \bar{y})^2}} \quad (6)$$

で与えられる。ここで、 $\bar{c}_i$ ,  $\bar{y}$  はそれぞれ形態素 n-gram  $i$  の時系列出現回数および時系列経済指標の算術平均を指す。ピアソンの積率相関係数  $Cor_i$  は

$$-1 \leq Cor_i \leq 1 \quad (7)$$

の範囲で値をとる。単に相関・相関係数というときは、ピアソンの積率相関係数を指すことが一般的である。また、 $|Cor_i|$  が 1 に近いとき相関が強い、 $|Cor_i|$  が 0 に近いとき相関が弱いという。そして、上記相関係数  $Cor_i$  を計算した後、一定の相関を持つ形態素 n-gram のみを厳選する。つまり、閾値を  $\theta_r$  とおくと、

$$|Cor_i| \geq \theta_r \quad (8)$$

を満たす形態素 n-gram  $i$  のみを抽出する。十分な次元を持つ特徴量を取得できるよう、閾値  $\theta_r$  は予測対象とする経済指標ごとに探索して定める。

以上により、予測性能に優れたデータとして、予測対象とする経済指標と相関がある時系列 n-gram

$$C' = \begin{pmatrix} C'_{t_{m-l}} \\ \vdots \\ C'_{t_{m-l+23}} \end{pmatrix} = \begin{pmatrix} c'_{t_{m-l},0} & \cdots & c'_{t_{m-l},I'-1} \\ \vdots & \ddots & \vdots \\ c'_{t_{m-l+23},0} & \cdots & c'_{t_{m-l+23},I'-1} \end{pmatrix} \quad (9)$$

が得られる。ただし、3.2.4 で削除して残る形態素 n-gram、かつ本項の式 (8) を満たす形態素 n-gram が  $I'$  種類 ( $I' \leq I$ ) 厳選されたとする。

### 3.3 主成分分析による次元削減

3.2 節で取得した特徴量  $C'$  は  $I'$  次元  $\times$  24 ヶ月間の行列で表される。特徴量  $C'$  を説明変数に用いて回帰分析を行う場合は、

- 次元の罫による予測精度悪化
- ランク落ちによる不適当な回帰式生成

といった問題が生じる可能性がある。この問題を回避するため、特徴量  $C'$  を主成分分析により次元削減する。主成分分析とは、多次元データにおいて変数間の相関が強いものを排除し、相関が弱い変数に合成していくことで次元数を削減する手法である。特徴量  $C'$  の次元  $I'$  を第  $d$  主成分まで次元削減したものを

$$X = \begin{pmatrix} X_{t_{m-l}} \\ \vdots \\ X_{t_{m-l+23}} \end{pmatrix} = \begin{pmatrix} x_{t_{m-l},0} & \cdots & x_{t_{m-l},d-1} \\ \vdots & \ddots & \vdots \\ x_{t_{m-l+23},0} & \cdots & x_{t_{m-l+23},d-1} \end{pmatrix} \quad (10)$$

とおく。次元  $d$  の値は予測対象とする経済指標によって予測誤差が小さくなるよう調整した。以下では、次元削減した特徴量  $X$  を回帰分析における説明変数として用いる。

### 3.4 重回帰分析による予測

経済指標を予測するにあたって重回帰分析を行う。重回帰分析では、説明変数には国会議事録から生成した特徴量  $X$  を、目的変数には予測対象とする経済指標の増減  $\Delta y_{t_{m+j}}$  ( $j = 0, 1, \dots, 23$ ) を用いる。経済指標の増減  $\Delta y_{t_{m+j}}$  は  $24 \times 1$  行列  $\Delta Y$  で表記し、

$$\Delta Y = \begin{pmatrix} \Delta y_{t_m} \\ \vdots \\ \Delta y_{t_{m+23}} \end{pmatrix} = \begin{pmatrix} y_{t_m} - y_{t_{m-1}} \\ \vdots \\ y_{t_{m+23}} - y_{t_{m+22}} \end{pmatrix} \quad (11)$$

により算出する。以上の説明変数  $X$  と目的変数  $\Delta Y$  を用いて予測モデルとする線形重回帰式

$$\Delta Y = a_0 + Xa + \varepsilon \quad (12)$$

を構築する。ただし、 $a_j$  は回帰係数、 $\varepsilon$  は平均 0、分散  $\sigma^2$  の正規分布に従う残差である。上記回帰式の当てはめを行った後、25 ヶ月目における特徴量  $X_{t_{m-l+24}}$  を説明変数として外挿予測を行う。外挿予測により算出される値  $\Delta y_{t_{m+24}}$  が 25 ヶ月目にあたる経済指標の増減予測値となる。

### 3.5 予測値の補正

予測誤差を縮小し、経済指標の推移も学習するため、3.4 項で導出した外挿予測値  $\Delta y_{t_{m-l+24}}$  を写像により補正し、さらに、時系列解析と組み合わせて予測値を算出した。

#### 3.5.1 偏差値を用いた写像

時系列データが急に大きな変動をとることは少ないものと考えられる。この観点から、外挿予測値の絶対値  $|\Delta y_{t_{m-l+24}}|$  が著しく大きい場合には値が小さくなるよう補正する。本手法では、時系列データから算出した偏差値を用いて外挿予測値の写像を行う。具体的には、まず、外挿予測値を含めた 6 ヶ月分の経済指標増減

$$\Delta y_{t_{m-l+j}} \quad (j = 19, 20, \dots, 24) \quad (13)$$

からなる時系列を考える。そして、時系列経済指標増減  $\Delta y_{t_{m-l+j}}$  ( $j = 19, 20, \dots, 24$ ) における外挿予測値  $\Delta y_{t_{m-l+24}}$  の偏差値  $Tscore_{t_{m-l+24}}$  を

$$Tscore_{t_{m-l+24}} = 50 + \frac{10(\Delta y_{t_{m-l+24}} - \overline{\Delta y_{t_{m-l+j}}})}{\sigma} \quad (14)$$

により求める。ただし、 $\overline{\Delta y_{t_{m-l+j}}}$  は  $\Delta y_{t_{m-l+j}}$  の算術平均、 $\sigma$  は時系列経済指標増減  $\Delta y_{t_{m-l+j}}$  の標準偏差である。次に、時系列経済指標増減  $\Delta y_{t_{m-l+j}}$  より 1 ヶ月前の経済指標増減

$$\Delta y_{t_{m-l+j-1}} \quad (j = 19, 20, \dots, 24) \quad (15)$$

からなる時系列を考える。最後に、時系列経済指標増減  $\Delta y_{t_{m-l+j-1}}$  により仮定される母集団から推計される偏差値  $Tscore_{t_{m-l+24}}$  を持つ値  $\Delta y'_{t_{m-l+24}}$  を

$$\Delta y'_{t_{m-l+24}} = \overline{\Delta y_{t_{m-l+j-1}}} + \frac{(Tscore_{t_{m-l+24}} - 50)\sigma'}{10} \quad (16)$$

により算出する。ただし、 $\overline{\Delta y_{t_{m-l+j-1}}}$  は  $\Delta y_{t_{m-l+j-1}}$  の算術平均、 $\sigma'$  は  $\Delta y_{t_{m-l+j-1}}$  の標準偏差である。以上により求めた  $\Delta y'_{t_{m-l+24}}$  を外挿予測値からの写像結果とする。

### 3.5.2 AR モデルの導入

本研究では、時系列解析を導入することで、国会議事録を学習するだけでなく過去の経済指標がたどった推移も学習することができることから、予測誤差縮小を見込む。以下では、3.5.1 で求めた写像結果  $\Delta y'_{t_{m-l+24}}$  と AR モデルによる予測値を組み合わせる。AR モデルは、時系列解析における代表的なモデルであり、

$$y_{t_{m-l}} = a_0 + \sum_{j=1}^p a_j y_{t_{m-l-j}} + \varepsilon \quad (17)$$

で表される回帰式に時系列データ  $y_{t_m}$  を当てはめてモデルを構築する。ただし、 $p$  は次数、 $a_0$  は切片、 $a_j$  は自己回帰係数、 $\varepsilon$  は残差である。本手法で用いる AR モデルでは次数  $p$  を 1 とし、24 ヶ月分の経済指標  $y_{t_{m-l+j}}$  ( $j = 0, 1, \dots, 23$ ) を用いて

$$y_{t_{m-l+j}} = a_0 + a_1 y_{t_{m-l+j-1}} + \varepsilon_j \quad (j = 1, \dots, 23) \quad (18)$$

によりモデル構築を行う。AR モデルによる経済指標の増減予測値  $\Delta yAR_{t_{m-l+24}}$  は AR モデルによる予測値  $yAR_{t_{m-l+24}}$  と経済指標  $y_{t_{m-l+23}}$  を用いて

$$\Delta yAR_{t_{m-l+24}} = yAR_{t_{m-l+24}} - y_{t_{m-l+23}} \quad (19)$$

により算出する。そして、算出した AR モデルによる増減予測値  $\Delta yAR_{t_{m-l+24}}$  を 3.5.1 で求めた増減予測値  $\Delta y'_{t_{m-l+24}}$  と内分点により組み合わせる。具体的には、AR モデルによる増減予測値  $\Delta yAR_{t_{m-l+24}}$  と国会議事録から導出した増減予測値  $\Delta y'_{t_{m-l+24}}$  を  $w : (1-w)$  に内分する点  $\Delta y''_{t_{m-l+24}}$  を、

$$\Delta y''_{t_{m-l+24}} = w \Delta y'_{t_{m-l+24}} + (1-w) \Delta yAR_{t_{m-l+24}} \quad (20)$$

$$0 < w < 1$$

により計算する。内分点を計算した値  $\Delta y''_{t_{m-l+24}}$  が本手法による増減予測値となる。内分比率  $w : (1-w)$  は予測誤差が小さくなるように予測対象とする経済指標によって調整する。

以上より、最終的な経済指標の予測値  $y_{t_{m-l+24}}$  は、時系列データと本手法による増減予測値を足し合わせて

$$y_{t_{m-l+24}} = y_{t_{m-l+23}} + \Delta y''_{t_{m-l+24}} \quad (21)$$

と計算する。

## 4. 実験・評価

本節では、3. 節で提案した手法にもとづき、実質 GDP (季節調整値) と消費者物価指数をナウキャストした実験について述べる。

### 4.1 評価方法

#### 4.1.1 評価基準

本稿で行った実験には、2011 年 1 月から 2012 年 12 月までの 24 ヶ月間に渡る経済指標を予測対象とした。テストデータを予測した 24 ヶ月分の予測値に対して、

- ピアソンの積率相関係数 (Pearson's product-moment correlation coefficient)
- 平均絶対誤差 (Mean Absolute Error)
- 平均絶対誤差率 (Mean Absolute Percentage Error)
- 平方平均二乗誤差 (Root Mean Squared Error)

を評価基準として採用し、それぞれ計算した。

#### 4.1.2 外挿予測による最適なパラメータの探索

3. 節で述べたように、本手法には様々なパラメータの取り方がある。予測誤差が小さくなる最適なパラメータは、トレーニングデータだけを用いた回帰分析における当てはまり (内挿予測) から判断することができる。しかし、回帰分析における当てはまりが改善するパラメータを特定したとしても、過学習によりテストデータを用いた外挿予測結果は悪化する可能性がある [13]。そこで、本研究では回帰分析の当てはまりによりパラメータを決定するのではなく、様々なパラメータで外挿予測を試行し、最適なパラメータを探索することとする。以後、様々なパラメータで外挿予測を試行した中で、最も予測誤差が小さかったパラメータを掲載する。そして、予測誤差が最小となったときのパラメータを適用したときの結果を図示する。

#### 4.1.3 ベースラインとなる手法との比較

本手法を経済指標のナウキャストに適用するにあたって、国会議事録が有用であることを調査するため AR モデルをベースラインとして選択し比較した。AR モデルによる予測を 24 回試行し、予測値を本手法と同様に算出した。また、本手法によるナウキャストが AR モデルによる予測よりも小さい誤差で実現できた場合、誤差の差分が有意であるかを対応のある t 検定により調査した。なお、有意差を判定する際に考慮する信頼区間は 95% とした。

### 4.2 実質 GDP のナウキャスト

3. 節で提案した手法を用いて実質 GDP (国内総生産、季節調整値) をナウキャストした実験結果を報告する。

#### 4.2.1 月次実質 GDP のナウキャスト

GDP は内閣府から発表されるが、四半期ごとに更新される指標となっている。そのため、月次実質 GDP をナウキャストする際に内閣府公表の実質 GDP をテストデータとして用いることはできない。そこで本研究では、日本経済研究センター [4] が公開している月次実質 GDP 推計値 (季節調整値) (注11) をトレーニングデータおよびテストデータとして用いた。

各種パラメータを探索的に設定し、実質 GDP のナウキャストを行った中で、平方平均二乗誤差  $RMSE$  が最小となったパラメータを表 3 に示す。表 3 にあるパラメータを設定して実質 GDP をナウキャストした結果を表 4 に示す。

(注11) : 日本経済研究センター : 月次 GDP (実質季節調整済系列実質国内総支出), <http://www.jcer.or.jp/research/gdp/index.html> (2015/1/6 アクセス)

また、t 検定により予測誤差に有意差が認められるか検定したところ、p-値が  $0.04900 < 0.05$  となった。よって、実質 GDP のナウキャストイングにおいて本手法が AR モデルより有意な差を持つ予測誤差で予測可能であることが示された。表 3 にあるパラメータを設定したときのナウキャストイング結果を図 1 に示す。

表 3 月次実質 GDP のナウキャストイングに設定したパラメータ

パラメータの種類	パラメータの値
特徴量	3-gram
ラグ $l$	10 ヶ月
相関 $ \theta_r $	0.6 以上
次元削減後の次元数 $d$	20 次元
AR モデル予測との内分点 $w : (1 - w)$	$w = 0.1$

表 4 月次実質 GDP のナウキャストイング結果

評価基準	提案手法	AR モデルによる予測
$Cor$	0.6525	0.5967
$MAE$	4047.61(十億円)	4492.92(十億円)
$MAPE$	0.7962(%)	0.8820(%)
$RMSE$	6381.38	6888.11

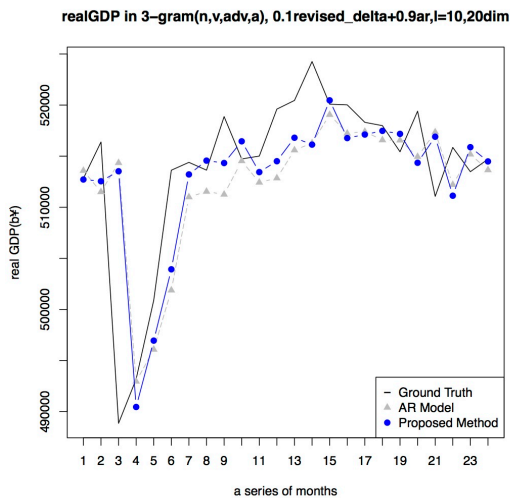


図 1 月次実質 GDP のナウキャストイング結果

#### 4.2.2 四半期別実質 GDP のナウキャストイング

本研究では、4.2.1 で示した月次実質 GDP を用いて四半期別実質 GDP を推計する。ここで、第  $k$  四半期実質 GDP を  $rGDP_{Qk}$ 、 $m$  月実質 GDP を  $rGDP_{Mm}$  とおき、式 (22a)~式 (22d) により四半期別実質 GDP  $rGDP_{Qk}$  を推計することとする。なお、パラメータは表 3 に示すものと同じ条件を適用した。提案手法と AR モデルそれぞれにより予測した 24 ヶ月分月次実質 GDP から、8 四半期実質 GDP を推計した際の評価を表 5 に示す。また、提案手法と AR モデルによる予測誤差の差異における有意性を t 検定により検証した。t 検定の結果、p-値

は  $0.0450 < 0.05$  となり、AR モデルより有意な差で本手法の予測誤差が小さいということが示された。四半期別実質 GDP を推計した結果を図 2 に示す。

$$\left\{ \begin{array}{l} nGDP_{Q1} = \frac{1}{3} \sum_{j=1}^3 (nGDP_{Mj}) \quad (22a) \\ nGDP_{Q1} = \frac{1}{3} \sum_{j=4}^6 (nGDP_{Mj}) \quad (22b) \\ nGDP_{Q1} = \frac{1}{3} \sum_{j=7}^9 (nGDP_{Mj}) \quad (22c) \\ nGDP_{Q1} = \frac{1}{3} \sum_{j=10}^{12} (nGDP_{Mj}) \quad (22d) \end{array} \right.$$

表 5 四半期別実質 GDP のナウキャストイング結果

評価基準	提案手法	AR モデルによる予測
$Cor$	0.8366	0.8053
$MAE$	3472.99(十億円)	4166.64(十億円)
$MAPE$	0.6781(%)	0.8124(%)
$RMSE$	3981.50	4555.14

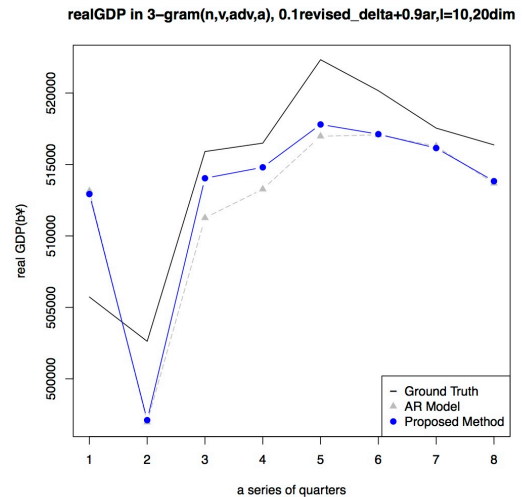


図 2 四半期別実質 GDP のナウキャストイング結果

#### 4.3 消費者物価指数のナウキャストイング

3. 節で提案した手法を用いて消費者物価指数をナウキャストイングした実験結果について述べる。トレーニングデータおよびテストデータとなる消費者物価指数は平成 22 年基準のラスパイレース指数を用いた<sup>(注12)</sup>。

各種パラメータを探索的に設定し消費者物価指数をナウキャストイングした中で、平方平均二乗誤差  $RMSE$  が最小となったパラメータを表 6 に示す。表 6 で設定したパラメータを適

(注12)：総務省統計局：消費者物価指数(中分類指数全国総合), [http://www.e-stat.go.jp/SG1/estat/List.do?bid=000001033702&cycode=0\(2015/1/6アクセス\)](http://www.e-stat.go.jp/SG1/estat/List.do?bid=000001033702&cycode=0(2015/1/6アクセス))



用して消費者物価指数のナウキャストイングを行った結果を表7に示す。また、t検定によりARモデルと比較して予測誤差に有意差が得られたか検定したところ、p値が0.03115 < 0.05となった。t検定の結果から、本手法が消費者物価指数をナウキャストイングする際、ARモデルより有意な差で予測可能であることが示せた。表6にあるパラメータを設定したときのナウキャストイング結果を図3に示す。

表6 消費者物価指数のナウキャストイングに設定したパラメータ

パラメータの種類	パラメータの値
特微量	2-gram と 3-gram の両方
ラグ $l$	4ヶ月
相関 $ \theta_r $	0.5 以上
次元削減後の次元数 $d$	15 次元
ARモデル予測との内分点 $w : (1-w)$	$w = 0.2$

表7 消費者物価指数のナウキャストイング結果

評価基準	提案手法	ARモデルによる予測
$Cor$	0.7865	0.7371
$MAE$	0.1431	0.1736
$MAPE$	0.1436(%)	0.1742(%)
$RMSE$	0.2004	0.2252

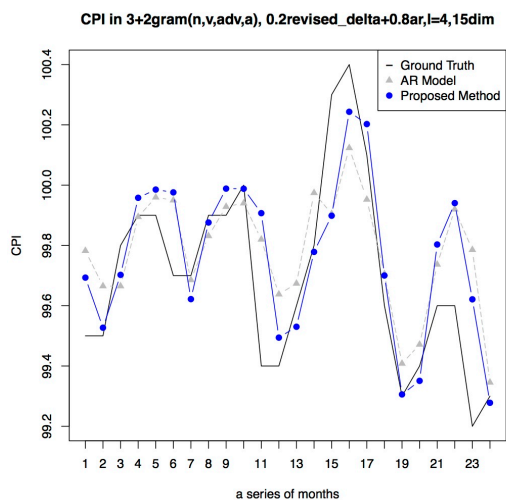


図3 消費者物価指数のナウキャストイング結果

## 5. まとめ

本稿では、国会議事録を用いて経済指標をナウキャストイングする手法を提案した。提案手法により複数の経済指標をナウキャストイングする実験を行った。実験の結果、実質GDP(季節調整値)を8四半期ナウキャストイングした場合は平方平均二乗誤差が3981.50、消費者物価指数を24ヶ月ナウキャストイングした場合の平方平均二乗誤差は0.2004となった。この結果を国会議事録に依存しないARモデルによる予測と比較し、

より小さい予測誤差でナウキャストイングすることが可能であることを示すことができた。

今後はより小さい誤差でナウキャストイングを行えるよう手法の分析・改善に取り組みたい。本手法における問題点として、24ヶ月という学習期間では回帰分析の当てはまりが十分でない可能性がある。また、国会での発言内容は、発言者によって社会に対する影響度が異なると考えられるが、本手法では発言者に関する情報は考慮していない。以上から、

- トレーニングデータを長期間とり、回帰分析の学習期間を拡大する。
  - 国会での発言者によって特徴量に重み付けを加味する。
- といった改善点が挙げられる。

## 参考文献

- [1] 総務省統計局：統計結果の公表情報，[http://www.stat.go.jp/data/kouhyou/e-stat\\_cpi2014.xml](http://www.stat.go.jp/data/kouhyou/e-stat_cpi2014.xml)(2014/12/13アクセス)
- [2] 内閣府：統計結果の公表情報，[http://www.esri.cao.go.jp/jp/stat/di/e-stat\\_di.xml](http://www.esri.cao.go.jp/jp/stat/di/e-stat_di.xml)(2014/12/13アクセス)
- [3] 内閣府：公表予定-内閣府，[http://www.esri.cao.go.jp/jp/sna/kouhyou/kouhyou\\_top.html](http://www.esri.cao.go.jp/jp/sna/kouhyou/kouhyou_top.html)(2014/12/13アクセス)
- [4] 日本経済研究センター：経済予測，<http://www.jcer.or.jp/research/index.html>(2014/12/13アクセス)
- [5] みずほ総合研究所：経済指標解説，<http://www.mizuho-ri.co.jp/publication/research/indicator/index.html>(2014/12/13アクセス)
- [6] 原 尚子，山根 渉太郎：“GDPのナウキャストイング(足もと予測)のための新たな月次推計手法：日本への応用例”，日本銀行調査統計局 Working Paper, 13-E-14, 2013年10月
- [7] 内閣府：平成26年度予算概算要求，[http://www.cao.go.jp/yosan/soshiki/h26/gaiyou\\_h26.pdf](http://www.cao.go.jp/yosan/soshiki/h26/gaiyou_h26.pdf)(2014/12/13アクセス)
- [8] 内閣府：平成27年度予算概算要求，[http://www.cao.go.jp/yosan/soshiki/h27/gaiyou\\_h27.pdf](http://www.cao.go.jp/yosan/soshiki/h27/gaiyou_h27.pdf)(2014/12/13アクセス)
- [9] 蔵本 貴之，和泉 潔，吉村 忍，石田 智也，中嶋 啓浩，松井 藤五郎，吉田 稔，中川 裕志：“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, pp.291-296, 2012年
- [10] 阿部 修人，外木 暁幸：“価格改定頻度及び特売と消費者物価指数の動向-大規模POSデータに基づく考察-”，一橋大学経済研究所物価研究センター Working Paper, No.15, 2007年
- [11] 渡辺 広太，渡辺 努：“スキャナーデータを用いた日次物価指数の計測”，東京大学金融教育研究センターワーキングペーパー，CARF-J-094, 2013年
- [12] 総務省統計局：統計局ホームページ/消費者物価指数に関するQ&A(回答)，<http://www.stat.go.jp/data/cpi/4-1.htm#G9>(2014/12/13アクセス)
- [13] Yahoo!JAPAN ビッグデータレポート：景気の「今」を把握することは可能か？，<http://event.yahoo.co.jp/bigdata/keiki/>(2014/12/13アクセス)
- [14] 内閣府：景気動向指数の利用の手引き，<http://www.esri.cao.go.jp/jp/stat/di/di3.html#link002>(2014/12/13アクセス)
- [15] Baiffi, Alberto, Roberto Golinelli and Giuseppe Parigi: Bridge Models to Forecast the Euro Area GDP, International Journal of Forecasting, Vol.20, pp.447-460 (2004)
- [16] Banbura, Marta, Domenico Giannone, Michele Modugno and Lucrezia Reichlin: Now-casting and the Real-time Data Flow, ECB Working Paper, No. 1564 (2013)