３－３　各種経済（景気）指標との関係

本項では、まず景気ウォッチャー調査の「Ⅰ.景気の現状に対する判断理由等」に対応するテキストデータ（表記の簡単のために、本章では判断理由欄と記すこととする）から、同期の現状判断 DI 及び先行き判断 DI の原数値を予測する分析を行う。すなわち､前項までの分析で景気指標との関係性があるとして抽出された有意単語 400 語の、特定年月における登場頻度を主たる説明変数とし、当該年月の現状判断 DI と先行き判断 DI の原数値を目的変数とした回帰分析（を拡張した異種混合学習）を行うことで、そのような予測を行う予測式が構成できないかの検討を行う。

加えて､各アンケート回答者の現状判断である「良くなった」､「やや良くなった」､「変わらない」、「やや悪くなった」、「悪くなった」を、100 点、75 点、50 点、25 点、0 点という数値とみなして目的変数とし、該当する回答の判断理由欄における有意単語 400 語の登場頻度を説明変数とした異種混合学習を行うことで、数値としてアンケート回答者の現状判断を予測し、それを平均することで DI 原数値を予測することができないかの検討も行う。さらに、先行き判断についても同様の検討を行う。

本項では大きく分けて 4 種類の分析 A～D を実施している。

表 3－3－1　３－３における分析の概要

	分析	条件設定	
景気 DI の予測	A	分析内容	有意単語 400 語の発生頻度による全国現状判断 DI の予測
		分析モデル単位	月次
		学習期間	2010 年 1 月～2012 年 12 月（36 か月）
		予測期間	2013 年 1 月～2013 年 12 月（12 か月）
		説明変数	パターン①：有意単語 400 語の総登場頻度
			パターン②：DI 値の過去の値
			パターン③：有意単語 400 語の総登場頻度＋DI 値の過去の値
		目的変数	現状判断 DI の原数値（全国）
	B	分析内容	有意単語 400 語の発生頻度による現状判断 DI の予測(地域別＋全国)
		分析モデル単位	月次
		学習期間	2010 年 1 月～2012 年 12 月
		予測期間	2013 年 1 月～2013 年 12 月
		説明変数	パターン①：有意単語 400 語の総登場頻度
			パターン②：DI 値の過去の値
			パターン③：有意単語 400 語の総登場頻度＋DI 値の

	分析	条件設定	
景気 DI の予測	B (続)		過去の値
		目的変数	現状判断 DI の原数値（地域別＋全国）
			※分析 A との相違点：有意単語の頻度ヒストグラムを地域別にも生成し、見かけのデータ数を増やすことで、オーバーフィッティングを回避する試み。
	C	分析内容	有意表現 400 語の発生頻度による先行き判断 DI の原数値（地域別＋全国）の分析
		分析モデル単位	月次
		学習期間	2010 年 1 月～2012 年 12 月
		予測期間	2013 年 1 月～2013 年 12 月
		説明変数	パターン①：有意単語 400 語の総登場頻度 パターン②：DI 値の過去の値 パターン③：有意単語 400 語の総登場頻度＋DI 値の過去の値
		目的変数	先行き判断 DI の原数値（地域別＋全国） ※分析 B と同じ枠組みで（現状判断ではなく）先行き DI について予測した
	D	分析内容	個人別の現状判断／先行き判断コメントに基づいた DI 値の予測
		学習期間	2010 年 1 月～2012 年 12 月のコメントからランダムで 5000 件抽出
		予測期間	2013 年 1 月～2013 年 12 月のコメント全体
		説明変数	D－1：景気の現状に対する判断コメント各件における有意単語 400 語の登場頻度 D－2：景気の先行きに対する判断コメント各件における有意単語 400 語の登場頻度
		目的変数	各判断コメント(現状/先行き)における 5 値の判定(「良くなった」=100、「やや良くなった」=75、「変わらない」=50、「やや悪くなった」=25、「悪くなった」=0) →これをもとに、各件の予測結果（上記数値換算値）を集計し現状(先行き)判断 DI 値予測値を算出

３－３－１　分析手順

分析Ａ～Ｄについてはいずれも、以下のステップにしたがっている。

① 説明変数、目的変数を定義する。
② データセットを学習期間と予測期間に分割する。
③ 学習データにおける「目的変数と説明変数の間の関係性モデル」を抽出する。
④ 当該モデルを使用した予測の精度の評価と考察。

各分析の詳細を本ステップに従って報告する。

３－３－２　分析Ａ：　有意単語の発生頻度による全国現状判断 DI の予測

３－３－２－１　説明変数、目的変数の定義

前項までの分析で抽出した有意単語 400 語の総登場頻度を説明変数とし、現状判断 DI の原数値（全国）を目的変数とする。これを分析Ａと呼ぶことにする。

また比較実験として、説明変数に当該 DI 値の過去 12 か月分の値、前月差分、3 か月平均を用いた分析、さらにそれらに上記の有意単語の頻度をあわせたものを説明変数として用いた分析も実施する。以下、表記の簡単のために、上記の有意単語を説明変数とする分析をパターン①、DI 値の過去の値等を説明変数とする分析をパターン②、合わせたものを説明変数とする分析をパターン③と記すこととする。

３－３－２－２　学習期間、予測期間の定義

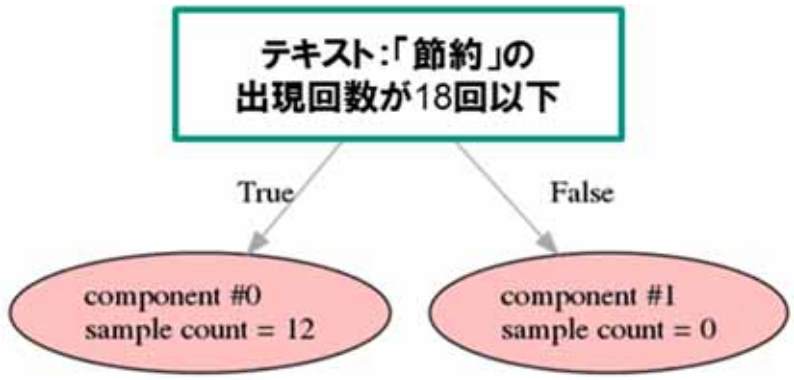
学習期間は 2010 年 1 月～2012 年 12 月とし、予測期間は 2013 年 1 月～2013 年 12 月としている。

３－３－２－３　「目的変数と説明変数の間の関係性モデル」の抽出

図 3－3－1 は、説明変数パターン①の分析Ａにおいて異種混合学習により抽出された、説明変数と目的変数の間の関係をグラフ化したものである。本分析においては、関係式が二つ（0 番と 1 番）が抽出され、各関係式がどのような条件のときに成立しているかを表したものが同図の上段、それぞれの関係式の係数を表したものが同図の下段である。以下、本分析の目的を鑑み、関係式を特に予測式と表記する場合もあるとする。上段においては、「節約」という単語の頻度が 18 回以下か否かで、0 番か 1 番の予測式が成立していることを示している。下段において、左に並んでいるのが有意単語であり、青いバーが 0 番目の予測式において当該単語の頻度に乗ずるべき係数（ただし 1 行目の bias は、当該関係式における定数項）、赤いバーが 1 番目の予測式におけるそれの大き

さをあらわしている。予測式 0 においては「激減」という単語の頻度が相対的に大きな負の係数を持ち、予測式 1 においては「陥る」という単語のそれが同様であることが見て取れる。

抽出されたこれらの関係性、あるいはその切り替え条件ともに、目的変数を予測するに直観的なものになっているとは言えず、オーバーフィッティング[1]の可能性を強く示唆するものとなっている。



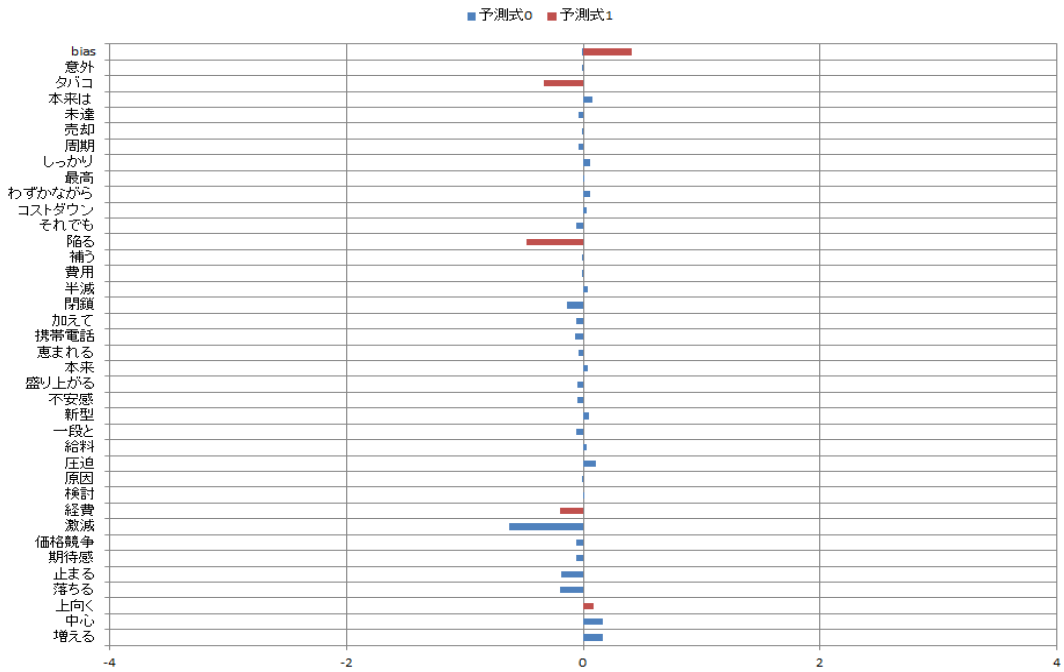
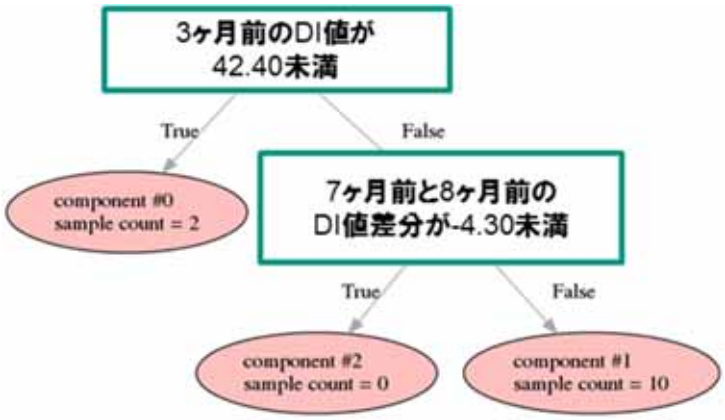


図 3－3－1　パターン①：テキスト特徴量のみ　（分析A）

[1] モデルの複雑さに対し学習データサンプルが少なすぎる場合などに生じる現象。学習期間にたまたま現れた説明変数の値を用いて目的変数を満たしてしまうが、予測期間においては精度が大きく低下する。

図 3－3－2 は、説明変数パターン②の分析Ａにおける、異種混合学習によって抽出された関係性をグラフ化したものである。上段の図の行名は説明変数の種別を表している（説明変数として使用した何か月前の DI 値、あるいはそれらの前月差分、あるいは平均を意味している）。ここでは 3 つの予測式が抽出されており、予測式 1 において 1 か月前の DI の値、および 5 か月前と 6 か月前の差分が正の係数を持つ以外は、すべての予測式のすべての係数が負の値を持っていることがわかる。



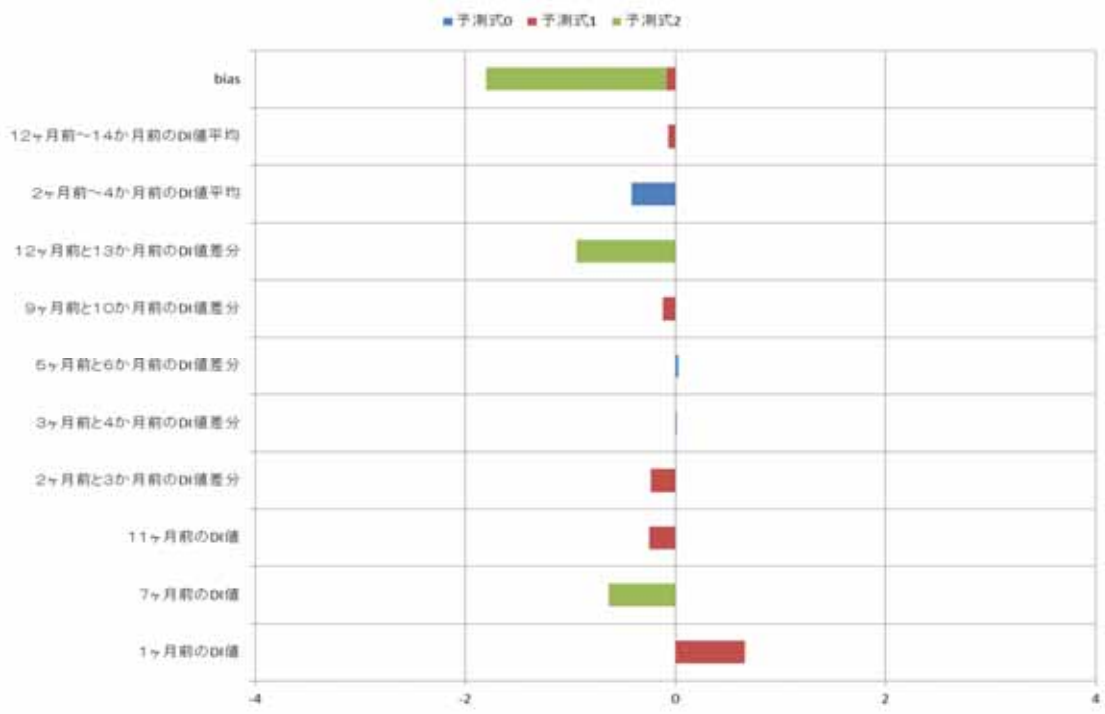
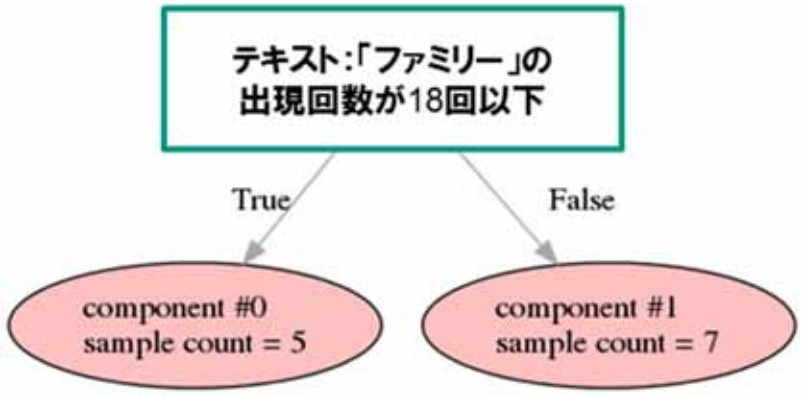


図 3－3－2　パターン②　目的変数関連のみ　（分析Ａ）

図 3－3－3 はパターン③の分析Ａに対する抽出結果である。説明変数はパターン①の有意単語 400 語の頻度ヒストグラムと、パターン②の現状判断 DI の 12 か月分の「値」、「前月差分」、「3 か月平均」の両方を用いることとなる。抽出された結果は、やはり直観にあうものとは言えず、オーバーフィッティングを示唆するものとなっている。



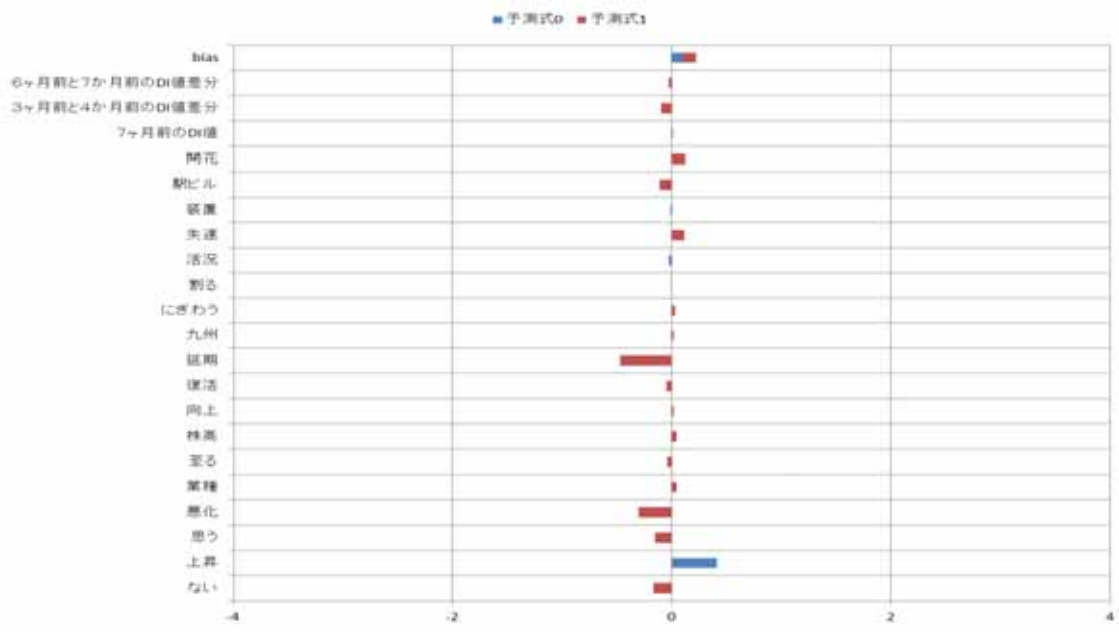


図 3－3－3　パターン③テキスト特徴量＋目的変数関連　（分析Ａ）

３－３－２－４　**予測精度の評価と考察**

図 3－3－4 及び図 3－3－5 は、抽出された予測式を使って行った DI の予測値と、実際の DI 値を比較したグラフである。どちらも黒い太線が実際の値、赤線が説明変数パターン①、緑線が説明変数パターン②、青線が説明変数パターン③を示しており、図 3－3－4 は学習期間における様子、図 3－3－5 は予測期間における様子である。また、表 3－3－2 はそれらの精度をまとめたものである。図と表の両方から読み取れる通り、判断理由欄を説明変数に用いたパターン①と③は、学習期間においては、非常に高い精度を示している（誤差 0%）。一方、予測期間においては、大まかな増減には追従している様子が図から見て取れるが、精度としては絶対平均誤差が 1.5 ないしは 2.6 程度と必ずしもよいとは言えない。この学習期間と予測期間の精度の差は、オーバーフィッティングを起こしていることを強く示唆している。なお、比較実験として行っている DI の過去値関係のみを用いたパターン②の結果と比較すると、テキストデータを用いた予測の精度は良くなってはいるが、予測期間のデータ数が 12 と非常に少ないために、特段の結論を強く示唆するものではない。

抽出された関係性が直観的ではないこと、学習期間と予測期間の精度が大きくかい離していることなどから、学習データの数が過小であることによるオーバーフィッティングが生じていることが、ほぼ確実であると考えられる。

一方、精度は高くないにせよ、予測期間において目的変数の予測値が実績値に一定の追随を見せている。これは、判断理由欄を適切に処理することにより、一定の説明力を持つ説明変数を構築できることを意味している。サンプル数を大幅に増やすことなどにより、オーバーフィッティングを解消できれば、予測期間においても高い精度を達成できる可能性はあることになる。

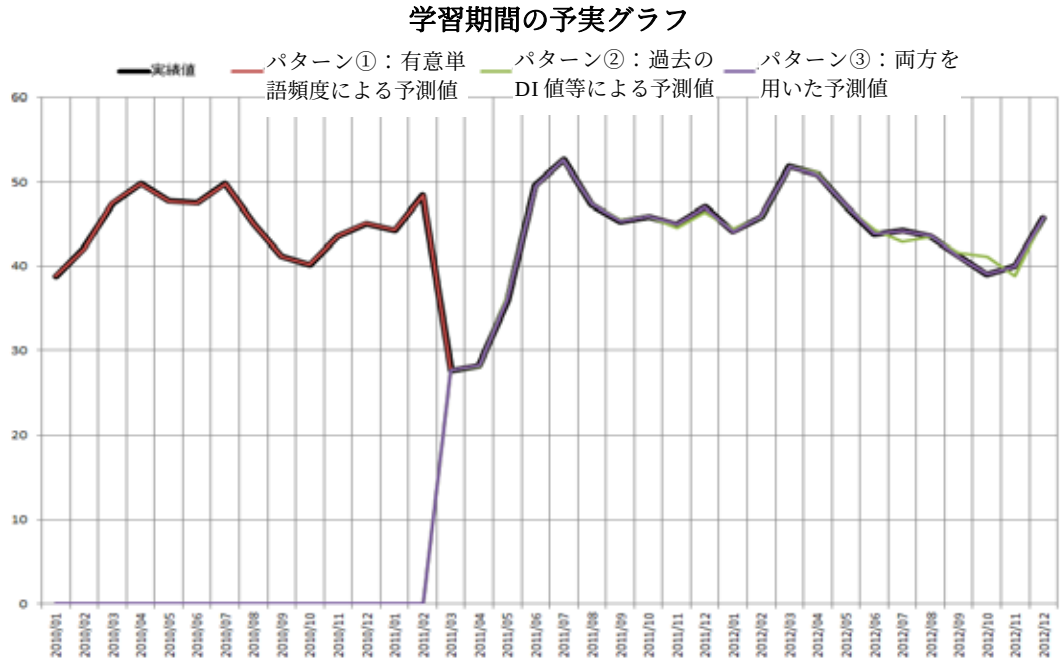


図 3－3－4　学習期間における実績値と予測値の比較グラフ　（分析Ａ）

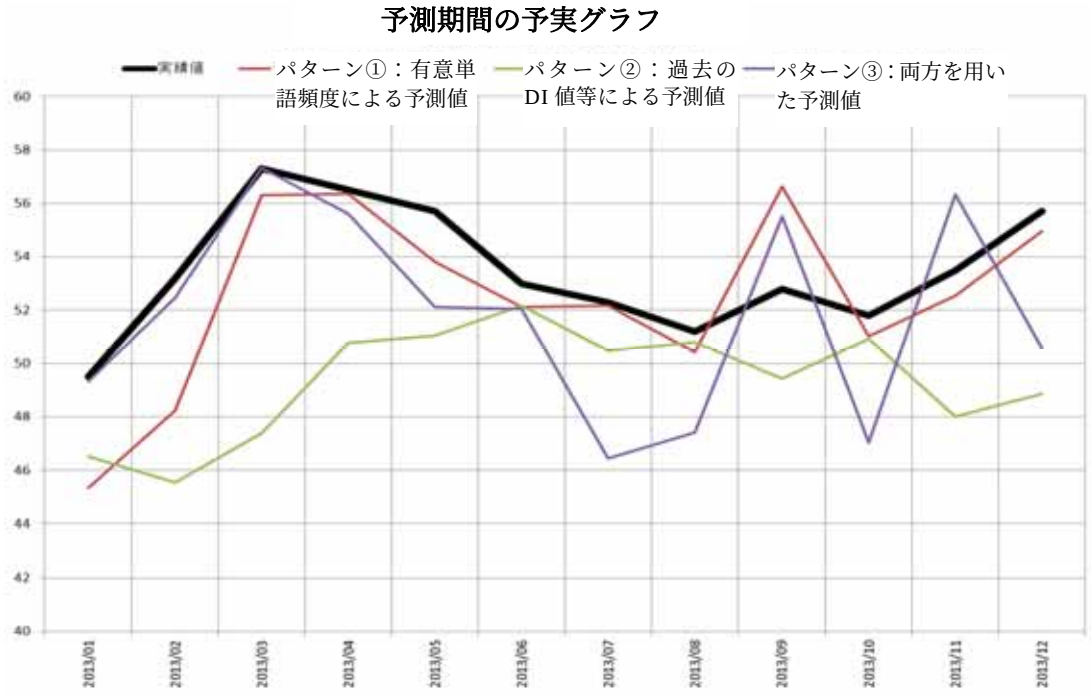


図 3－3－5　予測期間における実績値と予測値の比較グラフ　（分析Ａ）

表 3－3－2　精度（分析Ａ）

説明変数パターン	学習区間 平均絶対誤差	予測区間 平均絶対誤差
パターン①テキスト特徴量のみ	0.00	1.68
パターン②目的変数関連のみ	0.36	4.21
パターン③テキスト＋目的変数関連	0.00	2.62

説明変数パターン	学習区間 誤差率	予測区間 誤差率
パターン①テキスト特徴量のみ	0.0%	3.1%
パターン②目的変数関連のみ	0.8%	7.9%
パターン③テキスト＋目的変数関連	0.0%	4.9%

３－３－３　分析Ｂ：　有意単語の発生頻度による地域別現状判断 DI の予測

分析Ａにおいてオーバーフィッティングの可能性が強く示唆されたことを踏まえ、有意単語の頻度ヒストグラムを地域別にも生成し、見かけのデータ数を増やすことで、オーバーフィッティングを回避するトライアルを行う。特にここでは、元データで定義されていた 11 地域に加え、「全国」も一つの地域とみなし、それぞれが別の学習サンプルであると定義することで、データサンプル数を分析Ａの 12 倍に増やす。各地域における説明変数と目的変数の間の関係性が同一なのであれば、分析Ａに比してオーバーフィッティングが緩和されることが期待できる[2]。

３－３－３－１　説明変数、目的変数の定義

前章までの分析で抽出した有意単語 400 語の地域別登場頻度、および地域フラグ（どの地域に関する頻度かを表す変数）を説明変数とし、現状判断 DI の当該地域の原数値を目的変数とする。これを分析Ｂと呼ぶことにする。

比較実験のための説明変数のバリエーションは分析Ａと同じとする。すなわち、上記の有意単語と地域フラグを説明変数とする分析をパターン①、当該地域の DI 値の過去の値等を説明変数とする分析をパターン②、①と②を合わせたものを説明変数とする分析をパターン③と記すこととする。

３－３－３－２　学習期間、予測期間の定義

分析Ａと同じく、学習期間は 2010 年 1 月～2012 年 12 月とし、予測期間は 2013 年 1 月～2013 年 12 月とする。

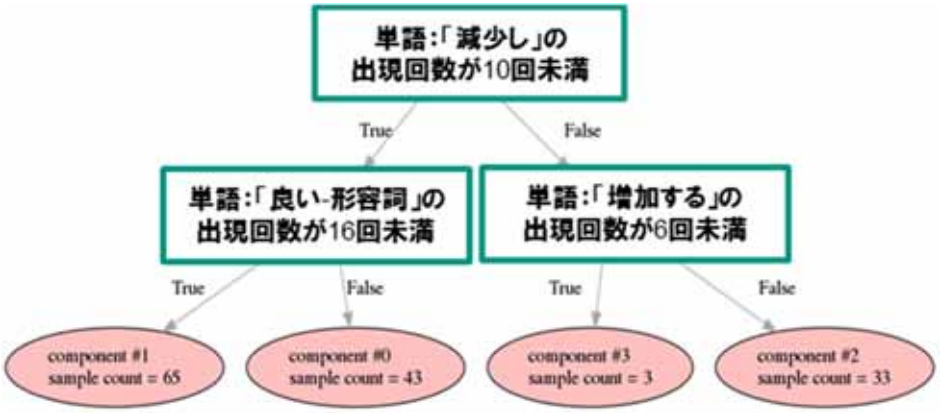
３－３－３－３　説明変数と目的変数の関係性モデルの抽出

図 3－3－6～図 3－3－8 は、分析Ｂにおいて抽出された関係性である（それぞれ説明変数パターン①～③）。残念ながら、抽出された関係性の直観的解釈性が向上したとは言えず、オーバーフィッティングの解消はみてとれなかった。

３－３－３－４　予測精度の評価と考察

図 3－3－9 および図 3－3－10 は「全国」の DI 値について予測結果と実際の値を比較したグラフである（各地域の結果グラフは資料 8 に掲載）。学習期間と予測期間における精度の乖離も改善しておらず、分析Ｂにおいて、オーバーフィッティングが特段の改善を見せたとはいえなかった。

[2] 分析Ｂ及び分析Ｃ、分析Ｄ（いずれも後述）においてはテキストの特徴量についても精度改善を目的とした若干の変更を加えている。



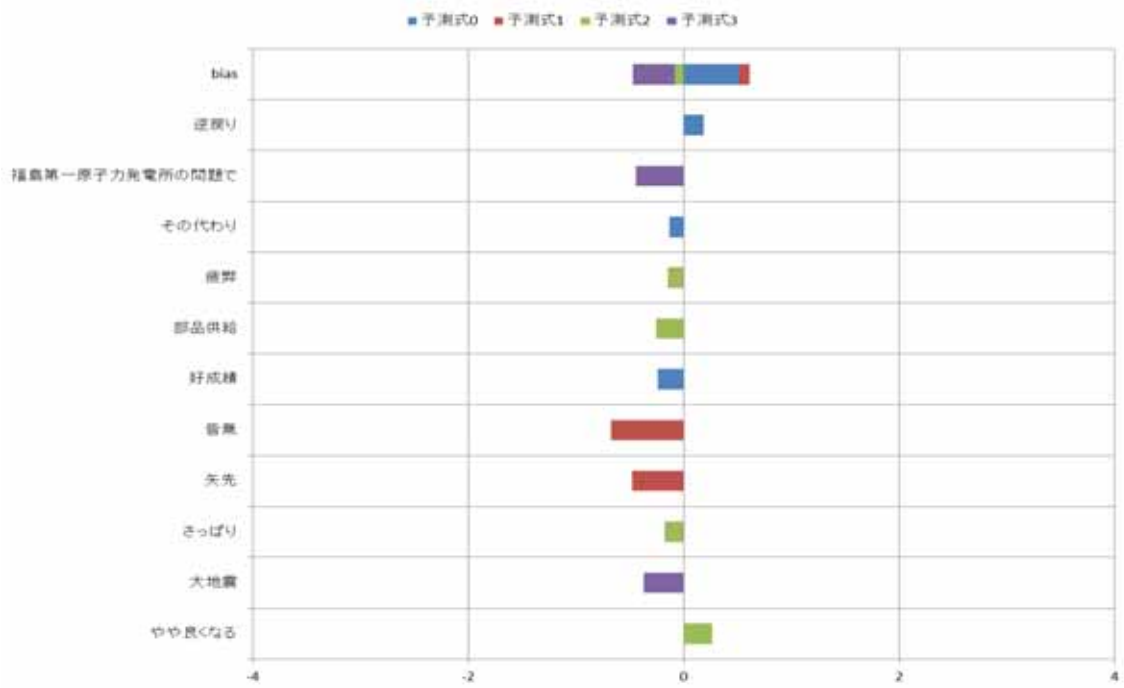
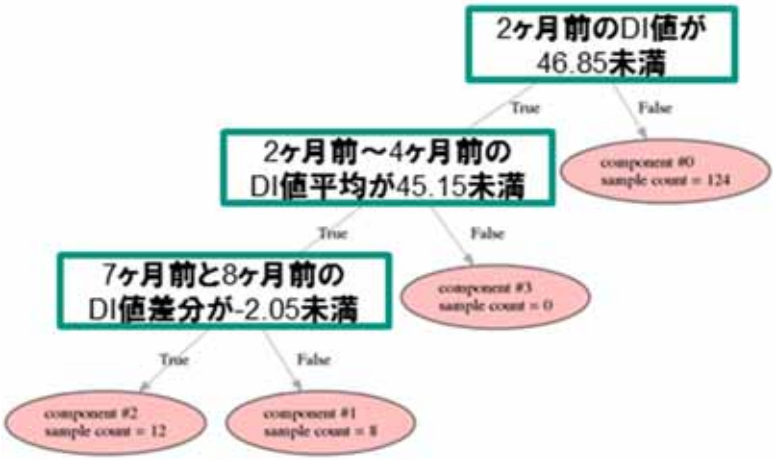


図 3－3－6　パターン①：テキスト特徴量のみ（分析Ｂ：全国）



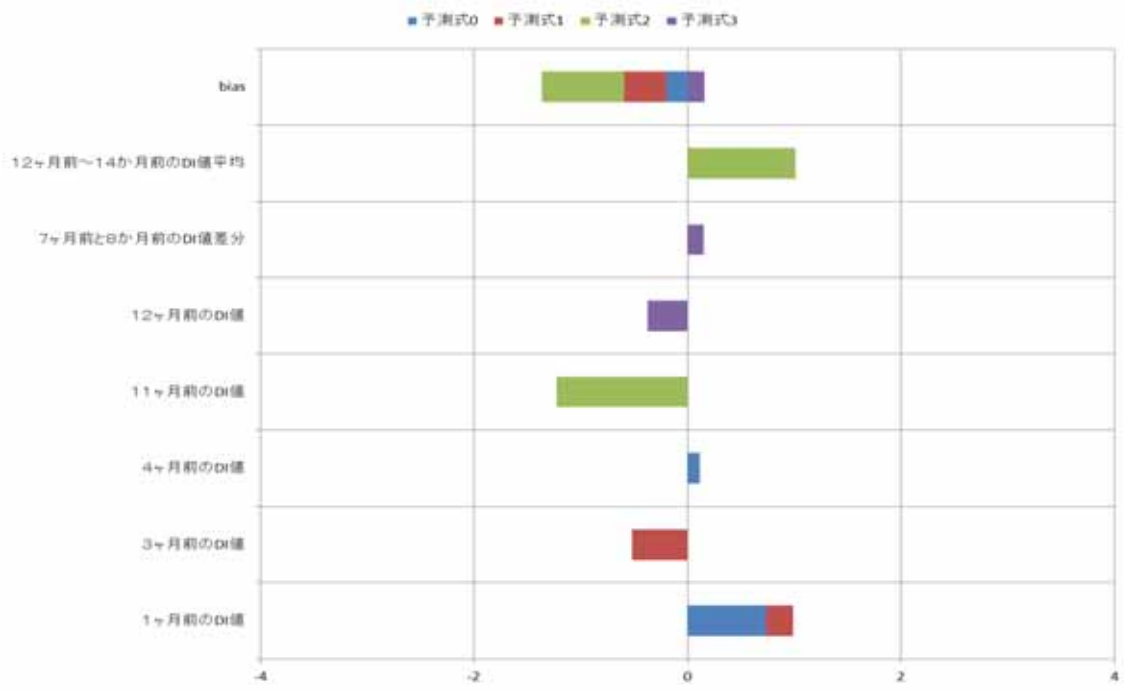
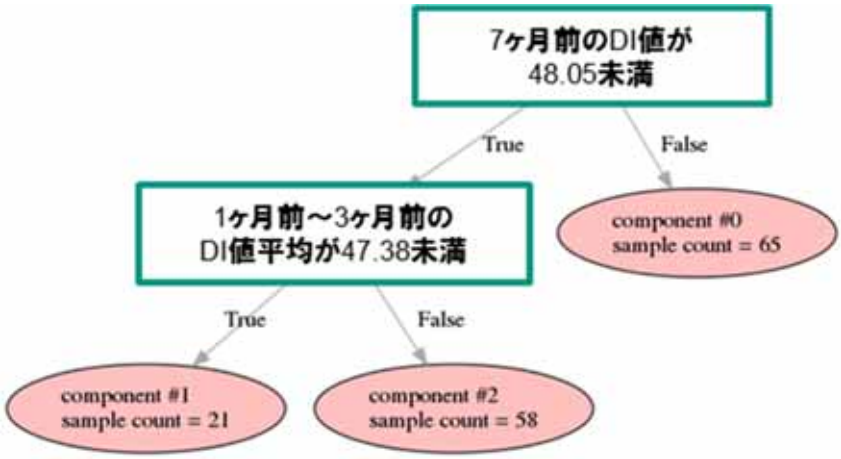


図 3－3－7　パターン②　目的変数関連のみ　（分析Ｂ：全国）



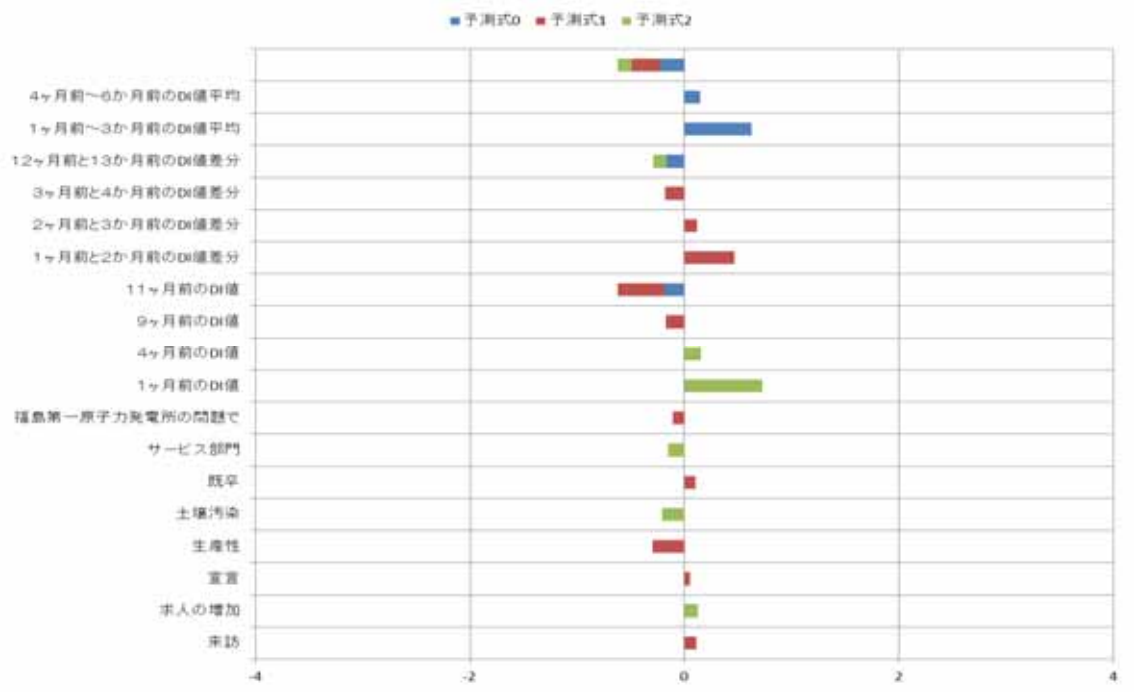


図 3－3－8　パターン③テキスト特徴量＋目的変数関連　（分析Ｂ：全国）

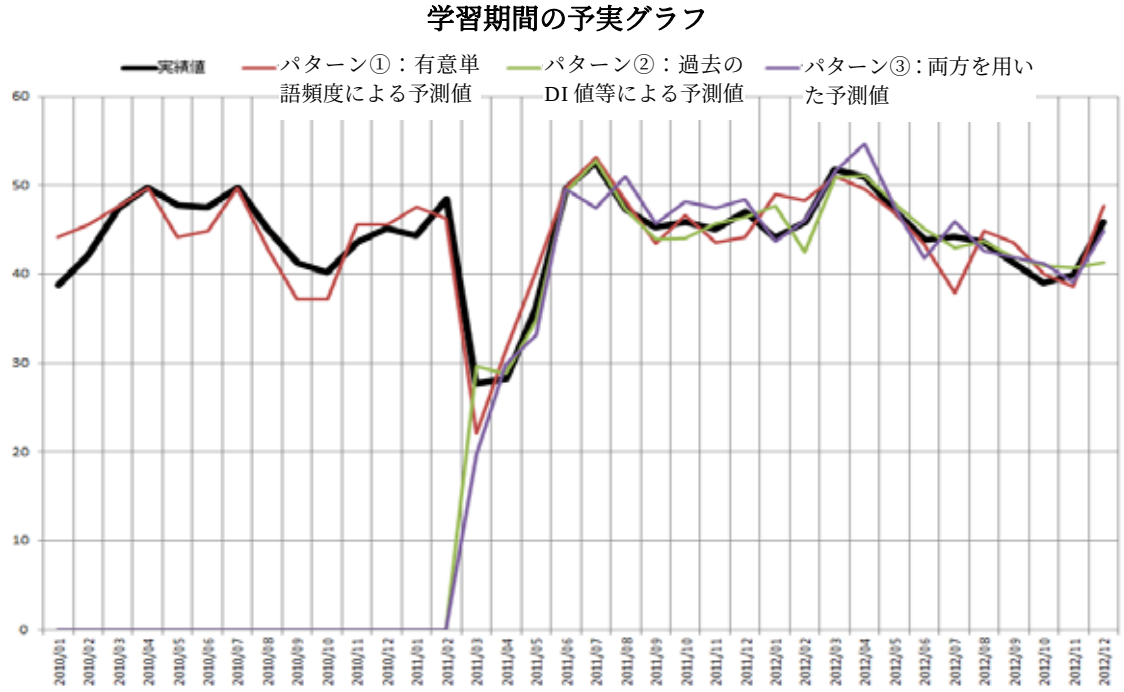


図 3－3－9　学習期間における実績値と予測値の比較グラフ　（分析Ｂ：全国）

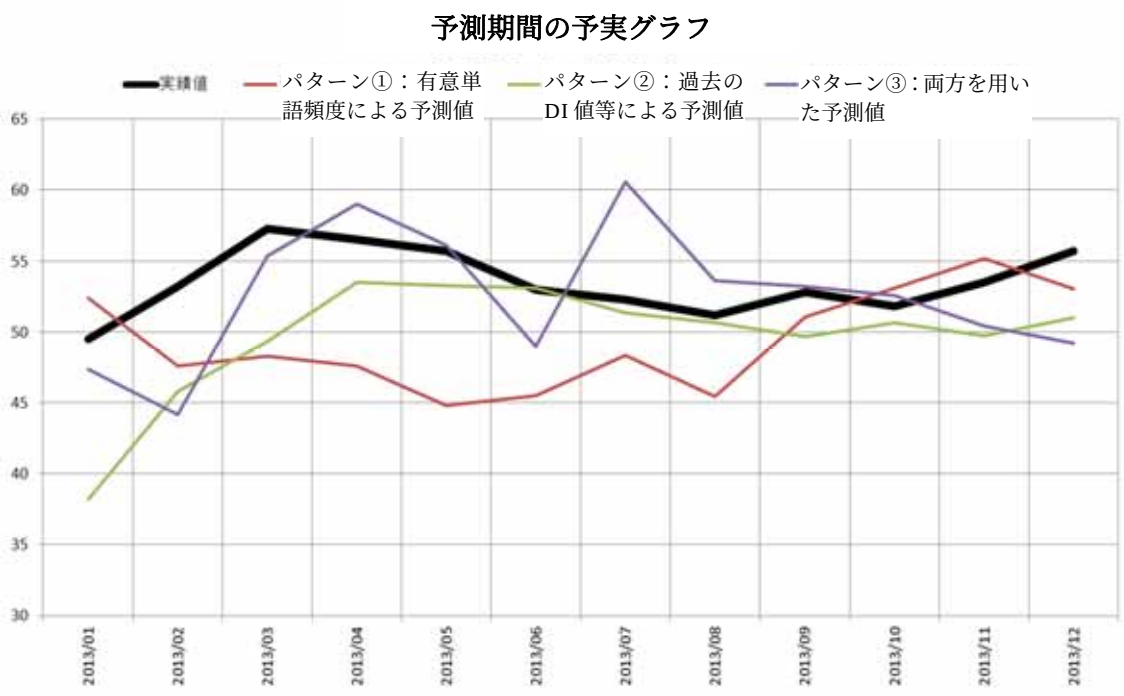


図 3－3－10　予測期間における実績値と予測値の比較グラフ　（分析Ｂ：全国）

表 3－3－3　精度（分析Ｂ）

説明変数パターン	学習区間 平均絶対誤差	予測区間 平均絶対誤差
パターン①テキスト特徴量のみ	3.01	9.72
パターン②目的変数関連のみ	1.28	3.87
パターン③テキスト＋目的変数関連	1.21	4.41

説明変数パターン	学習区間 誤差率	予測区間 誤差率
パターン①テキスト特徴量のみ	6.8%	18.2%
パターン②目的変数関連のみ	2.9%	7.2%
パターン③テキスト＋目的変数関連	2.7%	8.2%

３－３－４　分析Ｃ：先行き判断 DI の原数値（地域別＋全国）の分析

これまでは現状判断 DI の原数値を目的変数とする分析について説明してきたが、以降は先行き判断 DI の原数値(地域別＋全国)を目的変数とした分析について説明する。この分析については、以下、分析Ｃと記す。

３－３－４－１　説明変数・目的変数の定義

分析の手順については、現状判断 DI を目的変数とした分析Ｂとほぼ同じであり、前章までの分析で抽出した有意単語 400 語の登場頻度を説明変数とし、先行き判断 DI の原数値（地域別＋全国）を目的変数とする。これを分析Ｃと呼ぶこととする。分析Ａ・Ｂと同様に、説明変数の三つのバリエーション①～③に関して比較実験を行う。

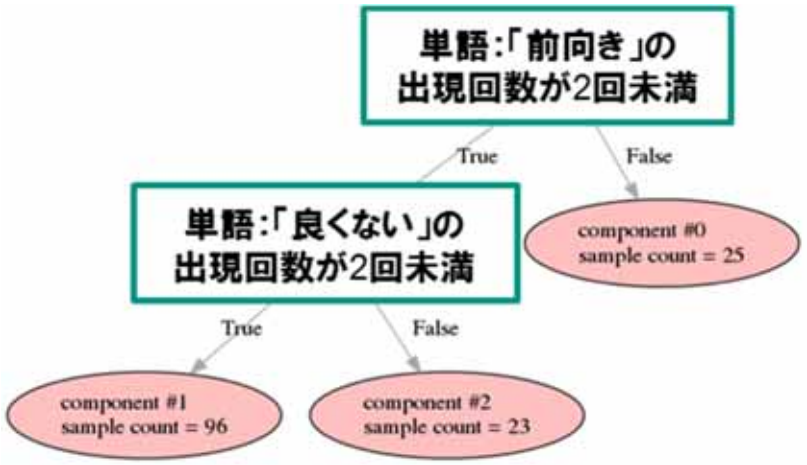
３－３－４－２　学習期間と予測期間の定義

学習期間は 2010 年 1 月～2012 年 12 月とし、予測期間は 2013 年 1 月～2013 年 12 月としてする。

３－３－４－３　関係性モデルの抽出、予測精度の評価と考察

図 3－3－11～図 3－3－13 は、先行き判断 DI を目的変数とする分析で抽出された関係性､図 3－3－14 および図 3－3－15 は予測結果と実際の値を比較したグラフである。

パターン①においては、「立たない」「真っ暗」「地震の影響」などの言葉が大きな負の係数を持ち、直観的解釈性は若干向上しているようにみえるが、学習期間と予測期間の精度の乖離が相変わらず非常に大きく、オーバーフィッティングの発生を強く示唆している。パターン③はパターン①に対して、予測期間の精度が若干改善しているが、予測式において強く効いている係数は、主に DI の過去値由来の説明変数にあり、テキストデータを用いた予測の可能性については、強い意味を示す結果とは言えない。



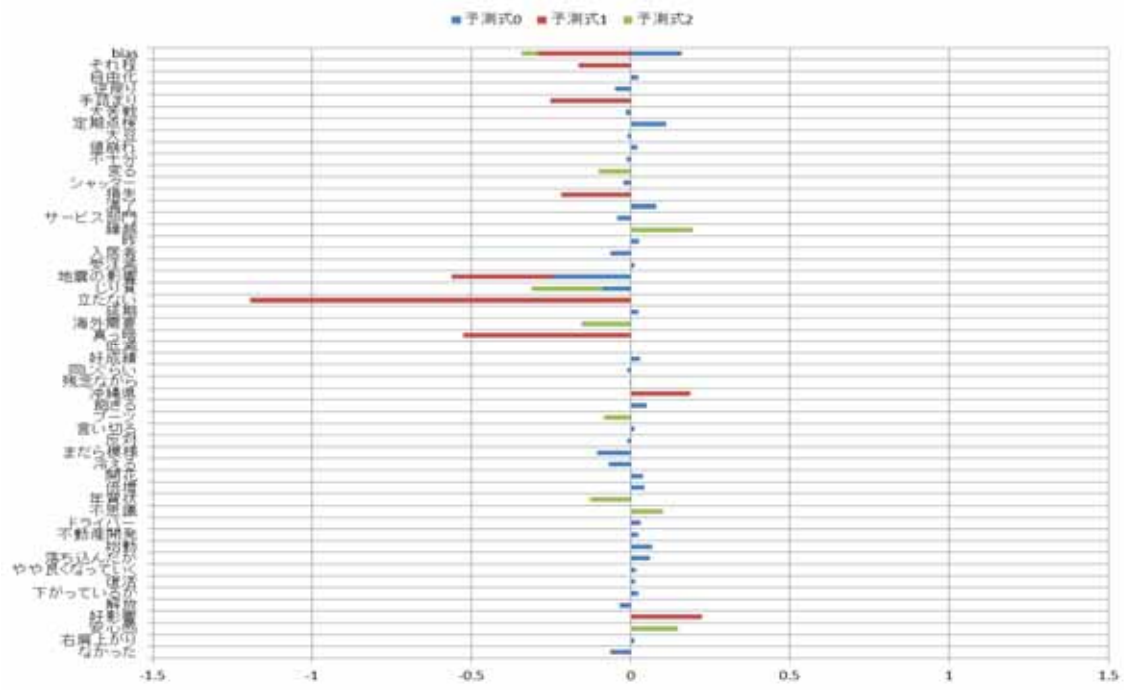
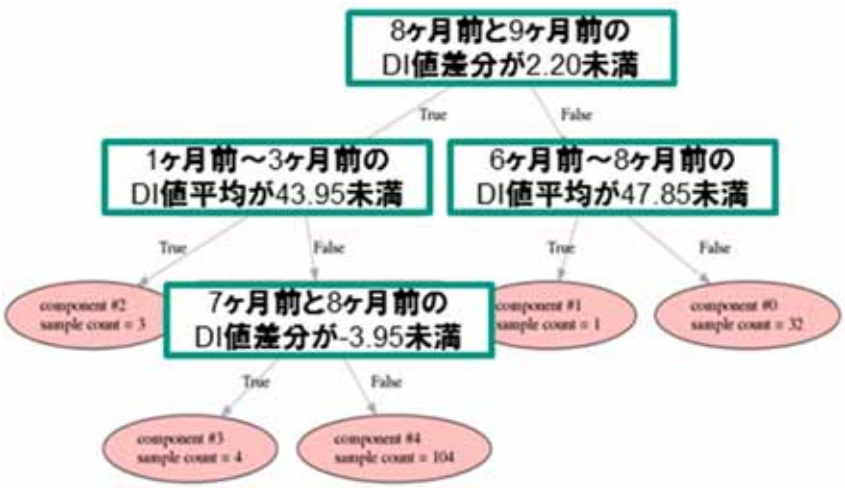


図 3－3－11　パターン①：テキスト特徴量のみ　（分析Ｃ：全国）



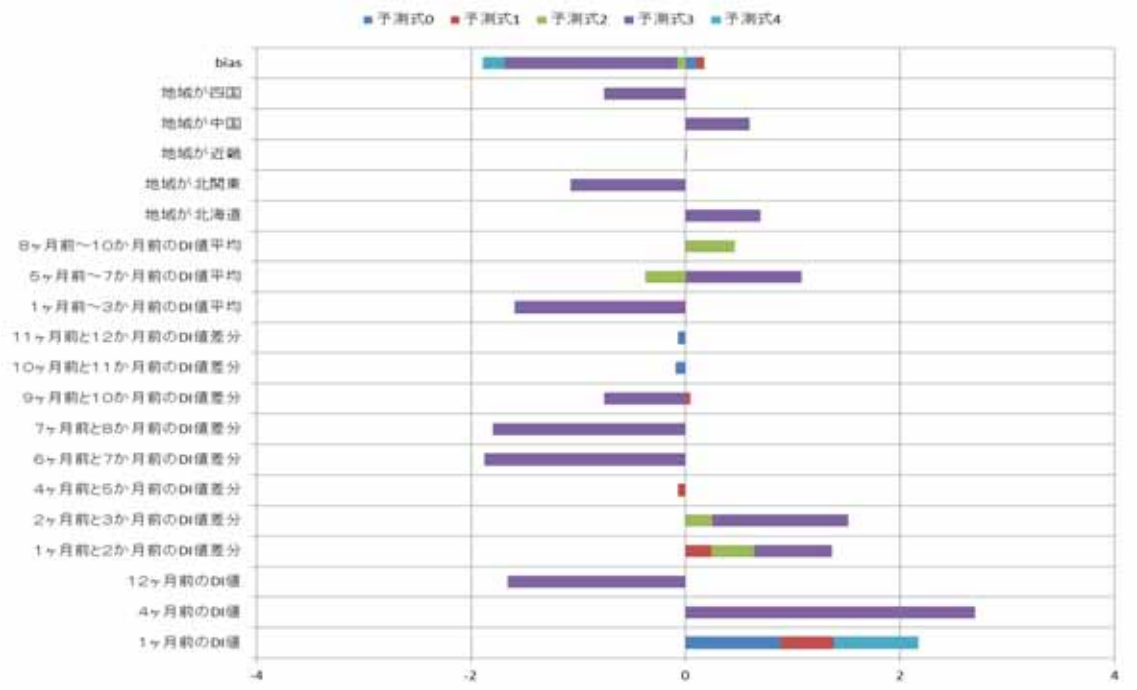
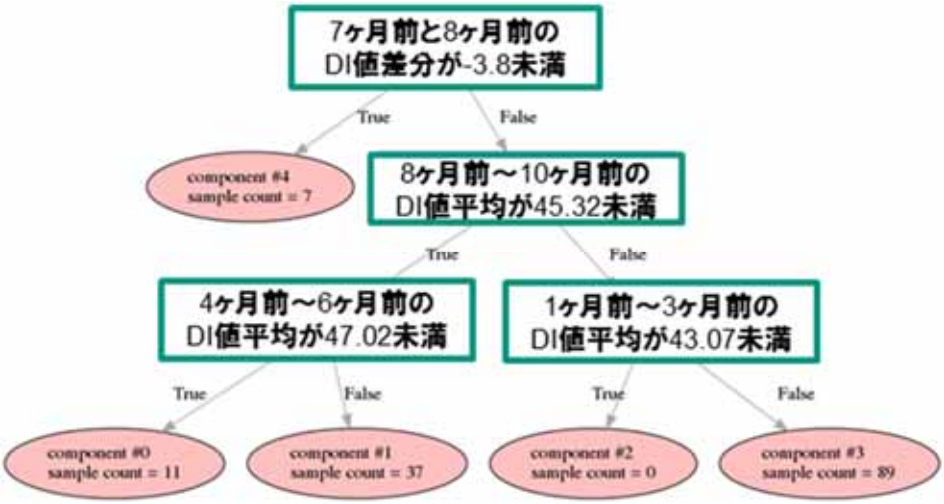


図 3－3－12　パターン②　目的変数関連のみ　（分析Ｃ：全国）



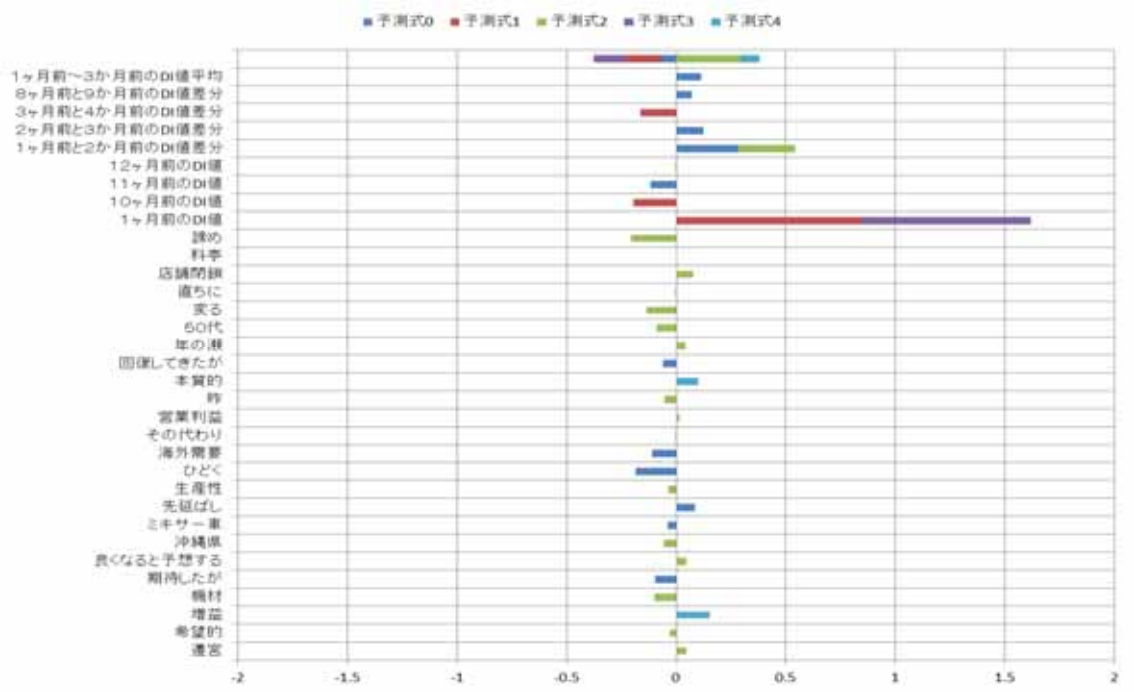


図 3－3－13　パターン③テキスト特徴量＋目的変数関連　（分析Ｃ：全国）

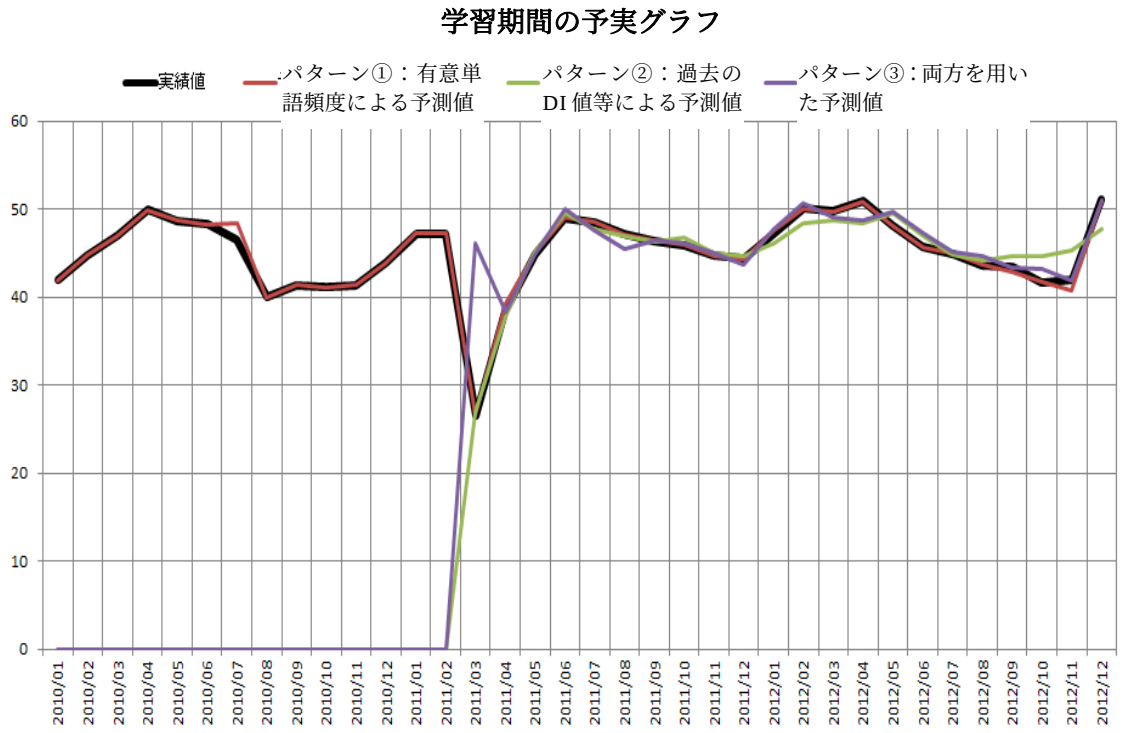


図 3－3－14　学習期間における実績値と予測値の比較グラフ　（分析Ｃ：全国）

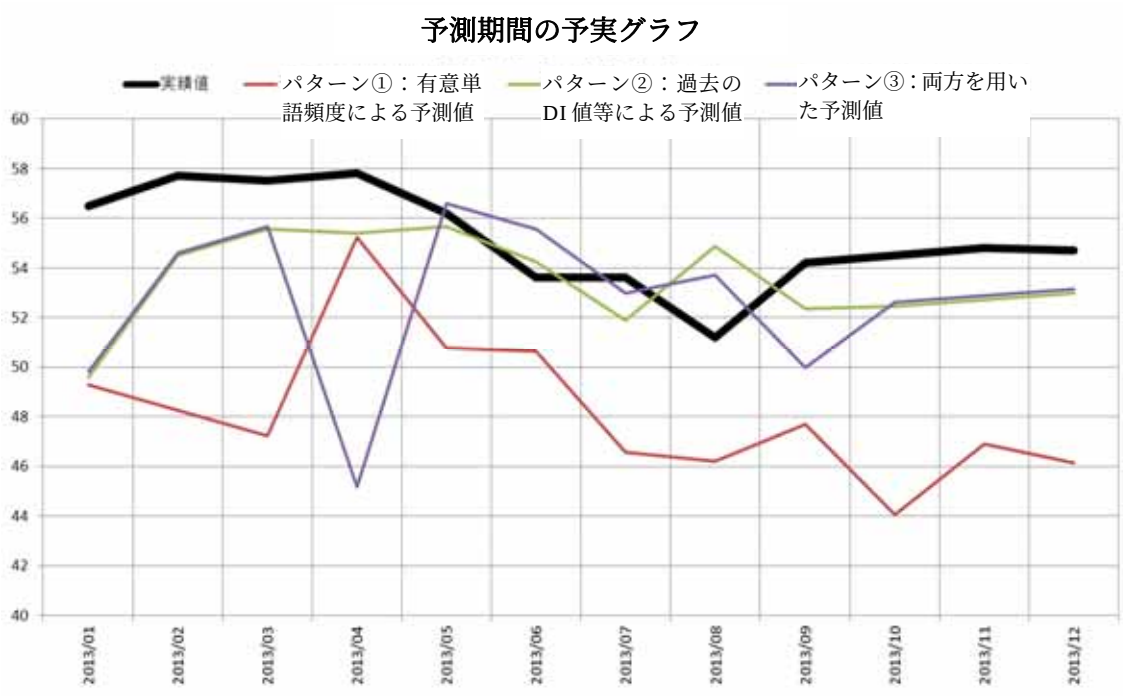


図 3－3－15　予測期間における実績値と予測値の比較グラフ　（分析Ｃ：全国）

表 3－3－4　精度（分析Ｃ）

説明変数パターン	学習区間 平均絶対誤差	予測区間 平均絶対誤差
パターン①テキスト特徴量のみ	0.13	6.94
パターン②目的変数関連のみ	1.10	2.39
パターン③テキスト＋目的変数関連	1.58	3.28

説明変数パターン	学習区間 誤差率	予測区間 誤差率
パターン①テキスト特徴量のみ	0.3%	12.6%
パターン②目的変数関連のみ	2.4%	4.3%
パターン③テキスト＋目的変数関連	3.5%	5.9%

３－３－５　分析Ｄ：個人別現状判断／先行き判断の予測によるＤＩ値の予測

前述の分析Ａにおいては、特定年月のコメントにおける有意単語の頻度を説明変数として、同期の現状判断 DI の値について大まかな増減は予測できる一方、サンプル数過小に起因したオーバーフィッティングが発生していることが強く示唆され、必ずしも予測精度は高くない結果が得られていた。見かけのサンプル数を増やした分析Ｂでも、同問題は解消しなかった。また分析Ｃにおいて、先行き DI の値の予測可能性についても検討を行ったが、分析Ａに比して特段の前向きな結果は得られなかった。

しかしながら、分析Ａでテキストデータから生成した情報の説明力に一定の確認が得られたため、別の方法でのオーバーフィッティング解消の試みとして、個人別コメントを 1 サンプルとして DI の値を予測する分析を試行する。すなわち、個人別コメントにそれぞれ付与されている、個人別の景気判断を目的変数として予測器を構築し、その予測結果の平均を用いて DI の値の予測値とする分析Ｄである。以下に詳細を述べる。

３－３－５－１　説明変数と目的変数の定義

説明変数は各判断理由欄における有意単語の登場回数、および該当する地域コードとする。目的関数は当該判断理由欄に付与されていた各アンケート回答者の現状判断である「良くなった」、「やや良くなった」、「変わらない」、「やや悪くなった」、「悪くなった」を、100 点、75 点、50 点、25 点、0 点という数値へ変換した「現状判断値」とする。

なお本分析は、「景気の現状に対する判断理由等」と対象とした分析Ｄ－１と「景気の先行きに対する判断理由」を対象とした分析Ｄ－２から成る。

３－３－５－２　学習データと予測データ

それぞれ以下のとおりとする。

分析Ｄ－１

学習データ：2010/1～2012/12 のアンケートからランダムで 5000 件を抽出

予測データ：2013/1～2013/12 のアンケート全件

分析Ｄ－２

学習データ：2010/1～2012/12 のアンケートからランダムで 5000 件を抽出

予測データ：2013/1～2013/12 のアンケート全件

なお、学習データに関して 5,000 件を抽出した理由は、異種混合学習エンジンの実行時間を抑え本分析業務の作業期間内で結果を得るためである。

（将来的には、件数を増加させて分析を行うことで結果が改善される可能性はある。）また、分析におけるサンプルバランスの調整を行うことでよりモデルを適正なものにで

きる可能性はあるが今回は行わなかった。これは将来の課題である。

表 3－3－5　分析Ｄ－１（現状判断分析）のデータ数

景気判断		学習			予測	
景気判断	数値化	分析対象サンプル数※ランダム抽出	分析対象サンプル数	非対象サンプル数※テキスト特徴量なし	分析対象サンプル数	非対象サンプル数※テキスト特徴量なし
良くなった	100	62	394	348	244	200
やや良くなった	75	933	5564	3243	2948	1965
変わらない	50	2266	12953	7146	4975	2462
やや悪くなった	25	1253	7325	4114	1551	774
悪くなった	0	486	2783	1788	309	201
計		5000	29019	16639	10027	5602

表 3－3－6　分析Ｄ－２（先行き判断分析）のデータ数

景気判断		学習			予測	
景気判断	数値化	分析対象サンプル数※ランダム抽出	分析対象サンプル数	非対象サンプル数※テキスト特徴量なし	分析対象サンプル数	非対象サンプル数※テキスト特徴量なし
良くなった	100	55	315	328	289	306
やや良くなった	75	973	5422	3579	3281	2787
変わらない	50	2516	13853	8695	4901	3435
やや悪くなった	25	1116	6436	4426	1270	982
悪くなった	0	340	2150	1847	242	233
計		5000	28176	18875	9983	7743

※分析を行うコメントのうち、テキスト特徴量があるもののみを分析対象サンプルとし、テキスト特徴量なしのコメントは非対象とした。

※学習データについては、分析対象サンプルからランダムに抽出した 5000 件のみを使用。

３－３－５－３　**関係性モデルの抽出**

また、図 3－3－16 および図 3－3－17 に、分析Ｄ－１とＤ－２の説明変数の影響度合い（回帰係数）を示す。なお、異種混合学習の特徴である複数の予測式の生成は、行われなかった。ただし、回帰係数は、絶対値が一定以上大きいもののみを抽出している。

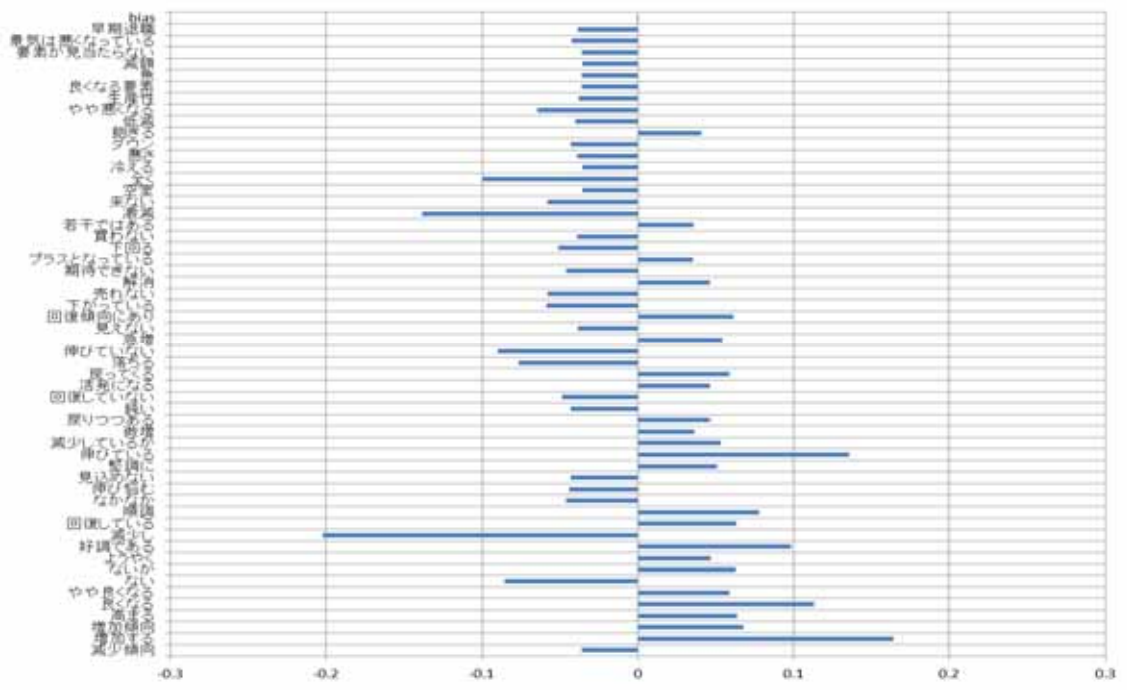


図 3－3－16　分析Ｄ－１の回帰係数

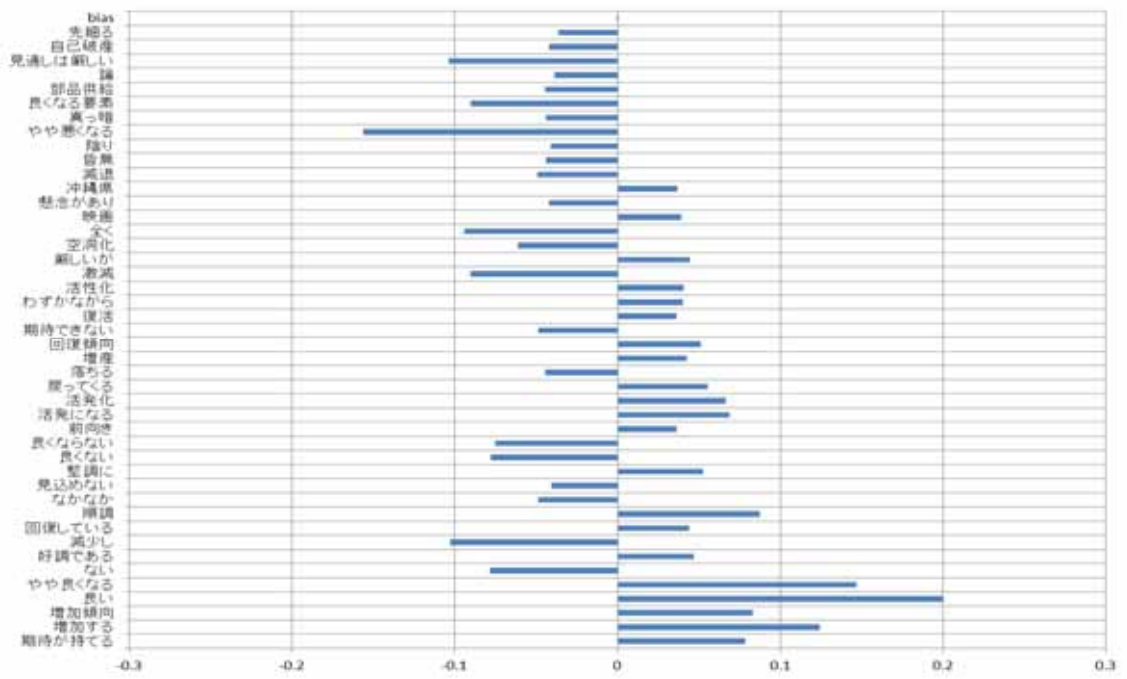


図 3－3－17　分析Ｄ－２の回帰係数

３－３－５－４　**予測精度の評価と考察**

表 3－3－7 は、各個人の学習データおよび予測データにおける精度である。

表 3－3－7　分析Ｄ－１およびＤ－２の精度

	学習データ平均絶対誤差	予測データ平均絶対誤差
分析Ｄ－１：現状判断	15.52	15.39
分析Ｄ－２：先行き判断	13.92	14.88

	学習データ平均誤差率	予測データ平均誤差率
分析Ｄ－１：現状判断	35.2%	28.9%
分析Ｄ－２：先行き判断	30.0%	26.9%

図 3－3－18 は、分析Ｄ－１の結果について、予測値・実績値それぞれについて、各月ごとに全国の全個人の結果を平均して算出し、グラフ化したものである。比較として、全国の DI 値も記している。

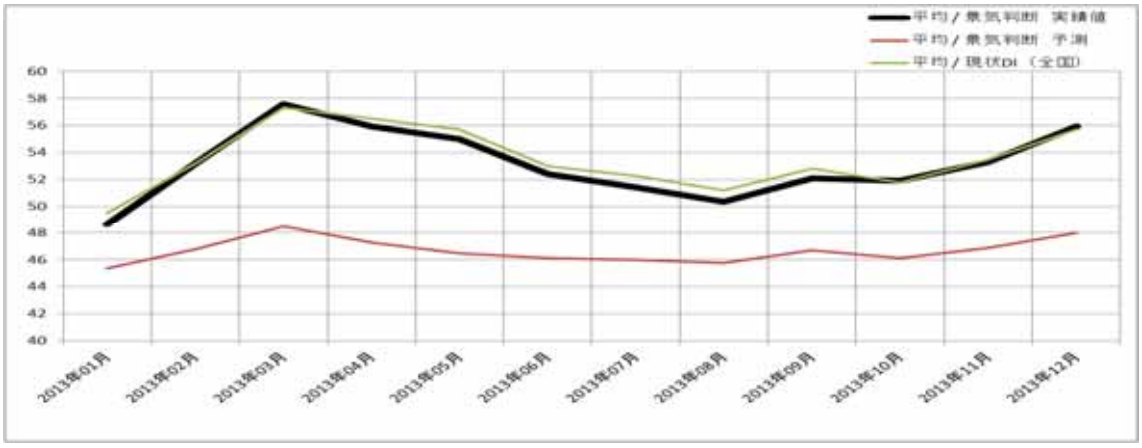


図 3－3－18　分析結果（分析Ｄ－１）に関する全国平均値

（注）上段グラフの凡例（タテ軸は DI 値）

「平均／景気判断　実績値」：内閣府が公表している当該地域の景気の現状判断理由・追加コメントのそれぞれに付随する 5 値判断（「良くなっている」「やや良くなっている」「変わらない」「やや悪くなっている」「悪くなっている」）をサ

ンプル対象についてそのまま集計し、DI 値に換算したもの

「平均／景気判断 予測」：景気の現状判断理由・追加コメントのテキスト特徴量を説明変数、当該コメントそれぞれに付随する 5 値判断を数値化したものを目的変数として分析した結果をサンプル対象について集計、平均したもの

「現状 DI」：内閣府が公表している当該地域の現状判断 DI 値

また、図 3－3－19 は、分析Ｄ－１の結果について、予測値・実績値のそれぞれの全国の個人の平均値を算出した結果について、前月との差分を算出してグラフ化したものである

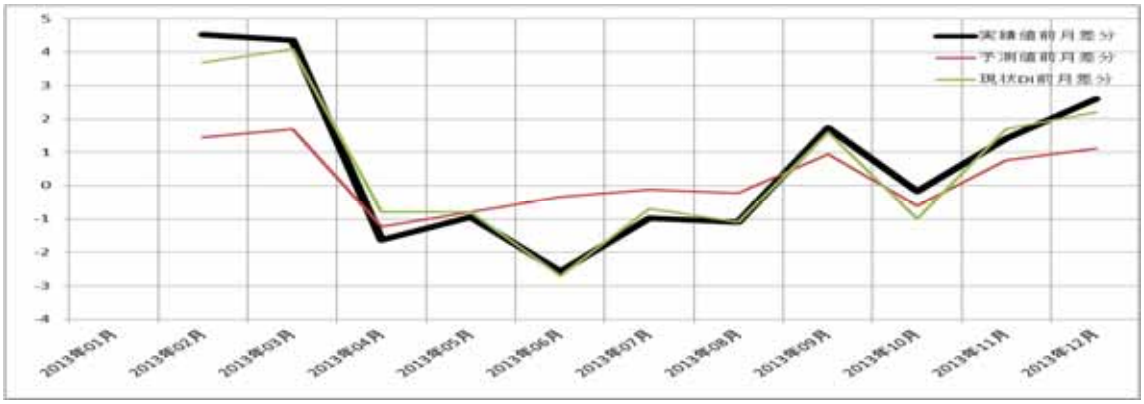


図 3－3－19　分析結果（分析Ｄ－１）に関する全国平均値の前月差分

同様に、図 3－3－20 および図 3－3－21 に分析Ｄ－２に関する全国平均値および全国平均値の前月差分を示す。

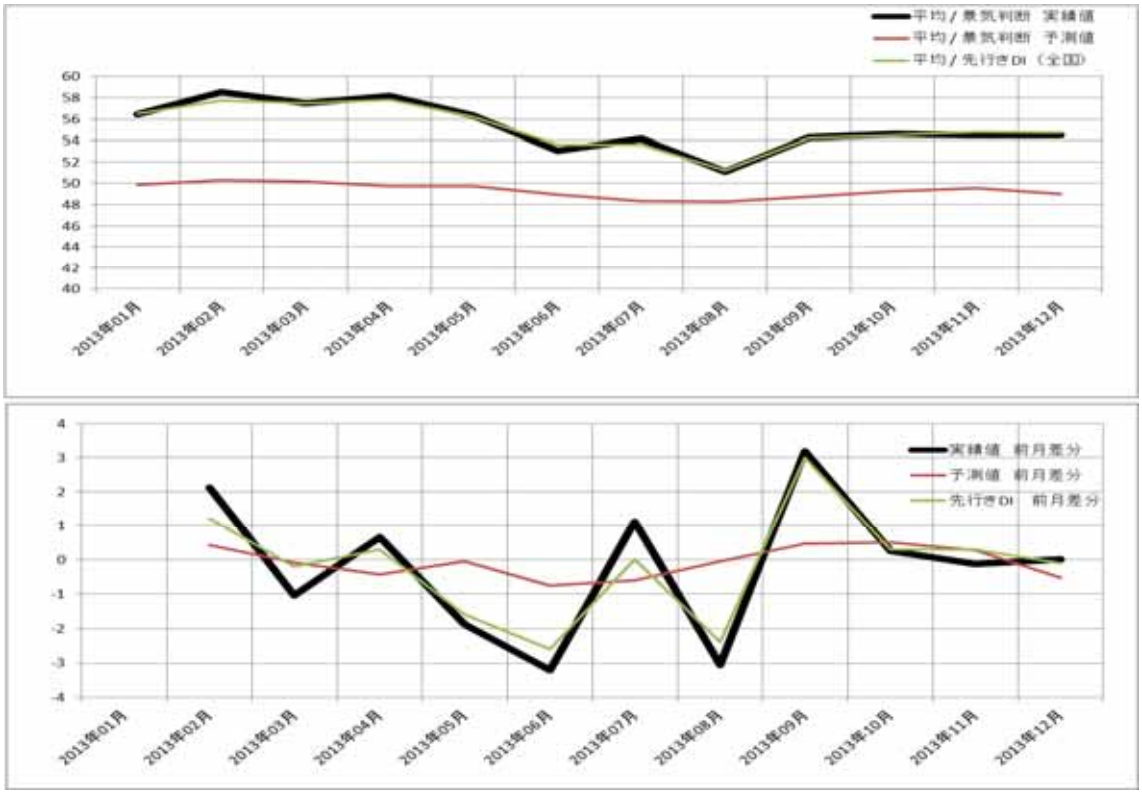


図 3－3－20　分析結果（分析Ｄ－２）に関する全国平均値
図 3－3－21　分析結果（分析Ｄ－２）に関する全国平均値の前月差分

また、前記は全国平均値であったが、地域別の結果を抽出した結果を別紙に示す。参考に、北関東の分析Ｄ－１の結果を図 3－3－22 および図 3－3－23 に示す。

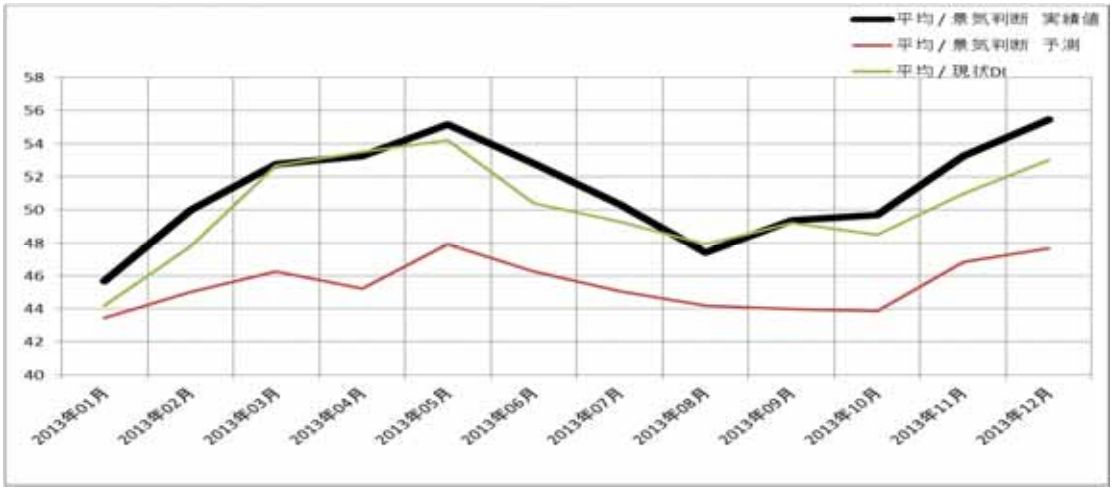


図 3－3－22　分析結果（分析Ｄ－１）に関する北関東の平均値

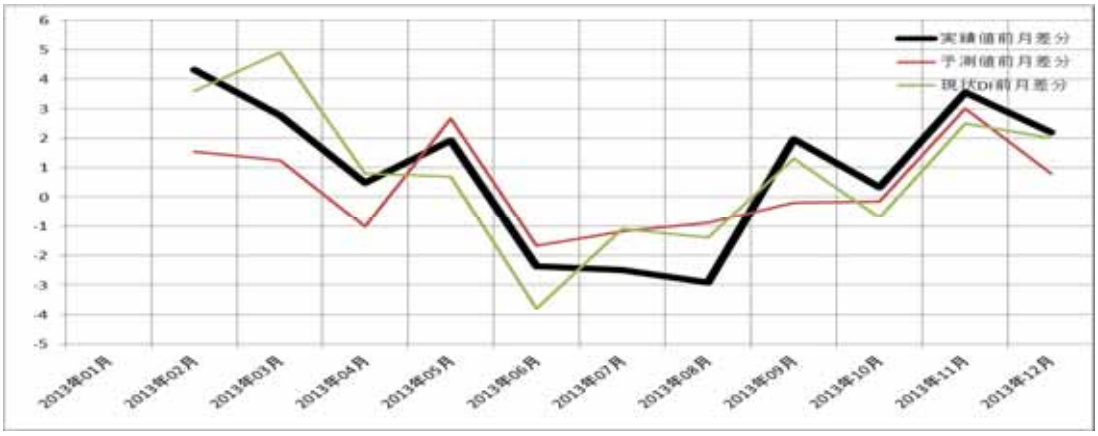


図 3－3－23　分析結果（分析Ｄ－１）に関する北関東の平均値の前月差分

分析Ｄを行った結果、テキスト特徴量を用いた DI 値予測結果の平均値は、図 3－3－18 のように実際の値に比べ低めに出る結果となっている。これは、表 3－3－5 および表 3－3－6 にあるように、学習データ内のデータ数が「やや悪くなった」「悪くなった」に偏っていることと関係していると考えられ、本件を解決するために、学習データ数を「100～0 点」のそれぞれのサンプル数が等しくなる（すなわち、景気ウォッチャー調査結果のコメントにおいて景気判断の 5 値のそれぞれを回答したサンプル数が全サンプル数に占める割合を、ランダムサンプリングの抽出比率として適用する）ように再サンプリングすることで改善される可能性があると考えられる。

また、前述のとおり値はサンプル偏りの関係で平均的に異なった値となっているが、前月との差分という観点にすると、ある程度の精度で差分を予測できていることがわかる。

また、分析Ｄ－１、分析Ｄ－２それぞれについて地域別にすると一部の地域ではなお差分の予測精度が高くなっていることがわかる。今回は、地域別のデータをすべて混ぜて分析を行ったが、地域別に分けての再分析などを行うことで、地域別に景気判断指標と影響があるテキスト特徴量を抽出できる可能性があると考えられる。

また、説明変数の回帰係数を見ると、テキスト特徴量のうち一部の 20～40 語程度に顕著に大きな重みを付けていることがわかる。このことから、テキスト特徴量のうち一部の語に着目することで、DI 値あるいは景気判断の前月比でみた上下が予測できる可能性があると考えられる。

なお、分析Ｄの内容は３－２－１０と同様のものであるが、分析Ｄの方が誤差が大きい結果となった。この違いは、テキスト特徴量として、分析Ｄの場合は有意表現 400 語を使用しているのに対し、３－２－１０の場合は有意表現抽出前の約 1 万語の単語・特徴表現を使用しているためと考えられる。

３－３－６　３－３全体のまとめと考察

今回行った分析において、景気ウォッチャー調査における特定年月のコメントを集団としてヒストグラム化し DI 値を予測しようとする場合、そのおおまかな傾向は予測することができるものの、サンプル数が少なすぎることでオーバーフィッティングが起きるという課題が浮かび上がった。

また全国に加えて地域別サンプルを構築し、見かけのサンプル数を増やす試みも行ったが、課題の解決には至らないことが分かった。

一方、分析Ｄの結果は、（見かけのサンプル数ではなく）、真のサンプル数を増やすことで事態が改善することを示唆する。例えば、個人別の景気判断を個人のコメントから予測し、それを用いて DI 値を予測した場合には、

- 予測値は DI の実績値より小さくなる傾向（負のバイアス）が出ているが前月差分値では予測精度が改善する
- 予測に強く効いているのは数十程度のキーワードである

などが観察されている。

これらは真のサンプル数増加により予測精度の改善可能性を支持する事項であり、サンプリング手法の改善などによる、さらなる精度向上の方法も複数想定できる

以上をまとめると、

- テキストマイニング技術を前提として、適切に収集されたテキストデータは景気判断の説明力を有し、予測に強く効くキーワードの数は必ずしも多くない
- ただし、十分なサンプル数が確保できないと、オーバーフィッティングの問題が発生する

ということになる。

これは、Twitter やネットニュースなどについて、多くない数のキーワードの発生頻度を随時追跡しておくことによって、ネット上での景気判断を随時予測できる可能性を示唆している。