第２章　有識者ヒアリングの実施結果

２－１　使用データや分析手法等に関する指摘事項

有識者ヒアリングにおいては様々な観点からの指摘、コメントを頂いたが、それらは概ね以下の項目に分類することができる。

２－１－１　利用可能な「ビッグデータ」について
２－１－２　ビッグデータを利用して経済、景気等の分析を行った事例
２－１－３　分析に使用するコンピュータハードウェア、ソフトウェア、プログラム等について
２－１－４　ビッグデータと経済、景気の両方を分析・研究している研究者等について
２－１－５　ビッグデータの分析に関する各種考察や定性的な知見
２－１－６　東日本大震災後の日本経済の変化について

このうち、２－１－６については別項２－２として記載することとし、本項では２－１－１から２－１－５のそれぞれについて有識者からの指摘事項の概要をまとめた。

２－１－１　利用可能な「ビッグデータ」について

一般に「ビッグデータ」とは、データベース上に整理・統合されているようなデータの他に、文章（整理されたものではないテキスト）や音声、動画、各種センサーの感知情報、通信ログなどの様々な情報・データで、なおかつ

・データ量が多い（volume）
・データの頻度が多い、更新速度が速い（velocity）
・データの種類が多様（variety）

の３要素をその特質とするものといわれている。

しかし、ビッグデータというのはデータを他の目的に利用するという性格のもの（渡辺努氏指摘事項）であることを考えると、今回分析の目的を達成できる可能性のある「ビッグデータ」の種類は自ずと絞られるものと推測される[1]。

有識者からのヒアリングにおいて、経済状況や景気循環の分析目的に利用が可能と考えられる、あるいは既に研究レベルで利用しているデータは以下の通りであった。

① Twitter の tweets（つぶやきのテキストデータ）、及び情報サービス会社がそれら

[1] 「ビッグデータ」の種類やデータの内容は今後増えていく、充実していくことが考えられ、一方で分析技術やコンピュータの処理能力も今後進歩していくであろうから、現在はある分析目的を達成するためにビッグデータは利用可能ではないと判断されたとしても、将来は利用可能となることは十分ありえる。また、データ入手・利用のための費用についても変化が起こり、例えば商業上利用可能ではなかったものが将来は利用可能になることも考えられる。

をキーワード毎に抽出、集計したデータ

② 各種のブログに掲載されているテキストデータ

③ Google などの検索サイトにおける検索履歴（ログ）、及びそれらをキーワード毎に抽出、集計したデータ

④ コンビニエンスストアやスーパーマーケットの POS データ（コンビニエンスストアチェーン単体の購入履歴、情報サービス会社が集計したスーパーマーケット店舗の POS データ）

⑤ 価格.com（家電製品等）、Suumo（不動産）などの価格情報サイトにおける価格データ、購買履歴データ[2]

⑥ ニュース配信会社が機関投資家向けに配信するビジネスニュースのテキスト

⑦ 民間調査会社の企業財務データ

⑧ 銀行の送金（振込）データ

⑨ 金融・株式市場における注文・取引データや価格データ

⑩ 中央銀行が発行する金融経済月報のテキスト

以上 10 種類のデータのうち、⑧は一般に入手または利用することが現時点では不可能であるが、そのほかのデータは有償、あるいは無償で入手または利用することが可能である。

ただし、有償による入手、利用に際しては、利用する側の主体あるいは目的によってデータの価格が大きく異なることがありえる点に留意が必要である。すなわち、研究目的で研究機関あるいは研究者がデータを利用する際にデータ出所元に対し支払う金額は、一般の企業が商用目的でデータを利用する際に支払う金額（定価）とは大きく異なる可能性があり、後者の方が価格が高い（それも金額の桁が異なる）ことが多いという指摘を有識者からは頂いた。

また、入手、利用可能なデータを加工の有無の観点から分類すると、

a. ニュース、テキスト、数値データ等をそのまま提供する「一次データ」、あるいは「生データ」

b. テキスト内に特定ワードが出現した回数や、数値データの集計値など、一次データを加工して利用しやすい形に変換した「二次データ」

の２種類がある。一般に、b.はデータを利用目的に合致するよう利用者が加工する手間がかからないという利点があるが、その分提供価格は高くなる傾向がある。

（資料１参照）

---

[2] この他、不動産取引情報の提供サイトとしては成約価格を基にした情報サイト「REINS」(レインズ) もある。データはデジタル化されたデータとして利用可能で、同データを使用した中古不動産価格の分析実例もある（資料３参照）。
http://www.contract.reins.or.jp/search/displayAreaConditionBLogic.do

利用可能なデータの量については、ビッグデータの分析が本格的に始まったとされる2000年代初頭と比較すると現在は飛躍的に増加している。現在では、数億件レベルのデータを分析対象とすることがごく一般的に行われている。（資料２参照）

２－１－２　ビッグデータを利用して経済、景気等の分析を行った事例

有識者からは、有識者本人の研究を含めさまざまな事例につき指摘を頂いた。
（有識者からの指摘に基づく事例のリストについては資料３参照）

指摘頂いた事例を、経済や景気等におけるどのような事象を分析、解明することを目指したかという目的別に分類すると、概ね以下の通りとなる（カッコ内は分析に利用したデータの種類）。

① 金融市場、不動産におけるバブルやインフレなどの兆候を早期発見する（為替取引データ、株価データ、インターネット上の不動産価格・取引データ）
② 商品販売の売れ行き（ブーム）の予測（コンビニエンスストアの販売記録、Twitterやブログのテキスト）
③ 急成長企業の成長メカニズムの解明（企業の財務諸表、特許に関するデータ）
④ 商品の価格の値崩れ、暴騰メカニズムの解明（インターネット上の家電等の価格、販売データ）
⑤ インフルエンザの流行現状の把握（インターネットの検索ログ）
⑥ 景気ウォッチャーDIの推計、予測（ブログのテキスト）
⑦ 景気や企業業績、株価の予測（機関投資家向けのビジネスニュースのテキスト、企業の財務諸表データ、銀行の取引データ）
⑧ 消費者物価指数速報の推計（スーパーマーケットのPOSデータ集計値）
⑨ 失業率（米国）の推計（インターネットの検索ログ）
⑩ 株式市場（平均株価）の騰落予測（Twitterのテキスト）
⑪ 国債利回りの推計、予測（日銀金融経済月報のテキスト）

また、これらの他にも、

⑫ 景気動向指数（一致指数）の推計、予測（インターネットの検索ログ）
⑬ ISM製造業景況指数（米国）の推計、予測（ニュース記事のテキスト）

などの事例がインターネット上で公開[3]されている。

上記の事例からは、分析の目的に応じ使用するデータの種類もある程度色分けされる傾向が全般にみてとれる。例えば、景気の推移を直接に示す指標である景気指数（あるいはDI）の推計や予測に際しては、インターネット上の検索ログやニュース記事、ブ

[3] 参照先は以下の通り。http://adv.yomiuri.co.jp/ojo/tokusyu/20130805/201308toku2.html
http://event.yahoo.co.jp/bigdata/keiki/

ログなどのテキストデータが利用されるケースが現在までのところ多いことがわかる。また、物価全般、あるいは対象品目の物価を公式統計よりも早く把握する目的では、インターネット上あるいは POS 集計による価格データが利用されている事例が多い。さらに、バブルや流行など、平時の経済や景気の推移と異なる動きをいち早く捉えるために、金融市場のデータやインターネット上の検索ログ、テキストデータといった各種のビッグデータが利用されている事例がみられることも注目に値する。

２－１－３　分析に使用するコンピュータハードウェア、ソフトウェア、プログラム等について

ビッグデータの分析、解析にはコンピュータの利用が不可欠である。しかし、ビッグデータの解析を実際に行おうとする際に、一般には以下のような基本的な情報が不足しているものと推察される。

・どのようなスペックのハードウェア、ソフトウェアを準備すれば分析が可能となるのか
・解析のためのソフトウェアは市販、あるいは無料で入手可能（フリーウェア）なのか
・もし独自に解析プログラムを作成する場合にはどのようなプログラム言語を使用するのが一般的、かつ効率的なのか

このため、有識者ヒアリングにおいては上記の実務的な情報においても可能な限り伺った（結果の集計は資料４を参照）。

断片的な情報ではあるが、ヒアリングからは大要以下の示唆が得られた。

・ハードウェアに関しては、分析の対象とするデータの量によって必要となるスペックは異なる。金融市場における取引データなどの膨大なデータ量を処理する場合にはマシンを並列させて計算をする、あるいはスーパーコンピュータを使用することがあるが、一般的なテキストの分析ならばパソコンで対応が可能。
・データの分析を具体的に実行する際に、分析目的に応じどのようにでも対応できるような汎用ソフトウェアは存在しないため、自前で作成するか、あるいは外注により作成をすることが必要。ただし、テキストマイニングや統計解析用のツールは市販されてはいる。
・形態素解析などのツールは一部無料（フリーウェア）で利用可能。ただし利用するには、Python や Java などのプログラム上にソフトウェアをのせないといけないので、自前で分析を行うにはプログラム言語を習得する必要がある。習得期間は大体半年程度。これに加え、もし自前で解析用のソフトウェアを作成する場合は、それに関する知識が必要となる。従い、ビッグデータの解析には人的並びに予算面の

リソース準備が不可欠。

２－１－４　ビッグデータと経済、景気の両方を分析・研究している研究者等について

「１－１ 調査目的」で前述した通り、ビッグデータ業務は盛んに行われているが、経済構造や景気局面の変化に対する研究としてはまだ十分ではない。本分析業務の成果を将来的に拡充し、経済構造や景気局面の変化をより早く、より正確に捉えていくためには、内閣府をはじめとした政府部門（官）だけではなく学術分野（学）、民間企業部門（産）を含めた知見、技術、そしてデータの集約や協力関係構築が不可欠とみられる。

上記の観点から、有識者ヒアリングにおいては
・経済や景気の分析、予測をする目的でビッグデータを分析・研究している人物
についての情報を伺った。

一般に、ビッグデータの分析や研究をしている人物、組織は
① 大学や研究所などの学術部門
② ビジネスインテリジェンス、ビジネスアナリティクス[4]の業務を提供する金融系、ICT 系などのベンダー
③ シンクタンク、コンサルタント会社
④ 政府系機関（公的機関）
に大別される。主に、①は研究目的、②と③は商業目的、そして④は例えば金融取引や国民生活の秩序維持などの個別の公共目的を達成のためにそれぞれ活動を行っていると考えられる。

しかし、上記の中で経済や景気の分析、予測をする目的でビッグデータの分析・研究をしている人物の数は有識者ヒアリングの結果をみても少ない模様である[5]。また、その多くは学術分野（上記①）の研究者に偏り、②や③などの民間部門に所属する人物（いわゆる「データサイエンティスト」と呼称される人）の数はさらに少ない模様である（資料５参照）。

---

[4]一般に、Business Intelligence とは、企業などの組織のデータを経営上などの意思決定に役立てる手法や技術を指す。同手法の中には、データ分析、データマイニング、テキストマイニングなどが含まれる。また、Business Analytics とは上記の手法を踏まえた将来予測や最適化問題解決の提案等を含めた概念。

[5]経済や景気の分析、予測をするという目的を考えると、関係する分野としては
1)経済、2)統計データ、3)人工知能（解析技術）
の３つが考えられるという指摘を和泉潔氏から頂いた。

２－１－５　ビッグデータの分析に関する各種考察や定性的な知見

有識者ヒアリングにおいては、上記２－１－１～２－１－４以外にもビッグデータの分析に関するさまざまな指摘、知見を頂いた。それらの概要をまとめると以下の通りである。（詳細は資料２参照）

〇データの種類による特性について

・景気の現状は語るが、将来については語らないのがブログの特徴なので、景気予測をするときにはブログは使いにくい。さらに、ブログには産業別や企業に関する情報は少ない。
一方、機関投資家向けのニュース、コラムを利用すれば、Twitter では得られない企業内部の情報や、社会のお金の流れを察知することが可能。

・東大物価指数は日次で存在するので、例えば東日本大震災前後の動きは月次の CPI よりもより詳細に把握可能。また、増税時や TV 番組放送後の価格の動きなど、イベント発生時の動向把握も可能。

〇分析の手法や手順、実施体制等について

・自然言語処理におけるネガ・ポジの判定を用い、ブログ等テキスト上の「よい」「悪い」を表す単語の出現回数（比率）を 2 つの説明変数として、景気ウォッチャー調査 DI 等との相関推計が可能。

・テキストの内容から地域、業種別の判定も可能になりつつあるが、景気調査と合わせるようなかたちでサンプリングのバランスをとることが必要。

・景気、経済の構成要素別にニュース等のテキストを抽出・分類する手法もある。

・例えばテキストデータを使って予測する場合、どういったデータを使うかによって手法、予測頻度（日次、月次等）も変わる。

・Tweets やブログの中には関連（重要）キーワード以外の、いわゆる「ゴミ」の部分が非常に多い。従って、Twitter のテキストデータから関連語を抜き出す際には、キーワード全てではなく、何か絞り込む（フィルタリングの）手法、アイデアが分析の肝になる。しかしそれがないときは、ちょっと少なくてもいいから、少数のテキストを使う。

・東大物価指数の作成の際、数値計算の再現性を重視している。すなわち、計算アルゴリズムは完全開示し、作成データは可能な範囲内ですべて開示。

・東大物価指数作成のコスト構成は、POS データの利用価格、データの配信料、作成作業にかかる人件費、プログラムの作成費用。これらを全て合計したとしても、人海戦術で行う集計作業と比較するとコストの桁が低いと思う。

・指数算出を継続するには、原データの出所元からの継続コミットの他に、研究資

金の継続が必要。さらに、将来の対象範囲・品目カバレッジの拡大と、それに伴う作業工数の確保を考えると、大学の研究室内だけで実施するのは限界がくる。アルゴリズムの開発までは大学内でよいとしても、その段階以降は企業やシンクタンク、政府のほうが上手にやれる可能性がある。

○ビッグデータによる景気の分析に関する示唆

・できるだけリアルタイムで、簡易に、可能な範囲で景気を把握するとしたら、例えば口コミをもとにした景気指数の推計を、既存のプログラムを応用して伸ばしてみるのが可能性の一つ。

・ブログや Twitter を分析するよりも、投資家向けの情報端末に流れているニュースの分析をしたほうが、景気のことはわかると思う。ブログや Twitter のテキストデータは、だれが書いているかというサンプリングの問題等が出てきて、最後の詰めが非常に難しい。もっとも、ビッグデータの持つ性質上、サンプリングバイアスが生じるのは止むを得ない。その限界を認識した上でサンプリングバイアスを補完し、精度を上げていく工夫が必要。

・景気ウォッチャー調査のデータでは、分析の際にテキスト分量が足りない可能性はある。

・ナウキャスティングを試行的に始めるのならば、まず、既に持っている景気ウォッチャー調査の少数のテキストデータに関して、キーワードの抽出とポジ・ネガ判定を行う。次に、ピックアップしたキーワードが、ブログもしくは Twitter においてポジ・ネガの文脈付きの回数として日次で何回出てきたかを調べる。

○ビッグデータによる因果関係の把握について

・例えば台風などの外的ショックが起こったとき、それを外的ショックとしてデータから外して景気判断することが妥当かどうかの判別は難しい。また、景気の良し悪しとデータの分析結果との間の因果関係の把握も非常に難しい。実務家のコメントをもらい、経済理論に基づき因果を予測し、モデル化をすることが望ましい。経済学と情報処理、自然言語処理を結び付けて分析しないと、実務に耐えられない。

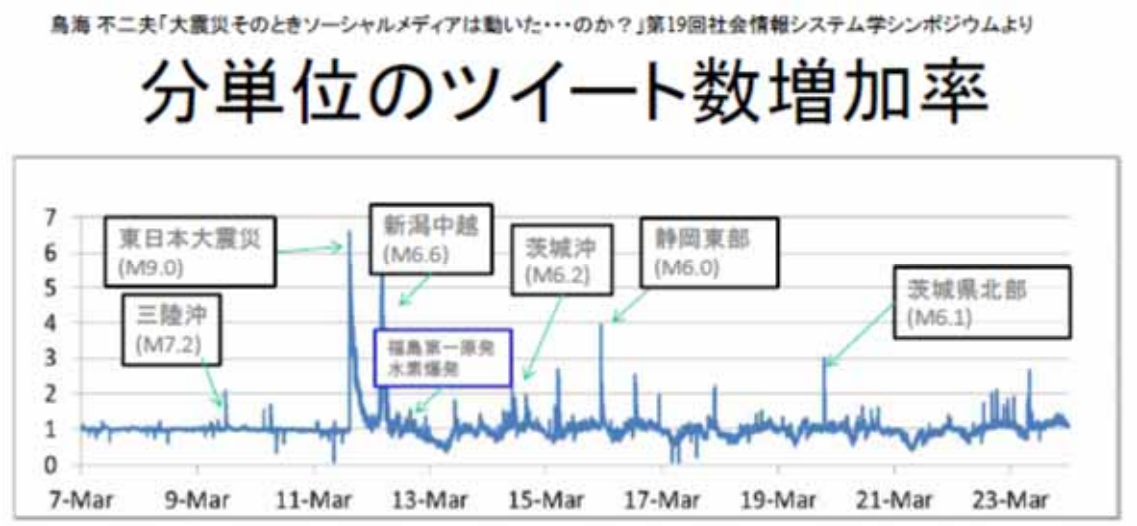
・ビッグデータを使って過去は分析できるが、外挿は非常に難しい。因果関係に結び付けて将来の推定という段階になったら、人間がデータの裏にある経済構造などのモデル、仮説を入れて見るべき。しかし、全く今までなかったイベントや事象が起こった場合は、いくら過去の事象をビッグデータで集めても、外挿するのは難しい。個人の経済行動のモデル、仮説を作ってから、将来に関しては、個人個人が集まった社会・経済全体がどう動くかということをシミュレーションしていく立場がこれから必要になると思う。

２－２　東日本大震災後の日本経済の変化に関する指摘事項

有識者ヒアリングにおいて、東日本大震災後における日本経済の構造や景況の変化に関する指摘は以下の２項目であった。

２－２－１　情報プラットフォームと情報ネットワークのルートにおける変化

東日本大震災前後の Twitter データを分析した結果[6]をみると、まず当然のことながら、Tweets それ自体の数が震災直後には急激に増えている。また、震災後に発生した比較的大きな規模の地震の直後においても一時的な急増がみられる。



（グラフの縦軸スケールは、震災前の同時刻の震災前平均 Tweet 数で正規化<normalize>したもの）

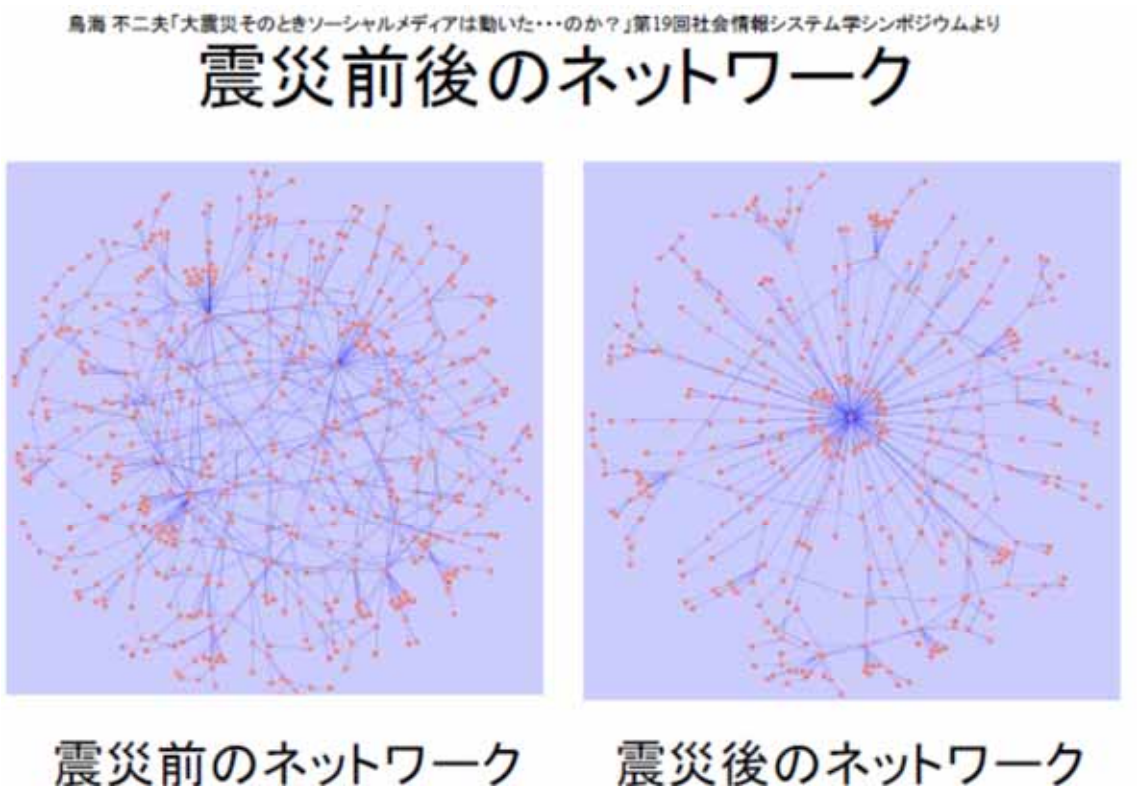
さらに、ユーザー間の「つながり」の形態をみると、

・震災前（通常時）は、日常のユーザー間のつながりを基盤とした、一般的な話題にリツイーツされるコミュニケーションの構造であったものが、

・震災後は NHK、消防庁、地震速報などの公的な情報源を中心として、それに情報を加える、もしくは詳細な情報を一般のユーザーが足していく、というような形で広がりを見せる、かなり集中度が高まったコミュニケーションの構造が現出した。（下図参照）

[6]第 19 回社会情報システム学シンポジウム「大震災・そのときソーシャルメディアは動いた…のか?.」鳥海不二夫、篠田孝祐、榊原剛史（2013.1.23）。Twitter のデータはユーザー数約 3000 万、tweets 数約 4 億。

震災時においては、特に被災地域を中心に通信機能の途絶、遅延が発生し、一般市民における情報収集手段・能力が低下したなかで、Twitter 等の SNS ツールが役立ったことはよく知られている。震災後におけるコミュニケーションツールの利用目的が情報収集にシフトし、そのため公的情報源を中心（ノード=結節点）としたいわば情報提供ツールに変貌したことがこの構造変化からは窺える。



（赤い点が Twitter のユーザー、赤い点をつなぐ線が Twitter によるコミュニケーションのつながりを示す）

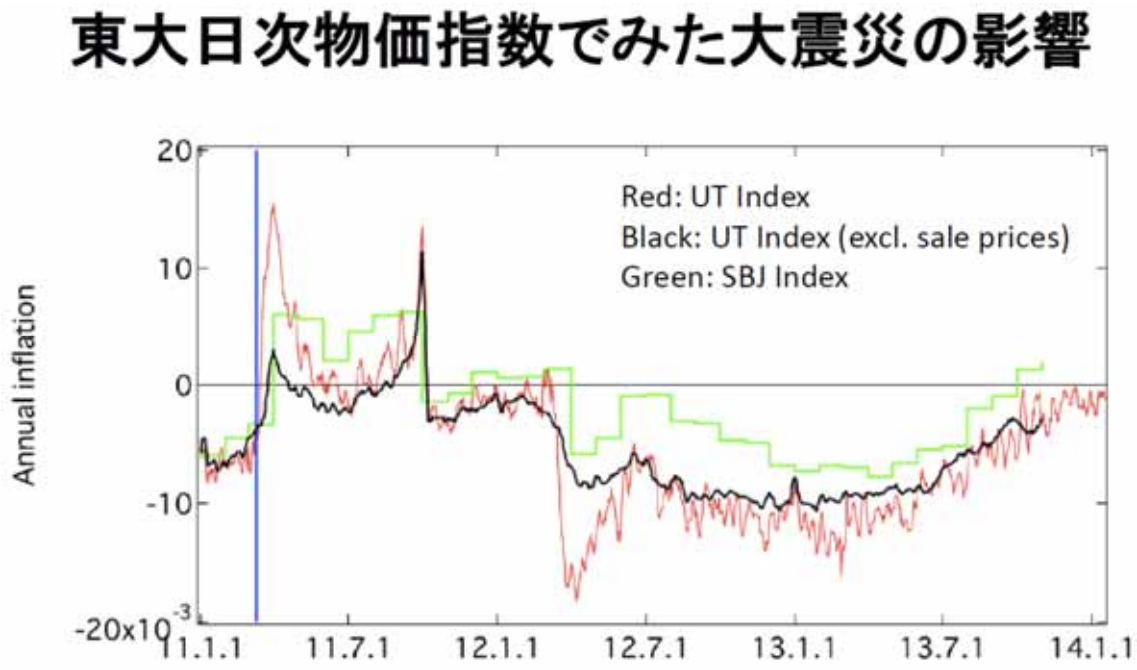
２－２－２　消費者物価の変動

「ビッグデータ」の中にはデータの更新、入力頻度が多いものがあり、その特性を上手に利用すると政府機関等が公表している統計よりもより詳細に、かつ異なる視点からデータを把握することが可能になる場合がある。

「東大日次物価指数」の分析結果によると、対象品目(170 品目。総務省・消費者物価指数は 588 品目)の物価上昇率(前年同日比)は、震災前はマイナス 0.5%前後だったところから、震災直後にはプラス 1.5%前後に急上昇している。しかし、その内訳をみると、定価(商品の通常販売価格)を引き上げたことによる値上げの幅は小さいことから、震災前後における約 2%ポイントの物価上昇のほとんどは特売によるものということが

わかる。すなわち、スーパーにおいては定期的に特売を実施しているのが通常期の販売活動と推測されるが、震災直後においては特売を実施した頻度が減少した結果、いわば消極的な意味での値上げが起きたことを示している。

また、この消極的値上げが解消し、通常期のパターンに戻る（下グラフにおいて、赤の折れ線と黒の折れ線の乖離幅が縮小する）までには数カ月という比較的長い期間を要したこともわかる。



(注)タテ軸は前年同日比の物価上昇率(1 目盛は 1%ポイント)

青色タテ線=震災発生時(2011 年 3 月 11 日)

(折れ線) 赤=東大物価指数(日次)

黒=東大物価指数(日次。特売分を除き、通常販売価格の商品のみを抽出したベース)

緑=総務省統計局公表の消費者物価指数(月次。対象品目を東大物価指数の対象品目に合わせたベース)