# 論文 / 著書情報
# Article / Book Information

| | |
|---|---|
| 題目(和文) | ソーシャルメディアにおける単語出現頻度時系列の解析とモデル化 |
| Title(English) | Empirical analysis and modeling of word frequency time series in social media |
| 著者(和文) | 岡田幸恵 |
| Author(English) | Yukie Sano |
| 出典(和文) | 学位:博士(理学),<br>学位授与機関:東京工業大学,<br>報告番号:甲第9265号,<br>授与年月日:2013年9月25日,<br>学位の種別:課程博士,<br>審査員:高安 美佐子,奥村 学,樺島 祥介,寺野 隆雄,小野 功 |
| Citation(English) | Degree:Doctor (Science),<br>Conferring organization: Tokyo Institute of Technology,<br>Report number:甲第9265号,<br>Conferred date:2013/9/25,<br>Degree Type:Course doctor,<br>Examiner:,,,, |
| 学位種別(和文) | 博士論文 |
| Type(English) | Doctoral Thesis |

# ソーシャルメディアにおける
# 単語出現頻度時系列の解析とモデル化

Empirical analysis and modeling of word frequency
time series in social media

東京工業大学
大学院総合理工学研究科
知能システム科学専攻

博士 (理学) 学位論文

岡田 (佐野)　幸恵
Yukie Sano

2013 年 9 月

# 目次

本論文は，以下の 3 論文の内容に基づいている．

[A] Yukie Sano and Misako Takayasu.
*Macroscopic and microscopic statistical properties observed in blog entries.*
**Journal of Economic Interaction and Coordination**, vol. 5 pp. 221–230 (2010).
主に 3 章「データ処理の前準備」.

[B] Yukie Sano, Kimmo Kaski, and Misako Takayasu.
*Statistics of collective human behaviors observed in blog entries.*
**Proceedings of the 9th Asia-Pacific Complex Systems Conference**,
pp. 195–198 (2009).
主に 4 章の「時系列からの異常値検出」.

[C] Yukie Sano, Kenta Yamada, Hayafumi Watanabe, Hideki Takayasu, and Misako Takayasu.
*Empirical analysis of collective human behavior for extraordinary events in the blogosphere.*
**Physical Review E**, vol. 87, 012805 (2013).
主に 5 章の「有限時間発散する時系列の解析とモデル化」.

## その他参考文献

[D] 高安美佐子編著
ソーシャルメディアの経済物理学 -ウェブから読み解く人間行動.
日本評論社 (2012).
第 3 章ブログデータの解析担当 (pp. 54–117).

[E] Yukie Sano, Hideki Takayasu, and Misako Takayasu.
*Zipf's Law and Heaps' Law Can Predict the Size of Potential Words.*
**Progress of Theoretical Physics Supplement**, vol. 194, pp. 202–209 (2012).
ブログなどの文章における Zipf 則と Heaps 則について調べた論文.

[F] 佐野幸恵，高安秀樹，高安美佐子
文書中の語彙ネットワークにおける Zipf 則と Heaps 則
**ネットワークが創発する知能研究会 (JWEIN12) 発表論文集** No.12007 (2012).
[E] の論文と関連し，単語ネットワークにおける Zipf 則と Heaps 則について調べた論文.

# 第 1 章

# 序論

データに基づく社会現象への科学的研究の背景と歴史を説明する．また，本研究が対象とするソーシャルメディアの日本での普及状況，実世界に期待されている役割について述べる．特に本研究では日本で最も普及しているソーシャルメディアの一つであるブログに注目する．そこで，多くの分野に渡って存在するブログに関連する研究を概観し，最後に本研究の目的と構成ついて述べる．

## 1.1 背景

### 1.1.1 データと実証科学

科学の発展には精緻なデータの収集と解析が不可欠である．そして，データに基づく新たな理論の構築と，データに基づく理論の検証は，実証科学の根幹をなしてきた．例えば 16 世紀の天文学者 Johannes Kepler は，Tycho Brahe による膨大で詳細な天体の観測データに基づき，天体の運行法則 (Kepler の法則) を導いた．Kepler の法則では，惑星は太陽を一つの焦点とする楕円軌道を描くこと (第一法則)，その軌道と太陽を結ぶ単位時間あたり描く面積は一定になること (第二法則)，惑星の公転周期の二乗は軌道の長半径の三乗に比例すること (第三法則) を示した．Kepler は宇宙を力学系という視点から捉えた最初の科学者であり，Kepler の発見が続く 17 世紀の Sir Isaac Newton による重力理論を打ち立てるきっかけとなった [1]．すなわち古典力学を創始し，微積分法を発明した近代物理学の祖である Sir Isaac Newton の偉大な功績も，Tycho Brahe によるデータから端を発しているといっても過言ではない．それでは，21 世紀の実証科学は，どのような新しいデータに基づき，法則を導き，理論を構築し，検証して行くべきであろうか．

現在の社会経済において ICT(Information and Communication Technology) の急速な発展と普及は無視することができない．ある調査会社の推計によれば，2006 年に人類

が送出したデータは約 160E バイト ($1.6 \times 10^{20}$ バイト) だったが，2011 年には約 2Z バイト ($2 \times 10^{21}$ バイト) になり，2016 年までにはその 4 倍の 8Z バイトに達すると予想されている [2]．つまり ICT の急速な発展は Kepler の時代から 20 世紀まで入手することが困難であった膨大な量の社会経済データを，低コストで記録・保持することを可能としたのである．

特に注目すべきは，個人の言動に関する社会経済データである．何時何分何秒，どの店舗で，何の商品が，どの商品と一緒に，どういう人に買われたか，といった細かな販売履歴情報と，仕入れや廃棄の情報まで含む小売業者の保持するデータ．インターネット上で何時何分何秒，どのエリアで，どんな検索単語と一緒に，単語が検索されたかを記録した検索エンジンの持つデータ．GPS 内蔵の携帯電話から送られてくる位置情報データ．改札口で電子的に記録される行動履歴，図書館の貸し出し履歴，ウェブサイトでの購入履歴や，アクセス履歴など，これまでの実証科学が研究対象とすることが困難であった新しいデータがあふれる時代となったのである．

この天文学的な量のデータから，Kepler のように新しい法則を導き，理論を構築し，それを検証する新しい方法論を築くことが，今後のさらなる科学の発展のために不可欠であると考えられる．しかし一方で，社会経済データの基礎となる人間は，自由な意志を持ち，自律的に行動し，時には説明困難な感情的な振る舞いをすることさえある．彼らの言動を精緻に観測し，解析した結果として，そこに何らかの法則性，数学的パターンを見いだせるのだろうか？

### 1.1.2 数理に基づく社会現象の理解

社会に潜む数学的な法則を社会経済データから探索し，モデル化して理解するという実証科学的挑戦は，経済物理学 (econophysics) や社会物理学 (sociophysics) と呼ばれる新しい研究分野が行なっている．そこでは，人間は意思を持たず，何らかの法則に支配された「社会的原子 (social atom)[3]」とも表現される無機質な個体として仮定される．その研究背景には，気体分子の運動量に代表される微視的 (ミクロ) 状態量と，温度に代表される巨視的 (マクロ) 状態量を結びつけ，世界を理解してきた統計物理学の類推 (アナロジー) で，社会を理解できないかという物理学者の期待が存在している．そういった期待を持つ物理学者は，個人を微視的状態量，社会を巨視的状態量とみなして社会現象を数学的側面から解明することに，新たな可能性を感じているのである [3, 4]．

数学的な視点で人間社会を記述しようとする試み自体の歴史は古く，17 世紀から存在した [5]．例えば Pierre-Simon Laplace は年ごとのパリの男女の出生率を時系列で比較し，その比率がいつもほぼ等しいことを発見し，この発見を神の英知の証であると論じた．また Laplace や Siméon Denis Poisson に学んだ 19 世紀のベルギーの天文学者

Adolphe Quételet も社会の統計に見られる数学的な法則に引きつけられた一人である．彼は社会調査データに関する統計学についての業績を残し，mechanical social science という新たな研究分野を提案した．統計調査に基づく病院管理の先駆けとなった Florence Nightingale もオックスフォードで彼の mechanical social science を学んだ．しかし 17 世紀から 19 世紀の科学者が対象としてきたのは，出生率や犯罪率などのマクロ社会統計であり，集計データの表層的な相関関係から社会現象のメカニズムを規範的に議論することに終始していた．

一方，経済物理学は外国為替時系列に代表される精緻で大規模に収集されたミクロ経済データに注目した物理学者によって，1990 年代後半に誕生した新しい分野 [6, 7] である．経済物理学は，経済データに基づいて科学的な統計性を議論する点では，計量経済学や金融工学と同様の関心を有する．しかしながら，経済物理学は，経済現象を物質科学と同様の視点で実証的に分析し，カオスやフラクタルに代表される統計物理学の理論や解析手法を用いる点で特長的である．その代表的な結果として，為替時系列の変動の大きさの分布が，従来の経済学で仮定されてきたような正規分布ではなく，もっと分布の裾野の広いベキ分布になることを発見したことが広く知られている．ベキ分布は，物理の破壊現象や相転移現象ともなじみが深く，統計物理学の分野ではそのベキ分布の背後にある因果メカニズムを取り扱った理論モデルも存在している．そのため，経済現象の因果メカニズムに関する数学的な法則を発見し，新たな理論を構築するために，経済物理学は新しい研究分野として大いに注目され，今まさに世界的な研究が蓄積されている．

### 1.1.3 インターネットとデータ

インターネットはデータの宝庫であり，そこに蓄積される情報量は日に日に増加している．例えば，ルータ間のトラフィックデータ，ウェブサイトへのアクセスデータ，さらに人々の書き込みデータまですべて，タイムスタンプ付きで精緻に記録されている．また，インターネットのブロードバンド化に伴い，買い物や電話でのコミュニケーションなど，これまで実世界で行われていたことが，インターネットの世界へと置き換わっている．その結果，インターネットの世界が実世界へと与える影響は日に日に大きくなり，大規模なデータを使った新たな社会科学の出現 [9] というだけに留まらず，インターネットの世界で起きていることを正確に把握することは急務となっている．実際に，アメリカ国防省では 2010 年の計画で，インターネットを含むサイバー空間は陸，海，空，宇宙と同様の国際公共財 (Global Commons) のひとつとして，アクセスを保証することが必要だと発表している．日本でも平成 24 年度の防衛白書に「サイバー空間をめぐる動向」という項目を設け，取り組みが進められている．さらに，2013 年 4 月には日本でインターネットを使った選挙活動も認められる法案が可決され，ますますインターネットの世界を流れる情

報を把握する重要性が高まり，その需要が拡大している．

**ソーシャルメディアの普及**

21 世紀になって，インターネットを基盤としたソーシャルメディア (social media) が日常生活に浸透してきている．ソーシャルメディアは，従来のマスメディアとは異なり，インターネットを基盤とすることから，双方向型で情報をやり取りすることができることが最大の特徴である．また，インターネットを基盤としたシステムは，分散処理されているいため，従来の電話などと異なり，通信の迂回路が多数存在し，災害時にも比較的復旧が早い．そのため，2011 年の東日本大震災時にも，多くの人がソーシャルメディアを使って，必要な人と連絡し合ったり，最新の情報や局地的な情報を得たりした [8].

ソーシャルメディアの中でも Twitter(詳細は第 2 章で述べる) は情報発信，その拡散共有が容易であることから，災害時に機能する情報媒体として注目されている．そのため，地震や洪水などの非常時におけるソーシャルメディア上での動きも盛んに研究されている [10, 11, 12, 13, 14]．図 1.1 は，2011 年 3 月 11 日の東日本大震災時における全投稿数の変化である．データは「東日本大震災ビックデータワークショップ-Project311-[*1]」を通じて Twitter Japan 株式会社から公式に提供されたものを使っている．実際の地震のデータ[*2]と同時刻で比較すると，本震直前は 1 秒あたりの投稿数は約 200 件であったものが，本震が収まってから急激に増え続け，わずか 3 分後には 1 秒あたり約 1200 件の投稿となっていることが分かる．

さらに，Twitter を使った実世界の記述は，緊急時の場合だけには収まらない．選挙の結果予測 [16, 17, 18, 19]，社会の雰囲気の定量化 [20, 21, 22]，商品や映画の興行収入の予測 [23, 24, 25] といった研究も 2010 年前後より急速に増えている．この Twitter は 1 回あたり 140 文字までの文章が投稿できる仕組みで，ミニブログとも呼ばれている．Twitter に代表される，一般の人が任意の書き込みをできるソーシャルメディアは，社会を大規模に，定量的に理解するためのデータとして近年，注目を集めている．

本研究では，ソーシャルメディアの中でも日本語のブログデータを用いているが，解析の結果得られた統計性は，英語などの他の言語や，ブログ以外の異なる媒体のデータでも観測できるものも多い．例えば，図 1.2 の左図は，Google Trend[*3]という検索サービスを使い，半角英語で「April Fool」という単語を検索した結果である．Google Trend は，検索したい単語を入力すると，過去のその単語の検索回数や，ニュースへ取り上げられた回数を使い，独自に重み付けされた指数を出力するサービスである．さらに，検索元や

[*1] https://sites.google.com/site/prj311/

[*2] http://www.seisvol.kishou.go.jp/eq/kyoshin/jishin/110311_tohokuchiho-taiheiyouoki/data/L311E4E1.csv (Accessed:2013.07.12)

[*3] http://www.google.co.jp/trends/

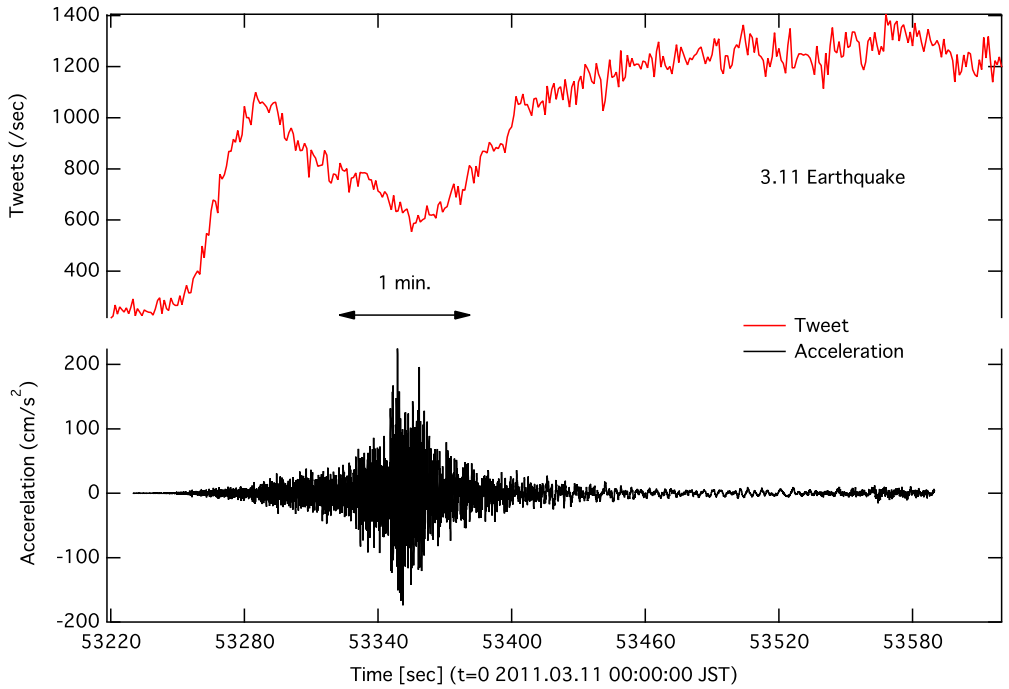


図 1.1　2011 年 3 月 11 日の東日本大震災時における Twitter への投稿数の変化 [15]. (上段)1 秒あたりの投稿数，(下段) 東京都千代田区大手町での南北方向の最大加速度 (gal = cm/s$^2$).

ニュースの情報源をもとに，世界中の任意のエリアを対象に絞り込んで検索することができる．図 1.2 の左図では，対象を全世界として，2008 年から 3 年分を 1 週間ごとに検索したところ，ちょうど 4 月 1 日のエイプリルフールを含む週で毎年，鋭いピークを持つことを確認できる．

図 1.2 の右図では，同じく 2008 年から 3 年間を対象に，本研究で扱った日本語のブログの世界で，日本語のカタカナで「エイプリルフール」と検索した結果の 1 週間ごとの時系列である．縦軸の数値は，全投稿数で除算する規格化を行った後の値になっている．図 1.2 の左図と同様に 4 月 1 日を含む週で鋭いピークを持つ．これらから，検索単語や対象とするエリア，言語は異なるが，ほぼ同じ変動を示していることを確認できる．

ブログと Twitter で比較しても，類似した統計性が得られることが分かってきている．「津波」等に代表される，大きなニュースに対して一斉に社会の関心が集まる場合，どちらの媒体でも関連する単語の書き込み数が増え，その後ベキ関数的に減少する (第 5 章の図 5.19 と図 5.20)．同一人物が 1 日に投稿する回数や，その時間間隔はブログと比較して Twitter の方が頻繁ではあるという違いはあるが，集団としてみた場合，そこには共通する統計性を見いだすことができる．このように人の集団としての振る舞いは，観測する媒体や国境，言語の枠を超え普遍的であることが期待できる．

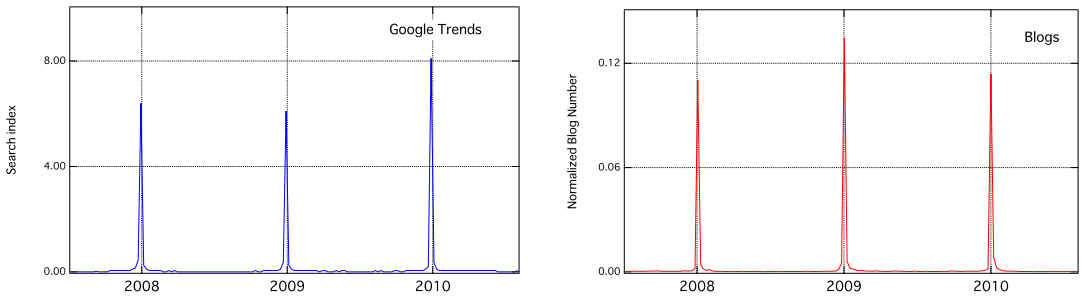


図 1.2 英語と日本語での検索結果の比較．どちらも同じタイミングで鋭いピークを持つ．(左図)Google Trend を使って検索した「April Fool」の結果．(右図) 日本語のブログで「エイプリルフール」を検索した結果．

**ブログの仕組みと取り巻く状況**

ブログは，ソーシャルメディアの中でも歴史が古く，2003 年頃から日本で広く普及し，最も利用されているものの一つである (図 1.3)．ブログ自体は 1990 年代後半にアメリカで始まった．他方，日本では，HTML(HyperText Markup Language) の知識のある人が，インターネット上に自分のウェブサイトを作り，その中でウェブ日記と呼ばれるものを残していた [26]．このウェブ日記とブログは親和性が高く，ブログが広く日本で普及している一因であると考えられる．また日本には，HTML などの知識がなくても，簡単に無料でブログを開設できるサービスを提供するブログサービス事業者も多数存在する．このことが，人々のブログ参入への敷居を下げ，多くの人がブログを利用するきっかけとなった．その結果，2006 年 3 月時点で日本のブログサービス事業者 53 社への登録者数は 868 万人[*4]存在する [27]．2007 年に発表されたのアメリカの Technolati 社の調査によると，世界中に存在するブログのうち日本語のブログが最も多い[*5]．さらに，日本ではインターネット利用者の 74% がブログを閲覧すると言われており [27]，この割合は韓国の 43%，米国の 27% と比較しても突出して高い．日本では，有名なスポーツ選手や芸能人などが個人のブログサイトを開設し，試合の結果から結婚や出産などの個人的な事項まで，ブログを使ってファンに直接発信している点を考えると，多数のブログ閲覧者が存在することは自然であろう．このように，日本においてブログは，ただの一個人の日記にとどまらず，時にはセンセーショナルなニュースリリースの場となったり，多くの人が身近に接する 21 世紀の新たな情報媒体であると言える．

[*4] ブログサイトは，基本的にはブログサービス事業者への申し込みフォーマットに記入するだけで簡単に開設できるため，必ずしもブロガー数が個人の数と対応しているとは限らない．

[*5] http://www.sifry.com/alerts/archives/000493.html (Accessed:2013.07.03)

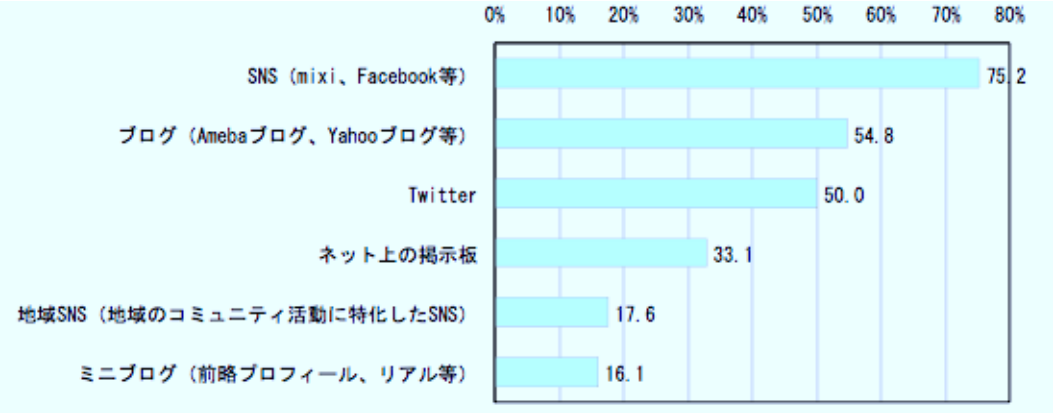


図 1.3　主なソーシャルメディアの利用内訳．アンケート対象者は $n = 1361$ 人で，ブログの利用割合は SNS に次いで高い．(総務省資料 [28] より抜粋．)

図 1.4 の上図は，ウェブブラウザを通して閲覧した実際のブログサイトである．ウェブブラウザを通じてみると，ブログは上図のように整形されて表示されている．しかし，ブログ記事のデータは下図のような HTML コードに似た，RSS(RDF Site Summary または Rich Site Summary) に代表される「フィード」によって記述されており，そのフィードを使ってブログ記事を容易に配信できる仕組みを持っている．ブログの読者は，あらかじめ，お気に入りのブログサイトのフィードを自分の RSS リーダーに登録しておくことで，各々のブログサイトを巡回しすることなく，ブログを読むことができる．主に XML(Extensible Markup Language) によって記述されるフィードは，ウェブサイトの内容を簡易に配信するために作られた仕組みで，情報の再利用が容易に行えるようにフォーマットが統一されている．この中にはブログ記事において，どこまでがブロガーが書いた本文なのか，記事がいつ発信されたかというタイムスタンプも含まれている．そのため，時間概念を伴った「生きた」テキストデータとしての利便性も高く，研究対象としても利用されている．本研究で扱うブログデータも，このフィードの仕組みを利用して大規模に収集されたものである [D].

**ブログに関する先行研究**

ブログは RSS フィードの仕組みの恩恵もあり，データとして収集が容易であることから，研究対象として，多方面より取り上げられている (図 1.5).

社会心理学において，多くの場合，ブログへの関心は「なぜ書くのか」「何を書いているのか」といったブロガー自身へと向けられることが多い [29]．調査の方法は，直接インタビュー [30, 31] や質問紙 [26, 29, 32, 33] を使った質的な研究が主である．こういった研究は歴史的には文化人類学に代表される民族学で発展し，エスノグラフィ (ethnography)




図 1.4　ブログサイトの例．赤枠部分が本文に相当する．(上図) ブラウザからみて整形されて表示されている場合と (下図) 同じページの HTML ソース．

研究と呼ばれるが，その研究の性質上，サンプル数は限られる．また，ブログに書き込まれた単語をコーパスとして捉え，特徴的な単語を分類することで，ブロガーの心理状態の変化を追うものもある [34, 35]．そして，ブロガーのサンプルを数多く集め，社会全体としての「幸福度」を定量化しようという試みもある [21]．

アメリカでの調査結果 [30] でも日本の調査結果 [32] でも，ブロガーがブログを開設して，継続している動機は，自己表現の場であったり，必要な情報を共有するためであったりと多様性がある．しかし，実際に書きこまれている内容はどちらも似ており，アメリカでも日本でも自分の知識や日常の記憶を整理するためであったと言われている [27, 35]．ブログ発祥の地アメリカでは，2001 年 9 月の同時多発テロ以降にブログが急速に普及したが，同時多発テロの後で，人々の感情がネガティブなものから元の状態へ戻るのには約 2 週間であったといった事例研究 [34] も 2004 年という比較的早い段階で行われている．

社会心理学者がブログに魅力を感じるのは，従来の質問紙調査による筆者 (ブロガー)

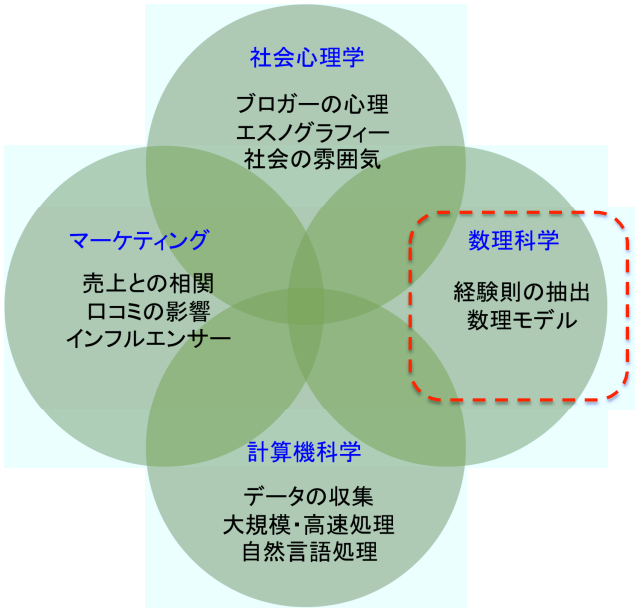


図 1.5 ブログに関する研究の概観．主に注目されるトピックを挙げている．

の「意識」と，実際のデータから得られる筆者の「行動」が同時に把握できるからである [29]．このような意識と行動の重層的なデータの入手は，インターネットによって初めて可能になったと言っても良い．人間の心理，行動をつぶさに観測したい社会心理学者にとってブログは貴重なデータの宝庫となっている．

マーケティングの分野ではインターネット登場以前から，消費者間での相互作用の結果生じる CGM(Consumer Generated Media) または口コミ (Word of Mouth, WOM) が，実際の商品の売り上げとどう関連するのかという研究が数多く行われてきた [47]．CGM で広告はどう評価されたのか，インフルエンサー，オピニオンリーダーやハブと呼ばれる，情報を拡散したり，影響力を持つ消費者は存在するのか，存在するとすればとどのくらい影響力を持っているのか [48]，といったことが議論されてきた．その延長線上で，ブログはオンライン上の口コミとして捉えられている．

インターネットの登場で，口コミの規模も大きくなり，商品のレビュー欄を筆頭に，口コミの形式も多様化した．こういったインターネット上の口コミはバズ (buzz) と呼ばれ，ブログもバズの一つとして注目されている．広告に代表される従来の伝統的なマーケティング手法と，バズの効果の違いは，その即効性と持続性という点でバズの方が優位であると言われている [49]．

ブログをはじめとするインターネット上に残った情報が，実世界の情報を反映すること

も多く，ブログの書き込み数変化から，実際のアーティストの楽曲売上や映画の興行収入を予測するという試みもなされている [23, 24, 25, 50].

ブログに限らず，データの中から自動的にいち早く異常値を検出することは，計算機科学の分野では盛んに行われてきた．その中で，ブログの世界においてそれらの手法が応用されるのは自然な流れであろう．オンラインニュースなどのデータストリームから，話題を自動的に抜き出すことは *Topic Detection and Tracking* (TDT)[36] と呼ばれ，ブログの普及以前から行われてきている．また，ブログに限らないテキストから新たな知見を発見するテキストマイニングの研究も盛んである．計算機の進化を背景にして，大規模にデータを収集することや [37]，バースト検出などの広く実用的な部分 [38, 39] で次々と成果を上げている．

計算機科学の分野ではブログのネットワークに注目した研究も多い．ブログには，トラックバックという機能があり，互いに引用し合うこともあり，その機能を使いネットワークを描画することも可能である．2004 年のアメリカ大統領選において，保守派とリベラル派の政治的なブログでどう違いがあったのかだけではなく，どう引用関係が起こったのか [40]，といった今までになかった研究がブログデータの出現により盛んに行われるようになっている．他にもネットワークの描画技術の普及もあり，ブログのネットワーク上や，その他 Twitter などのソーシャルメディアの上でどう情報カスケードが起きているのかを直接観測 [41, 42, 43, 44, 45, 46] する試みもある．

数理科学の分野では，ブログを人間の「集団」として捉え，その中に数理的パターンや普遍性を探索する傾向がある．本博士論文もこの数理科学の立ち位置に属する．例えば，大きな本震の後には，余震の発生する頻度が，本震からの経過時間からベキ関数的に減少していく「大森則」が知られているが [51]，この大森則と似た現象をブログの中でも指摘し，大きなイベントの後の書き込み数の減少がベキ関数的になることを報告したもの [C][52] もある．ブログの世界の中で，一気に情報が拡散する現象を，砂山の雪崩と同じように自己組織化と捉えて理解するものもある [53]．また，単語の出現頻度だけではなく，バースト性といわれる個人の行動に見られる非自明なパターンは，ブログ投稿パターンでも存在が指摘されている [A][54].

## 1.2 目的と構成

膨大で精緻なデータがあり，社会的な需要が高まっているにもかかわらず，インターネットの世界を読み解く方法に確立されたものは存在しない．その一方，高度に洗練された大衆受けするウェブサービスを投入する競争は熾烈で，今世紀のゴールドラッシュのような状況である．本研究では，広く応用可能な時系列データに関する解析手法を開発し，それを大規模ブログデータに適用し，人々の普遍的な行動パターンをモデル化して理解す

表 1.1　本研究で主に扱うブログにおける単語出現頻度時系列の分類.

| 分類 | 単語の代表例 | 主な特徴 | 扱う章 |
|---|---|---|---|
| 日常語 | 「また」「漸近線」 | ランダム投稿モデル | 4 章 |
| イベント語 | 「エイプリルフール」「海の日」 | ベキ関数 (ピーク前／後) | 5 章 |
| 日付 | 「10 月 18 日」「6 月 21 日」 | ベキ関数 (ピーク前／後) | 5 章 |
| ニュース語 | 「マイケル・ジャクソン」「津波」 | ベキ関数 (後) | 5 章 |
| 流行語 | 「KY」「Twitter」 | 指数関数 (主にピーク前) | 4 章の一部 |
| 季節流行語 | 「みかん」「ひまわり」 | 指数関数 (ピーク前／後) | 4 章の一部 |

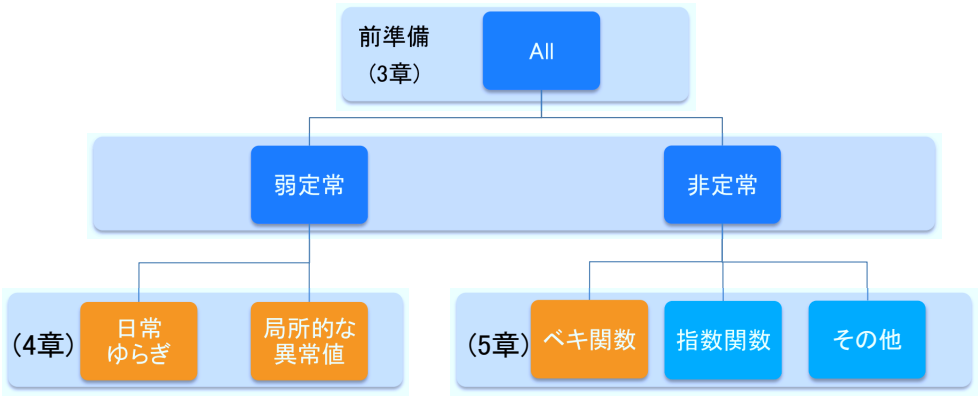


図 1.6　本研究の構成と主に取り上げる章.

ることを目的とする.

本研究の構成と結果は以下である. ブログにおける単語出現頻度時系列を扱った手順と対応する章を図 1.6, その結果の分類を表 1.1 に示した.

第 2 章でデータの具体的な説明をし, 行った必要な前準備について第 3 章で述べる. 次

の第 4 章と 5 章で，ブログの世界での単語の出現頻度の時系列を使った解析とモデル化を行う．それぞれの章では，注目する単語の定義や検索方法，解析手法と結果，その応用例を提示してまとめる．

第 4 章では，まず最初に「また」「そして」などのように日常的に使われる単語をここでは「日常語」と呼び，その統計的特性に注目する．ブログのみならず，インターネット上の単語の出現頻度に注目する場合，新たに生まれた流行語や注目のニュースに関連する語など明らかに非定常なもののみを扱う場合が多い．しかし，本研究では日常語からスタートして，その単語出現頻度のゆらぎに基づいて異常値を定義していくことが，最大の特徴である．日常語の出現頻度時系列における平均値と標準偏差の関係には，非自明なスケーリング則があり，それを再現するランダム投稿モデルを導入する．ランダム投稿モデルを元に，時系列からの局所的な異常値検出手法について説明する．

第 5 章では，第 4 章で検出した異常値の中から，日常とは逆にニュースや季節変動などに影響され，非日常的に使われる単語に注目する．非日常的に使われる単語は，地震などの突発的なニュースに関連する語から，季節に応じて緩やかなトレンドを持つ語まで様々であるが，第 5 章ではその中でも，時系列中にベキ関数的な変動を含むものに注目する．ベキ関数的な変動を含む単語は，「締め切り」を持つことが最大の特徴である．締め切りとは，「エイプリルフール」が毎年 4 月 1 日にあるように，あらかじめ与えられた日付を指す．そして，締め切りがある場合，ブログ上の時系列はその前後で数理的にベキ関数的な変動をする．また，地震や訃報などの突発的なニュースは，ニュース直後にブログへの書き込み数は跳ね上がり，その後，ベキ関数的に減少する．全く違う単語でも，ベキ関数的変動という点では共通で，その背後にあるメカニズムは何かを数理モデルを導入し，議論する．最後に，本研究のまとめを行い，将来の研究展望について述べる．

# 第 2 章

# データの説明

本章では，本研究で使う用語を説明し，扱ったブログデータの詳細を説明する．

## 2.1 用語の説明

はじめに，本研究で多用する用語を説明する．インターネットの普及と同時に新たに使われるようになった用語は数多く存在する．それらは主に英語での用語をそのままカタカナに直したものであり，必ずしも明確な定義が存在するわけではない．そこで，混乱を防ぐため，主な用語を列挙し，本研究中での使い方の定義をしておく．

- **インターネット (internet)** ルータを介した物理的な IP アドレスのつながり．基本的に双方向へ自由に行き来できるため，無向ネットワークになっている．
- **ウェブ (web)** インターネット上に存在するサイト (ページ) のつながり．IP アドレスで物理的につながっているからと言って，リンクがある訳ではない．意図しないとリンクが生成されないため，有向ネットワークになっている．
- **ソーシャルメディア (social media)** 利用者個人が情報を発信し，形成していくメディア．利用プラットフォームはインターネットに限ってはいないが，ローカルなケーブルテレビや FM 局[*1]を除くと，ほぼすべてのアプリケーションがインターネットをプラットフォームとしている．ソーシャルメディアの利用者は，10 代から 30 代までの若年層程高い．また若年層ほど，ブログと SNS など同時に使う複数利用の割合も高い [28]．2012 年時点での主なソーシャルメディアを表 2.1 に挙げた．
- **ブログ (blog)** ソーシャルメディアの一つ．個人がウェブ上に自由に開設し，そこに文章や写真を記事として投稿できるサイト．ウェブ (web) とログ (log) を組み合

[*1] 総務省の資料 [55] によると，地域の商業，行政情報や地元情報に特化し，放送エリアが地域 (市町村単位) に限定されるコミュニティ放送もソーシャルメディアに含まれている．

表 2.1　2012 年時点での主なソーシャルメディア．

| 分類 | 代表例 | 主な特徴 |
|---|---|---|
| ブログ | アメブロ，livedoor ブログ | 日記などの文章 |
| SNS | Facebook，mixi | 社会ネットワーク |
| 動画共有サイト | YouTube，ニコニコ動画 | 撮影した動画 |
| 写真共有サイト | Flicker，Picasa | 撮影した写真 |
| 情報共有サイト | Wikipedia，クックパッド | レシピなどの情報 |
| マイクロブログ | Twitter，mixi ボイス | ブログよりも短い文 |
| 掲示板 | 2 ちゃんねる，Yahoo!知恵袋 | Q&A など |
| ソーシャルゲーム | GREE，モバゲー | ウェブ上で行うゲーム |
| レビューサイト | 価格.com，食べログ | 商品レビュー |
| その他 | Forsquare，Amazon.com | 現在地など |

わせた造語としてウェブログができ，それを省略したブログ (blog) の名で一般に普及した．ブログの作者のことをブロガー (blogger) という．本研究はこのブログを対象にする．

- **ソーシャルネットワークワービス (social networking service, SNS)** ソーシャルメディアの一つ．個人が，ウェブ上に構築されたネットワーク内に自分の ID を登録し，他人と交流し合うサービス．代表例として世界中で利用者が 10 億人を突破した[*2]Facebook，日本で最大の規模を持つ mixi がある．
- **Twitter** ソーシャルメディアの一つ．アメリカの Twitter 社が独自に運営している．「ツイート」と呼ばれる文字数が 140 文字に制限された文章を投稿するブログのような仕組みで，ミニブログやマイクロブログ (microblogging) と分類されることもある．使用者間で参照し合う仕組み（フォロー／フォロワー）が明示されているという点でネットワーク的な側面も強く，SNS の 1 つと言われることもあるが，同社はそれを否定し，より広義な情報ネットワーク (information network) であるとしている[*3]．

[*2] 2012 年 10 月 4 日，創業者の Mark Zuckerberg が自身の Facebook のタイムラインで報告した．http://money.cnn.com/2012/10/04/technology/facebook-billion-users/index.html (Accessed:2013.07.03)

[*3] http://news.cnet.com/8301-19882_3-20112261-250/twitters-not-a-social-network/ (Accessed:2013.07.03)

## 2.2 データの説明

本研究では，株式会社ホットリンク[*4](以下，ホットリンク) の所有するデータベースに保存されたブログデータを用いる．ホットリンクでは，2006 年 11 月より 20 を超える日本の主要なブログプロバイダー (詳細は付録 A に記載)，２ちゃんねるや Yahoo!掲示板などの国内ウェブサイトへ書き込まれた公開データを収集し，データベース化して，検索ツール（口コミ@係長[*5]，電通バズリサーチ[*6]）として有料で顧客に提供している (図 2.2, 図 2.3)．検索ツールは，機能拡張を続けており，詳細設定で検索対象媒体や期間を指定できるだけではなく，書き込み内容からポジティブ，ネガティブの割合を判定したり，男女比を判定したり，共起語を絞り込んで出力することも可能である．さらに，近年はウェブへの書き込みデータだけではなく，テレビへの露出回数や，ウェブ上のニュースへの出現回数も合わせて集計することができる機能も追加された．ホットリンクの顧客は，主にウェブの書き込みを自社の広告効果測定や，新商品の評判分析等のマーケティング支援に使っている．

データの規模は，2013 年 5 月現在，ブログだけで 3200 万ブロガーから投稿された約 26 億 7000 万記事 (図 2.1)，掲示板などの書き込みを合わせると約 89 億記事を有する．

同様にブログを集約して顧客に提供するサービスは，国内外にも存在する．国内では代表的なものに株式会社きざしカンパニーが運営する「きざし[*7]」があるが，対象とできるのは 2013 年 6 月現在，約 1 億 6000 万記事で，ホットリンクのものよりは規模が小さい．海外ではアメリカにおいてブログ検索システムは盛んに商品化されており，代表的なものに Tailrank.com の運営する Spinn3r[*8]，テクノラティ (Technorati) 社が提供するブログ検索[*9]，市場調査会社であるニールセン (Nielsen) が提供する「BlogPulse[*10]」がある．ホットリンクでもアメリカのソーシャルメディアデータ供給会社である Gnip 社[*11]などとの業務提携を始めているが[*12]，日本語の大規模なブログデータを研究対象とするためには，ホットリンクのデータベースは最適であると言える．

ホットリンクのブログベータベースには，投稿された記事のタイムスタンプ，URL，記事の表題と本文が記録されており，ブログの URL を分析することで，ブロガーの区別や，

[*4] http://www.hottolink.co.jp
[*5] http://kakaricho.jp/
[*6] http://dbuzz.jp/
[*7] http://kizasi.jp/
[*8] http://spinn3r.com/
[*9] http://technorati.com/
[*10] 2012 年 1 月に終了．
[*11] http://gnip.com/
[*12] http://www.hottolink.co.jp/press/3727 (Accessed:2013.06.10)

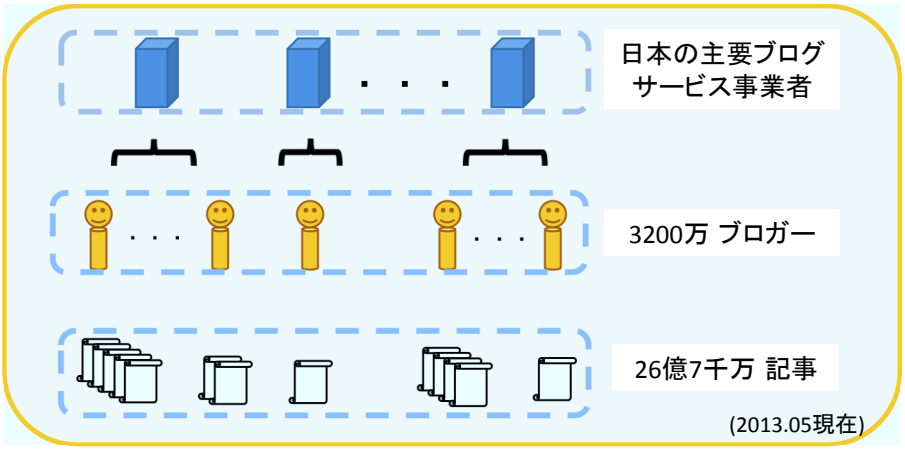


図 2.1　本研究で扱うブログデータの概観.


図 2.2　ホットリンクによる検索ツールの検索画面.

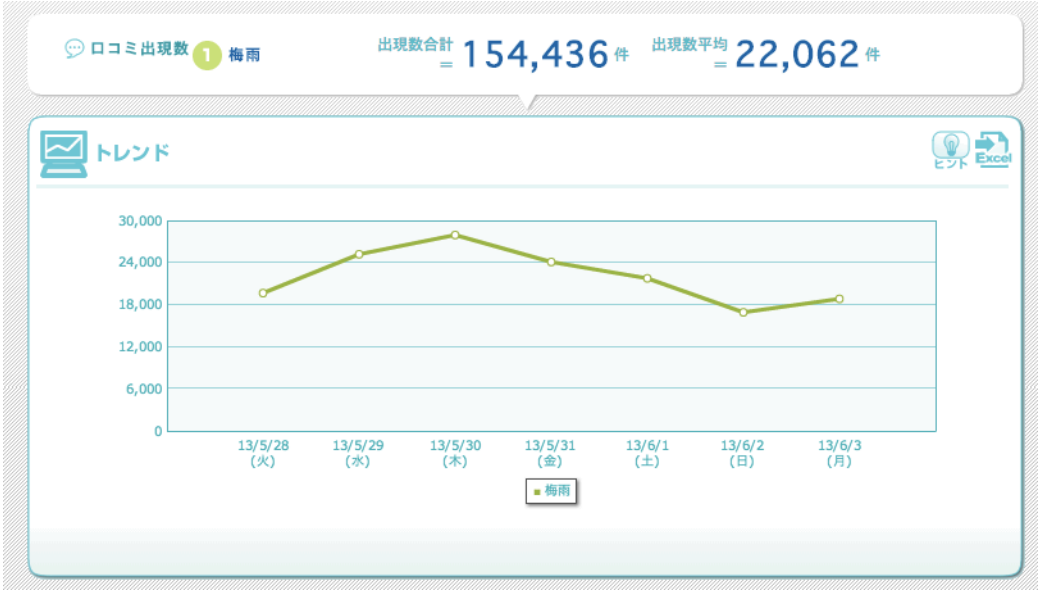


図 2.3　ホットリンクによる検察ツールの検索結果画面.

ブログサービス事業者の区別を付けることができる．ホットリンクでは 2009 年より有料で API(Application Programing Interface)[*13]機能を公開しており，任意の検索語を含む 1 日刻みのブログ数を容易に得ることができる．本研究でもこの API を使用して，検索単語を含む日ごとのブログ記事を「書き込み」と呼び，その出現頻度時系列を抽出している．

本研究で扱うデータは以下の二つである．

**ブログデータ 1**

本研究で主に扱うデータで，検索語を含む書き込み (ブログ記事) 数の 1 日刻みの出現頻度時系列．検索単語を含む書き込み数は，データベースに保存されている記事の本文中か記事の表題に検索単語を一度でも含むかどうかで決定している．そのため，同一記事に複数回，検索単語を含んでいても，1 と数える．他方，同一ブロガーが，同日に，検索単語を含む書き込みを複数投稿した場合は，別々に検索対象となるため，複数に数え上げる．タイムスタンプやブログ URL は検索対象にならない一方，記事の本文中に直接書き込まれている URL や，日付は検索対象になる．また，検索単語の全角，半角単語の区別はない．そのため半角文字で「AKB48」と検索しても全角文字で「ＡＫＢ４８」としても同じ結果となる．

1 日の区切りはタイムスタンプ通りの午前 0 時からの 24 時間であり，人の概日周期と

[*13] 利用者が簡単な手続きで，必要な機能を呼び出して扱うことが出来る．

は必ずしも一致していない．ブロガーの投稿周期は深夜 23 時にピークを迎える (詳細は第 3.3 章に記載)．このブロガーの投稿周期は，第 4 章で扱う毎日ほぼ同程度で現れる単語の場合には影響しないが，第 5 章で扱う 1 日単位で書き込み数が大きく変化する場合には影響を受ける．ブログデータ 1 に関し，ブロガーの投稿周期に対して必要な処理については第 3.3 章で説明する．

**ブログデータ 2**

ランダムサンプルした匿名化されたブロガー約 33 万人分の全投稿記事を含むデータ．ブロガー個人の行動履歴が分かる詳細データが必要な場合，公開 API では取得できないため，ホットリンクに依頼して，別途データベースにアクセスして取得したものを扱っている．詳細データには，ユーザ ID と記事本文が含まれるため，いつ，誰が，どういった記事を投稿したかの情報が得られる．このデータベースに対して検索をかける際には，unix の grep コマンドを用いて，パターンマッチしたかどうかで検索しているため，全角文字と半角文字の区別がある．

# 第 3 章

# データ処理の前準備

本章では，本研究においてブログデータ解析の前に行った処理について説明する．ブログデータは，実験環境が厳密に管理された元で得られるデータとは異なり，スパムなど外的なシステムに依存した様々なノイズを含んでいる．そこで，ブログにおける単語出現頻度時系列が本来持つ，ゆらぎや動的な成分を抽出するため，考えられるノイズを除去しておくことは，次章以降のデータ解析にとって不可欠な行程である．

## 3.1 スパムブログの除去

2006 年の調査[*1]で，世界中の言語の中で，最もシェアが大きいと指摘された日本語のブログであるが，同時に，スパムブログが多く存在することが知られている．スパムブログはスパム (spam) とブログ (blog) を組み合わせてスプログ (splog) とも呼ばれるが，ここでは単純にスパムと呼ぶ．ニフティ研究所が 2007 年 10 月から 2008 年 2 月の間に日本国内のブログ記事を，各月 10 万件ずつサンプリングして調べた結果，約 40% のブログがスパムだったと報告している[*2]．

スパムの目的は，アフェリエイト (affiliate) と呼ばれる成功報酬型広告の表示や，特定サイトへの誘導である．アフェリエイトはブログで本や食品などの商品を紹介し，最後に実物の購入を促し，成功すれば広告主がブロガーに報酬を支払う仕組みである．この時，実物の購入はなくとも，何らかの検索結果で偶然，ページが表示されただけでも，一定の閲覧 (ページビュー) があったとカウントされ，ブロガーに広告主から代金が支払われる．この仕組みを悪用したブロガーは，なるべく検索されやすい単語をちりばめたブログを大

[*1] http://www.sifry.com/alerts/archives/000493.html (Accessed: 2013.07.03)
[*2] http://gigazine.net/news/20080326_spam_blog/ (Accessed: 2013.04.30)

量に作成する．その結果，短期間に大量のスパムが発生し，ブログの検索にも悪影響を及ぼす．

ブログポータル事業者にとっては，スパムはシステムのリソースを無為に消費されるだけではなく，ブログポータルのブランドイメージを損ねる可能性もある．そこでスパムと認められた場合，通常，ブログポータル事業者はブログを強制的に閉鎖するなどの処置を取る．しかし，スパム作成者は，次々に新しいブログを開設しては，閉鎖を繰り返すため，結果として大量のスパムが，日本語のブログの世界に混在している．

コンピュータウィルスが，ウィルス対策ソフトから逃れて次々にバージョンアップするように，スパムも巧妙化しており，絶対的な排除方法はない．スパムの種類も，検索されやすい人名や商品名を組み合わせただけの「ワードサラダ」と呼ばれる，スパムと見分けやすいものから，他の情報ソースの内容の一部ずつをコピーしてつなげた「コピースパム」と呼ばれる，通常のブログに見間違うものまで多数存在する．

それらに対応する方法として，スパムのパターンを学習させた分析器を用いる機械学習による検出，類似投稿との比較検出，辞書によるフィルタリングなどそれぞれのスパムに対して必要な処置を取る．これらの処理をまとめたものがスパムフィルタ [D][56, 57, 58, 59] である．スパムフィルタには，スパム除去の処理を論理和で結んだカスケード接続や，並行に処理を行うパラレル接続がある．

スパムフィルタの評価には，適合率 ($precision$) と再現率 ($recall$) を用いるのが一般的である [59]．適合率は，スパムと判定されたものの中に，どれくらい本当にスパムを含むかを表す比率である．再現率は，検出すべきスパムのうち，実際には何件がスパムと判定されたかを表す比率である．例えば，200 件中に 80 件のスパムを含むベンチマーク用データがあったとする．フィルタ A ではスパムとして 50 件抽出し，50 件中すべてがスパムであった．フィルタ B ではスパムとして 150 件抽出し，そのうち 80 件がスパムであった．このとき，フィルタ A の適合率は $p_A = 50/50 = 1$，再現率は $r_A = 50/80 = 0.63$ となる．フィルタ B の適合率は $p_B = 80/150 = 0.53$，再現率は $r_B = 80/80 = 1$ となる．

適合率が高ければスパムをスパム，一般のブログはスパムではないと判定する．しかし，検出すべきスパムを見逃す可能性が出て，再現率と言う点では下がる．逆に，適合率が低ければ，相対的に再現率は向上するが，スパムと判定したものの中に，一般のブログが混入している可能性が高い．すなわち，適合率と再現率にはトレードオフの関係があるため，それらを調和平均した $F$ 値が使われる．

$$F\text{ 値} = \left[\frac{1}{2}\left(\frac{1}{precision} + \frac{1}{recall}\right)\right]^{-1} = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{3.1}$$

この $F$ 値が高い方が，バランスがとれた精度の良いフィルタであると判定するのが一般的である．先ほどの例の場合，フィルタ $A$ の場合は $F_A = 0.77$，フィルタ $B$ の場合は

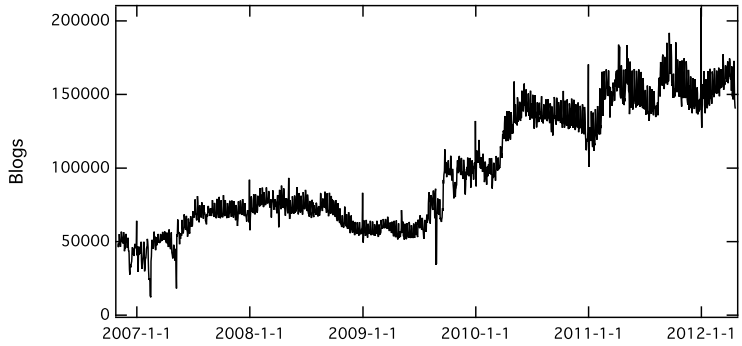


図 3.1　「また」の書き込み時系列 $w^{(\text{また})(t)}$．(規格化前)

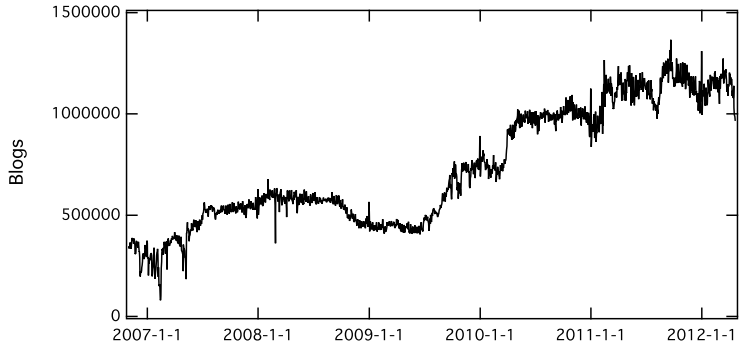


図 3.2　2006 年 11 月から 2012 年 4 月までの全数の時系列 $w(t)$.

$F_B = 0.69$ となるので，フィルタ A の方が高精度となる．

本研究においては，スパムの除去にはホットリンクの提供するスパムフィルタを用いる．スパムフィルタは検索画面 (図 2.2) と API 両方で設定ができる．ホットリンクの提供するスパムフィルタは，段階的に適合率の高い「強」，中程度の「中」，再現率の高い「弱」を提供している．本研究では，適合率と再現率のトレードオフの関係をうまく調整して，スパムの検出がバランスよく行われている「中」のスパムフィルターを用いてデータを取得している．

## 3.2　全数による規格化

### 3.2.1　全数除算による規格化

スパムを排除したとはいえ，厳密に管理された実験環境から得られるデータとは異なり，本研究で扱う観察者がコントロールできない社会現象から得られるデータは，他にも

様々なノイズを内包している．図 3.1 は接続詞「また」の書き込み時系列 $w^{(\text{また})}(t)$ である．図 3.2 は，単語にはよらない全投稿数 (以下，「全数」と呼ぶ) の時系列 $w(t)$ である．$w^{(\text{また})}(t)$ と $w(t)$ は，2009 年以降に全体的に数が増えており，「また」という単語の書き込みが増えていた理由は，全数の影響であると考えられる．実際に $w(t)$ が増えた時期は，ホットリンクがデータ収集対象とするブログプロバイダーを増やしたことと対応している．そこで，個々の単語の書き込みに注目したい場合は，時系列から全数からの影響を取り除く必要がある．

個々の時系列から，外的要因に起因するゆらぎ (全数のゆらぎ) と，内的要因に起因するゆらぎ (単語ごとに内包するゆらぎ) を切り分ける手法として，de Menezes と Barabási の手法が知られている [60]．彼らの手法は，全体 $w(t)$ が重複の無い $N$ 個の小さな部分 $k$ の時系列 $w^{(k)}(t)$ の総和からできていると仮定した上で，$k$ が全体に占める割合 $A^{(k)}$ を以下で定義した．

$$A^{(k)} = \frac{\sum_{t=1}^{T} w^{(k)}(t)}{\sum_{t=1}^{T}\sum_{k=1}^{N} w^{(k)}(t)} \tag{3.2}$$

ここで時系列の範囲は時刻 $t \in [1, T]$ である．時刻 $t$ において，部分 $k$ の外的要因に起因する部分 $w_{\text{ext}}^{(k)}(t)$ は，全体 $\sum_{k=1}^{N} w^{(k)}(t)$ に線形に比例するとして

$$w_{\text{ext}}^{(k)}(t) = A^{(k)} \sum_{k=1}^{N} w^{(k)}(t) \tag{3.3}$$

として定義し，内的要因に起因する部分 $w_{\text{int}}^{(k)}(t)$ は，$w^{(k)}(t)$ より $w_{\text{ext}}^{(k)}(t)$ を減算することで得られる．

$$w_{\text{int}}^{(k)}(t) = w^{(k)}(t) - w_{\text{ext}}^{(k)}(t) \tag{3.4}$$

彼らは，切り分けたそれぞれの $w_{\text{ext}}^{(k)}(t)$ と $w_{\text{int}}^{(k)}(t)$ の時系列の標準偏差 $\sigma_{\text{ext}}^{(k)}$，$\sigma_{\text{int}}^{(k)}$ より，それらの比 $\eta^{(k)} = \sigma_{\text{ext}}^{(k)} / \sigma_{\text{int}}^{(k)}$ を比較することで，外的要因が支配的になっている時系列かどうかを判定する手法を提案した．具体的には $\eta^{(k)} \gg 1$ であれば，外的要因が支配的，$\eta^{(k)} \ll 1$ であれば内的要因が支配的であると判断する．コロラドの高速道路の交通量の時系列，アメリカの河川の流量時系列，インターネット上のルーターを通過するパケットの流量の時系列，マイクロプロセッサ回路上の電流量などに適用し，高速道路や河川の場合は外的要因が支配的であるが，マイクロプロセッサやインターネットの場合は内的要因が支配的であることを示した．

彼らの手法は単純で分かりやすく，非定常性を含む時系列に幅広く適用が期待できる．しかし，本研究で扱うブログ時系列に対しては直接適用することができない．なぜなら彼らの手法では，全体を「重複のない」$N$ 個の部分の集合だと仮定している．一方，ブログ上の単語出現頻度時系列は，明らかに重複があり $N$ を定義できないためである．その結

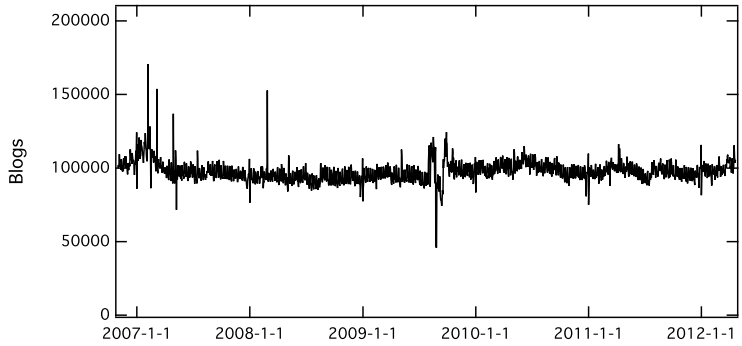


図 3.3　「また」の書き込み時系列 $w'^{(\text{また})}(t)$．(規格化後)　式 (3.5) を用いて規格化した場合．

果，式 (3.2) の $A^{(k)}$ を定義することもできない．ブログ時系列の場合，一つの記事の中で複数の単語が含まれており，考えられるすべての単語の時系列 $w^{(k)}(t)$ を数え上げたとすると，全数の時系列 $w(t)$ を上回る．$(w(t) \ll \sum_k^{\text{全ての単語}} w^{(k)}(t))$

そこで本研究では，以下のような規格化を考える．単語 $k$ の書き込み時系列 $w^{(k)}(t)$ に対し，単語によらない全投稿数時系列 $w(t)$ で除算し，規格化した時系列 $w'^{(k)}(t)$ は以下の式で定義する．

$$w'^{(k)}(t) = \frac{w^{(k)}(t)}{w(t)}\langle w \rangle \tag{3.5}$$

ここで $\langle w \rangle$ は，全期間 $T$ での $w(t)$ の平均値 $\langle w \rangle = \frac{1}{T}\sum_{t=1}^{T} w(t)$ であり，規格化の前後で $w^{(k)}(t)$ の数が保存されるように乗算している．図 3.3 は，式 (3.5) を用いて規格化した「また」を含む時系列である．規格化を行うと，時期によらず平均値の周りで出現頻度が揺らいでいることが分かる．以下，本研究では，特に断りのない限り式 (3.5) で規格化した時系列を扱う．

本研究で用いた規格化は全体に対する単語 $k$ の書き込み割合の時系列に直すことであり，単語そのものの変動の特徴を捉えることができる．このような処理は，検索クエリを使ったトレンドなどの抽出でも行われている．

### 3.2.2　全数との相関

図 3.4 は，1 日に約 1 件しか使われない単語「漸近線」の書き込み時系列である．図 3.1 で確認した，1 日に約 10 万件使われる単語「また」と比較すると，「漸近線」の時系列は全数の増減にはほとんど影響を受けず，全数の変動とは無関係に見える．そこで，全数 $w(t)$

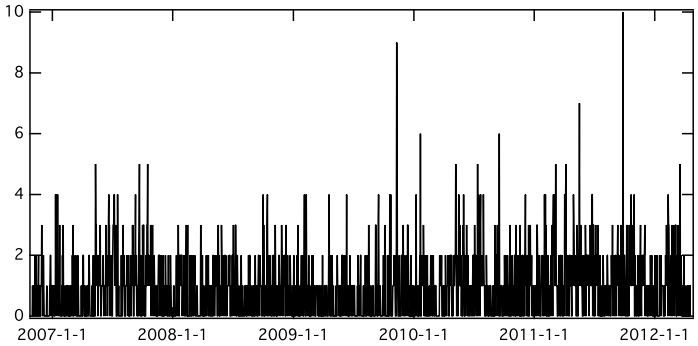


図 3.4　「漸近線」の書き込み時系列 $w^{(\text{漸近線})}(t)$．(規格化前)

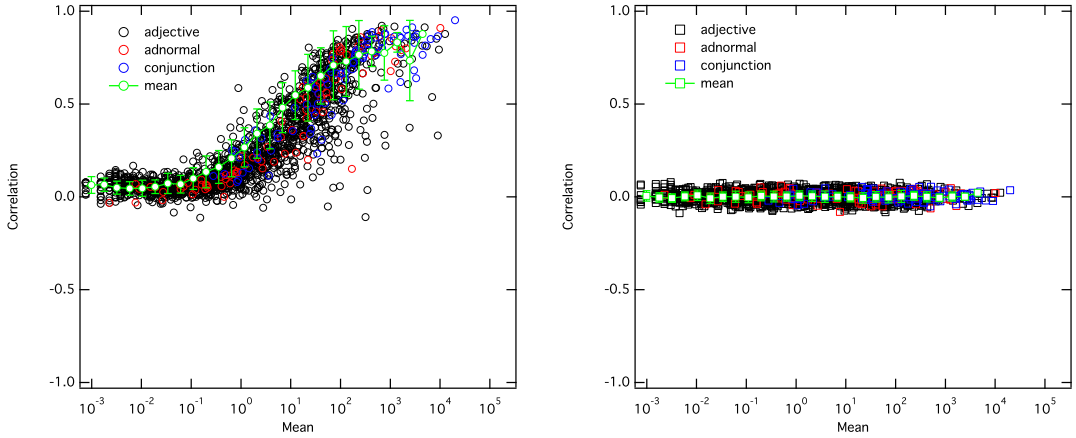


図 3.5　横軸に単語 $k$ の平均値 $\langle w^{(k)} \rangle$，縦軸に全数との相関 $\rho_{ww^{(k)}}$ を示した．黒印が形容詞 (1526 単語)，赤印が連体詞 (94 単語)，青印が接続詞 (92 単語)，白抜きの緑印が全体の平均値．平均値のばらつきが大きいため，横軸は対数スケールで表示している．(左図) 実際の時系列 $w^{(k)}(t)$ では，$\langle w^{(k)} \rangle$ が大きくなると，平均値に対して相関係数 $\rho_{ww^{(k)}}$ が非線形に増え，1 へと収束する．(右図)$w^{(k)}(t)$ において時刻 $t$ をランダムシャッフルした時系列の場合．平均値に対して全数との変動は無相関となる．

と単語 $k$ を含む時系列 $w^{(k)}(t)$ のピアソンの積率相関係数 $\rho_{ww^{(k)}}$ を以下で算出する．

$$\rho_{ww^{(k)}} = \frac{\langle (w(t) - \langle w \rangle) \left( w^{(k)}(t) - \langle w^{(k)} \rangle \right) \rangle}{\sigma_w \sigma_{w^{(k)}}} \tag{3.6}$$

分子の $\langle \ldots \rangle$ は，$w$ と $w^{(k)}$ の共分散 (covariance) である．$\sigma_w$ は全数時系列 $w(t)$ の標準偏差，$\sigma_w^{(k)}$ は単語 $k$ の時系列 $w^{(k)}$ の標準偏差である．期間は 2006 年 11 月 1 日からの 1317 日間を対象とした．「また」を含む時系列の平均値は $\langle w^{(\text{また})} \rangle = 72305$ で，相関係数は $\rho_{ww^{(\text{また})}} = 0.97$ である．他方，平均値が小さくなると，無相関に収束し，「漸近線」の場合は平均値は $\langle w^{(\text{漸近線})} \rangle = 0.94$ で，相関は $\rho_{ww^{(\text{漸近線})}} = 0.02$ である．

図 3.5 の左図は出現頻度時系列が動力学的な成分を持たず，平均値の周りで揺らぐ「日常語」(詳細は 4 章で定義する) の形容詞，連体詞，接続詞において，横軸に時系列の平均値 $\langle w^{(k)} \rangle$，縦軸に式 (3.6) の相関係数 $\rho_{ww^{(k)}}$ を示した．相関係数は，平均値が増えるにつれ増大し，$\langle w^{(k)} \rangle \sim \langle w \rangle$ の極限では $\rho_{ww^{(k)}} \to 1$ へと収束する．$\langle w^{(k)} \rangle$ が大きくなるとき，相関係数が 1 へと収束する様子は，平均値に対して線形にはらない．図 3.5 の右図は，比較のため，単語 $k$ を含む時系列において $w^{(k)}(t)$ の時刻 $t$ をランダムシャッフルした時系列と，実際の全数 $w(t)$ の時系列との相関係数を示した．図から明らかなように，全数の変動とランダムシャッフルした個別の単語出現頻度時系列は相関を持たない．

渡邊ら [61] により，全数の変動 $w(t)$ と単語 $k$ の書き込み時系列 $w^{(k)}$ の相関は，平均値が 1 になるように規格化した全数の変動 $\overline{w(t)} = w(t)/\langle w \rangle$ を，動力学的に変動する部分 $y(t)$ とゆらぎ部分 $\epsilon(t)$ に分けて考えることで記述できることが，以下のように解析的に示されている．

単語 $k$ の時系列 $w^{(k)}(t)$ の平均値は定常で，定数 $\langle w^{(k)} \rangle$ とすると，分散 $V[w^{(k)}]$ は，

$$V[w^{(k)}] = \langle \left( w^{(k)} \right)^2 \rangle - \langle w^{(k)} \rangle^2 \tag{3.7}$$

$$\simeq \langle w^{(k)} \rangle \left\{ 1 + \langle w^{(k)} \rangle (V[y] + V[\epsilon]) \right\} \tag{3.8}$$

と記述することができる．ここで，$\overline{w(t)} = y(t) + \epsilon(t)$ であり，$V[y]$ は $y(t)$ の時系列の分散 $V[y] = \langle y^2 \rangle - \langle y \rangle^2$，$V[\epsilon]$ は $\epsilon(t)$ の時系列の分散 $V[\epsilon] = \langle \epsilon^2 \rangle - \langle \epsilon \rangle^2$ である．ただし $V[\epsilon]$ は $V[y]$ と比較すると非常に小さいので，この項を無視すると，相関係数 $\rho_{ww^{(k)}}$ は以下のように記述することができる．

$$\rho_{ww^{(k)}} \simeq \langle w^{(k)} \rangle \sqrt{\frac{V[y]}{\langle w^{(k)} \rangle \left( 1 + \langle w^{(k)} \rangle V[y] \right)}} \tag{3.9}$$

もし，$\langle w^{(k)} \rangle$ が大きい場合は，$1 \ll \langle w^{(k)} \rangle V[y]$ より $\rho_{ww^{(k)}}$ は 1 へと漸近する．逆に $\langle w^{(k)} \rangle$ が小さく，0 に漸近する場合には，$\rho_{ww^{(k)}}$ も 0 へと漸近する (図 3.6)．

図 3.5 の左図から確認できるように，単語出現頻度時系列 $w^{(k)}(t)$ は，全数の時系列 $w(t)$ に対して非自明な正の相関を持っている．しかし，式 (3.5) で導入した規格化を行った後の時系列と，全数の時系列の間には緩やかな負の相関が現れる．式 (3.5) で規格化した時系列 $w'^{(k)}(t)$ と全数 $w(t)$ との相関係数を示したものが図 3.7 の左図である．図 3.5 の場合とは逆に，平均 $\langle w'^{(k)} \rangle$ が大きくなると相関係数 $\rho_{ww'^{(k)}}$ は負数へ偏ることが分かる．

例えば，1 日の出現頻度が低い「漸近線」のような単語の場合，全数 $w(t)$ で除算すると，全数の急激に減った日付では，規格化した後の単語 $k$ の出現頻度 $w'^{(k)}(t)$ が増えることもあり得ることからも明らかである．図 3.8 は $w^{(\text{漸近線})}(t)$ を式 (3.5) で規格化した場合

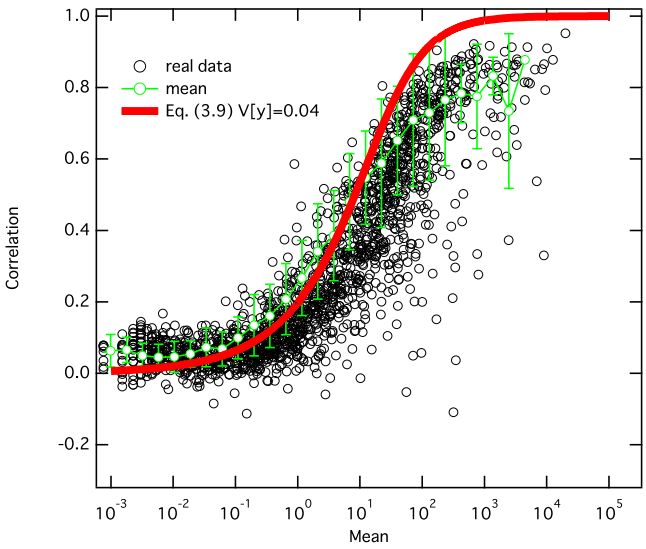


図 3.6　渡邊ら [61] による式 (3.9) のモデルによる $V[y] = 0.04$ の理論線 (赤) と実データ (黒) の比較.

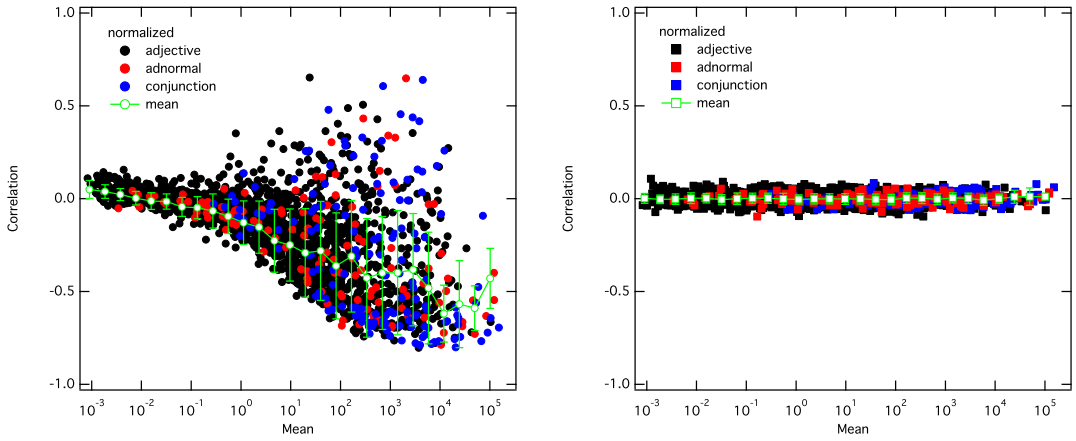


図 3.7　横軸に規格化した単語 $k$ の平均値 $\langle w'^{(k)} \rangle$，縦軸に全数との相関 $\rho_{ww'^{(k)}}$ を示した．黒印が形容詞，赤印が連体詞，青印が接続詞，白抜きの緑印が全体の平均値．(左図) 実際の時系列 $w'^{(k)}(t)$ では $\langle w'^{(k)} \rangle$ が大きくなると，$\rho_{ww'^{(k)}}$ は負数の絶対値が大きい方へと傾く．(右図)$w'^{(k)}(t)$ において $t$ をランダムシャッフルした時系列の場合.

の結果である．2007 年から 2009 年までの全数が少ない期間は元の時系列 (図 3.4) と比べると，微増している.

そのため，式 (3.6) の相関係数で重み付けして，

$$w''^{(k)}(t) = \rho_{ww^{(k)}} \frac{w^{(k)}(t)}{w(t)} \langle w \rangle + (1 - \rho_{ww^{(k)}})\, w^{(k)} \tag{3.10}$$

として規格化することによって，出現頻度の低い単語を規格化しすぎる場合を回避するこ

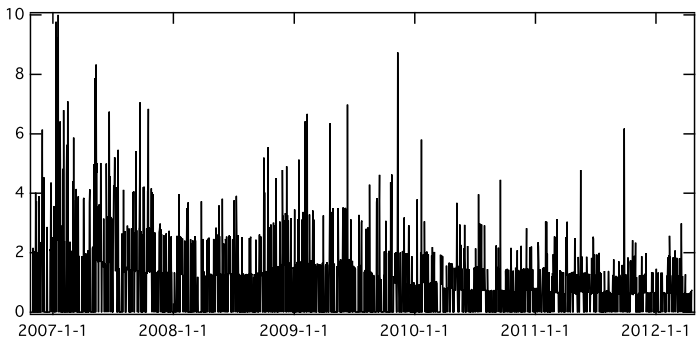


図 3.8　「漸近線」の書き込み時系列 $w'^{(\text{漸近線})}(t)$．(規格化後)　式 (3.5) を用いて規格化した場合．

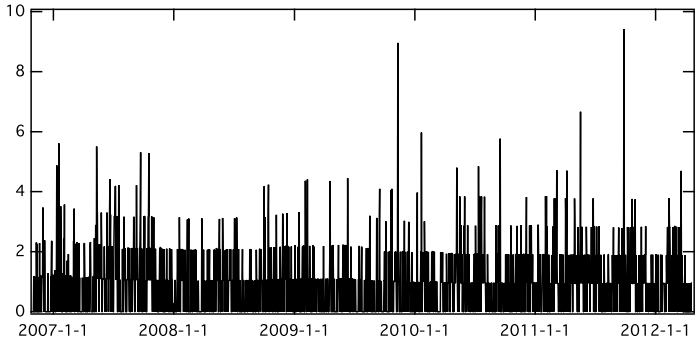


図 3.9　「漸近線」の書き込み時系列 $w''^{(\text{漸近線})}(t)$．(規格化後)　式 (3.10) を用いて相関係数で重み付けして規格化した場合．

ともできる．

図 3.9 は式 (3.10) を用いて相関係数で重み付けして規格化した場合の時系列である．この図 3.9 の相関係数で重み付けをした規格化の場合では，ほぼ全期間で同程度の出現頻度になっていることが分かる．しかしながら，相関係数で重み付けした時系列 $w''^{(k)}(t)$ と全数 $w(t)$ の相関係数 $\rho_{ww''^{(k)}}$ もまた，図 3.7 と同様の傾向が残っており，個々の単語の時系列 $w^{(k)}(t)$ から，全数の影響を取り除くことは容易でないことが分かる (図 3.10)．

除算だけで完全に全数の影響を取り除くことのできない理由として，全数 $w(t)$ を構成しているブロガーの内訳が常に変化していることが考えられる．この理由については第 4 章で詳しく述べる．また，相関係数で重み付けする式 (3.10) の方法は経験的な方法であり，理論的な根拠がある訳ではない．そこで本研究では，手順をなるべく簡素にとどめておくため，相関係数で重み付けした式 (3.10) の規格化ではなく，式 (3.5) の規格化を用いている．

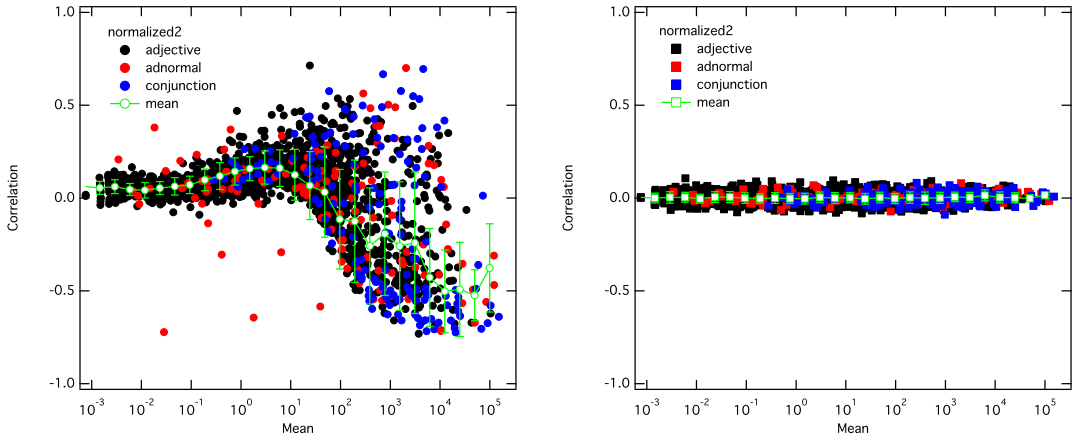


図 3.10　横軸に相関係数で重み付けして規格化した単語 $k$ の平均値 $\langle w''^{(k)} \rangle$，縦軸に全数との相関 $\rho_{ww''^{(k)}}$ を示した．黒印が形容詞，赤印が連体詞，青印が接続詞，白抜きの緑印が全体の平均値．(左図) 実際の時系列 $w''^{(k)}(t)$ では $\langle w''^{(k)} \rangle$ が大きくなると，$\rho_{ww'^{(k)}}$ は負数の絶対値が大きい方へと傾く．(右図)$w''^{(k)}(t)$ において $t$ をランダムシャッフルした時系列の場合．

## 3.3　周期性の除去

ブログ投稿において，1 日の概日周期 (circadian pattern) や 1 週間の周期性を観測することができる．オリンピックに関連する単語であれば，書き込み数は 4 年ごとの周期，「こどもの日」などの年中行事に関連する単語は 1 年ごとの周期を持つ．他にも「暑い」「寒い」といった単語も緩やかな 1 年周期を持つ．個別の単語ごとに周期性を取り上げるときりがないが，本節では，ブログの投稿行動自体が持つ 1 週間と 1 日ごとの周期性について述べ，それに対して行った処理について述べる．

ブログの他にも，携帯電話の通話記録 [62]，ウェブ上の百科事典である Wikipedia の編集履歴 [63]，また日本のコンビニエンスストアにおける購買行動でさえ，様々な周期性が見える．携帯電話の通話記録の研究からは，行動の周期性を取り除いた上でもまだ，人の行動はランダムとはいえずバースト性が残ることが指摘されている．Wikipedia の編集履歴では，およその IP アドレスや言語から国ごととの周期性の強さなども議論されて，それらが文化的背景に基づくものではないかと考察されている．単純な周期の解析だけでも議論すべき点は多いが，本研究では，ブログのもつ自明な周期は，取り除くべきノイズとして扱う．

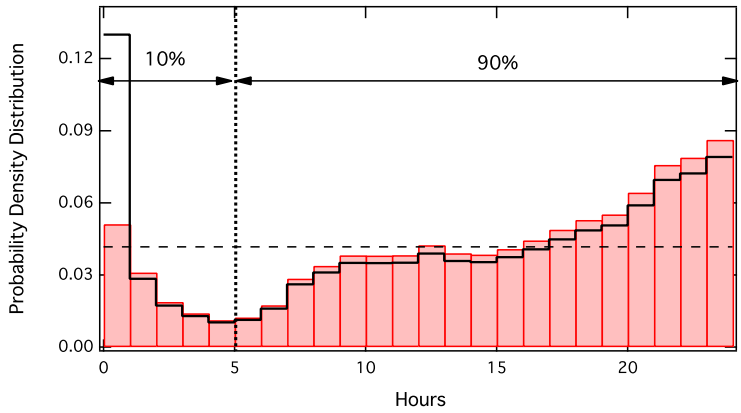


図 3.11　ブログ投稿における時間別の確率分布．点線は時間に関係なく，24 時間平等にブログが投稿されると仮定した場合．黒線は全データを対象とした場合，赤棒は 00:00:00 ちょうどのタイムスタンプのデータを取り除いた場合の結果．

### 3.3.1　1 日の周期

図 3.11 は秒単位のタイムスタンプ付きの詳細なブログデータ 2 から抽出した 10 万人分の投稿記事を用いて，2006 年 11 月 1 日から 2011 年 4 月 19 日までの 1631 日間で算出した 1 日の投稿における確率分布である．すべてのデータを対象とした場合，ブログサイトの設定によってはタイムスタンプの時間を 1 日刻みでしか記録していない場合もあり，この場合には，00:00:00 で扱われる (図 3.11 の実線)．そこで，00:00:00 のタイムスタンプを持つデータを取り除いて再度，投稿確率を算出すると図 3.11 の赤の棒グラフの結果となる．本研究では図 3.11 の赤の棒グラフの結果をブロガーの真の投稿周期として捉えて扱うこととする．24 時間すべての時間帯で等確率にブログが投稿されることを仮定すると，1 時間あたり $1/24 \simeq 4.2\%$ の確率で書き込まれるはずである (図 3.11 の点線)．しかし実際には，深夜に向けて増え続け，夜明けに向けての投稿が減少している．これは人間の 1 日の概日周期が反映されていると考えるのが自然であり，最も投稿が多い時間帯が 23 時台で次いで 22 時台，最も少ないものが午前 4 時台となっている．

このような概日周期は，第 4 章で取り上げる毎日ほぼ同程度で現れる単語であれば影響を受けない．しかし，第 5 章で取り上げるベキ関数的な変動をする時系列が影響を受ける．例えば毎年，4 月 1 日に世界中でイベントが起こる「エイプリルフール」の書き込み時系列の場合，4 月 1 日の前日，3 月 31 日の書き込み数よりも，1 日後の 4 月 2 日の書き込み数が多い (図 3.12 の□)．同様に日本の年中行事である「ひな祭り」や「七夕」の場合でもピーク前後で 1 日前と 1 日後の書き込み数を比較すると，常に 1 日後の書き込み数の方が多い．

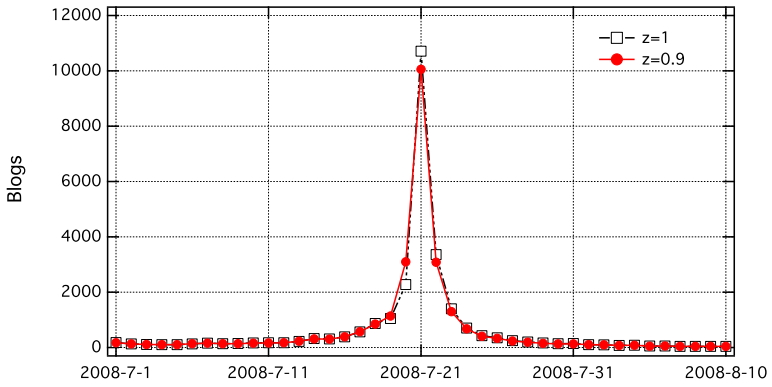


図 3.12　2008 年「海の日」の書き込み時系列 $w^{(\text{海の日})}(t)$．式 (3.11) による時間シフト前後による比較．□が $z = 1.0$ のシフトをしない場合 $w^{(\text{海の日})}_{\text{raw}}(t)$，●が $z = 0.9$ でシフトを行った場合 $w^{(\text{海の日})}_{\text{shift}}(t)$．

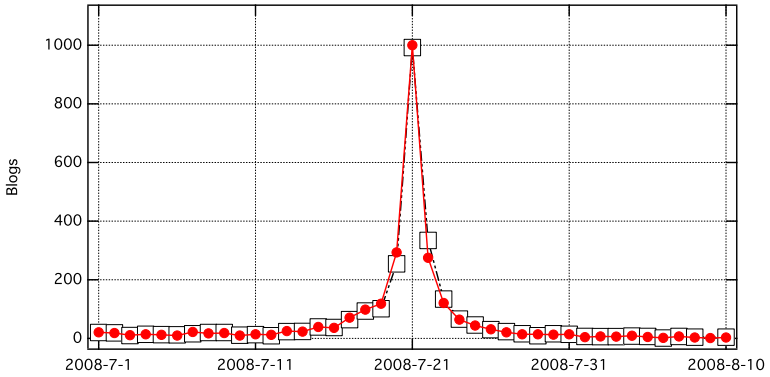


図 3.13　ブログデータ 2 を使って作成した 2008 年「海の日」の書き込み時系列 $w^{(\text{海の日})}(t)$．□が午前 0 時を 1 日の区切りにした場合，●が午前 5 時を 1 日の区切りにした場合．

これは，扱っている時系列データは，午前 0 時が 1 日の区切りとなっているため，イベント当日の午前 0 時を過ぎ，就寝前などの書き込みが，翌日の扱いになっていることが理由であると考えられる．そこで，図 3.11 で確認した概日周期を考えると，おそらくブロガーの就寝前であるイベント当日午前 0 時過ぎに書き込まれた記事は，ブロガーにとってはイベント当日であると考える方が自然である．

実際に，タイムスタンプが詳細なブログデータ 2 を用いて，ブログ投稿の 1 日を午前 5 時を起点とした 24 時間に設定し，時系列を作成すると，ピーク前後での書き込み数がほぼ同じとなることが確認できる．図 3.13 は 33 万人分のブログデータ 2 から作成した「海の日」の書き込み時系列 $w^{(\text{海の日})}(t)$ である．このデータには，ブロガー ID，投稿日時，

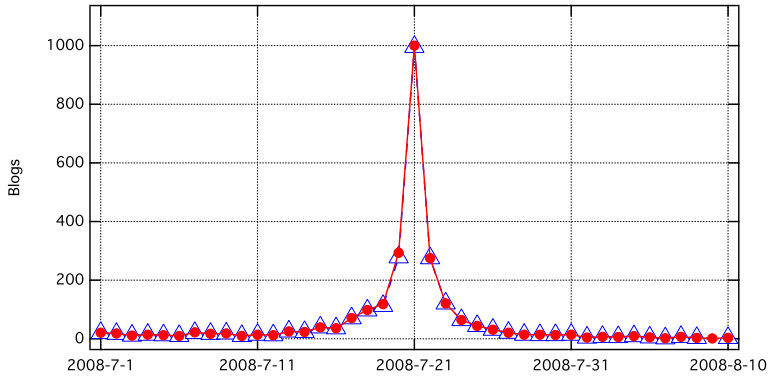


図 3.14　ブログデータ 2 を使って作成した 2008 年「海の日」の書き込み時系列 $w^{(\text{海の日})}(t)$．●が午前 5 時を 1 日の区切りにした場合．△は午前 5 時を 1 日の区切りにし，00:00:00 のデータはサンプルから除去した場合．どちらもほぼ同じ結果になる．

投稿記事すべての情報が含まれるため，投稿日時を午前 5 時を起点とした数に変更した時系列を作成できる．図 3.13 より時間の区切りを午前 5 時に変えた時系列の場合は，ピークの前後 1 日での投稿数ほぼ同程度になっていることを確認できる．また確認のため，同じくブログデータ 2 を用いて，正確なタイムスタンプでないと思われる 00:00:00 のデータをサンプルから取り除いて比較したのが図 3.14 である．どちらの場合でも，午前 5 時からの 24 時間に 1 日の区切りを変更すると，ピーク前後での書き込み数は同程度になっており，ピーク前後 1 日の書き込み数の差は，ブロガーの概日周期の影響であると考えることができる．

そこで，人間の概日周期の影響を取り除いたブログ投稿を見るために，1 日の始まりを朝 5 時に変更すればよいのだが，本研究で取得できるブログデータ 1 の制約上，1 日の時間区切りの変更は行えない．そこで以下のような処理を行う．

元の時系列 $w^{(k)}_{\text{raw}}(t)$ に対して，以下の時間シフトをした時系列 $w^{(k)}_{\text{shift}}(t)$ を用いる．

$$w^{(k)}_{\text{shift}}(t) = z w^{(k)}_{\text{raw}}(t) + (1-z) w^{(k)}_{\text{raw}}(t+1) \tag{3.11}$$

ここで $z$ は，1 日のブロガーの概日周期から決まる重みであり，5:00:00 から 23:59:59 までの 19 時間に投稿されるブログ数と，0:00:00 から 4:59:59 までの 5 時間に投稿されるブログ数の比がおよそ 9:1 となっていることから，$z = 0.9$ とした．時間シフトの処理により，ブロガーの概日周期を反映した時系列に変換することができる (図 3.12 の●)．特にピーク 1 日前後での書き込み数が時間シフトした後は，ほぼ同じ高さになっていることを確認できる．この処理は第 5 章での解析に導入した．ほぼ毎日，同程度書き込まれる単語であれば影響を受けないので，第 4 章では導入していない．

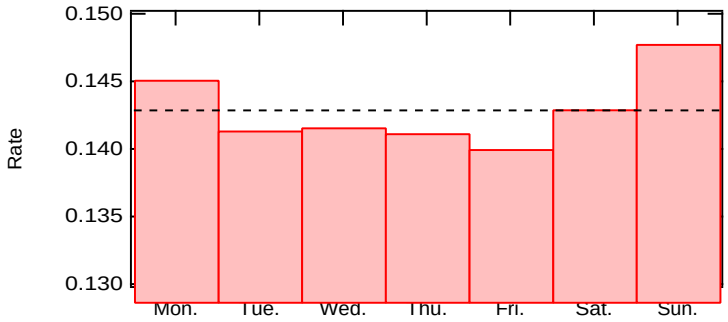


図 3.15　ブログ投稿における曜日別の確率分布．点線は曜日に関係なく，1 週間で等確率にブログが投稿されると仮定した場合．

### 3.3.2　1 週間の周期

図 3.15 に 1 週間におけるブログ投稿における確率分布を示した．1 週間では日曜に最も多く 14.8%，金曜に最も少ない 14.0% となる．週の周期性は 1 日の周期性に比べて大きな差は見られないが，曜日に関係なくブログが投稿されると仮定した場合 (図 3.15 の点線で 14.3% に相当) と比較すると，特に土曜から日曜にかけての週末に微増していることが分かる．このような投稿行動が持つ 1 週間の周期性は，前節で導入した，全数での除算によって取り除くことが出来る．

しかし，個々の単語で見た場合は，投稿周期よりもさらに強い周期性を持つ例を紹介をしておく．図 3.16 左下は検索単語「ドライブ」を含む時系列で，週末ごとに増加している．図 3.16 の右図は曜日ごとの書き込み確率で，年間を通じて週の周期性が強い．そこで月曜から日曜までの曜日の，「ドライブ」の書き込み確率で除算すると，1 週間ごとのでの周期的な変動が消え，「ドライブ」という単語が持つ変動を見通しやすくできる (図 3.16 左上)．ただし，本研究では，個別の単語の周期性には踏み込まないので，この処理は行わない．

## 3.4　その他のノイズ

本研究で扱うデータの性質上，考えられるノイズは前節までに挙げただけにはとどまらない．図 3.17 は，2009 年以降普及した携帯電話の新規機種であるスマートフォンの代表である「iPhone」と「Android」の書き込み時系列である．これらスマートフォンの普及と同時期の 2009 年以降にブログ上の出現頻度は急増しているが，2012 年 1 月 1 日，不

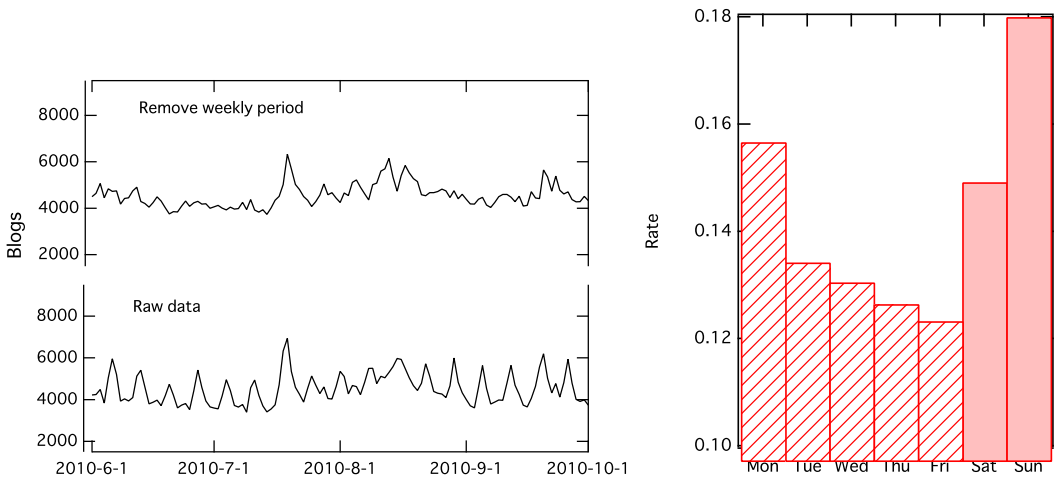


図 3.16　「ドライブ」の書き込み時系列．(左図) 上段は週の周期性除去後の時系列，下段は周期性除去前の時系列．(右図)1 週間の書き込み確率分布，週末に最も書き込まれやすい．

自然に激減している．全数で規格化してもこのノイズは消えないため，データベースを保持し，検索システムを提供しているホットリンクに確認したところ，データ検索システムの仕様変更がなされたことが明らかになった．iPhone や Android から投稿されたブログ記事の最後には，「iPhone からの投稿」「Android からの投稿」という文言が機種の初期設定で付加されている．こういった文言はブロガーが直接書き込んだ単語ではないため，この一連の単語列は検索対象から外す方が，本来の単語の出現頻度を見る際にはふさわしい，ということで 2012 年から検索対象外になった．この仕様変更の影響が時系列に現れている．

データの持つノイズに対し，唯一の対処方法があるわけではない．厳密に管理された実験環境から得られるデータとは異なり，特に社会現象に関するデータはノイズのような非定常性だけではなく，個別性が強い．そのため，データの性質をよく把握した上で解析を行うことが必要である．

## 3.5　まとめ

本章では，ブログにおける単語出現頻度時系列を扱うにあたり行ったノイズを除去する前処理について説明した．データには，ブログ特有の現象であるスパムや，ソーシャルメディアの普及の影響などが反映されており，本来調べたい単語の出現頻度の変動は，前処理抜きでは正しく把握することができない．

そこでまずはじめに，本研究が対象としている日本のブログの世界に多く存在している

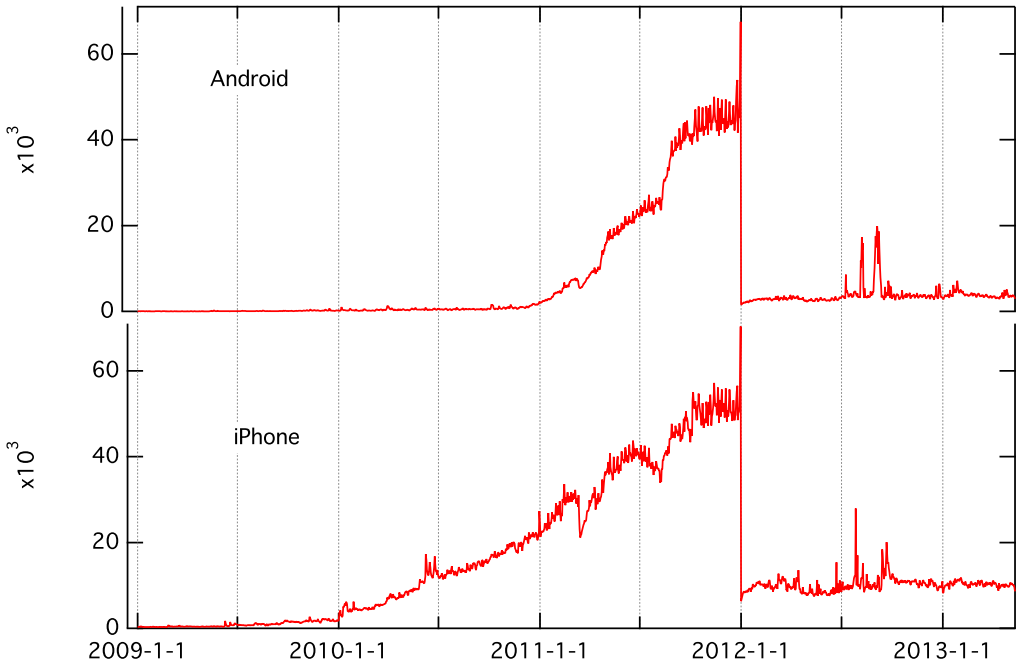


図 3.17　明らかなシステム仕様変更の影響が見られる場合．(上図)「Android」の書き込み時系列．(下図)「iPhone」の書き込み時系列．2012 年 1 月 1 日に仕様変更の影響がある．

と言われているスパムに関しては，既存のフィルタを用いることで対応した．除去したスパムの割合，スパムを含む場合の結果に関しては付録 D に述べた．

次に，単語 $k$ を含む書き込み時系列 $w^{(k)}(t)$ に注目する場合，単語によらないブログ自体の投稿数 $w(t)$ によるゆらぎの影響が無視できない．本研究では $w^{(k)}(t)$ を $w(t)$ で除算する規格化を行うことにより，ブログ自体の投稿数の影響を差し引くことを行った．ただし，$w^{(k)}(t)$ の平均値が大きくなるにつれ，非線形に全数との相関係数が大きくなる傾向があり，本研究で導入した，単純な除算だけでは全数変動からの影響を取りきれていないことを示唆している．

また，深夜に向けてブログ投稿が増加し，明け方に減少するというブロガーの概日周期の影響が大きいことも指摘した．本研究で扱う日次のブログ時系列は，午前 0 時が 1 日の区切りとなっているが，ブロガーの概日周期を 1 日の基準とするならば，午前 5 時を 1 日の区切りとする方が望ましい．しかし，容易に 1 日の区切りを変更した時系列を得ることができないため，ブロガーの概日周期による投稿確率を考慮した重み付けを行って時系列を再構築することで対応した．

# 第 4 章

# 時系列からの異常値検出

本章では，単語出現頻度時系列において，前章で導入したノイズ除去を行った後，日常的に使われる単語を「日常語」として定義し，そのゆらぎ (標準偏差) に注目する．日常語時系列の平均値と標準偏差には非自明なスケーリング則が成り立っており，そのスケーリング則を再現する簡単な確率モデル，ランダム投稿モデルを導入する．ランダム投稿モデルでは，平均値から，標準偏差を一意に算出することができるため，許容されるゆらぎからの逸脱を異常値として定量的に決定することができる．このランダム投稿モデルを用いて，異常値を検出する方法を導入し，実際の検出例を挙げる．実際の例から，ランダム投稿モデルが局所的な異常値を検出できることを示す．

## 4.1 導入：日常的に使われる語のゆらぎの重要性

日本語の語彙は，ブログで使われる新語などを含めると無数に存在する．一例として，日本語の形態素解析器で最も普及しているものの一つである MeCab[*1] に付属の辞書には約 42 万語 (423371 語) 存在する．自分の持つ語彙の中からブロガーは単語を選び，文章を作成し，書き込みを行なっている．そして，流行語などの特徴的な単語が注目される反面，ブロガーが日常的に使う単語の出現頻度時系列に，どのような統計的特性があるのかは，あまり注目されてこなかった．

一般の文章において日常的に使われる単語は多くの場合，その出現過程が独立であるという仮定の元，基本的な確率モデルであるポアソン過程が想定されてきた [38, 64]．文法などの制約条件はあるが，日常的に使われる単語の発生過程にポアソン過程を想定するのは自然であろう．しかし，2005 年に約 6.8 万の RSS フィードを使って集められたブログ

[*1] https://code.google.com/p/mecab/ (Accessed:2013.06.29)

を使って調べた結果では，1 日あたりの出現頻度が 1 回以下の単語の場合でも，特異なイベント (extreme events) が数多く起こりがちであり，ポアソン過程には従わないことが指摘されている [65].

そこで本章では，約 26 億記事以上の大規模なブログデータを用いて，出現頻度が低い単語から高い単語までの時系列のゆらぎについての解析を行い，統計的特性を明らかにする．そしてデータから経験則を抽出し，それを再現する確率モデルに基づいた異常値検出を行う．

日常的に使われる単語は，トレンドを持つ流行語やニュースに左右される単語と比較すると直接的な応用の機会が少ないため，これまでその統計性にはあまり注目されてこなかった．

しかし，最近になり，主に Twitter のデータを中心に社会の雰囲気を定量化するために形容詞や，何気ない名詞の出現頻度が注目されるようになった [21, 22, 66]．これらの単語には，日常的に使われる単語が多く含まれている．これらの研究は雰囲気から金融市場を予測したり，社会の幸福度を定量化することを目的としている．ウェブの世界で起こっていることと実世界に隔たりがあるのは明らかだが，不特定多数の人々が何気なく発する単語の集合は，応用可能性も高く，大きな魅力を持つ科学的分析の対象となっている．

## 4.2 「日常語」の抽出

出現頻度時系列中にトレンドや大きなピークを持たず，日常的に使われる単語を本研究では「日常語」と呼び，その統計的性質を調べる．日常語を定義するため，はじめに，日常語の候補となる単語の出現頻度時系列を網羅的に収集する．日常語の候補となる単語は，ニュースや，季節性変動の外的要因に影響を受けにくいものが望ましい．そこで，形態素解析器 MeCab に標準仕様で付属している IPA 辞書[*2]の単語の中から，形容詞，連体詞，接続詞を抽出する．形容詞の場合，仮定形や命令形など様々な活用形が辞書に含まれるため，全部で 27210 項目ある．その中から，形容詞の原型部分だけを対象とした 1796 単語を抽出する．解析対象期間である 2006 年 11 月 1 日から，2010 年 6 月 9 日までの 1317 日間で一度も使われなかった単語を除くと，全部で 1768 単語が残る．候補となった形容詞の中には「あいいれない」「相いれない」「相容れない」など同一の読みでも複数の書き方があり，これらは「表記ゆれ」と呼ばれるが，別の単語として扱っている．また，「暑い」「寒い」といった明らかな季節のトレンドを含むもの，2008 年当時の首相の発言で話題になり，その前後で出現数が 30 倍以上も増えた「さもしい」も含まれている．そ

[*2] http://code.google.com/p/mecab/downloads/detail?name=mecab-ipadic-2.7.0-20070801.tar.gz (Accessed:2012.10.31)

の他,「あくる」「由々しき」などに代表される連体詞は 135 単語,「すると」「ならば」などに代表される接続詞は 171 単語の時系列を収集することができる.

これら形容詞,連体詞,接続詞の時系列に対して第 2 章の式 (3.5) で定義した規格化を行った後, 弱定常性かどうかを確認するため, 単位根検定 (Unit root test) を行う [67, 68]. 単位根検定は,一階差の時系列 $\Delta y(t) \equiv y(t) - y(t-1)$ で,$\Delta y(t)$ が定常となる時系列に対して,

$$\Delta y(t) = (\gamma - 1)y(t-1) + \epsilon(t) \tag{4.1}$$

で表される $\gamma$ を,データから見積もり,$|\gamma| < 1$ かどうかを検定するものである.$\gamma = 1$ のときは単位根を持ち,ランダムウォークとなる.$\epsilon(t)$ はノイズ項である.例えば,データが $n$ 個与えられた場合,最小二乗法より $\gamma$ は

$$\tilde{\gamma} = \frac{\sum_{t=1}^{n} y(t-1)\Delta y(t)}{\sum_{t=1}^{n} y(t-1)^2} + 1 \tag{4.2}$$

と見積もることができ,有意に $|\tilde{\gamma}| < 1$ かどうかを検定すればよい.単位根検定では「モデルは単位根を持つ」という帰無仮説に対し,$p$ 値を計算する.そのため $p$ 値が小さい場合に,単位根を持つとは言えない,という意味で弱定常性が保証される.

ここでは, 単位根検定の一つである拡張 Dickey-Fuller(ADF) 検定 [69] を用いる.ADF 検定は,他の単位根検定の一つである Phillips-Perron 検定と比較すると,有限サンプルにおけるパフォーマンスで優れていると言われている [68].有意水準 5% 以下で単位根過程が棄却された,弱定常性の性質を持つ単語をここでは「日常語」として,次節でゆらぎの詳細を調べる.形容詞では 1749 単語,連体詞では 132 単語,接続詞では 154 単語がこの日常語として残る.もし式 (3.5) による規格化をしなかった場合には,ADF 検定後の日常語は形容詞は 1526 語,連体詞は 94 語,接続詞は 92 語しか残らない (詳細は付録 B).単位根過程が棄却されず,非定常と判定された単語のなかには,明らかな 1 年周期のトレンドを持つ「温かい」などの緩やかな季節変動をする形容詞が含まれている.

データ収集に用いた検索 API は,パータンマッチした文字列を検索しているだけなので,検索結果が本当に形容詞かどうかはチェックはしていない.そのため,平均出現数が最も大きい形容詞は「ない」で 1 日あたりの $\langle w^{(\text{ない})} \rangle = 275016$ となるが,これには助動詞の否定を表す「ない」も含まれているためだと考えられる.次に多い単語は「いい」で 1 日あたり $\langle w^{(\text{いい})} \rangle = 108846$ で,こちらも形容詞としての意味の他に,動詞の活用形としての「彼がいいました」などに含まれる「いい」がマッチしたために多いと考えられる.

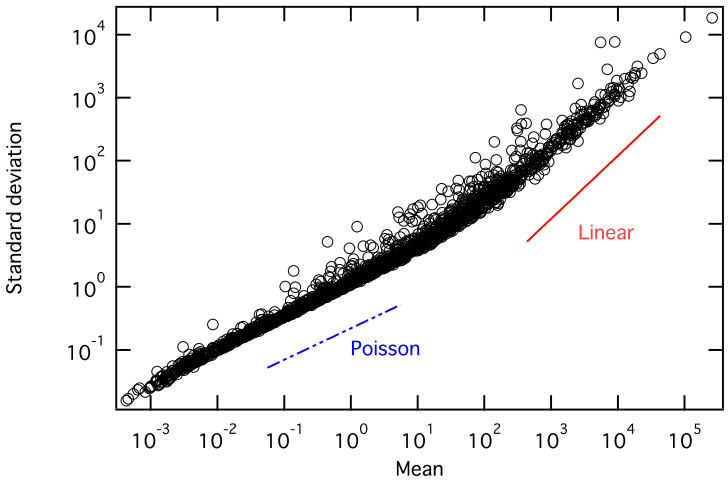


図 4.1　実際のブログ時系列において，日常語となった形容詞 1749 単語の平均値と標準偏差で散布図をとった例．一つの ∘ が一つの形容詞に対応する．スケールのばらつきが大きいため，両軸を対数で表示している．

## 4.3　平均値と標準偏差のスケーリング

図 4.1 に，前節で抽出し日常語となった形容詞時系列 1749 単語において単語 $k$ を変えながら，平均値 $\langle w^{(k)} \rangle$ を横軸に，標準偏差 $\sigma^{(k)}$ を縦軸に散布図にした．ある変数の大きさを変化させたときの，興味ある量の変換関係をスケーリングと言うが，図 4.1 より，平均値に対する標準偏差の非自明なスケーリング則を確認できる．

もし，個々の単語が独立に使われているのであれば，先行研究の仮定通り単語の出現過程は，ポアソン過程で記述できるはずである．ポアソン過程で記述できるのであれば，時系列の標準偏差 $\sigma^{(k)}$ は平均値の平方根で記述でき，$\sigma^{(k)} = \sqrt{\langle w^{(k)} \rangle}$ となるはずである．図 4.1 において，出現頻度が低く，$\langle w^{(k)} \rangle \simeq 10$ 以下の単語では，この関係を確認できる (図 4.1 の破線)．他方，平均値が大きくなるに従って，$\sigma^{(k)} = \sqrt{\langle w^{(k)} \rangle}$ からのずれは大きくなり，平均値が大きい領域では，標準偏差が平均値に対して線形に増え始める (図 4.1 の実線)．この二つの領域を，ブログにおける単語出現頻度の平均値と標準偏差の関係に見いだすことができる．

平均値と標準偏差のスケーリング則は，テイラーのスケーリング則 (Taylor's scaling law)[70] と呼ばれ，1960 年代より知られている．テイラーは平均値 $m$ と標準偏差 $s$ において以下のスケーリング則を見いだし，土壌内，地表，空気中に存在する昆虫，葉の上や羊に寄生するダニ，海中の魚の数などにおいて，以下の指数 $a$ と $b$ を調べた．

$$s^2 = am^b \tag{4.3}$$

指数 $b$ は凝集指数 (index of aggregation) として，機構に固有の指数である．$b$ が 0 に近い場合は，ほぼ定期的，ランダムの場合の $b = 1$ を経て，さらに 1 から大きくなるにつれ，高密度に凝集されると解釈する．

平均値と標準偏差のスケーリング則は，自然界から社会現象まで，一見何の共通点もなさそうな幅広い対象で観測されている．テイラーは，このスケーリング則を，集合の大きさとその標準偏差についてのスケーリング (Ensamble Fluctuation Scaling) として見いだした．例えば，鳥の群れの大きさ，ニューヨーク取引市場での取引量 [71] などがある．しかし，その後の研究で，時系列の平均値とその標準偏差についてのスケーリング (Temporal Fluctuation Scaling) として，河川の流量，アメリカの幹線道路の交通量，インターネット上のパケット流量の時系列の他，ニューヨーク取引市場での株価や取引量の時系列でも，同様のスケーリング則が報告されている [71, 72]．ブログにおける単語出現頻度時系列の平均値と標準偏差のスケーリング則も，この Tmporal Flucutation Scaling に分類される．

## 4.4 スケーリングを説明するランダム投稿モデル

### 4.4.1 先行研究 - ランダム拡散モデル

テイラーのスケーリング則を説明するためのモデルの 1 つに，ネットワークを使ったランダム拡散モデル (Random Diffusion Model)[72, 73] がある．ランダム拡散モデルは，任意のネットワーク上を複数のランダムウォーカーが移動し，ノードごとに単位時間あたりのランダムウォーカーの通過数を数えて，時系列として扱うモデルである．

モデルでは，ランダムウォーカーは，時間ステップごとに次のノードへと移動する．ランダムウォーカー同士の相互作用や，記憶はないと仮定する．また，ランダムウォーカーには寿命があり，ある時間ステップを過ぎると自然消滅する．さらに，ランダムウォーカーは時間ステップごとに新たに注入される．どのノードに注入されるかは，全ノードから等確率で選択される．

ランダム拡散モデルにおけるランダムウォーカーは，インターネット上を行き来するパケットを想像すると，理解しやすい．パケットは時間が経つと消滅し，消滅するまでは目的地に向かってルーター間を移動し続ける．また末端のコンピューターのユーザーによりリクエストが発生し，新たなパケット注入され続ける．

ランダム拡散モデルでは，観測時間あたりにノードを通過するランダムウォーカー数から時系列ができ，そこから平均値と標準偏差を計算できる．リンクの多いノードはランダムウォーカーが多く通過するため，平均値が大きい時系列ができ，リンクの少ないノードは通過するランダムウォーカーが少ないため，平均値の小さい時系列ができる．

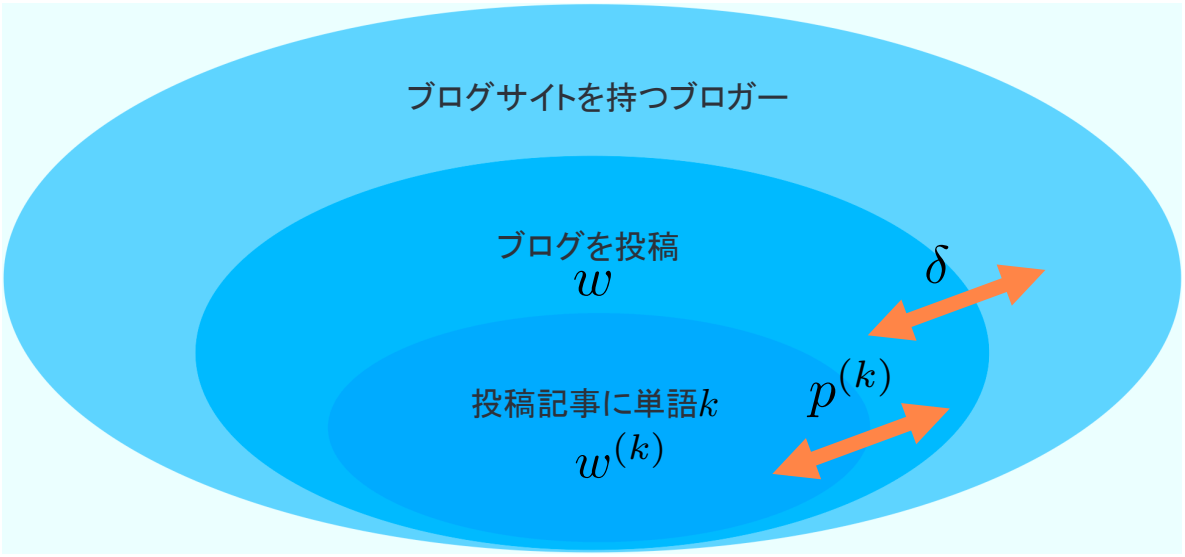


図 4.2　ランダム投稿モデルの概念図．モデルでは全数に関するゆらぎ $\delta$ と，単語ごとに確率 $p^{(k)}$ によって書き込まれるゆらぎのゆらぎの重ね合わせによって記述される．

ランダム拡散モデルは，解析的にも解くことができ，系全体 (ネットワーク全体) に存在するランダムウォーカーの数が，注入と消滅の総和によるゆらぎによって変動することで，時系列の平均値と標準偏差の間に，非自明なスケーリング則が生じることが示される [73]．解析的には，系全体に存在するランダムウォーカー数のゆらぎに一様分布を仮定していたが，数値実験では，ランダムウォーカー数のゆらぎに正規分布を仮定しても，一様分布の場合と同様の平均値と標準偏差のスケーリング則が得られることが指摘されている．

### 4.4.2　ランダム投稿モデル

本研究では，ランダム拡散モデルを，ブログの世界に置き換えたランダム投稿モデル (Random Posting Model) を導入する．ランダム拡散モデルではまずネットワークを仮定していたが，本ランダム投稿モデルはネットワークは考慮しない点で大きく異なる．

単語 $k$ ごとに書き込まれる確率を $p^{(k)}$ とする．$k$ が「しかし」「漸近線」などの個々の

単語に対応する．観測しているブログの世界全体で $w$ 人のブロガーがいると仮定する．単語 $k$ は他の単語との共起関係は考えず，過去の自分の書き込み回数からも無相関だとすると，$p^{(k)}$ は定数で表され，単語 $k$ を含むブログの期待値 $\langle w^{(k)} \rangle$ は全数に比例し，以下で与えられる．

$$\langle w^{(k)} \rangle = p^{(k)} w \tag{4.4}$$

期待値 $\langle w^{(k)} \rangle$ をパラメータとしたポアソン過程を考え，単語 $k$ が $n$ 回書き込まれる確率は以下のようになる．

$$P^{(k)}(n; \langle w^{(k)} \rangle) = \mathrm{e}^{-\langle w^{(k)} \rangle} \frac{\langle w^{(k)} \rangle^n}{n!} = \mathrm{e}^{-\left(p^{(k)} w\right)} \frac{\left(p^{(k)} w\right)^n}{n!} \tag{4.5}$$

次に，観測している全ブロガー数が時刻 $t$ によって変化する場合を $w(t)$ で考える．これは，ブロガーは単語 $k$ を書き込むか否かの決定の前に，単語によらず，ブログを投稿するか否かを決定することを想定している (図 4.2)．ここでは簡単のため，単語によらない全投稿数が $[w-\delta, w+\delta]$ の範囲で一様分布でゆらぐ場合を考える．すると単語 $k$ を含む書き込み数の期待値 $\langle w^{(k)} \rangle$ も $[p^{(k)}(w-\delta), p^{(k)}(w+\delta)]$ の範囲でゆらぐが，$\langle w^{(k)} \rangle$ は式 (4.4) と同じで，$\langle w^{(k)} \rangle = p^{(k)} w$ となる．

式 (4.5) でみた，単語 $k$ が $n$ 回書き込まれる確率は，全ブロガー数が $[w-\delta, w+\delta]$ の場合を等確率で足し上げれば良いので，

$$P^{(k)}(n) = \frac{1}{2\delta+1} \sum_{m=-\delta}^{2\delta} \mathrm{e}^{-\left[p^{(k)}(w+m)\right]} \frac{\left[p^{(k)}(w+m)\right]^n}{n!} \tag{4.6}$$

と表すことができる．これより二次のモーメントは

$$\langle {w^{(k)}}^2 \rangle = \sum_{n=0}^{\infty} n^2 P^{(k)}(n) = \langle w^{(k)} \rangle^2 \left[ 1 + \frac{\delta(\delta+1)}{3w^2} \right] + \langle w^{(k)} \rangle \tag{4.7}$$

より，$\delta \simeq \delta + 1$ とすると

$$ {\sigma^{(k)}}^2 = \langle {w^{(k)}}^2 \rangle - \langle w^{(k)} \rangle^2 = \langle w^{(k)} \rangle \left[ 1 + \frac{\langle w^{(k)} \rangle}{3} \left( \frac{\delta}{w} \right)^2 \right] \tag{4.8}$$

となる．式 (4.8) 式の右辺第二項にある全ブロガー数のゆらぎがないとき，$\delta = 0$ より，${\sigma^{(k)}}^2 = \langle w^{(k)} \rangle$ となり，ポアソン過程における平均値が分散に等しいことと一致する．他方，全ブロガー数 $w$ に対し，ゆらぎ $\delta$ が無視できなくなるとき，式 (4.8) において右辺第二項が支配的になり，標準偏差は平均値に対して線形に増加する ($\sigma^{(k)} \propto \langle w^{(k)} \rangle$) ことが示される．また，$\delta$ の大きさと全ブロガー数 $w$ の比 $\delta/w$ に対して $\langle w^{(k)} \rangle$ が大きいかどうかで，式 (4.8) の右辺第一項が支配的か，第二項が支配的かが決まる．

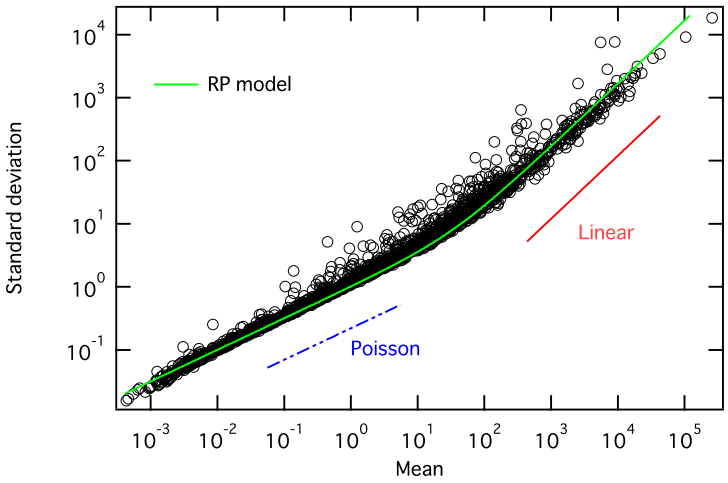


図 4.3　式 (4.8) の解 ($\frac{\delta}{w} = 0.29$) と，実際のブログデータの結果 (図 4.1) を重ねたもの．標準偏差が平均値の平方根に比例する領域から，線形に比例する領域まで記述できている．

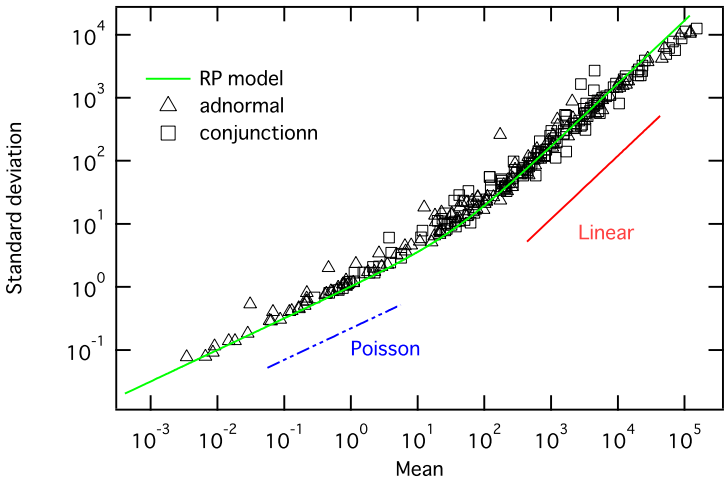


図 4.4　式 (4.8) の解 ($\frac{\delta}{w} = 0.29$) と，実際のブログデータの結果を重ねたもの．△が連体詞，□が接続詞に対応する．

図 4.3 は，式 (4.8) の解と，図 4.1 で確認した実際のブログデータの結果を重ねたもので，全領域を記述できていることが分かる．

連体詞と接続詞でも，形容詞の場合と同様の平均値と標準偏差のスケーリング則と，ランダム投稿モデルの理論線が実際のデータを再現できていることを確認できる (図 4.4)．

ランダム投稿モデルで重要な役割を果たすのは，観測している系に存在する全ブロガー数が，時間ステップごとにゆらぐ，という効果である．これは具体的には，単語によらない全ブログ投稿数のゆらぎに対応する．

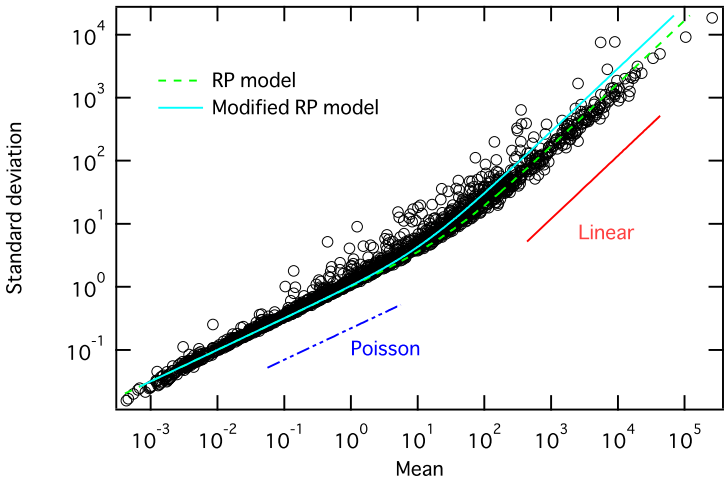


図 4.5　式 (4.8) の解 (破線) と式 (4.9) の解 (実線) の比較．どちらもパラメータは $\frac{\delta}{w} = 0.29$．○は形容詞を使った実際のデータ．

投稿される全ブログ数 $w(t)$ を実際のデータから確認しておく．$w(t)$ は，1 週間の周期性など，日々ゆらいでおり，平均値 $\langle w \rangle$ は約 70 万件，標準偏差 $\delta$ は約 20 万件である．この値を式 (4.8) に代入し，$\delta/w = 20$ 万$/70$ 万 を使って実際に計算してみると，右辺第二項が支配的になるのは，$\langle w_j \rangle = 3 \times (70\text{ 万}/20\text{ 万})^2 \simeq 37$ のあたりであることが分かる．これは図 4.1 で見た，ブログの平均書き込み数が約 10 件より大きくなると，標準偏差が平均値に対して線形に増え始めることともほぼ一致している．日常語の場合において，標準偏差が平均値に対して線形に増え始める現象は，全投稿数に由来するということが，ランダム投稿モデルから示唆された．

本節での計算は，全ブロガー数のゆらぎの分布に $[w-\delta, w+\delta]$ の一様分布を仮定したが，$p^{(k)}$ が非常に小さい場合と，大きい場合で分けることで，任意のゆらぎの場合に拡張しても同等の結果が得られる [C]．この場合の平均値と標準偏差のスケーリング則は，

$$\sigma^{(k)^2} = \langle w^{(k)} \rangle \left[ 1 + \left( \frac{\delta}{w} \right)^2 \langle w^{(k)} \rangle \right] \tag{4.9}$$

となる．この結果は，式 (4.8) の右辺第二項にあった $\frac{1}{3}$ の部分が省略されるだけで，式 (4.8) とほぼ同じである．図 4.5 に式 (4.8) と式 (4.9) において同じパラメータ $\frac{\delta}{w} = 0.29$ を使って比較したものを表示した．同じパラメータ $\frac{\delta}{w} = 0.29$ を用いると，任意の分布に拡張した式 (4.9) の方が，$\frac{1}{3}$ の分だけ標準偏差は大きめに見積もられる．式 (4.9) において $\frac{\delta}{w} = 0.17$ とすると，$\frac{\delta}{w} = 0.29$ の場合の式 (4.8) と同じ結果となる．

インターネットのパケット流量の時系列，マイクロプロセッサ内に流れるの微細な電流の時系列など，多くのブログ以外のデータで，平均値と標準偏差のスケーリング則をポア

ソン的な振る舞い ($\sigma \propto \sqrt{\langle w \rangle}$) をするのか，線形な関係 ($\sigma \propto \langle w \rangle$) になるのかどちらかに分類し，その結果，外的要因が支配的か，内的要因が支配的かどちらかに二分して議論されているのに対し，ブログデータは，平均値と標準偏差の関係がポアソン過程に従う領域だけではなく，線形になる領域の両方が，一つのデータの中に潜んでいる点で興味深い．これは単語が無数に存在するため，数多くの時系列を取得可能なブログデータを使った研究の長所であると言える．

本章での解析は，前章で導入した通り，あらかじめ全数で除算する前準備を行っている．しかし，前準備を行なっていても，全数のゆらぎの影響が残っている．(付録 B に全数で除算しない場合の結果を挙げた．全数で除算しない場合は，全数の影響が大きく $\delta$ の値が大きい) 式 (3.5) の全数で単純な除算をしただけでは，完全に全数の影響を取りきれないことは，規格化した後も出現頻度の大きさに応じて全数との相関が非自明に変化していることからも明らかである (図 3.7)．この理由として考えられるのが，ブロガーの不均質性である．

ランダム投稿モデルでは，式 (4.4) においてすべてのブロガーは，単語 $k$ に対して等確率で投稿すると仮定し，その和としての単語 $k$ の投稿確率 $p^{(k)}$ を用いていた．この仮定より得られた $\langle w^{(k)} \rangle$ を用いて，最終的な式 (4.8) を導いていた．しかし，本来ブロガーは均質ではなく，ブロガー $i$ ごとに単語 $k$ に対して異なる投稿確率 $p_i^{(k)}$ で行動しているはずである．さらに，時間ステップごとにブログ記事自体を投稿するブロガーも異なる．そのため，

$$w^{(k)}(t) = w(t) \sum_{i \in t \text{ での投稿ブロガー}} p_i^{(k)} \tag{4.10}$$

となり，式 (4.4) における $\langle w^{(k)} \rangle$ の値にも，ブロガーごとのゆらぎが反映される．しかし，式 (4.8) や式 (4.9) ではブロガーの不均質性の効果は無視している．そのため，全数での除算後の時系列にも，ブロガーの不均質性のゆらぎが残っていることが考えられる．個々の時系列から切り分け困難な全数の影響を取り除くことは，今後の重要な課題の一つである．

## 4.5　応用：異常値の検出

式 (4.8) と式 (4.9) の結果より，平均的な全数の値 $w$ に対するゆらぎ $\delta$ の比 $\frac{\delta}{w}$ だけで，平均値 $\langle w^{(k)} \rangle$ から標準偏差 $\sigma^{(k)}$ を導けることが明らかになった．そこで，様々な時系列に対して，式 (4.9) を適用し，許容されるゆらぎから大きく逸脱した値を「異常値」としてとらえ，得られた主な結果を紹介する．

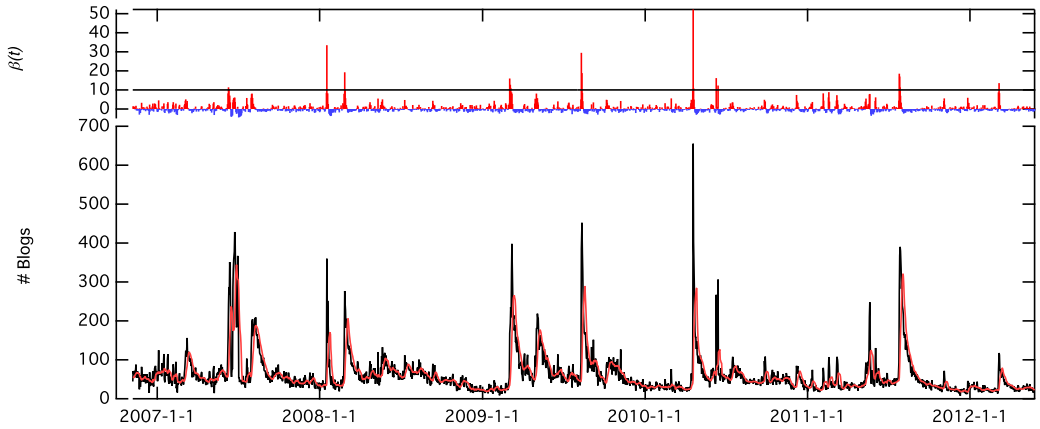


図 4.6　「生茶」という商品名の書き込み時系列．下段の実線が全数で規格化したブログ数 $w^{(k)}(t)$，破線が過去 7 日間の移動平均値 $\overline{w^{(k)}}(t)$．上段が実際のブログ数から見積もったゆらぎから逸脱した割合 $\beta(t)$.

### 4.5.1　検出方法

異常値の検出には，予測される書き込み数とそのゆらぎを用いる．ここでは予測される書き込み数は過去 $n$ 日間の移動平均値

$$\overline{w^{(k)}}(t) = \frac{1}{n}\sum_{m=1}^{n} w^{(k)}(t-m) \tag{4.11}$$

を用いる．移動平均値 $\overline{w^{(k)}}(t)$ に対して，式 (4.9) より標準偏差は

$$\sigma^{(k)}(t) = \sqrt{\overline{w^{(k)}}(t)\left[1+\left(\frac{\delta}{w}\right)^{2}\overline{w^{(k)}}(t)\right]} \tag{4.12}$$

となる．これを予測される書き込み数のゆらぎとする．実際の書き込み数 $w^{(k)}(t)$ のゆらぎからの逸脱度 $\beta(t)$ は以下によって定義する．

$$\beta(t) = \frac{w^{(k)}(t) - \overline{w^{(k)}(t)}}{\sigma^{(k)}(t)} \tag{4.13}$$

$\beta(t)$ は移動平均値からの乖離を算出しているため，負数もある．しかしここでは，何かイベントがあるごとに $\beta(t)$ が大きく増加する様子に注目するため，正数のみを扱う．

表 4.1　図 4.6 の「生茶」の書き込み時系列におけるピークの抽出例．$\beta(t)$ の大きかった上位 5 件．

| 順位 | 日付 | イベント | 逸脱度 $\beta(t)$ | 規格化前ブログ数 | 規格化後 |
|---|---|---|---|---|---|
| 1 | 2010.04.19 | CM リニューアル 1 | 52.2 | 763 | 655 |
| 2 | 2008.01.17 | CM リニューアル 2 | 33.5 | 261 | 340 |
| 3 | 2009.08.11 | キャンペーン 1 | 29.5 | 299 | 385 |
| 4 | 2008.02.26 | キャンペーン 2 | 19.2 | 216 | 277 |
| 5 | 2008.08.12 | 3 位と同じ | 18.8 | 339 | 452 |

### 4.5.2　検出した異常値

図 4.6 は「生茶」という商品名に注目し，下段に実際のブログ数 $w^{(\text{生茶})}(t)$ と，$n = 7$ 日間の移動平均値 $\overline{w^{(\text{生茶})}}(t)$，上段に逸脱度 $\beta(t)$ を表示した．$n = 7$ とした理由は，単語ごとに持ちうる週の周期性の影響を取り除くためである．表 4.1 には，$\beta(t)$ の大きい順に日付を抽出し，該当するイベントをブログ本文から読み取り，商品の公式ウェブサイトなどで確認して対応付けした．逸脱度 $\beta(t)$ は 2010 年 4 月 19 日に最大値をとり，この日付には人気のあるタレントによる新 CM 放送開始という話題があり，明らかな外的要因を対応づけることができる．その他の日付にも，それぞれ外的要因を確認できる (表 4.1)．

$\beta(t)$ に対して閾値の大きさをどうするかで，異常値の検出力を調整する事が出来る．閾値を低めにしておくと，多くの異常値が検出できるが，わずかなノイズでも反応してしまう．大きすぎると，検出したかった異常値を見逃してしまうかもしれない．また，異常値の出やすさは，ブログに書き込まれやすい単語やそうでない単語の個別性に大きく左右される．

図 4.7 は 2006 年 11 月 1 日から 2012 年 5 月 25 日までの 2033 日間での，式 (4.13) で定義した逸脱度 $\beta(t)$ の相補累積分布関数 (Complementary Cumulative Distribution Function, $CCDF$) である．$CCDF$ は，確率密度関数を下限から $x$ まで積分した累積分布関数 $P(\geq x)$ に対して，$CCDF = 1 - P(\geq x)$ で表示される．ランダム投稿モデルのパラメータは $\left(\frac{\delta}{w}\right)^2 = 0.17$ とし，単語は日本の主なペットボトル入りお茶の商品名「生茶」「爽健美茶」「綾鷹」「伊右衛門」「からだ巡り茶」を用いた．これらの商品名は図 4.6 の場合と同様，CM を断続的に行っており，その度にピークが現れる．また，検索対象となる単語は特徴的な名前であり，お茶の商品名以外の用途で使われることがないため，異常値検出用のサンプルとして実用的である．

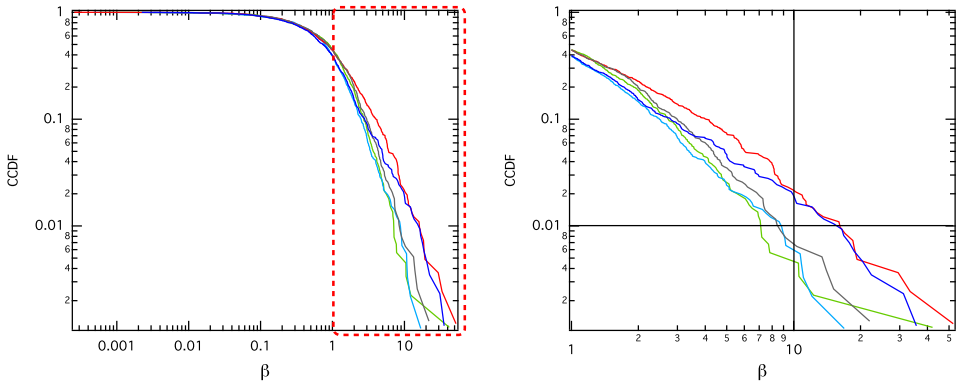


図 4.7　式 (4.13) で定義した $\beta$ の相補累積分布関数．左図は全範囲を表示したもの．右図は $\beta \geq 1$(左図の破線で囲った部分) の範囲を拡大したもの．どちらも両軸対数で示した．

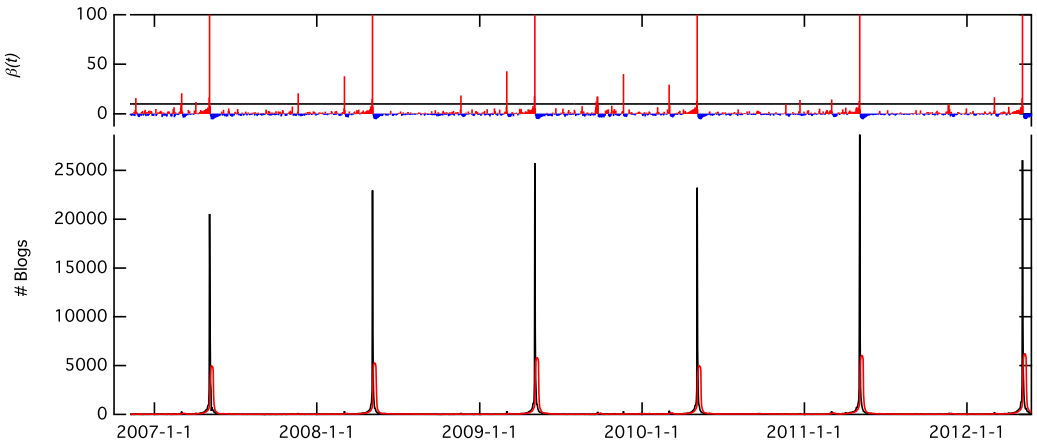


図 4.8　(下段)「こどもの日」の書き込み時系列 (黒)．赤線が過去 7 日間の移動平均線．(上段) 式 (4.13) で求めた逸脱度 $\beta(t)$．毎年，5 月 5 日のこどもの日に鋭いピークを持ち，逸脱度 $\beta(t)$ の値も 100 を超える．

図 4.7 の右図より，$\beta \geq 10$ とすると，およそ $CCDF$ の割合で 1% 以下の異常値検出ができることが分かる．

この他に，検出できる異常値に以下のようなものがある．

- **締め切りを持つ語**　「こどもの日」「母の日」など毎年あるイベントの名前で，日付が決まっている語．これらの単語は，イベント当日に鋭いピークを持ち，その前後でベキ関数的に増減する．図 4.8 は「こどもの日」の時系列とその逸脱度 $\beta(t)$ である．毎年，こどもの日である 5 月 5 日に書き込み数が増え，$\beta(t)$ の値も 100 を超える大きな値を持つ．また，3 月 3 日のひな祭り，さらに 11 月 20 日の世界こども

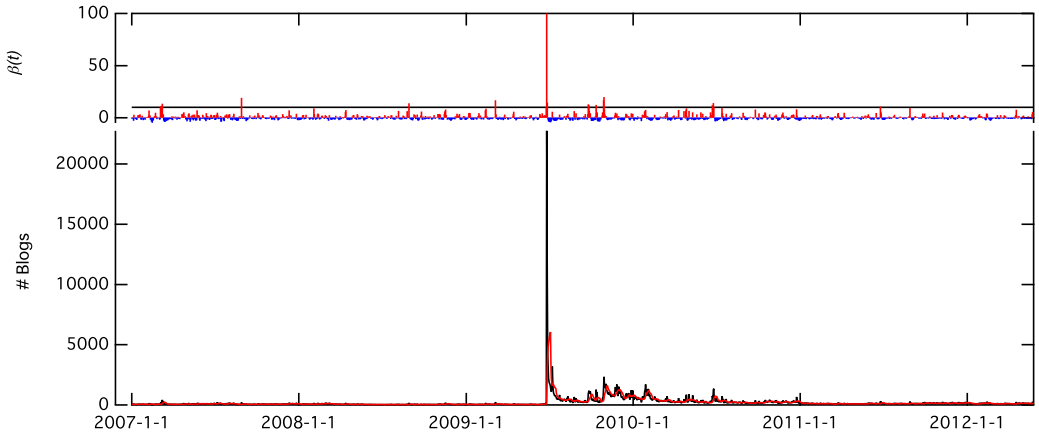


図 4.9　(下段)「マイケル・ジャクソン」の書き込み時系列 (黒)．赤線が過去 7 日間の移動平均線．(上段) 式 (4.13) で求めた逸脱度 $\beta(t)$．2009 年 6 月のマイケル・ジャクソン急逝のニュースにより，逸脱度 $\beta(t)$ の値が跳ね上がる．

の日でも「こどもの日」という単語が使われているため，それぞれ $\beta(t) \geq 10$ となるピークを持っている．これらの単語は第 5 章で詳細に解析する．

- **突発的なニュースに関連する語**　「マイケル・ジャクソン」「iPS 細胞」など訃報，受賞に代表される，突然入る大きなニュースに関連する語．これらの単語は，ニュースリリース日に最も大きなピークを持ち，その後ベキ関数的に書き込み数が減少する．図 4.9 は「マイケル・ジャクソン」の時系列とその逸脱度 $\beta(t)$ である．マイケル・ジャクソンの急逝ニュースが世界中を駆け巡った 2009 年 6 月 26 日に最大のピークを持つ．それ以前は 1 日あたりの書き込み数は約 50 件であったのに対し，ピーク時の値は約 2.3 万件，さらにピークから 2 年以上経った現在も 1 日あたり約 100 件の書き込みがある．これらの単語も第 5 章で詳細に解析する．

### 4.5.3　検出困難な場合

前節で検出した検出方法は，特に突発的な鋭いピークを検出するのに有効である．反面，本手法は，過去の移動平均を使っているため，緩やかなトレンドを持つ場合や，書き込み数が極端に少ない場合は検出が困難になる．

#### 日常語の場合

比較のため，日常語である「また」の時系列対して異常値検出を行ったのが図 4.10 である．日常語は，トレンドやピークを持たないため，$\beta(t)$ の値はどの期間でも 10 より小

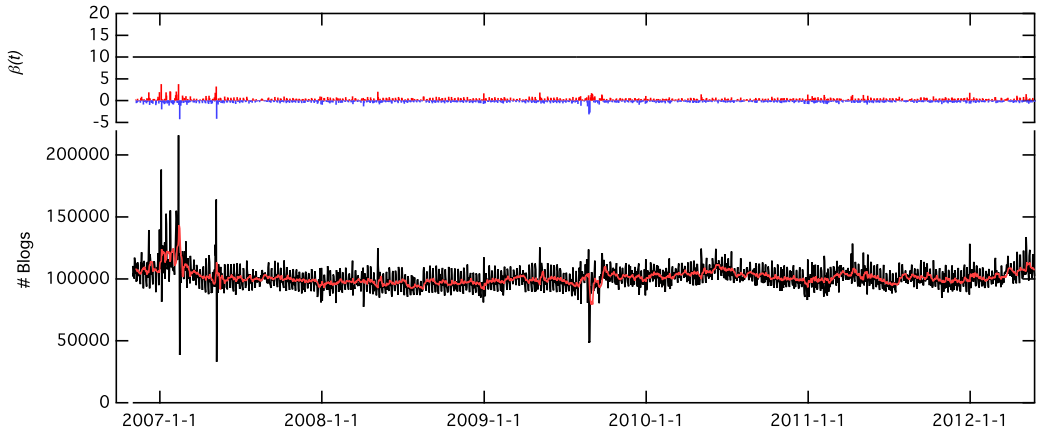


図 4.10　(下段)「また」の書き込み時系列 (黒)．赤線が過去 7 日間の移動平均線．(上段) 式 (4.13) で求めた逸脱度 $\beta(t)$．$\beta(t) \geq 10$ となる異常値は検出されない．

さいことが分かる．

**指数関数的に変動する場合**

図 4.11 は「なう」という単語含むブログ時系列である．「なう」は，Twitter において多用される表現で，「現在 (now)〜している」という意味で使われる．「なう」は日本での Twitter の普及と同時に多くの人に使われるようになり，ブログ上でも出現頻度が増えた．そのため，2010 年から 2011 年にかけては書き込みが増え続けている．しかし，$\beta(t)$ はほぼ 10 より小さい値に収まっている．

時系列中の毎週月曜にピークが現れるが，これはブログサービス事業者の一つで，大きなシェアを占めるアメーバブログが，「なうをまとめて自動投稿」という機能を提供しているためである．これは，ブロガーが一週間の間に投稿した「なう」が，まとめて翌日の記事に自動で投稿される機能で，この更新周期が月曜になっているため，毎週月曜にピークが現れている[*3]．

このように，明らかなトレンドを持つ場合には，過去の移動平均値も順次大きい値を取り続けるため，逸脱度 $\beta(t)$ を使った異常値検出は困難になる．しかし，同時にこれらの時系列は弱定常性が棄却される．図 4.11 で見た「なう」の場合にも ADF 検定で $p$ 値は 0.86 となり非定常と判定されている．これらの結果より，弱定常性の検定と本ランダム投稿モデルによるゆらぎを組み合わせることにより，精緻な異常値検出が期待できる．

また，トレンドを持つ単語は，最大値をとるピーク日を $t_c$ として，以下の指数関数的な

[*3] http://helps.ameba.jp/faq/now/7044/post_120.html (Accessed:2013.06.18)

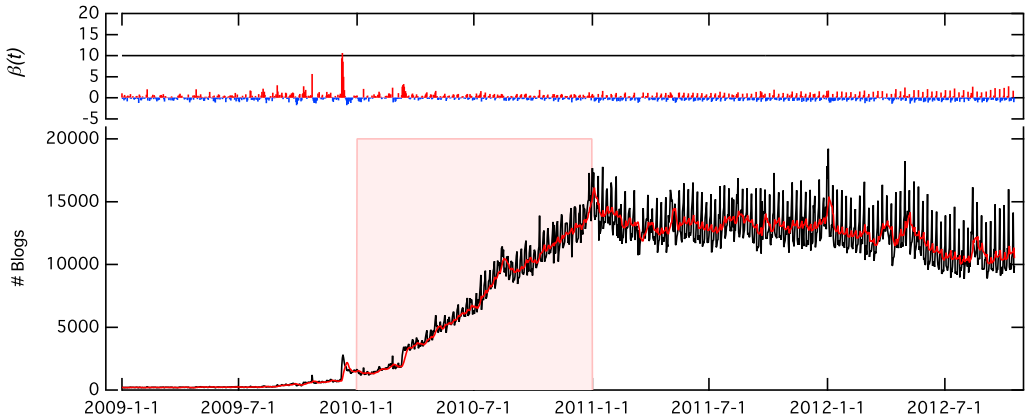


図 4.11　(下段)「なう」の書き込み時系列 (黒)．赤線が過去 7 日間の移動平均線．(上段) 式 (4.13) で求めた逸脱度 $\beta(t)$．2010 年から 2011 年にかけて書き込みが増え続けているが，$\beta(t) \geq 10$ となる期間はほとんどない．

変動で記述することができる．

$$w^{(k)}(t) = B^{(k)} \exp\left(-\frac{|t_c - t|}{\tau^{(k)}}\right) \tag{4.14}$$

$\tau^{(k)}$ は，時定数 (time constant) と呼ばれる指数関数の変動を特長づける重要なパラメータである．現在は指数関数は自然対数 e の底を使っているため，ピークから $1/\mathrm{e} \sim 0.37$ 倍になるまでにかかる日数に相当する．図 4.12 はの左図は，図 4.11 で確認した「なう」の書き込み時系列のピーク日の 1 年前からの期間を式 (4.14) で示したものである．片対数の図でほぼ直線的となっており，指数関数的な変動をしていることが分かる．

このように指数関数的に変動する単語は，他にも以下のようなものが挙げられる．

- **一過性の流行語**　「ＫＹ」「婚活」など，新語や社会的なブームの名前は指数関数的に変動する単語の一つである．図 4.13 は「ＫＹ」の時系列とその逸脱度 $\beta(t)$ を示した．2007 年ごろから緩やかに増え続けているため，局所的な異常値は検出されない．図 4.12 の右図は，図 4.13 の「KY」において網掛け部分を片対数で示した図である．片対数でほぼ直線的になっており，指数関数的に変動していることが分かる．
- **季節性の流行語**　「みかん」「ひまわり」など，旬をもつ野菜果実や花の名前は毎年，周期的に増減する (図 4.14)．図 4.15 は図 4.14 の「みかん」において毎年の網掛け部分を片対数で示した図である．こちらも毎年，片対数でほぼ直線的になっており，指数関数的に変動していることが分かる．

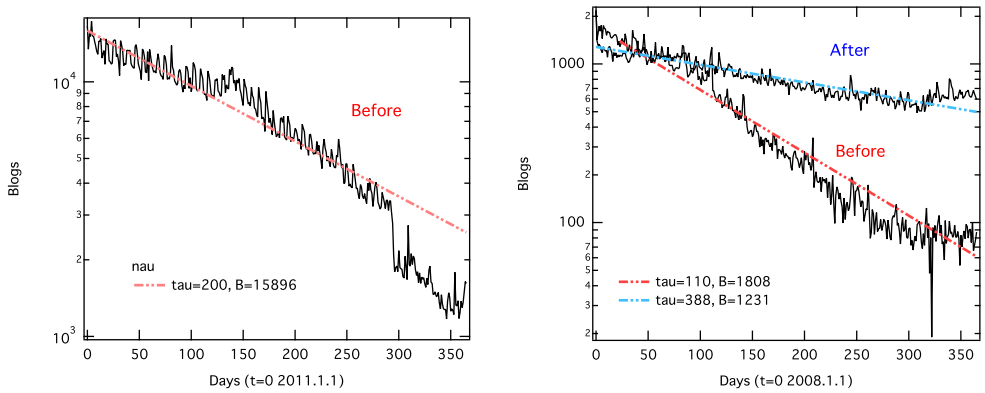


図 4.12　(左図)「なう」が最大値を取る 2011 年 1 月 1 日を $t_c$ として，$t_c$ 以前の 1 年間 (図 4.11 の網掛け部分) の変動をピークからの日数に対し，片対数で示したもの．破線は式 (4.14) において $\tau^{(\text{なう})} = 200$，$B^{(\text{なう})} = 15896$ を当てはめたもの．(右図)「KY」がピークを持つ 2008 年 1 月 1 日を $t_c$ として，$t_c$ の前後 1 年間 (図 4.13 の網掛け部分) の変動を片対数で示したもの．破線は式 (4.14) において $\tau^{(\text{KY, fore})} = 110$，$B^{(\text{KY, fore})} = 1808$，$\tau^{(\text{KY, after})} = 388$，$B^{(\text{KY, after})} = 1231$ を当てはめたもの．片対数でほぼ直線的になっており，指数関数的な変動をしていることが分かる．

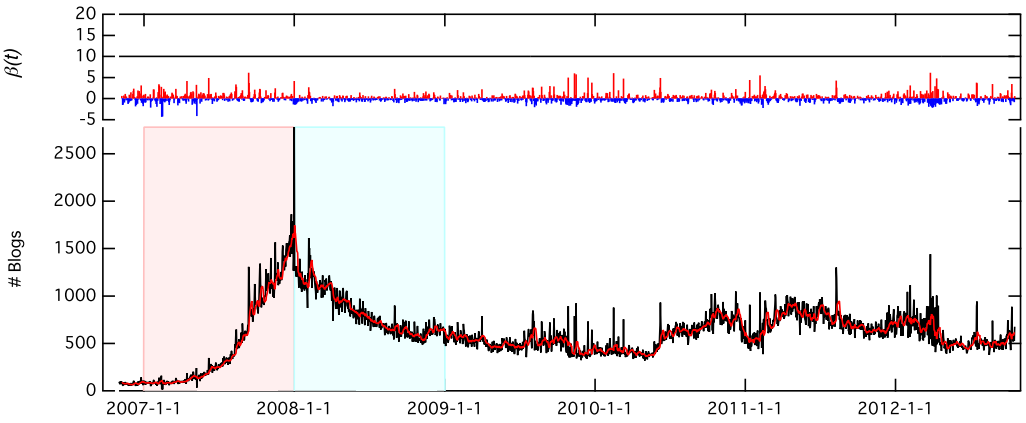


図 4.13　(下段)「K Y」の書き込み時系列 (黒)．赤線が過去 7 日間の移動平均線．(上段) 式 (4.13) で求めた逸脱度 $\beta(t)$．2007 年から 2008 年にかけて書き込みが増え続けるが，$\beta(t) \leq 10$ に収まっていることを確認できる．

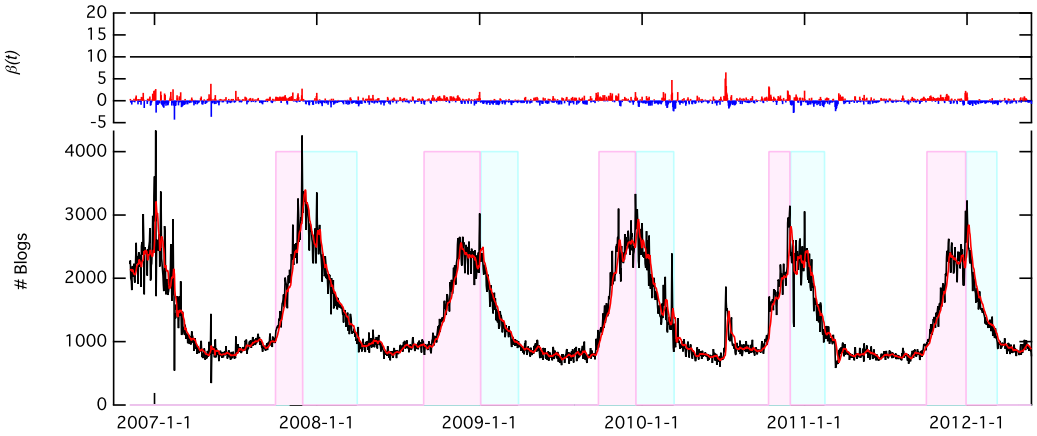


図 4.14　(下段)「みかん」の書き込み時系列 (黒)．赤線が過去 7 日間の移動平均線．(上段) 式 (4.13) で求めた逸脱度 $\beta(t)$．

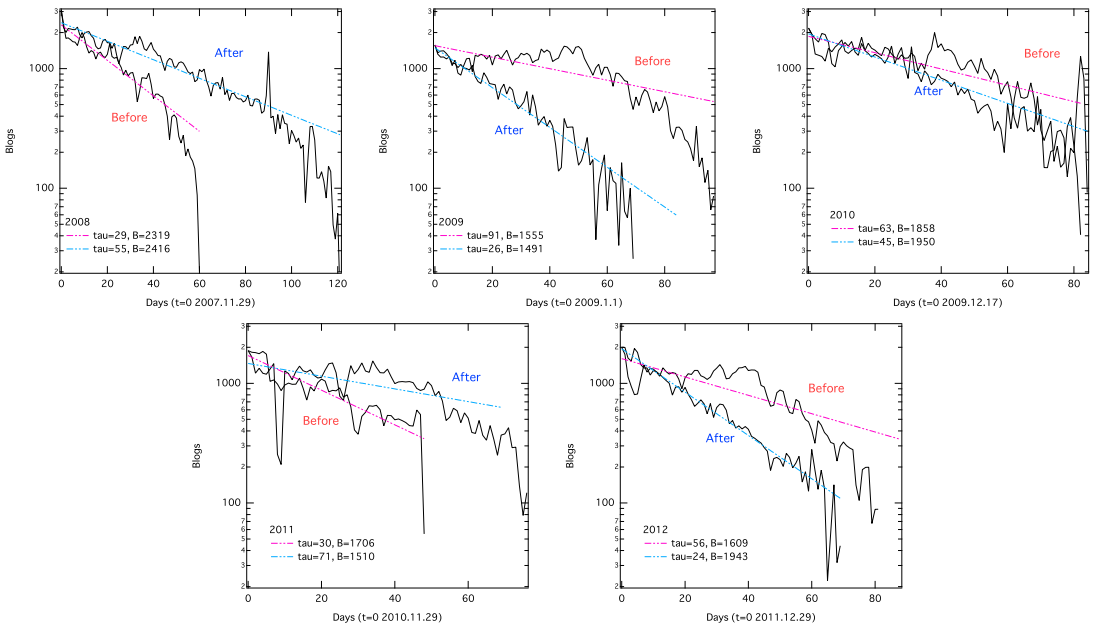


図 4.15　毎年「みかん」がピークを持つ日を $t_c$ として，ピーク前後で中央値を超え続けた期間 $n^{(\text{みかん})}$(図 4.11 の網掛け部分) の変動をピークからの日数に対し，片対数で示したもの．破線は式 (4.14) を当てはめた結果で，パラメータの値は表 4.2 に示した．

表 4.2　「みかん」のピーク前後の変動を式 (4.14) の指数関数に当てはめた時のパラメータ．$n^{(みかん)}$ は，式 (4.14) を当てはめた期間．

| | 2008 | | 2009 | | 2010 | | 2011 | | 2012 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Before | After | Before | After | Before | After | Before | After | Before | After |
| $t_c$ | 2007.11.29 | | 2009.1.1 | | 2009.12.17 | | 2010.11.29 | | 2011.12.29 | |
| $\tau^{(みかん)}$ (days) | 29 | 55 | 91 | 26 | 63 | 45 | 30 | 71 | 56 | 24 |
| $B^{(みかん)}$ | 2319 | 2416 | 1555 | 1491 | 1858 | 1950 | 1706 | 1510 | 1609 | 1943 |
| $n^{(みかん)}$ (days) | 60 | 121 | 97 | 84 | 82 | 84 | 48 | 76 | 88 | 69 |

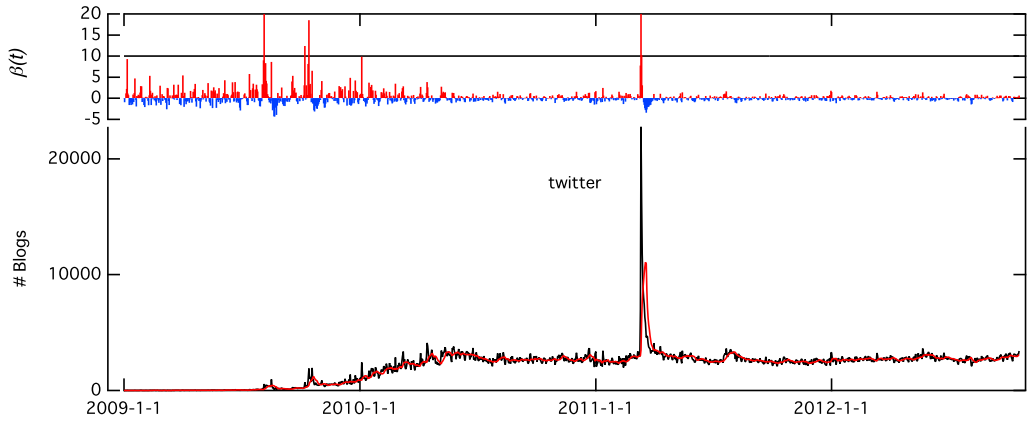


図 4.16　(下段)「Twitter」の書き込み時系列 (黒)．赤線が過去 7 日間の移動平均線．(上段) 式 (4.13) で求めた逸脱度 $\beta(t)$．2011 年 4 月に Twitter に関連するテレビ番組の放送が開始され話題を呼んだため鋭いピークを持つ．書き込みが少ない場合，$\beta(t)$ が大きな値を持つ．

### 投稿数が少ない場合

書き込み数が 1 日に 1 件や 2 件と非常に少ない場合，ゆらぎの影響で相対的に逸脱度が大きくなる．図 4.16 から図 4.18 は，それぞれ書き込み数が全くなかった状態から，広まってきた単語「Twitter」「Facebook」「iPS 細胞」である．1 日あたりの書き込みが 0 件や 1 件の場合で，逸脱度 $\beta(t)$ が大きく影響を受けていることが分かる．

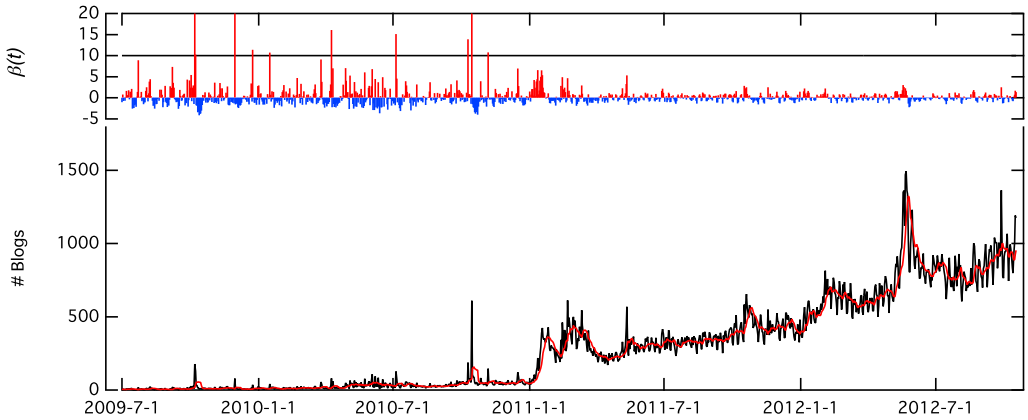


図 4.17　(下段)「Facebook」の書き込み時系列 (黒)．赤線が過去 7 日間の移動平均線．(上段) 式 (4.13) で求めた逸脱度 $\beta(t)$.

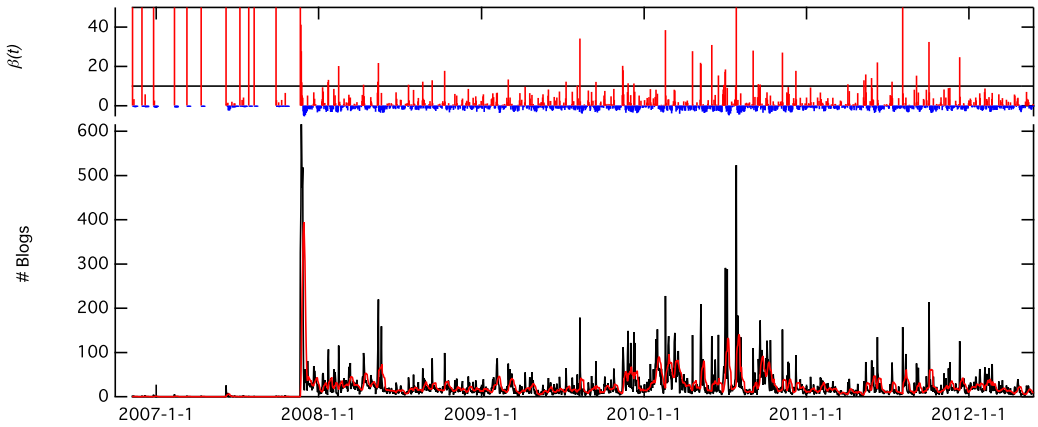


図 4.18　(下段)「iPS 細胞」の書き込み時系列 (黒)．赤線が過去 7 日間の移動平均線．(上段) 式 (4.13) で求めた逸脱度 $\beta(t)$.

## 4.6　まとめ

本章では，時系列からの異常値検出手法を確立するため，形容詞，連体詞，接続詞の出現頻度の時系列に対して弱定常性であると判定された語を「日常語」とし，そのゆらぎの統計性の解析を行った．その結果，時系列の平均値に応じたゆらぎ (標準偏差) のスケーリング則を指摘した．スケーリング則では，単語の出現過程において，その出現頻度が低い領域ではポアソン過程で記述できるが，出現頻度の高い領域ではゆらぎが平均値に対して非自明に線形に増加する．先行研究 [65] ではブログ上に現れる単語は，出現頻度によ

らず，単純なポアソン過程では記述できないとされていたが，本研究では，規格化などの前準備を行い，弱定常性の性質を持つ単語で絞り込むことによって，出現頻度の低い領域では，ポアソン過程に従うことを指摘した．平均値とゆらぎのスケーリング則は Taylor's scaling law として 1960 年代より知られており，ブログだけではなく生態系における個体数のゆらぎなどの自然現象でも観測することができる．そこで Taylor's scaling law を再現するランダム拡散モデルを始点に，ブログにおける時系列のゆらぎを再現するランダム投稿モデルを提案した．ランダム投稿モデルでは，注目する単語 $k$ を書き込むかというゆらぎの他，ブログを投稿するか否かのゆらぎの重ね合わせによって表現される．後者のブログ投稿自体のゆらぎは，出現頻度の高い単語に対して影響を強く及ぼし，その結果，標準偏差が平均値に対して線形に増えるスケーリング則が現れる．

ランダム投稿モデルでは，平均値から標準偏差を導くことができるため，それを応用して，実際の商品名等の時系列を用いた異常値検出を行うことができる．ランダム投稿モデルを用いた異常値検出法は，非定常な大きなトレンドを持つ単語の検出には不向きだが，局所的に鋭いピークを持つ単語のピークの検出に有効であることを示した．

# 第 5 章

# 有限時間発散する時系列の解析とモデル化

本章では，前章で検出した異常値を持つブログ時系列のうち，特にベキ関数的に変動をする単語の出現頻度時系列を取り上げる．ベキ関数的に変動する具体例として，「海の日」「バレンタインデー」など毎年イベント日が決まっている単語，「5 月 18 日」などの日付，「津波」など大きなニュースに関連する単語がある．これらの単語の出現頻度時系列は，ピークとなる日を発散点とした，任意の指数を持つベキ関数で表すことができる．この任意のベキ指数は，ブロガーが時間に対して非線形に反応する効果と，前日に投稿された書き込み数からのフィードバック効果の二つが反映されていることを，数理モデルから示す．最後に，ベキ関数的な変動を応用する例を取り上げる．

## 5.1 導入：ベキ関数が見られる現象

大森則 (Omori's law) は，大きな地震の後の余震回数が，本震からの経過時間に対してベキ関数的に減少していくという，よく知られた経験則である [51]．大森則は日本の地震学者である大森房吉によって 1890 年代に報告された．大森は本震からの経過時間に対して逆数で余震回数が減る，すなわち-1 のベキ指数を持つベキ関数的な変動としていたが，宇津によって後に任意のベキ指数を持つベキ関数に拡張された [74]．地震の中で，本震自体は，プレートにかかった力により生じた歪みが断層で解放されることによって起こる．余震は，本震時に解放されなかったエネルギーが，連鎖的に他の断層で起こることによって起こると考えられている．この連鎖的な反応の結果，余震回数はベキ関数的な変動をする．

このような連鎖的な反応がベキ関数的に起きる現象は，地震だけではなく，太陽フレア [75] や犬の肺の動き [76] や人間の心臓の鼓動パターン [77] にも見られる．

ベキ関数的な変動は，相転移現象とも関連が深い．一例として，自発的な磁化を持ち強磁性を示す物質が，臨界温度 (転移温度，キュリー温度)$T_c$ で急速に磁化を持たない常磁性の物質となる現象がある．このときの磁化 $m_s(T)$ は，温度 $T$ に対しベキ関数的に変化する．

$$m_s(T) \simeq (T_c - T)^{\beta} \tag{5.1}$$

臨界温度 $T_c$ は物質によって異なるが，$\beta$ で表される臨界指数は，多くの物質で共通の値を持つことが知られている．

地震や太陽フレア，磁性の変化は人間がコントロールすることのできない現象だが，個別の意思を持つ人間が構成する社会現象でも，ベキ関数的な現象が報告されている．代表的なものに，金融の世界で起こる通貨のハイパーインフレーション [78, 79]，株の暴騰・暴落現象 [80, 81, 82, 83] がある．これらは，余震の場合のように，人々の期待や不安が連鎖的に大きくなり，その結果，ベキ関数的な変動が生じているのではないかと指摘されている．類似の現象で，大きなニュースが起こり，その後の人々が関心がベキ関数的に減少する様子は，インターネットのポータルサイトに掲示された新しい話題へのアクセス数 [84]，論文のダウンロード数 [85] などにも見られ，人間の集団行動の背後にも何らかの普遍性が潜んでいることが期待されている．

動画サイト (YouTube) では，人々の注目を集めた後の再生回数の減り方が，内的，外的な要因によって異なる指数を持つベキ関数で記述できるということが報告された [86]．さらにそのベキ指数の違いは，その動画自体がもつ価値に依存するとしている．動画の価値は，ある期間内の再生回数に占める，ピーク時の再生回数で特徴づけることができる．ある期間内の再生回数のうち，ピーク時の再生回数が少ないほど，動画の価値は高いとする．すなわち，口コミなどの内的要因でじわじわと再生回数が増えた動画は，そのピーク前後の期間での再生回数も継続的に多いため，再生回数の減り方が緩やかで，ベキ指数の絶対値が小さくなる．一方，たまたまサイトのトップページ等に掲載されたなどの，外的要因で再生回数が増えた動画は，ピーク時の再生回数が多いが，その前後ではすぐに視聴されなくなるため，ベキ指数の絶対値は大きな値を持つ．

ウェブサイト (Amazon.com) での本の売り上げ順位変動もまた，内的，外的な要因により，その変動を異なる指数を持つベキ関数で記述できることが知られている [87]．例えば発行部数が大きい新聞の書評で紹介された本は，その直後，急激に販売部数が大きく跳ね上がりピークを持つ．その後は日が経つにつれ，販売部数が減少していく．これは外的要因による売り上げ変動の典型例で，ベキ関数で記述できる．他方，内的要因の代表例として，じわじわと口コミで人気が広がり，販売部数が伸びる本もある．この場合は，部数

の増え方と減り方が，外的の場合よりは絶対値の小さい指数を持つベキ関数的な変動をする．内的要因の場合でベキ指数が非常に小さい場合は，指数関数的であるという指摘 [89] もあるが，内的，外的の要因の違いで異なる変動を示すことは共通している．

オバマ大統領が初当選した，2008 年のアメリカ大統領選におけるブログデータの解析においてもベキ関数が観測されている [52].「サラ・ペイリン (Sarah Palin)」が共和党の副大統領候補に指名されたとき，彼女の名前を含むブログ数が急増し，その後ベキ関数的に減少した．これは [86] や [87] で提案された分類において，外的要因に相当する．対して，内的要因の例に「大統領就任式 (inauguration)」を含むブログ数が，ベキ関数的に増加，減少することが指摘された．内的，外的要因どちらの場合でもベキ指数の大きさは，絶対値で 0 から 2.5 の間の値をとる．また，内的要因のピーク前後のベキ指数にも強い相関 (相関係数 0.67) がある．さらに，大森則の他に，任意の検索語を含むブログ数を規模と定義し，その分布がベキ分布になることから，地震の規模 (マグニチュードの大きさ) とその発生回数の関係がベキ分布になるという Gutenberg-Richter 則と似ていることも指摘している．

締め切り前の人々の行動にもベキ関数的な変動が見られる．ある国際会議の主催者が，約 1500 人の参加者データを解析したところ，参加の締め切り日に向けて，累計の申し込み者数が非線形に増加することを指摘した [88]．この現象は，参加者の申し込み確率を，締め切り日までの残り日数に反比例すると仮定した数理モデルで再現できる．

これら，社会現象にあらわれるベキ関数を再現するモデルは，5.5 節で詳しく紹介する．

## 5.2 ベキ関数的な変動をする語

本章ではブログ変動の中でも，ベキ関数的な変動をするものを取り上げる．ベキ関数的な変動に共通する事項は，要因があった日，締め切り日などピークとなる特異日 $t_c$ があることである (図 5.1)．そこで，ピークを $t_c$ として，ピーク前の期間における単語 $k$ を含むブログ数変動 $w^{(k)}(t)$ を以下のベキ関数で記述する．

$$w^{(k)}(t) = A^{(k)}(t_c - t)^{-\alpha^{(k,\mathrm{fore})}} + w_0^{(k)} \tag{5.2}$$

ピーク $t_c$ 後の期間における変動を

$$w^{(k)}(t) = A^{(k)}(t - t_c)^{-\alpha^{(k,\mathrm{after})}} + w_0^{(k)} \tag{5.3}$$

と記述する．ベキ指数 $\alpha^{(k)}$ が，変動を特徴付ける．$\alpha^{(k)}$ の絶対値が大きいほど，変動は急激になり，ピーク前の変動であれば，ピーク直前に盛り上がることに対応し，ピーク後の変動であれば，すぐに忘れ去られると解釈できる．

ベキ関数的に変動する語を網羅的に集めるため，以下の 3 カテゴリーで単語を定義し，

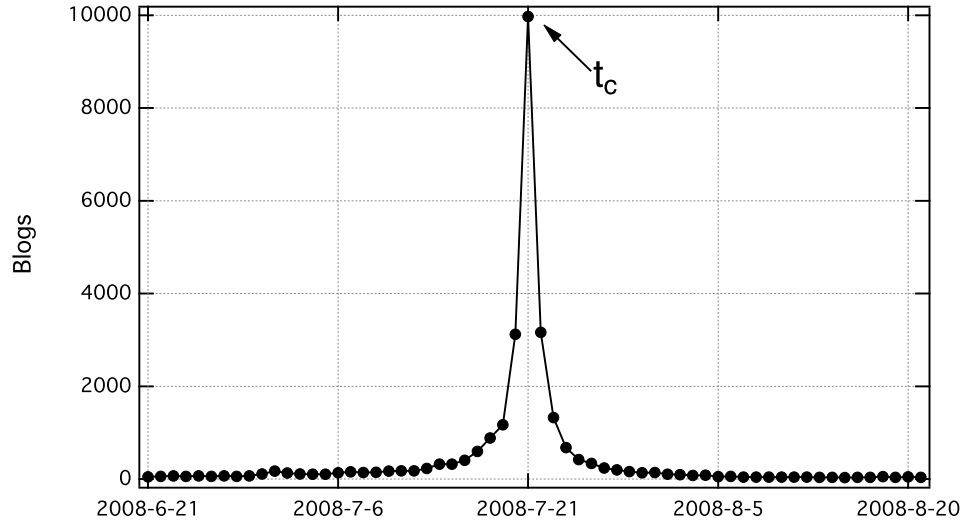


図 5.1　2008 年「海の日」の書き込み時系列．海の日当日である 2008 年 7 月 21 日にピークを持つ．

データを収集した．時系列は 2006 年 11 月 1 日から 2010 年 10 月 31 日を対象とした (詳細は付録 C).

- **イベント語**　「海の日」などの日本の祝日の名前と，主な年中行事の名前．ウェブ上の百科事典である wikipedia 日本語版[*1]の「国民の祝日」に掲示されている 14 の祝日名と，「日本の年中行事」に掲示されている行事のうち，クリスマスやバレンタインデーなど 16 のイベント名，合わせて 30 単語の時系列を収集した．4 年分のデータの中にはどのイベントも少なくとも大きな 4 つのピークが含まれる．ただし，祝日のうち．2007 年に祝日の日付が 4 月 29 日から 5 月 4 日に移動した「みどりの日」に関しては，時系列が明確なピークを持っていないため，サンプルからは外している．

- **日付**　「5 月 18 日」などの日付．地域行事の告知，拾得物の情報，誕生日，試験日など日付の入った情報がブログ記事本文では提供されている．それらの日付を検索すると，出現頻度の時系列はいつも検索日付当日にピークを持つ．(ブログ記事が持つタイムスタンプ自身は検索対象とならない．)

- **ニュース語**　「津波」「マイケル・ジャクソン」などの大きなニュースに関する単語．

[*1] http://ja.wikipedia.org/wiki/

2006 年 11 月から 2010 年 10 月の間に起こった，大きな地震の地名を wikipedia 日本語版の「地震の年表」から 11 単語抽出し，「地震」かつ「地名」で検索をした．もう 1 点は、2007 年から 2010 年にかけての訃報者名を同じく wikipedia 日本語版の「訃報 2007 年」などから年ごとに抽出して，人名を検索した．このリストには訃報者の名前が毎年約 500 名分掲示されているが，ブログに書き込まれるためには知名度が必要なので，その中で約半数にあたる，wikipedia 内に当人のページが存在する人に限って検索をした．このように絞り込んでも，ブログの出現頻度において，多くの場合，知名度は低く，平均的な出現頻度は 0 に近い．このような単語はパルス的に書き込みがあるのみで，変動を追うのは困難になるため，サンプルから除外している．さらに 2008 年のノーベル賞受賞者の 4 名分も対象としている．

## 5.3 モデルの見積もり方法

時系列をベキ関数的で記述するにあたり，統計的に有意にどの範囲までがベキ関数的だと言えるのか，そのときの有意確率はどう求めるのか，といった問題に対して確立された方法があるわけではない．そこで，本節では金融時系列解析において導入された，ブートストラップ法 [90] を元にした Preis らの手法 [83] をもとに，モデルを正確に見積もる手法を提案する．

はじめに，他の関数と比較して，データがベキ関数的な変動かを確認するため，非線形関数の代表例である指数関数と比較することを行う．そこで，以下のベキ関数と指数関数でそれぞれデータからパラメータを見積もる．

$$x^{(k)}(t) - x_0^{(k)} = A^{(k)}|t_c - t|^{-\alpha^{(k)}} \tag{5.4}$$

$$x^{(k)}(t) - x_0^{(k)} = B^{(k)} \exp\left(-\frac{|t_c - t|}{\tau^{(k)}}\right) \tag{5.5}$$

パラメータ $A^{(k)}$，$\alpha^{(k)}$，$B^{(k)}$，$\tau^{(k)}$ の見積もりには非線形最小二乗法の一つでよく知られている Gauss-Newton 法を用いる．$x_0^{(k)}$ は，平時の書き込み数で，平均値や中央値などが候補として考えることができるが，本研究ではピーク時の書き込み数の影響を受けにくくするため，全期間での中央値を用いる．

次に，二つのモデル式でどちらの方が，実際のデータに近いかをコロモゴロフ-スミルノフ検定 (Kolomogorov-Smirnov test，KS 検定)[*2]を用いて決定する．KS 検定での統計

[*2] 2 つの分布を累積値に直し，それらの差の絶対値によって行うノンパラメトリック検定，ベキ分布の検定で多く使われる．

検定量 $D$ は，以下で求める．

$$D = \max_{t\in[t_c\pm1,t_c\pm n]} \left| X^{(k)} - W^{(k)} \right| \tag{5.6}$$

ここで，$X^{(k)}$ は式 (5.4) と式 (5.5) で見積もったモデル式の累積数で，$W^{(k)}$ は実データの累積数である．モデルを適用する範囲は，ピーク前の場合 $[t_c - n, t_c - 1]$，ピーク後の場合 $[t_c + 1, t_c + n]$ とする．ベキ関数のモデルで見積もった時の KS 検定量 $D$ の方が小さい場合のみ，ベキ関数のモデルを候補として残す．

ベキ関数が候補として残った場合，ブートストラップ手法で $p$ 値を計算し，最終的にベキ関数のモデルを採用するかを決定する．まず乱数を用いて，人工的にベキ関数な変動をする時系列を生成する．乱数は正規分布から発生させ，正規分布の平均値はモデルの式 (5.4) より $x^{(k)}$，標準偏差 $\sigma$ はランダム投稿モデルの式 (4.9) に $x^{(k)}$ を代入して求めた $\sigma = \sqrt{x^{(k)}\left[1 + \left(\frac{\delta}{w}\right)^2 x^{(k)}\right]}$ を使う．ここで全数揺らぎに起因するパラメータは，$\frac{\delta}{w} = 0.08$ とした．これは形容詞のデータに対して式 (4.9) を直接当てはめて求めた値である．この結果，一つの人工的な時系列 $y_i^{(k)}(t)$ ができる．$y_i^{(k)}(t)$ と式 (5.4) の $x^{(k)}(t)$ で KS 検定量

$$D' = \max_{t\in[t_c\pm1,t_c\pm n]} \left| X^{(k)} - Y_i^{(k)} \right| \tag{5.7}$$

を算出する．$Y_i^{(k)}$ は $y_i^{(k)}(t)$ の累積数である．これを $i = 1 \cdots M$ まで $M$ 回繰り返し，$D \leq D'$ となる場合の数 $m$ を数え上げ，$q = \frac{m}{M}$ として有意水準として扱う．$M$ の値で有意水準の有効数字を決定できるが，ここでは 0.1% 水準まで求めれば十分なので $M = 1000$ とした．$D \leq D'$ となる場合が 100 の場合，$q = 100/1000 = 0.1$ としてベキ関数のモデルを採用する．$D \leq D'$ となる場合が 200 だと $q = 0.2$ とし，さらに実際のデータとモデルが有意に近いと解釈する．すなわち，ここで $p$ 値の代わりに使う $q$ は，通常の $p$ 値と異なり，$q$ が大きい場合に，モデルが実際のデータを記述できていると解釈する．

最後に，この手順を $(t_c \pm 1)$ 日目から $(t_c \pm n^{(k)})$ 日目まで，モデルを適用する期間を延長して行い，$q \geq 0.1$ であると判定された最長の期間を $n^{(k)}$ として採用した．ただし，モデルを適用するに際して，$n^{(k)}$ の最小値は 5 とした．

## 5.4　ベキ関数的な変動

図 5.2 は，「海の日」の書き込み時系列をベキ関数の式 (5.4) に当てはめた例である．時系列は式 (3.5) で規格化して，式 (3.11) で概日周期の除去を行なっている．実線が式 (5.4) のモデル線であり，印が実際のデータである．ピークである 2008 年 7 月 21 日を $t_c$ として，$t_c$ までの残り日数または経過日数を横軸に取り，両軸対数で表示してい

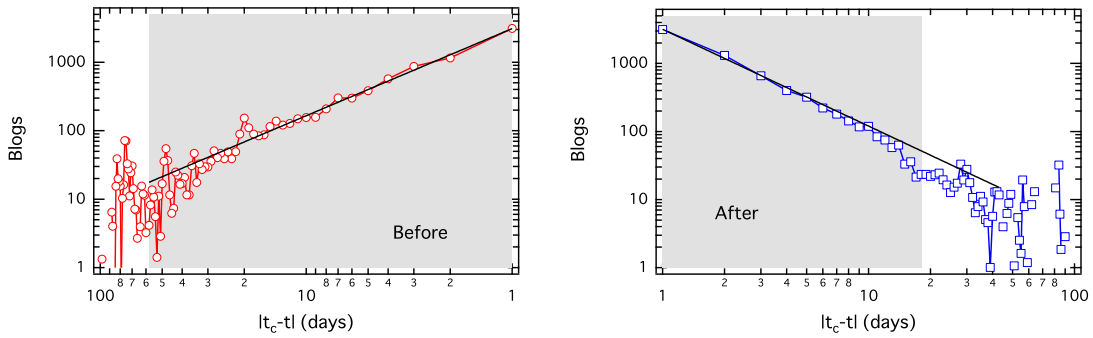


図 5.2　2008 年の「海の日」の書き込みに対するモデルの見積もり例 (両軸対数)．網掛けの部分はベキ関数的に変動する期間 ($n^{(k)}$ 日) である．(左図) ピーク前の変動．($\alpha^{(\text{海の日,fore})} = 1.27, A^{(\text{海の日,fore})} = 3100, n^{(\text{海の日,fore})} = 58$)(右図) ピーク後の変動．($\alpha^{(\text{海の日,after})} = 1.42, A^{(\text{海の日,after})} = 3171, n^{(\text{海の日,after})} = 18$)

る．ピーク前の場合，ベキ関数的に変動する期間 $n^{(\text{海の日,fore})}$ は 58 日間，ベキ指数は $\alpha^{(\text{海の日,fore})} = 1.27$ となった．ピーク後の場合，ベキ関数的に変動する期間 $n^{(\text{海の日,after})}$ は 18 日間，ベキ指数は $\alpha^{(\text{海の日,after})} = 1.42$ であり，$\alpha^{(\text{海の日,fore})}$ よりもやや大きい．

### 5.4.1　ベキ指数

表 5.1 に見積もったベキ指数の一覧，図 5.3 と図 5.4 にそれらの分布を示す．ベキ指数 $\alpha^{(k)}$ はピークの前後で共に $0.1 \leq \alpha^{(k)} \leq 2.5$ の値を取る．時系列ごとに $\alpha^{(k)}$ をピーク前後で比較すると，イベント語の場合，ピーク前後共に，ベキ関数と判定された 65 時系列のうち 38 サンプル，日付の場合は 603 時系列のうち 486 サンプルで $\alpha^{(k,\text{fore})} < \alpha^{(k,\text{after})}$ となる．ピーク前後での $\alpha^{(k)}$ 違いは，$t$ 検定による結果，イベント語の場合は有意ではないが，日付の場合は $p < 2 \times 10^{-16}$ で有意にピーク後の方が大きい．

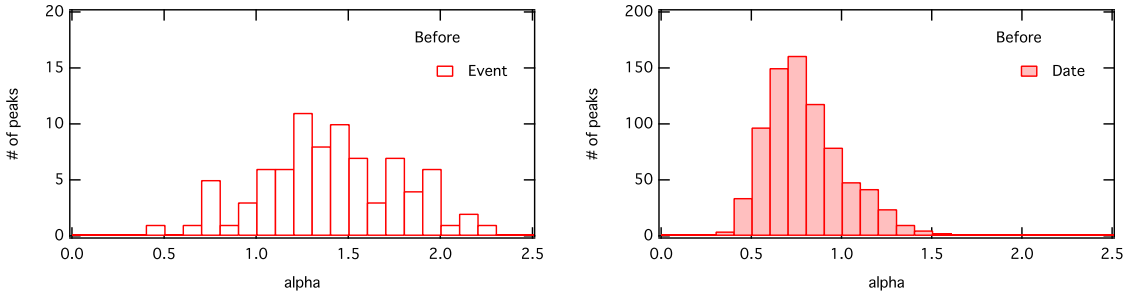


図 5.3　ピーク前のイベント語 (左図)，日付 (右図) のベキ指数 $\alpha^{(k)}$ の分布．

ニュース語の時系列はイベントや日付と違い，ピーク後のみベキ関数的な変動をする．

表 5.1　ベキ指数 $\alpha^{(k)}$ の平均値と標準偏差．カテゴリーによって偏りはあるが，$0.1 \leq \alpha^{(k)} \leq 2.5$ の値をとる．ベキ関数的に変動した期間 $n^{(k)}$ は中央値．

| | | $\alpha^{(k)}$ | $n^{(k)}$ (days) | # samples |
|---|---|---|---|---|
| Event | Before | 1.40 ± 0.38 | 10 | 83 |
| | After | 1.44 ± 0.28 | 16 | 91 |
| Date | Before | 0.79 ± 0.38 | 9 | 776 |
| | After | 1.11 ± 0.16 | 21 | 1229 |
| News | After | 1.09 ± 0.45 | 10 | 21 |
| All | Before | 0.85 ± 0.30 | 9 | 859 |
| | After | 1.13 ± 0.21 | 20 | 1341 |

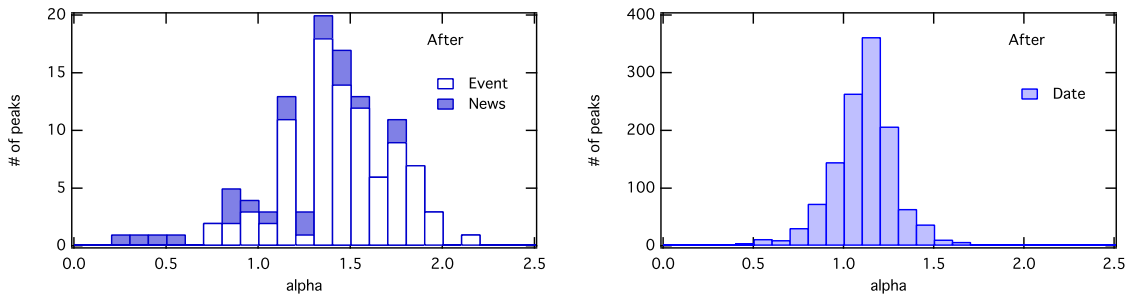


図 5.4　ピーク後のイベント語とニュース語 (左図)，日付 (右図) のベキ指数 $\alpha^{(k)}$ の分布．

さらに，ニュース語の時系列は TV などの外的要因によって大きく揺らぎ，前節で導入した解析手法でベキ関数が採用されるものが少ない．例えば，2009 年 6 月マイケル・ジャクソンが亡くなったという報道の後は，ニュース 10 日後の告別式のニュース，3 ヶ月後の映画の公開ニュース等，続けて外的なニュースが入り，その度に小さなピークが現れる (図 5.5)．そのため，ベキ指数の見積もりも，それら 2 番目，3 番目のピークの影響を大きく受け，ベキ関数が棄却されるか，ベキ指数の絶対値が小さな値を取りがちである．最終的にベキ関数を有意に観測できた 21 サンプルでのベキ指数の絶対値の平均は，1.09 となっている．

本の売り上げ時系列の先行研究 [87] と比較すると，ブログ上でベキ関数的に変動する単語の例は，外的要因があった場合の本の売り上げ変動のベキ指数 $\alpha \simeq 0.75$ に近い．内的要因の場合はベキ指数が $\alpha \simeq 0.4$ となるので，わずかにそちらに近いものがあるが，多くの単語でベキ指数の絶対値は 0.5 より大きい．また，さらに [87] では見られなかった $\alpha \geq 1$ も数多く観測できる．これはニュース語には明らかに外的要因が対応するが，イベ

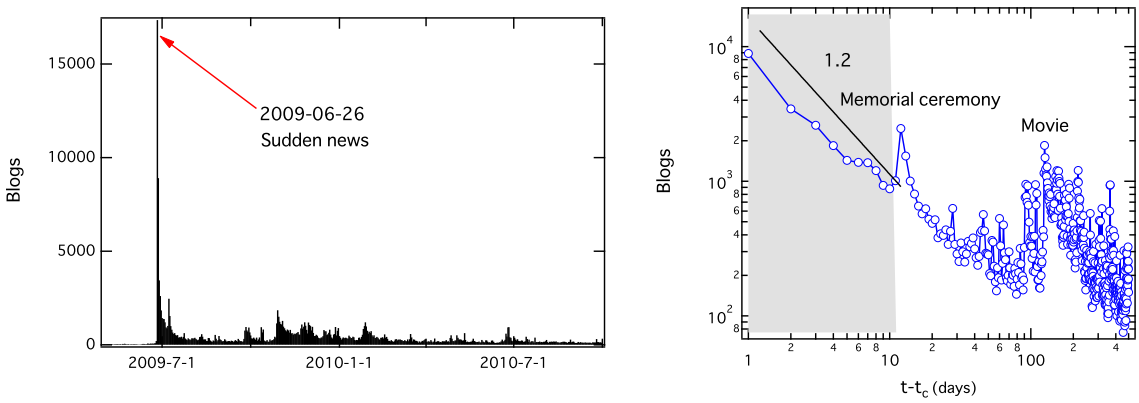


図 5.5　(左図)「マイケル・ジャクソン」の書き込み時系列．(右図)$t_c$ =2009.06.29 としたときの「マイケル・ジャクソン」の書き込み数変動．$\alpha^{(\mathrm{MJ,\ after})} = 1.2$，$n = 10$ となり，2 番目のピークの手前までのベキ関数が採用される．

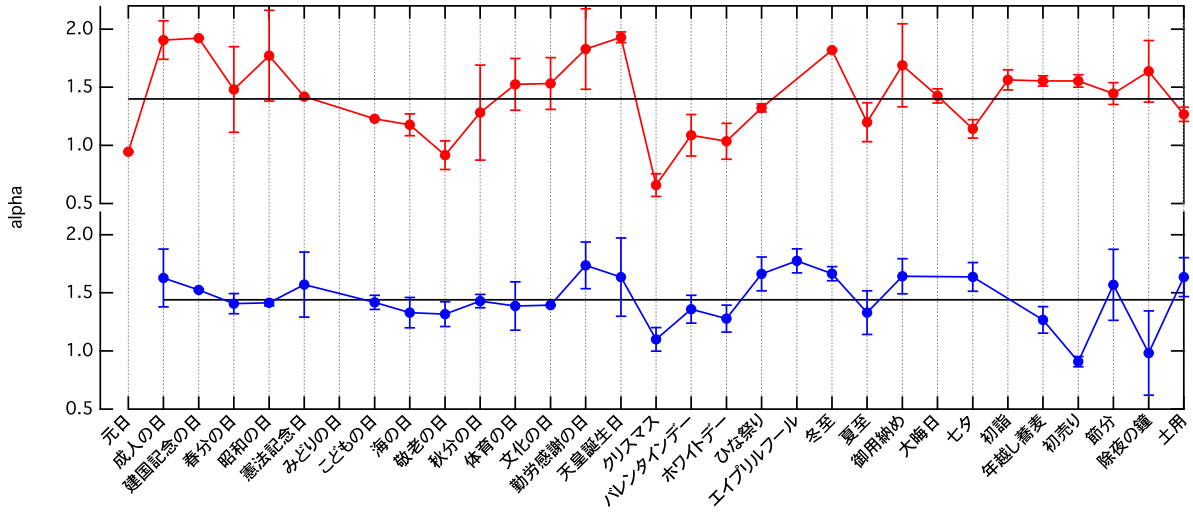


図 5.6　イベント語のピーク前 (上段) 後 (下段) でのベキ指数の比較．黒の実線はそれぞれベキ指数の平均値 $\alpha^{(\mathrm{fore})} = 1.40$ と $\alpha^{(\mathrm{after})} = 1.44$ に対応する．

ント語や日付もまた外的要因が作用していることを示唆している．

イベント語において，採用されたベキ指数の値を単語ごとに平均して示したものが図 5.6 である．データ点が無い部分は，ベキ関数が棄却され，サンプルがとれなかったことに起因する．イベント語は 4 年間分のデータからは，最大 4 点とれるので，それらの標準偏差からエラーバーは算出した．年中行事の「クリスマス」のみ，イベント前後どちらの場合でも突出してベキ指数は小さい．

### 5.4.2　ベキ関数的に変動する期間

ベキ関数的に変動すると判定された期間 $n^{(k)}$(日) の分布が図 5.7，図 5.8 である．$n^{(k)}$ の最小値は 5.3 節で設定した条件より，5 日となっている．図から明らかなように，ベキ関数的に振る舞う期間は短い場合が多く，全体として $n^{(k)}$ の中央値はピーク前の場合は 9 日，ピーク後の場合は 20 日である (表 5.1)．

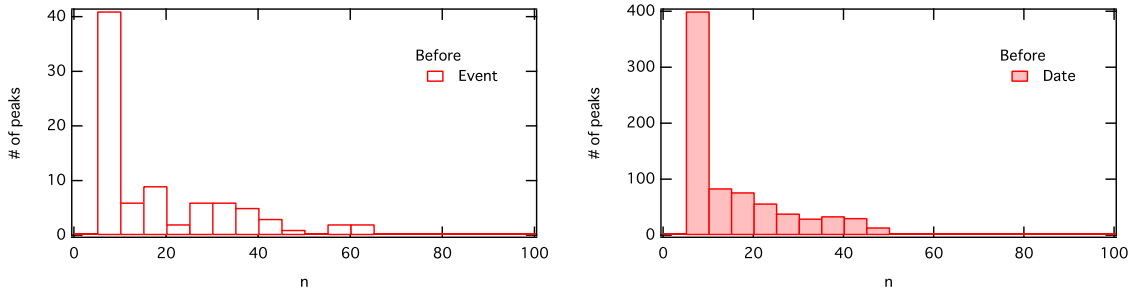


図 5.7　ピーク前のイベント語 (左図)，日付 (右図) のベキ関数的に変動する期間 $n^{(k)}$(日) の分布．

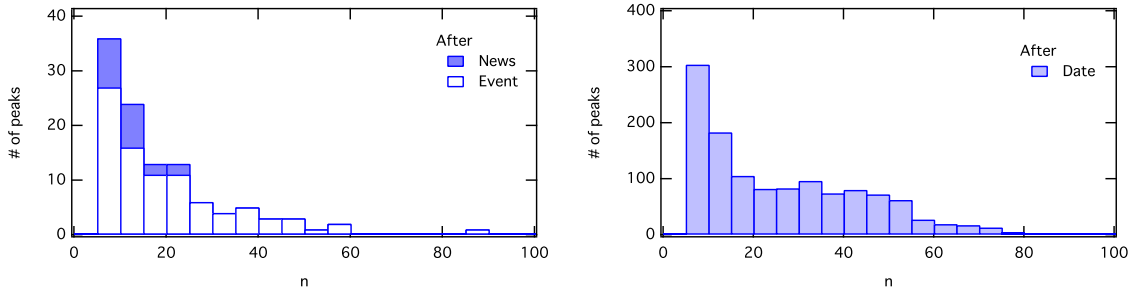


図 5.8　ピーク後のイベント語とニュース語 (左図)，日付 (右図) のベキ関数的に変動する期間 $n^{(k)}$(日) の分布．

式 (5.4) を延伸することで，およその $n^{(k)}$ を見積もることはできるが (5.6 節)，本解析の場合は統計的に有意にベキ関数と判定された場合に限定しているために，$n^{(k)}$ の値が小さい．

イベント語において，$n^{(k)}$ の日数を単語ごとに平均して示したものが図 5.9 である．データ点が無い部分は，ベキ関数が棄却され，サンプルがとれなかったことに起因する．

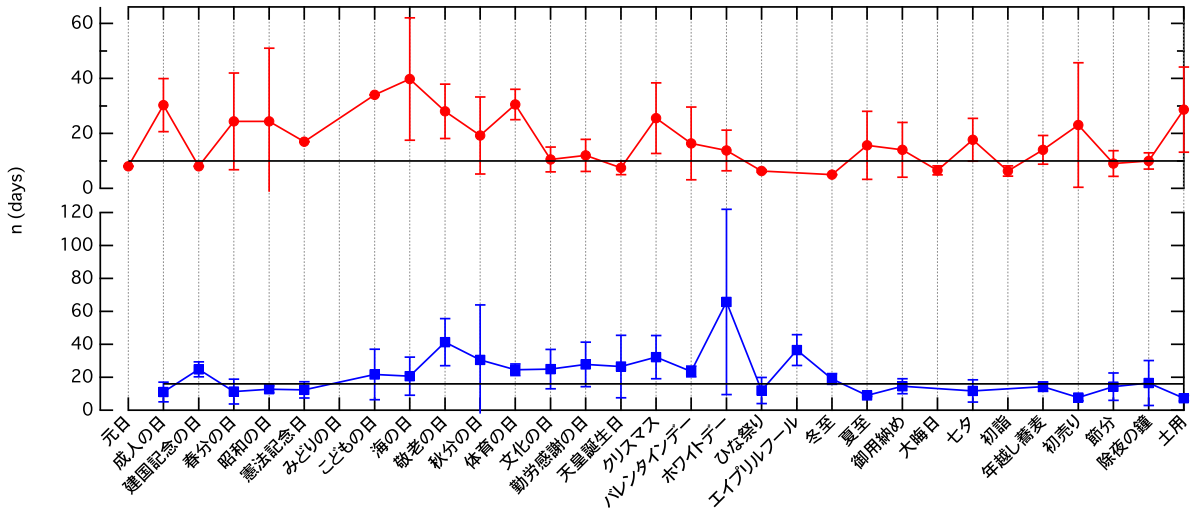


図 5.9　イベント語のピーク前 (上段) 後 (下段) でのベキ関数的に変動すると判定された期間 $n^{(k)}$(日) の比較．黒の実線はそれぞれの中央値 $n^{(\mathrm{fore})} = 10$ 日と $n^{(\mathrm{after})} = 16$ 日に対応する．

### 5.4.3　ピーク前後での比較

Klimek らの研究 [52] では，アメリカ大統領選に関するブログ記事のうち，内的要因で増えた 150 時系列において，ピーク前後のベキ指数を比較したところ相関係数が $0.67(p < 10^{-22})$ と報告されている．本研究においても，イベント語と日付において，同じピーク前後で両方がベキ関数が採用された時系列に絞り込んで比較を行ったのが図 5.10 である．比較可能なイベント語は 65 サンプル，日付は 603 サンプルでベキ指数 $\alpha^{(k)}$ 相関を確認したところ，式 (3.6) を使った相関係数 $\rho$ で，イベント語の場合 $\rho = 0.29$ ($p$=0.018)，日付の場合は $\rho = -0.10$ ($p$=0.018) となった．Klimek らの結果とは異なり，イベント語と日付において，ピーク前後のベキ指数の値に統計的に有意 ($p < 0.01$) な相関は見られない．

また，同様にベキ関数的に変動した期間 $n^{(k)}$ もピーク前後での相関を確認したが，イベント語の場合 $\rho = 0.28$ ($p$=0.023)，日付の場合は $\rho = -0.070$ ($p$=0.084) どちらも有意な相関は見られない (図 5.11)．

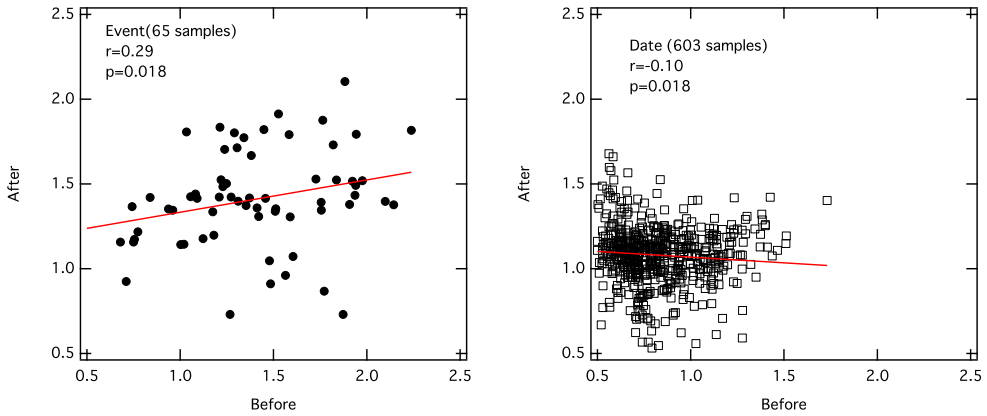


図 5.10　ピーク前後のベキ指数 $\alpha^{(k)}$ の比較．イベント語の場合 (左図) と日付の (右図) 場合．

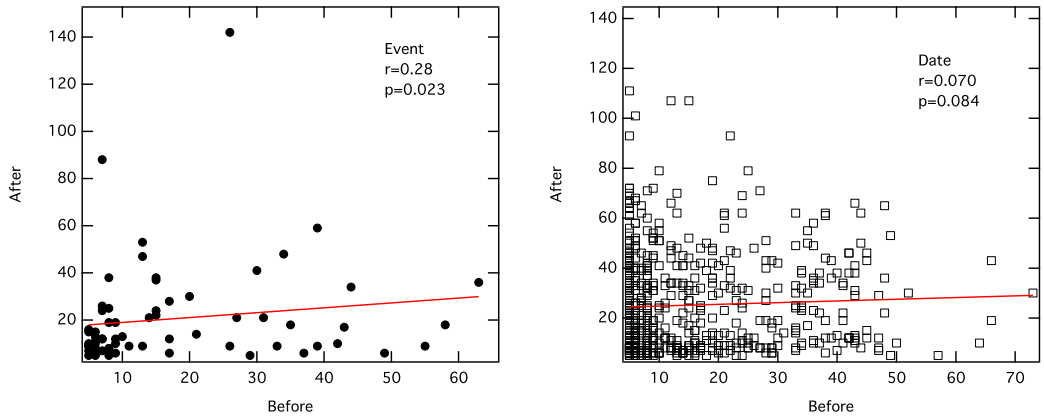


図 5.11　ピーク前後のベキ関数的な変動をする期間 $n^{(k)}$(日) の比較．イベント語の場合 (左図) と日付の (右図) 場合．

### 5.4.4　投稿者の内訳

では，誰がベキ関数的な変動に寄与しているのだろうか？詳細なブログデータ 2 から見ると，複数回書き込むリピーターではなく．次々に新規に参入してくるブロガーがベキ関数的な変動に寄与していることを明らかにできる．図 5.12 は，ブログデータ 2 を用いて，初回書き込み記事のみに注目して「海の日」の書き込み数を示したものである．初回書き込み記事だけでも，ベキ関数が現れる．

図 5.13 の上段は，詳細ブログデータ 2 を用いて，全投稿記事のうち，何割が新規参入者かを示した．図から明らかなように新規参入者の割合は多くの場合，半分を超えており，同一人物の複数回の書き込みよりも，新規加入者の書き込みが全体的にベキ関数に寄与していることが分かる．

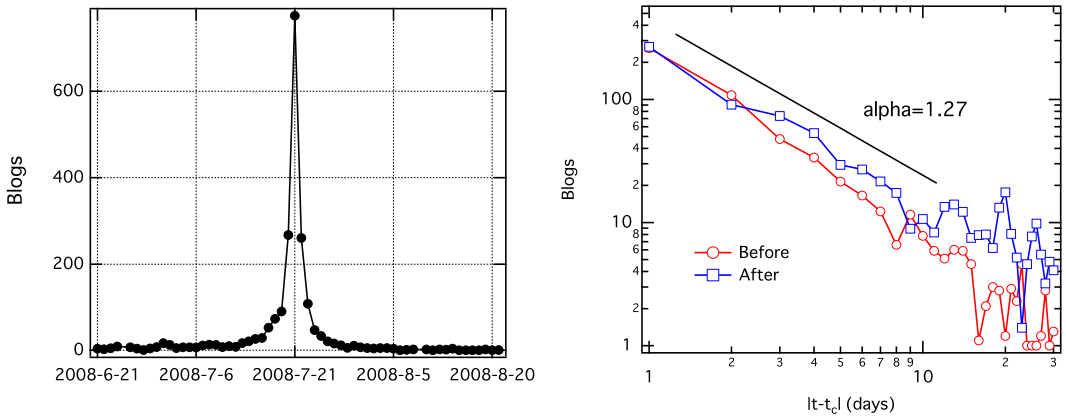


図 5.12　33 万人の詳細ブログデータ 2 を用いて初回書き込み者に限定した 2008 年の「海の日」の書き込み時系列．(左図) 両軸線形，(右図) 両軸対数による表示．初回書き込み者に限定しない場合 (図 5.2) と比較しても，ほぼ同じベキ指数が得られる．

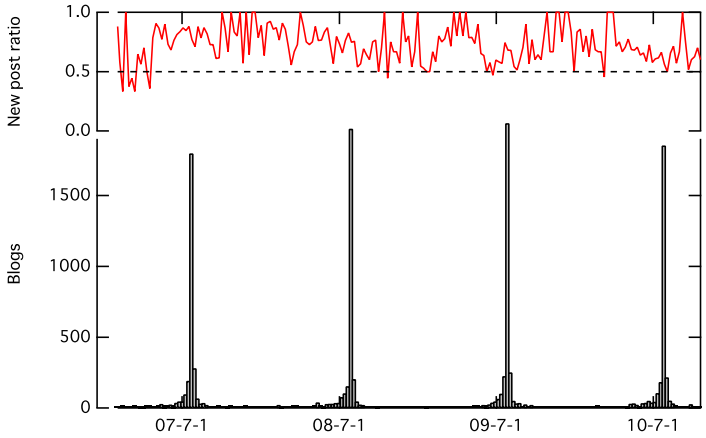


図 5.13　ブログデータ 2 で作成した「海の日」の書き込み時系列と新規参入者の割合．上段が新規参入者の割合．ほぼ全期間で 0.5 以上 (破線) になっており，新規書き込み者が大半を占めていることが分かる．

## 5.5　ベキ関数的変動を説明するモデル

### 5.5.1　先行研究

ピーク前のベキ関数的な変動を表すモデルに，時間に対する人の反応を，締め切りまでの残り時間に反比例で仮定するものがある．Alfi らは，自らが主催した国際会議への申し込み者数変動の時系列の解析の結果，締め切り日を $t_c$ として，時刻 $t$ での参加確率 $p(t)$

を以下のように定義することで，実際のデータを再現できることを示した [88].

$$p(t) = \frac{C}{(t - t_c)} \tag{5.8}$$

ここで $C$ は規格化定数であり，潜在的な申し込み者数に相当する．彼らは潜在的な申し込み者数 $C$ は，非常に大きいと仮定し，定数としている．本来ならば，潜在的な申し込み者数は，参加した時点で減って行くため $C(t)$ の減少関数になるはずである．しかし，彼らは $C$ を定数とする仮定をしつつ，実際の申し込み者数変動の再現を行い，締め切りに対して，人々が時間に対して非線形に反応していることを例示した．

ピーク後の飽きや忘却に関する人の変動を表すモデルに，人の記憶の減衰に任意のベキ関数を仮定するもの [86] がある．Crane らは基本的な人の記憶関数 (memory function)$A_{\mathrm{bare}}(t)$ を以下のように定義した．

$$A_{\mathrm{bare}}(t) \approx \frac{1}{(t - t_c)^{1+\theta}} \tag{5.9}$$

この仮定によるモデルは，インターネット上の動画サイト (YouTube) における再生回数の減少を記述するために用いられた．特に価値もない動画が偶然，トップページに表示されただけで再生回数が跳ね上がり，減少する場合を外的，逆に，動画サイト内で口コミ等でじわじわ人気がでて，ゆっくりと再生回数が増減する場合を内的と区別し，それらが異なるベキ指数 $\theta$ を持つという仮定がおかれた．

人の時間への反応が線形ではないという点では，1885 年エビングハウス (Ebbinghaus) の忘却曲線 (forgetting curve)[91] が心理学の分野でよく知られている．エビングハウスは，被験者に無意味な文字列を記憶させ，思い出せる文字数の時間変化を追った．実験では時刻 $t$ 経過後の記憶 $R(t)$ は指数関数的に減少すると結論している．

$$R(t) = \mathrm{e}^{-\frac{t}{S}} \tag{5.10}$$

ここで $S$ は記憶の強さで，エビングハウスのモデルは Crane ら [86] とは異なり指数関数を採用している．エビングハウスの実験の場合，記憶したのは無意味な文字列で，1 人の被験者での実験あったのに対し，Crane らの研究では，動画サイト内の多くの人の再生回数の重ね合わせである点で異なる．

### 5.5.2　ベキ関数的に変動するブログ時系列のモデル

前節までに見たベキ関数的に変動するブログの背後には，どのようなメカニズムがあるのだろうか．本節では，時系列からみる人間の集団としての振る舞いを，先行研究を元に簡単なモデルを考えることで理解を試みる．

ここで，われわれは，Alfi らのモデル [88] を元に，新たなモデルを提案する [C]．Alfi らのモデルだと，増加部分しか表すことができず，ベキ指数は 1 の場合しか記述できない．そこで，ブロガーの投稿行動に対して，われわれは Alfi らの効果を含め，以下の二つの効果を取り入れた．

1. **時間に対して非線形に反応する効果**
2. **書き込み数のフィードバック効果**

1. の時間に対して非線形に反応する効果は，ブロガーは自発的に「締め切り日にまでにブログを書き込む」，という心理的なプレッシャーを受けていると仮定して設定した．このような効果は「締め切り効果」と言われることがある．2. は，ブロガーは書き込む際に他のブログも参照していることが考えられる．そこで，すでに書き込まれたブログ数が多い程，書き込む確率が大きくなるフィードバック効果を考えた．

ブロガー $i$ が時刻 $t$ でピークとなる締め切り日 $t_c$ を持つ単語 $k$ を書き込む確率を $p_i^{(k)}(t)$ として，

$$\Delta p_i^{(k)}(t+1) \propto \frac{w^{(k)}(t)}{|t-t_c|} \tag{5.11}$$

と記述する．ここで $\Delta p_i^{(k)}(t+1) = p_i^{(k)}(t+1) - p_i^{(k)}(t)$ である．ブロガーの均質性を仮定すると，ブロガー $i$ の添字は省略でき，$w^{(k)}(t) \propto w(t) \cdot p^{(k)}(t)$ となる．時間幅 $\mathrm{d}t$ は 1 日であるので，ノイズ項 $f(t)$ を加え，$\alpha^{(k)}$ を比例定数とすると，締め切り前 $(t_c > t)$ の場合，

$$\frac{\mathrm{d}w^{(k)}(t)}{\mathrm{d}t} = \alpha^{(k)} \cdot \frac{w^{(k)}(t)}{(t_c - t)} + f(t) \tag{5.12}$$

という微分方程式の形に書き下せる．締め切り後 $(t_c < t)$ の場合は，減少することが自明なのであらかじめ負数の符号を付け，

$$\frac{\mathrm{d}w^{(k)}(t)}{\mathrm{d}t} = -\alpha^{(k)} \cdot \frac{w^{(k)}(t)}{(t - t_c)} + f(t) \tag{5.13}$$

となる．ノイズ項 $f(t)$ は無視して，これらの微分方程式を解くと，ピーク $t_c$ 前後で任意のベキ指数 $\alpha^{(k)}$ を持つブログ数変動を記述することができる．

$$w^{(k)}(t) \propto |t_c - t|^{-\alpha^{(k)}} \tag{5.14}$$

本モデルを実際のデータから直接確認する．式 (5.12) と式 (5.13) を離散型に直すと，締め切り前 $(t_c > t)$ 前は

$$\frac{\Delta w^{(k)}(t)}{w^{(k)}(t-1)} = \alpha^{(k)} \cdot \frac{1}{(t_c - t)} \tag{5.15}$$

締め切り後 $(t_c < t)$ の場合は，

$$\frac{\Delta w^{(k)}(t)}{w^{(k)}(t-1)} = -\alpha^{(k)} \cdot \frac{1}{(t-t_c)} \tag{5.16}$$

と書き下すことができる．$\Delta w^{(k)}(t) = w^{(k)}(t) - w^{(k)}(t-1)$ として，式 (5.15) と式 (5.16) をデータから直接確認したのが図 5.14 である．ベキ指数 $\alpha^{(k)}$ のわずかな違いで，大きく差が出てしまうため，ピーク前のエリア $(t < t_c)$ では，特に直前の書き込み数の変位では大きく差がでているが，その他の部分は，ピーク後のエリア $(t > t_c)$ を含め，理論線と実際のデータが似通っており，本モデルの妥当性を表している．

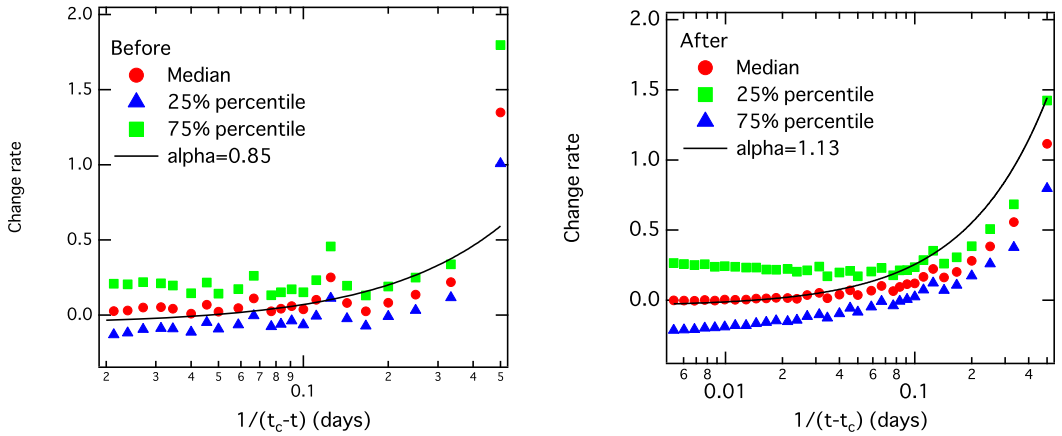


図 5.14　本提案モデル式 (5.15) と式 (5.16) をデータから直接確認した．実線が実データから見積もったベキ指数を持つ理論式に相当し，各点がデータから算出した中央値とパーセンタイル点に対応する．

## 5.6　応用

集団としての人々の行動が，ピークの前後ではベキ関数で書き下せる経験則を応用する例を三つ紹介する．

### 5.6.1　将来の書き込み数予測

将来の書き込み数予測は，式 (5.4) を用いて，なるべく誤差が小さくなるように，逐次的にベキ指数 $\alpha^{(k)}$ と切片 $A^{(k)}$ を見積もることで実現できる．

ブログ書き込み数は全数で規格化し，1 日の周期性を除去したのち，$[t_c - n_0, t_c - m]$ の範囲で式 (5.4) 式の $\alpha^{(k)}$ と $A^{(k)}$ を見積もる．ベキ関数のスタート地点 $t_c - n_0$ はここでは連続して 10 日間，中央値を超えた時点とする．

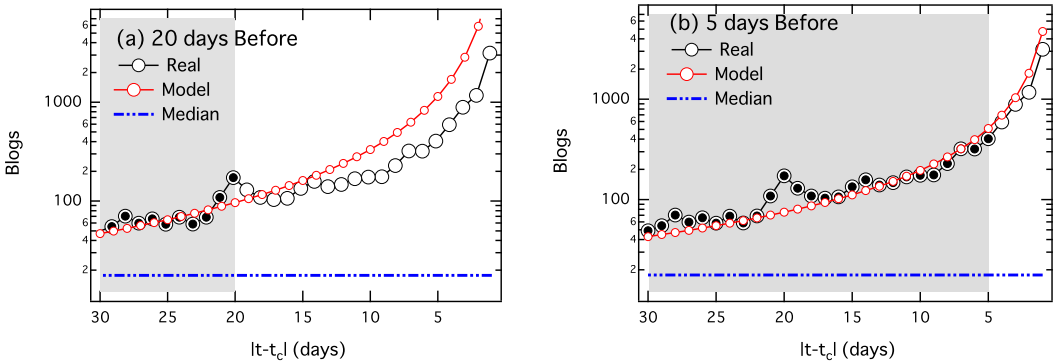


図 5.15　2008 年の「海の日」での書き込み数の予測例 (縦軸対数)．網掛けの部分は採用された見積もりに使った期間である．(右図) 海の日 ($t_c$) の 20 日前からの見積もり．(左図)$t_c$ の 5 日前からの見積もり．見積もりに使えるデータポイントが増えるほど，精度が上がる．

2008 年の「海の日」を例にすると，$t_c$ は 2008 年 7 月 21 日で，$n_0 = 76$ と判定されたので，$t_c$ から 76 日前の 2008 年 5 月 6 日からベキ関数の見積もりをスタートする．はじめは 2008 年 5 月 6 日から 2008 年 5 月 16 日までの 10 日間のデータを使い (すなわち $[t_c - 76, t_c - 67]$ の範囲)，$t_c$ は 2008 年 7 月 21 日となるようなベキ関数を見積もることを行う．しかし，この場合ベキ指数は収束せず見積もることができない．

そこで，見積もりに使うデータの範囲を，$m = 67, 66 \cdots, 1$ と 1 日ずつ逐次的に広げていくことで，徐々にベキ関数の $\alpha^{(k)}$，$A^{(k)}$ が収束していくことを確認できる．

図 5.15 は，図 5.1 で見た，「海の日」を含む書き込み数を事前予測した例である．2008 年は 7 月 21 日が海の日で，ピーク $t_c$ であるという情報は事前に与え，ベキ指数 $\alpha^{(\text{fore, 海の日})}$ と切片を $A^{(\text{fore, 海の日})}$ を見積もった．$t_c$ の 20 日前からの見積もり結果は $\alpha^{(\text{fore20days, 海の日})} = 1.78$ と $A^{(\text{fore20days, 海の日})} = 20222$ であり，5 日前からの見積もり結果は $\alpha^{(\text{fore5days, 海の日})} = 1.38$ と $A^{(\text{fore5days, 海の日})} = 4725$ である．最終的な結果は，$\alpha^{(\text{fore, 海の日})} = 1.27$ と $A^{(\text{fore, 海の日})} = 3100$ であったので，見積もりに使えるデータ数が多ければ，最終的に見積もられるベキ関数に漸近することを確認できる．

図 5.16 は，2007 年から 2011 年までの「海の日」の書き込み数データを用いて，見積もりに使ったデータ範囲 ($[t_c - n_0, t_c - m]$) を長くしていった時の，二乗誤差，ベキ指数 $\alpha^{(k)}$，切片 $A^{(k)}$ の変化である．$m = 1$ の時は，最終的なベキ関数の値となっている．どの年でも，見積もりに使える範囲の少ない，ピークから 50 日以上前から ($m \geq 50$) の範囲ではベキ指数が見積もりが困難であり，およそピークから 20 日以内 ($m \leq 20$) の範囲で $\alpha^{(k)}$ と $A^{(k)}$ が最終的な値に収束に収束する様子が分かる．同様にベキ関数のモデルが収束する様子を「こどもの日」で行ったものが図 5.17 である．見積もり範囲がピークの 20 日前程度にならないと，見積もることができないことが分かる．

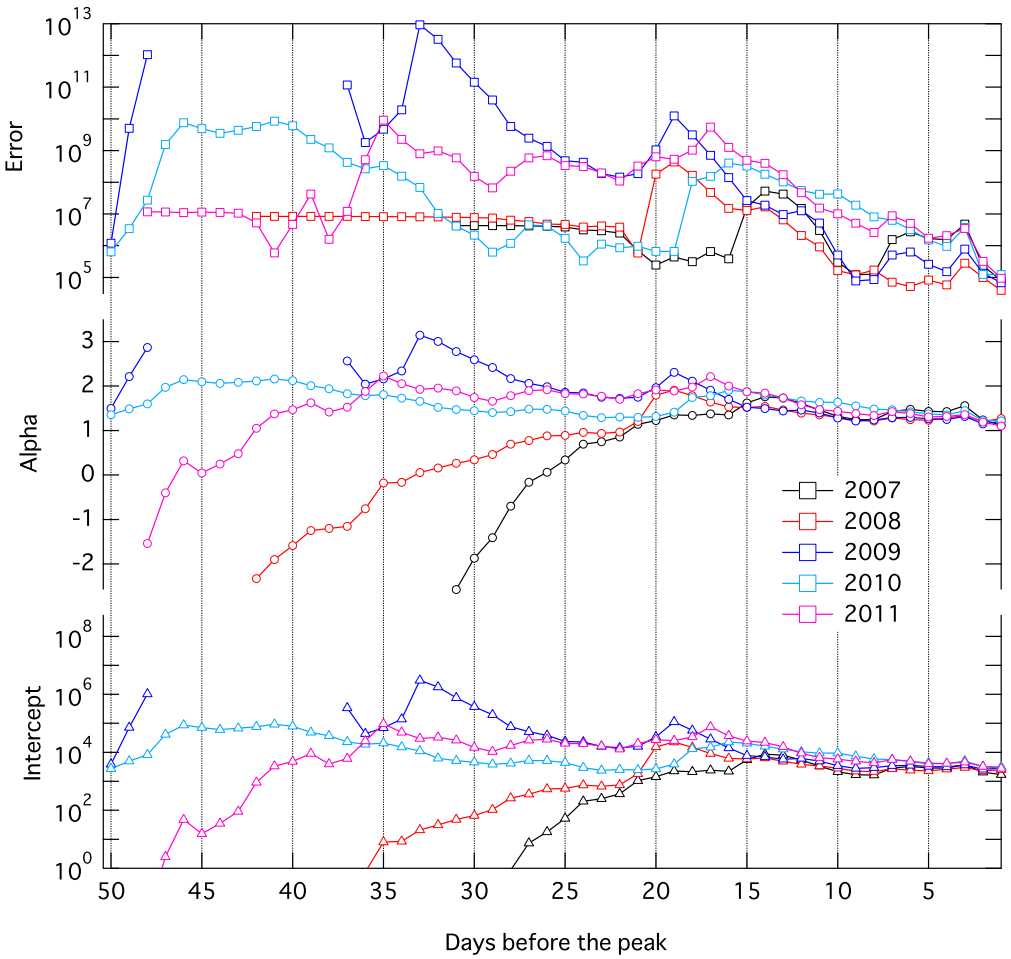


図 5.16　「海の日」における見積もり範囲を変えた場合の，見積もったベキ関数との二乗誤差 (上段)，ベキ指数 $\alpha^{(k)}$(中段)，切片 $A^{(k)}$(下段)．見積もり範囲がピーク $t_c$ に近づくにつれ二乗誤差も小さくなっていることが分かる．見積もりができなかった場合はデータ点が抜けている．

しかし，一般にわずかなベキ指数の違いが全体の書き込み数に大きく影響するため，ベキ関数的な変動は，予測が難しい．また本手法は，発散点となる $t_c$ は定数を事前に与えており，モデル上は $t_c$ での書き込み数は無限大に発散する．

### 5.6.2　平時に戻るまでの時間

式 (5.4) のモデルで，大きなニュースが入った後，どのくらいで平時の書き込み数に戻るかも予測することができる．図 5.18 は，「アーサー C. クラーク」という SF 作家が亡くなった際の実際の書き込みの変動と，ピークから 7 日後までの値から見積もった，ベキ指数 $\alpha^{(アーサー)} = 1.53$ と切片 $A^{(アーサー)} = 383$ をモデル式 (5.4) に代入した予

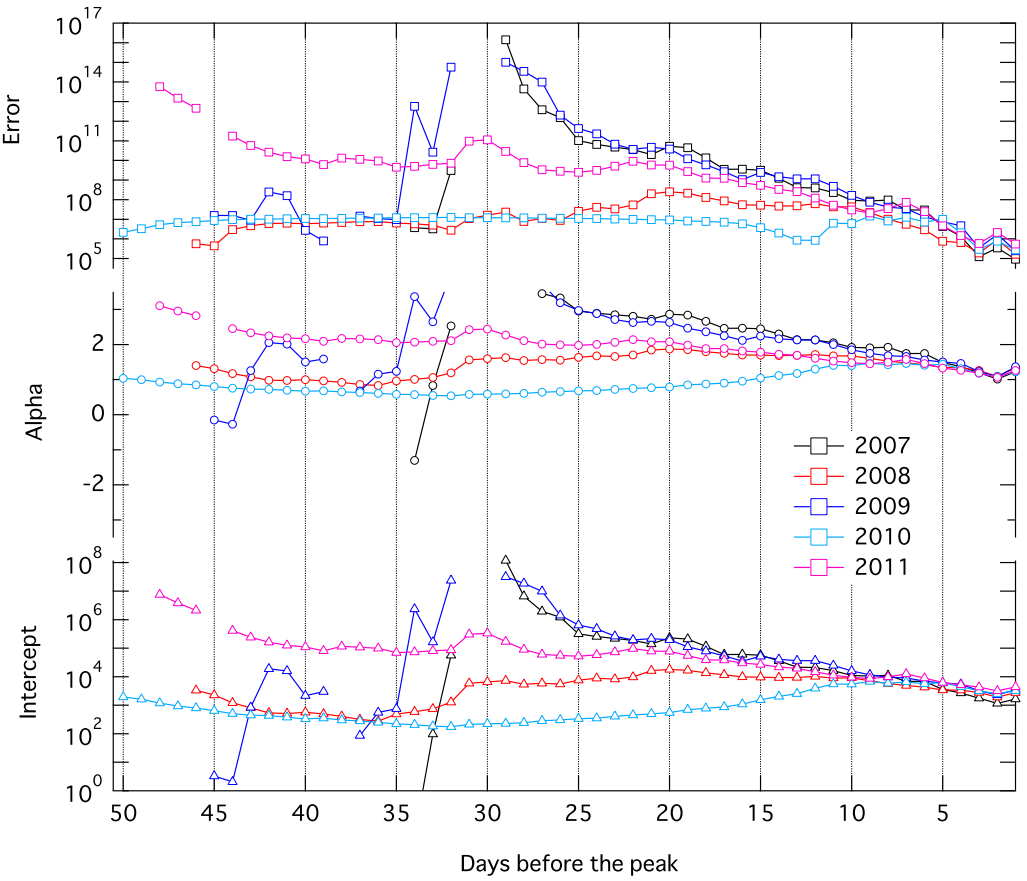


図 5.17　「こどもの日」における見積もり範囲を変えた場合の，見積もったベキ関数との二乗誤差 (上段)，ベキ指数 $\alpha^{(k)}$(中段)，切片 $A^{(k)}$(下段)．見積もり範囲がピーク $t_c$ に近づくにつれ二乗誤差も小さくなっていることが分かる．

測線である．亡くなる 1 週間前の平時の「アーサー.C. クラーク」の書き込みの中央値は，$w_0^{(アーサー)} = 1.9$ で，実線で示した．ベキ関数的に変動する書き込み数が，平時の $w_0^{(アーサー)}$ の値に戻るまではピークからの経過日数 $T^{(アーサー)} = t - t_c$ として，以下の式に当てはめ，

$$T^{(k)} = \left( \frac{w_0^{(k)}}{A^{(k)}} \right)^{-\frac{1}{\alpha^{(k)}}}, \tag{5.17}$$

より $T^{(アーサー)} \simeq 32$ 日程度かかることが予測される．実際の書き込み数はピークから 32 日後は $w^{(アーサー)}(t_c + 32) = 2.7$ となっており，実現された値をランダム投稿モデルのゆらぎの範囲内 $[0.5, 3.3]$ で予測できていることが分かる．

2011 年 3 月 11 日に起こった，東日本大震災の後の「津波」という単語でも同様の見積もりができる．2011 年の 3 月以前の平均的な書き込み数は 1 日約 300 件であった．式 (5.17) を使って，実際に平時に戻るまでの見積もり行うと，ベキ指数は $\alpha^{(津波)} = 0.67$ で

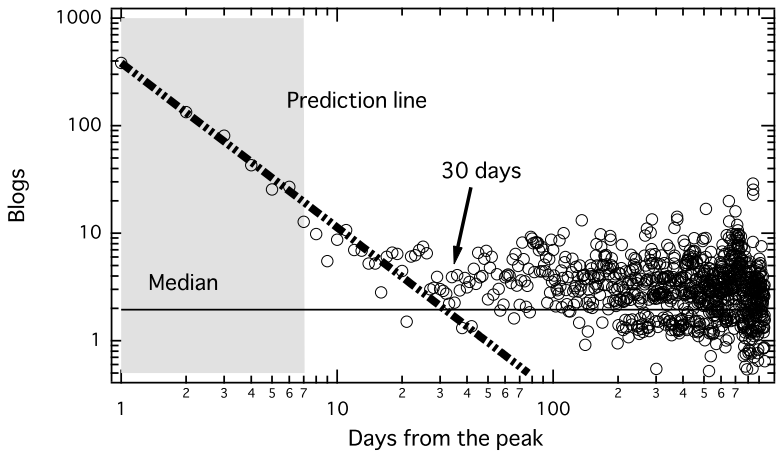


図 5.18　「アーサー C. クラーク」の書き込み時系列．ピークから 7 日目以降がモデルから予測した変動で，約 30 日後に実線で示した中央値である平時の書き込みに戻ることを予測している．

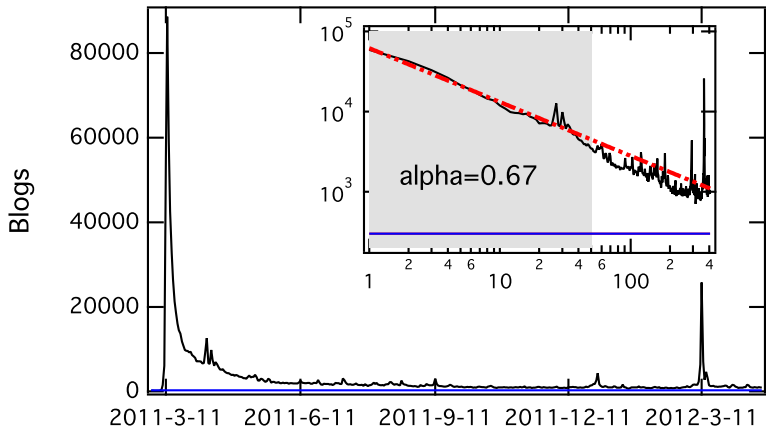


図 5.19　東日本大震災時の「津波」の書き込み時系列．(挿入図) 両軸対数による表示．$t_c$ を 2011 年 3 月 11 日とし，$\alpha^{(\text{津波})} = 0.67$ のベキ関数を赤の破線で示した．青線は，震災前の平均的な書き込み数である 300 件を表している．

あることから，平時に戻るには 8623 日 (約 23 年以上) かかることが予想されている．実際に，東日本大震災から 1 年近く経った後でも依然として震災前の 10 倍近くの 1 日 3000 件の書き込みがあり，「津波」は非常に特異な例であると言える．

「津波」の例では，ブログだけではなく Twitter における書き込み数でも同様の $\alpha < 1$ となるベキ指数が現れる．図 5.20 は，東日本大震災時に地震関連の単語を書き込んだ，1397783 ツイートを使って作成した 1 時間刻みの「津波」の出現頻度時系列である．時間刻みはブログの場合の 1 日とは異なるが，Twitter においてもベキ関数的な変動が現れ，

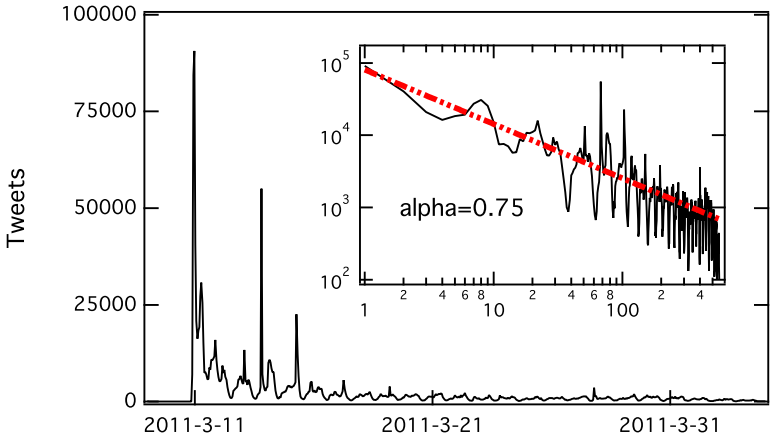


図 5.20　東日本大震災時の「津波」のツイート数変動．(挿入図) 両軸対数による表示．$t_c$ を 2011 年 3 月 11 日 15 時とし，$\alpha^{(\text{津波, Twitter})} = 0.75$.

その時のベキ指数は $\alpha^{(\text{津波, Twitter})} = 0.75$ となる．

### 5.6.3　非整数発散点の見積もり

これまでは発散点 $t_c$ の情報はあらかじめ与え，固定した上で，1 日刻みで解析を行った．本節では，5.4 節で見積もったベキ関数の延長線上に，非整数値の発散点 $t_c + \epsilon$ があると仮定して，$t_c$ からのわずかなずれ $\epsilon$ を見積もる方法を紹介する．非整数値の発散点を見積もることで，1 日の 24 時間よりも小さい時間刻みの中で，何時何分頃に発散点を迎えたのか見積もることが可能となる．

ここでは，式 (5.4) のモデル $x^{(k)}(t)$ に対して，わずかにずれた発散点 $t_c + \epsilon$ を仮定し，

$$x'^{(k)}(t) - x_0^{(k)} = A^{(k)}(|t_c - t| + \epsilon)^{-\alpha^{(k)}} \tag{5.18}$$

であるとする．ずれの大きさを表す $\epsilon$ は，発散点が $(t_c - 1, t_c + 1)$ の範囲に存在すると仮定して，$|\epsilon| < 1$ とする．すなわち $\epsilon = 24$ 時間 と考える．$\epsilon = 0$ の場合の $x'^{(k)}(t)$ は，式 (5.4) の $x^{(k)}(t)$ と一致する．

わずかなずれである $\epsilon$ を見積もる方法は，観測した全区間での $x^{(k)}(t)$ の中央値 $x_0^{(k)}$ と，5.4 節で見積もったパラメータ $\alpha^{(k)}$ と $A^{(k)}$ を持つベキ関数，および実測したピークの書き込み数 $w^{(k)}(t_c)$ を用いる．$w^{(k)}(t_c)$ を式 (5.18) の左辺第一項 $x'^{(k)}(t)$ に $t = t_c$ として代入し，

$$w^{(k)}(t_c) - x_0^{(k)} = A^{(k)}\epsilon^{-\alpha^{(k)}} \tag{5.19}$$

より，未知変数である $\epsilon$ は

$$\epsilon = \left( \frac{w^{(k)}(t_c) - x_0^{(k)}}{A^{(k)}} \right)^{-\frac{1}{\alpha^{(k)}}} \tag{5.20}$$

として一意に算出することができる．式 (5.20) 上は，$|\epsilon| > 1$ となることもあり得るが，ここでは $|\epsilon| < 1$ となる場合のみを考える．

これまでピーク日の書き込み数 $w^{(k)}(t_c)$ は，実際には有限の値であるにも関わらず，式 (5.4) では，モデル上 $x^{(k)}(t_c) = \infty$ となるため，$w^{(k)}(t_c)$ の情報は使っていなかった．しかし，本解析ではこの $w^{(k)}(t_c)$ の値を手がかりにして，非整数発散点 $t_c + \epsilon$ を探る．具体的には毎年 4 月 1 日にピークを持つ「エイプリルフール」であれば，発散点は，時間シフト後の 4 月 1 日 5 時 0 分から 4 月 2 日 4 時 59 分の中での一点で固定した扱いになっていたところを，式 (5.20) に，見積もったパラメータ値と，実測値を代入することでさらに細かい時間スケールで，4 月 1 日の何時何分に発散点を迎えているのかを見積もることができる．

**ピーク前 ($t_c > t$) のベキ関数の延長上の発散点**

ピーク前の増加関数側のベキ関数の発散点 $t_c$ は，元のモデル式 (5.4) の $x^{(k)}(t)$ に対して，わずかに遅れていることに相当する (図 5.21 の左図)．2008 月 7 日 21 日にピークを持つ「海の日」であれば，$t_c$ は，7 月 21 日の 17 時台から 7 月 22 日の 4 時台の範囲に存在していることになる．ピーク前の盛り上がりの最高点，という解釈になる．盛り上がりの最高点といっても，同時に降下し始めた期待である「飽き」も入り交じっている．

**ピーク後 ($t_c < t$) のベキ関数の延長上の発散点**

ピーク後の減少関数側のベキ関数の発散点 $t_c$ は，元のモデル式 (5.4) の $x^{(k)}(t)$ に対して，わずかに早まっていることに相当する (図 5.21 の右図)．2008 月 7 日 21 日にピークを持つ「海の日」であれば，$t_c$ は，7 月 21 日の 5 時台から 7 月 21 日の 16 時台の範囲に存在していることになる．

## 5.7 まとめ

本章では，第 4 章で導入したランダム投稿モデルを用いて，検出した異常値の中から，ベキ関数的な変動を含む時系列について詳細な解析と，モデル化を行った．ベキ関数的な変動は，物理現象では相転移ともなじみが深いが，ブログ上の単語出現頻度だけではなく，インターネット動画へのアクセス数の変動や，為替変動の時系列の中など，人間の集団に関わる現象の中に数多く報告されている．しかし，このようなデータ統計的に記述する

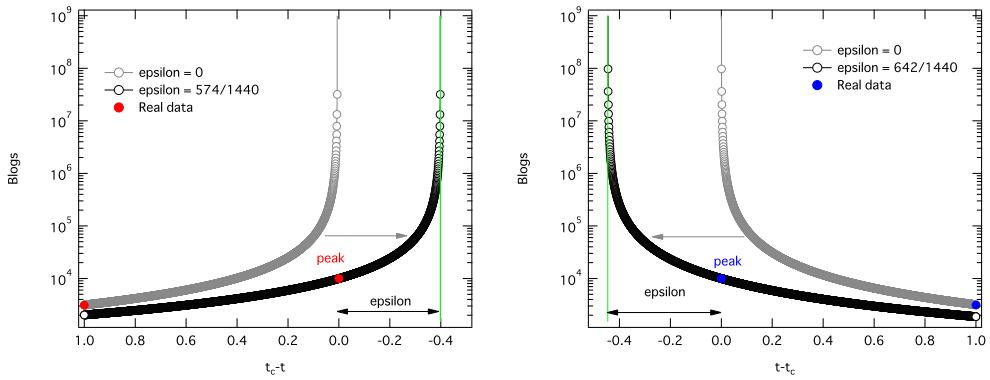


図 5.21　2008 年の「海の日」非整数発散点の予測例 (縦軸対数)．(左図) ピーク前 ($t_c > t$) のベキ関数の延長の発散点．(右図) ピーク後 ($t_c < t$) のベキ関数の延長の発散点．

ための確立された方法は存在しない．その理由は，ベキ関数の扱いにくさだけではなく，データごとの異質性や，ベキ関数的に変動する事象サンプル数の少なさにも起因する．

本研究では，為替変動の先行研究 [83] で使われた手法を元に，ブログ上でベキ関数的な変動をする時系列の，以下の値を見積もる方法を提案した．

1. **ベキ関数のパラメータ**　非線形最小二乗法にて求めたベキ指数 $\alpha^{(k)}$，切片 $A^{(k)}$
2. **有意水準** ブートストラップ法にて求めた通常の $p$ 値に相当する $q$
3. **ベキ関数的な変動をする区間**　$n \geq 5$ の区間で $q \geq 0.1$ となる最長の期間 $n^{(k)}$(日)

本手法は，ベキ関数的な変動だけではなく指数関数など他の関数の場合にも拡張が容易であり，また，ブログ以外の時系列にも適用が期待できる．

次に，本研究で開発した解析方法を，ベキ関数的な変動をするブログの日次時系列に応用した．経験的に分かっている事実として，ブログ上でベキ関数的な変動をする単語は，あらかじめピークとなる「締め切り日」を多くの人が分かっている場合に限られる．ここでは，それらの単語として以下の三つの分類でサンプルを収集した．

1. **イベント語**　「こどもの日」「バレンタインデー」などに代表される日本の祝日の名前や，有名な年中行事
2. **日付**　「5 月 18 日」などのブログ本文にブロガーが書き込む日付
3. **ニュース語**　「マイケル・ジャクソン」「津波」「iPS 細胞」などに代表される訃報や天災，また受賞などの社会に大きなインパクトを与えるニュースに関連する語

イベント語と日付の場合には，ピークの前後でベキ関数的に振る舞うが，ニュース語の場合は，ピークの後のみベキ関数的な振る舞いを観測できる．ベキ指数は単語によってもば

らつきがあるが，絶対値で 0.1 から 2.5 の値となる．同一時系列中での，ピークの前後でのベキ指数の関係は，有意な相関は観測できない．ニュース語の場合はベキ指数は平均すると 1 より大きいベキ指数を持つが，2011 年の東日本大震災で甚大な被害をもたらした「津波」に関しては $\alpha^{(津波)} = 0.67$ という小さい値となり，このことからもニュースが与えた大きな衝撃と影響を示唆している．

ベキ指数の違いは何に起因するのかを理解するため，本研究では簡単な数理モデルを提案した．モデルでは，単語 $k$ は，ピークとなる締め切り日 $t_c$ を持っていて，すべてのブロガーが $t_c$ がいつかを知っていると仮定した上で，単語 $k$ を書き込むブロガーに以下の二つの効果を加えることで，任意の指数を持つベキ関数的な振る舞いを示すことを示した．

1. **時間に対して非線形に反応する効果**
2. **書き込み数のフィードバック効果**

最後に，締め切り日を持つブログ書き込み数がベキ関数的に変動するという仮定を応用する例を三つ紹介した．

1. **将来の書き込み数予測**
2. **平常時の書き込み数に戻るまでにかかる時間**
3. **1 日以内の時間スケールでの発散点の見積もり**

# 第 6 章

# まとめ

本章では，これまでの本研究の結果をまとめる．そしてこの結果をふまえ，今後取り組むべき課題や将来の展望について述べる．

## 6.1 本博士論文のまとめ

前世紀末より急速に発展したコンピュータ技術によって，大規模に人間社会のデータが蓄積，解析され，そこに数理的なパターン (法則性) を見いだすことが可能となった．そこで，本研究では，2000 年以降急速に発展し，普及したウェブ上のソーシャルメディアの一つである「ブログ」に現れる単語の出現頻度時系列の解析とモデル化を行うことで，これら人間の集団におけるパターンについて議論してきた．

第 2 章では，用語の定義と，扱ったデータの説明を行った．本研究で主に扱ったデータは，検索単語を含む書き込みを，記事単位で数えた日次の時系列である．時系列の元になるデータベースは，日本国内で公開されている約 3200 万ブロガーの，約 26 億記事のデータが含まれる大規模なものである．また，必要に応じてそのデータからランダムサンプルした約 33 万人分の ID 付きの詳細データを用いている．

第 3 章では，本研究で行ったデータ解析における前処理について説明した．厳密に管理された実験環境から得られるデータとは異なり，ウェブという不特定多数の人が自由な意志で行動するプラットフォームから得られるデータは，予測の難しいノイズが数多く含まれている．特に，時系列の変動に大きく影響を与えるものとして，機械によって生成されるスパム，収集したブログ数自体の変動，自明な 1 日の周期がある．そこで，これらのノイズに対して行った対処方法について述べた．

第 4 章では，日常的にブログの世界に現れる語 (「日常語」) が持つ典型的なゆらぎを定義し，そこからの逸脱で時系列からの異常値検出を行った．日常語の代表例として，本

研究では形容詞，連体詞，接続詞に注目した．これらの中で，単位根過程が棄却されず，弱定常性の性質を持つ単語の出現頻度時系列における，平均値と標準偏差の間のスケーリング則を指摘した．このスケーリング則を説明するためのモデルとして，ランダム投稿モデルを提案した．ランダム投稿モデルでは，ブロガーが記事自体を投稿するか否かのゆらぎの他に，ブログを投稿したという条件下で，記事に検索語を書き込むか否か，という二つの過程におけるゆらぎを考慮することで，平均値と標準偏差の間に現れるスケーリング則を再現できることを数理的に示した．さらにランダム投稿モデルと移動平均を使って，時系列から局所的な異常値を検出する応用方法を紹介した．

第 5 章では，非定常な性質を持つブログ時系列のうち，ベキ関数的に変動するものに注目した解析とモデル化を行った．ベキ関数的に変動する単語の代表例は，多くの人があらかじめピークとなる日付を認識している，祝日名や年中行事名，さらにその日付そのものである．そこで，本研究ではそれらの解析手法を提案し，その手法を用いた詳細な解析を行った．これらの単語はピーク前にはベキ関数的に書き込みが増え，ピーク後にはベキ関数的に書き込みが減る．これらは絶対値で 0.1 から 2.5 の任意のベキ指数を持ち，ピーク前後のベキ指数には有意な相関は見られなかった．

さらに，突然の自然災害や，訃報などニュースに関連する語は，その第一報が入った当日，またはその翌日に大きなピークを持ち，その後のみベキ関数的に減少する．この時のベキ指数はニュースの衝撃と影響を反映しているといえるが，平均的にはおよそ絶対値で 1 となる．しかし，2011 年 3 月の東日本大震災に関連する「津波」などの単語は，例外的に絶対値が小さいベキ指数となり，2 年以上経った現在でも，書き込み数は震災前よりも多く，未だにその余波が残っていることを示唆している．

得られた様々な指数を持つベキ関数を再現するため，ブロガーが書き込みを行う際の確率の時間変化に，時間に対して非線形に反応する効果と，前日の書き込み数のフィードバック効果，という二つの効果を取り入れたモデルを提案した．数理的にこのモデルから，任意の指数を持つベキ関数を再現できることを示し，実際のデータからその妥当性を示した．ベキ関数的に変動する時系列を応用する例として，将来の書き込み数推定，ピークから平時に戻るまでの時間の推定，さらに実測したピーク時の値を使い，1 日以内のスケールでの発散点を見積もる手法を紹介した．

## 6.2　今後の展望

本研究ではブログのデータの中から特に，特定の単語に注目した時系列に注目したが，ブログ自体でもまだ明らかになっていないことは多い．例えば，人が何か行動する時の時間間隔には非自明な相関があり，「バースト」と呼ばれている [92, 93, 94]．多くのブロガーの行動にもバースト性が見られるが [A]，その行動特性を組み合わせた，ブロガーを

エージェントとしたミクロなモデルから，ベキ関数的，指数関数的に変動する時系列を再現するのは今後の課題である [95].

さらにブロガーの書いたテキストの統計的特性を調べることも今後の課題である．言語の経験則としてよく知られている Zipf 則 [96] や Heaps 則 [97] はブログでも成り立っていることをすでに確認しているが [E,F]，夏目漱石などの歴史的な職業作家と，21 世紀のブロガーの文章はどう違うのか，数理的な側面から明らかにする．人間の書く文章の特徴を数理的な側面から明らかにすることは，ソーシャルメディアの普及や，電子化された本が爆発的に増えた現在，研究が盛んに行われている分野であり [98, 99, 100, 101, 102, 103]，文章の起源や言語への理解を深めるのに役立つとされ，注目を集めている．

本研究では主にブログを扱った．しかし，インターネットの世界では次々に新しいメディアが台頭している．ミニブログと呼ばれる Twitter は，開始から数年で全世界での利用者は 5 億人，代表的な SNS の Facebook は利用者が 10 億人を突破した．さらに，インターネットの世界だけではなく，スマートフォンの台頭によって，実空間での人間の移動，行動履歴なども精緻に記録されている．現実的には，データはあっても，公開されていなかったり，技術的問題や法的規制で入手可能かは別問題である．しかし，これら幅広い人間行動に由来するデータにも，ブログと同じような統計性や数理的普遍性が見られるのかを確認するのは，今後の課題である．

本研究の最大の特徴は，データに基づいた人間の集団行動に見られる数理的普遍性の理解である．精緻な解析技術の確立や，大規模な実際のデータ解析を通して，その数理的側面の存在を指摘できた．しかし，人間の行動に，数理的な側面がなぜ現れるのかという根本的な問いにはまだ答えられていない．自由な意志を持って動いているはずの人間の集団に，数理的な側面がなぜ生じるのか，これらは，多種多様なデータの解析や数理モデルの構築を通し理解できることが期待できる．

# 謝辞

4 年以上勤めた会社を辞め，再び大学院で学ぶことに関して経済的，精神的な面から全面的に協力してくれた愛媛に住む両親と祖父母，離れて暮らす妹弟に感謝します．

本研究で用いたブログデータは株式会社ホットリンクより提供を受けています．株式会社ホットリンクの代表取締役社長の内山幸樹氏に感謝申し上げます．また，2007 年より継続しているブログに関する共同研究プロジェクトメンバーの皆様に感謝申し上げます．株式会社電通，特に関西支社の皆様，セーヨー・サンティ氏はじめとするホットリンクの皆様にはデータ提供だけではなく，様々なアドバイスを頂きました．

本学位論文の審査に加わっていただいた，奥村学先生，樺島祥介先生，寺野隆雄先生，小野功先生に感謝申し上げます．相馬亘先生，伴周一先生をはじめとする日本大学理工学部一般教育教室物理系列の皆様には辛抱強い励ましを頂き，ありがとうございました．筑波大学システム情報工学研究科の岡田幸彦先生には，主に第一章の原稿のチェックを頂き，ありがとうございました．

東京工業大学の高安研究室の卒業生を含む関係者の皆様には，大変お世話になりました．特に秘書の町田理香さんには事務的な面で数多くのサポートを頂きました．博士後期課程在籍中の田村光太郎さんは注意深く原稿のチェックをして頂きました．早稲田大学高等研究所の山田健太さん，ホットリンクの渡邊隼史さんには，本研究全体を通して，共同研究者として数々の議論や，有意義な情報を共有させて頂きました．ソニー CSL の高安秀樹先生にはゼミでいつも助言を頂きました．あらためて感謝申し上げます．これらのサポートが無ければ本研究を推し進めることできませんでした．

指導教官である高安美佐子先生，長きにわたり大変お世話になりました．研究のアドバイスのみならず，数々の与えていただいたチャンスに大変感謝しております，ありがとうございました．

# 付録 A

# データの詳細

ホットリンクでは，2006 年 11 月より日本の主要なブログサービス事業者のサイト，2 ちゃんねる，Yahoo!掲示板や教えて！ goo，Oracle Techonology Network などの国内ウェブサイトへの投稿を収集し，データベース化して，検索ツールを通じて顧客に提供している．顧客は主に，インターネット上の書き込みを広告の効果測定など自社のマーケティング活動に使っている．また，ホットリンクでは 2009 年より API 機能を公開しており，容易に任意の検索語を含むブログ数を得ることができる．同社が収集対象としている

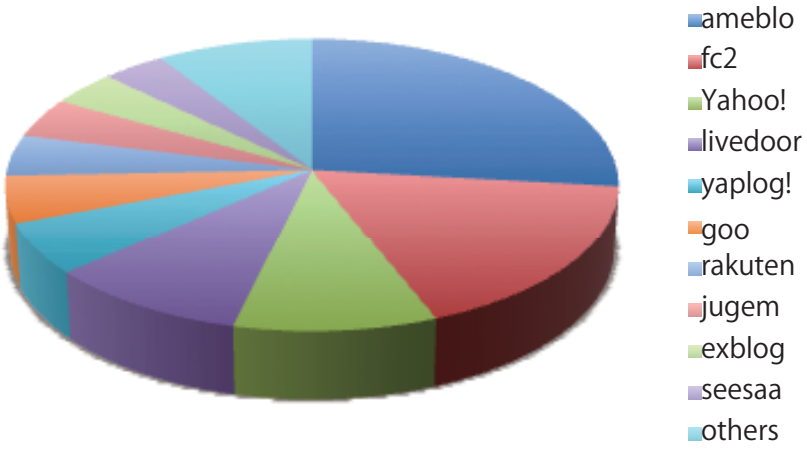


図 A.1　ランダムにサンプリングした，100000 人分のブログサイトの内訳．最も多いアメーバブログ (27%)，FC2 ブログ (17%)，Yahoo!ブログ (10%) で全体の半数以上を占める．

ブログデータのクロールする範囲は以下である．

- アメーバブログ [26.8%] (http://ameblo.jp/)
- FC2 ブログ [17.2%] (http://blog.fc2.com/)
- Yahoo!ブログ [9.7%] (http://blogs.yahoo.co.jp/)
- livedoor Blog [9.7%] (http://blog.livedoor.com/)
- ヤプログ！ [5.8%] (http://www.yaplog.jp/)
- goo ブログ [5.2%] (http://blog.goo.ne.jp/)
- 楽天広場 [4.5%](http://plaza.rakuten.co.jp/)
- ジュゲム [4.4%] (http://jugem.jp/)
- エキサイトブログ [3.9%] (http://www.exblog.jp/)
- Seesaa ブログ [3.9%] (http://blog.seesaa.jp/)
- ココログ [2.5%] (http://www.cocolog-nifty.com/)
- はてなダイアリー [1.9%] (http://d.hatena.ne.jp/)
- ウェブリブログ [1.3%]( http://webryblog.biglobe.ne.jp/)
- auOne ブログ[*1] [1.2%] (http://www.au.kddi.com/auone_blog/)
- So-net ブログ [0.8%] (http://blog.so-net.ne.jp/)
- CURURU[*2] [0.6%] (http://www.cururu.jp/)
- ドリコムブログ[*3] [0.3%] (http://blog.drecom.jp/)
- DTI ブログ [0.3%] (http://blog.dtiblog.com/)
- ドブログ[*4] [0.1%] (http://www.doblog.com/)
- LOVELOG [0.0%] (http://blog.dion.ne.jp/)
- 独自ドメインブログ [0.0%]

すでにサービスが終了したもの，また 2006 年 11 月以降に新たにクロール範囲に入ったものなど含まれる．ランダムに抽出した 100000 人のブログにおいて内訳を算出した割合を [ ] 内で表示した．有名人が多くアカウントを持ち，絵文字などのオプション機能も充実しているアメーバブログが最も割合が多い．この結果は，Alexa 社 (http://www.alexa.com/) による，2012 年 9 月現在の国内インターネットトラフィック上位が，9 位と 16 位アメーバブログ (ameblo.jp, ameba.jp)，5 位 FC2(fc2.com) となっていることとも呼応する結果となっている．

[*1] 2011 年 3 月 31 日サービス終了．
[*2] 2010 年 3 月 31 日サービス終了．
[*3] 2010 年 3 月 31 日サービス終了．
[*4] 2009 年 5 月 30 日サービス終了．

# 付録 B

# データ処理の前準備を行わない場合

## B.1　平均値と標準偏差のスケーリング

図 B.1 と図 B.2 には，第 4 章で確認した，平均値と標準偏差のスケーリング則を，式 (3.5) の全数による規格化を行う前の時系列で示した．全数による規格化前は，単位根検定で弱定常と判定され，日常語に分類されたものに絞り込むと形容詞は 1526 単語，連体詞は 94 単語，接続詞は 260 単語になる．規格化を行わない場合にも，平均値と標準偏差のスケーリング則が成り立つ．しかし，標準偏差は規格化を行った場合と比較すると大きい．

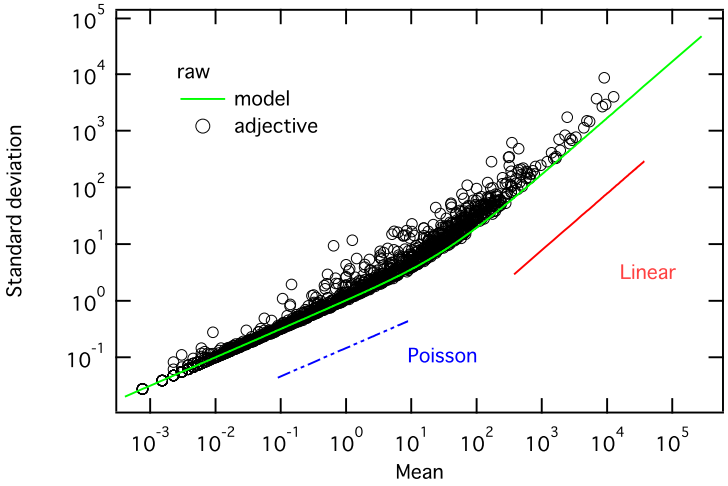


図 B.1　規格化を行わない時系列において，日常語の形容詞の平均値と標準偏差の散布図．実線は式 (4.8) の解 ($\frac{\delta}{w} = 0.29$) に対応する．

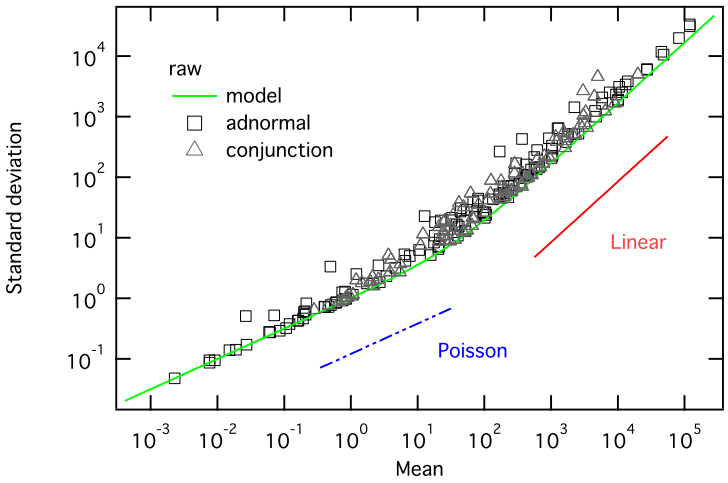


図 B.2　規格化を行わない時系列において，日常語の連体詞と接続詞の平均値と標準偏差の散布図．□が連体詞，△が接続詞．実線は式 (4.8) の解 ($\frac{\delta}{w} = 0.29$) に対応する．

## B.2　ベキ関数的な変動

図 B.3 は全数による規格化を行う前の「海の日」の書き込み時系列である．ピークである 2008 年 7 月 21 日を $t_c$ として，$t_c$ までの残り日数または経過日数を横軸に取り，両軸対数で表示している．第 5 章で導入した解析手法を用いると，ピーク前の場合ベキ関数的に変動する期間 $n^{(k)}$ は 28 日間，ベキ指数の値は $\alpha^{(\text{海の日, fore})} = 1.10$ となった．ピーク後の場合ベキ関数的に変動する期間 $n$ は 18 日間，ベキ指数の値は $\alpha^{(\text{海の日, after})} = 1.44$ であり，$\alpha^{(\text{海の日, fore})}$ よりもやや大きい値を取る．他の単語でもベキ指数の見積もりをし，ヒストグラムにしたのが図 B.4 から図 B.7 である．期間は 2006 年 11 月から 2010 年 10 月までの 4 年間である．表 B.1 にベキ指数 $\alpha^{(k)}$ の平均値とベキ関数的に変動する期間 $n^{(k)}$ の中央値を示した．

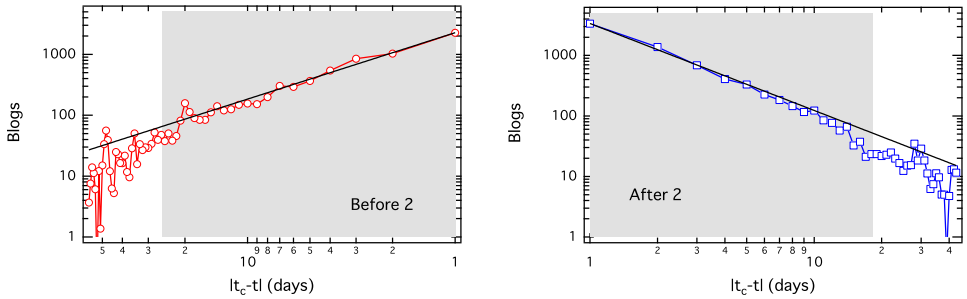


図 B.3　2008 年「海の日」の書き込み時系列 (両軸対数)．網掛けの部分はベキ関数的に変動する期間 ($n^{(k)}$ 日) である．(右図) ピーク前の推移．($\alpha^{(\text{海の日,fore})} = 1.10, A^{(\text{海の日,fore})} = 2273, n^{(\text{海の日,fore})} = 26$)(左図) ピーク後の推移．($\alpha^{(\text{海の日,after})} = 1.44, A^{(\text{海の日,after})} = 3369, n^{(\text{海の日,after})} = 18$)

表 B.1　規格化を行う前の時系列におけるベキ指数 $\alpha^{(k)}$ の絶対値．

| | | $\alpha^{(k)}$ | $n^{(k)}$ (days) | # samples |
|---|---|---|---|---|
| Event | Before | $1.21 \pm 0.38$ | 15.5 | 80 |
| | After | $1.48 \pm 0.28$ | 16 | 85 |
| Date | Before | $0.64 \pm 0.25$ | 11.5 | 418 |
| | After | $1.14 \pm 0.18$ | 22 | 1176 |
| News | After | $1.21 \pm 0.35$ | 11 | 18 |
| All | Before | $0.73 \pm 0.35$ | 12 | 498 |
| | After | $1.16 \pm 0.21$ | 21 | 1279 |

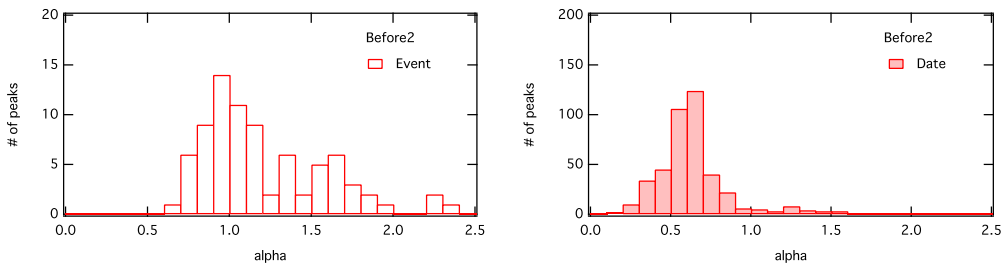


図 B.4　$t_c$ 前のイベント語 (左図)，日付 (右図) のベキ指数 $\alpha^{(k)}$ の分布．

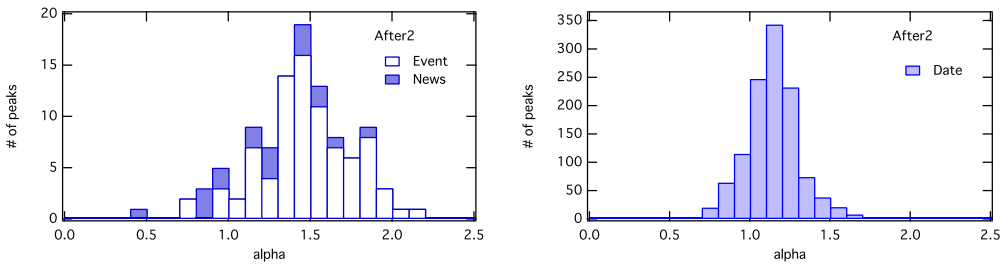


図 B.5　$t_c$ 後のイベント語とニュース語 (左図)，日付 (右図) のベキ指数 $\alpha^{(k)}$ の分布.

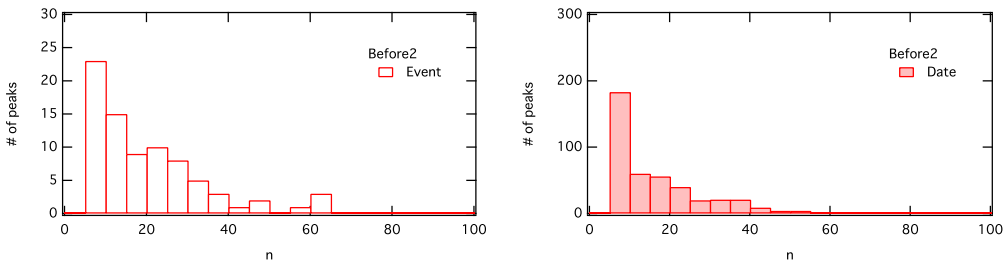


図 B.6　$t_c$ 前のイベント語 (左図)，日付 (右図) のベキ関数的に変動する期間 $n^{(k)}$ の分布.

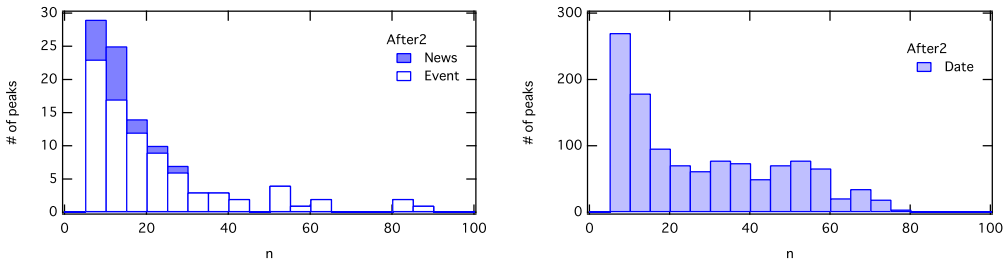


図 B.7　$t_c$ 後のイベント語とニュース語 (左図)，日付 (右図) のベキ関数的に変動する期間 $n^{(k)}$ の分布.

# 付録 C

# ベキ関数的に変動する単語

第 5 章で取り上げたベキ関数的に変動するイベント語とニュース語の詳細を以下に挙げる．

イベント単語は Wikipedia 日本語版の「国民の祝日」と「年中行事」から引用して取り上げた．祝日の一つである「みどりの日」は，法改正により 2006 年までの毎年 4 月 29 日から，2007 年より 5 月 4 日に変更となった．そのため日付が広く日本社会に浸透しきっていないと考えられる．特に 2007 年以降のブログ上は，出現頻度は 4 月 29 日と 5 月 4 日両方にピークを持つ．そのためモデル見積もりが困難となるため，サンプルから外している．

ニュース語は Wikipedia 日本語版の「訃報」の項と「地震の年表」から，ブログにある程度書き込みがなされる単語を取り上げた．その結果，ニュース語の候補となったのは以下の 55 単語である．

> 池田晶子，ウメ子，稲尾和久，アベフトシ，大浦みずき，阿部典史，川内康範，石立鉄男，市川崑，アイポッパー，神戸みゆき，臼井儀人，アドマイヤキッス，植木等，大原麗子，アストンマーチャン，加藤和彦，オディール，赤塚不二夫，アグネスタキオン，阿久悠，緒形拳，川田亜子，忌野清志郎，飯島愛，能登，中越，宮城内陸，千島列島，青海省，スマトラ，サモア，ラクイラ，ペルー，ハイチ，チリ，四川，盧武鉉，速水優，マイケル・ジャクソン，川村カオリ，赤池弘次，森繁久彌，スハルト，アーサー・C・クラーク，エドワード・ローレンツ，リチャード・ライト，アール・パーマー，マイケル・クライトン，シドニィ・シェルダン，ブライアン・アダムス，小林誠，益川敏英，南部陽一郎，下村脩

その中で，さらに最終的にベキ関数的に変動すると判定されたものを表 C に挙げた．地震に関しては，検索語は「地名」かつ「地震」の論理積を使って設定した．

表 C.1　ベキ関数的に出現頻度が変動するイベント語の一覧.

| # | 検索語 | 英名 | ピーク日 | 平均数 |
|---|---|---|---|---|
| 1 | 元日 | New Year's Day | 1 月 1 日 | 187.0 |
| 2 | 成人の日 | Coming of Age Day | 1 月の第 2 月曜日 | 44.6 |
| 3 | 建国記念の日 | National Foundation Day | 2 月 11 日 | 13.4 |
| 4 | 春分の日 | Vernal Equinox Day | 3 月 21 日 | 56.4 |
| 5 | 昭和の日 | Showa Day | 4 月 29 日 | 50.5 |
| 6 | 憲法記念日 | Constitution Memorial Day | 5 月 3 日 | 11.9 |
| 7 | こどもの日 | Children's Day | 5 月 5 日 | 143.9 |
| 8 | 海の日 | Marine Day | 7 月の第 3 月曜日 | 77.8 |
| 9 | 敬老の日 | Respect for the Aged Day | 9 月の第 3 月曜日 | 189.4 |
| 10 | 秋分の日 | Autumnal Equinox Day | 9 月 23 日 | 34.0 |
| 11 | 体育の日 | Health and Sports Day | １０月の第２月曜日 | 45.5 |
| 12 | 文化の日 | National Culture Day | 11 月 3 日 | 66.4 |
| 13 | 勤労感謝の日 | Labor Thanksgiving Day | 11 月 23 日 | 57.2 |
| 14 | 天皇誕生日 | The Emperor's Birthday | 12 月 23 日 | 40.7 |
| 15 | クリスマス | Christmas | 12 月 25 日 | 5805.0 |
| 16 | バレンタインデー | Valentine's Day | 2 月 14 日 | 500.9 |
| 17 | ホワイトデー | White Day | 3 月 14 日 | 584.1 |
| 18 | ひな祭り | Girls' Day | 3 月 3 日 | 272.5 |
| 19 | エイプリルフール | April Fool | 4 月 1 日 | 249.2 |
| 20 | 冬至 | Winter Solstice | 12 月 22 日ごろ | 122.2 |
| 21 | 夏至 | Summer Solstice | 6 月 21 日ごろ | 57.3 |
| 22 | 御用納め | Last Business Day of the Year | 12 月 28 日ごろ | 13.7 |
| 23 | 大晦日 | New Year's Eve | 12 月 31 日 | 830.3 |
| 24 | 七夕 | Star Festival | 7 月 7 日 | 553.2 |
| 25 | 初詣 | First Visits of the Year | 1 月 1 日 | 751.7 |
| 26 | 年越し蕎麦 | Year-crossing Noodles | 12 月 31 日 | 39.7 |
| 27 | 初売り | New Year opening sales | 1 月 2 日 | 175.3 |
| 28 | 節分 | Seasonal Watershed | 2 月 4 日ごろ | 432.2 |
| 29 | 除夜の鐘 | Bells on New Year's Eve | 12 月 31 日 | 60.5 |
| 30 | 土用 | Midsummer Day of the Ox | 7 月 20 日ごろ | 77.7 |

表 C.2　ベキ関数的に出現頻度が減少したニュース語の一覧.

| # | 検索語 | ピーク日 | 平均数 | ピーク時の書き込み数 |
|---|---|---|---|---|
| 1 | アベフトシ | 2009 年 07 月 22 日 | 7.0 | 1642 |
| 2 | 川内康範 | 2008 年 04 月 08 日 | 7.9 | 542 |
| 3 | 市川崑 | 2008 年 02 月 14 日 | 14.0 | 829 |
| 4 | 加藤和彦 | 2009 年 10 月 19 日 | 23.1 | 4928 |
| 5 | 忌野清志郎 | 2009 年 05 月 03 日 | 71.2 | 8728 |
| 6 | 飯島愛 | 2008 年 12 月 24 日 | 109.3 | 7232 |
| 7 | 盧武鉉 | 2009 年 05 月 23 日 | 12.2 | 441 |
| 8 | マイケル・ジャクソン | 2009 年 06 月 26 日 | 200.9 | 12406 |
| 9 | 森繁久彌 | 2009 年 11 月 12 日 | 10.2 | 1421 |
| 10 | スハルト | 2008 年 01 月 28 日 | 1.6 | 136 |
| 11 | アーサー・C・クラーク | 2008 年 03 月 19 日 | 4.1 | 590 |
| 12 | シドニィ・シェルダン | 2007 年 01 月 31 日 | 1.8 | 56 |
| 13 | ブライアン・アダムス | 2012 年 02 月 12 日 | 3.1 | 264 |
| 14 | 能登 && 地震 | 2007 年 03 月 25 日 | 18.6 | 3531 |
| 15 | 青海省 && 地震 | 2010 年 04 月 16 日 | 6.7 | 1367 |
| 16 | スマトラ && 地震 | 2012 年 04 月 12 日 | 16.1 | 1382 |
| 17 | ハイチ && 地震 | 2010 年 01 月 15 日 | 55.1 | 4887 |
| 18 | チリ && 地震 | 2010 年 02 月 28 日 | 55.9 | 14329 |
| 19 | 四川 && 地震 | 2008 年 05 月 14 日 | 59.3 | 4456 |
| 20 | 南部陽一郎 | 2008 年 10 月 08 日 | 3.8 | 905 |
| 21 | 下村脩 | 2008 年 10 月 09 日 | 4.8 | 774 |

# 付録 D

# スパムの影響

スパムの除去は, ホットリンクのスパムフィルタを用いており, このスパムフィルタの影響を確認しておく．図 D.1 は, 本研究で用いたスパムフィルタ導入前後での全数 $w(t)$ の比較である．スパムフィルタを導入しない場合の全数の平均値 $\langle w^{(\text{フィルタなし})} \rangle = 1124740$, フィルタを導入した場合の平均値 $\langle w^{(\text{フィルタあり})} \rangle = 749479$ であることから，平均して，スパムの割合はおよそ 3 割程度であることが分かる (図 D.1).

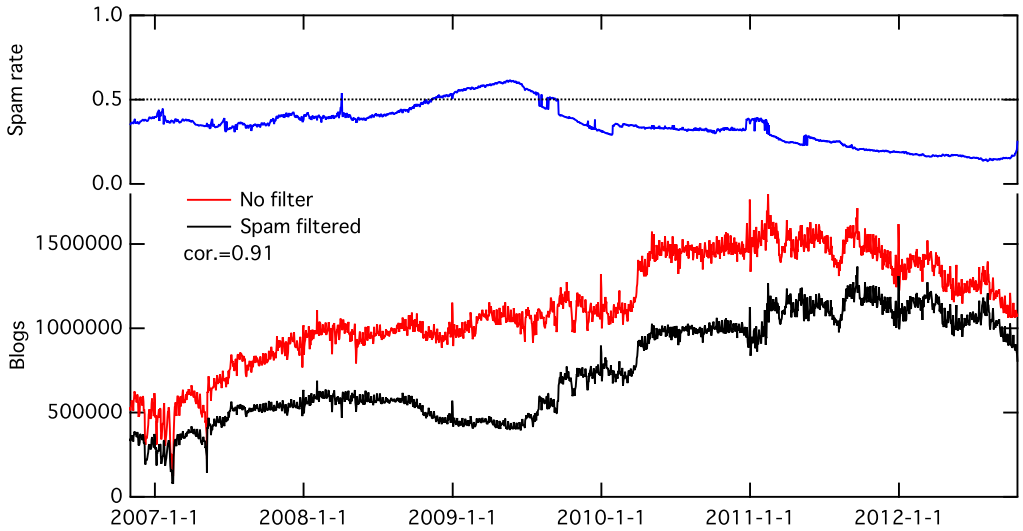


図 D.1　ブログ投稿におけるスパムの影響．(上段) スパムの混入率．クローリング強化前後の 2009 年でやや低くなっている．(下段) 黒が本研究で用いたスパムフィルタを用いた場合の時系列，赤がスパムフィルタを用いない場合の結果．これらの時系列は，式 (3.6) によるピアソンの積率相関係数で $\rho = 0.91$ である．

図 D.2 は，第 3 章に用いた例，接続詞「また」の書き込み時系列である．それぞれ，黒が本章で用いたスパム除去後の時系列，赤がスパムフィルタを用いない場合の時系列に

なっている．

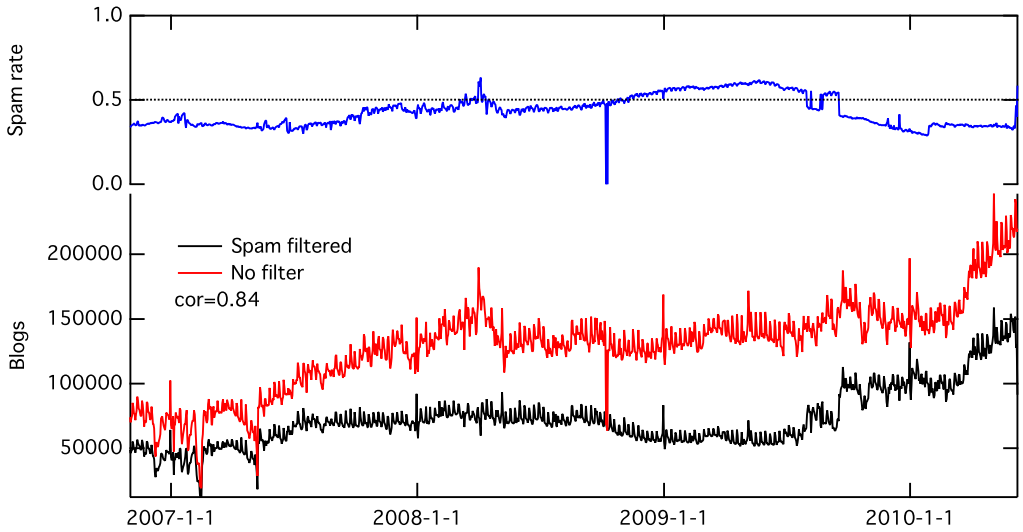


図 D.2　「また」の書き込みに対するスパムブログの影響．(上段) スパムの混入率．(下段) 黒が本研究で用いたスパムフィルタを用いた場合の時系列，赤がスパムフィルタを用いない場合の結果 (全数規格化前)．式 (3.6) によるピアソンの積率相関係数で $\rho = 0.84$ である．

次に第 4 章で取り上げた平均値と標準偏差のスケーリング則を，スパムフィルタを用いる前後で，形容詞，連体詞，接続詞のそれぞれの時系列で確認した (図 D.3)．今回は規格化や，単位根検定によって，弱定常性の単語に絞り込んではいないが，スパムフィルタの有無ではこのスケーリング則にはほぼ影響が出ていないことが分かる．

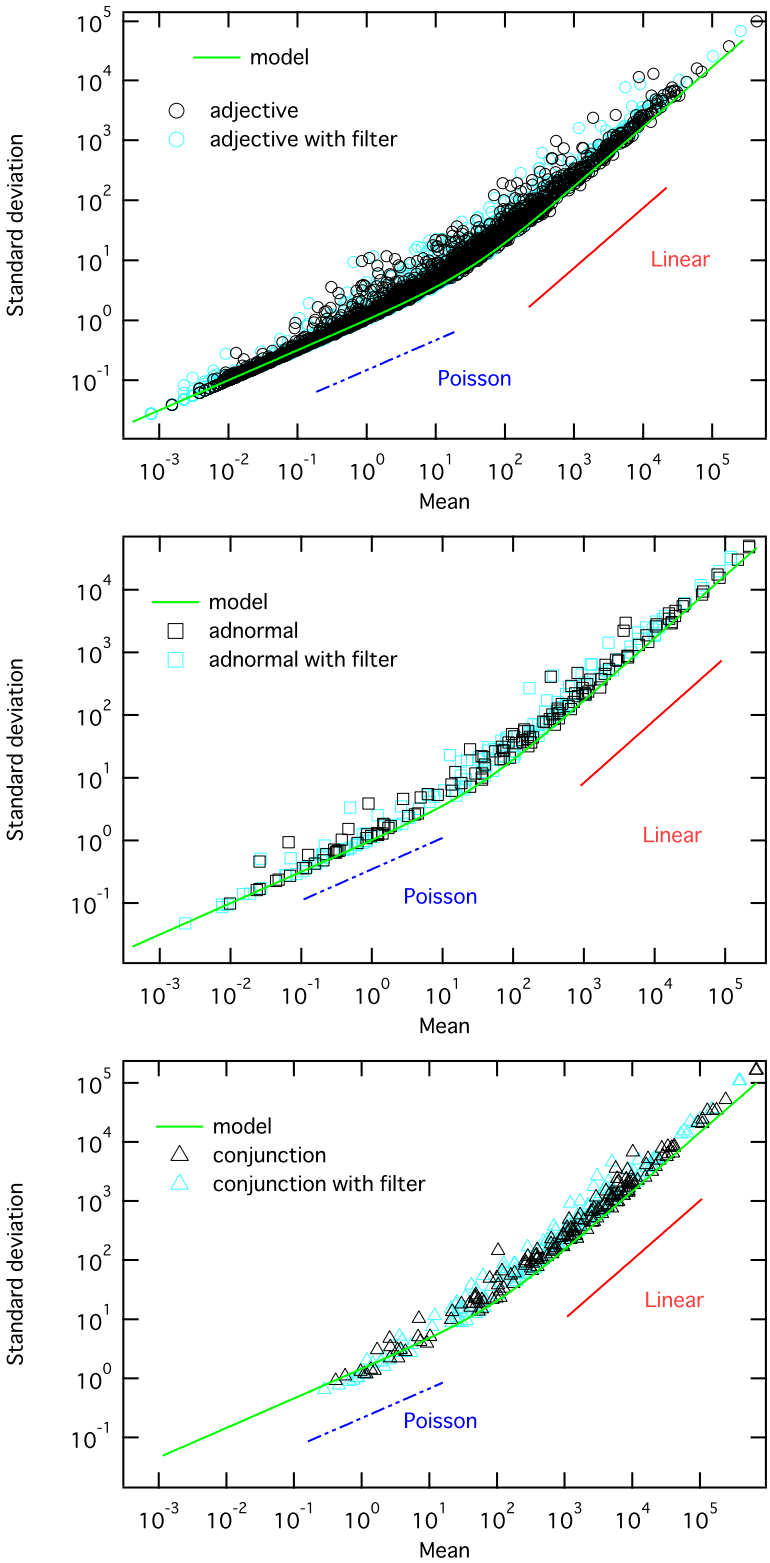


図 D.3　第 4 章で取り上げた平均値と標準偏差のスケーリング則．黒色がスパムフィルタを用いない場合の結果，水色が 4 章でみたスパム除去後の結果．上段から，形容詞，連体詞，接続詞．図中の実線は式 (4.8) の解 ($\frac{\delta}{w} = 0.29$) に対応する．

# 付録 E

# 時間に対して任意の指数で非線形に反応する場合

第 5 章では，以下の二つの仮定をおいたモデルでベキ関数的に変動する時系列を再現できることを示した．

1. **時間に対して非線形に反応する効果**
2. **書き込み数のフィードバック効果**

特に，1 の時間に対して非線形に反応する効果は，Alfi らが導入したモデル [88] と同様に，締め切り $t_c$ からの時間 $|t_c - t|$ の逆数に比例する効果としていた．では，単純な逆数ではなく，$|t_c - t|^{-\gamma}$ で表されるような任意のベキ指数 $\gamma$ で記述できる場合，結果はどうなるのであろうか．

ここでは，第 5 章で導入した，式 (5.11) において，時間に対して非線形に反応する項に $\gamma$ を導入し，ブロガー $i$ が時刻 $t$ でピークとなる締め切り日 $t_c$ を持つ単語 $k$ を書き込む確率の変化量 $\Delta p_i^{(k)}(t) \equiv p_i^{(k)}(t) - p_i^{(k)}(t-1)$ を，

$$\Delta p_i^{(k)}(t+1) \propto \frac{w^{(k)}(t)}{|t - t_c|^{\gamma^{(k)}}} \tag{E.1}$$

と記述する．ブロガーの均質性を仮定すると，ブロガー $i$ の添字は省略でき，$w^{(k)}(t) = w(t) \cdot p^{(k)}(t)$ となる．ノイズ項 $f(t)$ を加え，$\alpha^{(k)}$ を比例定数とすると，締め切り前 $(t_c > t)$ の場合，

$$\frac{\mathrm{d}w^{(k)}(t)}{\mathrm{d}t} = \alpha^{(k)} \cdot \frac{w^{(k)}(t)}{(t_c - t)^{\gamma^{(k)}}} + f(t) \tag{E.2}$$

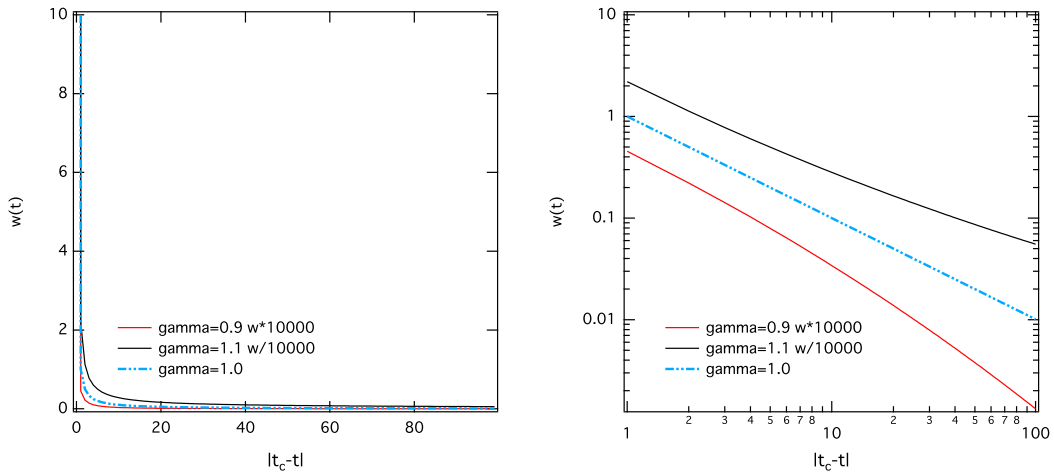


図 E.1 式 (E.4) において $\alpha^{(k)} = 1.0$ として $\gamma^{(k)} = 0.9$(赤) と $\gamma^{(k)} = 1.1$(黒) の場合の比較．破線は式 (5.14) の $w^{(k)}(t) = |t_c - t|^{-\alpha^{(k)}}$ を表す．左図は軸を線形表示とした場合，右図は両対数表示とした場合で，式 (E.4) の結果は比較しやすいようにそれぞれ，$\gamma^{(k)} = 0.9$ の場合 $w^{(k)}(t) \times 10^4$，$\gamma^{(k)} = 1.1$ の場合 $w^{(k)}(t) \times 10^{-4}$ としている．

という微分方程式の形に書き下せる．締め切り後 $(t_c < t)$ の場合は，

$$\frac{\mathrm{d}w^{(k)}(t)}{\mathrm{d}t} = -\alpha^{(k)} \cdot \frac{w^{(k)}(t)}{(t - t_c)^{\gamma^{(k)}}} + f(t) \tag{E.3}$$

となる．ノイズ項 $f(t)$ は無視して，これらの微分方程式を解くと以下の式が得られる．

$$w^{(k)}(t) = \exp\left[\frac{\alpha^{(k)}}{(\gamma^{(k)} - 1)} |t_c - t|^{-\gamma^{(k)}+1}\right] \tag{E.4}$$

ただし $\gamma^{(k)} \neq 1$ である．図 E.1 に $\alpha^{(k)} = 1$ に固定し，$\gamma^{(k)} = 0.9, 1.1$ の場合と，$\gamma^{(k)} = 1$ に対応する，$w^{(k)}(t) = |t_c - t|^{-\alpha^{(k)}}$ の結果を示した．図 E.2 と図 E.3 には．それぞれ $\gamma^{(k)}$ の値を 0.7 から 0.9，1.1 から 1.3 の範囲で変えた場合の式 (E.4) を示した．これらの結果よりわずかな $\gamma^{(k)}$ の違いが，結果に大きく影響することが分かる．

第 5 章の図 5.2 で確認した 2008 年「海の日」と同じデータを用いて，式 (E.4) を定数 $(A^{(k)})$ 倍した以下式で記述したのが図 E.4 である．

$$w^{(k)}(t) = A^{(k)} \exp\left[\frac{\alpha^{(k)}}{(\gamma^{(k)} - 1)} |t_c - t|^{-\gamma^{(k)}+1}\right] \tag{E.5}$$

$\gamma^{(k)}$ を考慮しない場合の結果と比較すると，ピーク前の場合は式 (5.2) による記述で，実データとモデルとの二乗和 $\chi^2 = 50252$ であったのに対し，式 (E.4) による記述で $\chi^2 = 68691$ であった．ピーク後の場合は式 (5.2) による記述で $\chi^2 = 25396$ であったのに対し，式 (E.4) による記述で $\chi^2 = 9806$ であった．ピーク後では $\gamma^{(k)}$ を導入し，パラ

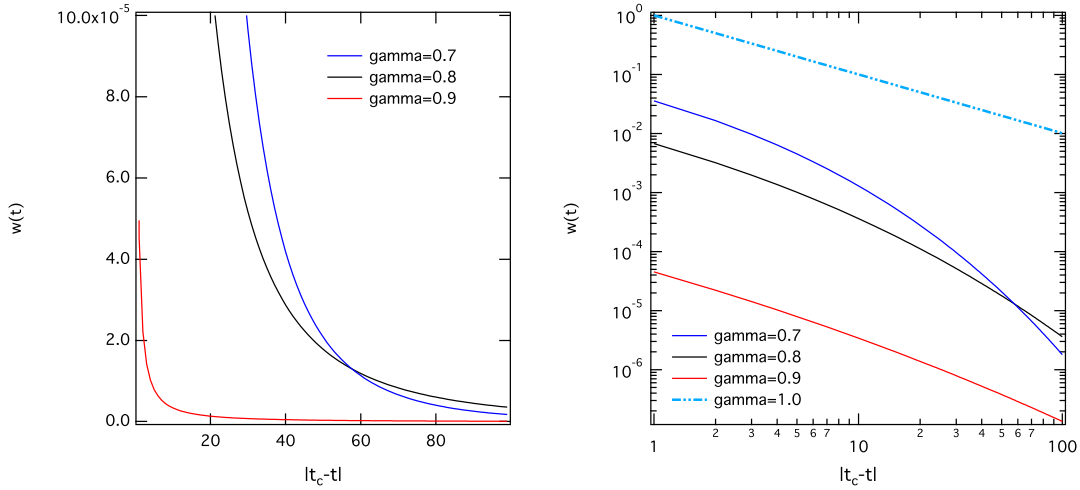


図 E.2　式 (E.4) において $\alpha^{(k)} = 1.0$ において $\gamma^{(k)} < 1$ の範囲で $\gamma^{(k)}$ を変化させた場合の比較．$\gamma^{(k)}$ は 0.7(青),0.8(黒),0.9(赤)．破線は式 (5.14) の $w^{(k)}(t) = |t_c - t|^{-\alpha^{(k)}}$ を表す．左図は軸を線形表示とした場合，右図は両対数表示とした場合．

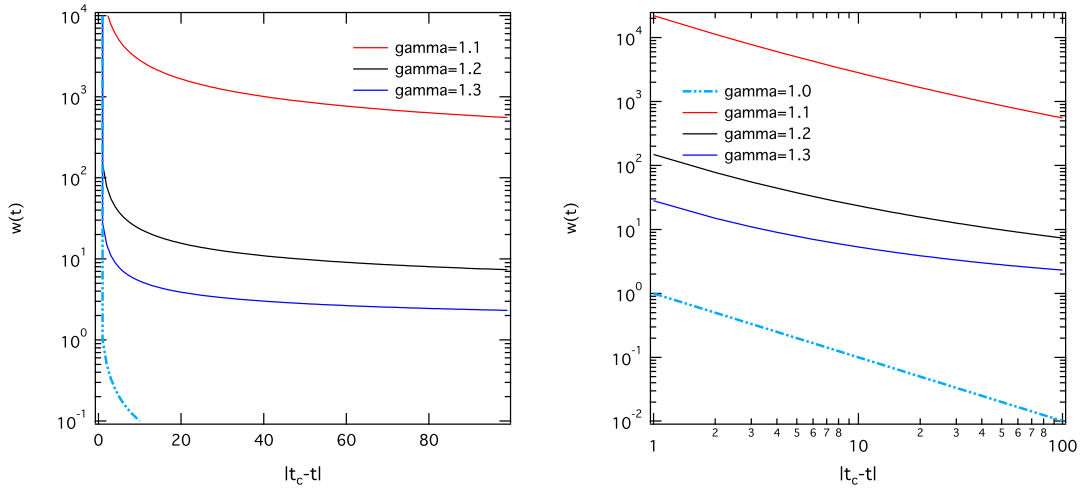


図 E.3　式 (E.4) において $\alpha^{(k)} = 1.0$ において $\gamma^{(k)} > 1$ の範囲で $\gamma^{(k)}$ を変化させた場合の比較．$\gamma^{(k)}$ は 1.1(青),1.2(黒),1.3(赤)．破線は式 (5.14) の $w^{(k)}(t) = |t_c - t|^{-\alpha^{(k)}}$ を表す．左図は縦軸を対数表示とした場合，右図は両対数表示とした場合．

メータが一つ多い式 (E.4) の方が，より実際のデータに近いことがわかるが，ピーク前の結果では逆に $\gamma^{(k)}$ の項が入らない場合の方が，実際のデータに近い．この理由は，式 (E.4) のモデル見積もりの際にパラメータが初期値に依存した，局所解に陥っている可能性が高い．$\gamma^{(k)}$ の導入前後で劇的に見積もり精度が上がる訳ではなく，パラメータ見積もりも困難になることを考え，本研究の解析では $\gamma^{(k)} = 1$ の時間の逆数で人間が反応する，という仮定を採用した．

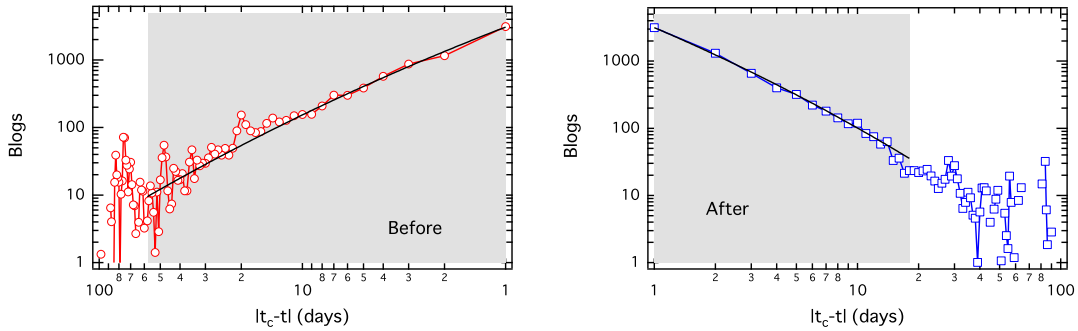


図 E.4　2008 年「海の日」の書き込み時系列において，式 (E.5) のモデルで記述した例．それぞれのパラメータは，ピーク前で $A^{(\text{fore, 海の日})} = 4.44 \times 10^8$，$\alpha^{(\text{fore, 海の日})} = 1.16$，$\gamma^{(\text{fore, 海の日})} = 0.902$，ピーク後で $A^{(\text{after, 海の日})} = 1.66 \times 10^8$，$\alpha^{(\text{after, 海の日})} = 1.30$，$\gamma^{(\text{after, 海の日})} = 0.880$．

# 参考文献


[1] D. Orrell. *Apollo's Arrow: The Science of Prediction and the Future of Everything.* (HarperCollins, Toronto, 2007).

[2] 平成 24 年度版情報通信白書. (2013).

[3] M. Buchanan. *The Social Atom: why the rich get richer, cheats get caught, and your neighbor usually looks like you.* (Bloomsbury, New York, 2007).

[4] C. Castellano, S. Fortunato, and V. Loreto. Statistical physics of social dynamics. *Rev. Mod. Phys.*, **81**, 591 (2009).

[5] P. Ball. The physical modelling of society: a historical perspective. *Physica A*, **314**, 1 (2002).

[6] R. N. Mantegna and H. E. Stanley. *Introduction to Econophysics: Correlations and Complexity in Finance.* (Cambridge University Press, Cambridge, 1999).

[7] 高安 秀樹, 高安 美佐子. **エコノフィジックス市場に潜む物理法則**. (日本経済新聞社, 2001).

[8] 執行 文子. 東日本大震災・ネットユーザーは ソーシャルメディアをどのように利用したのか. **放送研究と調査**, 8 月号 2 (2011).

[9] D. Lazer, A. Pentland, L. Adamic, S. Aral, Barabási, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, M. V. Alstyne. Computational Social Science. *Science*, **323**, 721 (2009).

[10] M. Mendoza, B. Poblete, and C. Castillo. Twitter Under Crisis: Can we trust what we RT? In *SOMA '10*, 71 (2010).

[11] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In *WWW '10*, 851 (2010).

[12] K. Starbird, L. Palen, A. L. Hughes, and S. Vieweg. Chatter on the Red:What Hazards Threat Reveals about the Social Life of Microblogged Information. In *CSCW '10*, 241 (2010).

[13] S. Verma, S. Vieweg, W. J. Corvey, L. Palen, J. H. Martin, M. Palmer, A. Schram, and K. M. Anderson. Natural Language Processing to the Rescue?: Extracting "Situational Awareness" Tweets During Mass Emergency. In *ICWSM '11*, 385 (2011).

[14] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen. Microblogging During Two Natural Hazards Events. In *CHI '10*, 1079 (2010).

[15] Y. Sano, K. Yamada, H. Watanabe, W. Miura, K. Sato, H. Takayasu, and M. Takayasu. Rumor Diffusion and Convergence during the 3.11 Earthquake. (in preparation).

[16] R. M. Bond, C. J. Fariss, J. J. Jones, A. D. I. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, **489**, 295 (2012).

[17] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *ICWSM '10*, 122 (2010).

[18] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Governance in Social Media: A Case Study of the Wikipedia Promotion Process. In *ICWSM '10*, 98 (2010).

[19] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *ICWSM '10*, 178 (2010).

[20] A. Chmiel, J. Sienkiewicz, M. Thelwall, G. Paltoglou, K. Buckley, A. Kappas, and J. A. Hołyst. Collective Emotions Online and Their Influence on Community Life. *PLoS ONE*, **6**, e22207 (2011).

[21] P. S. Dodds and C. M. Danforth. Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents. *J. Happiness. Stud.*, **11**, 441 (2010).

[22] S. A. Golder and M. W. Macy. Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. *Science*, **333**, 1878 (2011).

[23] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The Predictive Power of Online Chatter. In *KDD '05*, 78 (2005).

[24] F. Abel, E. Diaz-Aviles, N. Henze, D. Krause, and P. Siehndel. Analyzing the Blogosphere for Predicting the Success of Music and Movie Products. In *ASONAM '10*, 276 (2010).

[25] V. Dhar and E. A. Chang. Does Chatter Matter? The Impact of User-Generated Content on Music Sales. *J. Interat. Mark.*, **23**, 300 (2009).

[26] 山下 清美, 川浦　康至, 川上 善郎, 三浦 麻子. **ウェブログの心理学**. (NTT 出版, 2005).

[27] ブログの実態に関する調査研究. 総務省 情報通信政策研究所 (2009).

[28] 次世代 ICT 社会の実現がもたらす可能性に関する調査研究. 総務省 情報通信国際戦略局 情報通信経済室 (2011).

[29] 三浦 麻子, 森尾 博昭, 川浦　康至. **インターネット心理学のフロンティア**. (誠信書房, 2009).

[30] B. A. Nardi, D. J. Schiano, M. Gumbrecht, and L. Swartz. Why We Blog. *Commun. ACM*, **47**, 41 (2004).

[31] A. Lenhart and S. Fox. Bloggers: A portrait of the internet's new storytellers. Pew Internet & American Life Project (2006).

[32] A. Miura and K. Yamashita. Psychological and Social Influences on Blog Writing: An Online Survey of Blog Authors in Japan. *J. Comput.-Mediat. Comm.*, **12**, 1452 (2007).

[33] M. A. Stefanone and C.-Y. Jang. Writing for Friends and Family: The Interpersonal Nature of Blogs. *J. Comput.-Mediat. Comm.*, **13**, 123 (2007).

[34] M. A. Cohn, M. R. Mehl, and J. W. Pennebaker. Linguistic Markers of Psychological Change Surrounding September 11, 2001. *Psychol. Sci.*, **15**, 687 (2004).

[35] S. C. Herring, L. A. Scheidt, I. Kouper, and E. Wright. *Blogging, Citizenship, and the Future of Media*. (Routledge, London, 2007).

[36] J. Allan, J. G. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic Detection and Tracking Pilot Study Final Report. In *DARPA '98*, (1998).

[37] T. Nanno, T. Fujiki, Y. Suzuki, and M. Okumura. Automatically Collecting, Monitoring, and Mining Japanese Weblogs. In *WWW Alt. '04*, 320 (2004).

[38] J. Kleinberg. Bursty and Hierarchical Structure in Streams. *Data Min. Knowl. Disc.*, **7**, 373 (2003).

[39] 藤木 稔明, 南野 朋之, 鈴木 泰裕, 奥村 学. document stream における burst の発見. **情報処理学会研究報告**. **自然言語処理研究会報告**, 85 (2004).

[40] L. A. Adamic and N. Glance. The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. In *LinkKDD '05*, 36 (2005).

[41] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading Behavior in Large Blog Graphs. In *SDM '07*, (2007).

[42] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information Diffusion Through Blogspace. In *WWW '04*, 491 (2004).

[43] S. González-Bailón, J. Borge-Holthoefer, A. Rivero, and Y. Moreno. The Dy-

namics of Protest Recruitment through an Online Network. *Sci. Rep.*, **1**, 197 (2011).

[44] E. Bakshy, J. Hofman, W. A. Mason, and D. J. Watts. Everyone's an Influencer: Quantifying Influence on Twitter. In *WSDM '11*, 65 (2011).

[45] S. Myers, C. Zhu, and J. Leskovec. Information Diffusion and External Influence in Networks. In *KDD '12*, 33 (2012).

[46] L. Weng, A. Flammini, A. Vespignani, and F. Menczer. Competition among memes in a world with limited attention. *Sci. Rep.*, **2**, 335 (2012).

[47] 濱岡 豊, 里村 卓也. **消費者間の相互作用についての基礎研究**. (慶応義塾大学出版会, 2009).

[48] B. Lyons and K. Henderson. Opinion leadership in a computer-mediated environment. *J. Consum. Behav.*, **4**, 319 (2005).

[49] K. Niederhoffer, R. Mooth, D. Wiesenfeld, and J. Gordon. The Origin and Impact of CPG New-Product Buzz: Emerging Trends and Implications. *J. Adv. Res.*, **47**, 420 (2007).

[50] A. Ishii, H. Arakaki, N. Matsuda, S. Umemura, T. Urushidani, N. Yamagata, and N. Yoshida. The 'hit' phenomenon: a mathematical model of human dynamics interactions as a stochastic process. *New J. Phys.*, **14**, 063018 (2012).

[51] F. Omori. On the After-shocks of Earthquakes. *J. Coll. Sci. Imper. Univ. Tokyo*, **7**, 111 (1894).

[52] P. Klimek, W. Bayer, and S. Thurner. The blogosphere as an excitable social medium: Richter's and Omori's Law in media coverage. *Physica A*, **390**, 3870(2011).

[53] M. Mitrović and G. Paltoglou. Quantitative analysis of bloggers' collective behavior powered by emotions. *J. Stat. Mech.*, **2011**, P02005 (2011).

[54] M. Mitrović and B. Tadić. Bloggers behavior and emergent communities in Blog space. *Eur. Phys. J. B*, **73**, 293 (2010).

[55] ソーシャルメディアの利用実態に関する調査研究. 総務省 情報通信国際戦略局 情報通信経済室 (2010).

[56] Y. Sato, T. Utsuro, T. Fukuhara, Y. Kawada, Y. Murakami, H. Nakagawa, and N. Kando. Analysing Features of Japanese Splogs and Characteristics of Keywords. In *AIRWeb '08*, 33 (2008).

[57] Y. Sato, T. Utsuro, T. Fukuhara, Y. Kawada, Y. Murakami, H. Nakagawa, and N. Kando. Collecting and Analyzing Japanese Splogs based on Characteristics of Keywords. In *ICWSM '11*, 218 (2008).

[58] T. Takeda and A. Takasu. A splog Filtering Method Based on String Copy Detection. In *ICADIWT '08*, 543 (2008).

[59] 竹田 隆治. ブログタイトルに注目した splogger の判別手法. **情報処理学会研究報告. 自然言語処理研究会報告**, 89 (2009).

[60] M. A. de Menezes and A.-L. Barabási. Separating Internal and External Dynamics of Complex Systems. *Phys. Rev. Lett.*, **93**, 068701 (2004).

[61] H. Watanabe, Y. Sano, H. Takayasu, and M. Takayasu. The fluctuation scalings of Poisson processes with non-stationary shared parameters. (in preparation).

[62] H.-H. Jo, M. Karsai, J. Kertész, and K. K. Kaski. Circadian pattern and burstiness in mobile phone communication. *New J. Phys.*, **14**, 013055 (2012).

[63] T. Yasseri, R. Sumi, and J. Kertész. Circadian Patterns of Wikipedia Editorial Activity: A Demographic Analysis. *PLoS ONE*, **7**, e30091 (2012).

[64] B. Efron and R. Thisted. Estimating the Number of Unseen Species: How Many Words Did Shakespeare Know? *Biometrika*, **63**, 435 (1976).

[65] R. Lambiotte, M. Ausloos, and M. Thelwall. Word statistics in Blogs and RSS feeds. *J. Informetr.*, **1**, 277 (2007).

[66] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *J. Comput. Sci.*, **2**, 1 (2011).

[67] J. D. Hamilton. *The Time Series Analysis*. (Princeton University Press, 1994).

[68] R. Davidson and J. G. MacKinnon. *Econometric Theory and Methods*. (Oxford University Press, New York, 2004).

[69] S. E. Said and D. A. Dickey. Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order. *Biometrika*, **71**, 599 (1984).

[70] L. Taylor. Aggregation, Variance and the Mean. *Nature*, **189**, 732 (1961).

[71] Z. Eisler, I. Bartos, and J. Kertész. Fluctuation scaling in complex systems: Taylor's law and beyond. *Adv. Phys.*, **57**, 89 (2008).

[72] M. A. de Menezes and A.-L. Barabási. Fluctuations in Network Dynamics. *Phys. Rev. Lett.*, **92**, 028701 (2004).

[73] S. Meloni, J. Gómez-Gardeñes, V. Latora, and Y. Moreno. Scaling Breakdown in Flow Fluctuations on Complex Networks. *Phys. Rev. Lett.*, **100**, 208701 (2008).

[74] T. Utsu. A statistical study on the occurrence of aftershocks. *Geophys. Mag.*, **30**, 521 (1961).

[75] L. De Arcangelis, C. Godano, E. Lippiello, and M. Nicodemi. Universality in Solar Flare and Earthquake Occurrence. *Phys. Rev. Lett.*, **96**, 051102 (2006).

[76] B. Suki, A.-L. Barabási, Z. Hantos, F. Petak, and H. E. Stanley. Avalanches and power-law behaviour in lung inflation. *Nature*, **368**, 615 (1994).

[77] P. C. Ivanov, L. A. N. Amaral, A. L. Goldberger, S. Havlin, M. G. Rosenblum, Z. R. Struzik, and H. E. Stanley. Multifractality in human heartbeat dynamics. *Nature*, **399**, 461 (1999).

[78] T. Mizuno, M. Takayasu, and H. Takayasu. The mechanism of double exponential growth in hyper-inflation. *Physica A*, **308**, 411 (2002).

[79] D. Sornette, H. Takayasu, and W.-X. Zhou. Finite-time singularity signature of hyperinflation. *Physica A*, **325**, 492 (2003).

[80] N. Vandewalle, M. Ausloos, Ph. Boveroux, and A. Minguet. How the financial crash of October 1997 could have been predicted. *Eur. Phys. J. B*, **4**, 139 (1998).

[81] F. Lillo and R. N. Mantegna. Power-law relaxation in a complex system: Omori law after a financial market crash. *Phys. Rev. E*, **68**, 016119 (2003).

[82] A. M. Petersen, F. Wang, S. Havlin, and H. E. Stanley. Market dynamics immediately before and after financial shocks: Quantifying the Omori, productivity, and Bath laws. *Phys. Rev. E*, **82**, 036114 (2010).

[83] T. Preis, J. J. Schneider, and H. E. Stanley. Switching processes in financial markets. *Proc. Natl. Acad. Sci.*, **108**, 7674 (2011).

[84] Z. Dezsö, E. Almaas, A. Lukács, B. Rácz, I. Szakadát, and A.-L. Barabási. Dynamics of information access on the web. *Phys. Rev. E*, **73**, 066132 (2006).

[85] A. Johansen and D. Sornette. Download relaxation dynamics on the WWW following newspaper publication of URL. *Physica A*, **276**, 338 (2000).

[86] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proc. Natl. Acad. Sci.*, **105**, 15649 (2008).

[87] D. Sornette, F. Deschâtres, T. Gilbert, and Y. Ageon. Endogenous versus exogenous shocks in complex networks: An empirical test using book sale rankings. *Phys. Rev. Lett.*, **93**, 228701 (2004).

[88] V. Alfi, G. Parisi, and L. Pietronero. Conference registration: how people react to a deadline. *Nat. Phys.*, **3**, 746 (2007).

[89] R. Lambiotte and M. Ausloos. Endo- vs. exogenous shocks and relaxation rates in book and music "sales". *Physica A*, **362**, 485 (2006).

[90] 金 明哲, 村上 征勝, 永田 昌明. **言語と心理の統計**. (岩波書店, 2003).

[91] H. Ebbinghaus. *Memory: A Contribution to Experimental Psychology.* (Teachers college, Columbia university, New York, 1913).

[92] A.-L. Barabási. *Bursts: The Hidden Pattern Behind Everything We Do.* (Dut-

ton Adult, 2010).

[93] K.-I. Goh and A.-L. Barabási. Burstiness and memory in complex systems. *Europhys. Lett.*, **81**, 48002 (2008).

[94] A.-L. Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, **435**, 207 (2005).

[95] K. Yamada, Y. Sano, H. Takayasu, and M. Takayasu. Understanding General Human Behavior by Cyberspace Communication data. (in preparation).

[96] G. K. Zipf. *Human Behavior and the Principle of Least Effort*. (Addison-Wesley, Cambridge, 1949).

[97] H. S. Heaps. *Information retrieval: Computational and Theoretical Aspects*. (Academic Press, New York, 1978).

[98] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, The Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, **331**, 176 (2011).

[99] E. Lieberman, J.-B. Michel, J. Jackson, T. Tang, and M. A. Nowak. Quantifying the evolutionary dynamics of language. *Nature*, **449**, 713 (2007).

[100] M. Gerlach and E. G. Altmann. Stochastic Model for the Vocabulary Growth in Natural Languages. *Phys. Rev. X*, **3**, 021006 (2013).

[101] T. Yasseri, A. Kornai, and J. Kertész. A Practical Approach to Language Complexity: A Wikipedia Case Study. *PLoS ONE*, **7**, e48386 (2012).

[102] A. M. Petersen, J. N. Tenenbaum, S. Havlin, H. E. Stanley, and M. Perc. Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Sci. Rep.*, **2**, 943 (2012).

[103] A. M. Petersen, J. Tenenbaum, S. Havlin, and H. E. Stanley. Statistical Laws Governing Fluctuations in Word Use from Word Birth to Word Death. *Sci. Rep.*, **2**, 313 (2012).