Discussion Paper Series

RIEB

Kobe University

DP2017-16

Measuring Social Change Using Text Data: A Simple Distributional Approach

Takashi KAMIHIGASHI

Kazuhiro SEKI

Masahiko SHIBAMOTO

Revised July 5, 2017

Research Institute for Economics and Business Administration

Kobe University

2-1 Rokkodai, Nada, Kobe 657-8501 JAPAN

Measuring Social Change Using Text Data: A Simple Distributional Approach*

Takashi Kamihigashi† Kazuhiro Seki‡
Masahiko Shibamoto§

July 5, 2017

Abstract

This paper proposes a simple approach to measuring social change using text data. The approach is based on the idea that any significant change in a society should affect the distribution of the words used in the society. Essentially we use the total variation distance between the distributions of words in adjacent months as a measure of social change during the latter month. Based on text data from the Nikkei Newspaper from 1989 to 2015, the largest social change observed in Japan during this period took place in March 2011, the month of the Great East Japan Earthquake.

Keywords: Social change; text analytics; newspaper; keyword extraction; event detection

*Financial support from the Japan Society for the Promotion of Science ("Topic-Setting Program to Advanced Cutting-Edge Humanities and Social Sciences Research"; KAKENHI 15H05729) is gratefully acknowledged.

†RIEB (Research Institute for Economics and Business Administration), Kobe University, 2-1 Rokkodai, Nada, Kobe 657-8501 Japan. tkamihig@rieb.kobe-u.ac.jp

‡Faculty of Intelligence and Informatics, Konan University, Okamoto, Higashinada, Kobe, 658-8501 Japan; RIEB Research Fellow, RIEB, Kobe University, 2-1 Rokkodai, Nada, Kobe 657-8501 Japan. seki@konan-u.ac.jp

§RIEB, Kobe University, 2-1, Rokkodai, Nada, Kobe 657-8501 Japan. shibamoto@rieb.kobe-u.ac.jp

1 Introduction

Social change is difficult to measure. While various researchers have discussed the issues and difficulties involved in measuring social change (e.g., Garonna and Triacca, 1999; Livingstone, 2002; Goodwin, 2009; Phillips, 2011; Antadze and Westly, 2012), none seem to agree on a method for performing the measurement.

In this paper we propose a simple approach to measuring social change using text data. The approach is based on the idea that any significant change in a society should affect the distribution of the words used in the society. When a new system is introduced, for example, it comes with new words and expressions to describe it; new laws and institutions are often given new names; the advent of new technology such as information and communications technology often leads to the invention of numerous new terms.

The analysis in this paper uses text data from the Nikkei Newspaper from March 1989 to December 2015. The total number of words printed in the nationwide version of Nikkei over this period stayed fairly stable on a monthly basis, providing a proper environment to measure significant changes in the text data. Our measure of social change is based on the total variation distance between the distributions of words in adjacent months. Essentially we interpret a large change in the word distribution as an indication of a significant change occurring in the society.

This paper is only a preliminary attempt to measure social change based on this approach. Our purpose at this stage is to investigate whether this naive approach can lead to any meaningful result. As it turns out, our results show that large changes in the word distribution tend to take place upon the occurrence of major events likely to impact the society. The largest change in the word distribution during the sample period took place in March 2011, the month of the Great East Japan Earthquake.

While measurement of the social impacts of events was not our main purpose here, we found that a major change in the word distribution could be associated with a major social event by examining candidate keywords, i.e., words that sharply increase in usage in parallel with the change in the word distribution. Our method for selecting candidate keywords is somewhat similar to the approach proposed by Andrade and Valencia (1998) for automatic keyword extraction from scientific text.

Though we know of no longitudinal analysis linking changes in the word

distribution in newspapers with significant social events, periodic patterns of word frequencies in historical newspaper data over 87 years were analyzed in a recent study by Dzongang et al. (2016). In addition, much research has been done on detecting major topics in time-tagged streams of textual data, such as microblogs and news stories (Aggarwal and Subbian, 2012; Sayyadi et al., 2009; Yang et al., 1998). Swan and Allan (2000) made a seminal attempt to find clusters of terms indicative of major news topics. Their approach adopts classical hypothesis tests to assess the significance of term appearances on given dates and identifies named entities (e.g., proper names) and noun phrases appearing more frequently than statistically expected. It then further identifies co-occurrences of identified terms that together represent major topics. The resulting series of topics can be presented as a timeline of topics occurring over the period covered by the input textual data.

Topic models based on Latent Dirichlet Allocation (LDA) (Blei et al., 2003) have also been used for topic detection. LDA, an approach that regards word occurrences as a generative process from unobserved, hidden topics, allows one to estimate underlying topical mixtures in input text.[1] Several extensions of LDA have been proposed for the analysis of the temporal properties of text (AlSumait et al., 2008; Griffiths and Steyvers, 2004). One of the extensions, the "Topic over Time" model (Wang and McCallum, 2006), associates each topic with a continuous distribution over timestamps, enabling one to analyze how each of the discovered topics is distributed over time.[2]

Our approach differs from those mentioned above in that our primary interest is in measuring social change by computing the total variation distance between the word distributions in adjacent months without considering underlying events. It is only after identifying months when major changes in the word distribution occurred that we select candidate keywords and events to investigate the possible causes of the changes.

The rest of this paper is organized as follows. Section 2 formally presents our approach. Section 3 describes how we constructed the text data used in our analysis. Section 4 discusses the seasonal and trend-cycle adjustments applied to our data and identifies the 30 specific months in which the 30 largest changes in the word distribution took place. Section 5 discusses the candidate keywords and major events associated with the 10 largest changes

[1] Zhao et al. (2011) used LDA to detect topics in text data from the New York Times over a four months' period.

[2] See Atefeh and Khreich (2015), Goswami and Kumar (2016), Cordeiro and Gama (2016), and Hasan et al. (2017) for recent surveys on event detection in Twitter data.

in the word distribution. Section 6 concludes the paper by discussing possible issues and extensions for future research.

2 A Simple Distributional Approach

Let W be a finite set of distinct words. Let T be a set of time periods: for example, each $t \in T$ may be a month in a year. For each $t \in T$, we are given some text data, D_t, consisting of instances of words from W. For $w \in W$ and $t \in T$, let $n_{w,t}$ be the number of occurrences of word w in D_t. Let N_t be the total number of words contained in D_t:

$$N_t = \sum_{w \in W} n_{w,t}. \tag{2.1}$$

Let $s_{w,t}$ be the share of word w in D_t; more precisely,

$$s_{w,t} = \frac{n_{w,t}}{N_t}. \tag{2.2}$$

It follows from (2.1) and (2.2) that

$$\sum_{w \in W} s_{w,t} = 1. \tag{2.3}$$

Thus $\{s_{w,t}\}_{w \in W}$ can be considered a probability distribution over W. Let c_t be the (normalized) total variation distance between the distributions $\{s_{w,t-1}\}_{w \in W}$ and $\{s_{w,t}\}_{w \in W}$:

$$c_t = \frac{1}{2} \sum_{w \in W} |s_{w,t} - s_{w,t-1}|. \tag{2.4}$$

We call c_t the distributional change in period t. As $0 \leq c_t \leq 1$ by (2.4), it is the percentage change in the distribution of words from period $t-1$ to period t in the total variation sense

3 Data

Our data for this study was text data from the morning and evening editions of the Nikkei Newspaper from January 1982 to December 2015.[3] While this

[3]The data set was purchased from Nikkei Media Marketing, Inc.

newspaper publishes both nationwide and regional editions, we only used the former to keep our data stable, as the number of regional editions grew over time.

We processed the title and main body of each article using a Japanese morphological analyzer called MeCab,[4] to extract the surface forms of the words appearing in the articles. We then divided the text data into separate words. To facilitate our analysis, we discarded "words" of the following four types:

1. Words composed exclusively of numerals, either Roman or Chinese.

2. Words composed exclusively of symbols, i.e., containing no *kanji* (or Chinese) characters, no *hiragana* letters, no *katakana* letters,[5] and no alphabet letters.

3. Words composed exclusively of *hiragana* letters.

4. The names of the months (i.e., January, February, ..., December).

While the words of types 1, 2, and 3 are mostly uninformative, they outnumber content words in the raw data because the analyzer recognizes a potentially countless number of unique strings of numerals, symbols, or *hiragana* letters as distinct words.

Words of type 1 are merely numbers. Though some of these numbers may have special meanings, we applied the simple rule of removing all of them. Words of type 2 are not words in the usual sense: many are used to separate newspaper articles and paragraphs. Words of type 3 are mostly function words, the equivalents of articles, pronouns, prepositions, conjunctions, etc., in English. Meaningful nouns and verbs written entirely in *hiragana* are rare.

Words of type 4, the names of the months, are informative, but they appear in a highly seasonal way. The word "July," for example, is used in many articles printed in July. This increase in the frequency of "July" in July has no particular relevance for our purposes.

With the above words removed, our text data still contains a total of 850,634 distinct words. These words constitute the set W introduced at the beginning of Section 2. Our text data D_t for each time period $t =$ January

[4] `http://taku910.github.io/mecab/`

[5] *Hiragana* is the primary Japanese syllabary.

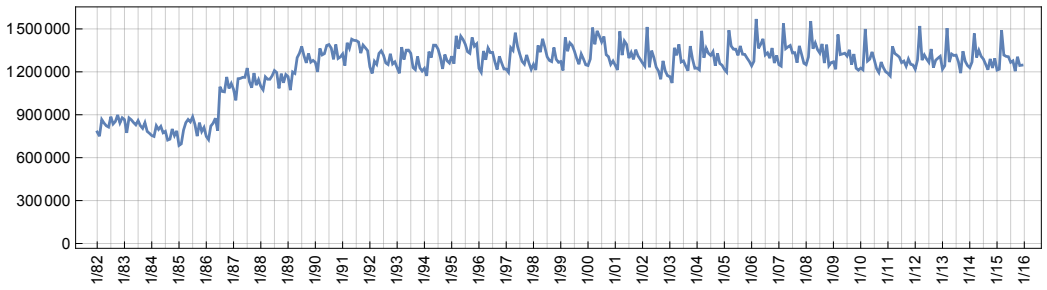


Figure 1: Total number of words in D_t (N_t)

1982, ..., December 2015 consists of possibly multiple instances of words from W.[6]

4 Seasonal and Trend-Cycle Adjustments

Figure 1 shows N_t (recall (2.1)), the total number of words contained in D_t, for t = January 1982, ..., December 2015. As the figure demonstrates, the total number of words more or less stabilizes after March 1989. Given the very high likelihood that the changes in the total number of words before March 1989 affected the distributional changes during the same period, we used only the data from March 1989 to December 2015 for our main analysis.

A number of apparent seasonal patterns emerge in Figure 1. Most prominently, we see a peak in the total number of words in March of every year after 1999, except in 2003. This upsurge in words seems to be related to the scheduling of the Japanese fiscal year, which starts in April. Having no particular interest in seasonal variations, we applied a seasonal adjustment to the time series for the distributional change c_t (recall (2.4)), as we explain below.

Figure 2(a) plots c_t for t = April 1989, ..., December 2015. Recall from (2.4) that c_t is the percentage change in the distribution of words from period $t-1$ to period t. The peak in March 2011 is immediately noticeable. Remarkably, more than 22% of the word distribution changed in March 2011, in the total variation sense. This distributional change can easily be associated

[6]In fact, we directly computed $\{n_{w,t}\}_{w\in W, t\in T}$ from the raw text data without explicitly constructing D_t.

with the Great East Japan Earthquake, which struck on March 11, 2011. Two other noticeable peaks appear in September 2000 and September 2008: the first can be associated with the Sydney 2000 Olympics; the second with the bankruptcy of Lehman Brothers. We discuss these events in more detail in Section 5.

To make a fair comparison of the magnitudes of distributional changes at different points in time, we applied some adjustments to the time series for the distributional change in Figure 2(a). To be specific, we used the X12-ARIMA program provided by the U.S. Census Bureau to decompose the series into a sum of seasonal, trend-cycle, and irregular components. According to Bee Dagum and Bianoconcini (2016, p. 95), "The X12ARIMA is today the most often applied seasonal adjustment method by statistical agencies. It was developed by Findley et al. [9] and is an enhanced version of the X11Arima method." Details on the X12-ARIMA program can be found in Findley et al. (1998), U.S. Census Bureau (2011), and Bee Dagum and Bianoconcini (2016).

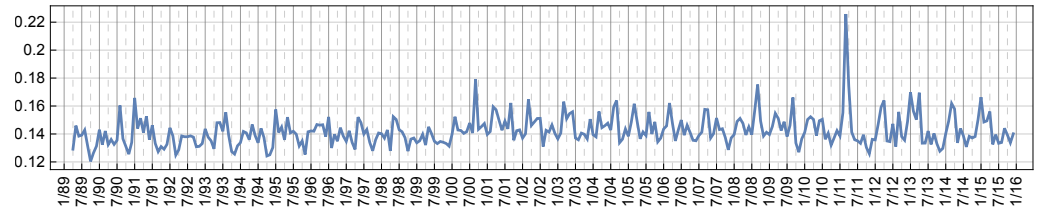
We ran the X12-ARIMA program by setting the regARIMA model in the order $(1,0,0)(0,1,1)_{12}$ with regressors for a constant and additive outliers. The program then automatically selected September 2000 and March 2011 as additive outliers. With these outliers, we ran the X11-ARIMA program in its default setting.

Figure 2(b) and (c) plot the estimated seasonal component and the seasonally adjusted series obtained by subtracting the former from the unadjusted series in Figure 2(a). The presence of seasonality was supported at the 0.1% significance level by a standard F-test and at the 1% level by a nonparametric F-test.

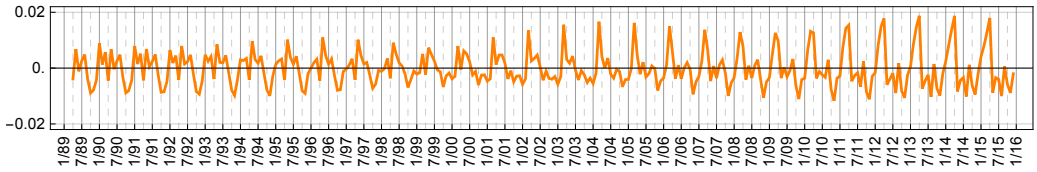
Figure 2(b) shows a gradual change in seasonal patterns over time. The seasonal component peaks in May from 1992 to 1999, in March from 2000 to 2010, and in April from 2011 to 2015. While it is unsurprising to observe the largest distributional change in around April, the start of the Japanese fiscal year, we have no clear explanation for the shift of the peak month over time.

Figure 2(d) plots the trend-cycle component, which was obtained by adjusting the trend-cycle component returned by the X12-ARIMA program by subtracting its mean. Observe that overall, the trend-cycle component takes larger values after July 2000 than before. This suggests that the word distribution changed more rapidly on a monthly basis after July 2000.

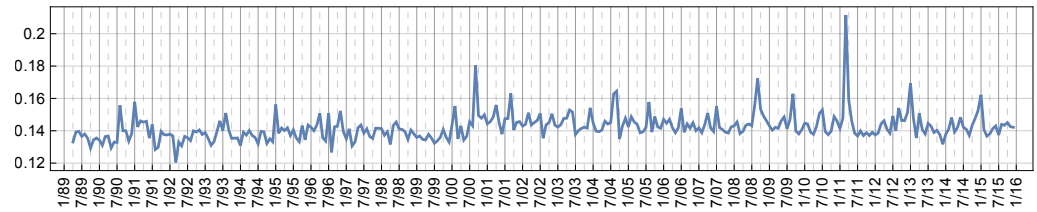
Finally, Figure 2(e) plots the fully adjusted series for the distributional



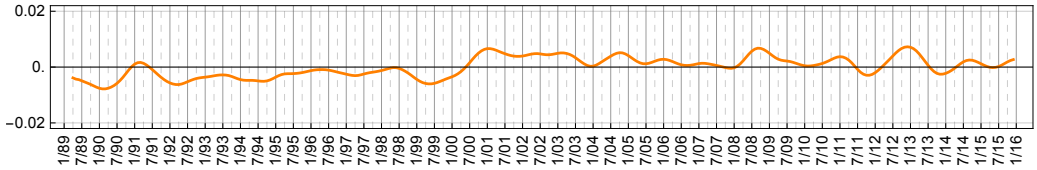
(a) distributional change (c_t) (unadjusted)



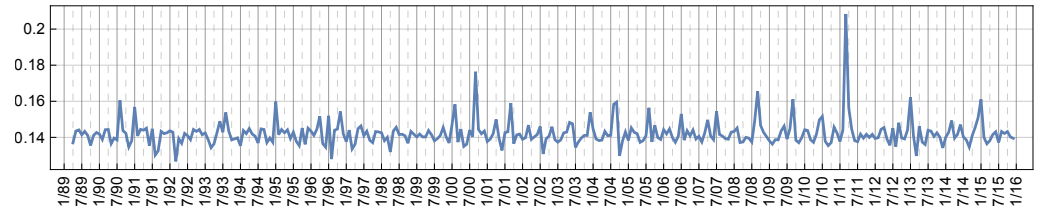
(b) seasonal component



(c) seasonally adjusted distributional change



(d) trend-cycle component



(e) fully adjusted distributional change

Figure 2: Decompositions of distributional change (c_t)

change, which was obtained by subtracting the trend-cycle component from the seasonally adjusted series. Our discussion in the subsequent sections is mostly based on this series.

Table 1 lists the months of the 30 largest peaks in the fully adjusted series along with the corresponding values of the seasonally adjusted and unadjusted series. Note that the order of the three highest peaks is common to the three series. The ranking for April 2011, however, changes from fourth in the unadjusted series to 14th in the fully adjusted series. The ranking for September 2004, meanwhile, changes from fifth in the seasonally adjusted series to ninth in the fully adjusted series.

5 Keyword Extraction and Event Association

Tables 2 and 3 list the top 30 words with the highest share gains for the first 10 entries listed in Table 1, which correspond to the 10 highest peaks in Figure 2(d). To be more precise, for each year-month pair t in the top 10 entries in Table 1, we ordered the words in W according to their share gains $s_{w,t} - s_{w,t-1}$ in decreasing order (recall (2.2)–(2.4)) and selected the top 30 words in this order. These words can be regarded as candidate keywords for the corresponding month.

The top 30 words for March 2011 in Table 2 include "disaster," "earthquake," "East Japan," "Fukushima," and "nuclear power plant." Thus the distributional change in this month can easily be associated with the Great East Japan Earthquake, which occurred on the 11th of that month.

Figure 3 plots the shares of the top 10 words for March 2011 (marked in the figure by solid orange lines). Panels (2), (3), (5), (8), and (10) suggest that the shares of many of the words in Figure 3 increased in March 2011 due to seasonal variations. Numerous personnel changes at firms and government agencies for the upcoming year are announced in March, in the lead-up to the start of the next Japanese fiscal year in April. Thus the shares of related words are expected to increase in March, when the announcements are issued. Note, however, that the contributions of such words to the distributional change in March are controlled to a degree by the seasonal adjustment discussed in Section 4.

Panels (6) and (9) in Figure 3 show sharp increases in the shares of the words "disaster" and "earthquake" in March 2011. The shares of these words also rose steeply in January 1995 (marked in the figure by dashed orange

	Year.Month	F.A.	S.A.	Unadjusted
1	2011.03	0.2083715	0.2115671	0.2258311
2	2000.09	0.1763706	0.1803774	0.1792017
3	2008.09	0.1655838	0.1723415	0.1753649
4	2013.01	0.1622175	0.1693156	0.1697610
5	2009.09	0.1611525	0.1628637	0.1661110
6	2015.01	0.1610688	0.1622630	0.1663304
7	1990.08	0.1604640	0.1556287	0.1604273
8	1995.01	0.1598087	0.1562096	0.1575174
9	2004.09	0.1595294	0.1644664	0.1639825
10	2001.09	0.1589841	0.1631122	0.1622019
11	2004.08	0.1583339	0.1628268	0.1592294
12	2000.02	0.1583126	0.1553046	0.1523556
13	1991.01	0.1568591	0.1579167	0.1658177
14	2011.04	0.1564243	0.1589229	0.1743079
15	2005.08	0.1563922	0.1577585	0.1558020
16	2007.07	0.1544525	0.1550223	0.1514092
17	1996.11	0.1543386	0.1522831	0.1445347
18	2003.12	0.1539111	0.1541982	0.1506215
19	1993.08	0.1538091	0.1509449	0.1554801
20	2006.07	0.1529792	0.1537956	0.1499701
21	1996.07	0.1519701	0.1508850	0.1522109
22	1996.04	0.1517000	0.1507942	0.1463097
23	2010.07	0.1515982	0.1528859	0.1504785
24	2014.12	0.1508709	0.1525682	0.1504125
25	2001.04	0.1499375	0.1559717	0.1572302
26	2007.04	0.1497330	0.1509308	0.1573038
27	2010.06	0.1496744	0.1506686	0.1494634
28	2008.08	0.1493830	0.1557984	0.1567156
29	2014.03	0.1493253	0.1481587	0.1618540
30	1993.06	0.1489330	0.1460998	0.1481500

Table 1: 30 highest peaks in the fully adjusted series (Figure 2(e)) and corresponding values of the seasonally adjusted and unadjusted series.

	Year.Month	Words with the 30 highest share gains
1	2011.03	city; headquarters; head (chief); prefecture; no.; disaster; day; sales; earthquake (jishin); enterprise; East Japan; same; Fukushima; cum; town; nuclear power plant; division; great earthquake; place; integration; time; evacuation; division head; blackout; earthquake (shinsai); nuclear; power plant; sub; damage; area
2	2000.09	division head; Olympic(s); head (chief); Sydney; table; Japan; female; headquarters; cum; male; second; crude oil; sales; athlete; place (rank); final; meter; preliminary round; class; number; company; U.S.; minute; no.; enterprise; meeting; collaboration; state (country); Europe; personnel
3	2008.09	finance; America; Mr./Ms.; market; head (chief); election; fund; Aso; headquarters; organization; president (of political party or BOJ); prime minister; management; crisis; same; bank; company; person; bankruptcy; securities; day; Lehman; enterprise; -ification; dollar; election; cum; president (of a company); incident; asset
4	2013.01	last year; year; people; yen; government; Algeria; firm; fiscal year; personnel; hostage; incident; world; tax; profit; period; plane; information; market; confirmation; countermeasure; America; terrorism; sales; Japanese; eye; person; increase; tax system; price level; cheaper; Nikki
5	2009.09	day; headquarters; head (chief); minister; Hatoyama; Mr./Ms.; enterprise; administration; sales; prime minister; division; same; meeting; personnel; finance; charge; cum; problem; leader; policy; strategy; press conference; international; America; economy; Democratic Party; cabinet member; gathering; treasury; execution

Table 2: Top 30 words with the highest share gains for each of months 1, . . . , 5 in Table 1

	Year.Month	Words with the 30 highest share gains
5	2015.01	year; last year; state (country); Islam(ic); day; America; Europe; terrorism; U.S.; middle; euro; yen; people; period; profit; Japan; Jordan; settlement of accounts; eye; sales; (Mr.) Goto; world; incident; intended for; government; fiscal year; group; radical; international; announcement
7	1990.08	Iraq; Kuwait; Middle East; crude oil; miliary; petroleum; state of affairs; invasion; people; Saudi; rise; Arab; price; U.N.; sanctions; day; imminence; dispatch; Jordan; market price; military affairs; Hussein; Saudi Arabia; year; Iran; increased production; dollar; minute; price increase; barrel
8	1995.01	earthquake; last year; Kobe; Hyogo; prefecture; city; Hanshin (Osaka-Kobe); disaster; great earthquake; damage; calamity; southern part; district; restoration; Mexico; countermeasure; area; day; reconstruction; person; earthquake; current; crisis; currency; market; Osaka; aid; evacuation; Japan; outbreak
9	2004.09	head (chief); headquarters; enterprise; same; sales; person; day; cum; Mr./Ms.; prime minister; personnel; division; meeting; development; sub; company; baseball team; suspicion; firm; fiance; incident; -fication; gender; securities; integration; minister; problem; management; type; side
10	2001.09	terrorism; America; division head; simultaneous; day; U.S.; head (chief); headquarters; deal; cum; economy; building; sales; market; aid; New York; same; division; retaliation; world; influence; insurance; Pakistan; Taliban; incident; attach; credit; branch; cooperation; Self-Defense Forces

Table 3: Top 30 words with the highest share gains for each of months 6, . . . , 10 in Table 1

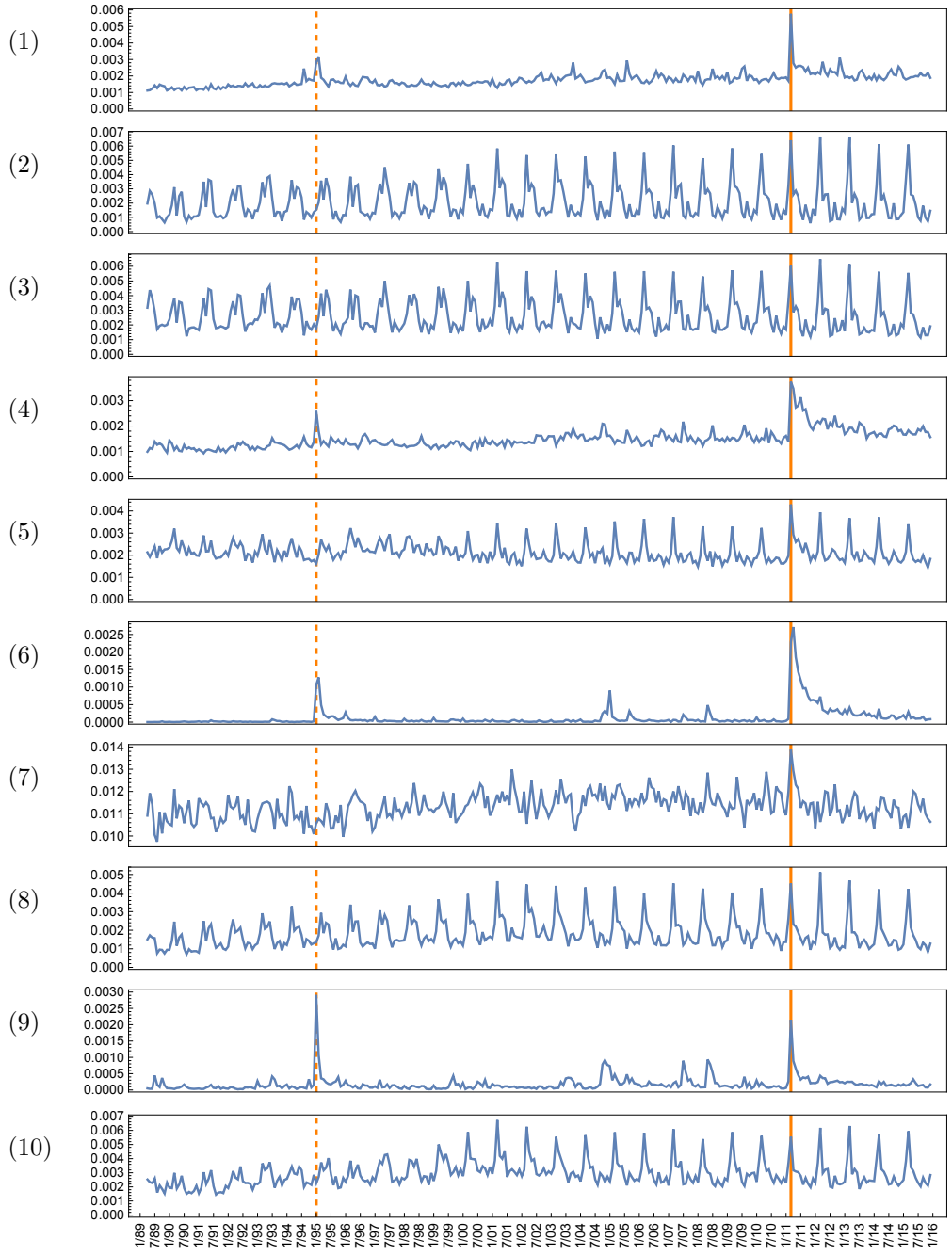


Figure 3: Shares of the top 10 words for March 2011: (1) city; (2) headquarters; (3) head (chief); (4) prefecture; (5) no.; (6) disaster; (7) day; (8) sales; (9) earthquake; (10) enterprise

lines). The upward spikes in January 1995 can be associated with the Great Hanshin-Awaji Earthquake, which struck on the 17th of that month.

Panels (1) and (4) in Figure 3 also show sharp increases in the shares of the words "city" and "prefecture" in January 1995, as well as in March 2011. This suggests that co-occurrences of such common words can be a useful indicator of an important event.

Figure 4 plots the shares of the top 10 words for September 2000 (marked by solid orange lines). Panels (2), (4), (7), and (10) show that the shares of the words "Olympic(s)," "Sydney," "female," and "male" sharply increased in September 2000. The distributional change in this month can thus be associated with the Sydney Olympics, which took place at the same time. As in Figure 3, the shares of some of the top 10 words in Figure 4 also increased in September 2000 due to seasonal variations, as can be seen in panels (1), (3), (8), and (9).

One might wonder if there was anything special about the Sydney Olympics versus the other summer Olympic games held from 1989 to 2015. The dashed orange lines in Figure 4 indicate the months when the other summer Olympic games were commenced. The precise periods of these events were as follows:

- Barcelona: 1992.07.25–1992.08.09
- Atlanta: 1996.07.19–1996.08.04
- Sydney: 2000.09.15–2000.10.01
- Athens: 2004.08.13–2004.08.29
- Beijing: 2008.08.08–2008.08.24
- London: 2012.07.24–2012.08.12

Panels (2), (7), and (10) in Figure 4 show sharp increases in the shares of the words "Olympic(s)," "female," and "male" in July 1996, August 2004, and August 2008, i.e., the months marking the start of the Olympic games in Atlanta, Athens, and Beijing, respectively. These Olympic events correspond to the 21st, the 11th, and the 28th places in Table 1, respectively. Indeed, Figure 2(c) shows that the fully adjusted series for the distributional change has significant peaks in July 1996, August 2004, and August 2008.

While the Olympic games in Athens and Beijing started and ended in August, those in Barcelona, Atlanta, and London started in July and ended

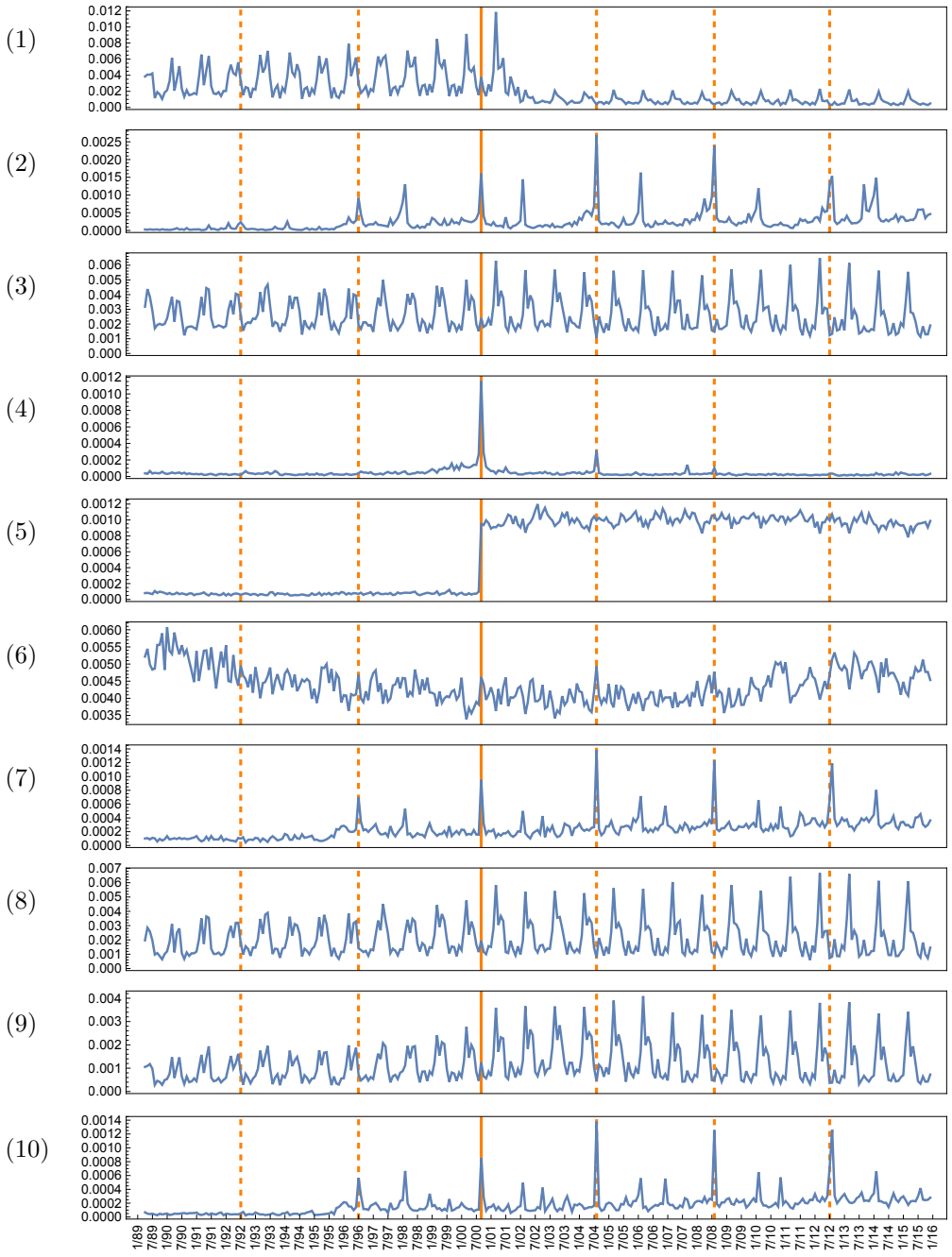


Figure 4: Shares of the top 10 words for September 2000: (1) division head; (2) Olympic(s); (3) head (chief); (4) Sydney; (5) table; (6) Japan; (7) female; (8) headquarters; (9) cum; (10) male

in August. Thus the influences of the latter on the Nikkei text data may have been divided between July and August. This may partly explain why the other three Olympic events are associated with higher peaks in the fully adjusted series for the distributional change. Moreover, in Panel (5) in Figure 4, we see that the share of the word "table" sharply rose in September 2000 and remained at a similar level thereafter. This rise may have been attributable to a change in Nikkei's policy on text data. Other changes in Nikkei's policy might have contributed to the distributional change in September 2000. They also might have helped distinguish the Sydney Olympics from the Olympic games in other cities.

Figure 5 plots the shares of the top 10 words for September 2008 (marked by solid orange lines). As mentioned above, the distributional change here can be associated with the Lehman Brothers bankruptcy filed on September 15, 2008 in the United State. Panels (1), (4), (7), and (10) show sharp rises in the shares of closely related words such as "finance," "market," "fund," and "organization" in September 2008. The shares of these words also sharply increased during the 1997 Asian financial crisis and 1998 Russian financial crisis (marked by dashed orange lines in Figure 5).

Figure 6 plots the shares of the top 10 words for January 2013 (marked by solid orange lines). The shares of the words "Algeria" and "hostage" dramatically increased in January 2013, indicating that the distributional change in this month can be associated with the Algeria hostage crisis. The literal translation of this crisis in Japanese, "Algeria hostage incident," explains the higher share of the word "incident" occurring in the same month. The dashed lines indicate the months of the 1996 Japanese embassy hostage crisis in Peru and 2004 foreign hostage crisis in Iraq. Similarities in the movements of the shares of certain words appear during these months.

Figure 7 plots the shares of the top 10 words for September 2009 (marked by solid orange lines). The distributional change in this month can be associated with the historic victory of the Democratic Party of Japan, led by Yukio Hatoyama, over the long-ruling Liberal Democratic Party (LDP) in the Japanese House of Representatives. The dashed orange lines indicate the three months in the 1990s when non-members of the LDP became prime minister, and the month when the LDP, led by Shinzo Abe, returned to power.

Figures 8–10 plot the shares of the top 10 words for January 2015, August 1990, and September 2001. We include September 2001 here for convenience even though in Tables 1 and 3, it appears below January 1995 and September

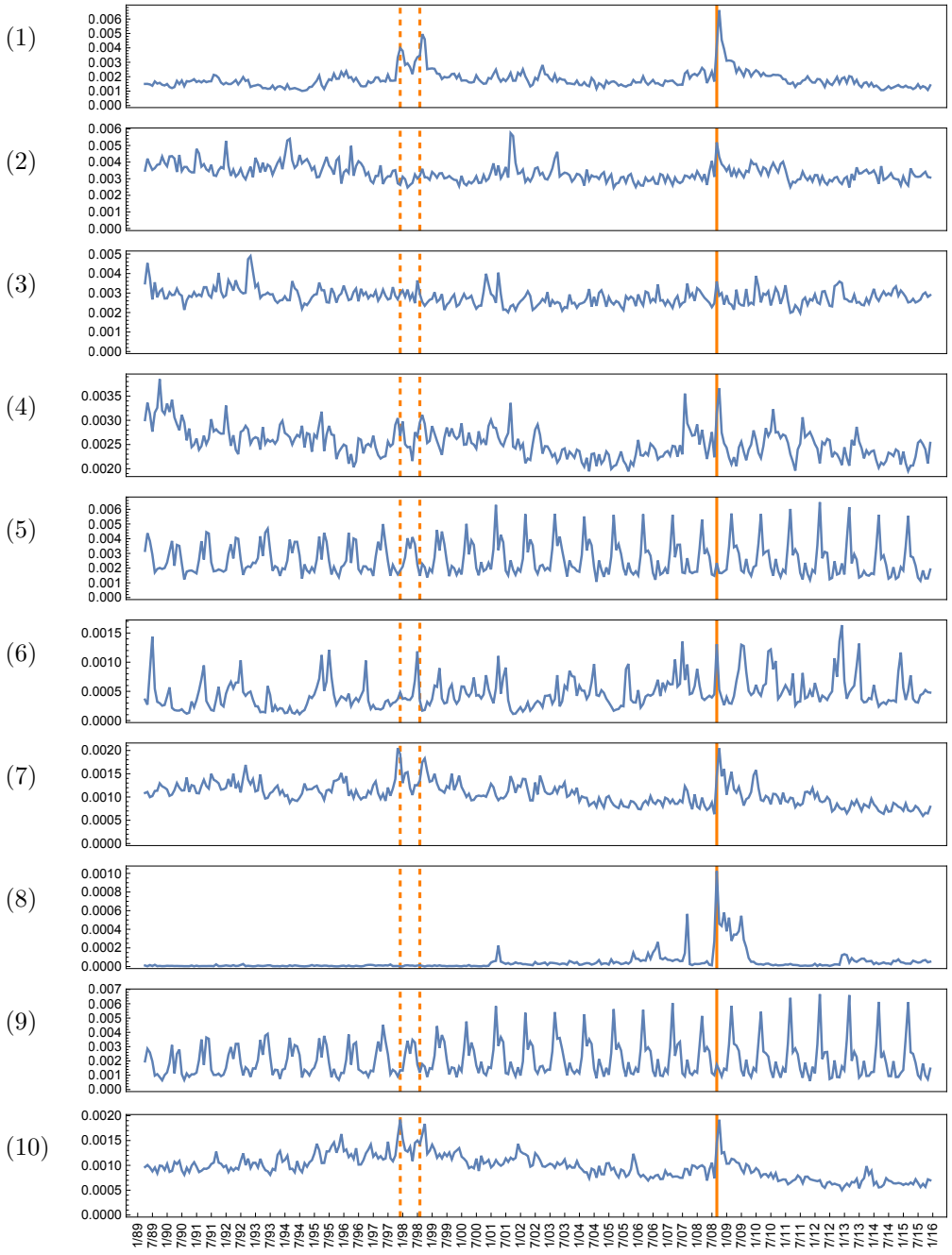


Figure 5: Shares of the top 10 words for September 2008: (1) finance (or financial); (2) America; (3) Mr./Ms.; (4) market; (5) head (chief); (6) election; (7) fund; (8) Aso; (9) headquarters; (10) organization

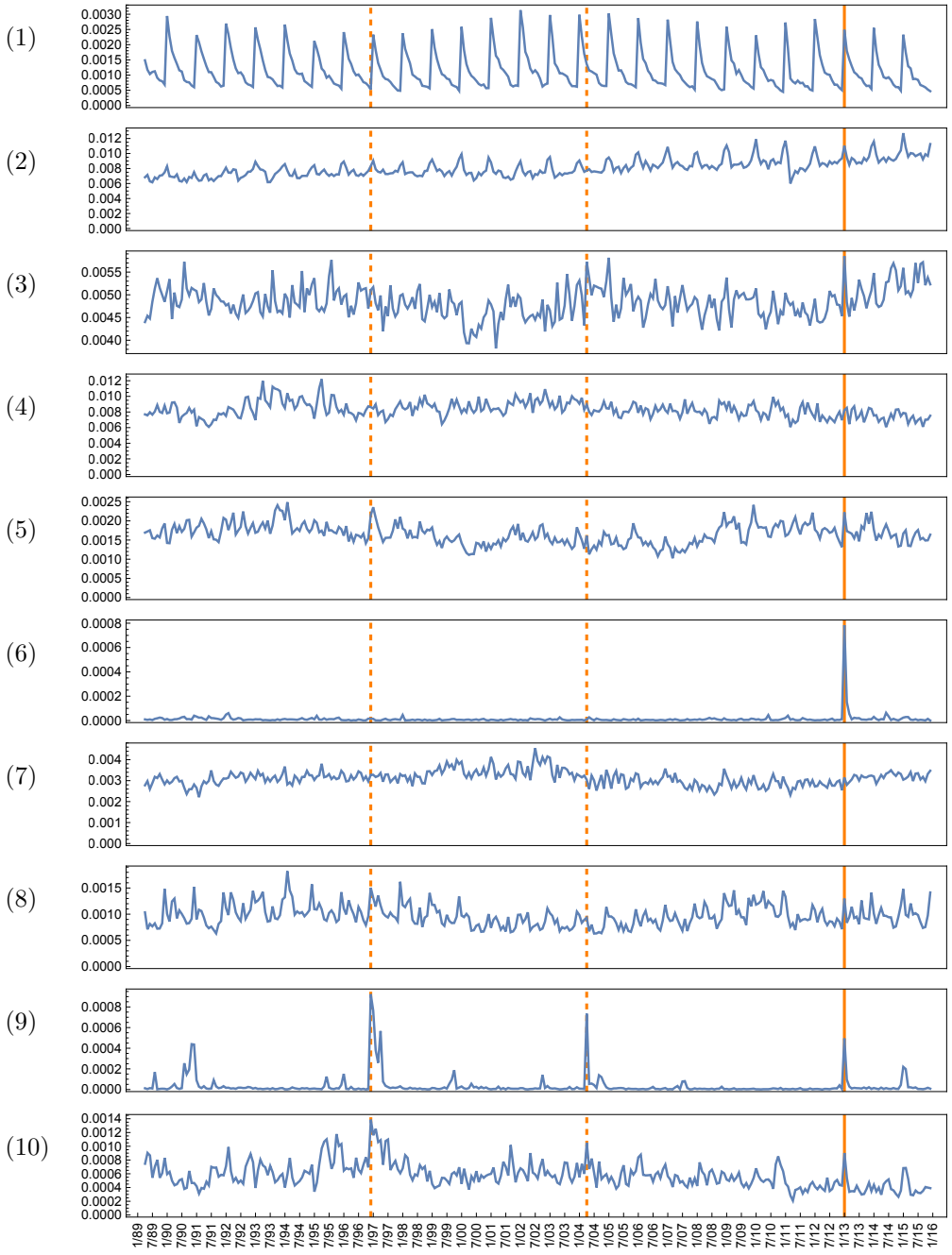


Figure 6: Shares of the top 10 words for January 2013: (1) last year; (2) year; (3) people; (4) yen; (5) government; (6) Algeria; (7) firm; (8) fiscal year; (9) hostage; (10) incident

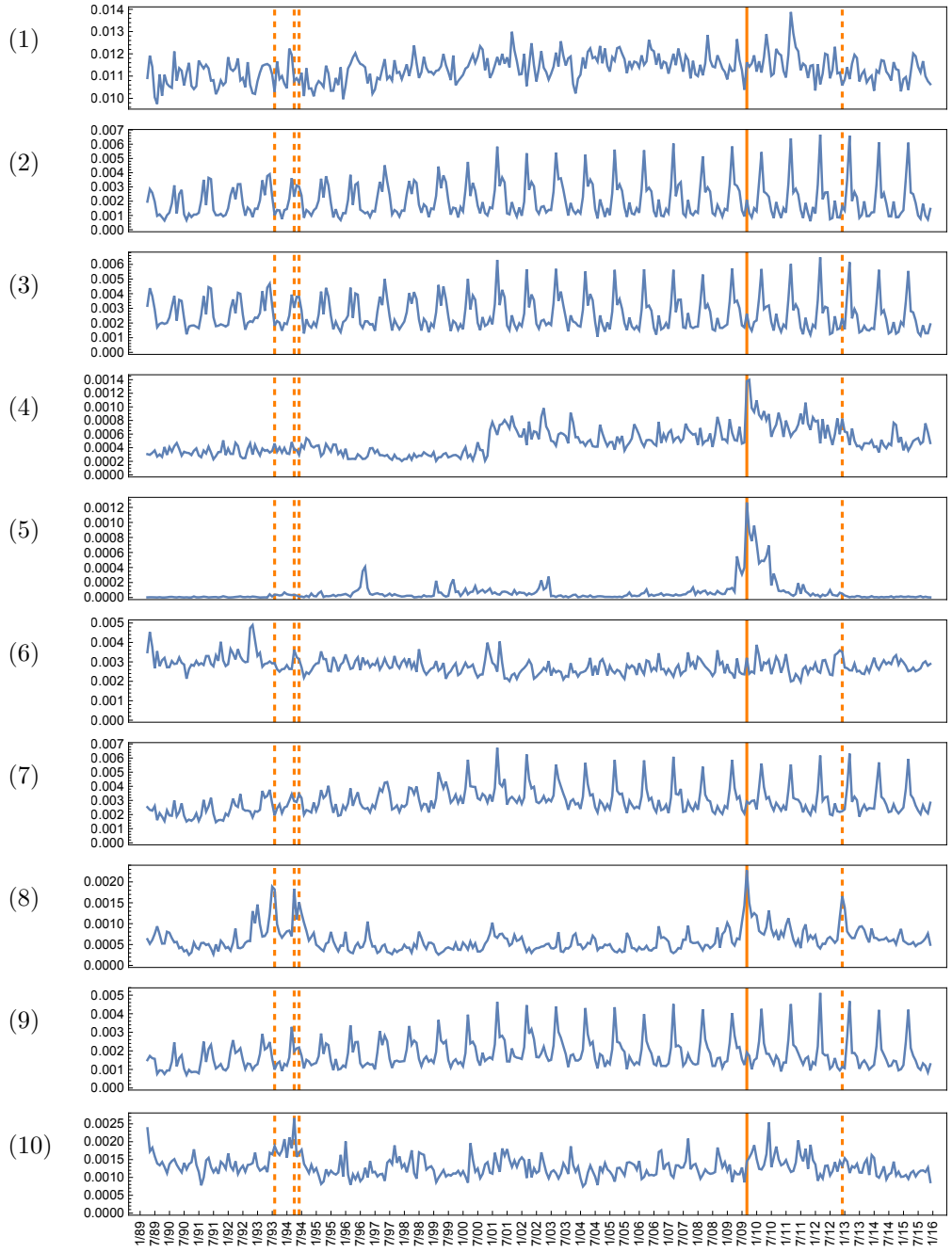


Figure 7: Shares of the top 10 words for September 2009: (1) day; (2) headquarters; (3) head (chief); (4) minister; (5) Hatoyama; (6) Mr./Ms.; (7) enterprise; (8) administration; (9) sales; (10) prime minister

2004, which we discuss below. In Figure 8, the solid orange lines indicate the Islamic State ransom incident in January 2015, while the dashed orange lines indicate the September 11 attacks in 2001. In Figure 9, the solid orange lines indicate the Iraqi invation of Kuwai, while the dashed orange lines indicate the Gulf War and the Iraq War. In Figure 10, the solid orange lines indicate the September 11 attacks. One can make various observations in these figures, as in Figures 3–7.

Figure 11 plots the shares of the top 10 words for January 1995. The solid lines in the figure indicate the Great Hanshin-Awaji Earthquake of January 17, 1995, while the dashed orange lines indicate the Great East Japan Earthquake. The word "earthquake" had a larger share in January 1995 than in March 2011, but the words "prefecture," "city," "disaster," "great earthquake," and "damage" all had larger shares in March 2011 than in January 1995. This is consistent with the pattern of damage caused by the Great East Japan Earthquake, which was larger overall but not a direct consequence of the earthquake itself.

Figure 12 plots the shares of the top 10 words for September 2004 (marked by solid orange lines). No specific event can be easily associated with the plot in this figure, which shows no dramatic increases in the shares of any words in September 2004. Figure 13 plots the shares of the next ten words. The share of one word in this figure, "baseball team," sharply increased in September 2004. Figure 14 plots the shares of the next 10 words, but again, the plot shows no dramatic increases in the shares of any words in September 2004. We see in Figure 15, however, that the share of the very next word, "strike," sharply increased in September 2004. Hence the distributional change in this month can be associated with the first strike by the professional baseball players in the Nippon Professional Baseball leagues. The strike was triggered by a heated debate on the possible restructuring of the baseball leagues. Other factors may of course have contributed to the distributional change in this month.

Table 4 recapitulates the major events associated with the top 10 peaks in the fully adjusted series for the distributional change in Figure 2(e).

6 Discussion

As discussed above, the top 30 words for a major peak in the fully adjusted series for the distributional change often include many "seasonal" words.

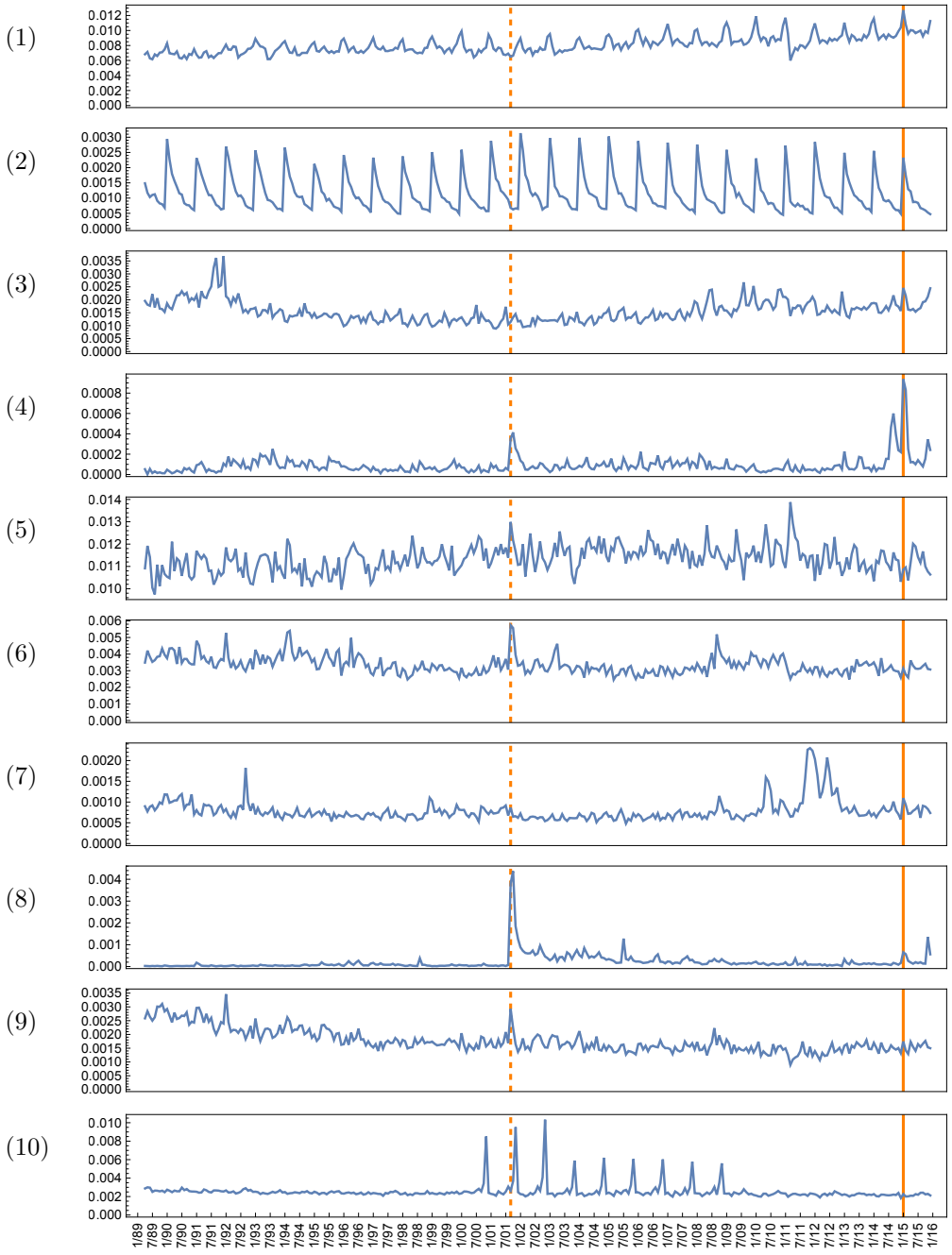


Figure 8: Shares of the top 10 words for January 2015 (1) year; (2) last year; (3) state (country); (4) Islam(ic); (5) day; (6) America; (7) Europe; (8) terrorism; (9) U.S.; (10) middle

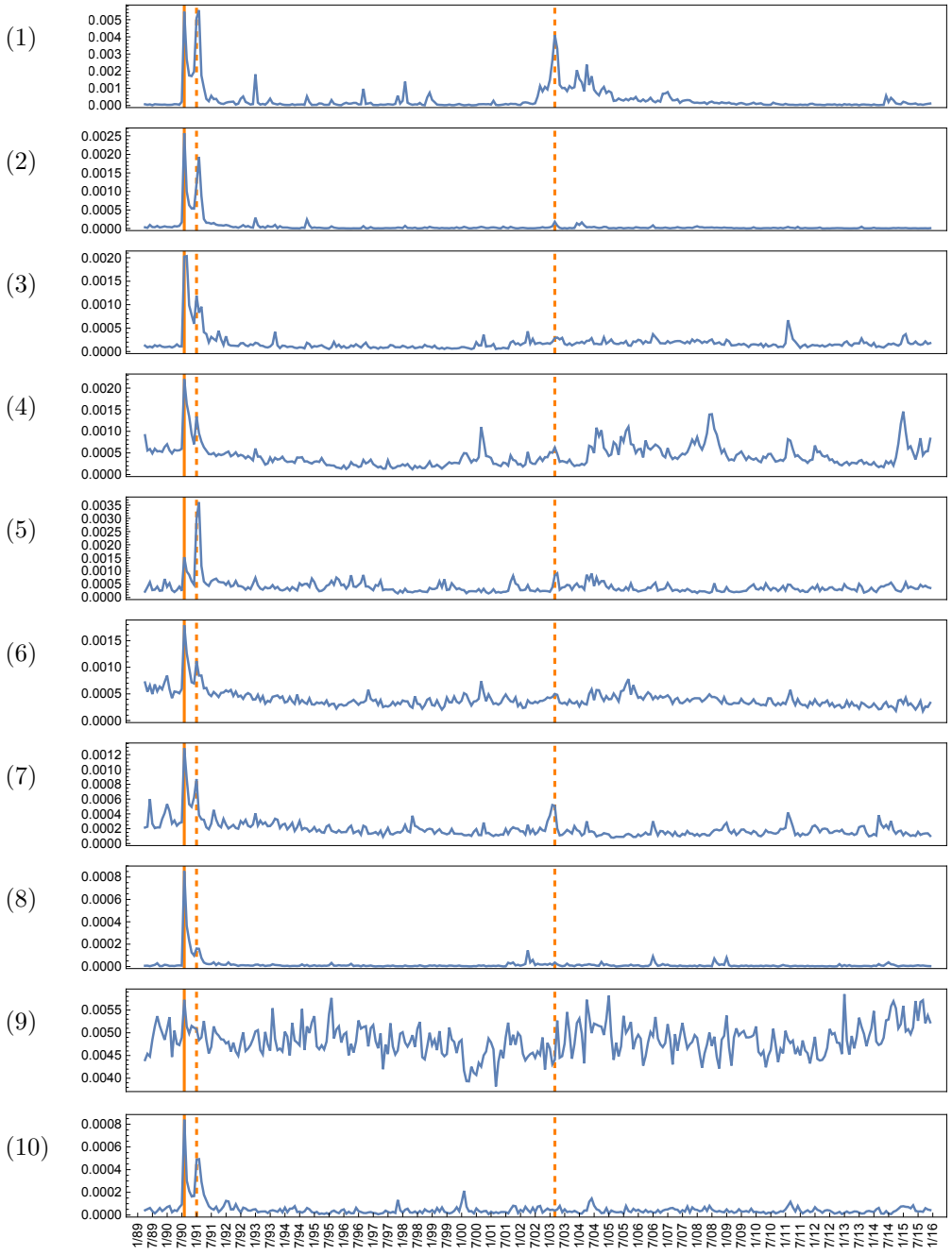


Figure 9: Shares of the top 10 words for August 1990: (1) Iraq; (2) Kuwait; (3) Middle East; (4) crude oil; (5) military; (6) petroleum; (7) state of affairs; (8) invasion; (9) people; (10) Saudi

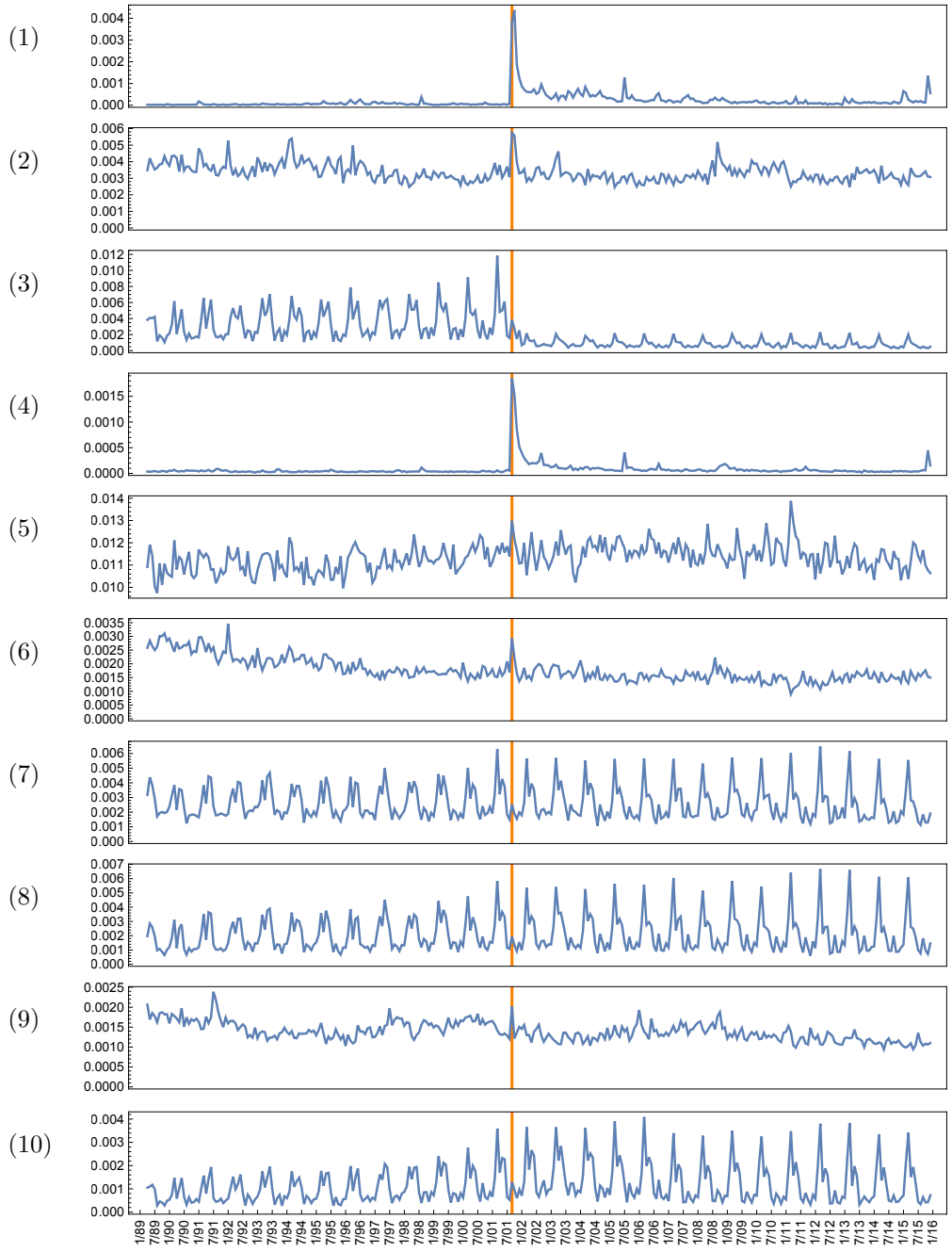


Figure 10: Shares of the top 10 words for September 2001: (1) terrorism; (2) America; (3) division head; (4) simultaneous; (5) day; (6) U.S.; (7) head (chief); (8) headquarters; (9) deal; (10) cum

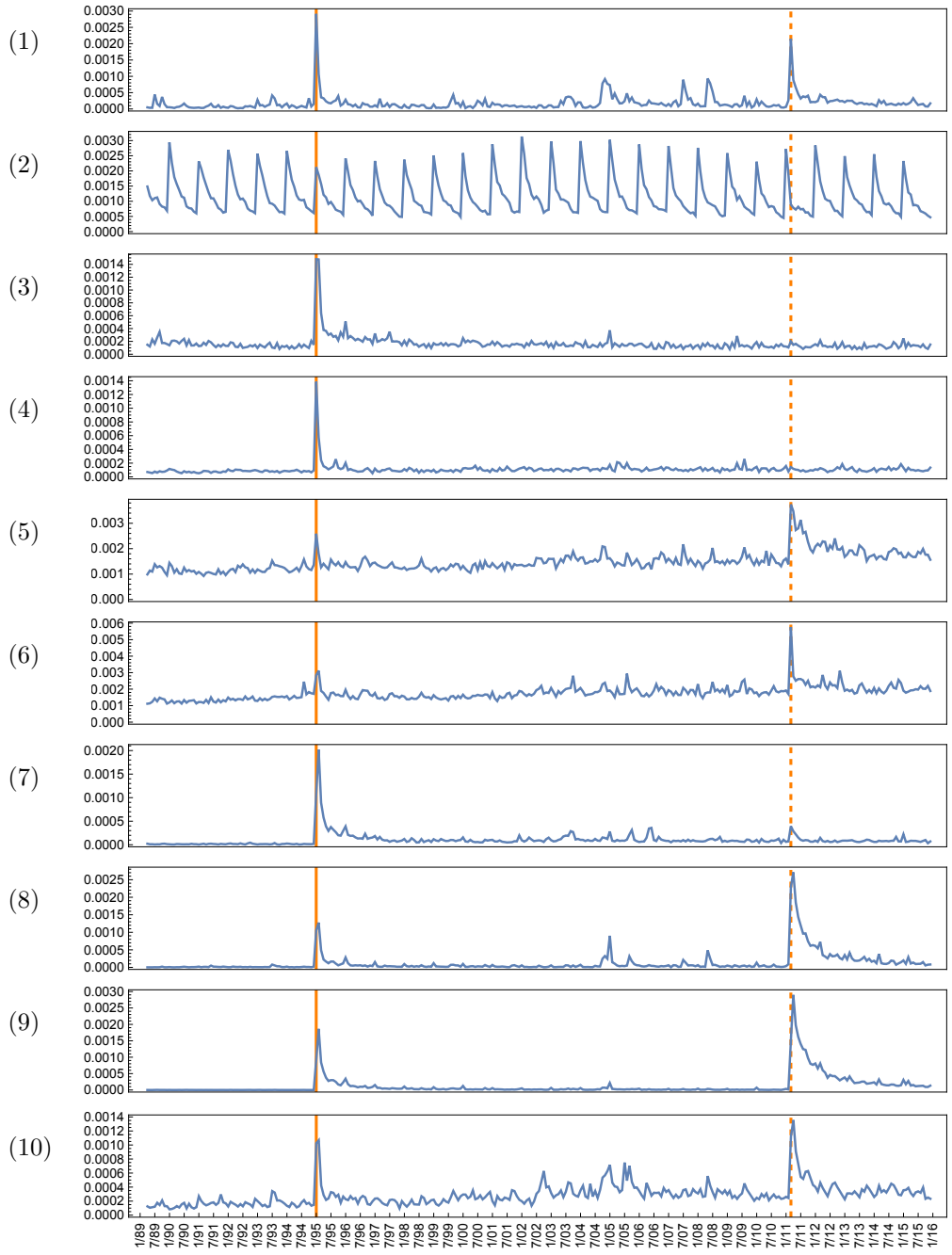


Figure 11: Shares of the top 10 words for January 1995: (1) earthquake; (2) last year; (3) Kobe; (4) Hyogo; (5) prefecture; (6) city; (7) Hanshin (Osaka-Kobe); (8) disaster; (9) great earthquake; (10) damage

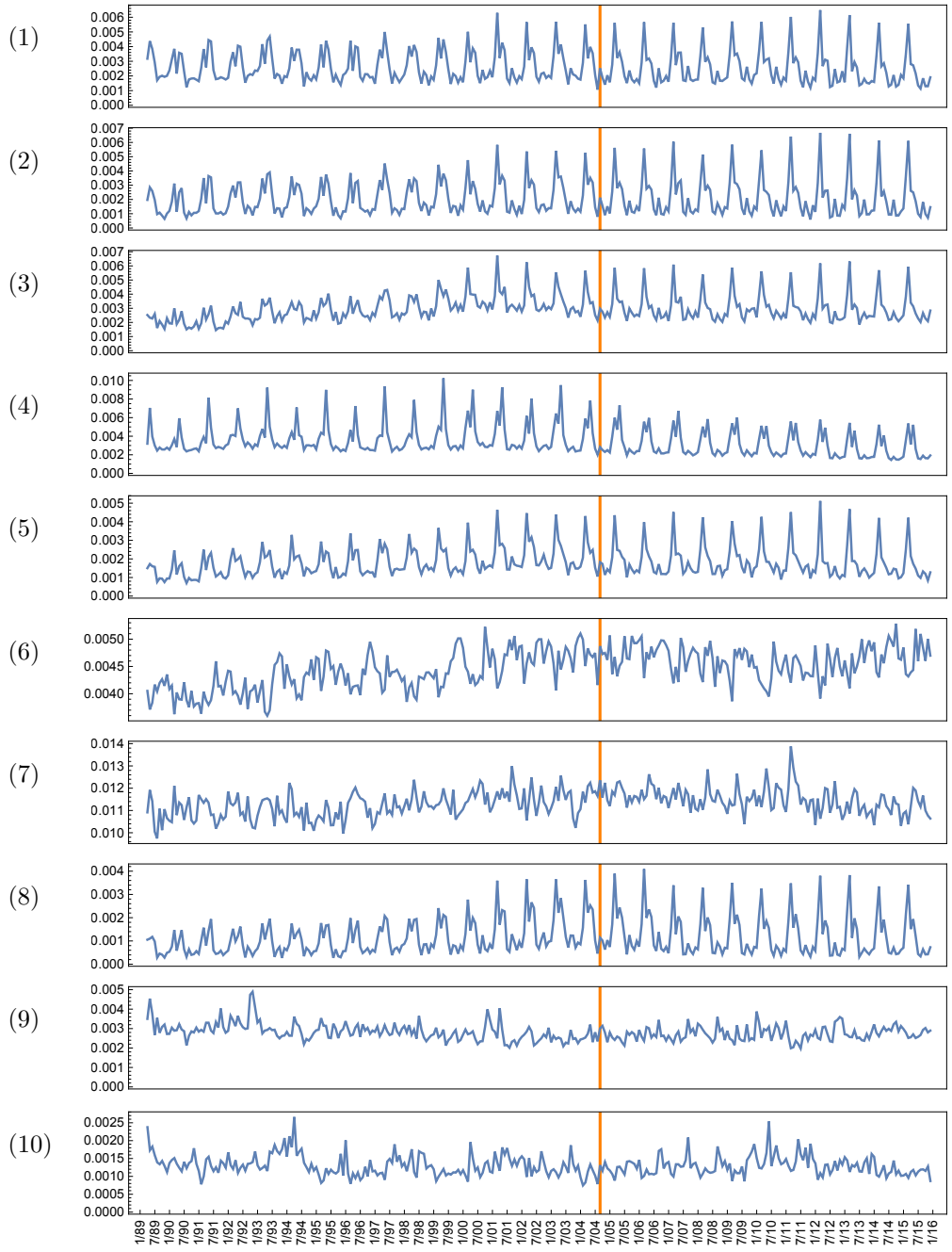


Figure 12: Shares of the top 10 words for September 2004: (1) head (chief); (2) headquarters; (3) enterprise; (4) same; (5) sales; (6) person; (7) day; (8) cum; (9) Mr./Ms.; (10) prime minister

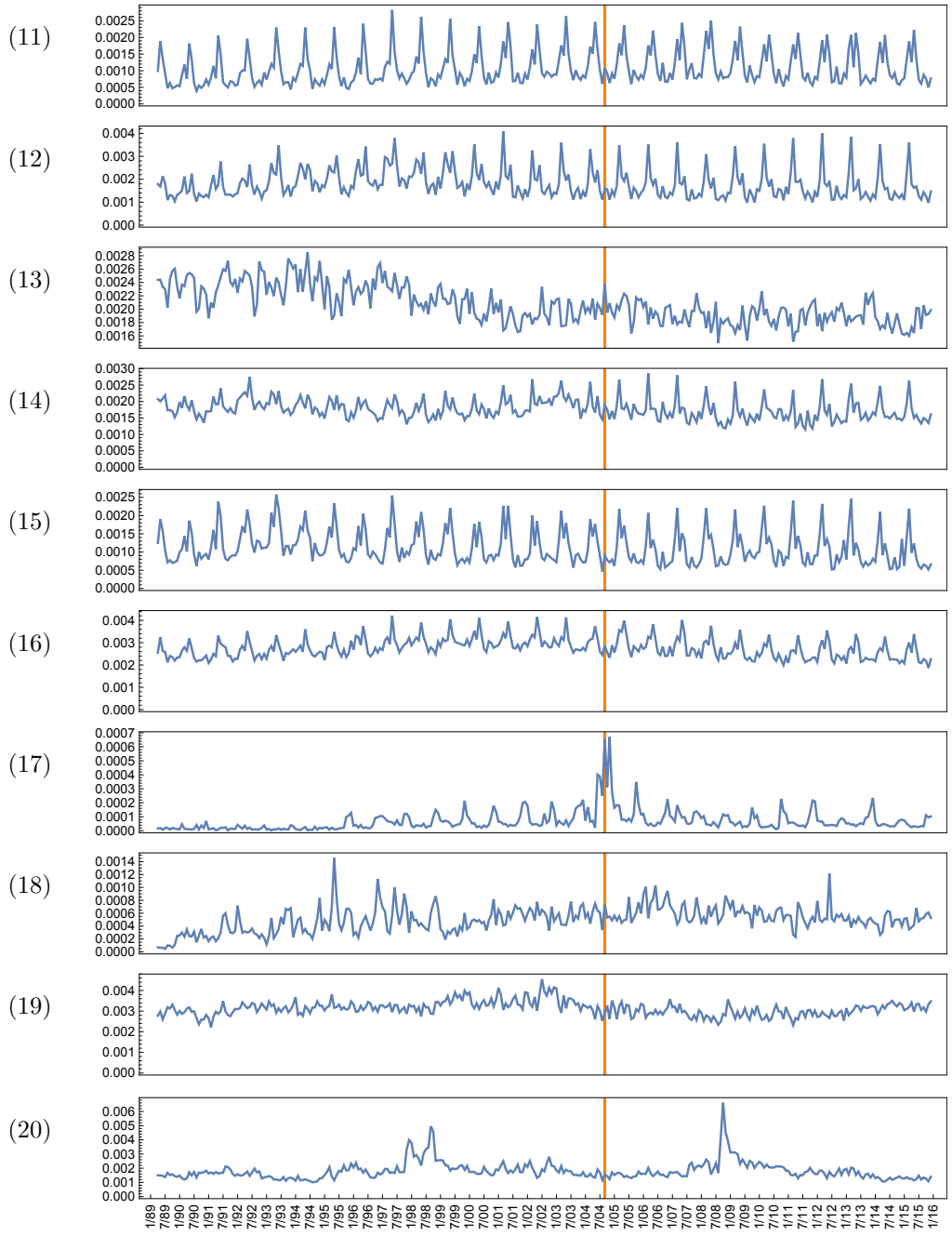


Figure 13: Shares of the top 10 words for September 2004: (11) personnel; (12) division; (13) meeting; (14) development; (15) sub; (16) company; (17) baseball team; (18) suspicion; (19) firm; (20) finance

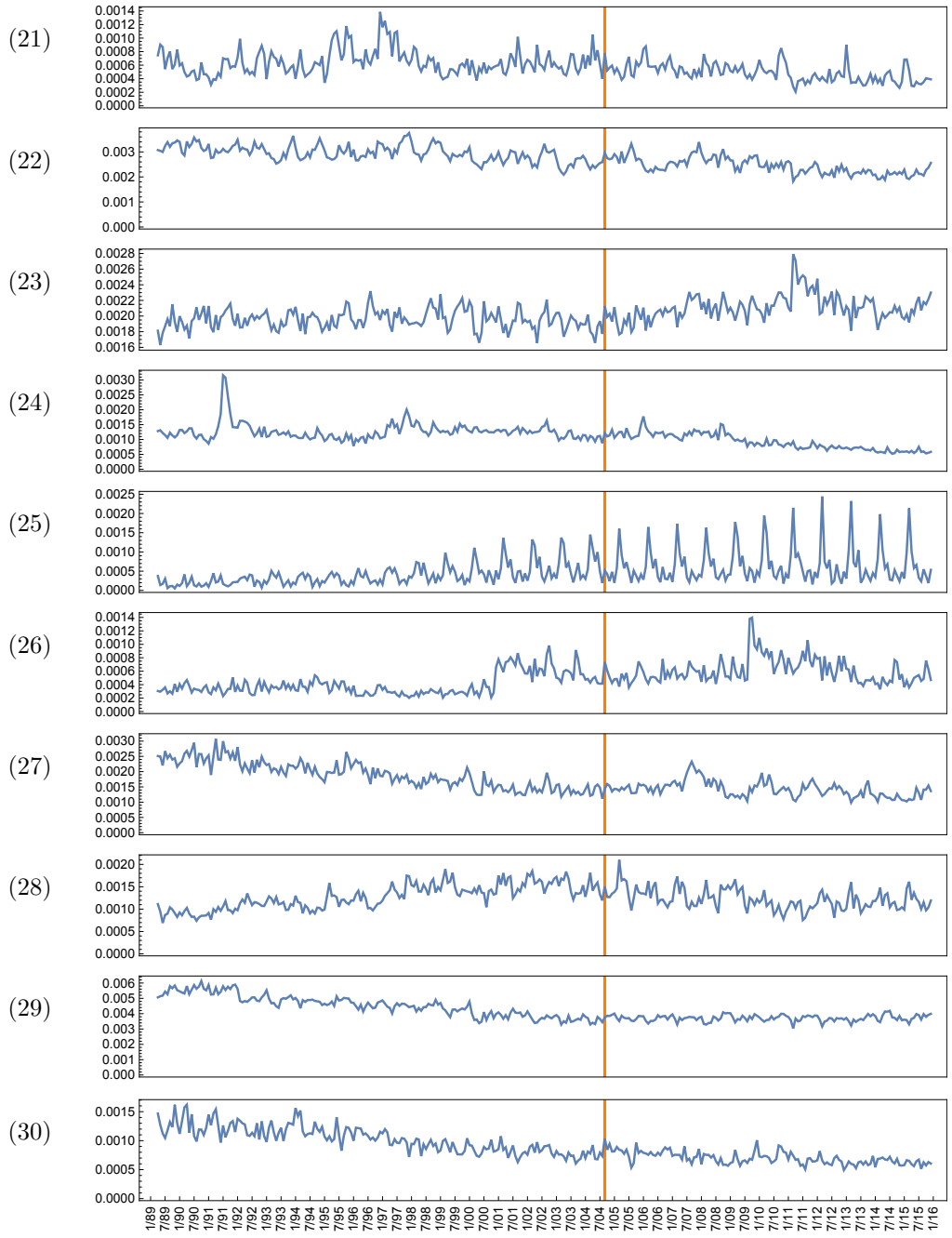


Figure 14: Shares of the top 10 words for September 2004: (21) incident; (22) -fication; (23) gender; (24) securities; (25) integration; (26) minister; (27) problem; (28) management; (29) type; (30) side

(31)

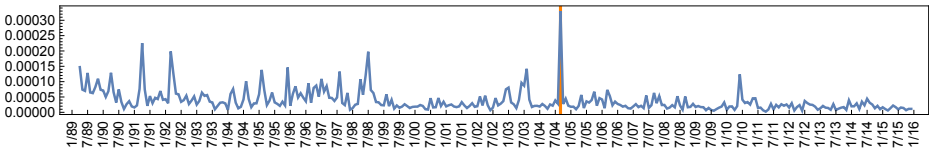


Figure 15: Share of "strike"

	Year.Month	Major event
1	2011.03.11	Great East Japan Earthquake
2	2000.09.15 –2000.10.01	Sydney Olympic Games
3	2008.09.15	Bankruptcy of Lehman Brothers
4	2013.01.16	Algeria hostage crisis
5	2009.08.30	Democratic Party becomes the ruling party
6	2015.01.20 –2015.01.30	Islamic state ransom incident
7	1990.08.02	Iraqi invasion of Kuwait
8	1995.01.17	Great Hanshin-Awaji Earthquake
9	2004.09.18 –2004.09.24	First strike by professional baseball players
10	2001.09.11	September 11 attacks

Table 4: Major events associated with the 10 highest peaks in Figure 2(e)

While a seasonal adjustment to the number of occurrences of each word would have been desirable throughout this paper, we applied a seasonal adjustment only to the unadjusted series for the distributional change (except for the names of the months, which we removed from the data).

While our main purpose was not to measure the social impacts of events, our results seem to measure such impacts fairly well. Since we used monthly data, however, the effect of an event on the word distribution depended upon the timing of the event within the month. We discussed this point in Section 5, in the context of summer Olympic events. An event occurring earlier in a given month can easily be expected to exert a larger influence on the text data of the month than an event occurring later in the month.

Although we mostly focused on major peaks in the fully adjusted series for the distributional change, they may only capture temporary changes in the word distribution. It remains unclear whether such changes indicate social change. Alternatively, we can also use the trend-cycle component in Figure 2(d) as a measure of social change. When we do so, Figure 2(d) suggests that overall societty changed more rapidly after July 2000 than before. This and the other issues mentioned above are left for future research.

References

Aggarwal, C.C., Subbian, K., 2012, Event detection in social streams, In: Ghosh, J., Liu, H., Davidson, I., Domeniconi, C., Kamath, C., eds., Proceedings of the 2012 SIAM International Conference on Data Mining, pp. 624–635.

AlSumait, L., Barbar'a, D., Domeniconi, C., 2008, On-line LDA: adaptive topic models for mining text streams with applications to topic detection and tracking, In: Giannotti, F., Gunopulos, D., Turini, F., Zaniolo, C., Ramakrishnan, N., Wu, X., eds., Eighth IEEE International Conference on Data Mining, pp. 3–12.

Andrade, M.A., Valencia, A., 1998, Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families, Bioinformatics 14, 600–607.

Antadze, N., Westley, F.R., 2012, Impact metrics for social innovation: barriers or bridges to radical change? Journal of Social Entrepreneurship 3, 133–150.

Atefeh, F., Khreich, W., 2015, A survey of techniques for event detection in Twitter, Computational Intelligence 31, 132–164.

Bee Dagum, E., Bianconcini, S, 2006, Seasonal Adjustment Methods and Real Time Trend-Cycle Estimation, Springer, Switzerland.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003, Latent Dirichlet allocation, Journal Machine Learning Research 3, 993–1022.

Cordeiro, M., Gama, J., 2016, Online social networks event detection: a survey, In: Michaelis, S., Piatkowski, N., Stolpe, M., eds., Solving Large Scale Learning Tasks: Challenges and Algorithms: Essays Dedicated to Katharina Morik on the Occasion of Her 60th Birthday, Springer International Publishing, Swtzerland, pp. 2-41.

Dzogang, F., Lansdall-Welfare, T., Team, F.N., Cristianini, N., 2016, Discovering periodic patterns in historical news, PloS one 11.11, e0165736.

Findley, D.F., Monsell, B.C., Bell, W.R., Otto, M.C., Chen, B.-C., 1998, New capabilities and methods of the X-12-ARIMA seasonal-adjustment program, Journal of Business & Economic Statistics 16, 127–152.

Garonna, P., Triacca, U., 1999, Social change: measurement and theory, International Statistical Review 67, 49–62.

Goodwin, R., 2009, Changing Raltions: Achieving Intimacy in a Time of Social Transition, Cambridge University Press, Cambridge UK.

Goswami, A., Kumar, A., 2016, A survey of event detection techniques in online social networks, Social Network Analysis and Mining 6, 107.

Griffiths, T.L., Steyvers, M., 2004, Finding scientific topics, Proceedings of the National Academy of Sciences 101, 5228–5235.

Hasan, M., Orgun, M.A., Schwitter, R., 2017, A survey on real-time event detection from the Twitter data stream, Journal of Information Science 2017, 1-21.

Livingstone, S., 2002, The changing social landscape, In: Lievouw, L.A., Livingstone, S., eds., Handbook of New Media: Social Shaping and Social Consequences of ICTs, Sage, London, pp. 17-21.

Phillips, F., 2011, The state of technological and social change: impressions, Technological Forecasting & Social Change 78, 1072–1078.

Sayyadi, H., Hurst, M., Maykov, A., 2009, Event detection and tracking in social streams, In: Proceedings of the International Conference on Weblogs and Social Media, pp. 311-314.

Swan, R., Allan, J., 2000, Automatic generation of overview timelines, In: Belkin, N.J., Leong, M.-K., Ingwersen, P., eds., Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 49–56.

U.S. Census Bureau, 2011, X-12-ARIMA Reference Manual, Version 0.3.

Wang, X., McCallum, A., 2006, Topics over Time: A non-Markov continuous-time model of topical trends, In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp. 424–433.

Yang, Y., Pierce, T., Carbonell, J., 1998, A study of retrospective and on-line event detection, In: Ungar, L., Craven, M., Gunopulos, D., eds., Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 28–36.

Zhao, W.X., Jiang, J., Weng, J., He, Jing, Lim, E.-P., Yan, H., Li, X., 2011, Comparing Twitter and traditional media using topic models, In: Clough, P., Foley, C. Gurrin, C., Jones, G.J.F., Kraaji, W., Lee, H., Murdoch., V., eds., Advances in Information Retrieval: 33rd European Conference on IR Reseaech, ECIR 2011, pp. 338–349.