한국통계학회논문집
2012, 19권, 6호, 771–780

DOI: http://dx.doi.org/10.5351/CKSS.2012.19.6.771

Statistical Properties of News Coverage Data

Eunju Lim[a], Kyu S. Hahn[1,b], Johan Lim[2,a], Myungsuk Kim[c], Jeongyeon Park[a], Jihee Yoon[a]

[a]Department of Statistics, Seoul National University
[b]Department of Communication, Seoul National University; [c]School of Business, Sogang University

Abstract

In the current analysis, we examine news coverage data widely used in media studies. News coverage data is usually time series data to capture the volume or the tone of the news media's coverage of a topic. We first describe the distributional properties of autoregressive conditionally heteroscadestic(ARCH) effects and compare two major American newspaper's coverage of U.S.-North Korea relations. Subsequently, we propose a change point detection model and apply it to the detection of major change points in the tone of American newspaper coverage of U.S.-North Korea relations.

Keywords: ARCH effects, mass media, dynamic programming, news coverage data, change point analysis.

1. 서론

미디어(media)는 "중간"이라는 그 어원에서 알수 있듯이 서로 일면식이 없는 사람들 사이에서 뉴스 또는 정보의 전달자 역할을 하는 개체를 일컫는다. 특히 여러 미디어들 중 대중매체(mass media) 또는 언론은 방송이나 신문매체와 같이 대량으로 뉴스를 생산 배포하는 산업을 일컫는 말로 사회가 복잡 다양해짐에 따라 대중의 의견인 여론(public opinion)의 전달자로서 그 중요성이 증대되어 왔다. 여론은 지난 수 십여 년간의 연구 결과가 보여주고 있듯이 그 주요 의제(agenda)가 매우 유동적이고 빠르게 등락을 거듭하고 이러한 등락은 현대 사회가 복잡해짐에 따라 언론매체에 의하여 주도되는 경향이 있다고 알려져 있다 (Lippmann, 1922). 즉, 언론의 보도가 어떠한 의제를 부각 시키느냐에 따라 해당 의제에 대하여 대중이 부여하는 중요도 또한 변하게 됨을 이야기 한다 (MacKuen과 Coombs, 1981; Erbring 등, 1980).

현대 사회에 있어서 언론의 의제 변화가 대중의 관심도에 영향을 끼친다는 것은 매우 중요한 시사점을 가진다. 이는 언론이 단순히 대중들 사이에서 의견을 전달하는 전달자의 입장을 넘어 대중과 의견을 주고받고 때로는 대중의 의제에 영향을 주는 독립된 개체임을 의미 한다. 이러한 독립된 개체로서의 역할이 가장 극명하게 나타난 예는 1992년 미국 대선이라고 할 것이다. 걸프전의 랠리 효과(rally effect)로 90%를 넘었던 부시대통령의 지지율이 종전 직후 언론보도의 초점이 '전쟁'에서 '경제'로 옮겨가면서 급격한 하락으로 이어졌고 결과적으로 역사상 처음으로 공화당 현직 대통령으로서 재선에 실패하게 된다. 이러한 예는 언론매체 시장이 덜 경쟁적인 한국에서도 종종 관측된다. 2008년의 경우

Support for this research was partially provided by the Korean Science Foundation (NRF-2010-330-B00028).
[1] Corresponding author: Kwanak-gu, Daehak-dong, Department of Communication, Seoul National University, Seoul 151-747, E-mail: kyuhahn@snu.ac.kr
[2] Corresponding author: Kwanak-gu, Daehak-dong, Department of Statistics, Seoul National University, Seoul 151-747, Korea. E-mail: johanlim@snu.ac.kr

MBC의 PD수첩이 미국산 소고기 수입 문제를 대중의 주요 의제로 이끌어 낸바있고, 또한 2010년의 도가니사건, 한·미·FTA 등도 대중매체가 여론에 영향을 준 대표적인 예로 생각할 수 있다.

이러한 중요성으로 인하여 언론 보도에 대한 많은 연구가 진행되어 왔다. 특히 언론 보도와 관련하여 보도(기사)의 어조(긍정/부정)와 보도의 양(기사에 사용된 단어의 수)을 분석의 중요한 정보로 사용한다. 이러한 요약 정보는 기사의 단어의 수를 세고 해당 기사의 어조에 따라 +/−의 부호를 부여함으로써 하나의 숫자로 표현하고 이러한 자료를 "언론보도자료"라 한다. 언론에 관한 기존의 연구들을 살펴보면 이러한 언론보도자료에 대한 단순 통계량에 의존하고 자료 자체의 분포적 성질을 포함한 통계적 고찰은 극히 제한적 이었다

본 연구에서는 언론 연구에 중요한 역할을 하는 언론보도자료에 대한 몇 가지 통계적 성질에 대하여 연구한다. 특히 본 연구는 기존 연구와는 다르게 언론보도자료가 시계열 자료임을 인식하고 시계열 모형하에서의 성질들에 대한 연구에 집중한다. 먼저 언론보도자료의 분포적 성질로 주변 분포(marginal distribution)의 두꺼운 꼬리(heavy tail) 현상 등을 포함한 arch(autoregressive conditional heteroscadestic)효과와 이에 대한 측도들을 살펴본다. 또한 이러한 분포적 성질들이 언론 매체들의 특정 주제(issue)에 대한 보도행태를 비교하는데 중요한 도구로 사용될 수 있음을 예제를 통하여 살펴본다. 다음으로 언론보도자료의 시계열적 특성에 보다 충실하여 변화점 모형을 적용하여 본다. 이러한 변화점 분석의 결과는 식별된 변화점에 대한 추가적인 내용 분석(context analysis)을 통하여 해당 언론사의 주요 사건(event)들에 대한 보도 성향을 이해하는데 도움을 준다.

본 논문은 다음과 같이 구성 되었다. 제 2절에서는 본 연구에서 사용하게 될 언론보도자료에 대하여 설명한다. 본 논문의 자료는 연구자들이 미국에서 공부하는 기간 동안 북한 핵실험으로 인하여 대북 문제가 미국 언론의 많은 조명을 받고 있었고 이를 계기로 자료를 수집하게 되었다. 제 3절에서는 언론보도자료의 분포적 성질로서 arch효과에 대하여 공부하고 이를 이용하여 언론사들의 사건 또는 의제에 대한 보도 태도를 비교하여 본다. 제 4절에서는 언론보도자료의 변화점 모형과 이를 추정하기 위한 동적프로그밍(dynamic programming)을 소개한다. 또한 이를 실제 자료에 적용하여 본다. 제 5절에서는 본 연구를 간단히 요약한다.

2. 자료

본 논문에서는 1992년 7월 1일부터 2004년 1월 14일 사이에 미국의 뉴욕 타임즈와 워싱턴 포스트지에 게재된 북한 관련 기사들의 분석을 통하여 언론보도자료의 통계적 특성을 살펴본다. 본 연구에서 신문을 선택한 이유는 신문의 구독률이 TV시청률에 비하여 상대적으로 낮으나 정치나 외교문제 등과 같은 경성 뉴스에 있어서는 여론 형성에 큰 영향력이 있음이 알려져 있다 (Curran 등, 2008). 특히 본 연구에서 살펴 보게 될 두 신문사는 미국에서 다른 언론 매체들의 의제를 설정하는데도 막대한 영향력을 미치는 매체임이 많은 실증연구를 통하여 알려져 있다 (Gans, 1980; Iyengar과 McGrady, 2007).

분석에 포함된 기사들은 렉서스-넥서스(Lexis-Nexis) 데이터 베이스에서 키워드 검색을 통해 수집되었다. 북한과 관련된 기사들을 추려 내기 위해 (1) 북한의 공식 국명("Democratic People's Republic of Korea")과 그 약자("DPRK"), (2) 북한의 수도명("Pyongyang"), 그리고 (3) "북한사람"에 해당하는 "North Korean(s)"을 키워드로 사용하였다. 이 세 가지 키워드가 헤드라인에 등장하는 기사가 1차 분석 대상으로 분류되었고, 이렇게 추려진 검색 결과를 기사의 내용을 확인하는 과정을 거쳐 실제 북한과 관련이 없는 기사들은 분석 대상에서 제외하였다. 실제로 1차 검색에서 추출되어 나온 상당수의 기사들이 본 분석과 관련이 없는 스포츠 등에 대한 기사였고 이들은 최종 분석 대상에서 제외되었다. 이런 검색절차를 거쳐 최종적으로 3,373개(뉴욕 타임즈: 1,942개, 워싱턴 포스트: 1,431개)의 북한 관련 헤드라인이 분석 대상으로 분류되었다.

총 3,373개의 기사들을 대상으로 여덟 명의 코더(coder)들이 다양한 정보를 자료화 했다. 여덟 명의 코더 전원은 미국 스탠포드 대학(Stanford University)의 학부 및 대학원생들로 모두 영어를 모국어로 사용하고 정치외교학 또는 국제정책을 전공하는 학생들로 구성되었다. 자료의 신뢰도를 높이기 위해 모든 코더들간의 데이터에 상당한 정도의 일치도가 성취될 때까지 충분한 훈련을 실시하였다. 단순한 몇 가지 변인들(기사 게재 일자, 섹션, 페이지 등) 외에 각 기사별로 몇 가지 코더의 주관적인 판단을 필요로 하는 변인들이 지수화 되었다. 우선, 각 기사를 (1) “북한 내부적 상황”, (2) “북한과 미국의 관계”, (3) “남북관계”, 그리고 (4) “북한과 기타국가와의 관계”에 대한 기사로 분류하였다. 이렇게 수집된 기사들을 토대로 북한관련 보도의 월 별 어조에 대한 측도를 만들기 위해 일단 뉴욕 타임즈와 워싱턴 포스트에 게재된 북한 관련 보도의 기사별 어조를 측정했다. 이를 위해 일단 각 기사에서 문단 별로 해당 문단이 북한에 대하여 “긍정적(또는 중립적)” 또는 “부정적” 내용인지의 여부를 입력한 후, 각 기사별로 “긍정적”, “부정적” 문단의 비율을 구했다. 다음으로 해당 기사의 “긍정적” 문단과 “부정적” 문단의 비율들에 그 기사의 총 단어숫자를 곱한 후, “긍정적” 단어 숫자에서 “부정적” 단어 숫자를 빼는 방식으로 기사별 어조를 계산하였다. 최종적으로 해당 월에 나온 모든 기사의 어조를 더하여 월별 어조를 계산하였다.

3. 분포적 성질: Arch효과

3.1. 정의

언론보도자료는 여러가지 측면에서 주식의 수익률에 관한 자료와 유사한 분포적 특성을 지닌다. 이러한 특성들 중 하나로 아래에서 간략하게 살펴볼 arch효과를 생각 할 수 있다.

일반적으로 arch효과를 하나의 문장으로 정의하기는 어려우나 굳이 정의하면 이름에서 알 수 있듯이 관측값의 오차항이 auto-regressive conditionally heteroscadestic(arch)한 특성을 지닌 시계열에서 나타나는 현상들을 통틀어 지칭하는 용어라 할 수 있다. Arch 오차항의 몇 가지 대표적인 현상들로는 우선 자료의 주변분포(marginal distribution), 즉 시간 축을 무시한 원 자료들의 분포가 0 근처에 집중되어 있고 두터운 꼬리를 가지는 현상이다. 이러한 현상은 첨도(kurtosis)를 통하여 확인 가능하고 arch오차항을 지닌 경우 첨도는 정규 분포의 첨도인 3보다 큰 값을 지니게 된다. 이러한 현상을 특별히 leptokurtotic 현상이라 부른다. Arch 오차항을 지닌 경우 나타나는 다른 현상으로는 변동성의 군집화(volatility clustering) 현상을 이야기 할 수 있다. 이는 자료의 분산이 시간에 따른 자기 상관을 지니는 현상을 일컫는 것으로 어느 시점에서 자료의 분산이 크면 뒤따른 시점들에서도 분산이 크게 나타나는 경향을 이야기한다.

이제 언론보도자료가 위에서 언급된 arch효과들을 지니는지 생각하여 볼 필요가 있다. 일반적으로 어떤 관심 의제와 관련하여 새로운 사건이 발생하지 않으면 해당 기간 동안 언론매체는 통상적인 최소의 양의 보도를 하게 되는 반면 해당 의제와 관련하여 중요한 사건이 발생하면 사건에 대한 배경 분석 등을 포함한 상당한 양의 보도가 나오게 된다. 이러한 현상은 arch효과의 존재가 처음으로 제기된 주식시장에서도 자주 나타나는 현상으로 특별한 호재나 악재가 없는 경우 주식의 수익률은 0 근처에서 변동하지만 주식시장에 최근에 문제가 된 유럽의 금융 위기와 같은 불확실성이 개입하게 되면 사소한 사건들에도 수익률이 크게 요동치는 현상이 발생한다. 이러한 현상의 결과로 주식 수익률의 주변 분포는 0 근처에서 높은 밀도를 보이고 두꺼운 꼬리를 지닌 분포적 특성을 지니게 된다. 마찬가지 이유로 언론보도자료 또한 주변 분포가 0 근처에서 높은 빈도를 갖게 되고 두꺼운 꼬리를 가지게 된다.

언론보도자료에 있어 변동성의 군집화 현상은 대중매체의 경우 특정한 사건에 대한 보도가 지면(section)을 옮겨가며 단 기간에 사라지지 않는 현상을 통하여 설명할 수 있다. 한 예로 지난해에 국내에서 논쟁이 되었던 장애인 교육 시설에 대한 의제(도가니 사건)의 경우 초기에는 연예면의 한 영화

에 대한 소개로 부터 논의가 시작 되었으나, 다음으로는 교육에 관련된 사회면으로, 최종적으로는 관련 법안에 대한 의제를 다루는 정치면으로 이동하며 오랜 기간 언론의 관심을 받게 되었다. 물론 언론보도자료도 시간에 따라 관측되는 시계열 자료이고 의제 자체의 시간적 상관성도 변동성의 군집화에 영향을 준다고 생각 할 수 있다.

3.2. 측도들

다음으로 앞에서 설명한 언론보도자료의 통계적 특성의 이해를 돕기 위한 세 가지 정량적 측도들에 대하여 살펴본다. 본 절에서 소개하는 측도들은 통계학에서 널리 사용되는 측도이고 본 연구자들이 새로이 제안하는 것은 아니다.

Arch효과의 통계적 측도로서 먼저 왜도(skewness)를 생각 할 수 있다. 이는 관심 의제에 대하여 보도의 성향(긍정/부정)을 나타내는 측도로 만일 어느 언론매체가 A라는 주제에 대하여 음의 왜도 값을 가졌다면 이는 덩치가 큰 부정적인 보도가 해당 매체에 자주 나타났음을 나타내고 결과적으로는 매체가 해당 주제에 대하여 부정적인 보도 행태를 지니고 있음을 보여준다고 할 수 있다.

두 번째 측도로서 첨도(kurtosis)를 생각 할 수 있다. 첨도는 원 자료의 꼬리의 두꺼운 정도를 나타내는 지수로 첨도의 값이 클수록 언론보도자료의 주변 분포가 중앙 부분은 뾰족하고 꼬리가 두터운 분포를 갖게 됨이 알려져 있다. 언론매체의 보도 자료가 특정 의제에 관하여 큰 첨도 값을 가졌다 함은 해당 매체가 관련 의제에 대하여 높은 중요도를 두고(또는 민감하게 반응하여) 관련된 사건이 발생 시 많은 양의 보도를 내보내고 있음을 이야기 한다.

마지막으로 변동성 군집화에 관련된 지수를 arch모형에서는 분산에 관련한 자기 회귀 모형을 이용하여 정의한다. 제안된 측도는 분산 성분의 자기 상관이 얼마나 빠른 속도로 0으로 가는지를 나타낸다. 다시 이야기하면 큰 변동성이 나타났을 때 이 변동성이 얼마나 오랜 기간 지속되는지를 나타내는 측도이다. 언론보도자료가 arch모형으로 부터 나왔다는 가정 하에서 변동성의 자기회귀식은

$$\sigma_t^2 = \alpha_0 + \alpha_1 y_{t-1}^2 + \alpha_2 y_{t-2}^2 + \cdots + \alpha_p y_{t-p}^2$$

로 표현되고 제안된 측도는 위 식의 귀무식

$$\phi(z) = 1 - \alpha_1 z - \alpha_2 z^2 - \cdots - \alpha_p z^p = 0$$

의 고유해(eigenvalue)

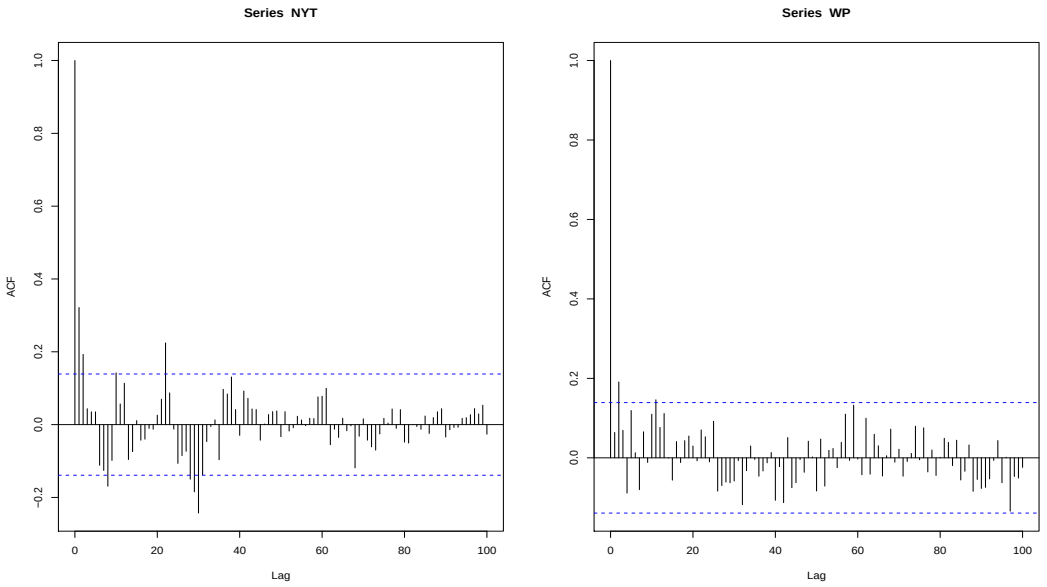
$$\lambda_1, \lambda_2, \ldots, \lambda_p$$

들의 절대값 중 가장 큰 값으로 정의한다. 여기서 분산 성분에 대한 자기상관계수가 상수 $c_1, c_2, \ldots, c_p$에 대하여

$$\gamma(k) \equiv \mathrm{cov}(y_t, y_{t+k}) = c_1 \lambda_1^k + c_2 \lambda_2^k + \cdots + c_k \lambda_p^k, \quad k = 0, 1, 2, \ldots$$

로 표현됨을 생각하면 자기상관계수가 0으로의 수렴하는 속도는 우리가 제안한 측도인 고유해의 절대치의 최대값에 의하여 결정됨을 알 수 있고 따라서 변동성의 군집화에 대한 적절한 측도임을 확인 할 수 있다.

위의 측도들 외에도 arch효과에 대한 여러 다양한 측도들이 존재하고 이는 (Tsay, 2009) 등과 같은 시계열 관련 서적을 살펴보면 쉽게 확인할 수 있다. 본 논문에서는 언론보도자료에 대하여 이러한 측도들이 사용될 수 있다는 사례를 보이고자 응용 연구자들이 쉽게 사용할 수 있는 세 가지 측도에 의거하여 논의를 진행한다.



(a) New York Times (b) Washington Post

Figure 1: *Auto-correlation function of New York Times and Washington Post data. The dotted lines are 90% confidence interval for white noise.*

3.3. 사례응용

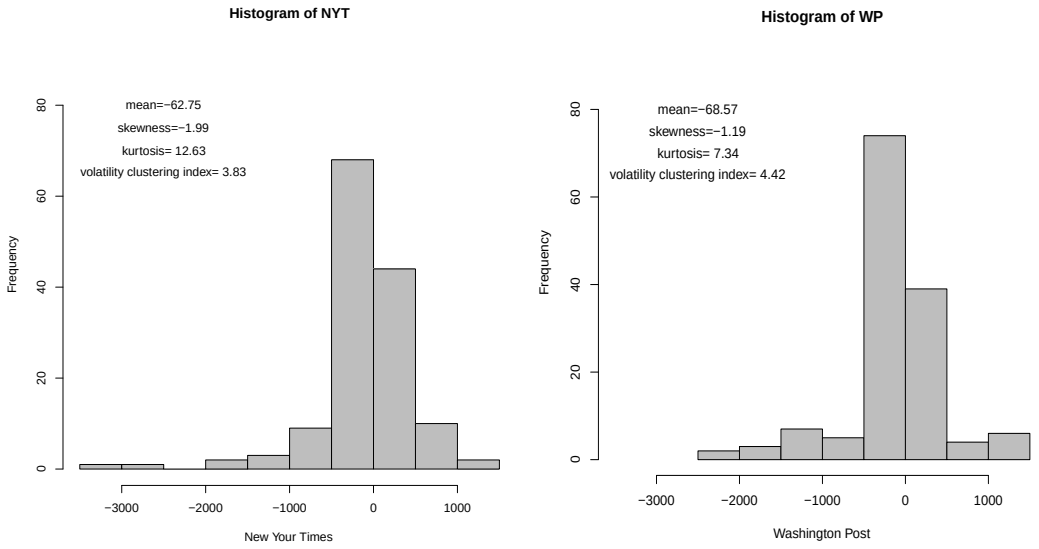
이 절에서는 앞서 소개된 세가지 측도들을 본 연구진이 수집한 실제 자료에 적용하여 보고자 한다.

본 분석을 시작하기 이전에 언론보도자료의 arch효과에 대한 탐색적 분석으로서 두 매체의 자료들의 자기상관계수를 계산하여 본다. Figure 1은 두 자료의 자기상관계수를 그림으로 표현한 것으로 두 자료 모두 긴 시간의 Lag에 대하여도 자기상관계수가 쉽게 없어지지 않음을 알 수 있다. 또한 두 자료에 대하여 ARCH모형을 적합시켜보면 워싱턴-포스트의 경우는 ARCH(1)모형이 뉴욕 타임즈 자료의 경우 ARCH(2)모형이 BIC(Bayeisan Information Criteria)기준을 잘 적합되고 모형의 계수들이 유의하게 나옴을 확인할 수 있다.

이제 구체적으로는 위 세 가지 측도를 활용하여 미국 유력 일간지인 워싱턴 포스트와 뉴욕 타임즈 간의 북한관련 보도에 대한 매체적 차이를 비교해 본다.

우선 평균을 비교하여 보면 뉴욕 타임즈(평균 = −62.75) 보다 워싱턴 포스트(평균 = −68.57)가 북미 관계를 상대적으로 부정적으로 보도했다는 것을 알 수 있으나 이 차이는 우리 측도상 불과 다섯 단어 정도의 차이여서 큰 의미를 부여하기는 어려웠다. 따라서 북한 관련 보도에 있어 두 언론사간의 어조의 차이는 극히 미미하고 이는 일반적으로 미국 언론에서 북한 관련 보도가 매우 사건 중심적이라는 점에서 어느 정도 예견된 일치성이라 할 수 있을 것이다. 즉, 보도 가치가 있는 북한 관련 사안이 한정적이기 때문에 두 언론사간의 전반적인 어조의 차이가 상당히 작으리라는 예측이 가능하다.

평균과는 다르게 왜도(skewness)와 첨도(kurtosis)에서의 차이를 고려할 경우 두 언론사간의 북한 관련 보도의 특성 차이가 좀 더 확연하게 드러났다. 구체적으로 살펴보면 왜도의 경우 뉴욕 타임즈(왜도 = −1.99)가 워싱턴 포스트(왜도 = −1.19)보다 두 배 가까운 음의 값을 가지고 이는 뉴욕 타임즈가 북한 관련 사안에 있어 많은 지면을 할애한 부정적인 보도를 워싱턴 포스트지에 비하여 자주 보도하고 있



(a) New York Times (b) Washington Post

Figure 2: *Arch effects*

음을 의미한다. 첨도의 경우 뉴욕 타임즈(첨도 = 12.63)가 워싱턴 포스트(첨도 = 7.34)보다 매우 큰 값을 가진 것을 확인할 수 있고 이 또한 뉴욕 타임즈의 주변분포의 꼬리 쪽 분포가 워싱턴 포스트보다 더 두텁다는 것을 의미한다. 즉, 북미 관계에 대한 보도에 있어서 뉴욕 타임즈의 경우 상대적으로 극단적인 어조를 띄는 기사의 숫자가 많았다는 것을 시사한다.

마지막으로 두 신문의 변동성 군집화 현상은 워싱턴 포스트가 (4.42) 뉴욕 타임즈보다 큰 것으로 (3.83) 나타났다. 이것은 두 신문 모두 북한 관련 사안에 대한 관심이 사건 중심적이고 일시적이나, 이러한 현상이 뉴욕 타임즈의 경우 특히 심했다는 것을 의미한다. 즉, 사안의 신선함이 하락할 때 뉴욕 타임즈의 경우 워싱턴 포스트보다 훨씬 더 빠른 속도로 보도의 양이 감소한다는 것을 의미한다. 이는 워싱턴 포스트의 경우 그 지리적 위치상 정치나 외교와 관계된 사안을 좀 더 비중 있게 다루는 반면 뉴욕 타임즈지는 보다 다양한 사안을 균형 있게 다루는 특성이 있음이 알려져 있다. 이러한 이유로 뉴욕 타임즈의 경우 기사의 관심 사안이 빠르게 이동하는 경향이 있고 이러한 경향이 변동성 군집화 지수에 반영되었다.

결론적으로 뉴욕 타임즈의 경우 북한 관련 사안이 발생했을 때 상대적으로 극단적인 어조의 기사를 많이 내는 반면 기사의 관심이 빠르게 다른 사안으로 옮겨가는 것을 확인할 수 있었다. 반면, 워싱턴 포스트의 경우 북한 관련 사안이 발생했을 때 뉴욕 타임즈보다 상대적으로 균형 잡힌 어조의 기사가 주류를 이루고 좀 더 긴 기간 동안 사안에 대한 관심을 유지하면서 심층적인 보도를 했다는 것을 알 수 있다.

4. 시간적 특성: 변화점 모형

앞 절에서는 arch효과에 대한 측도들이 특정 주제(issue)에 대한 여러 언론매체들의 보도 행태를 측정하고 이를 비교 분석하는데 유용하게 활용될 수 있음을 살펴보았다. 본 절에서는 언론보도자료에 변화점 모형을 적합시키고 이를 통하여 특정 매체가 사건(event)들에 대하여 가지는 보도의 특성을 살펴본다. 본 절의 분석 절차를 구체적으로 살펴보면 관심 주제(issue)에 대하여 언론보도자료에 변화점 모

형을 적용하고 검색된 변화점에서 발생한 사건들을 살펴봄으로써 매체의 사건별 보도 특성을 알아보고자 한다.

4.1. 변화점 모형

변화점의 탐색을 위하여 언론보도자료에 대하여 다음의 가정을 한다. 지정된 매체의 t번째 월의 자료를 y_t라 하면 y_t는 평균이 μ_t이고 분산이 σ_t^2인 분포를 따르고 μ_t와 σ_t^2는 시간에 따라 분할적-상수(piecewise constant)함수임을 가정한다.

변화점의 개수가 m개로 정하여져 있을 때 μ_t와 σ_t^2의 변화점을 추정하는 문제를 표현하여 보자. 먼저 $\mu[1:n] = (\mu_1, \ldots, \mu_n)$와 $\sigma[1:n] = (\sigma_1, \ldots, \sigma_n) \in \mathbf{R}^n$라 하고 $\mathcal{S}_k[1:n]$를 k개의 변화점을 지닌 $(\mu[1:n], \sigma[1:n])$들의 집합으로 정의한다. 이들 정의 하에 변화점의 탐색은 최적화 문제

$$\begin{aligned} &\text{minimize } \frac{1}{2}\sum_{t=1}^{n}\left\{\frac{(y_t-\mu_t)^2}{\sigma_t^2} + \log\sigma_t\right\} \\ &\text{subject to } (\mu[1:n], \sigma[1:n]) \in \mathcal{S}_m[1:n] \end{aligned} \tag{4.1}$$

의 해를 구하는 문제로 표현된다. 위의 변화점 모형은 변화점에 대하여 l_0-벌점(bounded complexity)을 가정한 모형으로 Friedrich 등 (2008)과 Boysen 등 (2009)에 의하여 연구되었다.

4.2. 알고리즘과 모형선택

이 절에서는 최적화 문제 (4.1)를 해결하기 위하여 동적 프로그램(dynamic programming)과 모형 선택 기준에 대하여 공부한다. 동적 프로그램에 대한 자세한 소개는 Bellman (1975)를 살펴보기 바란다.

문제 (4.1)를 풀기 위한 방법으로 동적프로그램을 생각 할 수 있다. 먼저 $\mathbf{J}_k^*(i,j)$를 $k-1$개의 변화점을 가지고 있는 자료점들 $y_i, \ldots, y_j$에 대한 문제 (4.1), 즉

$$\begin{aligned} &\text{minimize } \frac{1}{2}\sum_{t=i}^{j}\left\{\frac{(y_t-\mu_t)^2}{\sigma_t^2} + \log\sigma_t\right\} \\ &\text{subject to } (\mu[i:j], \sigma[i:j]) \in \mathcal{S}_k[i:j] \end{aligned} \tag{4.2}$$

의 최적값으로 정의한다. 이렇게 정의된 $\mathbf{J}_k^*(i,j)$들은 재귀적 관계식

$$\mathbf{J}_k^*(1,n) = \min_{1\le i\le n}\left\{\mathbf{J}_1^*(1,i) + \mathbf{J}_{k-2}^*(i+1,n)\right\}$$

을 만족시킨다. 여기서 $k = 1, \ldots, m$의 임의의 수가 될 수 있고 최적 값 $\mathbf{J}_1^*(i,j)$은 모든 i와 j에 대하여 표본 평균과 분산을 이용하여 쉽게 계산 할 수 있다.

이제 문제 (4.1)을 푸는 반복적 알고리즘을 설명하면 다음과 같다. 먼저 k_1을 $\mathbf{J}_m^*(1,n)$을 최소화 하는 최적 분할점이라고 가정하면 $\mathbf{J}_{m-1}^*(k_1,n)$을 최소화 하는 변화점 k_2를 구할 수 있고, 비슷한 방법으로 계속해서 k_{m-1}까지 구할 수 있다. 이렇게 계산된 $(k_1, k_2, \ldots, k_{m-1})$이 식 (4.1)의 해를 정의하는 변화점이 된다.

다음으로 변화점의 개수 m을 선택하는 과정에 대해서 설명하고자 한다. 문제 (4.1)과 관련하여 m을 선택하는 것은 유한 개의 포함 모형(nested model)들 가운데 최적의 모형을 찾는 문제이고 이를 위한 통상적인 방법은 정보량 기준(information criteria)을 이용하는 방법이다. 대표적인 정보량 기준들로는 Akaike (1974)의 Akaike Information Criteria, Schwarz (1978)의 Bayesian Information Criteria,

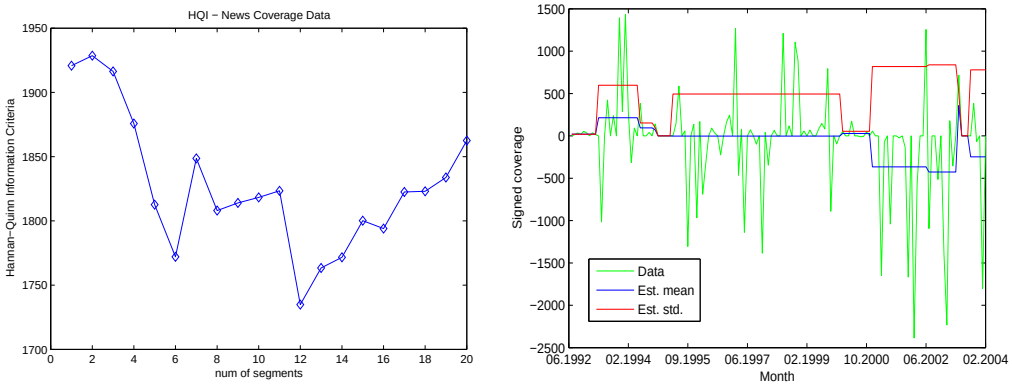


Figure 3: *Analysis of the Washington Post data.*

그리고 Hannan과 Quinn (1979)의 Hannan-Quinn Information Criteria(HQI)를 생각 할 수 있다. 본 절에서는 HQI를 사용하기로 한다. 정보량 기준 HQI는, $k = 1, \ldots, M$에 대해서,

$$\mathbf{HQI}(k) = -2 \max_{\theta \in S_k,} \log f(y_t; \theta_t) + 2k \log\log(n)$$

로 정의된다. 여기서 M은 우리가 얻을 수 있는 모형의 복잡도(model complexity)의 최대값을 의미하고 변화점 모형의 경우 최대 허용 변화점의 수이다. HQI는 복잡한 모형에 대하여 AIC보다는 강한 벌점을 BIC보다는 약한 벌점을 제공함으로써 선택된 모형은 복잡도(또는 모수의 수)에 있어서 두 정보량 기준들에 의하여 선택된 모형의 중간 정도이다.

4.3. 사례응용: 워싱턴-포스트

본 절에서는 앞에서 제안된 변화점 모형을 이용하여 워싱턴-포스트의 언론보도자료를 분석하였다. 변화점의 수 $m = 1, 2, \ldots, M = 20$에 대하여 식 (4.1)의 해를 구하고 이 결과를 이용하여 HQI를 계산하였다. Figure 3의 왼쪽 패널은 각 변화점의 수에 대한 HQI값을 그린 것인다.

앞 절에서 제안한 방법에 근거하면 평균이나 분산에서 총 12개의 변화점이 검색되었고 이들 중 평균에 큰 변화를 보인 1993년 3월, 1994년 5월, 2000년 10월과 2003년 2월을 주요 변화점으로 간주하고 변화점 주변의 사건들에 대한 추가적인 조사를 시행하였다.

우리 분석에 따르면 첫 변화점은 1993년 3월로 볼 수 있었다. 북한은 이 보다 조금 뒤인 1993년 5월에 미사일 발사에 성공함으로써 북미관계가 최악으로 치닫는 시발점이 된 바 있는데, 사실 이것은 치밀한 계획 하에 진행된 수 개월간에 걸친 이란과의 협상 및 공조의 결과물이었다. 이란과 북한은 일본 본토까지 타격이 가능한 장거리 미사일을 개발하기 위한 협력을 꾸준히 진행해 왔었으며 이러한 협력의 결과물로서 1993년 3월 처음으로 북한이 이란으로부터 미사일 발사대 제작에 필요한 원자재를 수입하기에 이르렀다. 그리고 북한은 곧 핵확산방지조약(NPT)에서 탈퇴를 선언함으로써 소위 제1차 북핵위기가 촉발된 바 있다. 미국 언론의 시각에서 이 때부터 이미 북미관계의 새로운 변화를 보여주는 중대한 사건으로 보여 졌으리라 추측해 볼 수 있다.

두 번째 변화점은 1994년 5월경으로 보여 지는데 이 시기는 북한의 미사일 발사실험으로 촉발된 일촉즉발의 위기상황에서 북한의 태도에 일련의 변화의 조짐이 나타나기 시작한 시점이라 할 수 있다. 1994년 5월, 북한당국은 처음으로 UN 핵 사찰단을 만나 핵무기의 원료가 될 수 있는 폐 핵연료봉의 관리와 감시를 위한 시스템을 구축하는 것에 대한 논의를 시작하는데 합의했다. 이 합의는 핵사찰과 관

련 북미간의 입장차가 상당한 접근을 보임에 따라 양국 당국자들 간의 고위급회담을 개최하는 것에 양측이 합의했다는 워싱턴발 보도가 나온 직후에 발표되었다. 우리 데이터에서 1994년 5월 전후의 뉴스 보도가 긍정적이면서 상당히 높은 변동폭을 보이는 것은 이런 맥락에서 해석이 가능할 것이다.

세 번째 변화점은 1994년 10월경이었는데 이 시점을 전후로 뉴스 보도의 어조는 중립적인 태도를 띄면서 상당히 폭의 변동폭을 보이는 것이 감지되었다. 이 시점을 전후한 북한관계 사건들을 살펴보면 1994년 10월 21일, 북한과 미국의 협상단은 북한이 핵개발 의심시설에 대한 특별 사찰을 거부하면서 시작된 18개월에 걸친 위기상황을 봉합하는데 공식적으로 합의하였다. 양국은 이미 8월에 합의를 위한 기본틀에 동의한 바 있고, 이것이 10월 합의의 기초가 되었다. 양국은 8월 이전에도 북한의 93년 3월 핵확산금지조약(NPT) 탈퇴로 촉발된 긴장상태를 타결하기 위한 부분적 합의를 도출해 낸 바 있었으나 최종적인 합의는 1994년 10월에 이르러서야 이뤄지게 되었다.

네 번째 변화점은 2000년 10월이었다. 2000년 10월, 매들린 올브라이트 당시 미국무부장관이 클린턴 대통령의 북한 방문 가능성을 논의하기 위해 이틀간 북한을 방문하는 역사적인 사건이 일어났다. 그러나, 올브라이트 장관은 북미관계 개선을 위해서는 북한의 핵무기 개발능력에 대한 투명성 제고가 선행되어야 한다는 점을 분명히 하였고, 결국 클린턴 대통령의 방북은 끝내 이뤄 지지 못했다. 또한, 비슷한 시기에 두 대의 미국 전투기가 실수로 남북한 국경을 침범하는 사건이 발생하여 양국 간의 긴장완화에 걸림돌이 되기도 했다. 실제로 이 네 번째 주요 변화점 이후 뉴스 보도의 어조가 매우 부정적으로 변한 것을 확인할 수 있다.

마지막 변화점은 2003년 2월이었는데 이것은 제2차 북핵위기와 일치한다고 할 수 있는데 2002년 가을부터 이미 제2차 북핵위기의 전조를 예시하는 일련의 사건들이 일어나기 시작했다. 좀 더 구체적으로 적시하자면 2002년 10월 16일 북한당국이 제임스 켈리 미 국무부 차관보에게 핵무기 개발 프로그램의 존재를 시인함으로써 제2차 북핵위기가 촉발되었다고 할 수 있다. 이에 대한 대응으로 2002년 11월에는 한반도 에너지 개발기구(KEDO)는 1994년 제네바 합의에 의해 지원했던 대북 중유 지원을 중단키로 결정했고, 이에 반발하여 북한은 12월 13일 핵시설 동결 해제를 선언했으며, 22일에는 핵시설 봉인 제거작업을 시작하면서 국제 원자력 기구(IAEA)의 감시 카메라를 제거하기에 이르렀다. 또 닷새 후에는 국제원자력기구(IAEA)에서 파견된 감시요원을 추방했다. 또한 2003년 1월 10일에는 북한은 핵확산금지조약(NPT) 탈퇴를 선언하면서 그 효력이 즉시 발효한다고 발표하면서 이 탈퇴선언은 1993년 이미 했던 핵확산금지조약 탈퇴를 임시정지 시킨 것이었을 뿐이기 때문에 공식절차인 “탈퇴 3개월 전 통보”는 불필요하다는 주장을 폈다. 급기야 북한은 2003년 2월 26일에 영변의 핵원자로를 재가동하기 시작하면서 현재까지도 해결의 실마리가 보이지 않고 있는 제2차 북핵위기가 본격화 되었다.

5. 결론

본 연구에서는 대중매체의 연구에 있어 중요한 도구가 되는 언론보도자료의 여러 통계적 성질에 대하여 살펴보았다. 특히 언론보도자료에 있어 금융 시계열자료에서 흔히 나타나는 arch효과와 변화점 모형에 대하여 살펴보고 이를 실제 북한 관련 실제 언론보도자료에 적용하여 보았다. 본 연구의 결과를 추후 한국의 주요 언론들 간의 비교나 또는 한국과 미국의 매체들 간의 비교에 응용할 수 있으리라 사료된다.

References

Akaike, H. (1974). A new look at the statistical model identification, *IEEE-Transactions on Automatic Control*, **19**, 716–723.

Bellman, R. (1975). *Dynamic Programming*, Princeton Univ Press, Princeton, New Jersey.

Boysen, L., Kempe, A., Liebsher, V., Munk, A. and Wittich, O. (2009). Consistencies and rates of convergence of jump-penalized least square estimators, *The Annals of Statistics*, **37**, 157–184.

Curran, J., Iyengar, S., Lund, A. B. and Salovaara-Moring, I. (2008). Media system, public knowledge and democracy: A comparative study, *European Journal of Communication*, **24**, 5–26.

Erbring, L., Goldenberg, E. N. and Miller, A. H. (1980). Front-page news and real-world cues: A new look at agenda-setting by the media, *American Journal of Political Science*, **24**, 16–49.

Friedrich, F., Kempe, A., Libesher, V. and Winkler, G. (2008). Complexity penalized M-estimation: Fast computation, *Journal of Computational and Graphical Statistics*, **17**, 1–24.

Gans, H. J. (1980). *Deciding What's News*, Vintage Books, New York.

Hannan, E. J. and Quinn, B. G. (1979). The Determination of the order of an auto-regression, *Journal of the Royal Statistical Society-Series B*, **41**, 190–195.

Iyengar, S. and McGrady, J. A. (2007). *Media Politics: A citizen's Guide*, W.W. Norton, New York.

Lippmann, W. (1922). *Public Opinion*, Free Press, New York.

MacKuen, M. J. and Coombs, S. L. (1981). *More Than News*, Sage Publications, Beverly Hills, C.A.

Schwarz, G. E. (1978). Estimating the dimension of a model, *The Annals of Statistics*, **6**, 461–464.

Tsay, R. S. (2009). *Analysis of Financial Time Series*, John Wiley & Sons, Hoboken, New Jersey.

2012년 5월 19일 접수; 2012년 9월 15일 수정; 2012년 9월 26일 채택