The Korean Journal of Applied Statistics (2013) DOI: http://dx.doi.org/10.5351/KJAS.2013.26.6.987
26(6), 987–998

Time Series Modelling of Air Quality in Korea: Long Range Dependence or Changes in Mean?

Changryong Baek[a,1]

[a]Department of Statistics, Sungkyunkwan University

(Received October 3, 2013; Revised October 28, 2013; Accepted November 14, 2013)

Abstract

This paper considers the statistical characteristics on the air quality (PM10) of Korea collected hourly in 2011. PM10 in Korea exhibits very strong correlations even for higher lags, namely, long range dependence. It is power-law tailed in marginal distribution, and generalized Pareto distribution successfully captures the thicker tail than log-normal distribution. However, slowly decaying autocorrelations may confuse practitioners since a non-stationary model (such as changes in mean) can produce spurious long term correlations for finite samples. We conduct a statistical testing procedure to distinguish two models and argue that the high persistency can be explained by non-stationary changes in mean model rather than long range dependent time series models.

Keywords: Long range dependence, power-law tail distribution, changes in mean.

1. 서론

산업화는 인간의 삶을 물질적으로 풍요하게 만들었지만 전 세계적인 인구 증가와 급격한 산업발달은 대기질(air quality)를 심각하게 악화시켰다. 이는 곧 인간의 건강을 위협하고 삶의 질을 떨어뜨리는 주요 요인이 되고 있다. 이러한 전 세계적인 관심은 안전한 대기질 관리로 이어지고 있으며 이에 발맞추어 우리 나라에서는 효율적인 대기질 관리를 위하여 전국에 대기 오염 물질 측정소를 설치 운영하여 기초적인 통계 자료를 실시간으로 수집을 하고 있다.

주요 대기오염물질로서는 미세먼지(PM10)를 1시간 간격으로 오존, 이산화질소, 일산화탄소, 아황산가스의 농도를 5분 간격으로 측정하고 있다. 이 논문에서는 특히 미세먼지인 PM10의 통계적 분석에 초점을 맞추고 있다. PM10(particulate matter)은 대기중에 존재하는 미세먼지 중 $10\mu m/m^3$ 보다 작은 미립자로 폐에 흡착되어 심각한 호흡기 질환 및 많은 건강 문제를 일으키고 있음이 알려져 있다. 이에 따라 PM과 사망률에 대한 연구가 활발히 진행되고 있는 상태이며 우리나라의 PM10은 선진국과 비교하여 크게는 3배 이상 높은 수준으로 특별한 관심이 필요하다.

이 논문에서는 대기 오염도를 측정하는 미세먼지 PM10의 시계열 자료분석을 통해 우리 나라에서 관측되는 PM10의 특징을 살펴본다. 기존 논문을 통해, 예를 들어 Windsor과 Toumi (2001), Varotsos 등 (2005) 그리고 Pan과 Chen (2008), 영국, 그리스, 미국, 대만에서의 PM10의 경우 강한 종속

[1]Assistant Professor, Department of Statistics, Sungkyunkwan University, Sungkyunkwan-ro, Jongno-gu, Seoul 110-745, Korea. E-mail: crbaek@skku.edu

성(long range dependence, persistency)과 함께 꼬리가 두터운 분포를 따름이 발표되었다. 본 논문에서는 2011년 매 시간 수집된 우리나라의 PM10 자료 분석을 통하여 위 두가지 특성에 대해서 살펴본다.

하지만, 선행연구에서는 과연 그러한 강한 종속성이 어떠한 원인에 의해서인지에 대한 통계적인 고찰이 이루어지고 있지 않고 있다. 또한 단순한 평균변화(changes in mean) 모형의 경우 유한 표본일 때 마치 강한 종속성을 띠고 있는 것처럼 보인다는 사실이 잘 알려져 있다. 이는 곧 현실적으로 확률모형을 세울 때 이 두가지의 구분이 매우 중요한 문제임을 알려준다. 즉, 만약 이러한 강한 상호관계를 설명하기 위하여 장기종속시계열 모형을 사용한다면, 예를 들어 FARIMA(p, d, q) 모형, 기본적으로 정상 시계열을 가정하게된다. 즉, 장기 예측값은 평균값으로 수렴하게 된다. 반면, 평균변화 모형은 비정상 모형이므로 예측값은 평균변화 모형에 따라 결정된다. 따라서, 어떠한 모형을 사용하느냐에 따라서 통계적인 가설 검증 및 예측값은 매우 다르게된다.

본 논문은 기본적인 미세먼지 시계열의 특징 탐색뿐만 아니라 이러한 강한 종속성을 어떠한 모형이 더 잘 설명하는지 근본적인 확률 모형(physical model)에 대한 논의를 심도있게 다룬다. 즉 매우 높은 시차에서도 사라지지 않은 높은 수준의 상관관계가 정상시계열인 장기종속시계열(long range dependnece; LRD) 모형에 의해서 생성된 것인지 아니면 외부 개입에 의한 비정상 구조변화모형(structural breaks, changes in mean)에 따른 그럴듯한 가짜 현상인지에 대해서 통계적 가설 검정을 살펴보도록 한다.

2. 장기종속시계열(Long range dependence)

장기범위종속(long range dependence; LRD) 시계열 혹은 장기 기억 보존(long memory) 시계열 $\{X_j, j \in \mathbb{Z}\}$이란 정상시계열(weakly stationary time series)로서 자기상관함수(autocorrelation function)가 시차(lag)에 따라서 다음과 같이

$$\gamma(h) = \mathrm{Corr}(X_0, X_h) \sim Ch^{2d-1}, \quad C > 0,\ d \in (0, .5). \tag{2.1}$$

멱함수(power-law) 형태로 서서히 감소하는 시계열이다. 수식 (2.1)에 나오는 모수 d를 LRD 계수 혹은 LRD 모수라고 부른다. 이러한 느린 감소로 인해서 LRD 시계열은 자기 상관함수들의 시차에 따른 합은 발산하게 된다. 즉,

$$\sum_{h=-\infty}^{\infty} \gamma(h) = +\infty.$$

이와 반대로 자기 상관함수들의 합이 절대수렴하는

$$\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty.$$

시계열을 단기범위종속(short range dependence; SRD) 시계열이라 칭한다. SRD 시계열의 대표적인 모형은 ARMA 모형이다.

또한 LRD 시계열의 스펙트럼 밀도 함수는 적절한 가정하에 다음과 같이 주어진다.

$$f(\lambda) \sim c|\lambda|^{-2d}, \quad \lambda \to 0,\ c > 0. \tag{2.2}$$

따라서 LRD 시계열은 원점 근처 주파수(frequency)에서 높은 밀도함수를 가지므로 시계열 그림(time plot)상에서 LRD 시계열은 마치 커다란 주기(period, cycle)를 갖는 듯한 패턴을 가지게 된다. 즉 이러

한 커다란 주기가 매우 높은 양의 상관관계로 이어지게되므로 LRD 시계열은 정상시계열이지만 높은 지속성(high persistency)을 갖는 시계열이다.

특별히 SRD 시계열의 경우 원점근방에서의 밀도함수가 유한하므로 LRD 모수 $d = 0$을 갖게된다. 그러므로 LRD 모수 d를 통해 LRD 혹은 SRD 시계열인지 구분할 수 있으며, 만약 $d = 1$인 경우에는 비정상 단위근(unit root)이 된다. 이처럼 LRD 모수는 시계열 모형의 특징을 잘 나타내주는 모수이고 정확한 LRD 모수의 추정은 시계열 연구에 있어서 매우 중요하다. 먼저 수식 (2.2)의 양변에 로그 변환(log-transformation) 을 취하게되면

$$\log f(\lambda) \approx \log c - 2d \log |\lambda|$$

란 선형관계식을 얻게된다. 따라서 관측자료 $X_1, \ldots, X_n$에 대해서 스펙트럼 밀도함수를 푸리에 주파수(Fourier frequency) $\omega_l = 2\pi l/n$, $l = 1, \ldots, [n]/2$에 따라 주기도(periodogram)

$$I_X(\omega_l) = \frac{1}{2\pi n} \left| \sum_{j=1}^{n} X_j e^{-ij\omega_l} \right|^2$$

를 이용하여 추정할 경우, 푸리에 주파수와 주기도가 log-log 그림에서 직선형태로 감소함을 알 수 있다. 이러한 선형 관계는 LRD 모수를 직선회귀에 의해서 추정할수 있게 하며 이를 Geweke와 Porter-Hudak(GPH) 추정량이라고 부른다. 좀 더 엄밀히 표현하면 GPH 추정량은

$$\hat{d}_{gph} = \frac{\sum_{l=1}^{m} (z_l - \bar{z}) \log I_X(\omega_l)}{\sum_{l=1}^{m} (z_l - \bar{z})^2}, \quad z_l = -2 \log \omega_l, \ \bar{z} = \sum_{l=1}^{m} \frac{z_l}{m} \tag{2.3}$$

으로 표현할 수 있고 m은 추정에 쓰인 푸리에 주파수의 개수이다. Robinson (1995b)에 따르면 GPH 추정량은 일치 추정량이고 또 점근적 정규성을 가진다

$$\sqrt{m}(\hat{d}_{gph} - d) \sim \mathcal{N}\left(0, \frac{\pi^2}{24}\right).$$

수식 (2.3)에서 보듯이 GPH 추정량은 m에 의존하게 된다. Geweke와 Porter-Hudak (1983)은 본래 $m = \sqrt{n}$을 사용하였으나 실제 응용에서는 m의 값에 따른 GPH 추정량들의 자취 그림을 통해서 안정된 값을 갖는 구간을 점검해 봄으로서 결정한다. 이론적으로는 정규분포 가정하에서 $m = O(n^{4/5})$에서 가장 좋다는 것이 밝혀졌다 (Hurvich 등, 1998).

최소제곱합(least squares estimation)에 기반한 GPH 추정량과 비교하여 국소 휘틀(local Whittle; LW) 근사에 의한 최대 우도 추정량(maximum likelihood estimator)은 다음과 같이 주어진다

$$\hat{d}_{lw} = \underset{d \in (0,1)}{\operatorname{argmin}} \ \log \left(\frac{1}{m} \sum_{l=1}^{m} \omega_l^{2d} I_X(\omega_l) \right) - 2d \frac{1}{m} \sum_{l=1}^{m} \log \omega_l. \tag{2.4}$$

Robinson (1995a)에 따르면 점근적으로 LW 추정량은 다음의 정규성을 가진다

$$\sqrt{m} \left(\hat{d}_{lw} - d \right) \sim \mathcal{N}\left(0, \frac{1}{4}\right).$$

따라서 LW 추정량이 GPH 추정량보다 분산이 약간 더 작음을 알 수 있다. 푸리에 주파수 개수 m에 대한 결정 방법은 GPH 추정량과 동일하게 자취 그림을 통해 결정한다.

3. 평균변화 모형(Changes in mean model)

본 연구에서는 각 구간에서 일정한 평균을 가지는 다음의 단순 평균변화 모형을 생각한다. 평균 변화점(break points) $k_1, \ldots, k_R$에 대하여 각각의 구간에서의 평균변화량 Δ_r, $r = 1, \ldots, R$ 및 SRD 오차 $\{\epsilon_t, t = 1, \ldots, n\}$에 대하여 다음과 같이 표현된다.

$$X_t = g(t) + \epsilon_t, \quad g(t) = \mu + \sum_{r=1}^{R} \Delta_r 1_{\{k_r < t \le n\}}. \tag{3.1}$$

즉, $(R+1)$개의 부분표본에 대해서 평균이 각각 다른 모형으로서, 주어진 표본 $\{X_1, \ldots, X_n\}$에 대하여 평균이 변화하는 변화점들 $k_1, \ldots, k_R$은 알려지지 않은 모수이다. 이러한 변화점을 찾기위한 방법으로는 CUSUM 통계량에 기반한 방법이 가장 많이 쓰이고 있다. 이해를 돕기위해 한 개의 변화점을 갖는 경우를 생각하자. CUSUM 통계량은 다음과 같이 정의된다.

$$\text{CUSUM}(k) = \frac{1}{\sqrt{n}} \left(\sum_{j=1}^{k} X_j - \frac{k}{n} \sum_{j=1}^{n} X_j \right) = \frac{k}{n} \left(1 - \frac{k}{n} \right) \sqrt{n} \left(\frac{1}{k} \sum_{j=1}^{k} X_j - \frac{1}{n-k} \sum_{j=k+1}^{n} X_j \right)$$

따라서, CUSUM(k)를 k번째 관측치를 기준으로 전후 평균의 차이에 대해 표본개수에 따른 조정을 해주는 값이라 볼 수 있다. 이러한 CUSUM 통계값에 대해서 평균변화점은 평균의 변화가 가장 크게 일어나는 값으로

$$\hat{k}_{CUSUM} = \underset{1 \le k \le n}{\text{argmax}} |\text{CUSUM}(k)|$$

으로 주어진다. CUSUM 통계량에 대해서 변화점이 없을 경우 브라우니안 브리지(Brownian Bridge) $\{B^0(t), t \in [0,1]\}$에 대하여 다음의 극한 분포를 가짐이 알려져 있다

$$\max_{1 \le k \le n} \frac{1}{\sigma} |\text{CUSUM}(k)| \xrightarrow{d} \sup_{0 \le t \le 1} \left| B^0(t) \right|.$$

여기에서 $\sigma^2 = \sum_{h=-\infty}^{\infty} \gamma(h)$이다. 따라서 표본 자기 상관함수 $\hat{\gamma}(h)$에 대해서 바틀렛추정치(Bartlett estimator)

$$\hat{\sigma^2} = \sum_{h=-q}^{q} \left(1 - \frac{h}{q+1} \right) \hat{\gamma}(h)$$

를 통하여 변화점의 존재 여부에 대한 가설 검증을 할 수 있다. 만약 $R \ge 2$일 경우 다중변화점이 되므로 이에 대한 추정은 이원분리(binary segmentation)을 이용한다. 간단히 설명하자면, 첫번째 변화점 추정량 $\hat{k}$에 대해서 전체 표본을 두개의 보조 표본(sub samples)으로 나뉜 뒤, 앞서 설명한 가설 검정 방법을 각각의 보조 표본에 적용하여 추가 변화점이 있는지 살펴보는 방법이다. 이렇게 찾은 변화점의 개수는 일치 추정량임이 잘 알려져 있다. 이러한 CUSUM을 비롯한 다양한 방법들이 지난 수십년 동안 연구되었다. 예를 들어, 평균변화에 다른 가중치를 주는 다음 변화점 추정량의 경우 조정된 CUSUM(adjusted CUSUM)이라 부르며

$$\hat{k}_{adjCUSUM} = \underset{1 \le k \le (n-1)}{\text{argmax}} \left| \left(\frac{k}{n} \left(1 - \frac{k}{n} \right) \right)^{-\frac{1}{2}} \text{CUSUM}(k) \right|$$

으로 주어진다. 또한 잔차를 이용한 CUSUM 통계량, 바틀렛 추정치를 개선한 다양한 추정방법등 많은 방법들이 제안되었다. 이론적인 성질들을 비롯한 변화점 추정에 대한 자세한 사항은 Csörgő와 Horváth (1997) 및 Baek과 Pipiras (2013) 를 참조한다.

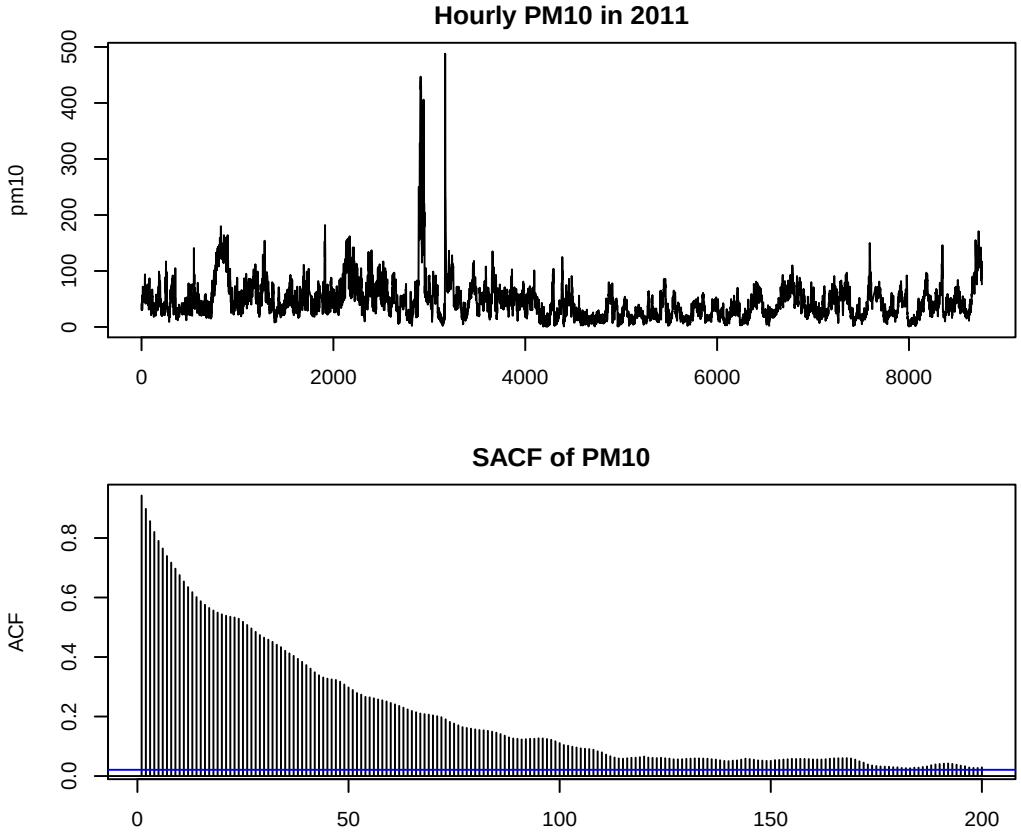

Figure 4.1. Time plot of PM10 (top) and sample autocorrelation plot (bottom).

4. 실증분석

본 장에서는 대기 오염도를 측정하는 중요한 지표중에 하나인 미세먼지 PM10를 2011년 한해동안 매시 관측한 시계열을 분석한다. 분석에 쓰인 자료는 2011년 국립환경과학원 확정자료이며 총 8760개($365 * 24$)의 관측자료이고 267개의 결측치에 대해서는 선형보간(linear interpolation)을 적용하였다.

4.1. 미세먼지 PM10의 장기범위종속성

먼저 Figure 4.1은 PM10의 시계열 그림 및 표본자기상관함수(sample autocorrelation function; SACF) 그림을 나타낸다. 먼저 첫 번째 시계열 그림에서 많은 관측값들이 군집을 이루어서 오랫동안 머무르는 경향을 확인할 수 있다. 예를 들어 600–700번째 관측값들은 $25\mu m/m^3$ 근처에서 변화를 보인 반면 800–900 번째 관측값들은 $125\mu m/m^3$에 머무르고 있으며 이러한 군집은 전체 시계열에서 나타나고 있다. 두 번째 그림은 보다 명확하게 높은 상관관계를 보여주고 있다. 시차 1에서 SACF는 .94로 1에 매우 가깝고, 시차 100의 경우에서도 .11로 매우 완만하게 감소하고 있음을 알 수 있다.

이러한 장기범위종속성은 스펙트럼 분석에서도 명확히 드러난다. Figure 4.2는 PM10 데이터에 대한 주기도(periodogram)와 주파수를 log-log 그림으로 나타낸것이다. 왼쪽 패널에서 살펴보듯이 뚜렷한

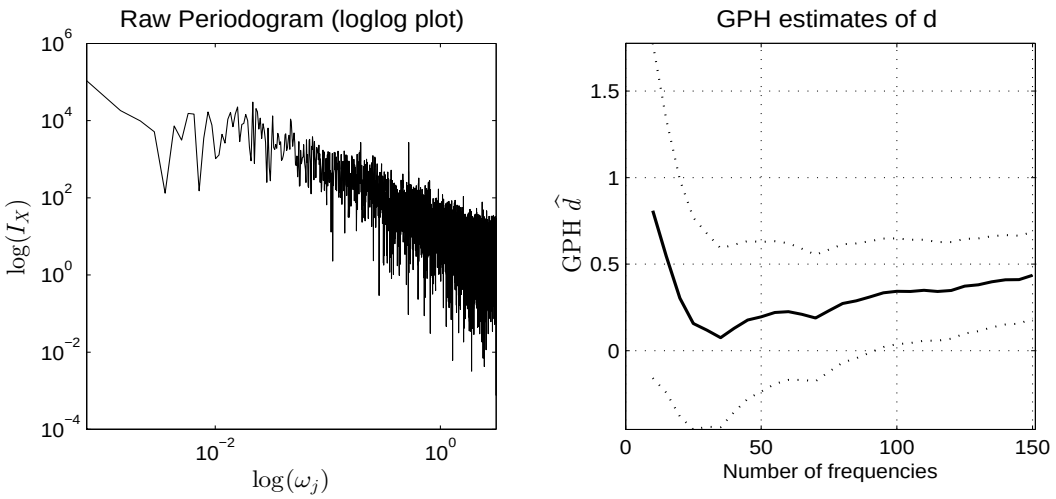

Figure 4.2. Raw periodogram (left) and GPH estimates of LRD parameter (right).

선형감소 추세가 나타나고 있다. 오른쪽 패널은 GPH 추정량을 이용한 LRD 모수의 주파수에 따른 자취 그림을 나타냈다. 사용 주파수가 $\sqrt{8970} \approx 100$ 근방에서 GPH 추정량 값은 대략 .34로 안정적으로 추정되고 있으며 정상시계열 범위인 $(0, .5)$에 있음을 알 수 있다.

결론적으로 표본자기상관함수 그림 및 주기도 그림, GPH 추정량을 통해서 2011년에 수집된 미세 먼지 PM10이 매우 강한 종속관계를 가지는 LRD 시계열임을 확인할 수 있다.

4.2. 미세먼지 PM10의 멱함수 꼬리분포

시계열 그림 Figure 4.1에서 대략 2900번째(4월초) 및 3100번째(5월초)에 두 개의 커다란 뾰족함(peakness)이 관찰되었다. 이는 우리나라에 발생한 황사의 영향으로(실제로 5월 1일부터 4일까지 전국에 황사주의보가 내려졌다) 발생한 이상점(outlier)로 판단된다. 하지만, 이러한 이상점을 제외하고 나더라도 전체적으로 많은 요철이 관측되고 있다. 이에 따라 PM10의 주변분포(marginal distribution)이 멱함수 형태의 꼬리분포(power-law tail distribution)를 가지는지 살펴보았다. 꼬리 지수(tail-index) α를 가지는 멱함수 꼬리 분포는 다음과 같이 표현된다. 꼬리 분포(tail distribution) $\overline{F}(x)$에 대해서

$$\overline{F}(x) := 1 - F(x) \sim cx^{-\alpha}, \quad c > 0,\ \alpha > 0. \tag{4.1}$$

가장 오래되고 직관적인 꼬리 지수에 대한 추정은 다음과 같다. 수식 (4.1)의 양변에 로그 변환을 취하면 큰 x값에 대해서

$$\log \overline{F}(x) \approx \log c - \alpha \log x$$

임을 알 수 있다. 순서통계량 $X_{(1)} \le X_{(2)}, \ldots, X_{(n)}$에 대하여 경험적 꼬리 분포함수(empirical tail distribution) $\overline{F}(X_{n-i+1}) \approx i/n$이므로 위 관계식은

$$\log\left(\frac{i}{n}\right) \approx \log c - \alpha X_{n-i+1}$$

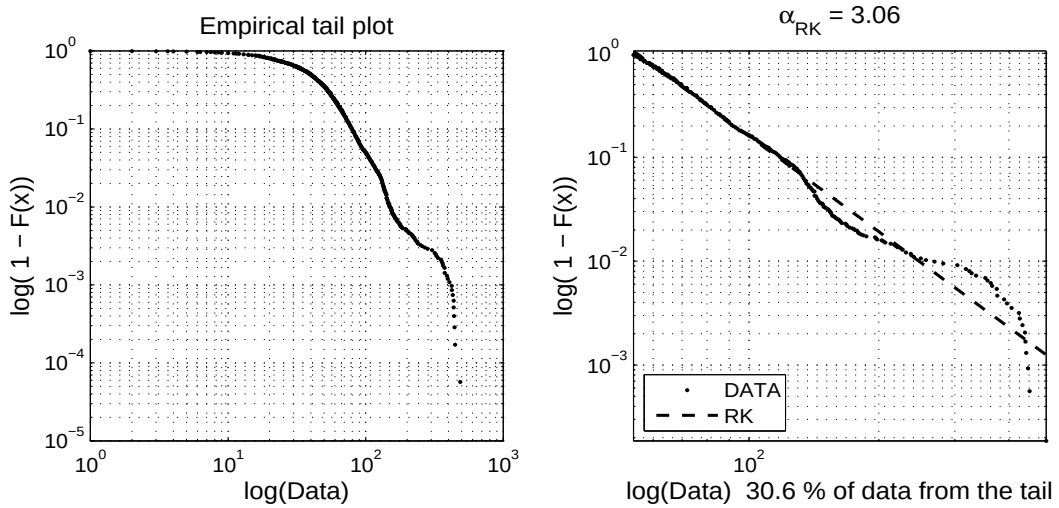

Figure 4.3. Empirical tail plot (left) and estimation of tail index (right).

임을 알 수 있고 따라서 경험적 분포함수(empirical distribution) 과 순서통계량을 기초하여 log-log 그림에서 선형적합을 통해서 꼬리 지수 α를 추정할 수 있다. 이 추정량을 순위기반(rank-based) 추정량 $\hat{\alpha}_{RK}$이라고 명하겠다. 여기에서도 추정에 쓰인 큰 순서통계량(upper order statistics)의 선택이 중요하고 앞선 LRD 모수의 추정처럼 여러 개의 값에 대해서 변화가 작아지는 값을 선택한다. 여러 꼬리 지수 추정에 대한 자세한 내용은 Baek과 Pipiras (2010)을 참고한다.

Figure 4.3에서 살펴보듯 PM10 관측치의 주변 분포가 멱함수 꼬리를 가짐을 알 수 있다. 꼬리 지수를 추정하기 위해서 대략 상위 30%의 큰 순서통계량들을 사용하였을 때 꼬리 지수가 안정되게 추정되며 대략 3에 가까움을 확인할 수 있었다. 또한 다른 꼬리 추정치, 예를 들어 Hill estimator도 비슷한 꼬리 지수 값을 보여주었다. 사실 멱함수 꼬리는 LRD 시계열에서 매우 잘 관찰되는 분포이며 $\alpha \in (1,2)$인 두꺼운 꼬리분포(heavy-tail distribution)은 LRD 시계열을 생성하는 물리적 모형을 규명짓는 중요한 역할을 한다 (Doukhan 등, 2003).

Pan과 Chen (2008)에서는 PM10의 주변분포가 로그-정규분포(Log Normal distribution)를 따름을 밝히고 FARIMA(Fractionally integrated autoregressive moving average)모형을 이용하여 LRD 모형을 세웠다. 여기에서는 Mandelbrot (1997)에 따라 로그-정규분포는 멱함수 꼬리를 갖는 분포와 매우 쉽게 혼동이 되어 주의가 필요함을 상기시킨다. 예를 들어서 Figure 4.4는 멱함수 꼬리 분포를 일반화 파레토 분포(generalized Pareto distribution)

$$\overline{F}(x) = \left(1 + \xi \frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}}, \quad x > \mu,\ \sigma > 0,\ \xi > 0$$

를 이용하여 모수적 추정을 할 경우와 로그-정규분포를 사용하여 추정을 할 경우 QQ-plot을 통해 비교하였다. 그림의 점선은 95% 신뢰대(confidence band)를 나타낸다. 오른쪽 패널에서 확인할 수 있듯이 로그-정규분포의 경우 큰 값들에 대해서 관측값이 신뢰대를 크게 벗어남을 알 수 있고, 이는 곧 주변함수의 주요한 특징인 두터운 꼬리를 로그-정규분포를 통해서는 적절히 반영할 수 없음을 의미한다. 반면 일반화 파레토 분포는 두터운 꼬리 분포를 대체적으로 잘 적합하고 있다.

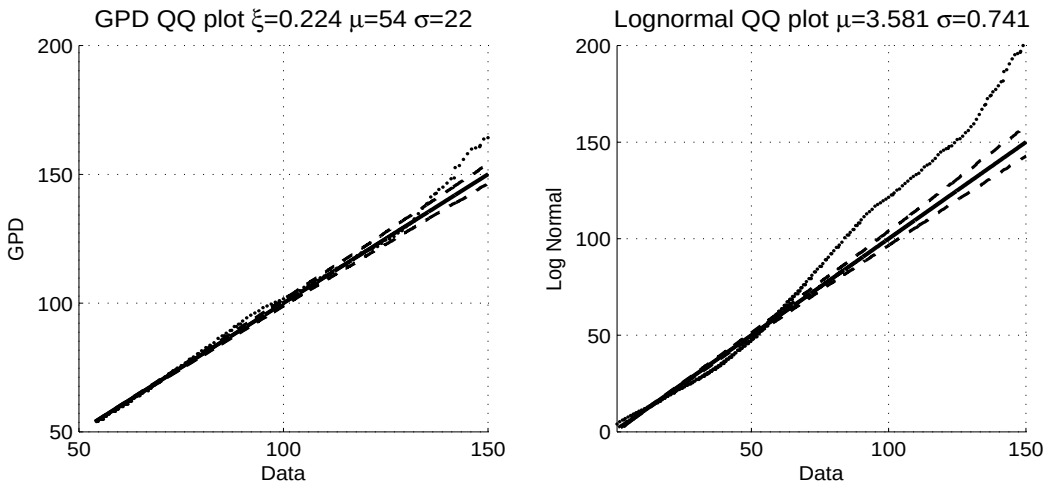

Figure 4.4. QQ-plots for parametric modelling of tail distribution.

4.3. LRD 모형 및 비정상 평균변화 모형

앞서 살펴보았듯이 LRD 모형은 원점 근처의 스펙트럼 밀도함수가 높기 때문에 시계열 그림에서는 마치 큰 사이클이 존재하는 것처럼 보인다. 따라서 유한 표본에 있어서 LRD 모형과 평균 변화 모형(changes in mean model)은 매우 유사하여 종종 큰 혼동을 야기한다. 하지만, 평균 변화모형은 비정상 시계열이지만 역시 유한 샘플에서 마치 LRD처럼 보이므로 정상 시계열로 착각하게 된다. 이 둘을 구분짓는 일은 모형 분석을 통해 물리적인 특성을 성찰하고자 하는 현실에서 매우 중요하며 시계열 분석에서 가장 중요한 미래값 예측에 있어서 어떠한 모형을 사용하느냐에 따라 그 예측값은 전혀 다르게 된다. 즉 LRD 모형은 정상시계열이므로 미래의 예측값은 시계열의 평균값으로 수렴하게 된 반면 비정상 평균모형의 경우 이러한 평균이 변화함을 의미하여 예측값에는 큰 차이가 있게된다. 보다 심도 있는 논의는 Diebold와 Inoue (2001) 그리고 Baek과 Pipiras (2012) 및 위 논문들에서 인용한 논문들을 참조한다.

Baek과 Pipiras (2012, 2013)에서는 잔차를 이용한 간단하면서도 LRD 시계열에 대해서 매우 검정력이 높은 통계방법을 제안하였다. LW 추정에 따른 가설 검정 방법(LW LRD test)은 다음과 같이 요약될 수 있다. 우리가 검정하는 가설은 다음과 같다.

$$H_0: \text{ 비정상 평균변화 모형 (3.1)}, \quad H_a: \text{ LRD 모형}.$$

(0단계): $\hat{R} = 0$이라 하자.

(1단계): 주어진 $\hat{R}$에 대해서 3장에서 설명한 조정된 CUSUM(adjusted CUSUM) 변화점 추정을 통해서 $g(t)$를 추정하고 잔차 $R(t) = X_t - \hat{g}(t)$를 구한다.

(2단계): 1단계에서 구한 잔차에 대해서 LW 추정을 통해 LRD 모수를 추정한다.

(3단계): 만약 $2\sqrt{m}\hat{d}_{lw} \geq z_\alpha$이면 $\hat{R}$을 1 증가시키고 (1단계)로 되돌아 간다.

(4단계): (3단계)에서 기각한 횟수가 최종 $\hat{R}_{LW}$이 되고 비정상 평균변화 모형일 경우 참값 R에 수렴하게 되며, LRD 모형의 경우 $P(\hat{R}_{LW} \to \infty) = 1$이 되므로 두 모형을 구별할 수 있다.

하지만, 실관측데이터의 관측개수는 유한개이므로 많은 다중변화점을 갖는 경우 위의 알고리즘은 작은 표본에서 높은 검정력을 기대하기 힘들다. 그래서 Baek과 Pipiras (2013)에서는 3장에서 설명한

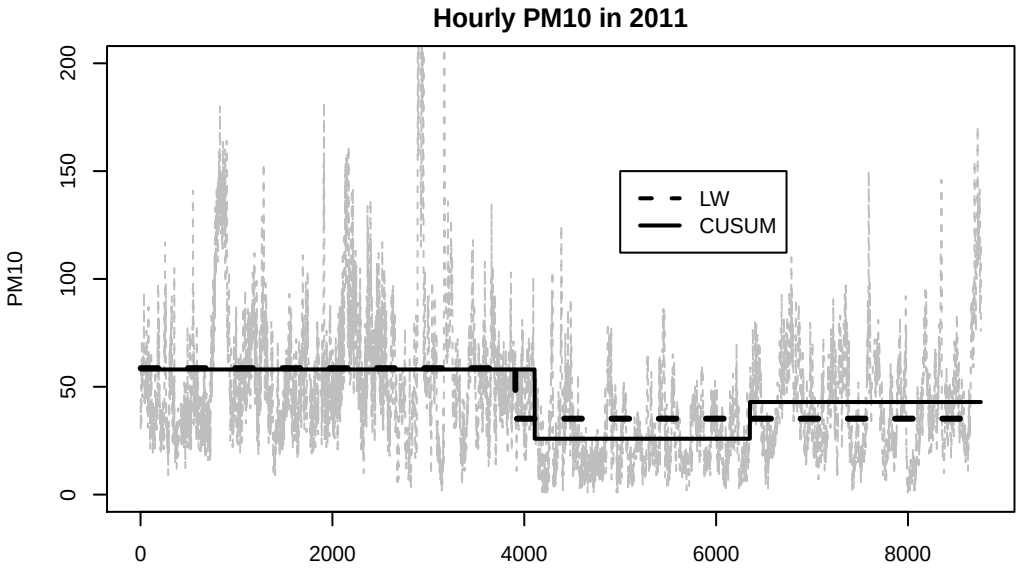

Figure 4.5. Estimated changes in mean models by LW and CUSUM.

CUSUM에 기반한 다중변화점추정 방법과의 상호 비교를 통해서 LW LRD test를 통해 얻어진 다중변화점 개수 비교를 통하여 LRD 시계열을 가려내는 방법을 제안하였고 많은 시뮬레이션과 실증분석을 통해서 그 타당성을 살펴보았다. 만약

$$\hat{R}_{CUSUM} \approx \hat{R}_{LW}$$

이면 평균변화모형을 따르며

$$\hat{R}_{CUSUM} \ll \hat{R}_{LW}$$

이면 LRD 시계열이다.

LW LRD test에 기반한 비정상 평균변화 모형과 LRD 시계열간의 가설검정 결과를 미세먼지 자료에 적용한 결과는 다음과 같다. Baek 과 Pipiras (2013)에서는 CUSUM에 기반한 4개의 다중변화점 추정방식, CUSUM, CUSUM-MAC, CUSUM-JX 그리고 CUSUM-RO을 사용하였다. 그 결과 CUSUM에 기반한 방법의 경우 두 개의 변화점 $\hat{k}_1 = 4108, \hat{k}_2 = 6352$을 추정하였다. LW LRD test에 기반한 변화점은 $\hat{k}_1 = 3907$로서 CUSUM 및 LW 방법 모두 한 개 혹은 두 개의 평균변화점을 찾아냈다. 변화점을 추정하는 방법이 근소하게 다르기 때문에 첫번째 변화점이 추정 방법에 따라 정확히 일치하지는 않으나 대략 6월 중순으로 거의 비슷함을 알 수 있다. 두번째 변화점은 CUSUM에 의한 방법론을 통해서만 찾아낸 변화점으로 대략 9월 중순이다. LW LRD test를 통해서 두 번째 변화점을 찾지 못한 것은 변화점을 찾는 추정방법의 문제가 아니라 변화점 개수를 찾는 가설검증의 차이다. 즉, CUSUM에서 찾은 변화점을 그대로 사용한다 할지라도 LW LRD test의 경우 첫번째 한 개의 변화점만을 유의하게 찾았다. Figure 4.5은 변화점에 따른 추정함수를 표본과 함께 보여주고 있다. 따라서 PM10 시계열의 경우 매우 천천히 감소하는 표본자기상관함수 및 원점 근처에서의 높은 밀도함수가 LRD 모형에 기반하기 보다는 비정상 평균변화 모형에서 오는 착시효과에 더 가깝다는 결론을 내릴 수 있다.

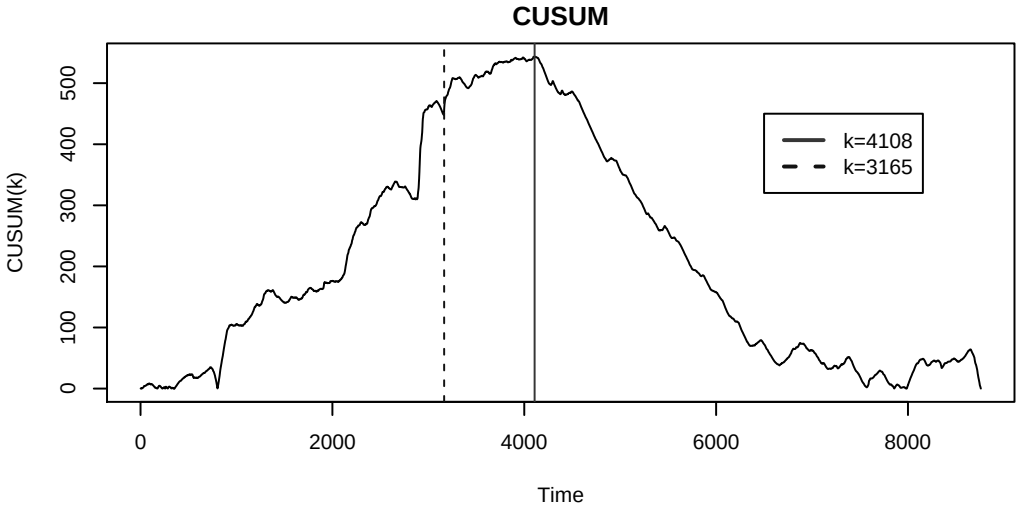

Figure 5.1. CUSUM(k) statistics for PM10.

5. 논의

이 논문에서는 대기질을 관리하는 주요 변량 중에 하나인 미세먼지의 특성에 대해서 살펴보았다. 분석 결과 미세먼지(PM10)는 매우 강한 양의 상관관계가 매우 큰 시차에서도 천천히 감소하는 LRD시계열의 특성을 보였으며 주변분포의 경우 로그변환을 취하더라도 사라지지 않을 정도의 꼬리가 두터운 멱함수형태의 꼬리 분포를 가짐을 알 수 있었다. 이는 한국의 미세먼지가 미국, 영국을 비롯한 다른 지역에서 관측되는 미세분포의 성질과 비슷함을 확인해주었다. 하지만, 한 걸음 더 나아가서 이러한 강한 종속성이 과연 정상 시계열인 LRD 시계열을 따르는 물리적 모형에 의한 것인지 아니면 비정상 구조변화모형이 유한 표본에서 가지는 유사 LRD 성질을 따른 것인지에 대한 검증도 함께 실시하였다. 그 결과 미세 먼지의 경우 LRD 시계열보다는 구조변화모형을 따름이 가설검정의 결과 드러났다.

첫 번째 평균변화점은 대략 6월 중순으로 또한 미세먼지의 수준이 큰 수준으로 감소하였다. 이러한 평균변화는 통상 6월 중순 혹은 말에 찾아오는 장마와 여름 기간 동안의 태풍의 영향으로 많은 비가 내려 미세 먼지가 씻겨 내려가 대기의 질이 향상되지 않았나 추측해본다. 즉 강수량이 미세먼지를 결정하는 중요한 인자가 될 수 있음을 의미한다. 두번째 평균변화점은 CUSUM에 의해서만 찾았으며 대략 9월 중순이다. 이는 곧 9월 들어서 태풍이 지나가고 강수량이 줄어듬에 따라 미세먼지의 수준이 다시 높아졌으리라 짐작해본다. 또한, Yang (2002) 의 결과에서도 보고됐듯이 미세 먼지는 대체로 여름에는 낮고 추운 겨울에 높아 기온이 미세먼지 수준을 결정하는 인자임을 알 수 있다. 하지만, 미세먼지의 자세한 변동폭을 설명하기에는 강수량과 온도가 부족함을 지적하였고 추후 많은 연구가 필요함을 주장하였다. 또한 두터운 꼬리 분포를 일반화 파레토분포를 통해 잘 적합할 수 있음을 볼 때 극한값이론에서 많이 쓰이는 POT(peak over threshold) 방법론을 통해서 강수량 및 온도를 비록한 여러 인자들에 대한 요인 분석을 통해 미세먼지를 보다 잘 이해할 수 있을 것이라 기대한다. 이렇게 찾아진 요인들에 대한 보다 체계적인 이해를 통해서 한국의 대기질을 향상시킬 수 있을 것이라 본다.

또 한가지 주목할 점은 봄철 황사에 대한 영향이 평균변화 모형에서는 반영되지 않았다. Figure 5.1은 CUSUM(k)의 통계량의 값을 나타내주는 그림이다. 여기에서 보듯이 황사가 시작된 점 $k = 3165$에서의 CUSUM 통계량값은 장마가 시작되는 지점 $k = 4108$의 CUSUM 통계량값 보다 작다. 두번째 변화

점을 찾는 경우에서도 CUSUM 통계량이 황사가 있는 시점에서는 최대값을 가지지 못한다. 이를 통해 살펴볼 때, 본 논문에서 고려한 단순 평균 변화 모형에 대해서 1년이라는 긴 시간에서는 황사의 효과가 일시적이였음을 알 수 있다. 대신 황사의 효과는 미세 먼지의 분포가 두터운 꼬리를 가짐에 더 많은 기여를 한 것으로 보인다.

References

Baek, C. and Pipiras, V. (2010). Estimation of parameters in heavy-tailed distribution when its second order tail exponent is known, *Journal of Statistical Planning and Inference*, **140**, 1957–1967.

Baek, C. and Pipiras, V. (2012). Statistical tests for a single change in mean against long-range dependence, *Journal of Time Series Analysis*, **33**, 131–151.

Baek, C. and Pipiras, V. (2013). On distinguishing multiple changes in mean and long-range dependence using local Whittle estimation, *submitted*, Available from: http://web.skku.edu/crbaek.

Csörgő, L. and Horváth, L. (1997). *Limit Theorems in Change-point Analysis*, Wiley & Sons Ltd., Chichester.

Doukhan, P., Oppenheim, G. and Taqqu, M. (2003). *Theory and Applications of Long-Range Dependence*, Birkhäuser Boston Inc., Boston.

Diebold, F. and Inoue, A. (2001). Long memory and regime switching, *Journal of Econometrics*, **105**, 131–159.

Geweke, J. and Porter-Hudak, S. (1983). The estimation and application of long memory time series models, *Journal of Time Series Analysis*, **4**, 221–238.

Hurvich, C., Deo, R. and Brodsky, J. (1998). The mean squared error of Geweke and Porter-Hudak's estimator of the memory parameter, *Journal of Time Series Analysis*, **19**, 19–46.

Mandelbrot, B. (1997). A case against the lognormal distribution, In *Fractals and Scaling in Finance*, Springer, New York.

Pan, J. and Chen, S. (2008). Monitoring long-memory air quality data using ARFIMA model, *Environmetrics*, **19**, 209–219.

Robinson, P. M. (1995a). Gaussian semiparametric estimation of long range dependence, *The Annals of Statistics*, **23**, 1630–1661.

Robinson, P. M. (1995b). Log-periodogram regression of time series with long range dependence, *The Annals of Statistics*, **23**, 1048–1072.

Varotsos, C., Ondov, J. and Efstathiou, M. (2005). Scaling properties of air pollution in Athens, Greece and Baltimore, Maryland, *Atmospheric Environment*, **39**, 1352–2310.

Yang, K.-L. (2002). Spatial and seasonal variation of PM10 mass concentrations in Taiwan, *Atmospheric Environment*, **36**, 3403–3411.

Windsor, H. and Toumi, R. (2001). Scaling and persistence of UK pollution, *Atmospheric Environment*, **35**, 4545–4556.

한국의 미세먼지 시계열 분석: 장기종속 시계열 혹은 비정상 평균변화모형?

백창룡[a,1]

[a]성균관대학교 통계학과

(2013년 10월 3일 접수, 2013년 10월 28일 수정, 2013년 11월 14일 채택)

요 약

이 논문에서는 한국의 대기질을 결정하는 중요한 수치인 미세먼지(PM10)에 대한 통계적 고찰을 한다. 2011년 매시 관찰된 자료 분석을 토대로 미세먼지가 매우 높은 시차에서도 강한 양의 상관관계를 가지는 장기 종속 시계열의 특징을 보임을 밝힌다. 또한 주변분포는 꼬리가 두터운 모형으로서 로그-정규분포보다는 일반화 파레토 분포가 훨씬 더 자료를 잘 적합함을 보인다. 하지만 이러한 높은 상관관계는 종종 단순한 평균변화 모형에 의한 그럴듯싸한 가짜 효과에 기인하기도 하여 통계모형을 세우는데 많은 혼동을 준다. 따라서 이 논문에서는 강한 종속성이 장기 종속 시계열에 의한 것인지 아니면 비정상 평균변화에 의한 것인지 근본적인 물리적 모형에 대한 논의를 통계적인 가설 검정을 통해 살펴본다. 그 결과 미세먼지의 강한 종속성은 구조변화에의한 착시 효과임을 밝힌다.

주요용어: 장기종속시계열, 두터운 꼬리분포, 평균변화모형.

[1](110-745) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과, 조교수. E-mail: crbaek@skku.edu