

Threshold Modelling of Spatial Extremes - Summer Rainfall of Korea

Seungyong Hwang^a · Hyemi Choi^{a,1}

^aDepartment of Statistics, Chonbuk National University

(Received June 11, 2014; Revised August 6, 2014; Accepted August 6, 2014)

Abstract

An adequate understanding and response to natural hazards such as heat wave, heavy rainfall and severe drought is required. We apply extreme value theory to analyze these abnormal weather phenomena. It is common for extremes in climatic data to be nonstationary in space and time. In this paper, we analyze summer rainfall data in South Korea using exceedance values over thresholds estimated by quantile regression with location information and time as covariates. We group weather stations in South Korea into 5 clusters and fit extreme value models to threshold exceedances for each cluster under the assumption of independence in space and time as well as estimates of uncertainty for spatial dependence as proposed in Northrop and Jonathan (2011).

Keywords: Extreme value modeling, quantile regression, exceedance over threshold.

1. 서론

폭염, 태풍, 홍수, 가뭄 등과 같은 이상기후적인 사건들은 매우 드물게 발생하지만 큰 피해를 초래할 수 있는 사건들이다. 생명, 안전과 재산에 주는 직접적인 피해뿐만 아니라 기반 시설의 확충과 보험료의 상승 등 간접적인 피해도 만만치 않다. 이에 적절히 대응하기 위하여 통계적인 분석을 통한 정확한 예측에 대한 필요성이 최근 들어 대두되고 있다. 예를 들어 홍수의 경우에는 연 최대 강수량과 같은 극단값의 분포를 추정하여 대응책을 마련해야 할 것이다. 이와 같은 중요성에도 불구하고 우리나라에서는 극단값 모형을 통한 자료 분석의 예가 적은 편이다.

본 논문에서는 우리나라 여름철 강수량을 극단값 모형으로 분석해 보고자 한다. 한국은 여름철 장마와 태풍으로 인한 집중호우로 홍수 피해가 자주 발생한다. 우리나라 지역별로 여름철 강수의 집중 시기가 다르며 최근 들어 강수량이 증가하는 경향이 있음이 알려져 있다 (Kim 등, 2008; Choi, 2004; Lee와 Kwon, 2004; Lee 등, 2012; Yoon과 Moon, 2014). 특히 Lee 등 (2012)와 Yoon과 Moon (2014)은 강수의 극단값 분석을 하였는데, 전자는 서울 지역 최대 강수량 자료를 사용하였고, 후자는 기준 강수량을 초과하는 강수 자료를 사용하였다. 강수량과 같은 기후환경자료들은 지정학적 위치와 관련이 있으므로,

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(NRF-2012R1A1A3010532).

¹Corresponding author: Department of Statistics (Institute of Applied Statistics), Chonbuk National University, Jeonbuk 561-756, Korea. E-mail: hchoi@jbnu.ac.kr

공간적인 종속성을 반영하고 관측되지 않는 다른 위치에서의 분포 추정도 가능케 하는 공간 모형이 고려되어야 한다. 그러나 앞서 언급한 연구들은 자료의 공간 종속성을 고려하지 않았다. Jung 등 (2008)은 2007년 7월부터 9월까지의 우리나라 강수량 자료를 두 가지 공간통계 분석방법(가중최소제곱방법, 잔차최대우도추정방법)을 이용하여 공간적 특성을 알 수 있는 모형을 찾아 비교하였다. 그러나 이 연구에서는 극단값 모형을 적용하지 않았다.

최근 공간 극단값(spatial extremes) 모형의 이론적 접근과 이를 적용한 사례연구가 활발하게 진행되고 있다 (Davison 등, 2012; 그 안에 소개되어 있는 여러 논문). 공간 극단값 모형으로 계층적인 베이지안 모형, 코플라(copula)와 최대값 안정 확률장(max-stable random field) 등 여러 방법들이 사용되고 있다. 우리나라 여름철 강수량 자료 역시 여러 지역에 설치된 관측소에서 수집된 자료이므로 공간 종속성이 존재한다. 이러한 공간 종속성을 자료 분석에 반영하기 위해서는 위에 언급된 바와 같은 공간통계에 대한 전문지식과 다소 복잡한 적합 과정들이 필요하다. 그런데 관측된 지점에서의 주변분포 추정이 주요 관심이라면, 공간 상관 모형을 명확하게 세우는 대신에 먼저 자료가 공간적으로 독립이라는 가정 하에 극단값 모형을 추정한 후 모수 추정량이나 검정 통계량의 분포에 공간 종속성을 반영하도록 할 수 있다. Northrop과 Jonathan (2011)은 허리케인 발생 시 멕시코만의 파고 자료의 분계점 초과값들을 사용하여 극단값 분석을 하였는데, 먼저 공간적 독립 가정 하에서 모형을 추정하였다. 그리고 추정량의 표준오차를 구하는 과정에서 Chandler와 Bate (2007)가 제안한 방법으로 공간 종속성이 반영되도록 하였다. 분계점 초과값 분석에서 분계점의 결정은 중요한 문제이다. 극단값 자료의 공변량에 의존하는 분계점 결정에 관한 여러 논의가 있었으나 (Smith, 1989; Coles, 2001) 일반화에 어려움이 있었다. 공간 자료의 경우 공간적 위치가 영향을 주게 되어 자료가 대개 정상성을 만족하지 않게 되는데, Northrop과 Jonathan (2011)은 관측 지점의 지리 정보를 공변량으로 하는 분위수 회귀(quantile regression)를 통하여 분계점 결정에 보다 합리적으로 접근하였다.

본 논문에서는 기존 연구결과를 바탕으로 1973년부터 2012년까지의 기간 중 우리나라의 강수량이 집중되는 5월부터 9월의 주 최대 강수량 자료를 이용하여 분계점 초과값 분석을 실시하였다. 효율적인 분석을 위하여 기간 내의 관측소별 최대 강수량과 각 관측소의 지리 정보를 이용하여 먼저 군집분석을 실시하였다. 분계점은 Northrop과 Jonathan (2011)과 같이 분위수 회귀를 통하여 결정하였으며, 이 때 관측 지점의 지리 정보(경도, 위도)와 시간을 공변량으로 고려하였고, 각 군집별로 분위수 회귀 모형을 적합하여 결정하였다. 본 연구의 모든 계산과 그림은 R로 수행되었다.

본 논문의 구성은 다음과 같다. 2절에서 극단값 분계점 모형을 소개하고, 3절에서 분석 자료의 소개와 탐색적인 분석 결과, 자료의 분계점 모형 적합 결과를 제시하고 있으며, 마지막으로 4절은 요약 및 해석, 향후 연구 내용으로 이루어져 있다.

2. 극단값 분계점 모형

확률변수 Y_1, \dots, Y_n 가 서로 독립이고 동일한 분포를 따를 때, Y_1, \dots, Y_n 의 최대값 M_n 에 대하여 다음과 같은 극단값 정리가 성립한다고 잘 알려져 있다. 적당한 상수 $\{a_n\} > 0$ 과 $\{b_n\} \subset \mathbb{R}$ 가 존재해서 $(M_n - b_n)/a_n$ 의 분포함수가 비퇴화(nondegenerate) 분포 G 로 수렴한다면, G 는 $\{y : 1 + \xi(y - \mu)/\sigma > 0\}$ 에서 정의된 일반화 극단값(generalized extreme value) 분포함수

$$G(y) = \exp \left\{ - \left[1 + \xi \left(\frac{z - \mu}{\sigma} \right) \right]^{-\frac{1}{\xi}} \right\} \quad (2.1)$$

이어야 한다. 여기에서 $\xi = 0$ 일 때 $G(y)$ 는 $\xi \rightarrow 0$ 일 때의 값으로 한다. 이러한 일반화 극단값 분포함수를 $GEV(\mu, \sigma, \xi)$ 로 나타내기로 한다. 모수 (μ, σ, ξ) 의 범위는 $\mu \in \mathbb{R}$, $\sigma > 0$, $\xi \in \mathbb{R}$ 이다. ξ 의

부호에 따라 분포의 꼬리 성질, 즉, 얼마나 두꺼운지 얼마나 빠른 속도로 0으로 수렴하는지 등이 결정되어 이에 대한 적절한 추정이 극단값 이론에서는 매우 중요한 문제이다. $\xi > 0$ 인 GEV 분포는 우측 꼬리가 다항적(polynomially) 추세로 수렴하여 두껍고 긴 꼬리를 가지고 있다. $\xi = 0$ 인 경우는 지수적(exponentially) 추세로 0으로 수렴하며, $\xi < 0$ 이면 짧은 꼬리를 가지는 분포이다.

최대값 자료보다 더 많은 자료를 사용할 수 있어 통계적으로 더 효율적인 다음과 같은 분계점 초과값 모형이 많이 이용되고 있다. 먼저 Y 들의 분포함수 F 의 상한값을 $y_F = \sup\{y : F(y) < 1\}$ (y_F 는 무한할 수 있음)으로 나타내기로 하자. 최대값 M_n 이 $GEV(\mu, \sigma, \xi)$ 로 수렴한다면, 정해진 분계점 u 에 대한 Y 의 초과값의 분포

$$F_u(y) = \Pr\{x > u + y | x > u\} = \frac{1 - F(u + y)}{1 - F(u)}, \quad y > 0$$

는 $u \rightarrow y_F$ 일 때, 일반화 파레토분포(generalized Pareto distribution) $GPD(\tilde{\sigma}, \xi)$ (Davison과 Smith, 1990),

$$H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-\frac{1}{\xi}} \quad (2.2)$$

로 수렴한다. 여기에서 $\tilde{\sigma}$ 는 $GEV(\mu, \sigma, \xi)$ 의 $\sigma + \xi(u - \mu)$ 와 같다. 분계점 초과값 모형에서 분계점 u 의 설정은 중요한 문제이다.

다음으로 최대값과 분계점 초과값 모형을 포괄하는 점과정(point process) 모형이 있다. 최대값 M_n 이 $GEV(\mu, \sigma, \xi)$ 로 수렴하면, 2차원 점과정 $\{(i/(n+1), (Y_i - b_n)/a_n)\}_{i=1}^n$ 은 분계점으로 정의된 영역 $[t_1, t_2] \times [u, \infty)$, $0 < t_1 < t_2 < 1$ 에서 비동질적인 포아송 과정(non-homogeneous Poisson process)으로 수렴한다 (더 자세한 내용은 Coles (2001) 참조). 이 결과를 이용하면 최대값 모형보다 모수 (μ, σ, ξ) 의 더 정밀한 추정이 가능하다. Northrop과 Jonathan (2011)은 허리케인 발생 시 멕시코만 특정 지역의 파도 높이의 분포 모형을 추정하기 위하여, 점과정 모형을 적용하였다.

2.1. Northrop과 Jonathan의 공간 극단값의 분계점 모형

대부분의 기후환경자료는 공간적인 비정상성을 나타내고 있다. 정상성을 가지는 경우 극단값 모형의 모수 (μ, σ, ξ) 나 분계점을 상수로 가정하지만, 비정상성을 가지는 경우 공간적 요소를 (μ, σ, ξ) 나 u 의 공변량으로 반영할 수 있을 것이다. Northrop과 Jonathan (2011)은 관측 지점의 지리 정보를 공변량으로 하는 분위수 회귀(quantile regression)를 통하여 먼저 분계점을 결정한 후, 점과정 모형의 모수들도 지리 정보를 공변량으로하여 추정하는 방법을 제안하였다. 모형 추정 시에는 먼저 자료가 공간적으로 독립이라는 가정하였으며, 공간적 종속성을 반영하기 위하여 Chandler와 Bate (2007)에 의해 제안된 방법으로 추정오차를 구하였다. 공간 분계점 모형을 간략히 소개하면 다음과 같다.

먼저 공변량 x 에서의 반응변수를 $Y(x)$ 로, 분계점을 $u(x)$, $\Pr(Y(x) > y^\tau(x)) = \tau$ 를 만족하는 분위수를 $y^\tau(x)$ 로 나타내기로 한다. 주어진 τ 에 대하여 $y^\tau(x)$ 를 공변량 x 의 함수로 모형화 한 후, Koenker와 Bassett (1978)에 의해 제안된 점검 함수(check function) 방법으로 추정하였다. 추정된 $\hat{y}^\tau(x)$ 를 점과정 모형의 분계점 $\hat{u}(x)$ 으로 사용하였다. 즉, 공변량 x 가 주어졌을 때, 초과 확률(exceedance probability) $p(x) = \Pr(Y(x) > u(x))$ 가 상수 τ 가 되도록 분위수 회귀를 통하여 분계점 $u(x)$ 를 선택한 것이다. 이 때 적절한 τ 의 선택이 중요한데, 주관성이 개입되는 것을 줄이기 위하여 Northrop과 Jonathan (2011)은 τ 를 다양하게 변화시켜가며 추정한 분계점의 초과값들을 사용하여 점과정 모형을 적합시킨 후 모수 추정치들이 안정되기 시작하는 τ 에 해당하는 분계점을 선택하였다.

추정된 분계점으로 결정되는 점과정 모형의 가능도 함수를 살펴보면 다음과 같다. 먼저 장소 i 의 j 번째 관측된 공변량과 반응변수값을 x_{ij} 과 y_{ij} 로 나타내고, 그에 해당하는 분계점을 u_{ij} , GEV 모수를 $\theta = \{(\mu_{ij}, \sigma_{ij}, \xi_{ij})\}$, 조건부 확률밀도함수를 $f_{ij}(y_{ij}|x_{ij}, \theta_{ij})$ 라고 하자. Northrop과 Jonathan (2011)은 먼저 공간 지점별 자료가 독립이라는 가정하에서 점과정의 가능도 함수

$$L_I(\theta) = \prod_j \prod_i f_{ij}(y_{ij}|x_{ij}; \theta) \quad (2.3)$$

$$\approx \prod_j \prod_i \exp \left\{ -\frac{1}{\lambda} \left[1 + \xi \left(\frac{u_{ij} - \mu_{ij}}{\sigma_{ij}} \right) \right]^{-\frac{1}{\xi}} \right\} \times \prod_j \prod_{i: y_{ij} > u_{ij}} \frac{1}{\sigma(x_{ij})} \left[1 + \xi \left(\frac{y_{ij} - \mu_{ij}}{\sigma_{ij}} \right) \right]^{-\frac{1}{\xi-1}}$$

를 구하였다. 여기에서 λ 는 본 논문의 경우 연평균 관측횟수를 나타내며, $1 + \xi(u_{ij} - \mu_{ij})/\sigma_{ij} > 0$, $\sigma_{ij} > 0$ 이고 $\xi_{ij} = \xi$ 로 가정하였다. 공변량 x_{ij} 에서 초과 확률 $p(x_{ij})$ 의 다음과 같은 모수와의 관계

$$p(x_{ij}) \approx \frac{1}{\lambda} \left[1 + \xi \left(\frac{u_{ij} - \mu_{ij}}{\sigma_{ij}} \right) \right]^{-\frac{1}{\xi}} \quad (2.4)$$

가 성립한다. 그리고 $p(x_{ij}) = \tau$ 가 만족되는 u_{ij} 에 대하여 분위수 회귀를 실시한다면,

$$u_{ij} \approx \mu_{ij} + c\sigma_{ij}$$

와 같은 관계가 성립한다. 여기에서 $c = [(\lambda\tau)^{-\xi} - 1]/\xi$ 이다. 모수 σ 의 모형은 μ 와 동일하거나 저차원의 공변량 다항 함수 형태로 가정하고, μ 와 u 는 동일한 차수의 다항 함수 모형을 설정하였다. Northrop과 Jonathan (2011)은 각 지점별 최대값을 3차원 그림으로 표현하여, 대략적으로 파악한 μ 모형을 u 추정에 사용하였다. 그리고 추정된 u 를 분계점으로 하는 $L_I(\theta)$ 을 최대화 하는 해 $\hat{\theta}$ 를 구한 후, Chandler와 Bate (2007)가 제안한 방법으로 모수의 추정오차를 수정하였다.

3. 자료 분석

3.1. 자료 소개

본 연구에서는 기상청에서 제공한 1973년부터 2012년의 5월~9월의 일 강수량 자료를 주 최대 강수량으로 요약하여 사용하였다. 해당 기간의 강수량 자료를 모두 보유하고 있으며 섬 지역에 위치하지 않은 55개의 관측소만을 고려하였다. 즉, 자료는 5월부터 9월은 20주로 이루어져 있으므로 각 관측소별 매년 20개의 관측값들로 구성되어 있다. 모든 관측 지점의 지리 정보는 위도, 경도이고 강수량의 측정 단위는 mm이다. 우리나라의 복잡하고 다양한 지형적인 특성상 기상 사건이 전국적으로 발생하는 경우는 찾기 어려워 Park과 Kim (2008)이 계보적 군집 방법을 이용하여 북부(경기, 강원, 충북), 중부(경북, 전북, 충남), 남부(전남, 경남) 지역, 총 3개의 군집으로 나누어 우리나라 강수량 자료를 분석한 예가 있다. 본 논문에서는 위도, 경도, 해발, 해안 인접도, 최대 강수량 총 5개의 변수를 이용하여 최적 분리 군집 방법으로 Figure 3.1과 같이 55개의 관측소를 5개의 군집으로 구분하였다. Park과 Kim (2008)에서 고려한 변수와 관측소의 수 등이 본 논문과 다르다. 군집 1은 대체로 중서부 해안 지역, 군집 2는 중부 내륙 지역, 군집 3은 동해안 지역, 군집 4는 남부 내륙 지역, 군집 5는 남서해안 지역에 해당한다고 볼 수 있다. Figure 3.2에서는 군집별로 최대 강수량의 추세와 분포를 살펴보았다. 군집 1, 2와 4 지역은 기존의 연구결과와 같이 연 최대 강수량이 증가하는 경향이 있음을 확인한 반면 군집 3과 5 지역은 뚜렷한 경향을 보이지 않았고, 상자 그림을 통하여 극단값 분포의 꼬리 성질을 나타내는 ξ 가 양의 값을 짐작할 수 있었다. 다음으로 Figure 3.3과 같이 내륙에 위치한 금산 관측소와 인접한 대전 관측소, 금산 관측소와 멀

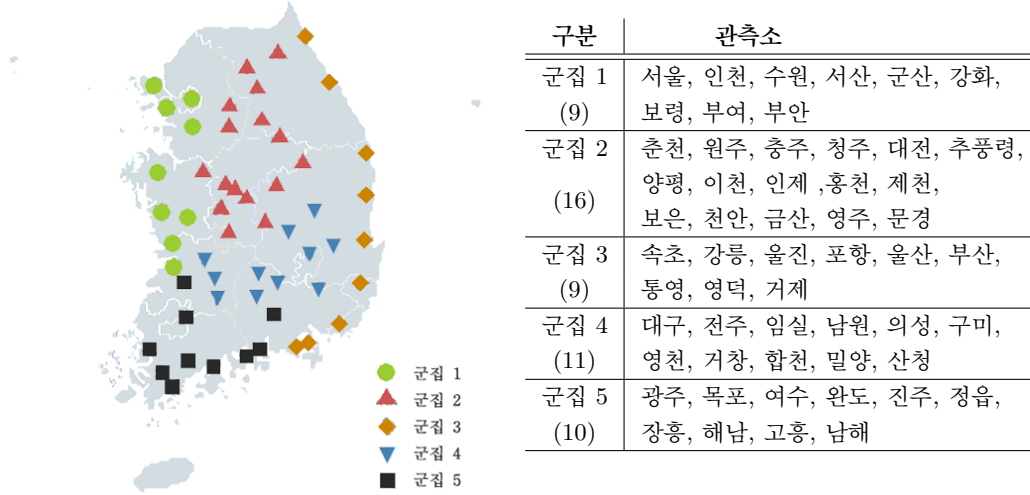


Figure 3.1. The result of cluster analysis

리 떨어져 있는 속초 관측소의 주 최대 강수량의 산점도와 상관계수를 살펴봄으로써 최대 강수량이 공간적인 연관이 있음을 확인할 수 있다. 금산 관측소의 주 최대 강수량은 멀리 떨어진 속초 관측소의 주 최대 강수량과 약한 양의 상관관계를 보인 반면, 인접 지역인 대전 관측소의 경우 매우 강한 양의 상관관계를 보이고 있다.

3.2. 공간적 극단값모형의 적합

위의 사전적인 탐색적 자료 분석을 통하여 한국의 (연도별, 주별) 최대 강수량이 관측소의 위치와 시간의 영향을 받음을 확인하였다. 따라서, 극단값 분포 모수 μ, σ 가 위도(l_x), 경도(l_y)와 시간(t)의 다항식, ξ 는 상수 모형으로 가지는 2.1절의 점과정 가능도 함수를 고려하고, 이 가능도 함수를 최대로 하는 모수 $\mu(l_x, l_y, t), \sigma(l_x, l_y, t), \xi$ 를 추정하였다. 모형 추정에 사용된 공변량의 값은 각각

$$\frac{l_x - (\max l_x + \min l_x)/2}{\max l_x - \min l_x}, \quad \frac{l_y - (\max l_y + \min l_y)/2}{\max l_y - \min l_y}, \quad \frac{t - (\max t + \min t)/2}{\max t - \min t}$$

로 변환된 값을 사용하였다.

Northrop과 Jonathan (2011)과 달리 시간도 공변량으로 사용하게 되어 자료의 시각화를 통한 μ 의 모형을 짐작하는데 어려움이 있었다. 본 논문에서는 각 관측소별 연 최대 강수량 자료를 l_x, l_y, t 의 다양한 차수 다항식의 μ 와 σ 에 대하여 $GEV(\mu, \sigma, \xi)$ 모형 적합한 후 그 결과를 이용하여 선택하였다. 지리 정보 (l_x, l_y)는 4차 다항 모형까지 시간 (t)은 2차 다항 모형까지 고려하여 로그 가능도를 계산하였다. 그리고, Chandler와 Bate (2007)가 제안한 방법으로 공간 종속성을 반영한 가능도비 검정 통계량을 이용하여 유의수준 0.05로 모형을 선택하였다. 연 최대 강수량 분포의 μ 는 군집 2, 3과 4에서 위도, 경도와 시간 모든 공변량과, 군집 1에서는 시간과, 군집 5에서는 위도, 경도와 선형 관계를 보여주고 있으며, σ 는 모든 군집에서 시간과는 선형 관계가 없음을 확인할 수 있었다. 따라서 분계점 u 의 분위수 회귀는 군집 2, 3과 4는 위도, 경도와 시간을 공변량으로 하는 선형 모형으로 군집 1은 시간의 선형 모형, 군집 5는 위도와 경도의 선형 모형으로 적합하였다. 분계점 초과 확률 τ 의 선택을 위하여 그 값을 변화 시

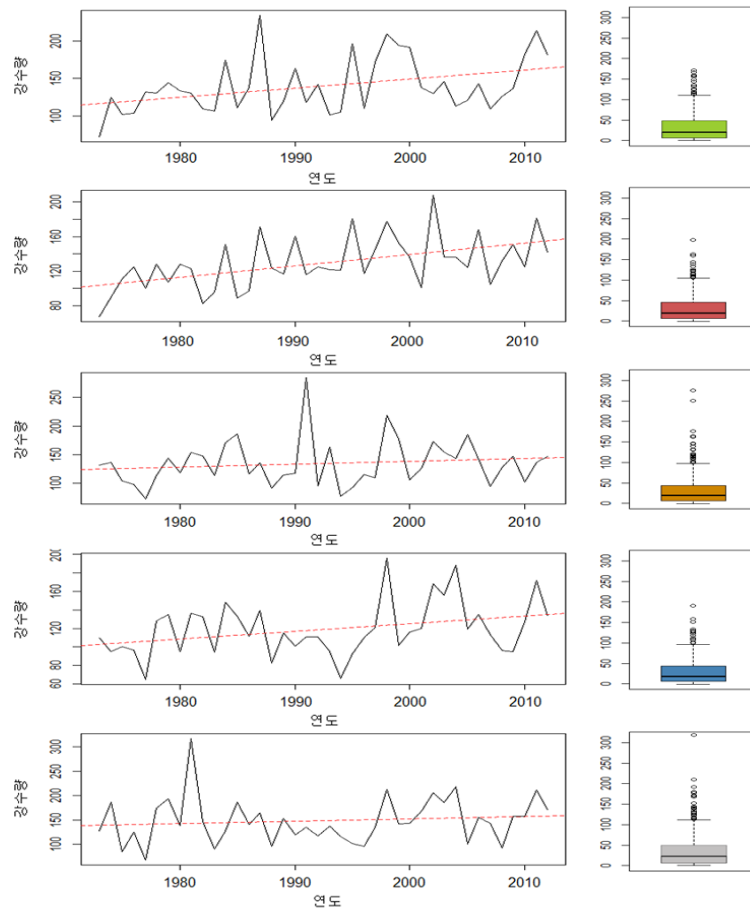


Figure 3.2. Left panel shows Time series plot of annual maximum rainfall for each cluster. Right panel shows Boxplot of weekly maximum rainfall for each cluster

켜가며, 각 τ -분위수 회귀분석을 실시하였다. 그리고 각 τ 에서 추정된 분계점을 초과하는 주 최대 강수량 분포 모수들을 Figure 3.4와 같이 그려보고 안정되기 시작하는 초과 확률 값을 선택하였다. 총 6개의 모수를 추정해야 하는 군집 3의 경우는 Figure 3.4에서 대략 $\tau = 0.38$ 부터 추정량이 안정화 되어가는 것을 볼 수 있다.

위와 같은 방법으로 모든 군집에서 초과 확률은 0.38로 선택되었다. Table 3.1은 선택된 초과 확률을 사용한 각 군집별 분위수 회귀 모형, 즉 분계점 선택 결과를 나타내고 있다. 추정된 분계점을 초과하는 값을 사용하여 각 군집별로 적합된 모형은 Table 3.2와 같다. 추정된 분계점 모형과 모수 μ 를 살펴보면, 남서해안 지역인 군집 5를 제외하고는 모든 관측 지역에서 시간에 따라 강수량의 증가 경향이 있음을 알 수 있다. 특히 남부 내륙 지역의 최대 강수량 증가 속도가 가장 빠르며, 중부 지역도 동해안 지역보다는 빠르게 증가하는 것으로 보인다. 위도가 높아질수록, 즉 북쪽으로 갈수록 전남, 남부 내륙과 동해안 지역은 최대 강수량이 작아지는 경향이 있는 반면, 북부 내륙 지역은 높아진다는 것을 알 수 있다. 경도가 높아질수록, 즉 동쪽으로 갈수록 내륙과 동부 지역은 최대 강수량이 작은 반면, 남서해안 지역, 즉 전남

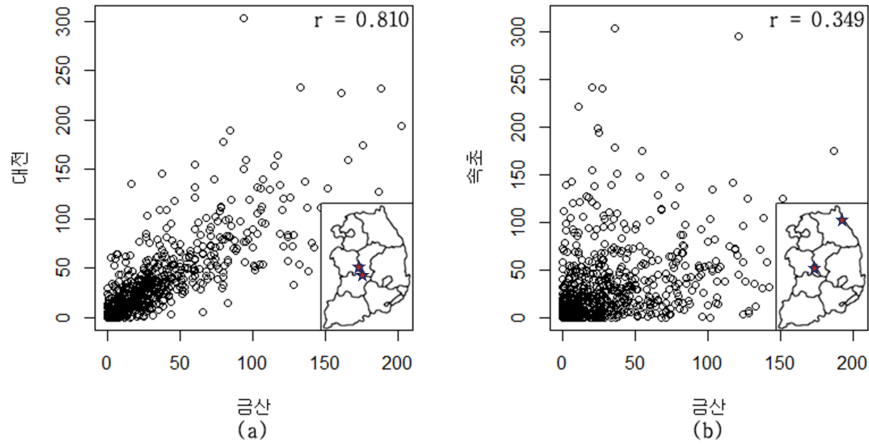


Figure 3.3. Scatter plot and correlation coefficient of weekly maximum rainfall at pairs of sites. ((a) contiguous sites (Keumsan and Daejeon), (b) distant sites (Keumsan and Sokcho))

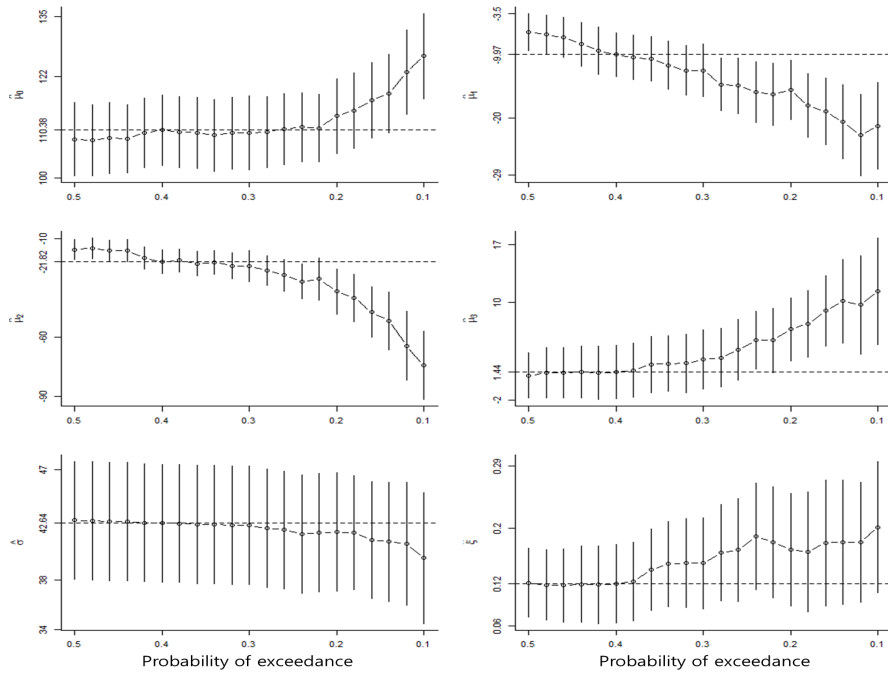


Figure 3.4. Maximum likelihood estimates against probability of exceedance for a point process model in Cluster 3. The vertical lines indicate approximate 95% confidence intervals, based on adjusted standard error. The horizontal dashed lines indicate the MLE for $\tau = 0.38$

지역은 동쪽이 더 높음을 알 수 있다. 서해안 지역은 관측지점의 위치의 영향을 받고 있지 않는 것으로 파악되었다. 모수 σ 는 우리나라 전 지역에서 시간의 영향은 받지 않는 것으로 추정되었고, 동서해안 지역을 제외한 내륙과 전남 지역은 경도와 위도와는 선형관계를 보이고 있다.

Table 3.1. Probability of exceedance and fitted quantile regression model of each cluster

군집	초과확률 τ	분계점 (\hat{u})
1	0.38	$27.675 + 3.142t$
2	0.38	$26.50 + 3.290l_x - 1.998l_y + 3.375t$
3	0.38	$34.476 - 11.018l_x - 23.935l_y + 2.184t$
4	0.38	$23.447 - 24.957l_x - 6.674l_y + 3.774t$
5	0.38	$33.185 - 3.835l_x + 18.946l_y$

Table 3.2. Maximum likelihood estimates and adjusted standard errors (in parentheses) for a point process model of each cluster.

군집	$\hat{\mu}$				$\hat{\sigma}$			$\hat{\xi}$
	상수	t	l_x	l_y	상수	l_x	l_y	상수
1	108.221 (1.874)	2.657 (1.174)			43.310 (1.322)			0.087 (0.020)
2	96.635 (1.977)	2.984 (0.858)	31.889 (8.909)	-13.087 (10.441)	35.904 (1.102)	14.756 (4.079)	-6.379 (4.846)	0.041 (0.014)
3	109.976 (2.626)	1.635 (1.064)	-10.397 (2.075)	-21.106 (5.135)	42.623 (1.350)			0.123 (0.020)
4	84.794 (3.432)	4.174 (0.962)	-73.076 (19.582)	-10.373 (7.978)	30.719 (1.758)	-25.032 (8.822)	-0.850 (3.622)	0.023 (0.017)
5	101.314 (7.569)		-59.260 (19.348)	44.882 (10.272)	36.883 (3.804)	-29.294 (9.387)	13.769 (4.974)	0.087 (0.019)

모수 ξ 는 앞 절에서 상자 그림을 통하여 짐작한 바와 같이 모든 군집에서 양의 값으로 추정이 되었다. 군집 4인 남부 내륙 지역에서 ξ 의 95% 신뢰구간이 $(-0.010, 0.056)$ 로 0의 값을 포함하고 있는 것을 제외하고는 다른 모든 군집에서 ξ 의 95% 신뢰구간이 0을 포함하고 있지 않는 것으로 나타났다. 즉, 전체적으로 우리나라 모든 지역에서 최대 강수량이 긴 꼬리 분포 형태임을 확인할 수 있었다. 특히 동해안 지역의 ξ 추정량은 상대적으로 매우 크다. 이 지역은 홍수와 같은 피해의 발생 가능성이 다른 곳에 비해 높을 수 있다.

본 논문에서는 모형의 적합도를 파악하기 위해 각 관측소별로 잔차확률도표(residual probability plot)를 그려 보았다. Figure 3.5는 각 군집 별로 두 곳의 잔차확률도표를 보여주고 있는데, 전반적으로 잘 적합되었음을 확인하였다.

4. 요약 및 결론

본 논문에서는 기상청에서 제공하는 일 강수 자료를 주 최대 강수량으로 요약하여 극단값 분석을 실시하였다. 우리나라의 복잡하고 다양한 지형적인 특성상 강수 사건이 전국적으로 발생하는 경우는 드물기 때문에 최적 분리 군집 방법을 통하여 5개의 군집으로 세분화한 후에 분석하였다. 자료가 공간적으로 독립이라는 가정 하에 극단값 모형을 추정된 후 모수 추정량이나 검정 통계량의 분포에 공간 종속성을 반영하였다. 따라서, 공간 상관 모형을 정확히 세우지 않고 추정된 결과를 관측되지 않은 지점까지 포함하여 해석할 수 없다는 한계가 있다. 그러나, 분석에 포함된 관측 지점의 수가 55개이고, 비교적 면적이 작은 나라인 우리나라 전반에 고루 흩어져 있으므로, 추정된 각 모형을 바탕으로 일반화하여 해석하면 다음과 같다.

각 관측 지점의 분위수 회귀를 통해 정해진 분계점으로 적합한 군집별 분계점 모형에서 모수 μ 의 모형

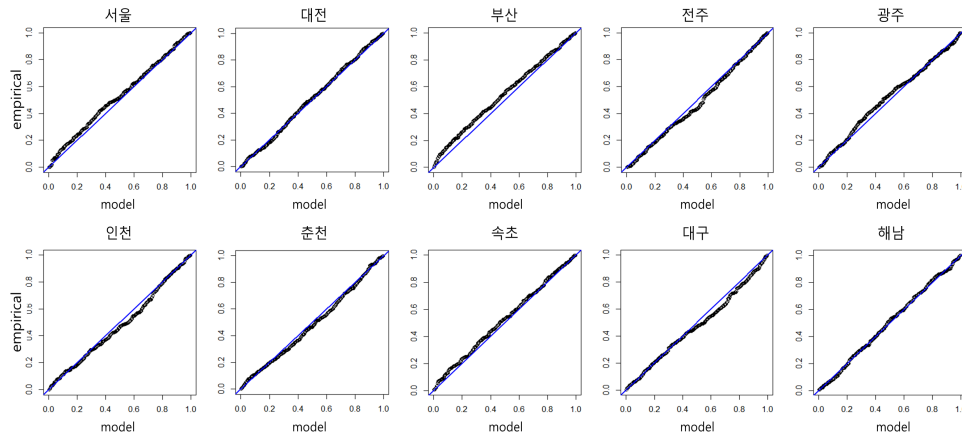


Figure 3.5. Residual probability plots of fitted models at two weather stations for each cluster

을 살펴보면, 군집 5를 제외한 모든 군집에서 시간에 따른 강수량의 증가 경향을 확인할 수 있었으며, 특히 남부 내륙 지역의 최대 강수량 증가 속도가 가장 빠르며, 중부 지역도 동해안 지역보다는 빠르게 증가하는 것으로 보인다. 그런데 각 관측소별로 분위수 초과 강수량의 시간과의 관계를 살펴본 Yoon과 Moon (2014)은 광주 관측소의 경우 선형 관계가 있음을 확인하고 있다. 위도가 높아질수록 전남, 남부 내륙과 동해안 지역은 최대 강수량이 작아지는 경향을 확인하였으며, 북부 내륙 지역 최대 강수량이 높아지는 것을 확인하였다. 경도가 높아질수록 내륙과 동해안 지역은 최대 강수량이 작아지며, 전남 지역은 증가함을 알 수 있다. 서해안 지역은 관측 지점의 위치의 영향을 받고 있지 않는 것으로 파악되었다. 모수 σ 는 우리나라 전 지역에서 시간의 영향은 받지 않는 것으로 추정되었고, 동서해안 지역을 제외한 내륙과 전남 지역은 경도와 위도와는 선형 관계를 보이고 있다. 모수 ξ 는 모든 군집에서 양의 값으로 추정되어 우리나라 대부분의 지역에서 주 최대 강수량이 긴 꼬리 분포 형태임을 확인할 수 있었다. 그중 특히 동해안 지역으로 구분된 군집 3지역의 ξ 값이 다른 지역에 비해 큰데 이 지역은 홍수와 같은 피해의 발생 가능성이 다른 지역에 비해 높을 것을 의미한다. 추정된 모형의 적합성을 잔차 확률 도표로 확인하였을 때 전반적으로 잘 적합된 모형이라는 결과를 얻었다.

References

- Chandler, R. E. and Bate, S. B. (2007). Inference for clustered data using the independence log-likelihood, *Biometrika*, **94**, 167–183.
- Choi, Y. E. (2004). Trends on temperature and precipitation extreme events in Korea, *Journal of the Korean Geographical Society*, **39**, 711–721.
- Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*, Springer, London.
- Davison, A. C., Padoan, S. A. and Ribatet, M. (2012). Statistical modeling of spatial extremes, *Statistical Science*, **27**, 161–186.
- Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds (with discussion), *Journal of the Royal Statistical Society Series B*, **52**, 393–442.
- Jung, J. Y., Jin, S. H. and Park, M. S. (2008). Precipitation analysis based on spatial linear regression model, *The Korean Journal of Applied Statistics*, **21**, 1096–1107.
- Kim, S. Y., Kwon, Y. A., Park, C. Y. and Chun, J. H. (2008). Regional characteristics of the monthly variability of precipitation from May to September over South Korea, *Journal of Climate Research*, **2**,

64–75.

- Koenker, R. and Bassett, G. Jr. (1978). Regression quantiles, *Econometrica*, **46**, 33–50.
- Lee, K. M., Baek, H. J. and Cho, C. H. (2012). Analysis of changes in extreme precipitation in Seoul using quantile regression, *Journal of Climate Research*, **3**, 199–209.
- Lee, S. H. and Kwon, W. T. (2004). A variation of summer rainfall in Korea, *Korean Geographical Society*, **39**, 819–832.
- Northrop, P. J. and Jonathan, P. (2011). Threshold modelling of spatially dependent non-stationary extremes, *Environmetrics*, **22**, 799–809.
- Park, M. S. and Kim, H. Y. (2008). Classification of precipitation data based on smoothed periodogram, *the Korean Journal of Applied Statistics*, **21**, 547–560.
- Smith R. L. (1989). Extreme value analysis of environmental time series : An application to trend detection in ground-level ozone, *Statistical Science*, **4**, 367–377.
- Yoon, S. K. and Moon, Y. I. (2014). The Recent increasing trends of exceedance rainfall thresholds over the Korean major cities, *Journal of the Korean Society of Civil Engineers*, **34**, 117–133.

공간 극단값의 분계점 모형 사례 연구 - 한국 여름철 강수량

황승용^a · 최혜미^{a,1}

^a전북대학교 통계학과

(2014년 6월 11일 접수, 2014년 8월 6일 수정, 2014년 8월 6일 채택)

요약

폭염, 폭우와 가뭄 등과 같은 이상 기후 현상에 대한 적절한 대응이 최근 많이 요구되고 있다. 이상 기후 현상을 분석하기 위해 극단값 분석 기법을 적용할 수 있는데, 본 논문에서는 한국의 여름철 강수량 자료(1973년부터 2012년까지의 5월부터 9월)를 분계점 초과값 모형으로 분석해보았다. 분계점은 한국의 기상관측소들을 5개의 군집으로 나누어, 각 군집별로 지리 정보와 시간을 공변량으로 하는 분위수 회귀 방법을 통하여 추정하였다. Northrop과 Jonathan (2011)과 같이 극단값들이 시공간적으로 독립이라고 가정하고 분석한 후, 추정오차와 검정 과정에 공간 종속성을 반영하였다.

주요용어: 극단값 모형, 분위수 회귀, 분계점 방법.

이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임 (NRF-2012R1A1A3010532).

¹교신저자: (561-756) 전북 전주시 덕진구 백제대로 567, 전북대학교 통계학과(응용통계연구소).

E-mail: hchoi@jbnu.ac.kr