Journal of Hydrology 538 (2016) 831–841

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Abrupt change point detection of annual maximum precipitation using fused lasso

Jong-June Jeon [a], Jang Hyun Sung [b], Eun-Sung Chung [c,*]

[a] Department of Statistics, University of Seoul, Seoul, Republic of Korea
[b] Ministry of Land, Infrastructure and Transport, Geum River Flood Control Office, Gongju-si, Chungcheongnam-do, Republic of Korea
[c] Department of Civil Engineering, Seoul National University of Science & Technology, Seoul 01811, Republic of Korea

ARTICLE INFO

Article history:
Received 4 March 2016
Received in revised form 18 April 2016
Accepted 20 April 2016
Available online 9 May 2016
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Sheng Yue, Associate Editor

Keywords:
Annual maximum precipitation
Bayesian change point analysis
Fused lasso penalty
GEV distribution
Skewness

SUMMARY

Because the widely used Bayesian change point analysis (BCPA) is generally applied to the normal distribution, it cannot be freely used to the annual maximum precipitations (AMP) in South Korea. Therefore, this study proposed the fused lasso penalty function to detect the change point of AMP which can be generally fitted by using the Generalized Extreme Value (GEV) distribution in South Korea. First, four numerical experiments are conducted to compare the detection performances between BCPA and fused lasso method. As a result, fused lasso shows the superiority of the data generated by GEV distribution having skewness. The fused lasso method is applied to 63 weather stations in South Korea and then 17 stations having any change points from BCPA and the GEV fused lasso are analyzed. Similar to the numerical analyses, the GEV fused lasso method can delicately detect the change point of AMPs. After the change point, the means of AMPs did not go back to the previous. Alternately, BCPA can be stated to find variation points not change points because the means returned to their original values as time progressed. Therefore, it can be concluded that the GEV fused lasso method detects the change points of non-stationary AMPs of South Korea. This study can be extended to more extreme distributions for various meteorological variables.

1. Introduction

The temporal variability of precipitation affects human societies throughout the world. Therefore, it is important to obtain definitive knowledge of the changes in the magnitude and frequency of precipitation, especially for extreme events (Durrans and Kirby, 2004). The magnitude and frequency of extreme rainfall events have increased in several areas, such as Japan, India, Australia (Iwashima and Yamamoto, 1993), China (Jiang et al., 2014), Senegal (Sarr et al., 2013), UAE (Ouarda et al., 2014) and South Korea (Jung et al., 2010; Sung et al., 2015). Moreover, climate change is currently affecting precipitation patterns throughout the world because higher average air temperatures result in higher evaporation rates, higher water vapor contents and consequently an accelerated hydrologic cycle (Menzel and Burger, 2002).

Many quantitative and non-stationary studies on temperature and precipitation have recently been conducted to provide clear evidences of climate change (Choi, 2004; Santos and Leite, 2009; Choi and Park, 2010). In addition, trends in hydro-meteorological data have been observed using various statistical methods (Gleick, 1989; Lettenmaier and Gan, 1990; Lettenmaier et al., 1994; McCabe and Hay, 1995; Lins and Slack, 1999; Douglas et al., 2000; Wong et al., 2006; O'Brien and Burn, 2014; He et al., 2015).

The trend analysis, one of the commonly used methods to detect nonstationarity, is based on the linear regression model (Visser and Molenaar, 1995; Haktanir and Citakglu, 2014). Because the regression model uses time-dependent covariates as predictors and projects the response variables via estimated regression coefficients, the model can incorporate many possible factors, including both climatological and geomorphological ones, for the analysis of the nonstationarity (Jaiswal et al., 2015).

Due to the flexibility of linear regression model, numerous statistical methods have been developed. For example, Strupczewski et al. (2001a, 2001b) and Strupczewski and Zdzisław (2001) proposed several regression models using Maximum Likelihood (ML) method. Katz et al. (2002) and Clarke (2002a, 2002b, 2002c) applied a linear regression model in the Generalized Extreme Value (GEV) distribution. Khaliq et al. (2006) gave several fruitful reviews of frequency analyses for nonstationary hydro-meteorological observations. However, it can be difficult to use the trend analysis

* Corresponding author.
E-mail address: eschung@seoultech.ac.kr (E.-S. Chung).

http://dx.doi.org/10.1016/j.jhydrol.2016.04.043
0022-1694/

stemming from the regression model to explain or detect an abrupt change, or multiple abrupt changes, caused by site-specific components or induced by climate change, because the regression model typically assumes a smooth model (eg. a linear model). Villarini et al. (2009) developed the nonparametric regression models, called of generalized additive models for location, scale and shape parameters (GAMLSS), for flood frequency analysis.

Abrupt changes are usually represented by change points, which are moments of discontinuities in a stochastic process. In hydrology and climatology, such abrupt changes have been empirically detected (e.g. Chung et al., 2011), and Bayesian Change Point Analysis (BCPA) is widely used (e.g. Sung et al., 2015). Under an assumed model, BCPA provides the posterior distribution of the occurrence of change points such that location and uncertainty of change points are obtained simultaneously. In addition, through the posterior distribution of parameters in the model, various inferences of random quantity (eg. return level) are available.

In spite of their usefulness, most methods for BCPA are studied within the normal distribution or Poisson distribution. In general, BCPA is restricted to the pre-assumed probability model or likelihood. If the likelihood is not fitted well, the result of BCPA can be unreliable. BCPA should be carefully applied to estimate the high return levels especially when the true underlying skewed distribution has thicker tail than normal distribution because BCPA can be useful to the normal distributions whose skewness and kurtosis are given zero. Hence a tail behavior of a heavy tailed distribution is not explained by the normal distribution. In this reason, most of probability model in extreme data analysis including Bayesian data analysis assume the heavy tail distribution for likelihood (Coles and Tawn, 1996; Stephenson and Tawn, 2004; Katz et al., 2002). That is, the likelihood function used in BCPA is severely misspecified such that BCPA can fail to obtain reliable return level estimates of extreme events in hydrological process. However, it is unknown whether BCPA can estimate means and detect change points in skewed and heavy tailed distribution.

Therefore, this paper focuses on the abrupt change detection of mean-estimates of a skewed distribution with multiple change points and claims that the BCPA also fails to detect change point and estimate means. In addition this study proposes a fused lasso method as an alternative to BCPA for extreme value distributions having positive or negative skewness. The fused lasso, first proposed by Tibshirani et al. (2005), is a regression model with a penalty function which can detect change points of the means of the normally distributed data. Furthermore, the fused lasso can detect multiple change points simultaneously in the observed data and can give a theoretically better estimator under regularity conditions (Rinaldo, 2009). Therefore, this study extends this fused lasso method to GEV distribution and shows the usefulness of the fused lasso through several numerical simulations. This study also applies the fused lasso to 6-h, 12-h, and 24-h annual maximum precipitations (AMPs) of the 63 weather stations in South Korea.

2. Methodology

The GEV distribution was found to be useful to fit the frequency of AMPs of South Korea (Park et al., 2011). Therefore, this study developed the GEV fused lasso method (described in Section 2.1) to find change points of the AMPs series of 63 weather stations in South Korea (see Fig. 1) to compare their performances for 6 h, 12 h, and 24 h AMPs. This study also compared two results from the fused lasso and BCPA (described in Section 2.2). 17 weather stations are selected, shown in Table 3, because they show any linear trends from MK trend analysis, or any change points from BCPA and fused lasso method.

2.1. Fused lasso method

2.1.1. GEV distribution with fused lasso penalty function

First, the ML method with constraints of parameters corresponding to the change point analysis is described in this section. For simplicity, a class of three parameter distributions with location (μ), scale (α) and shape (κ) is considered. This class of distribution is sufficiently large to cover most probability models used in hydrology. The normal distribution belongs to a class of distributions with location and scale parameters; the Log-Pearson type III (LP3), the GEV and the Generalized Pareto (GP) distributions belong to a class of distributions with three parameters related to their location, scale and shape. In these distributions, $f(\cdot;\mu,\alpha,\kappa)$ represents the Probability Density Function (PDF). Assume that the nonstationary model has a location parameter depending on time t (μ_t). Under the assumption that the observations x_t for $t = 1,\ldots, n$ are independently distributed, the likelihood function is defined by

$$L(\boldsymbol{\mu},\alpha,\kappa) = \prod_{t=1}^{n} f(x_t;\mu_t,\alpha,\kappa) \tag{1}$$

with $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_n)$. The maximizer of Eq. (1) with the constraint on $\mu_1 = \mu_2 = \cdots = \mu_n$ is equal to the Maximum Likelihood Estimator (MLE) in the stationary model, which is generally known as an efficient estimator under some regularity conditions (Smith, 1985). With a different constraint on μ_t for $t = 1,\ldots,n$, we obtain other estimators by maximizing Eq. (1). Let μ_1^* and μ_2^* be unknown true location parameters ($\mu_2^* \neq \mu_1^*$). Suppose that s is a change point such that $\mu_t = \mu_1^*$ for $1 \leqslant t \leqslant s-1$ and $\mu_t = \mu_2^*$ for $s \leqslant t \leqslant n$, then the MLE of $(\boldsymbol{\mu},\alpha,\kappa)$ is given by

$$\begin{aligned} (\hat{\boldsymbol{\mu}},\hat{\alpha},\hat{\kappa}) &= \underset{\mu,\alpha,\kappa}{\arg\max}\; L(\boldsymbol{\mu},\alpha,\kappa) \\ \text{subject to } \mu_1 &= \cdots = \mu_s \\ \mu_{s+1} &= \cdots = \mu_n \end{aligned} \tag{2}$$

If μ_s is the unique change point, the conventional MLE ignoring the change point cannot capture the changes of the mean such that the performance of estimation may become poor. Without considering the change point, the naïve estimator can severely underestimate or overestimate forecasting (e.g., return level). Practically, the statistical test is recommended to detect or monitor the changes. CUSUM test is one of the most popular methods to detect a change point (Sonali and Kumar, 2013; Whitcher et al., 2002). The CUSUM test was originally developed for statistical quality control in production process (Page, 1954). Using the test, we can estimate the parameters via two steps: (1) detection of change points and (2) estimation of parameters with reflecting the change of probability models. For example, we estimate the change point $\hat{s}$ by the CUSUM test and obtain the final estimator of $\boldsymbol{\mu}$ by maximizing (2). As an alternative method, we can consider a regularized regression method, which has been widely and deeply studied in fields of statistics and computer science.

The distribution of the extreme event with maximum value is expected to have the max-stability. It is well known that a distribution is max-stable if and only if it is the GEV distribution. The max-stability is a property satisfied by distribution for which the operation of taking sample maxima leads to an identical distribution, apart from a change of scale and location (Coles, 2001). In this reason, the GEV distribution has been widely used with the annual maximum value.

Here, a fused lasso method is used to estimate change points and parameters of the GEV distribution. The regularized empirical risk function is written as follows:

$$R_\lambda(\boldsymbol{\mu},\alpha,\kappa) = -\log L(\boldsymbol{\mu},\alpha,\kappa) + \lambda\sum_{t=2}^{n}|\mu_t - \mu_{t-1}| \quad (\lambda \geqslant 0) \tag{3}$$