Trend and Return Levels for Chicago Snowfall

Patrick Toman*

15 December 2020

Introduction

In the first 9 months of 2020, there have been 16 extreme weather or climate related events with monetary losses exceeding $1 billion dollars in the united states, tying the annual records set by 2011 and 2017 Smith (2020). One of the most disruptive events, particularly for urbanized areas, are extreme snow events. One question at the forefront of researchers and planners minds is whether or not the frequency and intensity of extreme winter storms is increasing. To that end, Lee and Lee proposed methods based on *generalized extreme value distributions* for modeling trend and return levels for extreme snow events in New York City using 56 years of annual snowfall data Lee and Lee (2020). Indeed, the authors found that their modeling framework was useful in modeling explaining extreme snow events in the city. Given Lee and Lee's report success, an obvious line of inquiry would be to assess if this model can be applied to a different geo-spatial location. With this in mind, this report takes the methods from Lee and Lee's paper and applied them to annual snowfall data for Chicago,Illinois, another city that is prone to crippling blizzards.

Data

The data used in this report are downloaded from National Centers for Environmental Information, a part of NOAA, using their Climate Data Online tool Environmental Information (NCEI) (2020). Three weather stations are selected from the Chicago Area: Midway, O'Hare, and Park Forest. Table 1, gives a geographic description of the three stations. The last 61 years of daily snowfall data from each of the four station is used in the analysis, starting from the dates July 1st,1960 to June 30th, 2020. Daily snowfall is defined as the maximum amount of that has accumulated prior to melting or settling for the day. The snowfall data from NCEI is measured in inches with the amounts being rounded to the nearest tenth of an inch with amounts less than 0.1 being recored as zeros. Furthermore, $\approx 8\%$ of the days in our analysis have non-trace snow amounts, we call these *snow events* in our analys. Since major snowfall events are tend to last multiple days, consecutive days with non-zero snowfall are merged to represent one single snow event in the data. In addition, snowfall observation refers to the accumulated snowfall associated with a given storm. Finally, a *snow year* is defined as a one-year period that starts on July 1st to June 30th. For instance, the snow records from July 1st 2012 and June 30th, 2013 would correspond to the 2012 snow year. Note that there are missing observations in the data. For Midway, $< 0.8\%$ of daily observations are missing, O'Hare is missing $\approx 0.9\%$ of observations and $\approx 8\%$ are missing for Park Forest. Most of the missingness appears to be concentrated in non-winter months, therefore, the issue of missing data is negligible.

*patrick.toman@uconn.edu; Ph.D. student at Department of Statistics, University of Connecticut.

Table 1: Station Summaries

Station	Full Station Name	Latitude	Longitude	Elevation
Chicago	CHICAGO MIDWAY AIRPORT 3 SW, IL US	41.74	-87.78	189.00
Park Forest	PARK FOREST, IL US	41.49	-87.68	216.40
O'Hare	CHICAGO OHARE INTERNATIONAL AIRPORT, IL US	41.96	-87.93	201.80

Methods

Block Maxima Methods for GEV

Let $X_1, \dots, X_k$ be independent and identically distributed IID random variables that have the common CDF $F(.)$. Next, let us define $M^{(k)} = max\{X_1, \dots, X_k\}$ as the maximum order statistic for a *block* of k these random variables. Suppose then that there are a set of constants $\{a_k\}$ and $\{b_k\}$ with $b_k > 0 \; \forall k$ so that scale $M^{(k)}$ such that

$$P\left(\frac{M^{(k)} - a_k}{b_k} \leq x\right) \rightarrow G(x) \text{ for } k \rightarrow \infty \tag{1}$$

where $G(.)$ is a non-degenerate distribution function. Then according to the Fisher-Tippett-Gnedenko theorem, $G(.)$ follows one of these distribution families: Gumbel,Frechet, or Weibull Lee and Lee (2020). Furthermore, these families can be further generalized using the *generalized extreme value* (GEV) distribution function

$$G(x) = exp\left\{-\left[1 + \xi(\frac{x-\mu}{\sigma})\right]_+^{\frac{-1}{\xi}}\right\} \tag{2}$$

where $z_+ = max\{z, 0\}, \mu \in \mathcal{R}, \sigma \in \mathcal{R}^+,$ and $\xi \in \mathcal{R}^+$. In application settings, block maxima methods rely on a sequence of of maximum order statistics from a CDF $F(.)$. If the block sizes are large enough, then the GEV theorem states that In our case, each snow year is treat as a block and then we extract the maximum snowfall statistic from each snowfall year for each station as our extreme event for that year.

Maximum Likelihood for GEV Models

Assume that we have $\xi \neq 0$. If $X_1, \dots, X_N \overset{iid}{\sim} GEV(\Theta)$ where $\Theta = (\mu, \sigma, \xi), -\infty < \mu, \xi < \infty, \sigma > 0$ is our vector of unknown parameters, then we have the log-likelihood for for $\mathbf{X} = \{X_1, \dots, X_N\}$ as

$$ln\left(\Theta|\mathbf{X}\right) = -Nln(\sigma) + \left(\frac{1}{\xi} + 1\right)\sum_{i=1}^{N}\left(1\xi\left(\frac{x_i - \mu}{\sigma}\right)\right) - \sum_{i=1}^{N}\left(1 + \xi\left(\frac{x_i - \mu}{\sigma}\right)\right) \tag{3}$$

assuming that $1 + \xi\left(\frac{x_i - \mu}{\sigma}\right) > 0$ cite GEV book here. Indeed, parameter estimates can be obtained by numerical estimation methods such as Newton-Raphson. Further information can be found in cite Hosking or Macleod Here. Expressions for the gradient can be found the appendix possibly remove. In this project, we use the standard *optim()* function found in the R programming language.

Return Levels

Some of the most important quantities in any extreme values analysis are the return levels. The return level for an associated return period of K years is the expected level that is to be exceeded on average once over the following K years. In the case of annual maximums we have the return level X_k for a period of K years defined as

$$X_k = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - \{-ln(1-K^{-1})\}^{-\xi}\right] \text{ if } \xi \neq 0 \\ \mu - \sigma\left[-ln(1-K^{-1})\right] \text{ if } \xi = 0 \end{cases} \tag{4}$$

Invoking the invariance property of MLEs, we can solve for the MLE of the return level $\hat{X}_k$ by simply substituting the MLEs for (μ, σ, ξ) attained through Newton-Raphson.

Smith's Method for Spatial and Temporal Dependence

Snowfall data collected from stations close in proximity to one another will inevtiably exhibit spatial correlation. To account for this spatial correlation in our standard errors, we can employ a method devised by Smith cite Smith where MLE estimates are attained under the standard IID assumption and then standard errors are corrected in a post-hoc fashion to account for spatial dependence. write in Smith's method for our context?. Furthermore, for block maxima methods, temporal depdence is typically not an issue since extreme snowfall events tend to be separated by a large gap of time. Indeed, we find that auto-correlations between consecutive years for all three stations are quite small put in results here.

Bootstrap CI for Return Levels

For GEV models, asymptotic standard errors attained via the delta method are often inadequate in estimating the sampling variability of MLE estimators for return levels cite here. Lee and Lee @trend_return_levels propose bootstrapping the confidence intervals for return levels using the standard percentile method where we use the $\alpha/2$ upper and lower quantiles of the bootstrapped return level estimates. Because the distribution of return levels tends to exhibit right-skew, therefore, standard bootstrap methods will be biased. To remedy this bias in bootstrap return level estimates, we can employ the *bias corrected and accelerated* (BCa) bootstrap methods introduced by cite Efron here. Suppose that we wish to generate B bootsrap estimates of return level X_k. The BCa method works by first calculating the bias correction estimate z_{BC}

$$z_bc = \Phi^{-1}\left(\frac{1}{B}\sum_{b=1}^{B} I\left(\hat{x}_K^{(b)} < \hat{x}_K\right)\right) \tag{5}$$

where $\Phi(.)$ is the standard normal CDF, $I(.)$ is the indicator function, and $\hat{x}_k$ denotes the MLE of the return level for period K using the original data. Next, we calculate the acceleration constant c_A which is defined as

$$c_A = \frac{\sum_{t=1}^{n}\left(\tilde{x}_K^{-t} - \tilde{x}_K\right)^3}{6\left[\sum_{t=1}^{n}\left(\tilde{x}_K^{(-t)} - \tilde{x}_K\right)^2\right]^{3/2}} \tag{6}$$

where $\tilde{x}_K^{(-t)}$ denotes the delete-1 jacknife estimate of x_K where we have deleted the t^{th} observation from the dataset and $\tilde{x}_K^{(-t)} = \sum_{t=1}^{n} \tilde{x}_K^{(-t)}$ cite givens and hoeting. Thus, the $(1-\alpha) * 100\%$ BCa is interval has the following quantiles as its upper (lower) endpoints given by

$$\Phi\left(z_{BC} + \frac{z_{BC} \pm z_{\alpha/2}}{1 - c_A(z_{BC} \pm z_{\alpha_2})}\right) \tag{7}$$

Model Fitting and Results

Let N_s denote the number of snowstorms that occured at station s during the study period from July 1st, 1960 to June 30th, 2020. If we let $\mathbf{X} = \{X_{s,1}, \ldots, X_{s,N_s}\}$ represent the the accumulated snowfall observed at station s where $X_{s,j},\ j \in 1, \ldots, N_S$ denotes the accumulated snowfall for snowstorm j during the study period. Next, if we denote $M_{s,t}$ denote the maximum for each snow year $t \in \{1960, \ldots, 2020\}$, then we have $M_{s,t} \sim GEV(\mu_s, \sigma_s, \xi_s)$. Using the block maxima methods described in the Methods section, we fit several different models and compare them using Akaike's Corrected Information Criterion (AICc) where $AICc = 2\left(p - ll(\hat{\Theta}) + \frac{p(p+1)}{n-p+1}\right)$ where p is the number of parameters, $ll(\hat{\Theta})$ denotes the value of the log-likelihood evaluated at the MLE, and n is the number of observations. Models with smaller AICc are preferred.

Initially, we build a full model in which we assume that each station has its own location, shape, and scale parameters. In other words, we have $\Theta_s = (\mu_s, \sigma_s, \xi_s), s = 1, 2, 3$. The AICc of this model is -884.1794. Next, we build a model where μ_s is fit individually for each station but σ and ξ are assumed to be the same for all three locations. For the reduced model, we find that the AICc is -895.6098. Therefore, we prefer the reduced model where μ_s's are assumed to be different for the stations but σ and ξ are assumed common.

Stationary Models

Having established that σ and ξ should be common to all three stations, three different stationary models are fit to the data in which we vary μ_s by iteratively merging the two closest μ_s's. Because we anticipate spatial dependence in the data, we correct the standard errors of each using Smith's method. put in a table of results

Stationary Models

Conclusions and Further Work

Environmental Information (NCEI), National Centers for. 2020. "Climate Data Online." *Climate Data Online (CDO) - the National Climatic Data Center's (NCDC) Climate Data Online (CDO)*. https://www.ncdc.noaa.gov/cdo-web/.

Lee, Mintaek, and Jaechoul Lee. 2020. "Trend and Return Level of Extreme Snow Events in New York City." *The American Statistician* 74 (3): 282–93. https://doi.org/10.1080/00031305.2019.1592780.

Smith, Adam. 2020. "NOAA National Centers for Environmental Information (Ncei) U.s. Billion-Dollar Weather and Climate Disasters." https://www.ncdc.noaa.gov/billions/.