April 2020

McGill University
MATH410: Introduction to Extreme Value Theory

Lambert De Monte (260746902)

Abstract

In this project, I intend to familiarize myself with extreme value theory through a study of the book *An Introduction to Statistical Modeling of Extreme Values* [2] by Coles and explore relevant statistical modeling packages using the R software. The final objective is to produce an analysis of the maxima of temperature in July in Montreal, Québec City, Toronto and Vancouver and the maxima of daily rain precipitations for the same cities using data from Evironment Canada from 1937 up to 2020.

Keywords: Extreme-Value Theory, Univariate Extremes, Temperature Extremes, Rain Precipitation Extremes, Canadian weather, R software.

1 Introduction

The principal intent of extreme value analysis is to provide reasonable quantitative descriptions of the unusual (small or large, termed extreme) values attained by the stochastic behavior of a process. In a more mathematical expression, extreme value theory aims to quantify M_n, the maximum of a process over an "n-observation" period. We define M_n as

$$M_n = max\{X_1, X_2, ..., X_n\}$$

where X_i is the i-th realization of the process X for $i \in \{1, 2, ...\}$.

In theory, if the distribution function F of X_i were known (and assuming independent and identical distributions for all i), one could derive the corresponding cumulative distribution function of M_n as

$$\begin{aligned} Pr\{M_n \leq z\} &= Pr\{X_1 \leq z, X_2 \leq z, ..., X_n \leq z\} \\ &= \prod_{i=1}^{n} Pr\{X_i \leq z\} \\ &= \{F(z)\}^n \end{aligned}$$

However, in practice F is unknown and the estimation of $Pr\{M_n \leq z\}$ is thus more subtle. Since small discrepancies in the estimation of F may lead to considerable ones in F^n, extreme value analysis suggests we should instead model the behavior of F^n as $n \to \infty$.

Through the attempt of extrapolating the observed extremes and arriving to a model generalization for the limits of their underlying distribution, one faces the so-called "extreme value paradigm". Indeed, the procedure's intent is to arrive to conclusions on potentially infinite maxima based solely on a finite amount of observations. This paradigm should be considered as a warning for all of the forthcoming analyses: the extrapolations made on the expected maxima to be observed should be considered as lower bounds, as there is no garantee that the finite amount of data observed captures enough information on the asymptotic behavior of the "true" underlying distribution.

2 Classical Extreme Value Theory

Modeling the behavior of $Pr\{M_n \leq z\} = \{F(z)\}^n$ is however not directly suitable in the limit as $n \to \infty$. Defining z_+ as the upper endpoint of F (i.e. $z_+ = min(z) : F(z) = 1$), we have that the cumulative distribution function of M_n degenerates to point mass on z_+. Indeed,

$$\lim_{n\to\infty} Pr\{M_n \leq z\} = \lim_{n\to\infty} \{F(z)\}^n = \begin{cases} 1 & \text{if } z = z_+ \\ 0 & \text{otherwise} \end{cases}$$

We therefore require a linear normalization on M_n, denoted M_n^* such that

$$M_n^* = \frac{M_n - a_n}{b_n}$$

where $\{a_n > 0\}$,$\{b_n\}$ are sequences of constants that stabilize the location and scale of M_n^* as $n \to \infty$ and we seek for a model of the limit distribution for M_n^*.

2.1 Extremal Types Theorem

If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$Pr\{M_n^* \leq z\} \to G(z) \text{ as } n \to \infty$$

where $G(z)$ is a non-degenerate distribution function, then $G(z)$ belongs to one of the following families:

I)
$$G \sim \text{Gumbel} : G(z) = e^{-e^{-\frac{z-a}{b}}}, -\infty < z < \infty$$

II)
$$G \sim \text{Frechet} : G(z) = \begin{cases} 0 & \text{z} \leq b \\ e^{-(\frac{z-a}{b})^{-\alpha}} & \text{z > b} \end{cases}$$

III)
$$G \sim \text{Weibull} : G(z) = \begin{cases} e^{-[-(\frac{z-a}{b})^{\alpha}]} & \text{z < b} \\ 1 & \text{z} \geq b \end{cases}$$

The above theorem implies that the rescaled sample maxima M_n^* converge in distribution to a random variable belonging to one of the Gumbel, Fréchet or Weibull distributions. Since these are limit distributions for the tails of the underlying $\prod_{i=1}^{n} Pr\{X_i \leq z\} = \{F(z)\}^n \approx G(z)$ distribution of the X_i's, it is understood that they must account for different behaviors of the extremes of X_i. Defining z_+ as the upper end-point of G, one makes the following observations regarding the choice of extremal type:

- If $G \sim$ Weibull, then $|z_+| < \infty$
- If $G \sim$ Fréchet, Gumbel, then $|z_+| = \infty$

In other words, we have that the Weibull distribution has a finite upper end-point, whereas the Fréchet and Gumbel distributions do not; the latter two are differentiated by the fact that their density respectively decay polynomially and exponentially.

2.2 Generalized Extreme Value Distribution

Originally, it was usual to model the extremes of a particular process by fitting its tail values to one of the extremal types discussed in section 2.1 according to the belief of the underlying behavior for the extremes. A better approach now consists of instead fitting the data to the Generalized Extreme Value (GEV) distribution, which presents itself as the name suggests, as a generalization of the Extremal Types. Indeed, we introduce the defining location (μ), scale (σ) and shape (ξ) parameters for the GEV(μ, σ, ξ) distribution:

$$G(z) = e^{-[1+\xi(\frac{z-\mu}{\sigma})^{-\frac{1}{\xi}}]}$$

where z is such that $1 + \xi(\frac{z-\mu}{\sigma}) > 0$.

We interpret the process' extremal behavior by analyzing the maximum likelihood estimates (μ, σ, ξ) of the GEV distribution fitted to the data. The shape parameter ξ segregates between the possible tail behaviors of a process according to the extremal types. Indeed, $\xi > 0$ and $\xi < 0$ correspond to the cases of Fréchet and Weibull family of distributions respectively. Further, the case $\xi = 0$, which should be interpreted as $\lim_{\xi \to 0} G(z)$, corresponds to the Gumbel family of distributions.

We therefore have that by fitting the GEV maximum likelihood estimates to the data, we obtain a value for ξ which directly implies which of the three extremal types is the most appropriate. Constructing confidence intervals for ξ using the delta method, we can test for the null hypothesis that $\xi = 0$, and thus that $G(z) \sim$ Gumbel.

The extremal types theorem therefore generalizes to its analogue for the GEV distribution:

> If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that
>
> $$Pr\{M_n^* \leq z\} \to G(z) \text{ as } n \to \infty$$
>
> where $G(z)$ is a non-degenerate distribution function, then $G(z)$ belongs the generalized extreme value family
>
> $$G(z) = e^{-\left[1+\xi(\frac{z-\mu}{\sigma})^{-\frac{1}{\xi}}\right]}$$
>
> where z is such that $1 + \xi(\frac{z-\mu}{\sigma}) > 0$ and $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$.

The theorem implies that $M_n^* = \frac{M_n - a_n}{b_n}$ can be modeled by the GEV family of distributions, but in practice the normalizing constants a_n and b_n are unknown and the model that we want is actually for M_n. It can however be shown that if $G(z)$ can me modeled by a GEV, then

$$G(z) = Pr\{\frac{M_n - a_n}{b_n}\}$$

and

$$Pr\{M_n \leq z\} = G(\frac{z - a_n}{b_n}) = \tilde{G}(z)$$

where $\tilde{G}(z)$ is a model for M_n with estimates $(\tilde{\mu}, \tilde{\sigma}, \tilde{\xi})$. Indeed, we have that $\tilde{G}(z)$ is a also member of the generalized extreme value family of distributions, with potentially different estimates.

2.3 Block Maxima Models

Using the theory from sections 2.1 and 2.2, we develop a first statistical approach to model the extremes of a process X. Suppose $X_1, X_2, ..., X_{m \cdot n} \in X$, we "block" the data into m bins each containing n observations and compute the maximum $M_{n,k}$ for each bin.

The choice of m, although sometimes enforced by a logical period like a regular year of $m = 365$ daily observations for instance, induces a bias-variance trade-off in the model. Indeed, choosing too small of a value for m can lead to selecting values that should not be considered as extremes, introducing bias to the model. Conversely, choosing m to be too large can lead to ignoring some values that should be considered as extremes, reducing the amount of data used to fit the model and thus inducing variance.

The binning operation yields the series of maxima $M_{n,1}, M_{n,2}, ..., M_{n,m}$ which we use to fit the GEV distribution and get the maximum likelihood estimates $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$. The resulting function $G(z)$ represents the fitted probability that an observation is smaller than z. Therefore, if $G(z_p) = 1 - p$, p is the probability that an n-observation period maximum exceeds z_p, and $1/p$ is the expected number of n-observation periods until the maximum observation exceeds z_p. We say that z_p is the return level and that $1/p$ is the return period. The return level z_p is thus a quantile of the n-observation maximum's distribution and is obtained by inverting the GEV distribution to obtain

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - y_p^{-\xi}\right] & \xi \neq 0 \\ \mu - \sigma log(y_p) & \xi = 0 \end{cases}$$

where $y_p = -log(1-p)$. A similar approach is used to obtain a model for the extremes of minimal values $\bar{M_n} = min\{X_1, X_2, ..., X_n\}$. Letting $Y_i = X_i$, We have that $\bar{M_n} = -M_n$, where $M_n = max\{Y_1, Y_2, ..., Y_n\}$. We then proceed as before and fit $M_{n,1}, M_{n,2}, ..., M_{n,m}$ to the GEV distribution.

2.3.1 Inference for return levels

Since the maximum likelihood estimate (say φ) of any function (say g) of another maximum likelihood estimate (say ϕ) is obtained by simple substitution (i.e. $\varphi = g(\phi)$), we have that the maximum likelihood estimate for z_p is

$$\hat{z_p} = \begin{cases} \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\left[1 - y_p^{-\hat{\xi}}\right] & \hat{\xi} \neq 0 \\ \hat{\mu} - \hat{\sigma} log(y_p) & \hat{\xi} = 0 \end{cases}$$

Using the approximate normality of the distribution of maximum likelihood estimators, we can compute the variance of z_p using the delta method. Thus, if V is the covariance matrix for the maximum likelihood estimates,

$$\mathrm{Var}(z_p) = \nabla_{z_p}^T V \nabla_{z_p}$$

where

$$\nabla_{z_p}^T = \left[\frac{\partial z_p}{\partial \mu}, \frac{\partial z_p}{\partial \sigma}, \frac{\partial z_p}{\partial \xi}\right] = \left[1, -\frac{(1-y_p)^{-\xi}}{\xi}, \sigma\frac{(1-y_p)^{-\xi}}{\xi^2} - \sigma\frac{log(y_p)}{\xi y_p^{\xi}}\right]$$

evaluated at $(\mu, \sigma, \xi) = (\hat{\mu}, \hat{\sigma}, \hat{\xi})$.

Further, in the case where $\hat{\xi} < 0$, it is also possible to make statistical inference on what is termed the "infinite-observation return period", or the upper end-point of the underlying distribution. That is, we use the return period $\lim_{p \to 0} {}^1\!/_p = \infty$, such that the return level $\lim_{p \to 0} \hat{z}_p = \hat{z}_0$ is

$$\hat{z}_0 = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}} \quad \text{and} \quad \nabla_{z_0}^T = \left[1, -\hat{\xi}^{-1}, \hat{\sigma}\hat{\xi}^{-2}\right]$$

again evaluated at $(\mu, \sigma, \xi) = (\hat{\mu}, \hat{\sigma}, \hat{\xi})$. In the case where $\hat{\xi} \geq 0$, the maximum likelihood estimate for $\hat{z}_0$ is infinity.

It remains that one should interpret the return levels with care, especially as they are computed for long return periods. Indeed, the normal approximation of the maximum likelihood distribution for the estimators may not be perfectly accurate. One way to obtain more accurate results for the normality of the maximum likelihood estimates, and therefore better inference on return levels is to use profile likelihoods.

For any $0 < p < 1$, it is possible to obtain confidence intervals using the profile likelihood by reparametrizing the GEV model such that z_p is one of the model parameters.

$$\mu = \begin{cases} z_p + \frac{\sigma}{\xi}\left[1 - y_p^{-\xi}\right] & \xi \neq 0 \\ z_p + \sigma log(y_p) & \xi = 0 \end{cases}$$

We then obtain the profile likelihood for z_p by maximizing the likelihood with respect to the remaining parameters.

Although the profile likelihood generally yields better normality for its estimates than simple maximum likelihood estimates, it remains that in both cases, the results for the estimates rely on the assumption that the model itself is correct.

Because of the uncertainty of the model correctness, the results for return levels and their estimated variance should be interpreted as lower bounds that could easily be surpassed.

2.3.2 Model Checking

Since it is not possible to make validations on the extrapolations of the fitted GEV distribution, it is usual to assess the goodness of a model with how well it references the actual data. Ordering each block maximum so that we have $z_{(1)}, z_{(2)}, ..., z_{(m)}$, we obtain the empirical distribution function evaluated at $z_{(i)}$

$$\tilde{G}(z_{(i)}) = \frac{i}{m+1}$$

and the model based estimates are

$$\hat{G}(z_{(i)}) = e^{-\left[1+\hat{\xi}\left(\frac{z_{(i)}-\hat{\mu}}{\hat{\sigma}}\right)\right]^{-\frac{1}{\hat{\xi}}}}$$

If the model fits the data well, the plot generated by $\left\{\left(\tilde{G}(z_{(i)}), \hat{G}(z_{(i)})\right), i = 1, 2, ..., m\right\}$ should lie on the unit-slope line $x = y$. However, the limits of both $\tilde{G}(z_{(i)})$ and $\hat{G}(z_{(i)})$ approach 1 as $z_{(i)}$ increases, which leads to a plot having a $(x = y)$-like look for large values of $z_{(i)}$. This is problematic for model assessment since we are usually interested in the large values of $z_{(i)}$.

One way to get a better portrait of the goodness of the model is to look at an alternate plot, the quantile plot $\left\{\left(\hat{G}^{-1}(\frac{i}{m+1}), z_{(i)}\right), i = 1, 2, ..., m\right\}$ where

$$\hat{G}^{-1}(\frac{i}{m+1}) = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\left[1 - \left[-log\left(\frac{i}{m+1}\right)\right]^{-\hat{\xi}}\right]$$

The above plot should again have a $(x = y)$-like linearity look, discrepancies between the two could imply that the model is not accurate for the data.

Another appropriate way to assess the goodness of the model is via the return level plot $\left\{\left(log(y_p), \hat{z}_p\right), 0 < p < 1\right\}$. The empirical estimates of the return level function should look similar to the return level plot for the fitted model.

2.4 R-Largest Order Statistic Model

One of the weaknesses of the Block Maxima models is the loss of information that happens for the extremes of the underlying distribution, caused by the use of only the maximum over a certain period. Indeed, there is no reason to believe that the second largest observation, up to the r^{th} largest values over the same period cannot be regarded as extremes too: ignoring these might very well lead to too simple of a model.

The r-largest order statistic models can thus be considered as generalizations of Block Maxima models in the sense that it accounts for more data. We extend the previous concept of $M_n = max\{X_1, X_2, ..., X_n\}$ to

$$M_n^{(k)} = k^{\text{th}} \text{ largest of } \{X_1, X_2, ..., X_n\}$$

and identify the limit behavior of $\lim_{n\to\infty} M_n^{(k)}$ for fixed k. We can thus define the vector of the r-largest order statistic $\boldsymbol{M}_n^{(r)}$ as

$$\boldsymbol{M}_n^{(r)} = \left(M_n^{(1)}, ..., M_n^{(r)}\right)$$

An adaptation of the extremal types theorem to the r-largest order statistic is as follows.

If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$Pr\{M_n^* \leq z\} \to G(z) \text{ as } n \to \infty$$

for some non-degenerate distribution funtion so that $G(z)$ is the GEV distribution function, then for fixed k,

$$Pr\left\{\frac{\boldsymbol{M}_n^{(k)} - a_n}{b_n} \leq z\right\} \to G_k(z)$$

on $\{z : 1 + \xi^{(z-\mu)}/\sigma > 0\}$ as $n \to \infty$ where

$$G(z) = e^{-\tau(z)} \sum_{s=0}^{k-1} \frac{\tau(z)^s}{s!}$$

with

$$\tau(z) = \left[1 + \xi(\frac{z-\mu}{\sigma})\right]^{-\frac{1}{\xi}}$$

The above theorem implies that normalizing each of the k^{th}-largest order statistic ($1 \leq k \leq r$) by the same sequences as the maximum $M_n^{(1)}$ yields the corresponding $G_k(z)$ limiting distributions with parameters corresponding for the ones of $M_n^{(1)}$.

There are however two main difficulties that arise with this model. Firstly, the resulting $G_k(z)$ gives a family for the approximate distribution of each of the components of $\boldsymbol{M}_n^{(r)}$, but not for its joint distribution. Secondly, each component depends on the others, as by construction, $M_n^{(1)} \geq M_n^{(2)} \geq ... \geq M_n^{(r)}$. These two facts make it hard to compute the probability of observing a specific manifestation of the vector $\boldsymbol{M}_n^{(r)}$.

The following theorem offers an approach to obtain the joint density function of the limit distribution.

If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$Pr\{M_n^* \leq z\} \to G(z) \text{ as } n \to \infty$$

for some non-degenerate distribution funtion so that $G(z)$ is the GEV distribution function, then for fixed r, the limiting distribution as $n \to \infty$ of

$$\tilde{\boldsymbol{M}}_n^{(r)} = \left(\frac{M_n^{(1)} - a_n}{b_n}, ..., \frac{M_n^{(r)} - a_n}{b_n}\right)$$

falls within the family having joint probability density function

$$f(z^{(1)}, ..., z^{(r)}) = \exp\left\{-\left[1 + \xi\left(\frac{z^{(r)} - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \prod_{k=1}^{r} \sigma^{-1} \left[1 + \xi\left(\frac{z^{(r)} - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}-1}$$

where $-\infty < \mu < \infty, \sigma > 0$ and $-\infty < \xi < \infty$; $z^{(r)} \leq z^{(r-1)} \leq ... \leq z^{(1)}$; and $z^{(k)} : 1 + \xi\left(z^{(k)} - \mu\right)/\sigma > 0$ for $k = 1, ..., r$.

The case $\xi = 0$ should again be interpreted as

$$\lim_{\xi \to 0} f(z^{(1)}, ..., z^{(r)}) = \exp\left\{-\exp\left[-\left(\frac{z^{(r)} - \mu}{\sigma}\right)\right]\right\} \prod_{k=1}^{r} \sigma^{-1} \exp\left[-\left(\frac{z^{(r)} - \mu}{\sigma}\right)\right]$$

It can also be seen that in the latter case, $r = 1$ reduces to the density of the Gumbel family.

2.4.1 Modeling the r Largest Order Statistic

Let X be a process such that $X_1, X_2, ..., X_{m \cdot n}$ are independently and identically distributed. The data are grouped into m blocks, each containing n observations. Let $z_i^{(k)}$ be the k^{th} largest observation of the process X for the block i and define the vector of the r largest values for the same block as

$$M_i^{(r_i)} = \left(z_i^{(1)}, ..., z_i^{(r_i)}\right)$$

for $i = 1, ..., m$ and $r_1 = ... = r_m = r$ for some r, unless some block has fewer data. As for Block Maxima models, the choice of m in the r Largest Order Statistic models has the same bias-variance trade-off, and the concept can also be extended to the choice of r. Indeed, small values of r lead to higher variance in the data used for the model, whereas larger values lead to bias that presents itself as a violation of the asymptotic support for the model. Instinctively, one

could say that too large of an r-value leads to incorporating non-extreme values to the model. The specific choice of r is usually determined based on model diagnostics.

From the previous equation for $f(z^{(1)}, ..., z^{(r)})$, we get that the likelihood for this model is

$$L(\mu, \sigma, \xi) = \prod_{i=1}^{m} f(z^{(1)}, ..., z^{(r_i)}) = \prod_{i=1}^{m} \left\{ \exp\left[-\left[1 + \xi \left(\frac{z^{(r_i)} - \mu}{\sigma} \right) \right]^{-\frac{1}{\xi}} \right] \prod_{k=1}^{r_i} \sigma^{-1} \left[1 + \xi \left(\frac{z^{(k)} - \mu}{\sigma} \right) \right]^{-\frac{1}{\xi}-1} \right\}$$

In the case when $\xi = 0$, the likelihood becomes

$$L(\mu, \sigma, \xi) = \prod_{i=1}^{m} f(z^{(1)}, ..., z^{(r_i)}) = \prod_{i=1}^{m} \left\{ \exp\left(-\exp\left[-\left(\frac{z^{(r_i)} - \mu}{\sigma} \right) \right] \right) \prod_{k=1}^{r_i} \sigma^{-1} \left[-\left(\frac{z^{(k)} - \mu}{\sigma} \right) \right] \right\}$$

The maximum likelihood estimates for the r Largest Order Statistic models can be obtained numerically and correspond to Block Maxima models' estimates. However, the estimates computed using the r largest order statistics are fitted with more data and thus contain more information: the precision of the computed estimates should be increased compared to those of the Block Maxima.

2.5 Threshold Models

A pretty obvious motivation for the improvements made from Block Maxima models to the r Largest Order Statistic models was the reduction of the data/information waste by using more than just the maximum over a certain period of observation.

The r Largest Order Statistic models however still have some flaws. Indeed, the choice made for r can lead to a loss of information too if the number of deemed extreme events in a certain block of observations is larger than those r largest. Further, the model still imposes a choice for the number of observations per block, which even for seemingly logical choices like yearly grouping can lead to a bias-variance trade-off. These are the main motivations for a possible choice of the Threshold models over the previously developed ones.

Assuming that we have observations $X_1, ..., X_n$ from a process X that are again independently and identically distributed, we now term "extreme" those that excess a certain threshold u. Denoting an arbitrary term X_i of the sequence by X, we model the probability that this observation exceeds the threshold u by an amount of y, given that it does exceeds u, i.e. $Pr\{X > u + y | X > u\}$. Assuming a marginal distribution function F for the X_i's, this probability is thus computed as

$$Pr\left\{ X > u + y \,\middle|\, X > u \right\} = \frac{Pr\{X > u + y, X > u\}}{Pr\{X > u\}} = \frac{1 - Pr\{X < u + y\}}{1 - Pr\{X < u\}} = \frac{1 - F(u + y)}{1 - F(u)}$$

As before, since the underlying F is unknown, we cannot directly compute this probability and we seek for a method to model the excesses of high threshold values.

2.5.1 General Pareto Distribution

The distributions that are generated by the equation

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

are members of the generalized Pareto family. The following theorem links the general Pareto distribution to the excesses of a threshold u.

> Let $X_1, X_2, ...$ be a sequence of independent random variables with common distribution function F, and let $M_n = max\{X_1, ..., X_n\}$. Denote an arbitrary term in the X_i sequence by X, and suppose that F satisfies

$$Pr\{M_n \leq z\} \approx G(z),$$

where

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

for some $\mu, \sigma > 0$ and ξ. Then for large enough u, the distribution function of $(X-u)$ conditional on $X > u$ is approximately

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

where $y > 0$ and $(1 + {}^{\xi y}/_{\tilde{\sigma}}) > 0$ with $\tilde{\sigma} = \sigma + \xi(u-\mu)$.

The above theorem implies that if the block maxima of the process X are distributed according to the generalized extreme value distribution, then the excesses of a threshold u have a corresponding approximate generalized Pareto distribution, with parameters (μ, σ, ξ) determined by the ones of the GEV distribution fitted to the block maxima. In fact, ξ is actually equal for both distributions since it is not affected by the block size and $\tilde{\sigma}$ remains constant as n varies: the change in μ and σ for different u's is self-compensating in the calculation of $\tilde{\sigma}$.

As for the GEV distribution, we see that the parameter ξ influences the behavior of the generalized Pareto distribution.

If $\xi < 0$, the excesses are upper bounded by $u - {}^{\tilde{\sigma}}/_{\xi}$.

If $\xi > 0$, the excesses are unbounded.

If $\xi = 0$, the excesses are unbounded and

$$H(y) = 1 - \exp\left(-\frac{y}{\tilde{\sigma}}\right), y > 0$$

2.5.2 Justification for the Generalized Pareto Model

Let X have distribution function F. Then for large enough n,

$$F^n(z) \approx \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}$$

for some $\mu, \sigma > 0$ and ξ. Therefore,

$$n\log(F(z)) \approx -\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}$$

Now, for large values of z, the Taylor expansion of $\log(F(z)) \approx -\{1-F(z)\}$ and thus

$$1 - F(u) = -\log(F(u)) = \frac{1}{n}\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}$$

and thus, similarly, for $y > 0$,

$$1 - F(u+y) = \frac{1}{n}\left[1+\xi\left(\frac{(u+y)-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}$$

The above two estimates for $1 - F(u)$ and $1 - F(u+y)$ allow us to compute

$$Pr\{X > u + y \,\Big|\, X > u\} = \frac{1 - F(u+y)}{1 - F(u)}$$

$$\approx \frac{(1/n)\left[1 + \xi\left(\frac{(u+y)-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}}{(1/n)\left[1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}}$$

$$= \left[1 + \frac{\xi(u+y-\mu)/\sigma}{\xi(u-\mu)/\sigma}\right]^{-\frac{1}{\xi}}$$

$$= \left[1 + \frac{\xi y}{\tilde{\sigma}}\right]^{-\frac{1}{\xi}}$$

where $\tilde{\sigma} = \sigma + \xi(u - \mu)$

2.5.3 Modeling Threshold Excesses

As before, we assume that the observed data we have, say $x_1, ..., x_n$, are independent draws from the same underlying distribution. Suppose we find a threshold u such that all greater observations can be considered extreme, we would then define these (all $x_i : x_i > u$) as exceedances and order them $x_{(1)}, ..., x_{(k)}$. We thus term threshold excesses all y_i such that $y_i = x_{(i)} - u$. The inference and the modeling of the threshold excesses is made by fitting the general Pareto distribution to these y_i.

The obvious question that arises with this modeling technique is the procedure for choosing u, as it clearly influences the number of excesses to be used. As for the previous model, we have that the choice of the threshold generates a bias-variance trade-off. In this case, too low of a threshold induces bias to the model, as it is likely that non-extreme observations end up being modeled. Conversely, using a large threshold may discriminate extreme values and keep only a small amount of data to be modeled, inducing variance to the model. It is thus evident that the choice of u is critical and should be taken with care. *Coles* develops two techniques for arriving to a choice of threshold.

The first method is based on the mean of the general Pareto distribution for a specific choice of u. Indeed, for a given u_0 such that the y_i are observations of a general Pareto distribution Y, and for $u > u_0$

$$\mathbb{E}(Y) = \mathbb{E}(X - u | x > u) = \frac{\sigma_u}{1 - \xi}$$

From the equation for $\tilde{\sigma} = \sigma + \xi(u - \mu)$ in section 2.5.2, we see that σ grows linearly with u, implying that

$$\mathbb{E}(Y) = \frac{\sigma_u}{1 - \xi} = \frac{\sigma_{u_0} + \xi u}{1 - \xi}$$

The mean of the observations of Y should thus grow linearly with u as well. It follows that data for which a generalized Pareto distribution fits well should yield a linear trend when plotting the mean of the threshold excesses against u. Such a plot is termed "mean residual life plot" and can be expressed as the set of points

$$\left\{ \left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} (x_i - u) \right) : u < x_{\max} \right\}$$

where $x_{(1)}, ..., x_{(n_u)}$ are the n_u observations that exceed the threshold u, and $x_{\max}$ is the largest of all observations from the process X. In practice, it is not always straight forward to find obvious linearity and it might be necessary to settle for weaker evidence of linearity. Indeed, selecting only the data that fit obvious linearity may lead to selecting too high of a threshold i.e. too few data points deemed extreme.

The second method is based on the parameter estimates for the model at various thresholds u. By the theorem for the excesses of a threshold u developed in section 2.5.1, we have that ξ is constant for $u > u_0$. Therefore, a good Threshold model should have its maximum likelihood parameter $\hat{\xi}$ be approximately constant for u above a critical u_0. Further, because of the linearity of $\tilde{\sigma} = \sigma + \xi(u - \mu)$, we know that for a given σ_{u_0} and for $u > u_0$, σ_u should be growing linearly, and thus that

$$\sigma_u = \sigma_{u_0} + \xi(u - u_0)$$

Reorganizing, we get

$$\sigma_u - \xi u = \sigma_{u_0} + \xi u_0 = \sigma^*$$

where clearly σ^* is a constant, and thus the quantity $\hat{\sigma}^* = \hat{\sigma}_u - \hat{\xi} u$ should remain constant too for $u > u_0$. The method thus consists of plotting $\hat{\sigma}^*$ and $\hat{\xi}$ against u and picking u_0 to be the minimal u that yields near constant plots. It is also possible to test for H_0: "Estimates $\hat{\sigma}^*$ and $\hat{\xi}$ are constant" using linear regression, where the confidence intervals for $\hat{\xi}$ are obtained from the covariance matrix V of the estimates and through the delta method for $\hat{\sigma}^*$.

$$\mathrm{Var}(\sigma^*) \approx \nabla_{\sigma^*}^T V \nabla_{\sigma^*}$$

where

$$\nabla_{\sigma^*}^T = \left[\frac{\partial \sigma^*}{\partial \sigma_u}, \frac{\partial \sigma^*}{\partial \xi} \right] = [1, -u]$$

Once a threshold has been chosen, it is possible to obtain the maximum likelihood estimates of the generalized Pareto distribution by numerical maximization of the log-likelihood for the model. That is, if $\xi \neq 0$, then

$$\ell(\sigma, \xi) = -k \log \sigma - (1 + \frac{1}{\xi}) \sum_{i=1}^{k} \log(1 + \frac{1 + \xi y_i}{\sigma})$$

provided $\xi \neq 0$ and $(1 + {}^{\xi y_i}/_{\sigma}) > 0$, otherwise $\ell(\sigma, \xi) = -\infty$. If $\xi = 0$, then

$$\ell(\sigma, \xi) = -k \log \sigma - \frac{1}{\sigma} \sum_{i=1}^{k} y_i$$

2.5.4 Return Levels

Assuming that a generalized Pareto distribution with parameters σ and ξ is a valid model for the exceedances of a given threshold u by the process X, then as discussed in section 2.5.2,

$$Pr\left\{ X > x \,\middle|\, X > u \right\} = \left[1 + \xi \left(\frac{x - u}{\sigma} \right) \right]^{-\frac{1}{\xi}}.$$

Now,

$$Pr\left\{ X > x \,\middle|\, X > u \right\} = \frac{Pr\{X > x, X > u\}}{Pr\{X > u\}} = \frac{Pr\{X > x\}}{Pr\{X > u\}}$$

Therefore, if we wish to obtain the probability that an observation from the process X exceeds some x, unconditionally of u, we need to multiply the above equation by $\zeta_u = Pr\{X > u\}$. Clearly, the number of observations that exceed u is distributed according to a binomial distribution, meaning that the best estimator for the probability for one observation to exceed u is $\zeta_u = {}^{k}/_{n}$, where k is the number of exceedances and n is the total number of observations. Thus,

$$Pr\{X > x\} = \zeta_u \left[1 + \xi \left(\frac{x - u}{\sigma} \right) \right]^{-\frac{1}{\xi}}.$$

This implies that for $Pr\{X > x_m\} = 1/m$, then x_m is the level that is expected to be exceeded once every m observations, and thus termed the m-observation return level. Solving for x_m in the equation above, we get

$$x_m = \begin{cases} u + \dfrac{\sigma}{\xi} \left[(m\zeta_u)^{\xi} - 1 \right] & \xi \neq 0 \\ u + \sigma \log(m\zeta_u) & \xi = 0. \end{cases}$$

2.6 Extremes of Dependent Sequences

Coles defines the notion of dependence according to the $D(u_n)$ condition. Letting $X_1, X_2, ...$ be a stationary series, then the process X is said to satisfy the $D(u_n)$ condition if for all $i_1 < ... < i_p < j_1 < ... < j_q$ such that $j_1 - i_p > l_n$

$$|Pr\{X_{i1} \leq u_n, ..., X_{ip} \leq u_n, X_{j1} \leq u_n, ..., X_{jq} \leq u_n\} - Pr\{X_{i1} \leq u_n, ..., X_{ip} \leq u_n\} Pr\{X_{j1} \leq u_n, ..., X_{jq} \leq u_n\}| \leq \alpha(n, l_n)$$

where $\alpha(n, l_n) \to 0$ for some sequence l_n such that $l_n/n \to 0$ as $n \to \infty$.

This criterion basically uses the fact that perfect independence would lead the difference in the two probabilities of the above equation to be 0. Thus, if the absolute value of the difference in probabilities tends to 0 as n increases, then the dependence of the series decreases fast enough to fit the $D(u_n)$ criterion.

This criterion is used in the following theorem which allows for an analysis of stationary sequences:

> If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that
>
> $$Pr\{M_n^* \leq z\} \to G(z) \text{ as } n \to \infty$$
>
> where $G(z)$ is a non-degenerate distribution function and the $D(u_n)$ criterion is satisfied with $u_n = a_n z + b_n$ for all $z \in \mathbb{R}$, then $G(z)$ belongs the generalized extreme value family of distributions.

The above theorem stipulates that provided that long-range dependence at extreme levels is weak, the data can be considered to satisfy the $D(u_n)$ criterion and that the fitted Block Maxima models yield a distribution that is a member of the same family of distributions as its related independent series. This loosens the assumptions of validity for Block Maxima models.

Additional care needs to be taken when dealing with Threshold Exceedances models. Dependent sequences, in the eventuality of an observation being extreme, have a higher probability of generating "neighbours" that are extremes too. Such neighbouring observations could be considered as clusters of extreme values that introduce bias in the marginal estimation of the underlying distribution of the X_i's by invalidating the log-likelihood maximization. De-clustering procedures can be used by identifying the clusters extreme values and keeping only the maximum of the cluster as an observation, and then fitting the generalized Pareto distribution to these cluster maxima. This procedure however introduces potential bias in the model, simply because of the arbitrary choices made in identifying the clusters.

Section 2.6 of this project, although not covered in the chapters 1 to 4 of *Coles*, felt like a relevant justification for the validity of the analyses to come. Autocorrelation Function plots for the time series generated by the observations for the maxima of temperatures in July in Montreal and of daily rain precipitations are found in Figures 17 and 18 of the Appendix. The ACF plot of the daily temperatures in July reveals significance in the correlation between observations at lags of 5 or less, which will for this project justify the use of Block Maxima models and Threshold models. The ACF plot of the maxima shows continuous correlation between observations, but only with a very small magnitude of approximately 0.025 or less. I will thus consider these correlations between observations to leave Block Maxima models unaffected. Further, such a small coefficient of correlation would make the operation of declustering difficult as the identification of the clusters in itself would be ambiguous. The following analyses thus treat the observations as results of stationary sequences that can be modeled with classical extreme value theory.

3 Analysis of Maxima of Temperature in July for Cities in Canada

In the following section, I perform an analysis of the maxima of temperature in July in Montreal, Québec City, Toronto and Vancouver.

3.1 Data Processing and Reproducibility

The data used for the subsequent analyses are obtained from the website of Environment Canada[1]. The downloaded data are partitioned by year in different *.csv* files. I developed a Python script (found in my GitHub repository[3]) to merge all the *.csv* files from a specific location into a single *.csv* file, which is then imported into R for the purpose of extreme value analysis. Further details on the exact procedure is provided in the README.*md* file of my repository[3].

The resulting data frame is filtered to keep only observations that match the criterion $\mathrm{Month} = 7$ and to remove *NA* values from the daily temperature maxima column. Since different weather stations may have been used to get data for a longer time period and to guarantee uniqueness in daily observations, the data is then grouped by specific date and only the maximum is kept. The justification for this choice is that we wish to provide return levels that are pessimistic in the sense that they have a higher probability of "trapping" the future maxima to be observed. A plotted example of the resulting data is illustrated in Figure 1.

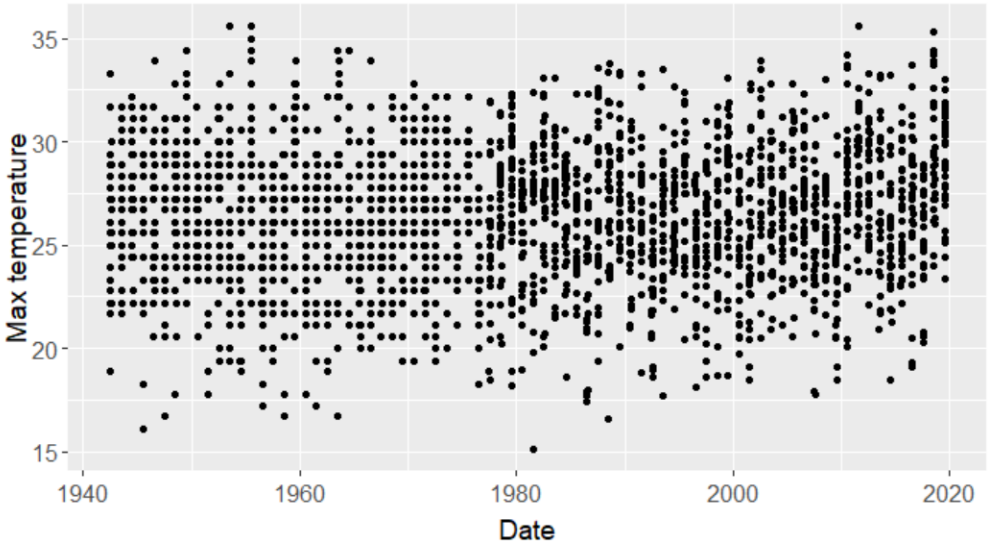

Figure 1: Daily Maxima in July in Montreal, Canada

An important remark needs to be made about the analyses to come: the basic assumption for all the models is that we do not observe significant trend in the data which should be considered in the computation of the long time return levels. A deeper analysis could eventually take into account the possible effects of climate change as a trend for all the data, or even simply as a trend in the extremes.

3.2 Analysis of the Maxima of Temperature in July in Montreal

Using the *fgev*() from the *evd* R package[5] and setting the block size m to 31 days (31 observations each year in July), we obtain the maximum likelihood estimates $[\hat{\mu}, \hat{\sigma}, \hat{\xi}] = [31.664, 1.737, -0.359]$ with corresponding 95% confidence interval (found in Table 4 in the Appendix) leading to the following confidence intervals:

$$\hat{\mu} \in [31.244, 32.085], \hat{\sigma} \in [1.440, 2.034], \hat{\xi} \in [-0.490, -0.229].$$

It follows that if the corresponding fitted GEV parameters yield an accurate model for the block maxima, then we can conclude that the maxima of temperature in July in Montreal are distributed according to a member of the Weibull family of distributions. Looking at the Quantile-Quantile plot in Figure 2, we observe fairly good linearity of the data with small residuals around the unit-slope line. This suggests that the model indeed seems to explain the distribution of the maxima well. Since there is evidence for the goodness of fit of this model, we can proceed to compute return levels, and in this case, even get an estimate for the infinite return level (since $\hat{\xi} < 0$, and thus $z_+ < \infty$).

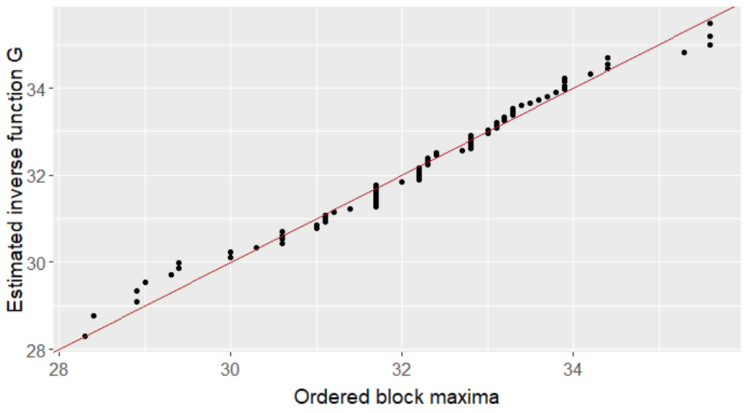

Figure 2: Quantile-Quantile plot for the GEV distribution fitted with the Maxima

As presented in Table 2 and in Figure 21 of the Appendix, we get from the generalized extreme value distribution fitted to the block maxima that the maximum likelihood estimates for the 10, 100, and "infinite" years return levels are respectively 34.345, 35.527, 36.499 Celsius degrees. For simplicity of comparison reasons, I plotted the infinite return level as the 0-year return level of Figures 21,22 and 23. The corresponding 95 % confidence intervals are respectively $[33.941, 34.749]$, $[34.996, 36.149]$ and $[35.255, 37.742]$.

The next logical approach to this analysis is to use the r-Largest Order Statistic models and compare the results. The case $r = 1$ obviously leads to the same results as the block maxima approach, since the likelihood for the models reduce to the same equations. Any small decimal discrepancy between both methods should be due to a difference in the optimizing technique, or the choice of the library. To compute the maximum likelihood estimates for the cases $r \geq 2$, I used the *ismev* R package[4]. The resulting fitted maximum likelihood estimates and corresponding 95% confidence intervals for the cases $r = 2$ and $r = 3$ are found in Table 4. This approach yields slightly different estimates for the generalized extreme value distribution than the Block Maxima's, and the Quantile-Quantile plots in Figure 3 for these estimates seem to reveal lack of fit in the new approach. In fact, the estimated $\hat{G}^{-1}(\frac{i}{m+1})$ seem to all lie above the unit slope line, slowly converging to it in the limit as the theory predicts even with models that do not fit well. It would imply that the fitted model predicts higher return levels than what should be expected in this case. This seems to be the case looking at the return levels from Table 2 and Figure 22 of the Appendix, but not with an alarmingly high difference, especially since the 95% confidence intervals for the same return levels overlap significantly between the above models.

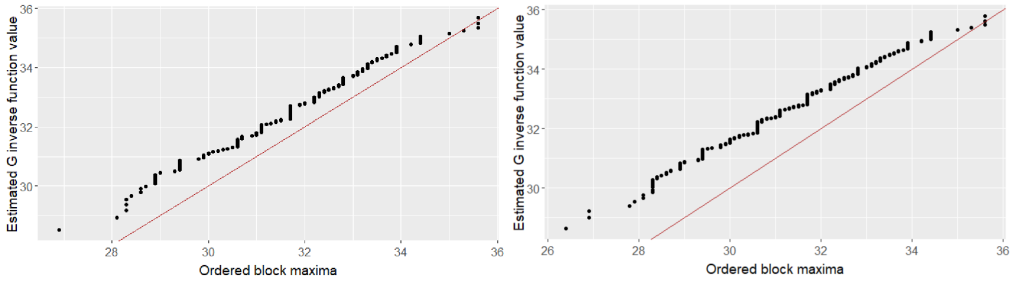

Figure 3: Quantile-Quantile plots for the GEV distribution fitted with the $r = 2$ (left) and $r = 3$ (right) Largest Order Statistic

The final approach is to analyze the threshold exceedances models for the same data. Using the *gpdFit*() function from the fExtremes R package[6], I obtained the maximum likelihood estimates $\hat{\sigma}$ and $\hat{\xi}$ with their respective 95% confidence intervals (found in Table 4) for various choices of thresholds on the interval $u \in [25, 32]$. The plots in Figure 4 show the fitted estimates against the value of u, as well as the least squares regression for the slope of the estimates above the threshold $u_0 = 28$.

The Normal QQ-plots of the residuals of the linear regression on the shape parameters (found in the Appendix, Figures 19 and 20) reveal approximate normality that allows for some inference on the slope. Although the normality is not perfectly respected, there seems to be significance in the slope (-0.294) of the scale parameter above the threshold $u_0 = 28$. Further, the p-value (0.620) for the slope (-0.002) of the shape parameter being very high, we do not reject the null hypothesis that the shape is indeed constant above the threshold $u = u_0$. Further, the mean residual life plot

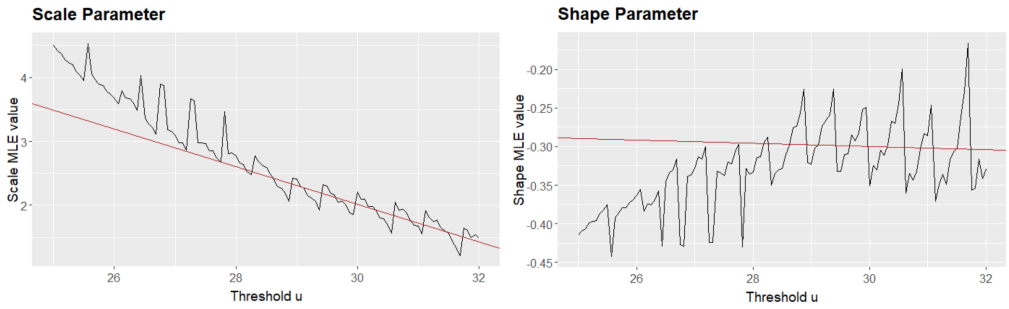

Figure 4: Linearity (in red) of the Scale and Shape Parameters for a Thresholds $u = 28$

for the model with $u = u_0$ in Figure 5 also has a linear trend, as expected for a model that fits the data well. We can conclude from the above results that the model appears to be a valid explanation for the maxima of temperature in July in Montreal and we can proceed to computing expected return levels. The model did seem to be accurate for higher choices of u as well, but I chose to keep the model $u = u_0$, 28 being the smallest choice such that the model holds. Indeed, lower thresholds, at the risk of introducing bias (which does not seem to be the case here), also reduce the variance as more data can be used for the "extremal" fit.

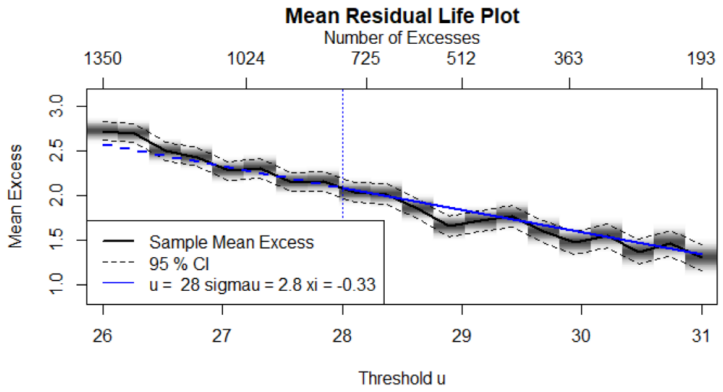

Figure 5: Mean Residual Life Plot for the Montreal Threshold Model

As seen in Table 2 and in Figure 23 of the Appendix, the maximum likelihood estimates of the 10 and 100 years return levels computed from the threshold exceedances models for $u \in \{28, 29, 30\}$ are in concordance with those of the block maxima and r-largest order statistic models. The 95% confidence intervals for the 10 and 100 years return levels of each threshold model do overlap significantly with those of the previous models. The case can thus be made that according to the data and based on the 95% confidence intervals, the different models agree on the expected return levels.

Table 1: Montreal Maxima Maximum Likelihood Estimates for the return levels (with 95% C.I.).

Model	**10 years** (°C)	**100 years** (°C)	**Infinite** (°C)
Block Maxima			
$m = 31$	34.345 ± 0.404	35.572 ± 0.577	36.499 ± 1.243
r-Largest Order Statistic			
$r = 1$	34.345 ± 0.404	35.572 ± 0.577	36.499 ± 1.243
$r = 2$	34.469 ± 0.338	35.560 ± 0.485	36.352 ± 0.908
$r = 3$	34.503 ± 0.324	35.570 ± 0.478	36.371 ± 0.864
Threshold Exceedances			
$u = 28$	34.525 ± 0.280	35.499 ± 0.404	
$u = 29$	34.535 ± 0.305	35.554 ± 0.468	
$u = 30$	34.572 ± 0.297	35.537 ± 0.432	

3.3 Analysis of the Maxima of Temperature in July in Other Canadian Cities

To arrive to the results presented in this section, the same techniques and R code were used. There is thus less to be said on the actual results and more about the comparison between the canadian cities.

As it was the case for Montreal, the quantile plots (Figure 6) for the Block Maxima model in Québec, Toronto and Vancouver reveal models that seem to fit to their respective maxima of temperature in July.

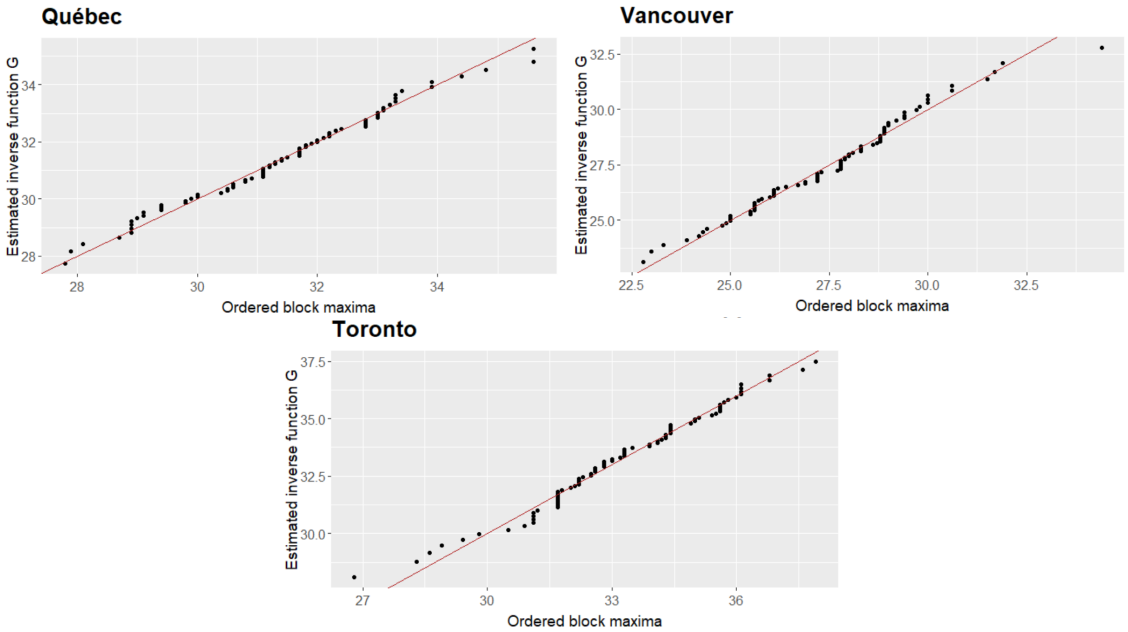

Figure 6: QQ-plots for the Block Maxima Models in Other Canadian Cities

Since the shape parameters are the same for the Block Maxima models in all four cities (i.e. $\hat{\xi} < 0$), we can again compute the same (including the infinite) return levels as in the Montreal case, and obtain the comparison of the maxima in Figure 7. The infinite return levels are plotted for simplicity as the 0-year return levels. We observe that the city that is the most prone to suffering from maxima in temperature is Toronto. Indeed, the 95% confidence interval is distinct from those of the other cities for very high return levels. They do overlap for the infinite return level, but not on the interval [0,100]. Québec and Montreal have highly overlapping 95% confidence intervals which make them not significantly distinguishable. This bears interest in itself knowing that both city are relatively close. Indeed, the converse statement being that maxima are significantly different could have been a red flag for our models. It would at least have been a topic to discuss with meteorologists before accepting the model. Finally, Vancouver seems to be the city that is the less likely to suffer from high temperatures in July.

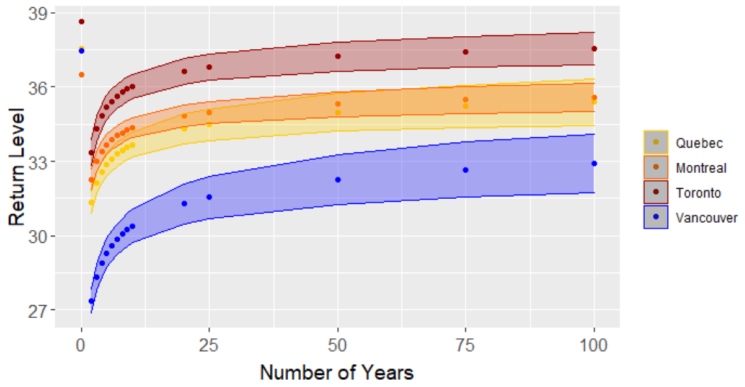

Figure 7: Block Maxima Return Levels for Maxima of temperatures of Cities in Canada in July

As it was the case for Montreal, the r-Largest Order Statistic models did not do as well for $r \geq 2$ with respect to the quantile plots, and so the discussion of the comparison of these model feels less relevant here. A better comparison is thus made through the threshold exceedances models.

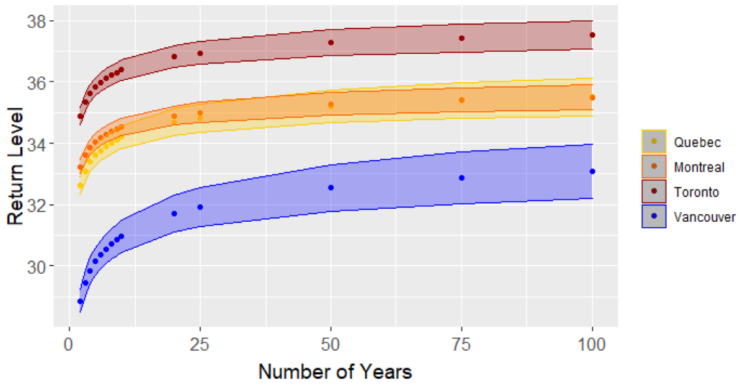

Figure 8: Threshold Exceedances Return Levels for Maxima of temperatures of Cities in Canada in July

Figure 8 displays the similar return level results obtained this time with Threshold Exceedances models. A superposition of the Block Maxima and Threshold Exceedances models is shown in the Appendix in Figure 24. It can be seen in both Figure 8 and 24 that Toronto has the highest expected return levels, with 95% confidence intervals distinct from the other cities in both models. Montreal and Québec again arrive in second position in terms of expected maxima return levels with highly overlapping 95% confidence intervals. Finally, Vancouver also has both its models agree significantly on the expected return levels, making the city the one that is less likely to suffer of high temperatures in July.

Table 2: Maximum Likelihood Estimates for the Return Levels in Other Canadian Cities (with 95% C.I.).

City	**10 years** (°C)	**100 years** (°C)	**Infinite** (°C)
Québec			
Block Maxima	33.670 ± 0.516	35.385 ± 0.922	37.538 ± 3.079
Threshold (u=28.5)	34.195 ± 0.380	35.508 ± 0.622	
Toronto			
Block Maxima	36.019 ± 0.495	37.551 ± 0.651	38.650 ± 1.319
Threshold (u=29.5)	36.385 ± 0.323	37.527 ± 0.470	
Vancouver			
Block Maxima	30.380 ± 0.678	32.914 ± 1.191	37.454 ± 5.316
Threshold (u=25)	30.972 ± 0.524	$33.089 \pm 0.0.884$	

4 Analysis of Maxima of Rain Precipitations for Cities in Canada

In the following section, I perform an analysis of daily maxima of rain precipitations in Montreal, Québec, Toronto and Vancouver.

4.1 Data Processing and Reproducibility

The initial data downloading and processing is the same as in Section 3.1. Once the .csv files have been merged, the imported data frames are now cleaned from null values only with respect to the relevant column *Total Rain (mm)*. The resulting daily rain precipitation maxima from 1943 to 2020 in Montreal are presented in Figure 9.

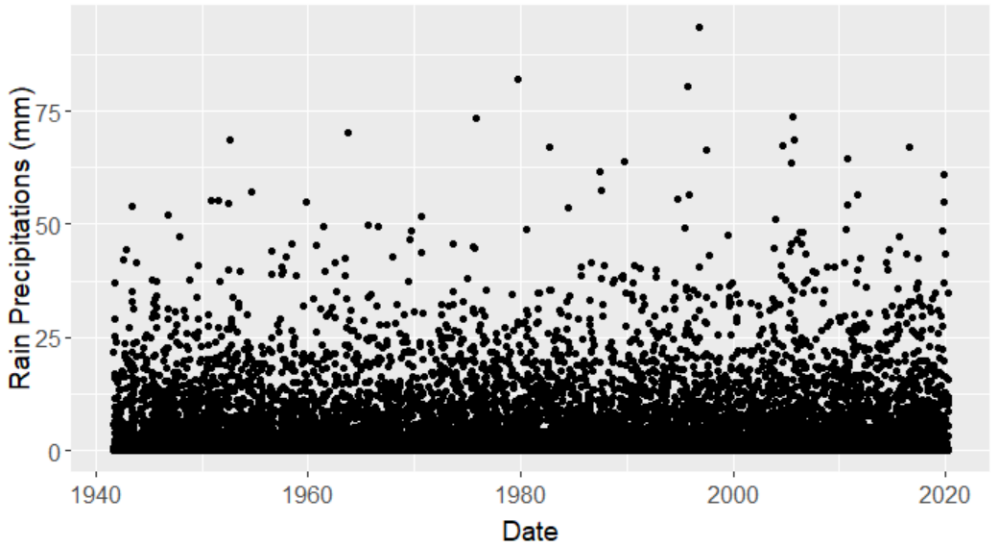

Figure 9: Daily Maxima of Rain Precipitations (mm) in Montreal, Canada

4.2 Analysis of the Maxima of daily maxima of rain precipitations in Montreal

As before, a first approach in the analysis of the extremes of rain precipitations is by the use of Block Maxima models. Since we are now considering the whole year and not just a month, the size of m is now 365; no observation is removed from the dataset unless it is null. Again using the *fgev*() function from the *evd* R package [5], we obtain the following maximum likelihood estimates: $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ = $(40.493, 11.718, -0.034)$ with respective 95% confidence intervals

$$\hat{\mu} = [37.592, 43.395], \hat{\sigma} = [9.623, 13.813], \hat{\xi} = [-0.200, 0.133].$$

Strictly looking at a sign function result on the maximum likelihood estimate of $\hat{\xi}$, the very binary test would attribute a Weibull distribution to the daily maxima of rain precipitation. However, from the magnitude of the parameter and its 95% confidence interval, we cannot reject the hypothesis that they are actually distributed according to a Gumbel ($\hat{\xi} = 0$) (or even a Fréchet ($\hat{\xi} = 0$) distribution). In this case, because of the proximity of $\hat{\xi}$ with 0, I fitted the model again optimizing the remaining parameters $\hat{\mu}$ and $\hat{\sigma}$ with the constraint that $\hat{\xi}$ equals zero. An ANOVA test on the original Block Maxima model and the one with forced $\hat{\xi} = 0$ yields a very high p-value (0.698), sign that we cannot reject the null hypothesis that the model with $\hat{\xi} = 0$ is an accurate simplification of the previous one. The new estimates become $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ = $(40.282, 11.597, 0)$ with corresponding 95% confidence intervals

$$\hat{\mu} = [37.603, 42.960], \hat{\sigma} = [9.618, 13.576], \hat{\xi} = [0, 0].$$

A comparison of the expected return levels is found in Figure 10.

Again, the next logical approach to this analysis is to examine the r-Largest Order Statistic models. Using the *ismev* R package[4], I obtained the maximum likelihood estimates displayed in Table 5 in the Appendix. We have again that in $rin\{2, 3, 4\}$, 0 is in the 95% confidence interval of the maximum likelihood estimate, reinforcing the belief that the maxima may be distributed according to a Gumbel distribution. The Quantile-Quantile plots for the $(r = 2, 3, 4)$-Largest

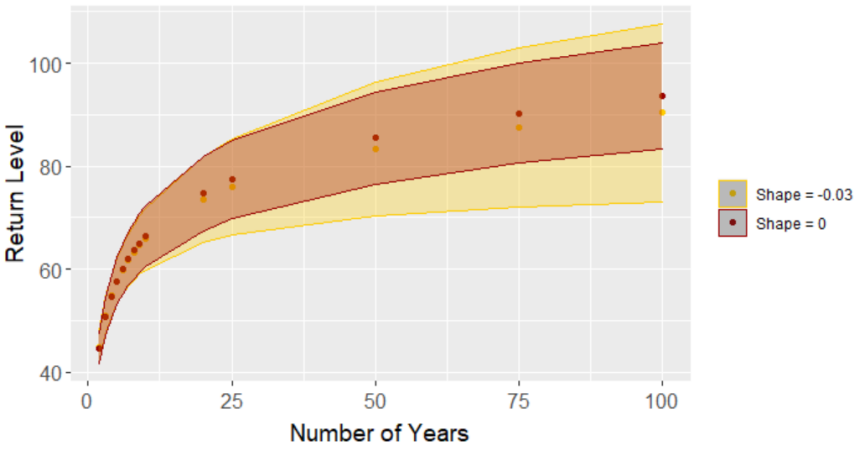

Figure 10: Comparison of the expected return levels with 95% confidence intervals for the models with free $\hat{\xi}$ and $\hat{\xi} = 0$.

Order Statistic models found in Figure 25 shows a seemingly good fit of the models to the block maxima, especially in the $r = 2$ case. The cases for $r \geq 5$ are not shown here because their Quantile-Quantile plots show departure from linearity, which implied that bias was induced to the model by modeling non-extreme values. The fitted expected return levels are shown in Figure 11 and in Table 3. The 95% confidence intervals are not plotted in Figure 11 for clarity purposes, but it is pretty clear from Table 3 that the confidence intervals do overlap, even with the Block Maxima fitted with the forced $\hat{\xi} = 0$.

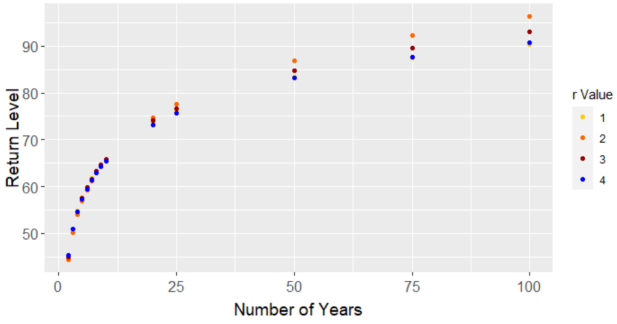

Figure 11: Expected return levels for maxima of rain precipitations (mm) in Montreal for r-Largest Order Statistic models.

The final approach is again the fitting of a generalized Pareto distribution to the threshold exceedances. I used the *gpdFit*() function from the fExtremes R package[6] to achieve this purpose. Although the model checking is not as satisfying and conclusive as was the case for the threshold exceedances models in the temperature analysis, I still get significant linearity in the scale parameter $\hat{\sigma}$ and a small (-0.015) but significant slope for the shape parameter $\hat{\xi}$ as u is varied.

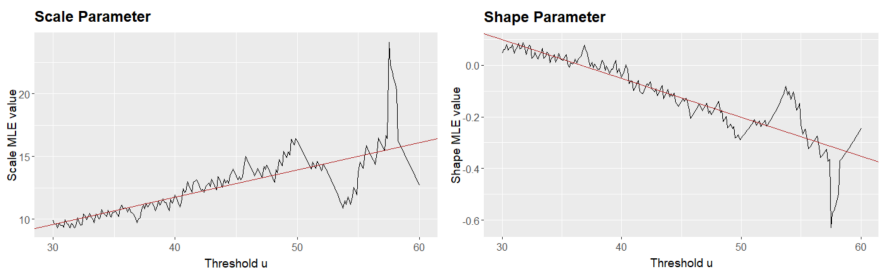

Figure 12: Linearity (in red) of the Scale and Shape Parameters for a Thresholds $u = 30$.

The resulting maximum likelihood estimates for the expected return levels computed for various u values are presented in Figure 26 in the Appendix. A comparison of the expected return levels computed with the various models above (found in Figure 13) shows nested 95% confidence intervals with very similar maximum likelihood estimate values. One could conclude that the models do seem to fit the extremes of rain precipitations well in Montreal.

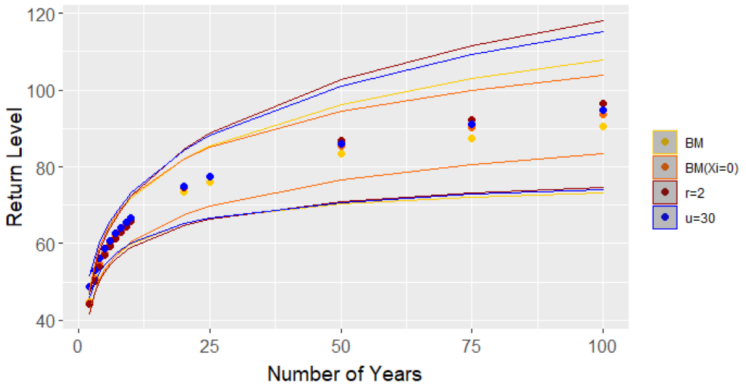

Figure 13: Comparison of the expected return levels of rain precipitations in Montreal (mm) for the best model within each approach.

Table 3: Montreal maximum likelihood estimates for the expected return levels in rain precipitations (mm) (with 95% C.I.).

Model	**10 years** (°C)	**100 years** (°C)
Block Maxima ($m = 365$)		
Free $\hat{\xi}$	65.887 ± 5.959	90.422 ± 17.278
Fixed $\hat{\xi} = 0$	66.380 ± 5.877	93.631 ± 10.269
r-Largest Order Statistic		
$r = 2$	65.731 ± 6.657	96.375 ± 21.751
$r = 3$	65.882 ± 6.054	93.013 ± 17.231
$r = 4$	65.394 ± 5.591	90.756 ± 14.693
Threshold Exceedances		
$u = 30$	66.647 ± 6.531	94.639 ± 20.575

4.3 Analysis of the Maxima of daily maxima of rain precipitations in Other Canadian Cities

In this section, the specific choices behind the models are a little less detailed and more attention is given to the comparison of results for Montreal, Québec, Toronto and Vancouver.

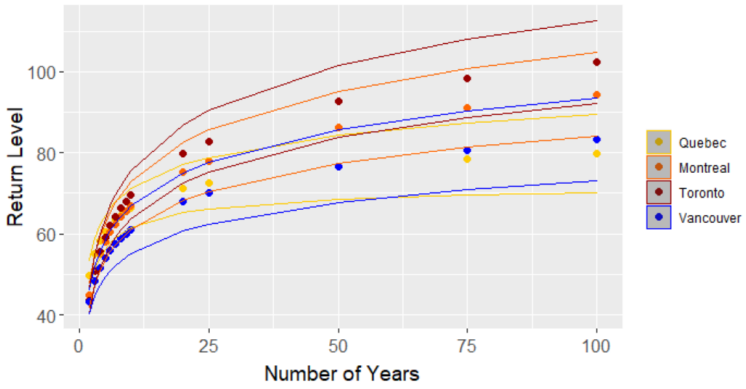

Figure 14: Expected return levels for maxima of rain precipitations (mm) in other canadian cities using Block Maxima models.

The same approach was used for each city, first fitting the Block Maxima model with block size $m = 365$. Since in all cases, 0 was in the 95% confidence intervals for the maximum likelihood estimate $\hat{\xi}$, I tested the model simplification $\hat{\xi} = 0$ with an ANOVA test with the null hypothesis that the latter model is an accurate simplification of the previous. The p-values for Toronto (0.233) and Vancouver (0.245) were high enough to not being able to reject the null hypothesis while the Québec maxima yielded a significant p-value (0.0495). Figure 14 thus contains the Block Maxima models fitted with forced $\hat{\xi} = 0$ for Montreal, Toronto and Vancouver while Québec was modeled with free $\hat{\xi}$.

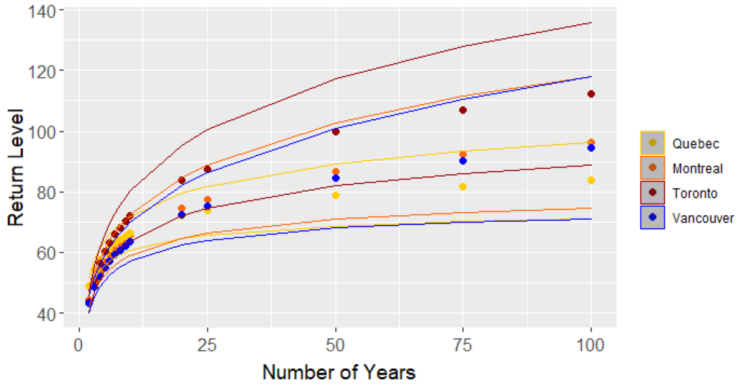

Figure 15: Expected return levels for maxima of rain precipitations (mm) in other canadian cities using r-Largest Order Statistic models.

Figure 15 shows the r-Largest Order Statistic models with the specific r value that yielded the best Quantile-Quantile plot on the ordered block maxima for a given city. As it can be seen in Figures 14, 15 and 16, ordering the maximum likelihood estimates for the expected return levels expresses that the city that seems to be the most likely to suffer from the larger daily precipitations is Toronto, followed by closely separated Montreal and Vancouver, and finally by Québec. However, the 95% confidence intervals notably overlap in all models, leading to the impossibility of a significant segregation between the return levels in the four cities.

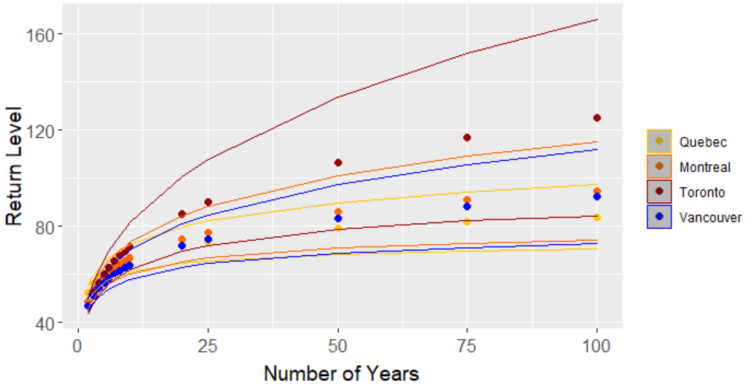

Figure 16: Expected return levels for maxima of rain precipitations (mm) in other canadian cities using Threshold Exceedances models.

5 Conclusion

Through this undergraduate project under the supervision of Professor Christian Genest, I have summarized the extreme value theory presented in chapters 1 to 4 of the textbook *An Introduction to Statistical Modeling of Extreme Values* by S. Coles [2] and applied classical results to the analysis of the maxima of temperature in July and daily rain precipitations in the cities of Montreal, Québec, Toronto and Vancouver.

I have come to the conclusion that the maxima of temperature across all four cities are distributed according to a Weibull distribution, which implies an upper end-point to the expected return levels. In other words, it is expected, since climate change or other trends were not detected in the data, that temperatures should never exceed some level in these cities.

I also noted a different behavior in the tails of the daily precipitations, closer to a Gumbel distribution which does not allow for infinite return level estimation. This implies that we should "expect the unexpected" and eventually observe some unprecedented levels of rain precipitations.

Although deeper analyses could be made, for instance taking into account dependence between observations at small lags of the time series or taking other environmental variables and transforming this into a multivariable analysis, I believe that the present project has been a very relevant and interesting introduction to extreme value theory and its concrete and computational practices in R.

Finally, the exact motivations for the choice of variables included in this project's analyses have remained dubious, as I have primarily focused on the statistical and computational aspects of such an analysis. It is however not hard to sympathize with the importance of similar analyses of temperature maxima on environmental data for regions affected by extreme drought, or conversely for regions affected by floods and where heavy rains are important predictors of sudden floods. I am thus particularly enthusiastic for the forthcoming statistical analyses of extremes that I will be able to pursue through master studies.

References

[1] Environment canada. https://climate.weather.gc.ca/historical_data/search_historic_data_e.html.

[2] S. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer, 2001.

[3] L. De Monte. Github repository. https://github.com/lambertdem/MATH410.

[4] J. E. Heffernan. ismev r package. https://cran.r-project.org/web/packages/ismev/ismev.pdf.

[5] A. Stephenson. evd r package. https://cran.r-project.org/web/packages/evd/evd.pdf.

[6] D. Wuertz. fextremes r package. https://cran.r-project.org/web/packages/fExtremes/fExtremes.pdf.

6 Appendix

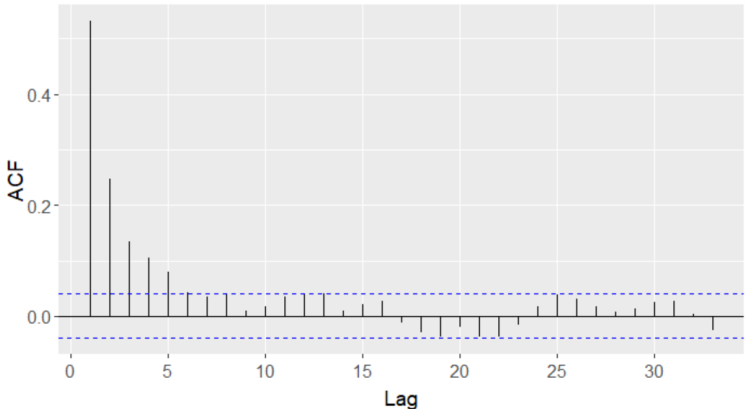

Figure 17: Autocorrelation Function plot of the daily temperatures in July time series in Montreal.

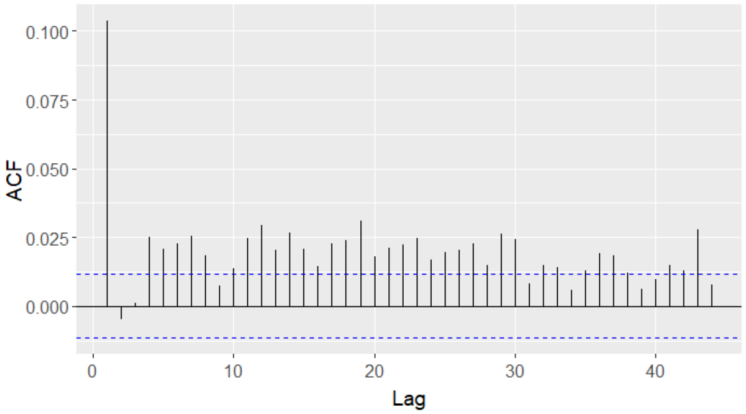

Figure 18: Autocorrelation Function plot of the daily rain precipitations time series in Montreal.

Table 4: Montreal maximum likelihood estimates for the GEV and gPd distributions (with 95% C.I.).

Model	**Location** ($\hat{\mu}$)	**Scale** ($\hat{\sigma}$)	**Shape** ($\hat{\xi}$)
Block Maxima			
$m = 31$	31.664 ± 0.421	1.737 ± 0.297	-0.359 ± 0.131
r-Largest Order Statistic			
$r = 2$	32.037 ± 0.332	1.590 ± 0.161	-0.368 ± 0.093
$r = 3$	32.171 ± 0.298	1.512 ± 0.122	-0.360 ± 0.078
Threshold Exceedances			
$u = 28$		2.780 ± 0.231	-0.333 ± 0.048
$u = 29$		2.413 ± 0.250	-0.323 ± 0.063
$u = 30$		2.214 ± 0.294	-0.350 ± 0.080

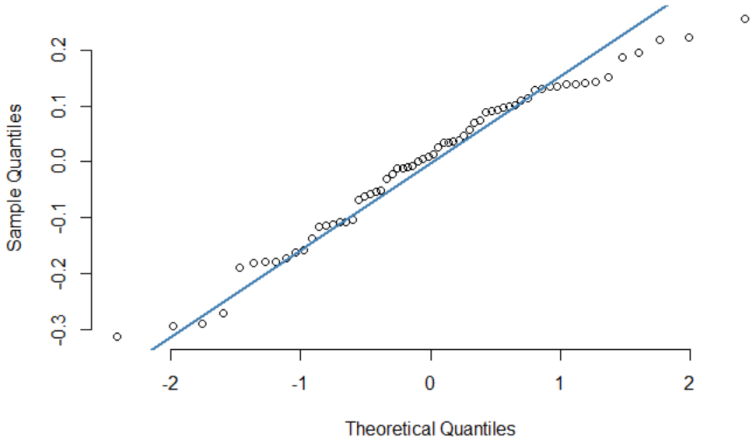

Figure 19: Normal QQ-plot for the residuals of the fitted scale estimate ($u = 28$).

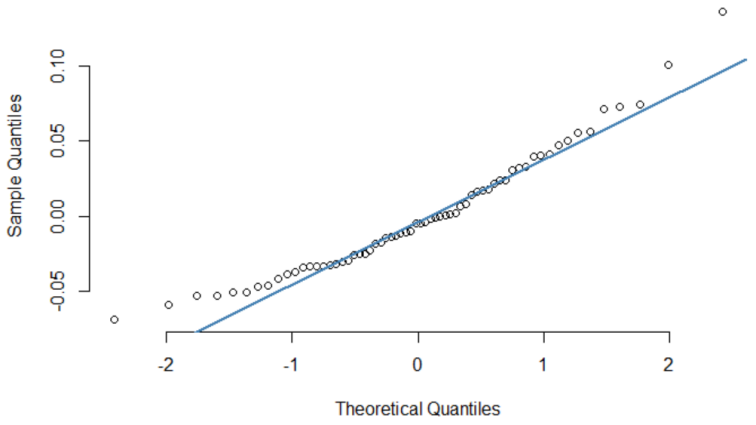

Figure 20: Normal QQ-plot for the residuals of the fitted shape estimate ($u = 28$).

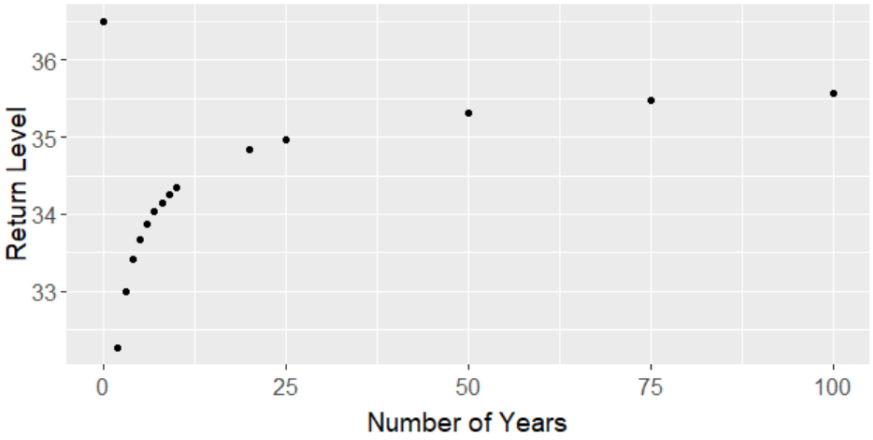

Figure 21: Block Maxima Return Levels for Maxima of Temperature in July in Montreal.

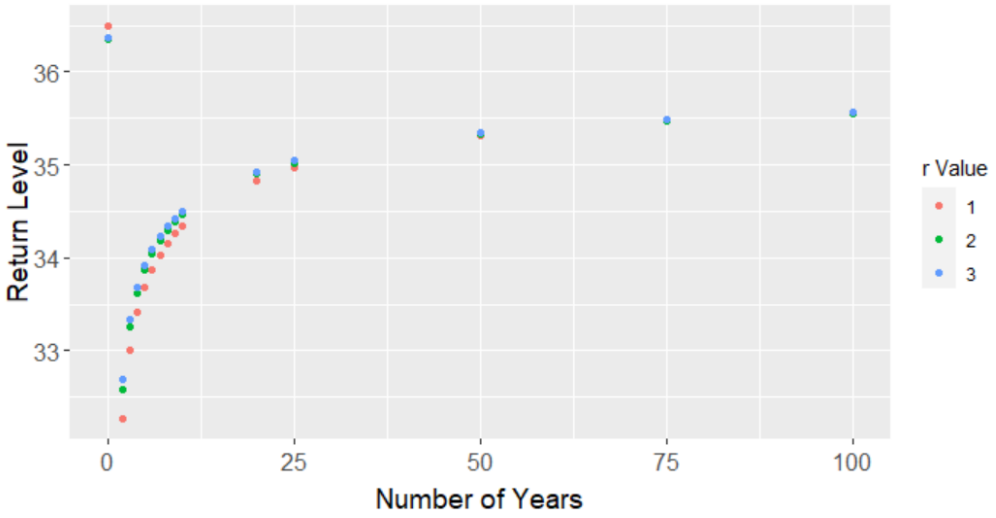

Figure 22: r-Largest Order Statistic Model Return Levels for Maxima of Temperature in July in Montreal.

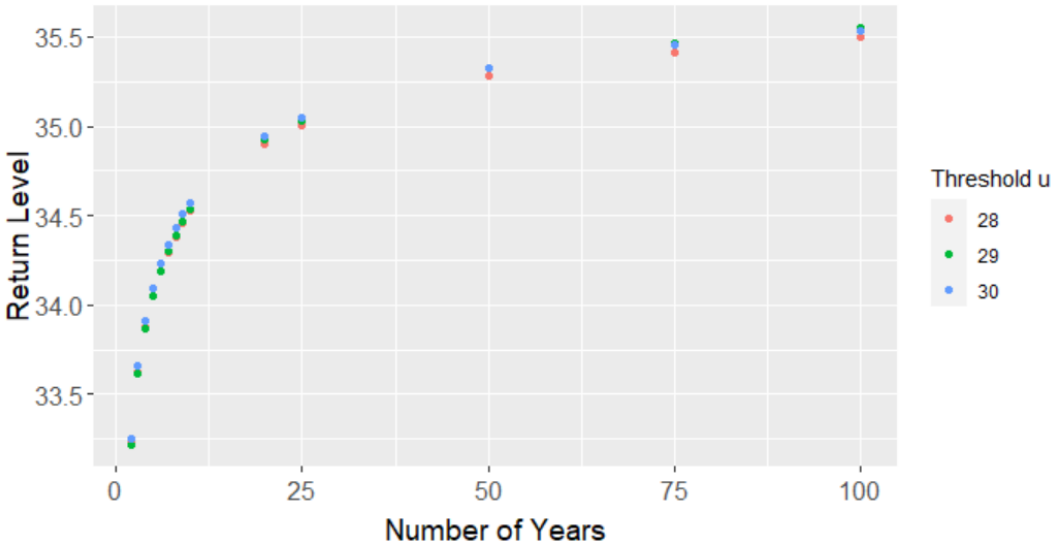

Figure 23: Threshold Models Return Levels for Maxima of Temperature in July in Montreal.

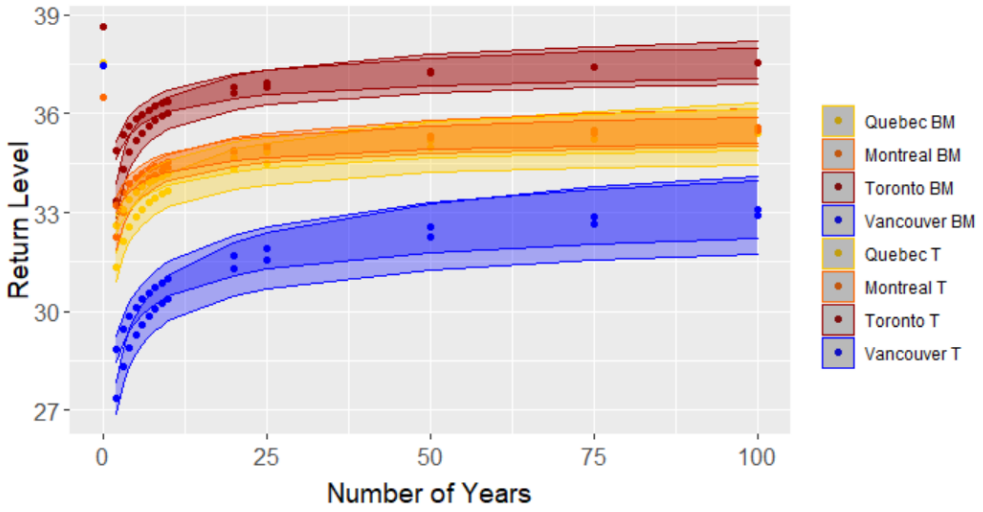

Figure 24: Return Levels for Maxima of temperatures of Cities in Canada in July (Block Maxima (BM) and Threshold (T) Models).

Table 5: Montreal Maxima Maximum Likelihood Estimates for the return levels (with 95% C.I.).

Model	**Location** ($\hat{\mu}$)	**Scale** ($\hat{\sigma}$)	**Shape** ($\hat{\xi}$)
Block Maxima ($m = 365$)			
Free $\hat{\xi}$	40.493 $\pm$ 2.901	11.718 $\pm$ 2.095	-0.034 $\pm$ 0.167
Fixed $\hat{\xi} = 0$	40.282 $\pm$ 2.679	11.597 $\pm$ 1.979	0
r-Largest Order Statistic			
$r = 2$	40.393 $\pm$ 2.157	10.470 $\pm$ 1.473	0.064 $\pm$ 0.136
$r = 3$	41.048 $\pm$ 2.086	10.794 $\pm$ 1.345	0.020 $\pm$ 0.108
$r = 4$	41.375 $\pm$ 1.969	10.615 $\pm$ 1.247	0.005 $\pm$ 0.089
Threshold Exceedances			
$u = 30$		9.895 $\pm$ 1.884	0.044 $\pm$ 0.143

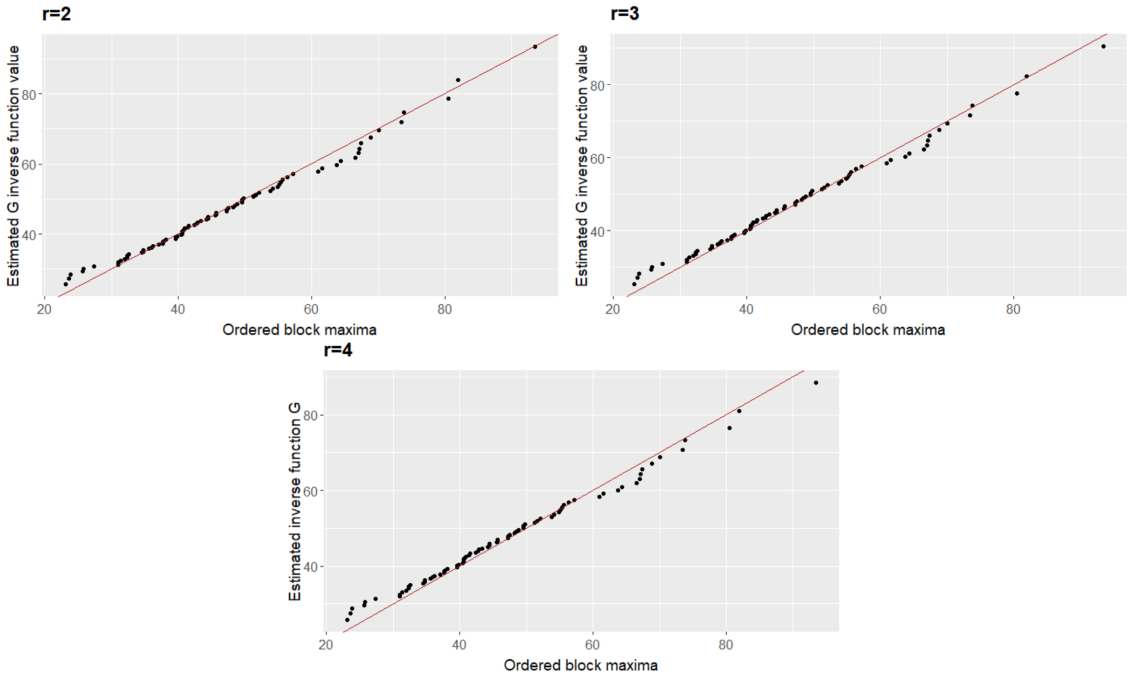

Figure 25: Quantile-Quantile plots for the r-Largest Order Statistic models for the maxima of rain precipitation in Montreal.

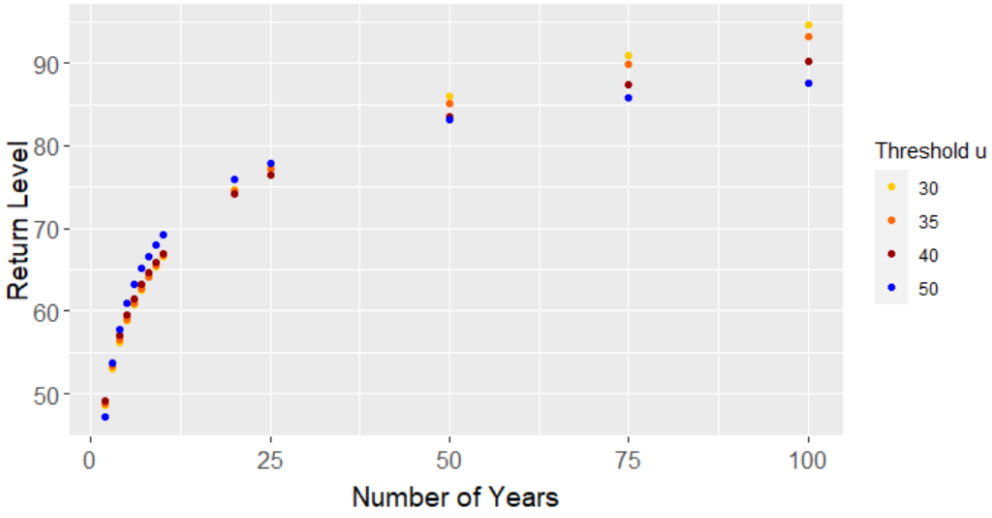

Figure 26: Return Levels for maxima of rain precipitations (mm) in Montreal.