Heavy-Tailed Series Analysis of the Air Quality Data: with Peaks over Threshold (POT) Approach

Project Report submitted

By

Anubhab Biswas

Reg.No 21MSMS02

Under the guidance of

Dr. BG Manjunath

The project submitted in partial fulfillment of the requirement for the degree of Master of Science in Statistics

From 2021 To 2023

At

UNIVERSITY OF HYDERABAD
SCHOOL OF MATHEMATICS AND STATISTICS
PROF. C R ROAD, GACHIBOWLI, HYDERABAD
TELANGANA, INDIA - 500046

Scanned with CamScanner

Certificate

This is to certify that the project work entitled "*Heavy-Tailed Series Analysis of the Air Quality Data: with Peaks over Threshold (POT) Approach*" submitted By Mr. *Anubhab Biswas* is a bonafide project work done under my supervision. It is being submitted in partial fulfillment of the requirements of degree in Master of Science in Statistics, University of Hyderabad.

(Signature)
Dr. BG Manjunath
School of Mathematics and Statistics
University of Hyderabad, Telangana

डॉ. बी.जी. मंजूनाथ / DR. B. G. MANJUNATH
सहायक प्रोफ़ेसर / Assistant Professor
गणित और सांख्यिकी संकाय
School of Mathematics & Statistics
हैदराबाद विश्वविद्यालय / University of Hyderabad
हैदराबाद / Hyderabad-500 046. तेलंगाना / T.S.

Dean
School of Mathematics and Statistics
University of Hyderabad

संकाय-अध्यक्ष / DEAN
गणित और सांख्यिकी संकाय
School of Mathematics & Statistics
दराबाद विश्वविद्यालय/University of Hyderabad
दराबाद/Hyderabad-500 046. तेलंगाना, T.S.

Date: 23-May-2022
Place: Hyderabad

Scanned with CamScanner

Declaration

I *Anubhab Biswas* hereby declared that the project work entitled "*Heavy-Tailed Series Analysis of the Air Quality Data: with Peaks over Threshold (POT) Approach*" is an original record of studied and bonafide work carried out by me under the guidance of *Dr. BG Manjunath*, School of Mathematics and Statistics, University of Hyderabad, Telangana, India and has not been submitted by me elsewhere for the award of any degree, diploma, title or recognition before.

Anubhab Biswas
23/05

Name of the Student: **Anubhab Biswas**
Roll No: **21MSMS02**
School of Mathematics and Statistics
University of Hyderabad, Telangana

Scanned with CamScanner

Acknowledgement

I take this opportunity to express my profound gratitude and deep regards to the following for their generous assistance and incorporation of fresh ideas in the completion of this Dissertation Paper.

I would like to thank my institution, University of Hyderabad, Hyderabad for giving me chance to do this project.

Most importantly, I would also like to express my sincere gratitude towards my project guide Dr. BG Manjunath for being an idealistic channel without whose paramount guidance and utmost care, this paper would not have been possible.

I would like to extend my gratitude to University Library, for having provided various reference books and magazines related to my project.

Lastly, I would like to thank each and every individual who directly or indirectly assisted me in the completion of this dissertation paper, especially my Parents and Peers, including my batch mates and seniors who supported me throughout my project.

Anubhab Biswas
23/05

Anubhab Biswas
21MSMS02
M.Sc. Statistics
School of Mathematics and Statistics
University of Hyderabad

Scanned with CamScanner

Heavy-Tailed series analysis of the Air Quality data: with Peaks over Threshold (POT) approach

Anubhab Biswas

June 7, 2023

Abstract

Objective: In this project, the main objective is to understand the limiting behavior of the AQI for PM2.5 both in terms of the worst case scenario and the average cluster time over a high threshold to understand how bad is the condition of Air pollution and how necessary it is to implement preventive measures in order to decrease health hazards.

Methodology: The study has been done on the Air Quality data of New Delhi to comment about the tail behavior of the underlying stochastic process in the background. Two approaches have been applied to the data - one with the regular ARMA-GARCH modeling by taking the empirical bootstrap distribution of the residuals. In the

other approach, the tail process and tail measure method has been applied, by taking the simplest distribution for modeling peaks over threshold - the Generalized Pareto distribution.

Conclusion: We have seen the behavior of the average time spent over threshold, the frequency of crossing the threshold from both the approaches and have concluded that, irrespective of the method applied, the AQI of New Delhi has been suffering serious issues since both the methods show a tendency to cross the high threshold set by the Govt. very often, at least once in every 2.5 hours! We have also observed the worst of the worst case (a high threshold to cross once in every 1 year if the process goes on as it is) turns out to be over 3500 in terms of AQI if we want to make sure that the process crosses this threshold once in every year in expectation! So, we strongly suggest that some serious measures to be taken to the air quality of New Delhi if we want to decrease the return level.

Contents

1 Introduction

Air pollution has been a major issue in India since the growth of metropolitan cities and factories all over the country. Many preventive measures have been suggested in the past to bring down the air quality index, for example, the odd-even car rule in New Delhi . But in spite of these measures as per the data of 2019, 21 out of 30 most polluted cities of the world have been from India[1]. Thus, it is of no surprise that the air pollution needs to be taken care of both statistically and realistically in order to prevent the deadly health hazards due to air pollution. That has been the major motivation behind this thesis.

1.1 How is AQI calculated?

The amount of air pollution is best measured by the Air Quality Index given by,

$$Air\,Quality\,Index\,(I) = \frac{I_{high} - I_{low}}{C_{high - C_{low}}}(C - C_{low}) + I_{low}$$

where $C : the\,pollutant\,concentration$

$C_{high} : the\,concentration\,breakpoint\ \geq C$

$C_{low} : the\,concentration\,breakpoint\ \leq C$

$I_{high} : index\,breakpoint\,corresponding\,to\,C_{high}$

$I_{low} : index\,breakpoint\,corresponding\,to\,C_{low}$

The Air Quality Index in India has the following categories[2]:

[1] [source : Wikipedia]

[2] Indian Express; Nov 6, 2022

0-50	50-100	100-200	200-300	300-400	400-500
Good	Satisfactory	Moderately polluted	Poor	Very Poor	Severe

The Central Control Room for Air Quality Management measures various kinds of particulate matters and harmful gases present in the air at different stations all over India. We will be focusing mainly on Delhi since its one of the most polluted places in terms of air quality and needs to be taken care of immediately. Our work is based on solely on the particulate matter PM2.5 which is considered to be the deadliest of all. The AQI cutoff set for the PM2.5 is 60 as per the Central Pollution Control Board.

1.2 Objective

We have hourly Air quality index data available on PM2.5 for 39 stations of Delhi. Since this is a time series data, each observation is mildly dependent on the previous ones. So, once the data goes above a given threshold (for example, a threshold 60 set by the pollution control board) it stays above the threshold for sometime, and then comes down. Thus, the data above some threshold looks like clusters. What we are interested here is, for how much time the air quality index stays above threshold, or how do the above-mentioned clusters behave in the long run i.e., limiting behavior of these clusters above the threshold. This has been dealt with by two different approaches viz. data-driven empirical bootstrap approach and the limiting distributional approach using tail process and tail measure which will be discussed in the methodologies section. We will also like to shed some light on how does the process behave over the sequence of thresholds as time increases and threshold increases, i.e., if the worst scenario (severe) happens, how long is it going to last in the long run. This final part will be proposed using just the distributional theory since the data-driven approach does not reach there. The data has been obtained from the official website of Central Pollution Control Board : cpcb.nic.in

1.3 Brief literature

Time series analysis for heavy-tailed data has been a topic of major interest since the last few decades. The initial work in terms of usual time series analysis was proposed by **Engle(1982)** where he introduced the conditional heteroscedastic model for the first time. **Bollerslev(1986)** refined the model by introducing the GARCH and since then a lot of modifications has been done in that model which is currently used a lot for modeling financial processes. Taking different Autoregressive models to model the mean has been combined with GARCH to model the second moment such as the Threshold Autoregressive (TAR) and Self-Exciting TAR (SETAR) by **Tong(1983)** where the time series is modeled differently below and above some threshold making it a regime switching model. The GARCH has also undergone various transformations such as the exponential GARCH (E-GARCH) by **Nelson(1991)** which is used to overcome some of the issues faced while modeling financial data. But none of the models can talk about the limiting behavior of the process or the extreme behavior of the process. This issue can be addressed by the Extreme Value Theory by considering the Peaks over Threshold method. The major work on this topic for a time series process has been done initially by **Pickands(1975)** where the Generalized

Pareto distribution was introduced to model exceedances over a high threshold. Since then, Leadbetter in his various papers has introduced results in order to understand the limiting distribution of the extremes for time series in the block maxima approach by introducing extremal index which captures the deviation from the limiting distribution for independent sequences. But for the PoT approach this has been extended by **Hsing(1988)** where he has proposed the Point process approach to model the limiting behavior of exceedances in terms of a compound Poisson Point process model. The most recent work for modeling extremes of a time series process is due to **Soulier(2021)** where the overall approach is to define a tail process and obtain the limiting behavior of that in terms of extremal index in a direct way along with the average cluster time. The limiting behavior of the original data coincides with it once some weak conditions are satisfied. For our data, we have applied the ARMA-GARCH and tail process approach using a GPD.

2 Methodologies

2.1 Time Series

The AQI data is a time series data. One of the major differences that makes a Time Series data different from other types of data is that the observations are not at all independent and identically distributed (iid) (Shumway(2000)). In fact, the current observations are mildly or heavily dependent on the previous ones. So the usual statistical inference does not work here and hence we cannot apply the general modeling techniques here to do estimation, hypothesis testing, or forecasting. The closest we can get to the iid condition is stationarity which we will define as follows:

A time series process $\{X_t : t = 0, 1, 2, ...\}$ is said to be **stationary** if:

1. $E(X_t) = \mu(constant)$
2. $V(X_t) = \sigma^2(constant)$ and,
3. $Cov(X_t, X_{t+h}) = Cov(X_s, X_{s+h}) = \gamma(h)$ is fixed for a given h (which is known as lag) for every s and t.

So, we have just sacrificed the independent condition of the data and this is the nearest we can go to the iid scenario. Most of the literature in time series is based on the assumption on stationarity and the part of the analysis we are interested in, will also consider the same.

Here we will be interested in a specific type of modeling of a time series data called the ARMA-GARCH. We will define them for a brief understanding of what we are about to do. The detailed scheme is provided in the section 2.5. We start with the definition of a white noise process. A time series process $\{X_t : t = 0, 1, 2, ...\}$ is said to be **white noise process** with mean 0 and variance σ^2 if, (i) $E(X_t) = 0$, (ii) $V(X_t) = \sigma^2(constant)$ and, (iii) $Cov(X_t, X_s) = 0$ for every $s \neq t$.

Now, a time series process $\{X_t\}$ with mean μ is said to be **ARMA(p,q)** , if the process can be expressed in the following way:

$$X_t = \mu + \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + ... + \theta_q \epsilon_{t-q}$$

where $Z_t \sim WNP(0, \sigma^2)$. It should be noted here that we call an ARMA(p,q) process to be **invertible** if, all the zeroes of $1 - \sum_{i=1}^{p} \phi_i y^i$ lies outside the unit circle. The invertible condition is important in the sense that, if a process is invertible, then the coefficients and hence the dependency on previous values will decrease with more lag. This is desirable since a process where dependency increases between time series variables which are far apart does not make any sense. Next we move on to the GARCH model.

A GARCH model **(Bollerslev(1986))** becomes important when the unconditional variance of the process remains constant, but the process at a time given the previous past has a conditional variance which is not constant. So, although the process has a constant unconditional variance, there are sudden bursts of high variability depending on the previous past and these bursts tend to appear in clusters. Economists call this phenomenon

volatility clustering. So if we consider a model which has both ARMA and GARCH components, it would stabilize both in mean and in variance by taking them into account. For a Time series process $\{X_t\}$ a typical $ARMA(p,q) - GARCH(p',q')$ representation is given by,

$$\zeta_t = \mu + \sum_{i=1}^{p} \phi_t \zeta_{t-i} + \epsilon_t + \sum_{j=1}^{q} \theta_t \epsilon_{t-j}$$

$$\epsilon_t = \sigma_t z_t$$

$$\sigma_t^2 = \omega + \sum_{i=1}^{p'} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{q'} \beta_j \sigma_{t-j}^2$$

where,ϵ_t is the white noise process, and further,σ_t^2 is the unconditional variance. The random variable Z_t is considered to be iid with mean 0 and variance 1. This concludes the initial things that we need to understand about time series processes. We move on to the extreme value theory for time series data in the next sections.

2.2 Time series of Extremes

The time series of Extremes is done in two different ways viz., (i) the block-maxima approach and (ii) the peaks over threshold approach just like the usual Extreme Value Theory. Although we will be working with the latter case here, we give a brief account of the block maxima method as well. It should be noted that the serial dependence nature of the time series affects both the magnitude and the qualitative behavior of the extremes. Hence, a modification on the standard method for extremes is required just like one required for a usual time series data. Extreme events in the real world where the data is a time series one exhibits an unique behavior – it always occurs in clusters. In our context, we can think of this in the following way — if the air quality becomes bad for some reason it will remain bad for some time before coming back to a satisfactory level. So, the first time it becomes bad (crosses the threshold) it influences the air quality of the next few time points.

2.3 Block maxima method

Suppose we have a stationary time series data $\{X_t, t = 0, 1, 2, ..., n\}$. We partition our data into k_n different blocks of sizes r_n, where $k_n = [\frac{n}{r_n}]$, where r_n is such that $r_n = o(n)$ as $n \to \infty$. Let us call those partitions J_j and $J = \{1, 2, ..., n\}$. So, $J_j = \{(j-1)r_n + 1, ..., jr_n\}$ where $j = 1(1)k_n$. Then if we define,

$$M_n = X_{(n)} = max\{X_1, ..., X_n\}$$

and

$$M_{r_n} = max\{X_{(j-1)r_n+1}, ..., X_{jr_n}\}$$

then under suitable conditions for a sequence of thresholds u_n,

$$P[M_n \leq u_n] = (P[M_{r_n} \leq u_n])^{k_n} + o(1)$$

as $n \to \infty$

Now, we define a condition called the $D(u_n)$ **condition** as follows:

For all $A_1 \in I_{1,l}(u_n)$, $A_2 \in I_{l+s,n}(u_n)$ and $1 \leq l \leq n-s$,

$$|P[A_1 \cap A_2] - P(A_1)P(A_2)| \leq \alpha(n,s)$$

and $\alpha(n, s_n) \to 0$ as $n \to \infty$ for some positive integer sequence s_n such that $s_n = o(n)$ where,

$$I_{j,k}(u_n) = \{\{M(I) \leq u_n\} : I \subseteq \{j, ..., k\}\}$$

The above condition actually tells us that if we consider the blocks to be of some short distance apart from each other, then as n increases, they can approximately become independent of each other. So it limits the long-range dependence in some sense. Then we can turn to the extremal limit theorem by **Leadbetter(1974)** to obtain the limiting distribution of extremes which suffice this condition. It says that, if the $D(u_n)$ condition holds with $u_n = a_n x + b_n \, \forall x$, then for our stationary sequence $\{X_t\}$,for which there exists constants $a_n > 0$ and $b_n \in \mathbb{R}$, and a non-degenerate distribution function G such that,

$$P[\frac{M_n - b_n}{a_n} \leq x] \overset{D}{\to} G(x)$$

then G is an Extreme Value distribution function.

Note that the above result is similar to the one we know for the extreme value distribution for independent sequences. In fact, it is closely related to the case of independent sequences as proved by **Leadbetter(1983)** in the following way:

If there exists sequences of constants $a_n > 0$ and $b_n \in \mathbb{R}$,and a non-degenerate distribution function $\widetilde{G}$ such that,

$$P[\frac{\widetilde{M_n} - b_n}{a_n} \leq x] \overset{D}{\to} \widetilde{G}(x)$$

as $n \to \infty$, and if $D(u_n)$ condition holds with $u_n = a_n x + b_n \, \forall x$, such that $\widetilde{G}(x) > 0$ and if $P[\frac{M_n - b_n}{a_n} \leq x]$ converges for some x, then,

$$P[\frac{M_n - b_n}{a_n} \leq x] \overset{D}{\to} G(x) := \widetilde{G}^{\theta}(x)$$

for some $\theta \in (0, 1)$. Here, $\widetilde{M_n}$ is the maxima for the associated independent sequence.

The above constant θ is known as the extremal index. It is easy to see that if $\theta = 1$, then it is basically the case of the independent sequence. Thus we can see that the limiting distribution of the extremes is not the same for the dependent time series data as of the one with independent sequences. But, under some mild conditions, they are closely related.

2.4 Two approaches for the Peaks over threshold (poT) method

In this project we are interested in two approaches for the peaks over threshold method and at the end we will be comparing their results. The **first**

approach is the ARMA-GARCH approach of fitting an ARMA-GARCH model to the data and get empirical distribution of the residuals from which we will generate a bootstrap procedure to get the estimates of the exceedance probability defined as, $EP(t) = P(X_t > 60) = \overline{F_{X_t}(60)} = 1 - F_{X_t}(60)$ at different time points t. Hence, this process is distribution-free in the sense that we will be using the bootstrap estimates of empirical distribution function – a non-parametric entity. We will also obtain a model for the number of times the process crosses the threshold and the amount of time it stays there using the estimates obtained from the ARMA-GARCH model.

The **second approach** is the tail process and tail measure approach where we will fit a generalized Pareto distribution to the data given that the first time point yields an extreme value, which acts as the tail measure. Now, once we have the fitted GPD, we will be obtaining the estimates of exceedance probability, extremal index and the behavior of the process above threshold. We will also comment on the worst case scenario by compiting the return level for the fitted GPD, something which cannot be done using the empirical ARMA-GARCH model. The detailed discussion is given in section 2.6. In the conclusion, we will also provide a brief comparison between the two methods.

2.5 Empirical bootstrap Approach

Let, $\{X_t\}$ be a random process that denotes the AQI of New Delhi for the t^{th} hour where $t = 0$ corresponds to 12:00 am on 1st January 2019. Since, this is a time series process, it obviously is expected to contain some kind of trend and seasonal component. If we want to study the stochastic behavior of the process, we have to get rid of these deterministic quantities. So, we shall write $X_t = m_t + \zeta_t$, where the term m_t takes care of the trend and seasonality component if its present in the process. The ζ_t term hence represents the actual stochastic component of our random process. We will be modelling this ζ_t,

We continue with the assumption that ζ_t follows a stationary, and invertible ARMA-GARCH model, since without that we cannot go further into the analysis part. With the assumption we can consider stabilizing and taking into account of the mean and the heteroscedastic behavior which is expected to be present in this kind of a data. After we are done with modeling, the next task will be to get an estimate of the exceddance probability where we take the threshold to be 60 which is set for our PM2.5 data. Recall that the exceedance probability at time t is defined as,

$$EP(t) = P[X_t > 60]$$

We will initially start with the following ARMA-GARCH model:

$$\zeta_t = \mu + \sum_{i=1}^{p} \phi_t \zeta_{t-i} + \epsilon_t + \sum_{j=1}^{q} \theta_t \epsilon_{t-j}$$

$$\epsilon_t = \sigma_t z_t \qquad 1.1$$

$$\sigma_t^2 = \omega + \sum_{i=1}^{p'} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{q'} \beta_j \sigma_{t-j}^2$$

where, ϵ_t is the white noise, and further, σ_t^2 is the unconditional variance. Z_t is considered to be iid with mean 0 and variance 1. Moreover $0 < \alpha_i, \beta_j < 1$, $\forall i = 1(1)p'$, $\forall j = 1(1)q'$

Here the assumptions that we have on the model are :

- m_t is differentiable and the derivative is continuous everywhere in the positive real line, i.e., the trend-seasonality component is smooth, which again implies that it can be locally approximated by a polynomial function.
- All the zeroes of $1 - \sum_{i=1}^{p} \phi_i y^i$ lies outside the unit circle and the same is true for all the zeroes of $1 - \sum_{j=1}^{q} \theta_j y^j$. This assumption takes care of our initial assumption that the ARMA process is stationary and invertible.
- $a > 0$ and $P \equiv \sum_{i=1}^{p'} \alpha_i + \sum_{j=1}^{q'} \beta_j < 1$, This ensures that the conditional variance i.e., σ_t is positive and weak-ergodic stationary.

Once we have the data, initially we have to get rid of the trend-cycle component with the help of a local smoother, thanks to our first assumption and hence obtain the irregular stochastic component. We will do this task by using the Nadaraya Watson Kernel smoothing with some suitable bandwidth as suggested by **D.Draghicescu**. The Nadaraya-Watson Kernel estimator at a time point t given by,

$$\widehat{m(t)} = \frac{\sum\limits_{i=1}^{n} K_h(t-t_i)y_i}{\sum\limits_{i=1}^{n} K_h(t-t_i)}$$

where $K_h(.)$ is an appropriate Kernel with bandwidth h.

Note that for the missing observations we have not applied any imputation methods since we have a huge chunk (from **1st January 2019, 00hrs** to **11th February 2023, 8pm**) of hourly AQI data, and not having observations for some hours doesn't really affect the results. But, it should be noted that we can always use imputation methods to obtain them. One of those has been suggested by using the trend-cycle Kernel smoother itself in (**Albano2(2020)**). Another way to do the imputation might be by taking the spatial estimates obtained from nearby weather stations at the corresponding time points. It should be noted that we could have used an ARIMA model and the method of differencing to model the stochastic component. But if one wants to do a spatial analysis as well to the data since we have observations from 39 different stations, then the differencing method is not applicable since we have to use a smoothing function for the trend in such a case as mentioned in (**D.Draghicescu**). So we stick to the local polynomial smoothing method to estimate the trend-cycle component.

After removing the trend-cycle we will be fitting an appropriate ARMA-GARCH to the stochastic component. We will be choosing the numbers

p, q, p', q' by calculating RMSE in terms of forecasting for different values of them. We will also be taking into account of the AIC while doing so. The next step is the estimation of the parameters and once we have these estimates, we will be bootstrapping from the empirical distribution of residuals to get our estimate of exceedance probability using the scheme suggested by (**Albano1(2020)**) given by the following algorithm:

1. Estimate the residuals and heteroscedasticity from the fitted model as:

$$\hat{\epsilon_i} = \hat{\zeta_t} - \mu - \sum_{i=1}^{p} \hat{\phi}_t \hat{\zeta_{t-i}} - \sum_{j=1}^{q} \hat{\theta}_t \hat{\epsilon_{t-j}}, \quad i = max\{p, q\} + 1, ..., n$$

$$\hat{\sigma_t^2} = \hat{\omega} + \sum_{i=1}^{p'} \hat{\alpha}_i \hat{\epsilon_{t-i}^2} + \sum_{j=1}^{q'} \hat{\beta}_j \hat{\sigma_{t-j}^2}, \quad i = max\{p, q, p', q'\} + 1, ..., n$$

2. For $i = max\{p, q, p', q'\} + 1, ..., n$, obtain

$$\hat{\eta_i} = \frac{\hat{\epsilon_i}}{\hat{\sigma_i}}$$

   which is a candidate estimate of z_i

3. But it should have 0 mean and unit variance. So, we have to standardize our estimate as,

$$\widetilde{\eta_i} = \frac{\hat{\eta_i} - \bar{\hat{\eta_i}}}{sd(\hat{\eta_i})}$$

4. Next we obtain the empirical distribution function of $\widetilde{\eta}_t$ as,

$$\mathcal{F}_T(x) := \frac{1}{T} \sum_{i=1}^{n} 1(\widetilde{\eta}_i \leq x)$$

We will bootstrap from this estimated empirical distribution. We generate a bootstrap process by taking B number of $\widetilde{\eta}$'s for each time point re-sampled from the above empirical distribution. Let us denote the j^{th} bootstrap observation of $\widetilde{\eta}$ at the t^{th} time point as η^*_{tj}. Then, we get the bootstrap estimate of the exceedance probability at each time point from the following algorithm:

1. Get the bootstrap residuals as, $\epsilon^*_{tj} = \widehat{\sigma}_t \eta^*_{tj}$
2. The bootstrap stochastic component is obtained as : $\zeta^*_{tj} = \widehat{\mu} + \sum\limits_{i=1}^{p} \hat{\phi}_t \hat{\zeta_{t-i}} + \sum\limits_{k=1}^{q} \hat{\theta}_t \hat{\epsilon_{t-k}} + \epsilon^*_{tj}$
3. Finally the bootstrap process is obtained as : $Y^*_{tj} = \widehat{m(t)} + \zeta^*_{tj}$

The bootstrap estimate of exceedance probability at time point t is given by, $\widehat{EP(t)} = \frac{1}{B} \sum\limits_{j=1}^{B} 1(Y^*_{tj} > 60)$. Once we have the estimate of exceedance probability, we can have an idea of how often the process crosses the given threshold set by the pollution board. Our immediate job is to see how well the exceedance probabilities act as a classifier for the original data if we take the event "over the threshold 60" as success and "below the threshold" as failure. We will use the ROC plot in order to comment on the performance of our estimated exceedance probability as a binary classifier for the mentioned scenario. The ROC curve plots the false positive rate vs the true positive rate where,

$$True\,positive\,rate = \frac{True\,positives}{True\,positives + False\,negatives}$$

and

$$False\,positive\,rate = \frac{False\,positives}{False\,positives + True\,negatives}$$

. So, it can be viewed as a plot of the proportion of correct predictions for the positive class vs the proportion of wrong ones in the negative class. So, in an ideal scenario, the first proportion should be 1 and the later one should be 0 and the plot should be at the co-ordinate $(0, 1)$. If we check the area under the curve for such a case, it should be 1. Hence, the closer the area under the curve (AUC) is to 1, the better it works as a binary classifier.

The above work concludes our non-parametric approach to see exceedance probability and get useful results to comment about the PM2.5 emission process by quantifying and fitting a curve to comment about the behavior of how likely it is to cross the given threshold, or how often it will be crossing threshold creating health hazard even if we increase the threshold upto severe deterioration of air quality index level(400). Our next approach will be to do the extremal analysis of the same data using the tail process approach by using the tail measure. It is through this measure that we will be able to comment on the limiting behavior of the underlying time series process along with the understanding of the scenario about when will the severe condition prevail making the place prone to health hazards in the long run.

2.6 Tail process and tail measure approach

In this section we will be discussing the tail process and tail measure approach to a heavy-tailed time series data. We begin the discussion by the fact that GARCH processes are associated with heavy-tailed properties and can be thought about being regularly varying. Then we move on to the study of regularly varying processes which in turn leads us to tail measures and tail processes. Once we have the tail processes, we can use them to understand the clusters (specifically the clusters above threshold in this case) and their limiting behavior.

2.6.1 From GARCH to regular variation

Note that we have used the GARCH model for modeling the AQI data since it exhibits conditional heteroscedastic behavior. As proved by (**Basrak(2002)**), the GARCH processes exhibit an important and unique property — their finite dimensional distributions are multivariate regularly varying due to which the higher dimensional moments of this processes do not exist. So, let us see what does this regular variation mean and what does it signify.

In the univariate setup, a function $f : [a, \infty) \to \mathbb{R}$ is said to be **regularly varying at infinity** if,

$$\underset{x\to\infty}{lim} \frac{f(tx)}{f(x)}$$

exists in $(0, \infty)$ for every $t > 0$. If f is a Lebesgue measurable function, defined on $[0, \infty)$ such that it is regularly varying for all t in a set of positive Lebesgue measure, then $\exists \gamma \in \mathbb{R}$, such that $\underset{x\to\infty}{lim} \frac{f(tx)}{f(x)} = t^{\gamma}\ \forall t > 0$. In such a case, f is said to regularly varying at infinity with index γ. Now a couple of points should be noted.

- If $\gamma = 0$, then we say f is slowly varying at infinity.
- As a result of the above nomenclature, it follows that, a function $f : [0, \infty) \to \mathbb{R}$ is regularly varying at infinity with index γ iff $\exists$ a slowly varying function l such that, $f(x) = x^{\gamma} l(x)$.

Next we define the **regular variation of sequences** in the following way: A sequence $\{c_n\}$ of positive numbers is said to be regularly varying if $\exists$ a function $\psi : [0, \infty) \to \mathbb{R}$ such that $\forall t > 0$, $\underset{n \to \infty}{lim} \frac{c_{[nt]}}{c_n} = \psi(t)$. Similar to that of functions, if the above is true then $\exists \rho \in \mathbb{R}$, such that $\psi(t) = t^{\rho}$and this implies that a function f defined as $f(x) = c_{[x]}$ is regularly varying with index ρ.

Now that we have the definition of regular variation for sequences, we can go on to define what does regular variation means in terms of a random variable. A random variable X defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is said to be regularly varying with tail index $\alpha > 0$ if $\overline{F_{|X|}(x)} = 1 - F_{|X|}(x)$ is regularly varying at infinity with index $-\alpha$ and $\exists p_X \in [0, 1]$ such that

$$\underset{x \to \infty}{lim} \frac{\mathbb{P}(X > x)}{\mathbb{P}(|X| > x)} = p_X$$

Here the parameter p_X is known as the **extremal skewness** of X.

Recall that a realization of a time series process consists of a random vector where the components are not independent. Hence the univariate definition of regular variation will not suffice there. We need a multivariate setup to get that. Suppose $\underset{\sim}{X}_{d\times 1}$ is a random vector. It is called regularly varying if there exists a non-zero $\mathcal{B}_0$(the class of all sets which are separated from $\underset{\sim}{0}$ i.e., included in the complement of an open neighborhood around the null map $\underset{\sim}{0}$)-boundedly finite measure $\nu_{\underset{\sim}{X}}$ on $\mathbb{R}\backslash\{\underset{\sim}{0}\}$(exponent measure of $\underset{\sim}{X}$), and a scaling sequence $\{c_n\}$ such that,

$$n\mathbb{P}(c_n^{-1}\underset{\sim}{X_n} \in .) \overset{\mathcal{B}-vaguely}{\longrightarrow} \nu_{\underset{\sim}{X}}$$

on $\mathbb{R}\backslash\{\underset{\sim}{0}\}$. The definitions of a $\mathcal{B}_0$-boundedly finite measure and vague convergence are given in **Appendix A**.

Now suppose that $\underset{\sim}{X}_{d\times 1}$ is a regularly varying random vector. Then there exists an $\alpha > 0$, a function $g : (0, \infty) \to (0, \infty)$ which is regularly varying at infinity with index α and a $\mathcal{B}_0$-boundedly finite non-zero measure $\nu_{\underset{\sim}{X}}$ such that $g(t)\mathbb{P}(t^{-1}\underset{\sim}{X}\in\cdot) \overset{\mathcal{B}-vaguely}{\longrightarrow} \nu_{\underset{\sim}{X}}$ as $t \to \infty$. The positive quantity α is known as the tail index of the random vector $\underset{\sim}{X}$. So, the characterization of a random vector in terms of regular variation is done by a tail index as well as an exponent measure.

Finally we are ready to define what is regular variation for a Time Series process which is heavy-tailed. We will be considering regularly varying $d - dimensional$ time series process that takes values in the space $\left(\mathbb{R}^d\right)^{\mathbb{Z}}$. The regular variation for this time series process is determined by the regular variation of it's distribution in the sequence space. If we can endow (see **Appendix A** for definition) the space $\left(\mathbb{R}^d\right)^{\mathbb{Z}}$ with some suitable topology, then these random vector which is regularly varying will have an exponent measure. This situation is equivalent to the exponent measures of the finite dimensional distributions of the time series. It turns out that A Time series $\{\underset{\sim}{X_t} : t \in \mathbb{Z}\}$ which takes values in $\mathbb{R}$ is said to be regularly varying with tail index α if the finite dimensional distributions are in a domain of attraction of a multivariate Frechet distribution with the same tail index and under the same scaling (**Soulier(2021)**). The extremal behavior of this regularly varying time series is determined by the two entities namely, **tail process** and **tail measure**. Before we move on to understand the concepts of tail measure and tail process there are a couple of important notes that we should keep in mind. Firstly we note that there is a fundamental difference between a Gaussian process and a regularly varying process. The first one is extremally independent as the extremal behavior of the finite dimensional distributions is similar to that of the one with independent components. In fact they are in $o(1)$ approximations of each other. But for a regularly varying process this is not true since its extremal behavior inherits serial dependence. Again, we should note that, in the context of stationarity the tail process comes into play. We can say that the tail process is like a weak limit of the finite dimensional distributions of $\underset{\sim}{X}$ given that $|\underset{\sim}{X_0}| > x$ as $x \to \infty$. Hence it is obvious that the usage of tail process is useless for a Gaussian process.

2.6.2 Tail measure and Tail process

Suppose that $\{\underset{\sim}{X_t} : t \in \mathbb{Z}\}$ is a time series which takes values in $\mathbb{R}$. If the process is regularly varying, then $\exists$ a non-decreasing sequence $\{a_n\}$ and a non-zero $\mathcal{B}^{(\infty)}$ boundedly finite measure $\underset{\sim}{\nu}$ on $\left(\mathbb{R}^d\right)^{\mathbb{Z}} \backslash \{0\}$ such that $n\mathbb{P}(a_n^{-1}\underset{\sim}{X} \in .) \overset{vaguely}{\longrightarrow} \underset{\sim}{\nu}$ in $\left(\mathbb{R}^d\right)^{\mathbb{Z}} \backslash \{0\}$ endowed with $\mathcal{B}^{(\infty)}$. For this scenario,

$\underset{\sim}{\nu}$ is known as the **tail measure (Kulik(2020))** of the time series X. For a tail measure $\underset{\sim}{\nu}$ the following hold:

- $\underset{\sim}{\nu}(\{\underset{\sim}{0}\}) = 0$
- $\underset{\sim}{\nu}(\{\underset{\sim}{y} \in \mathcal{D} = \left(\mathbb{R}^d\right)^{\mathbb{Z}} : |\underset{\sim}{y_0}| > 1\}) = 1$
- $\exists \alpha > 0 \ni \underset{\sim}{\nu}(tA) = t^{-\alpha}\underset{\sim}{\nu}(A)$ for every Borel set $A \in \mathcal{D}$

Again, here α is the tail index of the tail measure $\underset{\sim}{\nu}$. It follows from above that, $\underset{\sim}{\nu}(\{\underset{\sim}{y} \in \mathcal{D} = \left(\mathbb{R}^d\right)^{\mathbb{Z}} : |\underset{\sim}{y_i}| > 1\}) < \infty\ \forall i$ and that the tail measure is $\sigma - finite$. Hence, the tail measure $\underset{\sim}{\nu}$ restricted to the set $\{\underset{\sim}{y} \in \mathcal{D} = \left(\mathbb{R}^d\right)^{\mathbb{Z}} : |\underset{\sim}{y_0}| > 1\}$ can be thought of as a probability measure as it follows Kolmogorov's three axioms. This actually means that we can consider a D-valued random vector $\underset{\sim}{Y}$ on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with distribution $\underset{\sim}{\nu}(.\bigcap\{\underset{\sim}{y} \in \mathcal{D} : |\underset{\sim}{y_0}| > 1\})$. Then this random vector $\underset{\sim}{Y}$ is known as the **tail process** associated to the tail measure $\underset{\sim}{\nu}$. So, the distribution of a tail process is the tail measure restricted to the set described above. We should also note that the tail process actually describes the asymptotic behavior of a stationary time series given an extreme event at time point 0 that is, it is a weak limit of the finite dimensional distributions of X given that $|X| > x$ as $x \to \infty$. Thus, we can study the tail measure and the tail process without the reference to an underlying stochastic process. This is what makes this approach unique as no data-driven modeling can be of any help here, and only the distribution theory comes handy.

Now, if the tail measure is homogeneous, then $|\underset{\sim}{Y_0}|$ follows a Pareto distribution with tail index α and it is independent of the process known as the **spectral tail process** associated to the given tail measure defined as $= |\underset{\sim}{Y_0}|^{-1}\underset{\sim}{Y}$. The spectral tail process can be viewed as a spectral decomposition of the tail measure with respect to a pseudo-norm on $\mathcal{D}$. There exists a close relation between a shift-invariant tail measure and and tail process, since we will be able to recover the tail measure if it is shift-invariant from a tail process. In this context, there is another important entity known as the spectral process. A Borel measure ν on $\mathcal{D}$with tail index α is **shift-invariant** iff there exists a $\mathcal{D}$valued process $\underset{\sim}{Z}$ known as the **spectral process** such that,

1. $P(\underset{\sim}{Z} = \underset{\sim}{0}) = 0$
2. $E(|\underset{\sim}{Z_0}|^{\alpha}) = 1$
3. $\underset{\sim}{\nu} = \int\limits_0^{\infty} E[\delta_u \underset{\sim}{z}]\alpha u^{-\alpha-1} du$ [here $E[\delta_u](f) = E[f(U)]$ for a measurable map f]
4. $\forall a < b$, $t \in \mathbb{R}$, and bounded 0-homogeneous maps [see **Appendix A**] H on $\mathcal{D}$, $0 < E[\sup\limits_{a \leq t \leq b} |\underset{\sim}{Z_s}|^{\alpha}] < \infty$ and $E[|\underset{\sim}{Z_t}|^{\alpha} H(\underset{\sim}{Z})] = E[|\underset{\sim}{Z_0}|^{\alpha} H(B^T \underset{\sim}{Z})]$

As shown in **Soulier(2021)** a shift-invariant tail measure is entirely determined by its tail process $\underset{\sim}{Y}$ whose probability distribution $\mathbb{P}_Y$ can be written as $\mathbb{P}_Y = E[|\underset{\sim}{Z_0}|^{\alpha} \delta_{\frac{YZ}{|\underset{\sim}{Z_0}|^{\alpha}}}]$ where $\underset{\sim}{Z}$ is the spectral process, and Y follows a Pareto distribution with tail index α. An important note that should be made here is that although the tail process is unique in its distribution, a spectral tail process is not.

One necessary condition for the existence of such a unique shift-invariant tail measure is given below. Suppose $\underset{\sim}{Y}$ is a random element in $\mathcal{D}$ such that,

1. $\mathbb{P}(|\underset{\sim}{Y_0}| > 1) = 1$
2. the time change formula (given in **Appendix A**) holds
3. $\forall a < b, \int\limits_a^b E[\frac{1}{\int\limits_a^b 1\{|\underset{\sim}{Y_{t-s}}|>1\}dt}]ds < \infty$

then $\exists$ a unique shift-invariant tail measure $\underset{\sim}{\nu}$ such that the distribution of the tail process $\underset{\sim}{Y}$ is $\underset{\sim}{\nu}$ restricted to the set $\{\underset{\sim}{y} \in \mathcal{D} : |\underset{\sim}{y_0}| > 1\}$.

Keeping the above definitions and conditions in mind, we can summarize as proved in **Soulier(2021)** that when the condition C1 (see **Appendix B**) holds, the tail measure of the regularly varying process $\underset{\sim}{X}$ is supported on $\mathcal{D}$ and the associated tail process has almost sure cadlag paths (**Appendix A** for definition) and $\{\frac{\underset{\sim}{X}}{x} \mid |\underset{\sim}{X_0}| > x\}$ converges weakly to the tail process $\underset{\sim}{Y}$ as $x \longrightarrow \infty$ on $\mathcal{D}$ endowed with J_1-topology. [**Appendix A** for definition] We will be using this result in our statistical analysis using the tail process approach to the heavy-tailed AQI data.

Recall that in section 2.3 while describing the block maxima method, we have defined a term known as the extremal index, which plays a vital role when we have a dependent sequence of random variables such as in time series data. There are many ways to estimate this extremal index; but none of them are straightforward and they come with their own complications. But in the tail process approach to the heavy-tailed regularly varying time series data, this estimation can be done using the **exceedance functional**, a quantity which basically computes the frequency of the process being over threshold defined as,

$$\varepsilon(\underset{\sim}{Y}) = \int_{-\infty}^{\infty} 1\{|\underset{\sim}{Y_t}| > 1\}dt = \sum_{t=0}^{\infty} 1\{|\underset{\sim}{Y_t}| > 1\}$$

where $\underset{\sim}{Y}$ is the tail process associated with the regularly varying time series process. Note that, for a tail process $\underset{\sim}{Y}$, $\mathbb{P}(|\underset{\sim}{Y_0}| > 1) = 1$. Thus, $\varepsilon(\underset{\sim}{Y}) \geq 1$ almost surely. This fact ensures that $E[\frac{1}{\varepsilon(\underset{\sim}{Y})}] \leq 1$. The LHS of the last inequality is denoted by ϑ and is known as the **candidate extremal index**. It turns out that for max-stable processes and for the block maxima's extremal index, ϑ serves the purpose of being a suitable estimate. This comes in as very handy since this quantity is straightforward for computation purposes. In fact, we will show another even simpler way to compute the candidate extremal index as suggested by **(Soulier(2021)).**

A measurable map $\mathcal{I} : \mathcal{D} \longrightarrow \mathbb{R}$ is known as an **anchoring map** if,

1. $\mathcal{I}(\underset{\sim}{B}^t\underset{\sim}{Y}) = \mathcal{I}(\underset{\sim}{Y}) + t$ for every $t \in \mathbb{R}$
2. $|\underset{\sim}{Y_{\mathcal{I}(\underset{\sim}{Y})}}| > 1$ if $\mathcal{I}(\underset{\sim}{Y}) \in \mathbb{R}$

One important example of an anchoring map for a process indexed by $\mathbb{Z}$ is the first exceedance over 1 denoted as $\mathcal{I}_1$ given by,

$$\mathcal{I}_1(\underset{\sim}{Y}) = inf\{t : |\underset{\sim}{Y_t}| > 1\}$$

The candidate extremal index and any anchoring map is closely related as shown in **Soulier(2021)**. For a tail process $\underset{\sim}{Y}$such that, $\mathbb{P}(|\underset{\sim}{Y_0}| > 1) = 1$ and that satisfies the time change formula, if $P[\underset{|t|\longrightarrow\infty}{lim}|\underset{\sim}{Y_t}| = 0] = 1$, then for any anchoring map $\mathcal{I}$,

$$\vartheta = E[\frac{1}{\varepsilon(\underset{\sim}{Y})}] = P(\mathcal{I}(\underset{\sim}{Y}) = 0)$$

Thus, once we have the tail process, we just have to compute the probability of an anchoring map to be equal to zero. That in turn will serve the purpose of being the candidate extremal index and hence the estimate of the original extremal index.

2.6.3 Point processes

An important aspect of the extremal index is that, it is not only essential to understand the max-stable process and the block maxima GEV distribution for dependent sequences; but it also has a close relation with the clustering effect of the exceedances over high thresholds. This relation is given in terms of point process of exceedances and the asymptotic clustering over a high threshold can be tackled using the point process approach as well. Let us consider $\{X_i : i \in \mathcal{I}\}$ to be a representation of a location of points indexed by $\mathcal{I}$ occurring randomly in a state space $\mathcal{S}$.(**Beirlant(2004)**) A **point process** N counts the number of points in the region $\mathcal{S}$ and is defined as

$$N(A) = \sum_{i\in\mathcal{I}} 1(X_i \in A) \ \ , A \subseteq \mathcal{S}$$

The expected value of the point process is given by $E(N(A)) = \Lambda(A)$ and is known as the **intensity measure** associated to the point process. If $\mathcal{S}$ is

Euclidean or a subset of it, and if Λ has density function $\lambda : \mathcal{S} \longrightarrow [0, \infty)$ i.e., $\Lambda(A) = \int\limits_{A} \lambda(x)dx$, then λ is known as the intensity function of the process.

Now a point process with intensity measure λ is said to be a Poisson point process (PPP) if,

1. $\forall A \ni \Lambda(A) < \infty$, $N(A) \sim Poisson(\Lambda(A))$
2. $\forall k \in \mathbb{N}$, and $\forall A_1, ..., A_k \ni A_i \bigcap A_j = \phi$, $N(A_1), ..., N(A_k)$ are independent.

Moreover the process is homogeneous if $\lambda(x) = \lambda \, \forall x$. A marked point process counts for each point X_i a quantity Y_i i.e., $N(A) = \sum\limits_{i \in \mathcal{I}} Y_i 1(X_i \in A)$, where the marks $\{Y_i\}$ are identically distributed.

We now define the compound Poisson process which will be the main focus of discussion for the exceedance clusters. A compound Poisson process is a marked process such that X_i's occur according to a Poisson process independent of Y_i which are themselves iid. We denote this compound Poisson process as $CP(\lambda, \pi)$ where λ is the intensity function and π is the mark distribution. Now, a sequence of marked processes N_n on $\mathcal{S}$ converges in distribution to a marked point process N if $\forall k \in \mathbb{N}$, and $A_1, ..., A_k$,

$$\{N_n(A_i)\}_{i=1}^{k} \xrightarrow{\mathcal{D}} \{N(A_i)\}_{i=1}^{k}$$

As shown in **Hsing(1988)** suppose $\{X_t\}$ is a stationary time series process with extremal index $\theta > 0$, and $\exists$ a sequence of thresholds u_n such that $\Delta(u_n)$

holds [see **Appendix B**] and $n\bar{F}(u_n) \longrightarrow \tau \in (0, \infty)$. Let there exist s_n and r_n and a distribution π such that $s_n = o(r_n)$, $r_n = o(n)$, $n\alpha(n, s_n) = o(r_n)$ and $\pi_n(j) \longrightarrow \pi(j) \, \forall j \geq 1$ as $n \longrightarrow \infty$. Then $N_n \longrightarrow N \sim CP(\theta\tau, \pi)$. Here $\alpha(.,.)$ acts in a similar to the one we saw for $D(u_n)$ condition in section 2.3.

The point process we are interested in is very similar to the exceedance functional we defined in section 2.6.3. The point process is given by,

$$N_n(.) = \sum_{i \in \mathcal{I}} 1(\frac{i}{n} \in .)$$

where $\mathcal{I} := \{i : X_i > u_n, 1 \leq i \leq n\}$. This actually counts the times at which the sample $\{X_i\}_{i=1}^n$ exceeds the threshold u_n. Thus, with the help of the estimated extremal index, we get a good limiting process for the clusters of sample exceedances.

2.6.4 Application for the dataset

We have our time series dataset $\{X_t\}$ on the AQI of New Delhi at hour t where $t = 0$ corresponds to midnight, 1st January 2019. We will be applying the tail measure and tail process approach now. Here we have univariate data and hence we will confine ourselves to the case where $d = 1$. It should be noted that we have already assumed this data to follow a stationary ARMA-GARCH process. Hence, the regular variation is implied and the tail measure approach is reasonable in this case.

To get the estimate of the tail index, we will be using the Hill's estimator of the tail index for the parametric Generalized Pareto distribution which in turn will act as the candidate tail measure. We are using the GPD since it is the best option to estimate the tail and it is valid even for a non-iid scenario, something that we should take into account the data being a time series one. Now, by the properties of the tail measure, $\frac{X}{x} \mid |X_0| > x \overset{weakly}{\longrightarrow} Y$ in distribution, where the distribution of $Y$is given by our candidate tail measure the GPD. In our case, the threshold going to infinity is similar to the threshold being 400 (severe condition). Hence, at first, we take the transformation of the data as $\frac{X}{400}$ conditioned on the AQI at the first time point being more than 400. Next we fit the GPD given by

$$f(x|\mu,\sigma,\xi) = \frac{1}{\sigma}(1+\xi(\frac{x-\mu}{\sigma}))^{(-\frac{1}{\xi}-1)}$$

with $\mu \in \mathbb{R}, \sigma > 0, \xi \in \mathbb{R}$. The support of the distribution is

$$S = \begin{cases} \{x : x \geq \mu\}, & if\, \xi \geq 0 \\ \{x : \mu \leq x \leq \mu - \frac{\sigma}{\xi}\}, & if\, \xi < 0 \end{cases}$$

The shape parameter is obtained using the Hill's estimate in the following way:

The Hill's estimator is defined as,

$$H_{k,n} = \frac{1}{k}\sum_{j=1}^{k} log(X_{(n-j+1)}) - log(X_{(n-k)})$$

We will choose that k such that the plot of $(k, H_{k,n})$ is constant for some portion preferably in the tail region. Actually, it can be seen as the estimator of slope in the k last points of a Pareto QQ-plot while using constrained least squares where the regression line has to pass through $P = (-log(k+1), log(X_{n-k}))$.

Now once we have obtained the value of k from the graph, we have our estimated shape parameter as $H_{k,n}$ and the tail index $\widehat{\alpha} = 1/H_{k,n}$. If the estimate of the shape parameter is more than 0, then the other two parameters can be estimated in a straightforward way as $\widehat{\mu} = X_{(1)}$ and $\widehat{\sigma} = (\bar{X} - \widehat{\mu}) \times (1 - \widehat{\xi})$

So, we have our tail measure approximation to the stationary process as a GP distribution with tail index $\widehat{\alpha}$. Hence by the limiting properties, we can use this limiting process to obtain the entities we have obtained in our empirical bootstrap approach such as the amount of time the process stays over the threshold, the frequency of crossing the threshold etc. But the interesting point here is, we can even comment on the occurrence of a worst of the worst event and when is it likely to occur by taking the return period definition of the extreme value theory. This is not possible using the empirical approach.

The frequency of crossing increasing sequence of thresholds upto the extreme case (400/400=1) by the tail process is computed by calculating the exceedance probabilities for each threshold using the fitted GPD and plotting a curve for the same. We will also be looking into the return period and the return level. The **return period** is the average time between two extreme events, which in our case is crossing a threshold of 1 for the tail process. The **return level** is the level x such that it is expected to be exceeded once in r years. Hence it is actually $F^{-1}(1-\frac{1}{r})$.

Another important aspect of the tail measure approach is that, once we have the tail process, we can comment about the extremal index and its estimation can be done using the anchoring maps and the candidate extremal index as described in the section 2.6.2. We make use of the anchoring map $\mathcal{I}_1$ i.e., the first exceedance over 1 and obtain the estimate as:

$$\vartheta = P(\mathcal{I}_1(\underset{\sim}{Y}) = 0)$$

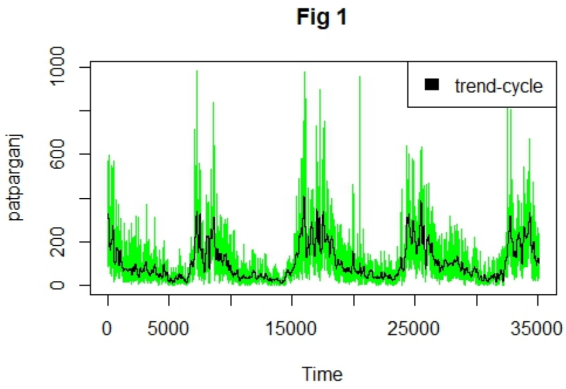
An interesting observation one can make here is that $P(\mathcal{I}_1(\underset{\sim}{Y}) = 0)$ is actually the same as $P(Y_1 > 1)$ and this can be obtained using the fitted GPD. Since this is the probability of the process in terms of the time-point the dependence structure of the process has no role here and it can solely be obtained by computing the probability for the fitted GPD.

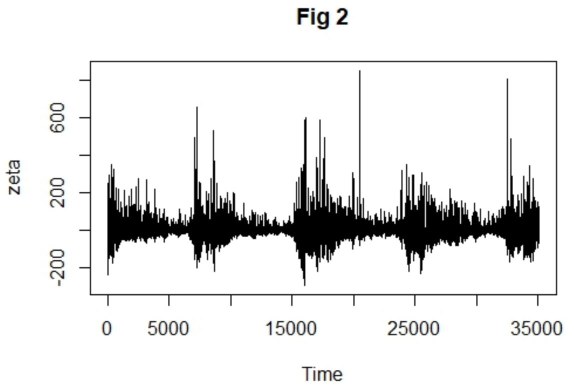
3 Statistical Analysis

3.1 Empirical bootstrap approach

We have 39 stations that measures the AQI (PM2.5) for Delhi. Our first motive is to conduct a time-series analysis for the stations by using an appropriate ARMA-GARCH model. Once we are able to fit the model, we will do the residual bootstrapping as explained in the section 2.5, from the empirical distribution of the residuals, to get our bootstrap estimate for the exceedance probability at different time points.

In order to get the stochastic component of the PM2.5 series, we need to get rid of the trend-cycle component at first. Here we have taken the Kernel to be Gaussian and the bandwidth as $h = 100$ as a compromise between underfitting and overfitting. The estimated trend-cycle component has been shown in Fig 1. Fig 2 shows the estimated stochastic component of the PM2.5 series after removing the trend-cycle.





From Fig. 2 it is evident that there is heterodcedasticity present in the stochastic component that we have got after de-trending. So, as explained in the "methologies" section, we should consider the ARMA-GARCH model for our modelling purpose. We will be selecting the ARMA-GARCH which gives the least RMSE in terms of forecasting.

For our data the result turns out to be an ARMA(1,0)-GARCH(1,1) model and hence the model that we are going to fit is:

$$\zeta_t = \mu + \phi_1 \zeta_{t-1} + \epsilon_t$$

$$\epsilon_t = \sigma_t z_t$$

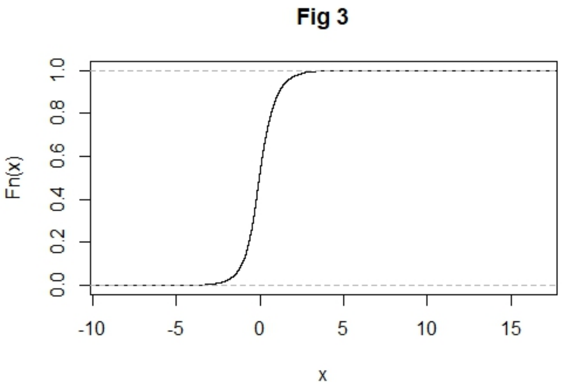
$$\sigma_t^2 = \omega + \alpha_1 \epsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

where the assumptions of Section "methodologies" hold.

From the QML estimation, we get the estimates given in Table 1.

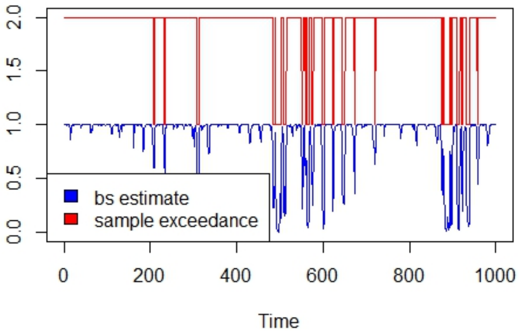
Parameter	Estimate	Std. Error
μ	-0.98	0.844
ϕ_1	0.93	0.002
ω	19.05	1.352
α_1	0.449	0.014
β_1	0.549	0.014

From the above estimates, we will get our residual estimates as, $\widehat{\epsilon_t} = \widehat{\zeta_t} - \widehat{\mu} + \widehat{\phi_1}\widehat{\zeta_{t-1}}\ for\,t > 1$ and their unconditional variances $\widehat{\sigma_t^2} = \widehat{\omega} + \widehat{\alpha_1}\widehat{\epsilon_{t-1}^2} + \widehat{\beta_1}\widehat{\sigma_{t-1}^2}\ for\,t > 1$. We are going to take $\widehat{\epsilon_1} = 0$ since it does not affect the values for the total process, as mentioned in "methodologies". After getting the estimates we standardize the residuals and obtain its empirical distribution function as shown in Fig 3.



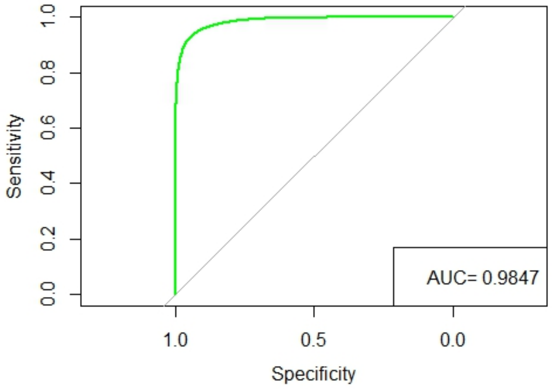
Our next job is to do residual bootstrapping to obtain the bootstrap estimate of the exceedance probability using the algorithm mentioned in Section 2.5. The results are shown in Fig 4 along with the original sample exceedances coded as indicator $I_t = \begin{cases} 1 & if\, X_t > C \\ 0 & otherwise \end{cases}$. It should be noted that here we choose $C = 60$, the cutoff set by the government for PM2.5, and compute the exceedances. Note that we have shifted the plot for sample exceedance in Fig 4 by one unit in the Y-axis for a better visualization and comparison with the bootstrap estimate. Again for the same purpose, we have taken the first 1000 observations as plotting the full curve will be cumbersome to visualize.

Fig 4



We will also be checking how does our bootstrap estimate of exceedance probability acts as a classifier for the exceedances. The ROC curve and AUC has been obtained in Fig 5. This shows that the AUC is 0.9847 with the ROC having an almost perfect shape. Hence our bootstrap estimate of exceedance probability can be considered good as a classifier for the exceedances.

Fig 5



The next question that we will be dealing with is what is the average behavior of the estimated process above the threshold and how does it change when we keep on increasing the threshold upto the worst case scenario i.e., above 400 (Severe) in terms of the average time spent over the threshold once it crosses it, and also how frequently it crosses the given threshold. We have exceeded the threshold from the cut-off set by the air pollution control board (60) to severe (400) for the above purpose. From Fig. 6,7 and 8 we can see that for the quantities of interest, viz., average time over threshold after crossing it, frequency of crossing the threshold, and the total time spent over threshold - have a quadratic behavior with a negative slope.

Fig 6

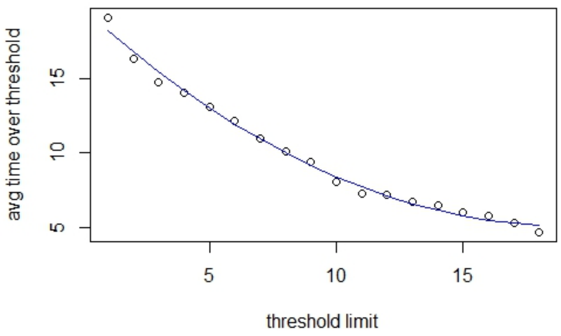


Fig 7

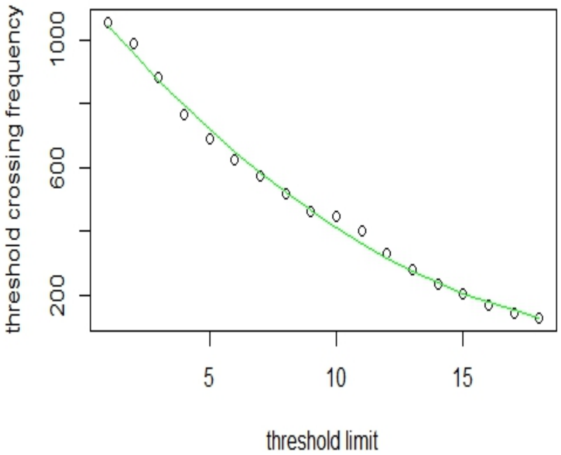
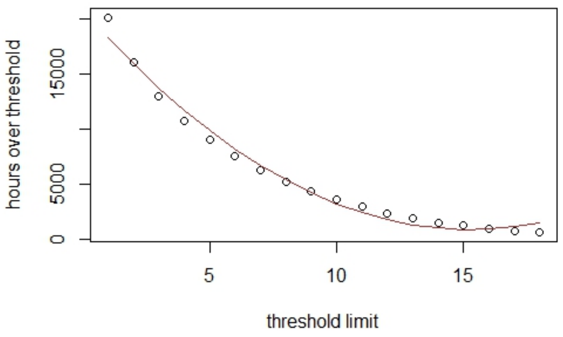


Fig 8



We have repeated the same procedure for another station Alipur, and achieved similar results using an ARMA(2,1)-GARCH(1,1) model. The parameter estimates are given in the following Table 2.

Parameter	Estimate	Std. Error
μ	-0.872	0.6201
ϕ_1	1.289	0.0137
ϕ_2	-0.384	0.013
θ_1	0.097	0.0154
ω	12.803	1.0625
α_1	0.35	0.0125
β_1	0.64	0.0135

The behavior of the bootstrap estimates of exceedances obtained from the same bootstrapping scheme applied for Patparganj station is shown in Fig 9. Fig 10 shows us that as a classifier for exceedances, it works very well with AUC 0.9867.

Fig 9

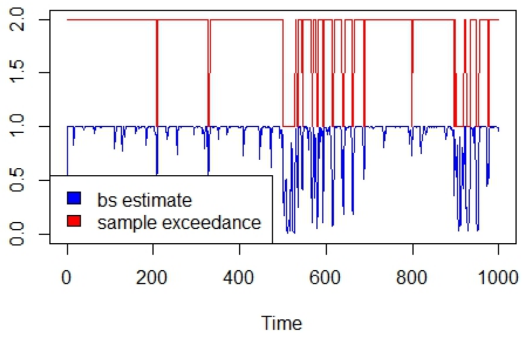
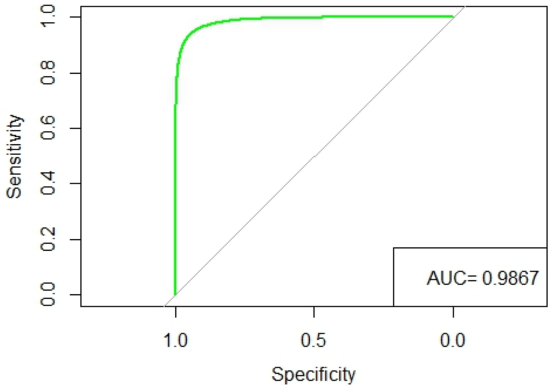


Fig 10



The average time spent over threshold, the frequency of crossing the threshold, and the total time spent over threshold when we increase the threshold from 60 to 400, show again a similar quadratic behavior with negative slope as illustrated in Fig 11, 12 and 13.

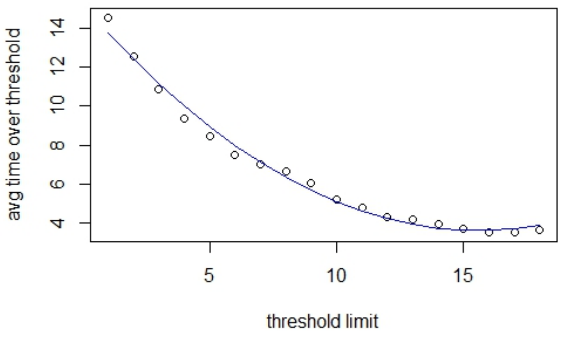


Fig 11

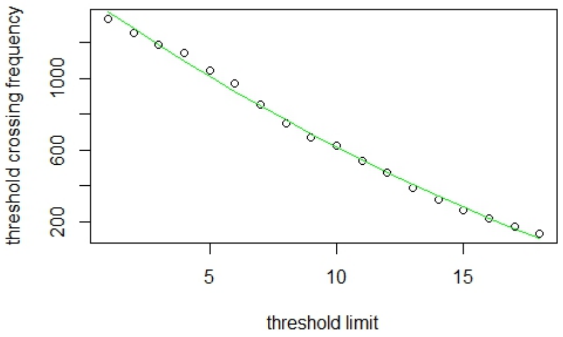


Fig 12

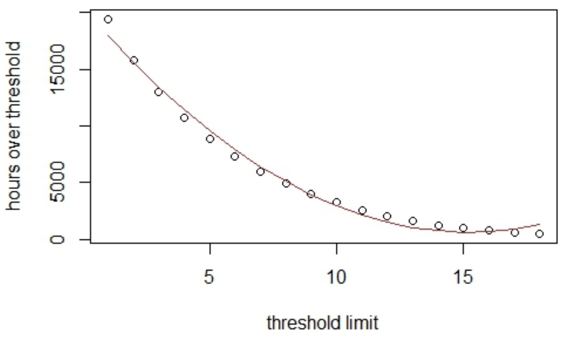


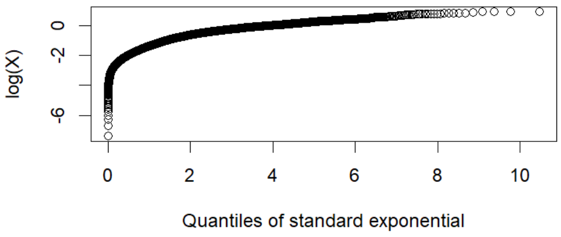
Fig 13

Now that we have modeled the behavior of the time series process over the threshold, we will move on to the greater question, i.e., what is the limiting behavior of the process and how to understand when will the worst of the worst, i.e., the process staying continuously over the threshold and creating extreme health hazards going to happen. For this part of the project we will be using the tail process approach by using tail measure.

3.2 Tail process approach

Again consider the hourly AQI data from the station Patparganj in New Delhi. Let $\{X_t\}$ denote the AQI (level of PM2.5) of Patparganj at hour t where $t = 0$ corresponds to midnight, 1st January 2019. We have already seen that it is a ARMA-GARCH process. Hence we can consider it to be regularly varying with tail index α. Let us try to see how the Pareto QQplot looks like for our data and we will be able to understand if we are going in the right path or not. The QQplot is given in Fig 14.

FIG 14



So the Pareto QQplot given in Fig 14 looks quite satisfactory as the tail of the log-transformed data is supposed to look like a straight line when plotted against quantiles of a standard exponential distribution. So we continue with the GPD as our tail measure. We move on to estimate the parameters of the GPD to get our fitted model which we will be using as tail measure. Note that we have standardized our process as $\{\frac{\underset{\sim}{X}}{400}\}$ since 400 acts as the extreme threshold for our data.

From the Pareto QQplot, we can see that the that the slope is almost constant in the region 4 to 6 for quantiles of std. exponential. Hence, we will try to find the best possible estimate of the tail index from that region. The Hill's estimator for the shape parameter is thus obtained as, $\widehat{\xi} = 0.32$ and hence the tail index turns out to be, $\widehat{\alpha} = 1/0.32 = 3.125$. As mentioned earlier, once we catch hold of the tail process, the underlying stochastic process is of no use. Hence, we will obtain the rest of the entities of interest with the help of the tail process only. The estimates of other parameters are obtained as, $\widehat{\mu} = 0.0006 \approx 0$ and $\widehat{\sigma} = 0.18$. The exceedance probability (unconditional) at any time point t is $P(X > 60) = P(\frac{X}{400} > \frac{60}{400}) = 0.665$ for the Govt. set threshold obtained from the fitted GP distribution. The candidate extremal index obtained from the fitted GPD is:

$$\vartheta = P(\mathcal{I}_1(\underset{\sim}{Y}) = 0) = P(Y_1 > 1) = 0.395$$

Hence, the average cluster time is obtained as, $1/\vartheta = 2.53$. So, the process goes above the extreme threshold (severe) once in every 2.53 hours on an average. This is not at all a good sign given that serious health hazards will occur in that condition. With the data in hand, for the given station, the 1 year return level is 8.14 for the normalized process. Hence the 1 year return level for the original process is $8.14 \times 400 = 3256.37$. So, if we want an extreme event (the process crossing a certain threshold) to happen once in a year, then we have to set that threshold for PM2.5 to be 3256.37! The plots for frequency of crossing the threshold, and the average time spent over threshold are given in FIG 15 and 16 respectively.

FIG 15

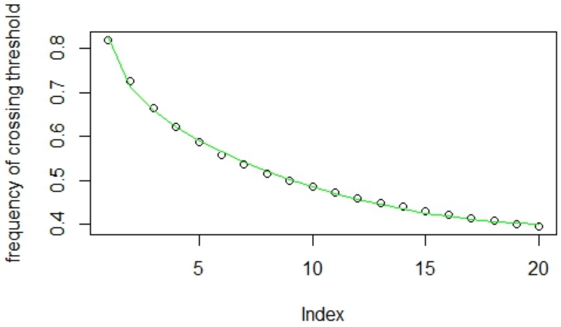
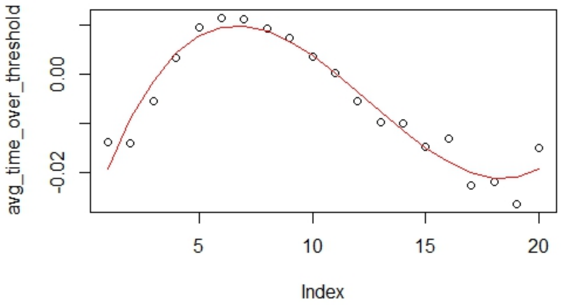


FIG 16



4 Conclusion

We have seen in both the methods that the condition of air quality is not very well in New Delhi as observed from the Patparganj station. We have used the bootstrap distribution of the residuals while fitting the empirical ARMA-GARCH model to obtain the exceedance probabilities. The bootstrap estimate of exceedance probabilities have done a good job as a binary classifier having an AUC of 0.9847. We have also seen the average behavior of the empirical process above the threshold, and they have generally showed a quadratic behavior with a negative slope. But the problem with this approach is that the time series model of ARMA-GARCH might be a very good model, but it cannot be used for a long-time forecast. Hence we cannot comment about return levels, or the limiting behavior of the process.

In the tail process approach, we have the advantage of using the distributional properties and hence can obtain the results that were not achieved while using the empirical method. Here we have fitted a generalized Pareto distribution as the tail measure and using its properties have obtained the candidate extremal index which can be used to comment about the limiting distribution of the max-stable process and to comment about the domain of attraction the extremes belong to. As discussed in section 2.3, the distribution of maxima of a dependent sequence like this one is heavily influenced by the extremal index θ, and that has been estimated as $\vartheta = \widehat{\theta} = 0.395$. We have also obtained the one-year return level to be an astonishing 3256.37 which is the level the process is expected to cross once in a year given the data we have. So, there are two possible solutions that can be made using this information - either the govt has to set a higher cutoff for the PM2.5 in New Delhi to get less extreme events, or we have to identify how to reduce the AQI by taking necessary precautions. This will be discussed in short in section 5. But one thing is obvious from the results, that the situation of air quality is not at all satisfactory since the extreme health conditions are expected to happen once in every 2.53 hours!

5 Extensions and possibilities

The area of research on the tail measure and tail process is still vast. We have just used some part of the whole literature on this topic that is available for our computations and worked on the application side of the problem. More interpretations on extremal index and on the tail processes can be done and that remains an open field of research. Even for the empirical modeling instead of using the ARMA-GARCH one can use the SETAR or the E-GARCH model and compare the results. The way to deal with clusters of extremes are not touched much in this project since we have used the tail process approach which does not demand much of de-clustering and the generalized Pareto distribution works well under moderate dependent data as well. But more work can be done using the de-clustering principle which can be compared with the results achieved here. We have also not taken into account of the possible covariates which can heavily influence the AQI. We have seen in this project that in both the approaches, the AQI situation of New Delhi is extremely poor the severe health hazard condition happening on an average once in 2.53 hours. So, if we can do a covariate study and identify which physical entities viz., temperature, wind speed, transport movement, industrial pollution are responsible for such a high AQI, then we might be able to provide a solution to reduce the AQI and achieve desirable results. We have also not touched the spatial analysis which would be a good extension to this project by taking data from all the stations of New Delhi and obtaining the results using spatial analysis. Thus, the scope of extending this project as well as the research topic is a vast region which can lead to better and more accurate results.

6 References

References

[1] Giuseppina Albano, Michele Rocca, and Cira Perna. *Estimating Exceedance Probability in Air Pollution Time Series*, pages 28–38. 10 2020.

[2] Giuseppina Albano, Michele Rocca, and Cira Perna. *Volatility Modelling for Air Pollution Time Series*, pages 176–184. 04 2020.

[3] Bojan Basrak, Richard A. Davis, and Thomas Mikosch. Regular variation of garch processes. *Stochastic Processes and their Applications*, 99(1):95–115, 2002.

[4] Bojan Basrak and Johan Segers. Regularly varying multivariate time series. *Stochastic Processes and their Applications*, 119(4):1055–1080, 2009.

[5] J. Beirlant, Y. Goegebeur, J. Segers, J.L. Teugels, D. De Waal, and C. Ferro. *Statistics of Extremes: Theory and Applications*. Wiley Series in Probability and Statistics. Wiley, 2004.

[6] Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986.

[7] Dana Draghicescu and Rosaria Ignaccolo. Modeling threshold exceedance probabilities of spatially correlated time series. *Electronic Journal of Statistics*, 3(none):149 – 164, 2009.

[8] Robert F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007, 1982.

[9] Husler J. Leadbetter M. R. Hsing, T. On the exceedance point process for a stationary sequence. *Probability Theory and Related Fields*, 78:97–112, 1988.

[10] J, Jacob and Mohammadipour Gishani, Azadeh. Threshold Detection in Autoregressive Non-linear Models, 2010. Student Paper.

[11] Anja Jansen. Spectral tail processes and max-stable approximations of multivariate regularly varying time series. *Stochastic Processes and their Applications*, 129(6):1993–2009, 2019.

[12] R. Kulik and P. Soulier. *Heavy-Tailed Time Series*. Springer Series in Operations Research and Financial Engineering. Springer New York, 2020.

[13] M. R. Leadbetter. On extreme values in stationary sequences. *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 28(4):289–303, 1974.

[14] M. R. Leadbetter. Extremes and local dependence in stationary sequences. *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 65(2):291–306, 1983.

[15] Daniel B. Nelson. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59(2):347–370, 1991.

[16] James Pickands. Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1):119–131, 1975.

[17] R.H. Shumway and D.S. Stoffer. *Time Series Analysis and Its Applications*. Springer texts in statistics. Springer, 2000.

[18] Philippe Soulier. The tail process and tail measure of continuous time regularly varying stochastic processes, 2021.

[19] H. Tong. *Threshold Models in Non-linear Time Series Analysis*. Lecture notes in statistics. Springer-Verlag, 1983.

[20] Seokhoon Yun. The distributions of cluster functionals of extreme events in a dth-order markov chain. *Journal of Applied Probability*, 37(1):29–44, 2000.

[21] Zhengjun Zhang. On studying extreme values and systematic risks with nonlinear time series models and tail dependence measures. *Statistical Theory and Related Fields*, 5(1):1–25, 2021.

7 Appendix

A. Definitions

1. $\mathcal{B}_0$-boundedly finite A measure ν is said to be $\mathcal{B}_0$-boundedly finite if $\nu(A) < \infty$ for all Borel-measurable sets A separated from $\underset{\sim}{0}$.

2. Vague convergence In order to define Vague convergence we need to define boundedness.

Let E be a set. A **boundedness** $\mathcal{B}$ on E is a collection of subsets of E, called bounded sets, with the following properties:

- (i) a finite union of bounded sets is bounded;
- (ii) a subset of a bounded set is a bounded set.

Let $(E,\mathcal{B})$ be a localized Polish space. A measure ν on E is said to be $\mathcal{B}$-boundedly finite if $\nu(A) < \infty$ for all $\mathcal{B}$-bounded Borel sets A. The set of $\mathcal{B}$-bounded finite measures on E is denoted $\mathcal{M}_{\mathcal{B}}$. A sequence $\nu_n, n \in \mathbb{N}$ of $\mathcal{B}$-boundedly finite measures **converges $\mathcal{B}$-vaguely** to a measure ν if $\underset{n\to\infty}{lim}\,\nu_n(f) = \nu(f)$ for all bounded continuous functions f with $\mathcal{B}$-bounded support.

3. Endowed topology A topological space is a set endowed with a structure called topology which allows defining continuous deformation of subspaces.

4. Homogeneous maps For a real number r, a map H on $\mathcal{D}$ is known as r-homogeneous if $H(t\underset{\sim}{y}) = t^r H(\underset{\sim}{y})$ for every positive number t and $\underset{\sim}{y} \in \mathcal{D}$.

5. Cadlag paths Let (M, d) be a metric space. Then a function f from $E \subset \mathbb{R}$ to M is known as a Cadlag path if $\forall t \in E$,

- (i) the left hand limit exists, $f(t-) = \underset{s\uparrow t}{lim} f(s)$
- (ii) it is right continuous, $f(t+) = \underset{s\downarrow t}{lim} f(s) = f(t)$

It is interesting to note that all cdfs irrespective of being continuous or not, are Cadlag paths.

6. Time change formula Let $\underset{\sim}{\nu}$ be a shift-invariant tail measure on $\mathcal{D}$ with tail index α and associated tail and spectral tail process $\underset{\sim}{Y}$ and $\mathbf{\Theta}$ respectively. For every non-negative measurable map $\mathcal{G}$ on $\mathcal{D}$ and $x > 0$,

$$\mathbb{E}(\mathcal{G}(\underset{\sim}{Y})1\{|\underset{\sim}{Y}_t| > x\}) = x^{-\alpha}\mathbb{E}(\mathcal{G}(x\underset{\sim}{B}^T\underset{\sim}{Y})1\{|x\underset{\sim}{Y}_{-t}| > 1\})$$

$$\mathbb{E}(\mathcal{G}(|\mathbf{\Theta}_t|^{-1}\mathbf{\Theta})|\mathbf{\Theta}_t|^{\alpha}) = \mathbb{E}(\mathcal{G}(\underset{\sim}{B}^T\mathbf{\Theta})1\{|\mathbf{\Theta}_{-t}| \neq 0\})$$

The above two conditions are equivalent and the second equality is known as the **time change formula** was obtained by **Basrak(2009)**.

B. Conditions

1. C1 condition Let $\underset{\sim}{X}$ be a stationary, stochastically continuous $\mathcal{D}$-valued process. Then, if $\underset{\sim}{X}$ is regularly varying in $\mathcal{D}$, then $\forall k \geq 1$ and $(s_1, ..., s_k) \in \mathbb{R}^k$, there exists a sequence a_n a non-zero measure $\nu_{s_1,...,s_k}$ on $\mathbb{R}^k\backslash\{\underset{\sim}{0}\}$ such that,

$$n\mathbb{P}((\frac{X_{s_1}}{a_n}, ..., \frac{X_{s_k}}{a_n}) \in .) \longrightarrow \nu_{s_1,...,s_k}$$

Moreover, for every $a < b$ and $\epsilon > 0$,

$$\underset{\delta\to 0}{lim}\,\underset{x\to\infty}{limsup}\frac{\mathbb{P}(\underset{\sim}{w}^T(\underset{\sim}{X}, a, b, \delta) > x\epsilon)}{\mathbb{P}(|\underset{\sim}{X_0}| > x)} = 0$$

where, $\underset{\sim}{w}^T(f, a, b, \eta) = \underset{(t_0,...,t_k)\in\mathcal{P}(a,b,\eta)}{inf}\ \underset{1\leq i\leq k}{sup}\ \underset{t_{i-1}\leq s,t<t_i}{sup} |f(s) - f(t)|$ and $\mathcal{P}(a, b, \eta)$ is the set of finite increasing sequences $(t_0, ..., t_k)$ with $k \geq 1$, $t_0 = a$ and $t_k = b$, and $inf(t_i - t_{i-1}) \geq \eta$

2. $\Delta(u_n)$ **condition** Let $\mathcal{F}_{j,k}(u_n)$ be the σ-algebra generated by $(\{X_i > u_n\} : j \le i \le k)$. Then $\forall$ $A_1 \in \mathcal{F}_{1,l}(u_n)$ and $A_2 \in \mathcal{F}_{l+s,n}(u_n)$, and $1 \le l \le n - s$,

$$|P(A_1 \bigcap A_2) - P(A_1)P(A_2)| \le \alpha(n, s)$$

and $\alpha(n, s_n) \longrightarrow 0$ as $n \longrightarrow \infty$ for some $s_n = o(n)$.