University of Venda
Creating Future Leaders

Stochastic Modelling of Daily Peak Electricity Demand Using Extreme Value Theory

by

Jerry Boano-Danquah

11622304

Research Dissertation for the
Master of Science Degree in Statistics

School of Mathematics and Natural Sciences
University of Venda
Thohoyandou, Limpopo
South Africa

Supervisor: Dr. C Sigauke

Co - Supervisor : Dr. K. A. Kyei

June 11, 2018

© University of Venda

Abstract

Daily peak electricity data from ESKOM, South African power utility company for the period, January 1997 to December 2013 consisting of 6209 observations were used in this dissertation. Since 1994, the increased electricity demand has led to sustainability issues in South Africa. In addition, the electricity demand continues to rise everyday due to a variety of driving factors. Considering this, if the electricity generating capacity in South Africa does not show potential signs of meeting the country's demands in the subsequent years, this may have a significant impact on the national grid causing it to operate in a risky and vulnerable state, leading to disturbances, such as load shedding as experienced during the past few years. In particular, it is of greater interest to have sufficient information about the extreme value of the stochastic load process in time for proper planning, designing the generation and distribution system, and the storage devices as these would ensure efficiency in the electrical energy in order to maintain discipline in the grid systems.

More importantly, electricity is an important commodity used mainly as a source of energy in industrial, residential and commercial sectors. Effective monitoring of electricity demand is of great importance because demand that exceeds maximum power generated will lead to power outage and load shedding. It is in the light of this that the study seeks to assess the frequency of occurrence of extreme peak electricity demand in order to come up with a full electricity demand distribution capable of managing uncertainties in the grid system.

In order to achieve stationarity in the daily peak electricity demand (DPED), we apply a penalized regression cubic smoothing spline to ensure the data is non-linearly detrended. The R package "evmix" is used to estimate the thresholds using the bounded corrected kernel density plot. The non-linear detrended datasets were divided into summer, spring, winter and autumn according to the calender dates in the Southern Hemisphere for frequency analysis. The data is declustered using Ferro and Segers automatic declustering method. The cluster maxima is extracted using the R package "evd". We fit Poisson GPD and stationary point process to the cluster maxima and the intensity function of the point process which measures the frequency of occurrence of the daily peak electricity demand per year is calculated for each dataset.

The formal goodness-of-fit test based on Cramer-Von Mises statistics and Anderson-Darling statistics supported the null hypothesis that each dataset follow Poisson GPD (σ, ξ) at 5 percent level of significance. The modelling framework, which is easily extensible to other peak load parameters, is based on the assumption that peak power follows a Poisson process. The parameters of the developed

© University of Venda

University of Venda
Creating Future Leaders

models were estimated using the Maximum Likelihood. The usual asymptotic properties underlying the Poisson GPD were satisfied by the model.

Keywords: Extreme Value Theory (EVT), Daily Peak Electricity Demand (DPED), Peaks-Over-Thresholds (POT), Poisson GPD, Maximun Likelihood Estimation(MLE).

© University of Venda

Declaration

I, Jerry Boano-Danquah declare that this research entitled "Stochastic Modelling of Daily Peak Electricity Demand Using Extreme Value Theory" is my original work and has not been submitted for any degree at any other university or institution. This dissertation does not contain other persons' writing unless specifically acknowledged and referenced accordingly.

Signed (Student): ..... Date: .....

© University of Venda

University of Venda
Creating Future Leaders

Contents

© University of Venda

University of Venda
Creating Future Leaders

© University of Venda

University of Venda
Creating Future Leaders

© University of Venda

University of Venda
Creating Future Leaders

© University of Venda

University of Venda
Creating Future Leaders

© University of Venda

University of Venda
Creating Future Leaders

List of Figures

© University of Venda

University of Venda
Creating Future Leaders

© University of Venda

University of Venda
Creating Future Leaders

© University of Venda

List of Tables

© University of Venda

University of Venda
Creating Future Leaders

© University of Venda

University of Venda
Creating Future Leaders

© University of Venda

University of Venda
Creating Future Leaders

Acknowledgements

I would like to thank the Almighty God who gave me the strength to pursue this program.

This work would not have been possible without the support, encouragement, patience, academic experience and assistance of many people.

I also thank my supervisor Dr. Caston Sigauke and co-supervisor Dr. Kwabena Kyei, the HOD of the Department of statistics for their support and directives throughout the study period.

Many thanks go to ESKOM for providing me with the DPED.

I really appreciate the National Research Fund (NRF), Univen branch for their fianancial support.

I also thank Mr. Emmanuel Numapur Gyamfi, GIMPA lecturer and a PhD student at the Department of statistics, Univen for his helpful suggestions.

Another thank you goes to all the postgraduate students in the Department of statistics, especially Mr. Emmanuel Antwi, Mr Albert Antwi, Emily, Retang, etc who made my time as a student very enjoyable. Special thanks go to Dr. Kabruse who did the proof reading.

The last but not the least, I would like to thank my first lady, Mrs. Helena Boano-Danquah and son, Jeremiah Boano-Danquah for their unflinching support and encouragement throughout the study period.

© University of Venda

Dedication

This dissertation is dedicated to my lovely wife, Mrs. Helena Boano-Danquah and son, Jeremiah Boano-Danqauh.

© University of Venda

University of Venda
Creating Future Leaders

1. Chapter 1

1.1 Introduction

This dissertation presents a stochastic modelling of daily peak electricity demand using extreme value theory, an application to South African Electricity data. Extreme value models are normally used to describe the distribution of unusual events.
Typically, an asymptotically motivated extreme value model is applied to approximate the tail of some population distribution. One of the key drawbacks associated with the application of extreme value models is the choice of "threshold" beyond which the asymptotically motivated model provides a reliable approximation to the tail of the population distribution. The threshold selection is a trade-off balance between bias and variance. Researchers should not choose the threshold too high so that there is insufficient data to reliably estimate the parameters of the model; i.e increasing the variance, but the threshold should be as high as possible, so that the asymptotic approximation is reliable; i.e little bias.

Extreme value theory (EVT) deals with the probabilistic and statistical issues in connection with very high or very low values in sequences of random variables and in stochastic processes. EVT, as a subject, is full of mathematical theory and also has a long history of applications in various areas. These include areas such as hydrology, finance, rainfall, temperature, wind speed, modelling natural phenomena such as extreme flood, earthquakes, volcanic actions, etc. In the insurance sector, EVT can be applied to model the amount of large insurance losses (claims recorded over a nominal threshold level), equity risks, day to day market risks, just to mention a few (Smith, 2003).

Embrechts *et al.,* (1997), give a detailed survey of the mathematical theory with an orientation towards applications in insurance and finance, while Coles (2001) focuses the theory on data analysis and statistical inference for extremes.

Available research has shown that, EVT has been applied to questions directed at the distribution of extremes. For example, what is the probability that a wind speed over a given level will occur in a given location during a given year? or the inverse problem of return levels? For example, what height of a river will be exceeded with a probability $\frac{1}{100}$ in a given year? This quantity is often called the 100-year

© University of Venda

University of Venda
Creating Future Leaders

return level; or what is the probability that electricity demand in peak periods (winter and summer) would exceed a given threshold at least twice in a specified time interval? Many new theories have been in existence in the past 30 years with full emphasis on the exceedances over high thresholds, the dependence among extreme events in various types of stochastic processes, and multivariate extremes. The conventional and best-known technique of analyzing extreme values is based on the extreme value limiting distributions. These distributions, pionered by Fisher and Tippett (1928), arise as limits developed for the distribution of maxima in samples of independent, identically distributed (iid) random variables. In an environmental setting, such as predicting extreme floods or sea levels, EVT can be used to model the annual maxima of the series, even though EVT can occasionally be applied to maxima over a different time period such as one month. The benchmark of these methods appears in Gumbel (1958), although there are a number of more current suggestions for fitting the distributions, such as Prescott and Walden, (1983).

A variety of techniques have been studied in recent years. Some researchers prefer to consider exceedances over high thresholds instead of maxima over fixed time periods. Looking at extreme value problems from this point of view may seem outdated, however, contemporary developments seem to have begun around 1970 with the "Peaks-Over-thresholds" or (POT) techniques (Todorovic & Zelenhasic, 1970), and further hypothesized in the English "Flood Studies Report"(NERC, 1975). Analogous to the above research, Hill and Pickands (1975), and Weissman (1978) proposed a mathematical development of techniques based on a certain number of extreme order statistics, and the Generalized Pareto distribution (GPD) as an appropriate distribution for threshold exceedances (Davison & Smith, 1990).

Load characterization plays a very important role in the planning and designing of electrical power systems. The problem is well noted and analyzed for any power system as peak demand may exceed the maximum power generated, which may lead to power outages and load shedding. Instability in the grid system can create more inconvenience especially in South Africa where about 95% of the citizens depend on electricity for domestic use. Peak electricity demand in a given season may be due to a range of uncertainties, for instance, underlying population growth, change in technological trend, economic situations, climate change/temperature-covariate, as well as the general randomness inherent in individual usage. It is also linked to some known calender effects due to the time of the day, day of the week, time of year, and public holidays. The peak loads demonstrate fast changes under these situations and a large degree of randomness, whose description needs a proper analysis by using Poisson

© University of Venda

University of Venda
Creating Future Leaders

GPD and stationary point process characterization of extremes.

1.2 Statement of the Problem

Electricity usage in South Africa is increasing continuously with the increase in population size, technology, electrical appliances and economic growth. Sigauke et al., (2012) use the Generalized Pareto Distribution to model the daily increases in peak electricity damand. The Electricity Supply Commission (ESKOM), South Africa's power utility company is expected to boost the supply of energy load to meet the demand (Inglesi, 2010). Due to this, new generating plants have been built in Medupi, Kusile and Ingula together with renewable energy transmission projects in the North grid, central grid, greater East London and Ingula scheme. Planners and decision-makers in the power utility companies like ESKOM face uncertainties in the demand of electricity because temperature is one of the main drivers of electricity demand.

According to Gencay and Selcuk (2004), Hahn, Meyer-Nieberg and Pickl (2009), most of the typical statistical techniques such as t-distributions , normal distributions, time series procedures which include the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) assume symmetric distributions and are therefore not suitable for modelling the tail behavior of the fat and heavy tailed distributions. This problem is catered for by using EVT distributions. Notwithstanding the numerous advantages of EVT, the common setback of the theory lies in the shortage of data for statistical modelling of extreme events (Coles, 2001).

The motivation behind this research stems from the need to assess the magnitude and frequency of occurrence of extreme peak electricity demand over a specified threshold using South African electricity data. Both grid companies and large consumers want to estimate their electricity consumption in the future, in order to ensure secured and economic system operation (Zhang *et al.,* 2012). Energy intensive enterprises (EIEs) in South Africa include steel , alumina and petrochemical, among others. Kou and Gao (2014), further claim that, in many cases, EIEs have their own generating plants, thus forming a grid. Notably, in many industrial countries, the electrical energy consumption in energy-intensive enterprises constitute a significant part of the country's total energy use (Kou & Gao, 2014).
Ability to estimate the frequency of ocurrences of extreme peak electricity demand can lead to probabilistic load forecasting which is a key point in effective operation and planning of power systems.This is because

© University of Venda

forecasting accuracy has significant impact on electric utilities and regulators. Occasionally, if electricity demand is overestimated it will cause abnormal operation, which lead to the startup of too many units causing needless level of reserve or excessive energy purchase, as well as substantial wasteful investment in the construction of excess power facilities. On the other hand, underestimation may result in risky operations and inadequate demand, leading to insufficient preparation of spinning reserve and this can cause the system to operate in a vulnerable manner.

1.3 Purpose of the study

1.3.1 Aim

The main aim of the study is to assesss the frequency of occurrence of extreme peak electricity demand using Poisson GPD and stationary point process characterization of extremes.

1.3.2 Objectives

The main objectives of the study are to:

1. estimate the frequency of ocurrence of extreme peak electricity demand using Poisson GPD,
2. detect extreme peak load over a sufficiently high threshold using a stationary GPD and determine whether or not there exists a trend in threshold excesses,
3. investigate the effect of modelling temporal dependence,
4. estimate frequency of occurrence of extreme peak load based on stationary point process characterization of extremes,
5. estimate the parameters of the identified EVT distribution using maximum likelihood estimation,
6. suggest areas for further research.

© University of Venda

University of Venda
Creating Future Leaders

1.4 Significance of the study

Information obtained from this study can help ESKOM in planning for future electricity generation facilities and transmission augmentation. This is very crucial because over a long period of time, management, decision makers and planners must embrace a probabilistic view of potential peak demand levels. The outcome of density forecasts is helpful in the estimation of the full probability distributions of the possible future values of the extreme peak electricity demand which is useful in the planning and decision-making process than point forecasts.

1.5 Scope of the study

A quadratic cubic spline is fitted to the daily peak electricity demand to ensure the data is detrended. A frequency analysis based on the time varying data is performed. The time varying datasets were divided into four seasons; winter (1st June to 31th August), spring (1st September to 30th November), summer (1st December to 28/29 February) and autumn (1st March to 31st May) according to the calender dates in the Southern Hemisphere and Poisson GPD is fitted to stationary dependent series. We declustered the sequences using Ferro and Segers (2003) declustering algorithm and extracted the cluster maxima using the R package "evd". We fit the Poisson GPD and stationary point process to the cluster maxima and estimated the intensity function of the point process which measures the frequency of the occurrence of the daily peak electricity demand per year for each dataset.

The parameters of the models were estimated using the method of maximum likelihood estimation (MLE) technique. In addition to this, 95% return levels have been predicted to forecast and capture the entire probability distribution of the frequency of the occurrence of daily peak electricity demands for each of the seasons in South Africa, which is in contrast to all the previous studies which had focused mainly on modelling the average daily peak electricity demand. Several model diagnostic tools and plots such as QQ plots, probability plots, density and return level plots have been used to check the validity and fit of the models as well as in choosing sufficiently high threshold for the peaks-over-thresholds approach.

Cramer-Von Mises and Anderson-Darling goodness of fit tests have been performed to check the efficiency of the model fits. The null hypothesis which states that each dataset follow Poisson GPD (σ, ξ) is tested against the alternative hypothesis which states that each dataset does not follow Poisson GPD (σ, ξ). The p-value of each test was greater than the 5% level of significance indicating that, we

© University of Venda

University of Venda
Creating Future Leaders

fail to reject the null hypothesis and concluded that each dataset followed Poisson GPD (σ, ξ).

1.6 Bootstrapping

This is the statistical technique mainly used to estimate the sampling distribution by simulating based on a good model of the data (Rochowicz, 2010). Notably, bootstrapping is important when inference about the data is to be made, based on a complex procedure which produces theoretically unavailable results (Davison & Kuonen, 2002).

© University of Venda

2. Chapter 2

2.1 Literature Review

2.1.1 Introduction

This chapter reviews some relevant research which have been conducted previously by other researchers in relation to the problem. Studies based on the use and applications of generalized extreme value distribution (GEVD), generalized Pareto distribution (GPD), Poisson generalized Pareto distribution (Poisson GPD) and point process characterisation of extremes are summarized. Several methods of analyzing extreme values are now known, most of which are based on the extreme value limit distributions or related families. The study reviews some of these methodologies and proposes some extensions based on the point-process view of high-level exceedances over a specified threshold. These ideas are illustrated with detailed analysis of the frequency of the occurrence of extreme peak electricity demand which is applied to data from ESKOM, South Africa's power utility company. According to Coles (2001), any framework of classical extreme value theory is concerned with the detection and analysis of extreme observations based on the behaviour of the annual block maxima of r largest order statistics. Most of the research articles reviewed in this study, however used the ordinary GEVD (for r=1) and the MLE for statistical inference on the target parameters. This research focused on the MLE for parameter estimation.

2.2 Applying the block maxima approach

Hor *et al.,* (2006), in their study outline the use of the block maxima approach to estimate the maximum load forecast errors (residuals), up to several decades ahead in order to assess the risk inherent in long-term electricity demand projection taking into account climate effects. The study discusses the maximum residuals in terms of return levels and return periods that would occur with a finite time series. Empirical results show that using a GEV model, the number of return observations exceeding a given return level is closely in line with what was predicted.

Hasan and Kassim (2012), model extreme temperature using GEVD at Penang, Malaysia. The study was based on the occurrences of the climatological disorder including floods and weather conditions

© University of Venda

University of Venda
Creating Future Leaders

such as maximum temperature that may negatively influence the living conditions of the Malaysian population at Penang. The GEVD as a classical approach of EVT is used to assess the frequencies and also to describe the stochastic behavior of maximum temperature as an extreme event.The study aimed at reaching informative results that may contribute towards the preparation of the community in relation to extreme weather conditions. The study used average daily temperatures captured in degree Fahrenheit over 32 years, from 22 meteorological stations at Penang. The study results emphasized the need for non-stationary GEVD model on 11 meteorological stations 8 of which showed the existence of trend.

Coles (2001) emphasizes that a GEVD that is fitted on r largest ordered observations in a block provides an improved frequency analysis based on block-maxima and block-minima as compared to the ordinary GEVD. Depending on the availability of extremes in the dataset, threshold-based models could handle the situation better than the block models. However, the EVA based on threshold exeedance models such as the GPD provide more reliable outcome since they tend to exploit as much as possible, the available data (Coles, 2001; Smith, 1989).

2.3 Applying the peaks-over-thresholds approach

There are two fundamental approaches for practical extreme value analysis. The first approach depends on deriving block maxima (minima) series as a preliminary step. In many situations, it is common practice and convenient to extract the annual maxima (minima), creating an "Annual Maxima Series" (AMS). The second approach depends on extracting, from a continuous record, the peak values reached at any period during which values exceed a certain predetermined threshold (falls below a certain threshold). This method is generally termed as the "Peaks-Over-Thresholds" method (POT).

With reference to the annual maxima series data, the analysis may partly depend on the outcome of the the theorem emphasized by Fisher and Tippett (1928), leading to the generalized extreme value distribution being selected for fitting. In practical situations however, numerous techniques exist which have been applied to select between a wider range of distributions. The theorem here applies to the limiting distributions for the minimum or the maximum of a very large collection of independent random variables from the same distribution. In situations were the number of relevant random events per year may be rather limited, it is not an uncommon practice that the analyses of observed annual maxima series data often results in distributions that are rather different from the generalized extreme value distribution (GEVD) being selected.

© University of Venda

University of Venda
Creating Future Leaders

With regards POT data, the analysis may involve fitting two distributions: one for the number of events in a time period considered and a second for the number or size of the exceedances over the chosen threshold. Some authors reserve the term 'POT method' for cases where the thresholds are non-random, and distinguish it from the case where one deals with exceedances of a random threshold.

2.3.1 Generalized Pareto distribution

Extreme value analysis of a large number of relatively short time series are in large demand in the area of environmental sciences and design. Fukutome, Liniger and Suveges (2014), present an automated method of the POT technique to extreme hourly precipitation in Switzerland. They emphasized in their modelling framework that the peaks-over-threshold approach provides a good fit of the generalized Pareto distribution (GPD) to independent exceedances above some sufficiently high threshold. In order to ensure independence among the threshold exceedances, the time series were pruned and the exceedances (separated by less than a fixed interval known as the run parameter) were considered as a cluster, and all but the cluster maxima were discarded. The study used the automation of an existing graphical approach for joint selection of threshold and run parameter (run parameter is the minimum separation interval between exceedances). The study analysed hourly precipitation at 59 stations of the MeteoSwiss observational network over the period from 1981 to 2010. According to the study, extreme hourly precipitation is linked to flash floods, mudslides, landslides and debris flows in the Alpine region, which can trigger avalanches, in the form of snow.

The study in conclusion proposed an automated procedure for selection of threshold and run parameter in the POT approach to extreme value analysis that was based on the graphical method developed by Suveges and Davidson (2010). The above procedure creates a subset of acceptable threshold-run parameter pairs and in this subset, chooses the pairs that create the largest number of clusters. Extreme events have the tendency to cluster and this shows dependency in the data. Dependence between high exceedances should be effected in the mean cluster size. According to Suveges and Davidson's study, lag dependence of hourly precipitation at extreme levels was calculated with the help of the dependence measures, in an attempt to put the findings into context. Suveges and Davidson (2010) made use of the information matrix test (IMT) to determine the likelihood of the limit law of the inter-exceedances times. The IMT gives a quantitative assessment of the compatibility of the threshold-run parameter pair with the two-dimensional extremal process. The use of the IMT is based on trial and error of all

© University of Venda

University of Venda
Creating Future Leaders

the combinations of the two parameters in a plausible range which results in a list of pairs that are not rejected at the 5% confidence level, that is with $IMT > 3.84$. The simulated proceedure proposed is pragmatic because only the subsets with the IMT values close to zero were selected to ensure small or less misspecification liability.

In the study, the GPD estimates, based on POT analysis of the cluster maxima resulting from the automated selection, bring to light many known features regarding precipitation in the Alpine region. Empirical observations from the study reveal that the signal due to thunderstorm activity was more visible in the seasonal frequency of events, rather than in their severity, as demonstrated by the return levels for high return periods.

The results of the study are consistent with the argument proposed by Fawcett and Walshaw (2007), that unnecessary large run parameters have negative results on the subsequent estimation of the GPD parameters. The IMT selection allows for lower run parameter as compared with the reference selection which allows fixed threshold and run parameter. This leads to higher return levels and in reality to more stringent design measures, a positive outcome considering the uncertainty associated with planning long-term structures based on only 30 years of data. Finally, the pattern of extreme hourly precipitation suggests that requirement for the observational network may be different for hourly rather than for daily precipitation.

According to Coles (2001), using the block approach in the presence of an entire time series appears as a waste of information because only the annual maxima or annual minima are taken into consideration. Due to this limitation, any analysis that makes use of POT leads to the GPD for threshold exceedances which has the ability to utilize as much of the available information as possible (Coles, 2001). The POT approach is also used to model extreme events. Under this technique, observations are called 'extreme values' if they exceed a defined value which is a sufficiently high threshold, after which the GPD is then fitted to the excesses above the defined threshold.

Cai and Campbell (2005), in their modelling framework apply the POT technique in the analysis of the rainfall over several weather stations that were scattered geographically in the Southwest Australia. The main aim of the study was to model the behavior of the time series of rainfall at the chosen meteorological stations above a sufficiently high threshold. The study used the mean excess plot for the threshold selection. The conditional distribution used by, Li *et al.,* (2005) in modelling the excess above the selected threshold was given by Balkema and De Haan (1974). The study applied a change point

© University of Venda

University of Venda
Creating Future Leaders

as an EVT technique in an extreme rainfall distribution and concluded that there had been substantial increase in the winter rainfall since the mid-twentieth century. The change in the winter extreme rainfall was quantified by estimating the tails behavior of the extreme rainfall distribution which confirmed the existence of spatial variation (Li *et al.,* 2005).

Kysely, Picek and Beranova (2010), model the occurrence of non-stationary extreme quantiles in the distribution of daily temperature. The approach is usually applied with a threshold that changes with respect to time. The authors observed the existence of trends within the observations and this leads to the violation of stationarity assumption thereby heading to a non-stationary GPD with time-varying parameters. The study made use of data for the period spanning from 1996 to 2010.

Bommier (2014), applied the non-stationary peak over threshold approach with time-varying scale in modelling environmental data. The author used the deviance statistics for model comparison and the study confirmed that among the two fitted models, the first model whereby the scale parameter σ and the shape parameter ξ were time homogeneous showed significant improvement more than the second model where only the scale parameter, σ is time dependent.

Eastoe and Tawn (2009), modelled non-stationary extremes with application to surface level ozone. In their modelling framework, an alternative approach that makes use of preprocessing methods was used to model the non-stationarity in the body of the process and thereafter, standard methods were used to model the extremes of the preprocessed data. A simulation study was used to demonstrate both the standard and the preprocessing methods in relation to a study of the ozone data. The analysis of the extremes of a series is important in situations where non-stationarity in the series is evident, especially in the context of environmental data sets. Eastoe and Tawn (2009) focussed on the analysis of the ozone data set which consists of the daily maxima of hourly concentrations of surface level ozone. The data set was obtained from a monitoring site in central Reading, UK, which form part of an automated air quality monitoring network that is managed under the auspices of the UK Government. According to the study, the apparent non-stationarity of the ozone data (mainly summer peaks and winter troughs) was due to the underlying process driving the ozone generation process.

The significance of the study lies in it explaining the changes in the extreme ozone levels based on the covariates relating to the precursor concentrations and meteorological conditions and hence coming

© University of Venda

University of Venda
Creating Future Leaders

out with a summary of the joint distribution of the extreme ozone levels under current conditions and for situations relating to future changes in patterns of emission and climate change. Eastoe and Tawn (2009), model the non-stationarity in the entire data set. The non-stationarity is then taken out of the data set, a technique that is called preprocessing, and the extremes of the preprocessed data are modelled by using the standard approach. What makes this technique unique is that, if the preprocessing is successful the extremes of the preprocessed series will have had most, if not all, of the non-stationarity removed and this enables a simple extreme value analysis of the preprocessed series to be done.

According to the findings of the study, the preprocessing approach is better because the reasons for non-stationarity in a data set are often intricately linked up with the mechanisms generating the underlying process, so modelling non-stationarity in the underlying process is more likely to capture the appropriate form of non-stationarity. The simulation study also confirms the benefits of exploiting the full structure of the mechanism, showing that the preprocessing approach is more likely to select the covariate model correctly than the standard approach. The preprocessing approach also produces a simpler and more efficient model fit and for hypothesis testing in the extreme value components of the models, the preprocessing approach changes the strategy of covariate model fitting, to be scientifically rational. The preprocessing approach also has the advantage of being simpler in terms of computation due to the fact that extremes of the transformed process Z_t are closer to stationarity than those of the original process Y_t. The study suggested that further work is necessary to investigate how to model the GPD parameters in the presence of covariates to retain the threshold stability property, so that it holds, even if the preprocessing methods suggests that the covariates are necessary.

The study is further supported by Smith (1989), who demonstrated that there exists a connection between the generalized extreme value distribution parameters and the GPD parameters through generalized point process results for the extreme values. All the parameters of the generalized extreme value distribution are threshold invariant, so exploiting this link may help. Based on this property, the study preferred the use of the generalized extreme value model instead of the GPD model for extreme value modelling.

Chan and Nadarajah (2015), suggest the use of the GPD to explain how extremes of electricity demand changes with time in the United Kingdom. The main idea underlying their study was to allow the GPD parameters to vary linearly and in a non-stationary sinusoidal manner. This was done in order to capture

© University of Venda

University of Venda
Creating Future Leaders

some patterns exhibited by electricity demand data. Initially, four different models were fitted, both stationary and non-stationary to investigate the significance of seasonality and trend.

Sugahara *et al.,* (2009), conducted research on an assessment of extreme value of daily rainfall in the city of Sao Paolo, over the period 1933-2005 using the GPD. The study used a time varying threshold, dependent on the day of the year, denoted as $q_p(t_d), t_d = 1, 2, ..., 365,$ composed of p-quantiles of daily rainfall amount obtained from all available data. Non-stationarity features present in the Sao-Paolo data , such as seasonality and long-term trend were incorporated as covariates.

2.3.2 Applying the point process characterization of extremes

The point process characterization of extremes introduced by Pickands (1971) has been used by several authors and researchers including Smith (1989), Coles (2001) and Beichelt (2006), who emphasized the importance of a point process technique in EVT modelling. The two main features of the Poison point process are singled-out: one, Poisson point process provides the analysis of extremes through merging the block and POT approach; two, the point process is more associated with the variations in the excesses above the threshold than the peak over threshold approach. Smith (1989), applied the EVT in modelling the environmental data that were captured in Houston, Texas. The non-stationary GEVD, non-stationary GPD as well as the point process were used. The aim of the study was to investigate the existence of trend in the data which consisted of hourly measurement of ozone from April 1973 to December 1986. The study aimed at estimating the frequency with which specified high levels are exceeded as well as the desire to determine whether or not there is any evidence of frequency changing over the period of the study.

A point-process viewpoint of high-level exceedances was proposed in this study as it illustrate the wide applicability of extreme-value procedures. Seasonality was inherent in the data and especially because extreme values tended to occur in clusters; clusters of high-level exceedances were identified with the intention of concentrating on cluster maxima for the rest of the data analysis. An arbitrary threshold and cluster intervals were chosen in order to identify clusters of exceedances. A cluster interval of 72 hours was finally adopted since this produced satisfactory results.

Smith (1989), concluded that there was a downward trend in the crossing rates at high levels. This indicates that the work of the regulatory bodies has a significant effect by at least reducing the frequency of the occurrences of high emissions and without such regulations, one could argue that there would be an increasing trend. The other evident feature from the analysis of the ozone data is that, unlike many

© University of Venda

University of Venda
Creating Future Leaders

extreme value problems whereby the shape parameter ξ is estimated to be close to zero, in the analysis of the ozone data, ξ however had significant positive values. In situations where the shape parameter ξ, has negative values, the data analysis becomes difficult because the tail of the distribution becomes long and precise estimation of the quantiles become difficult. This problem is discussed in detail by Davidson and Smith, (1990). Finally, Smith (1989) considered a single long tail series, without relating ozone emission to other covariates such us temperature or to data values at different sites, therefore he suggested that multivariate extremes should be a subject of more research and the methodology proposed in ozone data study be extended.

2.4 Applying extremal mixture models

2.4.1 Introduction

Extreme value theory has been applied to develop models for the behavior of the distribution functions in the tail and it has proven to be a reliable foundation for extrapolations (Hu, 2013). Within the context of conventional GPD applications, the data below the chosen threshold (non-extreme data) are discarded but only those inherently scarce extremal data will be fitted by the GPD models. The extremal mixture model consists of the bulk and tail model. The tail model is capable of describing the upper, lower or both tails based on which extremes are of interest to the researcher. Conventionally, the bulk model is a parametric, semi-parametric and non-parametric form as appropriate for the application whilst the tail model is a flexible threshold model (Hu, 2013).

Recently, interest in the application of extreme value mixture models by researchers have been on the rise as the models provide an automated approach both for the estimation of the threshold, quantification of the uncertainties arising from threshold choice and the use of all the data for parameter estimation in the inference (Hu, 2013).

2.4.2 Advantages of using extreme value mixture models

1. Extreme value mixture models have the capacity to capture both the bulk and tail distributions at the same time. It caters for all the available data without wasting any information. One main goal of extreme value mixture model is to choose a flexible bulk and tail model which fit to the non-extreme data as well as to the extreme data simultaneously (Hu, 2013).

© University of Venda

University of Venda
Creating Future Leaders

2. The choice of threshold is very problematic within the context of extreme value literature. There are situations which come with multiple suitable thresholds in some datasets. Choosing different threshold may result in different tail and corresponding with different return levels. Instead of using the traditional fixed threshold approach, these mixture models treat the threshold as a parameter to be estimated. Moreover, the inference method used in mixture models cater for all the uncertainties related to threshold estimation. The major advantage of the extreme value mixture models is that of taking into account the uncertainties associated with threshold selection, apart from providing a more natural way of threshold estimation (Hu, 2013).

3. Major work have been made in coming up with different types of bulk models. Particularly, many different semi-parametric or non-parametric statistical techniques have been suggested in litereature for bulk model, which seeks to describe the bulk model without assumptions on the model form. In situations where models are misspecified, the parametric bulk model doesn't give much flexibility compared with semi-parametric bulk models (Hu, 2013).

2.4.3 Disadvantages of using extreme value mixture models

1. The fundamental properties of extreme value mixture models are still not well understood and still need to be investigated further. These mixture models are unfortunately not commonly available in any existing software for practitioners to apply and develop further (Hu, 2013).

2. Another problem associated with mixture models is that they assume no influences between the bulk fit and the tail fit. Because both the bulk and the tail components share the threshold with each other, they are not independent, hence if the bulk fit strongly affects the tail fit, the threshold will also be affected. Under this circunmstance, it is not guaranteed that such a mixture model will provide a reliable quantile estimates. There is a lack of sensitive analysis of robustness issue of the bulk and the tail distribution in literature even though some discussions about robustness are available, a comprehensive study is still required (Hu, 2013).

3. Some extremal mixture models are discontinuous at the threshold.They do not have continuous density functions at the threshold and whether this is a problem for tail quantile estimation or whether the extra continuity constraints imposed would improve the model performances are still unclear in literature. In general, the advantage of a model without the continuity constraints is that it is more flexible compared to the model with the continuity constraints. However, it is

© University of Venda

University of Venda
Creating Future Leaders

important to have a continuous density function, by imposing the continuity constraints over the bulk and tail distribution , it has great impact on threshold estimation. Thus there is still a major concern about the robustness of both the bulk and tail fit (Hu, 2013).

2.5 Literature Review in Extremal Mixture Models

2.5.1 MacDonald et al., (2011)

MacDonald et al.,(2011), constructed a standard kernel density estimator as the bulk with a GPD or point process tail model. The main motivation of formulating such model is that they aim to provide a more flexible framework for extreme value analysis.

The distribution function is defined as:

$$[F(x|X,\lambda,u,\sigma_u,\xi,\phi_u) = \begin{cases} (1-\phi_u)\frac{H(x|X,\lambda)}{H(u|X,\lambda)} & x \leq u, \\ \\ (1-\phi_u) + \phi_u \times G(x|u,\sigma_u,\xi) & x > u, \end{cases} \tag{2.1}$$

where $H(.|X,\lambda)$ is the distribution function of the kernel density estimator and λ is the bandwidth.

MacDonald et al., (2011) also introduced a two tail model. This mixture model is contructed by splicing the standard kernel density estimator with two extreme value tail models. This two-tailed extreme value model also overcomes the inconsistency in the cross-validation likelihood estimation of the bandwidth for heavy tailed distribution. The parameter vector is $\Theta = (X, \lambda, u_1, \sigma_{ul}, \xi_l, \phi_{ul}, u_2, \sigma_{ur}, \xi_r, \phi_{ur})$.

The distribution function of two tailed mixture model is given by:

$$[F(x|\Theta) = \begin{cases} \phi_{u_l}(1 - G(-x| - u_l, \sigma_{u_l}, \xi_l)) & x < u_l, \\ \\ \phi_{u_l} + (1 - \phi_{u_l} - \phi_{u_r})\frac{H(x|X,\lambda) - H(u_l|X,\lambda)}{H(u_r|X,\lambda) - H(u_l|X,\lambda)} & u_l \leq x \leq u_r, \\ \\ (1 - \phi_{u_r}) + \phi_{u_r} G(x|u_r, \sigma_{u_r}, \xi_r) & x > u_r, \end{cases} \tag{2.2}$$

where ϕ_{u_l} and ϕ_r are estimated as the sample proportions less than the threshold u_l and above the threshold u_r respectively. $G(-x| - u_l, \sigma_{ul}, \xi_l)$ is the unconditional GPD function for $x < u_l$ and $G(x|u_r, \sigma_{ur}, \xi_r)$ is the unconditional GPD function for $x > u_r$. These two unconditional GPD functions can be represented by corresponding point process representation to remove the dependence between the threshold and the GPD scale parameter.

© University of Venda

University of Venda
Creating Future Leaders

2.5.2 MacDonald et al., (2013)

MacDonald et al., (2013) use both a boundary corrected kernel density estimator for bulk model which is cut at the threshold with a point process representation of the GPD tail model. What the boundary corrected kernel density estimator does is to reduce the bias inside of the kernel density estimator. There exists so many types of boundary corrected based kernel density estimators proposed in literature, notwithstanding this, many of these estimators can assume negative values at the boundary. MacDonald et al.,(2011) make use of the non-negative boundary corrected estimator which was proposed by Jones and Foster (1996) to handle the negative values close to the boundary.

The distribution function of the boundary corrected mixture model is given by:

$$F(x|X,\lambda_{BC},u,\sigma_u,\xi,\phi_u) = \begin{cases} (1-\phi_u)\frac{H_{BC}(x|X,\lambda_{BC})}{H_{BC}(u|X,\lambda_{BC})} & x \leq u, \\ \\ (1-\phi_u) + \phi_u \times G(x|u,\sigma_u,\xi) & x > u, \end{cases} \tag{2.3}$$

where $H_{BC}(x|X,\lambda_{BC})$ is the distribution function of the non-negative boundary corrected kernel density estimator and ϕ_u is estimated as the sample proportion of the data above u.

2.6 Formal goodness-of-fit tests for the generalized Pareto distribution

2.6.1 The Cramer-Von Mises statistics and the Anderson-Darling statistics

Choulakian and Stephens (2001) apply the Cramer-Von Mises statistics and the Anderson-Darling statistics to the the exceedances over given thresholds for 238 river flows in Canada. They consider the most practical situation where the parameters are unknown and demonstrated how the tests can be applied for threshold selection within the POT modelling framework. Choulakian and Stephens (2001) concluded that the tests are useful for threshold estimations and further researched on the approximation of the Poisson GPD to other distributions that might be used for long-tailed data.

© University of Venda

3. Chapter 3

3.1 Methodology

3.1.1 Introduction

This chapter presents the extreme value techniques that have been applied in this dissertation. Electricity data exhibit a large degree of non-stationarity, hence, the study covers stationary dependent series and the use of extremal mixture models where a non-parametric kernel density is fitted to a fixed threshold on the positive residuals above the time varying threshold.

The extremal index which is the measure of dependence or independence of the time series data is estimated. The data is declustered according to Ferro and Segers (2003) declustering technique and the cluster maxima are selected for both poisson GPD and point process analysis. The dissertation also considered the block maxima which has the tendency to minimize dependency in the data series.

In situation where $r = 1$, we have the usual block maxima. Due to the insufficient amount of data, researchers normally make use of r largest order statistics in extreme value analysis. This section also discusses threshold selection criteria. The modelling framework also explained the point process characterization of extremes together with extremes of clustered and dependent sequences, parameter estimates, model checking and model diagnosis.

3.2 Block maxima

Coles (2001), points out that for the properties of block maxima to hold, the data must have a weak long range dependence at extreme level such that the block maxima is distributed with the same family of distributions as in the case of dependent series. The block maxima approach has the tendency of making the effect of dependency in the data series (electricity data) very small (negligible) according to Coles (2001). The block maxima approach as a technique in EVT consists of dividing the observation period into a non-overlapping periods of equal duration. This technique focuses only on the maximum data values within each period and the theory is also applicable to block minima using the duality principle.

© University of Venda

University of Venda
Creating Future Leaders

3.3 Peaks-over-thresholds

Specifically, if x is a random variable representing the extreme daily peak electricity demand, and u is an arbitrary chosen threshold, then the random variable $x - u$ represents the value of the exceedances over the sufficiently high threshold u, provided that this threshold has been exceeded.

3.3.1 Generalized pareto distribution

Poisson GPD can be used for the modelling of the tails of distributions for data exceeding certain threshold. Let the distribution of X conditionally on exceeding some high threshold u (so $Y = X - u > 0$):

$F_u(y) = Pr\{Y \leq y | Y > 0\} = \frac{F(u+y)-F(u)}{1-F(u)}$

As $u \to \omega_F = sup\{x : F(x) < 1\}$, we often find a limit

$$F_u(y) \approx G(y; \sigma_u, \xi) \tag{3.1}$$

where G denotes the Generalized Pareto Distribution, Coles (2001).

$$G(y, \sigma, \xi) = 1 - (1 + \xi \frac{y}{\sigma})_+^{\frac{-1}{\xi}} \tag{3.2}$$

The Pareto and similar distributions have been used as models for long-tailed processes, with the robust link with the classical EVT being established by Pickands (1975). Thus there exists a relationship between the limit for sample maxima and limit results for exceedances over thresholds, which is quite extensively exploited in the contemporary statistical methods for extremes.
In the GPD modelling approach, the situation whereby $\xi > 0$ is long-tailed, for which $1 - G(x)$ decays at the same rate as $x^{\frac{-1}{\xi}}$ for large x. This is reminiscent of the normal case of the Pareto distribution, $G(x) = 1 - cx^{-\alpha}$, with $\xi = \frac{1}{\alpha}$. For $\xi = 0$, we may take the limit as $\xi \to 0$ to obtain

$G(y; \sigma, 0) = 1 - exp(-\frac{y}{\sigma})$

i.e. exponential distribution with mean σ. For $\xi < 0$, the distribution has finite upper endpoint at $u - \frac{\sigma}{\xi}$.

© University of Venda

University of Venda
Creating Future Leaders

Some other elementary results about the GPD are as follows:

$$\begin{aligned} E(Y) &= \frac{\sigma}{1-\xi}, (\xi < 1), \\ Var(Y) &= \frac{\sigma^2}{(1-\xi)^2(1-2\xi)}, (\xi < \frac{1}{2}), \\ E(Y-y|Y>y>0) &= \frac{\sigma+\xi(y)}{1-\xi}, (\xi < 1) \end{aligned} \tag{3.3}$$

Suppose we observe i.i.d. random variables $X_1, ..., X_n$, and observe the indices i for which $X_i > u$. Then if these indices are rescaled to points $\frac{i}{n}$, they can be viewed as a point process of rescaled exceedance times on [0,1]. If $n \to \infty$ and $1 - F(u) \to 0$, such that $n(1 - F(n)) \to \lambda (0 < \lambda < \infty)$; the process then converges weakly to a homogeneous Poisson process on [0,1], of intensity λ.
Motivated by the above, one can imagine a limiting form of the joint point process of exceedance times and excesses over the threshold, of the following form:

(a) The number, N, of exceedances of the level u in any one year has a Poisson distribution with mean λ;

(b) Conditionally on $N \geq 1$, the excess values $Y_1, ..., Y_N$ are i.i.d. from the GPD. This is called the Poisson GPD model.

3.3.2 Threshold selection

If the observations $x_1, ..., x_n$ represent the independently and identically distributed (i.i.d) average daily peak electricity demand and suppose u denote an arbitrarily and sufficiently high threshold, then the observations $x_1, ..., x_k$ are the k positive residuals above the sufficiently high threshold u if $x_i : x_i > u$ and the observations $y_j = x_j - u$, for $j = 1, ..., k$ are the threshold excesses (Coles, 2001). Coles (2001), also emphasizes that the selection of threshold processes is always a trade-off between the bias and variance and, if the chosen threshold is too small, it violates the asymptotic properties underlying the derivation of the GPD. However, if the chosen threshold is too high, the excesses $(x - u)$ above the threshold becomes too small to estimate the shape and the scale parameters, leading to high variance. Hence the threshold selection requires proper analysis to determine whether the limiting model provides a sufficiently good approximation versus the variance of the parameter estimates. There are numerous diagnostic models used to determine the threshold, such as, the Pareto quantile plot, mean excess plot, threshold stability plot, extremal mixture plots among others.

© University of Venda

University of Venda
Creating Future Leaders

3.3.3 Stationarizing the data

In order to ensure homogeneous processes, two main approaches were used to make the data stationary, i.e. non-linear detrending and differecing approaches. The graphs are found in the appendix 1 with the winter, spring, summer and autumn graphs superimposed on the non-linear detrended graphs.
The data is initially detrended using a penalized regression cubic smoothing spline given in the equation below:

$$\eta(t) = \sum_{i}^{n}(y_i - f(t_i))^2 + \lambda \int (f^{''}(t))^2 dt + u \tag{3.4}$$

where y_i denotes the daily peak electricity demand and λ is a smoothing parameter and $u \epsilon \Re$ is a shift factor which should be large enough to accommodate asymptotic conditions when GPD is fitted to the observations exceeding u. We use the extremal mixture models to determine the value of u and observations above the time varying threshold $\eta(t)$ are extracted without the shift factor. We estimate the positive shift factor u using extremal mixture models. Wang (2011) emphasized that the use of penalized cubic smoothing splines as a time-varying threshold has an attractive features such as deseasonalizing and detrending at one stroke, therefore penalized cubic smoothing splines have been incoporated in this study.
Since the electricity data is characterised with nonstationarity, we applied both non-seasonal and seasonal differencing techniques to keep the DPED in statistical equilibrium. The non-seasonal differencing technique is given in the equation below:

$$Z_t = \bigtriangledown y_t = (1 - B)y_t = y_t - y_{t-1} \tag{3.5}$$

where $y_t = DPED$ and $y_t - y_{t-1}$ represents the daily changes in electricity demand.
The seasonal differencing technique is given in the equation below:

$$S_t = \bigtriangledown^s y_t = (1 - B^s)y_t = y_t - y_{t-s} \tag{3.6}$$

3.3.4 The threshold stability plot

According to Hu and Scarrott (2013), GPD shows stability characteristics. When the estimates of the scale parameter (σ) and the shape parameter (ξ) are plotted against several threshold values u, we obtain a threshold stability plot which is taken as one of the threshold selection method since it has many advantages. The threshold stability plot provides a benchmark for determining a range of several thresholds for the invariance in the extremal index estimator.

© University of Venda

According to Heffernan and Southworth (2013), the threshold stability plots also provide the standard for the selection of the lowest threshold above which estimates of the extremal index are approximately constant. Bommier (2014), notes that if all the observations above the threshold (u) follow a Poisson GPD with the parameters ξ and σ_u, for any higher threshold (u_0), i.e $u_0 > u$, all the excesses will follow a Poisson GPD with ξ and scale parameter of:

$$\sigma_{u0} = \sigma_u + \xi(u_0 - u) \tag{3.7}$$

parametrizing the scale parameter σ_{u0} :

$$\sigma^* = \sigma_{u0} - \xi u_0 \tag{3.8}$$

Given u_0 as a sufficiently higher threshold then σ^* does not depend on (u_0). The threshold stability plot of the Poisson GPD means that the threshold (u) must be selected so that the shape parameter and the scale parameter remain constant, after taking the sampling variability into account.

3.4 The use of extremal mixture models

The extremal mixture models can be parametric, semi-parametric and non-parametric. The distribution function of the parametric form of the bulk component of the extreme value mixture model can be defined as:
Bulk Model Based Tail Fraction Approach,

$$F(x|u,\theta,\sigma_u,\xi,\phi_u) = \begin{cases} H(x|\theta) & x \leq u, \\ H(u|\theta) + [1 - H(u|\theta)] \times G(x|u,\sigma_u,\xi) & x > u, \end{cases} \tag{3.9}$$

Parameterised Tail Fraction Approach,

$$F(x|u,\theta,\sigma_u,\xi,\phi_u) = \begin{cases} (1-\phi_u) \times \frac{H(x|\theta)}{H(u|\theta)} & x \leq u, \\ (1-\phi_u) + \phi_u \times G(x|u,\sigma_u,\xi) & x > u, \end{cases} \tag{3.10}$$

where ϕ_u is the proportion of the data above the threshold u and $0 < \phi_u < 1$. $H(.|\theta)$ could be parametric distribution function such as gamma, Weibull or normal distribution function and θ is the

© University of Venda

University of Venda
Creating Future Leaders

parameter vector of the bulk distribution. $G(.|u, \sigma_u, \xi)$ denotes the GPD distribution function where u is the threshold, ξ is the shape parameter and σ_u is the scale parameter Hu (2013).

The latter technique in (3.8) explicitly makes it clear that when conditionally modelling the upper tail using the GPD, also requires the proportion of excesses to obtain the unconditional quantities of interest. The maximum likelihood estimator of this tail fraction parameter is of course the usual sample proportion. The former approach in (3.7) is included in (3.8) as a special case when ϕ_u is set to $1 - H(u|\theta)$ as emphasized in Behrens et al., (2004).

The study applies the non-parametric extremal mixture models to fit a fixed threshold on the positive residuals above the time-varying threshold. A kernel density is finally fitted to the bulk model and a Poisson GPD is fitted to the tails of the probability distributions.

MacDonald et al., (2011) and MacDonald et al., (2013) in their modeling framework came out with a non-parametric kernel density estimator based extreme value mixture models that was an extension of the work done by Tancredi et al., (2006).

3.4.1 Kernel GPD Model

MacDonald et al., (2011) constructed a standard kernel density estimator as the bulk model below the threshold with GPD for the tail model. The distribution function of the standard kernel GPD mixture model is as follows:

Bulk Model Tail Fraction Approach,

$$F(x|X, \lambda, u, \sigma_u, \xi, \phi_u) = \begin{cases} H(x|X, \lambda) & x \leq u, \\ (1 - \phi_u) + \phi_u \times G(x|u, \sigma_u, \xi) & x > u, \end{cases} \tag{3.11}$$

where $\phi_u = 1 - H(u|X, \lambda)$

Parameterised Tail Fraction Approach,

$$F(x|X, \lambda, u, \sigma_u, \xi, \phi_u) = \begin{cases} (1 - \phi_u)\frac{H(x|X,\lambda)}{H(u|X,\lambda)} & x \leq u, \\ (1 - \phi_u) + \phi_u \times G(x|u, \sigma_u, \xi) & x > u, \end{cases} \tag{3.12}$$

where $H(.|X, \lambda)$ is the distribution function of the kernel density estimator and $G(.|u, \sigma_u, \xi)$ is the distribution function of the GPD. The traditional kernel density estimator is given by:

© University of Venda

$$h(x;X,\lambda)=\frac{1}{n\lambda}\sum_{i=1}^{n}K(\frac{x-x_i}{\lambda}), \tag{3.13}$$

where $K(.)$ is the kernel function which is a symmetric and unimodal probability density function and lambda is the bandwith parameter. The kernel function usually satisfy the conditions bellow:

$K(x)\geq 0$ and $\int K(x)dx=1$

Two important practical issues arise with MLE for the kernel bandwith (Hu, 2013):

1. Cross-validation likelihood is required for the kernel density estimator bandwith parameter as the normal likelihood decays as emphasized in Habbema et al., (1974) and Duin (1976). The bandwith can be zero even if the cross-validation likelihood is used in cases whereby the data has rounded . In order to address this issue an option to add a small jitter to the data forms part of the fitting inputs, applying the jitter function to overcome the ties.

2. The bandwith is positively biased for heavy tailed population, generating a larger bandwith. This problem can be resolved by cutting the GPD to both the upper and lower tails using the likelihood and fitting function (gkg), since the bias emerges because both the upper and lower order statistics are joined together in the two tails.

3.4.2 Two Tailed Kernel GPD Model

The study adopts a two tailed mixture model of MacDonald et al., (2011), by splicing the standard kernel density estimator with extreme value tail models. This two-tailed extreme value mixture model helps to deal with the inconsistency in the cross-validation likelihood estimation of the bandwith for heavy tailed distributions.

The parameter vector is given by: $\Theta=(X,\lambda,u_1,\sigma_{ul},\xi_l,\phi_{u_l},u_2,\sigma_{u_r},\xi_r,\phi_{ur})$.

The distribution function of two tailed mixture model is given by:

Parameterised Tail Fraction Approach,

$$F(x|\Theta)=\begin{cases}\phi_{u_l}(1-G(-x|-u_l,\sigma_{u_l},\xi_l)) & x<u_l,\\ \phi_{u_l}+(1-\phi_{ul}-\phi_{u_r})\frac{H(x|X,\lambda)-H(u_l|X,\lambda)}{H(u_r|X,\lambda)-H(u_l|X,\lambda)} & u_l\leq x\leq u_r,\\ (1-\phi_{u_r})+\phi_{u_r}G(x|u_r,\sigma_{u_r},\xi_r) & x>u_r,\end{cases} \tag{3.14}$$

© University of Venda

University of Venda
Creating Future Leaders

where ϕ_{ul} and ϕ_{ur} are estimated as the sample proportions less than the threshold τ_l and above the threshold u_r respectively. $G(-x| - \tau_l, \sigma_{ul}, \xi_l)$ is the unconditional GPD function for $x < u_l$ and $G(x|u_r, \sigma_{ur}, \xi_r)$ is the unconditional GPD function for $x > u_r$.

Bulk Model Based Tailed Fraction Approach,

$$F(x|\Theta) = \begin{cases} \phi_{ul}(1 - G(-x| - u_l, \sigma_u, \xi_l)) & x < u_l, \\ H(x|X, \lambda) & u_l \leq x \leq u_r, \\ (1 - \phi_{u_r}) + \phi_{u_r} G(x|u_r, \sigma_{u_r}, \xi_r) & x > u_r, \end{cases} \tag{3.15}$$

where $\phi_{u_l} = H(u_l, X, \lambda)$ and $\phi_{u_r} = 1 - H(u_l, X, \lambda)$.

3.4.3 Boundary Corrected Kernel GPD Model

We also adopt the modelling approach of MacDonald et al., (2013) who combine a boundary corrected kernel density estimator for bulk model sliced at the threshold with a point process representation of the GPD tail model. The boundary correction kernel density estimator tries to reduce the inherent bias of the kernel density estimator.

The distribution function of the boundary corrected mixture model is given by:
Bulk Model Based Tail Fraction Approach,

$$F(x|X, \lambda_{BC}, u, \sigma_u, \xi, \phi_u) = \begin{cases} H_{BC}(x|X, \lambda_{BC}) & x \leq u, \\ (1 - \phi_u) + \phi_u \times G(x|u, \sigma_u, \xi) & x > u, \end{cases} \tag{3.16}$$

Parameterised Tail Fraction Approach,

$$F(x|X, \lambda_{BC}, u, \sigma_u, \xi, \phi_u) = \begin{cases} (1 - \phi_u)\frac{H_{BC}(x|X, \lambda_{BC})}{H_{BC}(u|X, \lambda_{BC})} & x \leq u, \\ (1 - \phi_u) + \phi_u \times G(x|u, \sigma_u, \xi) & x > u, \end{cases} \tag{3.17}$$

where $H_{BC}(x|X, \lambda_{BC})$ is the distribution function of the boundary correction kernel density estimator.

© University of Venda

University of Venda
Creating Future Leaders

3.5 Poisson GPD

The Poisson generalized Pareto distribution process is directly linked to the generalized extreme value distribution for annual maxima. Suppose the random variable x is such that $x > u$, the probability that the annual maximum of the Poisson GPD process is less than x is given by:

$$\begin{aligned} Pr\{maxY_i \leq x\} &= Pr\{N = 0\} + \sum_{n=1}^{\infty} Pr\{N = n, Y_1 \leq x, ...Y_n \leq x\} \\ &= e^{-\lambda} + \sum_{n=1}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} \times \{1 - (1 + \xi \frac{x-u}{\sigma})_+^{-1/\xi}\}^n \\ &= exp\{-\lambda(1 + \xi \frac{x-u}{\sigma})_+^{-1/\xi}\} \end{aligned} \tag{3.18}$$

If we substitute

$$\sigma = \psi + \xi(u - \mu), \lambda = (1 + \xi \frac{u-\mu}{\psi})^{\frac{-1}{\xi}}, \tag{3.19}$$

(3.16) reduces to the GEV form:

$$H(x) = exp\{-(1 + \xi \frac{x-\mu}{\psi})_+^{\frac{-1}{\xi}}\}, \tag{3.20}$$

Thus the GEV and GPD models are entirely consistent with one another above the threshold u, and (3.17) gives an explicit relationship between the two sets of parameters.

The Poisson model is closely linked to the POT model originally developed by hydrologists. In occassions with high levels of serial correlation the observations that exceed the chosen threshold do not occur singly but in groups or cluster, and in that case, the modelling framework is directly applied to the peak values within each cluster Smith (2003).

Another challenge is seasonal dependence, so the study extends the model to account for seasonality. The possible strategies according to Coles (2001), include:

(a) removal of seasonal trend before applying the threshold approach;

(b) applying the Poisson GPD model separately to each season;

(c) expanding the Poisson GPD model to include covaraites,

The above techniques have been comprehensively applied in past research of threshold methods. This study focuses on the application of Poisson GPD, Extremal Mixture models and stationary point process approaches.

We apply the Poisson Generalized Pareto Distribution (Poisson GPD) to show how it can be used to estimate the frequency of occurrences of extreme peak electricity demand per year. Assume that

© University of Venda

University of Venda
Creating Future Leaders

$Y_j, j = 1, 2, ..., n$ are i.i.d random variables and that we are interested in observing $Y_j > u$. We rescale the indices j to points $\frac{i}{n}$. These points are then seen as a point process of rescaled exceedance times on the interval [0, 1]. This process converges weakly to a homogeneous Poisson process with intensity $\theta > 0$, for $n \longrightarrow \infty$ and $(1 - F(u)) \longrightarrow 0$ such that $n(1 - F(u)) \longrightarrow \theta$. From this, it can be noted that the number of exceedances, denoted by n_u in some specified unit time, say a year, follows a Poisson distribution with mean θ and that $Y_1, Y_2, ..., Y_{nu}$ are i.i.d following a Poisson GPD Smith (2003).

Exceedances of time series data above a threshold usually occur in groups or clusters, resulting in dependencies. For energy processes such as electricity in particular, it is rarely the case that the probability of an extreme event is independent of time of the year. To minimize this dependence among the data series, we decluster the exceedances using the declustering method emphasized in Ferro and Segers (2003).

3.6 Modelling stationary dependent series

Since the exceedances occur in groups (clusters), we decluster the exceedances using the Ferro and Sengers (2003) intervals estimation method and fit a stationary point process model to the cluster maxima.

Suppose $\{Y_i\}$ denotes daily peak electricity demand which is a stationary process with univariate marginal distribution function $\mathbb{F}$ which has upper endpoint x^F. We define the extremes of Y_i to be the exceedances of a high threshold u, $u < x^F$. Pickands (1975) emphasized that as u approaches x^F, if the distribution of the excesses, $Y_i - u$, of u scaled as a function of u, converges to a non-degenerate limiting distribution, then that distribution must be the Generalised Pareto distribution (GPD). This motivates the use of the GPD as a statistical model for the exceedances of a high threshold u. The conditional survival function for the exceedances of u under the assumption that excesses follow a $GPD(\sigma, \xi)$ model is, for $y > 0$

$$Pr(Y > y + u | Y > u) = [1 + \frac{\xi y}{\sigma}]_{a_+}^{\frac{-1}{\xi}} \tag{3.21}$$

where $a_+ = max\{0, a\}$, $\sigma > 0$ and ξ are the scale and shape parameters respectively and the rate of exceedance of the threshold is determined by additional parameter of the tail model $\phi_u = Pr(Y > u)$. The theoretical justification of this model requires that ϕ_u is small, since unless F itself is GPD, the approximation to the tail of F the GDP holds only as τ approaches x^F.

Threshold-stability is a very important characteristic of GPD. If the conditional distribution of the

© University of Venda

University of Venda
Creating Future Leaders

exceedances of u is a GPD (σ, ξ), then for any level of υ, $u < \upsilon < x^F$, the conditional distribution of the exceedances of υ is a $GPD(\sigma, \xi)$ distribution, where $\sigma = \sigma + \xi(\upsilon - u)$. This result means that the form of the distribution of the threshold excceedances, including the shape parameter, is invariant to the selection of a higher threshold. In this study, we fit Poisson GPD using all exceedances assuming that the extreme peak electricity data is independent and afterwards account for dependance in the confidence interval evaluation by means of block bootstrap methods Davison and Hall (1993).

Using all the threshold exceedances, under the assumption of independence of extreme events the likelihood function for the stationary model is given by:

$$\mathbb{L}(\sigma, \xi, \phi_u) = \prod_{i=1}^{n} (1 - \phi_u)^{1-\mathbb{I}[y_i > u]} (\phi_u \sigma^{-1} [1 + \frac{\xi(y_i - u)}{\sigma}]_+^{\frac{-1}{\xi - 1}})^{\mathbb{I}[y_i > u]} \tag{3.22}$$

where $\mathbb{I}[y_i > u]$ is the indicator function taking the value of 1 when $y_i > u$ and zero otherwise. We compute the maximum likelihood estimate (MLE) for the rate parameter as $\phi_u = \frac{n_u}{n}$, where n_u is the number of exceedances of the threshold u.

3.6.1 Declustering algorithms and Extremal Index

Extreme values can occur in groups within the context of stationary dependent sequences. The initial stage in making statistical inferences is to identify the clusters in the data, an approach that is commonly known as declustering. Declustering is discussed in this chapter for univariate sequences. A variety of declustering techniques in comtemporary literature that are not satisfactory are well noted and an alternative that does not suffer from the same setback is suggested. The new approach relies on the original idea of the extremal index as a measure of depedence or independence between extreme observations. This also leads to new estimators for the extremal index and the cluster-size distribution that, unlike other estimators, do not require data to be declustered initially.

Since the exceedances occur in groups (clusters), we will decluster the exceedances using the Ferro and Segers (2003) intervals estimation method and fit a stationary point process model to the cluster maxima. We estimate the frequency and intensity of the cluster maxima using the maximum likelihood (ML) method, Ferro and Segers (2003).

© University of Venda

University of Venda
Creating Future Leaders

3.6.2 Univariate sequences

Consider $\{x_i\}_{i=1}^n$ as a sample from a stationary sequence of random variable with F as the marginal distribution function, and $\tilde{F} = 1-F$ as a survival function with upper end-point $\omega = sup\{x : F(x) < 1\}$. We aim to identify within this sample approximately independent clusters of extreme observations Ferro and Segers (2003). Within the context of univariate sequences (electricity data), an extreme value is any observation exceeding a chosen threshold u. For integers $0 \leq k < l$ and $n \geq 1$, put $M_{k,l} = max\{\xi_i := k+1, ..., l\}$ and $M_n = M_{0,n}$. The process $\{x_i\}_{i=1}^n$ is said to have extremal index $\theta \epsilon [0,1]$ if for each $u > 0$ there exists a sequence $\{u_i\}_{i=1}^n$ such that, as $i \longrightarrow \infty$,

(a) $n\tilde{F}(u_i) \longrightarrow u$ and

(b) $P(M_n \leq u_i) \longrightarrow exp(-\theta u)$; see Leadbetter et al., (1983).

If $\theta = 1$, then there is no dependence structure in the sequences of random variables. However, if $\theta < 1$, then there exists a degree of dependence, hence the sequences tend to cluster in the limit. The extremal index happens to be one of the features used for making inferences about the characteristics of such clusters. According to Leadbetter (1983), the extremal index is the reciprocal of the mean cluster size. It is important to identify independent clusters of exceedences above a high threshold, in order to assess for each cluster the main characteristic of interest and also to form estimates from these values.

The two main techniques used in literature to identify clusters defined from different estimators are the block and the runs declustering as emphasized in Leadbetter et al., (1989).

The runs declustering assumes that exceedances belong to the same cluster if they are separated by fewer than a certain number, the run length, of values below the threshold. The main drawback of these estimators as indicated by Hsing (1991) is the selection of the declustering parameters, which is largely arbitrary as the choice of run length usually has a significant inpact on the estimate of the cluster characteristic.

Investigation based on the point process of exceedance times emphasized by Hsing et al., (1988) shows that the asymptotic distribution of the interexceedance times belongs to a one-dimensional parametric family of distributions indexed by the extremal index. Extremal index can be estimated without declustering by simply equating theoretical moments of the limiting distribution to their empirical counterparts. We define an automatic declustering scheme that does not require a subjective choice of auxiliary parameter. Moreover, the declustering scheme supports a bootstrap technique for obtaining the confidence intervals on estimates of cluster characteristics that accounts for the uncertainty in the scheme's estimation Ferro and Segers (2003).

© University of Venda

3.6.3 Interexceedance times

We consider the limiting distribution of the times between exceedances of a threshold u by the process $\{x_i\}_{i=1}^n$. Suppose $T(u)$ is a random variable eqaul in distribution to:

$min\{n \geq 1 : \xi_{n+1} > u\}$ given $\xi_1 > u$,

i.e.

$P\{T(u) = n\} = P(M_{1,n} \leq u, \xi_{n+1} > u | \xi_1 > u)$ for $n \geq 1$,

alternatively,

$P\{T(u) > n\} = P(M_{1,n+1} \leq u | \xi_1 > u)$ for $n \geq 1$

We compute the asymptotic distribution of $T(u)$. The case of independent random variables $\{x_i\}_{i=1}^n$ is a straightforward. Clearly

$P\{T(u) > n\} = F(u)^n$ for $n \geq 1$,

so that, for $x > 0$,

$$P\left\{\tilde{F}(u)T(u) > x\right\} = P\left\{T(u) > \lfloor x/\tilde{F}(u) \rfloor\right\} = exp[\lfloor x/\tilde{F}(u) \rfloor log\{F(u)\}],$$

where $\lfloor x \rfloor$ denotes the integer part of x. If F has no atom at its end point ω then, since $log(1+\varepsilon) \sim \varepsilon$ as $\varepsilon \to 0$

we have

$$lim_{u \uparrow \varepsilon}[P\left\{\tilde{F}(u)T(u) > x\right\}] = exp(-x) \text{ for } x > 0.$$

So $\tilde{F}(u)T(u)$ is asymptotically standard exponentially distributed which is consistent with the outcome of Hsing et al., (1988) that the point process of exceedance times has a Poisson process limit.

We consider a general situation where the extremal index $\theta \epsilon [0, 1]$. The corresponding point process limit for the exceedance time is compound Poisson according to Hsing et al., (1988). This indicates that the limit distribution of the interexceedance times will be a mixture of an exponential distribution and a point mass on zero. Ferro and Segers (2003), emphasized that the extemal index plays a double role as θ is both the proportion of non-zero interexceedance times and also the reciprocal of the mean of the non-zero interexceedance times.

© University of Venda

3.6.4 Estimation of the extremal index

We describe an estimator for the extremal index that is based on the limit result of Ferro and Segers (2003).

Suppose that we have a sample of $\xi_1, ..., \xi_n$ and a high threshold u.

Let $N = N_n(u) = \sum_{i=1}^{n} I(\xi_i > u)$ be the number of observations exceeding u, and

let $1 \leq S_1 < ... < S_N \leq n$ be the exceedance times. The observed interexceedance times are $T_i = S_{i+1} - S_i$ for $i = 1, ..., N-1$.

Intervals estimator

The second moment of T_θ is $E(T_\theta^2) = \frac{2}{\theta}$, which we estimate to obtain a first estimator for θ

Let $\tilde{F}(u)$ be an estimator for $F(u)$. Then

$\tilde{\theta}_n(u) = \frac{2(N-1)}{\tilde{F}_n(u)^2 \sum_{i=1}^{N-1} T_i^2}$

For example , if $\tilde{F}_n(u) = \frac{N}{n}$, then we have

$\tilde{\theta}_n(u) = \frac{2n^2(N-1)}{N^2 \sum_{i=1}^{N-1} T_i^2}$

The first moment of T_θ is 1 so θ is related to the coefficient of variation, ν, of the interexceedance times by

$1 + \nu^2 = \frac{E(T_\theta^2)}{E(T_\theta^2)} = 2\theta^{-1}$.

The interexceedance times are overdispersed with clustering in the limit if and only if $\theta < 1$. This relationship motivates another estimator for θ,

$\tilde{\theta}_n(u) = \frac{2(\sum_{i=1}^{N-1} T_i)^2}{(N-1) \sum_{i=1}^{N-1} T_i^2}$

where we do not need to estimate $\tilde{F}(u)$.

A further improvement is possible if we consider a penultimate approximation to the limiting mixture distribution $(1-\theta)\varepsilon_0 + \theta_{\mu\theta}$ by setting $r_n = n$ in the proof of theorem 1 by Ferro and Segers (2003), then we see that the distribution of the exceedance times satisfies

$P\{T(u_n) > n\} = \theta F(u_n)^{n\theta} + o(1)$

An estimator for θ can be derived based on this relationsship. Let T denote a random variable on the positive integers whose distribution is given by

© University of Venda

$$P(T > n) = \theta p^{n\theta}; n \geq 1, \tag{3.23}$$

where $\theta \epsilon (0, 1]$ and $p \epsilon (0, 1)$ may be considered as $F(u_n)$. Note that

$\frac{2E(T)^2}{E(T^2)} = \frac{2\left\{1-(1-\theta)p^\theta\right\}^2}{2\theta p^\theta + \theta p^\theta (1-p^\theta) + (1-p^\theta)^2}$

$\frac{2E(T)^2}{E(T^2)} = \theta + \theta(2 - \frac{3\theta}{2})(1-p) + O\left\{(1-p)^2\right\}$ as $p \longrightarrow 1$,

which follows from

$E(T-1) = \sum_{n=1}^{\infty} P(T > n) = \theta p^\theta (1-p^\theta)^{-1}$,

$E\left\{\frac{T(T-1)}{2}\right\} = \sum_{n=1}^{\infty} nP(T > n) = \theta p^\theta (1-p^\theta)^{-2}$.

Therefore, the first-order bias of $\tilde{\theta}_n(u)$ at a threshold u is $\theta(2 - \frac{3\theta}{2})\tilde{F}(u)$

In contrast, the relationship

$$\frac{2\left\{E(T-1)\right\}^2}{E\left\{(T-1)(T-2)\right\}} = \theta \tag{3.24}$$

motivates the estimator

$\tilde{\theta}_n^*(u) = \frac{2\left\{\sum_{i=1}^{N-1}(T_i-1)\right\}^2}{(N-1)\sum_{i=1}^{N-1}(T_i-1)(T_i-2)}$,

where u is a sufficiently high threshold and T_i represent the interexceedance times. The extremal index, $\tilde{\theta}_n^*(u)$ measures the amount of declustering and $0 \leq \tilde{\theta}_n^*(u) \leq 1$, where $\frac{1}{\tilde{\theta}_n^*(u)}$ is the limiting mean cluster size. The first order bias of the estimator is 0. Compared with $\tilde{\theta}_n(u)$, the estimators for the first and second moments from which $\tilde{\theta}_n^*(u)$ is obtained have been shrunk towards 0. The smallest observed interexceedance times are greater than 0, while the limiting distribution of $(1-\theta)\varepsilon_0 + \theta_{\mu\theta}$ models them as zero. The estimator $\tilde{\theta}_n^*(u)$ makes sure that the role played by the smallest interexceedance times are indeed 0; however the larger interexceedance times are relatively unaffected.

It is also noted that $\tilde{\theta}_n^*(u)$ can assume values that are greater than 1 and is invalid if the largest interexceedance time is no greater than 2. In order to avoid these situation, we define

$$\tilde{\theta}_n(u) = \begin{cases} 1 \wedge \tilde{\theta}_n(u) & \text{if } \max\{T_i : 1 \leq i \leq N-1\} \leq 2, \\ 1 \wedge \tilde{\theta}_n^*(u) & \text{if } \max\{T_i : 1 \leq i \leq N-1\} < 2, \end{cases} \tag{3.25}$$

© University of Venda

University of Venda
Creating Future Leaders

which we term as the intervals estimator for the extremal index (Ferro and Segers , 2003).

Return level estimation

Extreme quantiles are estimated using the k-observation return level given by

$$y_{k,t} = u + \frac{\sigma}{\xi}[(k\phi_u)^{\xi} - 1] \tag{3.26}$$

The return period is in days, because the data is daily peak electricity demand and $y_{k,t}$ denotes the maximum value of y we expect to observe in k observations, with ϕ_u denoting the probability of exceeding the threshold u. By expressing σ in terms of other parameters in (3.22) we have

$$\sigma = \frac{(y_{k,t} - u)\xi}{(k\phi_u)^{\xi} - 1} \tag{3.27}$$

which is therefore used to calculate the confidence intervals of $y_{k,t}$.

3.6.5 Declustering and bootstrapping

Even though we have demonstrated how to estimate the extremal index without resorting to declustering, clusters still need to be identified in order to estimate other cluster characteristics. We explain how the limiting distribution of $(1-\theta)\varepsilon_0 + \theta_{\mu\theta}$ may be used to identify clusters without making an arbitrary choice. The limit theory also supports a bootstrap procedure for assessing uncertainties associated with parameter estimates. Because the limiting process of exceedance times is a compound Poisson process, we group interexceedance times into two types, independent intercluster times (between clusters) and independent sets of intracluster times (within clusters). The extremal index is the proportion of interexceedance times that may be regarded as intercluster times.

Suppose we observe N exceedance times, $S_1 < ... < S_N$, with interexceedance times $T_i = S_{i+1} - S_i$ for $i = 1, ..., N-1$. We can assume that the largest $C - 1 = \llcorner \theta N \lrcorner$ interexceedance times are approximately independent intercluster times that divide the remainder into approximately independent sets of intracluster times.

Precisely, if $T_{(c)}$ is the C^{th} largest interexceedance time and T_{ij} is the j^{th} interexceedance time to exceedance $T_{(c)}$, then $\{T_{ij}\}_{j=1}^{C-1}$ is a set of approximately independent intercluster times. In the case of ties, decrease C until $T_{(C-1)} > T_{(c)}$.

© University of Venda

University of Venda
Creating Future Leaders

Let $T_j = \{T_{ij-1+1}, ..., T_{ij-1}\}$, where $i_0 = 0$, $i_C = N$ and $T_j = \varnothing$ if $i_j = i_{j-1+1}$. Then $\{T_j\}_{j=1}^{C}$ is a collection of approximately independent sets of intracluster times. Moreover, each set T_j has associated with it a set of threshold exceedances $C_j = \{\xi_k : k \epsilon S_j\}$, where $S_j = \{S_{ij-1+1,...,S_{ij}}\}$.

This interpretation justifies a decomposition of the observed process into C clusters, where the jth cluster comprises the exceedances C_j. This is the same as runs declustering with run length $T_{(C)}$. We define C by replacing θ with an estimator $\tilde{\theta}$. Any estimator for θ may be used, however if we employ (3.21), then we have an entirely automatic declustering procedure justified by asymptotic theory.

3.7 A point process characterization of extremes

This statistical approach was introduced by Smith (1989), although the fundamental probability theory upon which it derives had already existed in literature. The modelling framework of Leadbetter (1983), illustrated that more light can be shared on point-process using the viewpoints of EVT. Under this modelling framework, the times at which high-threshold exceedances occur and the excess values over the chosen threshold are unified into one process based on a two-dimensional plot of exceedance times and values instead of considering the two events as separate processes. Here, the asymptotic theory of threshold exeedances shows that under suitable normalisation, this process behaves like a non-homogeneous Poisson process. As explained in Karr (1991) and emphasized by Resnick (2013), a point process is a stochastic model of points that are randomly scattered in some space, where the points may denote times of phenomena or location of objects that are characterized by a stochastic system. According to Coles (2001) and Smith (1989), a point process technique serves as a benchmark for unifying and extending EVT modelling based on both blocks and threshold methods in relation to high-level exceedances.

Let $\mathbf{Y} = \{Y_1, ..., Y_n\}$ be a sequence of daily peak electricity demand (DPED) with one observed covariate x such that $\mathbf{X} = \{X_1, ..., X_n\}$. Of these n observations, let n_u be the number of exceedances above a sufficiently high predetermined threshold, u. In this study, a time varying covariate will be used. Then the non-stationary likelihood function of the point process model over the region $[0, 1] \times (u, \infty)$ is given by:

© University of Venda

University of Venda
Creating Future Leaders

$$L(u,\boldsymbol{\theta};\mathbf{Y},\mathbf{X}) = \exp\left\{-\frac{1}{n}\sum_{i=1}^{n}\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}\times$$

$$\prod_{i:Y_i>u}\frac{1}{\sigma}\left[1+\xi\left(\frac{y_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}-1} \tag{3.28}$$

for $\xi \neq 0, \boldsymbol{\theta} = (\mu(x), \sigma(x), \xi(x))$ are constants. If the covariate x denotes the time varying effect, denoted by t then the model is written as follows:

The non-stationary models which are time varying are given below:

$$\mu(t) = \mu_0 + \mu_1 t$$
$$\log\sigma(t) = \sigma_0 + \sigma_1 t$$
$$\xi(t) = \xi_0 + \xi_1 t$$

The first component of the likelihood function represents the probability of getting n_u exceedances.

The stationary models which are time invariant are also given below:

$$\mu(t) = \mu$$
$$\sigma(t) = \sigma$$
$$\xi(t) = \xi$$

In this dissertation, we emphasized on stationary series. The daily peak electricity demand data has achieved statistical equilibrium (stationarity) through differencing and linear detrending.

3.8 A point process limit for extremes

If the marginal distribution function, F is followed by a series of random samples $Y_i, Y_2, ..., Y_n$, and if $M_n = max\{Y_i, Y_2, ..., Y_n\}$ is distributed with the GEVD as in (3.6) for the normalizing constants $\{a_n > 0\}$ and b_n, then the sequential point process T_n on R^2 is given by:

$$T_n = \{(\frac{i}{(n+1)}, \frac{(X_i - b_n)}{a_n}\}, i = 1, .., n \tag{3.29}$$

such that an axis of time passes through the closed interval $(0, 1)$; and the second point ensures stability in the occurrence of extremes as $n \to \infty$ such that on $[0, 1] \times [u, \infty]$, $K_n \to K$ as $n \to \infty$, where K

© University of Venda

University of Venda
Creating Future Leaders

is a heterogeneous Poisson process as emphasized in Karr (1991), Beichelt (2006) and Resnick (2013). For a higher threshold, u and for a given space of the form $A = [0,1] \times (u, \infty)$, all the values of T_n possess a p chance of happening within A, where

$$p = P\{\frac{(X_i - b_n)}{a_n} > u\} \approx \frac{1}{n}[1 + \xi(\frac{u-\mu}{\sigma})]^{-1/\xi} \tag{3.30}$$

As the binomial mass approaches the limiting Poisson distribution, then as $n \to \infty$, $T_n(A)$ obeys Poi$(\Lambda(A))$ such that for all spaces that satisfy $A = [t_1, t_2] \times (u, \infty)$, with $[t_1, t_2] \subset [0,1]$, the limiting distribution of $T_n(A)$ is also Poi$(\Lambda(A))$, where

$$\Lambda(A) = (t_2 - t_1[1 + \xi(\frac{u-\mu}{\sigma})]^{-1/\xi}$$

occurs as a homogeneous result of the process in the direction of time as emphasized in Karr (1991) and Beichelt (2006). Within the modelling framework of EVT, the Poisson process contains all the features of the ordinary GEVD model, the GEVD model for r largest ordered statistics, therefore the Poisson point process has become a robust alternative characterization for all the EVT models, according to Coles (2001). For the threshold models with

$$\Lambda(A_z) = \Lambda_1([t_1, t_2]) \times \Lambda_2([z, \infty))$$

valid for

$$\Lambda_1([t_1, t_2]) = (t_2 - t_1) \quad \text{and} \Lambda_2([z, \infty)) = [1 + \xi(\frac{x-\mu}{\sigma})]^{-1/\xi}$$

such that

$$\begin{aligned} P\{\frac{(X_i - b_n)}{a_n} > z \mid \frac{(X_i - b_n)}{a_n} > u\} &= \frac{\Lambda_2[z, \infty)}{\Lambda_2[u, \infty)} \\ &= [1 + \xi(\frac{z-u}{\tilde{\sigma}})]^{-1/\xi} \end{aligned} \tag{3.31}$$

with

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

© University of Venda

4. Chapter 4

4.1 Data Analysis

4.1.1 Introduction

4.2 Exploratory data analysis

Here we provide exploratory data analysis which includes the summary statistics, time series plots, density plots, normal quantile-quantile plots (QQ-plots), and Box plots.

Table 4.1 shows the five number summary statistics, the mean, the kurtosis and the skewness of the daily peak electricity data.

Minimum	Q_1	Q_2	Mean	Q_3	Maximum	Kurtosis	Skewness
17600	25660	29090	28580	31420	36660	2.295807	-0.2485498

Table 4.1: The five number summary statistics, Mean, Kurtosis and the Skewness of the DPED

The DPED data from the table has a mean value $\tilde{X} < Q_2$, indicating that the data is negatively skewed and the negative value of the skewness confirms it. The positive value of the kurtosis indicates heavy tails and peakness relative to the normal distribution.

Figure 4.1 represents the graphical distribution of the time series plot of the DPED.

© University of Venda

University of Venda
Creating Future Leaders

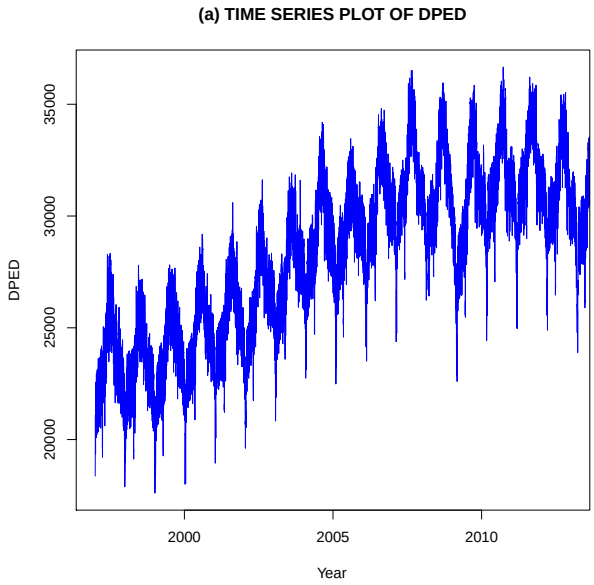

Figure 4.1: Time series plot of the DPED

Figure 4.1 is a time series plot of the daily peak electricity demand data with a time varying threshold which is a penalized cubic smoothing spline. Based on the generalised cross validation (GCV) criterion, a smoothing parameter $\lambda = 0.097$ was selected. An initial threshold is set at zero after fitting the time varying threshold and only the positive excesses above zero are considered. We fit a non-parametric extremal mixture model to the observations considered to determine a sufficiently high threshold and the exceedances are then declustered using the Ferro and Segers (2003) intervals estimator technique. A visual inspection of the time series plot in Figure 4.1 shows that the DPED is non-stationary and the plot indicates that data demonstrates high seasonality and a deterministic upward trend.

Figure 4.2 represents the density plot, histogram, box plot and the normal Q-Q plot of the DPED.

© University of Venda

University of Venda
Creating Future Leaders

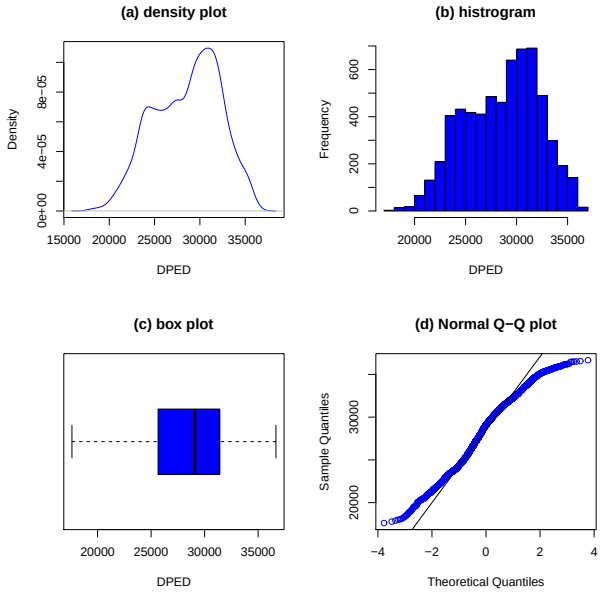

Figure 4.2: Density plot, histogram, box and QQ plots of the DPED

4.2.1 The density plot

The density plot confirms that the DPED data does not conform to the normal distribution curve as the visual inspection of the plot indicates three turning points on the graph, thereby deviating from the Guassian distribution.

4.2.2 Histogram

The output of the histogram also shows that the DPED data deviates from the normal distribution as the plot does not conform to the bell shape of the normal distribution.

4.2.3 The boxplot

The box-and-whisker plot shows that the DPED data is not symmetric as the spread of the data to the left of the median is not exactly or approximately equal to the spread of the data to the right of the median. The boxplot confirms that the DPED data does not follow the normal distribution.

© University of Venda

4.2.4 The QQ plot

The visual inspection of the qq plot indicates that when the DPED data was plotted against the ideal theoretical normal distribution, the output shows that all the points do not form a straight line, meaning that the data is not normally distributed. The output of the QQ line shows that the data has fat tails as some of the data points deviated from the QQ line.

The maximum likelihood estimates of the GEVD fit indicated that the location parameter, $\mu = 26925.7794$. The effect of the location parameter is to translate the graph relative to the normal distribution. It simply shifts the graph left or right on the horizontal axis. The scale parameter, $\sigma = 4206.6018$, has the effect of stretching out the graph. The shape parameter, $\xi = -0.4183864$, indicating that the DPED data conforms to the Weilbull distribution since $\xi < 0$.

© University of Venda

University of Venda
Creating Future Leaders

4.3 Generalized Pareto distribution frequency analysis

4.3.1 All non-linear detrended data

The DPED is initially detrended using a penalized regression cubic smoothing spline given in (3.4). The kernel density technique was used to estimate the value of the threshold, $u = 1489$ and all the positive residuals above the threshold were extracted for both yearly and monthly frequency analysis.

Figure 4.3 is the threshold estimation using the non-parametric extremal mixture model where a kernel density is fitted to the bulk model and Poisson GPD fitted to the tail of the distribution of all the non-linear detrended data with estimated threshold, $u = 1489$.

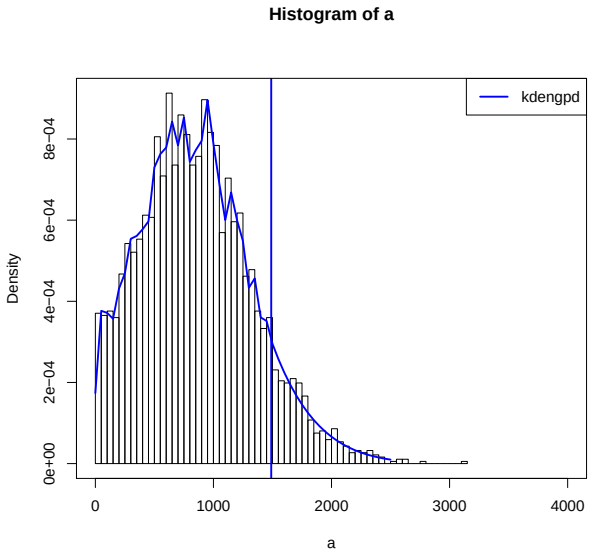

Figure 4.3: The threshold estimation using the non-parametric extremal mixture model where a kernel density is fitted to the bulk model and Poisson GPD fitted to the tail of the distribution of all the non-linear detrended data with estimated threshold, $u = 1489$.

© University of Venda

The Tables 4.2 and 4.3 show the distribution of the yearly frequency analysis of the threshold excesses of all the non-linear detrended data with number of exceedances 369.

Year	1997	1998	1999	2000	2001	2002	2003	2004
Frequency	16	12	11	14	17	22	29	30

Table 4.2: Yearly frequency analysis of all the non-linear detrended data above $u = 1489$.

Year	2005	2006	2007	2008	2009	2010	2011	2012	2013
Frequency	22	32	20	25	28	27	22	25	17

Table 4.3: Yearly frequency analysis of all the non-linear detreded data above $u = 1489$.

Figure 4.4 shows the histogram representing the yearly frequency analysis of all the non-linear detrended data above threshold, $u = 1489$

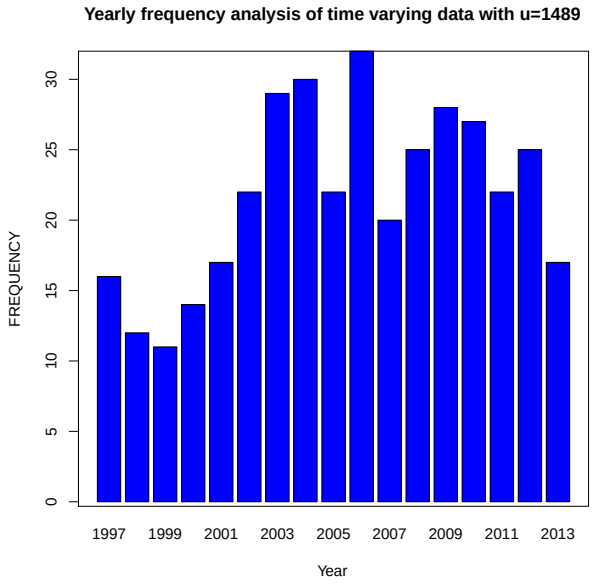

Figure 4.4: Histogram representing the yearly frequency analysis of all the non-linear detrended data above threshold, $u = 1489$.

© University of Venda

University of Venda
Creating Future Leaders

The Tables 4.4 and 4.5 show the distribution of the monthly frequency analysis of the threshold excesses of all the non-linear detrended data with number of exceedances 369.

Month	January	February	March	April	May	June
Frequency	56	0	18	51	44	30

Table 4.4: Monthly frequency analysis of all the non-linear detrended data above threshold, $u = 1489$.

Month	July	August	September	October	November	December
Frequency	27	41	27	16	4	55

Table 4.5: Monthly frequency analysis of all the non-linear detrended data above threshold, $u = 1489$.

Figure 4.5 shows the histogram representing the monthly frequency analysis of all the non-linear detrended data above threshold, $u = 1489$.

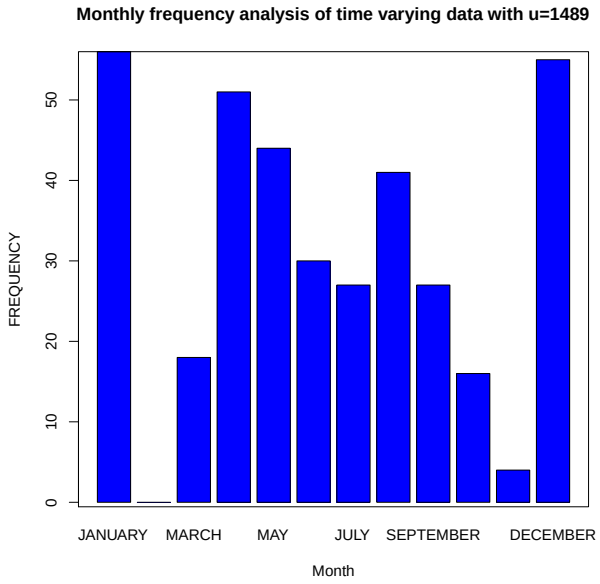

Figure 4.5: Histogram representing the monthly frequency analysis of all the non-linear detrended data above threshold, $u = 1489$.

© University of Venda

4.3.2 Non-linear detrended winter data

The linearly detrended data was also separated into winter and non-winter seasons. The threshold was estimated using the kernel density estimation and all the positive residuals above the threshold $u = 1358$ were extracted for both yearly and monthly frequency analysis.

Figure 4.6 is the threshold estimation using the non-parametric extremal mixture model where a kernel density is fitted to the bulk model and Poisson GPD fitted to the tail of the distribution of all the non-linear detrended winter data with estimated threshold, $u = 1358$.

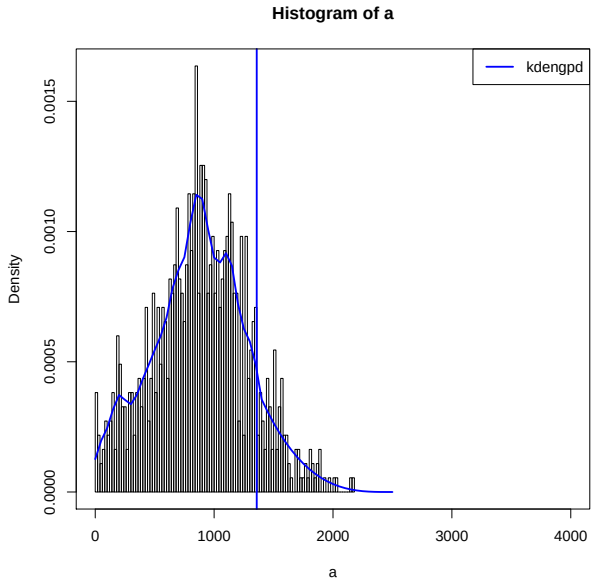

Figure 4.6: The threshold estimation using the non-parametric extremal mixture model where a kernel density is fitted to the bulk model and Poisson GPD fitted to the tail of the distribution of all the non-linear detrended winter data with estimated threshold, $u = 1358$.

© University of Venda

University of Venda
Creating Future Leaders

The Tables 4.6 and 4.7 show the distribution of the yearly frequency analysis of the threshold excesses of the detrended winter data with number of exceedances 102.

Year	1997	1998	1999	2000	2001	2002	2003	2004
Frequency	7	8	10	3	3	4	6	9

Table 4.6: Yearly frequency analysis of the time varying winter data above $u = 1358$.

Year	2005	2006	2007	2008	2009	2010	2011	2012	2013
Frequency	6	8	5	6	9	6	8	4	0

Table 4.7: Yearly frequency analysis of the time varying winter data above $u = 1358$.

Figure 4.7 shows the histogram representing the yearly frequency analysis of the time varying winter data above threshold, $u = 1358$.

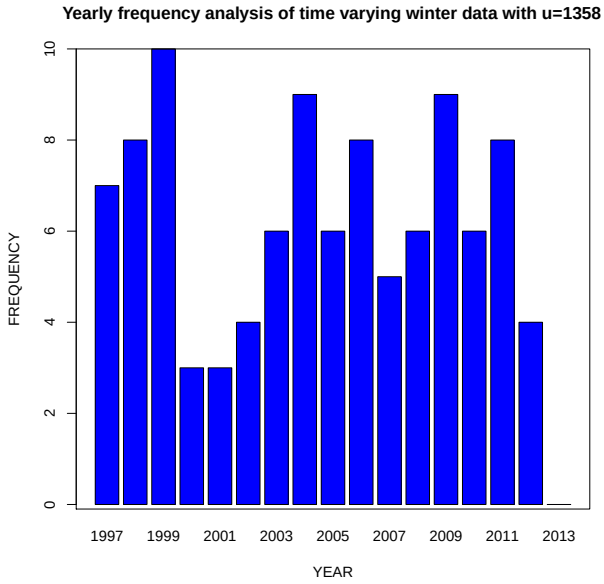

Figure 4.7: Histogram representing the yearly frequency analysis of the time varying winter data above threshold, $u = 1358$.

Table 4.8 shows the distribution of the monthly frequency analysis of the threshold excesses of the detrended winter data with number of exceedances 102.

© University of Venda

Month	June	July	August
Frequency	33	13	56

Table 4.8: Monthly frequency analysis of the time varying winter data above $u = 1358$.

Figure 4.8 shows the histogram representing the monthly frequency analysis of all the time varying winter data above threshold, $u = 1358$

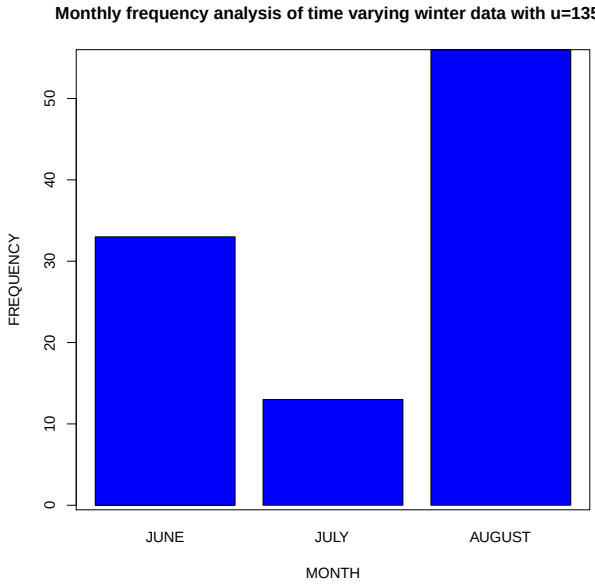

Figure 4.8: Histogram representing the monthly frequency analysis of all the time varying winter data above threshold, $u = 1358$.

© University of Venda

University of Venda
Creating Future Leaders

4.4 Fitting Poisson GPD to stationary dependent series

We decluster the data using Ferro and Segers (2003) automatic declustering to remove dependency by using an R package, "texmex". The declustering approach is applied to the non-linear detrended data.

4.4.1 Non-linear detrended data

As mentioned earlier, the declustering algorithm is extended to all the non-linear detrended data. The aim is to compare the values of the extremal index for both the differenced data and the time varying data. The dataset with higher extremal index indicates an independence process, hence we procceed to perform inference based on Poisson GPD.

4.4.2 All data

We decluster the sequences of all the time varying data using the Ferro and Segers (2003) interval method, with run length of 2 , 180 identified clusters and threshold, $u = 1489$. The length of the original series was 3711 with threshold exceedances of 380. The interval estimator of the extremal index $\theta = 0.476096$, indicating that the data is not independent as the value is closer to 0 than 1 and the positive residuals were extracted for analysis. The data is initially detrended using a penalized regression cubic smoothing spline given in the equation below:

$$\eta(t) = \sum_{i}^{n}(y_i - f(t_i))^2 + \lambda \int (f^{''}(t))^2 dt + u \tag{4.1}$$

where y_i denotes the daily peak electricity demand and λ is a smoothing parameter. We use the extremal mixture models to determine the value of u and observations above the time varying threshold $\eta(t)$ are extracted. We estimate u using extremal mixture models. Stoffer (2011) as well as Wang (2011) emphasized the use of penalized cubic smoothing splines as a time-varying threshold which has an attractive features such as deseasonalizing and detrending at one stroke. We therefore incorporate penalized cubic smoothing splines in this study.

The Figure 4.9 shows the time series plot of the time varying data for both positive and negative residuals.

© University of Venda

University of Venda
Creating Future Leaders

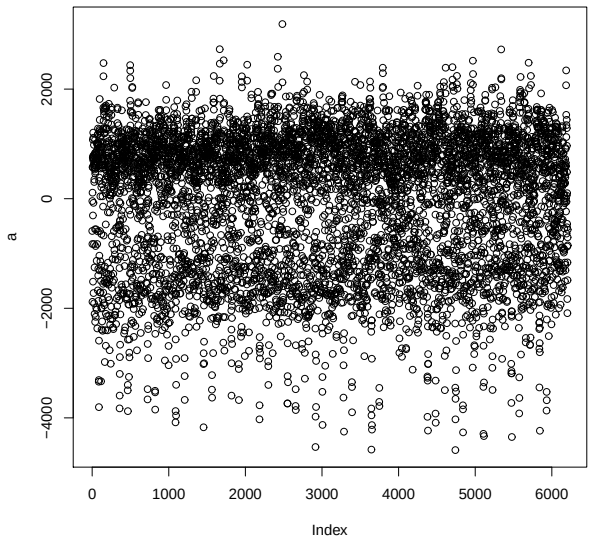

Figure 4.9: The time series plot of the time varying data

Figure 4.10 shows a time series plot of all the positive residuals extracted.

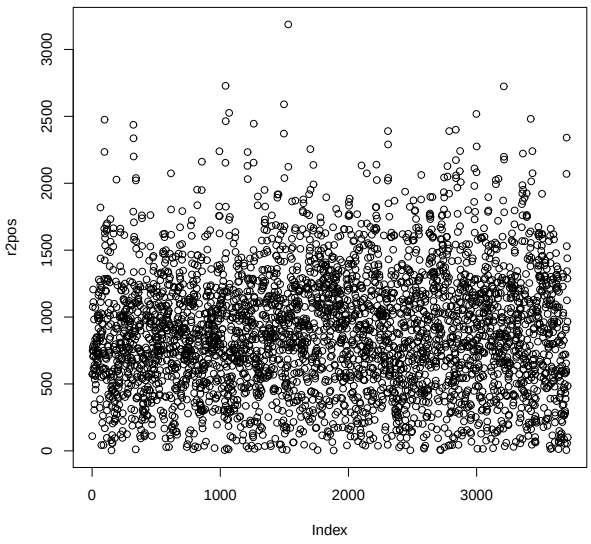

Figure 4.10: The time series plot of the positive residuals extracted

Figure 4.11 shows the time series plot of the positive residuals extracted after the declustering algorithm. The horizontal line indicates the threshold $u = 1489$. There exist a dependence among the time varying

© University of Venda

University of Venda
Creating Future Leaders

dataset as the interval estimator of the extremal index 0.476096 is closer to 0 than 1.

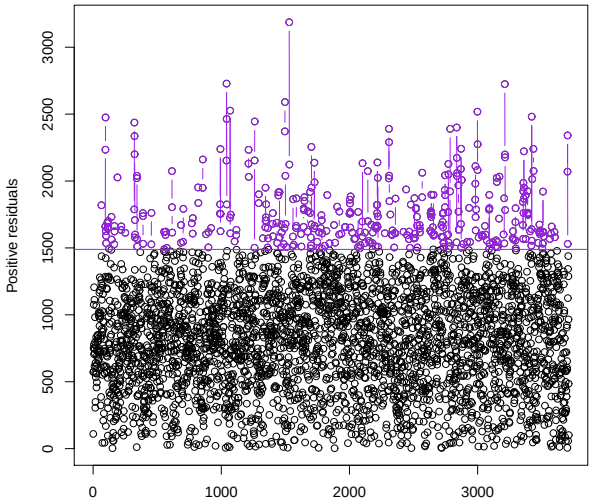

Figure 4.11: The time series plot of the positive residuals extracted after the declustering algorithm

We choose arbitrary thresholds of $u = 1490$, $u = 1495$ and $u = 1499$ and plot the interexceedance times against the standard exponential quantiles. Visual inspection of the graphs shows no significant difference in the plots as compared with the chosen threshold of $u = 1489$.

Figure 4.12 shows the plot of the interexceedance times against the standard exponential quantiles.

© University of Venda

University of Venda
Creating Future Leaders

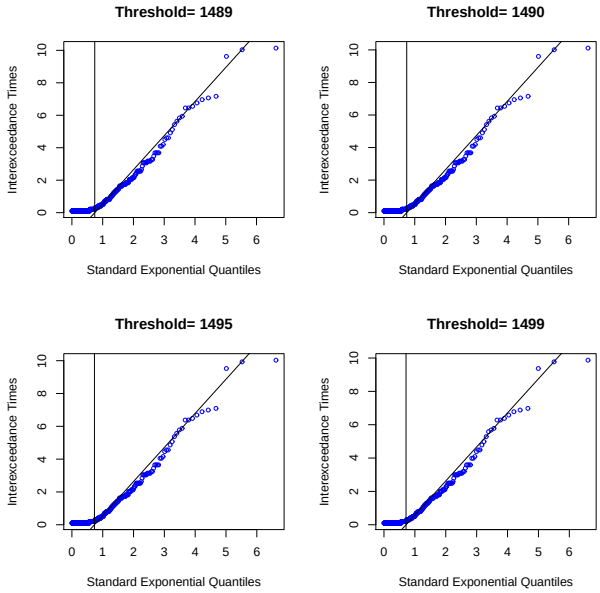

Figure 4.12: The plot of the interexceedance times against the standard exponential quantiles of the time varying data.

The summary of fit of the Generalized Pareto distribution to the cluster maxima is displayed in Table 4.9.

Threshold	Rate of exc	Log. Lik	AIC	σ	ξ
1489	0.0485	-1212.086	2428.172	5.79899 (0.10430)	-0.06476 (0.07319)

Table 4.9: Model fit by maximum likelihood

The coefficients are σ and ξ with their standard errors in parenthesis.
To ensure that the scale parameter is positive ($\sigma > 0$), we use the transformation $\sigma = e^{5.79899}$.

4.4.3 Formal goodness-of-fit test for all the non-linear detrended data

The formal goodness of fit tests based on Anderson-Darling test statistic A^2 and Crammer-Von Mises statistic W^2 were performed.
The hypothesis is formulated as follows: H_0: Data follow Poisson GPD(329.9661258; -0.06476)
H_1: Data do not follow Poisson GPD(329.9661258; -0.06476). The R package "eva" is used to fit

© University of Venda

Poisson GPD to cluster maxima of all the time varying data. Several simulations were performed based on the estimates from the summary fit to the cluster maxima.

The output of the Anderson-Darling test statistic A^2 is displayed in Table 4.10.

Test statistic	P-value	σ	ξ
0.3609138	0.6058275	331.52927177	-0.06072497

Table 4.10: Anderson-Darling goodness-of-fit test

The summary of fit based on Cramer-Von Mises statistic W^2 is shown in Table 4.11.

Test statistic	P-value	σ	ξ
0.04559185	0.6578228	331.52927177	-0.06072497

Table 4.11: Cramer-Von Mises goodness-of-fit test

Decision rule; reject H_0 if the $p-value < \alpha$, $\alpha = 0.05$. Since the p-values are greater than $\alpha = 0.05$, we fail to reject H_0 and conclude that, all the time varying data follow Poisson GPD(329.9661258; -0.06476). Similar tests were performed using thresholds $u = 1490$, $u = 1495$, $u = 1499$, but the chosen threshold $u = 1489$ is the best.

© University of Venda

The Generalized Pareto distribution fitted to the cluster maxima is shown in Figure 4.13 and the diagnostic plots below show a fairly good fit.

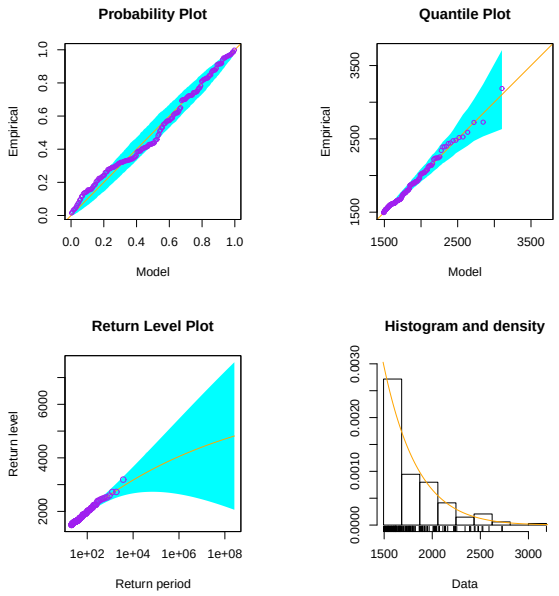

Figure 4.13: The diagnostic plots of the Poisson GPD model fitted to cluster maxima

In Table 4.12, we compare the parameter estimates of the Poisson GPD model fitted to the cluster maxima with those obtained by fitting the GPD to the original series.

Threshold	Rate of exc	Log. Lik	AIC	σ	ξ
1489	0.1024	-2510.115	5024.230	5.68109 (0.06900)	-0.07541 (0.04630)

Table 4.12: Model fit by maximum likelihood

The coefficients are σ and ξ with their standard errors in parenthesis.

© University of Venda

The Table 4.13 gives the summary (Parameter Bootstrap) of the Poisson GPD parameter uncertainties.

	σ	ξ
Original	5.79899038	-0.06475881
Bootstrap mean	5.81604203	-0.08572873
Bias	0.01705165	-0.02096993
SD	0.10738901	0.07756977
Bootstrap median	5.81602930	-0.08617193
Correlation		
	σ	ξ
σ	1.0000000	-0.7475342
ξ	-0.7475342	1.0000000

Table 4.13: Poisson GPD parameter uncertainties : Parameteric Bootstrap

Figure 4.14 is the plot of the Bootstrap density which gives us the empirical distribution from the plots.

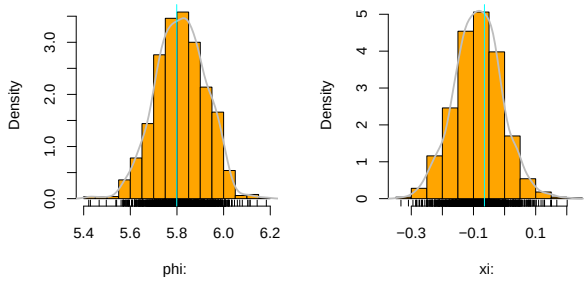

Figure 4.14: The Bootstrap plot of the time varying data

© University of Venda

Figure 4.15 gives the return level plot which shows the predicted return levels with 95% predictive intervals from 20 to 100 years at 10 years interval.

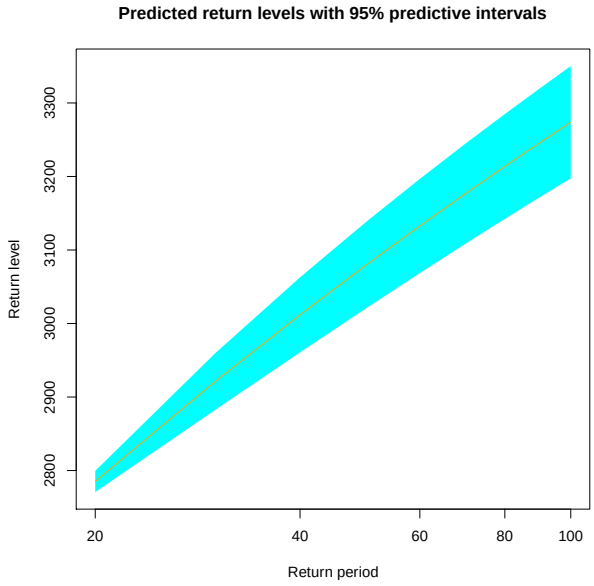

Figure 4.15: Return level estimation of the time varying data

Table 4.14 shows the predictive intervals for the return levels (RL):

Year	LCL (2.5%)	RL	UCL (97.5%)
20	2770.7	2785.1	2799.5
30	2882.4	2920.9	2959.3
40	2960.7	3011.7	3062.8
50	3020.5	3079.3	3138.1
60	3068.6	3132.7	3196.7
70	3108.4	3176.5	3244.6
80	3142.3	3213.6	3284.9
90	3171.7	3245.7	3319.7
100	3197.4	3273.8	3350.2

Table 4.14: 95% predictive interval for return levels

Inference from the table indicates that, we are 95% confident that, a peak electricity demand of 2785.1

© University of Venda

University of Venda
Creating Future Leaders

would be reached with a lower confident limit of 2770.69 and an upper confident limit of 2799.544 in 20 years time and so on.

4.4.4 Winter data

The time varying data is also divided into winter, summer, spring and autumn data. The estimated threshold, $u = 1358$, with the length of the original series is 919, the number of threshold exceedances being 167 and the interval estimator of extremal index being equal to 0.5770155 indicating an independent process as the value of the extremal index is closer to 1 than 0, run length of 1 with 87 identified clusters. Figure 4.16 shows the time series plot of the time varying winter data.

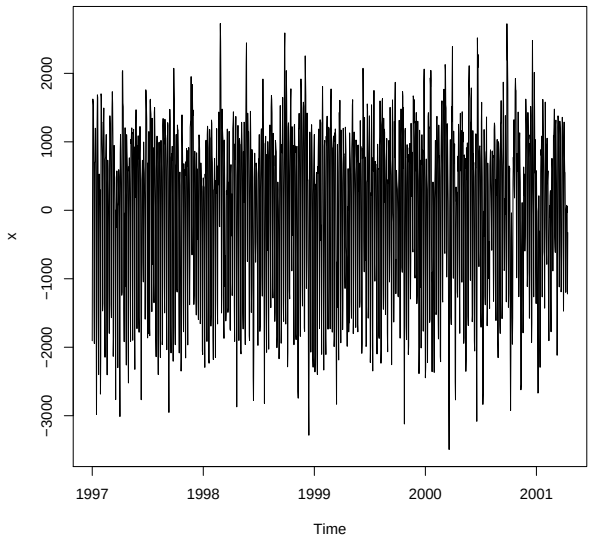

Figure 4.16: The time series plot of all the time varing winter data

© University of Venda

University of Venda
Creating Future Leaders

Figure 4.17 shows the time series plot of the positive residuals extracted after the declustering algorithm. The horizontal line indicates the threshold $u = 1358$. There exists some independence among the time varying winter dataset as the value of the extremal index 0.5770155 is closer to 1 than 0. The length of original series 919, declustering using the intervals method with run length 1 and 87 identified clusters.

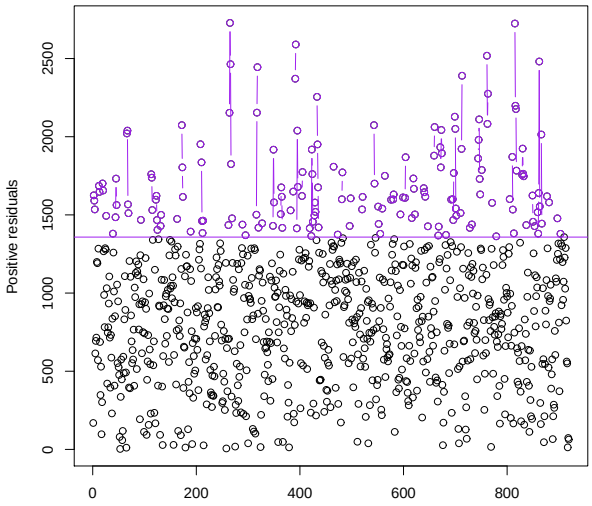

Figure 4.17: The plot of the positive residuals extracted from the declustered winter data.

We choose arbitrary thresholds of $u = 1360$, $u = 1365$ and $u = 1367$ and plot the interexceedance times against the standard exponential quantiles. Visual inspection of the graphs shows no significant difference in the plots as compared with the chosen threshold of $u = 1358$.

Figure 4.18 indicates the plot of the interexceedance times against the standard exponential quantiles.

© University of Venda

University of Venda
Creating Future Leaders

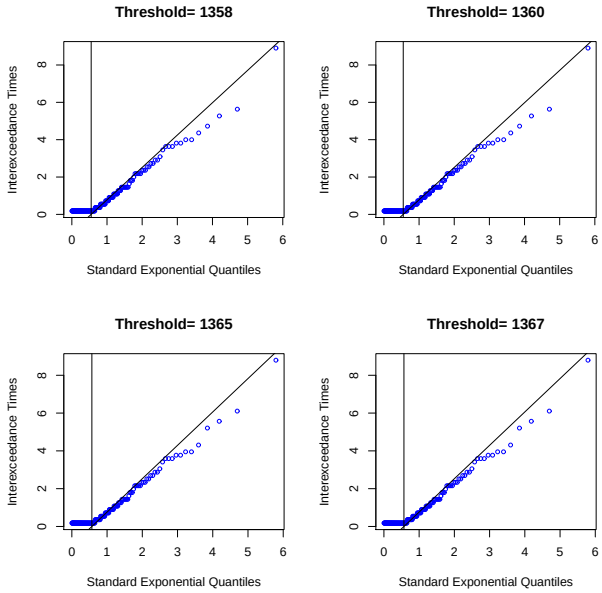

Figure 4.18: The plot of the interexceedance times against the standard exponential quantiles

The summary of fit of the Generalized Pareto distribution to the cluster maxima is displayed in Table 4.15.

Threshold	Rate of exc	Log. Lik	AIC	σ	ξ
1358	0.09467	-598.5341	1201.0683	6.0024 (0.1603)	-0.1227 (0.1199)

Table 4.15: Model fit by maximum likelihood

The coefficients are σ and ξ with their standard errors in parenthesis. To ensure that the scale parameter is positive ($\sigma > 0$), we use the transformation $\sigma = e^{6.0024}$.

4.4.5 Formal goodness-of-fit test for the non-linear detrended winter data

The formal goodness of fit tests based on Anderson-Darling test statistic A^2 and Crammer-Von Mises statistic W^2 were performed.

The hypothesis is formulated as follows: H_0: Data follow Poisson GPD(404.3981854; -0.1227)

H_1: Data do not follow Poisson GPD(404.3981854; -0.1227). The R package "eva" is used to fit Poisson GPD to cluster maxima of the time varying winter data. Several simulations were performed based on the estimates from the summary fit to the cluster maxima.

© University of Venda

The output of the Anderson-Darling test statistic A^2 is shown in Table 4.16.

Test statistic	P-value	σ	ξ
0.373295	0.5926508	407.5072279	-0.1214147

Table 4.16: Anderson-Darling goodness-of-fit test

The summary of fit based on Cramer-Von Mises statistic W^2 is displayed in Table 4.17.

Test statistic	P-value	σ	ξ
0.0834202	0.633723	407.5072279	-0.1214147

Table 4.17: Cramer-Von Mises goodness-of-fit test

Decision rule; reject H_0 if the $p-value < \alpha$, $\alpha = 0.05$. Since the p-values are greater than $\alpha = 0.05$, we fail to reject H_0 and conclude that, the time varying winter data follow Poisson GPD(404.3981854; -0.1227). Similar tests were performed using thresholds $u = 1360$, $u = 1365$, $u = 1367$, but the chosen threshold $u = 1358$ is the best.

© University of Venda

We fit a Generalized Pareto distribution to the cluster maxima and the diagnostic plots in Figure 4.19 shows a fairly good fit.

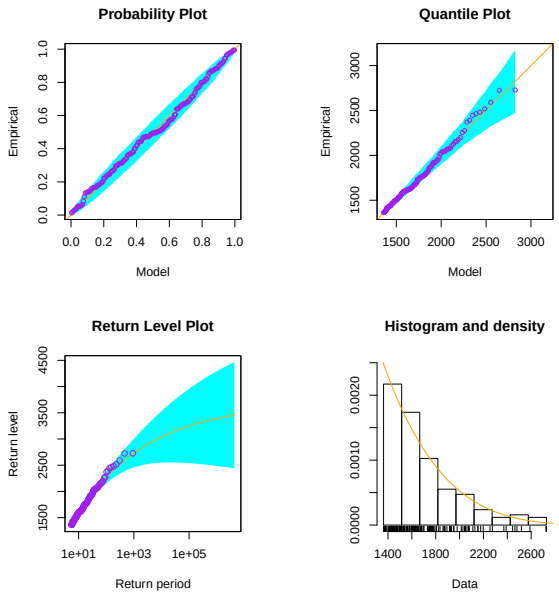

Figure 4.19: The diagnostic plots of the GPD model fitted to cluster maxima

We compare the parameter estimates of the Poisson GPD model fitted to the cluster maxima with those obtained by fitting the GPD to the original series in Table 4.18.

Threshold	Rate of exc	Log. Lik	AIC	σ	ξ
1358	0.1817	-1138.302	2280.603	5.98807 (0.10631)	-0.17177 (0.07412)

Table 4.18: Model fit by maximum likelihood

The coefficients are σ and ξ with their standard errors in parenthesis.

© University of Venda

Table 4.19 gives the summary (Parameter Bootstrap) of the GPD parameter uncertainties.

	σ	ξ
Original	6.0024063	-0.12270551
Bootstrap mean	6.02723337	-0.15931270
Bias	0.02482708	-0.03660719
SD	0.15795097	0.11093025
Bootstrap median	6.03063711	-0.16057734
Correlation		
	σ	ξ
σ	1.0000000	-0.7961523
ξ	-0.7961523	1.0000000

Table 4.19: GPD parameter uncertainties : Parameteric Bootstrap

The Bootstrap density which gives us the empirical distribution from the plots is shown in Figure 4.20.

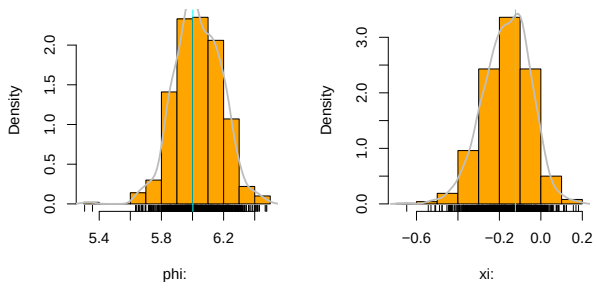

Figure 4.20: The Bootstrap plot of the time varying winter data

The return level plot which shows the predicted return levels with 95% predictive intervals from 20 to 100 years at 10 years interval is given in Figure 4.21.

© University of Venda

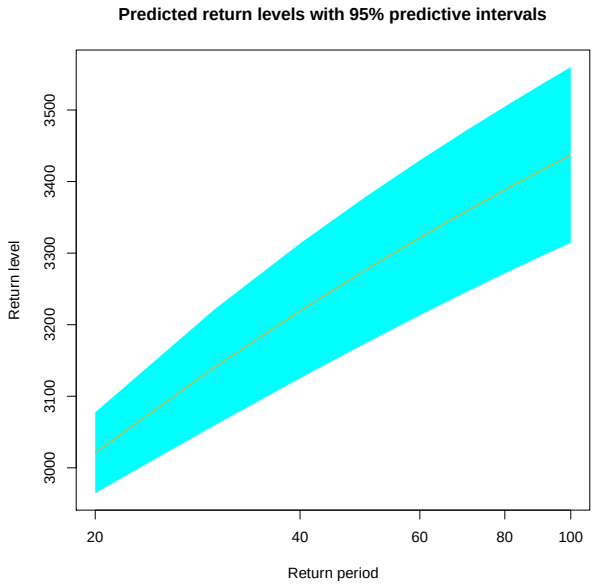

Figure 4.21: Return level estimation of the time varying winter data

Table 4.20 gives the predictive intervals for the return levels:

Year	LCL (2.5%)	RL	UCL (97.5%)
20	1542.327	1606.299	1670.271
30	1664.005	1754.206	1844.407
40	1750.364	1854.774	1959.183
50	1816.324	1930.371	2044.419
60	1868.977	1990.621	2112.264
70	1912.286	2040.52	2168.754
80	1948.699	2082.988	2217.277
90	1979.837	2119.874	2259.912
100	2152.422	2006.829	2298.014

Table 4.20: 95% predictive interval for return levels

Inference from the table indicates that, we are 95% confident that, a peak electricity demand of 1606.299 would be reached with a lower confident limit of 1542.327 and an upper confident limit of 1670.271 in 20 years time and so on.

© University of Venda

University of Venda
Creating Future Leaders

4.5 GPD frequency analysis of the non-linear detrended data

We split the time varying data into the four seasons according to the calender dates in the Southern Hermisphere as Summer, Autumn, Winter and Spring. The yearly and monthly frequency analysis are performed for the summer, spring and autumn data sets as the frequency analysis for the winter data has already been considered.

4.5.1 Summer data

We define the summer data according to the calender dates in the Southern Hemisphere spanning from the period 1 December to 28/29 February. Figure 4.22 below shows the time series plot of the summer data. Visual inspection indicates the data exhibit strong seasonality and stationarity, hence, the summer data is in statistical equilibrium.

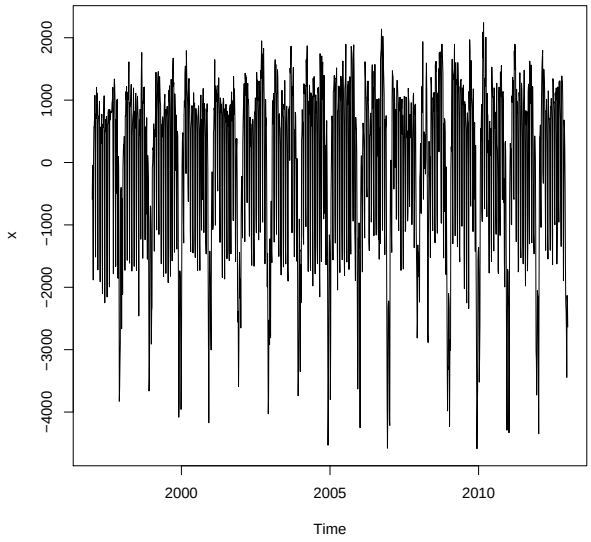

Figure 4.22: Time series plot of the time varying summer data

We fit a kernel density plot to the data using the R package "evmix" to determine the threshold for the summer data. Figure 4.23 indicates the kernel density plot with the vertical line indicating the estimated threshold.

© University of Venda

University of Venda
Creating Future Leaders

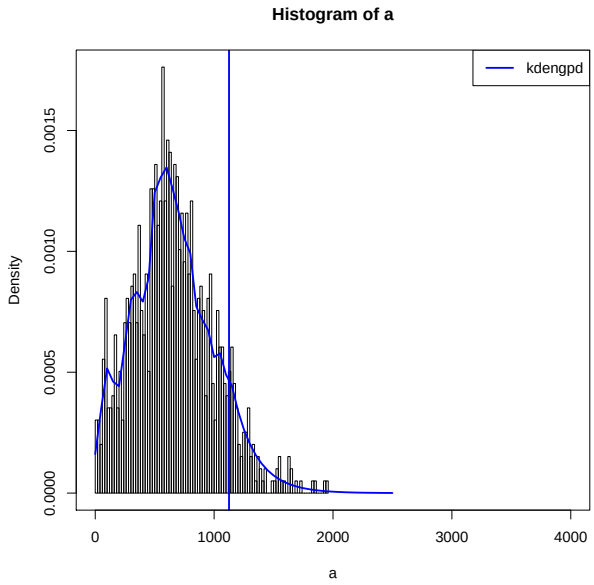

Figure 4.23: kernel density fitted to the time varying summer data, $u = 1125$

4.5.2 Fitting Poisson GPD to declustered non-linear detrended summer data

The data is declustered using Ferro and Segers (2003) interval method to remove dependency using the R package "texmex". We arbitrarily select thresholds $u = 1127$, $u = 1130$, $u = 1132$ and plot the interexceedance times against the standard exponenetial quantiles.
Visual inspection of the graphs in Figure 4.24 show no significant difference in the plots as compared with the selected threshold $u = 1125$.

The summary of fit of the Generalized Pareto distribution to the cluster maxima is displayed in Table 4.21.

Threshold	Rate of exc	Log. Lik	AIC	σ	ξ
1125	0.08177	-509.8857	1023.7714	6.27956 (0.14232)	-0.38938 (0.09372)

Table 4.21: Model fit by maximum likelihood

The coefficients are σ and ξ with their standard errors in parenthesis. To ensure that the scale parameter is positive ($\sigma > 0$), we use the transformation $\sigma = e^{6.27956}$.

© University of Venda

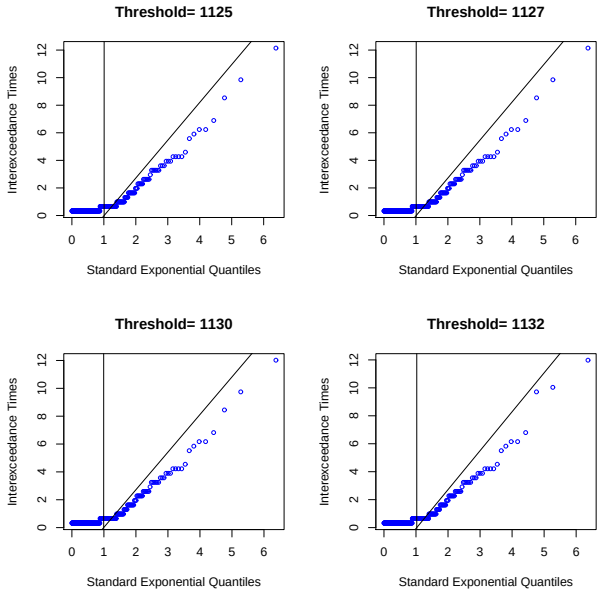

Figure 4.24: The plot of the interexceedance times against the standard exponential quantiles

4.5.3 Formal goodness-of-fit test for the non-linear detrended summer data

The formal goodness of fit tests based on Anderson-Darling test statistic A^2 and Crammer-Von Mises statistic W^2 were performed.

The hypothesis is formulated as follows: H_0: Data follow Poisson GPD(533.55384; -0.38938)

H_1: Data do not follow Poisson GPD(533.55384; -0.38938). The R package "eva" is used to fit Poisson GPD to cluster maxima of the time varying summer data. Several simulations were performed based on the estimates from the summary fit to the cluster maxima.

The summary of fit based on the Anderson-Darling test statistic A^2 is displayed in Table 4.22.

Test statistic	P-value	σ	ξ
0.4390814	0.5340341	547.0086503	-0.3986819

Table 4.22: Anderson-Darling goodness-of-fit test

© University of Venda

The summary of fit based on Cramer-Von Mises statistic W^2 is shown in Table 4.23.

Test statistic	P-value	σ	ξ
0.06172397	0.5412139	542.0086503	-0.3986819

Table 4.23: Cramer-Von Mises goodness-of-fit test

Decision rule; reject H_0 if the $p-value < \alpha$, $\alpha = 0.05$. Since the p-values are greater than $\alpha = 0.05$, we fail to reject H_0 and conclude that, the time varying summer data follow Poisson GPD(533.55384;-0.38938).

Similar tests were performed using thresholds $u = 1127$, $u = 1130$, $u = 1132$, but the chosen threshold $u = 1125$ is the best.

The Tables 4.24 and 4.25 show the distribution of the yearly frequency analysis of the threshold excesses of the time varying summer data with number of exceedances 97.

Year	1997	1998	1999	2000	2001	2002	2003	2004
Frequency	8	3	2	4	1	3	4	11

Table 4.24: Yearly frequency analysis of the time varying summer data above $u = 1125$

Year	2005	2006	2007	2008	2009	2010	2011	2012	2013
Frequency	13	7	3	7	10	7	8	4	2

Table 4.25: Yearly frequency analysis of the time varying summer data above $u = 1125$

© University of Venda

Figure 4.25 shows the histogram representing the yearly frequency analysis of the time varying summer data above threshold, $u = 1125$

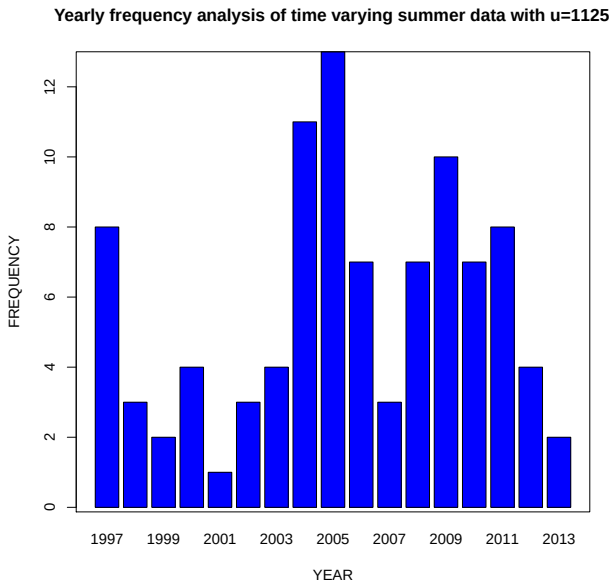

Figure 4.25: Histogram representing the yearly frequency analysis of the time varying summer data above threshold, $u = 1125$

Table 4.26 shows the distribution of the monthly frequency analysis of the threshold excesses of the time varying summer data with number of exceedances 97.

Month	January	February	December
Frequency	31	10	56

Table 4.26: Monthly frequency analysis of the time varying summer data above $u = 1125$

© University of Venda

Figure 4.26 is the histogram representing the monthly frequency analysis of the time varying summer data above threshold, $u = 1125$

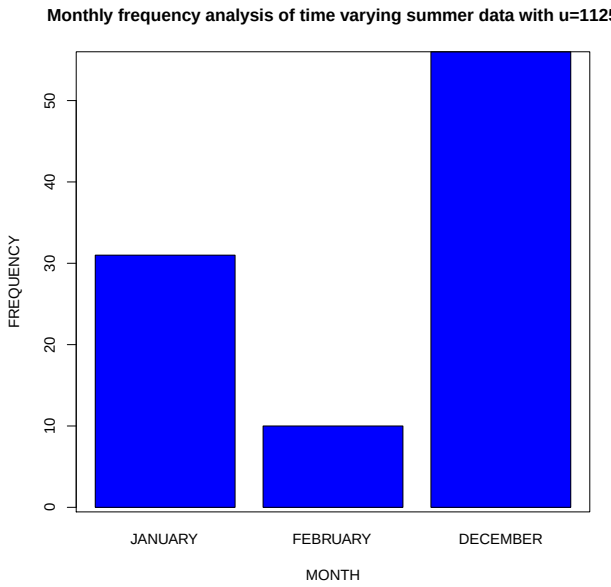

Figure 4.26: Histogram representing the monthly frequency analysis of the time varying summer data above threshold, $u = 1125$

4.5.4 Spring data

We define the spring data according to the calender dates in the Southern Hemisphere as the period spanning from 1 September to 30 November.

Figure 4.27 gives the time series plot of the spring data. Visual inspection indicates the data exhibit strong seasonality and stationarity, hence, the spring data is in statistical equilibrium.

© University of Venda

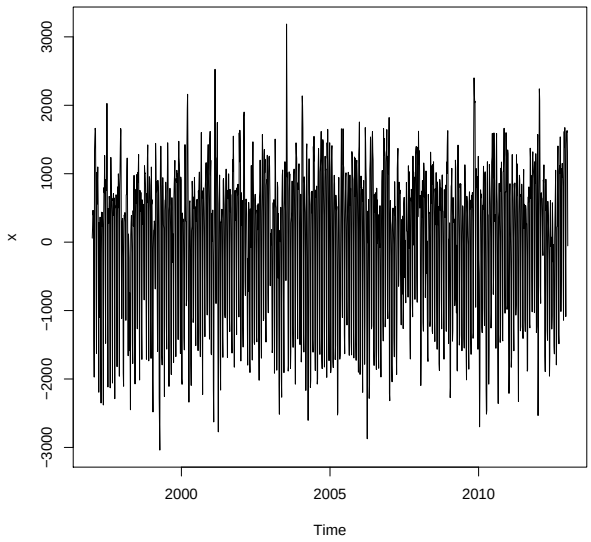

Figure 4.27: Time series plot of the time varying spring data

We fit a kernel density plot to the data using the R package "evmix" to determine the threshold for the spring data. The Figure 4.28 illustrates the kernel density plot with the vertical line indicating the estimated threshold.

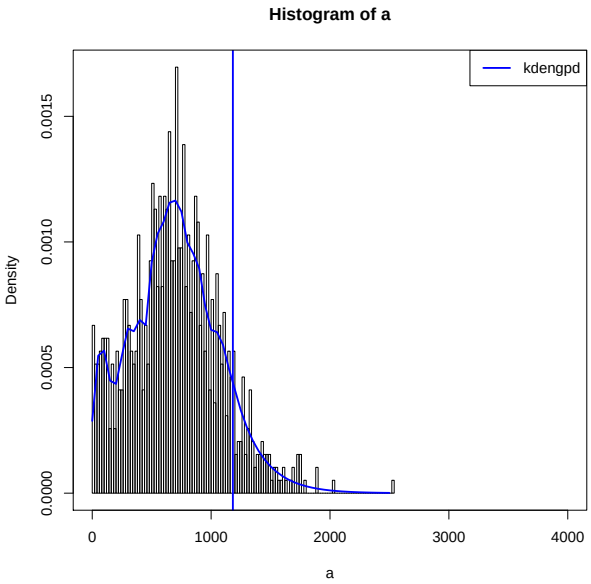

Figure 4.28: kernel density fitted to the time varying spring data, $u = 1183$

.

© University of Venda

4.5.5 Fitting Poisson GPD to declustered spring data data

The data is declustered using Ferro and Segers (2003) interval method to remove dependency using the R package "texmex". We arbitrarily select thresholds $u = 1185$, $u = 1187$, $u = 1189$ and plot the interexceedance times against the standard exponenetial quantiles.

Visual inspection of the graphs in Figure 4.29 show no significant difference in the plots as compared with the selected threshold $u = 1183$.

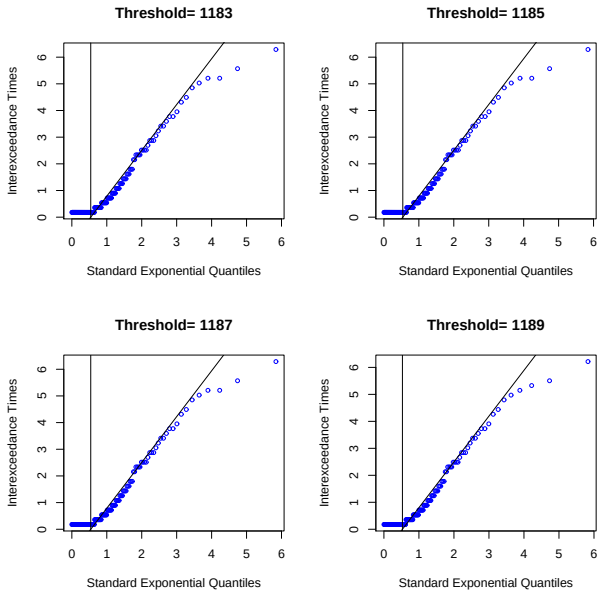

Figure 4.29: The plot of the interexceedance times against the standard exponential quantiles

The summary of fit of the Generalised Pareto distribution to the cluster maxima is displayed in Table 4.27.

Threshold	Rate of exc	Log. Lik	AIC	σ	ξ
1183	0.0945	-621.2989	1246.5977	5.87911 (0.13232)	-0.05186 (0.08093)

Table 4.27: Model fit by maximum likelihood

The coefficients are σ and ξ with their standard errors in parenthesis. To ensure that the scale parameter is positive ($\sigma > 0$), we use the transformation $\sigma = e^{5.87911}$.

© University of Venda

4.5.6 Formal goodness-of-fit test for the non-linear detrended spring data

The formal goodness of fit tests based on Anderson-Darling test statistic A^2 and Crammer-Von Mises statistic W^2 were performed.

The hypothesis is formulated as follows: H_0: Data follow Poisson GPD(357.4909332; -0.05186)

H_1: Data do not follow Poisson GPD(357.4909332; -0.05186). The R package "eva" is used to fit Poisson GPD to cluster maxima of the time varying spring data. Several simulations were performed based on the estimates from the summary fit to the cluster maxima.

The summary of fit based on the Anderson-Darling test statistic A^2 is displayed in Table 4.28.

Test statistic	P-value	σ	ξ
0.3721167	0.5989145	461.003856	-0.142979

Table 4.28: Anderson-Darling goodness-of-fit test

The summary of fit based on Cramer-Von Mises statistic W^2 is shown in Table 4.29.

Test statistic	P-value	σ	ξ
0.04807184	0.6411697	461.003856	-0.142979

Table 4.29: Cramer-Von Mises goodness-of-fit test

Decision rule; reject H_0 if the $p-value < \alpha$, $\alpha = 0.05$. Since the p-values are greater than $\alpha = 0.05$, we fail to reject H_0 and conclude that,the time varying spring data follow Poisson GPD(357.4909332; -0.05186).

Similar tests were performed using thresholds $u = 1185$, $u = 1187$, $u = 1189$, but the chosen threshold $u = 1183$ is the best.

The Tables 4.30 and 4.31 show the distribution of the yearly frequency analysis of the threshold excesses of the time varying spring data with number of exceedances 92.

Year	1997	1998	1999	2000	2001	2002	2003	2004
Frequency	4	1	5	6	6	4	6	15

Table 4.30: Yearly frequency analysis of the time varying spring data above $u = 1183$

© University of Venda

Year	2005	2006	2007	2008	2009	2010	2011	2012	2013
Frequency	8	8	1	2	8	8	4	5	1

Table 4.31: Yearly frequency analysis of the time varying spring data above $u = 1183$

Figure 4.30 is the histogram representing the yearly frequency analysis of the time varying spring data above threshold, $u = 1183$

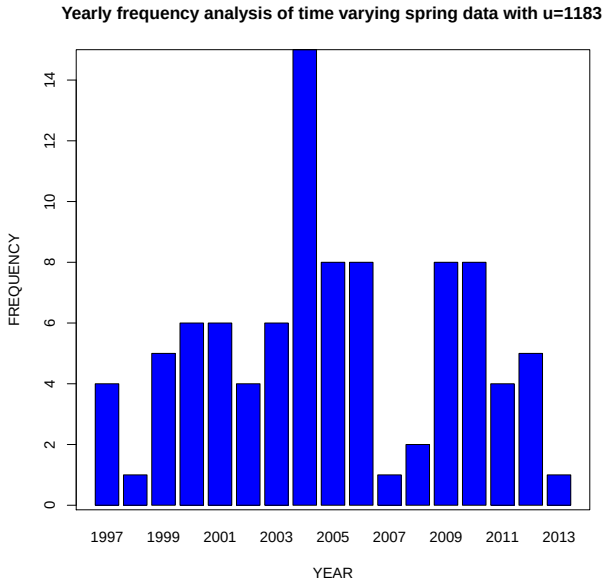

Figure 4.30: Histogram representing the yearly frequency analysis of the time varying spring data above threshold, $u = 1183$

The Table 4.32 gives the distribution of the monthly frequency analysis of the threshold excesses of the time varying spring data with number of exceedances 92.

Month	September	October	November
Frequency	44	20	28

Table 4.32: Monthly frequency analysis of the time varying summer data above $u = 1183$

Figure 4.31 is the histogram representing the monthly frequency analysis of the time varying spring data above threshold, $u = 1183$

© University of Venda

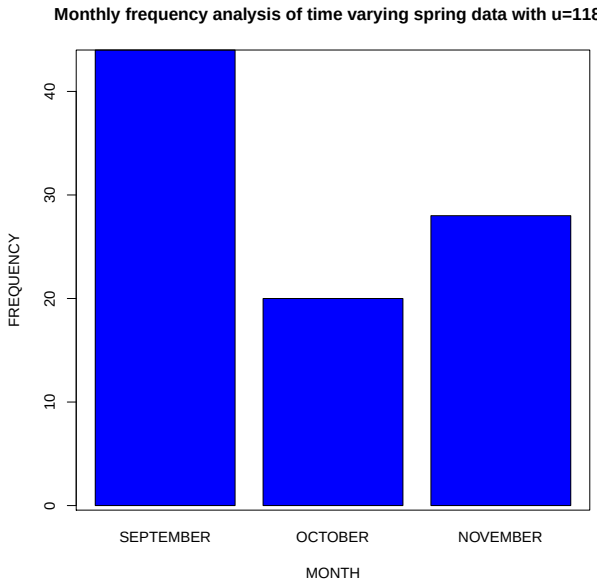

Figure 4.31: Histogram representing the monthly frequency analysis of the time varying spring data above threshold, $u = 1183$

4.5.7 Autumn data

We define the autumn data according to the calender dates in the Southern Hemisphere as the period spanning from 1 March to 31 May.

Figure 4.32 gives the time series plot of the autumn data. Visual inspection indicates the data exhibit strong seasonality and stationarity. Therefore the autumn data is in statistical equilibrium.

© University of Venda

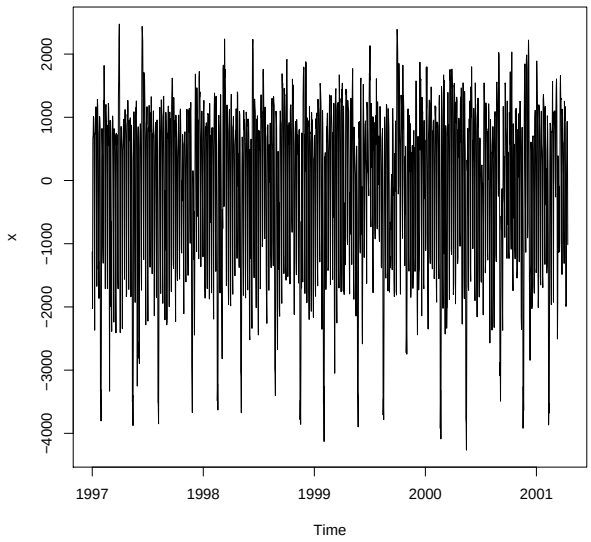

Figure 4.32: Time series plot of the time varying autumn data

We fit a kernel density plot to the data using the R package "evmix" to determine the threshold for the autumn data in Figure 4.33. The figure below shows the kernel density plot with the vertical line indicating the estimated threshold.

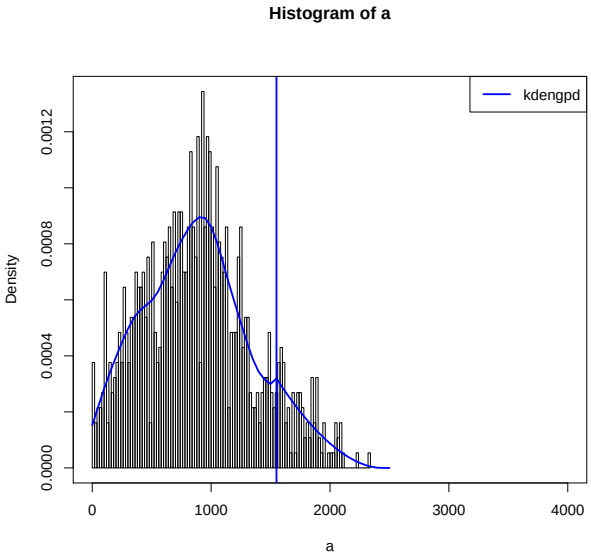

Figure 4.33: kernel density fitted to the time varying autumn data, with $u = 1550$

© University of Venda

University of Venda
Creating Future Leaders

4.5.8 Fitting Poisson GPD to declustered non-linear detrended autumn data

The data is declustered using Ferro and Segers (2003) interval method to remove dependency using the R package "texmex". We arbitrarily select thresholds $u = 1552$, $u = 1554$, $u = 1556$ and plot the interexceedance times against the standard exponenetial quantiles.

Visual inspection of the graphs in Figure 4.34 show no significant difference in the plots as compared with the selected threshold $u = 1550$.

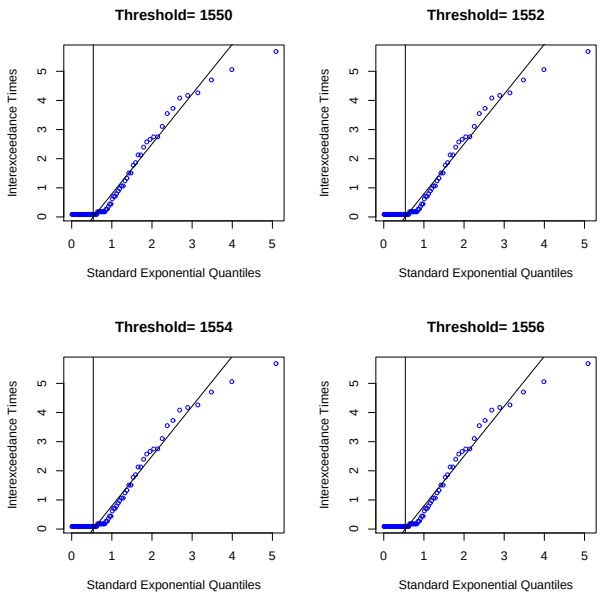

Figure 4.34: The plot of the interexceedance times against the standard exponential quantiles

The summary of fit of the Generalized Pareto distribution to the cluster maxima is displayed in Table 4.33.

Threshold	Rate of exc	Log. Lik	AIC	σ	ξ
1550	0.04762	-292.3795	588.7590	5.9388 (0.2159)	-0.2937 (0.1608)

Table 4.33: Model fit by maximum likelihood

The coefficients are σ and ξ with their standard errors in parenthesis. To ensure that the scale parameter is positive ($\sigma > 0$), we use the transformation $\sigma = e^{5.9388}$.

© University of Venda

4.5.9 Formal goodness-of-fit test for the autumn data

The formal goodness of fit tests based on Anderson-Darling test statistic A^2 and Crammer-Von Mises statistic W^2 were performed.

The hypothesis is formulated as follows: H_0: Data follow Poisson GPD(379.4792811; -0.2937)

H_1: Data do not follow Poisson GPD(379.4792811; -0.2937)

The R package "eva" is used to fit Poisson GPD to cluster maxima of the time varying autumn data. Several simulations were performed based on the estimates from the summary fit to the cluster maxima. The output based on the Anderson-Darling test statistic A^2 is displayed in Table 4.34.

Test statistic	P-value	σ	ξ
0.4054504	0.5705943	383.8005694	-0.2983685

Table 4.34: Anderson-Darling goodness-of-fit test

The summary of fit based on Cramer-Von Mises statistic W^2 is displayed in Table 4.35.

Test statistic	P-value	σ	ξ
0.05446445	0.5979021	383.8005694	-0.2983685

Table 4.35: Cramer-Von Mises goodness-of-fit test

Decision rule; reject H_0 if the $p-value < \alpha$, $\alpha = 0.05$. Since the p-values are greater than $\alpha = 0.05$, we fail to reject H_0 and conclude that,the time varying autumn data follow Poisson GPD(379.4792811; -0.2937).

Similar tests were performed using thresholds $u = 1552$, $u = 1554$, $u = 1556$, but the chosen threshold $u = 1550$ is the best.

The Tables 4.36 and 4.37 display the distribution of the yearly frequency analysis of the threshold excesses of the time varying autumn data with number of exceedances 93.

Year	1997	1998	1999	2000	2001	2002	2003	2004
Frequency	8	6	4	3	6	5	5	4

Table 4.36: Yearly frequency analysis of the time varying autumn data above $u = 1550$

© University of Venda

Year	2005	2006	2007	2008	2009	2010	2011	2012	2013
Frequency	6	4	6	7	8	5	4	8	4

Table 4.37: Yearly frequency analysis of the time varying autumn data above $u = 1550$

Figure 4.35 is the histogram representing the yearly frequency analysis of the time varying autumn data above threshold, $u = 1550$

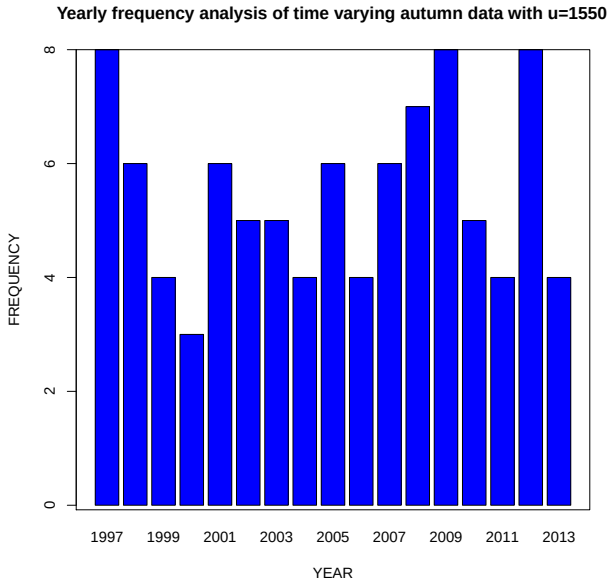

Figure 4.35: Histogram representing the yearly frequency analysis of the time varying autumn data above threshold, $u = 1550$

The Table 4.38 below shows the distribution of the monthly frequency analysis of the threshold excesses of the time varying autumn data with number of exceedances 93.

Month	March	April	May
Frequency	13	56	24

Table 4.38: Monthly frequency analysis of the time varying autumn data above $u = 1550$

Figure 4.36 is the histogram representing the monthly frequency analysis of the time varying autumn data above threshold, $u = 1550$

© University of Venda

University of Venda
Creating Future Leaders

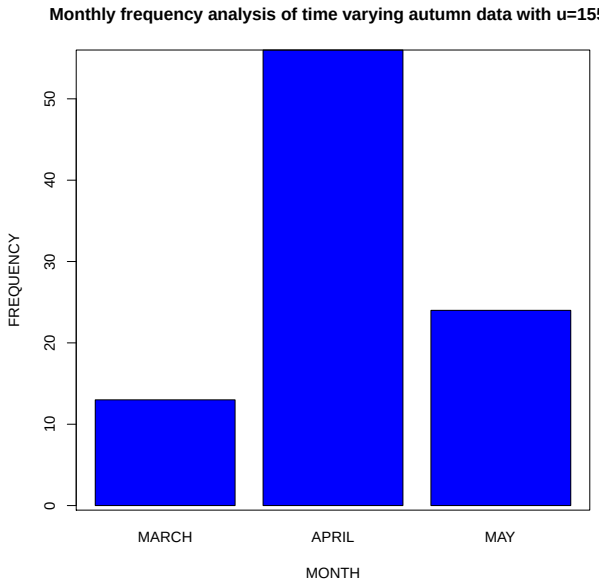

Figure 4.36: Histogram representing the monthly frequency analysis of the time varying autumn data above threshold, $u = 1550$

4.6 Poisson GPD and point process analysis of the non-linear detrended data

4.6.1 All data

Using automatic declustering method emphasized by Ferro and Segers (2003), all the time varying data is declustered. The length of the original series is 3711, with threshold $u = 1489$, number of threshold exceedances 380, interval estimator of extremal index $\theta = 0.476096$ indicating a fairly independent process as θ is approximately 0.5, run length of 2 and 180 identified clusters. The R package "evd" is used to extract the cluster maxima and a stationary point process is fitted to the cluster maxima using the R package "ismev". The Table 4.39 shows the maximum likelihood estimates with their errors in parenthesis.

Threshold	μ	σ	ξ
1489	3105.65508170(230.42563648)	224.91622944(80.88065629)	-0.06487778(0.07290805)

Table 4.39: Model fit by maximum likelihood

© University of Venda

The new scale parameter, σ after reparameterisation is estimated as:

$$\tilde{\sigma}^* = \tilde{\sigma} + \tilde{\xi}(u - \tilde{\mu}) = 224.91622944 - 0.06487778(1489 - 3105.65508170) = 329.801 \tag{4.2}$$

We estimate the intensity function of the point process which measures the frequency of the occurrence of the daily peak electricity demand per year as follows:

$$\tilde{\lambda} = \left(1 + \tilde{\xi}\frac{(u - \tilde{\mu})}{\tilde{\sigma}^*}\right)^{\frac{-1}{\tilde{\xi}}} \tag{4.3}$$

$$= \left(1 - 0.06487778\frac{(1489 - 3105.65508170)}{329.801}\right)^{15.413597} \tag{4.4}$$

$$= 70.543 \simeq 71 \tag{4.5}$$

Statistical inference of the value of the intensity function ($\tilde{\lambda} \simeq 71$), indicates that, daily peak electricity demand would be experienced approximately 71 days in a year (approximately 71 days of extreme daily changes in a year).

We fit Poisson GPD to the cluster maxima using the R package "ismev". The Table 4.40 shows the maximum likelihood estimate of the scale parameter and the shape parameter with their standard errors in parenthesis.

Threshold	nexc	NLLH	σ	ξ
1489	180	1212.086	329.90605856 (34.40827307)	-0.06487812 (0.07311627)

Table 4.40: Poisson GPD fitted to the cluster maxima of all the time varying data

© University of Venda

The Figure 4.37 illustrates the diagnostic plots of the Genaralised Pareto distribution fitted to the cluster maxima. Visual inspection of the plots shows a fairly good fit.

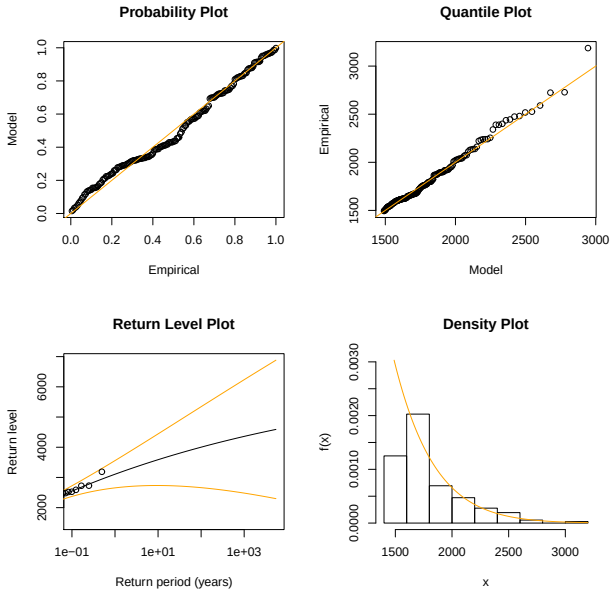

Figure 4.37: Diagnostic plot of the Genaralized Pareto distribution fitted to the cluster maxima

The Table 4.41 gives the summary (Parameter Bootstrap) of the Poisson GPD parameter uncertainties.

	σ	ξ
Original	5.79899038	-0.06475881
Bootstrap mean	5.81292634	-0.08203327
Bias	0.01393596	-0.01727446
SD	0.10526337	0.07775035
Bootstrap median	5.81780977	-0.08091802
Correlation		
	σ	ξ
σ	1.0000000	-0.7538874
ξ	-0.7538874	1.0000000

Table 4.41: Poisson GPD paramter uncertainties: Parameteric Bootstrap

© University of Venda

University of Venda
Creating Future Leaders

We plot the Bootstrap density which gives us the empirical distribution from the plots in Figure 4.38.

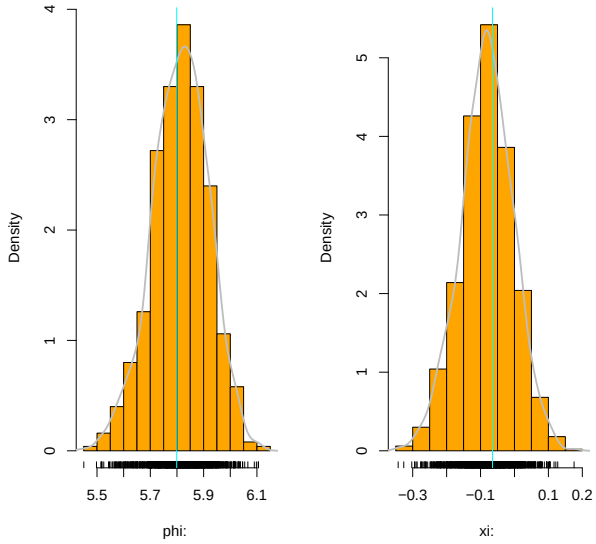

Figure 4.38: The Bootstrap plot of all the time varying data

We obtain the return level plot in Figure 4.39 which shows the predicted return levels with 95% predictive intervals from 20 to 100 years at 10 years interval.

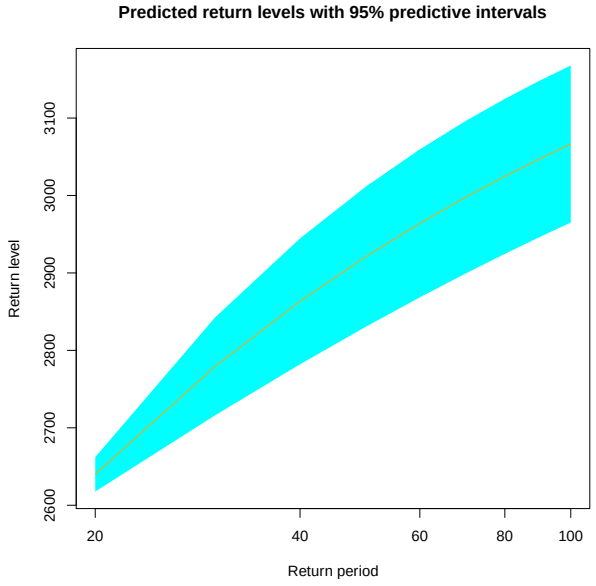

Figure 4.39: Return level estimation of all the time varying data

© University of Venda

University of Venda
Creating Future Leaders

The Table 4.42 shows the predictive intervals for the return levels:

Year	LCL (2.5%)	RL	UCL (97.5%)
20	2617.713	2639.913	2662.113
30	2716.118	2779.306	2842.494
40	2782.061	2863.309	2944.558
50	2830.611	2921.095	3011.579
60	2868.453	2964.021	3059.590
70	2899.102	2997.559	3096.015
80	2924.622	3024.714	3124.805
90	2946.321	3047.295	3148.269
100	2965.074	3066.464	3167.854

Table 4.42: 95% predictive interval for return levels

Inference from the table indicates that, we are 95% confident that, a peak electricity demand of 2639.913 would be reached with a lower confident limit of 2617.713 and an upper confident limit of 2662.113 in 20 years time and so on.

4.6.2 Winter data

We decluster the time varing winter data according to Ferro and Segers (2003). The length of the original series is 1235, with threshold $u = 1516$, number of threshold exceedances 168, interval estimator of extremal index $\theta = 0.5648498$ indicating an independent process as θ is closer to 1, run length of 1 and 89 identified clusters. The R package "evd" is used to extract the cluster maxima and a stationary point process is fitted to the cluster maxima using the R package "ismev". The Table 4.43 gives the maximum likelihood estimates with their errors in parenthesis.

Threshold	μ	σ	ξ
1516	2976.659910(270.5649550)	141.842347(83.4913244)	-0.174037(0.1214247)

Table 4.43: Model fit by maximum likelihood

© University of Venda

The new scale parameter, σ after reparameterisation is estimated as:

$$\tilde{\sigma}^* = \tilde{\sigma} + \tilde{\xi}(u - \tilde{\mu}) = 141.842347 - 0.174037(1516 - 2976.659910) = 396.051 \tag{4.6}$$

We estimate the intensity function of the point process which measures the frequency of the occurrence of the daily peak electricity demand per year as follows:

$$\tilde{\lambda} = \left(1 + \tilde{\xi}\frac{(u - \tilde{\mu})}{\tilde{\sigma}^*}\right)^{\frac{-1}{\tilde{\xi}}} \tag{4.7}$$

$$= \left(1 - 0.174037\frac{(1516 - 2976.659910)}{396.051}\right)^{5.745904606} \tag{4.8}$$

$$= 17.270272 \simeq 18 \tag{4.9}$$

Statistical inference of the value of the intensity function ($\tilde{\lambda} \simeq 18$), indicates that, daily peak electricity demand would be experienced approximately 18 days in a year (approximately 18 days of extreme daily changes in a year).

We fit Poisson GPD to the cluster maxima using the R package "ismev". The Table 4.44 indicates the maximum likelihood estimate of the scale parameter and the shape parameter with their standard errors in parenthesis.

Threshold	nexc	NLLH	σ	ξ
1516	89	605.8614	395.8867259 (63.8105427)	-0.1736926 (0.1228278)

Table 4.44: Poisson GPD fitted to the cluster maxima of the time varying winter data

© University of Venda

The Figure 4.40 shows the diagnostic plots of the Genaralised Pareto distribution fitted to the cluster maxima. Visual inspection of the plots shows a fairly good fit.

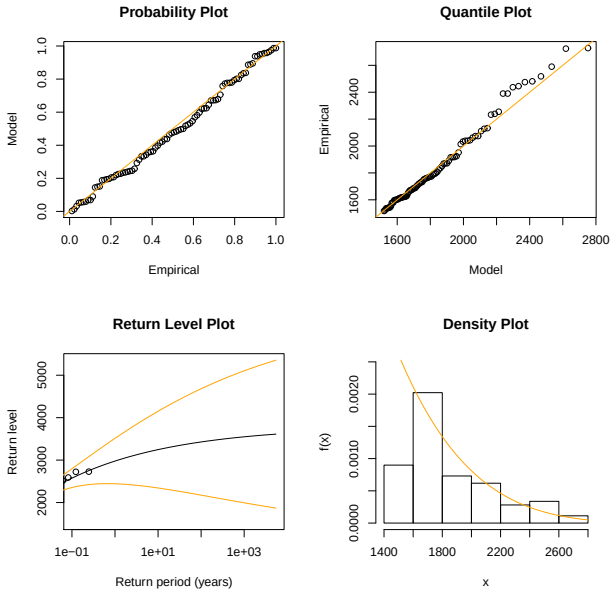

Figure 4.40: Diagnostic plot of the Genaralized Pareto distribution fitted to the cluster maxima

The Table 4.45 gives the summary (Parameter Bootstrap) of the GPD parameter uncertainties.

	σ	ξ
Original	5.98109442	-0.17362659
Bootstrap mean	6.01248106	-0.21645391
Bias	0.03138664	-0.04782732
SD	0.15053587	0.11050527
Bootstrap median	6.01162567	-0.20824132
Correlation		
	σ	ξ
σ	1.0000000	-0.8137344
ξ	-0.8137344	1.0000000

Table 4.45: Poisson GPD paramter uncertainties: Parameteric Bootstrap

© University of Venda

We plot the Bootstrap density which gives us the empirical distribution from the plots in Figure 4.41.

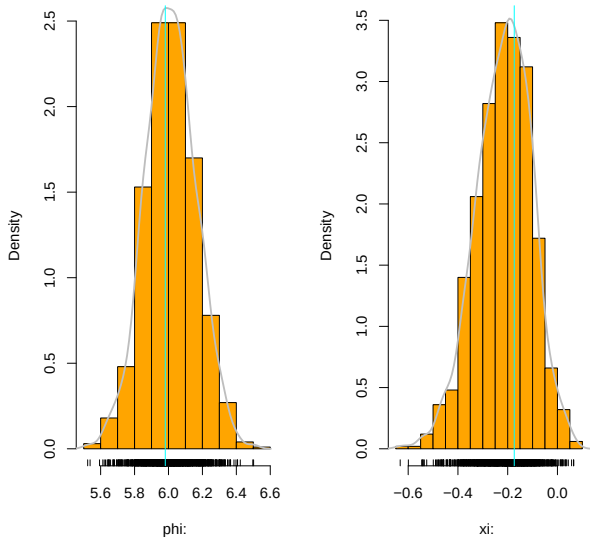

Figure 4.41: The Bootstrap plot of the time varying winter data

We obtain the return level plot in Figure 4.42 which indicates the predicted return levels with 95% predictive intervals from 20 to 100 years at 10 years interval.

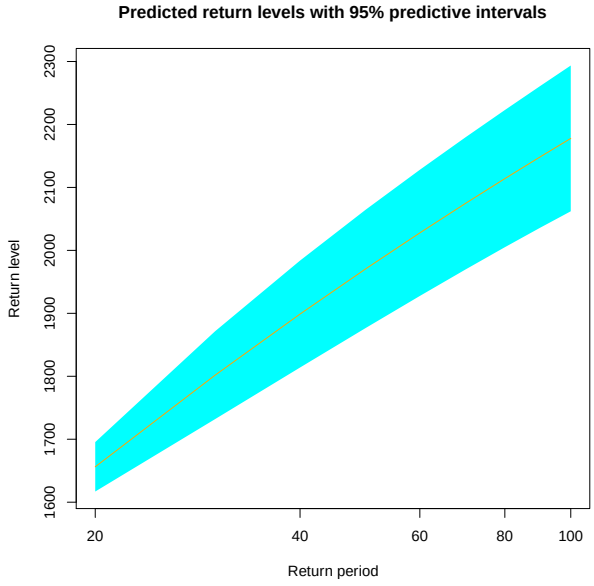

Figure 4.42: Return level estimation of all the time varying data

© University of Venda

The Table 4.46 givess the predictive intervals for the return levels:

Year	LCL (2.5%)	RL	UCL (97.5%)
20	1616.931	1656.212	1695.493
30	1731.903	1801.674	1871.458
40	1813.903	1898.843	1983.783
50	1876.937	1970.941	2064.946
60	1927.592	2027.811	2128.031
70	1969.532	2074.509	2179.486
80	2005.013	2113.963	2222.913
90	2035.524	2148.012	2260.500
100	2062.104	2177.886	2293.667

Table 4.46: 95% predictive interval for return levels

Inference from the table indicates that, we are 95% confident that, a peak electricity demand of 1656.212 would be reached with a lower confident limit of 1616.931 and an upper confident limit of 1695.493 in 20 years time and so on.

4.6.3 Summer data

We decluster the time varing summer data using the automatic declustering method. The length of the original series is 905, with threshold $u = 1125$, number of threshold exceedances 297, interval estimator of extremal index $\theta = 0.3645509$ indicating a dependent process as θ is closer to 0 than 1, run length of 2 and 74 identified clusters. We use the R package "evd" to extract the cluster maxima and fit stationary point process to the cluster maxima using the R package "ismev".

© University of Venda

The Table 4.47 givess the maximum likelihood estimates with their standard errors in parenthesis.

Threshold	μ	σ	ξ
1125	2357.5249670(98.96740374)	53.6232344(23.07456961)	-0.3894331(0.09339181)

Table 4.47: Model fit by maximum likelihood

The new scale parameter, σ after reparameterisation is estimated as:

$$\tilde{\sigma}^* = \tilde{\sigma} + \tilde{\xi}(u - \tilde{\mu}) = 53.6232344 - 0.3894331(1125 - 2357.5249670) = 533.609 \tag{4.10}$$

We estimate the intensity function of the point process which measures the frequency of the occurrence of the daily peak electricity demand per year as follows:

$$\tilde{\lambda} = \left(1 + \tilde{\xi}\frac{(u - \tilde{\mu})}{\tilde{\sigma}^*}\right)^{\frac{-1}{\tilde{\xi}}} \tag{4.11}$$

$$= \left(1 - 0.3894331\frac{(1125 - 2357.5249670)}{533.609}\right)^{2.567835} \tag{4.12}$$

$$= 5.194 \simeq 6 \tag{4.13}$$

Statistical inference of the value of the intensity function ($\tilde{\lambda} \simeq 6$), indicates that, daily peak electricity demand would be experienced approximately 6 days in a year (approximately 6 days of extreme daily changes in a year).

We fit Poisson GPD to the cluster maxima using the R package "ismev". The Table 4.48 shows the maximum likelihood estimate of the scale parameter and the shape parameter with their standard errors in parenthesis.

Threshold	nexc	NLLH	σ	ξ
1125	74	509.8857	533.79781 (75.98248731)	-0.38947 (0.09382287)

Table 4.48: Poisson GPD fitted to the cluster maxima of the time varying summer data

© University of Venda

The Figure 4.43 shows the diagnostic plots of the Genaralised Pareto distribution fitted to the cluster maxima. Visual inspection of the plots shows a fairly good fit.

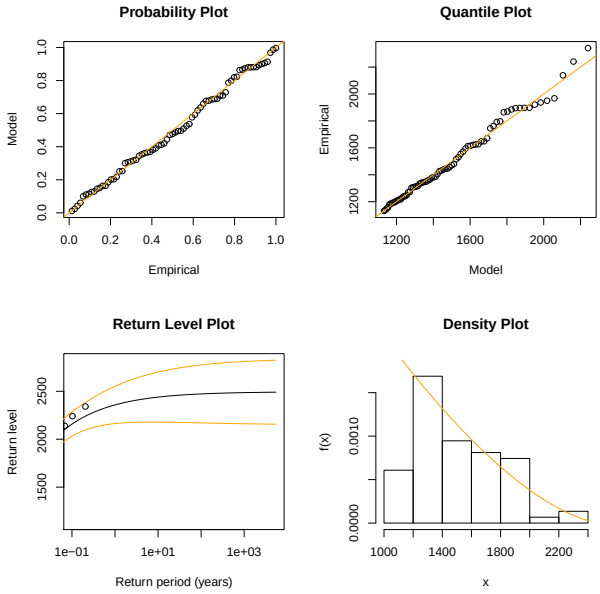

Figure 4.43: Diagnostic plot of the Genaralized Pareto distribution fitted to the cluster maxima

The Table 4.49 gives the summary (Parameter Bootstrap) of the Poisson GPD parameter uncertainties.

	σ	ξ
Original	6.27956451	-0.38938252
Bootstrap mean	6.33023340	-0.44997868
Bias	0.05066889	-0.06059616
SD	0.15198382	0.11662655
Bootstrap median	6.33108356	-0.44232894
Correlation		
	σ	ξ
σ	1.0000000	-0.8946438
ξ	-0.8946438	1.0000000

Table 4.49: Poisson GPD paramter uncertainties: Parameteric Bootstrap

© University of Venda

We plot the Bootstrap density which gives us the empirical distribution from the plots in Figure 4.44.

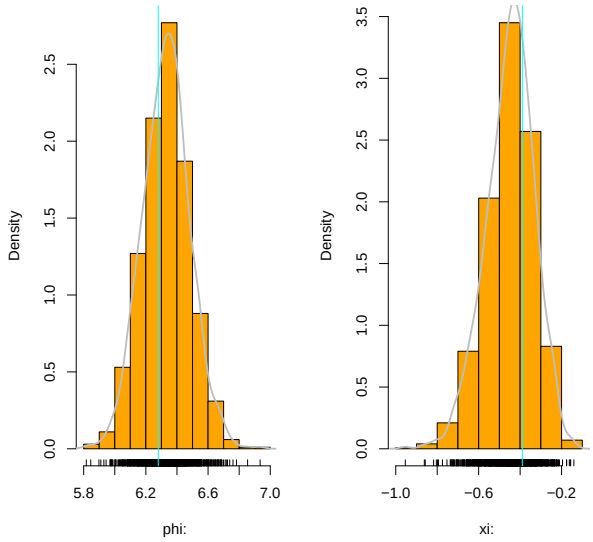

Figure 4.44: The Bootstrap plot of the time varying summer data

We obtain the return level plot in Figure 4.45 shows the predicted return levels with 95% predictive intervals from 20 to 100 years at 10 years interval.

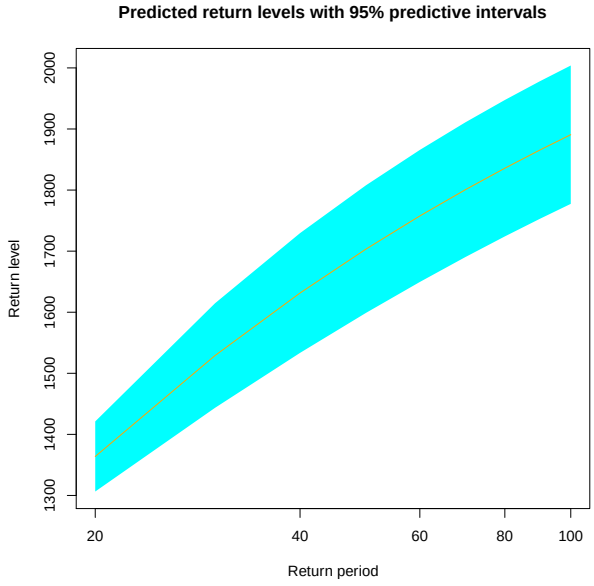

Figure 4.45: Return level estimation of the time varying summer data

© University of Venda

University of Venda
Creating Future Leaders

The Table 4.50 below shows the predictive intervals for the return levels:

Year	LCL (2.5%)	RL	UCL (97.5%)
20	1306.431	1363.835	1421.24
30	1443.483	1529.079	1614.674
40	1533.435	1631.467	1729.50
50	1598.897	1703.353	1807.808
60	1649.53	1757.623	1865.716
70	1690.308	1800.597	1910.885
80	1724.11	1835.793	1947.476
90	1752.742	1865.355	1977.967
100	1777.409	1890.674	2003.939

Table 4.50: 95% predictive interval for return levels

Inference from the table indicates that, we are 95% confident that, a peak electricity demand of 1363.835 would be reached with a lower confident limit of 1306.431 and an upper confident limit of 1421.24 in 20 years time and so on.

4.6.4 Spring data

We decluster the time varing summer data using the automatic declustering method. The length of the original series is 963, with threshold $u = 1183$, number of threshold exceedances 173, interval estimator of extremal index $\theta = 0.5821893$ indicating a independent process as θ is closer to 1 than 0, run length of 1 and 91 identified clusters. We use the R package "evd" to extract the cluster maxima and fit stationary point process to the cluster maxima using the R package "ismev". The Table 4.51 gives the maximum likelihood estimates with their standard errors in parenthesis.

Threshold	μ	σ	ξ
1183	3000.44048274(311.184367443)	263.26662120(105.87157663)	-0.05194042(0.08090203)

Table 4.51: Model fit by maximum likelihood

© University of Venda

The new scale parameter, σ after reparameterisation is estimated as:

$$\tilde{\sigma}^* = \tilde{\sigma} + \tilde{\xi}(u - \tilde{\mu}) = 263.26662120 - 0.05194042(1183 - 3000.44048274) = 357.6652 \quad (4.14)$$

We estimate the intensity function of the point process which measures the frequency of the occurrence of the daily peak electricity demand per year as follows:

$$\tilde{\lambda} = \left(1 + \tilde{\xi}\frac{(u - \tilde{\mu})}{\tilde{\sigma}^*}\right)^{\frac{-1}{\tilde{\xi}}} \quad (4.15)$$

$$= \left(1 - 0.05194042\frac{(1183 - 3000.44048274)}{357.6652}\right)^{19.2528} \quad (4.16)$$

$$= 90.8769 \simeq 91 \quad (4.17)$$

Statistical inference of the value of the intensity function ($\tilde{\lambda} \simeq 91$), indicates that, daily peak electricity demand would be experienced approximately 91 days in a year (approximately 91 days of extreme daily changes in a year).

We fit Poisson GPD to the cluster maxima using the R package "ismev". The Table 4.52 below shows the maximum likelihood estimate of the scale parameter and the shape parameter with their standard errors in parenthesis.

Threshold	nexc	NLLH	σ	ξ
1183	91	621.2989	357.5862770 (47.32278856)	-0.0518241 (0.08098523)

Table 4.52: Poisson GPD fitted to the cluster maxima of the time varying spring data

© University of Venda

The Figure 4.46 shows the diagnostic plots of the Genaralised Pareto distribution fitted to the cluster maxima. Visual inspection of the plots shows a fairly good fit.

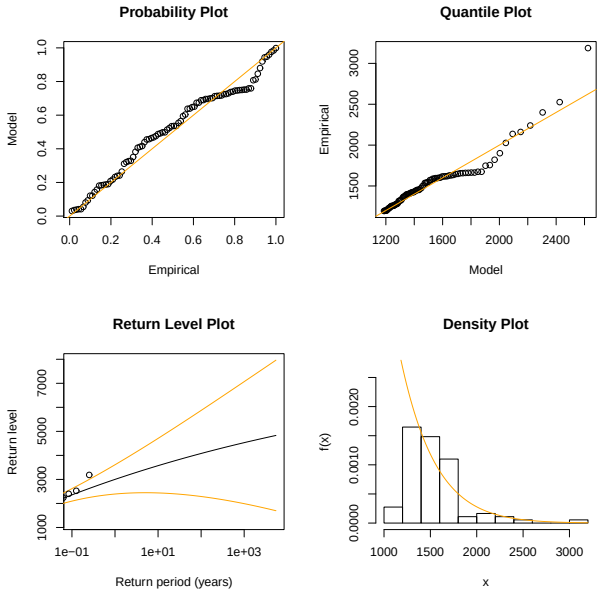

Figure 4.46: diagnostic plot of the Genaralized Pareto distribution fitted to the cluster maxima

The Table 4.53 gives the summary (Parameter Bootstrap) of the Poisson GPD parameter uncertainties.

	σ	ξ
Original	5.87911154	-0.05185662
Bootstrap mean	5.91073883	-0.08669630
Bias	0.03162729	-0.03483968
SD	0.15432972	0.11906930
Bootstrap median	5.90955590	-0.07771323
Correlation		
	σ	ξ
σ	1.0000000	-0.7728373
ξ	-0.7728373	1.0000000

Table 4.53: Poisson GPD parameter uncertainties: Parametric Bootstrap

© University of Venda

We plot the Bootstrap density which gives us the empirical distribution from the plots in Figure 4.47.

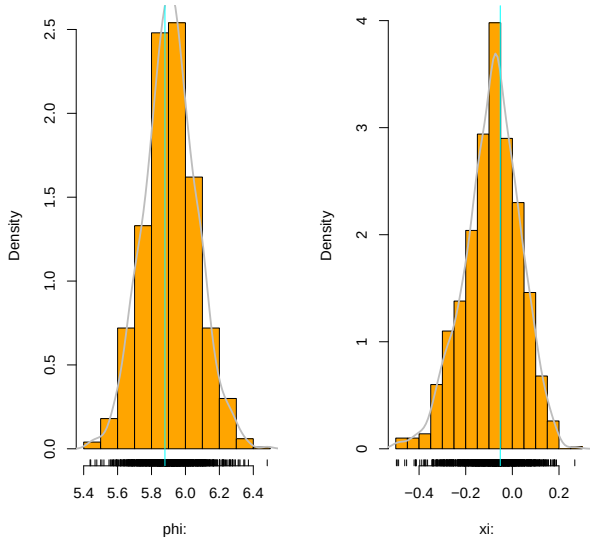

Figure 4.47: The Bootstrap plot of the time varying spring data

We obtain the return level plot in Figure 4.48 which shows the predicted return levels with 95% predictive intervals from 20 to 100 years at 10 years interval.

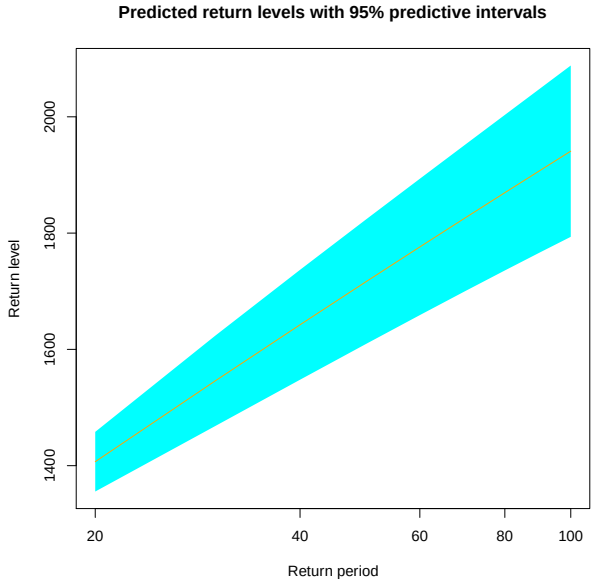

Figure 4.48: Return level estimation of the time varying spring data

© University of Venda

The Table 4.54 shows the predictive intervals for the return levels:

Year	LCL (2.5%)	RL	UCL (97.5%)
20	1355.535	1406.842	1458.15
30	1467.984	1545.622	1623.26
40	1547.836	1642.333	1736.83
50	1609.300	1716.361	1823.422
60	1658.945	1776.213	1893.481
70	1700.365	1826.378	1952.391
80	1735.74	1869.51	2003.279
90	1766.495	1907.307	2048.12
100	1793.609	1940.924	2088.238

Table 4.54: 95%predictive intervals for return levels

Inference from the table indicates that, we are 95% confident that, a peak electricity demand of 1406.842 would be reached with a lower confident limit of 1355.535 and an upper confident limit of 1458.15 in 20 years time and so on.

4.6.5 Autumn data

We decluster the time varing summer data using the automatic declustering method. The length of the original series is 924, with threshold $u = 1550$, number of threshold exceedances 82, interval estimator of extremal index $\theta = 0.5840964$ indicating a independent process as θ is closer to 1 than 0, run length of 1 and 44 identified clusters. We use the R package "evd" to extract the cluster maxima and fit stationary point process to the cluster maxima using the R package "ismev". The Table 4.55 gives the maximum likelihood estimates with their standard errors in parenthesis.

Threshold	μ	σ	ξ
1550	2614.0812560(197.9050440)	67.2162082(51.8117922)	-0.2933181(0.1615861)

Table 4.55: Model fit by maximum likelihood

© University of Venda

The new scale parameter, σ after reparameterisation is estimated as:

$$\tilde{\sigma}^* = \tilde{\sigma} + \tilde{\xi}(u - \tilde{\mu}) = 67.2162082 - 0.2933181(1550 - 2614.0812560) = 379.3305 \tag{4.18}$$

We estimate the intensity function of the point process which measures the frequency of the occurrence of the daily peak electricity demand per year as follows:

$$\tilde{\lambda} = \left(1 + \tilde{\xi}\frac{(u - \tilde{\mu})}{\tilde{\sigma}^*}\right)^{\frac{-1}{\tilde{\xi}}} \tag{4.19}$$

$$= \left(1 - 0.2933181\frac{(1550 - 2614.0812560)}{379.3305}\right)^{3.40927} \tag{4.20}$$

$$= 7.7434 \simeq 8 \tag{4.21}$$

Statistical inference of the value of the intensity function ($\tilde{\lambda} \simeq 8$), indicates that, daily peak electricity demand would be experienced approximately 8 days in a year (approximately 8 days of extreme daily changes in a year).

We fit Poisson GPD to the cluster maxima using the R package "ismev". The Table 4.56 below shows the maximum likelihood estimate of the scale parameter and the shape parameter with their standard errors in parenthesis.

Threshold	nexc	NLLH	σ	ξ
1550	44	292.3795	379.4640341 (81.9040782)	-0.2937848 (0.1606514)

Table 4.56: Poisson GPD fitted to the cluster maxima of the time varying spring data

© University of Venda

The Figure 4.49 shows the diagnostic plots of the Genaralised Pareto distribution fitted to the cluster maxima. Visual inspection of the plots shows a fairly good fit.

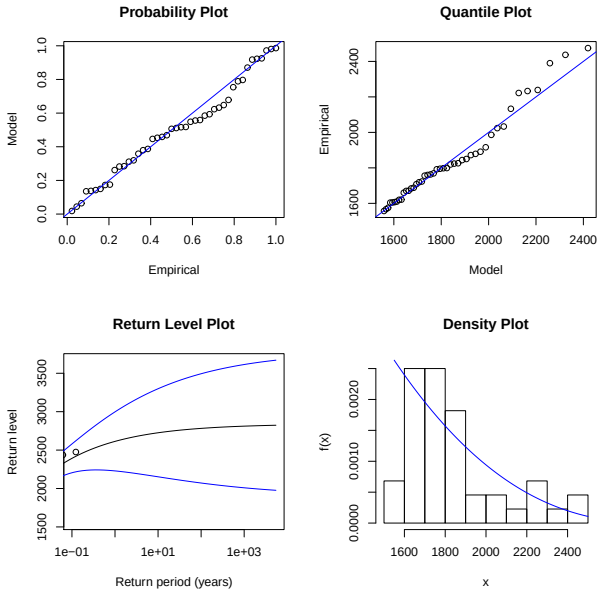

Figure 4.49: Diagnostic plot of the Genaralized Pareto distribution fitted to the cluster maxima

The Table 4.57 gives the summary (Parameter Bootstrap) of the Poisson GPD parameter uncertainties.

	σ	ξ
Original	6.00767585	-0.4689300
Bootstrap mean	6.10318140	-0.5929120
Bias	0.09550555	-0.1239820
SD	0.22576582	0.1921986
Bootstrap median	6.09990739	-0.5748507
Correlation		
	σ	ξ
σ	1.0000000	-0.9227822
ξ	-0.9228898	1.0000000

Table 4.57: Poisson GPD parameter uncertainties: Parametric Bootstrap

© University of Venda

We plot the Bootstrap density which gives us the empirical distribution from the plots in Figure 4.50.

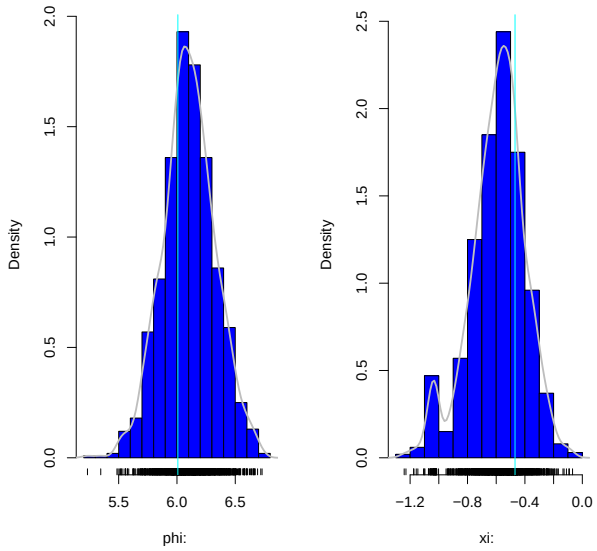

Figure 4.50: The Bootstrap plot of the time varying autumn data

We obtain the return level plot in Figure 4.51 which shows the predicted return levels with 95% predictive intervals from 20 to 100 years at 10 years interval.

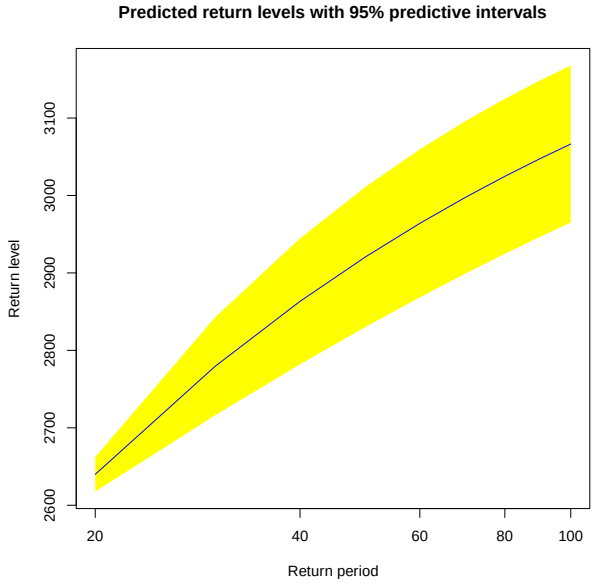

Figure 4.51: Return level estimation of the time varying autumn data

© University of Venda

The Table 4.58 shows the predictive intervals for the return levels:

Year	LCL (2.5%)	RL	UCL (97.5%)
20	2617.713	2639.913	2662.113
30	2716.118	2779.306	2842.494
40	2728.061	2863.309	2944.558
50	2830.611	2921.095	3011.579
60	2868.453	2964.021	3059.590
70	2899.102	2997.559	3096.015
80	2924.622	3024.714	3124.805
90	2946.321	3047.295	3148.269
100	2965.074	3066.464	3167.854

Table 4.58: 95% predictive intervals for return levels

Inference from the table indicates that, we are 95% confident that, a peak electricity demand of 2639.913 would be reached with a lower confident limit of 2617.713 and an upper confident limit of 2662.113 in 20 years time and so on.

© University of Venda

5. Chapter 5

5.1 Discussion

This study sets out to stochastically model the daily peak electricity demand using extreme value theory. This section of the chapter discusses the findings of the study based on the data analysis and the answers. To ensure the DPED is in statistical equilibrium, we apply a penalized regression cubic smoothing spline to remove all deterministic trend. High dependence was observed within the threshold excceedances, so we declustered the data according to Ferro and Segers (2003) interval method and the cluster maxima of each of the non-linear detrended datasets were extracted using the R package "evd". This was done to weaken the dependence structure of the threshold exceedances to keep the data relatively independent. We therefore fit Poisson GPD and stationary point process to the cluster maxima and the parameters were estimated using the maximum likelihood estimates. The negative values of the shape parameters of the model indicate that the tail distribution of the DPED belongs to the Weibull domain of attraction. All computations and graphics presented in this dissertation were performed using the R statistical programming langauge. Some of the packages were "tseries", "evd","fExtremes" ,"texmex", "ismev", "evmix". The codes used for the outputs in this dissertation are found in appendix G.
The empirical research findings are consistent with the main aim and objectives outlined in the study.

5.1.1 Genaralized Pareto distribution frequency analysis

All the chosen thresholds for the data analysis for the time varying or linear detrended data were estimated using the R package "evmix" through non-parametric extremal mixture model, where a kernel density was fitted to the time varying data.

5.1.2 All non-linear detrended data

The estimated threshold is 1489 with number of exceedances 369. The output of the yearly frequency analysis is shown in Tables 4.2, 4.3 and Figure 4.4. Inference from the output shows that the year 2006 experienced the highest frequency of 32, 2004 with 30, 2009 with 28 and the year 1999 with the lowest of 11.
We display the monthly frequency analysis of the data in Tables 4.4, 4.5 and Figure 4.5. From the

© University of Venda

University of Venda
Creating Future Leaders

output, January experienced the highest exceedances of 56, December with 55, April with 51 and February with none.

5.1.3 Non-linear detrended winter data

The threshold estimated is 1356 with number of exceedances being 102. The output of the yearly frequency analysis is shown in Tables 4.6, 4.7 and Figure 4.7. From the output, the year 1999 experienced the highest exceedance of 10, 2004 and 2009 with 9 times each and 2013 with none.
In Table 4.8 and Figure 4.8, the monthly frequency analysis are shown with the month of August having the highest frequency of 56, June 33 and July 13. The values are high as the demand for electricity is high in winter seasons.

5.1.4 Non-linear detrended summer data

The number of exceedances for the data is 97 with a threshold of 1125. The output for the yearly frequency analysis is displayed in Tables 4.24, 4.25 and Figure 4.25. From the outputs, the year 2005 experienced the highest frequency of 13, followed by the year 2004 with 11 and 2001 having the smallest of 1.
In Table 4.26 and Figure 4.26, we display the monthly frequency analysis where the month of December experienced the highest exceedances of 56, followed by January with 31 and lastly February with 10.

5.1.5 Non-linear detrended spring data

The estimated threshold is 1183 with the number of excceedances 92. The output of the yearly and monthly frequency analysis is displayed in Tables 4.30, 4.31 and Figure 4.30. The figures from the table indicate that the year 2004 experienced the the highest exceedances of 15, the years 2005, 2006, 2009, and 2010 having a tie of 8.
The monthly frequency distribution is shown in Table 4.32 and Figure 4.31. The month of september led by 44 threshold excesses, followed by November with 28 and October with the least of 20. During the spring seasons, South Africa experiences cold weather conditions causing increased usage of geysers and other heating appliances.

© University of Venda

University of Venda
Creating Future Leaders

5.1.6 Non-linear detrended autumn data

The value of the estimated threshold is 1550 with the number of exceedances being 93. The yearly frequency distribution is displayed in Tables 4.36, 4.37 and Figure 4.35. Inference from the output shows the years 1997, 2000, and 2009 experienced the same number of exceedances of 8 with the year 2000 having the least exceedances of 3.

The monthly frequency distribution of the autumn data is also shown in Table 4.38 and Figure 4.36. Careful analysis shows the month of April experiencing the highest with 56 excceedances, May with 24 and March with the lowest among the three with 13.

5.2 Poisson GPD and stationary point process analysis

5.2.1 All non-linear detrended data

Table 4.39 gives the summary statistics of the maximum likelihood estimates of the Poisson GPD and stationary point process fitted to the cluster maxima of all the time varying data. The reparameterised scale parameter is evaluated in (4.2), giving $\tilde{\sigma}^* = 329.801$. The associated intensity function which measures the frequency of the occurrence of the daily peak electricity demand per year is also computed in (4.5) as $\tilde{\lambda} \simeq 71$, indicating statistically that approximately 71 days of extreme daily changes occured in one year.

5.2.2 Non-linear detrended winter data

In Table 4.43, we display the output of the summary statistics of the maximum likelihood estimates of the Piosson GPD and stationary point process fitted to the cluster maxima of the time varying winter data. The reparameterised scale parameter is evaluated in (4.6), giving $\tilde{\sigma}^* = 396.051$. The associated intensity function which measures the frequency of the occurrence of the daily peak electricity demand per year is also computed in (4.9) as $\tilde{\lambda} \simeq 18$, indicating statistically that approximately 18 days of extreme daily changes occured in one year.

© University of Venda

University of Venda
Creating Future Leaders

5.2.3 Non-linear detrended summer data

We fit Poisson GPD and stationary point process to the cluster maxima of the time varying summer data and the maximum likelihood estimates are given in Table 4.47. The reparameterised scale parameter is evaluated in (4.10), giving $\tilde{\sigma}^* = 533.609$. The corresponding intensity function is calculated in (4.13) as $\tilde{\lambda} \simeq 6$, meaning that approximately 6 days of extreme daily changes occured in one year.

5.2.4 Non-linear detrended spring data

The maximum likelihood estimates obtained as a results of fitting the Poisson GPD and stationary point process to the cluster maxima of the time varying spring data are displayed in Table 4.51. The scale parameter is reparaterised and evaluated in (4.14) as $\tilde{\sigma}^* = 357.6652$.
The intensity function of the point process which measures the frequency of occurrence of the daily peak electricity demand per year is calculated in (4.17) as $\tilde{\lambda} \simeq 91$, meaning that approximately 91 days of extreme daily changes occured in one year.

5.2.5 Non-linear detrended autumn data

Table 4.55 gives the maximum likelihood estimates of the Piosson GPD and stationary point process fitted to the cluster maxima. The reparameterised scale parameter is calculated in (4.18), giving $\tilde{\sigma}^* = 379.3305$. The associated intensity function is calculated in (4.21) as $\tilde{\lambda} \simeq 8$, meaning that approximately 8 days of extreme daily changes occured in one year.

5.3 Contribution

This dissertation has added value to knowledge as existing research in literature focused on day-to-day changes based on differenced data sets. The research interest was actually focused on the extreme daily peaks derived from the non-linear detrended data and the frequency analysis were based on all the non-linear detrended data as well as the four seasons as winter, summer, spring and autumn based on the calender dates in the Southern Hemisphere.
Also, the knowledge about the intensity function of the point process coupled with the return level estimation of each of the datasets will meaningfully inform ESKOM, South African power utility company

© University of Venda

University of Venda
Creating Future Leaders

about effective planning and maintenance during the non-winter seasons to ensure stability in the grid system.

5.4 Conclusion and Recommendation for Future Work

5.4.1 Conclusion

The formal goodness of fit tests based on Anderson-Darling and Cramer-Von Mises test statistics proved that, each data set of the time varying data followed Poisson GPD with the associated estimates of both scale and shape parameters at 5% level of significant. This also means that the usual asymptotic properties underlying the Poisson GPD were satisfied by the model.

5.4.2 Recommendation for Future Work

Future researchers could consider using the discrete time Markov Chain (DTMC) and the Bayesian estimation approach to stochastically model the daily peak electricity damand. They could also consider including covariates such as temperature in the case of non-stationary dependent sequences as this study focused only the stationary dependent sequences without considering the non-stationary situations.

5.5 Limitations of the dissertation

The daily peak electricity data in South Africa were obtained from the different sectors of the economy namely, agricultural, commercial, residential and industrial. Under this situation, it is not certain to determine how significant each section of the economy is on daily demand. However, the extremal index of each of the differenced data sets were greater than that of the time varying data sets indicating that the differenced data are more independent, hence the frequency analysis based on the differenced data sets would be more reliable as compared to the analysis based on the varying data sets.

© University of Venda

References

[1] BALKEMA, A. A., & DE HAAN, L., (1974), Residual life time at great age, The Annals of Probability, pp. 792-804.

[2] BEHRENS, C. N., LOPES, H.F, & GAMERMAN, D., (2004), Bayesian analysis of extreme events with threshold estimation. ,Statistical modelling. 4(3), pp. 227-244.

[3] BEICHELT, F, (2006), Stochastic Processes in Science, Engineering and Finance, Chapman and Hall/CRC, pp. 410.

[4] BEIRLANT, J., GOEGEBEUR, Y., SEGERS, J., & TEUGELS, J., (2004), Statistics of Extremes: Theory and Applications, John Wiley and Sons, Ltd, pp. 490.

[5] BOMMIER, E., (2014), Peaks-Over Thresholds modelling of environmental data, Department of Mathematics, Uppsala University, Sweden: Technical report 33, 3 September.

[6] CAI W., & CAMPBELL E. P., (2005), Statistical modelling of extreme rainfall in South Western Australia.

[7] CHANG, S., & NADARAJAH, S., (2015), Extreme value analysis of electricity demand in the U.K., Applied Economics Letters, 22(15), pp. 1246-1251.

[8] CHOULAKIAN, V., & STEPHENS, M. A., (2001), Goodness-of-Fit Tests for the Generalized Pareto Distribution, Technometrics, 43 (4), pp. 478-484.

[9] COLES, S., (2001), An Introduction to Statistical Modelling of Extreme Values, Springer Series in Statistics Springer-Verlag London Ltd, pp. 208.

[10] COLES S. G., & POWELL, E. A., (1996), Bayesian methods in extreme value modelling, A review and new developments. International Statistical Review/ Revue International de Statistique, 64(1), pp. 119-136.

[11] COWLES, M.K., (2013), Applied Bayesian Statistics: with R and Open BUGS examples, Volume 98, Springer Science and Business Media.

[12] DAVISON, A.C., & HALL, P., (1993), *On Studentizing and blocking methods for implementing the bootstrap with dependent data, Aust. J. Statistist*, 35, pp. 215-224.

© University of Venda

University of Venda
Creating Future Leaders

[13] DAVISON, A., & KUONEN, D., (2002), An Introduction to the Bootstrap with Application in R, Statistical Computing and Statistical Graphics Newsletter, 13(1), pp. 6-11.

[14] DAVISON, A.C., & SMITH, R. L., (1990), *Models for exceedances over high thresholds, Journal of the Royal Statistical Society. Series B*, (Methodological), 52(3), pp. 393-442.

[15] DUIN, R.P.W., (1976), Choice of smoothing parameters for Parzen estimaters of probability density functions; IEEE Transactions on computers, 25(11), PP. 1175-1179.

[16] EASTOE, E. F., & TAWN, J. A., (2009), *Modelling non-stationary extremes with application to surface level ozone, Journal of the Royal Statistical Society*: Series C (Applied Statistics), 58(1), pp. 25-45.

[17] EMBRECHTS, P., KLUPPELBERG, C., & MIKOSCH, T., (1997), Modelling Extremal Events for Insurance and Finance, Berlin: Springer-Verlag.

[18] FAWCET, L., & WALSHAW, D., (2007), Improved estimation for temporary clustered extremes, Environmetrics, 18(2), pp. 173-188.

[19] FERRO, C. A., & SEGERS, J., (2003), *Inference for clusters of extreme values, Journal of the Royal Statistical Society*: Series B (Statistical Methodology), 65(2), pp. 545-556.

[20] FISHER, R. A., & TIPPETT, L. H. C., (1928), Limiting forms of the frequency distribution of the largest or smallest member of a sample, Proceedings of the Cambridge Philosophical Society, Volume 24(2). Cambridge University Press, pp.180.

[21] FUKUTOME, S., LINIGER, M.A., & SUVEGES, M., (2014), Automatic threshold and run parameter selection: A climatology for extreme hourly precipitation in Switzerland, Theoretical and Applied Climatology, 120(3), pp. 403-416.

[22] GENCAY, R., & SELCUK, F., (2004), *Extreme Value Theory and Value-at-Risk: Relative performance in emerging markets, International Journal of Forecasting*,20 (2),pp.287-303.

[23] GUMBEL, E.J., (1958), Statistics of Extremes, Columbia University Press.

[24] HAHN, H., MEYER-NIEBERG, S., & PICKL,S., (2009), *Electric load forecasting methods: Tools for decision making, European Journal of Operational Research*, 199(3), pp.902-907.

© University of Venda

University of Venda
Creating Future Leaders

[25] HASAN, A.R.N., & KASSIM, S., (2012), Modelling of Extreme Temperature Using Generalized Extreme Value (GEV) Distribution: A case study of Penang, Malaysia. In World Congress on Engineering, Volume 1, pp. 181-186.

[26] HEBBEMA, J.D.F., HERMANS, J., & VAN DEN BROEK, K., (1974), A stepwise discriminant analysis program using density estimation. Proceedings in Computational Statistics. Wien Physica Verlag. COMPSTAT.

[27] HEFFERNAN, J.E., & SOUTHWORTH, H., (2013), *Extreme value modelling of dependent series using R, Journal of Statistical Software*, 54(16), pp.1-25, R package version 2.1 http://www.jstatsoft.org/v54/i15/

[28] HILL, B. M., & PICKANDS, III, J., (1975), Statistical inference using extreme order statistics. The Annals of Statistics, 3(1), pp.119-131.

[29] HOR, C-L., WATSON, S.J., & MAJITHIA, S., (2006), Daily load forecasting and maximum demand estimation using ARIMA and GARCH, In Probabilistic Methods Applied to Power Systems 2006, PMAPS 2006, International conference on, pp. 1-6, IEEE.

[30] HSING ET AL., (1988), Probability Theory and Related Fields,78(1), pp. 97-112.

[31] HSING, T., (1991), On tail index estimation using dependent data. Ann Statist., volume 19, pp. 1547-1569.

[32] HU, Y & SCARROTT, C., (2013), *evmix:An R package for extreme value mixture modelling, threshold estimation and boundary corrected kernel density estimation. Journal of statistical Software*, 56(18),pp. 1-30, http: // www.math.canter.ac.nz/c.scarrott/evmix.

[33] HU, Y., (2013), Extreme Value Mixture Modelling with Simulation Study and Applications in Finance and Insurance, pp. 24.

[34] INGLESI, R., (2010), Aggregate electricity demand in South Africa: Conditional forecasts to 2030, Applied Energy, 87(1), pp. 197-204.

[35] JONES, M.C, & FOSTER, P.J, (1996), A simple nonnegative boundary correction method for kernel density estimation, Statistica Sinica, 6, pp. 1005-1013.

[36] KARR, R., (1991), Point Processes and their Statistical Inference, volume 7, Marcel Dekker, Inc. New York, pp. 499.

© University of Venda

University of Venda
Creating Future Leaders

[37] KOU, P., & GAO, F., (2014), *A sparse heteroscedastic model for the probabilistic load forecasting in energy-intensive enterprises, International Journal of Electrical Power and Energy Systems*, 55, pp. 144-154.

[38] KHULUSE, S. A., (2010), Modelling Heavy Rainfall Over Time and Space, Masters thesis, University of the Witwatersrand, Johannesburg, South Africa.

[39] KYSELY, J., PICEK, J., & BERANOVA, R., (2010), Estimating extremes in climate change simulations using the peaks-over-threshold method with a non-stationary threshold, Global and Planetary Change, 72(1), pp.55-68.

[40] LEADBETTER, M. R., (1983), Extremes and local dependence in stationary sequences, Probability Theory and Related Fields, 65(2), pp.291-306.

[41] LI, Y., CAI, W., & CAMPBELL, E., (2005), *Statistical modelling of extreme rainfall in Southwest Western Australia, Journal of Climate*, 18(16), pp.852-863.

[42] MACDONALD, A.E., SCARROTT, C. J., LEE, D.S., DARLOW, B., REALE, M., & RUSSELL, G., (2011), A flexible Extreme value mixture model. Computational Statistics and Data analysis, 55(6), pp.2137-2157.

[43] MACDONALD, A.E.,& SCARROTT, C. J., LEE, D.S., (2013), Boundary correction, consistency and robustness of kernel densities using extreme value theory: Available from http: www.math.canterbury.ac.nz/ c.scarrott.

[44] NERC, (1975), Flood studies report, Natural Environment Research Council (Vol 2, Meteorological Studies pp. 81)

[45] PICKANDS III, J., (1971), *The two-dimensional Poisson process and extremal processes, Journal of Applied Probability*, 8(4), pp.745-756.

[46] PRESCOTT, P., & WALDEN, A. T., (1983), *Maximum likelihood estimation of the parameters of the three-parameter generalised extreme value distribution from censored samples, Journal of Statistical computation and simulation*, 16, pp.241-250.

[47] RESNICK, S. I., (2013), Extreme values, Regular Variation and point Processes. Springer, pp. 319.

© University of Venda

University of Venda
Creating Future Leaders

[48] ROCHOWICZ JOHN A, JR, (2010), Bootstrapping analysis, Inferential Statistics and EXCEL, Spreadsheets in Education (ejSiE), 4(3), Article 4.

[49] SHUMWAY, R.H., & STOFFER, D.S., (2011), Time Series Analysis and Its Applications: With R Examples: Springer Science and Business Media, pp.596.

[50] SIGAUKE, C., CHIKOBVU, D., & VERSTER, A., (2012), *Modelling daily increases in peak electricity demand using the Generalized Pareto Distribution, South African Statistical Journal Proceedings: Proceedings of the 54th Annual Conference of the South African Statistical Association*: Congress 1, Sabinet Online, pp. 58-66.

[51] SMITH, R.L., (1985), Maximum likelihood estimation in a class of nonregular cases, Biometrika, 72, pp. 67-90.

[52] SMITH, R. L., (1987), Estimating tails of probability distributions, The Annals of Statistics, 15(3), pp.1174-1207.

[53] SMITH, R.L., (1989), Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone, Statistical Science,4(4), pp.367-377.

[54] SMITH, R. L., (2003), Statistics of Extremes, with applications in environment, insurance and finance, http:// www.stat.unc.edu/postscript/rs/semstatrls.ps, pp.8-18.

[55] STEPHENSON, A. G., & RIBATET, M., (2006), A user's guide to the evdbayes package (version 1.1).

[56] SUGAHARA, S., DA ROCHA R.P., & SILVEIRA R., (2009), *Non-stationary frequency analysis of extreme daily rainfall in Sao Paulo, Brazil. International Journal of Climatology*, 29, pp. 1339 - 1349.

[57] SUN, Z., & LI L., (2014), *Potential capability estimation for real time electricity demand response of sustainable manufacturing systems using Markov-Desicion Process, Journal of Cleaner Production*, 65, pp. 184-193.

[58] SUVEGES, M., & DAVISON, A. C., (2010), Model misspecification in Peaks Over Threshold analysis, The Annals of Applied Statistics, 4(1), 203-221.

[59] TANCREDI, A., ANDERSON, C., & O' HAGAN, A., (2006), Accounting for threshold uncertainty in extreme value estimation. Extremes, 9(2), pp. 87-106.

© University of Venda

University of Venda
Creating Future Leaders

[60] TODOROVIC, P., & ZELENHASIC, E., (1970), A stochastic model for Flood Analysis, Water Resources Research, 6(6), pp. 1641-1648.

[61] WANG, Y., (2011), Smoothing Splines: Methods and Applications. Chapman and Hall/CRC, pp. 354.

[62] WEISSMAN, I., (1978, JASA), *Estimation of parameters and large quantiles based on the k largest observations, Journal of the American Statistical Association*, 73, pp. 812-815.

[63] WIDEN, J., & WACKELGARD, E., (2010), A high-resolution stochastic model of domestic activity patterns and electricity demand, Applied Energy, 87, pp.1880-1892.

[64] WU, Z., HUANG, N. E., LONG, S. R., & PENG, C-K., (2007), On the trend, detrending, and variability of non-linear and non-stationary time series, Proceedings of the National Academy of Sciences, 104(38), pp.14889-14894.

[65] ZH ANG, H., WANG, Z., GAO, F., & ZHOU, D., (2012), A New Method for the Daily Electricity Consumption Forecast of Energy Intensive Enterprise, In Power and Energy Engineering Conference (APPEEC), Asia-Pacific, pp. 1-4, IEEE.

[66] ZHOU ET AL., (2004), Evidence for a significant urbanization effects on climate in China, Proceedings of the National Academy of Sciences, 101(26), pp. 9540-9544.

© University of Venda

6. Appendix

6.1 Appendix A

The plot of the daily peak electricity demand in mega watts against the cluster maxima of the non-linear detrended autumn data is shown in Fiigure 6.1 below where the autumn graph is superimposed on the non-linear detrended graph.

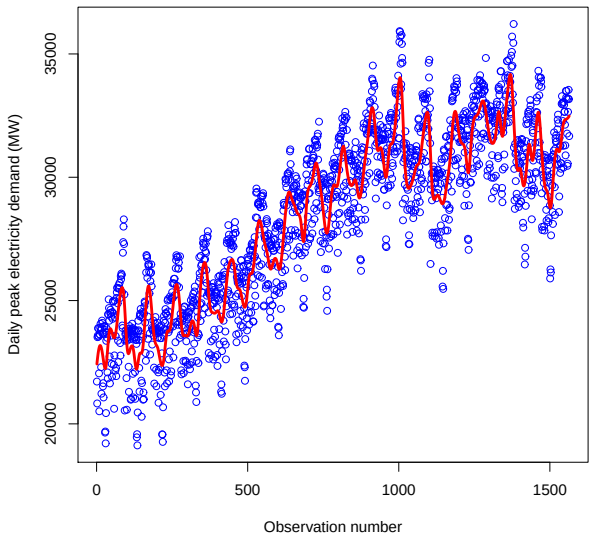

Figure 6.1: Plot of the DPED against the cluster maxima of the non-linear detrended autumn data.

© University of Venda

University of Venda
Creating Future Leaders

6.2 Appendix B

We display in Figure 6.2 the plot of the daily peak electricity demand in mega watts against the cluster maxima of the non-linear detrended spring data. The spring graph is superimposed on the non-linear detrended graph.

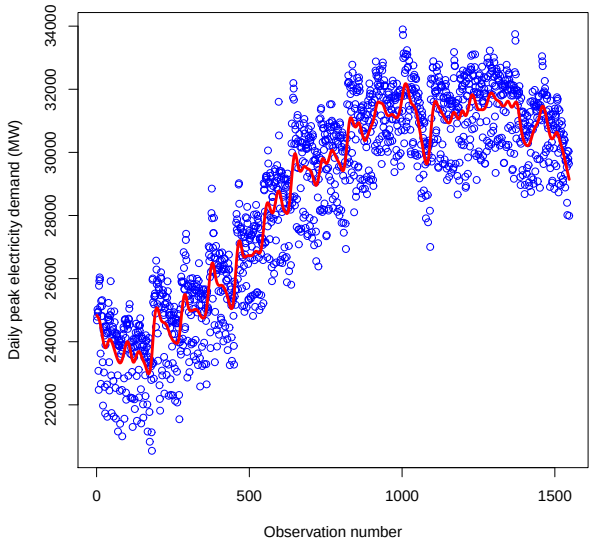

Figure 6.2: Plot of the DPED against the cluster maxima of non-linear detrended spring data.

© University of Venda

University of Venda
Creating Future Leaders

6.3 Appendix C

Figure 6.3 shows the gragh of the daily peak electricity demand in mega watts against the cluster maxima of the non-linear detrended summer data where the summer graph is superimposed on the non-linear detrended graph.

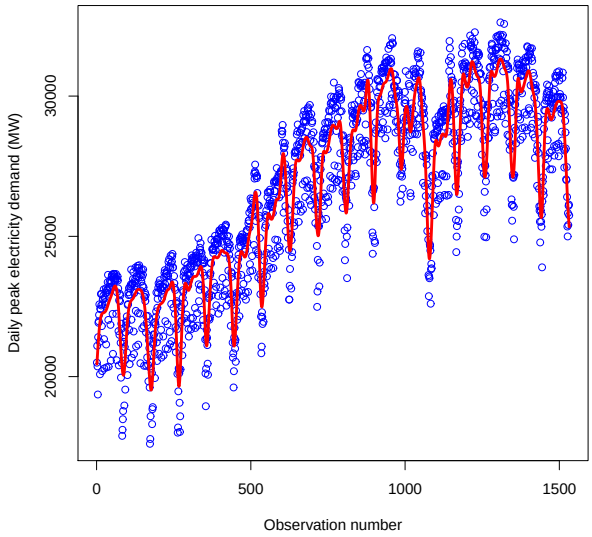

Figure 6.3: Plot of the DPED against the cluster maxima of non-linear detrended summer data.

© University of Venda

University of Venda
Creating Future Leaders

6.4 Appendix D

We graph the daily peak electricity demand in mega watts plotted against the cluster maxima of the non-linear detrended winter data in Figure 6.4.

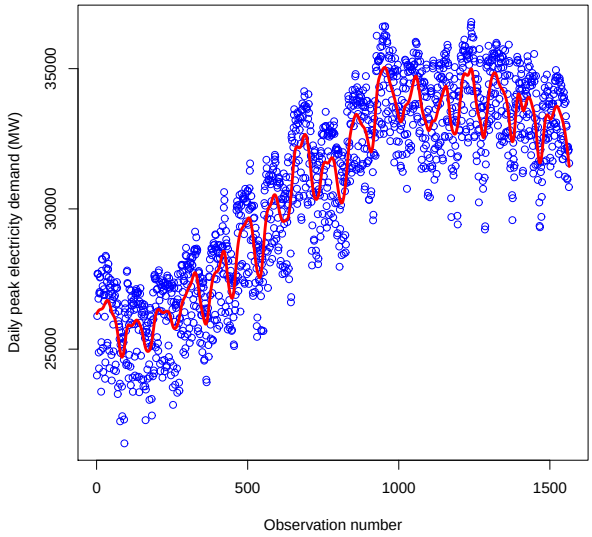

Figure 6.4: Plot of the DPED against the cluster maxima of non-linear detrended winter data.

© University of Venda

University of Venda
Creating Future Leaders

6.5 Appendix E

Figure 6.5 shows the plots of the positive residuals and the density plot of the non-linear detrended winter data.

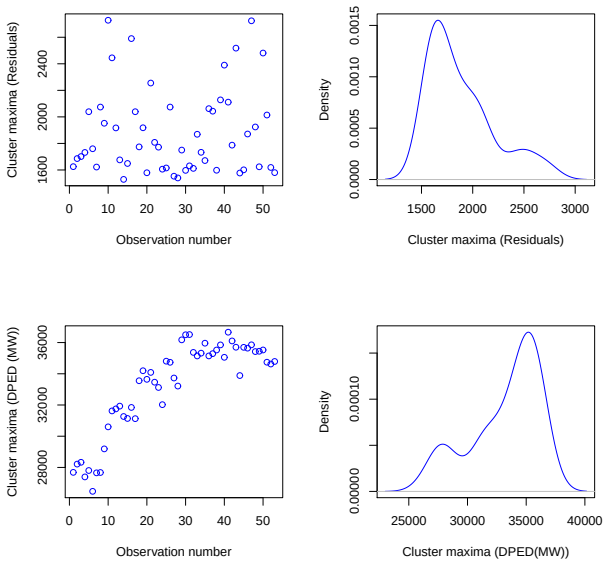

Figure 6.5: Plot of the positive residuals and density plot of the non-linear detrended winter data.

© University of Venda

6.6 Appendix F: Some R codes

```
EXPLORATORY DATA ANALYSIS

attach(windata)

head(windata)

win.graph()

par(mfrow=c(2,2))

x=DPED

plot(x, type="p", ylab="Winter daily peak electricity demand (MW)",col="blue",

     xlab="Observation number", main="(a) Winter DPED")

plot(density(x), col="blue",

main="(b) Density", xlab="Winter daily peak electricity demand (MW) ")

qqnorm(x,col="blue",main="(c) Normal QQ plot")

qqline(x)

boxplot(x,col="blue",main="(d) Box plot")
```

© University of Venda

```
NONLINEAR DETRENDING

attach(windata)

head(windata)

win.graph()

x=DPED

plot(x, type="p", ylab="Daily peak electricity demand (MW)",col="blue",

     xlab="Observation number")

r=smooth.spline(time(x), x)

r

lines(smooth.spline(time(x), x, spar= 0.3076149), lwd=3,col="red")

r2=residuals((smooth.spline(time(DPED), DPED, spar= 0.3076149)))

write.table(r2,"~/residwindata.txt",sep="\t")

NOLTR1
attach(windata)
head(windata)
win.graph()
plot(residwindata)
```

© University of Venda

```
EXTREMAL MIXTURE MODELS

library(evmix)

set.seed(1234)

a=residwindata[residwindata>0]

win.graph()

plot(a)

fit = fkdengpd(a, phiu = FALSE, std.err = FALSE)

hist(a,breaks=100, freq = FALSE, "Positive detrended SDPED (MW)",xlim = c(0,3500))

aa = seq(0, 2500, 50)

lines(aa, dkdengpd(aa, a, fit$lambda, fit$u, fit$sigmau, fit$xi, fit$phiu), col="blue", lwd

abline(v = fit$u, col="blue", lwd = 2)

legend("topright", "kdengpd", col = "blue",lty = 1, lwd = 2)

box()

fit
```

© University of Venda

University of Venda
Creating Future Leaders


```
FITTING GPD TO DECLUSTERED DATA

library(texmex)

palette(c("black", "purple", "cyan", "orange"))

set.seed(20120118)

r2pos = a

plot(r2pos)

library(texmex)

palette(c("black", "purple", "cyan", "orange"))

set.seed(20120118)

decluster the sequence by using the automatic declustering method

win.graph()

ei <- extremalIndex(r2pos, threshold = 1516)

ei

dc <- declust(ei)

par(mfrow=c(1,1))

plot(dc,col="blue", xlab="Observation number", ylab="Positive residuals")
```

© University of Venda

```
dc

dc <- declust(r2pos, threshold = 1516)

SENSITIVITY ANALYSIS

par(mfrow = c(2, 2))

ei <- extremalIndex(r2pos, threshold = 1516)

plot(ei,col="blue")

ei <- extremalIndex(r2pos, threshold = 1500)

plot(ei,col="blue")

ei <- extremalIndex(r2pos, threshold = 1550)

plot(ei,col="blue")

ei <- extremalIndex(r2pos, threshold = 1600)

plot(ei,col="blue")

FITTING THE GENERALIZED PARETO DISTRIBUTION TO CLUSTER MAXIMA

resid.gpd <- evm(dc)

resid.gpd

par(mfrow = c(2, 2))
```

© University of Venda

```
plot(resid.gpd)

GOODNESS OF FIT OF THE GPD USING THE ANDERSON-DARLING TEST

library(eva)

win.graph()

attach(cmaxDifAlldata)

head(cmaxDifAlldata)

Fit data

mle_fit <- gpdFit(cmax, threshold = 2718, method = "mle")

mle_fit

plot(mle_fit)

Goodness of fit tests A & D, Cramwer-von Mises,

set.seed(1234)

x <- rgpd(1500, loc = 0, scale = 357.3291, shape =  -0.1668)

gpdAd(x)

CRAMER-VON MISES GOODNESS-OF-FIT TEST FOR THE GENERALIZED PARETO (GPD) DISTRIBUTION

gpdCvm(x)
```

© University of Venda

```
GPD PARAMETER UNCERTAINTY: PARAMETRIC BOOTSRAP

boot <- evmBoot(evm(dc), trace = 1001)

summary(boot)

par(mfrow = c(1,2))

plot(boot)

RETURN LEVEL ESTIMATION

win.graph()

portRL <- predict(resid.gpd, M = seq(20, 100, by = 10), ci.fit = TRUE)

plot(portRL, main="Predicted return levels with 95% predictive intervals")

Prediction Intervals for return levels

portRL[c(1, 10, 20)]

IDENTIFY CLUSTERS OF EXCEEDANCES USING EVD

attach(windata)

head(windata)

win.graph()

library(evd)
```

© University of Venda

```
c = clusters(r2pos,1516, 1, cmax = TRUE)

write.table(c,"~/clusterspp.txt",sep="\t")

FITTING STATIONARY POINT PROCESS TO CLUSTER MAXIMA USING ISMEV

attach(clustermax)

head(clustermax)

win.graph()

library(ismev)

z=pp.fit(cmax,1516)

pp.diag(z)

y=gpd.fit(cmax,1489)

Fitting the GPD to cluster maxima using ismev

gpd.diag(y)

attach(wintercmax)

head(wintercmax)

win.graph()

par(mfrow=c(2,2))
```

© University of Venda

```
plot(cmax,col="blue",xlab="Observation number", ylab="Cluster maxima (Residuals)",main="")

plot(density(cmax),col="blue",xlab="Cluster maxima (Residuals)",main="")

plot(DPED,col="blue",xlab="Observation number", ylab="Cluster maxima (DPED (MW))",main="")

plot(density(DPED),col="blue",xlab="Cluster maxima (DPED(MW))",main="")
```

© University of Venda