Extreme Value Analysis for Air Pollution Data with Non-Gaussian State Space Modeling

Utashi Fujimoto

School of Mathematics and Statistics
University of Sheffield

The University Of Sheffield.

Dissertation submitted as part of the requirements for the award of
MSc in Statistics, University of Sheffield, 2020–2021

Contents

Lay Summary of the Dissertation

Due to the growing evidences of anthropogenic climate change, the issues of air pollution has gained an increasing attention over the last decades. The effective and scientific management of air pollution therefore became the subject of a public interest, and various applications of mathematical modeling have been discussed in literature. Particularly, the statistical modeling is a popular class of mathematical modeling, allowing us to investigate the underlying dynamics of natural and social phenomena. Konishi and Kitagawa (2008) suggests three purposes of statistical modeling: description of stochastic structures, extraction of information, and predictions. The main advantage of statistical modeling is that all of these aspects can be discussed on the basis of probability, which is generally easier for a human to understand.

The ordinary methods of statistical modeling, however, may not be sufficient to meet certain demands in practice. For example, when analysing environmental data, it is oftentimes more important to focus on the extreme situations rather than the ordinary. The reason for this may be that only extreme instances are relevant for policy makers, who would like to set environmental targets required to be met at the bare minimum. Extreme Value Theory, the unique branch of statistics focusing on statistically rare events, can provide a useful modeling framework in such a context. The main objective of the dissertation is to examine the utility of the theory in analysing air pollution data. Though the theory provides several different modeling approaches, throughout the analysis we focus on the two most common models: block maxima model and peaks-over threshold (POT) model. These models are also referred to as generalised extreme value (GEV) model and general Pareto (GP) model, respectively, when the emphasis is on the probabilistic aspect.

When applying these models to real-world data, it is essential to take care of the fact that the process of data generation may not be homogeneous over time, known as the non-stationality in statistical terminology. In fact, this is true of the data used in the analysis, and ignoring this would violate the model assumption and thus significantly reduce the reliability of results. The dissertation discusses two different approaches in tending to this problem. The earlier chapters discuss the classic method where the data is transformed to ensure the model assumption. The result indicates reasonable success of modeling, if not satisfactory. The issue here is that this approach comes at the cost of decreased

interpretability due to the transformations. The later chapters discuss more advanced modeling method under the framework of Bayesian statistics, which is another theory of statistics based on the Bayesian interpretation of probability. This approach enables a more complex structure of the model, obviating the necessity of data transformation. The results are satisfactory in that the models considerably improves its performance in explaining the data and that the interpretation of results are quite straightforward.

Chapter 1

Introduction

1.1 Origin and History of Extreme Value Theory

The pioneer of extreme value theory is said to be Leonard Tippet, whose interest was to strengthen cotton thread as he worked at the British Cotton Industry Research Association. During his research, he discovered that the strength of thread is ultimately determined by its weakest component or fiber - what is now called weakest link principle. Unlike the traditional discipline of statistics, extreme value theory seeks to quantify the stochastic behavior of a system at its unusually large or small incidents (e.g. maximum or minimum values observed during the given period of time), rather than its mean or medium. Inspired by Tippet's work, Gumbel (1958) confirmed the methodologies applicable to various kinds of engineering problems, whose work was then considered to be the most reliable reference for quite some time.

The theoretical development had to wait for a while until it gained sufficient attention. According to Beirlant et al. (2005), the 1970 doctoral dissertation by L. de Haan was the first piece of work which properly investigated the probabilistic and stochastic property of extreme values, in a manner analogous to the central limit theorem. The substantial groundwork on the theory has been accomplished by early 2000s leaving only several highly technical issues to be resolved. Coles (2001) provides an excellent summary of the relevant works completed in the past along with their key results, also laying out versatile frameworks for analyzing extreme values. The reviews of relevant theorems shown in section 3.1 and 4.1 refer to the authour's work.

As implied above, early application of extreme value theory was mainly in engineering, but it was soon applied to even wider range of fields including finance, environmental science and telecommunication, etc. As of late, researcher's interest seems to be centered around its applicability to more modern topics such as image recognition, anomaly detection and climate change, etc.

1.2 Main Objective

The ultimate objective of the project is to apply the analytical framework provided by the theory to the air pollution data and assess the performance of each method, while investigating the possible biases that may arise in practice. The use of the theory is often motivated by the fact that only extreme instances are relevant for policy makers, who would like to set environmental targets required to be met at the bare minimum. To be clear, the primary interest is not necessarily to gain new insights regarding the air pollution in the UK, but in the applicability of the theory to analysing such data. (However, any discovery regarding the topic that merits a mention will also be discussed along the way.) To this end, following several analytical targets are set for the project:

Firstly, find a model that explains (or fits) data well. This is investigated mainly by a set of diagnostic plots conventionally used in the theory, with slight modifications as required for more complex modeling method. Particularly, return level plot provides an important insight for data, presenting the theoretical magnitude of anomalies in given return period, which can then be compared with the realisations associated with the same period.

Secondly, find the distribution of data with estimates of relevant parameters. The diagnostic plots previously mentioned are established visualisation techniques and quite an informative tools for checking model fit. Generally, however, evaluating any statistical models entails the discussion of parameter inference. Particularly, in the context of Extreme Value Analysis (EVA), a shape parameter, often denoted as ξ, is a primary interest of statistical inference. This is because the sign of this parameter determines the family of distribution and thus the tail behaviour of such distribution. In addition, for models that assume time-varying parameters, identifying the trend of data is a relevant task here. Interpretation of parameters will be included in each chapter as required. Further, quantification of uncertainties with respect to each of these parameter estimates is of great interest. This can be achieved either by the conventional methods utilising the asymptotic property or by alternative methods used in Bayesian inference.

Lastly, predictions of pollution levels will also be made when possible. This can be in some sense done via the return level plots mentioned above. More complex models discussed in later chapters can predict the observations with quantified uncertainties associated with it. The result of these predictions may be better understood in the context of criteria provided by The Air Quality Strategy for England, Scotland, Wales and Northern Ireland (`https://www.gov.uk/government/publications`), which sets out air quality objectives policy options to further improve air quality in the UK. The latest executive summary reports that the current objectives for all air pollutants in over 99 percent of UK are met. Discussing the results obtained through this project in comparison to this claim may be of interest although this is again not the primary objective of the analysis.

1.3 Structure of the Dissertation

First, the result of basic exploratory data analysis (EDA) will be provided in chapter 2. In EVA, as often the case with any statistical discipline, a construction of a model is predicated on the assumption that the random variables are independently and identically distributed (i.i.d), which would render the validity of the models questionable if violated. In fact, this assumption is rarely realistic in practice, and a simple EDA is sufficient to reveal that our data is no exception. In this context, we attempt to specify the characteristics of data that needs to be addressed prior to the statistical modeling.

There is a set of modeling methods developed in Extreme Value Theory that are often used in practice. This includes modeling a distribution of maxima observed in a particular given interval, known as block maxima model. The choice of length for this interval can be arbitrary, but monthly or annual maxima is often used in practice. This modelling method is discussed in Chapter 3. Another common method is Peak-Over Threshold (POT) model, which sets a particular threshold and models the distribution of exceedances. This model is discussed in Chapter 4. Note that some of the issues caused by the false i.i.d assumption can be resolved by relatively simple modifications (e.g. differencing or declustering), after which these classic modeling methods can be implemented. However, this approach is often not entirely satisfactory, especially if there is residual non-stationarity. That is, a more sophisticated modelling approach would be required if one wishes to overcome this issue successfully.

In Chapter 5, Bayesian approach is discussed. This chapter is essentially designed to provide a nice segue to Chapter 6, clarifying the preliminary settings necessary for the advanced modeling method discussed there. Particularly, conducting HMC sampling (with Stan) for the block maxima or POT model requires an extra procedure of implementing the user-defined function. Further, Bayesian inference on its own light is quite beneficial in analysing the data since it gives us another way to quantify the bias associated with parameter inference, which is relatively easier to interpret. This aspect of Bayesian inference is also discussed in this chapter. Finally, in Chapter 6, state space model is discussed. The obvious benefit of the method is a grater flexibility in expressing the model, which can even describe the stochastic process of non stationary data if successfully implemented.

Each chapter begins with a brief summary of theoretical background as required, followed by the result of model implementations. This includes model diagnostics, estimate of parameters and other relevant method-specific topics.

1.4 About Data

The data is obtained from Automatic Urban and Rural Network (AURN), the UK's largest automatic network used for compliance reporting against the Am-

bient Air Quality Directives, according to the UK government website (`https://uk-air.defra.gov.uk/networks/network-info?view=aurn`). The network consists of automatic air quality monitoring stations situated in more than 250 locations across the UK, each measuring the concentration levels of wide variations of pollutants. The network thus provides an abundant source of information regarding the air pollution in the UK.

The openair project received funding from the UK National Environment Research Council (NERC) over 10 years ago, aiming to provide a dedicated set of easily accessible, open-source tools for analysing air quality data. The dataset used in the project can be downloaded via the R package openair. At the time of writing, the data is available from February 1973 to September 2021 although some of the stations could have come in to operation after the others. The type of pollutant recorded differs depending on locations, while some information (wind speed, wind direction, air temperature, etc.) is recorded for all the locations. The data also contains missing values, which needs to be treated carefully to decrease the bias for later statistical inference.

For the purpose of the project, we focus on two pollutants i.e. sulfur dioxide (SO_2) and ozone (O_3), observed at the single location London N. Kensington. This particular choice of pollutants is based on the observation that the two pollutant appear to well represent the common characteristic of the data seen among the recorded pollutants. This will be briefly discussed in Chapter 2 as well.

Chapter 2

Exploratory Data Analysis

2.1 Selection of pollutants

This chapter conducts the basic EDA prior to discussing the statistical modeling. Given the objective of analysis, it is sensible to focus on the small subset of data even though there are 8 different pollutants recorded at the chosen location. The following discussion will be thus focused on SO_2 and O_3. The reason for this choice is that these two pollutants adequately illustrate the characteristics of the data crucial in later modeling: trend, seasonality and variation in variance. These will soon be discussed in the following section in more detail.

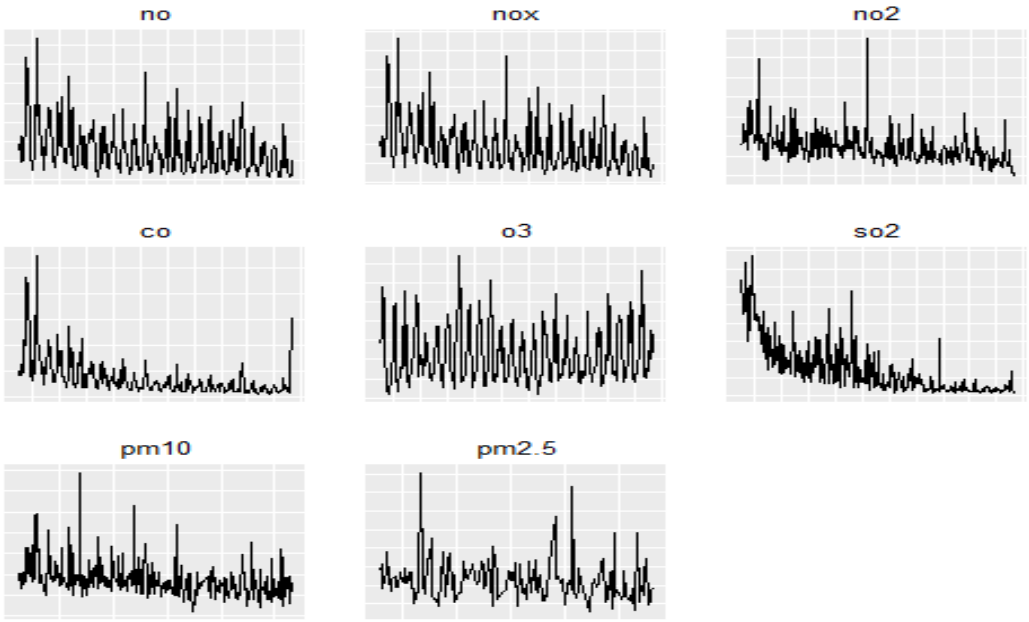


Figure 2.1: The time series plot of monthly maxima for all the pollutants recorded at the station in London N. Kensington. A careful observation concludes that SO_2 and O_3 are effective choice for discussing the modeling with.

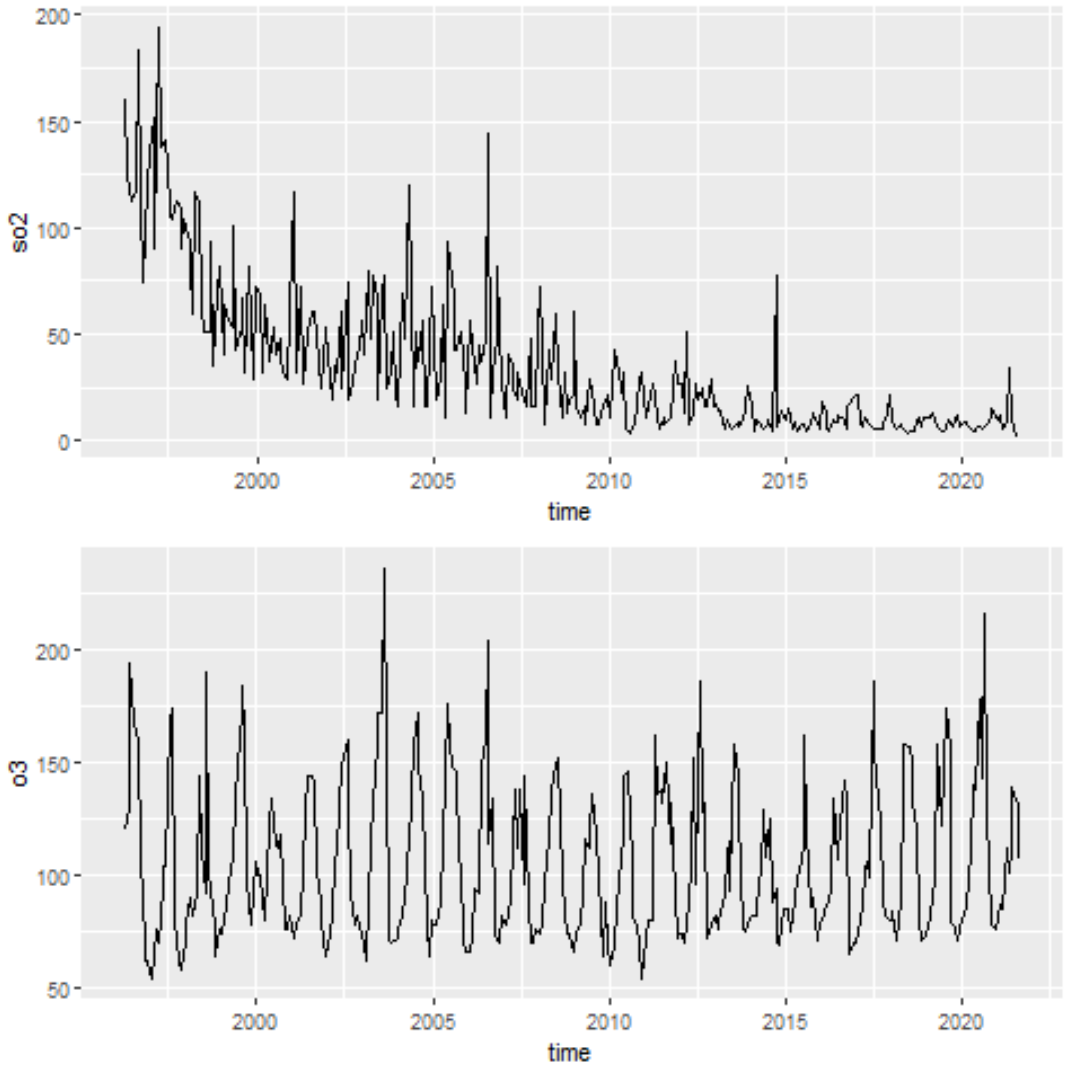


Figure 2.2: The time series plot of monthly maxima for SO_2 and O_3, respectively. Though not shown here, the original time series contains more than 200,000 observations of the air concentration level of pollutants recorded hourly every day over 25 years.

2.2 Air pollution data and i.i.d assumption

The data is time series, meaning that each record is associated with a particular point in time axis. Note that the following discussions consider the different time series depending on the modeling approach of interest; a model may be fitted to the original time series or to another distinct time series of monthly maxima, extracted from the original. Additionally, further transformations will be implemented to deal with certain issues, which will later be discussed in more detail as requird.

Figure 2.2 shows the time series plot of monthly maxima for the chosen pollutants i.e. SO_2 and O_3. Importantly, neither of these series meets the i.i.d.

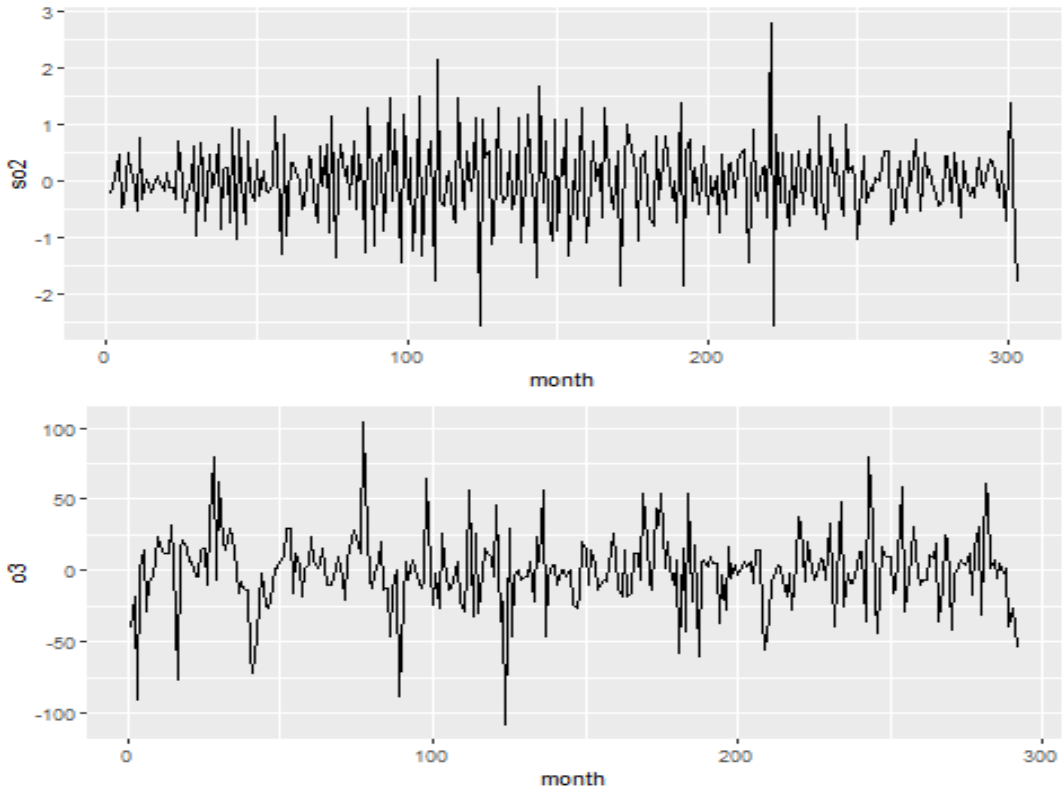


Figure 2.3: The transformed time series of monthly maxima for SO_2 and O_3, respectively. For SO_2, the logarithmic transformation is used to stabilise the variance of the series, after which the 1st order differencing is applied to remove the trend. As for O_3, the data is differenced with lag = 12, which appears to removes the seasonality of the original time series to some degree.

assumption; for SO_2, there is a visible down-ward trend over time, and the structure of variance appears to be time-variant. As for O_3, there appear to be a cyclic pattern likely to be caused by seasonal effect although there is no overwhelming evidence for any trend over time. The issue here is that the EVA framework is based on i.i.d. assumption, as will be discussed in following chapters in more detail. In other words, the effective analysis of extreme values involves taking care of the fact the the assumption is not met.

2.3 Box-Jenkins framework

It is of our interest to consider Box-Jenkins framework, though it is not particularly designed to analyse extreme values, as it provides some useful methodologies to deal with the non-i.i.d. time series. The idea of the framework is that the data will first be transformed in such way that the resulting series is said to be stationary. Then the set of conventional models will be fitted to this series to carry out usual parameter inference and predictions. Figure 2.3 shows the time series of monthly maxima for SO_2 and O_3 after appropriate transformations.

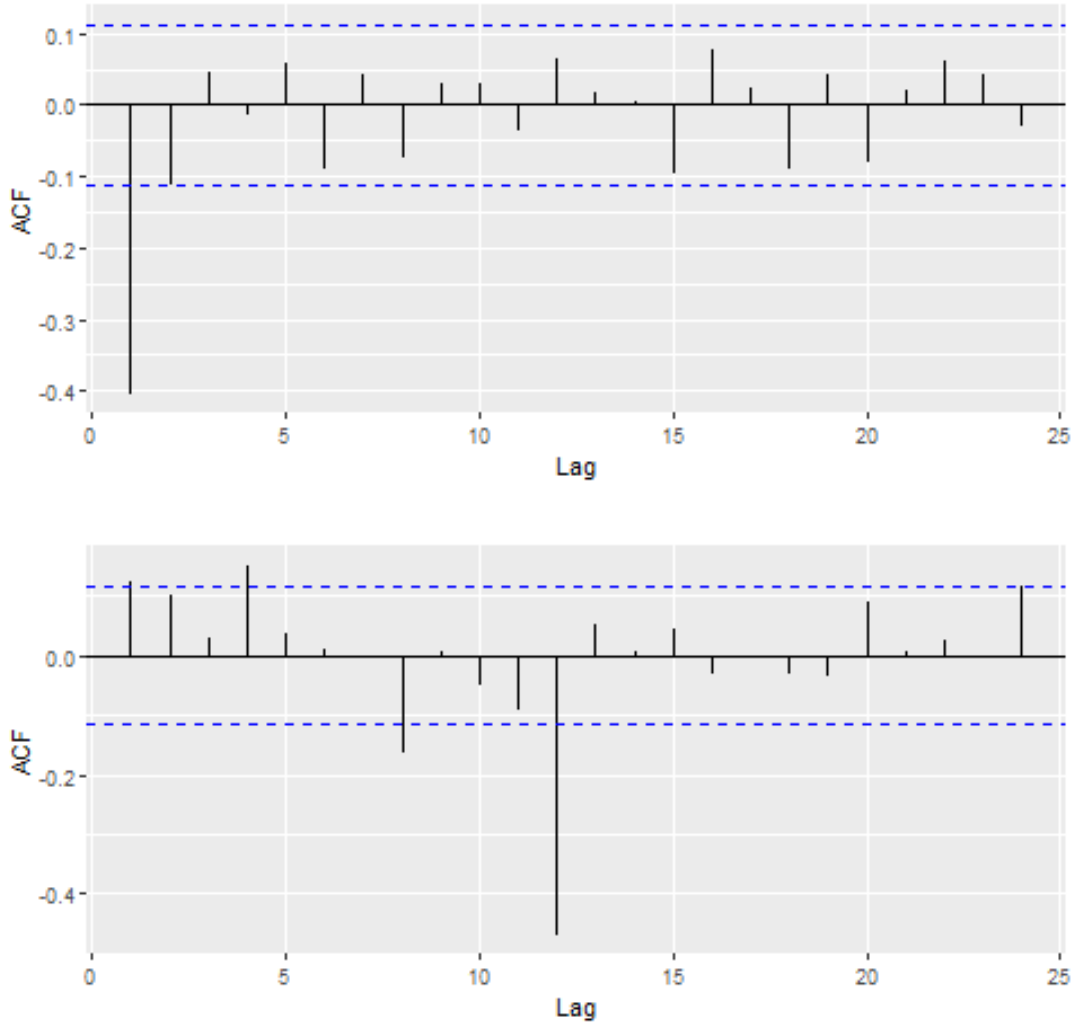


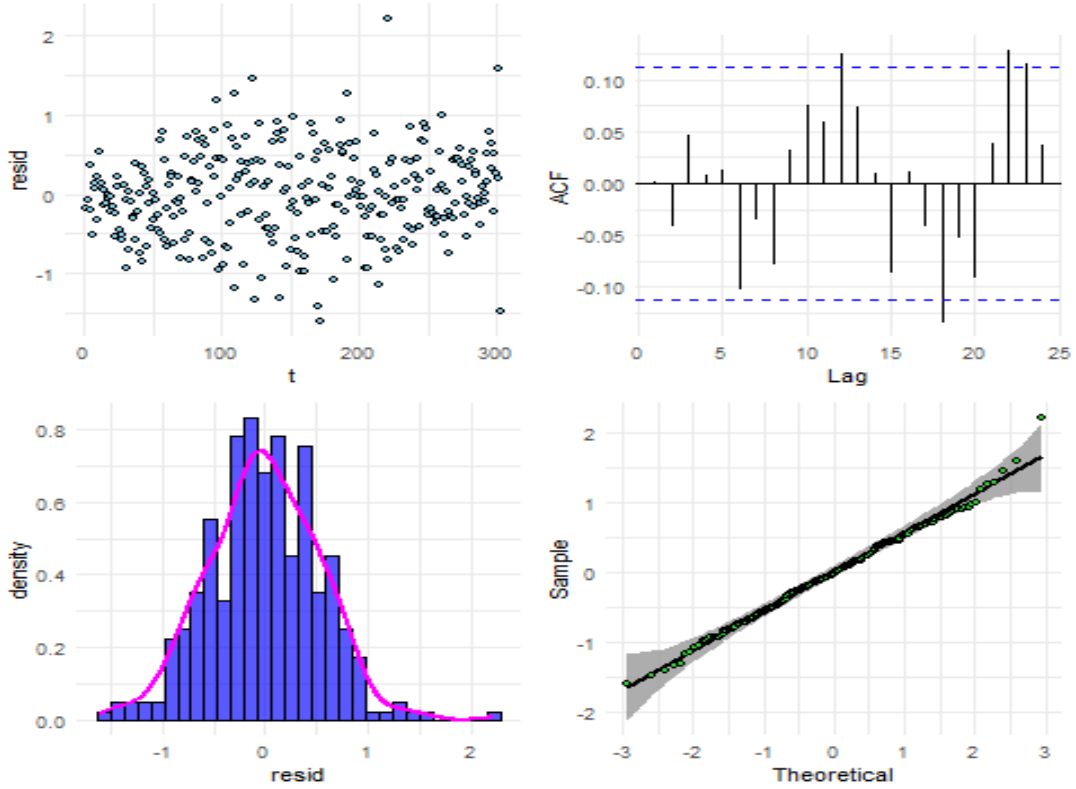
Figure 2.4: ACF for the transformed time series of monthly maxima for SO_2 and O_3, respectively. The plot indicates the significant auto-correlation at lag = 1 for SO_2 and lag = {4, 8, 12} for O_3.

Note, as can be seen in Figure 2.4, that neither of these transformed series can be considered as complete i.i.d. process. However, as explained, we can assume that these transformed data are now stationary, and then the classic modeling methods of EVA can be applied although minor modifications are necessary to allow for the correlation of the observations. This is in fact the main approach discussed in Chapter 3 and 4.

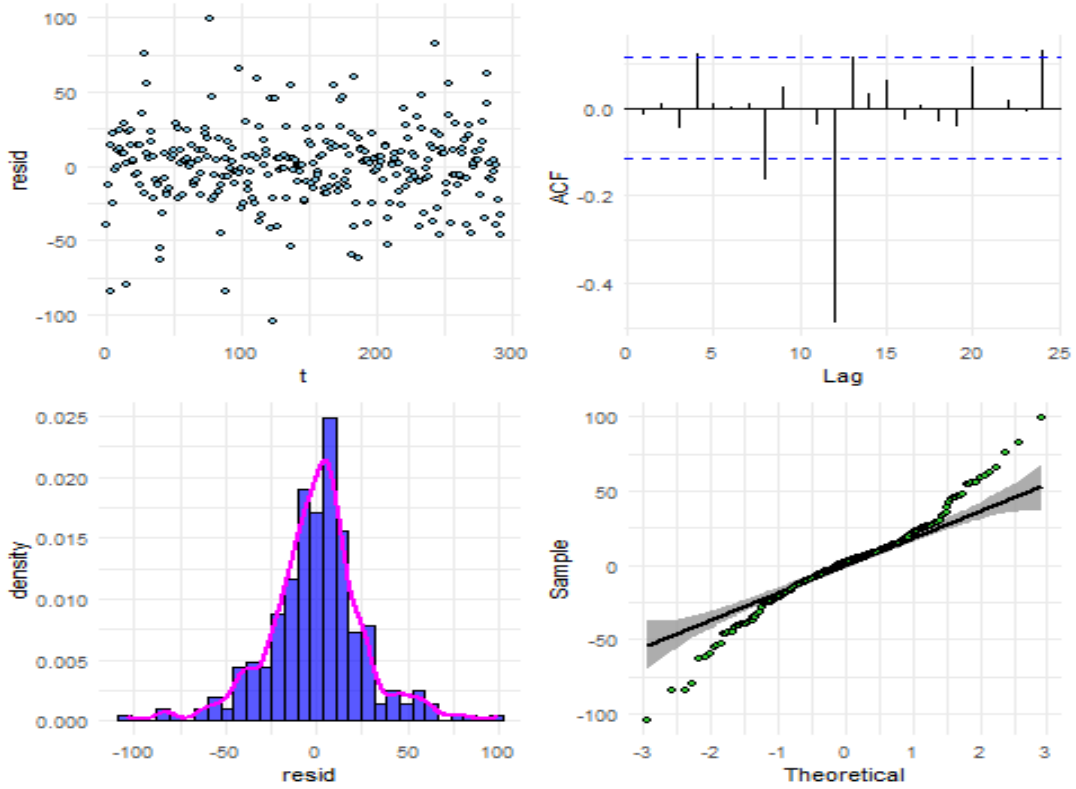
It may be of our interest to carry on with the framework and implement its usual model as a naive approach even though this is outside the context of EVA. In R, *auto.arima()* function is used to suggest the optimal model and estimate its parameter though the detail is omitted here. The diagnostic plots for error analysis is shown in Figure 2.5. For SO_2, surprisingly, these diagnostic

plots indicate the good fit of the model, except that there appears to be a visible structure in its residuals. This is possibly because the logarithmic transformation of the series successfully ensures the normality of residuals, which is generally assumed in the ARMA model. For O_3, the remaining seasonality seems to affect the performance of the model, and more importantly, the residuals are not normally distributed.

Besides the inadequate model fit, the main issue here is that after multiple transformations, the distribution of the original data is not quite tractable (the inverse-transformation of logarithm is quite straightforward, but the differencing requires an extra care), which is not ideal as this undermines one of our end-goals. Further, ARMA model, with its simple underlying distributional assumption, does not provide any insight regarding the stochastic behaviour of extreme values. These certainly give us the motivation to improve our approach, given the end-goals outlined in the previous chapter. In following chapters, a series of modeling methods developed in the context of EVA will be discussed.



(a) SO_2



(b) O_3

Figure 2.5: Error Analysis for ARIMA model

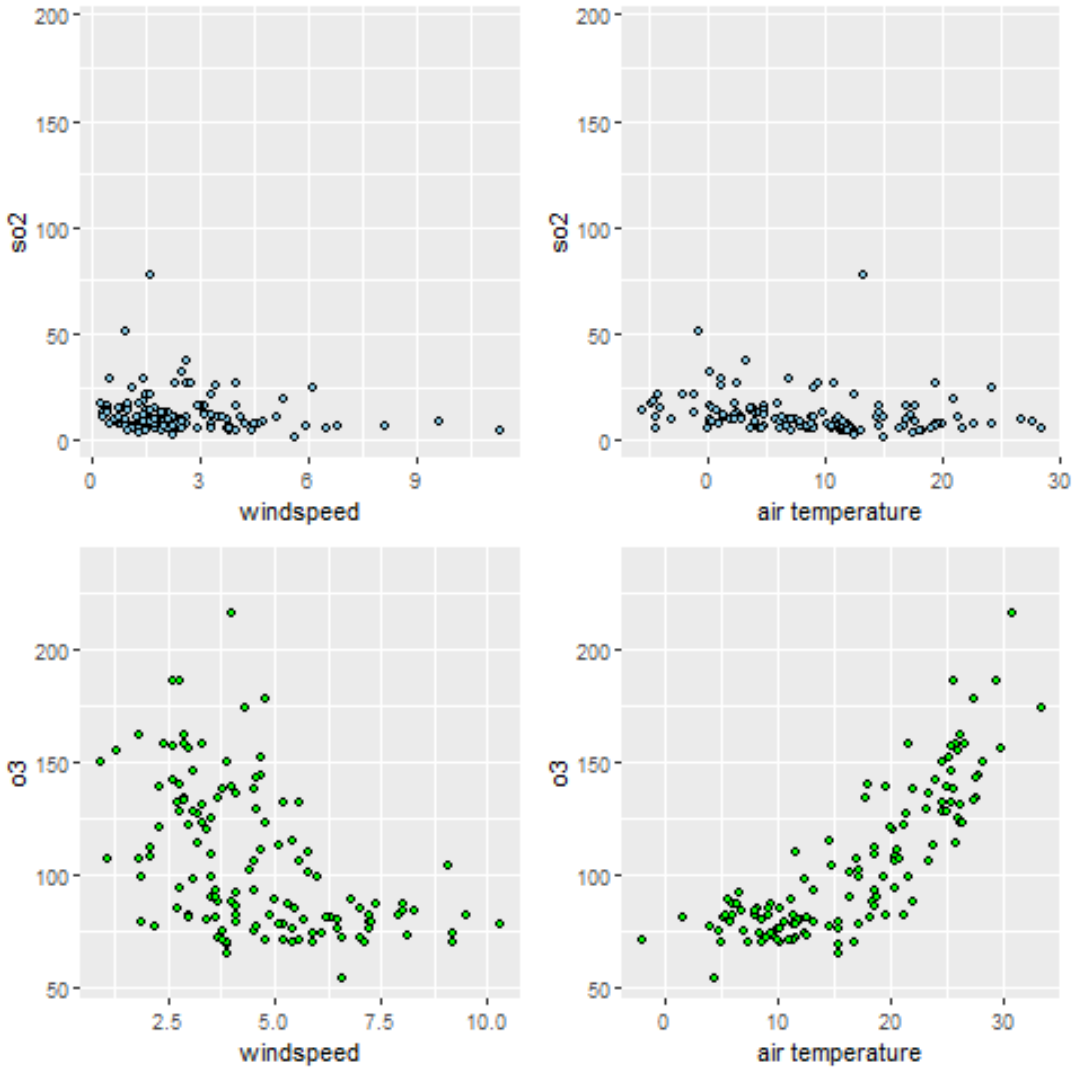


Figure 2.6: The two covariates (windspeed and air temperature) plotted against the observed values of SO_2 and O_3, respectively. It is known that wind has a dispersing effect for pollutants, resulting in smaller values of the target variables with higher wind speed. Additionally, there is a strongly positive correlation between air temperature and the concentration level for O_3, which is consistent with the cyclic pattern observed in Figure 2.2. This is not true, however, for SO_2, where increasing temperature seems to lower the value ever so slightly.

2.4 Covariates

As mentioned above, besides the concentration level of each pollutants (i.e. our target variable), the data set also contains a set of information as covariates. In Figure 2.6, the values of SO_2 and O_3 are plotted against the relevant variables.

Clearly, there are notable relations between them, suggesting that these covariates could be included in our models, if necessary. In fact, in Chapter 6, air temperature is used as covariate in an attempt to identify the seasonal compo-

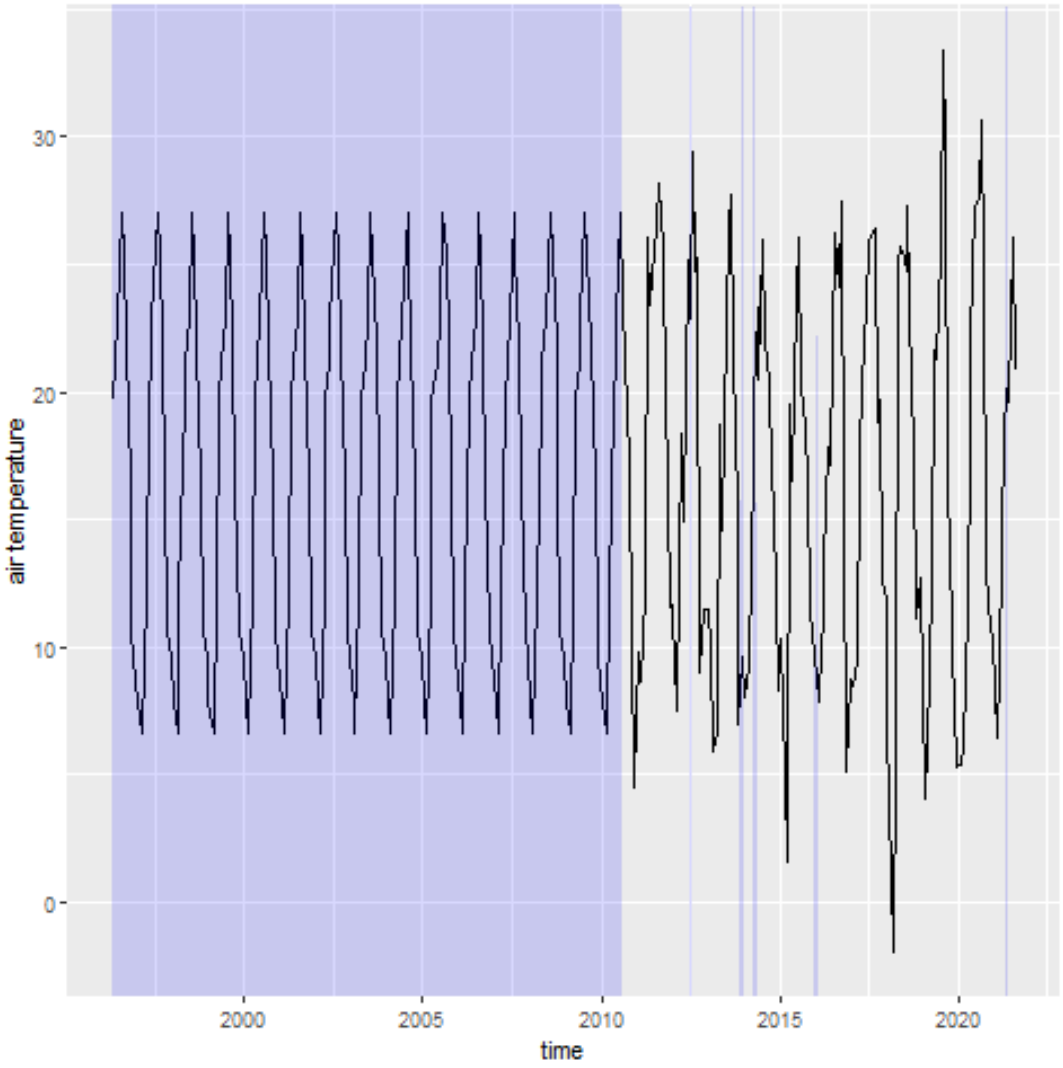


Figure 2.7: The missing observations are shown in the blue shaded region, which is quite large; out of the 304 observations there are 177 missing values. Due to the use of naive imputation method, variance of the data appears to be underestimated. The imputation nevertheless appears to reproduce the cyclic pattern of the data which is may be of use in explaining the seasonal effect.

nent of the stochastic process into the model.

The issue here, however, is that the air temperature contains quite a few missing values; the air temperature has not been recorded until August of 2010, and there are a few missing values afterwards. While there are several different methods to handle missing values, the simplest is to impute them with mean. Figure 2.7 shows the time plot of air temperature with missing values imputed with monthly mean. A more sophisticated measure such as multiple imputation should be implemented to address this issue although it is not considered here for the sake of simplicity.

Chapter 3

Block Maxima Model

3.1 Review of theory

The block maxima model is the method for analysing the statistical behavior of extreme values. By dividing the data into blocks, the model focuses on the maximum of a stochastic process measured on a regular time scale. It can be referred to as GEV model when the emphasis is on its underlying distribution.

We first briefly review the most important theorem when implementing the block maxima model. Consider,

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

where $X_1, ..., X_n$ is a sequence of i.i.d random variables whose common distribution function is denoted as F. Then, if there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that:

$$\Pr\left\{\frac{M_n - b_n}{a_n} \leq z\right\} \to \mathrm{G}(z) \quad \text{as n} \to \infty$$

for a non-degenerate distribution function G, then G is a member of the generalized extreme value (GEV) family of distribution

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

defined on $\{z : 1 + \xi(z - \mu)/\sigma > 0\}$, where $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$. The apparent difficulty is that since F is often unknown in practice, normalising terms $\{a_n > 0\}$ and $\{b_n\}$ are not known. However, assuming the statement above, it follows

$$\Pr\{M_n^* \leq z\} \approx \mathrm{G}(z) \quad \text{for large n,}$$

where $M_n^* = (M_n - b_n)/a_n$. Equivalently,

$$\begin{aligned}\Pr\{M_n \leq z\} &\approx \mathrm{G}(z) \\ &= \mathrm{G}^*(z),\end{aligned}$$

where G* is another member of the GEV family. In other words, the distribution of M_n itself can also be approximated by a different member of the same family. Since the parameter and its distributions are estimated anyway, not knowing the normalising constants is hardly an issue in practice.

3.2 Model specification

As shown above, the GEV distribution has three parameters to be estimated: location parameter μ, scale parameter σ and shape parameter ξ. As mentioned, ξ is of a particular interest as it determines the type of distribution and therefore the tail behavior of random variable. This chapter considers a simple GEV model

$$y_t \sim \mathrm{GEV}(\mu, \sigma, \xi) \quad \text{for t} = 1, ..., n.$$

Note that y_t represents an observation at point t of the monthly maxima time series. Recall first that the random variables are assumed to be i.i.d. as shown in section 3.1, and second that this assumption is clearly not met as discussed in Chapter 2. Thus, the above model is fitted to transformed monthly maxima time series, rather than the original. That is,

$$\begin{aligned}\text{for } \mathrm{SO}_2 \quad & y_t = \log x_t - \log x_{t-1}, \\ \text{for } \mathrm{O}_3 \quad & y_t = x_t - x_{t-12},\end{aligned}$$

where x_t represents a component of original monthly time series, respectively.

Technically, these treatments are not sufficient to ensure the i.i.d assumption. However, as shown in Coles (2001), provided that extreme events are close to independent at times that are far enough apart, the method of parameter inference is known to be practically unaffected by the stationarity of the data. Thus, the parameter estimates are obtained by maximizing the likelihood as usual.

3.3 Model diagnostics

In EVA in general, there is a set of diagnostic plots commonly used to investigate a goodness of the fit of model: quantile to quantile plot i.e. Q-Q plot, probability plot and return level plot, etc. Q-Q plot and probability plot are quite useful in that they can be constructed regardless of the underlying distribution, and

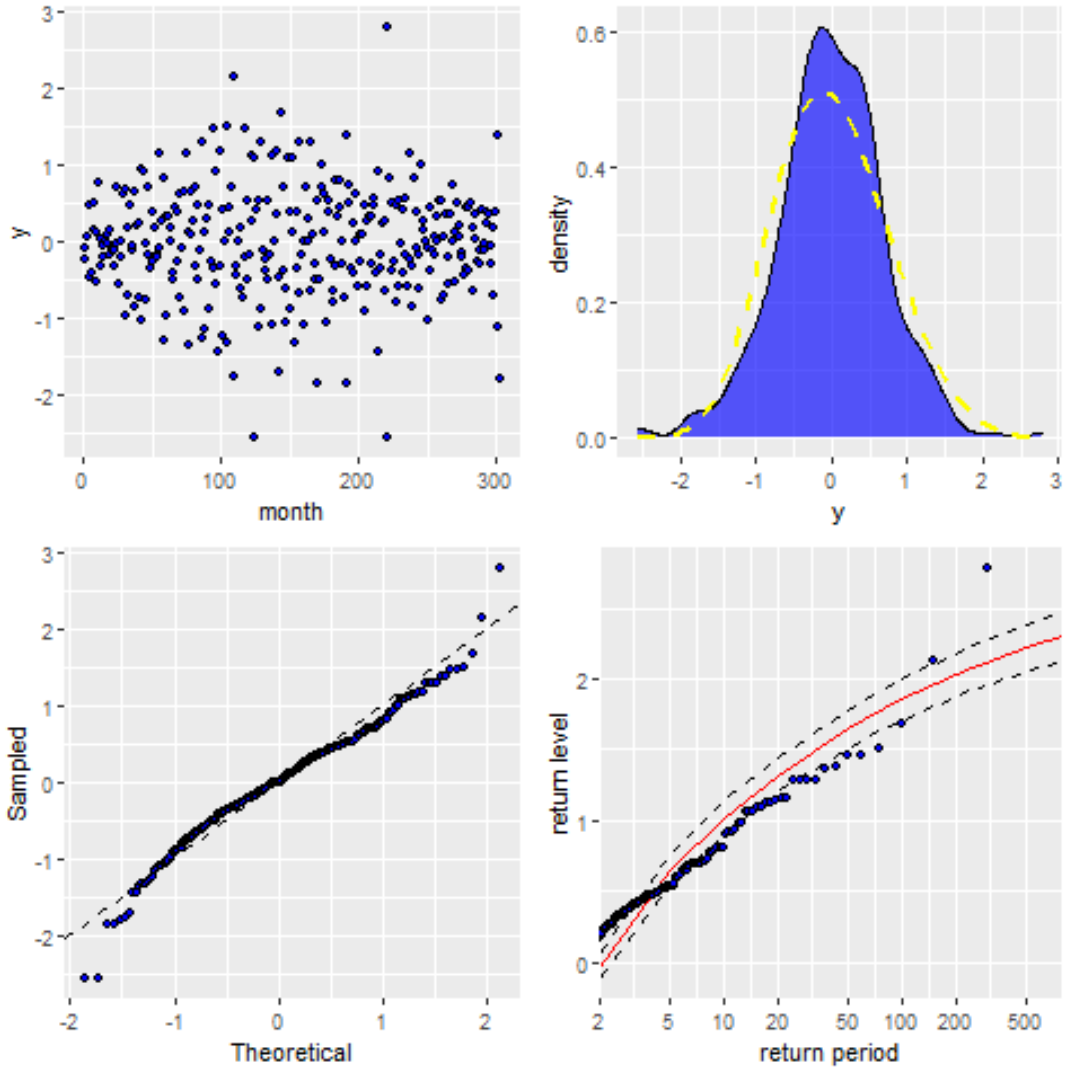


Figure 3.1: The diagnostic plots for the block maxima model fitted to transformed SO_2. The scatter plot can be used to check if there is any varying structure in the stochastic process. The rest of the plot can be used to visually assess the model fit.

its interpretation is incredibly simple; if the model fit is good, the observations should be seen aligned linearly on the plot. The return level plot, as mentioned, is a key visualising technique in EVA (see Appendix A for more detail). For the models, the diagnostic plot consists of 4 individual plots: scatter (time series) plot, density plot, Q-Q plot and return level plot.

Figure 3.1 shows the diagnostic plots for the block maxima model fitted to SO_2. The Q-Q plot and density plot both indicate a reasonably good fit of the model, except that the observations present heavier tails than suggested by the theoretical distribution. The return level plot, however, implies that the model does not exactly capture the behaviour of extreme values. Given $\hat{\xi} < 0$ (see 3.4 for detail), it is estimated that the return level is convex and has finite bound,

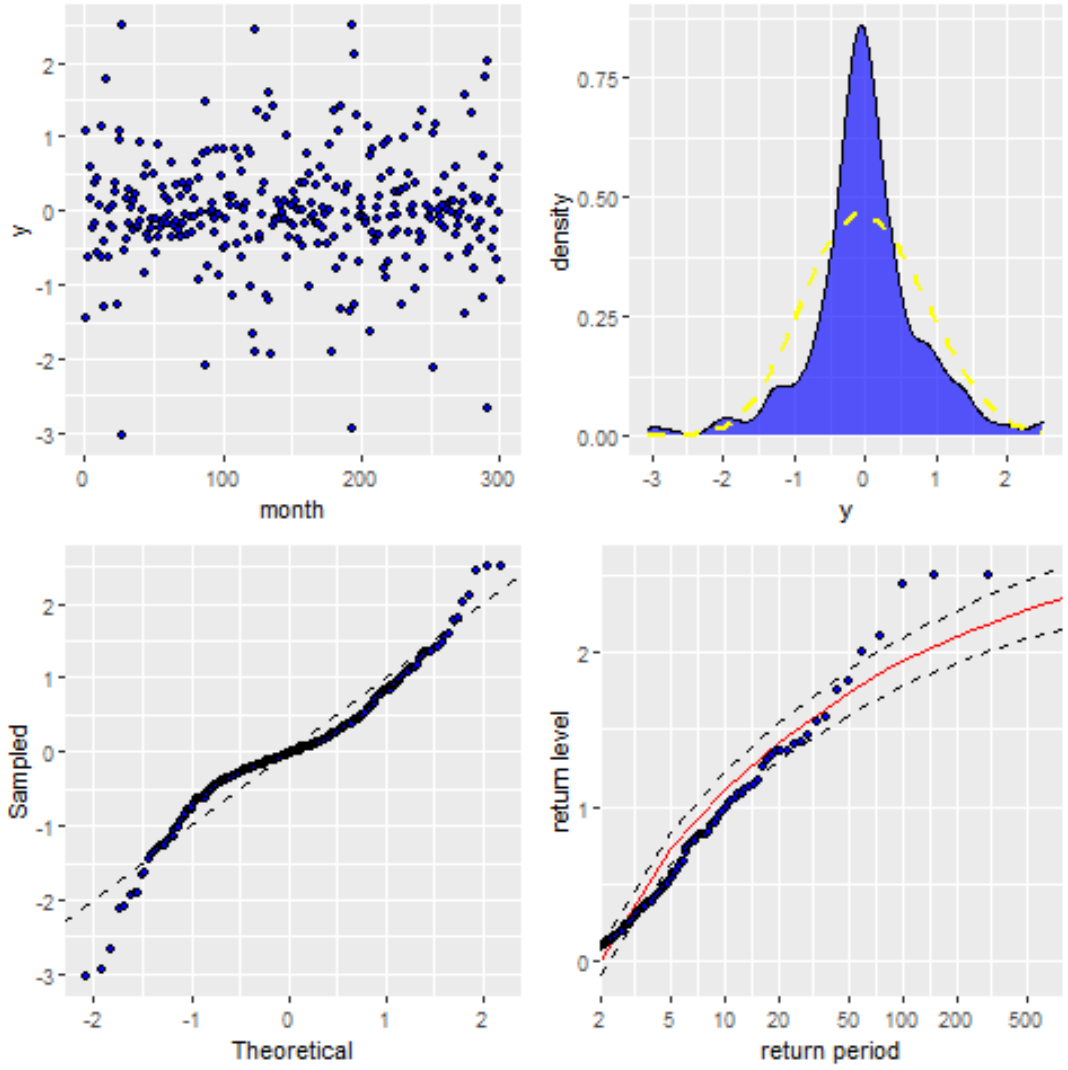


Figure 3.2: The diagnostic plots for the GEV model fitted to transformed O_3

but this does not appear to be strictly supported by the data obtained. In addition, the simple observation of the time series plot may lead to the same concern given that the variance of the data seems to become greater, presenting a funnel like structure.

Figure 3.2 shows the diagnostic plots for the block maxima model fitted to O_3. Compared to SO_2, the diagnostic plots indicate a poorer fit of the model. Neither the Q-Q plot nor the density plot suggests an agreement between the theoretical and sample quantiles. This is most likely due to the residual seasonal effect of the data, which remains strong even after differencing. In other words, given the characteristic of the data, the usual likelihood approach is not reliable in this case.

3.4 Parameter inference and uncertainty

As mentioned, the estimates are given by maximising the likelihood as usual. The confidence interval (CI) is constructed again in the ordinary way, i.e. by utilising the asymptotic normality of MLE. There are many R packages developed to achieve these very easily, and *extRemes* introduced in Gilleland and Katz (2016) is mainly used for the project. The result of parameter inference is given in the table below.

	SO_2			O_3		
	0.025	0.5	0.975	0.025	0.5	0.975
$\hat{\mu}$	-0.380	-0.291 (0.045)	-0.203	-0.387	-0.288 (0.050)	-0.190
$\hat{\sigma}$	0.681	0.738 (0.029)	0.795	0.757	0.819 (0.032)	0.882
$\hat{\xi}$	-0.249	-0.218 (0.016)	-0.187	-0.291	-0.253 (0.020)	-0.216
AIC		679.907			726.948	

Table 3.1: Parameter estimates and confidence intervals of block maxima model. The standard error of these estimates are given inside parentheses.

Given the results, with confidence $\hat{\xi}$ is estimated to be below 0 for both data if the models are correct. This is reflected in the bounded return levels shown in diagnostic plots discussed in section 3.3. The standard errors of parameter estimates are relatively small compared to the value of estimates.

It is important to note that the there are two distinctive sources of uncertainty when analysing the extreme values. Firstly, a poor approximation leads to a bias in the estimates. The approximation is expected to be good only when the number of observations in each block, denoted as n in 3.1, is large enough. There are more than 700 observations for the data, the precise number depending on the number of days in a particular month from which each monthly maxima is extracted. Empirically, this is expected to give enough sample size for a good approximation. Secondly, the parameter estimates are subject to bias when the number of observations used in inference is small. For the transformed data used in this chapter, there are 303 and 301 monthly maxima for SO_2 and O_3, respectively. This, in virtue of the Central Limit Theorem (CLT), gives a reasonably small margin of error as shown in Table 3.1. There is always an issue of trade-off in the source of uncertainties as taking the large n consequently leads to a small number of observations. This can be addressed by setting the threshold rather than dividing the data into block, the method known as POT model, which is discussed in Chapter 4. Alternatively, we could decrease the bias in inference by using r largest observations instead of single maximum.

3.5 Conclusion

Given the results in this chapter, it appears that transforming the data prior to model fitting to deal with the false i.i.d. assumption, an approach analogous to Box-Jenkins framework, is equally reasonable in the analysis of extreme value. Apparently, the effective transformation of the time series that ensures i.i.d. assumption or stationarity of data leads to a better performance of the model. Especially, the treatment of seasonality is quite important as the usual likelihood approach cannot be justified if it remains persistent over a long period of time. In general, the performance of the block maxima model is not quite satisfactory. Particularly, the model fails to explain the tail behaviour of the distribution.

The same issue of intractability of original distribution, discussed in section 2.3, is still present in this approach. After all, even if a model performs well, the choice of GEV distribution as an underlying stochastic process can be pointless if it cannot be linked to the original time series, which is of our interest. The more advanced method with greater flexibility in modeling structure can address this issue, which will be discussed in Chapter 6.

Chapter 4

Peak-Over Threshold Model

4.1 Review of theory

The POT model is an alternative method for statistical analysis of extreme values. Without splitting the data into blocks, the model focuses on the exceedance after an appropriate threshold is set. It can be referred to as generalised Pareto (GP) model when the emphasis is on its underlying distribution.

Again, we briefly review the main theorem. Consider

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

where $X_1, ..., X_n$ is a sequence of i.d.d random variables whose common distribution function is denoted as F. Denote an arbitrary term in the above sequence as X, suppose F satisfies, as shown in section 3.1, that

$$\Pr\{M_n \leq z\} \approx \mathrm{G}(z) \quad \text{for large n,}$$

where

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

for some μ, $\sigma > 0$ and ξ.

Then, for large enough u, the distribution function of $y = (X - u)$, conditioned that $X > u$, is approximately

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

defined on $\{y : y > 0 \text{ and } (1 + \xi y/\tilde{\sigma}) > 0\}$, where

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

4.2 Model specification

As mentioned above, the underlying distribution of POT model is GP distribution, which is often parameterised by scale parameter σ ($\tilde{\sigma}$ in the above theorem) and shape parameter ξ. As stated in section 4.1, these are associated with corresponding parameters of block maxima model. Then the POT model we consider in this chapter can be expressed as

$$y_t|u \sim \text{GP}(\sigma, \xi) \quad \text{for t} = 1, ..., n,$$

where the arbitrary threshold is denoted as u. Importantly, though it is not the target of statistical inference, the choice of u requires extra care for the POT model to be valid. The detail of this is discussed in section 4.3.

There are several things to note about y_t. Firstly, y_t here represents all the data, not just monthly maxima; the POT model seeks to increase the efficiency of inference by using all the exceedances instead of dividing the data into blocks and therefore discarding the potential extreme values. Secondly, for the same reason explained in 3.2, the time series of SO_2 undergoes the appropriate transformations. As for O_3, however, the data is not transformed in this chapter. This is because the seasonality does not on its own undermine the asymptotic condition if the POT model is fitted to cluster maxima (recall that O_3 does not present any structural variation except seasonality). This relates to additional modification known as declustering, commonly implemented with POT model to tend the stationarity of the data. Section 4.4. discusses this in more detail.

4.3 Threshold selection

There are several ways suggested to choose the effective value for threshold u, but the general consensus is that lowest value of u should be chosen (as this allows larger number of samples to be used for inference) while the approximation condition is ensured.

The most common tool to achieve this is known as mean residual life (MRL) plot, which is used for the models in this chapter. The plot is produced as the locus of the points

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

where $x_{(1)}, ..., x_{(n)}$ consists of n_u observations exceeding the chosen u. In practice, allowing for the confidence interval, one should look at the evidence of linearity as theoretically this is known to be true provided that the approximation condition is satisfied.

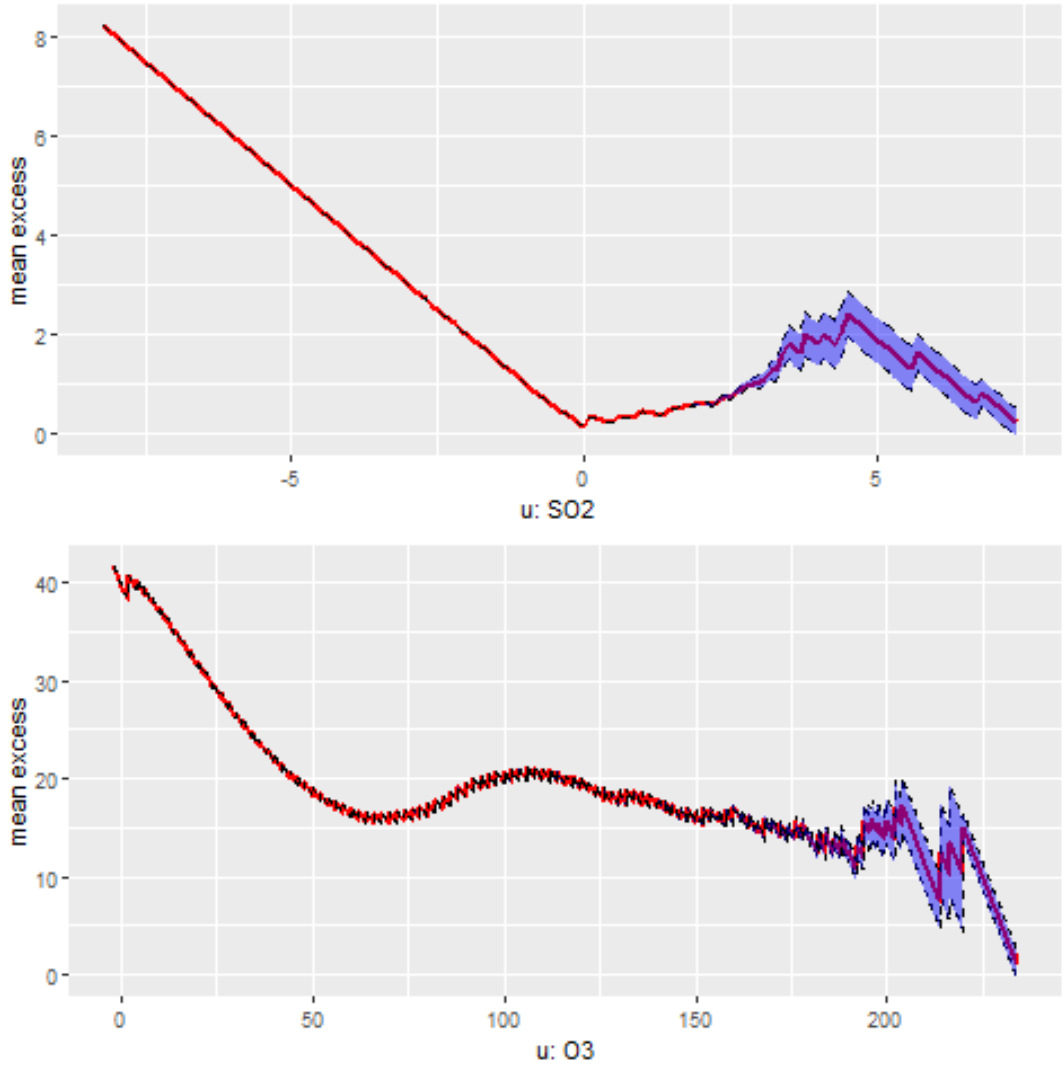


Figure 4.1: The MRL Plot of SO_2 and O_3: The confidence interval given is supported by the CLT.

In Figure 4.1 the MRL plot for SO_2 and O_3 is provided. According to the criteria stated above, the reasonable choice of threshold value for SO_2 may be 1.25. This gives the threshold exceedance rate of $\approx 0.536\%$, resulting with n observations. As for O_3, the reasonable choice of threshold value may be 2.5. This gives the threshold exceedance rate of $\approx 1.049\%$, resulting with 2158 observations.

The interpretation of the MRL plot is not straightforward. In fact, past literature argue that the subjectivity in threshold choice can undermine the exact reproduciblity. The relevant discussions are well summarised in Caballero-Megido et al. (2017). In practice, the GP model is usually fitted at a range of thresholds while monitoring the stability of parameter estimates. The plot used to this end is known as the parameter stability plot. Although the detail is omitted here, the above choice of thresholds by MRL plot appears to be reliable after taking

into account the stabilities of parameter estimates.

4.4 Declustering

As repeatedly mentioned, i.i.d assumption of the random variables are not exactly met even after transformation, resulting with stationary time series at best. For the block maxima model, this was not an issue as discussed in 3.2. This is not the case, however, with POT model, which requires a treatment known as declustering prior to model fitting.

The idea of declustering is that by combing neighboring exceedances into one cluster, each of resulting cluster maxima is then independent from one another. The number of observations below threshold needed to divide one cluster from another, commonly denoted as r, is often chosen arbitrarily. For the data, we set $r = 24 \times 30 = 720$, which is based on the assumption that the observations within the same month tends to be in the same cluster.

4.5 Model diagnostics

The same diagnostic plots shown in section 3.3 are used for the POT models. The difference is that all observations are included in the time plot, with those which exceed the threshold highlighted by red. Note that the rest of the plots are constructed based on the exceedance only. Additionally, to examine the effect of declustering, the POT models are fitted to both the original and declustered data so that the diagnostic plots can be compared to see the differences.

Figure 4.3 shows the diagnostic plots for SO_2, with (a) showing the result when the POT model fitted to the original data and (b) to the declustered data. Overall, declustering does not appear to affect the inference greatly, which is natural given that the auto-correlation for SO_2 does not persist over a long period of time as shown in 2.2. Both of the Q-Q plots suggest a mild agreement of the two quantiles although the upper tail of the distribution is not well explained by the models. As for the return level plot, compared to the corresponding block maxima model in Chapter 3, there seems to be some improvements in performance of the model, if not entirely satisfactory.

Figure 4.4 shows the diagnostic plots for O_3. Unlike SO_2, there is a significant difference in plot (a) and (b), which is again natural given that O_3 has a lingering auto-correlation possibly due to the seasonal effect. The comparison of these plots is insightful as it reveals the different sources of uncertainty having an effect on the model fit. The Q-Q plots suggest an overall agreement of the two quantiles with a few points in upper tails that seem to slightly deviate from the theoretical values. Notice that the declustering appears to compromise the model fit in Q-Q plot and density plot due to the decreased number of observations used for estimating the parameters. In fact, with the chosen value of u and r, the sample size decreases from 2158 to 54 as a result of declustering.

The return level plot, however, indicates a considerable improvement as the quality of approximation is ensured by the declustering. This trade-off of bias is analogous to the choice of block size in block maxima model.

4.6 Parameter inference and uncertainty

The same method of the parameter inference as previously implemented in Chapter 3 is used. In Table 4.1 and 4.2, the results are presented for both original and declusterd data. For more rigorous analysis, several more pairs of u and r should be investigated. As discussed in section 4.5, the comparison of the original and declusterd data again illustrates the difference sources of uncertainty and trade-off of bias.

	Original			Declusterd		
	0.025	0.5	0.975	0.025	0.5	0.975
$\hat{\sigma}$	0.261	0.291 (0.016)	0.322	0.230	0.436 (0.043)	0.641
$\hat{\xi}$	0.188	0.273 (0.105)	0.358	0.107	0.516 (0.208)	0.924
AIC		77.970			72.487	

Table 4.1: The result of parameter inference for SO_2. The number of observations used for inference is 943 for the original data and 54 for the declusterd data, respectively. Overall, the results are similar. The standard error for each parameter estimate is larger when the model is applied to the declusterd data, widening the confidence interval as a consequence. The AIC of the models becomes slightly smaller as a result of declustering.

	Original			Declusterd		
	0.025	0.5	0.975	0.025	0.5	0.975
$\hat{\sigma}$	22.326	23.591 (0.645)	24.856	37.772	57.189 (9.907)	76.607
$\hat{\xi}$	-0.163	-0.130 (0.017)	-0.096	-0.621	-0.404 (0.111)	-0.188
AIC		17401.840			440.364	

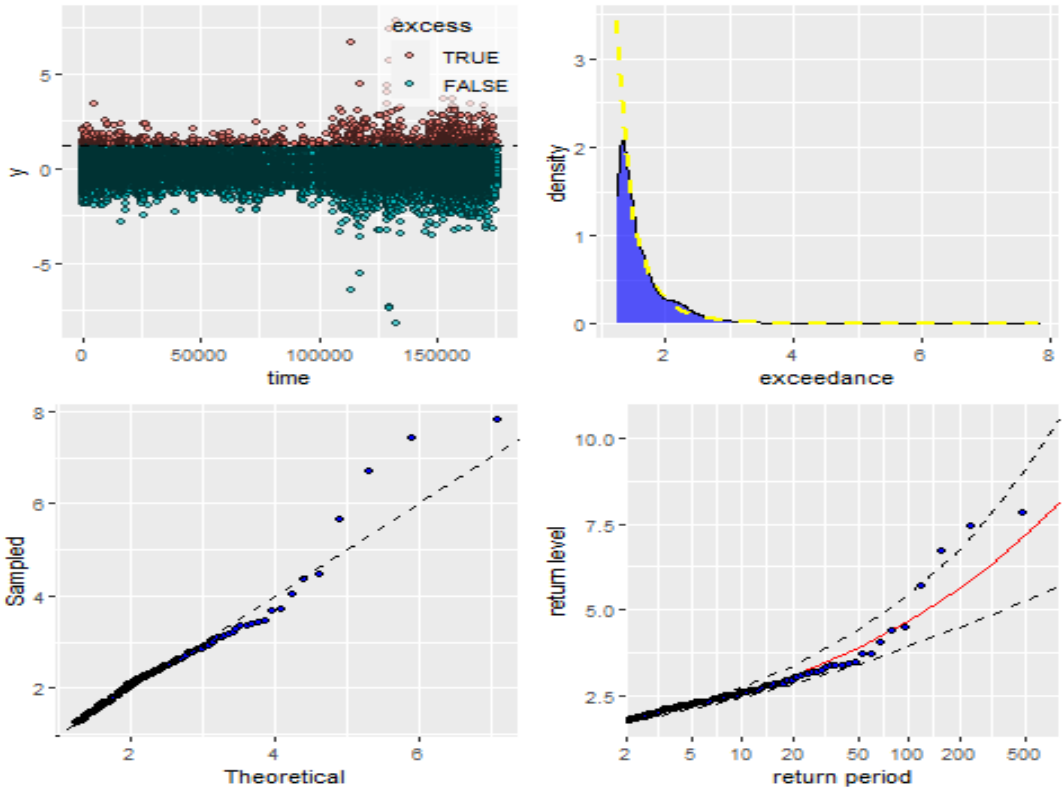
Table 4.2: The result of parameter inference for O_3. The number of observations used for inference is 2158 for the original data and 57 for the declusterd data, respectively. The results are clearly different. As a result of declustering, the standard error increases by a factor of 15 for scale parameter and 6.5 for shape parameter, severely widening the confidence interval. The large AIC for the original data is alarming, which shows an considerable reduction after declustering.

It appears that the seasonal effect in the data severely undermines the asymptotic condition, and thus the importance of declustering is indubitable for such a case. Comparing the AIC between the two models can be instructive. Since the same models are fitted to the different data, the AIC cannot be referred to in an orthodox way of the model selection (see Apendix B). Nevertheless, the considerable reduction that in AIC is cautionary, given that the Q-Q plot and density plot appear to indicate good fit of the model. In other words, one should not only rely on the diagnostic plots to assess the model fit. This also showcases the efficacy of return level plot as diagnostic tool of a model. The disadvantage of declustering is that the large value of r would severely decrease the effective sample size, leading to large bias of parameter inference. This becomes an issue if the auto-correlation of data is strong, as the case with O_3, because it inevitably requires the choice of r to be large for an effective declustering.

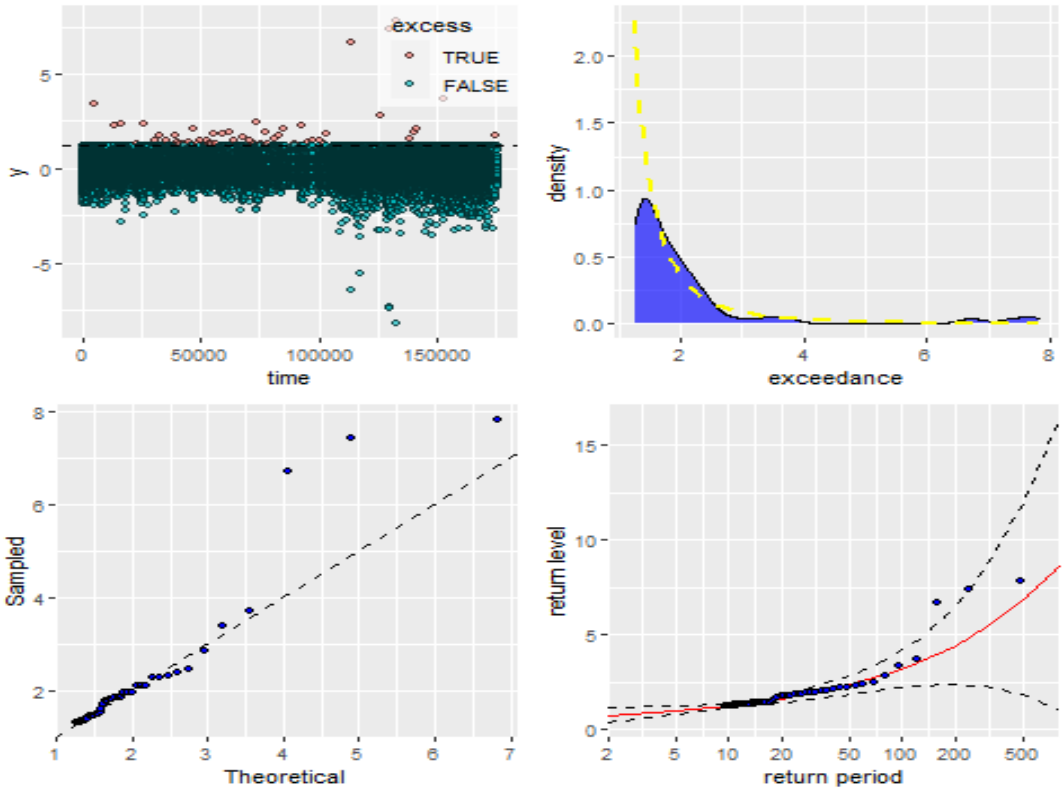
4.7 Conclusion

In this chapter, we discussed the POT model to analyse the stochastic process of extreme values. The same transformation approach discussed in Chapter 3 is partly used to deal with the false i.i.d assumption. The result indicates that the POT model better explains the data compared to block maxima model, provided that the appropriate care for the stationarity is taken. Particularly, the improvement seen in modeling O_3 with seasonal effects is remarkable as the declustering ensures the asymptotic condition of theory, although this comes at the cost of severely decreased sample size and therefore biased parameter estimates.

The POT model yet leaves room for improvements. Besides the remaining issue of intractability (see section 3.5), the modeling method does not provide a perfect solution for the data with time-variant structure. Particularly, as can be seen in the diagnostic plots, the model fails to explain the abrupt increase in variance observed in the transformed SO_2 even after the declustering is implemented. This again gives rise to more advanced modeling method that can address the aforementioned issues simultaneously. This will be fully discussed in Chapter 6.

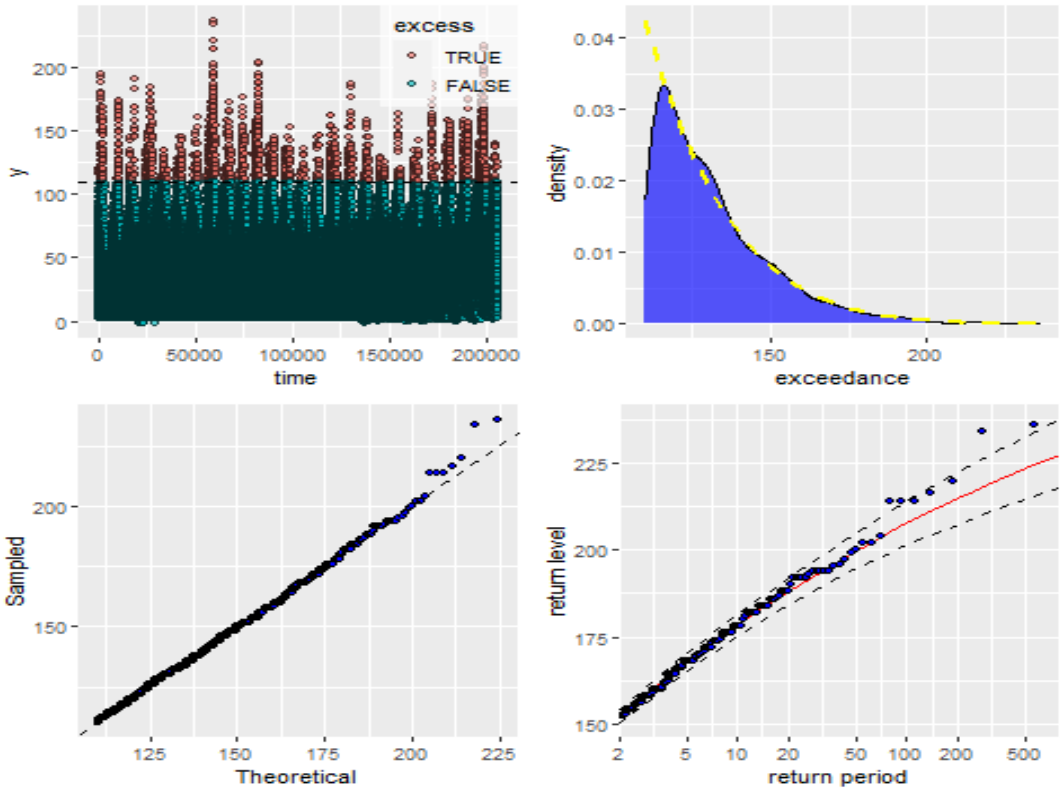


(a) Original

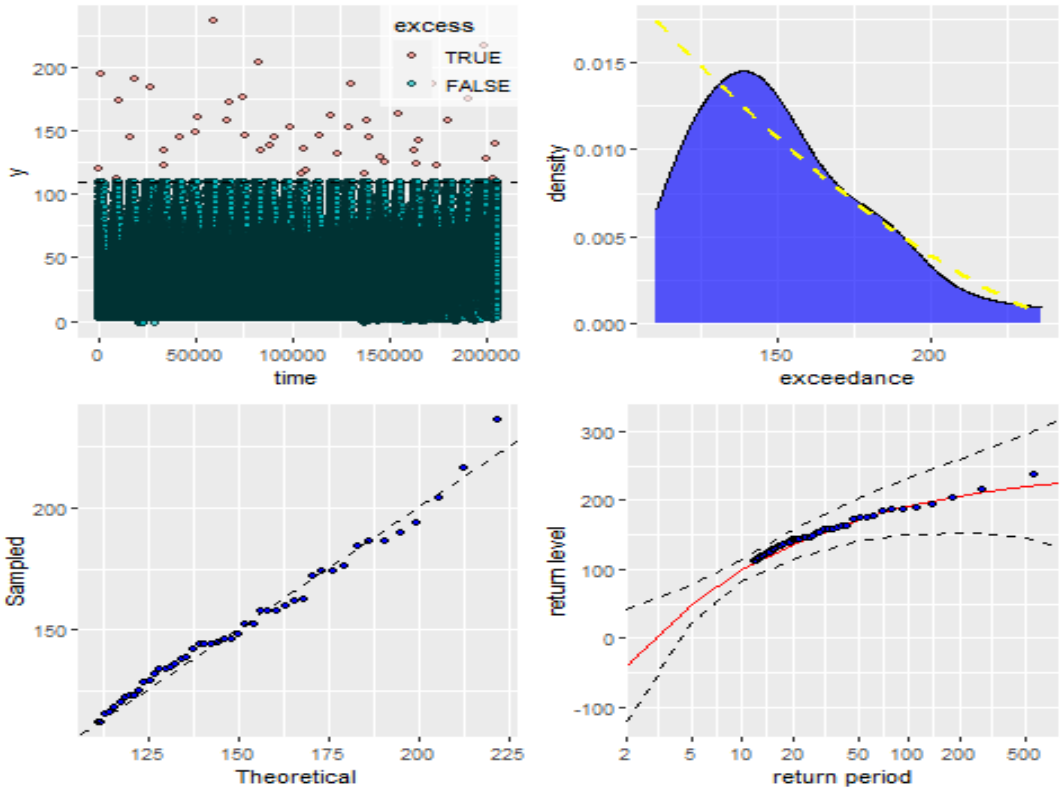


(b) Declusterd

Figure 4.2: Diagnostic plots for SO_2



(a) Orginal



(b) Declusterd

Figure 4.3: Diagnostic plots for O_3

Chapter 5

Bayesian Statistical Modeling

5.1 Motivation

So far, two classic models are fitted to the data after appropriate transformations and modifications to take care of the practical issues, which accomplished reasonably successful results. There is, however, still room for improvements in model fit of these methods. The block maxima model discussed in Chapter 3 does not fit to the data quite well, especially for O_3 with a residual seasonal effect. Although the POT model in Chapter 4 shows definite improvements, it fails to explain the upper tail behaviour of SO_2. All these issues are likely to be caused by the residual non-stationarity of the data that is not sufficiently dealt with by the multiple transmutations implemented prior to the model fitting. In other words, the more successful modeling method comes with more sophisticated techniques to deal with non-stationary time series. Further, as repeatedly stated, the major disadvantage of this approach is the difficulty in tracking the stochastic process of the original data. That is, this approach comes at the cost of decreased interpretability. Therefore, identifying the underlying distribution, trend, seasonal effect and other structure of data become challenging tasks, all of which are naturally of great interest.

These evident issues provide the motivation to implement a more sophisticated modeling approach. A state space model, for instance, is quite an appealing option as it produces highly interpretable outputs while effectively taking care of non-stationarity of data. This advanced modeling method may be discussed under the framework of Bayesian statistical modeling as parameter inference for such model, assuming more complex model structure, often requires Bayesian approach. In addition, another incentive in the use of Bayesian inference is to provide an alternative approach in quantifying the uncertainties of parameter inference; the posterior distributions of parameters will be provided enabling more straightforward interpretation, although this aspect of Bayesian inference is not confined in the context of state space model.

5.2 User-defined function in Stan

Stan is a commonly used probabilistic language, gaining as of late an increasing recognition among the scientific community ever since the series of Bayesian approach in statistical analysis became more popular. In contrast to the classical methods such as Kalman Filter, the parameter estimates can be obtained directly through HMC sampling, which enables more complex, flexible model structure. Stan provides a variety of built-in probability density (or mass) functions whereby one can construct a statistical model. The GEV distribution family, however, the most common family of distribution established in extreme value theory, is unfortunately not available. However, Stan provides a user-defined function feature to implement such distributions on demand, allowing more flexible statistical modeling of extreme values.

An incredibly helpful online notebook by Vehtari (2017) can be found in the community website of Stan, where the author suggests the set of user-defined functions to implement POT model with underlying GP distribution. The source code is referenced upon defining the following two functions, which corresponds to GEV distribution.

_lpdf

_lpdf function defines a log likelihood rather than the log PDF itself. Consider the log likelihood of GEV distribution when $\xi \neq 0$:

$$\begin{aligned}
\log\mathcal{L} &= \log\prod \mathrm{f}(y|\mu,\sigma,\xi) \\
&= \log\prod\left[\frac{1}{\sigma}\left(1+\xi z\right)^{-(\frac{1}{\xi}+1)}\exp\left\{-(1+\xi z)^{-\frac{1}{\xi}}\right\}\right] \\
&= \sum\left[-(\frac{1}{\xi}+1)\log(1+\xi z)-(1+\xi z)^{-\frac{1}{\xi}}\right]-\mathrm{N}\log\sigma,
\end{aligned}$$

where $z = \frac{y-\mu}{\sigma}$. The subset of GEV family with $\xi = 0$ can be interpreted as limit of PDF above as $\xi \to 0$, that is

$$\mathrm{f}(y|\mu,\sigma,\xi) = \frac{1}{\sigma}\exp\left\{\frac{-(x-\mu)}{\sigma}\right\}^{(\xi+1)}\exp\left\{-\exp\left\{\frac{-(x-\mu)}{\sigma}\right\}\right\}.$$

Given this, the log likelihood above when $\xi = 0$ simplifies to

$$\log\mathcal{L} = \sum\left[(\xi+1)z+\exp(-z)\right]-\mathrm{N}\log\sigma.$$

Note that the above likelihood assumes the relevant parameters to be time-invariant. That is, if one wishes to model otherwise an appropriate modification is necessary, which is rather straightforward.

_rng

_rng function generates random numbers given the GEV distribution specified above, iteratively with sampled parameters. The usual inversion method is used here. Consider the inverse CDF, which is given by

$$\mathrm{F}_X^{-1} = \mu + \frac{\sigma}{\xi}\left(\left\{\log\frac{1}{X}\right\}^{-\xi} - 1\right)$$

when $\xi \neq 0$, and

$$\mathrm{F}_X^{-1} = \mu - \sigma\log\left(-\log X\right)$$

when $\xi = 0$. Thus, a random number from the target from the target GEV distribution can be obtained by generating uniform random number i.e. $X \sim \mathrm{U}(0, 1)$.

Sampling constraint

In Stan, it is possible to declare the range of sampling with respect to each parameter. In practice, however, the sampling can occur beyond the declared constraint, rendering the parameter inference unreliable. To take care of this issue, each function is required to be implemented in such a way that the samplings will be rejected in case they take place beyond constraints determined by the support of PDF, for GEV distribution, that is

$$\begin{cases} x \in \left[\mu - \frac{\sigma}{\xi}, +\infty\right) & \text{if } \xi > 0, \\ x \in (-\infty, +\infty) & \text{if } \xi = 0, \\ x \in \left(-\infty, \mu - \frac{\sigma}{\xi}\right] & \text{if } \xi < 0. \end{cases}$$

After simple algebra, the set of constraints regarding the parameter can be obtained from above, which is reflected in each user-defined functions by using the default reject function.

5.3 Model testing

To ensure that the user-defined functions are properly implemented, it would be reasonable to conduct a couple of tests. In this section, a simple test model

$$y \sim \text{GEV}(\mu, \sigma, \xi),$$

which is the exact same model implemented in Chapter 3, is used for this purpose. This model is fitted to the same transformed SO_2 data (see 3.2) so as to make the comparison of results straightforward although this model itself is not particularly designed to overcome the issues discussed earlier.

Convergence diagnostics

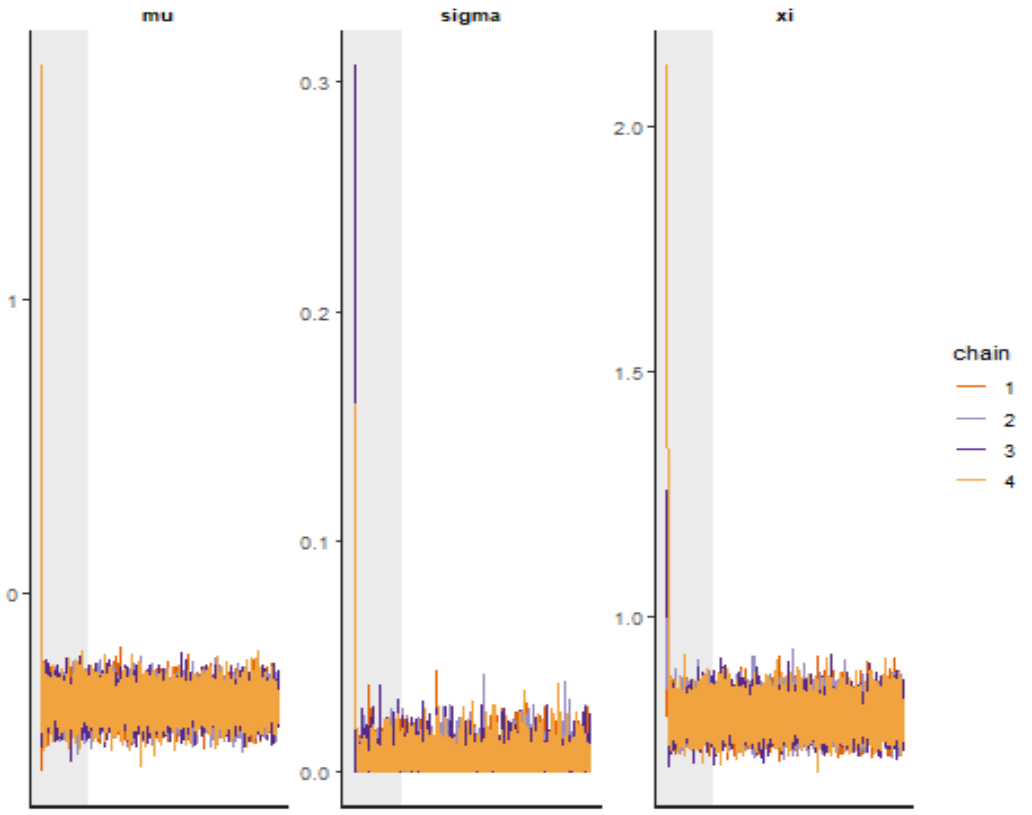


Figure 5.1: The trace plot of sampling with respect to three parameters. The four chains of samplings are conducted, each with 5000 iterations and 1000 warm-ups, which is shaded by the gray bar on left. All samplers appear to converge very quickly.

First, it is sensible to check the convergence of the sampling chain before discussing the parameter inference. The usual trace plot is shown in Figure 5.1. which indicates the successful convergence. Rhat is close enough to 1 for all parameters, suggesting that the sampling chains have mixed.

Comparison of parameter inference

Since the convergence of sampling chain is confirmed, it is possible to discuss the result of parameter inference. It is insightful to make a comparison of estimates between the usual likelihood approach discussed in previous chapters and Bayesian inference performed by Stan.

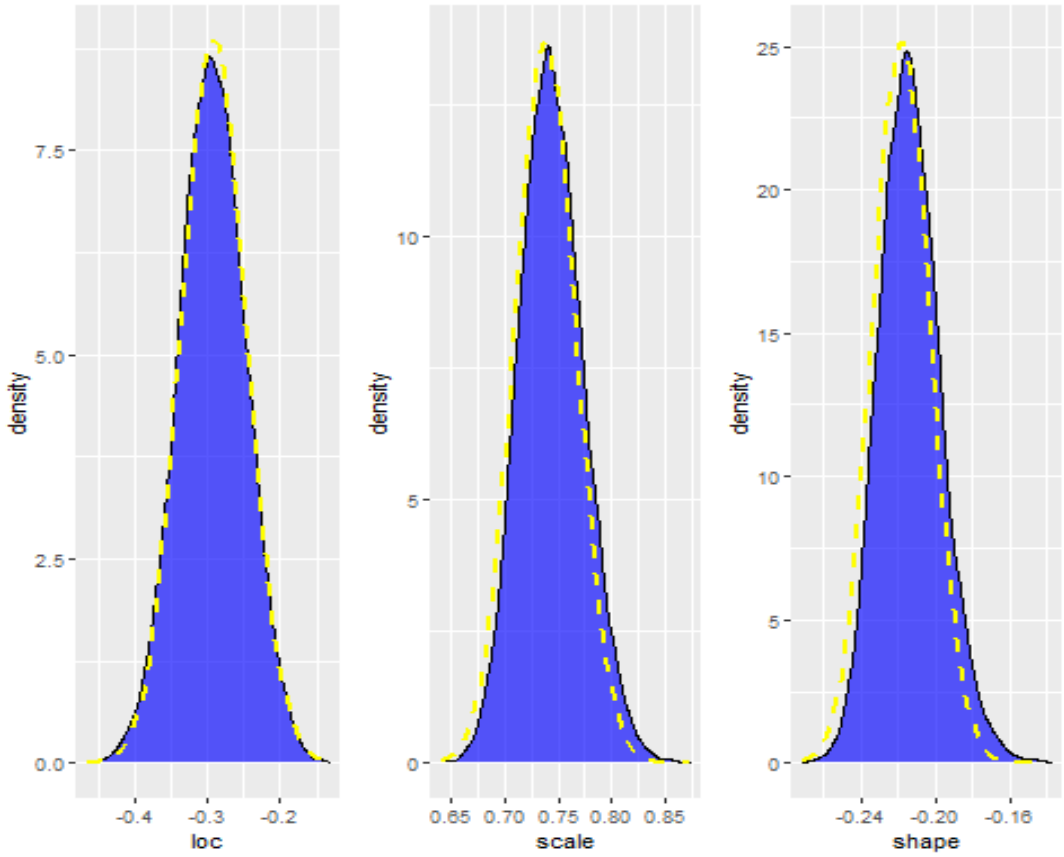


Figure 5.2: Comparison of the parameter distributions obtained by MLE approach and Bayesian inference. The yellow dotted line shows the asymptotic parameter distribution obtained by MLE and its standard error. The approximation is expected to be good with $n = 304$. The blue density plot shows the posterior distribution of parameters constructed by HMC samplings.

It is clear from Figure 5.2 that the implementation of the user-defined function is successful as the results obtained by Stan appear to be consistent with those of earlier analysis. There are slight differences in the interval of estimates given, although the credible interval obtained by Bayesian inference is easier to interpret, which is an advantage.

5.4 Conclusion

In this chapter, several preliminary topics are discussed upon proceeding to a more sophisticated model i.e. state space model, discussed in the final chapter.

This modeling approach seeks to address the issues discussed in section 5.1 and are implemented via the user-defined function shown in section 5.2.

Chapter 6

State Space Model

6.1 Introduction to state space modeling

6.1.1 Ad-hoc nature of state space modeling

In this chapter, we consider the state space model in an attempt to overcome the issues that are highlighted in section 5.1. Given the ad-hoc nature of the state space model, it is important that the characteristics of the data are carefully reflected in a structure of each model, which then should be optimally parameterised for the particular case. For this reason, the models in this chapter are constructed separately for each data i.e. SO_2 and O_3. There are three models discussed throughout the chapter, denoted as Model A (for SO_2), Model B-1 and Model B-2 (for O_3), respectively.

6.1.2 Modeling non-stationary extreme values

As discussed in the previous chapter, one of the primary incentives in using the state space model is to model a non-stationary stochastic process (without transforming the data). In fact, this has been the subject of literature concerning the EVA of environmental data partly due to the increasing awareness of anthropogenic climate change. Achieving this by state space model often amounts to assuming the model parameters to be time-variant.

It is instructive to briefly discuss the conventional approaches to modeling the non-stationary extreme values. Apparently, the most popular way to implement time-variant parameters is to introduce appropriate covariates. For example, typically to account for trend of the data, the location parameter is often assumed to be a linear function of time covariate; Cheng et al. (2014) evaluates the software package Non-stationary Extreme Value Analysis (NEVA), in which the model assumes $\mu_t = \mu_1 t + \mu_0$. Similarly, in order to allow for the seasonal effect of the data, one can introduce seasonal covariates, the choice of which is typically supported by the empirical evidence reported by relevant studies. Ouarda

and Charron (2019) refers to a variety of climate indices (e.g. the Southern Oscillation Index (SOI), the Pacific Decadal Oscillation (PDO) and the Atlantic Multi-decadal Oscillation (AMO), etc.) as useful covariates for modeling the extreme temperatures with seasonal effect. Further, an alternative approach is to use annually block maxima instead of monthly, which often obviates the necessity for modeling the seasonal effect if it occurs in the annual cycle. An obvious disadvantage in this approach is that it considerably decrease the effective sample size, jeopardising the reliability of inference based on traditional likelihood approach. The use of covariates have several advantages. First, parameter inference can be conducted within the framework of traditional likelihood approach. Towler et al. (2010) fits the conditional GEV model to the data using the R package *extRemes*, the same package used in Chapter 3 and 4. It is then possible to use the conventional model selection criteria, such as the likelihood ratio test (LRT) or AIC, which enable the statistical assessment as to whether certain covaraites increase the model performance.

The models used in the following discussion are implemented by Stan, which allows direct samplings from the target posterior as briefly explained in section 5.2. Stan does not limit the model structure to that of covaraite approach, which is clearly an advantage when implementing state space models that will be suggested momentarily. Note that it is generally difficult to use the aforementioned model selection methods because these are based on the likelihood of the model, which tends to be quite complex and difficult to evaluate. This, however, is barely an issue here because the presence of the trend or seasonal effect is indubitable from the simple EDA. The assessment of model fit can still be conducted via the modified diagnostic plots, which will be discussed in the next section.

6.1.3 Modification in Model Diagnostics

The diagnostic plots used in Chapter 3 and 4 are based on the estimated parameter. This poses a slight issue because the models in this chapter assume several time-variant parameters as discussed in the previous section. Coles (2001) suggests a modification method in the diagnostic plots, after which the model fit can be assessed in the same manner as previously. The idea of modification is to standardise the data by the obtained parameters so as to allow for the non-homogeneous distribution. Suppose, for example, the data is fitted to a GEV model with time-variant parameters i.e.

$$y_t \sim \text{GEV}\left(\mu_t, \sigma_t, \xi_t\right).$$

Then, the y_t can be standardised to be

$$\hat{z}_t = \frac{1}{\hat{\xi}_t} \log\left\{1 + \hat{\xi}_t \left(\frac{y_t - \hat{\mu}_t}{\hat{\sigma}_t}\right)\right\},$$

which can be compared to the standard Gumbel distribution with CDF

$$\Pr(Z \leq z) = \exp\left\{-e^{-z}\right\}.$$

This is a special case of GEV distribution, where $\mu = 0$, $\sigma = 1$, and $\xi = 0$, interpreted as limit of the CDF as $\xi \to 0$.

6.2 SO_2

6.2.1 Model specification

Recall the time series plot of SO_2 shown in Figure 2.1. The characteristic of SO_2 is an evident downward trend overt time and decreasing variance, both of which should be incorporated into the modeling structure. Note that all models in this chapter are to be fitted to the original time series, denoted as y_t, without any transformations or modifications. After trial and error, we suggest

Model A: GEV model with time-variant parameters:

$$\begin{aligned}
\mu_t &= \mu_{t-1} + \epsilon_t, \\
\sigma_t &= \beta_1 \exp\{-\beta_2 t\}, \\
\epsilon_t &\sim \mathrm{N}(0, \gamma), \\
y_t &\sim \mathrm{GEV}\left(\mu_t, \sigma_t, \xi\right).
\end{aligned}$$

In Model A, there are two time-variant parameters i.e. μ_t and σ_t. The parameterisation of σ_t is inspired by the observation of time series plot, which indicates exponential decay. The shape parameter ξ is assumed to be constant following the traditional approach although Ouarda and Charron (2019) argues that this may not be evidently justifiable. In the implementation of Stan, no particular priors are declared for β_1, β_2 and γ, meaning that Uniform prior is used for these parameters. This model is essentially a block maxima model discussed in Chapter 3 with time-varying parameters.

Intuitively, it appears sensible to incorporate a noise factor in the stochastic process of σ_t, but this may lead to a problematic posterior. The reason for this is that mean of GEV distribution is given by

$$\begin{cases}
\mu + \sigma(g_1 - 1)/\xi & \text{if } \xi \neq 0, \xi < 1 \\
\mu + \sigma\gamma & \text{if } \xi = 0 \\
\infty & \text{if } \xi \geq 1
\end{cases},$$

where γ is Euler's constant and $g_k = \Gamma(1 - k\xi)$. Since σ_t does have an effect on the mean of the stochastic process, including a noise term might lead to

colinearility. In fact, the sampling chains do not appear to converge successfully when the model includes noise term in σ_t.

6.2.2 Convergence diagnostics

Figure 6.1 shows the usual trace plot for the convergence check. The plot indicates the successful convergence. Rhat is close enough to 1 for all parameters, suggesting that the sampling chains have mixed. n_eff is low for γ and for the μ_t and σ_t in the beginning, which indicates that the samplings are not effectively conducted in this particular regions of parameter space even though the corresponding Rhat suggests that the sample chains have mixed eventually. Overall, these diagnostic values indicates the effective samplings. These values are partly included in the Table 6.1 with the result of parameter inference.

6.2.3 Model diagnostics

Figure 6.2 shows the diagnostic plots for Model A. As explained in section 6.1.3, the diagnostic plots are modified to allow for the lack of homogeneity in the distributional assumption of the model. For standardisation of the data, the parameter estimates obtained via HMC samplings are used, part of which is shown in 6.2.4. The result is quite satisfactory. There are three observations above 5.0 which may be considered as outliers. The plots otherwise suggest a remarkable improvement in model fit compared to the block maxima model discussed in Chapter 3.

6.2.4 Parameter inference and uncertainty

As discussed in Chapter 5, the parameter inference for the state space model is conducted by HMC sampling. The result of this is shown in Table 6.1.

	Estimates		95% CI			Convergence	
	mean	sd	0.025	0.5	0.975	n_eff	Rhat
$\hat{\mu}_1$	88.205 (0.465)	0.121	65.041	88.016	112.253	674.210	1.008
$\hat{\mu}_{304}$	2.799 (0.015)	1.202	0.109	2.909	4.869	6444.044	1.000
$\hat{\sigma}_1$	37.221 (0.066)	4.707	28.901	36.936	47.209	4996.952	1.000
$\hat{\sigma}_{304}$	1.942 (0.005)	0.305	1.397	1.928	2.592	4382.355	1.001
$\hat{\xi}$	0.238 (0.001)	0.064	0.126	0.234	0.374	5546.418	1.000
$\hat{\gamma}$	1.855 (0.020)	0.347	1.231	1.832	2.584	295.276	1.021
$\hat{\beta}_1$	37.590 (0.068)	4.779	29.148	37.290	47.752	5007.877	1.000
$\hat{\beta}_2$	0.010 (0.000)	0.001	0.008	0.010	0.011	5817.634	1.000

Table 6.1: The result of parameter inference for the state space model fitted to SO_2. For the time-varying parameter μ_t and σ_t, only the estimates at the first and the last points are shown.

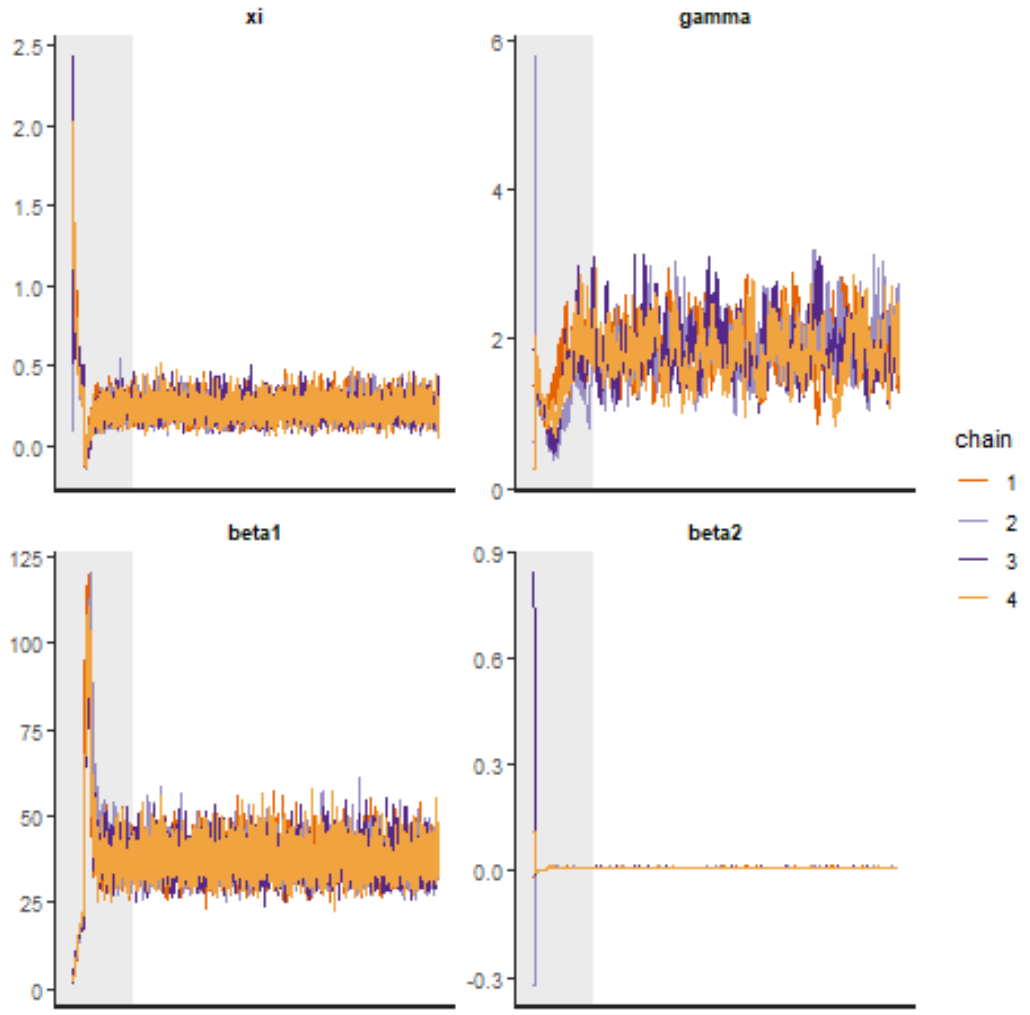


Figure 6.1: The trace plot of sampling with respect to four parameters ξ, γ, β_1 and β_2. The four chains of samplings are conducted, each with 5000 iterations and 1000 warm-ups.

Considering the similarity between the mean and 50% quantiles of the samplings, though the visualisation of posterior distributions are omitted here, for most of the parameters there is no significant difference between mean and mode, implying the uni-modality.

6.2.5 Prediction

Another advantage of state space model is the ease in constructing the prediction interval, which is shown in Figure 6.3. By using the random generator function defined in section 5.2, the prediction interval can be obtained for each point in the time series using the corresponding estimates of μ_t, σ_t and ξ. Again, the result is satisfactory in that the model appears to precisely capture the

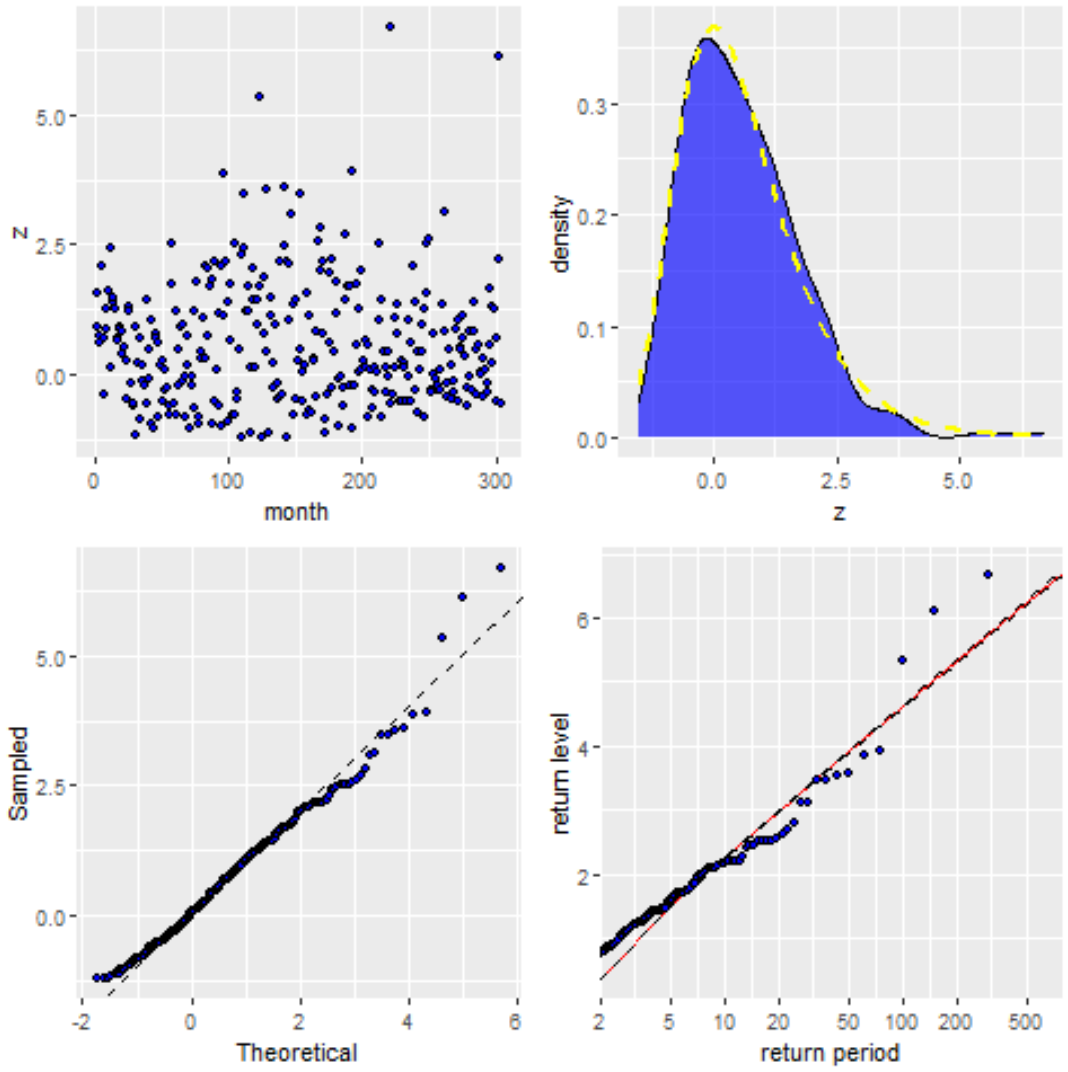


Figure 6.2: The diagnostic plots for the state space model with underlying GEV distribution. There is no apparent structure in the time series plot, indicating the successful modeling of time-variant components. The density plot indicates the satisfying agreement while the Q-Q plot implies the possible presence of outliers. The return level plot is shown without the confidence interval.

non-stationary stochastic process of extreme values with almost all observations falling into the calculated prediction interval. The three possible outliers noted in section 6.2.3 can be seen here as well.

In contrast the previous approach in earlier chapters, the interpretation is straightforward since the model is fitted to the original monthly maxima time series. That is, the extreme values of SO_2 follow GEV distribution whose parameter changes over time as per the equations given in 6.2.1. Unlike the typical approach mentioned in section 6.1.2, Model A incorporates the trend via the time-variant scale parameter σ_t instead of the location parameter μ_t. This allows us to si-

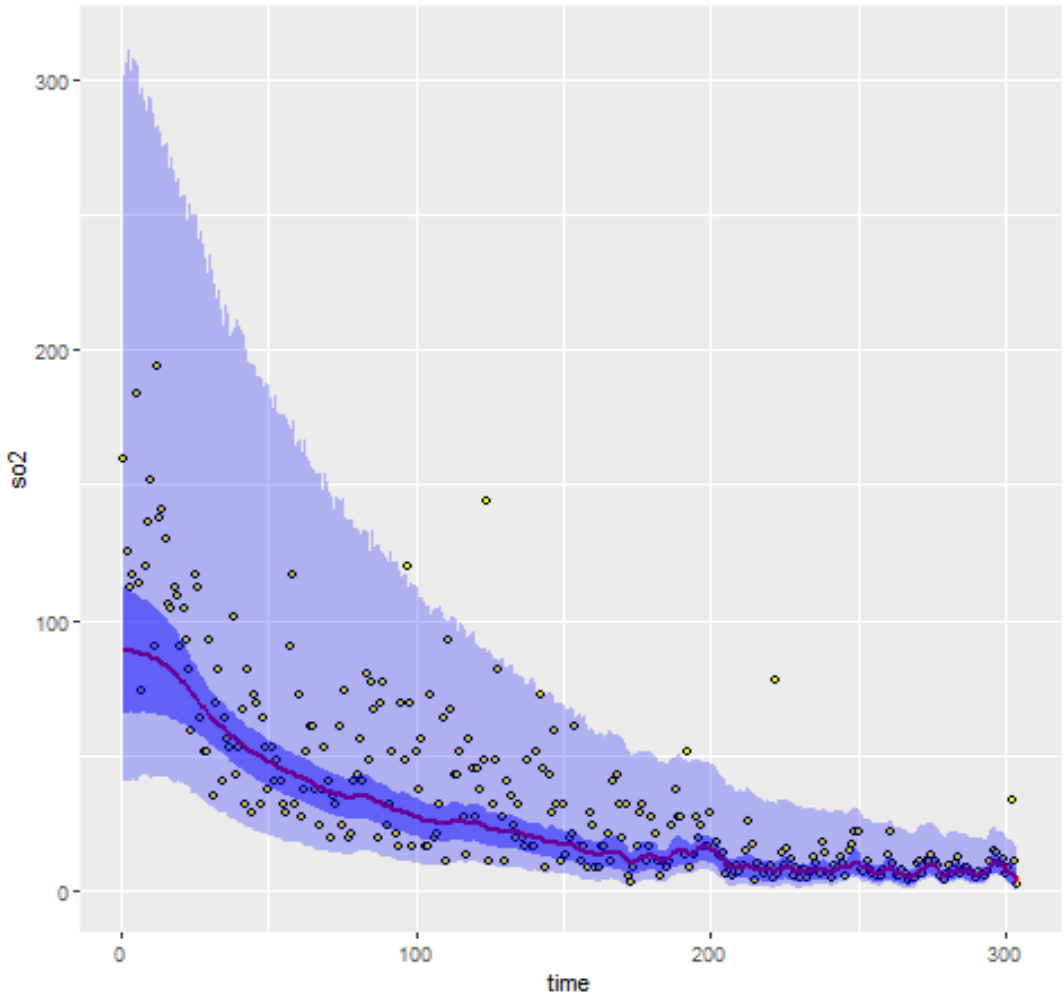


Figure 6.3: The prediction interval given by the state space model along with the observations of SO_2 (monthly maxima). The red line is the estimated mean process of GEV distribution (see section 6.2.1) with prediction interval shaded with darker blue. The prediction interval of the entire stochastic process is shaded with lighter blue.

multaneously model the trend and varying structure of variance. Interestingly, the simple EDA reveals that the tendency of decreasing variance can be seen in the almost all the pollutants recorded in the same site including CO, NO_2 and pm10, etc, suggesting that the model can be widely applicable to analysing the air pollution in general.

6.3 O_3

6.3.1 Model specification

Recall the time series plot of O_3 shown in Figure 2.1. The characteristic of O_3 is cyclic pattern due to the seasonal effect. It should also be noted that the data shows little sign for the trend over time or variation in the variance structure. It turns out that the state space modeling of extreme values with seasonal effect is quite a challenging task. Two different models are discussed in this chapter to address the practical issues that arises in this attempt, which will later be discussed in full detail. The first model we suggest is

Model B-1: GEV model with seasonal effect:

$$\begin{aligned} \mu_t &= \mu_{t-1} + s_t + \epsilon_t, \\ \sum_{i=0}^{11} s_{t-i} &\sim N(0, \gamma_s), \\ \epsilon_t &\sim \mathrm{N}(0, \gamma), \\ y_t &\sim \mathrm{GEV}\left(\mu_t, \sigma, \xi\right). \end{aligned}$$

The second line is based on the idea that the seasonal variation occurs in the 12 month cycle and that the variation of this seasonal effect itself is normally distributed. σ and ξ is assumed to be constant, and the default Uniform prior is chosen for γ and γ_s. The second model we suggest is

Model B-2: GEV model with air temperature as covariate:

$$\begin{aligned} \mu_t &= s_t + \epsilon_t, \\ s_t &= \beta x_t, \\ \epsilon_t &\sim \mathrm{N}(0, \gamma), \\ y_t &\sim \mathrm{GEV}\left(\mu_t, \sigma, \xi\right), \end{aligned}$$

where x_t is the air temperature measured at the time of observations t. (As mentioned in section 2.3, the covariate contains missing values, which are imputed by monthly mean.) In this model, following the conventional approach, the time-variant location parameter μ_t is conditioned on the seasonal term s_t, which is a linear function of coveriate with single coefficient β to be estimated. Again, σ and ξ is assumed to be constant, and again the default Uniform prior is chosen for γ. Compared to Model B-1, there is a reduction in dimension of parameter space, which is expected to make the samplings less challenging.

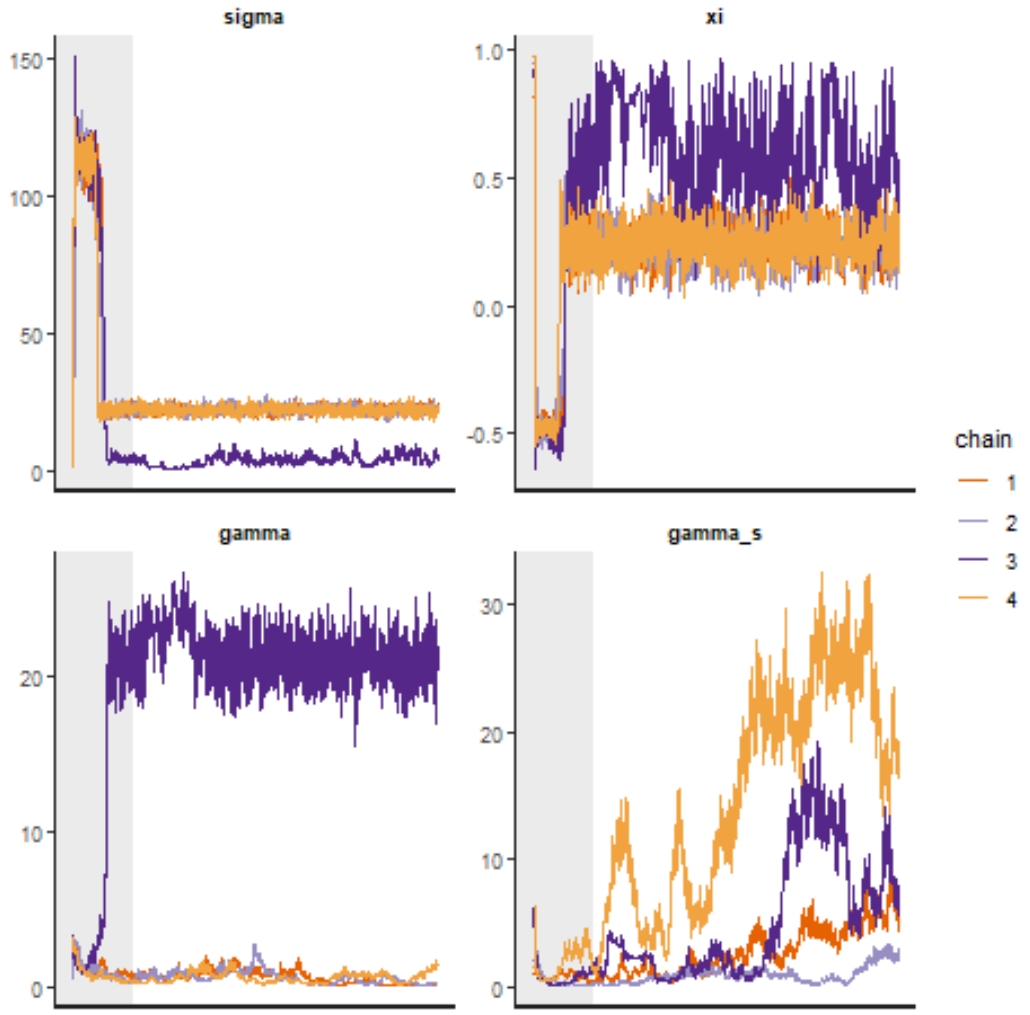


Figure 6.4: The trace plot of sampling for Model B-1, with respect to four parameters σ, ξ, γ and γ_s. The four chains of samplings are conducted, each with 5000 iterations and 1000 warm-ups.

6.3.2 Convergence diagnostics

Figure 6.4 shows the trace plots for Model B-1. The plot does not indicate the successful convergence of HMC samplings. The largest value of Rhat is 3.48, and therefore the parameter estimates obtained by the sampling is not to be trusted. The plots, however, do evoke some insights that could be worth reflecting upon, which is discussed in detail in section 6.3.3.

Figure 6.5 shows the trace plot of Model B-2. Although the plots imply the successful convergence of samplings, the warm-ups for σ and β appear to have taken longer than initially anticipated. In addition, the other diagnostic indicators present several issues which are fully discussed in section 6.3.5.

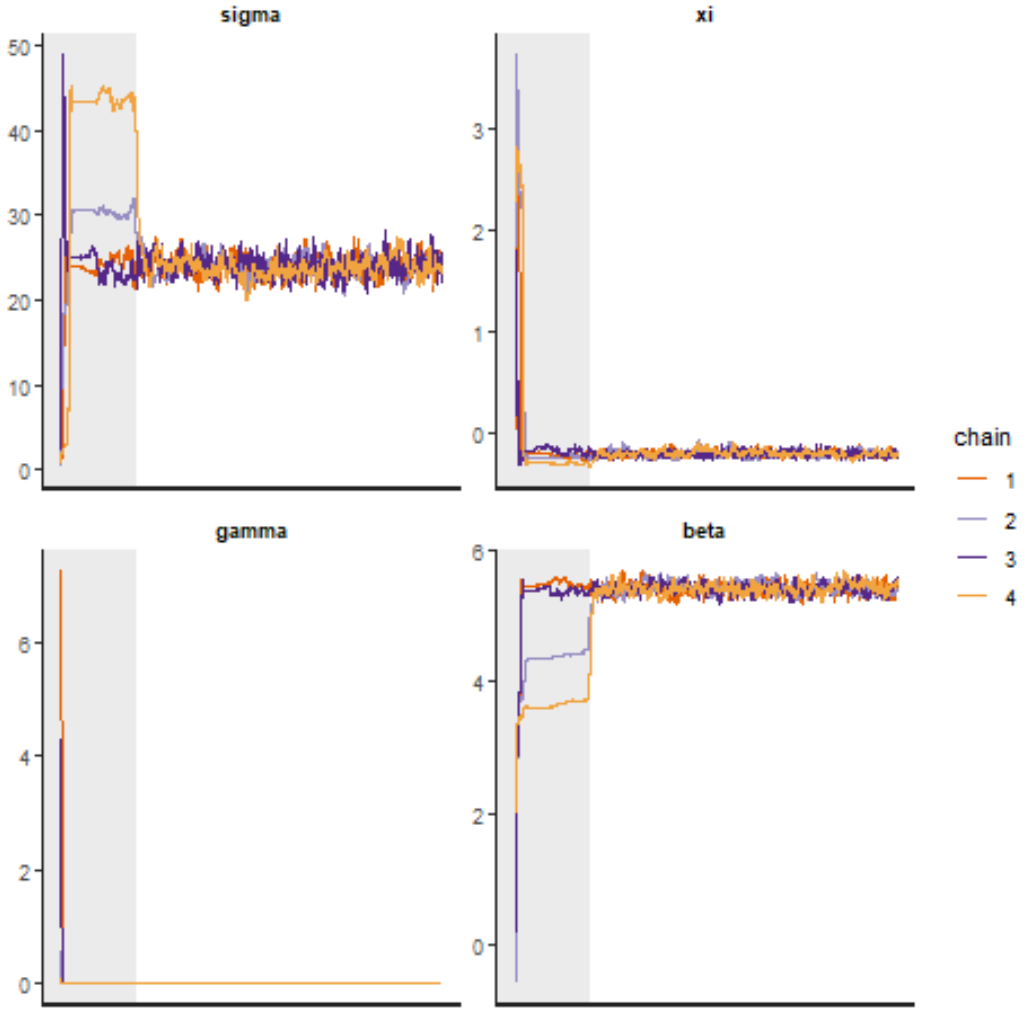


Figure 6.5: The trace plot of sampling for Model B-2, with respect to four parameters σ, ξ, γ and β. The four chains of samplings are conducted, each with 5000 iterations and 1000 warm-ups.

6.3.3 Shape parameter and difficulty in samplings

Though the implementation of Model B-1 is not quite successful, the trace plot sheds light on the root cause of difficulty in the samplings i.e. the shape parameter ξ. It is known that ξ has an determining effect on the behaviour of GEV distribution, affecting the mean, variance and tail behaviour, etc. In fact, deviation of Chain 2 is clearly affecting the sampling of other parameters in the same chain, showcasing that the uncertainty in ξ cause a detrimental effect on samplings and parameter inference.

Given this, any prior information that may narrow down the probable range of this parameter should be utilised, if at all possible. For example, Calafat and Marcos (2020) conducts the extensive EDA to gain some prior knowledge

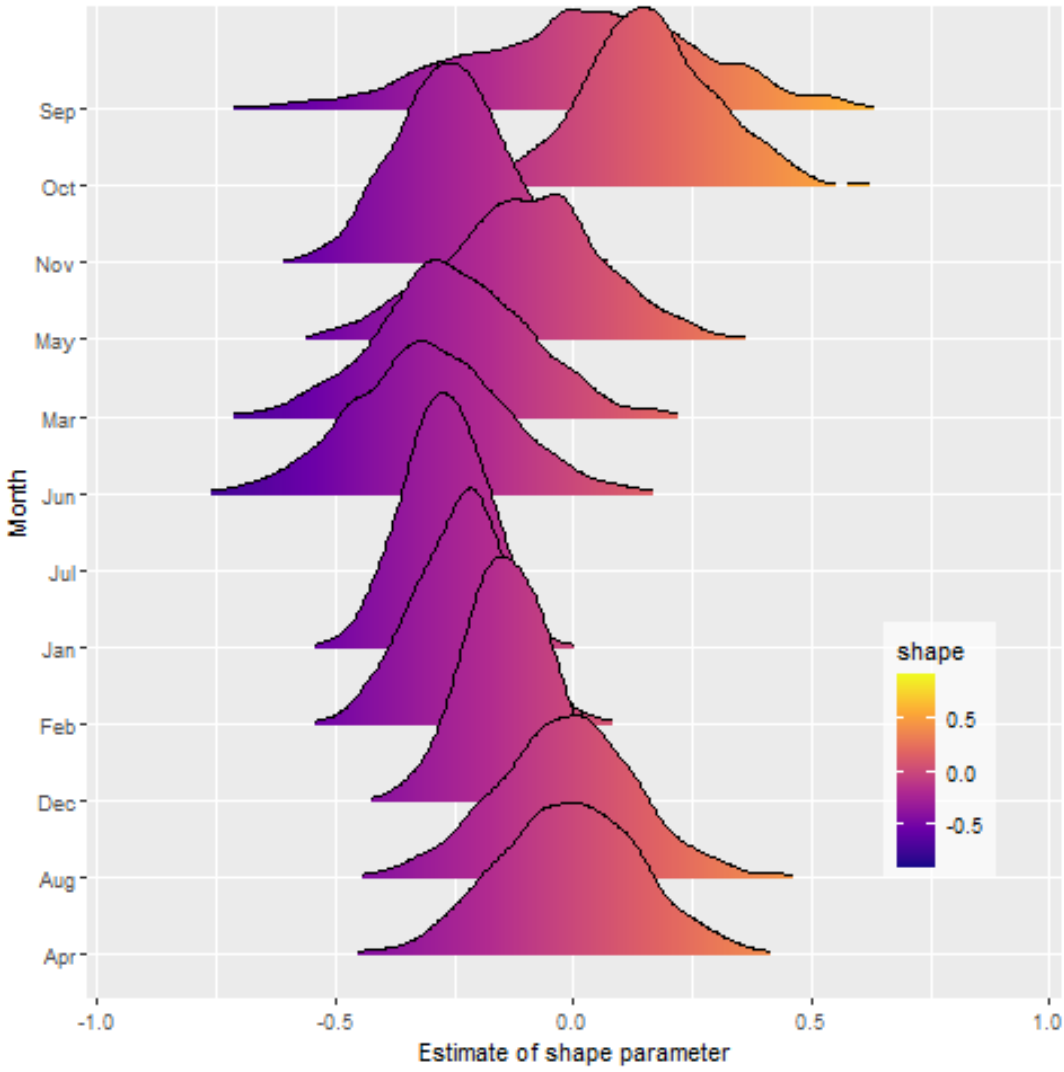


Figure 6.6: Posterior distributions of $\hat{\xi}$ with respect to each month. The estimates are subject to bias as the sample size for each model is reduced to around 25 observations after splitting the data into 12 monthly division. The estimate for July data is outside the graph marking 662.2, which is unlikely to be true. The plot nevertheless gives an idea regarding the probable range of ξ.

about parameters in order to justify their model assumptions that ξ is spatially constant. The similar EDA, though not equally thorough, can be conducted for our model; Figure 6.6 shows the 95% CI of ξ estimated by GEV model fitted to monthly data, respectively. The underlying assumption here is that within the same month O_3 nearly satisfies i.i.d assumption, which justifies the model fitting without transforming the data. This gives a fair guess that ξ sits in the interval between $[-0.75, 0.75]$. Unfortunately, even after including this knowledge as a parameter constraint, the samplings for Model B-1 are not effectively conducted. Indeed, it may be the fact that there is much density assigned to both sides of the parameter space across the critical region that makes the samplings challenging.

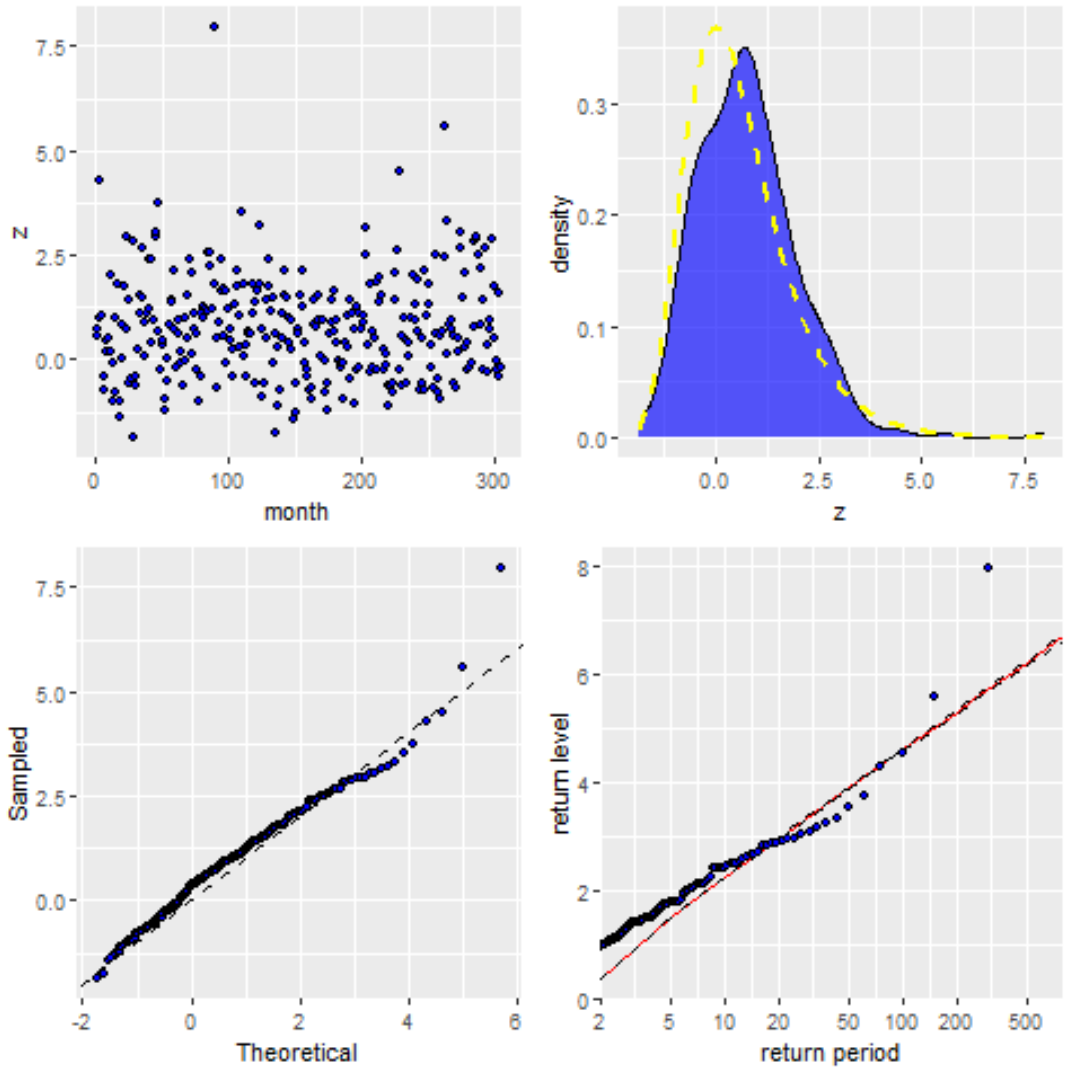


Figure 6.7: The diagnostic plots for Model B-2. There is no apparent structure in the time series plot, indicating the successful modeling of time-variant components. Both of the density plot and Q-Q plots indicate the reasonable agreement. The return level plot again shows a reasonable linearity except a few cases at high return period, which may be considered as outliers.

6.3.4 Model diagnostics

Given the convergence diagnostic in section 6.3.2, there is little point in further discussing the model diagnostics or parameter inference with regard to Model B-1. Thus the following discussions only concerns Model B-2. The diagnostic plots for model fit is provided in Figure 6.7.

The model fit again shows remarkable improvement compared to the corresponding model in Chapter 3. Additionally, there is great chance that the slight disagreement that can be seen in the diagnostics plots are attributed to the bias caused by the imputation.

6.3.5 Parameter inference and uncertainty

Table 6.2 provides the result of parameter inference for Model B-2.

	Estimates		95% CI			Convergence	
	mean	sd	0.025	0.5	0.975	n_eff	Rhat
$\hat{\mu}_1$	106.744 (0.119)	1.625	103.679	106.674	110.046	184.153	1.015
$\hat{\mu}_{305}$	112.514 (0.126)	1.713	109.283	112.440	115.994	184.153	1.015
$\hat{\sigma}$	23.964 (0.082)	1.179	21.789	23.932	26.385	205.788	1.007
$\hat{\xi}$	-0.190 (0.002)	0.030	-0.241	-0.192	-0.124	186.565	1.007
$\hat{\gamma}$	0.000 (NaN)	0.000	0.000	0.000	0.000	NaN	NaN
$\hat{\beta}$	5.409 (0.006)	0.082	5.254	5.406	5.577	184.1533	1.015

Table 6.2: The result of parameter inference for Model B-2 fitted to O_3. For the time-varying parameter μ_t, only the estimates at the first and the last points are shown.

As mentioned in section 6.3.2, there are a couple of concerns regarding the indicator of convergence diagnostics. Firstly, n_eff shows an undesirably low value for all parameters given the total sampling size of 16,000. Although it is no definite criteria, n_eff less than one tenth of the total sample size may indicate the ineffective samplings. Secondly, n_eff and Rhat for γ, the scale parameter of noise process, produce NaN. However, this is because the parameter is estimated to be almost 0, not necessarily due to the unsuccessful samplings. The estimate itself is most likely to be biased as a result of the naive imputation; as shown in section 6.3.1, the location parameter μ_t is affected by the value of single covariate (i.e. air temperature) whose variance is unduly low during the period where the value is imputed with monthly mean. It is then natural for γ to be underestimated.

6.3.6 Prediction

Lastly, the prediction interval obtained by the estimated parameters is shown in Figure 6.8. Again, the result is quite satisfactory. Given that the estimate $\gamma \approx 0$ is possibly biased, the predictive interval may be narrower than reality. It also presents a nearly homogeneous cyclic pattern during the period of mean imputation. There is one observation that appears to be an outlier, which corresponds to the outlook of the diagnostic plots given in section 6.3.4.

The observations nevertheless mostly fall within the interval, suggesting that the covariate is highly effective in modeling the target variable O_3. In fact, there is a plenty of literature such as Coates et al. (2016) that provides the evidence of relation between the air temperature and the production of O_3. The same simple EDA plots as shown in Figure 2.6. are useful to investigate the impact of covariates on a particular pollutant, which then can be incorporated to model the seasonality if it seems appropriate. Though not shown here, there

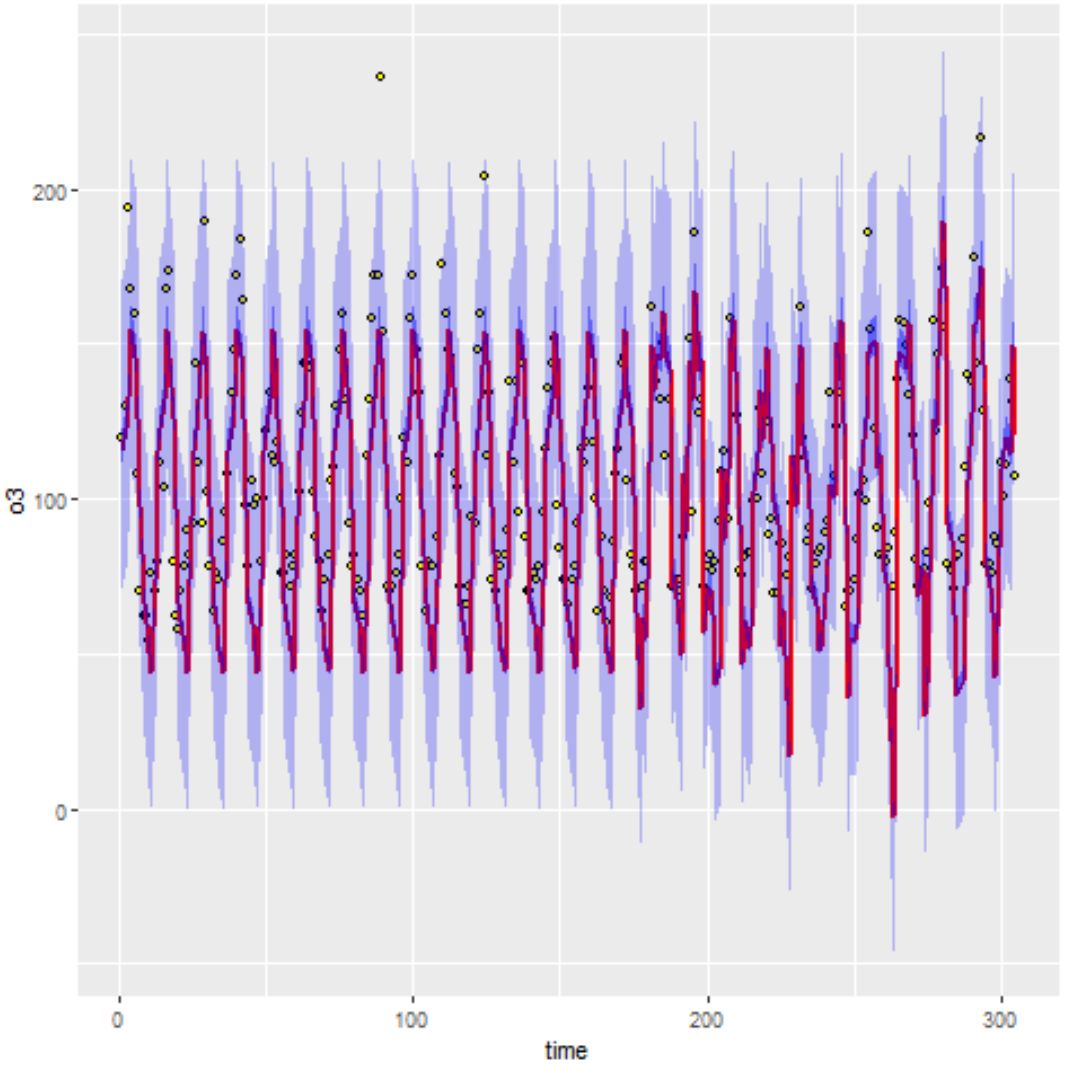


Figure 6.8: The prediction interval given by Model B-2 along with the observations of O_3 (monthly maxima). The red line is the estimated mean process of GEV distribution (see 6.2) with prediction interval shaded with darker blue. The prediction interval of the entire stochastic process is shaded with lighter blue.

are several other pollutants observed in the site that appear to correlate with air temperature. This again suggests that the model can generally be applied to a wide range of pollutants, provided that the appropriate seasonal covariates are introduced.

6.4 Conclusion

In this chapter, we explored the state space modeling as an advanced method to model the non-stationary stochastic process of extreme values. The modeling is conducted using the HMC sampling implemented via Stan. The results indicate the considerable utility of the modeling approach presenting the clear improvements from the perspective of model fit and interpretablity, compared to the previous approach discussed in earlier chapters.

Along with the preliminary discussion in Chapter 5, this chapter effectively highlights the methodology of EVA within the framework of Bayesian statistical modeling, which can be applicable when implementing the models with even more complex structures. We faced computational difficulties in sampling from posteriors when the model included seasonal effects. Although the issue can be resolved by introducing the appropriate covariate i.e. air temperature, further investigating the alternative modeling structure is of great interest. Additionally, integrating the spatial information into the modeling structure merits a great deal of consideration especially given the abundance of the data set (see section 1.4). Calafat and Marcos (2020) uses the spatio-temporal model that successfully incorporates spatial dependence of the GEV parameters achieving the remarkable reduction in bias compared to site-by-site modeling.

Appendix A

Return level plot

The return level plot is the visualisation technique commonly used in EVA, and it is particularly useful for both model diagnostics and extrapolation.

The construction of the plot is quite straightforward. Recall from Chapter 3 the CDF of GEV distribution

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

Inverting (A.1) yields

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}[1-\{-\log(1-p)\}]^{-\xi} & \text{for } \xi \neq 0, \\ \mu - \sigma\log\{-\log(1-p)\} & \text{for } \xi = 0, \end{cases}, \tag{A.2}$$

where $G(z_p) = 1-p$. Here z_p is often referred to as return level associated with return period $1/p$. If the GEV model is fitted to monthly maxima, then z_p is exceeded in any particular month with probability of p or once in $1/p$ months. The return level plot is return levels plotted against corresponding return period. As shown in Coles (2001), the confidence interval of return levels can be obtained by using the delta method or the profile likelihood.

The relation between the plot and the shape parameter ξ is worth noting. If $\xi < 0$ the plot is convex with asymptotic limit as $p \to 0$ at $\mu - \sigma/\xi$. If $\xi > 0$ the plot is concave and has no finite bound. The plot is linear if $\xi = 0$. This is one of the reason why the estimate of ξ is the primary interest of statistical inference in EVA.

Similarly, if we consider GP distribution with its CDF given by

$$\mathrm{H}(y) = 1 - \left(1 + \xi \frac{y}{\sigma}\right)^{-\frac{1}{\xi}}, \tag{A.3}$$

where $y = (X - u)$ conditioned that $X > u$, it then follows that

$$\begin{aligned} \Pr(X > x | X > u) &= \left[1 + \xi \left(\frac{x-u}{\sigma}\right)\right]^{-\frac{1}{\xi}} \\ \Rightarrow \Pr(X > x) &= \zeta_u \left[1 + \xi \left(\frac{x-u}{\sigma}\right)\right]^{-\frac{1}{\xi}}, \end{aligned} \tag{A.4}$$

where $\zeta_u = \Pr\{X > u\}$. Given this, the return level that is exceeded on average once every m year is the solution of

$$\zeta_u \left[1 + \xi \left(\frac{x_m - u}{\sigma}\right)\right]^{-\frac{1}{\xi}} = \frac{1}{m}. \tag{A.5}$$

By rearranging (A.4) we obtain

$$x_m = \begin{cases} \mu + \frac{\sigma}{\xi}[(m\zeta_u)^{\xi} - 1] & \text{for } \xi \neq 0, \\ \mu + \sigma \log(m\zeta_u) & \text{for } \xi = 0. \end{cases} \tag{A.6}$$

x_m is said to be the m-observations return level. Again plotting x_m against the corresponding return periods yields the same return level plot that is based on GEV distribution. The resulting plot can be interpreted in the same manner.

Appendix B

AIC

The Akaike information criterion (AIC) is named after Hirotsugu Akaike, a Japanese statistician, who formulated it in Akaike (1974). It is founded on information theory and often used as a criteria for model selection. The AIC is defined to be the unbiased estimator of average mean log-likelihood of the model.

B.1 K-L divergence and MLE

The derivation of AIC can be effectively discussed in the context of the Kullback-Leibler (K-L) divergence. K-L divergence is a measure of how one probability distribution is different from another, reference probability distribution. It is defined as

$$\begin{aligned} \mathrm{K}(g; f) &= \mathbb{E} \log \frac{g(X)}{f(X)} \\ &= \int_{-\infty}^{\infty} \log \left\{ \frac{g(X)}{f(X)} \right\} g(x)\, dx, \\ &= \int_{-\infty}^{\infty} \log \{g(X)\}\, g(x)\, dx - \int_{-\infty}^{\infty} \log \{f(X)\}\, g(x)\, dx, \end{aligned} \tag{B.1}$$

where $f(X)$ may represent a model distribution and $g(X)$ a target distribution. In the last line of (B.1), the second term

$$\int_{-\infty}^{\infty} \log \{f(X)\}\, g(x)\, dx \tag{B.2}$$

is known as mean log-likelihood. For the model approximation to be good, the parameter inference for $f(X)$ should be conducted such that $\mathrm{K}(g; f)$ is minimised. This task simply amounts to maximising (B.2) since the other term is a

given constant and thus irrelevant when evaluating $\mathrm{K}(g; f)$.

The issue here is that true value of (B.2) cannot be known since $g(X)$ is often unknown in practice. It is then natural to estimate this by using the data obtained. Recall the law of large Number (LLN);

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X]\right| \geq \epsilon\right) \to 0. \tag{B.3}$$

Then, in the virtue of the LLN, (B.2) can be approximated as

$$\begin{aligned}\int_{-\infty}^{\infty} \log\{f(X)\}\, g(x)\, dx &= \mathbb{E}\log f(X) \\ &\approx \frac{1}{n}\sum_{i=1}^{n} \log f(x_i).\end{aligned} \tag{B.4}$$

Notice that $\sum_{i=1}^{n} \log f(x_i)$ in the last term of (B.4) is nothing but the log-likelihood. In other words, minimising $\mathrm{K}(g; f)$ is asymptotically equivalent to maximising the log-likelihood, linking the maximum likelihood estimator (MLE) to K-L divergence.

B.2 Average mean log-likelihood

In the previous section, we see how K-L divergence is linked to the maximum likelihood method. That is, if a single model with a set of parameter is given, K-L divergence of the model can be asymptotically minimised by choosing the parameters such that the log likelihood of the model is maximised. Following discussion concerns the case where multiple candidate models are given and thus the model selection is the subject of interest. To discuss this we introduce average mean log-likelihood.

We consider the i.i.d random variables $\boldsymbol{X} = \{X_1, ..., X_n\}$ with common distribution $g(\cdot) = f(\cdot|\boldsymbol{\theta}^*)$, where $\boldsymbol{\theta}^*$ denotes a vector of true parameters. Let the realisation from this distribution denoted as $\boldsymbol{x} = \{x_1, ..., x_n\}$. Suppose that the model f can be determined by up to K parameters i.e. $\boldsymbol{\theta}^* = (\theta_1^*, ..., \theta_K^*)$. Then the parameter space for this model can be conditioned such that the model f only has k parameters i.e.

$$\boldsymbol{\Theta}_k = \{\boldsymbol{\theta} \in \Theta_K | \theta_{k+1} = ... = \theta_{K-1} = \theta_K = 0\}. \tag{B.5}$$

Then there must be the smallest value of k s.t. $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}_k$, and such k is said to be the true number of parameters. When the data $\boldsymbol{x}$ is obtained, the model f with k parameters has the log-likelihood of

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(x_i|\boldsymbol{\theta}), \tag{B.6}$$

whereby the MLE

$$l(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}) = \max_{\theta \in \Theta_k} l(\boldsymbol{\theta}) \tag{B.7}$$

can be obtained. Using a random variable Z with identical distribution with but independent from $\boldsymbol{X}$, the mean likelihood of the model $f(|\boldsymbol{\theta})$ is given by

$$\begin{aligned} \mathbb{E}_z\{\log f(Z|\boldsymbol{\theta})\} &= \int \log\left\{f(z|\boldsymbol{\theta})\right\} f(z|\boldsymbol{\theta}^*)\, dz. \\ \Rightarrow l^*(\boldsymbol{\theta}) &= n\, \mathbb{E}_z\{\log f(Z|\boldsymbol{\theta})\} \end{aligned} \tag{B.8}$$

Here, $l^*(\boldsymbol{\theta})$ denotes the mean log-likelihood of the model with parameter $\boldsymbol{\theta}$ times n. As discussed in section B.1, the model approximation is successful when $l^*(\boldsymbol{\theta})$ takes a large value. Note that $l^*(\boldsymbol{\theta})$ itself is random variable depending on the data $\boldsymbol{x}$. Thus by taking expectation with respect to $\boldsymbol{x}$, we define the average mean log-likelihood of the model as

$$\begin{aligned} l_n^*(k) &= \mathbb{E}_X[l^*(\hat{\boldsymbol{\theta}}_k)] \\ &= \int l^*(\hat{\boldsymbol{\theta}}_k) \prod_{i=1}^{n} g(x_i) d\boldsymbol{x}, \end{aligned} \tag{B.9}$$

where $l^*(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}})$ denotes the mean log-likelihood of the maximum likelihood model times n. $l_n^*(k)$ is quantity only dependent on the number of observations, true model $g(X)$ and the model used for approximation. Given the argument so far, the model selection can be conducted such that the model with largest value of $l_n^*(k)$ gets selected.

B.3 AIC

Though the proof is omitted here, it is known that $l(\hat{\boldsymbol{\theta}}_K) - K$ is asymptotically the unbiased estimator for the average log-likelihood estimator discussed in the previous section. The AIC is defined as

$$\mathbf{AIC} = -2l(\hat{\boldsymbol{\theta}}_K) + 2K. \tag{B.10}$$

That is, in the model selection the model with smallest value of AIC gets selected.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. (2005). *Statistics of Extremes: theory and applications*. Chichester.

Caballero-Megido, C., Hillier, J., Wyncoll, D., Bosher, L., and Gouldby, B. (2017). Technical note: comparison of methods for threshold selection for extreme sea levels. *Journal of Flood Risk Management*, 11(2):127–140.

Calafat, F. and Marcos, M. (2020). Probabilistic reanalysis of storm extremes in europe. *Proceedings of the National Academy of Sciences*, 117(4):1877–1883.

Cheng, L., AghaKouchak, A., Gilleland, E., and Katz, R. (2014). Non-stationary extreme value analysis in a changing climate. *Climate Change*, 127(2):353–369.

Coates, J., Mar, K., Ojha, N., and Butler, T. (2016). The influence of temperature on ozone production under varying NO_x conditions - a modelling study. *Atmospheric Chemistry and Physics*, 16(18):11601–11615.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Value*. Springer.

Gilleland, E. and Katz, R. W. (2016). extRemes 2.0: An extreme value analysis package in R. *Journal of Statistical Software*, 72(8):1–39.

Gumbel, E. (1958). *Statistics of Extremes*. Columbia University Press.

Konishi, S. and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. Springer.

Ouarda, T. and Charron, C. (2019). Changes in the distribution of hydro-climate extremes in a non-stationary framework. *Scientific Reports*, 9(1).

Towler, E., Rajagopalan, B., Gilleland, E., Summers, R., Yates, D., and Katz, R. (2010). Modeling hydrologic and water quality extremes in a changing climate: A statistical approach based on extreme value theory. *Water Resources Research*, 46(11).

Vehtari, A. (2017). Extreme value analysis and user defined probability functions in stan. https://mc-stan.org/users/documentation/case-studies/gpareto_functions.html. Accessed: 2021-08-1.