Volcanic hazard assessment for the Alaska region using extreme value theory

E. Ramona Stefanescu*

January, 2012

1 Introduction

The size of a volcanic eruption is an essential component of risk assessment and extreme value models are a natural candidate for modeling such phenomena. Volcanic eruptions represent a serious threat on the people leaving near a volcano, especially when their perception of risk is negatively influenced by a large repose time, or by the lack of clear evidence of major past activity [6, 7]. In such regions, various criteria need to be taken into account in taking evacuation or relocation decisions: the morphology of the volcano, the likelihood of a future eruption, and the plausible values of an eruption magnitude. Most of the time, direct application of database of volcanic events is complicated by an under-recording of the historical events. In the literature was reported that some volcanoes show stationarity patterns of activity [4], while others show time-dependent eruption rates. Nevertheless, combining the eruptions of large groups of volcanoes generates a definite homogeneous Poissonian behavior. Under this assumption we will use the dataset for the volcanoes situated in the Alaska region. The Global Volcanism Program database for Volcanoes of Alaska currently contains 93 volcanoes, with historical volcanic activity reported starting with 1741.

2 Historical records of volcanic eruptions in the Alaska region

A volcanic eruption is characterized by its impact or effect on the region, total mass or energy release and the rate of mass. The Volcanic Explosivity Index (VEI) is the quantity that characterizes eruption based on these parameters. In this study we will use this VEI to estimate the volcanic hazard of future eruptions. Figure 2 shows the data used and it includes the eruptions documented between 1900 and 2010. From the figure we see that the rate of activity in the last 30 years is less than in the previous 70 years, which it is in

*Department of Mechanical and Aerospace Department, University at Buffalo, (ers32@buffalo.edu).

agreement with the finding that the rate of activity has been more or less constant over the period. Also for Fig 1 we can conclude that the frequency of the events decreases as their size or magnitude increases.

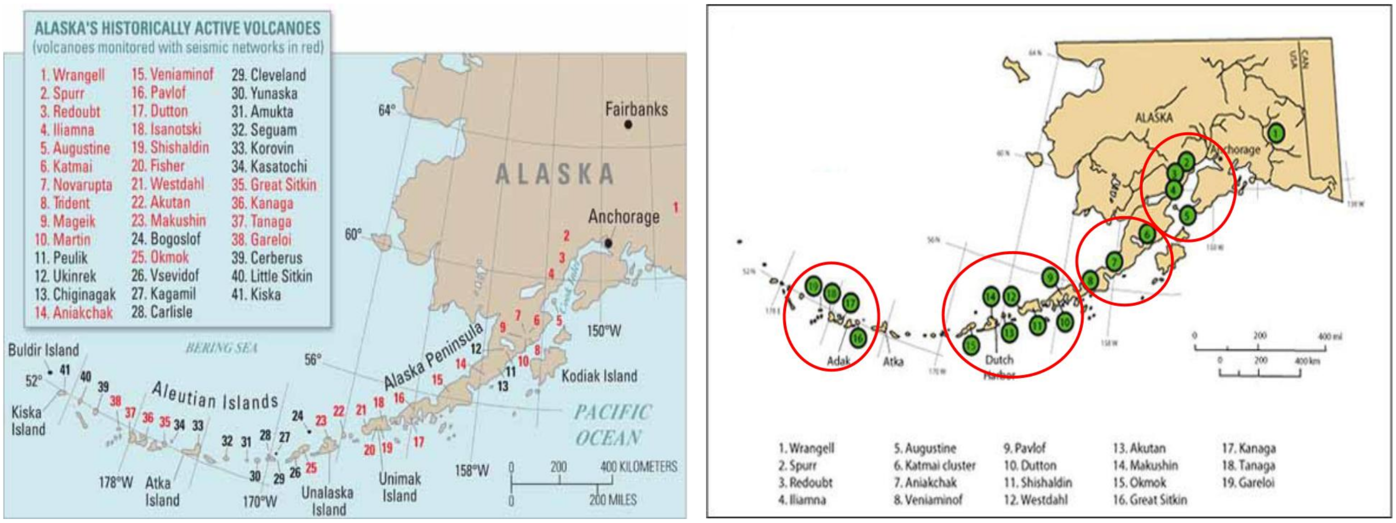

Figure 1: a) Alaska's active volcanoes (USGS) b) Cluster based on the geophysics (AVO)

3 Estimation of volcanic hazard using extreme-event statistics

3.1 Overview

Many of the extreme value models are assuming an underlying process consisting of a sequence of independent random variables. However, for the type of data to which extreme value models are commonly applied, temporal independence is usually an unrealistic assumption. The most natural generalization of a sequence of independent random variables is to a stationary series. Stationarity, which is a more realistic assumption for many physical processes, corresponds to a series whose variables may be mutually dependent, but whose stochastic properties are homogeneous through time. Usually, extreme events are close to independent at times that are far enough apart. It was observed that the occurrence of the volcanic eruptions is similar to that of earthquakes in the sense that the frequency of the events decreases as their size increases. It was found that the distribution of magnitudes in a region can be described by the frequency-magnitude distribution of earthquakes:

$$\log N = a - b \cdot M \tag{3.1}$$

where N is the cumulative number of earthquakes with magnitude $\geq$ M, and a and b are constants that describe the power law decay of occurrences with increasing magnitude over a given time interval. It was estimated the values $q = 3.494$ and $b = 0.789$ for the global volcanic activity based on the historical eruption data in the VEI range $3 - 6$. Later, it was integrated eruption data for various time intervals: 20, 200, 1000 and 2000 yr.

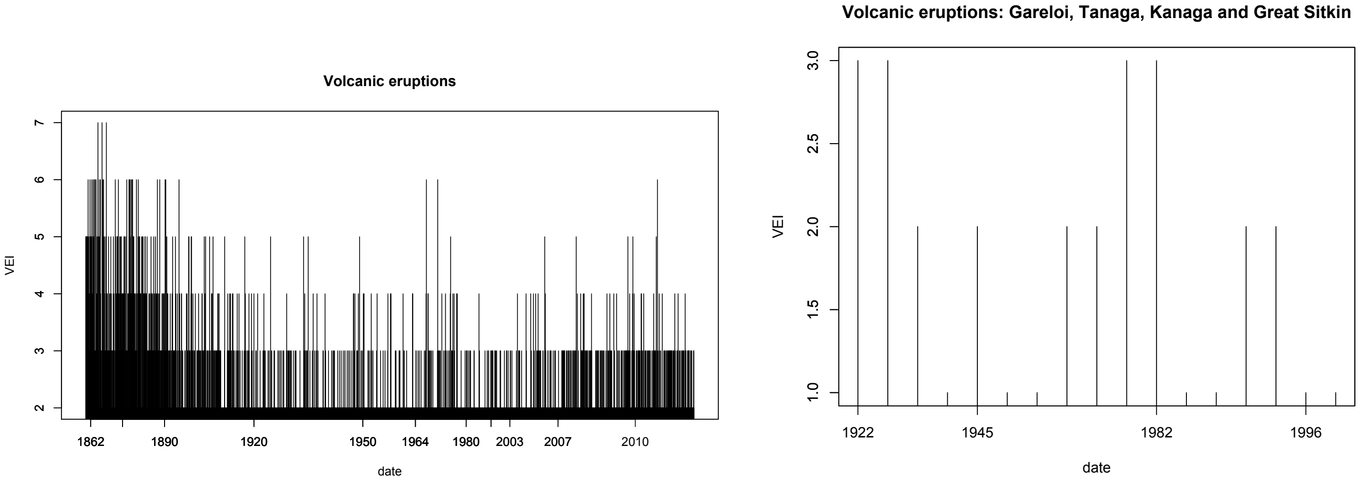

Figure 2: Volcanic activity from 1860 to 2012 b) Volcanic activity of cluster 1

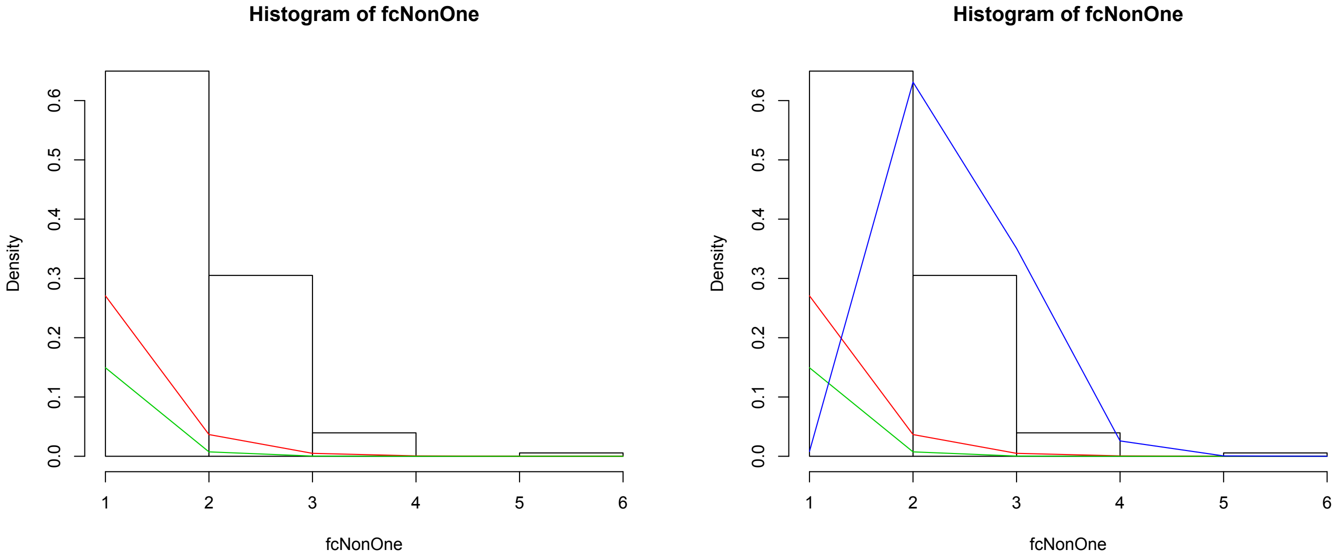

Figure 3: a) Fitted gamma distribution b) Optimal fitted gamma distribution

Let's consider a series of IID random variables $X_1, X_2, \ldots, X_n$ with distribution function

F. The maximum of these random variables is M_n. Then if there exists constants $a_n > 0$ and b_n such that:

$$Pr\left(\frac{M_n - b_n}{a_n} \leq x\right) \rightarrow G(x) \tag{3.2}$$

as $n \rightarrow \infty$ and G is a non-degenerate distribution then G belongs to the family of extreme value distributions [2].

Generalised Extreme Value (GEV) distribution

The GEV distribution is the limiting distribution of the minimum or maximum of a series of IID random variables. This distribution combines three extreme values distributions (Gumbel, Fr$\acute{e}$chet and Weibull) and characterizes them with different values of ξ which is the shape parameter of the GEV distribution [3, 5].

$$G(x;\mu,\sigma,\xi) = exp\left\{\left[-1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_{+}^{1/\xi}\right\} \tag{3.3}$$

where $1 + \xi\dfrac{x-\mu}{\sigma} > 0$, the location parameter $\mu \in \mathcal{R}$ and the scale parameter $\sigma > 0$.

The three different distributions (Gumbel, Fr$\acute{e}$chet and Weibull) can be distinguished based on the choice of the shape parameter ξ which determines the tail of the distribution:

- If $\xi > 0$, the distribution is Fr$\acute{e}$chet which has a heavy upper tail allowing for an increased probability of extreme values.
- If $\xi = 0$, the distribution is Gumbel which has an exponential tail.
- If $\xi < 0$, the distribution is Weibull which has a tail with a finite upper limit.

Often when examining extreme data we are interested in asking the question, "How often can we expect to observe extreme events?" For example how often do we expect a volcano to erupt and how big will be the eruption. In order to answer these questions we can calculate the return level for a given return period. The return period is the amount of time we expect to wait before observing an extreme event, and the return level refers to the intensity of the event that occurs within the return period [1].

The return level z_p for a return period $1/p$ is calculated from the GEV distribution where $0 < p < 1$ is the $1-p$ quantile of the GEV distribution.

$$\hat{z_p} = \begin{cases} \hat{\mu} -]\dfrac{\hat{\sigma}}{\hat{\xi}}\left[1 - \{-log(1-p)\}^{-\hat{\xi}}\right] & for\ \hat{\xi} \neq 0 \\ \hat{\mu} - \hat{\sigma}\log\{-\log(1-p)\} & for\ \hat{\xi} = 0 \end{cases} \tag{3.4}$$

where μ, σ and ξ are the maximum likelihood values as found from fitting the GEV distribution to the data.

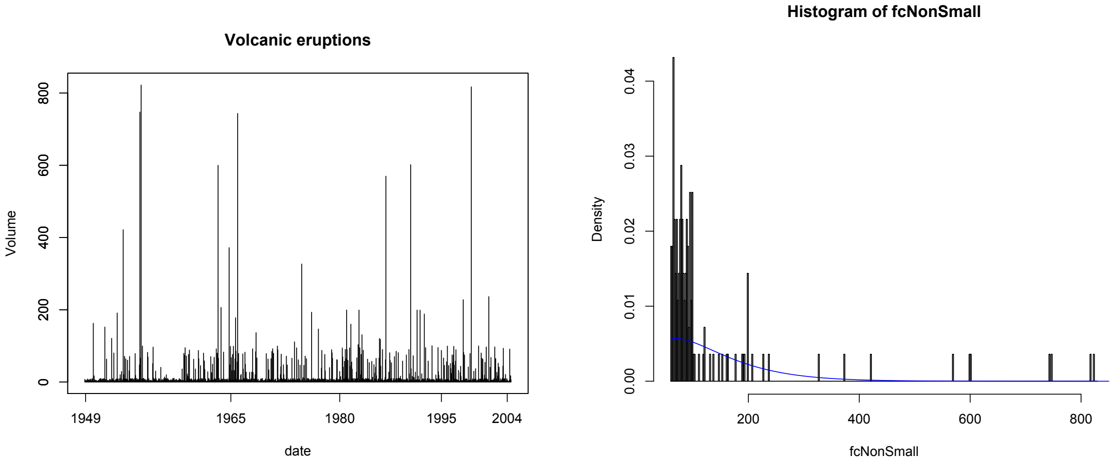

Figure 4: Made-up volume data from VEI b) Fitted gamma for volume $> 60m^3$

4 Methodology

Based on this observation we are using two different approaches in estimating the volcanic hazard: univariate block-maximum approaches and threshold-exceedance approaches.

4.1 Block maxima

Block maxima data is data which is separated into blocks of days, months or years and from these blocks the maximum of each block is recorded. The first step consist in blocking the data into sequences of n observations, n being sufficiently large. The maxima Z_i of each block i is calculated and the GEV distribution is fitted to this series of block maxima $Z_1, Z_2, \dots, Z_m$. The choice of the length of blocks implies a trade off between bias and variance. When the length of the block is small, then the approximation of the distribution by the limit is quite poor and this is leading to bias in estimation and extrapolation. Long blocks on the other hand generate only few data leading to large estimation variance.

The method most commonly used to estimate the parameters is the likelihood method. By denoting $Z_1, Z_2, \dots, Z_m$ the block maxima and under the assumptions that they are independent variables having a GEV distribution, the log-likelihood for the GEV when $\xi \neq 0$ is:

$$l(\mu, \sigma, \xi) = -m \log \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{m} \log\left(1 + \xi \frac{z_i - \mu}{\sigma}\right) - \sum_{i=1}^{m} \left(1 + \xi \frac{z_i - \mu}{\sigma}\right)^{-1/\xi} \tag{4.1}$$

provided that $\left(1 + \xi \dfrac{z_i - \mu}{\sigma}\right) > 0$ for $i = 1, \dots, m$. When this condition is not satisfied then

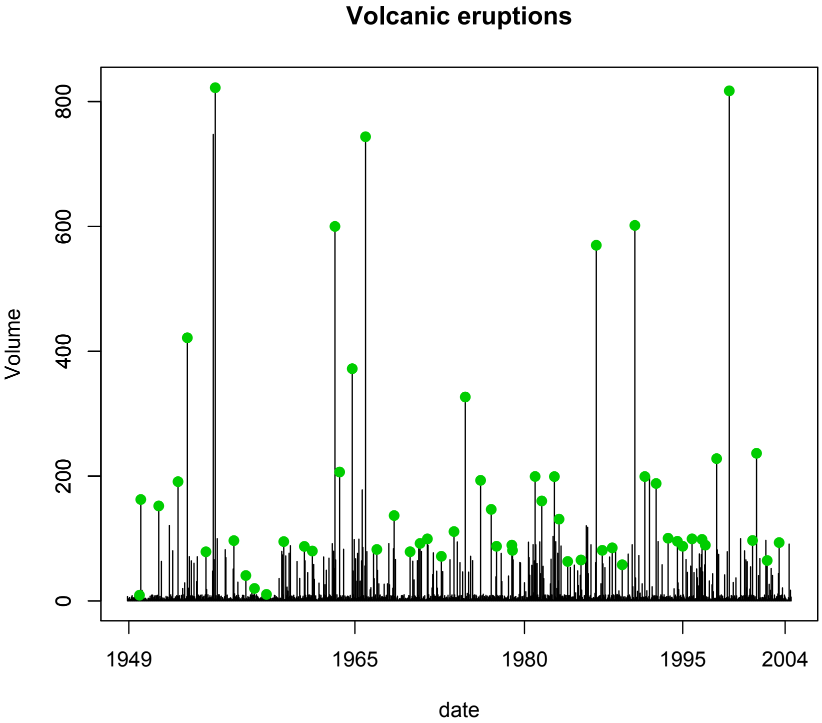

Figure 5: Maximum eruption values over 1 year period

the likelihood is zero and the log-likelihood is:

$$l(\mu,\sigma) = -m\log\sigma - \sum_{i=1}^{m}\left(\frac{z_i-\mu}{\sigma}\right) - \sum_{i=1}^{m} exp\left(-\frac{z_i-\mu}{\sigma}\right) \tag{4.2}$$

By maximizing these log-likelihood functions, we obtain the maximum likelihood estimates $(\hat{\mu},\hat{\sigma},\hat{\xi})$. The optimization is made using numerical optimization algorithms.

4.2 Graphical model checking

Though it is impossible to check the validity of an extrapolation based on the GEV model, assessment can be done with reference to the observed data.

Probability plot

A probability plot is a comparison of the empirical and fitted distribution functions. The empirical distribution function evaluated in the i-th ordered block maximum, $Z_{(i)}$, is $\tilde{G}(Z_{(i)}) =$

$i/(m+1)$, and the fitted distribution function in the same point is:

$$\hat{G}(Z_{(i)}) = exp\left\{-\left(1+\hat{\xi}\left(\frac{z_{(i)}-\hat{\mu}}{\hat{\sigma}}\right)\right)^{-1/\xi}\right\} \tag{4.3}$$

In order to have a good model it is necessary that $\tilde{G}(z_{(i)}) = \hat{G}(z_{(i)})$. In practice the plot of points $\left(\tilde{G}(z_{(i)}), \hat{G}(z_{(i)})\right)$, $i = 1, \ldots, m$, should lie close to the first diagonal.

Quantile plot

The quantile plot is a representation of the points $\left(\hat{G}^{-1}(i/(m+1)), z_{(i)}\right)$, $i = 1, \ldots, m$, where

$$\hat{G}^{-1}(i/(m+1)) = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\left(1-\left(-\log\frac{i}{m+1}\right)^{-\hat{\xi}}\right),\ i = 1, \ldots, m \tag{4.4}$$

In the ideal situation the plot should show a linear function. Departure from linearity in the quantile plot also indicate model failure.

Return level plot

The return level plot represents the points $(\log y_p, \hat{z_p})$, $0 < p < 1$. Confidence intervals are usually added to this plot to increase its information.

4.3 Peak over threshold

Modeling only block maxima implies that we waste a lot of data if a detailed recording of the studied phenomenon is available. Another approach consists in considering for the analysis those data that are viewed as extreme observations, e.g. those data that surpass a threshold level u. The stochastic behavior of these excesses over u is studied. So, given $X1, \ldots, X_n$ a sequence of independent and identically distributed random variables, having distribution function F, we are interested in the conditional probability $F_u(y) = P(X \leq u + y | X > u)$, this is $F_u(y) = \dfrac{F(u+y) - F(u)}{1 - F(u)}$.

If $X1, \ldots, X_n$ is a sequence of independent and identically distributed random variables with a common distribution function F, and $M_n = \max\{X_1, \ldots, X_n\}$ satisfying the conditions to be approximated by GEV, that is, for large n:

$$\Pr\{M_n \leq z\} \approx G(z),\ where\ G(z) = exp\left\{-\left(1+\xi\frac{z-\mu}{\sigma}\right)^{-1/\xi}\right\} \tag{4.5}$$

Then, for large enough u, the distribution function of $(X - u)$, conditioned on $X < u$, is approximately given by:

$$H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi} - the\ Generalized\ Pareto\ Distribution \tag{4.6}$$

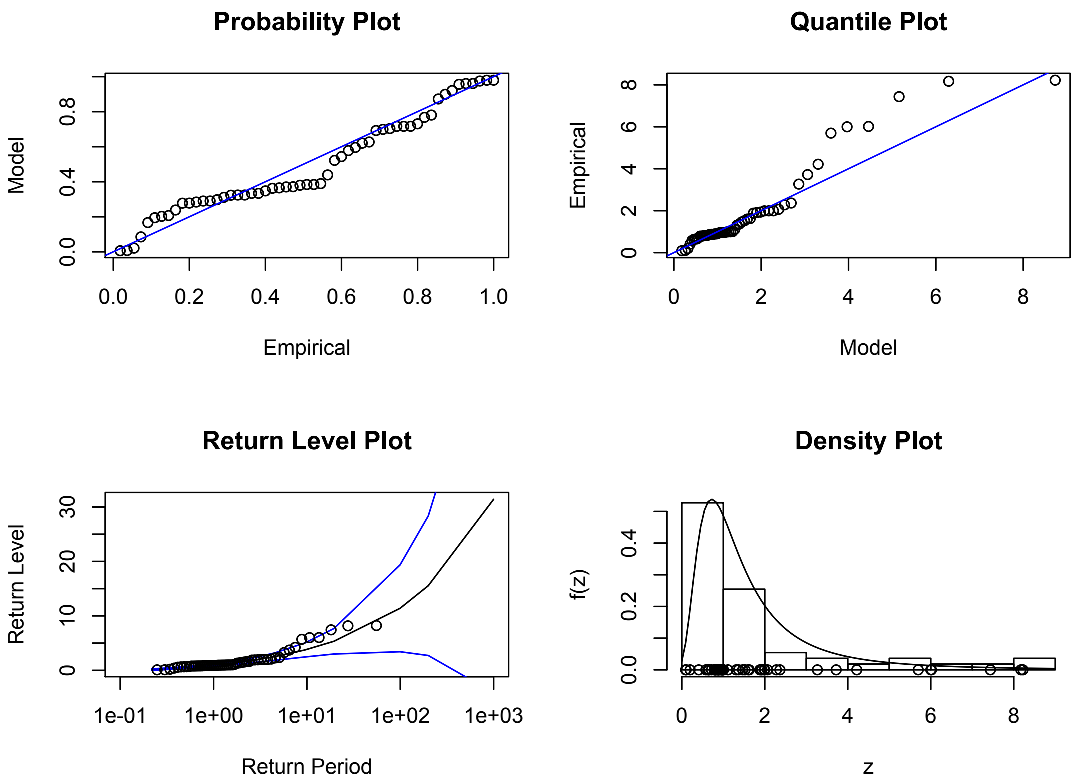

Figure 6: Maximum-likelihood fitting for the generalized extreme value distribution

defined on $\{y/y > 0 \; and \; (1 + \xi y/\tilde{\sigma}) > 0\}$, and where $\tilde{\sigma} = \sigma + \xi(u - \mu)$. The parameters of the Generalized Pareto Distribution (GDP) are uniquely determined by the parameters of the associated GEV distribution of block maxima. This implies that if we change the size of blocks in the GEV analysis then the parameter ξ remains unperturbed while the parameters μ and σ change but compensating their values to provide a fixed value for $\tilde{\sigma}$. The issue of choosing the threshold is similar to that of selecting the size of a block in the sense that both imply a balance between bias and variance. A low level leads to failure in the asymptotic approximation of the model and a high level provides few observations and the high variance.

Once the threshold has been estimated, the next step is to estimate the parameters of the GDP by maximum likelihood. If we denote by $y_1, \ldots, y_k$ the k excesses over the threshold, the log-likelihood function, in the case that ξ is not zero, is:

$$l(\sigma, \xi) = -k \log \sigma - (1 + 1/\xi) \sum_{i=1} k \log(1 + \xi y_i/\sigma), \tag{4.7}$$

when $(1 + \xi y_i/\sigma) > 0$, in other case $l(\sigma, \xi) = -\infty$.

In the case $\xi = 0$ the log-likelihood is $l(\sigma) = -k \log(\sigma) - \sigma^{-1} \sum_{i=1}^{k} y_i$

Probability plots, quantile plots and return plots are used for assessing the quality of a fitted generalized Pareto model. Assuming a threshold u, ordered excesses $y_{(1)}, \ldots, y_{(k)}$ and

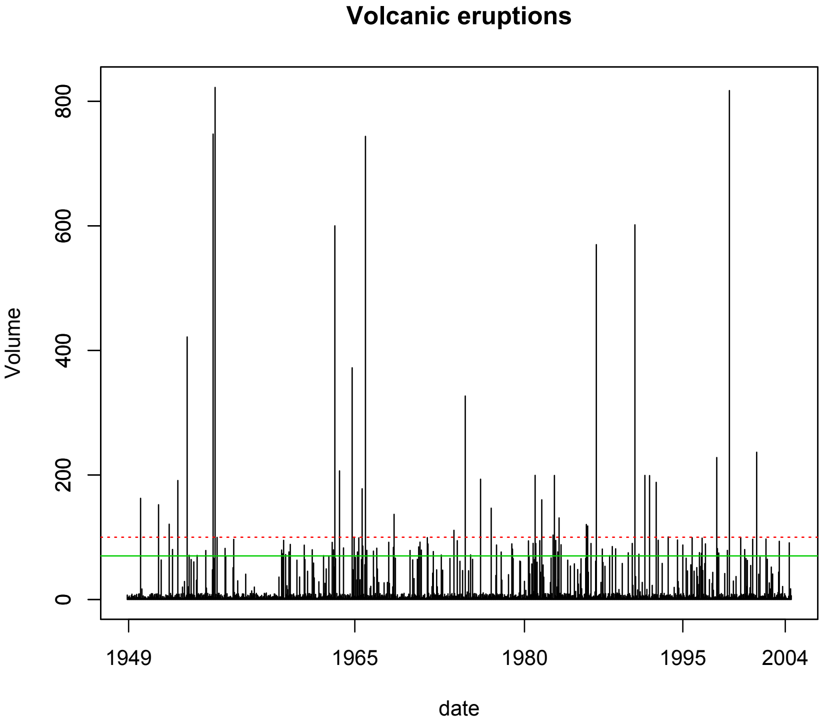

Figure 7: The threshold is set for 70 (green) and 100 (red)

an estimated model $\hat{H}$ for the GDP, then we have:

- *Probability plot*: It represents the points $\left(i/(k+1), \hat{H}(y_{(i)})\right)$, $i = 1, \ldots, k$.

- *Quantile plot*: It represents the points $\left(\hat{H}^{-1}(i/(k+1)), y_{(i)}\right)$, $i = 1, \ldots, k$. When the model is valid, in both plots the points are almost linearly placed.

  When $\hat{\xi} \neq 0$ the estimations are: $\hat{H}(y) = 1 - \left(1 + \frac{\hat{\xi} y}{\hat{\sigma}}\right)^{-1/\xi}$ and $\hat{H}^{-1}(p) = \frac{\hat{\sigma}}{\hat{\xi}}\left((1-p)^{-\xi} - 1\right)$.

  When $\hat{\xi} = 0$ the expression are: $\hat{H}(y) = 1 - \exp\left(-\frac{y}{\hat{\sigma}}\right)$ and $\hat{H}^{-1}(p) = -\hat{\sigma}\ln(1-p)$.

- *Return level plot*
  Denoting by $\delta_u = \Pr(X > u)$ and from the conditional distribution $\Pr\{X > x/X > u\} = \left(1 + \xi\frac{(x-u)}{\sigma}\right)^{-1/\xi}$ we obtain that:

$$\Pr\{X > x\} = \delta_u\left(1 + \xi\frac{(x-u)}{\sigma}\right)^{-1/\xi} \tag{4.8}$$

Hence, the level x_m that is exceeded on average once every m observations id the solution of $\frac{1}{m} = \delta_u \left(1 + \xi \frac{(x_m - u)}{\sigma}\right)^{-1/\xi}$, which is $x_m = u + \frac{\sigma}{\xi}\left((m\delta_u)^{\xi} - 1\right)$. In the case $\xi = 0$ the return level is $x_m = u + \sigma \log(m\delta_u)$, again for m enough large.

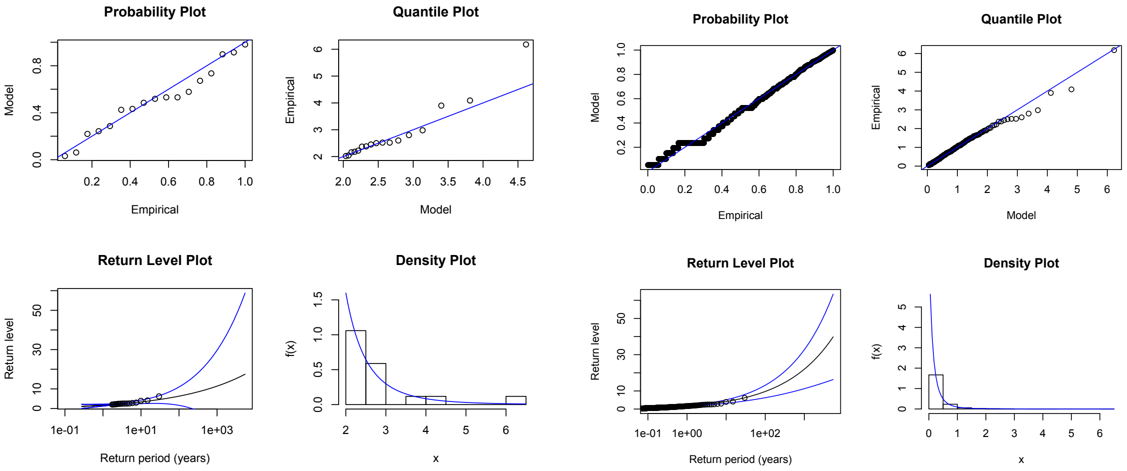

Figure 8: Peak over threshold approach a) threshold equals 70 b) threshold equals 100

References

[1] E. Castillo, *Extreme value theory in engineering. Statistical Modeling and Decision Science*, Academic Press New York, 1988.

[2] S. Coles and R. Sparks, *Extreme value methods for modeling historical series of large volcanic magnitudes*, Statistics in volcanology, 1 (2006), pp. 47–56.

[3] L. De Haan and A. Ferreira, *Extreme value theory*, Springer Science+ Business Media, 2006.

[4] C. Furlan, *Extreme value methods for modelling historical series of large volcanic magnitudes*, Statistical Modelling, 10 (2010), pp. 113–132.

[5] M. Parkinson, *The extreme value method for estimating the variance of the rate of return*, Journal of Business, (1980), pp. 61–65.

[6] M. Sheridan, A. Patra, K. Dalbey, and B. Hubbard, *Probabilistic digital hazard maps for avalanches and massive pyroclastic flows using TITAN2D*, in Groppelli G. and Viereck-Goette, L., editors, Stratigraphy and Geology of Volcanic Areas , Geological Society of America Special Paper, 464 (2010), pp. 281–291.

[7] R. Sobradelo, J. Martí, A. Mendoza-Rosas, and G. Gómez, *Volcanic hazard assessment for the canary islands(spain) using extreme value theory*, Natural Hazards and Earth System Sciences, 11 (2011), pp. 2741–2753.