REVSTAT – Statistical Journal
Volume 17, Number 2, April 2019, 163-186

MODELING LARGE VALUES OF SYSTOLIC BLOOD PRESSURE IN THE PORTUGUESE POPULATION

Authors: C. P. Caetano
– Departamento de Estatística e Investigação Operacional,
Faculdade de Ciências da Universidade de Lisboa
Portugal
caetanoconstantino@gmail.com

P. de Zea Bermudez
– Departamento de Estatística e Investigação Operacional
Faculdade de Ciências da Universidade de Lisboa, and CEAUL
Portugal
pcbermudez@fc.ul.pt

Received: October 2018 Revised: January 2019 Accepted: March 2019

Abstract:

- It has been well stated that high values of blood pressure constitute a risk factor for cardiovascular diseases [20], with the latter being the number one death cause in Portugal. The main interest of the present study is to model the high values of systolic blood pressure in the individuals of the population who are most at risk, i.e., the elderly. This group frequently suffers from a specific type of hypertension pathology, known as isolated systolic hypertension. With that purpose the *Peaks Over Threshold* methodology was applied, which consists in fitting a generalized Pareto distribution to the excesses above a sufficiently high threshold. The model will be able to estimate extreme quantiles and tail probabilities.

Key-Words:

- *Extreme value theory; Bayesian statistics; threshold models; threshold selection; multiple testing for ordered hypotheses.*

AMS Subject Classification:

- 60G70, 62C10, 62P10.

1. INTRODUCTION

Extreme events can be defined as low frequency episodes of some random process. For example, floods transpire when the water level of some water body exceeds an uncommonly high threshold. Classical statistical methodologies are not suited to treat this kind of data, since they aim to make predictions about future behavior of the phenomena under study based on the most common events, i.e., classical statistics uses the central data to infer on future behavior by fitting the data to models based on asymptotic central limit like results. Such approach might be considered too overly simplified to infer on rare events. Hence the extreme value analysis paradigm arose out of the necessity to address the situations that fall on this scope. It offers well suited statistical methodologies to describe the tail behavior of the distribution underlying the observed data.

Extreme value theory (EVT) has been applied to a large assortment of different areas, ranging from meteorology, hydrology and environment to insurance, among others. There are still not many contributions of EVT to medical data (see e.g. [13]) even though EVT has recently been applied to Public Health problems ([30]) and to disease early detection (see [18] and [8]). The *Peaks Over Threshold* (POT) approach is a widely used EVT methodology. It aims to fit a generalized Pareto distribution (GPD) to the excesses (or exceedances) above a sufficiently high threshold, see [26] and [3].

One of the most strenuous point using POT is the selection of the threshold, i.e., the value over which the asymptotic model is fitted. In this work, we apply several classic techniques, such as the mean excess function and also some recent methods, see [24], [2] and [10].

Models obtained by using EVT techniques are able to extrapolate beyond the observed data and also enable extreme quantile estimation. In fact, most commonly we aim to estimate exceedance probabilities, $P(X > z) = p$, for some random variable X and a very small probability p and also determine the value x_q $(q = 1/m)$ such that x_q is exceeded, in mean, once every m observations.

In this work, the large values of the systolic blood pressure (SBP) will be modeled by means of the *Peaks Over Threshold*. Several GPD models will be fitted to a group of individuals who suffer from a specific type of hypertension, termed isolated systolic hypertension (ISH). Extreme quantiles and exceedance probabilities will be estimated. We also analyze the consequences to the models that result from the discretization of the continuous SBP variable.

The details of the problem to be studied are presented in section 2. Section 3 contains the results of the exploratory data analysis. The core issue of this paper is presented in section 4 and deals with the modeling of the extreme SBP observed in elderly individuals. In this section the issues related with the quantization of the data are also addressed. The paper ends with some comments and conclusions which constitute section 5.

2. DESCRIPTION OF THE PROBLEM

Hypertension, also known as high blood pressure, is described as an abnormal pressure on the blood vessels caused by blood flow. As blood is pumped throughout the body, the blood vessels are impacted by this flow, thus creating blood pressure and blood vessel tension. The higher the tension, the stronger the effort the heart must exert in order to pump the blood. Diagnosing hypertension is performed by measuring two blood pressure markers. Systolic blood pressure is the tension measured by the compliance of the blood vessels to the blood flow during a heartbeat. Diastolic blood pressure (DBP) is the tension measured between heartbeats.

According to the World Health Organization, hypertension is a global public health issue. It is highly associated with incidents of heart disease, stroke, kidney failure, premature mortality and disability. It is also a risk factor associated with the leading death causes in Portugal. Hypertension has been linked to unhealthy diets, sedentary lifestyle, drug abuse and tobacco use, see [20].

With the goal of addressing this public health issue, the Portuguese National Association of Pharmacies (ANF) developed a campaign in 2005 through their Department of Pharmaceutical Care to study the risk factors associated with the leading death causes in the country. As a consequence, information regarding $n = 40065$ individuals that volunteered to join the study was registered. The variables recorded are presented in Table 1.

Variable	**Categories/Units**	**Variable**	**Categories/Units**
Gender	male/female	Age	years
District	(see Figure 4)	Smoking habits	yes/no
Body mass index (BMI)	kg/m^2	Fasting blood glucose level (FBG)	mg/dL
Systolic blood pressure	mmHg	Blood glucose level at random time (BGRT)	mg/dL
Triglyceride level	mg/dL	Diastolic blood pressure	mmHg
Physician visit	yes/no	Total cholesterol level	mg/dL

Table 1: Recorded variables and corresponding units of measurement or categories.

In a previous study the extreme levels of total cholesterol were modelled by Zea de Bermudez and Mendes [13]. In this article we apply the aforementioned *Peaks Over Threshold* methodology to the elderly individuals who suffer from isolated systolic hypertension, which are characterized by having **diastolic blood pressure $<$ 90 mmHg and systolic blood pressure $\geq$ 140 mmHg**. This group is of interest since there is a known relationship between the age of the individuals and the prevalence of ISH [4]. Moreover, it makes up the bulk of the hypertensive individuals contained in the database. The classification categories in terms of blood pressure conditions are presented in Table 2 (guidelines of the Portuguese Hypertension Society). The goal is to fit a GPD to the SBP excesses above a sufficiently high threshold u, and subsequently estimate tail probabilities and extreme quantiles.

Category	**SBP**		**DBP**
Optimal	< 120	and	< 80
Normal	120-129	and/or	80-84
Normal high	130-139	and/or	85-89
First Degree Hypertension	140-159	and/or	90-99
Second Degree Hypertension	160-179	and/or	100-109
Third Degree Hypertension	≥ 180	and/or	≥ 110
Isolated Systolic Hypertension	≥ 140	and	< 90

Table 2: Categories of blood pressure in mmHg* (Portuguese Hypertension Society guidelines).

2.1. The generalized Pareto distribution

The generalized Pareto distribution constitutes a fitting model for threshold exceedances, see [3] and [26]. Let $Y_1, Y_2, ..., Y_n$, be a sequence of i.i.d. random variables. The cumulative distribution function of the excesses $X = Y - u$, given that $Y > u$ for a sufficiently high threshold u, is approximately given by:

$$F(x) = \begin{cases} 1 - \left(1 + \frac{kx}{\sigma}\right)^{-\frac{1}{k}} & k \in \mathbb{R} \setminus \{0\}, \\ 1 - e^{-\frac{x}{\sigma}} & k = 0, \end{cases} \tag{2.1}$$

with shape parameter k, $-\infty < k < \infty$ and scale parameter σ, $\sigma > 0$.

This distribution has support $\big\{x \in \mathbb{R} : x > 0\big\}$ for $k \geq 0$ and support $\big\{x \in \mathbb{R} : 0 < x < -\frac{\sigma}{k}\big\}$ for $k < 0$. In practice the most suited threshold is the smallest value that still provides an adequate model fit to the data. The generalized Pareto distribution will be denoted by GPD(k,σ).

* Hypertension is categorized by the highest value of either SBP or DBP, the isolated systolic hypertension category should be classified by first, second and third degree according to the values of SBP in each category.

3. EXPLORATORY DATA ANALYSIS

In terms of blood pressure, the $n = 40065$ individuals are classified in one of the following groups:

- Group 1 – Individuals with both blood markers higher than the standard values (DBP $>$ 90 and SBP $>$ 140)
- Group 2 – Individuals suffering from isolated systolic hypertension (DBP $<$ 90 and SBP $\geq$140)
- Group 3 – Healthy individuals (DBP $<$ 90 and SBP $<$ 140)
- Group 4 – Individuals suffering from diastolic hypertension (DBP $>$ 90 and SBP $<$ 140)

There are also 3380 individuals with omitted information about these variables.

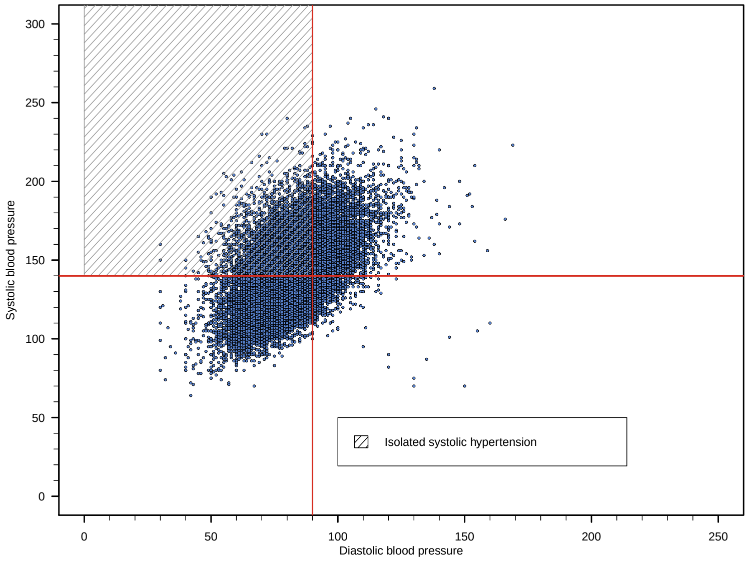

Figure 1: Systolic blood pressure *vs.* diastolic blood pressure for Portuguese voluntary pharmacy attendees.

Figure 1 results from plotting the SBP versus the DBP for the Portuguese voluntary pharmacy attendees, where the aforementioned stratification can clearly be seen. It suggests some linear correlation with positive slope between the two blood markers. The red horizontal and vertical lines convey the accepted limits over which an individual is considered from suffering an hypertension-type pathology, as illustrated by Table 2. It would also be interesting to study the extreme values of both variables, DBP and SBP. Although, a considerable amount of literature exists about bivariate extreme value analysis, see [21], it is not the focus of this study. From this point onward we will concentrate on the exploratory analysis of the SBP in individuals who suffer from ISH ($n = 9996$) - note that this data has lower bound equal to 140 mmHg.

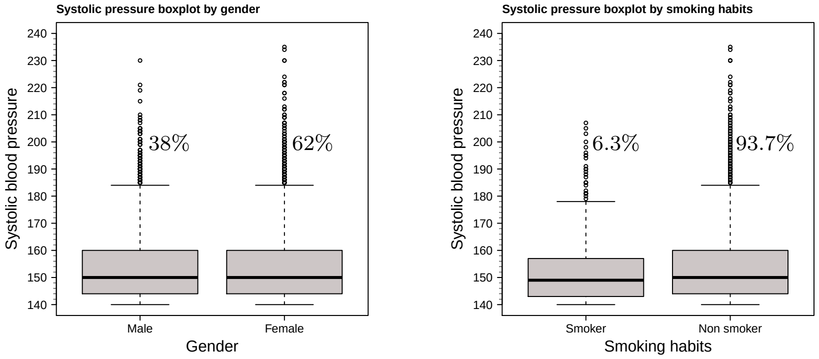

Figure 2: Systolic blood pressure boxplots by gender (left) and by tobacco consumption (right).

Figure 2 illustrates the SBP boxplots by gender and tobacco consumption. It can be seen that women seem to have more and higher extreme values of SBP than men. The boxplots produced for the smoking habits seems to indicate that those who smoke have, in the overall, lower values of SBP than those who do not. However, no credible conclusion about this relation can be derived from these boxplots since there are several confounding factors. For instance, out of the 9586 individuals with recorded smoking habits, only 6.3% are smokers. Moreover, this boxplot includes men and women, young and old individuals, which might also influence this outcome.

Age	**Min**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**	**n**	**Prop**
< 25	140.0	140.8	145.0	149.2	150.5	206.0	56	0.0059
25-34	140.0	141.0	146.0	149.2	152.0	212.0	171	0.0180
35-44	140.0	142.0	147.0	149.2	152.0	207.0	315	0.0331
45-54	140.0	142.0	148.0	149.7	154.0	203.0	803	0.0844
55-64	140.0	143.0	149.0	152.0	158.0	213.0	2075	0.2180
65-74	140.0	144.0	150.0	154.1	160.0	240.0	3613	0.3796
≥ 75	140.0	145.0	154.0	157.2	165.0	235.0	2486	0.2611
NA's							477	

Table 3: Summary statistics of the systolic blood pressure by age in Portuguese voluntary pharmacy attendees who suffer from isolated systolic hypertension.

One factor that has been shown to be highly associated with high values of SBP is age, see [27] and [4]. Table 3 presents a summary of the SBP variable in an array of different age strata. The individuals are not equally distributed by age stratum. It can be seen that the bulk of the observations lie above the 55 years old group. This might be the result of selecting individuals who suffer from

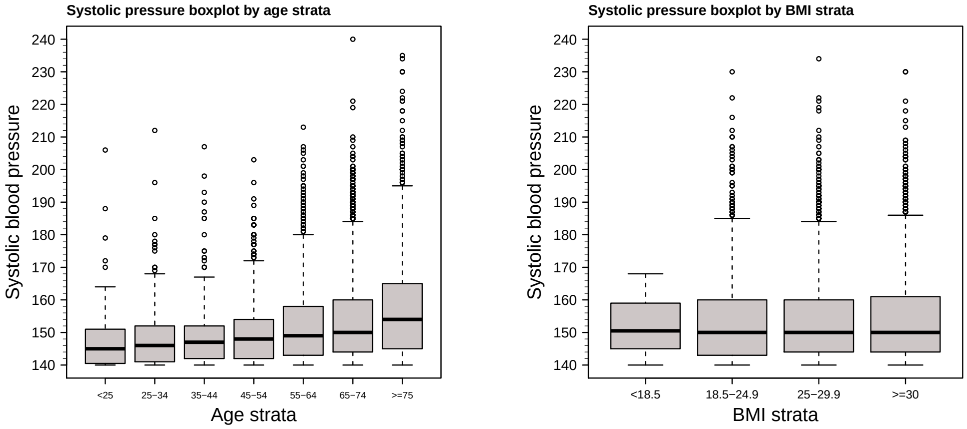

Figure 3: Systolic blood pressure boxplots by age strata (left) and by BMI strata (right).

ISH. As previously mentioned, this pathology is known to be more common in the elderly. We can see a steady rise of SBP values as age goes up. This can also be observed in Figure 3 (left) where it is most apparent that older individuals tend to have, in the overall, higher values of SBP. As mentioned before, SBP is the tension the blood flow produces on the blood vessels during a heartbeat. As a person gets older, he/she tends to lose blood vessel elasticity thus increasing the tension generated by the blood flow.

Underweight	**Normal Weight**	**Overweight**	**Obese**
< 18.5	18.5-24.9	25-29.9	> 30.0

Table 4: BMI classes.

Table 4 illustrates the different classes of body mass index. Figure 3 (right) presents the boxplots of the SBP values for each BMI category. The percentage of individuals which fall in each of the BMI strata is not even - 0.27%, 19.18%, 47.18% and 33.37% for underweight, normal, overweight and obese categories, respectively. Maybe the prevalence of ISH is higher in individuals with high BMI. The low number of observations in the underweight category might also be due to the fact that underweight people tend to have lower values of SBP, hence underweight individuals exceeding 140 mmHg are rare. Not taking into account the underweight stratum, there is little to no difference in the SBP between each class. This suggests that BMI by itself may not be sufficient to account for high levels of systolic blood pressure.

Our next interest is to compare the Portuguese districts and autonomous regions in terms of SBP. Figure 4 illustrates the boxplots of the values of the SBP observed in individuals who suffer from ISH by district and autonomous regions. One curious phenomena is that higher population density districts yield higher maximum values, i.e., the largest value is observed in Porto, which is the

District/ Autonomous Region	**n**	**District/ Autonomous Region**	**n**
Viseu	320	Vila Real	299
Viana do Castelo	120	Setúbal	788
Santarém	550	Porto	1590
Portalegre	92	Lisboa	2248
Leiria	467	Açores	88
Ilha da Madeira	221	Guarda	103
Faro	227	Évora	186
Coimbra	441	Castelo Branco	200
Bragança	276	Braga	810
Beja	117	Aveiro	736
NA	117		

Table 5: Number of voluntary attendees per district/autonomous region, suffering from ISH.

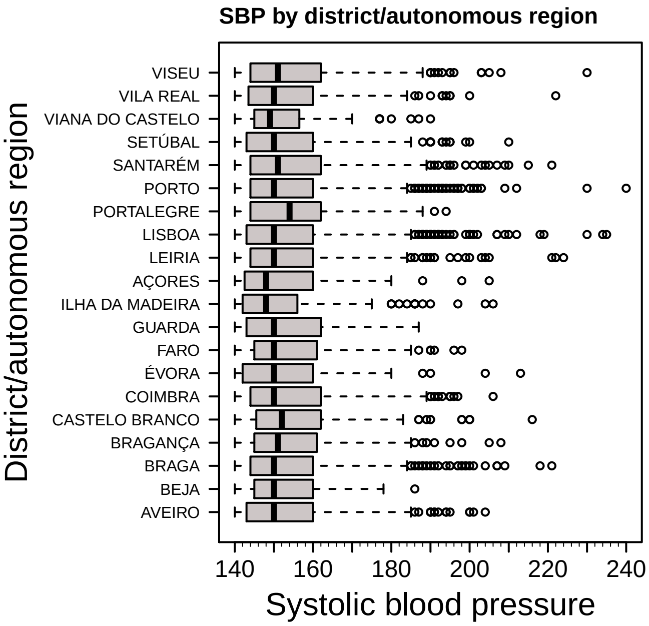

Figure 4: Systolic blood pressure by Portuguese district/autonomous region.

second most populated Portuguese district, followed by Lisboa, the most populated district with a maximum equal to 235 mmHg. Some other high population density districts with extreme maximum values are: Braga with maximum 229 mmHg, the third most populated Portuguese district and Viseu with maximum 230 mmHg. The previous mentioned districts also were the ones that supplied the largest samples, as seen in Table 5, (with the exception of Viseu) which might also be the cause for such high maximum values when compared with other districts with smaller sample size. Regarding the median values of systolic blood pressure, the autonomous regions of Açores and Madeira, which are the only non-continental Portuguese regions, provided lower median systolic blood pressure values than any individual district from mainland. The remaining districts are similar, with the exceptions of Portalegre that has a slightly higher median than the rest. In [29] the authors study the relationship between diet, leisure activity, BMI and serum cholesterol. Such analysis would also be well suited for the Portuguese districts and autonomous regions, since an individual's diet and lifestyle varies geographically. The modeling of extreme values of SBP of individuals that suffer of ISH by district and autonomous regions can be seen in [5].

4. MODELING EXTREME SYSTOLIC BLOOD PRESSURE VALUES IN THE ELDERLY

In this section we propose models for the extremes of systolic blood pressure of elderly individuals (age ≥ 55) who suffer from ISH. Out of the 9996 individuals suffering from ISH a total of 9519 have documented age. There are several reasons that justify our interest in this study. First, the exploratory data analysis shows that SBP values somewhat change between age strata (as seen in Figure 3 (left)), suggesting that as a person ages, his or her systolic blood pressure tends to rise. This phenomena is also well known in the literature, see [27]. Second, the elderly make up the bulk of the observations. Additionally 86% of the SBP readings were recorded in people aged 55 and older.

We begin by addressing the quantized structure of the data. The methods to model extreme values were constructed for continuous variables, hence some methods might not perform well when applied to a highly discretized data set. We quote [2] regarding the performance of the goodness-of-fit tests using a quantized data set: 'Quantization pushes the null distribution of the Anderson-Darling statistic to the right; the *p*-value obtained by positioning the observed statistic with the quantized data to the null distribution from continuous data is smaller than it should be'. Thus we may be led to reject a certain model that in fact was fit for the data.

By performing an exploratory analysis of individuals in this study (individuals aged 55 and older, suffering from ISH) we encountered two issues with the data, specifically the quantized structure of the data and the high frequencies of *rounded* numbers. Figure 5 illustrates this issue. The systolic blood pressure values of 140, 150, 160, 170, 180, 190 and 200 mmHg have higher frequencies than their 'neighbors'. The reason for this behavior is unknown, though one might assume that it was the result of biased approximations or perhaps the devices used to measured the blood pressure were not precise enough. The most common way to deal with this problem is to *shake* the sample distribution, by that we mean considering each value censured in an interval. For some observed value x_{obs}, its true value x belongs to an interval $[x_{obs} - \delta, x_{obs} + \delta]$, $\delta \in \mathbb{R}^+$. We can choose how

x is distributed in this interval. For example x can be equally distributed in the interval, or it may have a higher probability to be close to the observed value, x_{obs}. The former can be constructed by generating a set of random values from a continuous Uniform distribution with parameters $a = -\delta$ and $b = \delta$, $\delta \in \mathbb{R}^+$, and adding them to each observed value, while the latter can be obtained by generating values from a beta distribution with parameters $\alpha = \beta = \delta$, $\delta > 0$, location parameter 0.5 and scale parameter 1, hence taking values in $[-0.5, 0.5]$. This second alternative will result in a milder *shakeup* of the data when compared to the first, since it is more likely that the generated values will be close to 0. We would like to notice that this technique is used in several studies, see [2], and it is usually applied in order to obtain a smoother empirical distribution. It is important to underline that technically the data is being altered and hence usually a mild jitter is considered.

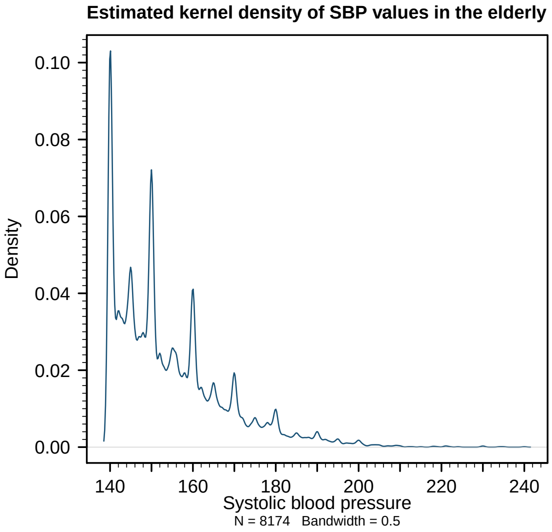

Figure 5: Kernel density function of the systolic blood pressure values of elderly individuals who suffer from ISH.

4.1. Jitter and non-jitter extreme value models for systolic blood pressure in elderly individuals who suffer from isolated systolic hypertension

We aim to produce produce three extreme value models for the SBP measured in individuals that satisfied the aforementioned criteria, using three distinct data sets.

1. Unaltered Data
2. Data + Uniform$(-1.5, 1.5)$

3. Data + Beta$(10, 10, -0.5, 0.5)$

We want to ascertain if there are differences in the models created from the three data sets mentioned above. Also, we would like to assess how these models hold up regarding their predicting capabilities.

Using the R function *rbeta* we generated a random sample with size 8174 from a beta distribution with parameters $\alpha = 10$, $\beta = 10$, location parameter 0.5 and scale parameter 1. We also generated a sample of size 8174 from the continuous uniform distribution with parameters $a = -1.5$ and $b = 1.5$ using the R function $runif$.

We then created two new data sets by adding each sample to the data. Note that by adding these simulated samples to the SBP values of the elderly individuals who suffer from ISH we got some values below 140, that were not considered in the subsequent analysis. Let's investigate how both jitters altered the data.

Figures 6 and 7 illustrate the histograms and kernel densities, respectively, of the non-jitter data and jitter data. Note that, as expected, both jitters seem to *smooth out* the sample distribution, as seen on Figure 7. The histograms show that there are less frequency differences between neighboring classes. It is important to point out that although there appears to be a slight difference between the jitter data and the non-jitter data, the summary statistics of these data sets seem not to differ much, as seen in Table 6.

Data	**Min**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**	**n**
Non-jitter	140.0	145.1	151.9	155.6	162.0	240.00	8174
Unif-jitter	140.0	145.6	151.7	155.7	162.0	238.50	7593
Beta-jitter	140.0	145.2	151.9	155.8	162.0	239.98	7628

Table 6: Summary statistics of the systolic blood pressure by age in the uniform jitter-data, beta-jitter data and non-jitter data.

Next we propose a sequence of possible threshold candidates: 140, 150, 160, 170, 180, 190 and 200 mmHg. Figures 8, 9 and 10 present the exponential QQ-plots and histograms for the data above each candidate threshold for the non-jitter, uniform-jitter and beta-jitter data, respectively. For values above 170 mmHg the exponential model seems to adequately fit the 3 cases. Furthermore, the associated histograms display a tail decay indicating that an exponential model could give an adequate fit.

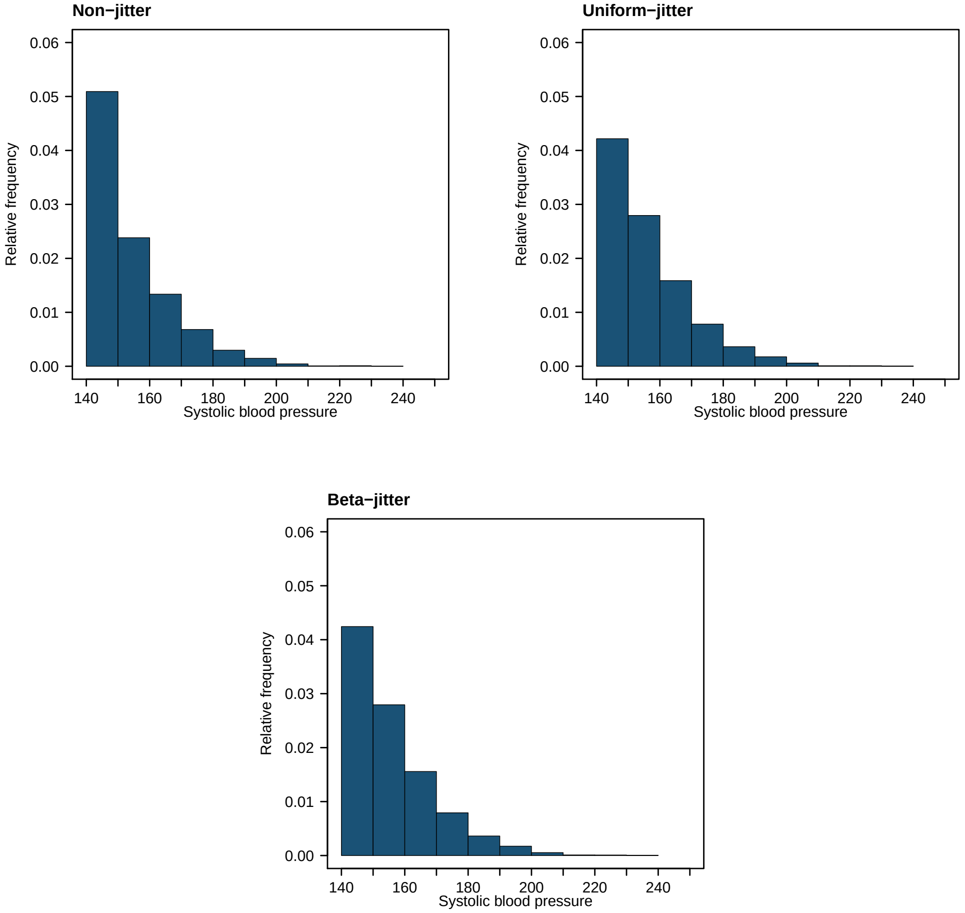

Figure 6: Histograms for the non-jitter data (top left) and jitter data using the uniform (top right) and beta (bottom) distributions.

4.1.1. Threshold selection analysis

We now start the procedure of selecting adequate thresholds for each case, with the goal of fitting GPD(k,σ) models for the excesses above each threshold. We start by plotting the empirical mean residual life function (MEF), see [10]. This function should have a linear behavior for some high value of systolic blood pressure, see [10]. Using the R package *eva* and its function *mrlPlot* we plotted the empirical MEF for each data set. Figure 11 illustrates the results. Although the plots are not very easy to analyze, they seem to indicate that for values between 180 mmHg and 200 mmHg, the function appears to have a linear behavior, implying that an appropriate threshold could lie between these two values.

Next we present the results for the Bayesian threshold selection method using measures of surprise for each data set, see [24]. Figure 12 presents the

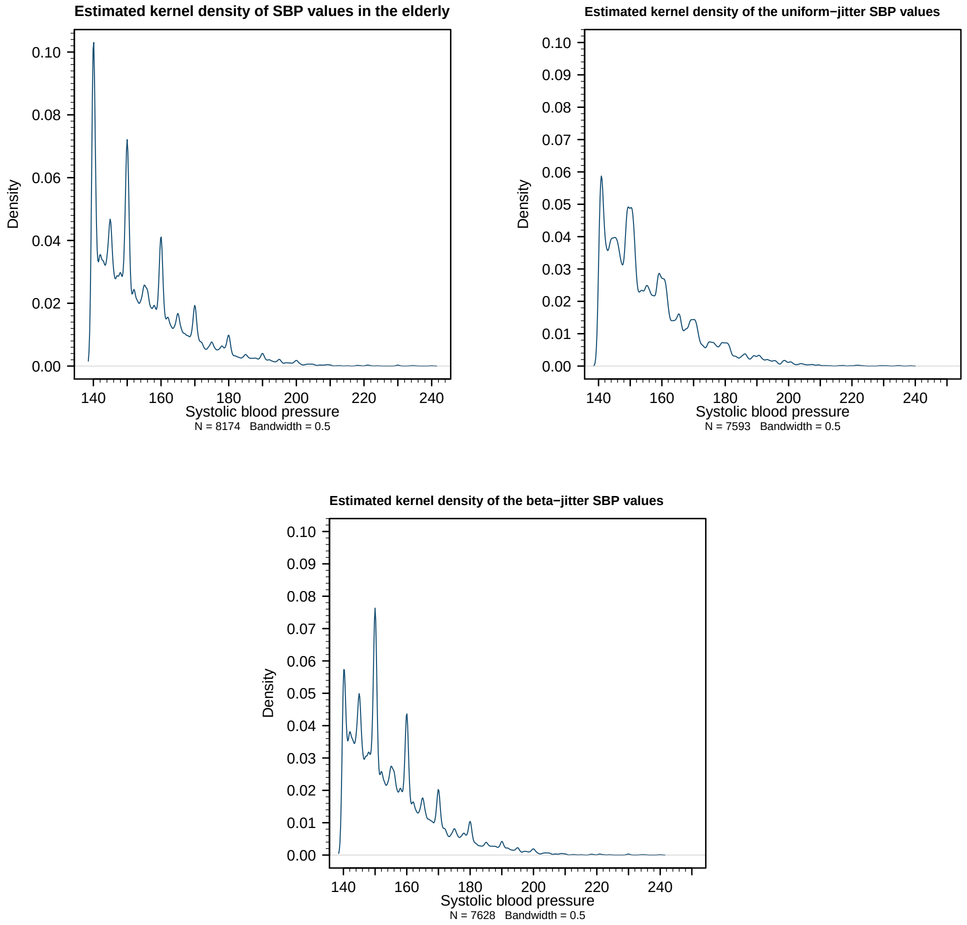

Figure 7: Estimated kernel densities for the non-jitter (top left) and jitter data using the uniform (top right) and beta (bottom) distributions.

predictive p-values obtained by considering the previous sequence of threshold candidates, for the non-jitter data, uniform-jitter data and beta-jitter data, respectively. The test statistic used is the likelihood, one of the possibilities recommended by [24]. For each threshold we sampled 5000 times the predictive posterior distribution. Then, we proceeded to compute the p-values. This process was repeated 30 times and presented in Figure 12 by a boxplot at each threshold. The p-value obtained per threshold can be interpreted as evidence against the GPD model when it is close to 0 or 1. Furthermore p-values close to 0.5 can be understood as showing less incompatibility with the GPD model [25], [24].

Figure 12 (top left) shows the method applied to the non-jitter data set. It only manifests less incompatibility with the GPD model for high threshold values, i.e., $190 < u < 200$ mmHg. Moreover, the p-values demonstrate a switch in surprise when more data is considered, i.e., when we introduce data below

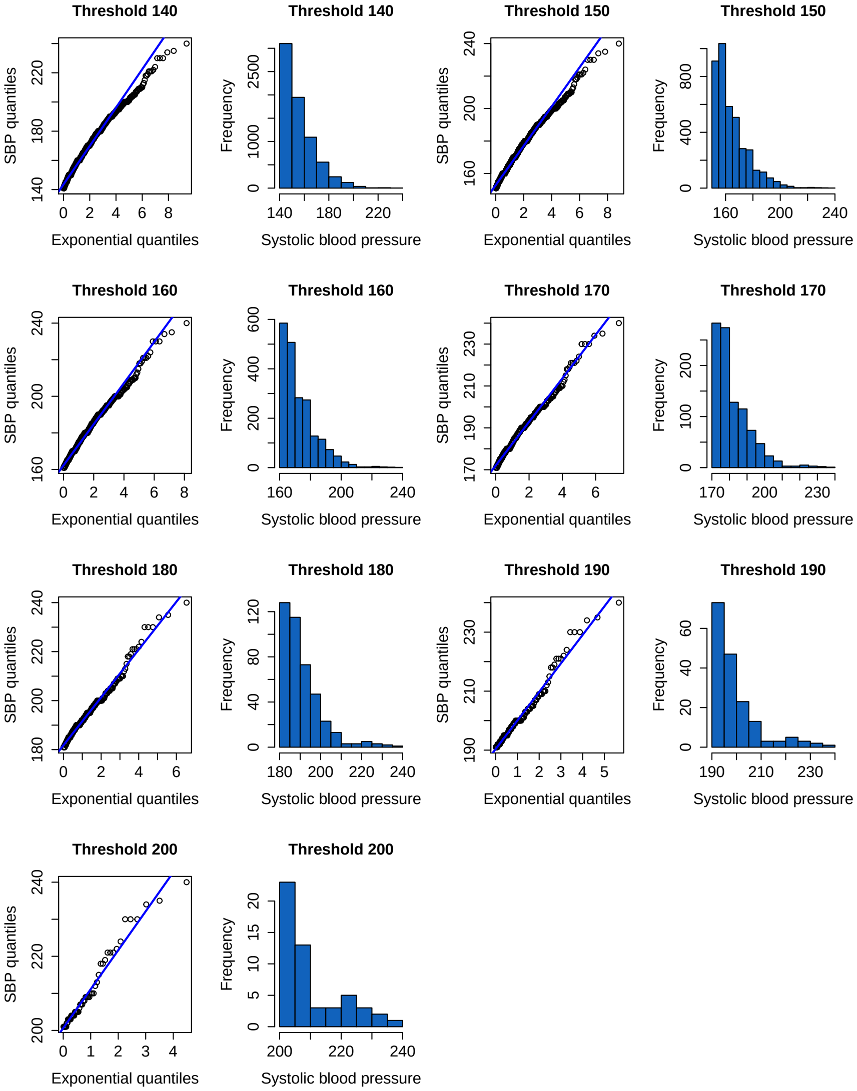

Figure 8: Exponential QQplots and histograms for each threshold candidate for the non-jitter data.

190 mmHg, the p-values move away from 0.5 and tend to 0, suggesting more incompatibility with the GPD model. On the other hand, the p-values obtained for both jitter cases do not seem to change a great deal until the data below 150 is considered. This method seems sensible to the jitter process, even for the case of the mild beta jitter, since it produces overall higher p-values in both jitter cases. Based on the output we are led to select a high threshold value for the non-jitter case, i.e., a value between 190 mmHg and 200 mmHg. Both jitter cases seem to indicate that 150 mmHg is an acceptable threshold, since there is a change in surprise from 140 mmHg to 150 mmHg, meaning that the predictive p-value obtained from 150 mmHg is closer to 0.5 than the one obtained from 140 mmHg. Furthermore, for the remaining threshold values, the obtained predictive p-values

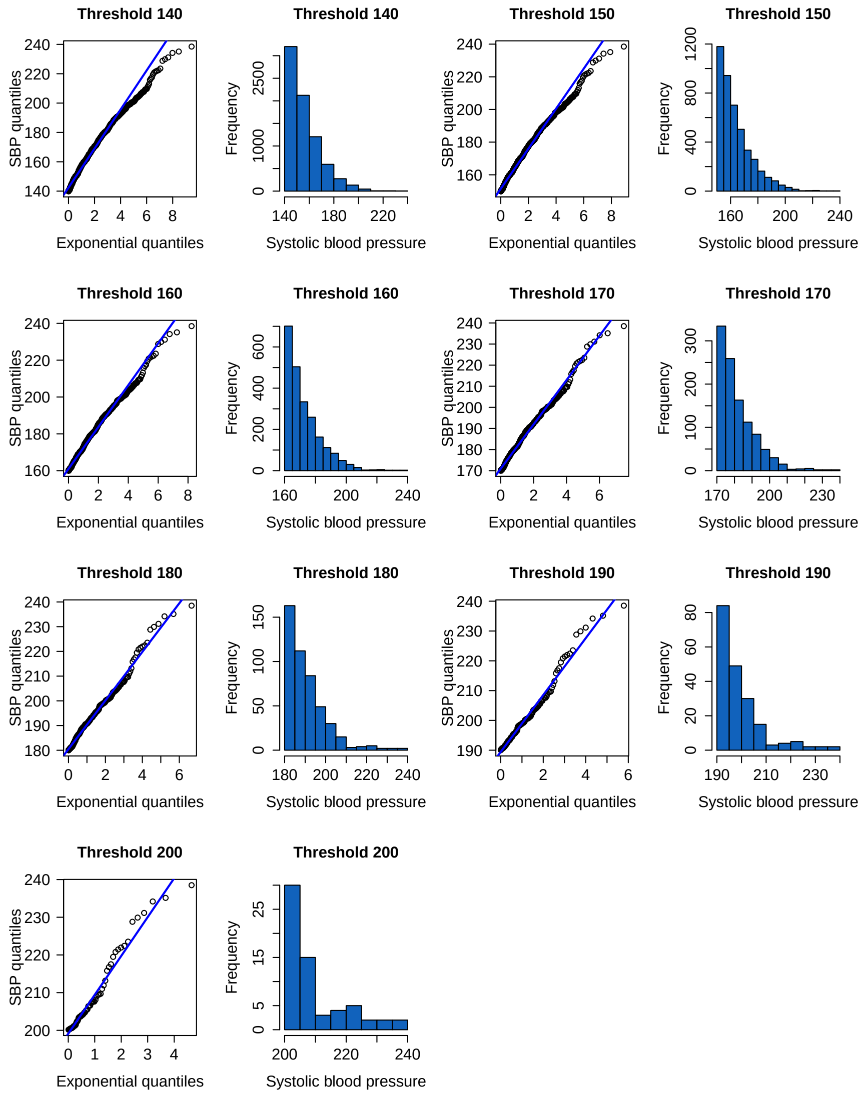

Figure 9: Exponential QQplots and histograms for each threshold candidate for the uniform-jitter data.

do not appear to change too much.

Next, we present the automated threshold selection method using goodness-of-fit tests for each of the previously mentioned data sets. We will adopt the ForwardStop rule outlined in [2] and [17]. Let $u_1, u_2, ..., u_m$ be a sequence of increasingly ordered threshold candidates for a given data set, and consider that $H_0^1, H_0^2, ..., H_0^m$, are m null test hypotheses such that for some $1 \leq i \leq m$, the ith null hypothesis is defined as H_0^i : the excesses over u_i come from a generalized Pareto distribution. Let $p_1, p_2, ..., p_m$ be the p-values obtained using the Cramér-Von Mises goodness-of-fit test for each sample. The FowardStop rule is given

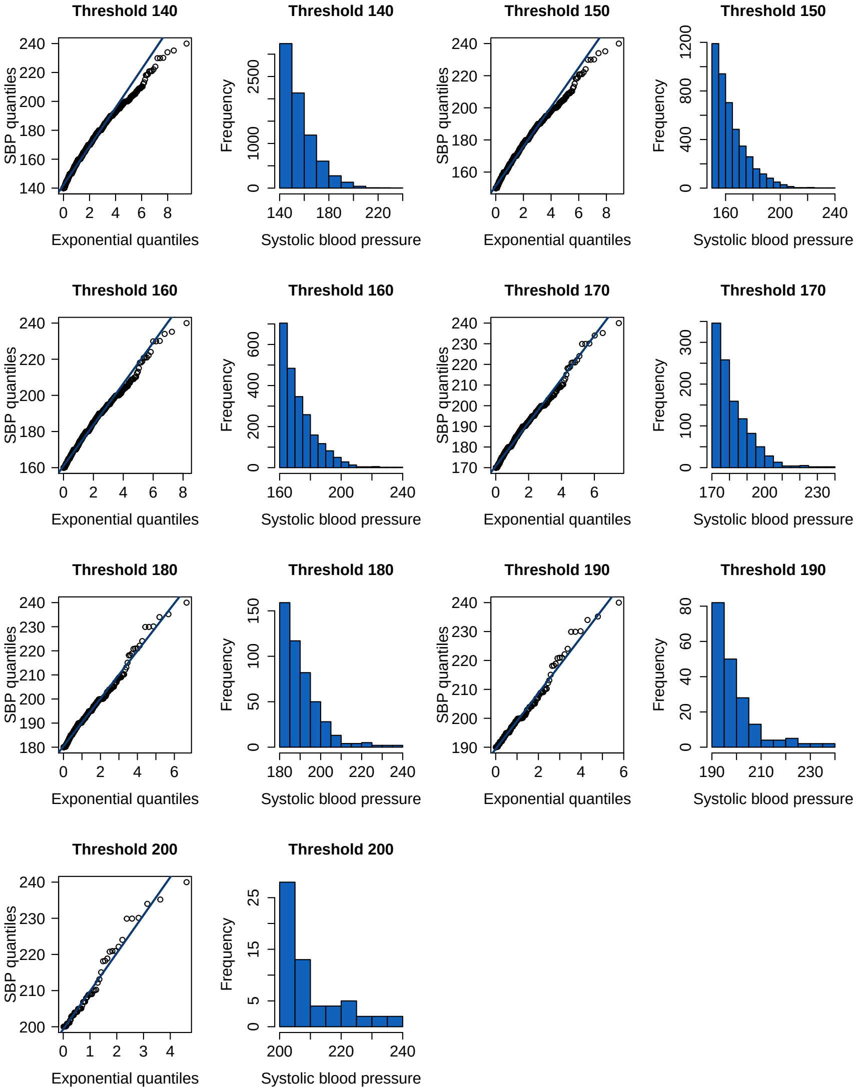

Figure 10: Exponential QQplots and histograms for each threshold candidate for the beta-jitter data.

by

$$\hat{\imath} = max\Big\{i \in \{1,...,m\} : -\frac{1}{i}\sum_{j=1}^{i} log(1-p_j) \le \alpha\Big\}, \tag{4.1}$$

where α is the significance level. The method consists on computing the p-values at each threshold, starting from the smallest until (4.1) is satisfied. Once $\hat{\imath}$ is obtained we reject H_i for some $i = 1, ..., \hat{\imath}$ thereby not rejecting the null hypothesis at $\hat{\imath} + 1$ and accepting the threshold associated with $H_0^{\hat{\imath}+1}$.

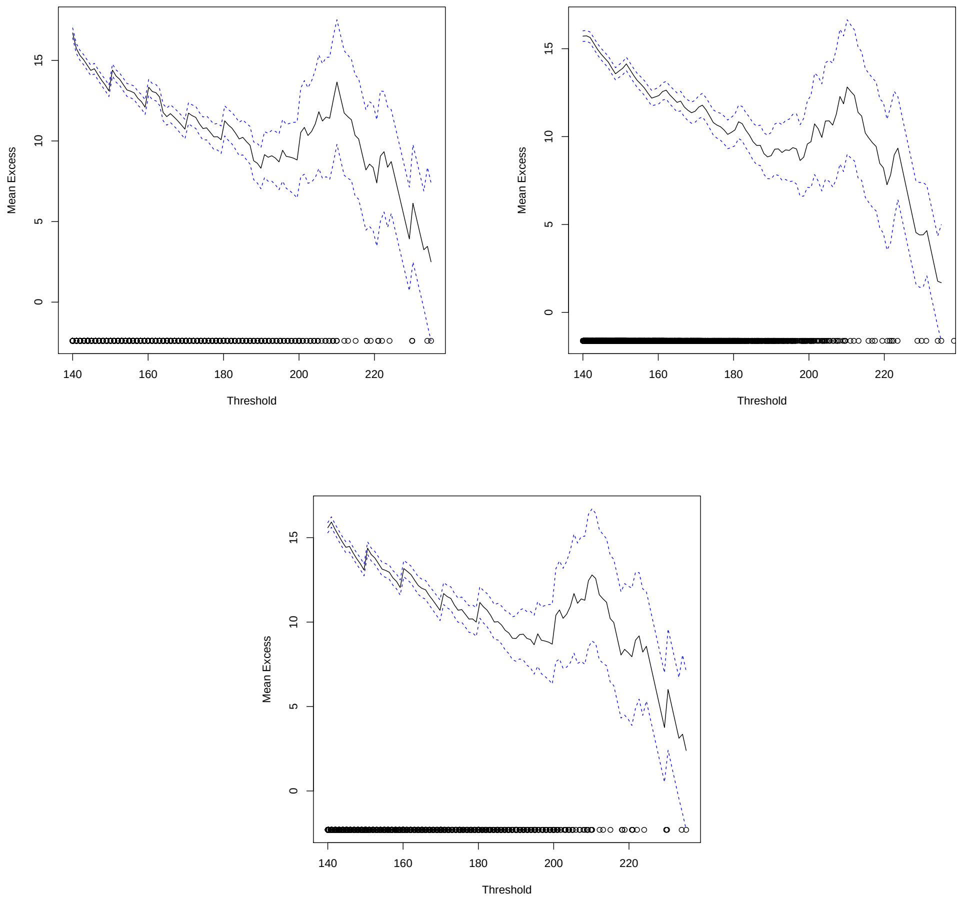

Figure 11: Mean residual life function for the non-jitter data (top left), uniform-jitter data (top right) and beta jitter data (bottom).

threshold	**num.above**	***p*-values**	**fowardstop**	**statistic**
140	7113	2.4221e-47	∼0	3.3111
150	4012	1.3233e-46	∼0	5.3901
160	2065	1.1008e-06	3.66926e-07	0.6684
170	973	1.4698e-05	3.9497e-06	0.5456
180	416	1.4609e-03	2.9556e-04	0.3260
190	173	6.8810e-02	1.2128e-02	0.1320
200	53	1.6762e-01	3.6604e-02	0.0988

Table 7: Results of the automated threshold selection using the Cramér-Von Mises goodness-of-fit tests for the non-jitter data set.

Table 7 illustrates the results of the FowardStop rule for the non-jitter data using the R package *eva* as outlined in [2]. The results show that we should re-

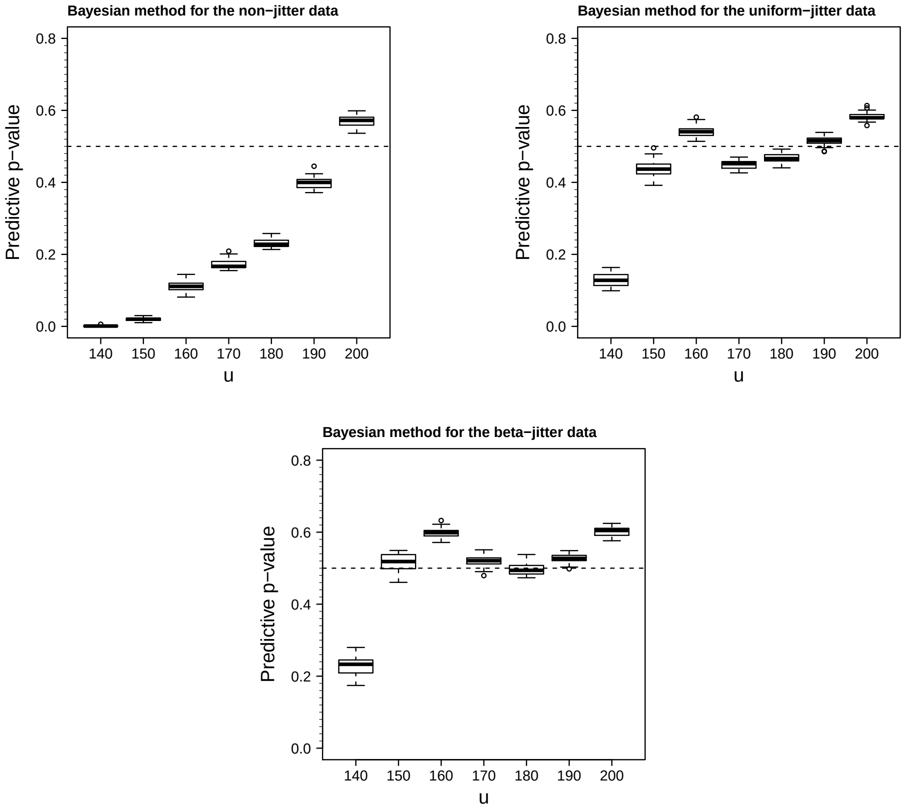

Figure 12: Bayesian threshold selection method using measure of surprise for the non-jitter (top left), uniform-jitter (top right) and beta-jitter data (bottom). The value u represents the SBP threshold.

ject the first five hypotheses at $\alpha = 0.01$ and select 190 mmHg as the adequate threshold, since the fifth test is the last test where the FowardStop mean, indicated in (4.1), is still below 0.01. We would like to point out that these results are in accordance with the results obtained from the Bayesian threshold selection method. Table 8 shows the results of the FowardStop rule for the uniform-jitter data set. Here the rule proposes a lower threshold. Effectively, 190 mmHg is the first threshold that produces a p-value above 0.01. However, this p-value is much larger than 0.01, which suggests that a proper threshold might lie between 180 mmHg and 190 mmHg.

Table 9 shows the FowardStop rule results for the beta-jitter data set. Here the first 5 hypotheses are rejected at α=0.01, suggesting that 190 is an adequate threshold.

Finally, $u = 190$ mmHg was the threshold selected for the three cases. The parameters, k and σ of the GPD fitted models for each data set are presented in Table 10.

threshold	**num.above**	***p*-values**	**fowardstop**	**statistic**
140	7593	1.4383e-67	∼0	7.9165
150	4391	2.5836e-08	1.2918e-08	0.8708
160	2269	1.4901e-03	4.9709e-04	0.3264
170	1064	2.0113e-03	8.7615e-04	0.3118
180	471	3.5537e-03	1.4129e-03	0.2763
190	196	5.7393e-01	1.4337e-01	0.0496
200	63	1.5124e-01	1.4631e-01	0.0960

Table 8: Results of the automated threshold selection using the Cramér-Von Mises goodness-of-fit tests for the uniform-jitter data set.

threshold	**num.above**	***p*-values**	**fowardstop**	**statistic**
140	7628	5.2077e-25	∼0	2.8071
150	4391	1.9948e-18	∼0	1.9746
160	2260	5.7836e-09	1.9279e-09	0.9210
170	1072	1.2669e-06	3.1817e-07	0.6539
180	468	1.3996e-04	2.8249e-05	0.4265
190	227	1.2325e-01	2.1946e-02	0.1095
200	74	1.8077e-01	4.7295e-02	0.0930

Table 9: Results of the automated threshold selection using the Cramér-Von Mises goodness-of-fit tests for the beta-jitter data set.

	u	N	n	$\hat{k}$	95% CI for k	σ	95% CI for σ	max	$endpoint$	$-log(L)$
M_I	190	8174	173	-0.049	(-0.190,0.093)	10.50	(8.34,12.65)	240	406.08	571.34
M_{II}	190	7593	196	0.062	(-0.097,0.222)	8.37	(6.60,10.15)	238.50	$*$	624.78
M_{III}	190	7628	192	0.062	(-0.100,0.224)	8.47	(6.65,10.29)	239.98	$*$	614.13

Table 10: GPD(k,σ) models fitted to the non-jitter (M_I), uniform-jitter (M_{II}) and beta-jitter (M_{III}) data. ($*$) indicates the support does not have an upper finite boundary.

The results presented in Table 10 were obtained using the *gpd.fit* function from the *ismev* R package.

Table 10 shows that the estimates of k are very close to zero, which indicates that the GPD might be reduced to the exponential model. Although the 95% confidence intervals contain zero, they are skewed to the left for the model fitted to the original data and to the right in the other two cases.

In order to evaluate whether the three GPD(k,σ) models can be reduced to the more parsimonious exponential model, a hypothesis test was performed. The null hypothesis $H_0 : k = 0$ was tested against $H_1 : k \neq 0$ using the likelihood ratio test. Under H_0,

$$T = 2\big(l_{M_1}(x) - l_{M_2}(x)\big) \sim \chi^2_1,$$

where in this case l_{M_1} is the log-likelihood function for the GPD(k,σ) model and

$l_{M_2}(x)$ is the log-likelihood function for the exponential model. The results are presented in Table 11. The large p-values obtained support that the exponential model should be selected in the three cases.

Model	l_{M_1}	$l_{M_2}(x)$	T	p
Non-jitter	-571.3383	-571.7379	0.7992268	0.3713246
Uniform-jitter	-624.7814	-625.5246	1.486305	0.2227907
Beta-jitter	-614.1304	-614.8442	1.427657	0.2321473

Table 11: Results of the deviance test for non-jitter model, uniform-jitter model and beta jitter-model.

Table 12 presents a comparison between some empirical quantiles and model quantiles. It is important to note that this comparison does not serve as a true accuracy measure of the model performance since the extreme empirical quantiles were calculated with a very small number of observations. The empirical quantile estimation was obtained using the R function *quantile*. The uniform-jitter model and the beta-jitter model supplied highly accurate quantile estimates when compared to the empirical ones.

Model	**Empirical**	**Model**	**IC 95%**	**Empirical**	**Model**	**IC 95%**
	$q_{0.99}$	$q_{0.99}$	$q_{0.99}$	$q_{0.995}$	$q_{0.995}$	$q_{0.995}$
Non-jitter	198.00	197.51	(196.38,198.63)	203.00	204.45	(202.29,206.60)
Unif-jitter	198.63	198.47	(197.28,199.66)	203.98	204.66	(202.60,206.71)
Beta-jitter	198.69	198.33	(197.15,199.52)	203.95	204.59	(202.52,206.66)

Table 12: Extreme quantiles for the uniform-jitter model, beta-jitter model and non-jitter model using the exponential model.

Table 12 shows that the fitted models perform in a very similar way in terms of extreme quantile estimation, irrespective of the data set being used. This result is quite unexpected and it shows that the discretization of the SBP readings did not, after all, produce significant inaccuracies.

Next we extrapolate on the likelihood of observing an individual with SBP value higher than the maximum value observed in each data set using the non-jitter, uniform-jitter and beta-jitter exponential models.

- Non-jitter model: $P(\widehat{SBP > 240}) = 0.000143$
- Uniform-jitter model: $P(\widehat{SBP > 238.5}) = 0.000113$
- Beta-jitter model: $P(\widehat{SBP > 239.98}) = 0.000099$

The resulting probability from the non-jitter model is higher than both probabilities produced by the jitter models, which is a consequence of the non-jitter and jitter models providing dissimilar scale parameter estimates, $\hat{\sigma}$ =10.01

for the non-jitter model and $\hat{\sigma}$ =8.93 and $\hat{\sigma}$ =9.03 for the uniform-jitter and beta-jitter models, respectively. Moreover, the resulting probabilities from the jitter models yielded similar results. The deflated estimates of the scale parameters for the jittering models might be a consequence of the jittering process considered.

5. CONCLUSION

Preliminary analysis of the resulting jitter data sets demonstrate that we have been successful in *breaking* the discrete feature of the recorded data, see Figure 7. Moreover, the jittering process did not alter the data a great deal, as described in Table 6.

The threshold $u = 190$ mmHg was in the end selected for each case and subsequently the models were fitted to the data. Table 10 displays the estimated parameters for the model. Although the fitted k is negative for the non-jitter data and positive for both cases of jitter data, all the values are very close to zero reflecting an exponential tail. In fact, the 95% confidence intervals for the shape parameter, in each case, includes 0. This conjecture is further investigated by applying the deviance test. The results indicate that there are no significant differences between the GPD and the exponential distribution for each case, as displayed in Table 11. Future work could be developed using other jittering distributions. For example, a stronger jitter could be applied to the values with higher than normal absolute frequencies and a milder jitter to the remaining data.

ACKNOWLEDGMENTS

The second author was financially supported by FCT - Fundação para a Ciência e a Tecnologia, Portugal, through the projects UID/MAT/00006/2013, UID/MAT/00006/2019, and PTDC/MAT-STA/28649/2017.

The authors thank the Department of Pharmaceutical Care Services of the Portuguese National Association of Pharmacies for allowing the use of the data that enabled this work.

The authors would also like to thank Lee, J., Fan, Y. and Sisson, S. A. [24] for courteously providing the R code used to apply the Bayesian method for threshold selection.

The authors also thank Dr. Eduardo Gomes da Silva, head of the Serviço de Medicina 3.2 of the Hospital dos Capuchos, Lisbon, for his availability to answer their questions about blood pressure and its pathologies.

The authors thank the referee for the careful reading of the paper and for the valuable suggestions that definitely improved the paper.

REFERENCES

[1] Apostol, T.M. (1967). *Calculus: One-Variable Calculus, with an Introduction to Linear Algebra*, John Wiley & Sons, New York.

[2] Bader, B.; Yan, J. and Zhang, X. (2018). Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate, *Annals of Applied Statistics*, **12**, 1, 310–329.

[3] Balkema, A.A. and de Hann, L. (1974). Residual life time at great age, *Annals of Probability*, **2**, 5, 792–804.

[4] Bavishi, C.; Goel, S. and Messerli, F.H. (2016). Isolated systolic hypertension: An update after sprint, *The American Journal of Medicine*, **129**, 12, 1251–1258.

[5] Caetano, C.P. (2018). An application of extreme value theory in medical sciences, MSc Thesis, University of Lisbon, Lisbon.

[6] Casella, G. and Berger, R.L. (2002). *Statistical Inference (2nd ed.)*, Duxbury/Thomson Learning, Pacific Grove, California.

[7] Castillo, E. and Hadi, Ali S. (1997). Fitting the generalized Pareto distribution to data, *Journal of the American Statistical Association*, **92**, 440, 1609–1620.

[8] Chen, J.; Lei, X.; Zhang, L. and Peng, B. (2015). Using Extreme Value Theory Approaches to Forecast the Probability of Outbreak of Highly Pathogenic Influenza in Zhejiang, China, *PloS one*, **10**, 2.

[9] Choulakian, V. and Stephens, M.A. (2001). Goodness-of-fit tests for the generalized Pareto distribution, *Technometrics*, **43**, 4, 478–484.

[10] Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*, Springer-Verlag, London.

[11] de Zea Bermudez, P. and Kotz, S. (2010). Parameter estimation of the generalized Pareto distribution - part I, *Journal of Statistical Planning and Inference*, **140**, 6, 1353–1373.

[12] de Zea Bermudez, P. and Kotz, Samuel (2010). Parameter estimation of the generalized Pareto distribution - part II, *Journal of Statistical Planning and Inference*, **140**, 6, 1374–1388.

[13] de Zea Bermudez, P. and Mendes, Z. (2012). Extreme value theory in medical sciences: Modeling total high cholesterol levels, *Journal of Statistical Theory and Practice*, **6**, 3, 468–491.

[14] DuMouchel, W.H. (1983). Estimating the stable index α in order to measure tail thickness: A critique, *Annals of Statistics*, **11**, 4, 1019–1031.

[15] Gonzaga, C.C.; Sousa, M.G. and Amodeo, C. (2009). Fisiopatologia da hipertensão sistólica isolada, *Revista Brasileira da Hipertensão*, **16**, 1, 10–14.

[16] Grimshaw, S.D. (1993). Computing maximum likelihood estimates for the generalized Pareto distribution, *Technometrics*, **35**, 2, 185–191.

[17] G'Sell, M.G.; Wager, S.; Chouldechova, A. and Tibshirani, R. (2015). Sequential selection procedures and false discovery rate control, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **78**, 2, 423–444.

[18] Guillou, A.; Kratz, M. and Le Strat, Y. (2014). An extreme value theory approach for the early detection of time clusters. A simulation-based assessment and an illustration to the surveillance of Salmonella, *Statistics in Medicine*. **33**, 28, 5015–5027.

[19] Gumbel, E.J. (1935). Les valeurs extrêmes des distributions statistiques, *Annales de l'Institut Henri Poincaré*, **5**, 2, 115–158.

[20] Hajar, R. (2016). Framingham contribution to cardiovascular disease, *Heart Views: The Official Journal of the Gulf Heart Association*, **17**, 2, 78–81.

[21] Heffernan, J.E. and Tawn, J.A. (2004). A conditional approach for multivariate extreme values (with discussion), *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **66**, 3, 497–546.

[22] Hosking, J.R.M. and Wallis, J.R. (1987). Parameter and quantile estimation for the generalized Pareto distribution, *Technometrics*, **29**, 3, 339–349.

[23] Kass, R.E. and Raftery, A.E. (1995). Bayes factors, *Journal of the American Statistical Association*, **90**, 430, 773–795.

[24] Lee, J.; Fan, Y. and Sisson, S. (2015). Bayesian threshold selection for extremal models using measures of surprise, *Computational Statistics & Data Analysis*, **85**, 84–99.

[25] Meng, X.-L. (1994). Posterior predictive p-values, *Annals of Statistics*, **22**, 3, 1142–1160.

[26] Pickands, J.III (1975). Statistical inference using extreme order statistics, *Annals of Statistics*, **3**, 1, 119–131.

[27] Pinto, E. (2007). Blood pressure and ageing, *Postgraduate Medical Journal*, **83**, 976, 109–114.

[28] Scarrott, C. and MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification, *REVSTAT - Statistical Journal*, **10**, 1, 33–60.

[29] Schröder, H.; Marrugat, J.; Elosua, R. and Covas, M.I. (2003). Relationship between body mass index, serum cholesterol, leisure-time physical activity, and diet in Mediterranean Southern-Europe population, *British Journal of Nutrition*, **90**, 2, 431–439.

[30] Thomas, M.; Lemaitre, M.; Wilson, M.L.; Viboud, C.; Yordanov, Y.; Wackernagel, H. and Carrat, F. (2016). Applications of extreme value theory in public health, *PLoS One*, **11**, 7.

[31] Paulino, C.D.; Amaral Turkman, M.A.; Murteira, B. and Silva, G.L. (2018). *Estatística Bayesiana, 2ª edição*, Fundação Calouste Gulbenkian, Lisboa.

[32] Wakabayashi, I. (2004). Relationships of body mass index with blood pressure and serum cholesterol concentrations at different ages, *Aging Clinical and Experimental Research*, **16**, 6, 461–466.