



INSTITUTO NACIONAL DE ESTATÍSTICA  
STATISTICS PORTUGAL

# REVSTAT

## Statistical Journal

Special issue on Biometry  
Celebrating the 3rd Portuguese-Galician Meeting of Biometry



**Guest Editors:**

Giovani Silva  
Inês Sousa  
Lisete Sousa  
Javier Roca Pardiñas

Volume 17, No.2

April 2019

# REVSTAT

Statistical Journal

## Catálogo Recomendada

**REVSTAT.** Lisboa, 2003-  
Revstat : statistical journal / ed. Instituto Nacional  
de Estatística. - Vol. 1, 2003- . - Lisboa I.N.E.,  
2003- . - 30 cm  
Trimestral. - Continuação de : Revista de Estatística =  
ISSN 0873-4275. - edição exclusivamente em inglês  
ISSN 1645-6726

## CREDITS

### - EDITOR-IN-CHIEF

- *Isabel Fraga Alves*

### - CO-EDITOR

- *Giovani L. Silva*

### - ASSOCIATE EDITORS

- *Marília Antunes*
- *Barry ARNOLD (2019)*
- *Narayanaswamy Balakrishnan*
- *Jan Beirlant*
- *Graciela Boente (2019-2020)*
- *Paula Brito*
- *Vanda Inácio de Carvalho*
- *Arthur Charpentier*
- *Valérie Chavez-Demoulin*
- *David Conesa*
- *Charmaine Dean*
- *Jorge Milhazes Freitas*
- *Alan Gelfand*
- *Stéphane Girard*
- *Wenceslao Gonzalez-Manteiga*
- *Marie Kratz*
- *Victor Leiva*
- *Maria Nazaré Mendes-Lopes*
- *Fernando Moural*
- *John Nolan*
- *Paulo Eduardo Oliveira*
- *Pedro Oliveira*
- *Carlos Daniel Paulino (2019-2021)*
- *Arthur Pewsey*
- *Gilbert Saporta*
- *Alexandra M. Schmidt*
- *Julio Singer*

- *Manuel Scotto*

- *Lisete Sousa*

- *Milan Stehlik*

- *María Dolores Ugarte*

### - EXECUTIVE EDITOR

- *José A. Pinto Martins*

### - FORMER EXECUTIVE EDITOR

- *Maria José Carrilho*

- *Ferreira da Cunha*

### - SECRETARY

- *Liliana Martins*

### - PUBLISHER

- *Instituto Nacional de Estatística, I.P. (INE, I.P.)*  
*Av. António José de Almeida, 2*  
*1000-043 LISBOA*  
*PORTUGAL*  
*Tel.: + 351 21 842 61 00*  
*Fax: + 351 21 845 40 84*  
*Web site: <http://www.ine.pt>*  
*Customer Support Service*  
*+ 351 218 440 695*

### - COVER DESIGN

- *Mário Bouçadas, designed on the stain glass window at INE by the painter Abel Manta*

### - LAYOUT AND GRAPHIC DESIGN

- *Carlos Perpétuo*

### - PRINTING

- *Instituto Nacional de Estatística, I.P.*

### - EDITION

- *140 copies*

### - LEGAL DEPOSIT REGISTRATION

- *N.º 191915/03*

### - PRICE [VAT included]

- *€ 9,00*



## Editorial of Special Issue on Biometry

This Special Issue on Biometry arose from the III Portuguese-Galician Meeting of Biometry (**EBio2018**), jointly organized by the Portuguese Statistical Society (SPE) - Biometry Section - and by the Galician Society for the Advancement of Statistics and Operations Research (SGAPEIO), and hosted by the University of Aveiro (Portugal), which took place from 28 to 30 June 2018.

**EBio2018** was preceded by I Portuguese-Galician Meeting of Biometry, I Portuguese-Galician Meeting on Ecological and Environmental Statistics and II Galician-Portuguese Meeting of Biometry with applications to the Health Sciences, Ecology and Environmental Sciences, held in Braga (2013), Vila Real (2014) and Santiago de Compostela (2016), respectively. These meetings aim, namely, to expand the field of action of both societies to new circles of the biometry community, and to promote the exchange and to intensify the relationships within each community and between the statistical communities.

According to the International Biometric Society, “The term **Biometrics/Biometry** has been used since early in the 20th century to refer to the field of development of statistical and mathematical methods applicable to data analysis problems in the biological sciences. Statistical methods for the analysis of data from agricultural field experiments to compare the yields of different varieties of wheat, for the analysis of data from human clinical trials evaluating the relative effectiveness of competing therapies for disease, or for the analysis of data from environmental studies on the effects of air or water pollution on the appearance of human disease in a region or country are all examples of problems that would fall under the umbrella of **Biometrics/Biometry** as the term has been historically used”.

Based on the increasing importance of the areas mentioned in the above definition, we decided to challenge **EBio2018** participants to submit their contributed papers to REVSTAT - Statistical Journal - for a Special Issue on Biometry. We intended to promote the dissemination of the latest advances in the development and application of statistical and mathematical methods in Biology, Medicine, Ecology, Psychology, Pharmacology, Agriculture, Environment and other Health and Life Sciences.

After a peer-review process, six manuscripts were accepted for publication in this issue covering the following topics: i) Model risks of extreme events in population dynamics, ii) Peaks over threshold methods to estimate extreme quantiles and probabilities related to hypertension pathology, iii) Assessing extreme value conditions motivated by two real environmental problems, iv) Parameters estimation of HIV dynamic models, v) Accuracy measures for binary classification in the selection of the optimal cut-point, vi) Joint modelling of longitudinal and competing risks clinical data.

Finally, we (guest editors) would like to thank all authors for their contributions and all the anonymous reviewers who helped to prepare this special issue. Furthermore, we are grateful to the past and current Editors-in-Chief of REVSTAT - Statistical Journal - for agreeing to publish this special issue, as well as to all members of the scientific and organizing committees who worked to make **EBio2018** a very interesting event on the field of Statistical Models in Biometry.

GIOVANI L. SILVA  
DMIST & CEAUL, Universidade de Lisboa  
giovani.silva@tecnico.ulisboa.pt

LISETE SOUSA  
DEIO & CEAUL, Universidade de Lisboa  
lmsousa@fc.ul.pt

INÊS SOUSA  
DMA, Universidade do Minho  
isousa@math.uminho.pt

JAVIER ROCA PARDIÑAS  
DEIO, Universidade de Vigo  
roca@uvigo.es

# INDEX

## **Modeling risk of extreme events in generalized Verhulst models**

*M. Fátima Brilhante, M. Ivette Gomes and Dinis Pestana* ..... 145

## **Modeling large values of systolic blood pressure in the portuguese population**

*Constantino P. Caetano and Patrícia de Zea Bermudez* ..... 163

## **Testing conditions and estimating parameters in extreme value theory: application to environmental data**

*Helena Penalva, Dora Prata Gomes, M. Manuela Neves and Sandra Nunes* .....187

## **On the parameters estimation of HIV dynamic models**

*Diana Rocha, Sónia Gouveia, Carla Pinto, Manuel Scotto, João Nuno Tavares, Emília Valadas and Luís Filipe Caldeira* ..... 209

## **Accuracy measures for binary classification based on a quantitative variable**

*Rui Santos, Miguel Felgueiras, João Paulo Martins and Liliana Ferreira* ..... 223

## **Joint modelling of longitudinal and competing risks data in clinical research**

*Laetitia Teixeira, Inês Sousa, Anabela Rodrigues and Denisa Mendonça* ..... 245



---

---

## MODELING RISK OF EXTREME EVENTS IN GENERALIZED VERHULST MODELS

---

---

Authors: M. FÁTIMA BRILHANTE

– Faculdade de Ciências e Tecnologia, Universidade dos Açores,  
Centro de Estatística e Aplicações da Universidade de Lisboa  
Portugal  
maria.fa.brilhante@uac.pt

M. IVETTE GOMES

– Centro de Estatística e Aplicações da Universidade de Lisboa,  
Instituto de Investigação Científica Bento da Rocha Cabral  
Portugal  
migomes@fc.ul.pt

DINIS PESTANA

– Centro de Estatística e Aplicações da Universidade de Lisboa,  
Instituto de Investigação Científica Bento da Rocha Cabral  
Portugal  
dinis.pestana@fc.ul.pt

Received: October 2018

Revised: January 2019

Accepted: March 2019

Abstract:

- A very popular model in population dynamics, which has been around since the first half of the nineteenth century, is the Verhulst logistic model. However, some limitations of this model have provided grounds to propose more sophisticated growth models using, for instance, the former as a basis. Since the Verhulst model and some generalizations of it are closely connected to extreme value distributions, either max-geometric-stable or max-stable, we show that the parameter attached to the retroaction factor of these generalized models establishes, on its own, which extreme value distribution is adequate to model risks of extreme events in population dynamics.

Key-Words:

- *Extreme value distributions; generalized Verhulst models; growth and retroaction parameters; max-geometric-stable distributions; population dynamics.*

AMS Subject Classification:

- 60G70, 92D25.





---

## 1. INTRODUCTION

---

*“It is generally agreed that the specific growth rate [...] declines as density increases, and hence that the form of the population curve with time in a limited system has a sigmoid shape. Of the many proposed models only one, the logistic of Verhulst (1838) [...] is widely used. It is presented in most current ecology texts and is incorporated into almost all fish and game management theories. Such tacit acceptance probably derives from its mathematical simplicity and biological clarity.”*

Smith (1963)

Let  $N(t)$  be the size of a population and  $R(t)$  the amount of available resources at time  $t$ . It is reasonable to relate  $R(t)$  and  $N(t)$  by the differential equation

$$(1.1) \quad \frac{d}{dt}R(t) = -\eta \frac{d}{dt}N(t),$$

with  $\eta$  representing the amount of resources consumed to yield a new population unit. The solution of (1.1) is

$$R(t) = \eta(K - N(t)) = R(0) - \eta N(t),$$

and hence  $K = R(0)/\eta > 0$  is the carrying capacity, *i.e.* the limiting size the population may reach without disruptive effects on the availability of resources.

On the other hand, it also makes sense to consider that the population growth rate is proportional to the amount of available resources, namely

$$\frac{\frac{d}{dt}N(t)}{N(t)} = \mu R(t).$$

Therefore,

$$(1.2) \quad \frac{d}{dt}N(t) = \rho N(t) \left(1 - \frac{N(t)}{K}\right),$$

where  $\rho = \mu R(0) > 0$  is the malthusian intrinsic growth rate, or growth rate *per capita*. In the right side of equation (1.2),  $N(t)$  is considered to be the growth factor and  $1 - N(t)/K$  the retroaction factor, which is responsible for curbing down population growth to sustainable levels. The solution of (1.2), known as the Verhulst model (Verhulst [16]), is

$$(1.3) \quad N(t) = \frac{KN(0)}{N(0) + (K - N(0))e^{-\rho t}},$$

which belongs to the logistic family of functions, hence the name logistic model ( $N(0)$  is the initial population size).

On some occasions, it is more convenient to express the Verhulst logistic equation (1.2) as a function of the population density  $\delta(t) = N(t)/K$ , namely

$$(1.4) \quad \frac{d}{dt}\delta(t) = \rho\delta(t)(1 - \delta(t)).$$

The solution of (1.4) is

$$\delta(t) = \frac{1}{1 + \exp(-\rho t)},$$

which is a member of the logistic family of distributions. As pointed out in Smith [13], over time the population curve will have a sigmoid shape, which is typical of continuous distribution functions.

In spite of its popularity, the Verhulst model has some limitations. For instance, one limitation is only being suitable for modeling sustainable growth, or modeling stable populations, in the sense that the population sizes are maintained at sustainable levels. Therefore, over the years the Verhulst model has been used as a building block for other more sophisticated models, some of which allowing the possibility of modeling different types of unrestricted population growth.

Many of newer models state that either  $\frac{d}{dt}N(t)$  or  $\frac{d}{dt}\ln N(t)$  is a decreasing function of the population density (as does the Verhulst model). As an example, we have the family of models based on the Box-Cox family of transformations (Box and Cox [3])

$$(1.5) \quad \frac{d}{dt}\ln N(t) = \rho \frac{1 - \left(\frac{N(t)}{K}\right)^\nu}{\nu} \Leftrightarrow \begin{cases} \frac{d}{dt}N(t) = \rho N(t) \frac{1 - \left(\frac{N(t)}{K}\right)^\nu}{\nu} & , \nu > 0, \\ \frac{d}{dt}N(t) = \rho N(t) \left(-\ln\left(\frac{N(t)}{K}\right)\right) & , \nu = 0 \end{cases},$$

which contains the Verhulst model as a special case ( $\nu = 1$ ). The subfamily in (1.5) for  $\nu > 0$  was considered in Richards [12], and the solution for  $\nu = 0$  is

$$N(t) = K \exp\left(\ln\left(\frac{N(0)}{K}\right) \exp(-\rho t)\right),$$

which is commonly known in population dynamics as the Gompertz growth model. This model is proportional to the Gumbel distribution, a well known extreme value (EV) distribution for maxima, and has been used, for instance, to model the growth of cancer tumors. Note that the Gumbel distribution has the functional form

$$(1.6) \quad \Lambda(x; \lambda, \delta) = \exp(-\exp(-(x - \lambda)/\delta)), \quad x \in \mathbb{R}, (\lambda, \delta) \in \mathbb{R} \times \mathbb{R}^+,$$

where  $\lambda$  and  $\delta$  are, respectively, location and scale parameters.

A natural extension of Verhulst's equation (1.2) is the Blumberg hiperlogistic equation (Blumberg [2])

$$(1.7) \quad \frac{d}{dt}N(t) = \rho (N(t))^\alpha \left(1 - \frac{N(t)}{K}\right)^\beta, \quad \alpha, \beta > 0.$$

However, Blumberg’s equation does not contain a closed form analytical solution, except for some values of the parameters  $\alpha$  and  $\beta$ . For example, if  $\alpha + \beta = 2$  (and  $K = 1$ ), the solution of (1.7) belongs to the class of max-geometric-stable distributions, where the shape parameter is a function of the retroaction parameter  $\beta$  in (1.7).

On the other hand, Brillhante *et al.* [4] extended the subfamily in (1.5) for  $\nu = 0$  by considering

$$(1.8) \quad \frac{d}{dt}N(t) = \rho N(t) \left( -\ln \left( \frac{N(t)}{K} \right) \right)^{1+\xi}, \quad \xi \in \mathbb{R}.$$

Those authors showed that the solution of (1.8), when  $K = 1$ , belongs to the general EV (GEV) family of distributions for maxima, with the functional form

$$(1.9) \quad G_\xi(x; \lambda, \delta) = \exp \left( -(1 + \xi(x - \lambda)/\delta)^{-1/\xi} \right), \quad 1 + \xi(x - \lambda)/\delta > 0,$$

where  $\xi \in \mathbb{R}$ ,  $\lambda \in \mathbb{R}$  and  $\delta \in \mathbb{R}^+$  are, respectively, shape, location and scale parameters. Observe that equation (1.8) can also be considered a generalization of Verhulst’s logistic equation, since  $1 - N(t)/K$  is a linear approximation of  $-\ln(N(t)/K)$ , due to the fact that  $N(t)/K \rightarrow 1$ , as  $t \rightarrow \infty$ . The effect of replacing  $1 - N(t)/K$  by  $-\ln(N(t)/K)$  in (1.2) is that we shall have a weaker control over population growth than before. This weaker control effect can easily be explained by noticing that if  $x \in (0, 1)$ ,  $1 - x$  is proportional to the density function of the minimum  $U_{1:2} = \min(U_1, U_2)$  and  $-\ln x$  is the density function of the product  $U_1U_2$ , where  $U_1$  and  $U_2$  are two independent standard uniform random variables, and thus the stochastic ordering  $U_1U_2 \preceq U_{1:2}$  holds true.

An even more general differential equation for population dynamics, based on the BetaBoop family of densities, was considered in Brillhante *et al.* [5], namely

$$(1.10) \quad \frac{d}{dt}N(t) = \rho (N(t))^\alpha [ -\ln(1 - N(t)) ]^\beta (1 - N(t))^\gamma ( -\ln N(t) )^\delta,$$

where  $\alpha, \beta, \gamma, \delta > 0$ . The previous equation includes equations (1.7) and (1.8) as special cases, but goes even further by allowing simultaneously two different growth factors depicted in  $(N(t))^\alpha$  and  $[ -\ln(1 - N(t)) ]^\beta$ , as well as two different environmental retroaction factors indicated by  $(1 - N(t))^\gamma$  and  $( -\ln N(t) )^\delta$ . Observe now that the growth factor  $N(t)$  can be considered as a linear approximation of the growth factor  $-\ln(1 - N(t))$ , but with the latter stimulating more growth than the former. However, equation (1.10) does not contain a closed form analytical solution, unless for some special combinations of the parameters  $\alpha$ ,  $\beta$ ,  $\gamma$  and  $\delta$ . For more information on other population growth models, cf. Lotka [10], Tsoularis [14] and Tsoularis and Wallace [15].

The EV distributions that arise as solutions to the Verhulst and some generalized Verhulst equations seem to indicate that there is a close connection between population dynamics and forms of extreme stability. This was our motivation to investigate what kind of relationship is indeed present. Therefore, in Section 2, we shall show that the parameter attached to the retroaction factor of some generalized Verhulst equations determines, on its own, and in most situations, which EV distribution for maxima is suitable to model risks of extreme events in population dynamics. Finally, in Section 3, some overall comments are further provided.

---

## 2. EXTREME STABILITY IN SOME GENERALIZED VERHULST MODELS

---



---

### 2.1. Some basic facts in extreme value theory

---

In extreme value theory (EVT) the logistic distribution, which arises as the solution of Verhulst's normalized logistic equation (1.4), is known to be one of three types of max-geometric-stable distributions, the other two being the log-logistic and backward log-logistic distributions (Rachev and Resnick [11]).

**Definition 2.1.** A distribution function  $H$  is a max-geometric-stable distribution if for all  $\theta \in (0, 1)$ , there exist real numbers  $a_\theta = a(\theta) > 0$  and  $b_\theta = b(\theta)$  such that

$$H(a_\theta x + b_\theta) = \frac{\theta H(x)}{1 - (1 - \theta)H(x)}.$$

Basically, if  $\{X_n\}_{n \in \mathbb{N}}$  is a sequence of independent and identically distributed random variables and  $X_{N:N} = \max(X_1, \dots, X_N)$  is the random maximum, where  $N$  is a geometric random variable of mean  $1/\theta$ , independent of each  $X_n$ , then, as  $\theta \rightarrow 0$ , max-geometric-stable distributions are the only possible non-degenerate limiting distributions for sequences of linearly normalized random maxima  $(X_{N:N} - b_\theta)/a_\theta$ .

Another well known fact in EVT is that GEV distributions for maxima, defined in (1.9), are the unique max-stable distributions.

**Definition 2.2.** A distribution function  $G$  is a max-stable distribution if for all  $n \in \mathbb{N}$ , there exist real numbers  $\alpha_n = \alpha(n) > 0$  and  $\beta_n = \beta(n)$  such that

$$G^n(\alpha_n x + \beta_n) = G(x).$$

In other words, and as  $n \rightarrow \infty$ , max-stable distributions are the only possible non-degenerate limiting distributions for sequences of linearly normalized

maxima  $(X_{n:n} - b_n)/a_n$ , with  $X_{n:n} = \max(X_1, \dots, X_n)$ , for  $\{X_n\}_{n \in \mathbb{N}}$  a sequence of independent and identically distributed random variables (Gnedenko [7]), or more generally, for stationary weakly dependent random variables with distribution function  $F$  (Leadbetter *et al.* [9]). If the aforementioned non-degenerate limit exists, we then say that  $F$  is in the domain of attraction for maxima of  $G_\xi$ , in (1.9).

Initially, in Gnedenko's seminal paper, there appeared three types of max-stable distributions, which can indeed be combined into a single family, the GEV family of distributions for maxima in (1.9). In particular, if  $\xi > 0$ , we have the so-called Fréchet distribution, if  $\xi < 0$ , we obtain the Weibull distribution for maxima and if  $\xi = 0$ , we get the Gumbel distribution, already defined in (1.6), by taking the limit of (1.9) as  $\xi \rightarrow 0$ . The shape parameter  $\xi$  in (1.9) is the extreme value index (EVI), a very important parameter associated with extreme events.

**Remark 2.1.** From the relation between the minimum and the maximum, namely  $\min(X_1, \dots, X_n) = -\max(-X_1, \dots, -X_n)$ , we get the GEV distribution for minima, defined by  $G_\xi^*(x; \lambda, \delta) = 1 - G_\xi(-x; \lambda, \delta)$ , with  $G_\xi(x; \lambda, \delta)$  given in (1.9).

In Section 1 we mentioned that, when  $K = 1$ , the GEV distribution for maxima appears as the solution of equation (1.8), which can be regarded as a generalized Verhulst equation. This makes some sense because there is a strong connection between max-geometric-stable and max-stable distributions. More precisely, if  $G_\xi$  represents the distribution function of a GEV distribution for maxima, with EVI  $\xi$ , and  $H = H_\xi$  represents the distribution function of a max-geometric-stable distribution, we have

$$H_\xi(x; \lambda, \delta) = \frac{1}{1 - \ln G_\xi(x; \lambda, \delta)} = \frac{1}{1 + (1 + \xi(x - \lambda)/\delta)^{-1/\xi}}, \quad 1 + \xi(x - \lambda)/\delta > 0,$$

with  $(\xi, \lambda, \delta) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$ . Therefore, we have a close relationship between the log-logistic and Fréchet distributions ( $\xi > 0$ ), between the backward log-logistic and Weibull for maxima distributions ( $\xi < 0$ ) and between the logistic and Gumbel distributions ( $\xi = 0$ ).

In the next subsection we shall be particularly interested in investigating which EV distribution is adequate to model risks of extreme events in population dynamics, when using some generalized Verhulst models. To this end, we recall one of the first order condition for establishing domains of attraction for maxima (or simply max-domains of attraction). In particular, we shall work with the first order condition given in de Haan [6], which is equivalent to the first order condition given in Gnedenko [7].

We say that a distribution function  $F$  belongs to the max-domain of attraction of a GEV distribution  $G_\xi$ , and we use the notation  $F \in \mathcal{D}_M(G_\xi)$ , if, and only if,

$$(2.1) \quad \lim_{t \rightarrow \infty} \frac{U(tx) - U(t)}{a(t)} = \begin{cases} \frac{x^\xi - 1}{\xi} & , \xi \neq 0 \\ \ln x & , \xi = 0 \end{cases}, \quad x > 0,$$

where  $U(t) = F^{\leftarrow}(1 - \frac{1}{t})$ ,  $t \geq 1$ , is the reciprocal tail quantile function,  $F^{\leftarrow}(y) = \inf\{x : F(x) \geq y\}$  is the generalized inverse function of  $F$  and  $a(\cdot)$  is an adequate positive function.

If  $\xi \neq 0$ , sometimes it is more convenient to consider the following conditions instead, which are equivalent to (2.1):

- a) If  $\xi > 0$ , we can choose  $a(t) = U(t)$ , in (2.1), and then  $F \in \mathcal{D}_M(G_{\xi > 0})$  if, and only if,  $\lim_{t \rightarrow \infty} \frac{U(tx)}{U(t)} = x^\xi$  for  $x > 0$ ;
- b) If  $\xi < 0$ ,  $U(\infty) < \infty$  and  $\lim_{t \rightarrow \infty} \frac{U(\infty) - U(t)}{a(t)} = -1/\xi$ : Then  $F \in \mathcal{D}_M(G_{\xi < 0})$  if, and only if,  $\lim_{t \rightarrow \infty} \frac{U(\infty) - U(tx)}{U(\infty) - U(t)} = x^\xi$ , for  $x > 0$ .

**Remark 2.2.** In spite of the close connection between max-geometric-stable and max-stable distributions, the former class does have their own set of characterizations for domains of attraction. However,  $H_\xi \in \mathcal{D}_M(G_\xi)$ . Indeed, and with the notation  $\bar{F} = 1 - F$  for the right tail function,  $\bar{H}_\xi = -\ln G_\xi / (1 - \ln G_\xi) \sim \bar{G}_\xi / (1 - \ln G_\xi)$ .

---

## 2.2. EV distributions in generalized Verhulst models

---

Let us consider again the Blumberg hiperlogistic equation

$$(2.2) \quad \frac{d}{dt} N(t) = \rho (N(t))^\alpha (1 - N(t))^\beta, \quad \alpha, \beta > 0.$$

Henceforth, we shall assume that  $K = 1$  in all differential equations in order to get a normalized solution, *i.e.* a distribution function  $N$ .

If  $\alpha \notin \mathbb{N}$ , the solution of (2.2) satisfies the equation

$$(2.3) \quad \frac{(N(t))^{1-\alpha}}{1-\alpha} {}_2F_1(1-\alpha, \beta; 2-\alpha; N(t)) = \rho t + C,$$

where  ${}_2F_1(a, b; c; z) = \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{(c)_n} \frac{z^n}{n!}$ , with  $(x)_n = x(x+1) \cdots (x+n-1)$ , is the hypergeometric function and  $C$  is a real number. Without loss of generality, we can assume that  $C = 0$ .

Since  ${}_2F_1(a, b; b; z) = (1 - z)^{-a}$ , it follows that if  $\alpha + \beta = 2$  ( $\beta = 2 - \alpha$ ), we get the closed form analytical solution

$$N(t) = \frac{1}{1 + ((1 - \alpha)\rho t)^{-1/(1-\alpha)}} = \frac{1}{1 + \left(1 + (1 - \alpha)\left(\rho t - \frac{1}{1-\alpha}\right)\right)^{-1/(1-\alpha)}}$$

for (2.2), which belongs to the max-geometric-stable family of distributions, with an EVI  $\xi = 1 - \alpha = \beta - 1$ . Consequently,  $N \in \mathcal{D}_{\mathcal{M}}(G_{\xi=1-\alpha=\beta-1})$  (see **Remark 2.2**).

**Remark 2.3.** Observe that the Verhulst logistic equation (1.2), which is just the Blumberg hiperlogistic equation in (2.2) for  $\alpha = \beta = 1$ , satisfies the condition  $\alpha + \beta = 2$ , assumed in Theorem 2.1, with  $\xi = 1 - \alpha = \beta - 1 = 0$ .

We shall next generalize the result above on the basis of the reciprocal tail quantile function associated with the solution that comes out of (2.3) when  $\alpha \notin \mathbb{N}$ , which is given by

$$U(t) = N^{\leftarrow}\left(1 - \frac{1}{t}\right) = \frac{1}{\rho} \left( \frac{1}{1 - \alpha} \left(1 - \frac{1}{t}\right)^{1-\alpha} {}_2F_1\left(1 - \alpha, \beta; 2 - \alpha; 1 - \frac{1}{t}\right) - C \right).$$

Hence, if  $\beta < 1$ , we have  $U(\infty) < \infty$  and if  $\beta \geq 1$ ,  $U(\infty) = \infty$ . These results follow from the properties of the hypergeometric function, namely  ${}_2F_1(a, b; c; 1) < \infty$  if  $a + b - c < 0$  and  ${}_2F_1(a, b; c; 1) = \infty$  if  $a + b - c \geq 0$ .

We first state:

**Theorem 2.1.** *If  $\alpha \notin \mathbb{N}$  in the Blumberg hiperlogistic equation (2.2), then  $N \in \mathcal{D}_{\mathcal{M}}(G_{\xi=\beta-1})$ .*

**Proof:** a) For  $\beta < 1$ ,

$$\begin{aligned} \lim_{t \rightarrow \infty} \frac{U(\infty) - U(tx)}{U(\infty) - U(t)} &= \\ &= \lim_{t \rightarrow \infty} \frac{{}_2F_1\left(1 - \alpha, \beta; 2 - \alpha; 1\right) - \left(1 - \frac{1}{tx}\right)^{1-\alpha} {}_2F_1\left(1 - \alpha, \beta; 2 - \alpha; 1 - \frac{1}{tx}\right)}{{}_2F_1\left(1 - \alpha, \beta; 2 - \alpha; 1\right) - \left(1 - \frac{1}{t}\right)^{1-\alpha} {}_2F_1\left(1 - \alpha, \beta; 2 - \alpha; 1 - \frac{1}{t}\right)} \\ &= x^{\beta-1} \lim_{t \rightarrow \infty} \left( \frac{1 - \frac{1}{tx}}{1 - \frac{1}{t}} \right)^{-\alpha} = x^{\beta-1}. \end{aligned}$$

Therefore,  $N \in \mathcal{D}_{\mathcal{M}}(G_{\xi=\beta-1 < 0})$ . To obtain the limit above we took into consideration the fact that

$$\frac{\partial}{\partial t} \left(1 - \frac{1}{t}\right)^{1-\alpha} {}_2F_1\left(1 - \alpha, \beta; 2 - \alpha; 1 - \frac{1}{t}\right) = (1 - \alpha) \left(1 - \frac{1}{t}\right)^{-\alpha} t^{\beta-2}.$$



b) If  $\beta > 1$ ,

$$\begin{aligned} \lim_{t \rightarrow \infty} \frac{U(tx)}{U(t)} &= \lim_{t \rightarrow \infty} \frac{\frac{1}{1-\alpha} \left(1 - \frac{1}{tx}\right)^{1-\alpha} {}_2F_1\left(1-\alpha, \beta; 2-\alpha; 1 - \frac{1}{tx}\right) - C}{\frac{1}{1-\alpha} \left(1 - \frac{1}{t}\right)^{1-\alpha} {}_2F_1\left(1-\alpha, \beta; 2-\alpha; 1 - \frac{1}{t}\right) - C} \\ &= x^{\beta-1} \lim_{t \rightarrow \infty} \left(\frac{1 - \frac{1}{tx}}{1 - \frac{1}{t}}\right)^{-\alpha} = x^{\beta-1}. \end{aligned}$$

Consequently,  $N \in \mathcal{D}_{\mathcal{M}}(G_{\xi=\beta-1>0})$ .

c) If  $\beta = 1$ ,

$$\begin{aligned} \lim_{t \rightarrow \infty} (U(tx) - U(t)) &= \frac{1}{(1-\alpha)\rho} \left[ \left(1 - \frac{1}{tx}\right)^{1-\alpha} {}_2F_1\left(1-\alpha, 1; 2-\alpha, 1 - \frac{1}{tx}\right) - \right. \\ &\quad \left. - \left(1 - \frac{1}{t}\right)^{1-\alpha} {}_2F_1\left(1-\alpha, 1; 2-\alpha, 1 - \frac{1}{t}\right) \right] \\ &= \frac{\ln x}{\rho}. \end{aligned}$$

Thus, if we consider  $a(t) = 1/\rho > 0$ , we have

$$\lim_{t \rightarrow \infty} \frac{U(tx) - U(t)}{a(t)} = \ln x,$$

which means that  $N \in \mathcal{D}_{\mathcal{M}}(G_{\xi=\beta-1=0})$ .

**Note 2.1.** The previous limit was obtained with the help of the software Mathematica, since there are series expansions involved and the use of relations between contiguous hypergeometric functions.

□

We next state:

**Theorem 2.2.** If  $\alpha, \beta \in \mathbb{N}$  in the Blumberg hiperlogistic equation (2.2), then we also get  $N \in \mathcal{D}_{\mathcal{M}}(G_{\xi=\beta-1})$ .

**Proof:** a) If  $\alpha = n = 2, 3, \dots$  and  $\beta = 1$ , the solution of (2.2) satisfies now the equation

$$\sum_{k=2}^n \frac{1}{1-k} \frac{1}{(N(t))^{k-1}} + \ln\left(\frac{N(t)}{1-N(t)}\right) = \rho t + C.$$

Hence, the reciprocal tail quantile function associated with the solution is, in this case,

$$U(t) = \frac{1}{\rho} \left( \sum_{k=2}^n \frac{1}{1-k} \frac{1}{\left(1 - \frac{1}{t}\right)^{k-1}} + \ln(t-1) - C \right),$$

with  $U(\infty) = \infty$ . It is quite straightforward to prove that

$$\lim_{t \rightarrow \infty} \frac{U(tx) - U(t)}{1/\rho} = \ln x,$$

from which follows that  $N \in \mathcal{D}_{\mathcal{M}}(G_{\xi=\beta-1=0})$ .

- b) When  $\alpha = 1$  and  $\beta = m = 2, 3, \dots$ , we have a solution satisfying the equation

$$\sum_{k=2}^m \frac{1}{k-1} \frac{1}{(1-N(t))^{k-1}} + \ln \left( \frac{N(t)}{1-N(t)} \right) = \rho t + C,$$

which in turn yields the reciprocal tail quantile function

$$U(t) = \frac{1}{\rho} \left( \sum_{k=2}^m \frac{1}{k-1} t^{k-1} + \ln(t-1) - C \right),$$

with  $U(\infty) = \infty$ . It is also quite straightforward to prove that

$$\lim_{t \rightarrow \infty} \frac{U(tx)}{U(t)} = x^{m-1},$$

which means that  $N \in \mathcal{D}_{\mathcal{M}}(G_{\xi=\beta-1>0})$ .

- c) For the more general case  $\alpha = n = 2, 3, \dots$  and  $\beta = m = 2, 3, \dots$ , we get a solution that verifies the equation

$$\sum_{k=2}^n \frac{a_k}{1-k} \frac{1}{(N(t))^{k-1}} + \sum_{j=2}^m \frac{b_j}{j-1} \frac{1}{(1-N(t))^{j-1}} + A \ln \left( \frac{N(t)}{1-N(t)} \right) = \rho t + C,$$

where the  $a_k$  and  $b_j$ 's are real numbers and  $A > 0$ . The reciprocal tail quantile function is now defined by

$$U(t) = \frac{1}{\rho} \left( \sum_{k=2}^n \frac{a_k}{1-k} \frac{1}{(1-\frac{1}{t})^{k-1}} + \sum_{j=2}^m \frac{b_j}{j-1} t^{j-1} + A \ln(t-1) - C \right),$$

with  $U(\infty) = \infty$ . It easily follows that

$$\lim_{t \rightarrow \infty} \frac{U(tx)}{U(t)} = x^{m-1},$$

which means, once again, that  $N \in \mathcal{D}_{\mathcal{M}}(G_{\xi=\beta-1>0})$ .

□

**Remark 2.4.** All previous results lead us to conjecture that for all  $\alpha, \beta > 0$ , the solution of equation (2.2) will be in the max-domain of attraction of a GEV distribution with an EVI  $\xi = \beta - 1$ , where  $\beta$  is the retroaction parameter. However, the case  $\alpha = 2, 3, \dots$  and  $\beta \notin \mathbb{N}$  is still left to be proved.

Unfortunately, we cannot use equation (2.3) because the hypergeometric function diverges for the parameters involved. So far we were not able to obtain a general equation as the ones obtained for different scenarios of  $\alpha, \beta \in \mathbb{N}$ . Nevertheless, the results above, and all particular cases we have tried, indicate that we should have a solution  $N \in \mathcal{D}_{\mathcal{M}}(G_{\xi=\beta-1})$ . This seems very likely, since it holds true for  $\beta \in \mathbb{N}$  and there is no apparent reason why it should not also hold for  $\beta \notin \mathbb{N}$ .

For example, if  $\alpha = 2$  and  $\beta = 1/2$ , the solution of (2.2) satisfies the equation

$$-\frac{\sqrt{1-N(t)}}{N(t)} - \operatorname{arctanh}\left(\sqrt{1-N(t)}\right) = \rho t + C,$$

which yields the reciprocal tail quantile function

$$U(t) = -\frac{1}{\rho} \left( \frac{\sqrt{\frac{1}{t}}}{1-\frac{1}{t}} + \operatorname{arctanh}\left(\sqrt{\frac{1}{t}}\right) + C \right),$$

with  $U(\infty) < \infty$ . Given that

$$\lim_{t \rightarrow \infty} \frac{U(\infty) - U(tx)}{U(\infty) - U(t)} = x^{-1/2} = x^{1/2-1},$$

we have  $N \in \mathcal{D}_{\mathcal{M}}(G_{\xi=-1/2})$ . On the other hand, if, for instance,  $\alpha = 3$  and  $\beta = 3/2$ , the solution of (2.2) satisfies now the equation

$$\frac{-2 - 5N(t) + 15(N(t))^2}{4(N(t))^2\sqrt{1-N(t)}} - \frac{15}{4} \operatorname{arctanh}\left(\sqrt{1-N(t)}\right) = \rho t + C,$$

from which the reciprocal tail quantile function is

$$U(t) = \frac{1}{\rho} \left( \frac{-2 - 5\left(1-\frac{1}{t}\right) + 15\left(1-\frac{1}{t}\right)^2}{4\left(1-\frac{1}{t}\right)^2\sqrt{\frac{1}{t}}} - \frac{15}{4} \operatorname{arctanh}\left(\sqrt{\frac{1}{t}}\right) - C \right),$$

with  $U(\infty) = \infty$ . Since

$$\lim_{t \rightarrow \infty} \frac{U(tx)}{U(t)} = x^{1/2} = x^{3/2-1},$$

we conclude that  $N \in \mathcal{D}_{\mathcal{M}}(G_{\xi=1/2})$ .

**Remark 2.5.** In Blumberg's hiperlogistic equation (2.2) we are not considering the possibility of an absent growth or retroaction factor, *i.e.*  $\alpha = 0$  or  $\beta = 0$ . For example, if we assume that  $\alpha = 0$  in (2.2), it is interesting to see that the solution is (for  $C = 0$ )

$$N(t) = 1 - ((\beta - 1)\rho t)^{-1/(\beta-1)} = 1 - \left( 1 + (\beta - 1) \left( \rho t - \frac{1}{\beta - 1} \right) \right)^{-1/(\beta-1)},$$

which is a member of the generalized Pareto (GP) family of distributions with shape parameter  $\beta - 1$ .

The GP family of distributions has the functional form

$$(2.4) \quad F_\xi(x; \lambda, \delta) = 1 - (1 + \xi(x - \lambda)/\delta)^{-1/\xi}, \quad 1 + \xi(x - \lambda)/\delta > 0, \quad x > \lambda,$$

with  $(\xi, \lambda, \delta) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$ . Once more,  $\xi$ ,  $\lambda$  and  $\delta$  are shape, location and scale parameters, respectively. The GP family defined in (2.4) combines into a single family three families of distributions, namely the exponential family, which is the limiting case of (2.4) as  $\xi \rightarrow 0$ , the classical Pareto family ( $\xi > 0$ ) and the so-called Pareto type II family ( $\xi < 0$ ). Note that the uniform distribution, which is the solution of equation (2.2) for  $\alpha = \beta = 0$ , is a GP distribution when  $\xi = -1$ .

In EVT, GP distributions also play an important role, more precisely, in modeling peaks over high thresholds. In fact, if  $X$  is a random variable with distribution function  $F$ , GP distributions arise as the limiting distribution for the distribution of conditional excesses  $X - u | X > u$ , as  $u \rightarrow x^F$ , where  $x^F = \sup\{x : F(x) < 1\}$  is the right endpoint of the underlying model  $F$ .

In a population dynamics context what matters to know is that  $F_\xi \in \mathcal{D}_M(G_\xi)$ . Therefore, when dealing with the case  $\alpha = 0$  in equation (2.2), we have a solution  $N \in \mathcal{D}_M(G_{\xi=\beta-1})$ . Note also that if  $\beta = 0$  in (2.2), the solution is now

$$N(t) = ((\alpha - 1)(-\rho t))^{-1/(\alpha-1)},$$

which is of the type  $1 - F_\xi(-x, \lambda, \delta)$ , with  $F_\xi$  defined in (2.4), and reminding us of the relation between GEV distributions for minima and for maxima, namely  $G_\xi^*(x; \lambda, \delta) = 1 - G_\xi(-x; \lambda, \delta)$ . In particular, if  $\alpha = 1$ , we get as solution  $N(t) = \exp(\rho t)$ , i.e. an exponential growth.

Let us next consider the differential equation

$$(2.5) \quad \frac{d}{dt}N(t) = \rho (N(t))^\alpha (-\ln N(t))^\beta, \quad \alpha, \beta > 0,$$

which generalizes equation (1.8) considered in Brillhante *et al.* [4]. We have now the validity of the following:

**Theorem 2.3.** *If  $N$  is the solution of the differential equation (2.5), then  $N \in \mathcal{D}_M(G_{\xi=\beta-1})$ .*

**Proof:** If  $\alpha = 1$  (and  $\beta > 0$ ), we get the closed form analytical solution (for  $C = 0$ ),

$$N(t) = \exp\left(-((\beta-1)\rho t)^{-1/(\beta-1)}\right) = \exp\left(-\left(1 + (\beta-1)\left(\rho t - \frac{1}{\beta-1}\right)\right)^{-1/(\beta-1)}\right),$$

which is a GEV distribution for maxima with an EVI  $\xi = \beta - 1$ , i.e.  $N \in \mathcal{D}_M(G_{\xi=\beta-1})$ .

On the other hand, if  $\alpha \neq 1$  and:

a)  $\beta < 1$ , the solution satisfies the equation

$$(2.6) \quad (1 - \alpha)^{\beta-1} \Gamma(1 - \beta, (\alpha - 1) \ln N(t)) = \rho t + C,$$

where  $\Gamma(a, z) = \int_z^\infty t^{a-1} e^{-t} dt$ ,  $a > 0$ , is the incomplete gamma function, or equivalently, the solution satisfies the equation

$$(2.7) \quad -(1 - \alpha)^{\beta-1} \gamma(1 - \beta, (\alpha - 1) \ln N(t)) = \rho t + C,$$

where  $\gamma(a, z) = \int_0^z t^{a-1} e^{-t} dt = \Gamma(a) - \Gamma(a, z)$  is another type of incomplete gamma function and  $\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$ ,  $a > 0$ , is the (complete) gamma function.

The reciprocal tail quantile function associated with (2.6) is

$$U(t) = \frac{1}{\rho} \left( (1 - \alpha)^{\beta-1} \Gamma\left(1 - \beta, (\alpha - 1) \ln\left(1 - \frac{1}{t}\right)\right) - C \right),$$

with  $U(\infty) < \infty$ . In this case,

$$\lim_{t \rightarrow \infty} \frac{U(\infty) - U(tx)}{U(\infty) - U(t)} = \frac{1}{x} \lim_{t \rightarrow \infty} \left( \frac{\ln\left(1 - \frac{1}{tx}\right)}{\ln\left(1 - \frac{1}{t}\right)} \right)^{-\beta} \left( \frac{t - \frac{1}{x}}{t - 1} \right)^{-\alpha} = x^{\beta-1},$$

since

$$\frac{\partial}{\partial t} \Gamma\left(1 - \beta, (\alpha - 1) \ln\left(1 - \frac{1}{t}\right)\right) = -\frac{(\alpha - 1)^{1-\beta} \left(1 - \frac{1}{t}\right)^{-\alpha} \left(\ln\left(1 - \frac{1}{t}\right)\right)^{-\beta}}{t^2}.$$

Therefore, we have  $N \in \mathcal{D}_{\mathcal{M}}(G_{\xi=\beta-1<0})$ .

b)  $\beta > 1$ , the solution satisfies now the equation

$$\frac{(-\ln N(t))^{1-\beta}}{\beta - 1} {}_1F_1(1 - \beta, 2 - \beta; (1 - \alpha) \ln N(t)) = \rho t + C,$$

where  ${}_1F_1(a, b; z) = \sum_{n=0}^\infty \frac{(a)_n z^n}{(b)_n n!}$  is the confluent hypergeometric function. In this case we are using equation (2.7) because if  $a < 0$ ,  $\gamma(a, z) = \frac{z^a}{a} {}_1F_1(a, a + 1; -z)$ .

The reciprocal tail quantile function is

$$U(t) = \frac{1}{\rho} \left( \frac{(-\ln\left(1 - \frac{1}{t}\right))^{1-\beta}}{\beta - 1} {}_1F_1\left(1 - \beta, 2 - \beta; (1 - \alpha) \ln\left(1 - \frac{1}{t}\right)\right) - C \right),$$

with  $U(\infty) = \infty$ , since we have  ${}_1F_1(a, b; 0) = 1$ . Therefore, without loss of generality, if  $C = 0$ ,

$$\begin{aligned} \lim_{t \rightarrow \infty} \frac{U(tx)}{U(t)} &= \lim_{t \rightarrow \infty} \left( \frac{\ln\left(1 - \frac{1}{t}\right)}{\ln\left(1 - \frac{1}{tx}\right)} \right)^{\beta-1} \frac{{}_1F_1\left(1 - \beta, 2 - \beta; (1 - \alpha) \ln\left(1 - \frac{1}{tx}\right)\right)}{{}_1F_1\left(1 - \beta, 2 - \beta; (1 - \alpha) \ln\left(1 - \frac{1}{t}\right)\right)} \\ &= x^{\beta-1}. \end{aligned}$$

Hence,  $N \in \mathcal{D}_{\mathcal{M}}(G_{\xi=\beta-1>0})$ .

c)  $\beta = 1$ , the solution satisfies the equation

$$-\text{Ei}((1 - \alpha) \ln N(t)) = \rho t + C,$$

where  $\text{Ei}(x) = -\int_{-x}^{\infty} \frac{e^{-t}}{t} dt$ ,  $x > 0$ , is the exponential integral function. Thus, the reciprocal tail quantile function is

$$U(t) = -\frac{1}{\rho} \left( \text{Ei}((1 - \alpha) \ln N(t)) + C \right),$$

with  $U(\infty) = \infty$ . Since the exponential integral function series expansion is

$$\text{Ei}(x) = \gamma + \ln x + \sum_{n=1}^{\infty} \frac{x^n}{n!n},$$

where  $\gamma = 0.57721\dots$  is Euler's constant (cf. Abramowitz and Stegun [1]), it follows that

$$\lim_{t \rightarrow \infty} (U(tx) - U(t)) = \frac{1}{\rho} \lim_{t \rightarrow \infty} \ln \left( \frac{\ln(1 - \frac{1}{t})}{\ln(1 - \frac{1}{tx})} \right) = \frac{\ln x}{\rho}.$$

Therefore,

$$\lim_{t \rightarrow \infty} \frac{U(tx) - U(t)}{1/\rho} = \ln x,$$

meaning that  $N \in \mathcal{D}_{\mathcal{M}}(G_{\xi=\beta-1=0})$ .

□

As proved above, the retroaction parameter  $\beta$  of the generalized Verhulst equation (2.5) is the only parameter that establishes which GEV distribution for maxima is adequate to model the risk of extreme events in population dynamics, with the EVI being equal to  $\beta - 1$ . We saw earlier that, for a large variety of situations, this also happens to be the case when using equation (2.2). Now, this might seem at first sight a bit strange, in the sense that the growth parameter  $\alpha$  has no involvement whatsoever in establishing the limit distribution. However, this apparent “abnormality” can be explained by noticing that we are working with normalized equations, and therefore getting normalized solutions, meaning that  $N(t) \in (0, 1)$ . In light of this, we have  $(1 - N(t))^\beta \rightarrow 0$  as  $\beta \rightarrow \infty$ , which in this context is translated into a weaker control on population growth, and therefore the possibility of occurrence of more extreme events. This situation will also be mirrored in the case of working with the retroaction factor  $(-\ln N(t))^\beta$ .

**Remark 2.6.** It is also interesting to see that the solution of the sub-family of models defined in (1.5) for  $\nu > 0$ , i.e.

$$(2.8) \quad \frac{d}{dt} N(t) = \rho N(t) \frac{1 - (N(t))^\nu}{\nu},$$

and which contains the Verhulst logistic equation as a special case ( $\nu = 1$ ), belongs to the max-domain of attraction of a GEV distribution, with EVI  $\xi = 0$ . In fact, we already know that the solution for  $\nu = 1$  is a member of the logistic family of distributions, and which in turn belongs to the max-domain of attraction of a GEV distribution with EVI  $\xi = 0$ . But, more generally, the solution of (2.8) satisfies the equation

$$\nu \ln N(t) - \ln(1 - (N(t))^\nu) = \rho t + C,$$

and therefore the reciprocal tail quantile function associated is

$$U(t) = \frac{1}{\rho} \left[ \nu \ln \left( 1 - \frac{1}{t} \right) - \ln \left( 1 - \left( 1 - \frac{1}{t} \right)^\nu \right) - C \right],$$

with  $U(\infty) = \infty$ . Since

$$\lim_{t \rightarrow \infty} \frac{U(tx) - U(t)}{1/\rho} = \ln x,$$

it follows that  $N \in \mathcal{D}_{\mathcal{M}}(G_{\xi=0})$ .

---

### 3. COMMENTS AND FURTHER RESULTS

---

As mentioned in Section 1,  $N(t)$  is a linear approximation of  $-\ln(1 - N(t))$ , with  $N(t) \in (0, 1)$ . So a valid question is, what happens if  $N(t)$  is replaced by  $-\ln(1 - N(t))$  in (2.2)? In other words, what kind of solution do we get for the generalized Verhulst equation

$$(3.1) \quad \frac{d}{dt} N(t) = \rho \left[ -\ln(1 - N(t)) \right]^\alpha (1 - N(t))^\beta, \quad \alpha, \beta > 0?$$

What happens is that the roles between  $\beta$  and  $\alpha$  are switched, in the sense that now the growth parameter  $\alpha$  establishes, on its own, which EV distribution for minima, not for maxima, is at stake.

In fact, if  $\beta = 1$ , the solution of (3.1) is a GEV distribution for minima  $G_\xi^*$ , with  $\xi = \alpha - 1$ . As an immediate consequence of the close connection between maxima and minima, there are only three types of stable distributions for minima, namely the Fréchet for minima ( $\xi > 0$ ), the Weibull ( $\xi < 0$ ) and the Gumbel for minima ( $\xi = 0$ ). On the other hand, if  $\beta \neq 1$ , the solution of (3.1) will belong to the min-domain of attraction of a  $G_\xi^*$ , with  $\xi = \alpha - 1$ . Note that in this new setting we can still have uncontrolled population growth, although this growth will be somehow restricted to minimum levels, due to “lack of space” to accommodate more explosive population growths.

An interesting and open topic of research lies now on the estimation of  $\beta$  on the basis of the estimation of  $\xi$ , or the other way round, the estimation of

$\xi$  on the basis of the estimation of  $\beta$ . (For the estimation of  $\xi$ , see the recent overview on statistical EVT by Gomes and Guillou [8], among others.) In fact, an adequate estimation of the parameter  $\beta$  is fundamental so that the generalized Verhulst models considered here can be applicable to real data. This is the case, since we have established that the retroaction parameter  $\beta$  is the only parameter that determines which GEV distribution for maxima is appropriate to model the risk of extreme large events in population dynamics, with the EVI (for maxima) being equal to  $\beta - 1$ . A similar comment applies to the growth parameter  $\alpha$  and the modeling of extreme small events in population dynamics, with the EVI for minima being then equal to  $\alpha - 1$ . Tsoularis and Wallace [15] investigated how the inflection point of the population growth curve is related to its malthusian growth and retroaction parameters, and their results may be exploited in this context.

---

## ACKNOWLEDGMENTS

---

Research funded by FCT - Fundação para a Ciência e Tecnologia, Portugal, Projects UID/MAT/00006/2013 and UID/MAT/00006/2019.

---

## REFERENCES

---

- [1] ABRAMOWITZ, M. and STEGUN, I.A. (1970). *Handbook of Mathematical Functions*, Dover, New York.
- [2] BLUMBERG, A.A. (1968). Logistic growth rate functions, *Journal of Theoretical Biology*, **21**, 42–44.
- [3] BOX, G.E.P. and COX, D.R. (1964). An analysis of transformations, *Journal of the Royal Statistical Society. Series B (Methodological)*, **26**, 2, 211–252.
- [4] BRILHANTE, M.F., GOMES, M.I. and PESTANA, D. (2011). BetaBoop brings in chaos, *CMSim—Chaotic Modeling and Simulation Journal*, **1**, 39–50.
- [5] BRILHANTE, M.F., GOMES, M.I. and PESTANA, D. (2012). Extensions of Verhulst model in population dynamics and extremes, *CMSim—Chaotic Modeling and Simulation Journal*, **4**, 575–591.
- [6] DE HAAN, L. (1984). *Slow variation and characterization of domains of attraction*. In “Statistical Extremes and Applications” (J. Tiago de Oliveira, Ed.), D. Reidel, Dordrecht, 31–48.
- [7] GNEDENKO, B.V. (1943). Sur la distribution limite du terme maximum d’une série aléatoire, *Annals of Mathematics*, **44**, 6, 423–453.
- [8] GOMES, M.I. and GUILLOU, A. (2015). Extreme value theory and statistics of univariate extremes: A review, *International Statistical Review*, **83**, 2, 263–292.



- [9] LEADBETTER, M.R., LINDGREN, G. and ROOTZÉN, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*, Springer, New York.
- [10] LOKTA, A.J. (1956). *Elements of Mathematical Biology*, Dover, New York.
- [11] RACHEV, S.T. and RESNICK, S. (1991). Max-geometric infinite divisibility and stability, *Communications in Statistics. Stochastic Models*, **7**, 191–218.
- [12] RICHARDS, F.J. (1959). A flexible growth function for empirical use, *Journal of Experimental Botany*, **10**, 29, 290–300.
- [13] SMITH, F.E. (1963). Population dynamics in daphnia magna and a new model for population growth, *Ecology*, **44**, 4, 651–663.
- [14] TSOULARIS, A. (2001). Analysis of logistic growth models, *Research Letters in the Information and Mathematical Sciences*, **2**, 23–46.
- [15] TSOULARIS, A. and WALLACE, J. (2002). Analysis of logistic growth models, *Mathematical Biosciences*, **179**, 1, 21–55.
- [16] VERHULST, P.F. (1838). Notice sur la loi que la population poursuit dans son accroissement, *Correspondance Mathématique et Physique*, **10**, 113–121.

---

---

## 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 <sup>2</sup>	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	$\geq 180$	and/or	$\geq 110$
Isolated Systolic Hypertension	$\geq 140$	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  $Y_1, Y_2, \dots, Y_n$ , 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:

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

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

This distribution has support  $\{x \in \mathbb{R} : x > 0\}$  for  $k \geq 0$  and support  $\{x \in \mathbb{R} : 0 < x < -\frac{\sigma}{k}\}$  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  $GPD(k, \sigma)$ .

---

\* 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.

---

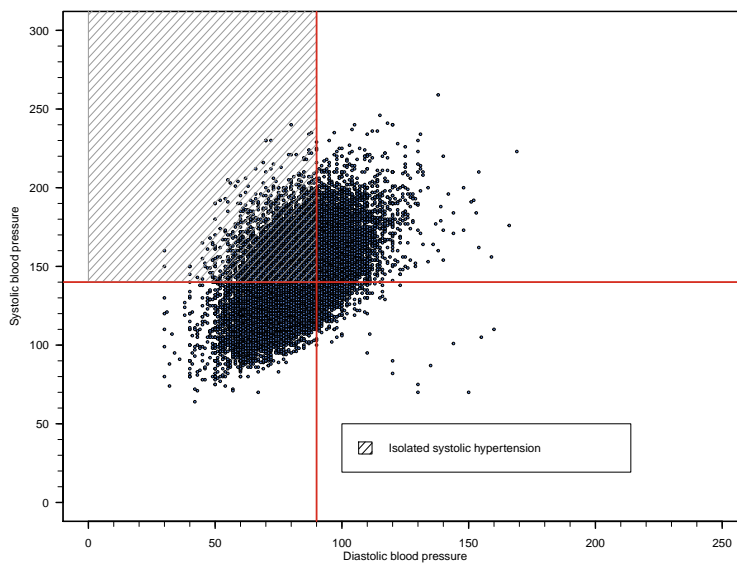
### 3. EXPLORATORY DATA ANALYSIS

---

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  $SBP < 140$ )

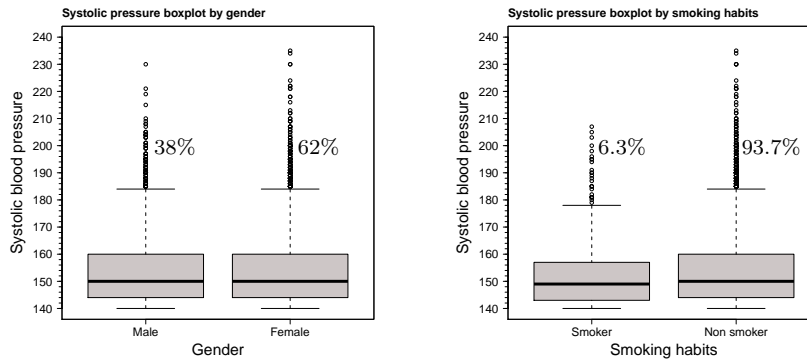
There are also 3380 individuals with omitted information about these variables.



**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).

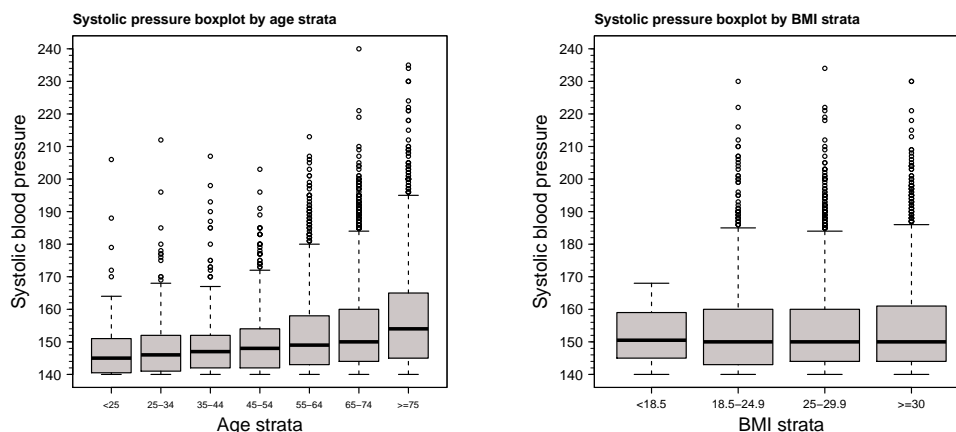
Figure 2 illustrates the SBP boxplots by gender and tobacco consumption. It can be seen that women seem to have more and higher extreme values of SBP than men. The boxplots produced for the smoking habits seems to indicate that those who smoke have, in the overall, lower values of SBP than those who 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
$\geq 75$	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

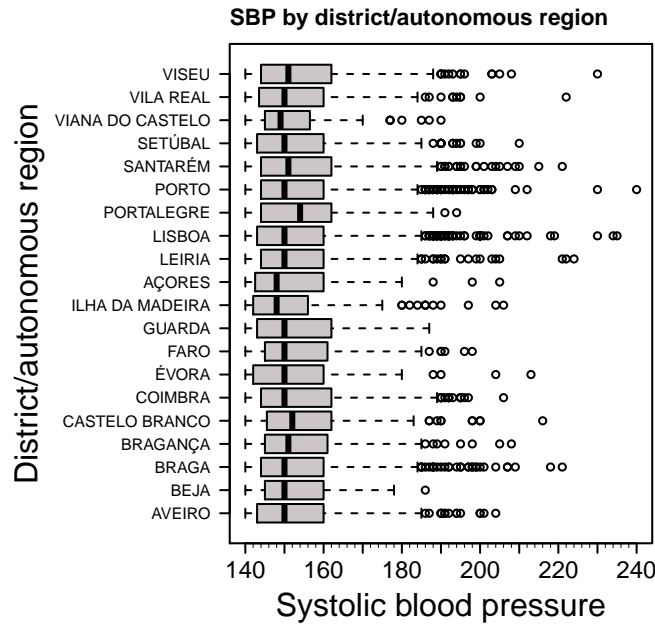
**Table 4:** BMI classes.

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

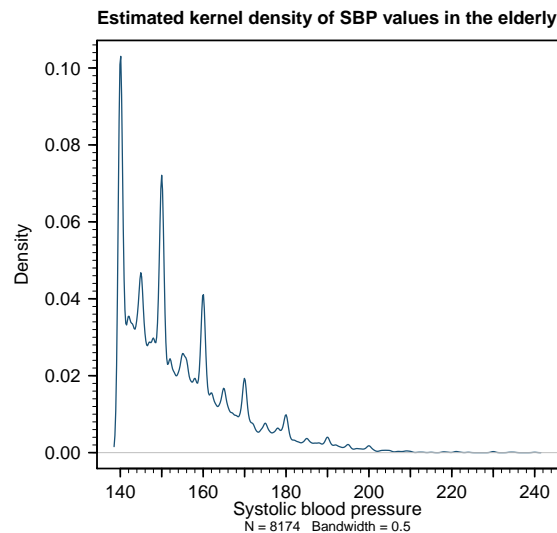
---

In this section we propose models for the extremes of systolic blood pressure of elderly individuals (age  $\geq 55$ ) 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 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*.

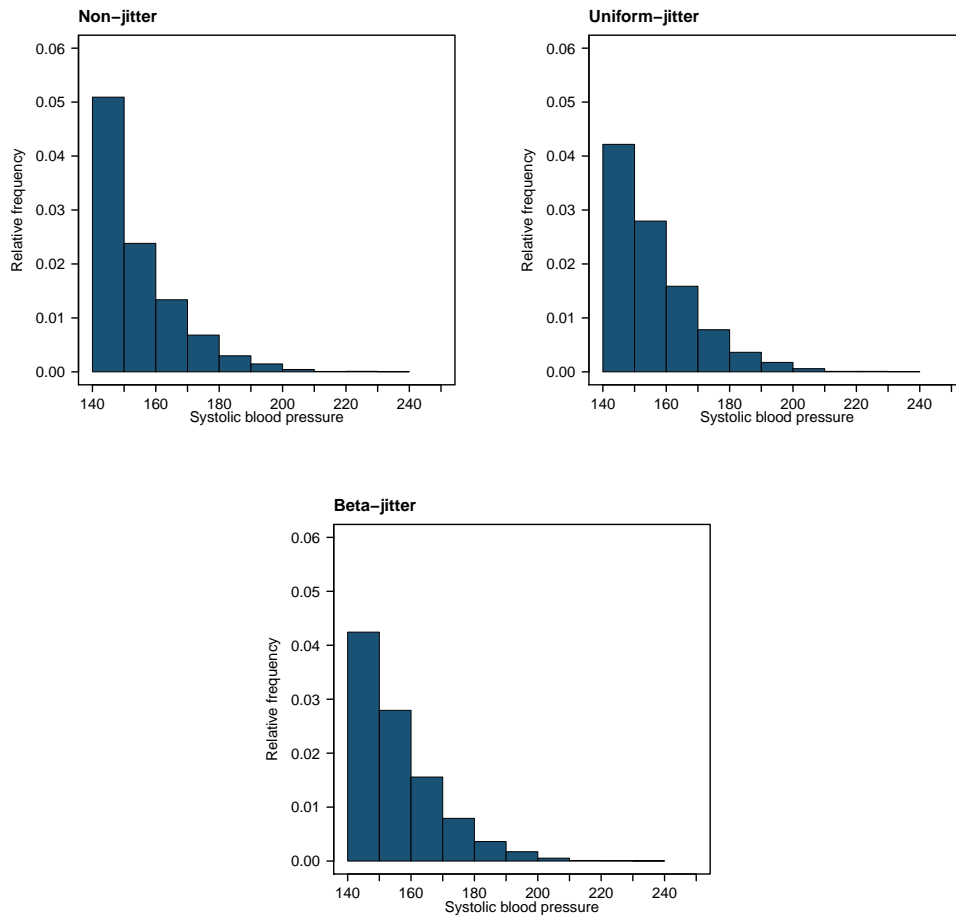
We then created two new data sets by adding each sample to the data. Note that by adding these simulated samples to the SBP values of the elderly individuals who suffer from ISH we got some values below 140, that were not considered in the subsequent analysis. Let's investigate how both jitters altered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

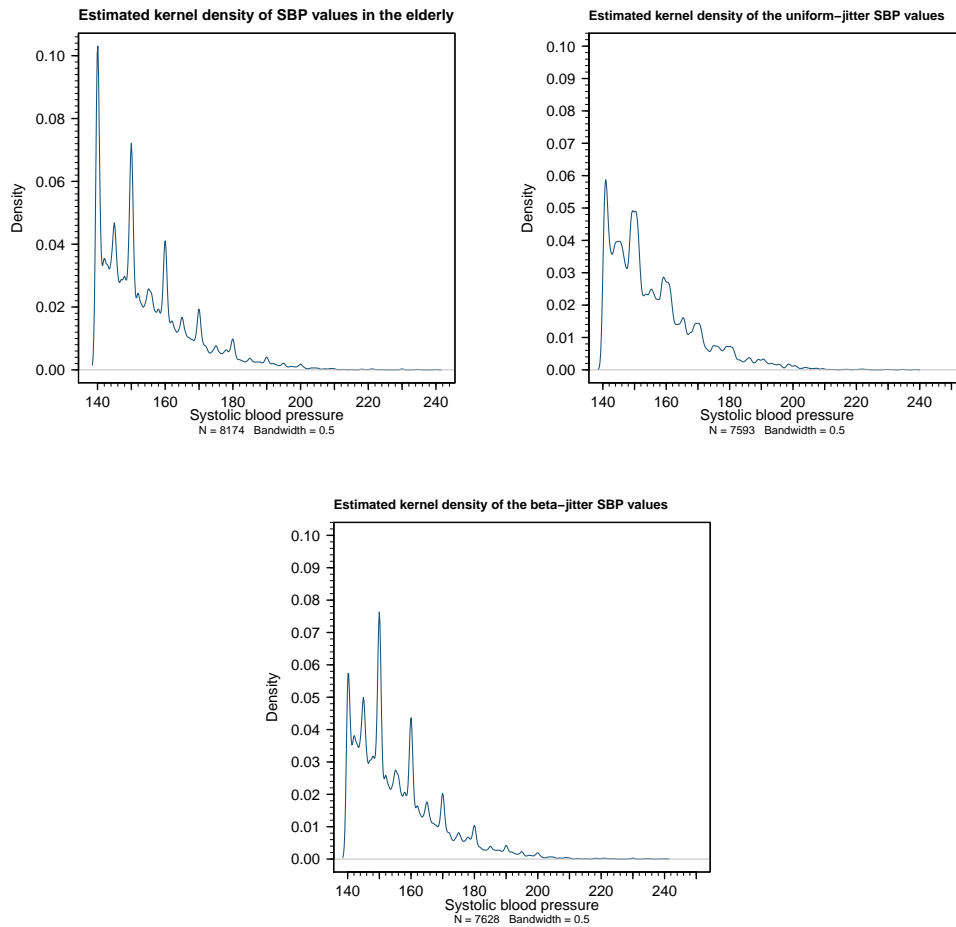
---

#### 4.1.1. Threshold selection analysis

---

We now start the procedure of selecting adequate thresholds for each case, with the goal of fitting  $GPD(k, \sigma)$  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.

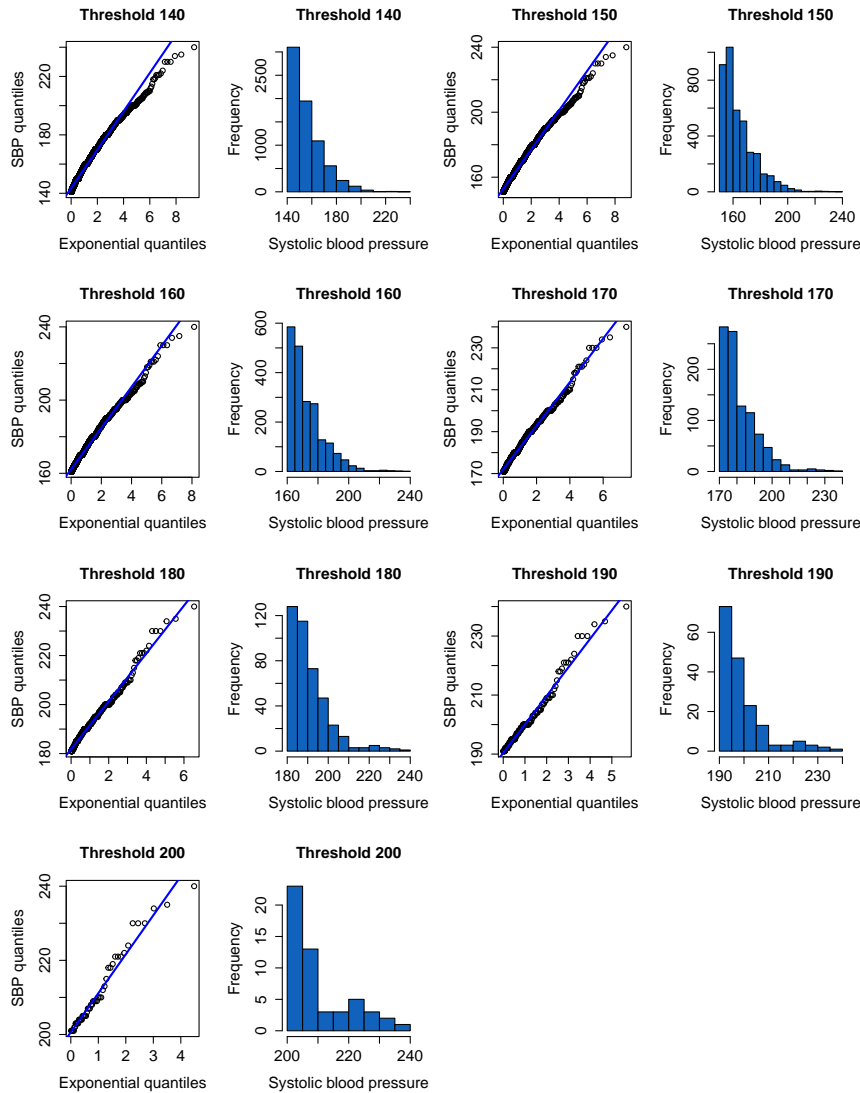
Next we present the results for the Bayesian threshold selection method using measures of surprise for each data set, see [24]. Figure 12 presents the



**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].

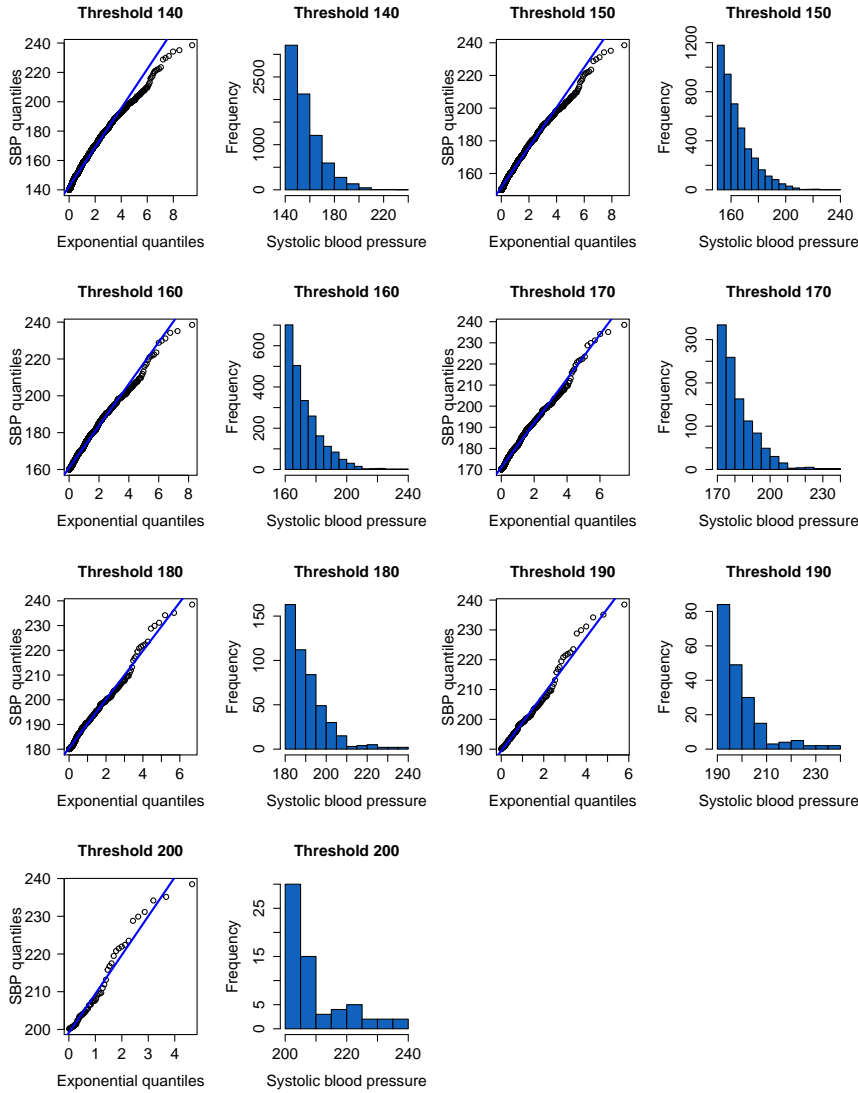
Figure 12 (top left) shows the method applied to the non-jitter data set. It only manifests less incompatibility with the GPD model for high threshold values, i.e.,  $190 < u < 200$  mmHg. Moreover, the  $p$ -values demonstrate a switch in surprise when more data is considered, i.e., when we introduce 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  $u_1, u_2, \dots, u_m$  be a sequence of increasingly ordered threshold candidates for a given data set, and consider that  $H_0^1, H_0^2, \dots, H_0^m$ , are  $m$  null test hypotheses such that for some  $1 \leq i \leq m$ , the  $i$ th null hypothesis is defined as  $H_0^i$ : the excesses over  $u_i$  come from a generalized Pareto distribution. Let  $p_1, p_2, \dots, p_m$  be the  $p$ -values obtained using the Cramér-Von Mises goodness-of-fit test for each sample. The ForwardStop rule is given

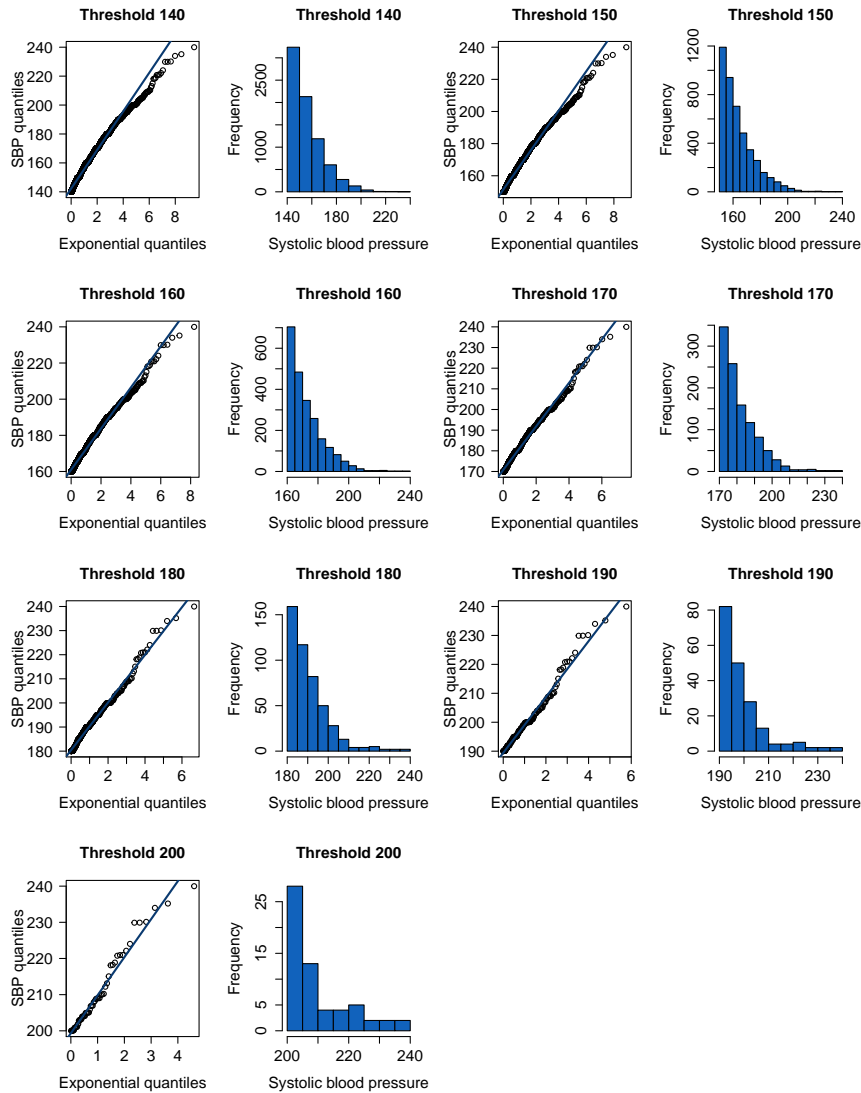
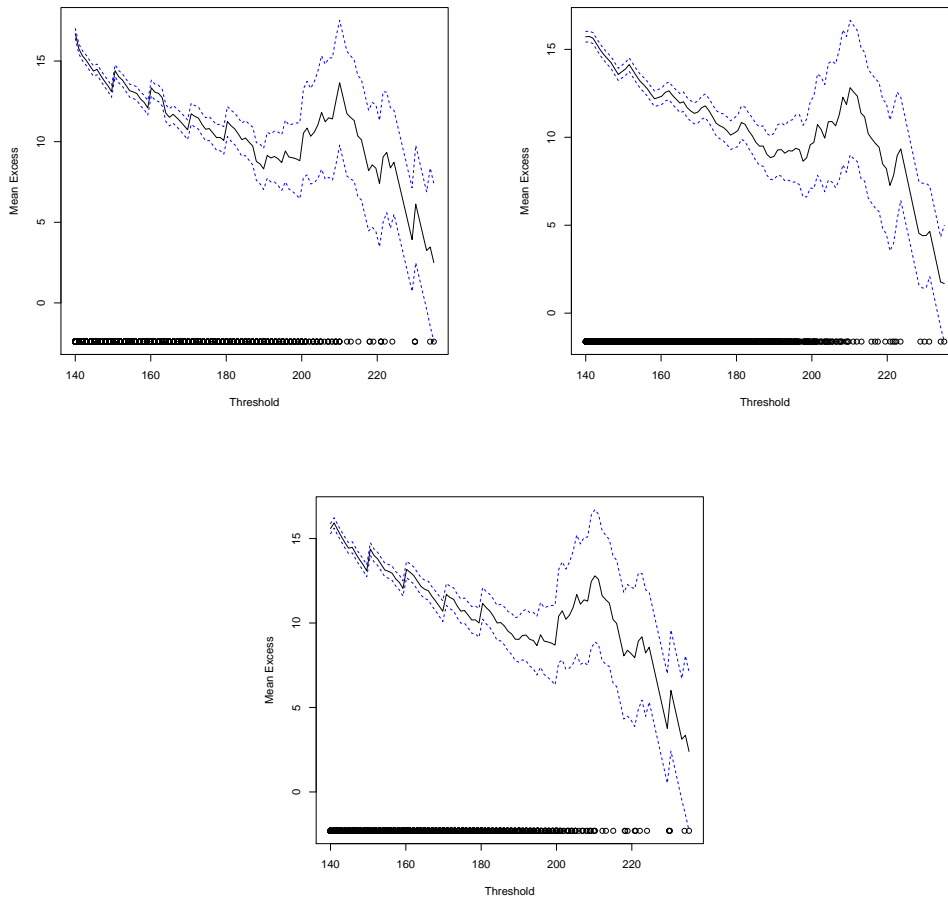


Figure 10: Exponential QQplots and histograms for each threshold candidate for the beta-jitter data.

by

$$(4.1) \quad \hat{i} = \max \left\{ i \in \{1, \dots, m\} : -\frac{1}{i} \sum_{j=1}^i \log(1 - p_j) \leq \alpha \right\},$$

where  $\alpha$  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  $\hat{i}$  is obtained we reject  $H_i$  for some  $i = 1, \dots, \hat{i}$  thereby not rejecting the null hypothesis at  $\hat{i} + 1$  and accepting the threshold associated with  $H_0^{\hat{i}+1}$ .

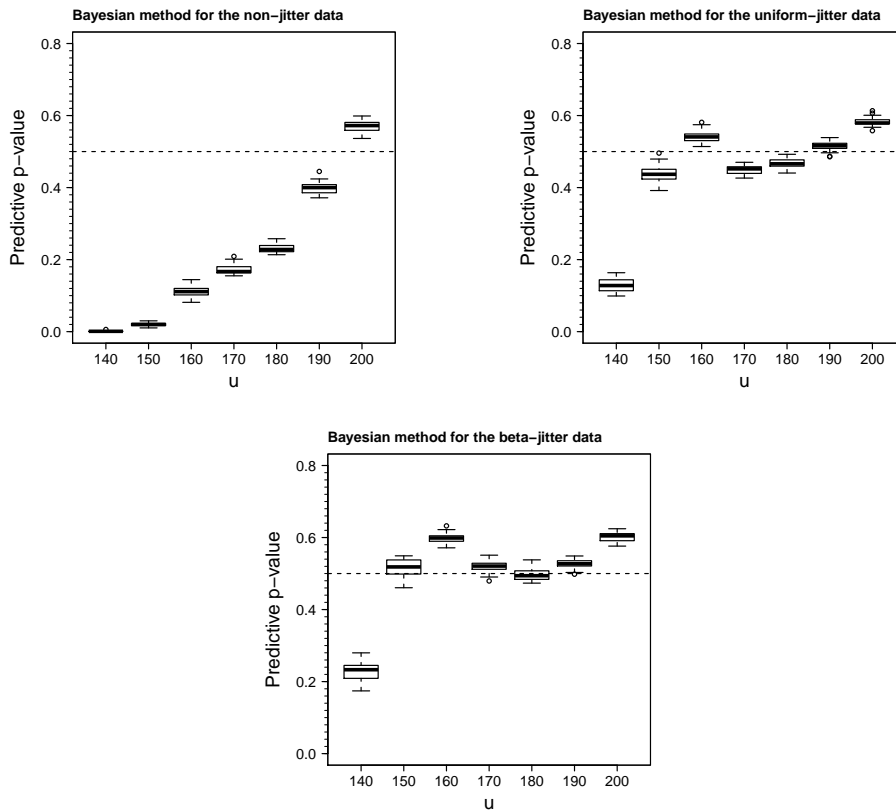


**Figure 11:** Mean residual life function for the non-jitter data (top left), uniform-jitter data (top right) and beta jitter data (bottom).

threshold	num.above	$p$ -values	fowardstop	statistic
140	7113	2.4221e-47	$\sim 0$	3.3111
150	4012	1.3233e-46	$\sim 0$	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  $\alpha=0.01$ , suggesting that 190 is an adequate threshold.

Finally,  $u = 190$  mmHg was the threshold selected for the three cases. The parameters,  $k$  and  $\sigma$  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 $k$	$\sigma$	95% CI for $\sigma$	$max$	$endpoint$	$-\log(L)$
$M_I$	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( $k, \sigma$ ) 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( $k, \sigma$ ) 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  $H_0$ ,

$$T = 2(l_{M_1}(x) - l_{M_2}(x)) \sim \chi_1^2,$$

where in this case  $l_{M_1}$  is the log-likelihood function for the GPD( $k, \sigma$ ) 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  $\alpha$  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, *Ageing Clinical and Experimental Research*, **16**, 6, 461–466.

---

---

## TESTING CONDITIONS AND ESTIMATING PARAMETERS IN EXTREME VALUE THEORY: APPLICATION TO ENVIRONMENTAL DATA

---

---

- Authors: HELENA PENALVA  
– Escola de Ciências Empresariais do Instituto Politécnico de Setúbal, and  
CEAUL, Universidade de Lisboa, Portugal  
`helena.penalva@esce.ips.pt`
- DORA PRATA GOMES  
– Faculdade de Ciências e Tecnologia and CMA/FCT,  
Universidade Nova de Lisboa, Portugal  
`dsrp@fct.unl.pt`
- M. MANUELA NEVES  
– Instituto Superior de Agronomia, and CEAUL,  
Universidade de Lisboa, Portugal  
`manela@isa.ulisboa.pt`
- SANDRA NUNES  
– Escola de Ciências Empresariais do Instituto Politécnico de Setúbal, and  
CMA/FCT, Universidade Nova de Lisboa, Portugal  
`sandra.nunes@esce.ips.pt`

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

Abstract:

- *Extreme Value Theory* has been asserting itself as one of the most important statistical theories for the applied sciences providing a solid theoretical basis for deriving statistical models describing extreme or even rare events. The efficiency of the inference and estimation procedures depends on the tail shape of the distribution underlying the data. In this work we will present a review of tests for assessing extreme value conditions and for the choice of the extreme value domain. Motivated by two real environmental problems we will apply those tests showing the need of performing such tests for choosing the most appropriate parameter estimation methods.

Key-Words:

- *Environmental data; extreme values; heavy-tailed distributions; semi-parametric estimation; statistical testing.*

AMS Subject Classification:

- 62G32, 62E20, 62G10.



---

## 1. MOTIVATION AND INTRODUCTION

---

Extreme Value Theory (EVT) is concerned with the behaviour of extreme values, i.e. values occurring at the tails of a probability distribution. Society, human life, etc. tend to adapt to near-normal conditions, and these conditions tend to produce fairly minimal impacts. In contrast, unusual and extreme conditions can have a substantial impact despite, by definition, occurring in a very low proportion of times. EVT is the branch of probability and statistics dedicated to characterizing the very low or quite high values of a variable, *the tail of the distribution*. EVT had its beginnings in the early to middle part of XX century and Emil Gumbel was the pioneer in applications of statistics of extremes. In *Statistics of Extremes* [23], he presents several applications of EVT on real world problems in engineering and in meteorological phenomena. In this book appear the first applications in hydrology.

Results in EVT rely on certain assumptions. However in some situations they can be not fulfilled. So, before dealing with an application, it is important to have an *a priori* knowledge on whether the underlying distribution verifies those assumptions. On the other hand statistical inference procedures should be performed according to the most adequate domain of attraction for the underlying distribution. So, tests for extreme value conditions and for the choice of the tail must be done before the application of any inferential procedure.

The motivation for this work came from a first study in Neves *et al.* [34] and Penalva *et al.* [36] presenting a review of tests and parameter estimation procedures applied to the daily mean flow discharge rate in the hydrometric station of Fragas da Torre in the river Paiva. The data were collected from 1946/47 to 2005/2006, i.e., 60 years of data. In Penalva *et al.* [36] we drew the attention for the need of a previous analysis for assessing extreme value conditions and for the choice of the extreme value domain, in order to choose the more adequate parameter estimators. We will review briefly the analysis already performed considering the data now available during 66 years, 1946/2012 and using, for comparison, two recent classes of estimators of the tail index of the extreme value distribution, introduced in Penalva *et al.* [37] and Gomes *et al.* [21].

The procedures proposed are also applied and commented to another data set referring to burned areas of wildfires in Portugal during 33 years (1984–2016).

So, the aim of this work is to perform an univariate extreme value analysis illustrating and reviewing tests on the extreme value condition and on the statistical choice of the tail of the underlying distribution. This should be the first step in order to choose the more adequate estimators. Some recent estimators of the tail index are also compared.

The paper proceeds as follows. Section 2 contains the main results that are

the basis of the theoretical background. In Section 3 the exploratory analysis of the first case-study aforementioned is performed, parametric and semi-parametric statistical approaches in EVT are briefly reviewed and first estimates of the main parameters are presented. In Section 4, statistical testing procedures for extreme value conditions and for choosing the tail are presented and applied to the data. Section 5 is dedicated to perform the study and estimation in a second case-study, showing the adequate procedure of performing the study. Finally Section 6 presents a first practical application on the effect of taking into consideration or not the choice of the tail of the underlying distribution and consequently the adequate EVI estimation. For the first case study, where estimation discrepancies were detected when the choice of the tail was made previously or not, high quantiles are estimated. A few comments on some other parameters that could be considered and the work in progress finish this section.

---

## 2. THEORETICAL BACKGROUND

---

Let us assume that we have a sample  $(X_1, \dots, X_n)$  of independent and identically distributed (iid) or possibly stationary, weakly dependent random variables from an unknown cumulative distribution function (cdf)  $F$ . Let us consider the notation  $(X_{1:n} \leq X_{2:n} \leq \dots \leq X_{n:n})$  for the sample of ascending order statistics associated to that sample.

The interest is focused on the distribution of the maxima, that is,  $M_n := \max(X_1, \dots, X_n)$ , for which we have

$$(2.1) \quad \begin{aligned} \mathbb{P}(M_n \leq x) &= \mathbb{P}(X_1 \leq x, \dots, X_n \leq x) \\ &= \mathbb{P}(X_1 \leq x) \dots \mathbb{P}(X_n \leq x) = F^n(x). \end{aligned}$$

We often deal with the maxima, given the “kind of symmetry”,  $\min(X_1, \dots, X_n) = -\max(-X_1, \dots, -X_n)$ .

This problem has similarities to that one of determining the distribution of  $S_n = \sum_{i=1}^n X_i$ . Obviously  $S_n$  and possibly  $M_n$  may tend to infinity, and their distribution is a degenerate one. The *central limit theorem* gives an answer to this problem under some conditions, showing that the normal distribution is obtained as the non-degenerate limit of  $S_n$  properly normalized by  $E[S_n]$  and  $\sqrt{Var[S_n]}$ .

As  $n$  goes to  $\infty$ , the distribution  $F^n$  in (2.1) has a trivial limit: 0, if  $F(x) < 1$  and 1, if  $F(x) = 1$ . So the idea for  $M_n$  was the same: first subtract a  $n$ -dependent constant, then rescale by a  $n$ -dependent factor. The first question is then whether one can find two sequences,  $\{a_n\} \in \mathbb{R}^+$  and  $\{b_n\} \in \mathbb{R}$  and a non-trivial distribution function,  $G$ , such that  $\lim_{n \rightarrow \infty} \mathbb{P}((M_n - b_n)/a_n \leq x) = G(x)$ .

First results on the  $G$  distribution are due to Fréchet [17], Fisher and Tippet [12], Gumbel [22] and von Mises [40]. But were Gnedenko [19] and de

Haan [24] who gave conditions for the existence of those sequences  $\{a_n\} \in \mathbb{R}^+$  and  $\{b_n\} \in \mathbb{R}$  such that when  $n \rightarrow \infty$  and  $\forall x \in \mathbb{R}$ ,

$$(2.2) \quad \lim_{n \rightarrow \infty} \mathbb{P} \left( \frac{M_n - b_n}{a_n} \leq x \right) = \lim_{n \rightarrow \infty} F^n(a_n x + b_n) = \text{EV}_\xi(x).$$

$\text{EV}_\xi$  is a nondegenerate distribution function, denoted as the Extreme Value cdf, given by

$$(2.3) \quad \text{EV}_\xi(x) = \begin{cases} \exp[-(1 + \xi x)^{-1/\xi}], & 1 + \xi x > 0 \text{ if } \xi \neq 0 \\ \exp[-\exp(-x)], & x \in \mathbb{R} \quad \text{if } \xi = 0. \end{cases}$$

When the above limit holds we say that  $F$  is in the domain of attraction (for maxima) of  $\text{EV}_\xi$  and write  $F \in \mathcal{D}_M(\text{EV}_\xi)$ .

The shape parameter  $\xi$ , in (2.3), is called the *extreme value index* (EVI) and it is the primary parameter of interest in EVT analysis. The  $\text{EV}_\xi$  incorporates the three (Fisher-Tippett) types: Gumbel, with  $\xi = 0$ , the right tail of  $F$  is of an exponential type; Fréchet with  $\xi > 0$ , the right tail is heavy, of a negative polynomial type, and  $F$  has an infinite right endpoint and Weibull with  $\xi < 0$ , the right tail is light, and  $F$  has a finite right endpoint ( $x^* < +\infty$ ).

These models can also incorporate location ( $\lambda$ ) and scale ( $\delta > 0$ ) parameters, and in this case, the EV cdf is given by,

$$(2.4) \quad \text{EV}_\xi(x; \lambda, \delta) \equiv \text{EV}_\xi((x - \lambda)/\delta).$$

We may then consider, when the sample size  $n \rightarrow \infty$ , the approximation

$$P[M_n \leq x] = F^n(x) \approx \text{EV}_\xi((x - b_n)/a_n).$$

---

### 3. FIRST CASE-STUDY – A REVIEW

---

The source of river Paiva is in the Serra de Leomil in the North of Portugal and it is a tributary of the river Douro, with a watershed area of approximately 700 Km. The discharge rate study of this river is a matter of major importance since it is one of the main alternatives to the river Douro as source of water supply in the south of Oporto region. The data are daily mean flow discharge rate values (m<sup>3</sup>/s) from 1 October, 1946 to 30 September, 2012 - collected from the “SNIRH: Sistema Nacional de Informação dos Recursos Hídricos”.

The descriptive study of these data revealed a tail heavier than that of the normal. Results in Table 1 are similar to those in Penalva *et al.* [36].

EVT has been developed under two frameworks. The first one is the parametric framework, that considers a class of models associated to the limiting

min	1st Qu.	Median	Mean	3rd Qu.	max
0.00	9.11	17.1	34.4	37.3	920.0
	n	Skewness	Kurtosis	St Dev	
	11946	4.14	27.13	50.26	

**Table 1:** Descriptive statistics for daily mean flow discharge rate values.

behaviour of the maxima, given in (2.2). The main assumption behind the parametric approach is that estimators are calculated considering the data following, approximately, an exact EV probability distribution function, defined by a number of parameters. In this approach several methodologies have been developed for estimating parameters: Block Maxima; Largest Observations; Peaks Over Threshold, to refer the most well known.

In the semi-parametric framework, the only assumption made is that the limit in (2.2) holds, i.e., that the underlying distribution verifies the extreme value condition. The EVI,  $\xi$ , that appears in (2.3), plays the central role in this framework. Under this approach several EVI-estimators have been developed. Some of the most relevant and also the most recent ones will be used here in the estimation.

As an illustration of parametric approaches to estimate EVT parameters, only the *Block Maxima* (BM) approach will be considered in this work. Other procedures can be seen in Penalva *et al.* [36].

---

### 3.1. The Block Maxima (BM) method

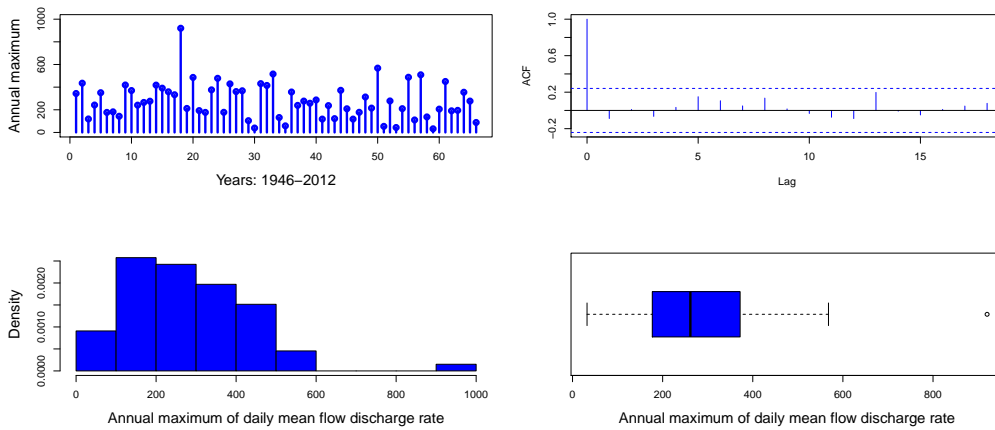
---

The so-called *Block Maxima* (BM), *Annual Maxima* or *Gumbel's* method is the first parametric approach for modelling extremes, Gumbel [23]. In this approach the  $n$ -sized sample is splitted into  $m$  sub-samples (usually  $m$  corresponds to the number of the observed years) of size  $l$  ( $n = m \times l$ ) for a sufficiently large  $l$ .  $EV_\xi$  or one of the models, Gumbel, Fréchet or Weibull, with unknown  $\xi \in \mathbb{R}$ ,  $\lambda \in \mathbb{R}$  or  $\delta \in \mathbb{R}^+$  are then fitted to the  $m$  maxima values of the  $m$  sub-samples.

Table 2 and Figure 1 show a very light positive asymmetry and kurtosis. It is also reasonable to consider data not correlated.

min	1st Qu.	Median	Mean	3rd Qu.	max
32.2	177.25	261.5	279.24	371.5	920.0
	m	Skewness	Kurtosis	St Dev	
	66	0.99	2.308	157.17	

**Table 2:** Basic descriptive statistics for the maximum values in each year.



**Figure 1:** Plots of the maximum value in each year, the partial autocorrelation function, the histogram and the boxplot.

Maximum likelihood estimates and standard errors were easily obtained using the `evd` package in  $\mathbb{R}$  software [38].

$\hat{\xi}$	$\hat{\lambda}$	$\hat{\delta}$
-0.03 (0.08)	207.74 (17.52)	127.11 (12.72)

**Table 3:** Maximum likelihood estimates and standard errors (in parenthesis).

---

### 3.2. Semi-parametric estimators

---

In this framework we do not need to fit a specific parametric model based on scale, shape and location parameters. We construct an EVI-estimator based on the largest  $k$  top observations, with  $k$  intermediate, i.e. such that  $k = k_n \rightarrow \infty$  and  $k/n \rightarrow 0$ , as  $n \rightarrow \infty$ , assuming only that the model  $F$  underlying the data is in  $\mathcal{D}_{\mathcal{M}}(\text{EV}_{\xi})$ , in specific sub-domains of  $\mathcal{D}_{\mathcal{M}}(\text{EV}_{\xi})$ , with  $\text{EV}_{\xi}$  provided in (2.3).

Most estimators show a strong dependence on that value  $k$ . They usually present: a small bias and a high variance for small values of  $k$ ; bias increases and variance decreases when  $k$  increases; the need of looking for an adequate value of  $k$  for which we have a minimum Mean Square Error. Thus, an intensive research has been performed trying to obtain estimators overcoming these difficulties. Currently there are several different EVI-estimators, so we decide to present and compare here a very few. Here we will illustrate the application of the following estimators: the classical Hill estimator, Hill [27], and a recent class of estimators, the *Lehmer mean-of-order- $p$*  ( $L_p$ ) estimators, Penalva *et al.* [37] and Penalva [35],



both defined for  $\xi > 0$ . Two of the estimators developed for  $\xi \in \mathbb{R}$  are here considered: the Moment estimator, Dekkers *et al.* [8] and the Mixed Moment estimator, Fraga Alves *et al.* [16].

Recently, Caeiro *et al.* [4] introduced a class of *reduced bias* EVI-estimators. This class can not only reduce the bias of the classical estimators but also do not increase the asymptotic variance of the estimators, for adequate levels of  $k$  and adequate estimation of parameters of second-order  $(\beta, \rho) \in (\mathbb{R}, \mathbb{R}^-)$ . These are the scale and the shape second-order parameters, controlling the rate of first-order convergence, and necessary for establishing distributional properties of the estimators. Details on second-order conditions can be found in Beirlant *et al.* [2], de Haan and Ferreira [25] and Fraga Alves *et al.* [15], among others. Those estimators are then denoted *minimum-variance reduced biased* (MVRB) EVI-estimators. We will consider two of those estimators, one based on the Hill and the other on the  $L_p$  estimators, see Gomes *et al.* [21].

Let  $X_{1:n} \leq X_{2:n} \leq \dots \leq X_{n:n}$  be the order statistics associated to the sample  $(X_1, X_2, \dots, X_n)$ .

Let us define the log-excesses as  $V_{ik} := \ln X_{n-i+1:n} - \ln X_{n-k:n}$ , and  $M_{k,n}^{(l)} := \frac{1}{k} \sum_{i=1}^k [V_{ik}]^l$ , for  $l \in \mathbb{R} \setminus \{0\}$ , and  $L_{k,n}^{(r)} := \frac{1}{k} \sum_{i=1}^k \left[1 - \frac{X_{n-k:n}}{X_{n-i+1:n}}\right]^r$ , for  $r \geq 1$ .

The aforementioned estimators have the functional definitions:

- The Hill estimator, H, defined for  $\xi > 0$ , as

$$(3.1) \quad \widehat{\xi}^H(k) \equiv H(k) := \frac{1}{k} \sum_{i=1}^k V_{ik}, \quad k = 1, 2, \dots, n-1.$$

- The Moment estimator, M, defined for  $\xi \in \mathbb{R}$ , as

$$(3.2) \quad \widehat{\xi}_{k,n}^M := M_{k,n}^{(1)} + 1 - \frac{1}{2} \left(1 - \frac{(M_{k,n}^{(1)})^2}{M_{k,n}^{(2)}}\right)^{-1}, \quad k = 1, 2, \dots, n-1.$$

- The Mixed Moment estimator, MM, defined for  $\xi \in \mathbb{R}$ , as

$$(3.3) \quad \widehat{\xi}_{k,n}^{MM} := \frac{\widehat{\varphi}_{k,n} - 1}{1 + 2 \min(\widehat{\varphi}_{k,n} - 1, 0)}, \quad k = 1, 2, \dots, n-1,$$

where

$$\widehat{\varphi}_{k,n} := \frac{M_{k,n}^{(1)} - L_{k,n}^{(1)}}{(L_{k,n}^{(1)})^2}.$$

- The class of *Lehmer mean-of-order-p* ( $L_p$ ) estimators, defined for  $\xi > 0$  and

$p > 0$ , as

$$(3.4) \quad \widehat{\xi}^{\mathcal{L}}(k) \equiv L_p(k) := \frac{1}{p} \frac{\sum_{i=1}^k V_{ik}^p}{\sum_{i=1}^k V_{ik}^{p-1}}, \quad k = 1, 2, \dots, n-1, \quad [L_1(k) \equiv H(k)].$$

- The class of *corrected-Hill* (CH) EVI-estimators, defined by

$$(3.5) \quad \text{CH}(k) := H(k) \left( 1 - \hat{\beta}(n/k)^{\hat{\rho}} / (1 - \hat{\rho}) \right), \quad k = 1, 2, \dots, n-1,$$

where  $H(k)$  is the Hill estimator and  $\hat{\beta}$  e  $\hat{\rho}$  are consistent estimators of parameters  $\beta$  e  $\rho$ . The use of  $\text{CH}(k)$  enables us to eliminate the dominant component of bias of the H EVI-estimator,  $H(k)$ , keeping its asymptotic variance.

- More generally than the class in (3.5), we shall now also consider the direct reduction of the dominant bias component of  $L_p(k)$ , in (3.4), working with the RB Lehmer's EVI-estimators, Gomes *et al.* [21], defined by

$$(3.6) \quad L_p^{\text{RB}}(k) := L_p(k) \left( 1 - \hat{\beta}(n/k)^{\hat{\rho}} / (1 - \hat{\rho})^p \right), \quad k = 1, 2, \dots, n-1,$$

$$[L_1^{\text{RB}} \equiv \text{CH in (3.5)]}$$

Figure 2 shows the sample paths of estimates obtained when using the aforementioned estimators.

Values of  $p$  in  $L_p(k)$  and  $L_p^{\text{RB}}(k)$  were chosen using criteria given in Penalva [35].

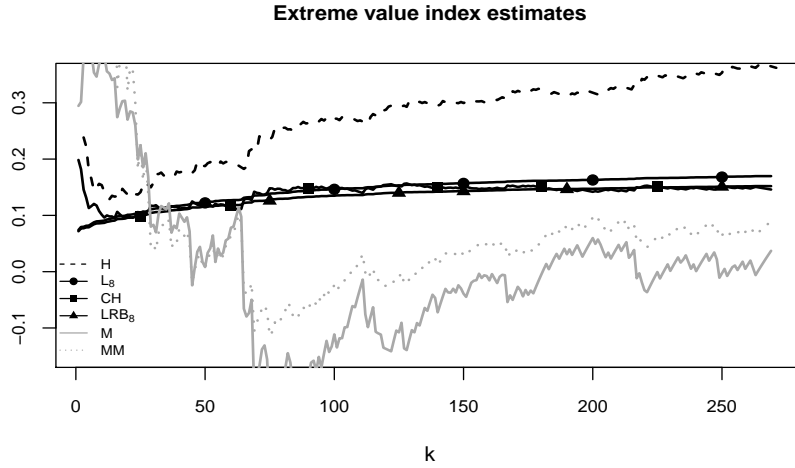
The discrepancies observed, already noticed in Penalva *et al.* [36], regarding the results of the above EVI-estimators and also compared with the results obtained under the parametric approaches claim for tests on extreme value domain of attraction. This emphasizes the care to be taken with the choice of the estimators, because even having very nice and stable paths, if conditions of their applicability are not verified, they may not stabilize near the true value of the parameter.

---

#### 4. TESTING CONDITIONS IN EVT LIMITING RESULTS

---

In any of the above procedures it is assumed that the underlying cdf  $F$  belongs to  $\mathcal{D}_{\mathcal{M}}(EV_{\xi})$ , for a appropriate value of  $\xi$ , or it is in specific sub-domains of  $\mathcal{D}_{\mathcal{M}}(EV_{\xi})$ . This condition is known as the *extreme value condition*.



**Figure 2:** Sample paths of the EVI-estimates considered.

---

#### 4.1. Testing the extreme value condition

---

It is then important, before any application, to check the assumption:

$$(4.1) \quad H_0 : F \in \mathcal{D}_{\mathcal{M}}(EV_{\xi}) \text{ for some } \xi \in \mathbb{R}.$$

Some tests for the hypothesis  $H_0$  are available, such as those in Dietrich *et al.* [9], Drees *et al.* [10] and Hüsler and Li [28].

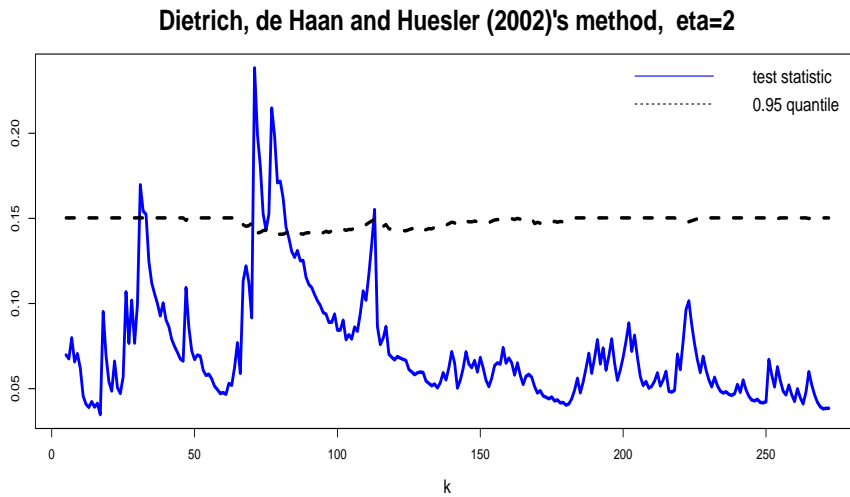
Let  $X_1, X_2, \dots, X_n$  be iid random variables with cdf  $F$  and suppose that some additional second order conditions hold then, for  $\eta > 0$ , Dietrich *et al.* [9] introduced the test statistic written as

$$(4.2) \quad E_n := k \int_0^1 \left( \frac{\log X_{n-[kt],n} - \log X_{n-k,n}}{\hat{\xi}_+} - \frac{t^{-\hat{\xi}_-} - 1}{\hat{\xi}_-} (1 - \hat{\xi}_-) \right)^2 t^{\eta} dt,$$

where  $k$  is again an intermediate sequence such that  $k = k_n \rightarrow \infty$ ,  $k/n \rightarrow 0$  and  $k^{1/2}A(n/k) \rightarrow 0$  as  $n \rightarrow \infty$  and  $A$  is related to the second order condition already referred to and  $\hat{\xi}_+$  and  $\hat{\xi}_-$  are the moment estimators, Dekkers *et al.* [8], of  $\xi_+ := \max(0, \xi)$  and  $\xi_- := \min(0, \xi)$ .

Hüsler and Li [28] present an algorithm for testing  $H_0$  using the test statistic  $E_n$  in (4.2). They have carried out an extensive simulation study with guidelines for obtaining the value of  $\eta$  and have provided quite accuracy tables for the quantiles  $\chi_{1-\alpha}$  of the variable limiting of  $E_n$ , see Hüsler and Li [28] for details. Values of  $E_n$  are compared with values of  $\chi_{1-\alpha}$ : if  $E_n > \chi_{1-\alpha}$  hypothesis  $H_0$  is rejected with a type I error  $\alpha$ . Otherwise there is no reason to reject  $H_0$ .

For our data, the application of the test based on (4.2), provided values of the test statistic smaller than the corresponding asymptotic 0.95–quantile for a large range of  $k$ –values. So, since the sample path of test statistic is almost always outside the rejection region, except for a small range of  $k$ , we find no evidence to reject the null hypothesis, see Figure 3.



**Figure 3:** Plot of the sample paths for the E-test, based on the test statistic in (4.2) and the corresponding quantile. Available sample size  $n = 11946$ .

See also Hüsler and Li [28], Neves and Fraga Alves [32] and Penalva *et al.* [36] for a description of other tests.

---

#### 4.2. Statistical choice of extreme domains of attraction

---

Once the hypothesis  $H_0 : F \in \mathcal{D}_{\mathcal{M}}(EV_{\xi})$  is not rejected, it is of major importance to decide for the type of the tail, i.e., the natural hypothesis testing are now:

$$(4.3) \quad H_0 : F \in \mathcal{D}_{\mathcal{M}}(EV_0) \quad vs \quad H_1 : F \in \mathcal{D}_{\mathcal{M}}(EV_{\xi})_{\xi \neq 0},$$

or against the one-sided alternatives

$$F \in \mathcal{D}_{\mathcal{M}}(EV_{\xi})_{\xi < 0} \quad or \quad F \in \mathcal{D}_{\mathcal{M}}(EV_{\xi})_{\xi > 0}.$$

This is known as the *statistical choice of extreme domains of attraction*.

Under the semiparametric framework, several tests have been proposed in literature, among which we can mention: Galambos [18], Castillo *et al.* [5];

Hasofer and Wang [26]; Falk [11]; Correia and Neves [7], that considered the Hasofer and Wang statistic and presented a slight modification. An extensive simulation study has been performed in Fraga Alves and Gomes [13], Marohn [29, 30], Fraga Alves [14] and Segers and Teugels [39]. Castillo *et al.* [5] considered tests to distinguish between polynomial and exponential tails, based on properties of the *coefficient of variation* (CV).

Neves and Fraga Alves [32, 33] studied the following tests statistics, that will be here applied.

The **Ratio-test**:

$$(4.4) \quad R_n^*(k) := \frac{X_{n:n} - X_{n-k:n}}{\frac{1}{k} \sum_{i=1}^k (X_{n-i+1:n} - X_{n-k:n})} - \log k \xrightarrow[n \rightarrow \infty]{d} \text{EV}_0.$$

The **Gt-test**:

$$(4.5) \quad G_n(k) := \frac{\frac{1}{k} \sum_{i=1}^k (X_{n-i+1:n} - X_{n-k:n})^2}{\left(\frac{1}{k} \sum_{i=1}^k X_{n-i+1:n} - X_{n-k:n}\right)^2},$$

and

$$G_n^*(k) = \sqrt{k/4} (G_n(k) - 2) \xrightarrow[n \rightarrow \infty]{d} N(0, 1).$$

The **HW-test**:

$$(4.6) \quad W_n(k) := \frac{1}{k} \left[ 1 - \frac{G_n(k) - 2}{1 + (G_n(k) - 2)} \right],$$

and

$$W_n^*(k) = \sqrt{k/4} (kW_n(k) - 1) \xrightarrow[n \rightarrow \infty]{d} N(0, 1).$$

For the two-sided tests  $R^*$ ,  $G^*$  or  $W^*$ , the null hypothesis is rejected if  $R^*(G^*)(W^*) < \chi_{\alpha/2}$  or  $R^*(G^*)(W^*) > \chi_{1-\alpha/2}$ , where  $\chi_p$  is the  $p$  probability quantile of the corresponding limiting distribution.

For the one-sided tests, the null hypothesis is rejected in favour of either unilateral alternatives, for example, for  $R_n^*$ ,

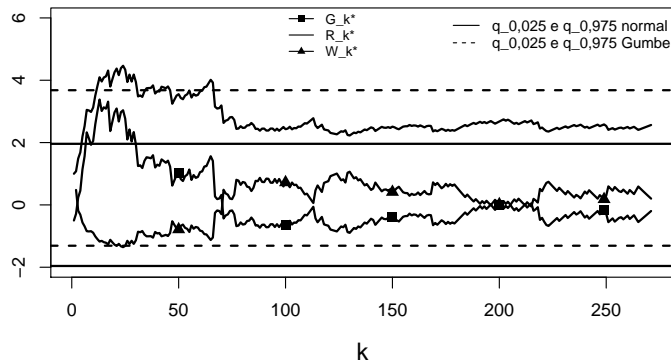
$$H_1^l : F \in D_M(\text{EV}_\xi)_{\xi < 0} \quad \text{or} \quad H_1^r : F \in D_M(\text{EV}_\xi)_{\xi > 0},$$

if

$$R_n^*(k) < \chi_\alpha \quad \text{or} \quad R_n^*(k) > \chi_{1-\alpha}.$$

Figure 4 illustrates the application of those tests.

These tests suggest the non rejection of the null hypothesis, leading us to consider that the underlying distribution of the data are in the domain of attraction of the Gumbel distribution.



**Figure 4:** Sample paths of the statistics  $R_n^*$ , with the associated quantiles  $\chi_{0.025}$  and  $\chi_{0.975}$  for the standard Gumbel distribution in dashed lines, and the  $G_n^*$  and  $W_n^*$  sample paths statistics, with the associated quantiles of the standard normal distribution in solid lines.

Such as we have already pointed out in Penalva *et al.* [36], with fewer years of data, we think that this explains the discrepancy observed in Figure 2, where were plotted sample paths of very well behaved EVI-estimators, but not adequate to the tail of the data under study. We claim again for the need of performing at first the tests described and illustrated briefly in this Section.

---

## 5. SECOND CASE-STUDY – THE ANALYSIS

---

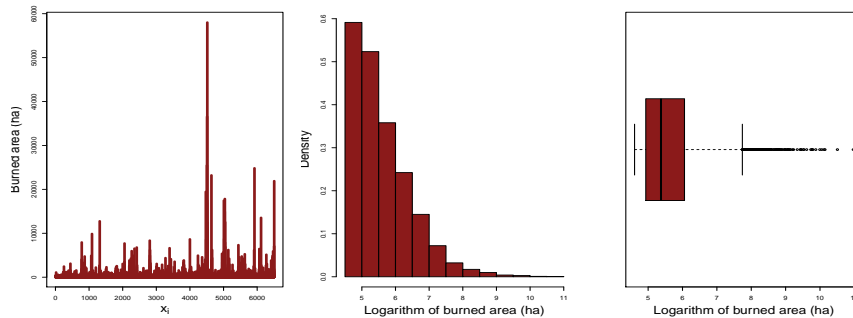
The second set of data analysed in this work, and also studied in Gomes *et al.* [20] based on a shorter period of time, consists of the burned area (ha), in Portugal, related to each of the wildfires occurred in a period from 1984 to 2016, exceeding 100 ha, making a total of 6507 observations. The data analysed here do not seem to have a significant temporal structure. This new data set is used to illustrate what we have just commented.

The main results of a graphical and descriptive analysis are shown in Table 4 and in Figure 5. Tables and graphics provide evidence on the heaviness of the right tail. Notice that similar conclusions were obtained by Beirlant *et al.* [1], for data analysis of burned area of wildfires exceeding 100 ha, recorded in Portugal from 1990 till 2003 ( $n = 2627$ ).

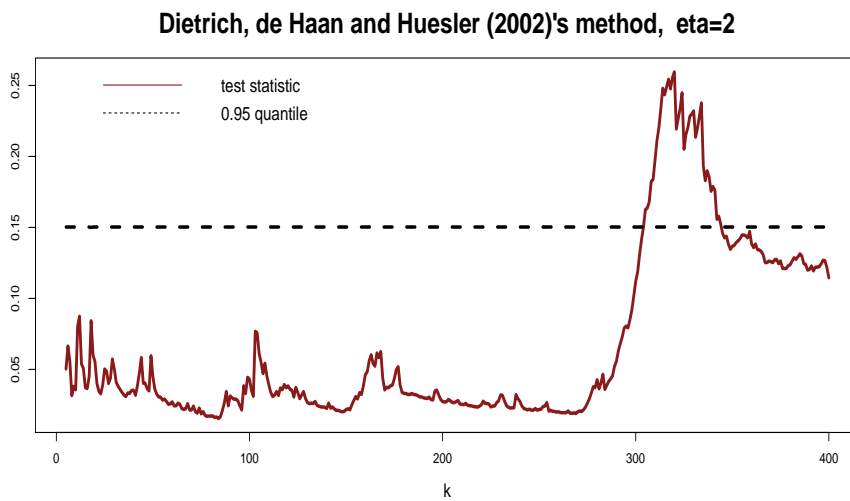
See in Figure 6 the application of the test to the extreme value condition, based on (4.2). We find no evidence to reject the null hypothesis, i.e.,  $F \in \mathcal{D}_{\mathcal{M}}(EV_{\xi})$ .

min	1st Qu.	Median	Mean	3rd Qu.	max
100	138.55	215.81	485.35	427.51	58012.75
n		Skewness	Kurtosis	St Dev	
6507		19.01	568.90	1407.58	

**Table 4:** Descriptive statistics for burned area of wildfires exceeding 100ha.



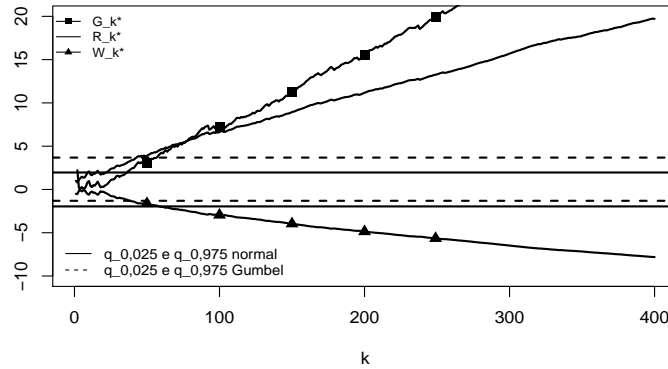
**Figure 5:** Plot of burned areas, histogram and boxplot, for wildfires, exceeding 100 ha.



**Figure 6:** Plot of the sample paths for the E-test, based on (4.2) statistics, with the corresponding quantile. Available sample size  $n = 6507$ .

The tests to the statistical choice of the tail, such as was described and presented in Subsection 4.2, produced now the plots presented in Figure 7. Those tests suggest the rejection of the null hypothesis, leading us to consider that the underlying distribution of the data is in the domain of attraction of the Fréchet

distribution.

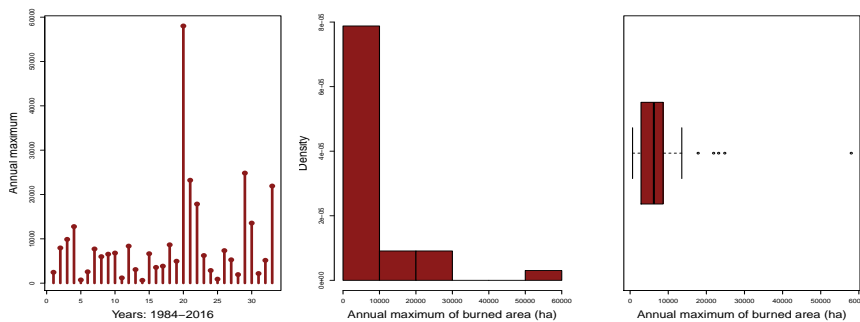


**Figure 7:** Sample paths of the statistics  $R_n^*$ , with the associated quantiles  $\chi_{0.025}$  and  $\chi_{0.975}$  for the standard Gumbel distribution in dashed lines, and the  $G_n^*$  and  $W_n^*$  sample paths statistics, with the associated quantiles of the standard normal distribution in solid lines.

Here we will consider again, in the BM methodology, blocks as the years of observations,  $m = 33$ . Figure 8 and Table 5 were obtained for the burned area of wildfires exceeding 100 ha.

min	1st Qu.	Median	Mean	3rd Qu.	max
641.33	2860.10	6235.83	8956.80	8652.43	58012.75
	m	Skewness	Kurtosis	St Dev	
	33	2.90	9.85	10889.31	

**Table 5:** Basic descriptive statistics for maximum values in each year.



**Figure 8:** Maximum value of burned areas in each year, histogram and boxplot.



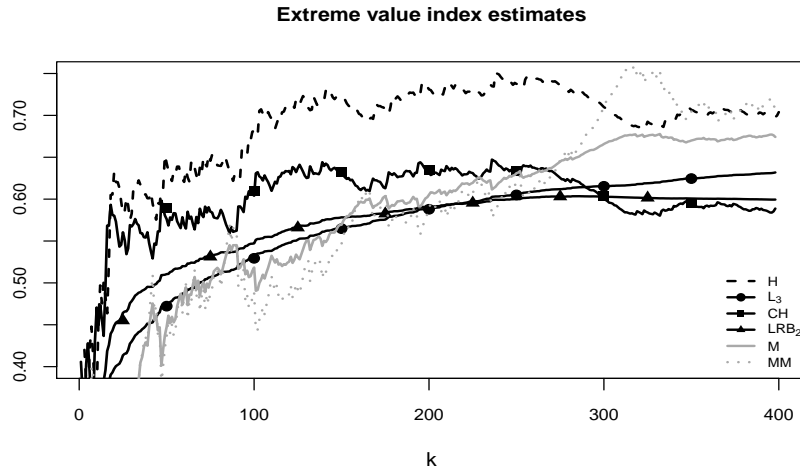
Below are given the estimates of the main parameters.

$\hat{\xi}$	$\hat{\lambda}$	$\hat{\delta}$
0.52 (0.21)	4007.95 (754.45)	3599.50 (727.93)

**Table 6:** Maximum likelihood estimates (standard errors in parenthesis).

The  $\xi$  estimate corroborates the first idea pointing that the data present clearly a tail heavier than that one of the first case-study.

Figure 9 shows the sample paths of estimates obtained using the aforementioned estimators. Values of  $p$  in  $L_p(k)$  and  $L_p^{\text{RB}}(k)$  were also chosen using criteria given in Penalva [35]. A quick analysis of the sample paths of the EVI-estimates allow us to consider as  $\hat{\xi}$  a value between 0.55 and 0.65, which is also in agreement with a heavy tail detected for the underlying cdf  $F$  and with the result obtained under the parametric approach.



**Figure 9:** Sample paths of the EVI-estimates considered.

---

## 6. FIRST COMMENTS ON PRACTICAL EFFECTS OF MISSING THOSE TESTS. A FEW COMMENTS

---

We showed, with this work based on the two case-studies, that the realization of tests on the extreme value conditions and on the statistical choice of the tail of the underlying distribution are with no doubt the first step to properly apply the several estimation approaches and to choose the more adequate estimators.

A first illustration of the practical effects in the estimation of other important parameters when the choice of the tail is performed or not *a priori*, is presented. It is well known how an accurate EVI estimation is important because it dominates the tail behaviour of a distribution. However in several situations, such as risk management or catastrophic situations, where human lives can be in danger, in addition to modelling the tails, other parameters are of the major importance to be estimated, such as extreme quantiles, return levels or return periods of the distribution of the process at risk. For the first case study, high quantiles were estimated.

While it is true that EVI determines the asymptotic behaviour of the tail and the quantiles of a distribution, other parameters (for example, scale and location) are no less important for an accurate estimation of quantiles, see Matthys and Beirlant [31] and Caeiro and Gomes [3], among others.

In the first example studied, Section 3., and in the parametric approach, a negative value, although very close to zero, was obtained for  $\hat{\xi}$ . Now considering the location and scale parameters estimates and by inverting the  $EV_{\xi}$  cdf in (2.3), for  $\xi \neq 0$ , the extreme quantiles, for very small values of  $p$ , can be easily estimated as

$$(6.1) \quad \hat{\chi}_{1-p} := \hat{\lambda} - \frac{\hat{\delta}}{\hat{\xi}} \left[ 1 - (-\ln(1-p))^{-\hat{\xi}} \right].$$

For example, for  $p = 0.01, 0.001, 0.0001$ , the corresponding quantile estimates are  $\hat{\chi}_{0.99} = 753.9114$ ;  $\hat{\chi}_{0.999} = 1000.7254$  and  $\hat{\chi}_{0.9999} = 1230.6420$ .

In the semi-parametric framework, and using the estimates displayed in Figure 2 that show a more stable sample path (and also the Hill estimates as reference), as usually is done, high quantile estimates, also for the previous values of  $p$  were calculated.

It was used the moment estimator described in Matthys and Beirlant [31], subsection 2.3, defined as:

$$(6.2) \quad \hat{\chi}_{1-p,k+1}^{\hat{\xi}} := X_{n-k:n} \hat{a}_{n,k+1}^{\hat{\xi}} \frac{c_n^{\hat{\xi}} - 1}{\hat{\xi}}; \quad c_n := \frac{k}{np} \quad \text{for } k < n$$

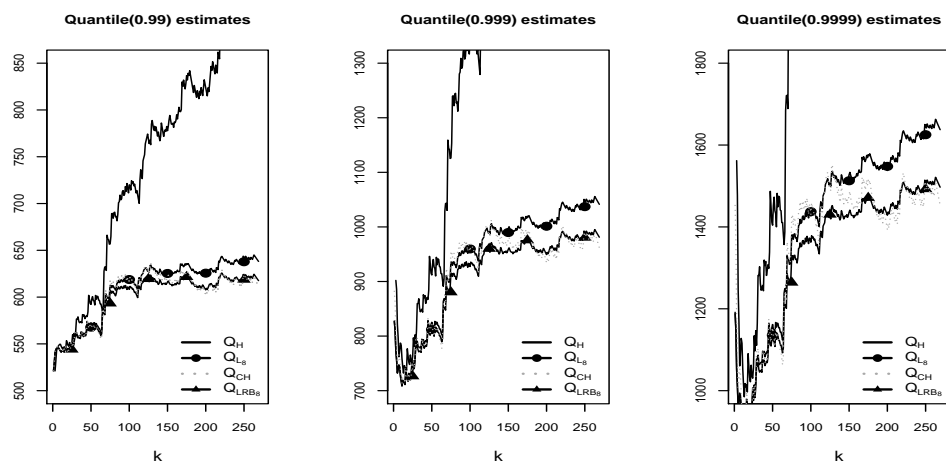
with

$$\hat{a}_{n,k+1}^{\hat{\xi}} = \frac{X_{n-k:n} H}{\rho_1(\hat{\xi})}, \quad \rho_1(\xi) = \begin{cases} 1 & \text{for } \xi \geq 0 \\ 1/(1-\xi) & \text{for } \xi < 0. \end{cases}$$

where  $\hat{\xi}$  is a consistent estimator of  $\xi$ . Here the H, L<sub>8</sub>, CH and LRB<sub>8</sub> estimates, displayed in Figure 2, were used in (6.2).

Figure 10 shows the paths of  $\hat{\chi}_{0.99}(k)$ ,  $\hat{\chi}_{0.999}(k)$  and  $\hat{\chi}_{0.9999}(k)$ .

However, if we have first performed the statistical test in (4.3), we were led not to reject the null hypothesis so we will consider  $\xi = 0$ . In this case the



**Figure 10:** Sample paths of the quantiles estimates.

extreme quantiles can be estimated under the approach aforementioned, based on the inversion of the  $EV_\xi$  cdf in (2.3), for  $\xi = 0$ , i.e.

$$(6.3) \quad \hat{\chi}_{1-p}(k) := \hat{\lambda} - \hat{\delta} \ln(-\ln(1-p)),$$

and for the previous values of  $p$  we will obtain  $\hat{\chi}_{0.99} = 788.3877$ ;  $\hat{\chi}_{0.999} = 1079.6836$  and  $\hat{\chi}_{0.9999} = 1370.4655$ .

We see that the quantiles estimates show large discrepancies among the procedures used. It is then advisable to perform a careful choice of the tail and also of the EVI-estimators in which the quantile estimates are based. This is out of scope of this article and an important topic for future research.

The next challenge is modelling and estimating clusters of extreme values since they are linked with incidences and durations of catastrophic phenomena. Here, an important parameter comes into play, the extremal index  $\theta$ , that characterizes the degree of local dependence in the extremes of a stationary sequence. It needs to be adequately estimated, not only by itself but because its influence on other relevant parameters, such as a high quantile. Ignoring  $\theta$  may lead to an underestimation of marginal quantile of  $F$  and an overestimation of quantiles of the EV.

---

## ACKNOWLEDGMENTS

---

The authors are grateful to the two anonymous referees for their careful reviews and helpful suggestions and comments, which have highly improved the final version of this article.

This work has been supported by **FCT**—Fundação para a Ciência e a Tecnologia, Portugal, through the projects UID/MAT/00006/2013 (CEAUL), UID/MAT/00006/2019 (CEAUL) and UID/MAT/0297/2013 (CMA/UNL).

We also thank our colleagues José Miguel Cardoso Pereira and Ana Sá for providing us with the data on burned areas.

---

## REFERENCES

---

- [1] BEIRLANT, J.; FRAGA ALVES, M.I. and GOMES, M.I. (2016). Tail fitting for truncated and non-truncated Pareto-type distributions, *Extremes*, **19**, 429–462.
- [2] BEIRLANT, J.; GOEGBEUR, Y.; SEGERS, J. and TEUGELS, J. (2004). *Statistics of Extremes. Theory and Applications*. Wiley.
- [3] CAEIRO, F. and GOMES, M.I. (2008). Minimum-variance reduced-bias tail index and high quantile estimation. *REVSTAT - Statistical Journal*, **6**, 1–20.
- [4] CAEIRO, F.; GOMES, M.I. and PESTANA, D.D. (2005). Direct reduction of bias of the classical Hill estimator, *REVSTAT - Statistical Journal*, **3**, 111–136.
- [5] CASTILLO, J. DEL; DAOUDI, J. and LOCKHART, R. (2014). Methods to distinguish between polynomial and exponential tails. *Scandinavian Journal of Statistics*, **41**, 382–393.
- [6] CASTILLO, E.; GALAMBOS, J. and SARABIA, J.M. (1989). *The selection of the domain of attraction of an extreme value distribution from a set of data..* In “Extreme value theory (Oberwolfach, 1987) – Lecture Notes in Statistics” (J. Hüslér and R.-D. Reiss, Eds.), Springer, Berlin-Heidelberg, **51**, 181–190.
- [7] CORREIA, A.L. and NEVES, M. (1996). *Escolha estatística em modelos extremos—testes de ajustamento*. In “Bom Senso e Sensibilidade” (J. Branco, P. Gomes and J. Prata, Eds.), Actas do III Congresso Anual da Sociedade Portuguesa de Estatística, Edições Salamandra, 223–236.
- [8] DEKKERS, A.L.M.; EINMAHL, J.H.J. and DE HAAN, L. (1989). A moment estimator for the index of an extreme-value distribution, *Annals of Statistics*, **17**, 4, 1833–1855.
- [9] DIETRICH, D.; DE HAAN, L. and HÜSLER, J. (2002). Testing extreme value conditions, *Extremes*, **5**, 1, 71–85.
- [10] DREES, H.; DE HAAN, L. and LI, D. (2006). Approximations to the tail empirical distribution function with application to testing extreme value conditions, *Journal of Statistical Planning and Inference*, **136**, 3498–3538.
- [11] FALK, M. (1995). On testing the extreme value index via the POT-method, *Annals of Statistics*, **23**, 2013–2035.
- [12] FISHER, R.A. and TIPPETT, L.H.C. (1928). Limiting forms of the frequency distributions of the largest or smallest member of a sample, *Proceedings of the Cambridge Philosophical Society*, **24**, 180–190.

- [13] FRAGA ALVES, M.I. and GOMES, M.I. (1996). Statistical choice of extreme value domains of attraction – a comparative analysis. *Communications in Statistics – Theory and Methods*, **25**, 4, 789–811.
- [14] FRAGA ALVES, M.I. (1999). Asymptotic distribution of Gumbel statistic in a semi-parametric approach, *Portugaliae Mathematica*, **56**, 3, 282–298.
- [15] FRAGA ALVES, M.I., GOMES, M.I., DE HAAN, L. and NEVES, C. (2007). A note on second order conditions in extreme value theory: linking general and heavy tails conditions. *REVSTAT - Statistical Journal*, **5**, 3, 285–305.
- [16] FRAGA ALVES, M.I.; GOMES, M.I.; DE HAAN, L. and NEVES, C. (2009). Mixed moment estimator and location invariant alternatives, *Extremes*, **12**, 149–185.
- [17] FRÉCHET, M. (1927). Sur la loi de probabilité de l'écart maximum, *Annales de la Société Polonaise de Mathématique (Cracovie)*, **6**, 93–116.
- [18] GALAMBOS, J. (1982). A statistical test for extreme value distributions. In “Non-parametric Statistical Inference” (B.V. Gnedenko *et al.*, ed.), North Holland, Amsterdam, 221–230.
- [19] GNEDENKO, B. V. (1943). Sur la distribution limite d'une série aléatoire, *Annals of Mathematics*, **44**, 423–453.
- [20] GOMES, M.I.; FIGUEIREDO, F. and NEVES, M.M. (2012). Adaptive estimation of heavy right tails: resampling-based methods in action, *Extremes*, **15**, 463–489.
- [21] GOMES, M.I.; PENALVA, H., CAEIRO, F. and NEVES, M.M. (2016). Nonreduced versus reduced-bias estimators of the extreme value index-efficiency and robustness. In “COMPSTAT 2016 22<sup>nd</sup> International Conference on Computational Statistics” (A. Colubi, A. Blanco and C. Gatu, Eds), 279–290.
- [22] GUMBEL, E.J. (1935). Les valeurs extrêmes des distributions statistiques, *Annales de l'institut Henri Poincaré*, **5**, 2, 115–158.
- [23] GUMBEL, E.J. (1958, 2004). *Statistics of Extremes*, Columbia University Press, New York.
- [24] DE HAAN, L. (1970). *On Regular Variation and its Applications to the Weak Convergence of Sample Extremes*, Mathematical Centre Tract 32, Amsterdam, Dordrecht: D. Reidel.
- [25] DE HAAN, L. and FERREIRA, A. (2006). *Extreme Value Theory: An Introduction*. Springer Science+Business Media, LLC, New York.
- [26] HASOFER, A. and WANG, J.Z. (1992). A test for extreme value domain of attraction. *Journal of the American Statistical Association*, **87**, 171–177.
- [27] HILL, B.M. (1975). A simple general approach to inference about the tail of a distribution, *Annals of Statistics*, **3**, 1163–1174.
- [28] HÜSLER, J. and LI, D. (2006). On testing extreme value conditions, *Extremes*, **9**, 69–86.
- [29] MAROHN, F. (1998a). An adaptive efficient test for Gumbel domain of attraction. *Scandinavian Journal of Statistics*, **25**, 311–324.
- [30] MAROHN, F. (1998b). Testing the Gumbel hypothesis via the POT-method. *Extremes*, **1**, 2, 191–213.
- [31] MATTHYS, G. and BEIRLANT, J. (2003). Estimating the extreme value index and high quantiles with exponential regression models, *Statistica Sinica*, **13**, 853–880.

- [32] NEVES, C. and FRAGA ALVES, M.I. (2007). Semi-parametric approach to the Hasofer-Wang and Greenwood statistics in extremes, *Test*, **16**, 297–313.
- [33] NEVES, C. and FRAGA ALVES, M.I. (2008). Testing extreme value conditions – an overview and recent approaches, *REVSTAT - Statistical Journal*, **6**, 1, 83–100.
- [34] NEVES, M.M.; PENALVA, H. and NUNES, S. (2015). *Extreme value analysis of river levels in a hydrometric station in the North of Portugal*. In “Current Topics on Risk Analysis: ICRA6 and RISK 2015 Conference proceedings” (M. Guillén, A. Juan, H. Ramalhinho, I. Serra and C. Serrat, Edts.), 533–538.
- [35] PENALVA, H. (2017). *Contributos Computacionais e Metodológicos na Estimação do Índice de Valores Extremos*. Tese de Doutoramento, ISA - Universidade de Lisboa, Portugal. <http://hdl.handle.net/10400.5/14946>
- [36] PENALVA, H.; NUNES, S. and NEVES, M. (2016). Extreme Value Analysis – a brief overview with an application to flow discharge rate data in a hydrometric station in the north of Portugal. *REVSTAT - Statistical Journal*, **14**, 2, 193–215.
- [37] PENALVA, H.; CAEIRO, F.; GOMES, M.I. and NEVES, M. (2016). *An Efficient Naive Generalization of the Hill Estimator – Discrepancy between Asymptotic and Finite Sample Behaviour*. Notas e Comunicações CEAUL 02/2016.
- [38] R CORE TEAM (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>.
- [39] SEGERS, J. and TEUGELS, J. (2000). Testing the Gumbel hypothesis by Galton’s ratio, *Extremes*, **3**, 3, 291–303.
- [40] VON MISES, R. (1936). La distribution de la plus grande de n valeurs., *American Mathematical Society*, Reprinted in Selected Papers Volumen II, Providence, R.I. (1954), 271–294.



---

---

## ON THE PARAMETERS ESTIMATION OF HIV DYNAMIC MODELS \*

---

---

Authors: DIANA ROCHA

– Center for R&D in Mathematics and Applications (CIDMA),  
University of Aveiro, Portugal  
`diana.isa.rocha@ua.pt`

SÓNIA GOUVEIA

– Institute of Electronics and Informatics Engineering of Aveiro (IEETA)  
and CIDMA, University of Aveiro  
`sonia.gouveia@ua.pt`

CARLA PINTO

– Centre of Mathematics of the University of Porto (CMUP) and  
School of Engineering, Polytechnic of Porto  
`cpinto@fc.up.pt`

MANUEL SCOTTO

– Center for Computational and Stochastic Mathematics (CEMAT) and  
Department of Mathematics, IST, University of Lisbon  
`manuel.scotto@tecnico.ulisboa.pt`

JOÃO NUNO TAVARES

– Centre of Mathematics of the University of Porto (CMUP)  
`jntavar@fc.up.pt`

EMÍLIA VALADAS AND LUÍS FILIPE CALDEIRA<sup>†</sup>

– Infectious Disease Service, Hospital Santa Maria, Lisbon (HSM/CHLN)  
`evaladas@medicina.ulisboa.pt` and  
`luis.caldeira@chln.min-saude.pt`

Received: October 2018

Revised: January 2019

Accepted: March 2019

Abstract:

- This work proposes an estimation method to obtain the optimal parameter estimates of a mathematical model, from a set of CD4<sup>+</sup>T values collected in a HIV patient. To this end, the following scheme is adopted: the first step consists in selecting an initial estimate for the model's parameters as that having minimum square error, from a set of uniform randomly generated candidates. In the second step, the initial solution is refined by an optimization algorithm with constraints and bounds (imposed by physiology), resulting on the optimal estimate. The proposed method is validated through a simulation study and illustrated with an application to a real data set of CD4<sup>+</sup>T cells counts for several HIV patients.

---

\*The opinions expressed in this text are those of the authors and do not necessarily reflect the views of any organization.

<sup>†</sup> Deceased January 29, 2019.



## Key-Words:

- *Parameter estimation; nonlinear programming; mathematical models; human immunodeficiency virus (HIV).*

## AMS Subject Classification:

- 49A05, 78B26.

---

## 1. INTRODUCTION

---

In the clinical follow-up of a HIV/AIDS patient, the viral load values and the CD4<sup>+</sup>T cells count, observed over time, constitute a set of non-equally spaced observations. In general, no information between consultations is available. In this context, it is clinically relevant to develop methods able to obtain a more complete description of the individual time evolution, either between consecutive consultations or for the prediction of evolution or disease progression. Viral dynamic models can be formulated through a system of nonlinear ordinary differential equations, which enables to describe the temporal evolution of the clinical parameters of a HIV patient [2, 10, 11]. In the last decades, a few literature studies show applications and developments in statistical methodologies for model inference including those based on Bayesian inference [6]. Briefly, the Bayesian approach incorporates non-informative prior distributions and yet the corresponding algorithms require initial estimates for model's parameters in order to carry out the iterative updates of the parameters. For the estimation of this initial estimates, the most commonly used approaches in practice are based on nonlinear least squares [7, 8]. In this context, this work presents a nonlinear programming approach to obtain the optimal estimates for the parameters of a HIV dynamic model. Our proposal differs from previous approaches in the fact that we add restrictions on the optimal estimate so that it verifies an equal contribution of negative and positive deviations from observations. Furthermore, the optimal estimate is restricted to be in-between lower and upper physiological bounds. Note that the least square methods are implemented as optimization problems requiring initial solutions to start the iterations. To cope with this limitation, we consider as initial solution the minimum square error solution from a set of 1000 uniform randomly-generated candidates on a uniform distribution delimited by the lower and upper physiological bounds. Therefore, the proposed method is fully automatic and does not require any other information to provide the optimal estimate of the model's parameters besides the data.

This paper is organized as follows: the methods concerning the description of the mathematical HIV model and the estimation approach developed to obtain the initial conditions for the model's parameters are presented in Section 2. The estimation approach is illustrated with simulated data that mimics the individual temporal trajectories of three HIV patients. The simulation strategy is described in Section 3, whereas the results on simulation and on real data from six HIV patients [12] are presented in Section 4. The selected patients were chosen according to some conditions, namely having started an antiretroviral treatment at the beginning of the trial [12]. Finally, Section 5 is devoted to conclusions.

---

## 2. METHODS

---

The dynamics of the HIV/AIDS infection is described through the mathematical model presented in Section 2.1. Model's parameters are estimated from the nonlinear programming approach described in Section 2.2. The developed algorithms and other software code used in this work were implemented in MATLAB<sup>TM</sup> (version R2015a), The Mathworks Inc., MA, USA.

---

### 2.1. Mathematical model

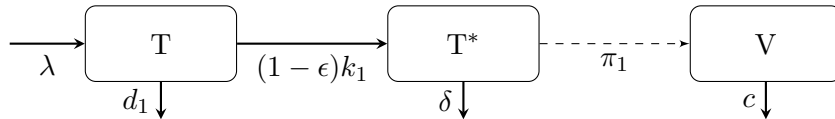
---

The mathematical model considered in this work translates known physiological relationships between viral load and CD4<sup>+</sup>T cells and incorporates parameters having clinical interpretation. We consider a modified version of the mathematical model in Stafford *et al.* [14] for the dynamics of HIV/AIDS infection, including an additional parameter  $\epsilon$  that denotes the effectiveness of the antiretroviral therapy [9]. The model is represented as

$$(2.1) \quad \begin{aligned} \frac{dT(t)}{dt} &= \lambda - d_1 T(t) - (1 - \epsilon)k_1 T(t)V(t), \\ \frac{dT^*(t)}{dt} &= (1 - \epsilon)k_1 T(t)V(t) - \delta T^*(t), \\ \frac{dV(t)}{dt} &= \pi_1 T^*(t) - cV(t), \end{aligned}$$

where the state variables are the viral load  $V(t)$  and the number of CD4<sup>+</sup>T cells defined as  $CD4(t) = T(t) + T^*(t)$ , with  $T(t)$  and  $T^*(t)$  representing the number of uninfected and infected CD4<sup>+</sup>T cells, respectively. Furthermore, for simplicity in notation we denote  $(T(0), T^*(0), V(0)) = (T_0, T_0^*, V_0)$  as the initial condition of the model. Along with the states variables, the mathematical model also incorporates parameters with clinical interpretation, namely  $\theta = (d_1, \epsilon, k_1, \delta, \pi_1, c)$ , with definition and units listed in Table 1.

The mathematical model in (2.1) can be alternatively defined from the flow-chart displayed in Figure 1.



**Figure 1:** Schematic diagram of the model (2.1).

The chart presents a compartmental description of the model that translates the evolution of the disease at the patient level. Within each compartment

Parameter	Definition	Units
$d_1$	difference between rate loss from cell death and rate gain due to cell division	day <sup>-1</sup>
$\lambda = T_0 d_1$	proliferation rate of uninfected target cells	cells ml <sup>-1</sup> day <sup>-1</sup>
$\epsilon$	effectiveness of therapy	
$k_1$	infectivity rate	ml day <sup>-1</sup>
$\delta$	death rate of infected cells	day <sup>-1</sup>
$\pi_1$	average number of virions produced by a single infected cell	day <sup>-1</sup>
$c$	clearance rate of free virions	day <sup>-1</sup>

**Table 1:** Definition and units of the parameters included in (2.1).

there are CD4<sup>+</sup>T cells (non-infected or infected) or viral load. In this representation, these units can move between compartments. For instance, the susceptible CD4<sup>+</sup>T cells in compartment  $T$  move to compartment  $T^*$  (infected cells) after being infected with HIV at a rate equal to  $(1 - \epsilon)k_1$ , and infected CD4<sup>+</sup>T cells of compartment  $T^*$  die at rate  $\delta$ .

---

## 2.2. Nonlinear programming

---

The parameters in  $\theta$  can be estimated from a set of  $CD4(t)$  values collected in one HIV patient at its clinical follow-up appointments over time. Let  $CD4(t_i)$  be the observed number of CD4<sup>+</sup>T cells at time  $t_i, i = 1, 2, \dots, n$ . Furthermore, define  $\widehat{CD4}(t_i) = T(t_i) + T^*(t_i)$  as the estimate of  $CD4(t_i)$  provided by the mathematical model (2.1). The *optimal* parameter estimates, say  $\widehat{\theta}$ , can be obtained by minimizing the square error between model estimates and observed CD4 values. In accordance with other literature studies [6], we considered a log10-transformation on the parameters to ensure their positiveness and to stabilize the  $CD4(t)$  variance. Thus, the nonlinear programming algorithm can be formulated as

$$\begin{aligned}
 & \text{minimize} && f(\theta) = \sum_{i=1}^n (\widehat{CD4}(t_i) - CD4(t_i))^2 = \sum_{i=1}^n e_{t_i}^2 \\
 (2.2) \quad & \text{subject to} && \sum_{i=1}^n e_{t_i} = 0 \\
 & \text{and} && \mathbf{lb} \leq \theta \leq \mathbf{ub}
 \end{aligned}$$

where the restriction guarantees that  $\widehat{\theta}$  verifies equal contribution of negative and positive deviations from observations. Also,  $\widehat{\theta}$  is restricted to physiological lower and upper bounds, respectively  $\mathbf{lb} = (0.01, 0, 10^{-11}, 0.24, 50, 2.39)$  and  $\mathbf{ub} = (0.02, 1, 10^{-5}, 0.7, 10000, 23)$  [5, 14]. This optimization procedure was implemented with the MATLAB<sup>TM</sup> function *fmincon*, that starts at an initial solution

$\theta^*$  to find a minimizer  $\hat{\theta}$  of  $f(\theta)$  subject to the above-mentioned restrictions and bounds. The initial solution  $\theta^*$  was obtained as that minimizing  $f(\theta)$  in a set of 1000 candidates randomly generated from a multivariate uniform distribution on  $\mathbf{lb}$  and  $\mathbf{ub}$ .

The HIV dynamic model (2.1) was implemented with MATLAB<sup>TM</sup> function *ode45*. This function makes use of an explicit Runge-Kutta formula, namely the Dormand-Prince pair [4], that computes the solution at time  $t_k$  based on the solution at time  $t_{k-1}$ . Furthermore, when the integration is considered in a time span, the algorithm runs with a variable time step for efficient computation. In this case, temporal resampling is needed to obtain the solutions at specific  $t_i, i = 1, 2, \dots, n$  (continuous time). Alternatively, the solver can provide the solution at requested time points  $t_i$  with its own built-in interpolation algorithm (discrete time). The differences between continuous/discrete time solutions were used to determine if differences between solutions evaluated at  $\theta$  and at  $\hat{\theta}$  are numerically relevant.

---

### 3. SIMULATED DATA

---

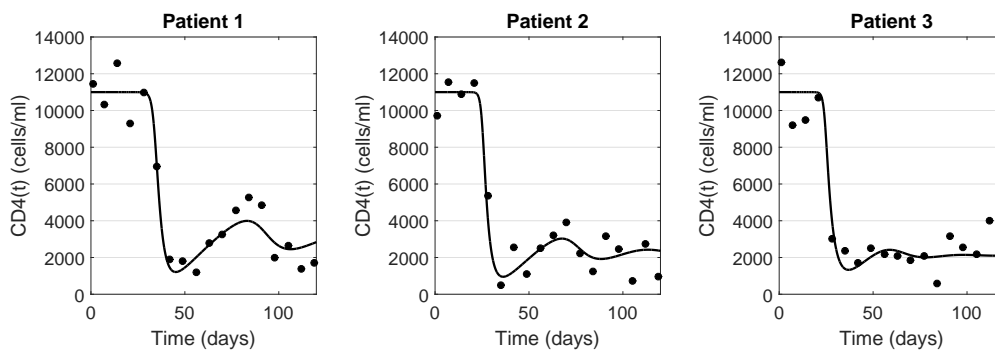
The estimation procedure described above is illustrated through a simulation study. In this work, regularly spaced  $CD4(t)$  and  $V(t)$  observations are obtained within the interval  $[0, 120]$ (days), by numerical Runge-Kutta integration of Equation (2.1). Note that simulating data for regularly spaced observations is not a limitation, as the model (2.1) can, in the same way, be applied to obtain non-equally spaced measurements. We reproduce the evolution of three HIV patients with parameters  $\theta_0$  presented in Table 2 [14]. Moreover, we considered the initial conditions  $(T_0, T_0^*, V_0) = (11 \times 10^3, 0, 10^{-6})$  with units  $(\frac{cells}{ml}, \frac{cells}{ml}, \frac{copies}{ml})$ , respectively, that mimics a condition with a large initial number of uninfected cells  $T_0$  and low values for the initial number of infected cells  $T_0^*$  and viral load  $V_0$ .

Patient	$d_1$	$k_1$	$\delta$	$\pi_1$	$c$
1	0.013	$0.46 \times 10^{-6}$	0.40	980	3
2	0.012	$0.75 \times 10^{-6}$	0.39	790	3
3	0.017	$0.80 \times 10^{-6}$	0.31	730	3

**Table 2:** Parameter values used for the simulation of 100 replicas for 3 patients [14], in a total of 300 simulations.  $\epsilon = 0$  is considered. Further description of the parameters can be found in Table 1.

Within this setting, we obtain a set of  $n = 18$  observations representing the temporal trajectory of each patient in a clinical follow-up every 7 days ( $t_i \in \{0, 7, 14, 21, 28, \dots, 119\}$ ,  $i = 1, 2, \dots, 18$  and  $t_1 = 0$  is the time instant of the first  $CD4^+$ T observation of the patient). Afterwards, 300 replicas (100 replicas for each patient) of that trajectory are randomly generated, by adding an error

to the CD4<sup>+</sup>T values, in accordance with the fact that laboratory CD4<sup>+</sup>T measurements have an error of about 20% of the measured value (i.e.  $e \sim N(0, \sigma_e^2)$ ) [15]. Note that the quadratic deviation (of the realizations) of  $e$  from zero is such that  $\sum_{i=1}^n e_{t_i}^2 = f(\theta_0) \approx \sigma_e^2(n - 1)$ , as  $\theta_0$  is the simulation reference. For each replica, we obtain  $\hat{\theta}_0$  as the solution of the optimization problem. For the purpose of illustration, Figure 2 shows one replica of each patient and highlights the similarities and differences between patients.



**Figure 2:**  $CD4(t)$  trajectory over time for the reference patients with the  $\theta$  parameters in Table 2. The circles represent the observations obtained for one replica of that patient [13].

After infection with HIV, there is an acute phase characterized by an accentuated decay of the number of CD4<sup>+</sup>T cells, since they are HIV preferred target. This can be observed, in the graphs, between 30 and 40 days approximately. The immune system tries to fight the virus by producing antibodies. After this phase, the chronic phase of infection starts, defined by the body recovery. It is observed a slight increase in the number of CD4<sup>+</sup>T cells. This feature is shared for the three patients although the minimum and the maximum values of the CD4<sup>+</sup>T cells vary between patients, before the CD4<sup>+</sup>T cells reach an almost constant value. Biologically, since the CD4<sup>+</sup>T cells play a key role in the immune response to pathogens, the differences in those values (namely, the minima) may induce the development of more severe infections, e.g., certain types of cancers and non-AIDS diseases.

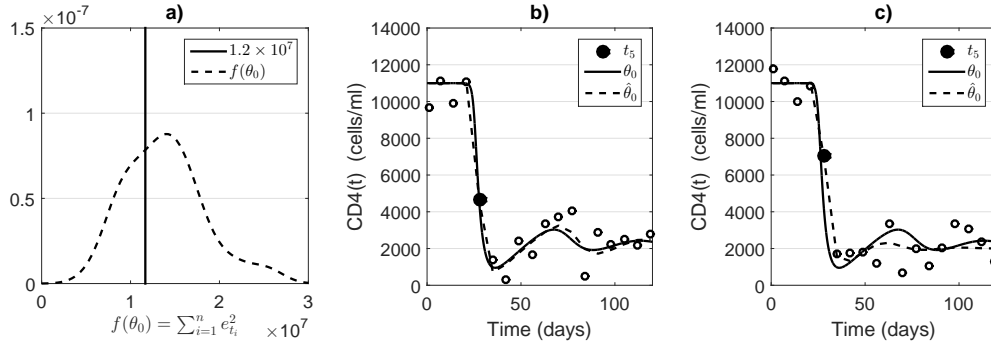
---

#### 4. RESULTS

---

In this section, the results are detailed for the simulations produced for patient 2 (see corresponding set of reference parameters  $\theta_0$  in Table 2). The performance evaluation of the model with respect to simulated data was assessed by  $f(\theta)$  either appraised for  $\theta_0$  (the reference simulation parameters) or  $\hat{\theta}_0$  (the parameters estimated from simulated data). The function  $f(\theta)$  translates the goodness-of-fit of the model-based observations with respect to simulated data

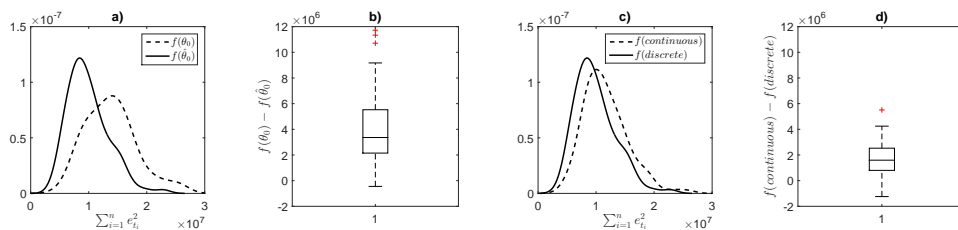
(see equation 2.2). Figure 3(a) shows the distribution of  $f(\boldsymbol{\theta}_0)$  for the 100 replicas of patient 2, where it is possible to observe that the values obtained for all replicas are centered around that evaluated for the reference parameters in  $\boldsymbol{\theta}_0$ . Figures 3(b–c) display the CD4 trajectory lines obtained from  $\boldsymbol{\theta}_0$  and from  $\hat{\boldsymbol{\theta}}_0$  for two replicas with  $f(\boldsymbol{\theta}_0)$  close to and higher than the reference value  $f(\hat{\boldsymbol{\theta}}_0)$ , respectively. As is illustrated in Figure 3(b), the estimation procedure provided similar curves for  $f(\boldsymbol{\theta}_0)$  close to  $f(\hat{\boldsymbol{\theta}}_0)$ . Moreover, as presented in Figure 3(c) for a replica with  $f(\boldsymbol{\theta}_0)$  higher than  $f(\hat{\boldsymbol{\theta}}_0)$  ( $f(\boldsymbol{\theta}_0) = 1.6 \times 10^7$  and  $f(\hat{\boldsymbol{\theta}}_0) = 1.0 \times 10^7$ ), there is a relevant improvement of fit from  $\boldsymbol{\theta}_0$  to  $\hat{\boldsymbol{\theta}}_0$ , as  $\hat{\boldsymbol{\theta}}_0$  produces a curve which is clearly more adjusted to the simulated data than that obtained with  $\boldsymbol{\theta}_0$ . Figure 3(c) also suggests that the observations do not contribute equally to the performance increase e.g. residuals at high derivative values (black dots) are increased for  $f(\boldsymbol{\theta}_0)$  and reduced when  $\boldsymbol{\theta}_0$  is replaced by  $\hat{\boldsymbol{\theta}}_0$ .



**Figure 3:** (a) Distribution of  $f(\boldsymbol{\theta}_0)$  evaluated for 100 replicas of patient 2 (i.e.  $\mathbf{s}_e^2(n-1)$  where  $\hat{e}$  are the residuals of the model with parameters  $\boldsymbol{\theta}_0$ , for each replica). The vertical line locates  $\sigma_e^2(n-1) = 1.2 \times 10^7$  used in the simulation. (b–c) CD4 trajectory line from  $\boldsymbol{\theta}_0$  and optimized  $\hat{\boldsymbol{\theta}}_0$  for two different replicas: (b)  $f(\boldsymbol{\theta}_0) = f(\hat{\boldsymbol{\theta}}_0) = 1.2 \times 10^7$  and (c)  $f(\boldsymbol{\theta}_0) = 1.6 \times 10^7$  and  $f(\hat{\boldsymbol{\theta}}_0) = 1.0 \times 10^7$ . The circles represent the simulated observations and the black dot highlights time  $t_5$ .

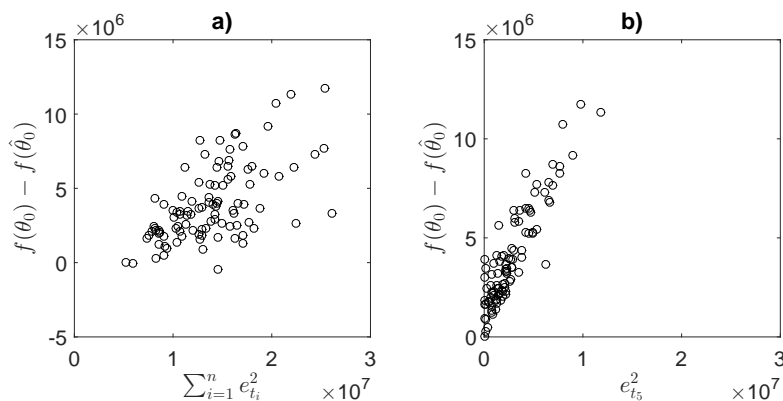
Figure 4 further compares the modeling results for the replicas for patient 2. As observed in Figures 4(a–b), the distribution of  $f(\hat{\boldsymbol{\theta}}_0)$  is more shifted towards the small deviations than  $f(\boldsymbol{\theta}_0)$  and  $f(\boldsymbol{\theta}_0) - f(\hat{\boldsymbol{\theta}}_0)$  is positive for almost all replicas, thus evidencing that lower squared errors are achieved for  $\hat{\boldsymbol{\theta}}_0$ . Moreover, as illustrated in Figures 4(c–d), the  $f(\boldsymbol{\theta}_0) - f(\hat{\boldsymbol{\theta}}_0)$  differences become higher than those obtained by choosing continuous/discrete time option for the model numerical resolution. This suggests that differences between  $f(\boldsymbol{\theta}_0)$  and  $f(\hat{\boldsymbol{\theta}}_0)$  are indeed relevant.

The result illustrated in Figure 3(c) suggested that the observations do not contribute equally to the performance increase with special emphasis on high derivative CD4 values. Figure 5(a) shows the association between performance increase of  $\hat{\boldsymbol{\theta}}_0$  with respect to  $\boldsymbol{\theta}_0$ , as measured by  $f(\boldsymbol{\theta}_0) - f(\hat{\boldsymbol{\theta}}_0)$ , and the dispersion



**Figure 4:** (a) Distribution of  $f(\theta_0)$  and  $f(\hat{\theta}_0)$  for 100 replicas of patient 2. (b) Boxplot of the paired differences  $f(\theta_0) - f(\hat{\theta}_0)$ . (c-d) Same representation as (a-b) for  $f(\hat{\theta}_0)$  and continuous/discrete time.

of the residuals introduced in the simulation process. The correlation turns out to be moderate for this patient ( $r = 0.60$ ). The effect of the residual at each time  $t_i$  was further investigated, by computing the correlation between  $f(\theta_0) - f(\hat{\theta}_0)$  and the squared residual value at time  $t_i$ . Figure 5(b) shows a high correlation between  $e_{t_5}^2$  and performance increase ( $r = 0.91$ ), where higher  $e_{t_5}^2$  values are associated with higher performance improvement. Furthermore, note that the large part of the residuals dispersion is due to the contribution of  $e_{t_5}$ . This analysis corroborates that the observations do not contribute equally to the performance increase. In this case,  $t_5$  corresponds to the time point with the largest residual values for  $\theta_0$  and highest derivate in the CD4 curve (Figures 3(b-c)).

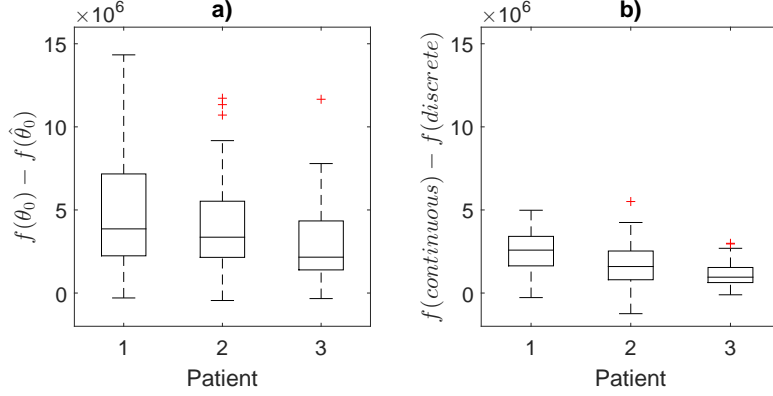


**Figure 5:** Dispersion diagram of  $f(\theta_0) - f(\hat{\theta}_0)$  as a function of **a)**  $\sum_{i=1}^n e_{t_i}^2$  and **b)**  $e_{t_5}^2$ , the time that maximizes correlation between  $f(\theta_0) - f(\hat{\theta}_0)$  and  $t_i, i = 1, 2, \dots, n$ . For the remaining time points the absolute correlation was  $< 0.20$ . Each dot represents one of the 100 replicas for patient 2.

An overall comparison of the 100 replicas simulated for the 3 patients (in a total of 300 replicas) is presented in Figure 6. As observed in Figure 6(a), the distribution of  $f(\hat{\theta}_0)$  is more shifted towards the small deviations than  $f(\theta_0)$  for all patients such that  $f(\theta_0) - f(\hat{\theta}_0)$  is positive for almost all replicas. Again, that lower squared errors are achieved for  $\hat{\theta}_0$  for all patients. Moreover, as illustrated

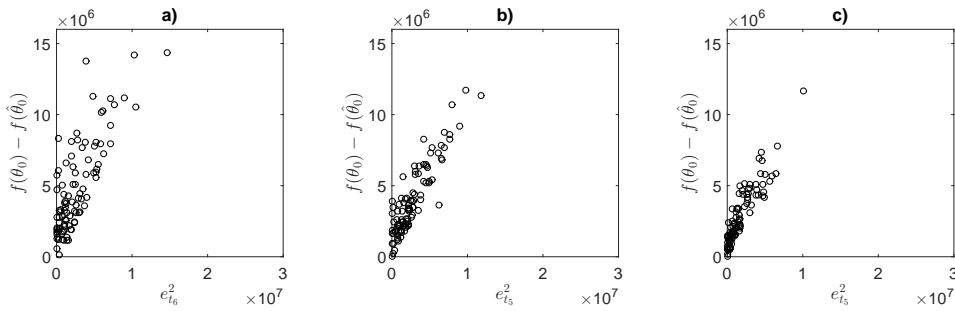


in Figure 6(b), the differences  $f(\theta_0) - f(\hat{\theta}_0)$  are higher than those obtained by choosing continuous/discrete time option on the model numerical resolution. This suggests that differences between  $f(\theta_0)$  and  $f(\hat{\theta}_0)$  are indeed relevant regardless of the simulated patient.



**Figure 6:** Boxplots of the paired differences (a)  $f(\theta_0) - f(\hat{\theta}_0)$  (b)  $f(\hat{\theta}_0)$  and continuous/discrete time, for each patient.

Finally, the contribution of the different observations to the performance increase is shown in Figure 7 for the 3 patients. Again, the results point out that the correlation between  $f(\theta_0) - f(\hat{\theta}_0)$  and the squared residual value at a given time  $t_i$  is highest for the time point with the largest residual values for  $\theta_0$  and highest derivate in the CD4 curve. The maximum correlation between these variables reach 0.86 for patient 1 and 0.90 for patient 3. For the remaining time points, the absolute correlation is lower than 0.2 for all patients.



**Figure 7:** Dispersion diagram of  $f(\theta_0) - f(\hat{\theta}_0)$  as a function of the squared error evaluated for the time that maximizes correlation between these variables. Data for patient (a) 1, (b) 2 and (c) 3.

In this work, the methods were also applied to a real data set of CD4<sup>+</sup>T cells count from six HIV patients [12]. The patients involved in the trial were chosen according to some conditions, namely being infected either by HIV-1 or HIV-2 type virus, being naive of any treatment at the beginning of the trial, not having hepatitis B or C virus co-infections during the 6 months before the inclu-

sion into the trial and having started an antiretroviral treatment at the beginning of the trial (therefore, we considered  $\epsilon > 0$  in equation 2.1). The follow-up of the CD4<sup>+</sup>T cells data for all patients can be found in the study of Rivadeneira et al. [12]. Table 3 presents the follow-up of Patient 6.

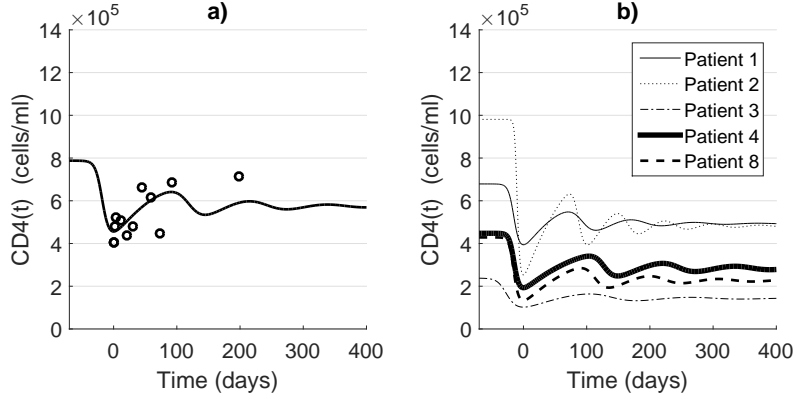
Day	0	1	2	4	11	21	31	44	59	74	92	199
CD4	405	403	480	522	510	436	479	661	615	445	686	716

**Table 3:** CD4<sup>+</sup>T cells count ( $\frac{cells}{mm^3}$ ) per time of observation (days) for Patient 6 [12]. The data for the remaining patients used in this work correspond to Patients 1, 2, 3, 4 and 8 and can be obtained in the paper of Rivadeneira and colleagues [12].

Since these patients start an antiretroviral treatment at the beginning of the trial, it is necessary to estimate the time of the infection. To this end, we start the optimization from the minimum value of  $\widehat{CD4}(t_i) = T(t_i) + T^*(t_i)$ , where  $\widehat{CD4}(t_i)$  is the estimate of  $CD4(t_i)$  provided by the mathematical model (2.1) at time  $t_i, i = 1, 2, \dots, n$ . Moreover, we also need to estimate the initial value  $T_0$  for that specific patient. This can be done by estimating additionally the parameter  $T_0$  (besides the vector of parameters  $\hat{\theta}$ ) with lower and upper bounds given by  $lb = 100$  and  $ub = 1200000$ , respectively [3].

Figure 8(a) shows the estimated trajectory for Patient 6, with units'  $\frac{cells}{ml}$  ( $\frac{cells}{mm^3} = \frac{cells}{ml} \times 10^3$ ), obtained from the optimal estimates  $\hat{\theta}_6 = (0.013, 0.908, 3.2 \times 10^{-9}, 0.693, 9999.999, 2.390)$  and initial number of uninfected cells  $\widehat{T}_0^6 = 787728 \frac{cells}{ml}$ . The results indicate that the effectiveness of therapy is approximately 91% ( $\hat{\epsilon}_6 = 0.908$ ) and that the infected cells die at a rate of 0.693 per day ( $\hat{\delta}_6 = 0.693$ ). Also, the analysis of the curve suggests that Patient 6 was infected around 68 days before being included in the trial ( $t = 0$ ). Figure 8 (b) resumes the results obtained for Patients 1, 2, 3, 4 and 8 [12] and illustrates the inter-subject variability of HIV individual patterns before and during antiretroviral treatment. The effectiveness of therapy  $\hat{\epsilon}$  is above 90% for all patients whereas the daily death rate of infected cells  $\hat{\delta}$  varies between 0.24 and 0.69. The analysis of the curves provides an estimate time of infection of approximately 28 days for Patient 2, 49 days for Patients 4 and 8 and 55 days for the remaining patients. Finally,  $\widehat{T}_0$  varies between 237902 (Patient 3) and 980902 (Patient 2). In all cases,  $f(\hat{\theta})$  varies between  $0.001 \times 10^7$  and  $0.021 \times 10^7$  which is lower than that observed in the simulated data. Such result is expected because in real data there are no points of very large residuals like in the simulation condition (e.g.  $t_5$  or  $t_6$  in the simulation condition, depending of the patient). Therefore, the results suggest that methods' performance in real data is not worse than that in simulated replicas of the same patient. This is an important result because the simulated data is drawn from the mathematical model, on the contrary of the real data, and thus good performances in terms of goodness-of-fit are expected for the curves estimated from the simulated data. Note that the order of magnitude of Figure 8 is different from that of Figure 2, since CD4 values vary between 100 and 1200000 [3]. Thus, we conclude that the patients of the simulation are worse

off than the patients in the study of Rivadeneira et al. [12].



**Figure 8:** (a)  $CD4(t)$  trajectory over time for Patient 6 with the  $\hat{\theta}_6$  parameters. The circles represent the observations for Patient 6 in Table 3. (b) Same representation for Patients 1, 2, 3, 4 and 8 [12].

---

## 5. CONCLUSION

---

This work addresses the problem of estimating the parameters of a HIV dynamic model from a set of observations. Our method considers the minimization of the square error between model estimates and observed CD4 values, with a restriction that guarantees that the optimal solution  $\hat{\theta}_0$  verifies equal contribution of negative and positive deviations from observations. Furthermore, the  $\hat{\theta}_0$  estimates are restricted to lower and upper physiological bounds, which allows us to obtain a fully automatic method, in which it is not necessary to introduce an initial condition. The proposed method is validated via a data simulated with reference parameters  $\theta_0$  to mimic 3 different patients. The results indicate that the replacement of  $\theta_0$  by  $\hat{\theta}_0$  decreases the fit error in a value that is greater than the difference between the fit errors obtained in the continuous and in the discrete options on the model numerical resolution. Therefore, the performance increase when replacing  $\theta_0$  by  $\hat{\theta}_0$  is numerically relevant. Finally, the algorithm provides adequate  $\hat{\theta}_0$  estimates (i.e. with low fit error to simulated and to real data), which enables a proper characterization of the temporal trajectory of a HIV patient.

---

## ACKNOWLEDGMENTS

---

This work was partially funded by the Foundation for Science and Technology, FCT, through national (MEC) and European structural (FEDER) funds,

in the scope of UID/MAT/04106/2019 (CIDMA/UA), UID/CEC/00127/2019 (IEETA/UA), UID/Multi/04621/2019 (CEMAT/UL) and UID/MAT/00144/2019 (CMUP/UP) projects. Diana Rocha acknowledges the FCT grant with reference SFRH/BD/107889/2015.

The revision of this paper has been overshadowed by Dr. Caldeira's departure on January 2019. As the Infectious Disease Service Director, he always welcomed our scientific collaboration with great enthusiasm. In sorrow, we dedicate this work to his memory.

---

## REFERENCES

---

- [1] BOFILL, M.; JANOSSY, G.; LEE, C.A.; MACDONALD-BURNS, D.; PHILLIPS, A.N.; SABIN C.; TIMMS A.; JOHNSON, M.A. and KERNOFF, P.B. (1992). Laboratory control values for CD4 and CD8 T lymphocytes. Implications for HIV-1 diagnosis, *Clinical & Experimental Immunology*, **88**, 2, 243–252.
- [2] BONHOEFFER, S.; MAY, R.M.; SHAW, G.M. and NOWAK, M.A. (1997). Virus dynamics and drug therapy, *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 6971–6976.
- [3] CONWAY, J.M. and PERELSON, A.S. (2015). Post-treatment control of HIV infection, *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 17, 5467–5472.
- [4] DORMAND, J.R. and PRINCE, P.J. (1980). A family of embedded Runge-Kutta formulae, *Journal of Computational and Applied Mathematics*, **6**, 19–26.
- [5] HADJIANDREOU, M.M.; CONEJEROS, R. and WILSON, D.I. (2009). Long-term HIV dynamics subject to continuous therapy and structured treatment interruptions, *Chemical Engineering Science*, **64**, 1600–1617.
- [6] HUANG, Y.; WU, H. and ACOSTA, E.P. (2010). Hierarchical Bayesian inference for HIV dynamic differential equation models incorporating multiple treatment factors, *Biometrical Journal*, **52**, 470–486.
- [7] LIANG, H.; MIAO, H. and WU, H. (2010). Estimation of constant and time-varying dynamic parameters of HIV infection in a nonlinear differential equation model, *The Annals of Applied Statistics*, **4**, 460–483.
- [8] LUO, R., PIOVOSO, M.J., MARTINEZ-PICADO, J. and ZURAKOWSKI, R. (2012). HIV Model parameter estimates from interruption trial data including drug efficacy and reservoir dynamics, *PLoS ONE*, **7**, 7, e40198.
- [9] NOWAK, M.A. and MAY, R.M. (2000). *Virus Dynamics: Mathematical Principles of Immunology and Virology*, Oxford University Press, Oxford.
- [10] PERELSON, A.S. (2002). Modelling viral and immune system dynamics, *Nature Reviews Immunology*, **2**, 28–36.
- [11] PERELSON, A.S.; KIRSCHNER, D.E. and BOER, R. (1993). Dynamics of HIV infection of CD4<sup>+</sup>T cells, *Mathematical Biosciences*, **114**, 81–125.

- [12] RIVADENEIRA, P.S.; MOOG, C.H.; STAN, G.B.; COSTANZA, V.; BRUNET, C.; RAFFI, F.; FERRÉ, V.; MHAWEJ, M.J.; BIAFORE, F.; OUATTARA, D.A.; ERNST, D.; FONTENEAU, R. and XIA, X. (2014). Mathematical modeling of HIV dynamics after antiretroviral therapy initiation: a clinical research study, *AIDS Research and Human Retroviruses*, **30**, 831–834
- [13] ROCHA, D.; SCOTTO, M.; PINTO, C.; TAVARES, J. and GOUVEIA, S. (2019). Simulation study of HIV temporal patterns using Bayesian methodology, *Springer Proceedings in Mathematics & Statistics*, Springer (Ed.) (to appear).
- [14] STAFFORD, M.A., COREYA, L., CAO, Y., DAARDD, E.S., HOB, D.D. and PERELSON, A.S. (2000). Modeling plasma virus concentration during primary HIV infection, *Journal of Theoretical Biology*, **203**, 285–301.
- [15] WHITBY, L.; WHITBY, A.; FLETCHER, M.; HELBERT, M.; REILLY, J.T. and BARNETT, D. (2013). Comparison of methodological data measurement limits in CD4<sup>+</sup>T lymphocyte flow cytometric enumeration and their clinical impact on HIV management, *Cytometry Part B (Clinical Cytometry)*, **84**, 4, 248–254.

---

---

## ACCURACY MEASURES FOR BINARY CLASSIFICATION BASED ON A QUANTITATIVE VARIABLE

---

---

- Authors: RUI SANTOS  
– School of Technology and Management, Polytechnic Institute of Leiria  
CEAUL – Centre of Statistics and Applications, Portugal  
rui.santos@ipleiria.pt
- MIGUEL FELGUEIRAS  
– School of Technology and Management, Polytechnic Institute of Leiria  
CEAUL – Centre of Statistics and Applications  
Centre of Applied Research in Management and Economics, Portugal  
mfelg@ipleiria.pt
- JOÃO PAULO MARTINS  
– School of Technology and Management, Polytechnic Institute of Leiria  
CEAUL – Centre of Statistics and Applications, Portugal  
jpmartins@ipleiria.pt
- LILIANA FERREIRA  
– School of Technology and Management, Polytechnic Institute of Leiria  
CMAFCIO – Centre for Mathematics, Fundamental Applications and  
Operations Research, Portugal  
liliana.ferreira@ipleiria.pt

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

### Abstract:

- The identification of the right methodology to perform binary classification based on an observed quantitative variable is usually a complex choice. Thus, the use of appropriate accuracy measures is crucial. In fact, the ROC curve reveals a lot of information about the accuracy of the applied methodology for all the possible values of the cut-point. In particular, the integral and partial areas under the ROC curve are widely used. The  $\phi$  index, in which sensitivity equals specificity, may also be applied. Nevertheless, the accuracy at one specific cut-point may be sufficient to assess the accuracy in some applications. Therefore, different ways to define the optimal cut-point may be applied, such as the maximization of the Youden index, the maximization of the concordance probability or the minimization of the distance to the point with absence of misclassification. To compare the adequacy of these measures, a simulation study was performed under different scenarios. The results highlight the advantages and disadvantages of each procedure and advise the use of the  $\phi$  index.

### Key-Words:

- *Binary classification; cut-point; ROC curve; sensitivity; specificity; simulation.*

AMS Subject Classification:

- 62P10, 92D30.

---

## 1. INTRODUCTION

---

Assume that an infection with prevalence rate  $p$  is affecting a population with  $N$  individuals. Let  $X_i$ , with  $i = 1, \dots, N$ , be  $N$  independent Bernoulli trials ( $X_i \sim \text{Ber}(p)$ ) with probability  $p$ , where the random variable (r.v.)  $X_i$  denotes the presence ( $X_i = 1$ ) or the absence ( $X_i = 0$ ) of the infection in the  $i$ -th individual. In addition, let  $Y_i$  represents the value of a diagnostic test performed by the  $i$ -th individual, characterized by the distribution  $D_0$  with parameter vector  $\theta_0$  if  $X_i = 0$  and by the distribution  $D_1$  with parameter vector  $\theta_1$  if  $X_i = 1$ , for  $i = 1, \dots, N$ . Finally, let  $t$  be the cut-point of the binary classification (healthy versus infected) based on the observation of the r.v.  $Y_i$ . Under these conditions we can define the following classification rule:

- If  $Y_i \leq t \Rightarrow X_i^-$  (a negative result, i.e. the individual is classified as healthy);
- If  $Y_i > t \Rightarrow X_i^+$  (a positive result, i.e. the individual is classified as infected).

As a matter of fact, the opposite inequalities can also be applied. Nevertheless, the reasoning is exactly the same and, therefore, we will restrict this presentation to the previously described situation.

The intention is to perform a diagnostic test to achieve a binary classification (e.g. healthy versus infected) based on the observed value of the quantitative variable  $Y_i$ . Nonetheless, almost all tests may result in misclassification due the occurrence of false negative or false positive results. Thus, it is essential to assess the performance of the applied binary classification procedure. The most common measure to evaluate the performance is the area under the Receiver Operating Characteristic (ROC) curve (AUC) [32], but it evaluates all possible cut-points, even those that are clinically unsuitable [7]. The partial AUC (pAUC) has been attracting the attention in medical issues [1, 2, 10] as well as in decision making and machine learning applications [16, 17] since it focus on a suitable range of interest for the true positive (or negative) rate [14]. Nevertheless, the partial AUC has some limitations in the application on ROC curves that cross the diagonal line, which are quite frequent in practice. Thus, there are still some contraindications for its widespread, regardless of some new proposals to overcome this problem (e.g., [30, 33]). And, to the best of our knowledge, there is no simulation study that allows to identify the existence of regions in which the computation of pAUC is suitable even in those cases. Moreover, despite its advantages over AUC, the pAUC continues to be unknown to many who apply binary classification procedures based on a quantitative variable. Hence, the main goal of this paper is to compare the usual measures of accuracy in binary classification in order to identify the most appropriate, completing the works already presented in [26, 25].

The main accuracy measures for binary classification based on a quantitative variable are presented in Section 2. In Section 3, a simulation study is



performed in order to compare those measures under different scenarios. All results were computed by the  $\text{\textcircled{R}}$  software using different distributions as well as diverse sample sizes. Finally, the main conclusions are outlined in Section 4.

---

## 2. DIAGNOSTIC ACCURACY MEASURES

---

The usual accuracy measures for classification can be computed for each possible value for the cut-point  $t$ , namely the specificity  $\varphi_e$  (or true negative fraction) that corresponds to the probability of obtaining a negative result in a healthy individual, i.e.

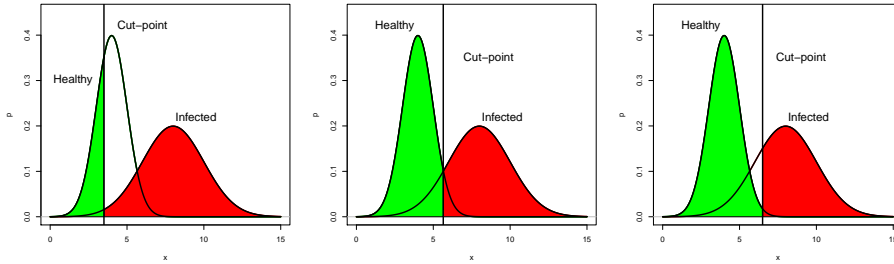
$$P(X_i^- | X_i = 0) = P(Y_i \leq t | X_i = 0) = F_{D_0}(t),$$

where  $F_D$  denotes the distribution function of the distribution  $D$ . Similarly, the sensitivity  $\varphi_s$  (true positive fraction) corresponds to the probability of getting a positive result in an infected individual, i.e.

$$P(X_i^+ | X_i = 1) = P(Y_i > t | X_i = 1) = 1 - F_{D_1}(t) = \bar{F}_{D_1}(t),$$

where  $\bar{F}_D$  denotes the survival function of the distribution  $D$ .

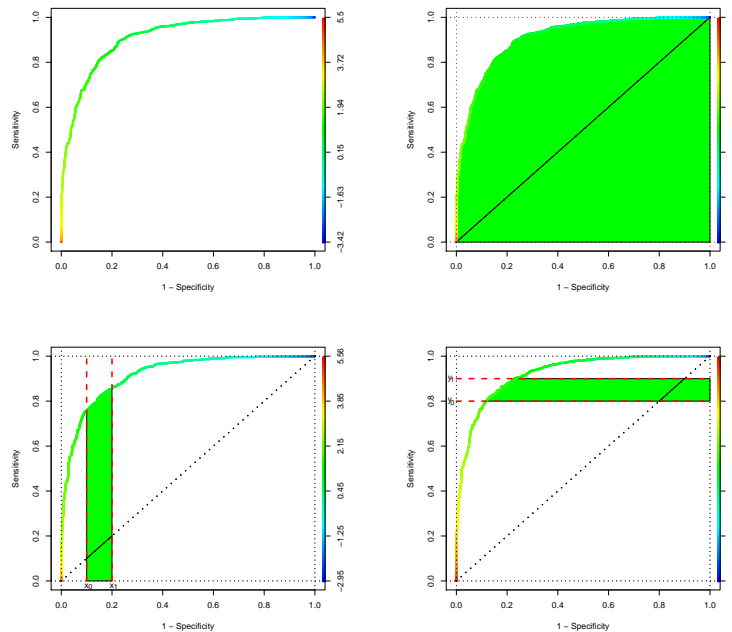
Note that the probabilities  $\varphi_s$  and  $\varphi_e$  depend on the value  $t$  considered for the cut-point and are inversely correlated, since increasing one of them implies decreasing the other when the same classification test is performed. Figure 1 uses the densities of the healthy (distribution  $D_0$ ) and of the infected (distribution  $D_1$ ) individuals to emphasize the changes in the sensitivity and specificity when different values for the cut-point are applied. The three graphs show the decreasing sensitivity (shaded area represented on the right of the cut-point) and the increasing specificity (shaded area represented on the left of the cut-point) as the value of the cut-point increases.



**Figure 1:** Sensitivity versus specificity on the use of different cut-points.

## 2.1. The receiver operating characteristic (ROC) curve

The Receiver Operating Characteristic (ROC) curve allows to visualize the evolution of  $\varphi_s$  and  $\varphi_e$  when the cut-point goes through all possible values, from the point in which all individuals are classified as infected to the other extreme where all individuals are classified as healthy. Therefore, this curve reveals all pairs  $(1 - \varphi_e, \varphi_s)$  which are also usually denoted by  $(x, \text{ROC}(x))$ . For this reason, the ROC curve is often used to identify the optimal cut-point of a binary classification methodology, as well as to compare the performance of different methodologies [5, 6, 9, 15, 18, 32, 35]. The first graph of Figure 2 displays an example of a ROC curve.



**Figure 2:** The ROC curve (top left) and the integral (top right) and partial (bottom) areas under the curve.

All ROC curves start in the point  $(0, 0)$  where all individuals are classified as healthy, and therefore  $\varphi_e = 1$  and  $\varphi_s = 0$ ; and finish on the opposite situation, i.e. where all individuals are classified as infected,  $\varphi_e = 0$  and  $\varphi_s = 1$ . The segment  $1 - \varphi_e = \varphi_s$  connecting these two points represents a random classification without using the information of  $Y_i$ , where the probability of classifying any individual as infected is equal to  $\varphi_s$ . Note that the accuracy in any point below this segment would increase if the classification of all individuals were simply changed. The other two vertices of the ROC plane correspond to the remaining extreme cases, the ideal point  $(0, 1)$  with absence of misclassification  $\varphi_e = \varphi_s = 1$ ; and the point  $(1, 0)$  in which every individual is misclassified  $\varphi_e = \varphi_s = 0$ .

---

### 2.1.1. The entire area under the ROC curve — AUC

---

The most widely used measure of accuracy is the area under the ROC curve (AUC). The second graph of Figure 2 shows the integral area under the ROC curve. It represents the mean value of  $\varphi_s$  for all possible values of  $\varphi_e$ . It can also be interpreted as the probability of correctly classifying a pair when the r.v.  $Y_i$  is continuous, where 0.5 means unreliability, as in a random classification, and 1 corresponds to the perfect classification (absence of misclassification). The value of the area is also related to the Wilcoxon-Mann-Whitney statistic, allowing to make inference about the ROC curve [18, 35]. The AUC is possibly the most commonly used measure to assess the diagnostic accuracy of a binary classification methodology [6, 18, 32]. However, this measure takes into account all possible values for the cut-point, even those that are unsuitable in practice because it generates very low specificity or sensitivity levels. This is the main drawback of this measure, although it summarizes the entire ROC curve it includes values which are not clinically relevant. In fact, these values should be neglected, otherwise they may interfere in the choice of the best methodology. Moreover, usually only a specific cut-point is applied.

---

### 2.1.2. The standardized partial area under the ROC curve — spAUC

---

The partial area under the ROC curve (pAUC) can be used to evaluate the performance at the interest cut-point values, for which the methodology performs satisfactorily [3, 8, 14, 13, 31, 35]. These values usually correspond to high specificity values, but can also be applied to high sensitivity values. The pAUC over the high specificity range  $[1 - x_1, 1 - x_0]$  can be defined as

$$\text{pAUC}(x_0, x_1) = \int_{x_0}^{x_1} \text{ROC}(x) \, dx,$$

which corresponds to the area of the shaded region in the bottom left chart of Figure 2. It analyses the  $\varphi_s$  when we fix the  $\varphi_e$  in a range of interest. However, in some applications the goal is to evaluate the  $\varphi_e$  when the  $\varphi_s$  is significant. In this cases the area is on the right of the ROC curve (see bottom right graph of Figure 2). Thus, we can compute the pAUC over the high sensitive range  $[y_0, y_1]$  using

$$\text{pAUC}(y_0, y_1) = \int_{y_0}^{y_1} 1 - \text{ROC}^{-1}(y) \, dy,$$

where  $\text{ROC}^{-1}$  denotes the generalized inverse function of the function ROC. The  $\text{pAUC}(y_0, y_1)$  corresponds to the area of the shaded region in the fourth chart of Figure 2. This latter case does not correspond properly to the area below the curve and perhaps the most appropriate designation would be the area on the right of the curve instead of the area under the curve.

In both cases, the pAUC verifies  $\text{pAUC}(0, 1) = \text{AUC}$  and

$$\frac{1}{2}(x_1^2 - x_0^2) \leq \text{pAUC}(x_0, x_1) \leq x_1 - x_0, \quad 0 \leq x_0 \leq x_1 \leq 1.$$

Hence, in order to be interpreted analogously to AUC, pAUC can be standardized by

$$\text{spAUC}(x_0, x_1) = \frac{1}{2} \left( 1 + \frac{\text{pAUC}(x_0, x_1) - \frac{1}{2}(x_1^2 - x_0^2)}{x_1 - x_0 - \frac{1}{2}(x_1^2 - x_0^2)} \right).$$

Thus, spAUC varies between 0.5 (random classification) and 1 (absence of misclassification). Nevertheless, the use of pAUC or spAUC requires the definition of the range of interest  $[x_0, x_1]$  or  $[y_0, y_1]$ . Usually  $[x_0, x_1]$  corresponds to  $[0, x_1]$ , i.e. the highest values for specificity, and spAUC can be seen as (approximately) the average of  $\varphi_s$  when specificity ranges in  $[1 - x_1, 1]$ . Similarly,  $[y_0, y_1]$  commonly corresponds to the highest values for the sensitivity, i.e.  $[y_0, 1]$ , and spAUC can be seen as (approximately) the average of  $\varphi_e$  when sensitivity ranges in  $[y_0, 1]$ .

Note that the use of spAUC introduces some difficulties to solve issues related to the arbitrariness of choosing the range of interest. Furthermore, some authors highlight the loss of information, claiming a loss of statistical precision as compared with inferences based on the entire AUC [7, 35].

## 2.2. The $\phi$ index

The use of the probability  $\phi$ , which verifies  $\varphi_s = \varphi_e = \phi$  for some cut-point, to measure the performance of diagnostic tests in the context of compound tests is advised in [23, 24]. In fact, it corresponds to the intersection of the ROC curve with the straight line  $\varphi_s = \varphi_e$ , as Figure 3 illustrates. If this value does not exist, as in the use of count distributions or small samples, the distance between  $\varphi_s$  and  $\varphi_e$  shall be minimized and  $\phi = \frac{\varphi_s + \varphi_e}{2}$ .

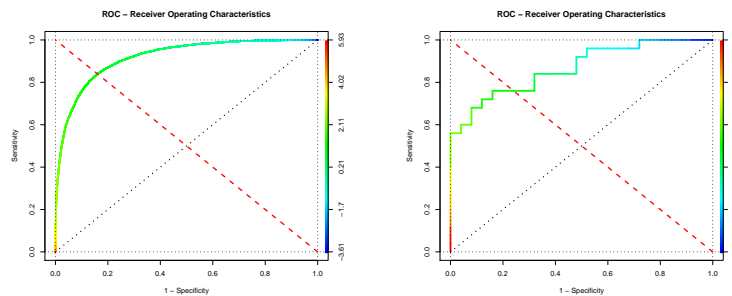


Figure 3: The  $\phi$  index.

In the simulations performed in Section 3, computations of spAUC over the range  $[\phi - 0.05, \min\{\phi + 0.05, 1\}]$  for both specificity and sensitivity are used.

The idea was to consider not only significant values for both measures but also to use a small range (the largest range is equal to 0.1).

---

### 2.3. The optimal cut-point

---

In practical issues only a single cut-point is usually applied. Thus, the knowledge of the accuracy at this specific point may be sufficient to assess the classification methodology. Therefore, the selection of the optimal cut-point is a complex decision that depends on several factors. For example, the severity of the infection and the risk of not diagnosing the infection may clearly encourage the choice of a high sensitivity and somehow neglect the specificity level. In the opposite direction, the side effects of the treatment and the treatment cost may favour the use of a high specificity and disregard the sensitivity level. Hence, it may be important to decide between sensitivity or specificity in the selection of the cut-point, because its determination implies a compromise between these two measures. Nonetheless, in the absence of clinical factors that lead to the choice of one of these measures over the other, some criterion of optimization can be applied to choose the optimal cut-point. In fact, there are several available methodologies in the literature to obtain the optimal cut-point value [4, 11, 19, 22, 29, 34, 36], such as the maximization of the Youden index, the minimization of the distance to the point with absence of misclassification and the maximization of the concordance probability.

---

#### 2.3.1. The Youden index — YI

---

One way of determining the cut-point is to choose the point that maximizes the Youden index (YI) defined by [4, 11, 20, 27, 34]

$$\text{YI} = \varphi_e + \varphi_s - 1 = F_{D_0}(t) - F_{D_1}(t).$$

Geometrically, it corresponds to the point on the ROC curve in which the vertical distance is greater from the line  $1 - \varphi_e = \varphi_s$ , i.e. the difference between  $\varphi_s$  and  $1 - \varphi_e$ , as the first chart of Figure 4 shows. It also corresponds to the point  $t$  which maximizes the sum  $\varphi_e + \varphi_s$  and, thus, maximizes the distance between  $F_{D_0}(t)$  (true positive rate) and  $F_{D_1}(t)$  (false positive rate).

---

#### 2.3.2. The closest-to-(0, 1) criteria — DI

---

As previously stated, the point (0, 1) corresponds to the perfect classification procedure where all individuals are well classified. Therefore, we intend to

be as close as possible to this situation. Hence, the minimization of the Euclidean distance to the ideal point  $(0, 1)$ , with  $\varphi_e = \varphi_s = 1$ , is another criteria to choose the best cut-point [11, 19, 27, 29], i.e. minimizing

$$D = \sqrt{(1 - \varphi_e)^2 + (1 - \varphi_s)^2} = \sqrt{F_{D_0}^2(t) + F_{D_1}^2(t)}.$$

The second chart of Figure 4 illustrates this procedure. However, in order to compare with the other measures, in the simulations performed in Section 3 it will be used the maximization of

$$DI = 1 - D = 1 - \sqrt{F_{D_0}^2(t) + F_{D_1}^2(t)},$$

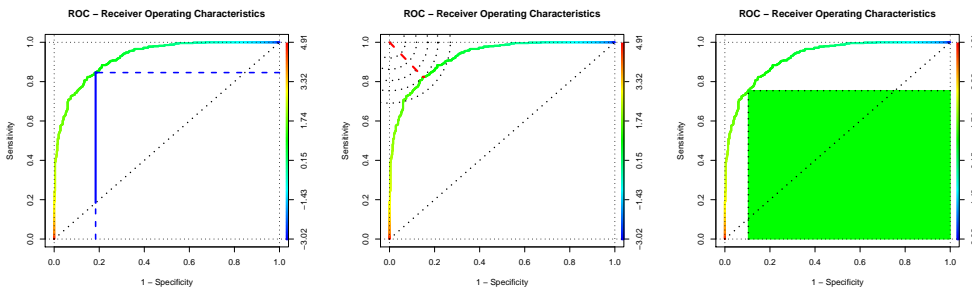
which corresponds to the minimization of  $D$  and provides the point on the ROC curve that is the closest to the ideal case  $(0, 1)$ . With this transformation all the measures to select the cut-point take values in the range  $[0, 1]$  and increase with the improvement of the accuracy of classification.

### 2.3.3. The concordance probability method — CP

When the r.v.  $Y_i$  is continuous, the AUC can be interpreted as the concordance probability. But, when  $Y_i$  is not continuous (discrete or ordinal) [11, 12] advocate the use of the concordance probability for a quantitative variable given by the product of sensitivity and specificity, i.e.

$$CP = \varphi_e \varphi_s = F_{D_0}(t) \overline{F}_{D_1}(t).$$

The maximization of the CP can be used to define the cut-point. The third chart of Figure 4 shows the area of the rectangle which corresponds to the CP value. Thus, covering all the points on the ROC curve as the upper left vertex of the rectangle, we intend to determine the rectangle with maximum area.




**Figure 4:** The optimal cut-point using the Youden index (left), the distance to the ideal point (center) and the concordance probability (right).

---

### 3. AN ACCURACY COMPARISON BY SIMULATION

---

In most cases, the focus in the application of a diagnostic test is the evaluation of the accuracy for a single cut-point, which shall be the best one for our purposes. Thus, it is indeed critical to compare the differences between the area AUC and the partial area spAUC under the ROC curve as well as the index  $\phi$ , and to realize if a greater value in these accuracy measures is sufficient to ensure a good accuracy in the selected cut-point, considering the cut-points obtained by the application of the three procedures provided in Subsection 2.3.

Hence, a simulation study was performed through the  software using the `ROCR` and `pROC` packages [21, 28]. All scenarios were analysed using  $10^3$  replicas and the following accuracy measures were computed:

- AUC – the entire area under the ROC curve;
- $SP_{90}$ ,  $SP_{75}$ ,  $SP_{50}$  – spAUC computed over the specificity range  $[0.9, 1]$ ,  $[0.75, 1]$  and  $[0.5, 1]$ , respectively;
- $SE_{90}$ ,  $SE_{75}$ ,  $SE_{50}$  – spAUC computed over the sensitivity range  $[0.9, 1]$ ,  $[0.75, 1]$  and  $[0.5, 1]$ , respectively;
- $\phi$  (or Phi) – the  $\phi$  index;
- $SP_{\phi}$  (or SPPhi) – spAUC computed over the specificity range  $[\phi - 0.05, \min\{\phi + 0.05, 1\}]$ ;
- $SE_{\phi}$  (or SEPhi) – spAUC computed over the sensitivity range  $[\phi - 0.05, \min\{\phi + 0.05, 1\}]$ ;
- YI – the maximum Youden index;
- DI – the maximum of 1-D where D denotes the distance to the ideal point  $(0, 1)$ ;
- CP – the maximum of the concordance probability.

In order to compare the obtained results in these measures, the Spearman's rank correlation coefficients were computed to assess monotonic relationships between them. Therefore, these correlations evaluate if the rank of the accuracy in each model is made in the same way using different measures. Note that all those measures vary in  $[0, 1]$  and increase with the improvement of the accuracy.

For the test design, diverse sample sizes were applied using equal number of infected and healthy individuals, i.e.  $n_0 = n_1 = n \in \{50, 100, 250, 500, 1000\}$ . The restriction  $n_0 = n_1$  only aims to achieve the same accuracy in the estimation of the sensitivity (only infected individuals are analysed) and specificity (only

healthy individuals are used). Besides, different distributions for the characterization of the infected and healthy individuals were considered in the simulations, both discrete and continuous. In order to simplify the presentation of the obtained results, we will restrict to the cases where the two subpopulations have the same distribution  $D_0 = D_1$  but with different values for the parameter vectors, i.e.  $\theta_0 \neq \theta_1$ . This restriction aims to simplify the interpretation of the results. Moreover, to minimize and simplify the discussion of the main conclusions, only the results obtained with some of the most applied distributions will be shown since they include the most usual shapes of ROC curves. In particular, the following distributions were used:

- Normal, with  $\mu_0 = 0$ ,  $\sigma_0 = 1$ ,  $\mu_1 = 2$  and  $\sigma_1 \in \{2/3, 1, 1.5, 2, 3\}$ ;
- Gamma, with  $\alpha_0 = 2$ ,  $\beta_0 = 1$ ,  $\alpha_1 \in \{6, 9, 12\}$  and  $\beta_1 \in \{1, 3\}$ ;
- Binomial, with  $p_0 = 0.25$  and  $p_1 \in \{0.3, 0.4, 0.5\}$ ;
- Geometric, with  $p_0 = 0.2$  and  $p_1 \in \{0.1, 0.02\}$ .

The main goal is to evaluate the association between those accuracy measures and, therefore, to assess whether those measures are able to evaluate the same criterion of accuracy.

---

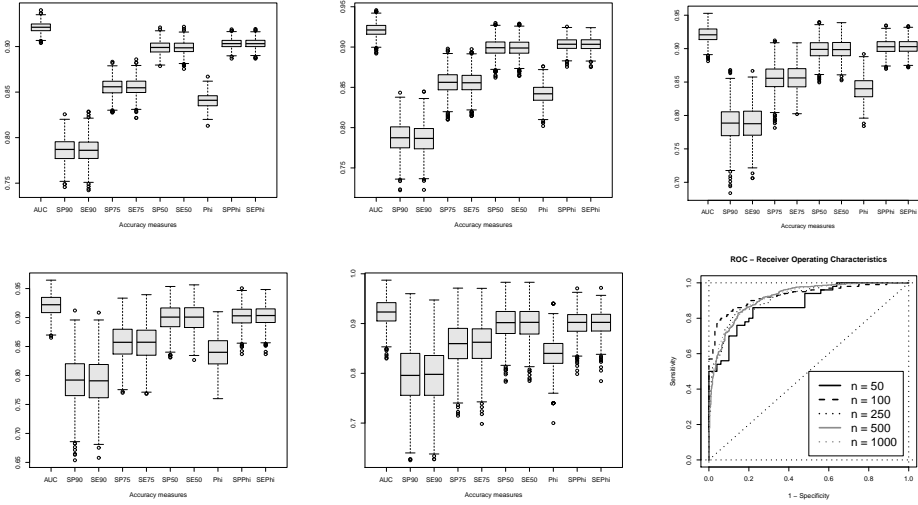
### 3.1. The sample dimension

---

Let us consider that the r.v.  $Y_i$  has Normal distribution with standard deviation  $\sigma = 1$  and mean  $\mu_0 = 0$  in a healthy individual and  $\mu_1 = 2$  in an infected individual. Figure 5 contains the boxplot for different sample sizes  $n \in \{50, 100, 250, 500, 1000\}$  of the applied diagnostic accuracy measures which do not depend on the cut-off value. As expected, the median seems to be always the same, but the range of variation and the interquartile range decrease with the increasing of the sample size. Besides, due to the symmetry of the ROC curves around the line  $\varphi_e = \varphi_s$ , the partial areas over the specificity have the same behaviour as the partial areas over the sensitivity, converging to the AUC when the range of interest increases. In the last chart some ROC curves obtained with different sample sizes are plotted to illustrate that the ROC curve becomes smoother as  $n$  increases.

Table 1 provides the Spearman's rank correlation coefficients between all the computed measures when the sample dimension is  $n = 1000$  (upper triangular matrix) and when the sample dimension is  $n = 50$  (lower triangular matrix). The results do not seem to have significant differences between the values obtained with  $n = 50$  and  $n = 1000$ . The correlation between the partial areas and the entire area seems to increase when the interval of interest increases and converges





**Figure 5:**  $D_0(\theta_0) = N(0, 1)$  versus  $D_1(\theta_1) = N(2, 1)$  with different sample dimensions:  $n = 1000$  (top left),  $n = 500$  (top middle),  $n = 250$  (top right),  $n = 100$  (bottom left),  $n = 50$  (bottom middle), and ROC curves (bottom right).

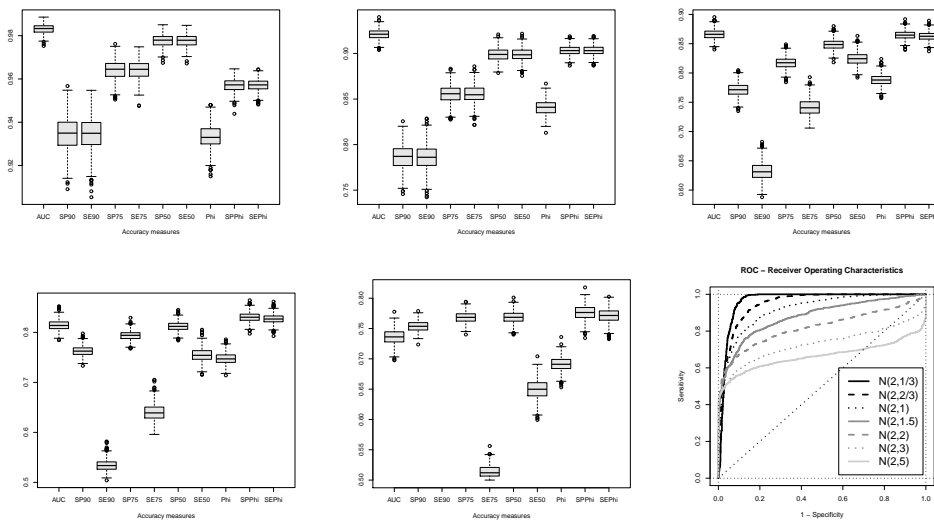
to all of the support  $[0, 1]$  as expected. The partial areas  $SP_\phi$  and  $SE_\phi$  exhibit significant correlation with AUC albeit having lower range in its computation. Moreover, the  $\phi$  index clearly reveals higher correlation with the measures YI, DI and CP used to set the best cut-point. Besides, the measures YI, DI and CP are strongly correlated with each other.

	AUC	SP <sub>90</sub>	SE <sub>90</sub>	SP <sub>75</sub>	SE <sub>75</sub>	SP <sub>50</sub>	SE <sub>50</sub>	$\phi$	SP $_\phi$	SE $_\phi$	YI	DI	CP
AUC	1	.784	.784	.928	.927	.988	.988	.815	.869	.864	.841	.828	.837
SP <sub>90</sub>	.791	1	.326	.912	.515	.818	.713	.550	.628	.575	.571	.560	.567
SE <sub>90</sub>	.780	.329	1	.516	.910	.712	.817	.561	.584	.628	.585	.572	.581
SP <sub>75</sub>	.935	.912	.531	1	.758	.959	.893	.810	.874	.841	.834	.823	.830
SE <sub>75</sub>	.925	.525	.915	.766	1	.892	.957	.818	.848	.872	.844	.831	.840
SP <sub>50</sub>	.991	.818	.718	.960	.896	1	.972	.831	.890	.880	.857	.844	.852
SE <sub>50</sub>	.989	.722	.817	.902	.957	.975	1	.834	.885	.885	.862	.848	.858
$\phi$	.823	.558	.592	.811	.830	.838	.842	1	.894	.890	.958	.977	.965
SP $_\phi$	.866	.668	.576	.870	.816	.884	.876	.810	1	.981	.948	.931	.944
SE $_\phi$	.862	.559	.690	.807	.881	.869	.885	.818	.882	1	.944	.927	.940
YI	.884	.634	.654	.865	.877	.896	.902	.913	.887	.890	1	.987	.998
DI	.863	.594	.629	.848	.868	.877	.884	.957	.888	.894	.975	1	.993
CP	.878	.619	.646	.860	.876	.891	.897	.937	.887	.891	.994	.989	1

**Table 1:** Spearman’s rank correlation coefficient, with  $n = 1000$  (upper triangular matrix) versus  $n = 50$  (lower triangular matrix).

**3.2. Normal distribution with different standard deviation**

Let us now consider that the r.v.  $Y_i$  has Normal distribution with standard deviation  $\sigma_0 = 1$  and mean  $\mu_0 = 0$  in a healthy individual and  $\mu_1 = 2$  in an infected individual. The standard deviation in an infected individual varies in  $\sigma_1 \in \{2/3, 1, 1.5, 2, 3\}$  and we are collecting samples with size  $n = 1000$ . Obviously, the accuracy will get worse with the increase of  $\sigma_1$ . For  $\sigma_1 \in \{2/3, 1\}$  the partial areas over the sensitivity are similar to the partial areas over the specificity (see Figure 6). Nevertheless, for  $\sigma_1 \in \{1.5, 2\}$  the boxplots are quite different and for  $\sigma_1 = 3$  the boxplot of  $SE_{90}$  is not even shown. Hence, this case reveals problems on the computation of the partial area over a range of high sensitivity. If we observe the last chart of Figure 6, for the worst plotted ROC curve the spAUC computed over the sensitivity range  $[0.9, 1]$  would be lower (even after standardization) than 0.5 and, therefore, it is even worse than the random classification. Consequently, this measure is not shown. Note, also, that the worst ROC curves are not symmetric around  $\varphi_e = \varphi_s$  and consequently the partial areas over the specificity have different behaviour comparing with the partial areas over the sensitivity. However, the partial areas over a neighbourhood of  $\phi$  do not seem to have any problems in assessing accuracy.



**Figure 6:**  $D_0(\theta_0) = N(0, 1)$  versus  $D_1(\theta_1) = N(2, 2/3)$  (top left),  $N(2, 1)$  (top middle),  $N(2, 1.5)$  (top right),  $N(2, 2)$  (bottom left),  $N(2, 3)$  (bottom middle), and ROC curves (bottom right), with  $n = 1000$ .

Table 2 displays the Spearman’s rank correlation coefficients between all the computed measures when the r.v.  $Y_i$  is characterized by  $N(0, 1)$  for a healthy individual and characterized by  $N(2, 2/3)$  (upper triangular matrix) and  $N(2, 2)$  (lower triangular matrix) for an infected individual. There seems to be some

differences between the results obtained in these two situations, but the main conclusions appear to be the same. The correlation between the partial areas and the entire area continues to increase when the interval of interest increases to the support  $[0, 1]$  and the  $\phi$  index continues to reveal quite strong correlations with the measures YI, DI and CP. Even though, when  $\sigma_1 = 2$  these correlations are lower.

	AUC	SP <sub>90</sub>	SE <sub>90</sub>	SP <sub>75</sub>	SE <sub>75</sub>	SP <sub>50</sub>	SE <sub>50</sub>	$\phi$	SP <sub><math>\phi</math></sub>	SE <sub><math>\phi</math></sub>	YI	DI	CP
AUC	1	.890	.876	.969	.970	.997	.997	.771	.616	.599	.798	.786	.796
SP <sub>90</sub>	.690	1	.597	.953	.801	.904	.878	.736	.715	.402	.703	.751	.761
SE <sub>90</sub>	.721	.289	1	.779	.942	.860	.888	.753	.359	.656	.776	.767	.775
SP <sub>75</sub>	.847	.906	.392	1	.919	.980	.963	.794	.673	.578	.824	.811	.823
SE <sub>75</sub>	.952	.484	.825	.655	1	.963	.980	.807	.573	.637	.835	.823	.833
SP <sub>50</sub>	.947	.793	.505	.945	.821	1	.993	.777	.629	.604	.805	.793	.803
SE <sub>50</sub>	.998	.659	.730	.830	.960	.939	1	.780	.616	.605	.807	.795	.805
$\phi$	.857	.579	.451	.815	.765	.909	.861	1	.286	.282	.959	.979	.963
SP <sub><math>\phi</math></sub>	.885	.603	.473	.842	.790	.936	.888	.982	1	.595	.403	.354	.394
SE <sub><math>\phi</math></sub>	.896	.606	.510	.813	.822	.924	.898	.843	.911	1	.405	.354	.396
YI	.793	.821	.367	.947	.608	.887	.783	.757	.785	.770	1	.990	.999
DI	.849	.656	.432	.897	.711	.919	.849	.908	.934	.886	.859	1	.992
CP	.825	.735	.397	.935	.657	.909	.820	.831	.859	.835	.949	.942	1

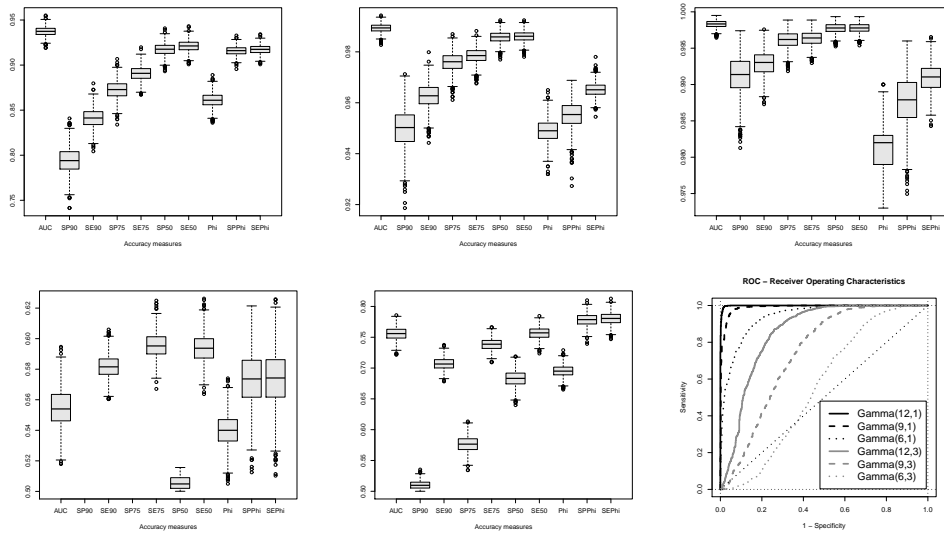
**Table 2:** Spearman's rank correlation coefficient, with  $n = 1000$ ,  $N(0, 1)$  versus  $N(2, 2/3)$  (upper triangular matrix) and  $N(2, 2)$  (lower triangular matrix).

---

### 3.3. Gamma distribution

---

Figure 7 and Table 3 show the results when the r.v.  $Y_i$  has Gamma distribution with  $\alpha_0 = 2$ ,  $\beta_0 = 1$  for a healthy individual, and  $\alpha_1 \in \{6, 9, 12\}$  and  $\beta_1 \in \{1, 3\}$  for an infected individual. The boxplots of the partial areas SP<sub>90</sub> and SP<sub>75</sub> relative to  $D_1(\theta_1) = \text{Gamma}(6, 3)$  are not shown in Figure 7. It reveals problems on the computation of the partial area over a range of high specificity. If we observe the graph with the ROC curves, the curve with the worst performance is below the line of random classification in the high specificity values. Thus, the standardized partial area under the ROC curve would be lower than 0.5 (accuracy worse than in random classification). As in previous case, some of the ROC curves are not symmetric around  $\varphi_e = \varphi_s$  and, therefore, the partial areas over the specificity are quite different from the partial areas over the sensitivity. Moreover, the partial areas over a neighbourhood of  $\phi$  seem to continue to assess accuracy without revealing any problem, regardless of whether they are being computed over the sensitivity or over the specificity.



**Figure 7:**  $D_0(\theta_0) = \text{Gamma}(2,1)$  versus  $D_1(\theta_1) = \text{Gamma}(6,1)$  (top left),  $\text{Gamma}(9,1)$  (top middle),  $\text{Gamma}(12,1)$  (top right),  $\text{Gamma}(6,3)$  (bottom left),  $\text{Gamma}(9,3)$  (bottom middle), and ROC curves (bottom right), with  $n = 1000$ .

Table 3 displays the Spearman’s rank correlation coefficients between all the computed measures when the r.v.  $Y_i$  is characterized by  $\text{Gamma}(2,1)$  for a healthy individual and characterized by  $\text{Gamma}(12,1)$  (upper triangular matrix) and  $\text{Gamma}(9,3)$  (lower triangular matrix) for an infected individual. In the  $\text{Gamma}(12,1)$  case, the AUC and the different spAUC are strongly correlated, but the rank correlations between AUC or any of the spAUC and the measures YI, DI and CP are not that significant. In fact, the index  $\phi$  is the only accuracy measure that reveals strong correlations with these indexes to select the optimal cut-point, albeit these correlations are not so significant in the  $\text{Gamma}(9,3)$  case.

---

### 3.4. Discrete distributions

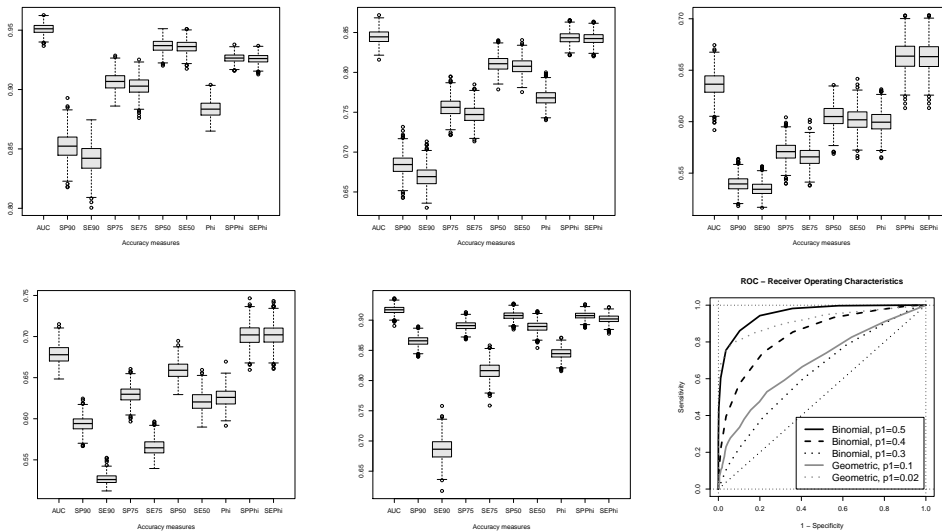
---

In these last scenarios, two count distributions are analysed, the Binomial with  $n$  trials and success probability  $p$  ( $B(n, p)$ ) and the Geometric distribution with probability  $p$  ( $G(p)$ ). Hence, in the first scenario let the r.v.  $Y_i$  have  $B(20, p)$  with  $p_0 = 0.25$  for a healthy individual and  $p_1 \in \{0.5, 0.4, 0.3\}$  for an infected individual. In the second scenario, the r.v.  $Y_i$  is characterized by  $G(p)$  where  $p_0 = 0.2$  for a healthy individual and  $p_1 \in \{0.1, 0.02\}$  for an infected individual. The results in both scenarios do not reveal any problem in the calculation of any

	AUC	SP <sub>90</sub>	SE <sub>90</sub>	SP <sub>75</sub>	SE <sub>75</sub>	SP <sub>50</sub>	SE <sub>50</sub>	$\phi$	SP $_{\phi}$	SE $_{\phi}$	YI	DI	CP
AUC	1	.994	.858	1.000	.966	1.000	.997	.638	.979	.802	.654	.656	.655
SP <sub>90</sub>	.592	1	.836	.995	.956	.994	.990	.639	.990	.772	.652	.656	.653
SE <sub>90</sub>	.657	.183	1	.856	.933	.858	.870	.811	.770	.984	.832	.830	.833
SP <sub>75</sub>	.875	.778	.335	1	.965	1.000	.997	.639	.981	.800	.655	.657	.656
SE <sub>75</sub>	.809	.250	.910	.465	1	.966	.975	.709	.923	.881	.727	.727	.728
SP <sub>50</sub>	.993	.619	.571	.905	.751	1	.997	.638	.979	.802	.654	.656	.655
SE <sub>50</sub>	.934	.357	.784	.660	.935	.904	1	.648	.974	.814	.666	.666	.667
$\phi$	.859	.351	.540	.650	.736	.859	.892	1	.545	.794	.924	.958	.936
SP $_{\phi}$	.899	.380	.552	.707	.759	.900	.916	.867	1	.697	.566	.566	.566
SE $_{\phi}$	.886	.359	.558	.671	.761	.886	.920	.979	.935	1	.832	.822	.831
YI	.769	.230	.790	.442	.947	.730	.890	.709	.733	.733	1	.981	.998
DI	.849	.311	.627	.578	.860	.837	.925	.878	.901	.909	.841	1	.989
CP	.821	.276	.693	.518	.916	.798	.921	.800	.833	.829	.927	.948	1

**Table 3:** Spearman’s rank correlation coefficient, with  $n = 1000$ , Gamma(2, 1) versus Gamma(12, 1) (upper triangular matrix) and Gamma(9, 3) (lower triangular matrix).

of the spAUC, despite some of the ROC curves being asymmetric around the line  $\varphi_e = \varphi_s$ . Thus, in some cases the partial areas over the specificity assume different values when compared with the partial areas over the sensitivity, but all measures were computed in the analysed cases.



**Figure 8:** Binomial and Geometric distributions with  $n = 1000$ : B(20,  $p$ ),  $p_0 = \frac{1}{4}$ ,  $p_1 = .5$  (top left), B(20,  $p$ ),  $p_0 = \frac{1}{4}$ ,  $p_1 = .4$  (top middle), B(20,  $p$ ),  $p_0 = \frac{1}{4}$ ,  $p_1 = .3$  (top right), G( $p$ ),  $p_0 = .2$ ,  $p_1 = .1$  (bottom left), G( $p$ ),  $p_0 = .2$ ,  $p_1 = .02$  (bottom middle), and ROC curves (bottom right).

Table 4 shows the Spearman’s rank correlation coefficients between all the computed measures when the r.v.  $Y_i$  is characterized by  $B(20, 0.25)$  for a healthy individual and by  $B(20, 0.5)$  for an infected individual (upper triangular matrix), and  $Geometric(0.2)$  for a healthy individual versus  $Geometric(0.1)$  for an infected individual (lower triangular matrix). The results using count distributions appear to be similar to those previously obtained with the use of continuous distributions. Thus, the  $\phi$  index continues to present very significant correlations with the measures YI, DI and CP, higher in the Binomial case than in the Geometric case.

	AUC	SP <sub>90</sub>	SE <sub>90</sub>	SP <sub>75</sub>	SE <sub>75</sub>	SP <sub>50</sub>	SE <sub>50</sub>	$\phi$	SP $_{\phi}$	SE $_{\phi}$	YI	DI	CP
AUC	1	.822	.837	.941	.954	.991	.994	.798	.905	.870	.799	.799	.798
SP <sub>90</sub>	.571	1	.410	.934	.641	.851	.783	.703	.799	.599	.704	.672	.699
SE <sub>90</sub>	.598	.126	1	.628	.928	.791	.860	.562	.670	.810	.563	.576	.566
SP <sub>75</sub>	.786	.860	.235	1	.839	.966	.923	.853	.942	.830	.854	.834	.851
SE <sub>75</sub>	.786	.203	.885	.352	1	.934	.973	.801	.886	.916	.801	.802	.802
SP <sub>50</sub>	.936	.689	.351	.909	.533	1	.983	.823	.929	.881	.823	.812	.822
SE <sub>50</sub>	.955	.357	.687	.582	.883	.816	1	.810	.915	.889	.811	.803	.810
$\phi$	.818	.318	.375	.556	.564	.806	.830	1	.920	.804	1.00	.990	1.00
SP $_{\phi}$	.893	.397	.387	.641	.596	.892	.889	.924	1	.910	.921	.897	.918
SE $_{\phi}$	.888	.414	.378	.649	.589	.890	.877	.847	.983	1	.805	.801	.805
YI	.854	.541	.310	.844	.462	.933	.745	.744	.829	.829	1	.990	1.00
DI	.862	.405	.344	.668	.532	.891	.835	.881	.960	.952	.869	1	.993
CP	.865	.422	.339	.697	.522	.904	.827	.863	.947	.941	.901	.990	1

**Table 4:** Spearman’s rank correlation coefficient, with  $n = 1000$ ,  $B(20, 0.25)$  versus  $B(20, 0.5)$  (upper triangular matrix) and  $Geometric(0.2)$  versus  $Geometric(0.1)$  (lower triangular matrix).

---

### 3.5. Sensitivity and specificity on the optimal cut-point

---

The first quartile  $q_1$  and the third quartile  $q_3$  of the sensitivity  $\varphi_s$  and of the specificity  $\varphi_e$  on the cut-points selected by the application of the YI, DI and CP criteria are displayed on Table 5. It is also shown  $q_1$  and  $q_3$  of the  $\phi$  index, in which  $\varphi_s = \varphi_e$  or, at least, its distance is minimized and  $\phi = \frac{\varphi_s + \varphi_e}{2}$ . The results clearly stand out the diverge accuracy levels obtained when the cut-points are set by YI, DI, and CP. Moreover, the results suggest that these differences may occur in any sense, i.e. none of these measures gives priority to sensitivity or to specificity in relation to the others measures. For example, the cut-point selected by the YI generates better sensitivity (consequently worse specificity) in the cases  $\Gamma(\cdot, 3)$  but generates worse sensitivity (and better specificity) in the  $N(2, 1.5)$  or  $G(0.10)$  cases. On the other hand, the cut-point selected by

the DI criterion generates better sensitivity (consequently worse specificity) in the cases  $N(2, \cdot)$  but generates worse sensitivity (and better specificity) in the  $\text{Gamma}(\cdot, 3)$  cases. Therefore, the accuracy of the cut-points selected through these procedures must be evaluated and compared in each application.

Let us also point out that the  $\phi$  index can also be used to select the cut-point in any application, as the results displayed in Table 5 prove. In this case the balance between sensitivity and specificity is a priority (these measures are the same, or at least very close), with a clear reduction in the variation of these measurements (as we can ascertain by comparing the interquartile range).

	YI				DI				CP				$\phi$	
	$\varphi_s$		$\varphi_e$		$\varphi_s$		$\varphi_e$		$\varphi_s$		$\varphi_e$		$\varphi_s = \varphi_e$	
	q1	q3	q1	q3	q1	q3	q1	q3	q1	q3	q1	q3	q1	q3
$n=50$	.800	.900	.840	.920	.820	.885	.840	.900	.820	.900	.840	.905	.820	.860
$n=100$	.810	.890	.840	.910	.830	.880	.830	.880	.820	.890	.830	.890	.820	.860
$n=250$	.816	.876	.828	.888	.832	.864	.832	.868	.828	.872	.828	.876	.828	.852
$n=500$	.822	.868	.828	.874	.834	.860	.834	.860	.830	.866	.828	.866	.834	.850
$n=1000$	.825	.861	.829	.865	.833	.855	.834	.855	.830	.860	.830	.860	.835	.846
$N(2, 2/3)$	.927	.944	.928	.946	.931	.942	.929	.941	.929	.944	.927	.944	.930	.937
$N(2, 1.5)$	.710	.751	.849	.887	.758	.782	.808	.836	.733	.768	.829	.865	.782	.794
$N(2, 2)$	.631	.672	.880	.914	.701	.725	.803	.833	.667	.700	.843	.879	.740	.755
$N(2, 3)$	.554	.587	.923	.949	.631	.655	.810	.843	.592	.621	.875	.910	.684	.699
$\text{Ga}(6, 1)$	.869	.898	.832	.864	.864	.882	.847	.866	.869	.895	.835	.864	.856	.866
$\text{Ga}(9, 1)$	.952	.967	.937	.952	.950	.961	.942	.953	.953	.966	.938	.952	.946	.952
$\text{Ga}(12, 1)$	.982	.989	.977	.985	.982	.987	.979	.984	.982	.989	.977	.984	.979	.983
$\text{Ga}(6, 3)$	.854	.906	.283	.338	.642	.693	.454	.489	.668	.726	.434	.474	.533	.547
$\text{Ga}(9, 3)$	.852	.894	.555	.601	.756	.792	.641	.668	.797	.840	.608	.642	.689	.701
$\text{Ga}(12, 3)$	.889	.922	.718	.755	.840	.867	.766	.787	.875	.905	.736	.766	.795	.808
$B(50, .5)$	.861	.877	.892	.904	.861	.876	.892	.904	.861	.876	.892	.904	.861	.876
$B(50, .4)$	.741	.759	.778	.795	.741	.759	.778	.795	.741	.759	.778	.795	.741	.759
$B(50, .3)$	.571	.597	.606	.631	.572	.593	.606	.627	.572	.593	.606	.627	.572	.593
$G(.10)$	.470	.546	.733	.802	.581	.606	.658	.683	.570	.602	.663	.696	.608	.655
$G(.02)$	.774	.804	.918	.943	.815	.833	.875	.898	.787	.813	.906	.932	.839	.852

**Table 5:** First and third quartiles of sensitivity and specificity on the optimal cut-point.

---

#### 4. CONCLUSION — FINAL REMARKS

---

In most situations AUC, spAUC and  $\phi$  are strongly correlated and, therefore, seem to be able to evaluate the same criterion of accuracy. Nevertheless, AUC shows less variability than spAUC, mainly on small samples and in cases with worse accuracy. Moreover, spAUC with sensitivity or specificity over  $[\phi - 0.05, \min\{\phi + 0.05, 1\}]$  shows less variability than over  $[0.9, 1]$ ,  $[0.75, 1]$  or even  $[0.5, 1]$ , albeit assessing a smaller range. In some cases, it is not possible

to compute the spAUC over a range of sensitivity using the Normal distribution and over a range of specificity using the Gamma distribution. Actually, in some situations the spAUC seems to provide better results when computed over a range of specificity (rather than sensitivity), but the opposite may also occur in other cases. However, the partial areas computed over a neighbourhood of  $\phi$  do not seem to have any problem in assessing accuracy even when the ROC curve crosses the diagonal line and, therefore, it enables to overcome the main drawback usually identified in the application of the spAUC. Furthermore, the  $\phi$  index has higher correlation with YI, DI, CP than AUC or any of the computed spAUC. In fact, the  $\phi$  index seems to be the measure with higher rank correlation with the sensitivity and specificity of the optimal cut-point selected by the use of any of the analysed optimization criteria. Additionally, this index can also be applied to select the optimal cut-point, ensuring a balance between sensitivity and specificity. Accordingly, this index seems to perform better in the evaluation of the most appropriate model as well as in the selection of the optimal cut-point. Finally, it is equally important to point out that the cut-points set by YI, DI, and CP can, in some cases, be quite different and generate significantly distinct accuracy measures. Hence, in each application their performances should be evaluated and the selected cut-points compared.

In fact, the variability of the diagnostic accuracy measures in simulations under the same scenario is quite high and, therefore, the obtained estimates do not always reveal the true accuracy of the applied classification procedure. Hence, new estimation techniques for these measures (or other measures) must be investigated in order to minimize this variability and to achieve more robust estimates, for example applying bootstrap or other resampling techniques.

---

## ACKNOWLEDGMENTS

---

This work has been funded by FCT - Fundação Nacional para a Ciência e Tecnologia, Portugal, through the projects UID/MAT/00006/2013, UID/MAT/04561/2013, UID/MAT/00006/2019 and UID/MAT/04561/2019.

---

## REFERENCES

---

- [1] ALLEN, F.; BEHAN, F.; KHODAK, A.; IORIO, F.; YUSA, K.; GARNETT, M. and PARTS, L. (2019). JACKS: joint analysis of CRISPR/Cas9 knock-out screens, *Genome Research*, available online since January 23, 2019.
- [2] CHEN, M.L.; DAVIT, B.; LIONBERGER R.; WAHBA Z.; AHN H.Y. and YU L.X. (2011). Using partial area for evaluation of bioavailability and bioequivalence, *Pharmaceutical Research*, **28**, 8, 1939–1947.



- [3] DODD, L.E. and PEPE, M.S. (2003). Partial AUC estimation and regression, *Biometrics*, **59**, 3, 614–623.
- [4] FLUSS, R.; FARAGGI, D. and REISER, B. (2005). Estimation of the Youden Index and its associated cutoff point, *Biometrical Journal*, **47**, 4, 458–472.
- [5] HAJIAN-TILAKI, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation, *Caspian Journal of Internal Medicine*, **4**, 2, 627–635.
- [6] HANLEY, J.A. and MCNEIL, J.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology*, **143**, 29–36.
- [7] HANLEY, J.A. (1989). Receiver operating characteristic (ROC) methodology: state of the art, *Critical Reviews in Diagnostic Imaging*, **29**, 307–335.
- [8] JIANG, Y.; METZ, C.E. and NISHIKAWA, R.M. (1996). A receiver operating characteristic partial area index for highly sensitive diagnostic tests, *Radiology*, **201**, 745–750.
- [9] KRAZANOWSKI, W.J. and HAND, D.J. (2009). *ROC Curves for Continuous Data*, CRC press, New York.
- [10] LEE, L.H.N.; CHOI, C.; GERSHKOVICH, P.; BARR, A.M.; HONER, W.G. and PROCYSHYN, R.M. (2016). Proposing the use of partial AUC as an adjunctive measure in establishing bioequivalence between deltoid and gluteal administration of long-acting injectable antipsychotics, *European Journal of Drug Metabolism and Pharmacokinetics*, **41**, 659–664.
- [11] LIU, X. (2012). Classification accuracy and cut point selection, *Statistics in Medicine*, **31**, 2676–2686.
- [12] LIU, X. and JIN Z. (2007). Item reduction in a scale for screening, *Statistics in Medicine*, **26**, 23, 4311–4327.
- [13] MA, H.; BANDOS, A. and GUR, D. (2015). On the use of partial area under the ROC curve for comparison of two diagnostic tests, *Biometrical Journal*, **57**, 304–320.
- [14] MA, H.; BANDOS, A.; ROCKETTE, H. and GUR, D. (2013). On use of partial area under the ROC curve for evaluation of diagnostic performance, *Statistics in Medicine*, **32**, 3449–3458.
- [15] METZ, C.E. (2008). ROC analysis in medical imaging: A tutorial review of the literature, *Radiological Physics and Technology*, **1**, 2–12.
- [16] NARASIMHAN, H. and AGARWAL, S. (2013). A structural SVM based approach for optimizing partial AUC, *Proceedings of the 30th International Conference on Machine Learning*, **28**.
- [17] NARASIMHAN, H. and AGARWAL, S. (2017). Support vector algorithms for optimizing the partial area under the ROC curve, *Neural Computation*, **29**, 7, 1919–1963.
- [18] PEPE, M.S. (2003). *Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press, Oxford.
- [19] PERKINS, N.J. and SCHISTERMAN, E.F. (2006). The inconsistency of “optimal” cut-points using two ROC based criteria, *American Journal of Epidemiology*, **53**, 670–675.

- [20] POWERS, D. (2011). Evaluation: from precision, recall and F-score to ROC, informedness, markedness & correlation, *Journal of Machine Learning Research*, **2**, 37–63.
- [21] ROBIN, X.; TURCK, N.; HAINARD, A.; TIBERTI, N.; LISACEK, F.; SANCHEZ, J.C. and MÜLLER, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves, *Bioinformatics*, **12**, 1–8.
- [22] ROTA, M. and ANTOLINI, L. (2014). Finding the optimal cut-point for Gaussian and Gamma distributed biomarkers, *Computational Statistics and Data Analysis*, **69**, 1–14.
- [23] SANTOS, R.; MARTINS, J.P. and FELGUEIRAS, M. (2015). *An overview of quantitative continuous compound tests*. In “Dynamics, Games and Science” (J.P. Bourguignon, R. Jeltsch, A. Pinto and M. Viana, Eds.), CIM Series in Mathematical Sciences **1**, 627–641.
- [24] SANTOS, R.; FELGUEIRAS, M. and MARTINS, J.P. (2015). *Discrete compound tests and Dorfman’s methodology in the presence of misclassification*. In “Theory and Practice of Risk Assessment” (C.P. Kitsos, T. Oliveira, A. Rigas and S. Gulati, Eds.), Springer Proceedings in Mathematics & Statistics **136**, 85–98.
- [25] SANTOS, R.; MARTINS, J.P.; FELGUEIRAS, M. and FERREIRA, L. (2017). *Binary classification based on a quantitative variable – an accuracy comparison by simulation*. In “Proceedings of 17th International Conference Computational and Mathematical Methods in Science and Engineering” (J. Vigo-Aguiar, Eds.), 1883–1886.
- [26] SANTOS, R.; MARTINS, J.P.; FELGUEIRAS, M. and FERREIRA, L. (2018). *Medidas de fiabilidade de classificação binária com base numa variável quantitativa – uma comparação via simulação*. In “Livro de Atas do III Encontro Luso-Galaico de Biometria” (M. Monteiro, A. Freitas, L. Teixeira and M. Costa, Eds.), Sociedade Portuguesa de Estatística, 86–89.
- [27] SCHISTERMAN, E.F.; PERKINS N.J.; LIU A. and BONDELL H. (2005). Optimal cut-point and its corresponding Youden index to discriminate individuals using pooled blood samples, *Epidemiology*, **16**, 1, 73–81.
- [28] SING, T.; SANDER, O.; BEERENWINKEL, N.; and LENGAUER, T. (2005). ROCr: visualizing classifier performance in R, *Bioinformatics*, **21**, 20, 3940–3941.
- [29] UNAL, I. (2017). Defining an optimal cut-point value in ROC analysis: an alternative approach, *Computational and Mathematical Methods in Medicine*, Article ID 3762651, 14 pages.
- [30] VIVO, J.-M.; FRANCO, M. and VICARI, D. (2018). Rethinking an ROC partial area index for evaluating the classification performance at a high specificity range, *Advances in Data Analysis and Classification*, **12**, 683–704.
- [31] WALTER, S.D. (2005). The partial area under the summary ROC curve, *Statistics in Medicine*, **53**, 2025–2040.
- [32] WITTEN, E. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation, *Caspian Journal of Internal Medicine*, **4**, 627–635.
- [33] YANG H.; LU K.; LYU, X. and HU, F. (2019). Two-way partial AUC and its properties, *Statistical Methods in Medical Research*, **28**, 1, 184–195.
- [34] YODEN, W.J. (1950). Index for rating diagnostic tests, *Cancer*, **3**, 32–35.

- [35] ZHOU, X.H.; OBUCHOWSKI, N.A. and MCCLISH, D.K. (2002). *Statistical Methods in Diagnostic Medicine*, Wiley & Sons, New York.
- [36] ZOU, K.H.; YU, C.R.; LIU, K.; CARLSSON, M.O. and CABRERA, J. (2013). Optimal thresholds by maximizing or minimizing various metrics via ROC-type analysis, *Academic Radiology*, **20**, 7, 807–815.

---

---

## JOINT MODELLING OF LONGITUDINAL AND COMPETING RISKS DATA IN CLINICAL RESEARCH

---

---

- Authors: LAETITIA TEIXEIRA  
– Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto  
CINTESIS - Centro de Investigação em Tecnologias e Serviços de Saúde,  
EPIUnit, Instituto de Saúde Pública, Universidade do Porto, Portugal  
lcteixeira@icbas.up.pt
- INÊS SOUSA  
– Departamento de Matemática e Aplicações, Universidade do Minho  
Centro de Biologia Molecular e Ambiental (CBMA), Portugal  
isousa@math.uminho.pt
- ANABELA RODRIGUES  
– Departamento de Nefrologia, Centro Hospital do Porto, Hospital Geral  
de Santo António, Instituto de Ciências Biomédicas Abel Salazar,  
Universidade do Porto, Portugal  
rodrigues.anabela2016@gmail.com
- DENISA MENDONÇA  
– Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto  
EPIUnit, Instituto de Saúde Pública, Universidade do Porto, Portugal  
dvmendon@icbas.up.pt

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

### Abstract:

- Joint modelling of longitudinal and survival data has received much attention in the recent years and is becoming increasingly used in clinical studies. When the longitudinal outcome and survival endpoints are associated, the many well-established models with different specifications proposed to analyse separately longitudinal and time-to-event outcomes are not suitable to analyse such data and a joint modelling approach is required. Although some joint models were adapted in order to allow for competing endpoints, this methodology has not been widely disseminated. The present study has as main objective to model jointly longitudinal and survival data in a competing risk context, discussing the different parameterisations of systematic implementations of these models in the R, using a real data set as an example for the comparison between the different model approaches. The relevance of this issue is associated with the need to draw attention of the users of this statistical software to the different interpretations of model parameters when fitting these models. To reinforce the relevance of these models in clinical research, we give an example of a data set on peritoneal dialysis that was analysed in this context, where death/transfer to haemodialysis was the event of interest and renal transplant was the competing event. Joint modelling results were also compared to separate analysis for these data.

## Key-Words:

- *Competing risks; joint modelling; longitudinal data; peritoneal dialysis; time-to-event data.*

## AMS Subject Classification:

- 62N, 62P10.

---

## 1. INTRODUCTION

---

In many clinical research studies, it is relevant to simultaneously analyse information on a longitudinal repeatedly registered biomarker and on the time to a specific outcome event. Furthermore, more than one outcome event may occur. In these situations, when longitudinal and time-to-event outcomes are associated, a joint modelling approach taking competing risks into account is required to correctly analyse such data [23, 25].

Although, in cross-sectional clinical studies only one measure of each clinical parameter (often the baseline) is used to guide medical decisions, the use of additional information on repeated measures of clinical parameters allows a better understanding of the disease progression or treatment benefits [1]. In this type of longitudinal studies, the analysis of repeated measures of clinical parameters may be supplemented with information about the time at which an event of interest has occurred, that is, survival data (also designated by time-to-event data).

With the purpose of analysing separately longitudinal and survival data, methods such as linear mixed model [4] and Cox proportional hazard model [3], respectively, are well-established. However, when longitudinal measurements are correlated with time-to-event (i.e., in the presence of informative censoring - when the reason for censoring is related to the study outcomes), when repeated measures are measured with error and/or when some missing values are present a joint modelling approach is required [23]. These aspects that realistically characterize observed data lead to biased inferences if naive separate methods are applied [5, 7, 9, 12, 20, 23].

Therefore, in joint modelling methodology several objectives may be formulated, according to the main focus of the analysis [5, 10]: (i) to analyse the time-to-event outcome, taking into account the effect of a longitudinal outcome as endogenous time-dependent covariate measured with error, (ii) to analyse the longitudinal outcome in the presence of informative (non-random) dropout time and (iii) to analyse effects of covariates of interest on both type of outcomes (longitudinal and time-to-event) simultaneously.

Despite joint modelling of longitudinal and survival data is becoming increasingly popular [2, 18, 24], joint modelling in competing risk framework has not been widely used in medical context. Given the complexity of the joint modelling approach in the presence of competing risks, several limitations can be enumerated, namely the small number of models implemented in statistical software and the restriction associated to the number of shared random effects to be integrated out in the likelihood function due to computational limitations [14].

Several authors have suggested extensions of joint models so that they could be applied in a competing risks problem, such as Elashoff et al. [6, 7], Williamson et al. [25], Li et al. [13] and Rizopoulos [19]. The approaches differ

according to parameterisation, joint likelihood function and estimation method considered. Up to now there are only two statistical packages, available in the CRAN repository, that implement systematically two different parameterisations, the JM package [19] and the `joineR` package [25].

A very recent review of several implementations of joint modelling was published [11], which summarize four published models, which have software available for model estimation. Each model features a different hazard function, latent association structure between the longitudinal and survival submodels, estimation approach and software implementation. The models described were applied to a trial of anti-epileptic drugs. However, in this work we further discuss the packages `joineR` and JM, namely the different interpretations of the model coefficients and the application in another clinical area.

Peritoneal dialysis is one of the main renal replacement therapy. The progression of end-stage renal disease patients included in a peritoneal dialysis program is monitored with regular control visits where several clinical parameters are recorded, as well as the time until the occurrence of relevant endpoints. Then, as in many other clinical research areas, in addition to the baseline characteristics, peritoneal dialysis patient data present two different types of outcomes: (i) longitudinal outcome, composed by clinical parameters measured at several time points (such as albumin), and (ii) time-to-event outcome, composed by the follow-up time until the occurrence of an event of interest. In the specific case of peritoneal dialysis patients, it is only possible to observe the first outcome event (and consequently the first time-to-event) from a set of possible competing events: death/transfer to haemodialysis and renal transplant. For this reason, we are in a competing risks framework [22].

As referred above, the focus of the present study is on the two approaches to joint model, which are the only ones implemented in common statistical software (R) for systematic use by any users, (1) JM package by Rizopoulos [19] and (2) `joineR` package by Williamson et al. [25]. In practice these two implementations of joint models correspond to different parameterisations with different parameter interpretations. With this work we emphasize that it is important to discuss at this stage the differences between the two joint models, since interpretation of model parameters are different, and confusing interpretations may occur. Notice that, using a real data set as an example, we want to analyse the differences of the results when using the two model approaches and make interpretations on the results. It is not our purpose to go further about the performance of the two approaches. Additionally, the implementation of these approaches allows us to illustrate the relevance of the joint modelling methodology in the evaluation of a peritoneal dialysis program.

The objective of this present study is threefold: i) to jointly model longitudinal and survival data in a competing risks framework; ii) to discuss different parameterisations of systematic implementations of these models in the available R statistical software; iii) to analyse data on peritoneal dialysis program under

a joint modelling approach with competing risks and compare results with those from separated longitudinal and survival analysis.

In the next section we review the theory of joint modelling, with focus on the competing risks approach. The third section presents the results of the analysis of the peritoneal dialysis dataset. Finally, a discussion and conclusion compose the last section.

---

## 2. JOINT MODELLING IN THE PRESENCE OF COMPETING RISKS

---

The joint modelling approach takes into account the association between the survival and longitudinal process, determining simultaneously the parameter estimates for both processes [14]. Different models can be considered, differing on the decomposition of the joint likelihood of the longitudinal and survival processes and on the submodels formulation for each outcome. The models most commonly used are selection models, pattern-mixture models and random effects models, and each model providing different information [21]. The two parameterisations considered in this work are classified as random effects models, where the survival process is assumed to be associated with the longitudinal process through shared random effects. In the presence of competing risks, the survival submodel needs to take into account the presence of several possible endpoints. In order to model jointly a longitudinal and a time-to-event outcome in the presence of competing risks some approaches are presented below.

According to the focus of the analysis, different specifications of the joint model might be considered, which corresponds to different parameterisations of the model, taking us to different interpretations of the model parameters.

When the focus is on the survival process and the interest is to analyse the effect of an endogenous time-dependent covariate (for example a clinical parameter such as albumin measured along time) on the time until an event of interest (for example, death), the time-dependent cause-specific hazard regression model usually used in competing risk survival analysis is not appropriate. Results obtained from this model may be substantially biased since longitudinal measures are measured with error [5, 6]. In these situations, the fundamental idea is to construct a suitable model to describe the evolution in time for the longitudinal outcome, and then to use this estimated evolution as time-dependent covariate in the survival model, considering a jointly estimation [1].

Alternatively, when the focus is on the longitudinal process (for example, of some clinical parameter such as albumin), the joint modelling approach is required when missing observations of the longitudinal outcome may be related with the endpoint observed (i.e. in the presence of informative censoring). The use of a joint modelling approach reduces the bias in the estimates [14].



Additionally, if the focus is on both processes, the aim of the model is on inference regarding the strength of the link between the two processes [14, 16].

Let  $y_i(t)$  be the observed value of a longitudinal response for the subject  $i$  at time point  $t$ , measured with error. Let  $T_i$  and  $C_i$  be the failure and non-informative censoring times and  $k$  the event observed of a set of  $K$  possible events ( $k = 1, \dots, K$ ). The event indicator is given by  $\delta_i = \{I(T_i \leq C_i), k\}$ , where  $\delta_i = 0$  if non-informative censoring occurs.

---

## 2.1. JM package

---

The JM package that implements the parameterisation proposed by Rizopoulos [19] was adapted to a competing risks problem [19]. This approach considers a linear mixed effects submodel for the longitudinal outcome and a relative risk submodel for each possible competing event. This model allows to quantify the effect of a longitudinal covariate in the time-to-event outcome, particularly when the longitudinal covariate is measured with error [14].

Consider  $m_i(t)$  the true and unobserved value of the longitudinal outcome  $y_i(t)$  at time  $t$ . In order to measure the effect of an endogenous covariate on the risk for an event,  $m_i(t)$  needs to be estimated. Furthermore, the complete history of the true unobserved longitudinal process up to time point  $t$ ,  $M_i(t) = \{m_i(s), 0 \leq s < t\}$ , is successfully reconstructed using the available measurements  $y_i = \{y_i(t), t = 1, \dots, n_i\}$  of each subject (where  $n_i$  represents the number of longitudinal measurements for each subject  $i$ ) and a set of modelling assumptions. A linear mixed effects model is considered to describe the subject-specific longitudinal evolutions and it is defined as:

$$(2.1) \quad y_i(t|x_{1i}, W_{1i}) = x_{1i}(t)^T \beta_1 + W_{1i}(t) + \varepsilon_i(t) = m_i(t) + \varepsilon_i(t)$$

where  $\beta_1$  denotes the vector of the unknown fixed effects parameters,  $x_{1i}(t)$  denotes row vectors of the design matrix for the fixed effects and  $\varepsilon_i(t)$  is the measurement error term with variance  $\sigma^2$  ( $\varepsilon_i(t) \sim N(0, \sigma^2)$ ).  $W_{1i}(t)$  is the value at time  $t$  of an unobserved zero-mean Gaussian random process.

To quantify the effect of  $m_i(t)$  on the risk for an event,  $\lambda_i$ , the authors proposed the use of a relative risk model:

$$(2.2) \quad \lambda_i(t|M_i(t), x_{2i}) = \lambda_0(t) \exp\{x_{2i}^T \beta_2' + \alpha m_i(t)\}$$

where  $\lambda_0(t)$  denotes the baseline risk function and  $x_2$  is a vector of baseline covariates with a corresponding vector of regression coefficients  $\beta_2'$ . Parameter  $\alpha$  quantifies the effect of the underlying longitudinal outcome on the risk for an event:  $\exp(\alpha)$  denotes the relative increase in the risk for an event at time  $t$  that results from one unit increase in  $m_i(t)$  at the same time point, adjusting for the remaining exploratory variables in the model.

In the presence of competing risks the notation for the survival submodel needs to be adapted. Then, for the event  $k$ , the standard relative risk model can be defined as:

$$(2.3) \quad \lambda_{ki}(t|M_i(t), x_{2i}) = \lambda_{0k}(t) \exp\{x_{2i}^T \beta'_{2k} + (\alpha_1 + \alpha_k)m_i(t)\}$$

where  $k = 1$  represents the event of interest and  $k = 2, \dots, K$  the competing events,  $\lambda_{0k}(t)$  denotes the baseline risk function and  $x_2$  is a vector of baseline covariates with a corresponding vector of regression coefficients  $\beta'_2$ .  $\alpha_1$  quantifies the effect of the underlying longitudinal outcome on the risk for the event of interest and  $\alpha_2, \dots, \alpha_K$  quantifies the additional effect of the underlying longitudinal outcome on the risk for the respective competing event. In this model, each of  $\beta'_{2k}$  is interpreted as the effect of each explanatory variable on the relative risk of event  $k$  after adjusting for the effect of the longitudinal response, which might also include the effect of the same explanatory variable. Then, the overall effect of a covariate on the hazard might be decomposed into the direct effect (survival submodel) and the indirect effect (longitudinal submodel) [11].

The estimation method proposed in this approach is the maximum likelihood considering a joint distribution of the observed outcomes  $\{T_i, \delta_i, y_i\}$ . This joint distribution is defined assuming that the vector of time-independent random effects  $W_{1i}$  underlies both the longitudinal and survival processes (the random effects account for both the association between the longitudinal and event outcomes, and the correlation between the repeated measurements in the longitudinal process). The likelihood function is given by

$$(2.4) \quad p(T_i, \delta_i, y_i|W_{1i}; \theta) = p(T_i, \delta_i|W_{1i}; \theta)p(y_i|W_{1i}; \theta)$$

and

$$(2.5) \quad p(y_i|W_{1i}; \theta) = \prod_{j=1}^{n_i} p\{y_i(t_{ij})|W_{1i}; \theta\}$$

where  $\theta = (\theta_t^T, \theta_y^T, \theta_{W_1}^T)$  denotes the full parameter vector, with  $\theta_t$  denoting the parameters for the event time outcome,  $\theta_y$  the parameters for the longitudinal outcome and  $\theta_{W_1}$  the parameters of the random-effects covariance matrix. Additionally, it is assumed that given the observed history, the censoring mechanism and the visiting process are independent of the true event times and future longitudinal measurements.

In the presence of competing risks, the likelihood part for the event process takes the form:

$$(2.6) \quad p(T_i, \delta_i|W_{1i}; \theta_t, \beta_1) = \prod_{k=1}^K [\lambda_{0k}(T_i) \exp\{x_{2i}^T \beta'_{2k} + \alpha_k m_i(T_i)\}]^{I(\delta_i=k)} \\ \times \exp\left(-\sum_{k=1}^K \int_0^{T_i} \lambda_{0k}(s) \exp\{x_{2i}^T \beta'_{2k} + \alpha_k m_i(s)\} ds\right)$$

The Expectation-Maximization (EM) algorithm was used to maximize the log-likelihood function  $l(\theta) = \sum_{i=1}^n \log p(T_i, \delta_i, y_i; \theta)$ . The aim is to find the parameter values  $\hat{\theta}$  that maximize the observed data log-likelihood  $l(\theta)$ , but by maximizing instead the expected value of the complete data log-likelihood (treating random effects as missing data). Additional information of this approach can be found in Rizopoulos [19].

---

## 2.2. `joiner` package

---

Williamson et al. [25] proposed a competing risks random-effects joint model fitting a cause-specific hazard submodel (allowing for competing risks) with a separate latent association between longitudinal measurements and each event [25]. The idea behind this model is to analyse data arising from competing survival and longitudinal processes simultaneously exploiting dependencies between the components. Given that the main focus of this approach is the link between longitudinal and survival processes, the association between these two processes is represented through shared latent random effects. For example, for a shared latent random effect model, this association is achieved through the inclusion of the longitudinal random intercept and/or random slope terms into the survival process model [14].

A Gaussian linear model is assumed for longitudinal response  $y(t)$  at time  $t$  (longitudinal submodel):

$$(2.7) \quad y_i(t|x_{1i}, W_{1i}) = x_{1i}(t)^T \beta_1 + W_{1i}(t) + \varepsilon_i(t)$$

where  $\beta_1$  denotes the vector of the unknown fixed effects parameters,  $x_{1i}(t)$  denotes row vectors of the design matrix for the fixed effects,  $W_{1i}(t)$  the value at time  $t$  of an unobserved zero-mean Gaussian random process and  $\varepsilon_i(t)$  denotes zero-mean Gaussian measurement error with variance  $\sigma^2$ .

The difference between the JM and `joiner` approaches is in the survival submodel. Survival time is associated with the longitudinal response through a second zero-mean latent Gaussian process  $W_{2i}(t)$ , correlated with  $W_{1i}(t)$ . A semi-parametric proportional hazards model is assumed conditioned to  $W_{2i}(t)$ , with hazard  $\lambda_i$  defined as:

$$(2.8) \quad \lambda_i(t|x_{2i}, W_{2i}) = \lambda_0(t) \exp\{x_{2i}^T \beta_2 + W_{2i}(t)\}$$

where  $\lambda_0(t)$  is an unspecified baseline hazard and  $x_2$  is a vector of baseline covariates with a corresponding vector of regression coefficients  $\beta_2$ . The longitudinal and survival processes are assumed to be conditionally independent given  $W_1$  and  $W_2$ , usually considered as a linear combination of Gaussian random effects [25]. If the two processes  $W_1$  and  $W_2$  were independent, we would be in the presence of two separate analyses (longitudinal and survival). Though, being  $W_1$  and  $W_2$  related with each other, their correlation will drive the association between longitudinal and survival processes.

This model was extended in order to include competing risks. In this case, a cause-specific hazard submodel with a separate latent association between longitudinal measurements and each possible event was considered. The longitudinal submodel remains the same type of model considered in the joint model without competing risks. On the other hand, one survival submodel for each competing risks is considered. Thus the survival submodel for cause  $k$  is defined as:

$$(2.9) \quad \lambda_{ki}(t|x_{2i}, W_{2i}) = \lambda_{0k}(t) \exp\{x_{2i}^T \beta_{2k} + W_{2ki}(t)\}$$

where  $\lambda_{0k}(t)$ ,  $k = 1, 2, \dots, K$ , are unspecified baseline hazard functions,  $x_2$  is a vector of baseline covariates and  $W_{2k}(t)$ ,  $k = 1, 2, \dots, K$ , are zero-mean latent Gaussian processes. In this case, it is assumed that  $W_{2k}(t) = \gamma_k W_1(t)$ , i.e.,  $W_1$  and  $W_2$  are proportional. The parameter  $\gamma_k$  indicates the level of association between the two components, i.e, quantify the effect of the unobserved stochastic process  $W_1$  on the risk for the event  $k$ . Longitudinal responses and competing risks survival times are assumed to be conditionally independent given  $W_1$  and  $W_2$ . In this parameterisation of the joint model the coefficient  $\beta_{2k}$  corresponds to the total effect of each explanatory variable on the relative risk of event  $k$ , after adjusting for an unobserved Gaussian process that do not include fixed effects. This different interpretation can be contrasted with the one previously given to  $\beta'_{2k}$  in the JM package.

The likelihood function for observed data is factorized as the product of the marginal distribution of  $y$  and the conditional distributions of competing events  $\eta \in (1, \dots, K)$  given the observed values of  $y$ .

Considering  $\theta$  the combined vector of unknown parameters and  $L_y(y, \theta)$  the standard likelihood corresponding to the marginal multivariate normal distribution of  $y$ . Conditional on latent processes  $W_{2k}(t)$ , the competing risks are independent of themselves and of the measurements  $y$ . The likelihood function is given by:

$$(2.10) \quad L(y, \theta, \eta) = L_y(y, \theta) \prod_{k=1}^K L_{\eta|y,k}$$

where

$$(2.11) \quad L_{\eta|y,k} = E_{W_{2k}|y} \{L_{\eta|W_{2k}}(\theta, \eta = k|W_{2k})\}$$

in which the conditional likelihood for each competing event,  $L_{\eta|W_{2k}}(\theta, \eta = k|W_{2k})$  captures any likelihood contribution arising from the number of longitudinal measurements observed before the  $k$ th competing event. In this model parameterisation, it is assumed that there is an unobserved process  $W_1$  that drives both  $y$  and risk for event,  $\lambda_k$ . The effect of covariates in hazard is both direct and overall [11].

In order to maximize the likelihood of the observed data and estimate the parameters of interest, EM algorithm is used, similarly as the JM approach. More details of this approach can be founded in Williamson et al. [25], Diggle et al. [5], Henderson et al. [10] and in Philipson et al. [17].

---

### **3. PERITONEAL DIALYSIS DATA**

---

In order to compare the two model specifications presented above and discuss the interpretation of model parameters, the methods, JM and *joineR*, were used to analyse peritoneal dialysis data. For the *joineR*, the extension to accommodate competing risks was requested directly to the author since our analyse started before the formal library to perform this analysis became available at September 2017.

Along the permanence in peritoneal dialysis program, different types of information concerning the patients and their health condition are collected. Firstly, information about baseline characteristics of the patients such as sex and age is considered. During the follow-up, albumin is usually recorded in each control visit (usually one per month). Finally, the event that forced the patient to abandon the treatment program (death, transfer to haemodialysis and renal transplantation) and the respective follow-up time are also reported given their clinical relevance. Then, due to the diversity of information resulting of the motorization of these patients, efficient and powerful regression models, such as joint models for longitudinal and time-to-event outcomes are required to analyse such data.

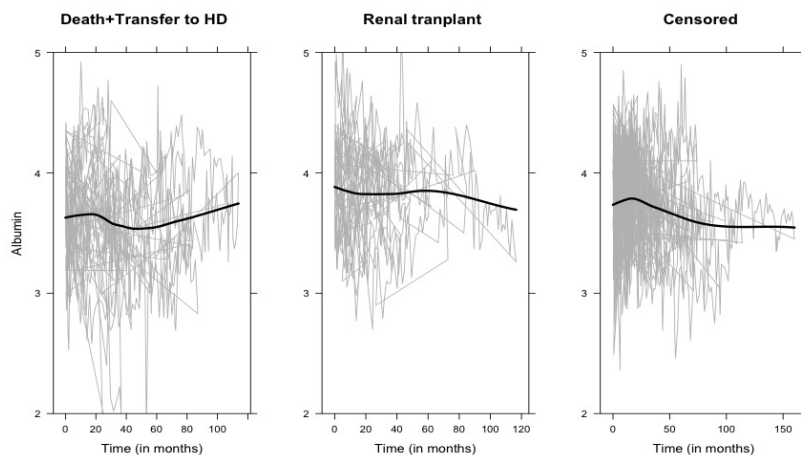
The sample of this study comprises patients included in the peritoneal dialysis program of the Peritoneal Dialysis Unit, Nephrology Department, Hospital Geral de Santo António, Centro Hospitalar do Porto, Porto, Portugal. The sample is composed by 160 patients who started peritoneal dialysis therapy between October 1999 and February 2013. Sex and age were considered as baseline covariates. Serum albumin level is an important clinical parameter for end-stage renal patients and it is used to assess the health status of patients in dialysis [15]. Low albumin level is associated with kidney failure. The number of measures and the time between measures differed for each patient. Combined survival, characterized by the combined event death/transfer to haemodialysis, represents an important indicator for the evaluation of a peritoneal dialysis program. Then, in this application, this combined event was considered as the event of interest and renal transplantation as the competing risk event. Registry data collection and analysis was submitted to ethical appreciation and approved by the National Commission of Data Protection, which is the national supervisory authority for personal data control.

Females represented 51.9% (n=83) of the total sample (n=160), which has an overall mean age of 47.9 years (sd=14.4 years). Thirty patients (18.8%) had diabetes. The median of follow-up time was 27.4 months (IQR: 12.8-49.0 months). Considering the longitudinal outcome, the number of measures of albumin varied among patients, with a minimum of 1 observation and a maximum of 60 observations. The median of observations per patient was 13 (IQR: 6-23 observations). The mean score of albumin was 3.7 g/dL (sd=0.4 g/dL) for a total of 3129 observations. Considering the time-to-event outcome, 53 (33.1%) patients experienced

the event of interest (death or transfer to haemodialysis) and 41 (25.7%) the competing risk (renal transplant). Survival times were censored for 66 (41.2%) patients who were still active on the peritoneal dialysis program at the end of the study.

### 3.1. Exploratory analysis

A *spaghetti plot* showing the albumin individual progressions (grey lines) of the longitudinal response for the different competing events is presented in Figure 1. The black lines in Figure 1 represent a smooth spline of all observation points in the same plot.

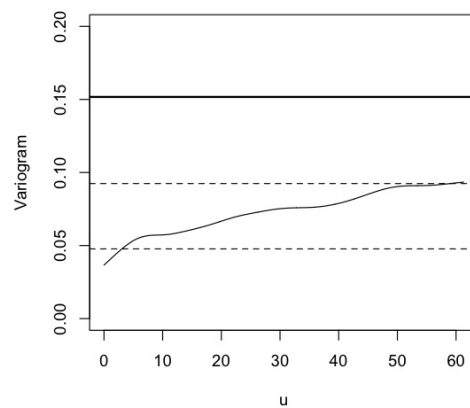


**Figure 1:** Smooth spline empirical mean of albumin evolution for the three subset of events: death/transfer to haemodialysis, renal transplant and censored.

Considering Figure 1, we verify that the mean of albumin score differs slightly according to the final event observed, showing a possible association between longitudinal albumin evolution and survival endpoint. Then, the analysis requires a joint modelling approach.

An estimate of the empirical variogram  $\gamma(u)$  is presented in Figure 2. The diagram shows both the basic quantities  $(u_{ijk}, v_{ijk})$ , where  $v_{ijk} = \frac{1}{2}(r_{ij} - r_{ik})^2$  is calculated from observed half-squared differences between pairs of residuals, of an ordinary least squares model (considering albumin as dependent variables and gender, age and time as independent variables), and  $u_{ijk} = t_{ij} - t_{jk}$  the corresponding time-differences, and the kernel smooth estimate of  $\gamma(u)$ . To accentuate the shape of the smooth estimate, the vertical axis was truncated at 0.2. The variogram smoothly increases with lag corresponding to a decreasing correlation as

observations are separated in time. The horizontal line represents the variogram-based estimate of the process variance, which is substantially large than the value of the sample variogram, indicating that the positive correlation remains at arbitrarily large time separations. The empirical variogram of residuals, after fitting the data for an ordinary least squares model, allows us to understand the correlation structure of the longitudinal data. From Figure 2, we can see that the total variance in the data can be decomposed into three variance components, variance between and within subjects and measurement error. Therefore, a longitudinal approach shows to be adequate for these data.



**Figure 2:** Empirical Variogram.

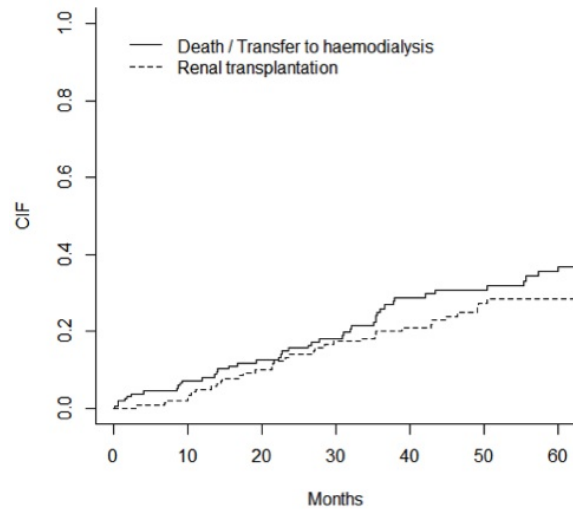
The cumulative incidence curves [8] give a global idea about the survival process. Figure 3 summarizes the cumulative incidence estimates for the two possible events taking competing risks into account (the time axis were halted at 60 months because the proportion of patients free of an event, but still in follow-up, becomes small). The probability of death/transfer to haemodialysis is always higher than the probability of renal transplantation. For example, the probabilities of death/transfer to haemodialysis by 1, 2 and 3 years after starting peritoneal dialysis were 0.08, 0.16 and 0.26 respectively and by the same time points the probabilities of renal transplantation were 0.05, 0.14 and 0.20.

---

### 3.2. Joint modelling

---

With the purpose of evaluating the relationship between longitudinal albumin scores and death/transfer to haemodialysis, in the presence of the competing risk renal transplantation, two joint model specifications implemented in the



**Figure 3:** Cumulative incidence curves for death/transfer to haemodialysis (solid line) and for renal transplantation (dotted line).

software R were analysed: the JM package proposed by Rizopoulos [19] and the `joiner` package proposed by Williamson et al. [25], both adapted to a competing risk situation. Furthermore, the parameters estimates and their standard errors using the joint modelling specifications were compared to those obtained with the independent models, a linear mixed model for the longitudinal outcome and a time-dependent Cox model with competing risks for the survival outcome.

For the two joint model specifications discussed above, a linear mixed-effects model was assumed for the longitudinal albumin outcome, with evolution in time for each patient with different average effects per sex and age. For notation simplification the individual index  $i$  and time index  $j$  were dropped. The longitudinal submodel used was defined as (see equations (2.1) and (2.7)):

$$(3.1) \quad y(t) = m(t) + \varepsilon(t) = \beta_0 + \beta_{11}Sex + \beta_{12}Age + \beta_{13}time + b_0 + b_1time + \varepsilon(t)$$

where  $y$  represents the albumin score and  $\beta_{11}$ ,  $\beta_{12}$  and  $\beta_{13}$  represent the parameters of the fixed-effects part composed by the main effect of sex, age and time, respectively. The unobserved zero-mean Gaussian random process  $W_1(t)$  as in Equation (2.1) and (2.7) is, in this case, a linear combination of a random intercept  $b_0$  and a random slope  $b_1$ . That is,  $(b_0, b_1)$  has bivariate Gaussian distribution with variances  $\sigma^2(b_0)$  and  $\sigma^2(b_1)$ , respectively, and correlation  $\rho$ .

Notice that, from Figure 2 the empirical variogram indicates the need to include a random effect at subject level ( $b_0$ , included) but also a possible Gaussian stochastic process with a time correlation structure, as well as a random noise ( $\varepsilon(t)$ , included). However, we have fitted a model without a Gaussian stochastic process because none of the systematic implementations, JM and `joiner`, allow to include such a term in the model. This is due to the computational implemen-



tation involved to estimate parameters of such a model. In particular, we would have to integrate out a continuous Gaussian process in any time points observed.

In practice, this is one of the arguments of the importance of having systematic implementations of any statistical model to be fitted by any users. Though, it implies correct interpretation and use of these statistical models. Therefore, a random intercept effect and a random slope effect were included in the model for peritoneal dialysis data, because this is the possible way to incorporate a time-dependent correlation structure within patient, which was indicated by the variogram, in any of the model implementations.

For the event process, the two approaches presented in this study have different formulation.

For the JM joint model, two cause-specific relative risks models were assumed, one for each possible event (see equation (2.3)):

$$(3.2) \quad \begin{cases} \lambda_1(t) = \lambda_{01}(t) \exp\{\beta'_{211}Sex + \beta'_{212}Age + \alpha_1 m(t)\} \\ \lambda_2(t) = \lambda_{02}(t) \exp\{\beta'_{221}Sex + \beta'_{222}Age + (\alpha_1 + \alpha_2)m(t)\} \end{cases}$$

The parameters  $\beta'_{211}$ ,  $\beta'_{212}$  and  $\alpha_1$  denote the direct effects of sex, age, and albumin, respectively, on the risk for death/transfer to haemodialysis and the parameters  $\beta'_{221}$  and  $\beta'_{222}$  denote the effects of sex and age, respectively, on the risk for renal transplantation. The parameter  $\alpha_2$  corresponds to the additional effect of the albumin score on the renal transplantation.

Considering the joineR joint model, a semi-parametric cause-specific hazard model for each event was assumed (see equation (2.9)):

$$(3.3) \quad \begin{cases} \lambda_1(t) = \lambda_{01}(t) \exp\{\beta_{211}Sex + \beta_{212}Age + \gamma_1 W_{21}(t)\} \\ \lambda_2(t) = \lambda_{02}(t) \exp\{\beta_{221}Sex + \beta_{222}Age + \gamma_2 W_{22}(t)\} \end{cases}$$

The parameters  $\beta_{211}$ ,  $\beta_{212}$  and  $\gamma_1$  denote the effects of sex, age, and albumin in the underlying unobserved process  $W_1$ , respectively, on the risk for death/transfer to haemodialysis while the parameters  $\beta_{221}$ ,  $\beta_{222}$  and  $\gamma_2$  denote the effects of sex, age, and albumin in the underlying unobserved process  $W_2$ , respectively, on the risk for renal transplantation. This approach has as focus the link between the two longitudinal and survival processes. Therefore, the association between these processes is represented through shared latent random effects, achieved through the inclusion of the longitudinal random intercept ( $b_0$ ) and random slope ( $b_1$ ) terms into the survival process.

The parameters estimates and respective p-value using joint modelling approaches are presented in Table 1. For both approaches, standard error of the parameter estimates were obtained by refitting the models to 500 bootstrap samples generated using the original data. The bootstrap sampling was performed with replacement.

Among the joint models fitted, despite the formulation of the longitudinal submodel was the same, the association structures and method of estimation used can have different influences on the longitudinal submodel estimates [11]. In this application, similar results were obtained for both approaches, with a decrease in the albumin score along time.

Considering the results obtained for the survival submodel with the JM package [19], we verify an association between albumin and the risk of death/transfer to haemodialysis ( $\hat{\alpha}_1 = -1.24, p = 0.011$ ), meaning that a unit decrease in the marker corresponds to a  $\exp(-(-1.24)) = 3.5$ -fold increase in the risk for death/transfer to haemodialysis, controlling for the remaining factors in the model. No association between albumin and the risk of renal transplantation was found ( $\hat{\alpha}_1 + \hat{\alpha}_2 = 0.54$  ( $se = 0.47$ ),  $p = 0.250$ ). Younger patients have a statistically significant higher hazard of getting a renal transplant (hazard ratio for one year decrease in age equals  $\exp(-(-0.041)) = 1.04$  ( $p < 0.001$ ). The direct effect of age in the hazard must be interpreted by also adjusted for the age-specific effect on albumin (longitudinal submodel). The log-likelihood from this joint model was -730.2515.

Results based on the `joiner` package [25] show a significantly  $\hat{\gamma}_1$  estimate indicating that albumin score is positively associated with time to death/transfer to haemodialysis. However, no evidence of association between albumin and time to renal transplantation was found ( $\hat{\gamma}_2 = 0.28, p = 0.625$ ). As expected, the estimates of the association parameters for the two competing events have opposite signs given that these two events have opposite reasons for discontinuation of therapy. Age (direct effect) was identified as statistically significant risk factor for renal transplantation (higher ages present lower hazard of renal transplantation), but not for death/transfer to haemodialysis. The log-likelihood from this joint model was -603.9029.

---

### 3.3. Separate analysis

---

Comparison of the parameters estimated and their standard errors from the joint model with the naive independent approach (independent linear mixed model and cause-specific hazard model), presented in Table 2, shows the differences of approaches. Results obtained for longitudinal outcome were similar. However, different results were obtained for time-to-event outcome. In separate analysis, sex was a significant factor for both events ( $HR = 1.41$  ( $p < 0.001$ ) for event death/transfer to haemodialysis and  $HR = 1.42$  ( $p < 0.001$ ) for event renal transplantation). Additionally, albumin (considered as time-dependent covariate) was a statistically significant factor for the event renal transplant ( $HR = 1.57, p < 0.001$ ).

	JM package		joiner package	
	Coefficient (se)	p	Coefficient (se)	p
<b>Longitudinal model</b>				
Fixed effects				
Intercept	3.88 (0.11)	< 0.001	3.88 (0.12)	< 0.001
Sex (male)	0.24 (0.062)	< 0.001	0.24 (0.078)	0.002
Age	-0.0052 (0.002)	0.014	-0.0051 (0.002)	0.025
Time	-0.0015 (0.0018)	0.400	-0.0014 (0.0007)	0.069
<b>Survival submodel</b>				
Event of interest (D/TH)				
Sex (male)	0.41 (0.33)	0.209	0.12 (0.28)	0.649
Age	-0.012 (0.011)	0.278	-0.007 (0.010)	0.502
Association coefficient	-1.24 (0.49)	0.011	-1.41 (0.50)	0.005
Competing risk (RT)				
Sex (male)	0.51 (0.40)	0.204	0.62 (0.37)	0.091
Age	-0.041 (0.012)	< 0.001	-0.048 (0.013)	< 0.001
Association coefficient	0.54* (0.47)	0.250	0.28 (0.59)	0.625
$\hat{\sigma}(\varepsilon)$	0.0524		0.0524	
$\hat{\sigma}(\hat{b}_0)$	0.147		0.159	
$\hat{\sigma}(\hat{b}_1)$	0.000117		0.000123	
$\hat{\rho}$	-0.388		-0.346	
Log-likelihood	-730.2515		-603.9029	

**Table 1:** Parameter estimates for joint models fitted to albumin (longitudinal outcome) and time to peritoneal dialysis treatment failure (survival outcome) in the presence of competing risks. \* indicates  $\alpha_1 + \alpha_2$ . joiner package: Williamson et al. [24]; JM package: Rizopoulos [19]. D/TH: Death/Transfer to haemodialysis; RT: Renal transplantation.

---

#### 4. DISCUSSION/CONCLUSION

---

It is very common to find clinical studies with both longitudinal measurements and event times. These measures are recorded on the participant of the study during follow-up time. Joint models are appropriate when interest lies in the association between a longitudinal covariate measured with error in a survival analysis or when accounting for event-dependent dropout in a longitudinal analysis. Several simulation studies have shown that joint model could be substantially more efficient than the separate analysis [6] because these models use

		Separate analysis	
		Coefficient (se)	p
<b>Longitudinal model</b>	Fixed effects		
	Intercept	3.88 (0.11)	< 0.001
	Sex (male)	0.24 (0.062)	< 0.001
	Age	-0.005 (0.002)	0.014
	Time	-0.0013 (0.0011)	0.270
<b>Survival model</b>	Event of interest (D/TH)		
	Sex (male)	0.34 (0.070)	< 0.001
	Age	0.022 (0.0025)	0.371
	Albumin	-0.58 (0.086)	< 0.001
	Competing risk (RT)		
	Sex (male)	0.35 (0.085)	< 0.001
	Age	-0.039 (0.0031)	< 0.001
	Albumin	0.45 (0.10)	< 0.001

**Table 2:** Parameter estimates longitudinal and survival model fitted separately, considering albumin as longitudinal outcome and time to peritoneal dialysis treatment failure as survival outcome in the presence of competing risks. Longitudinal model: linear mixed model; survival model: cause-specific. Cox proportional hazard model with time-dependent covariate. D/TH: Death/Transfer to haemodialysis; RT: Renal transplantation.

information from both outcomes. The literature about this theme is vast, and some review paper [14, 16, 21, 23] present and discuss different type of joint models focused on a single event with non-informative censoring. However, the majority of these models have only one event for the time-to-event outcome, excluding the possibility of observing competing risks. A very recent review paper described four approaches of joint models of longitudinal and survival data in the presence of competing risks, with application to an epilepsy drug randomized controlled trial. However, despite the recent methodological developments in the field of joint modelling of competing risks and longitudinal data, they remain still limited options for fitting these models in standard statistical software programs [11].

This work represents as far as we know the first study in the peritoneal dialysis area using joint modelling approach of longitudinal and survival data taking competing risks into account. The results obtained with this methodology produced new information about peritoneal dialysis program. Specifically, with this model it is possible to evaluate the association between the two processes, which cannot be obtained with standard survival models, contributing for a better knowledge of peritoneal dialysis program resulting in better management of the treatment program.

The development of different parameterisations with different perspectives and focuses allows to obtain different conclusions, and the choice of the model was related with the main clinical objective defined. Therefore, the three main objectives formulated in the context of joint modelling could be related with three main clinical research questions: (i) to evaluate the impact of albumin level on the combined survival, given that albumin level were recorded with measurement error; (ii) to analyse longitudinal evolution of albumin clinical parameter, given that lower levels of albumin may be associated with higher risk of mortality and morbidity (consequently less fit for renal transplant), i.e., in the presence of informative censoring; (iii) to evaluate the association between the progression of albumin level and combined survival and the identification of factors that influenced both outcomes.

In this paper, two parameterisations of a random shared effects joint model were compared considering an example in peritoneal dialysis. These two approaches are focused in distinct aspects. The parameterisation implemented in JM focuses mainly on the influence of a longitudinal variable measured with error in the estimation of the survival submodel. In this case, it is possible to quantify the effect of the longitudinal outcome in the survival hazard. On the other hand, the parameterisation implemented in *joineR* focuses mainly on the link between the processes, considering shared latent random effects to represent the correlation between longitudinal and survival process [14]. For this reason, the evaluation of the effect of an unobserved condition, shared between longitudinal and survival, in the hazard is possible using this parameterisation.

The two parameterisations presented provided complementary conclusions, given that they have different focus/objectives. The JM package was used to build a joint model when the focus is on a patient's survival and the inaccuracies in estimating albumin score. The *joineR* package was used to investigate the effect of a patient's changing albumin levels linking the longitudinal and survival processes through latent random effects. Although the two parameterisations present some differences relatively to the formulation, the modelling method of the baseline function and the survival submodel, the results had shown an evident relationship between the two processes in both approaches. This fact justifies the need for a joint modelling approach, and the advantages of the use of this methodology is highlighted when comparing results with separate analysis. Different conclusions were obtained considering separate analysis or a joint analysis, as shown in the previous section. Considering independent approaches, the focus is on the effect on parameters estimates and their standards errors ignoring the link between the longitudinal and survival processes and the longitudinal response measured with error within the survival process [14]. For separate analysis the effects of covariates that are significant, became not significant when a joint analysis approach is done. This might be due to variability that is being overestimated in a separate analysis, which is due to association between the two processes, longitudinal and survival. When this is taking into account this effect disappears.

In conclusion, joint modelling for longitudinal and time-to-event outcomes

in the presence of competing risks is useful in different areas of applications when the interest is the evaluation of the relationship between these two types of outcomes. In clinical studies diverse information about the patient is collected along a disease stages or treatment duration, and these models become an appropriate approach. Then it is necessary to alert clinicians for the implications and the advantages of a proper data collection and a correct data analysis.

---

## ACKNOWLEDGMENTS

---

The Authors would like to thanks R. Kolamunnage-Dona and D. Rizopoulos for the R codes provided.

---

## REFERENCES

---

- [1] ANDRINOPOULOU, E.; RIZOPOULOS, D.; JIN, R.; BOGERS, A.; LESAFFRE, E. and TAKKENBERG, J. (2012). An introduction to mixed models and joint modeling: Analysis of valve function over time, *Annals of Thoracic Surgery*, **93**, 6, 1765–1772.
- [2] BARRETT, J.; AND DIGGLE, P.; HENDERSON, R. and TAYLOR-ROBINSON, D. (2015). Joint modelling of repeated measurements and time-to-event outcomes: Flexible model specification and exact likelihood inference, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **77**, 1, 131–148.
- [3] COX, D.R. (1972). Regression models and life-tables, *Journal of the Royal Statistical Society: Series B (Methodological)*, **77**, 187–220.
- [4] DIGGLE, P.; HEAGERTY, P.; LIANG, K.-Y. and ZEGER, S.L. (2002). *Analysis of Longitudinal Data*, 2nd edition, Oxford University Press, Oxford.
- [5] DIGGLE, P.J.; SOUSA, I. and CHETWYND, A.G. (2008). Joint modelling of repeated measurements and time-to-event outcomes: The fourth Armitage lecture, *Statistics in Medicine*, **27**, 16, 2981–2998.
- [6] ELASHOFF, R.M.; LI, G. and LI, N. (2007). An approach to joint analysis of longitudinal measurements and competing risks failure time data, *Statistics in Medicine*, **26**, 14, 2813–2835.
- [7] ELASHOFF, R.M.; LI, G. and LI, N. (2008). A joint model for longitudinal measurements and survival data in the presence of multiple failure types, *Biometrics*, **64**, 3, 762–771.
- [8] GRAY, R.J. (1988). A class of K-sample tests for comparing the cumulative incidence of a competing risk, *The Annals of Statistics*, **16**, 3, 1141–1154.
- [9] GUO, X. and CARLIN, B.P. (1988). Separate and joint modeling of longitudinal and event time data using standard computer packages, *The American Statistician*, **58**, 1, 16–24.

- [10] HENDERSON, R.; DIGGLE, P. and DOBSON, A. (2000). Joint modelling of longitudinal measurements and event time data, *Biostatistics*, **1**, 4, 465–480.
- [11] HICKEY, G.L.; PHILIPSON, P.; JORGENSEN, A. and KOLAMUNNAGE-DONA, R. (2018). A comparison of joint models for longitudinal and competing risks data, with application to an epilepsy drug randomized controlled trial, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **181**, 4, 1105–1123.
- [12] HOGAN, J.W. and LAIRD, N.M. (2018). Model-based approaches to analysing incomplete longitudinal and failure time data, *Statistics in Medicine*, **16**, 3, 259–272.
- [13] LI, N.; ELASHOFF, R.M. and LI, G. (2009). Robust joint modeling of longitudinal measurements and competing risks failure time data, *Biometrical Journal*, **51**, 1, 19–30.
- [14] MCCRINK, L.M.; MARSHALL, A.H. and CAIRNS, K.J. (2013). Advances in joint modelling: a review of recent developments with application to the survival of end stage renal disease patients, *International Statistical Review*, **81**, 2, 249–269.
- [15] MEHROTRA, R.; DUONG, U.; JIWAKANON, S.; KOVESDY, C.P.; MORAN, J.; KOPPLE, J.D. and KALANTAR-ZADEH, K. (2011). Serum albumin as a predictor of mortality in peritoneal dialysis: Comparisons with hemodialysis, *American Journal of Kidney Diseases*, **58**, 3, 418–428.
- [16] NEUHAUS, A.; AUGUSTIN, T.; HEUMANN, C. and DAUMER, D. (2009). A review on joint models in biometrical research, *Journal of Statistical Theory and Practice*, **3**, 4, 855–868.
- [17] PHILIPSON, P.; DIGGLE, P.; SOUSA, I.; KOLAMUNNAGE-DONA, R.; WILLIAMSON, P. and HENDERSON, R. (2012). *joiner*: Joint modelling of repeated measurements and time-to-event data, R package version 1.0.1.
- [18] PROUST-LIMA, C.; SÉNE, M.; TAYLOR, J.M.G. and JACQMIN-GADDA, H. (2014). Joint latent class models for longitudinal and time-to-event data: A review, *Statistical Methods in Medical Research*, **23**, 1, 74–90.
- [19] RIZOPOULOS, D. (2012). *Joint Models for Longitudinal and Time-To-Event Data: With Applications in R*, CRC Press, Boca Raton.
- [20] SAVILLE, B.; HERRING, A. and KOCH, G. (2010). A robust method for comparing two treatments in a confirmatory clinical trial via multivariate time-to-event methods that jointly incorporate information from longitudinal and time-to-event data, *Statistics in Medicine*, **29**, 1, 75–85.
- [21] SOUSA, I. (2011). A review on joint modelling of longitudinal measurements and time-to-event, *REVSTAT - Statistical Journal*, **9**, 1, 57–81.
- [22] TEIXEIRA, L.; RODRIGUES, A.; CARVALHO, M.J.; CABRITA, A. and MENDONÇA, D. (2013). Modelling competing risks in nephrology research: an example in peritoneal dialysis, *BMC Nephrology*, **14**, 1, 110.
- [23] TSIATIS, A.A. and DAVIDIAN, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview, *Statistica Sinica*, **14**, 3, 809–834.
- [24] WALDMANN, E.; TAYLOR-ROBINSON, D.; KLEIN, N.; KNEIB, T.; PRESSLER, T.; SCHMID, M. and MAYR, A. (2017). Boosting joint models for longitudinal and time-to-event data, *Biometrical Journal*, **59**, 6, 1104–1121.
- [25] WILLIAMSON, P.R.; KOLAMUNNAGE-DONA, R.; PHILIPSON, P. and MARSON, A.G. (2008). Joint modelling of longitudinal and competing risks data, *Statistics in Medicine*, **27**, 30, 6426–6438.

# REVSTAT – STATISTICAL JOURNAL

## Background

Statistics Portugal (INE, I.P.), well aware of how vital a statistical culture is in understanding most phenomena in the present-day world, and of its responsibility in disseminating statistical knowledge, started the publication of the scientific statistical journal *Revista de Estatística*, in Portuguese, publishing three times a year papers containing original research results, and application studies, namely in the economic, social and demographic fields.

In 1998 it was decided to publish papers also in English. This step has been taken to achieve a larger diffusion, and to encourage foreign contributors to submit their work.

At the time, the Editorial Board was mainly composed by Portuguese university professors, being now composed by national and international university professors, and this has been the first step aimed at changing the character of *Revista de Estatística* from a national to an international scientific journal.

In 2001, the *Revista de Estatística* published three volumes special issue containing extended abstracts of the invited contributed papers presented at the 23rd European Meeting of Statisticians.

The name of the Journal has been changed to REVSTAT - STATISTICAL JOURNAL, published in English, with a prestigious international editorial board, hoping to become one more place where scientists may feel proud of publishing their research results.

- The editorial policy will focus on publishing research articles at the highest level in the domains of Probability and Statistics with emphasis on the originality and importance of the research.
- All research articles will be refereed by at least two persons, one from the Editorial Board and another external.



— The only working language allowed will be English. — Four volumes are scheduled for publication, one in January, one in April, one in July and the other in October.

### **Aims and Scope**

The aim of REVSTAT is to publish articles of high scientific content, in English, developing innovative statistical scientific methods and introducing original research, grounded in substantive problems.

REVSTAT covers all branches of Probability and Statistics. Surveys of important areas of research in the field are also welcome.

### **Abstract and Indexing Services**

The REVSTAT is covered by the following abstracting/indexing services:

- Current Index to Statistics
- Google Scholar
- Mathematical Reviews
- Science Citation Index Expanded
- Zentralblatt für Mathematic

### **Instructions to Authors, special-issue editors and publishers**

The articles should be written in English and may be submitted in two different ways:

- By sending the paper in PDF format to the Executive Editor (revstat@ine.pt) and to one of the two Editors or Associate Editors, whose opinion the author wants to be taken into account, together to the following e-mail address: revstat@fc.ul.pt

- By sending the paper in PDF format to the Executive Editor (revstat@ine.pt), together with the corresponding PDF or PostScript file to the following e-mail address: revstat@fc.ul.pt.

Submission of a paper means that it contains original work that has not been nor is about to be published elsewhere in any form.

Manuscripts (text, tables and figures) should be typed only in black on one side, in double-spacing, with a left margin of at least 3 cm and with less than 30 pages. The first page should include the name, institution and address of the author(s) and a summary of less than one hundred words, followed by a maximum of six key words and the AMS 2000 subject classification.

Authors are obliged to write the final version of accepted papers using LaTeX, in the REVSTAT style. This style (REVSTAT.sty), and examples file (REVSTAT.tex), which may be download to PC Windows System (Zip format), Macintosh, Linux and Solaris Systems (StuffIt format), and Mackintosh System (BinHex Format), are available in the REVSTAT link of the Statistics Portugal Website: <http://www.ine.pt/revstat/inicio.html>

Additional information for the authors may be obtained in the above link.

### **Accepted papers**

Authors of accepted papers are requested to provide the LaTeX files and also a postscript (PS) or an acrobat (PDF) file of the paper to the Secretary of REVSTAT: revstat@ine.pt.

Such e-mail message should include the author(s)'s name, mentioning that it has been accepted by REVSTAT.

The authors should also mention if encapsulated postscript figure files were included, and submit electronics figures separately in .tiff, .gif, .eps or .ps format. Figures must be a minimum of 300 dpi.

## **Copyright**

Upon acceptance of an article, the author(s) will be asked to transfer copyright of the article to the publisher, Statistics Portugal, in order to ensure the widest possible dissemination of information, namely through the Statistics Portugal website (<http://www.ine.pt>).

After assigning copyright, authors may use their own material in other publications provided that REVSTAT is acknowledged as the original place of publication. The Executive Editor of the Journal must be notified in writing in advance.

# Editorial Board

## Editor-in-Chief

**Isabel Fraga Alves**, University of Lisbon, Portugal

## Co-Editor

**Giovani L. Silva**, University of Lisbon, Portugal

## Associate Editors

**Marília Antunes**, University of Lisbon, Portugal

**Barry Arnold (2019)**, University of California, USA

**Narayanaswamy Balakrishnan**, McMaster University, Canada

**Jan Beirlant**, Katholieke Universiteit Leuven, Belgium

**Graciela Boente (2019-2020)**, University of Buenos Aires, Argentina

**Paula Brito**, University of Porto, Portugal

**Vanda Inácio de Carvalho**, University of Edinburgh, UK

**Arthur Charpentier**, Université du Québec à Montréal, Canada

**Valérie Chavez-Demoulin**, University of Lausanne, Switzerland

**David Conesa**, University of Valencia, Spain

**Charmaine Dean**, University of Waterloo, Canada

**Jorge Milhazes Freitas**, University of Porto, Portugal

**Alan Gelfand**, Duke University, USA

**Stéphane Girard**, Inria Grenoble Rhône-Alpes, France

**Wenceslao Gonzalez-Manteiga**, University of Santiago de Compostela, Spain

**Marie Kratz**, ESSEC Business School, France

**Victor Leiva**, Pontificia Universidad Católica de Valparaíso, Chile

**Maria Nazaré Mendes-Lopes**, University of Coimbra, Portugal

**Fernando Moura**, Federal University of Rio de Janeiro, Brazil

**John Nolan**, American University, USA

**Paulo Eduardo Oliveira**, University of Coimbra, Portugal

**Pedro Oliveira**, University of Porto, Portugal

**Carlos Daniel Paulino (2019-2021)**, University of Lisbon, Portugal

**Arthur Pewsey**, University of Extremadura, Spain

**Gilbert Saporta**, Conservatoire National des Arts et Métiers, France

**Alexandra M. Schmidt**, McGill University, Canada

**Julio Singer**, University of Sao Paulo, Brazil

**Manuel Scotto**, University of Lisbon, Portugal

**Lisete Sousa**, University of Lisbon, Portugal

**Milan Stehlík**, University of Valparaíso, Chile and LIT-JK University Linz, Austria

**María Dolores Ugarte**, Public University of Navarre, Spain

## Executive Editor

**José A. Pinto Martins**, Statistics Portugal

## Former Executive Editors

**Maria José Carrilho**, Statistics Portugal (2005-2015)

**Ferreira da Cunha**, Statistics Portugal (2003–2005)

## Secretary

**Liliana Martins**, Statistics Portugal