XXII CONGRESSO SPE 2015
SOCIEDADE PORTUGUESA DE ESTATÍSTICA

Estatística: Progressos e Aplicações

Atas do XXII Congresso da Sociedade Portuguesa de Estatística

Editores:
Clara Cordeiro, Conceição Ribeiro, Carlos Sousa, Maria Helena Gonçalves, Nelson Antunes e Maria Eduarda Silva.

UAlg
UNIVERSIDADE DO ALGARVE

SPE
SOCIEDADE PORTUGUESA DE ESTATÍSTICA

ESTATÍSTICA:
Progressos e Aplicações

Atas do XXII Congresso da
Sociedade Portuguesa de Estatística

Olhão, 07 a 10 de outubro de 2015

Editores
Clara Cordeiro
Conceição Ribeiro
Carlos Sousa
Maria Helena Gonçalves
Nelson Antunes
Maria Eduarda Silva

Novembro, 2016
Edições SPE

Editores: Clara Cordeiro, Conceição Ribeiro, Carlos Sousa, Maria Helena Gonçalves, Nelson Antunes e Maria Eduarda Silva

Título: Estatística: Progressos e Aplicações
Atas do XXII Congresso da Sociedade Portuguesa de Estatística

Editora: Sociedade Portuguesa de Estatística

Conceção Gráfica da Capa: Ludovico Silva, Gabinete de Comunicação e Protocolo da Universidade do Algarve

Impressão: Instituto Nacional de Estatística

Tiragem: 200 Exemplares

ISBN: 978-972-8890-39-1

Depósito Legal: 417937/16

Análise de Sobrevivência e Valores Extremos em R

Ana Maria Abreu
FCEE, Universidade da Madeira e CIMA, *abreu@staff.uma.pt*

Délia Gouveia-Reis
FCEE, Universidade da Madeira e CEAUL, *delia@uma.pt*

Palavras–chave: Análise de sobrevivência, *software* R, valores extremos

Resumo: O *software* R é uma ferramenta extremamente útil para a investigação estatística. No entanto, a proliferação de bibliotecas (na ordem dos milhares) dificulta o rápido e eficiente acesso a todas as possibilidades em cada uma das áreas desta ciência. Uma forma de limitar esta procura é aceder à *task view* correspondente, se existir. Pelos motivos descritos, neste trabalho procura-se compilar informação relevante nas áreas de análise de sobrevivência e de valores extremos, de modo a minimizar as dificuldades referidas. A abordagem na análise de sobrevivência, que possui *task view*, será sobretudo através de exemplos. Nos valores extremos será dada uma visão geral do que existe, uma vez que nesta área não há *task view*.

1 Introdução

O R é uma linguagem que surge pela criação da *R Foundation for Statistical Computing* [7], com o objetivo de fornecer uma ferramenta gratuita e de utilização livre, para o tratamento e análise de dados e para a construção de gráficos. Em 1993, Robert Gentleman e Ross Ihaka, na Universidade de Auckland, deram origem à linguagem R e tornaram-na *open source* em 1995.
O R é uma ferramenta bastante abrangente, com boas capacidades ao nível da programação e um conjunto bastante vasto (e em cons-

tante crescimento) de bibliotecas (livrarias) que acrescentam inúmeras potencialidades à já poderosa versão base do R. O *download* do R é gratuito e pode ser feito a partir da página principal do *R Project for Statistical Computing* em http://www.r-project.org/ ou do *Comprehensive R Archive Network* (CRAN) em http://cran.r-project.org/. Uma biblioteca muito utilizada é o *R Commander* (abreviadamente Rcmdr, desenvolvido em 2003, por John Fox) pois possui um interface gráfico que torna a interação com o utilizador muito mais amigável do que na consola do R. Além de possuir menus, permite a escrita de código e engloba as restantes funcionalidades existentes no R original. Existem ainda os *plugins* do *R Commander* que adicionam funcionalidades aos menus.

Uma forma de tornar eficiente a utilização do R, consiste em aceder à *task view* correspondente à área em estudo, se existir, pois estas *task views* são excelentes guias para encontrar as bibliotecas e funções adequadas ao propósito do investigador. Atualmente existem 33 *task views*, abrangendo áreas tão diversas como inferência bayesiana, ensaios clínicos, genética, otimização e programação matemática, análise de sobrevivência, séries temporais, entre outras. Contudo, ainda há várias áreas para as quais não existe esta funcionalidade, como sejam, valores extremos, análise em componentes principais, modelos com equações estruturais ou controlo de qualidade. Assim, neste trabalho irá ser feita uma breve revisão das bibliotecas existentes para a análise de sobrevivência e para os valores extremos, procurando contribuir para um eficiente acesso às potencialidades do R nestas áreas. A abordagem na análise de sobrevivência ([5, 9]) será sobretudo através de exemplos. Nos valores extremos ([1, 3]) será dada uma visão geral do que existe, uma vez que nesta área não há *task view*.

2 Análise de Sobrevivência

A análise de sobrevivência é uma das áreas da estatística que possui *task view* no R (https://cran.r-project.org/web/views/Survival.html),

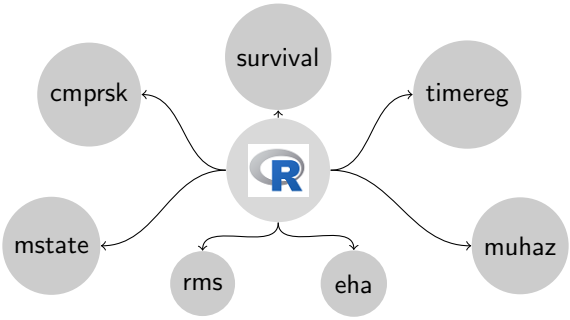

Figura 1: Bibliotecas *core*.

a qual se encontra organizada por temas e por ordem alfabética das bibliotecas. Importa notar desde logo que esta *task view* não esgota todas as funcionalidades que existem no R para esta área, mas cobre a maior parte. Allignol e Latouche (os responsáveis pela manutenção da *task view*) identificam sete bibliotecas *core* (Figura 1), sendo a survival, a rms e a eha as mais abrangentes.

Para além destes, há muitos outros (mais de cem), cujas particularidades seriam impossíveis de enumerar aqui de forma suficientemente abreviada. De qualquer modo, a já referida organização por temas permite uma escolha rápida da biblioteca apropriada para o objetivo pretendido como, por exemplo, a estimação da função de sobrevivência, a realização de testes ou a obtenção de modelos de regressão (bibliotecas survival, rms, eha e timereg) ou ainda a estimação da função de risco (bibliotecas rms, eha e muhaz). As bibliotecas cmprsk e mstate são mais específicas; referem-se, respetivamente, a modelos de riscos competitivos e a modelos multiestado.

O que se segue são pequenos exemplos de representações gráficas usuais (função de sobrevivência e função de risco) com alguns pormenores extra. A amostra aleatória, de dimensão 100, usada nos exemplos que se seguem, foi gerada da seguinte forma: os tempos de vida através da distribuição exponencial de parâmetro $\lambda = 1$, a

censura através da distribuição uniforme no intervalo 0.5 e 2, a idade através da distribuição normal de parâmetros $\mu = 60$ e $\sigma = 8$, e os estratos através de uma amostra aleatória de valores de 1 a 3. O respetivo código é o que se segue:

```
set.seed(123)
stime < − rexp(100) * 10
cens < − runif(100,.5,2) * 10
sevent < − as.numeric(stime <= cens)
stime < − pmin(stime, cens)
strat < − sample(1:3, 100, replace=TRUE)
idade < − rnorm(100,60,8)
dd < − data.frame("surv.time"=stime, "surv.event"=sevent, "strat"=strat,
"idade"=idade)
ddweights < − array(1, dim=nrow(dd))
```

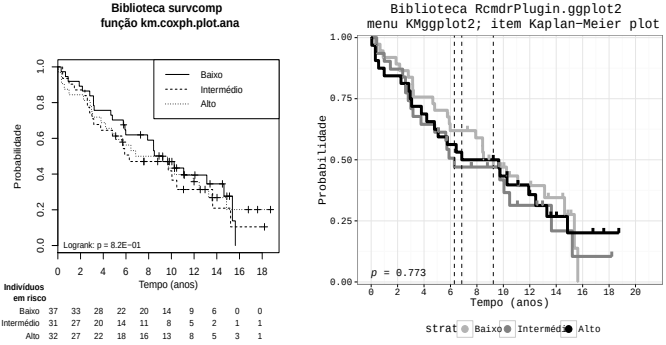

Figura 2: Estimativas de Kaplan-Meier da função de sobrevivência.

Exemplo 2.1 *Tendo presente que qualquer gráfico no R pode ser melhorado através da alteração do seu código base, a Figura 2 exibe dois gráficos da estimativa de Kaplan-Meier com uma alteração mínima do seu código (essencialmente a tradução para português).*

A particularidade da biblioteca survcomp *diz respeito não apenas ao facto de exibir o valor de prova resultante do teste logrank para a igualdade das funções de sobrevivência, mas sobretudo por mostrar o número de indivíduos em risco, para cada categoria, nos valores indicados na escala do tempo. Já no que diz respeito ao plugin* RcmdrPlugin.KMggplot2 *da biblioteca* RCommander *(que também indica o valor p), a principal inovação é a indicação do valor da mediana do tempo de vida através das retas verticais.*

Outro tipo de gráfico muito útil na análise de sobrevivência é o da função de risco pois, além de descrever o risco ao longo do tempo, ajuda na escolha da distribuição para a variável aleatória que representa o tempo de vida. O Exemplo 2.2 refere-se a duas representações desta função.

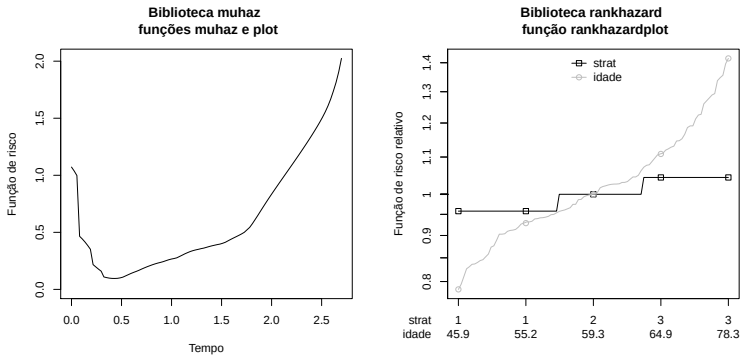

Figura 3: Funções de risco.

Exemplo 2.2 *A biblioteca* muhaz *permite que a representação da função de risco seja feita de um modo bastante simples (Figura 3), recorrendo apenas às funções* muhaz *e* plot*.*
Uma outra perspetiva interessante e inovadora da função de risco, é a que se obtém através da função rankhazardplot*, fornecida pela*

biblioteca rankhazard*, [4]. Esta abordagem indica o risco não ao longo do tempo mas ao longo dos valores das covariáveis presentes no modelo. Assim, num mesmo gráfico é possível observar o risco relativo para cada covariável e representar os valores das covariáveis no eixo horizontal (mínimo,* Q_1*, mediana,* Q_3 *e máximo). Concretamente, em relação às covariáveis apresentadas, verifica-se que a diferença entre os níveis 1 ("Baixo") e 2 ("Intermédio") é sensivelmente a mesma que entre os níveis 2 ("Intermédio") e 3 ("Alto") da covariável "strat" e que, em relação à covariável "idade", o risco diminui com a idade. Note-se que, quando a covariável é quantitativa, o valor de referência (correspondente ao indivíduo padrão) que é considerado por regra é o correspondente à mediana, como acontece com a covariável "idade".*

Por último, mas não menos importante, apresenta-se um outro gráfico de utilização frequente pois permite uma análise visual preliminar da proporcionalidade das funções de risco.

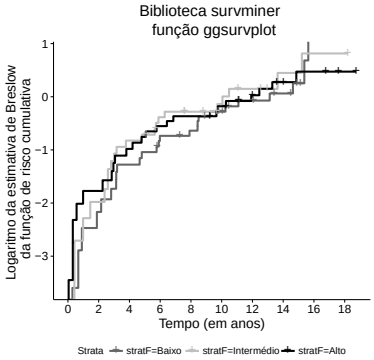

Figura 4: Logaritmo da estimativa de Breslow da função de risco cumulativa para os três estratos.

Exemplo 2.3 *Na Figura 4, o tempo é representado no eixo das abcissas e o logaritmo da estimativa de Breslow da função de risco*

cumulativa no eixo das ordenadas. Este gráfico é obtido duma forma simples através da biblioteca survminer*, utilizando a função* ggsurvplot*. O cruzamento das curvas dá indicação de não haver proporcionalidade das correspondentes funções de risco.*

As particularidades de natureza mais algébrica não foram aqui abordadas. No entanto, só a título de exemplo, refira-se que a função ConvertWeibull da biblioteca SurvRegCensCov faz a conversão da parametrização das estimativas dos parâmetros do modelo Weibull.

3 Valores Extremos

A crescente aplicabilidade da teoria dos valores extremos já era bem visível em 2006, ano no qual surgiu a elaboração de um estado da arte [10] sobre o *software* utilizado no estudo de valores extremos. Os seus autores, Stephenson e Gilleland, além de outros *softwares*, indicam algumas bibliotecas do R tais como ismev, evir, evd e evdBayes. Além disso, apresentam também a biblioteca extRemes como sendo essencialmente um interface gráfico do ismev e referem que muitas das funções da biblioteca fExtremes se baseiam em funções das bibliotecas ismev, evir e evd. Mais recentemente, Gilleland *et al.* [2] direcionam essa escolha para o R, pelo facto de ser o *software* que continha, em 2013, a maior variedade de metodologias na área de valores extremos. De entre estas metodologias, a dos máximos anuais foi a escolhida no estudo efetuado por Penalva *et al.* [6], para exemplificar uma análise de valores extremos no R. Nesse estudo, as autoras mencionam as bibliotecas evir, fExtremes e evdBayes e realizam uma descrição das bibliotecas ismev e evd. Até ao momento não existe qualquer *task view* exclusivamente dedicada à teoria dos valores extremos que facilite o acesso às bibliotecas e funcionalidades apropriadas, mas as *task views* Bayesian, Distributions, Environmetrics, Finance, Spatial incluem pelo menos uma das bibliotecas indicadas na Figura 5.

A existência de algumas funções relativas à teoria dos valores extremos motivou a referência de outras bibliotecas tanto nos esta-

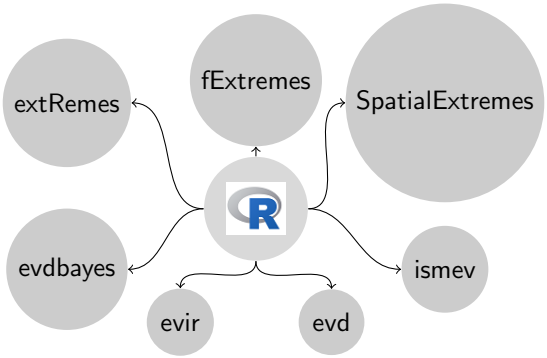

Figura 5: Bibliotecas em *task views*.

dos da arte já referidos como no *website* mantido por Eric Gilleland (http://www.ral.ucar.edu/ ericg/softextreme.php), embora essas bibliotecas não sejam exclusivamente dedicadas à aplicação de metodologias nesta área. Alguns exemplos são as bibliotecas copula, fitdistrplus e RandomFields. A biblioteca fitdistrplus estabelece a ligação mais visível entre as áreas da análise de sobrevivência e dos valores extremos. De facto, ao carregar a biblioteca fitdistrplus, automaticamente também é carregada a biblioteca survival. Por outro lado existem também bibliotecas no R destinadas a conteúdos muito específicos na área de valores extremos tais como as bibliotecas bgeva e spatial.gev.bma. A primeira disponibiliza uma função para modelos de regressão para extremos bivariados enquanto que a segunda permite ajustar um modelo espacial hierárquico a valores extremos. Apesar de mais gerais, as bibliotecas evt0, evmix, MCMC4Extremes e Renext estão focalizadas em certas metodologias específicas da área de valores extremos. A biblioteca evt0 é a única que aborda a metodologia PORT (*Peaks Over Random Threshold*) de entre as bibliotecas do R (pelo menos do conhecimento das autoras). Esta biblioteca, não indicada no *website* mantido por Eric Gilleland, é um produto da escola portuguesa de valores extremos que permite determinar o

índice de valores extremos γ por meio do estimador MOP (média de ordem p). Além disso, esta biblioteca (que requer a biblioteca evd) permite também obter as estimativas para γ pelos estimadores dos momentos, dos momentos mistos e generalizado de Hill. A biblioteca evmix fornece funções para a modelação mista de valores extremos, para a estimação do limiar u e para estimadores de densidade pelo método do núcleo. Apesar de esta biblioteca não requerer qualquer das bibliotecas mencionadas, os seus criadores indicam que existe uma razoável consistência com as funções base da biblioteca evd. O Exemplo 3.1 refere-se a uma dessas funções, cuja aplicação originou os dois gráficos da Figura 6.

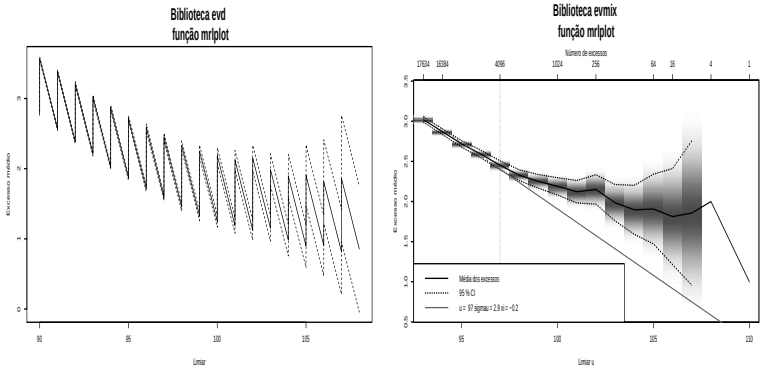

Figura 6: Gráficos de vida residual média.

Exemplo 3.1 *A Figura 6 exibe dois gráficos de vida residual média com uma alteração mínima do seu código (tradução para português e alteração da escala de cores). Os dados utilizados, já analisados por outros autores no contexto da metodologia POT [8], correspondem a idades de mulheres nascidas por volta do ano 1900 que morreram no ano de 1993.*

A particularidade da função mrlplot *da biblioteca* evmix *diz respeito não apenas ao facto de exibir um eixo superior com o número de*

excessos, mas sobretudo por mostrar um valor de referência para o limiar u. Além da indicação gráfica (linha vertical) e numérica de um limiar u, as estimativas de máxima verosimilhança dos parâmetros de escala e forma da correspondente distribuição generalizada de Pareto são também apresentadas. Esta informação pode ainda ser apresentada para três valores a considerar para o limiar u, facilitaando a interpretação do gráfico e a comparação de estimativas. A título exemplificativo, na Figura 7 é também indicado o valor sugerido para o limiar u pelos autores Reiss e Thomas [8].

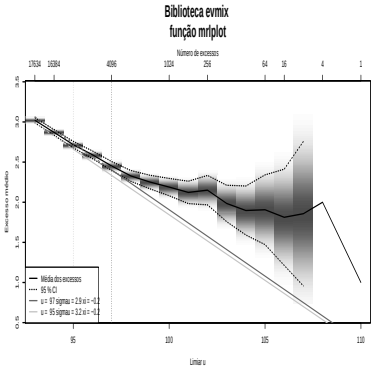

Figura 7: Gráfico de vida residual média com $u = 95$ e $u = 97$.

A biblioteca MCMC4Extremes partilha a metodologia Bayesiana com a biblioteca menos recente evdBayes mas requer a biblioteca evir em vez da biblioteca evd. A biblioteca Renext apareceu em 2010 e inclui a implementação de funções relativas ao denominado *méthode du renouvellement*. Esta abordagem surge como uma generalização da abordagem clássica POT (*Peaks Over Threshold*) ao permitir que os excessos em relação a um dado limiar u sigam uma distribuição de probabilidade diferente da distribuição de Pareto. Para esta biblioteca, que também requer a biblioteca evd, existe um interface gráfico denominado de RenextGUI.

A biblioteca extRemes, que inicialmente foi interface gráfico da biblioteca ismev, tornou-se numa biblioteca de valores extremos por si só, tendo sido criado um interface gráfico para algumas das suas funções, denominado de in2extRemes, cujo tutorial pode ser encontrado no link http://www.ral.ucar.edu/staff/ericg/extRemes. Outra biblioteca é a intitulada *texmex* (*Statistical modelling of extreme values*), que pode ser utilizada na modelação quer de máximos quer de excessos. Além disso, esta biblioteca tem a particularidade de ter funcionalidades específicas da abordagem bayesiana e da análise multivariada de valores extremos. Muito recentemente (fevereiro de 2016), surgiu uma nova biblioteca intitulada *Multivariate Extreme Value Distributions* (mev), a qual é inteiramente direcionada para o estudo dos valores extremos multivariados. Além da implementação de métodos de seleção do limiar, esta biblioteca permite ainda a simulação de processos max-estáveis. A biblioteca eva (*Extreme Value Analysis with Goodness-of-Fit Testing*), que surgiu também recentemente (dezembro de 2015), possui a particularidade de incluir testes de ajustamento para a escolha tanto do limiar u na metodologia de excessos de nível, como do número de observações k na metodologia das maiores observações.
Além das bibliotecas já mencionadas, existem outras mais gerais que englobam várias áreas da estatística e probabilidades, entre as quais a análise dos valores extremos (lmomco e VGAM, por exemplo). Existem ainda outras mais focalizadas numa área de aplicação em que recorrem a esta teoria como sejam, por exemplo, as bibliotecas actuar e QRM.

4 Conclusão

O R é tão dinâmico que qualquer trabalho sobre ele é inevitavelmente incompleto e um pouco desatualizado. Exemplo disso é o facto de à data do início da escrita deste artigo (junho de 2015) haver 6730 bibliotecas e atualmente (setembro de 2016) já haver 9202. Mas precisamente por essa razão, entende-se que uma sistematização

periódica por áreas pode se revelar útil, de modo a que as inúmeras potencialidades deste *software* possam ser plenamente aproveitadas. Embora a biblioteca Rcmdr não conste na *task view* para a análise de sobrevivência, recomenda-se o seu uso pois assim é possível usar os *plugins* relativos a esta área que lhe estão associados: RcmdrPlugin.survival, RcmdrPlugin.KMggplot2, RcmdrPlugin.EZR e RcmdrPlugin.NMBU. Esta abordagem torna a interação mais simples e mantém a mais valia da escrita do código, realidade que é válida para as restantes áreas. A existência da *task view* é uma grande vantagem pois rapidamente se identificam as bibliotecas existentes para áreas específicas, como sejam, por exemplo, modelos multiestado, sobrevivência relativa, modelos de efeitos aleatórios, modelos bayesianos, simulação, entre outros. Além disso, todos as bibliotecas podem ser instaladas simultaneamente, em vez de uma a uma, bastando para tal instalar a biblioteca ctv e proceder de acordo com as instruções existentes em https://cran.r-project.org/web/views/. Neste breve trabalho, tentou-se mostrar algumas das peculiaridades que distinguem este *software* de outros mais comerciais, através das inovações que apresenta nos gráficos mais usuais desta área. No entanto, as particularidades de natureza mais algébrica não foram abordadas.

Neste trabalho, fez-se também uma revisão das bibliotecas do R que podem ser aplicadas na análise de valores extremos. Procurou-se assim organizar uma coletânea de informação sobre estas bibliotecas, tendo como linhas de orientação a sua abrangência relativamente às metodologias da área, as suas interligações e as suas particularidades. Nesta área recomenda-se igualmente a biblioteca Rcmdr em detrimento dos interfaces gráficos in2extRemes e RenextGUI pois permite a utilização simultânea de uma ou mais das bibliotecas mencionadas num mesmo ambiente amigável. Se em 2013 o R já era o *software* que continha a maior variedade de metodologias na área de valores extremos, atualmente esse facto é ainda mais evidente dado o surgimento de novas bibliotecas, bem como o aperfeiçoamento das já existentes. Seria pois bastante útil reunir e organizar as bibliotecas do R sobre análise de valores extremos numa *task view*, segundo

tópicos que permitissem um fácil acesso e manuseamento da grande quantidade de funcionalidades existentes. Essa é a nossa proposta de trabalho futuro.
Em conclusão, trabalhar com o R é estar preparado para uma constante descoberta, acompanhada por muitos momentos de satisfação intercalados por alguns de frustração.

Agradecimentos

Este trabalho é parcialmente financiado por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito dos projetos UID/MAT/04674/2013 (CIMA) e UID/MAT/00006/2013 (CEAUL).

Referências

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

[2] Gilleland, E., Ribatet, M., Stephenson, A. (2013). A software review extreme value analysis. *Extremes* 16, 103–119.

[3] Gomes, M.I., Fraga Alves, M.I., Neves, C. (2013). *Análise de Valores Extremos: uma Introdução*. Edições SPE, Lisboa.

[4] Karvanen, J., Harrell Jr., F.E. (2009). Visualizing covariates in proportional hazards model. *Statistics in Medicine* 28, 1957–1966.

[5] Klein, J.P., Moeschberger, M.L. (1998). *Survival Analysis. Techniques for Censored and Truncated Data*, 2^a impressão. Springer, New York.

[6] Penalva, H., Neves, M., Nunes, S. (2013). Topics in data analysis using R in extreme value theory. *Metodološki zvezki* 10, 17–29.

[7] R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, URL: http://www.R-project.org/

[8] Reiss, R.D., Thomas, M. (2007). *Statistical Analysis of Extreme Values: With Applications to Insurance, Finance, Hydrology and Other Fields*. Birkhäuser, Basel.

[9] Rocha, C., Papoila, A.L. (2009). *Análise de Sobrevivência*. Edições SPE, Lisboa.

[10] Stephenson, A., Gilleland, E. (2006). Software for the analysis of extreme events: The current state and future directions. *Extremes* 8, 87–109.