ISSN 1645-6726

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

REVSTAT
Statistical Journal

Special issue on

«Risk Analysis: Challenges and Applications»

Guest Editors:
Christos P. Kitsos
Teresa A. Oliveira
Milan Stehlík

Volume 14, No.2
April 2016

REVSTAT

Statistical Journal

Catalogação Recomendada

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

CREDITS

- **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
    (National network) : 808 201 808
    Other networks: + 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**
  - *150 copies*
- **LEGAL DEPOSIT REGISTRATION**
  - *N.º 191915/03*
- **PRICE [VAT included]**
  - *€ 9,00*

FOREWORD

This special issue of *Revstat — Statistical Journal* presents selected papers on Risk Analysis, discussing recent developments, challenges and applications in several areas. For the development of Risk Analysis the year 1983 deserves a special notation as it was published that particular year the (i) British Royal Society Risk Assessment: Report of a Royal Society Study Group, Royal Society, London, and (ii) National Research Council, Risk assessment in the Federal Government: Managing the Process, National Academy Press, Washington, DC. At the first stage of Risk Assessment development, it was the typical way to work and study mainly human cancer risk (see Edler L., Kitsos, C.P (2005) for more than 1000 references listed mainly for Human Risk on Cancer). Still there are some particular differences between the Risk Assessment as it is developed in the European Union and the way United States defines it. The U.S Environmental Protection Agency (E.P.A) has published a number of valuable Guidelines for Carcinogen Risk Assessment since 1985. From that time until currently, the topic of Risk Analysis has been assisting an increasing popularity, offering many challenges in prudent Risk Assessment. The management of Risks to Human Health is based on two principles:

1. A Risk is accepted if it is sufficiently small to be considered in some way null or negligible known as the "de minimis principle".
2. The Risk benefit balance principle: which implies that a Risk is accepted if the obtained benefit largely justifies its acceptance.

The terms *Risk and Hazard* need to be clarified, sometimes, before their use. Certainly are different and are adopted for different qualitative methods. Needless to say, Risk Analysis comprises a number of Life Sciences, as well as Chemistry (and therefore Environmental and Food Science Problems, and Industry). Recently Globalization is based to the rapid development of economics and communication. Therefore Management and Economics need a particular approach through Risk Analysis. But, it is clear to us that, all lines of though under Risk Analysis need a strong Statistical background. We tried to serve this line of thought in this special issue of *Revstat — Statistical Journal* entitled *"Risk Analysis: Challenges and Applications"* and we hope it will stimulate and provide a huge interaction between researchers in several fields, once is clear that providing for minimizing risks is of general utmost importance.

In the first chapter, the Michaelis–Menten model (MM) is explored and this model is very well known for playing a very important role in pharmacokinetics. An analytical method for the nonlinear least squares estimation of the MM is introduced and it is proved that the MM model has not a unique parameter estimation, there is not a unique optimal experimental design and it might not have a unique D-optimal design. In chapter two, the authors study the impact of skewness on risk analysis by considering the product of two normally distributed variables. The moment generating function was obtained and several simulations were performed using software R. The paper focus an interesting topic that may have significant applications in several areas. Due to its prevalence and mortality,

a cancer diagnosis is one of the main fears of the general public. The discrimination of tissue for mammary cancer is an important topic, so the recent advances in this area are given in chapter three. The authors discuss statistical distributions of fractal dimensions for both mammary cancer and mastopathy and they conduct a multifractal analysis on the basis of a wavelet based approach. An interesting discussion was also provided with focus on alternative cancer therapy and cancer prevention. Credit scoring and credit risk are very important tools of financial risk management. In chapter four, the authors consider three different techniques applicable in the context of credit scoring when the event under study is rare and therefore we have to cope with unbalanced data. Practical application to balance sheets indicators of small and medium-sized enterprises and their legal status is given. Risk is a basic slogan of insurance and there is high importance to know how insurer can/may price a risk for which there is no history. In chapter five, the authors show which main mechanisms are needed to capture the tariff model of a related insurance minimizing the risk involved. Car insurance industry applications are presented. Ecological modelling and in particular water and hydrometric extremes are very important applications of statistical extremes theory. In chapter six authors present a readable survey on several tests and parameter estimation procedures available in recent literature. The application of these methods is provided to daily mean flow discharge rate values in the hydrometric station of Fragas da Torre in the river Paiva. The generalised extreme value (GEV) distribution is currently used to fit to environmental time series of extreme values, such as annual maxima and minima of temperatures. In the last, chapter seven, the authors present GEV distribution on a case study of temperature extremes in a mountainous area of Greece, emphasizing that searching through alternative distributions also adds an extra layer of uncertainty to the model selection procedure.

As it is testified by the articles in this special issue, the Applications are driving force for considering Risk Analysis as an integrated discipline in practically every scientific field. The particular Statistical background, covering the Risk Analysis approach, is that we try to improve in this volume.

We thank the *Revstat* that gave us such an opportunity, to collect selected papers in the Risk Analysis field and publish to an excellent journal. Our special thanks are addressed to Professor Ivette Gomes, editor-in-chief of *Revstat — Statistical Journal*, for her support.

Finally, on behalf of the Editorial Board we would like to thank the authors and the anonymous reviewers for their precious contribution to this special issue.

Christos P. Kitsos
Department of Informatics,
Technological Educational
Institute of Athens,
Greece
xkitsos@teiath.gr

Teresa A. Oliveira
Department of Sciences
and Technology,
Universidade Aberta,
Portugal
teresa.oliveira@uab.pt

Milan Stehlík
Institute of Statistics,
University of Valparaíso, Chile
Department of Applied Statistics,
Johannes Kepler University Linz, Austria
mlnstehlik@gmail.com

INDEX

Abstracted/indexed in: *Current Index to Statistics, DOAJ, Google Scholar, Journal Citation Reports/Science Edition, Mathematical Reviews, Science Citation Index Expanded®, SCOPUS* and *Zentralblatt für Mathematic.*

REVSTAT – Statistical Journal
Volume 14, Number 2, April 2016, 101–118

ESTIMATION ASPECTS OF THE MICHAELIS–MENTEN MODEL

Authors: Thomas L. Toulias
– Boulevard du Souverain 292, Auderghem 1160, Brussels, Belgium
th.toulias@gmail.com

Christos P. Kitsos
– Department of Informatics, Technological Educational Institute of Athens, Greece
xkitsos@teiath.gr

Received: September 2015 Revised: February 2016 Accepted: February 2016

Abstract:

- This paper studies the Michaelis–Menten model (MM), which plays an important role in pharmacokinetics, from a theoretical as well as a computational point of view. An analytical method for the nonlinear least squares estimation of the MM is introduced. It is proved that the MM model has not a unique parameter estimation (through the nonlinear least squares), and there is not a unique optimal experimental design and might not have a unique D-optimal design. An iterative process, based on the Sequential approach, is also introduced and tested on various data sets for the MM model. A different approach is also discussed which provides an initial estimate that increases the convergence rate of the Fully Sequential approach. Several examples demonstrate the provided methods.

Key-Words:

- *Michaelis–Menten model; optimal design; nonlinear least squares; fully sequential method.*

AMS Subject Classification:

- 62K05, 93E24, 62H12.

1. INTRODUCTION

A general theory for enzyme kinetics was firstly developed by Michaelis and Menten [17] in their pioneering work, where the metabolism of an agent is described by a reaction rate. The basic toxicokinetic model of metabolism is a Michaelis–Menten (MM) model. Briefly, when an enzyme E is combined reversibly with a substrate S to form an enzyme-substrate complex ES, which can be dissociate or proceed to the product P, the following enzyme-substrate reaction scheme

$$E + S \underset{k_1}{\overset{k_2}{\rightleftharpoons}} ES \xrightarrow{k_3} E + P \,, \tag{1.1}$$

is assumed, with k_1, k_2 and k_3 being the associated rate constants. We let $K_{\rm M} := (k_2 + k_3)/k_1$, known as the MM constant, and $V_{\max} := k_3 C_{\rm T}$, where $C_{\rm T}$ is the total enzyme concentration. Then, a plot of the initial velocity of reaction V against the concentration of substrate $C_{\rm S}$, will provide the MM rectangular hyperbola of the form

$$V \,=\, V(C_{\rm S}) \,=\, V(C_{\rm S};\boldsymbol{\theta}) \,:=\, \frac{V_{\max}\, C_{\rm S}}{K_{\rm M} + C_{\rm S}} \,, \tag{1.2}$$

where the parameters' vector $\boldsymbol{\theta} := (V_{\max}, K_{\rm M}) \in \Theta \subseteq \mathbb{R}^2$, and Θ being the parameter space which is assumed compact when sequential approaches are applied. It is understood that the appropriate estimate of $\boldsymbol{\theta}$ is very essential to eliminate the Risk on the enzyme-substrate reaction scheme as in (1.1). The hidden Risk is strongly related to the appropriate real estimate, as it is proved bellow, a number of estimates might exist, even not real. Therefore the estimate $\boldsymbol{\theta}$ clarifies the Risk Analysis for the toxicokinetic model of metabolism under investigation, so essential in pharmacokinetics.

In practice, we have n readings for the reaction's initial velocity $V_i := V(C_{{\rm S},i};\boldsymbol{\theta})$ corresponding to n substrate concentration values $C_{{\rm S},i}$, $i = 1, 2, ..., n$. That is, only the stochastic model of the form $y_i := V_i + e_i$, $i = 1, 2, ..., n$ is obtained, as the readings V_i are associated with noise; see Kitsos [14, 13]. In principle, $C_{\rm S} > 0$ and hence $K_{\rm M} > 0$, while usually the reaction velocity $V(C_{\rm S}) > 0$.

Different linear transformations have been suggested, [4, 1], to estimate (with a linear regression fit) the involved parameters, $V_{\max}$ and $K_{\rm M}$, as the input variable $C_{\rm S}$ and the response V are curvilinearly related. The usual linear transformations are: the Eadie–Hofstee (EH), the Hanes–Wolf (HW), the Lineweaver–Burk or "double reciprocal" (LB), and the "inverse" Eadie–Hofstee (iEH) linearizations, which are formulated by

$$V \,=\, V_{\max} - K_{\rm M}\frac{V}{C_{\rm S}} \,, \qquad \text{(EH)} \tag{1.3a}$$

$$\frac{C_{\rm S}}{V} \,=\, \frac{K_{\rm M}}{V_{\max}} + \frac{1}{V_{\max}}\, C_{\rm S} \,, \qquad \text{(HW)} \tag{1.3b}$$

$$\frac{1}{V} \;=\; \frac{1}{V_{\max}} + \frac{K_{\mathrm{M}}}{V_{\max}}\,\frac{1}{C_{\mathrm{S}}}\,, \qquad \text{(LB)} \tag{1.3c}$$

$$C_{\mathrm{S}} \;=\; -K_{\mathrm{M}} + V_{\max}\,\frac{C_{\mathrm{S}}}{V}\,, \qquad \text{(iEH)} \tag{1.3d}$$

respectively, with the double reciprocal being the most popular. The Hanes–Wolf linearization has been shown in a very early work by Dowd and Riggs [4], as the most efficient. Endvenyi and Chan [6] and Currie [3] discussed the heteroscedasticity in the MM model. Endvenyi and Chan [6] assumed that the error variance is proportional to the mean. To avoid heteroscedasticity problems in the "linearized" models (1.3a)–(1.3d) we should suggest to solve the model's Normal Equations and get the least square estimators when n readings of V and C_{S} are given, i.e. to solve

$$\sum_{i=1}^{n}\left(V_i - \frac{V_{\max}C_{\mathrm{S},i}}{K_{\mathrm{M}}+C_{\mathrm{S},i}}\right)\nabla_{\boldsymbol{\theta}}V(C_{\mathrm{S},i};\boldsymbol{\theta}) \;=\; 0\,, \tag{1.4}$$

with V as in (1.2), while

$$\nabla_{\boldsymbol{\theta}}V \;=\; \left(\frac{\partial V}{\partial V_{\max}},\,\frac{\partial V}{\partial K_{\mathrm{M}}}\right)^{\mathrm{T}} = \left(\frac{C_{\mathrm{S}}}{K_{\mathrm{M}}+C_{\mathrm{S}}},\,-\frac{V_{\max}K_{\mathrm{M}}C_{\mathrm{S}}}{(K_{\mathrm{M}}+C_{\mathrm{S}})^2}\right)^{\mathrm{T}}. \tag{1.5}$$

Hence,

$$\frac{\partial S}{\partial V_{\max}} \;=\; \sum_{i=1}^{n}\left(V_i - \frac{V_{\max}C_{\mathrm{S},i}}{K_{\mathrm{M}}+C_{\mathrm{S},i}}\right)\frac{C_{\mathrm{S},i}}{K_{\mathrm{M}}+C_{\mathrm{S},i}} \;=\; 0\,, \tag{1.6}$$

$$\frac{\partial S}{\partial K_{\mathrm{M}}} \;=\; \sum_{i=1}^{n}\left(V_i - \frac{V_{\max}C_{\mathrm{S},i}}{K_{\mathrm{M}}+C_{\mathrm{S},i}}\right)\frac{V_{\max}C_{\mathrm{S},i}}{(K_{\mathrm{M}}+C_{\mathrm{S},i})^2} \;=\; 0\,. \tag{1.7}$$

Then, the two Normal Equations can be easily obtained and solved numerically so that the estimates $\hat{\boldsymbol{\theta}} = (\hat{V}_{\max}, \hat{K}_{\mathrm{M}})$ are obtained. For a single observation the Fisher's information matrix is $(\nabla V)^{\mathrm{T}}(\nabla V)$. Therefore, the average-per-observation information matrix $\mathbf{M}(\boldsymbol{\theta},\xi)$ is evaluated as

$$\sigma^{-2}\,n\,\mathbf{M}(\boldsymbol{\theta},\xi) \;=\; \sum_{i=1}^{n}\begin{pmatrix} C_{\mathrm{S},i}^2\,\tau_i^2 & -V_{\max}C_{\mathrm{S},i}^2\,\tau_i^3 \\ -V_{\max}C_{\mathrm{S},i}^2\,\tau_i^3 & V_{\max}C_{\mathrm{S},i}^2\,\tau_i^4 \end{pmatrix}, \tag{1.8}$$

with $\tau_i = 1/(K_{\mathrm{M}}+C_{\mathrm{S},i})$, $i = 1, 2, ..., n$. Thus, the 2×2 variance-covariance matrix is approximately equal to

$$\mathbf{C} \;=\; \mathbf{C}(\hat{\boldsymbol{\theta}},\xi) \;=\; \big(n\mathbf{M}(\hat{\boldsymbol{\theta}},\xi)\big)^{-1}. \tag{1.9}$$

Hence, we can derive asymptotic approximate confidence intervals for the involved parameters, and we can work for optimally criteria, with the D-optimal design being the most applicable; see [13] for details. Notice that, due to the fact that the MM model is partially nonlinear,[9], the D-optimal design depends only on the K_{M} parameter; see [14] for details.

2. NONLINEAR LEAST SQUARES FOR THE MM MODEL

In this Section an analytical method for the Nonlinear Least Square (NLLS) estimation of the MM model is introduced and discussed. In particular, the following Theorem provides a compact analytic methodology for the "actual" NLLS estimation of the MM model's parameters.

Recall that the Sum of Squares of Errors *sse* is given by $sse = sse(\boldsymbol{\theta}) := \sum_{i=1}^{n}[V_i - V(C_{\mathrm{S},i};\boldsymbol{\theta})]^2$, $n > 2$, where $\boldsymbol{\theta}$ is a vector of the MM model parameters $(V_{\max}, K_{\mathrm{M}})$, while $(C_{\mathrm{S},i}), (V_i) \in \mathbb{R}^n$ are the data vectors for the substrate concentration and the reaction's velocity respectively. The estimated parameters $\hat{\theta}_1 = \hat{V}_{\max}$ and $\hat{\theta}_2 = \hat{K}_{\mathrm{M}}$, from the estimated vector $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2)$, are obtained when $sse(\hat{\boldsymbol{\theta}}) = \min\{sse(\boldsymbol{\theta})\}_{\boldsymbol{\theta}\in\Theta}$, $\Theta \in \mathbb{R}^2$ compact, i.e. when $\nabla S = 0$, or equivalently, when the normal equations (1.4) are satisfied. Recall also that the mean absolute relative error $mare = mare(\hat{\boldsymbol{\theta}}) := \mathrm{E}\big(\big|\{V_i - V(C_{\mathrm{S},i};\hat{\boldsymbol{\theta}})\}/V_i\big|\big)$, is also evaluated, see Example 2.1 below, while, due to [10], there is always a solution of the normal equations.

The following Theorem as it is stated and proved, provides evidence that the MM model has not unique Least Square Estimates. As we already mentioned in the introduction this creates a further investigation for the Risk Analysis under study, as the appropriate, among a number of estimates, has to be chosen. We discuss the proposed strategy in the sequence of this paper.

Theorem 2.1. *The NLLS estimators $\hat{V}_{\max}$ and $\hat{K}_{\mathrm{M}}$ of the MM model are the ones among the $K \le 4n-5$ possible estimates' vectors $\hat{\boldsymbol{\theta}}_k = (\hat{V}_{\max;k}, \hat{K}_{\mathrm{M};k})$, $k = 1, 2, ..., K$, with $sse(\hat{V}_{\max}, \hat{K}_{\mathrm{M}}) := \min\{sse(\hat{V}_{\max;k}, K_{\mathrm{M};k})\}_{k=1,2,...,K}$, where*

$$\hat{V}_{\max;k} = \left(\sum_{i=1}^{n} \frac{V_i C_{\mathrm{S},i}}{\hat{K}_{\mathrm{M};k} + C_{\mathrm{S},i}}\right)\left[\sum_{i=1}^{n}\left(\frac{C_{\mathrm{S},i}}{\hat{K}_{\mathrm{M};k} + C_{\mathrm{S},i}}\right)^2\right]^{-1}, \quad k = 1, 2, ..., K\,, \tag{2.1}$$

while the estimated $\hat{K}_{\mathrm{M};k}$ values are the $K \le 4n-5$ real roots of the following $(4n-5)$-degree polynomial of ϑ_2:

$$P(\vartheta_2) := \sum_{\substack{i,j=1\\(i\neq j)}}^{n} V_i C_{\mathrm{S},i} C_{\mathrm{S},j} (C_{\mathrm{S},j} - C_{\mathrm{S},i})(\vartheta_2 + C_{\mathrm{S},i})^2 (\vartheta_2 + C_{\mathrm{S},j}) \prod_{(i,j\neq)\,k=1}^{n} (\vartheta_2 + C_{\mathrm{S},k})^4\,. \tag{2.2}$$

Proof: Solving the first normal equation (1.6) with respect to $V_{\max}$ and substituting to the second one (1.6) we obtain

$$\left(\sum_{i=1}^{n} \frac{V_i x_i}{(K_{\mathrm{M}} + x_i)^2}\right)\sum_{i=1}^{n}\left(\frac{x_i}{K_{\mathrm{M}} + x_i}\right)^2 = \left(\sum_{i=1}^{n} \frac{V_i x_i}{K_{\mathrm{M}} + x_i}\right)\left(\sum_{i=1}^{n} \frac{x_i^2}{(K_{\mathrm{M}} + x_i)^3}\right),$$

where $x_i := C_{\mathrm{S},i}$, $i = 1, 2, ..., n$. Thus,

$$\left[\sum_{i=1}^{n} u_i x_i Q_i(2)\right]\left[\sum_{j=1}^{n} x_j^2 Q_j(2)\right] = \left[\sum_{i=1}^{n} u_i x_i Q_i(1)\right]\left[\sum_{j=1}^{n} x_j^2 Q_j(3)\right],$$

where $Q_i(d) := \prod_{(i\neq)\,m=1}^{n}(K_{\mathrm{M}} + C_{\mathrm{S},m})^d$, $i = 1, 2, ..., n$, $d = 1, 2, 3$. With some algebra, the above relation can be written as

$$\sum_{\substack{i,j=1\\(i\neq j)}}^{n} u_i x_i x_j^2 \left\{ s_i^2 s_j^2 \left(\prod_{(i,j\neq)\,m=1}^{n} s_m^4\right) - s_i^3 s_j \left(\prod_{(i,j\neq)\,m=1}^{n} s_m^4\right)\right\} = 0\,, \tag{2.3}$$

where $s_m := K_{\mathrm{M}} + C_{\mathrm{S},m}$, $m = 1, 2, ..., n$. Finally, the solution with respect to K_{M} of the above equation corresponds to the roots of the polynomial P of $\vartheta_2 := K_{\mathrm{M}}$, as in (2.2), as then the requested K_{M} $(= \vartheta_2)$ values would satisfy the normal equations (1.6) and (1.7). For each of the $K \leq 4n - 5$ real-valued root of (2.2), i.e. for each possible estimate $\hat{K}_{\mathrm{M};k}$, the corresponding $\hat{V}_{\max;k}$, $k = 1, 2, ..., K$, estimate is then obtained through (2.1), which is the solution of the (1.6) with respect to $V_{\max}$.

The solutions of the normal equations (1.6) and (1.7), through the roots of the polynomial (2.2), provide $4n-5$ possible estimates of $\boldsymbol{\theta}$. As this number is odd there is always at least one real root of (2.2). Therefore, at least one real critical point of the least square objective function S exists, which may yield at least one estimate $\hat{\boldsymbol{\theta}} = (\hat{\vartheta}_1, \hat{\vartheta}_2) := (\hat{V}_{\max}, \hat{K}_{\mathrm{M}})$ for the MM model. For a working example see [18]. However, when all data points are collinear, the nonlinear least squares estimate cannot exist, see [11] or [8]. Therefore, various different NLLS estimates may exist (among the K real-valued $\hat{\boldsymbol{\theta}}_k = (\hat{V}_{\max;k}, \hat{K}_{\mathrm{M};k})$, $k = 1, 2, ..., K$) which (locally) minimizes the sum of squares $S = sse(\hat{\boldsymbol{\theta}})$. The problem the experimenter has then to face is which of the real roots ϑ $(= \hat{K}_{\mathrm{M}})$ of $P = P(\vartheta)$, as in (2.2), can be chosen as the MM model's NLLS estimate $\hat{K}_{\mathrm{M}}$. Among the K real-valued (of the total $4n - 5$) candidate estimates, the experimenter can always choose the one which provides the minimum sum of squares, i.e.

$$sse(\hat{\boldsymbol{\theta}}_{\mathrm{NLLS}}) := sse(\hat{V}_{\max}, \hat{K}_{\mathrm{M}}) = \min\left\{sse(\hat{\boldsymbol{\theta}}_k)\right\}_{k=1,2,...,K}, \tag{2.4}$$

as the NLLS estimates' vector $\hat{\boldsymbol{\theta}}_{\mathrm{NLLS}}$ would then be a global minimum for the MM model's sum of squares. □

As the degree $(4n-5)$ of the polynomial $P(\vartheta_2)$ in Theorem 2.1 is odd, there is always at least one (real-valued) estimates' vector $\hat{\boldsymbol{\theta}}$ for the MM model. Thus, as more than one estimates's vectors can exist, then more than one corresponding average-per-observation information matrices are possible. Therefore, the design might not be considered as unique, as Biebler in [2] has been noticed. See also [16]. The criterion we suggest (the minimum *sse*) actually offers the design with the minimum variance, i.e. it corresponds to the D-optimal one.

The following Example provides a comparative presentation of the linear and the "actual" nonlinear estimation for the MM model.

Example 2.1. The treated case of the Puromicin data set was adopted, as in Bates and Watts [1, Table A1.3], for the comparison of the Linear Least Squares (LLS) estimations, through the linearizations as in (1.3a)–(1.3d), and the analytic NLLS estimation of Theorem 2.1. Table 1 provides the evaluated LLS estimates $\hat{V}_{\max}$ and $\hat{K}_{\mathrm{M}}$ (obtained by linear regression), which correspond to the Eadie–Hofstee (EH), Hanes–Wolf (HW), Lineweaver–Burk (LB), and the "inverse" Eadie–Hofstee (iEH) linearization methods, together with the NLLS estimates that provide the minimum sum of squares (among all the possible pairs of NLLS estimates obtained through Theorem 2.1). Their R^2 coefficient as well as their corresponding *sse* and *mare* (%) errors are also presented. All the involved calculations in Table 1 as well as the corresponding Figure 1 were done by using MATLAB as a programming tool.

Table 1: Comparison between the LLS and the analytic NLLS estimation.

Est. Method	$\hat{V}_{\max}$	$\hat{K}_{\mathrm{M}}$	R^2	*sse*	*mare* (%)
LLS (EH)	193.867711	0.043524	0.9282	184.69	10.74
LLS (HW)	216.216899	0.067915	0.9603	102.05	6.94
LLS (LB)	195.802709	0.048407	0.9378	160.05	9.19
LLS (iEH)	215.773203	0.067068	0.9606	101.45	6.99
NLLS	212.683743	0.064121	0.9613	99.62	6.99

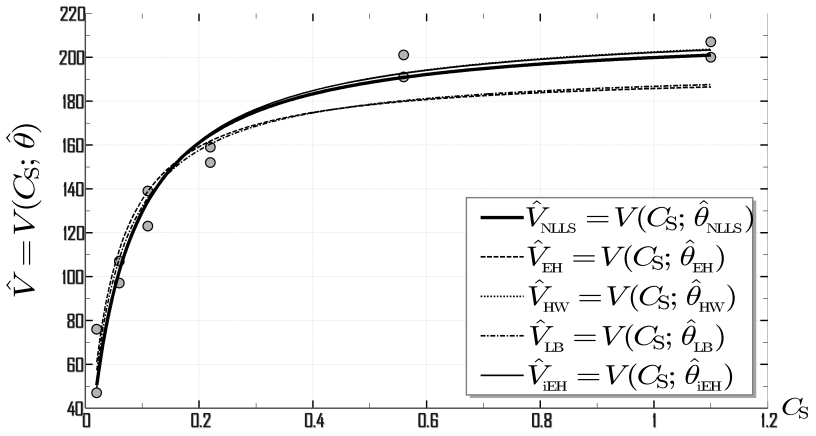

Figure 1: Visual comparison between the predicted NLLS model $\hat{V}_{\mathrm{NLLS}}$ against the four predicted LLS models.

It is clear that the NLLS estimation provides a "better" estimate than the LLS ones, in terms of the corresponding R^2 coefficients, the sum of squared errors

sse and the mean absolute relative errors *mare*. Notice that, the estimation through the iEH linearization approximates better the analytic NLLS estimation, as adopts the least sum of squared error *sse*, and almost identical *mare* error among the other three LLS estimation.

Figure 1 provides a graphic comparison between the linear and the nonlinear least squares estimation for the MM model (using the Puromicin data set), by depicting the estimated NLLS model $\hat{V}_{\mathrm{NLLS}} := V(C_{\mathrm{S}};\hat{\boldsymbol{\theta}}_{\mathrm{NLLS}})$ against with the four LSS estimated models $\hat{V}_{\mathrm{EH}} := V(C_{\mathrm{S}};\hat{\boldsymbol{\theta}}_{\mathrm{EH}})$, $\hat{V}_{\mathrm{HW}} := V(C_{\mathrm{S}};\hat{\boldsymbol{\theta}}_{\mathrm{HW}})$, $\hat{V}_{\mathrm{LB}} := V(C_{\mathrm{S}};\hat{\boldsymbol{\theta}}_{\mathrm{LB}})$ and $\hat{V}_{\mathrm{iEH}} := V(C_{\mathrm{S}};\hat{\boldsymbol{\theta}}_{\mathrm{iEH}})$. The estimates' vectors $\hat{\boldsymbol{\theta}}_{\mathrm{NLLS}}$, $\hat{\boldsymbol{\theta}}_{\mathrm{EH}}$, $\hat{\boldsymbol{\theta}}_{\mathrm{HW}}$, $\hat{\boldsymbol{\theta}}_{\mathrm{LB}}$ and $\hat{\boldsymbol{\theta}}_{\mathrm{iEH}}$ are provided by the corresponding vectors $(\hat{V}_{\max},\hat{K}_{\mathrm{M}})$ of Table 1. Note that the iEH and the HW linearizations provide almost similar LLS models (the dotted $\hat{V}_{\mathrm{HW}}$ curve is very close to the thin solid $\hat{V}_{\mathrm{iEH}}$ curve), both very close to the 'actual' NLLS model.

3. SEQUENTIAL GAUSS–NEWTON ESTIMATION

The Fully Sequential (FS) method has been discussed by Ford *et al.* in [7] and Kitsos in [15, 12] for the nonlinear experimental design problem. Expanding the FS method, in this Section we introduce and investigate the general case of the Batch Sequential (BS) iterative scheme, for the MM model estimation.

For the estimation of the parameter $\boldsymbol{\theta}\in\Theta$, $\Theta\subseteq\mathbb{R}^q$ compact, of any model $\eta=\eta(x;\boldsymbol{\theta})$ in general, recall the known definition of sum of squares $sse = sse(\boldsymbol{\theta})$,

$$sse(\boldsymbol{\theta}) = sse(\boldsymbol{\theta};\mathbf{x}) := \big\|\boldsymbol{\eta}-\eta(\mathbf{x};\boldsymbol{\theta})\big\|^2 = \sum_{i=1}^{n}\big[\eta_i-\eta(x_i;\boldsymbol{\theta})\big]^2\,, \tag{3.1}$$

where $\mathbf{x}:=(x_i)\in\mathbb{R}^n$, $\boldsymbol{\eta}:=(\eta_i)\in\mathbb{R}^n$, and $\|\cdot\|$ denotes the usual $\mathscr{L}^2$-norm. The estimated parameters' vector $\hat{\boldsymbol{\theta}}=(\hat{\vartheta}_1,\hat{\vartheta}_2)$ is obtained when $sse(\hat{\boldsymbol{\theta}})=\min\{sse(\boldsymbol{\theta})\}_{\boldsymbol{\theta}\in\Theta}$, i.e. when $\nabla S=0$.

Recall also the iterative GN method, for the parameter estimation of the general nonlinear model described by $\eta=\eta(x;\boldsymbol{\theta})$; see also [5]. In the GN iterative procedure a series of estimates $\hat{\boldsymbol{\theta}}_N\in\Theta$ is produced where the next estimates' vector $\hat{\boldsymbol{\theta}}_{N+1}$ is derived from the previous one $\hat{\boldsymbol{\theta}}_N$. When the sequence converges to a vector, then this vector is a possible NLLS estimates' vector, say $\hat{\boldsymbol{\theta}}_{\mathrm{NLLS}}\in\Theta$, of the model η, i.e. $\nabla sse(\hat{\boldsymbol{\theta}}_{\mathrm{NLLS}})=0$. The GN iterative scheme is described in a compact form, by

$$\hat{\boldsymbol{\theta}}_{N+1} = \hat{\boldsymbol{\theta}}_N - \mathbf{H}_S^{-1}(\hat{\boldsymbol{\theta}}_N)\,\nabla sse(\hat{\boldsymbol{\theta}}_N)\,,\qquad N\in\mathbb{N}^*:=\mathbb{N}\backslash\{0\}\,, \tag{3.2}$$

for a given initial estimates' vector $\hat{\boldsymbol{\theta}}_0=(\hat{\vartheta}_1^0,\hat{\vartheta}_2^0)\in\Theta$, where $\mathbf{H}_S^{-1}(\hat{\boldsymbol{\theta}}_N)$ is the inverse of the Hessian matrix of the sum of squares function $sse(\hat{\boldsymbol{\theta}}_N)$ as in (3.1).

The k-Batch Sequential approach (k-BS), $k \in \mathbb{N}^*$, presented here, is based on the GN method, and it consists of N_k steps in total. On the N-th step of the k-BS method, a number of GN iterations are performed on the chunk of Nk observations (from the total n), which always starts form the first and ends to the Nk-th observation. In particular, the total number of iterations N_k is given by

$$N_k := \begin{cases} n/k\,, & \text{when } \ n/k \in \mathbb{N}^*, \ \text{ and } \ k \neq 1\,, \\ [n/k]+1\,, & \text{when } \ n/k \notin \mathbb{N}^*, \ \text{ and } \ k \neq 1\,, \\ n-1\,, & \text{when } \ k = 1\,, \end{cases} \tag{3.3}$$

where $[\cdot]$ denotes the integer part of a real number and n is the number of observations. Note that, for the application of the BS iterative process, the data pairs $\{(x_i, \eta_i)\}_{i=1,2,...,n}$ are "entered" sequentially into the BS process. Therefore, instead of the usual sum of squares S, as in (3.1), the k-BS approach utilizes a partial sum of squares (p.s.s.) S_N (used for the corresponding GN iterations on every step N of the k-BS method), which is the sum of squares calculated only for the specific chunk of Nk observations on the N-th step of the method. Thus, the S_N is defined as

$$S_N(\hat{\boldsymbol{\theta}}) := \sum_{i=1}^{kN} \Big[V_i - V(C_{\mathrm{S},i}; \hat{\boldsymbol{\theta}}) \Big]^2, \qquad N = 1, 2, ..., N_k = n/k\,, \tag{3.4}$$

provided that $n/k \in \mathbb{N}$, with $k \neq 1$. For the case of $n/k \notin \mathbb{N}^*$ (and $k \neq 1$), the p.s.s. S_N is defined as in (3.4) for $N = 1, 2, ..., [n/k]$, while for the last step $N_k = [n/k] + 1$, the $S_{N=N_k}$ is defined by

$$S_N(\hat{\boldsymbol{\theta}}) := \sum_{i=1}^{n} \Big[V_i - V(C_{\mathrm{S},i}; \hat{\boldsymbol{\theta}}) \Big]^2. \tag{3.5}$$

For the special case of the 1-BS method, the p.s.s. S_N is defined by

$$S_N(\hat{\boldsymbol{\theta}}) := \sum_{i=1}^{N+1} \Big[V_i - V(C_{\mathrm{S},i}; \hat{\boldsymbol{\theta}}) \Big]^2, \qquad N = 1, 2, ..., N_1 = n-1\,. \tag{3.6}$$

Finally, the GN iterative procedures at each step N (of the total N_k steps) of the k-BS method, either converge or can be stopped (after a given maximum number of GN iterations) to some estimated parameters' vector, say $\hat{\boldsymbol{\theta}}_N$. This $\hat{\boldsymbol{\theta}}_N$ is then considered as the initial vector for the GN iterations of the next step $N+1$ of the k-BS method. Hence, an initial parameters' vector $\hat{\boldsymbol{\theta}}_0$ is then needed in order to begin the GN iterations of first step $N = 1$ of the k-BS method.

Notice that, for a set of $n = mk$, $m \in \mathbb{N}^*$, observations, the p.s.s. S_N (for the GN iterations on the N-th step of the k-BS scheme, $k \neq 1$) is calculated through the summation of successively $k, 2k, 3k, ..., mk = n$ terms. For a set of odd number of observations, say $n = mk + q$ with $\mathbb{N}^* \ni q < m$, the p.s.s. S_N

is calculated through the summation of successively $k, 2k, 3k, ..., mk, n$ terms. For the 1-BS case, the corresponding p.s.s. S_N summation is performed with successively $2, 3, ..., n$ terms. Notice also that the 2-BS scheme coincides with the FS scheme, as the p.s.s. S_N summation uses $2, 4, 6, ..., n$ terms.

The above description of the k-BS iterative method can be formulated into the following algorithm.

Algorithm 3.1. Consider an initial vector $\hat{\boldsymbol{\theta}}_0$ for the estimation of the model $\eta = \eta(x; \boldsymbol{\theta}) = \eta(x; \vartheta_1, \vartheta_2)$. On every step $N = 1, 2, ..., N_k$ of the k-BS method, a GN iterative process is applied, using the p.s.s. S_N for a given maximum number of iterations, say J. Then, a series of vectors is produced, say $\hat{\boldsymbol{\theta}}_{N,1}, \hat{\boldsymbol{\theta}}_{N,2}, ..., \hat{\boldsymbol{\theta}}_{N,J}$. The next estimate $\hat{\boldsymbol{\theta}}_{N+1}$ is then considered to be the last current estimate

$$\hat{\boldsymbol{\theta}}_{N+1} = \hat{\boldsymbol{\theta}}_{N,J}\,, \qquad N = 0, 1, 2, ..., N_k\,,$$

where the vectors $\hat{\boldsymbol{\theta}}_{N,j}$, $j = 1, 2, ..., J$, are described by the GN iterative scheme

$$\hat{\boldsymbol{\theta}}_{N,j+1} = \hat{\boldsymbol{\theta}}_{N,j} - \mathbf{H}^{-1}_{S_N}(\hat{\boldsymbol{\theta}}_{N,j})\nabla S_N(\hat{\boldsymbol{\theta}}_{N,j})\,, \quad j = 0,1, ..., J_N \leq J\,, \ \text{ with } \hat{\boldsymbol{\theta}}_{0,0} := \hat{\boldsymbol{\theta}}_0\,. \tag{3.7}$$

For every step N, the index $J_N \leq J$ is the one for which the GN process converges (when it does), i.e. when the convergence error, say e_N, of the estimate $\hat{\boldsymbol{\theta}}_{N,J_N} = (\hat{\vartheta}^1_{N,J_N}, \hat{\vartheta}^2_{N,J_N})$ is smaller or equal than a given threshold error e, i.e.

$$e_N := \max\Big\{\big|\hat{\vartheta}^1_{N,J_N} - \hat{\vartheta}^1_{N,J_N-1}\big|,\, \big|\hat{\vartheta}^2_{N,J_N} - \hat{\vartheta}^2_{N,J_N-1}\big|\Big\} \leq e\,, \quad N = 1, 2, ..., N_k\,. \tag{3.8}$$

Otherwise, when the convergence fails, the GN process stops at the J-th GN iteration (i.e. when $j = J$). For the next $N + 1$ step, as an initial vector $\hat{\boldsymbol{\theta}}_{N+1,0}$ for the new GN iteration, we consider the last estimate of the previous GN process, i.e. $\hat{\boldsymbol{\theta}}_{N+1,0} = \hat{\boldsymbol{\theta}}_{N,J_N}$, $N = 1, 2, ..., N_k$.

The following Example applies the FS iterative scheme, i.e. the 2-BS iterative scheme as in Algorithm 3.1.

Example 3.1. The Puromycin-treated data set, as in Example 2.1, consists of $n = 12$ observations where the $C_{\mathrm{S},i}$, $i = 1, 2, ..., 6$, readings are repeated. For this Example we consider the subset of the $n = 6$ non-replicated observations of the Puromycin data set. Let $e = 10^{-4}$ be the convergence error threshold of the 2-BS method, while as initial estimates' vector guess we adopt $\hat{\boldsymbol{\theta}}_0 = (100, 0)$. The first sub-Table of the Table 2 provides the last J_N-th GN convergent estimates' vector $\hat{\boldsymbol{\theta}}_{N,J_N}$, for each of the (total three) steps $N = 1, 2, 3$ of the 2-BS method. Recall that the total number of steps of the 2-BS algorithm for this data set is $N_{k=2} := n/k = 6/2 = 3$. Notice that $0 + 7 + 7 + 5 = 19$ GN iteration are needed, in total, to obtain $\hat{\boldsymbol{\theta}}_{\mathrm{NLLS}}$ with accuracy $< 10^{-5}$. Moreover, the GN processes, for every step N of the 2-BS scheme, do not have to converge at a

given error convergence e. For example, reducing the maximum number of the GN iterations to be, say $J=5$, we obtain also $\hat{\boldsymbol{\theta}}_{\mathrm{NLLS}}$ with the same accuracy $<10^{-5}$, but this time $0+5+5+5=15$ GN iteration are needed in total, see the middle sub-Table of Table 2.

Table 2: Convergence of various 2-BS processes, for the MM model estimation.

N	J_N	$\hat{V}_{\max}$	$\hat{K}_{\mathrm{M}}$	e_N	ε_N	R^2	*mare* (%)
• *Maximum # of GN iterations:* $J=10$							
0	0	100.	0.	—	2.5e+5	−1.6531	17.34
1	7	112.549618	0.009618	3.89e−8	0.	1.	0.
2	7	172.241997	0.034754	5.36e−10	7.64e−11	0.8739	10.63
3	5	**210.857219**	**0.062575**	7.03e−6	1.01e−8	0.9360	9.341
• *Maximum # of GN iterations:* $J=5$							
0	0	100.	0.	—	2.5e+5	−1.6531	17.34
1	5	112.548698	0.009618	0.143	2.18	1.	0.00043
2	5	172.241882	0.034754	0.0539	0.218	0.8739	10.63
3	5	**210.857219**	**0.062575**	7.03e−6	1.01e−8	0.9360	9.341
• $\hat{\boldsymbol{\theta}}_0=(V_{\max},K_{\mathrm{M}})$, $J=10$							
0	0	112.549618	0.009618	—	0.	1.	0.
1	1	112.549618	0.009618	0.	0.	1.	0.
2	7	172.241997	0.034754	5.36e−10	7.64e−11	0.8739	10.63
3	5	**210.857219**	**0.062575**	7.03e−6	1.01e−8	0.9360	9.341

The initial estimate guess the experimenter provides, plays a crucial role for the convergence of the general BS methodology. In order to address this issue, the 2-BS scheme can be applied adopting as initial parameters' vector $\hat{\boldsymbol{\theta}}_0$ the solution vector $(V_{\max},K_{\mathrm{M}})$ of the two MM model relations $V_i=V(C_{\mathrm{S},i};V_{\max},K_{\mathrm{M}})$, $i=1,2$. These relations can be solved analytically in the form of $(V_{\max},K_{\mathrm{M}})=(V_1+dV_1,\ dC_{\mathrm{S},2})$, $d:=C_{\mathrm{S},2}(V_2-V_1)/(V_1C_{\mathrm{S},2}-V_2C_{\mathrm{S},1})$. As a result, the first GN process (for the step $N=1$ of the 2-BS approach) will always converge at its first GN iteration $j=1=J_1$. This is due to the fact that the initial GN iteration process (where only the first two observations are involved) will surely converges to the only solution $(V_{\max},K_{\mathrm{M}})$, as above, which is now provided by the suggested initial vector $\hat{\boldsymbol{\theta}}_0$. This suggested $\hat{\boldsymbol{\theta}}_0$ it turns to be the convergent estimate $\hat{\boldsymbol{\theta}}_{1,J_1}$ of the first (and only) GN iteration for first step $N=1$, of the 2-BS process. Therefore, with the above proposed $\hat{\boldsymbol{\theta}}_0$ there is no need for guessing an initial vector, to feed the 2-BS process, that might not converge. This is true at least at the first step of the 2-BS approach. The last sub-Table of Table 2 provides the last J_N-th GN convergent estimates $\hat{\boldsymbol{\theta}}_{N,J_N}$ for every step $N=1,2,3$ of the 2-BS process, adopting as $\hat{\boldsymbol{\theta}}_0$ the proposed solution $(V_{\max},K_{\mathrm{M}})$ as discussed above. Table 2 also presents the convergence error e_N, the solution error distance $\varepsilon_N:=\|\nabla sse(\hat{\boldsymbol{\theta}}_N)\|$ as well as the R^2 coefficient for each estimated model $\hat{V}_N:=V(C_{\mathrm{S}};\hat{\boldsymbol{\theta}}_{N,J_N})$. The digits in bold represents the accurate digits of the NLLS estimates.

From our experience, due to the sequential nature of the BS methodology, the behaviour of the BS approach it might depend on the order in which the data entered into the sequential process. As the MM model is a strictly monotonous function, it is generally preferred that the $C_{\mathrm{S},i}$ readings of the data set $\{(C_{\mathrm{S},i}, V_i)\}_{i=1,2,...,n}$ are maintain this strictly monotonic pattern as they entered sequentially into the BS algorithm. However, the problem might arises when the BS process is applied on a "replicated" data set, i.e. a data set showing multiple values (two or more) of V_i for each $C_{\mathrm{S},i}$ value. Such data set is the Puromycin-treated data set which consisted of two V_i's readings for every (out of six) $C_{\mathrm{S},i}$ value. In particular, the problem is occurred when we adopt the MM solution vector $(V_{\max}, K_{\mathrm{M}})$ to play the role of the initial estimates' vector $\hat{\boldsymbol{\theta}}_0$, as we suggested earlier. Unfortunately, the solution vector $(V_{\max}, K_{\mathrm{M}})$ of the two MM relations $V_i = V(C_{\mathrm{S},i}; V_{\max}, K_{\mathrm{M}})$, $i = 1, 2$, does not exist (as $C_{\mathrm{S},1} = C_{\mathrm{S},2}$ in this data set). To avoid this problem we can apply a "higher order" BS process, as the 4-BS.

The following Example demonstrates the above discussion.

Example 3.2. The 4-BS iterative process is applied for the Puromycin data set as well as for the Enzyme Velocity (EV) data set in [11, pg. 242], which are both having replicated (double) values of C_{S} readings. Let $e = 10^{-4}$ be the threshold for the convergence error. For an initial guess $\hat{\boldsymbol{\theta}}_0$ we obtain firstly all the solutions $(V_{\max}, K_{\mathrm{M}})$ between the two MM model relations $V_i = V(C_{\mathrm{S},i}; V_{\max}, K_{\mathrm{M}})$ and $V_j = V(C_{\mathrm{S},j}; V_{\max}, K_{\mathrm{M}})$, for all $i \neq j$, $i, j = 1, 2, ..., n$. Then we adopt as $\hat{\boldsymbol{\theta}}_0$ that specific solution vector $(V_{\max}, K_{\mathrm{M}})$ which provides the minimum sum of squares for the corresponding MM model, i.e.

$$sse(\hat{\boldsymbol{\theta}}_0) = \min\Big\{sse(\boldsymbol{\theta}):\ \boldsymbol{\theta} \text{ is the solution of } V_k = V(C_{\mathrm{S},k}; \boldsymbol{\theta}),\ k \in \{i,j\}\Big\}_{i,j\in\{1,2,3,4\}},$$

or equivalently

$$sse(\hat{\boldsymbol{\theta}}_0) = \min\Big\{sse(\vartheta_1, \vartheta_2):\ \vartheta_1 = V_i(1 + d_{ij}),\ \vartheta_2 = C_{\mathrm{S},i}\, d_{ij}\Big\}_{i,j\in\{1,2,3,4\}}, \tag{3.9}$$

where $d_{ij} := C_{\mathrm{S},j}(V_j - V_i)/(V_i C_{\mathrm{S},j} - V_j C_{\mathrm{S},i})$, $i, j = 1, 2, 3, 4$. As the initial estimate $\hat{\boldsymbol{\theta}}_0$ is used for the application of the 4-BS process, it should to be obtained by using the first 4 observations (recall that the GN process at the first step of the 4-BS method is performed using the first four observations), and thus $\hat{\boldsymbol{\theta}}_0$ should satisfy (3.9) where $n := 4$.

Table 3 presents the results of the 4-BS approach for the Puromycin and the EV data sets, where the presented estimates $\hat{\boldsymbol{\theta}}_{N,J_N} = (\hat{V}_{\max}, \hat{K}_{\mathrm{M}})$ are calculated with only $J = 5$ maximum number of GN iterations on every step $N = 1, 2, 3$ (of the 4-BS algorithm). Notice the remarkable accuracies, of less than $10^{-7} < e$ (achieved in total $0 + 5 + 5 + 5 = 15$ GN iterations) for the requested $\hat{\boldsymbol{\theta}}_{\mathrm{NLLS}}$, for

the Puromycin data set, and of less than $10^{-5} < e$ for the EV data set (achieved in total $0+8+8+5=21$ GN iterations). The digits in bold represents the accurate digits of the NLLS estimates.

Table 3: Convergence of the 4-BS processes, for the MM model estimation.

N	J_N	$\hat{V}_{\max}$	$\hat{K}_{\mathrm{M}}$	e_N	ε_N	R^2	*mare* (%)
• *Puromycin-treated data set,* $J=5$							
0	0	134.413223	0.015372	—	1.53e+5	0.5542	18.
1	5	152.072336	0.029454	0.114	0.343	0.7771	14.94
2	5	184.624253	0.044340	8.13e−6	6.97e−9	0.9330	9.198
3	5	**212.683743**	**0.064121**	3.09e−8	5.09e−11	0.9613	7.
• *EV data set,* $J=8$							
0	0	0.001957	−0.169663	—	0.0176	−0.8679	24.79
1	8	0.028507	0.301639	0.144	3.01e−5	0.0422	29.16
2	8	0.085433	1.314163	0.00445	3.76e−9	0.4085	27.51
3	5	**0.10564**3	**1.70269**0	1.05e−6	6.08e−15	0.9379	21.71

If the 2-BS (or in general the k-BS) process fails to converge for data sets with replicated observations we then propose a practical way, which is demonstrated in the following, for the computational improvement of the 2-BS process when it is applied on such data sets.

Any data set showing replications can be re-arranged in order to help the k-BS method to converge. With this re-arrangement of the Puromycin data set (which contains two V_i readings for each $C_{\mathrm{S},i}$ value), the 2-BS (or the 1-BS) process can now be applied adopting as initial vector $\hat{\boldsymbol{\theta}}_0$ the MM solution vector (using the first two observations) as we did in Example 3.1. Recall that this $\hat{\boldsymbol{\theta}}_0$ cannot be calculated in the case of the original (non-modified) Puromycin data set. The suggestion for helping the performance of the calculations is that we first split the Puromycin data set (and any data set that contains two readings for the depending variable for each value of the independent variable), say $\mathfrak{P}$, into two subsets, say $\mathfrak{P}_1$ and $\mathfrak{P}_2$. Each set contains the two "non-replicated" parts of the original data set and sorted in an increasing order of the C_{S} values, i.e.

$$\mathfrak{P}_1 := \big\{(C_{\mathrm{S},i}, V_i)\big\}_{i=1,3,5,\ldots,11} \quad \text{and} \quad \mathfrak{P}_2 := \big\{(C_{\mathrm{S},i}, V_i)\big\}_{i=2,4,6,\ldots,12} \,. \tag{3.10}$$

Note that $C_{\mathrm{S},i+2} > C_{\mathrm{S},i}$ for $i = 1, 3, ..., 9$ and $i = 2, 4, ..., 10$. In order to re-join them back into a single data set (of 12 observations), with a "smooth" transition from the (increasing) $C_{\mathrm{S},i}$ values of $\mathfrak{P}_1$ data set to the (also increasing) $C_{\mathrm{S},i}$ values of $\mathfrak{P}_2$, we adopt the $\mathfrak{P}_1$ data set as is, and then the observations of $\mathfrak{P}_2$ are put in the reversed (decreasing) order, i.e.

$$\mathfrak{P} = \big\{(C_{\mathrm{S},i}, V_i)\big\}_{i=1,3,5,\ldots,11,12,10,8,\ldots,2} \,. \tag{3.11}$$

The 2-BS process can then be applied, with convergence error threshold $e = 10^{-4}$, while we adopt the MM solution $(V_{\max}, K_{\mathrm{M}})$, as in Example 3.1, as the initial estimates' vector $\hat{\boldsymbol{\theta}}_0$. The first sub-Table of Table 4 provides the GN convergent J_N-th estimates $\hat{\boldsymbol{\theta}}_{N,J_N}$, that calculated with maximum number of GN iterations being only $J = 3$ (at every step $N = 1, 2, ..., 6$ of the 2-BS process). The resulted accuracy of the obtained NLLS estimate $(\hat{V}_{\max}, \hat{K}_{\mathrm{M}})$ is less than $10^{-5} < e$. Similarly, for the "replicated" EV data set, the accuracy for the NLLS estimates is less than $10^{-6} < e$; see the corresponding calculation on the second sub-Table of Table 4. The digits in bold represents again the accurate digits of the NLLS estimates.

Table 4: Convergence of the 2-BS processes for the MM model estimation of the re-arranged P and EV data sets.

N	J_N	$\hat{V}_{\max}$	$\hat{K}_{\mathrm{M}}$	e_N	ε_N	R^2	*mare* (%)
• *Puromycin-treated data set,* $J = 3$							
0	0	112.549618	0.009618	—	0.	1.	0.
1	1	112.549618	0.009618	0.	0.	1.	0.
2	3	170.999898	0.033674	5.76	2.07e+3	0.8736	10.56
3	3	210.839180	0.062539	0.7	75.7	0.9360	9.345
4	3	214.144962	0.064599	1.46e−5	4.73e−8	0.9480	7.739
5	3	212.825859	0.064726	7.6e−6	1.64e−8	0.9389	7.22
6	3	**212.683743**	**0.064121**	1.28e−6	7.47e−10	0.9613	7.
• *EV data set,* $J = 3$							
0	0	0.005853	−0.065438	—	2.33e−17	1.	0.
1	1	0.005853	−0.065438	2.6e−18	1.3e−18	1.	4.e−14
2	3	0.017440	0.158092	0.102	0.000647	0.6203	14.26
3	3	0.078120	1.285978	0.458	0.00152	0.9553	19.51
4	3	0.126137	2.411911	0.325	9.89e−6	0.9655	14.05
5	3	0.104970	1.694986	0.272	0.000761	0.9419	21.76
6	3	**0.105643**	**1.702690**	2.93e−7	2.52e−16	0.9379	21.71

The re-arrangement, as in (3.11), of the data set (which affects the order in which the observations are sequentially inserted into the BS process) can also improve the performance of the BS process even for initial guesses $\hat{\boldsymbol{\theta}}_0$ for which the 2-BS, or even the 1-BS, process normally could not converge. The following Example demonstrates this improvement.

Example 3.3. Considering the initial estimates' guess $\hat{\boldsymbol{\theta}}_0 = (80, 0)$ and letting $e = 10^{-4}$ to be the convergence error threshold, the 2-BS (as well as the 1-BS process) fails to converge, when it is applied on the original Puromycin data set. However, the 2-BS process converges, to the requested NLLS estimate, when the data set is re-arranged, as described in (3.11), even with few $(J = 3)$ allowed GN iterations at every step of 2-BS, or 1-BS, process. Table 5 presents the performance of the 2-BS approach (first sub-Table) as well as of the 1-BS approach

(second sub-Table). The accuracy of the obtained NLLS estimates derived from the 2-BS process is less than $10^{-5} < e$, while the 1-BS process results an accuracy less than $10^{-4} < e$. The digits in bold represents also here the accurate digits of the NLLS estimates.

Table 5: Convergence of the 1-BS and the 2-BS processes, applied on the re-arranged Puromycin data set.

N	J_N	$\hat{V}_{\max}$	$\hat{K}_{\mathrm{M}}$	e_N	ε_N	R^2	*mare* (%)
• *2-BS process,* $J = 3$							
0	0	80.	0.	—	1.33e+4	−0.3832	11.39
1	3	112.410518	0.009554	2.04	275.	1.	0.062
2	3	170.980951	0.033657	5.8	2.1e+3	0.8736	10.56
3	3	210.839081	0.062538	0.702	76.2	0.9360	9.345
4	3	214.144962	0.064599	1.46e−5	4.74e−8	0.9480	7.739
5	3	212.825859	0.064726	7.6e−6	1.64e−8	0.9389	7.22
6	3	**212.683743**	**0.064121**	1.28e−6	7.47e−10	0.9613	7.
• *1-BS process,* $J = 3$							
0	0	80.	0.	—	1.33e+4	−0.3832	11.39
1	3	112.410518	0.009554	2.04	275.	1.	0.062
2	3	136.051800	0.017418	0.563	25.3	0.8983	5.843
3	3	172.114585	0.034640	1.93	221.	0.8739	10.62
4	3	199.818136	0.053452	0.398	13.1	0.9129	10.46
5	3	210.857217	0.062575	0.00718	0.0089	0.9360	9.341
6	3	211.078437	0.062765	9.91e−10	8.75e−12	0.9471	8.015
7	3	214.144962	0.064599	9.91e−06	2.13e−8	0.9480	7.739
8	3	213.560104	0.067269	0.00066	0.000118	0.9397	7.37
9	3	212.825859	0.064726	0.000225	1.56e−5	0.9389	7.22
10	3	212.086616	0.062938	5.28e−5	1.07e−6	0.9440	6.93
11	3	**212.683743**	**0.064121**	1.02e−5	4.46e−8	0.9613	7.

4. DISCUSSION

Certain aspects of the MM model, so essential in Risk Analysis as far as to form an enzyme-substrate complex especially to pharmacokinetics studies, were discussed in this paper, either theoretical (see Theorem 2.1) or computational (see the provided examples in Section 3). As far as the optimal design is concerned, recall Kitsos [14] and (1.8), the design depends on the nonlinear term K_{M}. When the D-optimal design problem was viewed from the MM model perspective it can be formed into the following compact form:

If $C_{\mathrm{S}} \in (0, C_{\mathrm{U}}]$ the locally D-optimal design at $K_{\mathrm{M}} = K_0$ which allocates the half of the observations V with optimum concentration

$$C_{\mathrm{S}}^{\mathrm{opt}} = \frac{K_0 C_{\mathrm{U}}}{2K_0 + U}\,, \tag{4.1}$$

with C_{U} being the maximum allowable substrate concentration. If $C_{\mathrm{U}} \gg K_0$ the locally D-optimal design ξ is

$$\xi^* = \begin{pmatrix} C_{\mathrm{U}} & K_0 \\ 0.5 & 0.5 \end{pmatrix}. \tag{4.2}$$

The corresponding value of the determinant of the D-optimal design is

$$d = \frac{V_{\max}^2 C_{\mathrm{U}}^6}{16 K_0^2 (K_0 + C_{\mathrm{U}})^6}\ . \tag{4.3}$$

See also Endrenyi and Chan [6]. If $C_{\mathrm{S}} \in [C_{\mathrm{L}}, C_{\mathrm{U}}]$ then the optimum C_{S}, through (4.1), is given by

$$C_{\mathrm{S}}^{\mathrm{opt}} = \frac{K_0 C_{\mathrm{U}}}{2K_0 + (C_{\mathrm{U}} - C_{\mathrm{L}})}\ , \qquad K_0 > 0, \quad 0 < C_{\mathrm{L}} < C_{\mathrm{U}}\ . \tag{4.4}$$

From the above relations and the average-per-observation information matrix as in (1.8) it is clear, due to Theorem 2.1, that there might be more than one NLLS estimates. This is a new point of view of the design and actually a crucial one. One could choose as "best" among the D-optimal designs the one which provides minimum value of the corresponding (4.3) evaluation, which is a common situation. Therefore, there might be (locally) D-optimal designs corresponding to the analytical real-valued NLLS estimates. The final adopted D-optimal design can be chosen in principle, as noted also above, to be the one that provides minimum $\det\big(\mathbf{M}(\hat{\boldsymbol{\theta}}, \xi)\big)$. It is clear from relations (4.3) and (4.4) that the right choice of the existent different values for the parameter $\boldsymbol{\theta}$ is essential for the Risk Analysis study under investigation. It is why we provide static or sequential design approach to reach the appropriate selected real value for $\boldsymbol{\theta}$. It is therefore crucial what we prove: there is always one real value, and thus the Risk Analysis can always proceeded. How we proceed on Risk Analysis were more than one real value for $\boldsymbol{\theta}$ exists, has been extensively discussed, see Theorem 2.1 and the Examples.

An analytic methodology for the nonlinear least squares estimation (NLLS) was also introduced and compared against the four known linearization technics. The analytic formulation of this method indicated that the NLLS estimation of the MM model was, in general, not unique. Moreover, an iterative scheme for the NLLS estimation was also introduced, called the Batch Sequential (BS) process, and tested in various cases of data sets which showing readings replication or not. Despite that the BS is an iterative process, meaning that an initial estimates' guess is needed, a different approach was discussed and tested which provides an initial estimate that increases the convergence performance of the BS algorithm. Finally, certain examples demonstrate all the proposed methods.

ACKNOWLEDGMENTS

We would like to thank the referees for their useful comments which improve this paper.

REFERENCES

[1] Bates, D.M. and Watts, D.G. (1998). *Nonlinear-Regression Analysis and its Applications*, Oliver & Boyd, Edinburgh.

[2] Biebler, K.E.; Schreiber, K. and and Bode, R. (2009). *Solutions to the Michaelis–Menten kinetics are not necessarily unique*. In "AIP Conference Proceedings", 1148, 37 (doi: 10.1063/1.3225321).

[3] Currie, D.J. (1982). Estimating Michaelis–Menten parameters: bias, variance and experimental design, *Biometrics*, **38**, 907–919.

[4] Dowd, J.E. and Riggs, D.S. (1965). A comparison of estimates of Michaelis–Menten kinetic constants from various linear transformations, *J. of Biological Chemistry*, **240**, 863–869.

[5] Draper, N.R. and Smith, H. (1998). *Applied Regression Analysis, 3rd ed.*, Wiley.

[6] Endrenyi, L. and Chan, F.Y. (1981). Optimal design of experiments for the estimation of precise hyperbolic kinetic and binding parameters, *J. Theor. Biol.*, **90**, 241–263.

[7] Ford, I.; Titterington, D.M. and Kitsos, C.P. (1989). Recent advances in nonlinear experimental design, *Technometrics*, **31**(1), 49–60.

[8] Hadeler, K.; Jukić, D. and Sabo, K. (2007). Least squares problems for Michaelis–Menten kinetics, *Mathematical Methods in the Applied Sciences*, **30**(11), 1231–1241.

[9] Hill, P.D.H. (1980). D-optimal designs for partially nonlinear regression models, *Technometrics*, **22**, 275–276.

[10] Jennrich, R.J. (1969). Asymptotic properties of nonlinear least squares estimators, *Ann. Math. Stat.*, **40**, 633–643.

[11] Jukić, D.; Sabo, K. and Scitovski, R. (2007). Total least squares fitting Michaelis–Menten enzyme kinetic model function, *Journal of Computational and Applied Mathematics*, **201**, 230–246.

[12] Kitsos, C.P. (2013). *Optimal Experimental Design for Non-Linear Models*, Springer (ISSN: 2191-544X).

[13] Kitsos, C.P. (2012). *Cancer Bioassays: A Statistical Approach*, Lampert Acad. Pub.

[14] Kitsos, C.P. (2001). Design aspects for the Michaelis–Menten model, *Biometrical Letters*, **38**, 53–66.

[15] Kitsos, C.P. (1989). Fully sequential procedures in nonlinear design problems, *Comput. Stat. and Data Analysis*, **8**, 13–19.

[16] Kitsos, C.P. and Toulias, T.L. (2015). *On the Michaelis–Menten model.* In "Proceedings of the ICRA6 International Conference" (M. Guillen, A. Juan, H. Ramanhinho, I. Serra and C. Serrat, Eds.), 449–456.

[17] Michaelis, L. and Menten, M.L. (1913). Kinetics for intertase action, *Biochemische Zeitung*, **49**, 333–369.

[18] Toulias, T.L. and Kitsos, C.P. (2016). Fitting the Michaelis–Menten model, *Journal Of Computational And Applied Mathematics*, **296**, 303–319.

REVSTAT – Statistical Journal
Volume 14, Number 2, April 2016, 119–138

SKEWNESS INTO THE PRODUCT OF TWO NORMALLY DISTRIBUTED VARIABLES AND THE RISK CONSEQUENCES

Authors: Amilcar Oliveira
– Center of Statistics and Applications (CEAUL), University of Lisbon, and Department of Sciences and Technology, Universidade Aberta, Portugal
amilcar.oliveira@uab.pt

Teresa A. Oliveira
– Center of Statistics and Applications (CEAUL), University of Lisbon, and Department of Sciences and Technology, Universidade Aberta, Portugal
teresa.oliveira@uab.pt

Antonio Seijas-Macias
– Department of Applied Economics II, Universidade da Coruña, and Universidad Nacional de Educación a Distancia (UNED), Spain
antonio.smacias@udc.es

Received: September 2015 Revised: February 2016 Accepted: February 2016

Abstract:

- The analysis of skewness is an essential tool for decision-making since it can be used as an indicator on risk assessment. It is well known that negative skewed distributions lead to negative outcomes, while a positive skewness usually leads to good scenarios and consequently minimizes risks. In this work the impact of skewness on risk analysis will be explored, considering data obtained from the product of two normally distributed variables. In fact, modelling this product using a normal distribution is not a correct approach once skewness in many cases is different from zero. By ignoring this, the researcher will obtain a model understating the risk of highly skewed variables and moreover, for too skewed variables most of common tests in parametric inference cannot be used. In practice, the behaviour of the skewness considering the product of two normal variables is explored as a function of the distributions parameters: mean, variance and inverse of the coefficient variation. Using a measurement error model, the consequences of skewness presence on risk analysis are evaluated by considering several simulations and visualization tools using R software ([10]).

Key-Words:

- *product of normal variables; inverse coefficient of variation; skewness; probability risk analysis; measurement error model.*

AMS Subject Classification:

- 62E17, 62E10.

1. INTRODUCTION

Consequences for the presence of skewness are very important, especially in risk analysis. When distribution of the expected value is skewed produces distortions on the decisions of the risk-neutral decision maker. In Mumpower and McClelland ([9]) authors analyse the consequences on a model of random measurement error. Another example, in valuation risk of assets, the risk-averse investors prefer positive skewness Krans and Litzenberg ([8], Harvey and Siddique [6]) and the effect of skewness on the R^2 of the model has influenced over the predictability of the model of assets Cochran ([4]).

Our objective in this paper is to study the relation between skewness of the distribution of a product of two normal variables and the parameters of these normal distributions. Our work has two focus: from a theoretical point of view using the moment-generating function, and through several simulations, using Monte-Carlo methods we estimate the skewness of the product of two variables.

Distribution of the product of normal variables is an open problem in statistics. First work has been undertaken by Craig ([3]), in his early paper, who was actually the first to determinate the algebraic form of the moment-generating function of the product. In Aroian and Taneja ([2]) proved the approximation of the product using the standardized Pearson type III distribution. But nowadays, the problem is not closed; although the product of two normal variables is not, in general, normally distributed; however, under some conditions, it is showed that the distribution of the product can be approximated by means of a Normal distribution Aroian and Taneja ([2]). The presence of the product of normal variables is well-known in Risk analysis Hayya and Ferrara ([7]), where functional relationships concerning two normally distributed variables (correlated or non-correlated) are encountered.

There are several test to estimate the normality of a sample, but for large size sample results are not always correct Deb and Sefton ([5]). The most accurate test for large size is skewness test. In this paper, we use the moment-generating function for analysing the value of skewness for a product of two normally distributed variables. We considered the influence of three parameters from the two distributions: mean, variance and correlation. Using the formula for skewness, we can calculate the value of the skewness for the product as a function of two set of parameters: First, where the mean, the variance and correlation between the two distributions are used for calculations. The second one is formed the inverse of the coefficient of variation for each distribution and the correlation.

At section 2, the moment-generating function for a product of two normally distributed variables is introduced. The formulas for three parameters of the product: mean, variance (standard deviation) and skewness are studied and

the evolution of skewness for the product of two normal variables is analysed in Section 3. Several cases are considered: taking into account first, the presence of correlation between both variables is assumed; second, the two normally distributed variables are uncorrelated. The influence of the parameters, mean and standard deviation of the two variables is analysed. The graphical visualization of the results is incorporated. In Section 4, an analysis of the effect of skewness for a model of random measurement error is introduced. Finally, Section 5 contains conclusions of the paper.

2. MOMENTS OF THE PRODUCT OF TWO NORMAL VARIABLES

Let X and Y be two normal probability functions, with means μ_x and μ_y and standard deviations σ_x and σ_y, respectively, r the coefficient of correlation and the inverses of the coefficient of variation, for the two variables, are: $\rho_x = \frac{\mu_x}{\sigma_x}$ and $\rho_y = \frac{\mu_y}{\sigma_y}$.

Craig ([3]) determined the moments, seminvariants, and the moment generating function of $z = \frac{x\,y}{\sigma_x\,\sigma_y}$. The moment generating function of z, $M_z(t)$ is:

$$M_z(t) = \frac{\exp \frac{\left(\rho_x^2+\rho_y^2-2r\,\rho_x\,\rho_y\right)t^2 + 2\rho_x\,\rho_y\,t}{2\left(1-(1+r)t\right)\left(1+(1-r)t\right)}}{\sqrt{\left(1-(1+r)t\right)\left(1+(1-r)t\right)}}\,, \tag{2.1}$$

where t is the order of the moment.

Let μ_z and σ_z be the mean and the standard deviation of z. Values of mean and standard deviation and skewness of z are calculated as (see ([3]) and ([1]):

$$\mu_z = \rho_x\,\rho_y + r\,, \tag{2.2}$$

$$\sigma_z = \sqrt{\rho_x^2 + \rho_y^2 + 2r\,\rho_x\,\rho_y + 1 + r^2}\,, \tag{2.3}$$

$$\alpha_3 = \frac{2\left(3\,\rho_x\,\rho_y + r^3 + 3\,\rho_x\,\rho_y\,r^2 + 3\,r\left(\rho_x^2+\rho_y^2+1\right)\right)}{\left(\rho_x^2+\rho_y^2+r^2+2\,\rho_x\,\rho_y\,r+1\right)^{3/2}}\,. \tag{2.4}$$

An alternative approach, without using the inverse of the coefficient of variation, can be obtained.

Proposition 2.1. *Let $x \sim N(\mu_x, \sigma_x^2)$ and $y \sim N(\mu_y, \sigma_y^2)$ be two normal variables with correlation r . Defining $x = x_0 + z_1$ and $y = x_0 + z_2$, where*

$$\begin{bmatrix} x_0 \\ z_1 \\ z_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ \mu_x \\ \mu_y \end{bmatrix} \begin{bmatrix} r\,\sigma_x\,\sigma_y & 0 & 0 \\ 0 & \sigma_x^2 - r\,\sigma_x\,\sigma_y & 0 \\ 0 & 0 & \sigma_y^2 - r\,\sigma_x\,\sigma_y \end{bmatrix} \right), \tag{2.5}$$

the two variables x and y are decomposed into independent summands, one of which is shared between them. Then, the moment-generating function of $z = x\,y = (x_0 + z_1)\,(x_0 + z_2)$ is

$$M_z[t] = \frac{\exp \frac{t\left(t\mu_y^2\sigma_x^2 + t\mu_x^2\sigma_y^2 - 2\,\mu_x\,\mu_y(-1+t\,r\,\sigma_x\,\sigma_y)\right)}{2+2\,t\,\sigma_x\,\sigma_y\left(-2r+t(-1+r^2)\,\sigma_x\,\sigma_y\right)}}{\sqrt{1+t\,\sigma_x\,\sigma_y\left(-2r+t(-1+r^2)\,\sigma_x\,\sigma_y\right)}}\,. \tag{2.6}$$

Proof: The moment-generating function of $z = x\,y$ is the same that $z = (x_0 + z_1)\,(x_0 + z_2)$, that is the product of two independent variables, then we know that

$$M_z[t] = \int_{-\infty}^{\infty} e^{tz} f(z)\,\mathrm{d}z\,. \tag{2.7}$$

The join probability density function (pdf) $f(z)$ could be written as the product of the independent three marginal pdf of the variables,

$$f(z) = f_{x_0}(x_0)\,f_{z_1}(z_1)\,f_{z_2}(z_2) = \frac{\exp\left(-\frac{x_0^2}{2\,r\,\sigma_x\,\sigma_y} - \frac{(z_1-\mu_x)^2}{2(\sigma_x^2-r\,\sigma_x\,\sigma_y)} - \frac{(z_2-\mu_x)^2}{2\left(\sigma_y^2-r\,\sigma_x\,\sigma_y\right)}\right)}{2\sqrt{2}\,\pi^{3/2}\sqrt{r\,\sigma_x\,\sigma_y}\,\sqrt{\sigma_x^2-r\,\sigma_x\,\sigma_y}\,\sqrt{\sigma_y^2-r\,\sigma_x\,\sigma_y}}\,. \tag{2.8}$$

Then,

$$\begin{aligned} M_z[t] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{tz} f_{x_0}(x_0)\,f_{z_1}(z_1)\,f_{z_2}(z_2)\;\mathrm{d}x_0\,\mathrm{d}z_1\,\mathrm{d}z_2 \\ &= \frac{1}{2\sqrt{2}\,\pi^{3/2}\sqrt{r\,\sigma_x\,\sigma_y}\,\sqrt{\sigma_x^2-r\,\sigma_x\,\sigma_y}\,\sqrt{\sigma_y^2-r\,\sigma_x\,\sigma_y}} \\ &\quad\cdot \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{t(x_0+z_1)(x_0+z_2) - \frac{x_0^2}{2\,r\,\sigma_x\,\sigma_y} - \frac{(z_1-\mu_x)^2}{2(\sigma_x^2-r\,\sigma_x\,\sigma_y)} - \frac{(z_2-\mu_y)^2}{2(\sigma_y^2-r\,\sigma_x\,\sigma_y)}}\;\mathrm{d}x_0\,\mathrm{d}z_1\,\mathrm{d}z_2 \end{aligned} \tag{2.9}$$

where we have the following assumptions: $t \in \mathbb{Z}$ non negative, $\sigma_x > 0$, $\sigma_y > 0$, $\sigma_x, \sigma_y, \mu_x, \mu_y$ all of then real numbers and $-1 \le r \le 1$. Solving the integral (2.9) results for theses assumptions,

$$\frac{\sqrt{\sigma_y\left(-(r\sigma_x-\sigma_y)\right)}\,\sqrt{-r\sigma_y(r\sigma_y-\sigma_x)}\,\exp\left(\frac{t\left(\mu_x^2\sigma_y^2 t - 2\mu_x\,\mu_y(r\sigma_x\sigma_y t-1) + m u_y^2\sigma_x^2 t\right)}{2\left((r-1)\sigma_x\sigma_y t-1\right)\left((r+1)\sigma_x\sigma_y t-1\right)}\right)}{\sqrt{\sigma_y(\sigma_y-r\sigma_x)}\,\sqrt{r\sigma_y(\sigma_x-r\sigma_y)}\,\sqrt{(r^2-1)\,\sigma_x^2\sigma_y^2 t^2 - 2r\sigma_x\sigma_y t + 1}}\,. \tag{2.10}$$

Then, simplifying this expression in (2.10) results (2.6), as the expression of the moment-generating function of the product $z = x\,y$. □

Derivatives of order i of (2.6) provides moments of order i, for $i = 1, 2, ...$ The moments of the distribution of the product of two normal variables are calculated:

1. Mean: First derivative of the moment generating (2.6) function respect t for $t = 0$:

$$E[z] = \mu_z = \mu_x \mu_y + r \sigma_x \sigma_y \ ; \tag{2.11}$$

2. Variance: Difference between second moment and first moment up to 2

$$\mathrm{Var}[z] = \sigma_z^2 = \mu_y^2 \sigma_x^2 + 2 \mu_x \mu_y r \sigma_x \sigma_y + \left(\mu_x^2 + (1 + r^2) \sigma_x^2\right) \sigma_y^2 \ ; \tag{2.12}$$

3. Skewness: Quotient between the third central moment and the second central moment up to the 3/2.

$$\alpha_3[z] = \frac{\left(6 \mu_y^2 r \sigma_x^3 \sigma_y + 6 \mu_x \mu_y (1 + r^2) \sigma_x^2 \sigma_y^2 + 2 r \left(3 \mu_x^2 + (3 + r^2) \sigma_x^3 \sigma_y\right) \sigma_y^2\right)}{\left(\mu_y^2 \sigma_x^2 + 2 \mu_x \mu_y r \sigma_x \sigma_y + \left(\mu_x^2 + (1 + r^2) \sigma_x^2\right) \sigma_y^2\right)^{3/2}} . \tag{2.13}$$

Attention will be paid to the evolution of skewness. Skewness for a normal distributed variable would be zero. For a product of two normal variables, skewness zero would be a proof of normality, other values should be carefully analysed.

3. EVOLUTION OF SKEWNESS OF THE PRODUCT OF TWO NORMAL VARIABLES

In the previous section the formula for skewness of the product of two normally distributed variables was introduced (see equation 2.13). Obviously, there are five factors that have a certain influence into the value of skewness: mean and standard deviation of each of the variables into product and the correlation between them.

3.1. Product of two correlated normally distributed variables

According Proposition 1, we can formulated several specific cases for the product of two normally distributed variables, and study evolution of the skewness.

Corollary 3.1. *For the product of two correlated normally distributed variables, three cases are considered:*

a) *Product of two standard normal distributions $N(0, 1)$ with $r = 1$. This a very special case, the product of this two variables follows a Chi-Square with one degree of freedom. In this case, skewness of product is $2\sqrt{2}$ than is bigger than zero and equals the theoretical value of skewness for the Chi-Square distribution with 1 degree of freedom ($\sqrt{8}$).*

b) *Two normal variables with same mean* $\mu_x = \mu_y = \mu$ *and same unit standard deviation* $\sigma_x = \sigma_y = 1$. *In this case we considered different values for correlation* r*:*

$$\alpha_3[z] = \frac{2\left(3\mu^2 r + 3\mu^2(1+r^2) + r(3+3\mu^2+r^2)\right)}{\left(1+2\mu^2+2\mu^2 r+r^2\right)^{3/2}}\,. \tag{3.1}$$

In Figure 1, evolution of the skewness is represented for two factors: mean and correlation. When the mean of two variables is zero, skewness is a increasing function of the correlation. When $\mu = 0$ *and* $r = 1$ *we have the Chi-Square case.When* r *tends to zero, then* $\alpha_3[z]$ *tends* 0 *when ratio (inverse of the coefficient of variation)* $\frac{\mu}{\sigma} = 0$*, but as the ratio increases the skewness rises rapidly, until it is at its maximum when the inverse of the coefficient of variation is one, then* $\mu = \sigma = 1$.

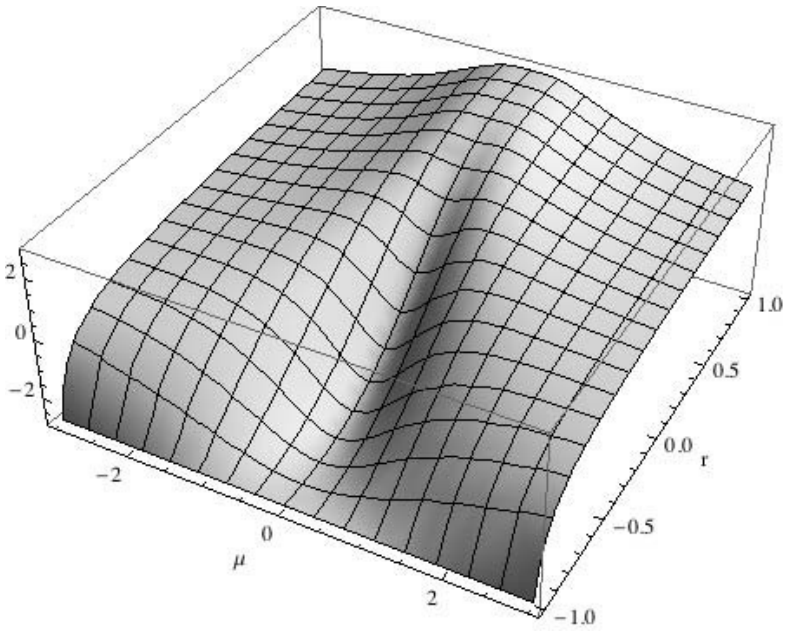

Figure 1: Skewness for product two normal variables same mean and standard deviation equals to 1.

c) *Two normal variables with the same mean* $\mu_x = \mu_y = \mu$ *and the same standard deviation* $\sigma_x = \sigma_y = \sigma$. *Considering two distribution with the same parameters (mean and standard deviation). In general, skewness tends to zero as ratio tends to infinity. The closer* $|r|$ *is to one the slower the approach of skewness to zero. When* $r = 1$ *then we have skewness of a Chi-Square Distribution* $(2\sqrt{2})$ *(see Figure 2).*

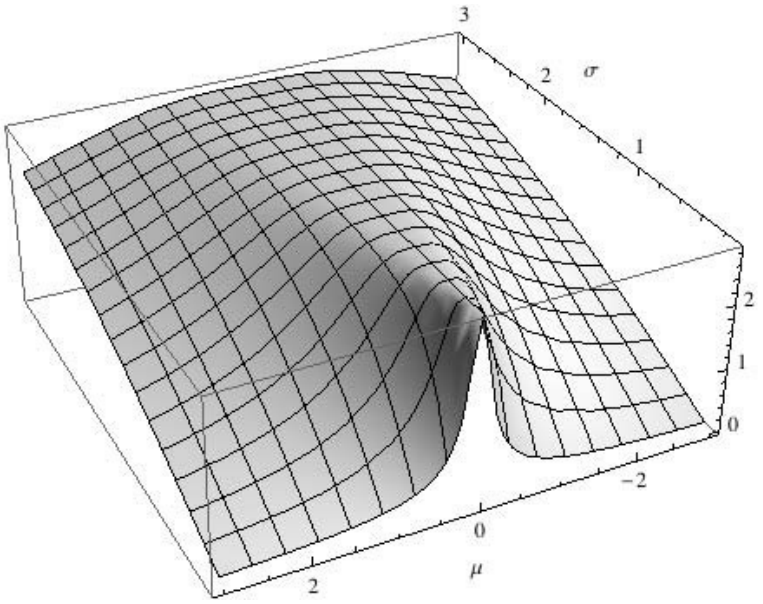

Figure 2: Skewness for product two normal variables $r = 1$.

When considering different mean or different standard deviation, the evolution of skewness presents more variability as a function of the values of the parameters, but in a rough inspection of the graphics we are identify several aspects:

1. If we consider the same standard deviation and different values for the mean, skewness zero is very common when σ tends to zero and the means are different.
2. When standard deviation increases the skewness increases for the same values of correlation and mean.
3. If we consider the same mean, skewness is very common, and only when the standard deviation are lower, skewness zero exists.
4. In general, the presence of correlation has effect on the presence of skewness, skewness zero or lower is more common when r tends to zero.

Now, we are going to study the skewness of uncorrelated distributions ($r = 0$).

3.2. Product of two uncorrelated normal distributions

Two uncorrelated normal variables ($r = 0$) are now considered. When the two variables are uncorrelated $r = 0$ and the value of skewness is a function of

only 4 parameters (means and variances), then:

$$\alpha_3[z] = \frac{6\,\mu_x\,\mu_y\,\sigma_x^2\,\sigma_y^2}{\left(\mu_y^2\,\sigma_x^2 + (\mu_x^2 + \sigma_x^2)\,\sigma_y^2\right)^{3/2}}\,. \tag{3.2}$$

3.2.1. Influence of inverse of the coefficient of variation

For the moment-generating function of the product of these two variables equation (2.6) is used, considering the inverse of the coefficient of variation $\rho_x = \frac{\mu_x}{\sigma_x}$ and $\rho_y = \frac{\mu_y}{\sigma_y}$. Now, skewness of the product of variables $z = x\,y$ will be:

$$\alpha_3[z] = \frac{6\,\rho_x\,\rho_y}{\left(\rho_y^2 + \rho_x^2 + 1\right)^{3/2}}\,. \tag{3.3}$$

Skewness depends on ratios ρ_x and ρ_y. In Figure 3, the skewness as a function of this ratios is illustrated.

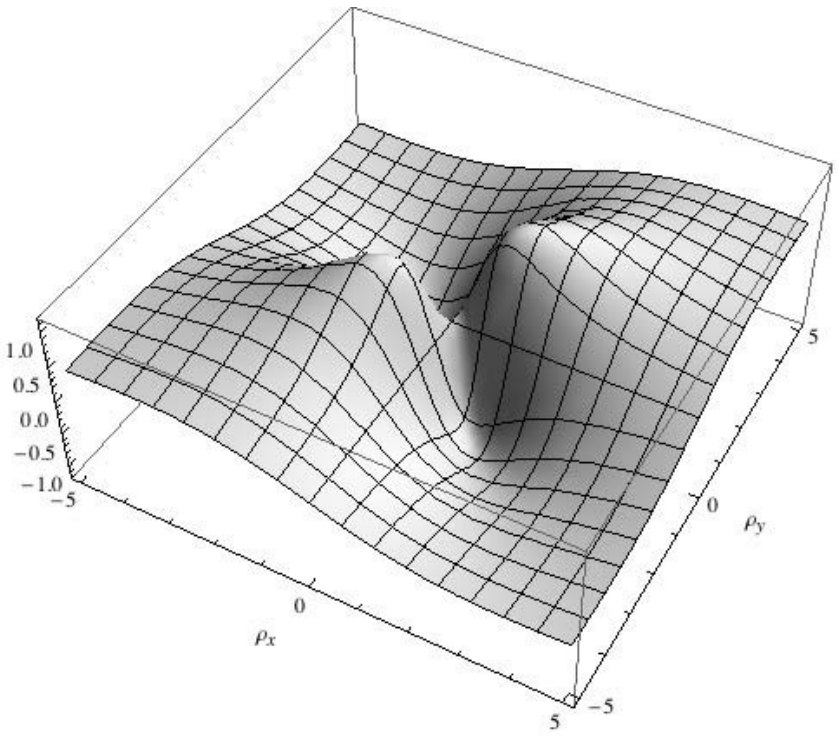

Figure 3: Skewness as a function of ratios ρ_x and ρ_y.

When $\rho_x \to 0$ and $\rho_y \to 0$ then $\alpha_3[z] \to 0$, that is true too, when one of them tends to zero and the other doesn't tend to infinity. The approach to the normal distribution for the product of two variables will be influenced by the existence

of large ρ_x, ρ_y values. Obviously, as the skewness decreases the approximation to normal distribution improves. Differentiating $\alpha_3[z]$ with respect to two variables, ρ_x and ρ_y, equals zero and solving we find the extreme points of skewness:

$$\left(\frac{\partial\alpha_3[z]}{\partial\rho_x}, \frac{\partial\alpha_3[z]}{\partial\rho_y}\right) = \left(\frac{6\rho_y\left(1-2\rho_x^2+\rho_y^2\right)}{\left(1+\rho_x^2+\rho_y^2\right)^{5/2}}, \frac{6\rho_x\left(1-2\rho_y^2+\rho_x^2\right)}{\left(1+\rho_x^2+\rho_y^2\right)^{5/2}}\right). \tag{3.4}$$

There are five points: $(0,0)$, $(1,1)$, $(-1,1)$, $(1,-1)$, $(-1,-1)$. Looking for the extreme values we are considering the Hessian matrix $H(\rho_x,\rho_y)$ say equals to:

$$\begin{pmatrix} \frac{18\rho_x\rho_y\left(2\rho_x^2-3(1+\rho_y^2)\right)}{(1+\rho_x^2+\rho_y^2)^{7/2}} & -\frac{6\left(1-2\rho_x^4+\rho_y^2+2\rho_y^4+\rho_x^2(1-11\rho_y^2)\right)}{(1+\rho_x^2+\rho_y^2)^{7/2}} \\ -\frac{6\left(1-2\rho_x^4+\rho_y^2+2\rho_y^4+\rho_x^2(1-11\rho_y^2)\right)}{(1+\rho_x^2+\rho_y^2)^{7/2}} & -\frac{18\rho_x\rho_y\left(3+\rho_x\rho_y(3+3\rho_x^2-2\rho_y^2)\right)}{(1+\rho_x^2+\rho_y^2)^{7/2}} \end{pmatrix}. \tag{3.5}$$

Then, we have: in $H(0,0)$ there is a saddle point; in $H(1,1)$ and $(H-1,-1)$ there are maxima with value $\frac{2}{\sqrt{3}}$ and in $H(-1,1)$ and $H(1,-1)$ there minima with value $-\frac{2}{\sqrt{3}}$. Thus, the skewness of z is largest when $|\mu_x| = \sigma_x$ and $|\mu_y| = \sigma_y$. If x and y have equal standard deviation (σ), the skewness of the product will be largest when $\mu_x = \mu_y = \sigma$. Figure 4 represents different values of skewness for combination of points for ρ_x and ρ_y from -5 until 5, these values are represented at "x"-axe. (At the center of graph are values with $\rho_x = \rho_y = 0$).

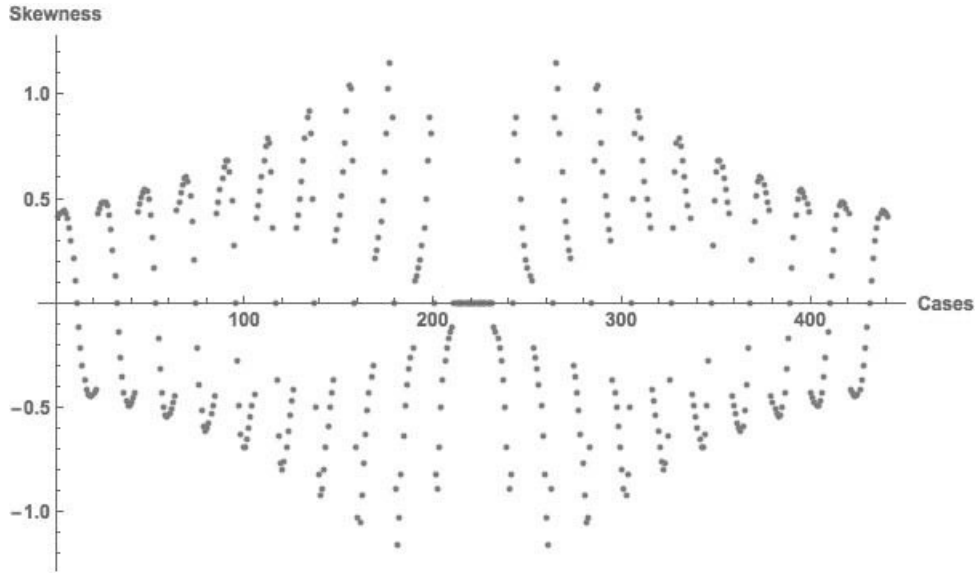

Figure 4: Skewness combination ρ_x and ρ_y in $[-5,5]$.

Table 1 resumes values of skewness for several products of two normal variables. The theoretical value, according to (3.3) and value for a simulation using Monte-Carlo Simulation for the product of two normal variables with 1.000.000 points is presented:

Table 1: Skewness for product of two normal variables.

Parameters	**Theoretical Skewness**	**Skewness**
$\mu_x = 1,\ \mu_y = 0.5,\ \sigma = 1$	0.88889	0.883946
$\mu_x = 5,\ \mu_y = 0.5,\ \sigma = 1$	0.111531	0.109204
$\mu = 1,\ \sigma_x = 1,\ \sigma_y = 0.5$	0.816497	0.811504
$\mu = 1,\ \sigma_x = 1,\ \sigma_y = 5$	0.411847	0.404888

The effect of the inverse of the coefficient of variation is correct for the two first rows in Table 1. In the first row we have two distributions with $\rho_x = 1$, $\rho_y = 0.5$ and skewness is high (0.88); in the second one the ratios for two distributions are higher ($\rho_x = 5,\ \rho_y = 0.5$) and skewness is lower (0.11). But this tendency is not correct when we consider examples in row 3 and 4 in Table 1. In row 3, we have $\rho_x = 1,\ \rho_y = 2$, these values are higher than $\rho_x = 1,\ \rho_y = 0.2$, values in row 4, but the evolution of skewness is inverse. Then, there is a inverse dependency between ρ and skewness but influence of μ is very important and may change the tendency.

A previous analysis of the influence of the value of mean and variance and their effect over skewness was considered. In next section, a more exhaustive study of this influence is explored.

Graph evolution of skewness as a function of mean and standard deviation is represented in Figure 5. In the first graph, we consider $\mu_x = 1,\ \mu_y, \sigma_y = 1$ and we depict skewness *vs.* σ_x, for three different values of $\mu_y = 0.5, 1, 5$. When σ_x increase, skewness increase until a point where decrease, higher values of μ_y produce faster decreasing for skewness. In this cases, there is a direct relation between skewness and ratio $\frac{\mu}{\sigma}$. At the second one, we consider $\mu_y = 1,\ \sigma_x = 1,$ σ_y and we depict skewness *vs.* μ_x, for three different values of $\sigma_y = 0.2, 1, 2$. The effect is very similar to the latter, but now: when σ is lower the decreasing of skewness is smaller than for higher values of σ, that is, the inverse effect between skewness and ratio inverse of the coefficient of variation $\frac{\mu}{\sigma}$.

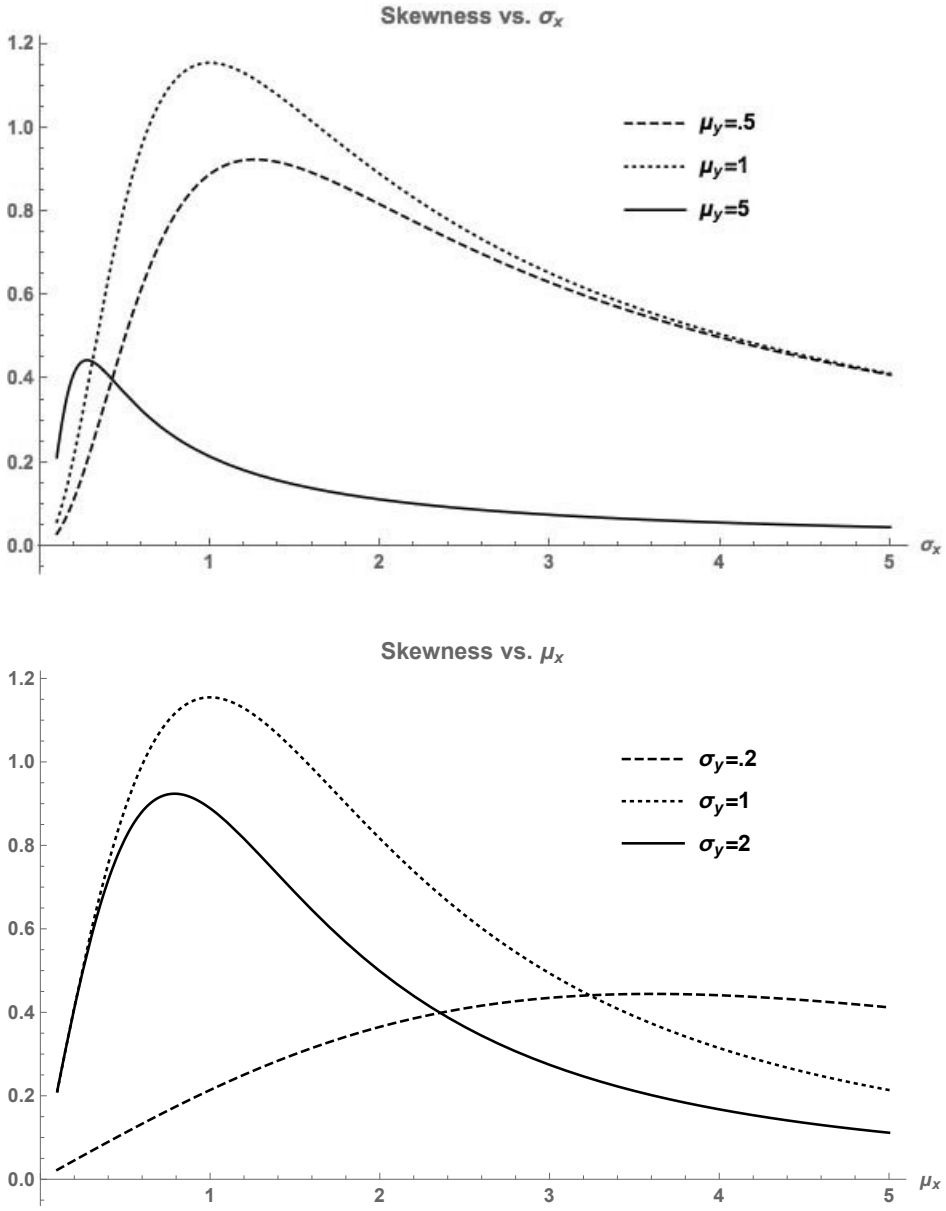

Figure 5: Skewness.

3.2.2. Influence of parameters: mean and variance

Skewness for the product of two normal functions is a function of the values of parameters of the distributions. In the previous section it was show that the influence of the ratio (inverse of the coefficient of variation) was direct, so that the higher the value of it, there was a lower value of skewness, but this influence

appeared nuanced if we considered the particular values for the parameters of the distribution. There is a strong dependence on the value of the standard deviation, so that the higher the value faster the skewness approaches zero, which contradicts, in part, that a greater ratio lower value of skewness. On the other hand also the average value had influence that the higher the average value also decreases the value of skewness.

From the moment-generating function as in (2.6), skewness (Sk) of a product of two normally distributed variables $x \sim N(\mu_x, \sigma_x)$ and $y \sim N(\mu_y, \sigma_y)$ is a function of the parameters of the two uncorrelated distributions:

$$Sk = \frac{6\,\mu_x\,\mu_y\,\sigma_x^2\,\sigma_y^2}{\left(\mu_y^2\sigma_x^2 + (\sigma_x^2 + \mu_x^2)\,\sigma_y^2\right)^{3/2}}\,. \tag{3.6}$$

When this value is zero, then skewness is zero, as a normal distribution. In the following cases, skewness of the product of two normal variables is zero, $Sk = 0$:

$$\mu_x\,\mu_y\,\sigma_x \neq 0 \quad \text{and} \quad \sigma_y = 0\,; \tag{3.7}$$

$$\mu_x\,\mu_y \neq 0\,, \quad \sigma_y \neq 0 \quad \text{and} \quad \sigma_x = 0\,; \tag{3.8}$$

$$-2\sigma_x\sqrt{\mu_y^2 + \sigma_y^2} = 0\,, \quad \mu_y^2\sigma_x + \sigma_x\sigma_y^2 \neq 0 \quad \text{and} \quad \mu_x = 0\,; \tag{3.9}$$

$$\mu_x \neq 0\,, \quad \mu_x^2\sigma_y + \sigma_x^2\sigma_y \neq 0 \quad \text{and} \quad \mu_y = 0\,. \tag{3.10}$$

Equations (3.8) and (3.9) represent Y and X as constants, then both cases are not normal variables. For the other two cases, we can use the normal distribution as an approach of the product. For another situations, the product of two normal variables is skewed and normal distribution doesn't represent a good approach for the product.

The evolution of Sk (3.6) follows the evolution of the equation (3.3). There are three saddle point at $(\sigma_x = 0)$, $(\sigma_y = 0)$ and $(\mu_x = \mu_y = 0)$, two minima at $(\mu_x = \sigma_x,\, \mu_y = -\sigma_y)$ and $(\mu_x = -\sigma_x,\, \mu_y = \sigma_y)$, and two maxima at $(\mu_x = \sigma_x,\, \mu_y = \sigma_y)$ and $(\mu_x = -\sigma_x,\, \mu_y = -\sigma_y)$. Range for skewness is in $\left[-\frac{2}{\sqrt{3}}, \frac{2}{\sqrt{3}}\right]$.

Figure 6 represents values of skewness for several combinations of values of parameters, the central part of the graph corresponds to values of μ near zero for both distributions

The same effect appears in Figure 7. The product of two normal variables, where $\mu \in [-2, 2]$ and $\sigma \in [0.1, 5]$ was considered here. Upper figure represents values from μ_x and μ_y, down figure represents values from σ_x and σ_y. Skewness tends to zero when both means (μ_x, μ_y) are lower or when at least one of the variances is small.

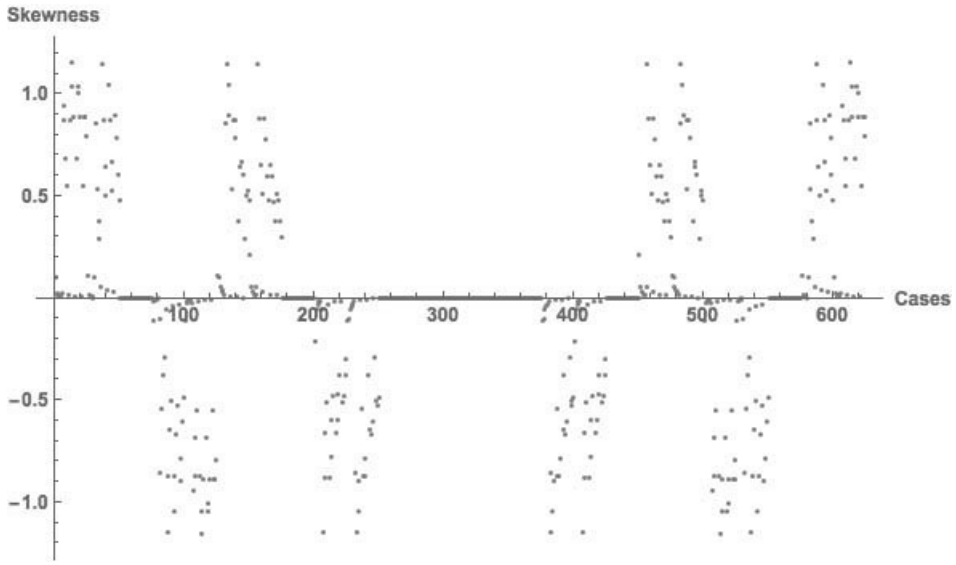

Figure 6: Skewness (Sk) for the product of two normal variables with $\mu \in \{-2, -1, 0, 1, 2\}$ and $\sigma \in \{0.1, 1.1, 2.1, 3.1, 4.1\}$.

In Figure 8, different values of parameters are considered and evolution of skewness is represented: we observe the existence of a tendency, when parameters are high values (> 1), the skewness tends to zero. Values for distribution represented in this figures are: $\sigma_x \in \{0.25, 1, 1.75, 2.5\}$, $\mu_x \in [0, 2]$, $\mu_y \in [0, 2]$, $\sigma_y \in \{0.25, 1.25, 2.25\}$. In columns is represented evolution of σ_x and σ_y in rows.

So far, the evolution of the skewness from a theoretical point of view according to the formula obtained through the moment generating function was analysed. In this last part, a number of real examples calculated using the Monte-Carlo method to simulate two uncorrelated normal distributions was considered. A total of 1,000,000 points were generated for each and their respective products were obtained. With the data thus generated, were calculated: the mean, variance and skewness of the product distribution. The results are shown in Figure 9.

The distributions used in the simulation had the following characteristics: $X \sim (\mu_x, \sigma_x)$ with values for $\mu_x = \{0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 2.25, 2.5\}$ and $\sigma_x = \{0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 2.25\}$ and $Y \sim (\mu_y, \sigma_y)$ for the same values for the parameters that variable X. In the Figure 9, evolution of the graph represent the evolution of the parameters, at first cases we have small values for the four parameters and then, they grow following the sequence: σ_y, μ_y, σ_x and μ_x.

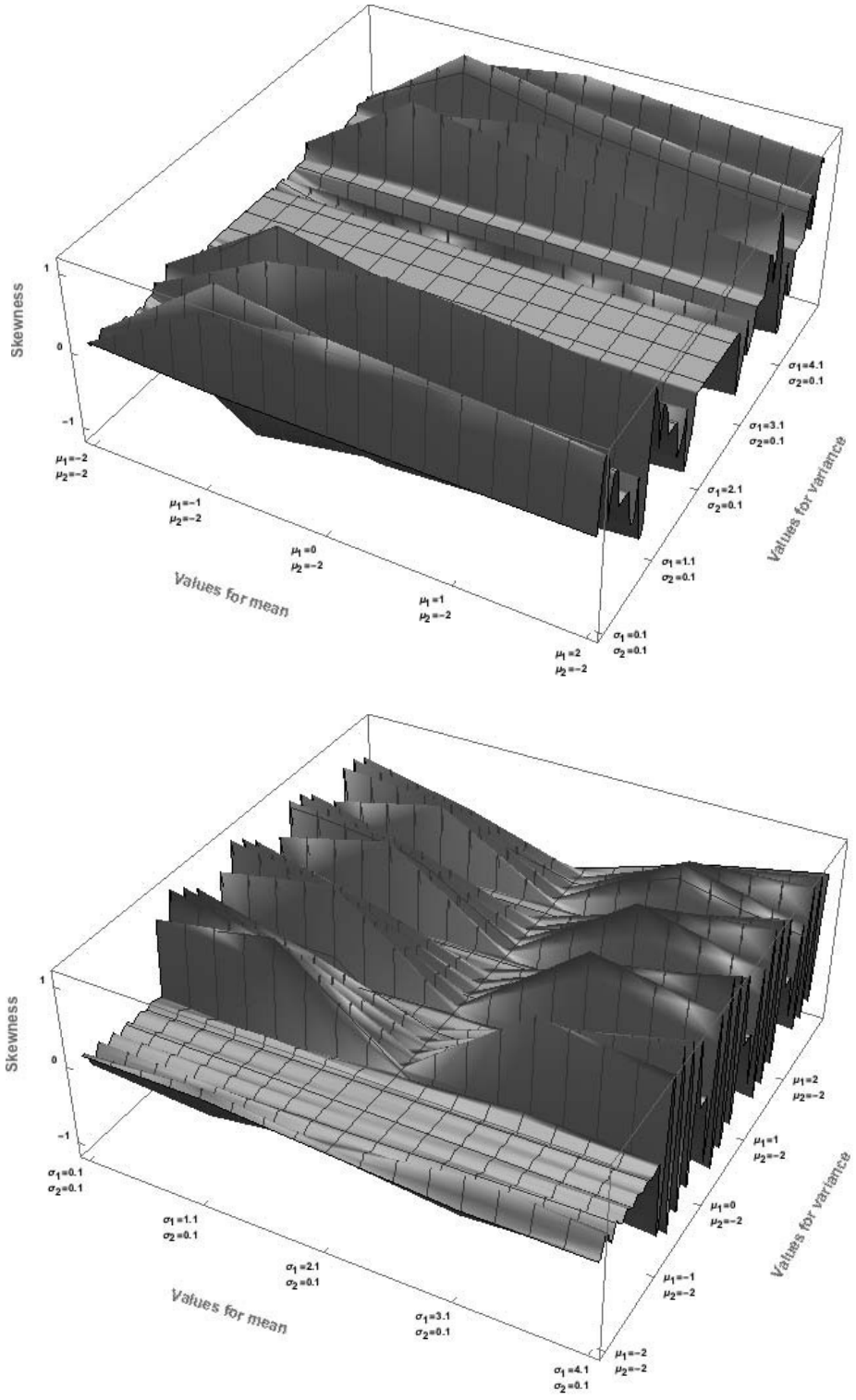

Figure 7: Skewness product of two normal variables.

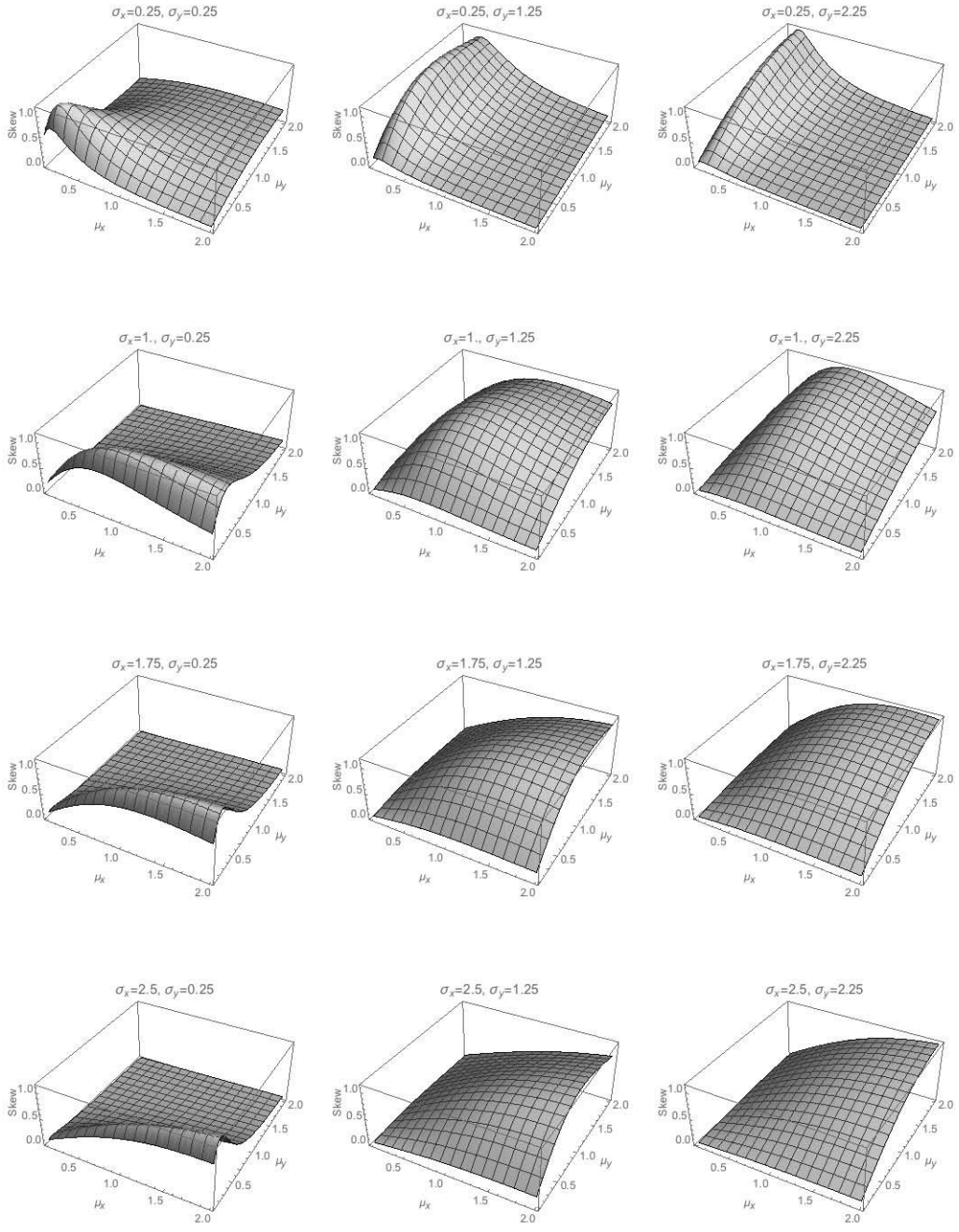

Figure 8: Evolution of skewness for the product of two normal variables.

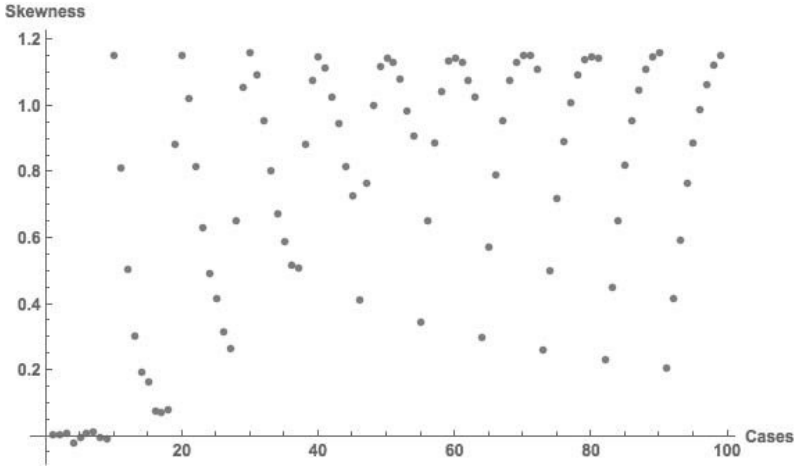

Figure 9: Evolution of skewness for the product of two normal variables — Monte-Carlo simulation.

Skewness distribution is slightly different from the expected distribution, and there are only few products where the value is zero or less than 0.1. It appears to be a tendency that the value of skewness is small when all parameters of the distributions are small and when there exists some high values for some parameter. However, when all the parameters are big, skewness grow up.

4. SKEWNESS. CONSEQUENCES ON RISK

As it is referred in Mumpower and McClelland ([9]) skewed distributions of risk estimates amplify the "winner's curse" so that the estimated risk premium for low-probability events is likely to be lower than the normative value. An application of this, is the Probabilistic risk analysis (PRA), where the probability of an event is based on frequency data, physical measurement or expert judgement. In such situation the existence of some type of error is obvious. A simple decision analysis situation could be a lottery of the form pay-off V' with probability p', otherwise 0 with probability $(1 - p')$. Assume that the decision-maker is an unbiased, valid estimator of p and V; however, those estimates are not perfectly reliable, there is an associate noise or random error. Then we define:

$$p = p' + e_p \ , \qquad V = V' + e_V \ , \tag{4.1}$$

where e_V represents the error of the variable.

So thus, p and V are the observed values for decision-maker, if the error term is symmetrically distributed so that, both variables could follow a normal distribution with mean V' and variance S_V^2 and p' and S_p^2, respectively.

At this situation, the decision-maker has two branches into the tree: one is the lottery with expected value: $EV = pV$ and the other is a constant value C. Let $C = EV$.

We consider an example Mumpower and McClelland ([9]) involving a decision point with two options. The first consists of a simple lottery of the payoff $V \sim N(4,1)$ with probability $p \sim N(0.4, 0.01)$; the second one consists of a certainty $C = 1.6$. We define $EV = p * V$, as a product of two uncorrelated normal distributions. Applying the formulas for the moments we have:

1. Mean: $E[EV] = 0.44 = 1.6 = C$;
2. Variance: $V(EV) = 0.33$;
3. Skewness: $\alpha_3(EV) = 0.506$.

Then, the distribution of the product is positively skewed and as a consequence, even though the expected value of EV and the value of C are the same, it is more likely that any randomly drawn single estimate, EV will be lower in value than C. At this situation, a risk-neutral decision-making who bases the choice by comparing both values, will select C with a probability greater than 0.5 since $p(C > EV) > 0.5$. In our example, this value is exactly $p(C > EV) = 0.5339$. Value of the median of the EV distribution is $1.55122 < 1.6$, lower that its expectation value.

Skewness is determined by the inverse of the coefficient of variation of the two variables V and p. When the inverse of the coefficient de variation is reduced then skewness will be high, and increasing the inverse of the coefficient of variation a lower value for skewness is obtained. Then, decreasing the level of measurement error this will not necessarily reduce the level of skewness.

From (3.3), value of skewness as a function of the inverse of the coefficient of variation is

$$\alpha_3(\rho_p, \rho_V) = \frac{6\,\rho_p\,\rho_V}{\left(\rho_p^2 + \rho_V^2 + 1\right)^{3/2}}\,. \tag{4.2}$$

For our example, value of $\rho_V = \frac{4}{1} = 4$ that is equal to $\rho_p = \frac{0.4}{0.1} = 4$. Considering $\rho_V = 4$ as constant, varying ρ_p and using first derivative of (4.2) we obtain:

$$\frac{\partial \alpha_3}{\partial \rho_p} = \frac{24.}{\left(\rho_p^2 + 17.\right)^{3/2}} - \frac{72.\,\rho_p^2}{\left(\rho_p^2 + 17.\right)^{5/2}}\,. \tag{4.3}$$

Skewness (4.2) has two maxima at $\rho_p = 2.91548$ and $\rho_p = -2.91548$, then is an increasing function in $(0, 2.91548)$ and then it decreases. Thus if the ρ_p increases, skewness could be increased.

For ρ_V conclusions are the same for the symmetry of the expression (3.3) with respect to both values of the inverse of the coefficient of variation.

These results are in the following proposition.

Proposition 4.1. *Let $V \sim N(\mu_V, \sigma_V)$ and $p \sim N(\mu_p, \sigma_p)$ be two normal uncorrelated distributions. We consider the variable product $EV = pV$, the skewness of EV is a increasing function for ρ_p and for ρ_V, where $\rho_p = \frac{\mu_p}{\sigma_p}$ and $\rho_V = \frac{\mu_V}{\sigma_V}$ are the inverse of the coefficient of variation.*

Skewness in the product of normal distributions has several implications on the risk analysis:

- The mean value is not the mode of the distribution.
- Risk-neutral decision-making choice using the median value of the product distribution and then, the probability of choice C as greater than 0.5.
- Decreasing the level of measurement error will not necessarily reduce the level of skewness.
- For the product of two uncorrelated normal variables there exists a maximum value for skewness $\left(\left|\frac{2}{\sqrt{3}}\right|\right)$.

5. CONCLUSIONS

The product of two normally distributed variables is an open question. This product could be considered a normally distributed variable in specific circumstances. Craig ([3]) and other authors later calculate the moment-generating function for the product and consider that the product is normally distributed when the inverse of coefficient of variation is high-valued.

In this paper we calculate the moment-generating function and we have used it in order to estimate the skewness of product distribution. When the value of the mean of at least one variable is high (> 1) then skewness is low, and when the value of mean is low, skewness is high. This relation lead to several questions that we have shown.

When the variables are correlated skewness is very common and normality of the product is not hold. When the variables are uncorrelated, the existence of skewness is a function of mean and variance of the two distributions. When the inverse of the coefficient of variation is high, skewness is low, but only when the variance is not low; the existence of a variable with a low value for variance produces the existence of skewness into the product of variables.

Skewed distributions of joint probability estimates and expected value estimates can affect the risk and consequently the choices of decision-makers. We have shown the effects in a particular probability risk analysis model.

ACKNOWLEDGMENTS

This research was partially sponsored by national funds through the Fundação Nacional para a Ciência e Tecnologia, Portugal – FCT under the project UID/MAT/00006/2013 and by the sabbaticals scholarships SFRH/BSAB/113669/2015 and SFRH/BSAB/113590/2015. We also acknowledge the valuable suggestions from the referees.

REFERENCES

[1] Aroian, L.A. (1947). The probability function of a product of two normal distributed variables, *Annals of Mathematical Statistics*, **18**, 256–271.

[2] Aroian, L.A.; Taneja, V.S. and Cornwell, L.W. (1978). Mathematical forms of the distribution of the product of two normal variables, *Communication in Statistics – Theory and Methods*, **7**(2), 164–172.

[3] Craig, Cecil C. (1936). On the frequency function of xy, *Annals of Mathematical Society*, **7**, 1–15.

[4] Cochrane, J.H. (2006). The Dog that Did not Bark: A Defense of Return Predictability. Working Paper, University of Chicago.

[5] Deb, P. and Sefton, M. (1996). The distribution of a Lagrange multiplier test of normality, *Economic Letters*, **51**, 123–130.

[6] Harvey, C.R. and Siddique, A. (2000). Conditional skewness in asset pricing tests, *Journal of Finance*, **55**, 1263–1295.

[7] Hayya, Jack C. and Ferrara, William L. (1972). On normal approximations of the frequency functions of standard forms where the main variables are normally distributed, *Management Science*, **13**(2), 173–186.

[8] Kraus, A. and Litzenberg, R.H. (1975). Measurement error, skewness and risk analysis: coping with the long tail of the distribution, *Journal of Finance*, **31**, 1085–1099.

[9] Mumpower, Jerryl L. and McClelland, Gary (2002). Skewness preference and the valuation of risk assets, *Risk Analysis*, **22**(2), 277–290.

[10] R Core Team (2015). A language and environment for statistical computing, *R Foundation for Statistical Computing*, Vienna, Austria.

REVSTAT – Statistical Journal
Volume 14, Number 2, April 2016, 139–155

FRACTAL BASED CANCER MODELLING

Authors: Milan Stehlík
– Institute of Statistics,
University of Valparaíso, Chile
Department of Applied Statistics,
Johannes Kepler University Linz, Austria
mlnstehlik@gmail.com

Philipp Hermann
– Department of Applied Statistics,
Johannes Kepler University Linz, Austria
philipp.hermann@jku.at

Orietta Nicolis
– Institute of Statistics,
University of Valparaiso, Chile
orietta.nicolis@uv.cl

Received: October 2015 Revised: February 2016 Accepted: February 2016

Abstract:

- Fractal hypothesis is both challenging and technical issue of mammary cancer. We conduct a simple discrimination on the basis of box-counting dimension. Moreover, we discuss on statistical distributions of fractal dimensions for both mammary cancer and mastopathy. Thereby, we detect significant differences in the underlying distribution between the two groups. A multifractal analysis on the basis of a wavelet based approach has been conducted. Discussion on alternative cancer therapy and cancer prevention is provided.

Key-Words:

- *mammary cancer; multifractal analysis; box-counting dimension; distribution fit; discrimination; alternative cancer therapy.*

AMS Subject Classification:

- 62F03.

1. INTRODUCTION

When we consider fractal based cancer diagnostic, many times a statistical procedure to assess the fractal dimension is needed. We shall look for some analytical tools to discriminate between cancer and healthy ranges of fractal dimensions of tissues (see [3, 19]). Fractal dimension may also help for early diagnosis of breast cancer, which is the key for breast cancer survival. Breast cancer, hereafter described as mammary cancer, is the most common cancer in women. The algebraic and topologic properties of cancer growth are available via appropriate set structure, e.g. bornology (see [20, 21]) or topology (see [28]). Here we illustrate some issues on discrimination between mammary cancer (mamca) and mastopathic (masto) tissues, which is follow-up of study of [13]. The data contains 391 histological images of mammary ($n = 192$) and mastopathic ($n = 199$) tissue, which were used to compute the box-counting dimension by means of ImageJ software [1]. We refer to [12] or [13] for more details how the fractal dimension was obtained. A modelling procedure for mammary cancer and mastopathy on the basis of randomized fractals has been introduced in [12], showing that this flexible model can reconstruct the development of the tissue of both, cancer and mastopathy. This approach allows to measure the fractal dimension with the aid of box-counting dimension, in order to observe the development of the tissue over time as well as to discriminate between these two groups.

Mammogram or sonogram examinations have been used as a first step in cases of breast cancer suspicion. Since biopsy, which is an invasive surgical operation imposing psychological and physiological stress for patients, has to be used to confirm the disease to date, other diagnostic tools with accurate diagnostic rates are of interest to be developed. Recently, computer aided diagnosis systems (CAD) are frequently investigated by researchers, see [4] among others, however, we discriminate between mastopathy and cancer on the basis of statistical differences (e.g. in terms of underlying distributions) in the fractal dimension of the two groups.

2. SIMPLE DISCRIMINATION BETWEEN MASTOPATHY AND MAMMARY CANCER BASED ON THE BOX-COUNTING DIMENSION

We consider boxplots in Figure 1 in order to have a graphical comparison between the two groups. Therein, the box-counting dimensions seem to be on average lower for mammary cancer tissue in addition that some candidates for outliers are apparent in the lower boundaries.

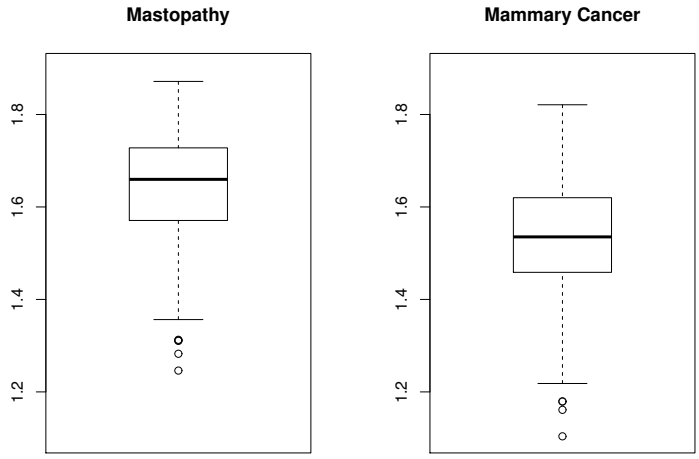

Figure 1: Boxplot of the groups mastopathy (left) and mammary cancer (right).

If we will follow the simple concept that higher dimension is more risky, the issue is that we will arrive with this dataset to some sort of contradiction. When we make a simple clustering based on ordering the box-counting dimension and decide to tell that more risky tissue has a box-counting dimension bigger than the median (1.5972) and non-risky tissue is below, then we only classified 135 of mamca and 60 of masto below. Recall that 199 observations contain the characteristic mastopathy and 192 observations mammary cancer. Even using the arithmetic mean of 1.587391 decreases the number of classified tissues to 128 for mamca and 56 for masto. Based on this simple example we can conclude that we need a more sophisticated procedure based on the box-counting dimension to discriminate between the two groups and we should take more detailed characteristics of the tissue into account. In extremal case there is no possibility to develop automatic clustering based on box-counting dimension, which could avoid histological expert examination.

Figure 2 indicates that using the only single box-counting dimension establishes inverse problems, which are ill posed. Loosely saying we need a continuous dimension spectrum, e.g. multi-fractal dimension spectra. It has already been used in breast cancer discrimination by [6, 10, 22]. A multifractal system is a generalization of a fractal system in which a single exponent (the fractal dimension) is not enough to describe its dynamics; instead, a continuous spectrum of exponents (the so-called singularity spectrum) is needed. This also relates to Tweedie exponential dispersion models, which, as a special case, contain both normal and gamma distributions. This is further justification for these two simple distributional families: in the case of our empirical data we have found a strong deviation from normality for mastopathy, and therefore we used gamma $G(\alpha, \beta)$ and Weibull $W(k, \lambda)$ distribution. In contrast to that mammary cancer data is also tested for normal distribution in the following.

The distribution of the ordered observations of the different groups is highlighted in Figure 2. Apparently, mammary cancer tissue have on average lower dimensions (dashed line) compared to mastopathological tissue (solid line). These conclusions were already recognizable due to the comparison of the mean as well as the interquartile-distance within the boxplots.

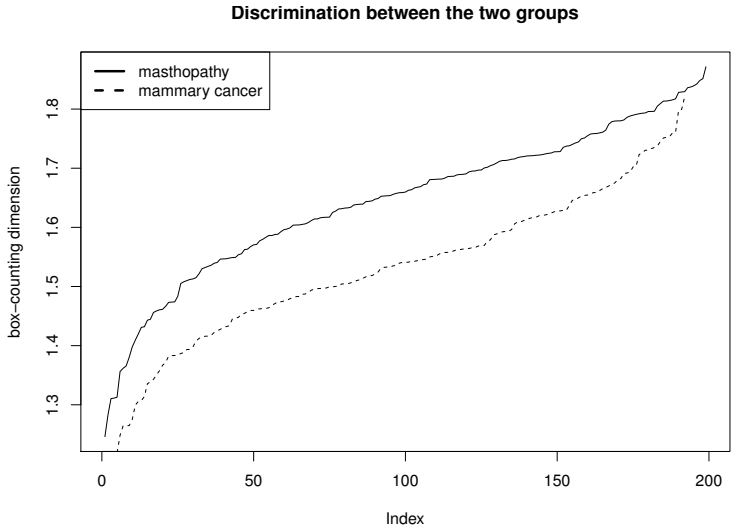

Figure 2: Plot of the dimensions discriminated between the groups mastopathy and mammary cancer.

3. TESTING FOR DISTRIBUTIONS OF THE GROUPS

Separating the data and testing for distributional fit of the groups may lead to further information on group discrimination. Therefore, maximum likelihood procedures have been conducted in order to estimate fitting parameters for gamma and Weibull distribution for fractal dimension of the mastopathy. Estimation for gamma distribution yields a shape equal to 162.58 and scale equal to 98.85, which results in a p-value of 0.15 by usage of Kolmogorov–Smirnov-test (KS-test). For Weibull distribution the two distribution forming parameters were estimated as 16.20 and 1.70. Testing with those parameters gives a p-value of 0.96. Fitting the distributions with estimated parameters in addition to the histogram is plotted on the left part of Figure 3. A better fit of the mammary cancer data with normal distribution has been seen in previous calculations. ML-estimations are computed in order to continue the testing procedure with gamma ($\hat{\alpha} = 129.50$ and $\hat{\beta} = 84.76$ results in $p = 0.43$) and Weibull ($\hat{k} = 13.23$ and $\hat{\lambda} = 1.59$ gives $p = 0.20$) distributions. In addition to that mean (1.53) and variance (0.017) are computed to fit normal distribution ($p = 0.66$). These p-values show that

gamma, Weibull, and normal distribution may not be rejected to fit mammary cancer box-counting dimensions. The right plot of Figure 3 shows the fit for the fractal dimension of mammary cancer data with the parameter estimates given above. Therein, gamma distribution is presented as solid line, Weibull distribution as dotted line and normal distribution is visualized with a dash-dotted line.

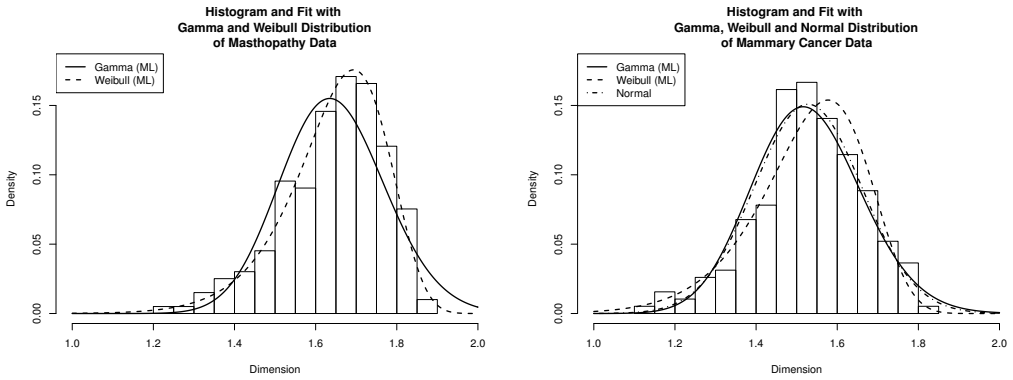

Figure 3: Fit of mastopathy (left) and mammary cancer (right) groups separately.

Table 1 provides the shape and scale parameters for both gamma and Weibull for the complete data as well as both groups separately. Note that parameters for normal distribution are not provided due to lack of comparability, since only for mammary cancer data this distribution was not rejected.

Table 1: Computation of p-values for gamma and Weibull distribution with shape and scale parameter for both, complete data and separated by groups mastopathy and mammary cancer.

Distribution	**Group**	**shape**	**scale**	p**-value**
Gamma	All	120.84	76.12	0.22
	Masto	162.58	98.85	0.15
	Mamca	129.50	84.76	0.43
Weibull	All	13.25	1.65	0.39
	Masto	16.20	1.7	0.96
	Mamca	13.23	1.59	0.20

Note that for the total data we received estimator of shape estimator 120.84 and a scale estimator of 76.12 for gamma distribution. KS-test of this set of parameters results in 0.22. Moreover, ML estimation for Weibull distribution gives the estimators 13.25 and 1.65 with a corresponding p-value of 0.39. Therefore,

one can see that discriminating between the groups yields differences in terms of underlying distribution parameters. Adjusting the data for outliers results in negligible differences in the estimates.

In the following we will apply that the sum of squared independent standard normal distributed random variables follows Chi-Squared distribution. We will assume:

- The differences between the curves are standard normal distributed. Therefore, Shapiro–Wilk-test can be used. On the given dataset it results in a rejection of the null-hypothesis of standard normal distribution. Hence, there is another possibility to justify the condition in order that usage of Chi-Squared distribution is allowed.
- The squared differences are Chi-Square distributed with one degree of freedom.

Computing the sum of the squared differences delivers a value equal to 5.38. A Chi-Square-test was accomplished to test whether we can distinguish between two groups within the data. The p-value of the distribution function of the Chi-Squared distribution with 199 degrees of freedom is approximately one. This p-value is another proof that the two groups are different. Furthermore, we made a standardization (by subtracting the mean and dividing by the standard error) of the previously calculated differences. The distribution function at the sum of standardized squared differences of 198 and 199 degrees of freedom is 0.49331. Hence, this p-value does not yield enough support to reject the null hypothesis of differences between the groups. However, the property of the data (only positive values) as well as high flexibility of gamma distribution leads us to hypothesis for gamma distribution. We simulated in order to maximize p-values with changes in shape and scale parameters of gamma distribution. By reducing the shape parameter and in contrast to that increasing the rate (reducing the scale parameter), Table 2 shows a convergence to higher p-values.

Table 2: Simulation of p-values with given shape and scale parameter.

shape	**scale**	p**-value**
0.45491680	$\frac{1}{0.45729929}$	$8.354 \cdot 10^{-5}$
0.48	$\frac{1}{0.4573}$	$3.7 \cdot 10^{-12}$
0.44	$\frac{1}{0.4673}$	$1.65 \cdot 10^{-11}$
0.425	$\frac{1}{0.48}$	$1 \cdot 10^{-9}$
0.42	$\frac{1}{0.48}$	0.0049
0.425	$\frac{1}{0.48}$	0.0449
0.43	$\frac{1}{0.48}$	0.0097
0.42	$\frac{1}{0.485}$	0.1272
0.415	$\frac{1}{0.485}$	0.0996

We can see with the aid of KS-test that we will find a rather good fit for specific values of the parameters. Thereby, the shape parameter of 0.415 and a scale parameter of $1/0.485$ delivered an accurate p-value of 0.0996. The test with a shape parameter of 0.42 delivered an even better p-value of 0.1272. Therefore, it can be assumed that the standardized differences are gamma distributed with a shape parameter lying in between the range [0.415, 0.42] and the scale parameter close to 2.06 ($\frac{1}{0.485}$). Therefore, we compare the standardized differences with generated random variables of a gamma distribution, with a shape parameter of 0.415 and a scale parameter of 2.062 in Figure 4.

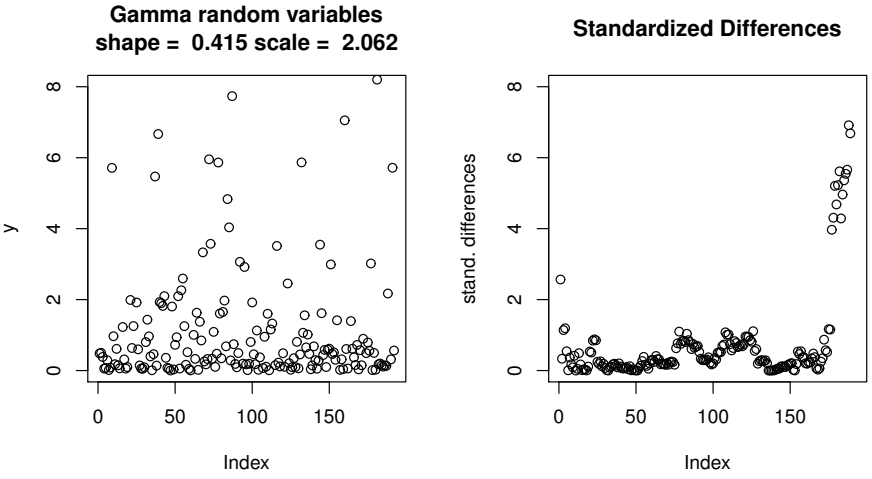

Figure 4: Comparison of standardized differences with random variables of a gamma distribution.

Shapiro–Wilk tests deliver a p-value for mammary cancer tissue of 0.0452 and a value smaller than 0.001 was obtained for mastopathic tissue. Hence, for a significance level of 95% both p-values are too small to state that the box-counting dimension of mammary cancer tissue or mathopathic tissue is normal distributed. QQ-Plots in Figure 5 are another indication, that mastopathy is not normal distributed, but normal distribution of the dimensions of mammary cancer should not be rejected without further analysis. Indeed, the lower quantiles differ significantly from the comparative line in the range of -3 to about -1.5 of the theoretical quantiles. Therefore, outlier detection for mammary cancer data with usage of box-plot rule has been performed. These computations are performed with $q_{0.25} - 1.5 \cdot IQR$ and $q_{0.75} + 1.5 \cdot IQR$, where IQR is the interquantile range as $q_{0.75} - q_{0.25}$ and q_α is the α-quantile. Four candidates for outliers from the lower end of the data were obtained and removed in order to yield useful information on the distributional behavior of mammary cancer tissue. The according p-value has significantly increased up to 0.5716 and therefore, it can be assumed that the modified data is normal distributed. Another indication for normality of

this group are histogram and QQ-Plot of the modified data in the second row of Figure 5. Both of the plots suggest that the modified box counting dimension of mammary cancer tissue is normal distributed.

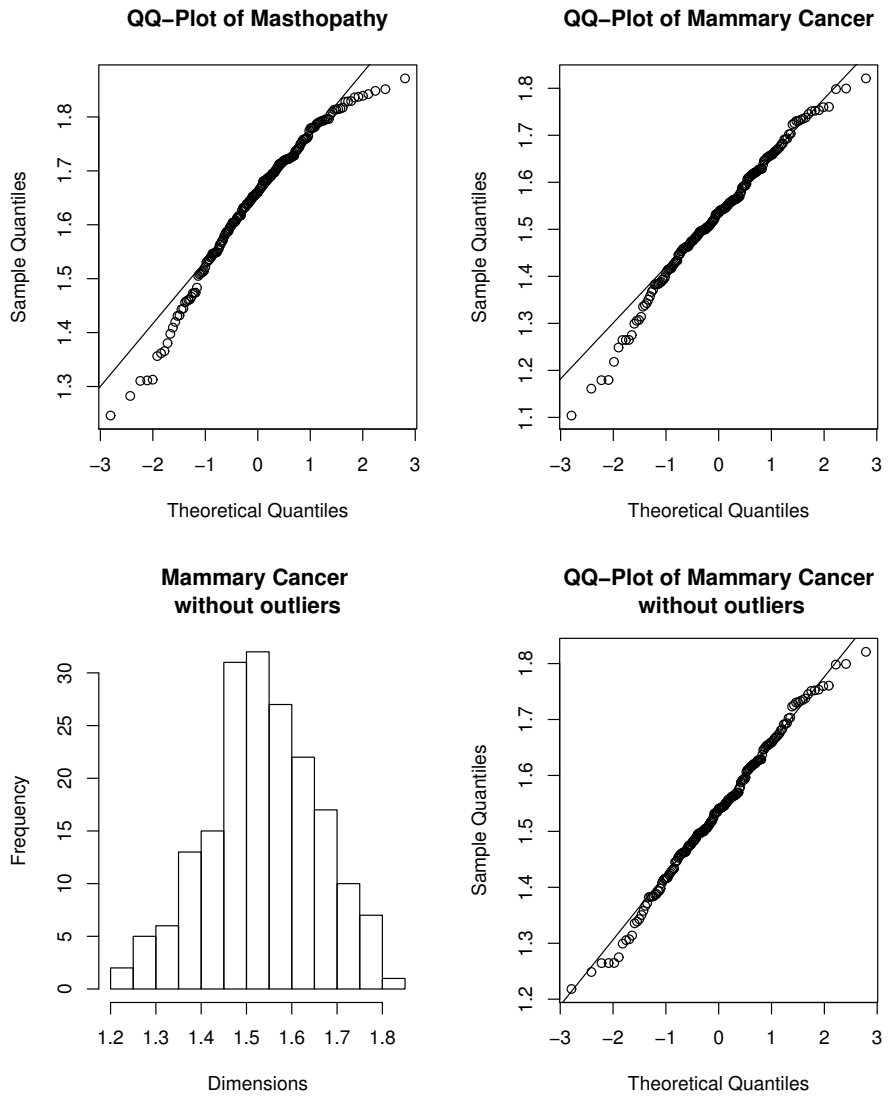

Figure 5: Top row: QQ-Plot of the groups mastopathy (left) and mammary cancer (right). Bottom row: Histogram and QQ-Plot of mammary cancer data without outliers ($n = 188$).

Robust normality testing procedures have been applied to both groups. Therefore, data has been truncated in the lower boundaries, such that only tissue higher than threshold ε has been taken into account. Shapiro–Wilk tests have been used to compute p-values for the fit of normal distribution. The development of p-values can be found in Figure 6, where especially modified box-counting

dimensions of mammary cancer tissue can be seen as normally distributed in contrast to mastopathic tissue box-counting dimensions. This test approach unfolds the different behaviour of the box-counting dimension with respect to normality.

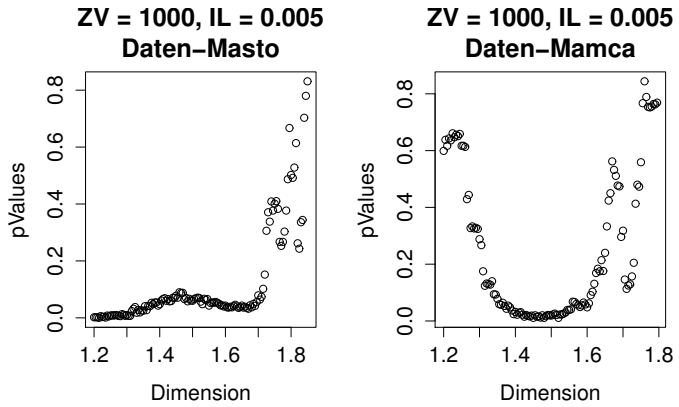

Figure 6: Test for normal distribution with truncated data for both groups.

Therefore, truncation of the data from the lower boundaries reveals mammary cancer box-counting dimension is more robust with respect to normality than mastopathic tissue.

4. MULTIFRACTAL ANALYSIS OF MAMMOGRAPHY: A WAVELET BASED APPROACH

Multifractal analysis is concerned with describing the local singular behavior of measures or functions in a Geometrical and Statistical fashion. It was first introduced by Mandelbrot in the context of turbulence (see [17, 18]) even if the term "multifractal", was successively proposed by [9].

Multifractal structures have been found in various contexts. Most prominently in studies of turbulence, stock market exchange rates, geophysics and recently also in traffic, introducing fruitful and novel aspects to the mentioned fields. The basic concept of multifractal analysis is to assess fractal dimensions of self-similar structures with varying regularities and to produce the distribution of indices of regularity, which constitutes the multifractal spectrum (MFS). The multifractal formalism relates the MFS to the partition function measuring high-order dependencies in the data. In the following we will describe the wavelet-based multifractal spectrum (WMFS) proposed by [11, 23, 24] and we will apply it to a sample of mammographic images. The advantages of using the wavelet-based MFS are availability of fast algorithms for wavelet transform, the

locality of wavelet representations in both time and scale, and intrinsic dyadic self-similarity of basis functions. The multifractal formalism is based on the concepts of the partition function which can be defined in terms of wavelet coefficients as

$$T(q) = \lim_{j \to -\infty} \log_2 E|d_{j,k}|^q \,, \tag{4.1}$$

where $d_{j,k}$ is the wavelet coefficient at level j and location k, and q is the order of moments. We emphasize that q is a real number within a certain range covering the negative numbers as well [11]. Even though (4.1) is very informative, the singularity measure is not explicit. It was proposed in [11] that the local singularity strength could be measured in terms of wavelet coefficients as:

$$\alpha(t) = \lim_{k2^j \to t} \frac{1}{j} \log_2 |d_{j,k}| \,, \tag{4.2}$$

where $d_{j,k}$ is the normalized wavelet coefficient at scale j and location k. The local singularity strength measure (4.2) converges to the local Hölder index of the process at time t. Small values of $\alpha(t)$ reflect more irregular behavior at time t. Any inhomogeneous process has a collection of local singularity strength measures and their distribution $f(\alpha)$ forms the MFS. A useful tool to estimate the MFS is through the Legendre as follows

$$f_L(\alpha) = \inf_q \big\{ q\,\alpha - T(q) \big\} \,. \tag{4.3}$$

It can be shown that $f_L(\alpha)$ converges to the true MFS by using the theory of large deviations [8]. If we rearrange (4.1), it becomes

$$E|d_{j,k}|^q \sim 2^{jT(q)} \qquad \text{as} \quad j \to -\infty \,. \tag{4.4}$$

A standard linear regression can be used to estimate the partition function $T(q)$ since the values $E|d_{j,k}|^q$ could be easily obtained by the moment-matching method.

Let $\widehat{S_j(q)} = \frac{1}{2^j} \sum_{k=1}^{N2^{-j}} |d_{j,k}|^q$ be the empirical q^{th} moment of the wavelet coefficients (N is the length of the time series). By applying the Central Limit Theorem, $\widehat{S}_j(q) \to E|d_{j,k}|^q$ as $N \to \infty$. Then, using the scaling property of the wavelet coefficients given by $d_{j,k} = 2^{jH} d_{0,k}$, we have that $\widehat{S}_j(q)$ is asymptotically normal with mean $2^{jT(q)} E|d_{0,0}|^q$ and variance $\sigma^2_{j,q} = \frac{2^{2jT(q)} \operatorname{Var}|d_{0,0}|^q}{2^{-j}N}$ (see, [11]). Considering the logarithm transformation of $\widehat{S_j(q)}$ we can write

$$\log_2 \widehat{S_j(q)} = j\,T(q) + \varepsilon_j \,, \tag{4.5}$$

where the error term ε_j is introduced from the moment matching method when replacing the true moments with the empirical ones. Using approximation theorems (see, [26]) one can prove that the $\log_2 \widehat{S_j(q)}$ is asymptotical normal with

mean and variance described by [11]. The ordinary least square (OLS) estimator gives the estimation of the partition function,

$$\widehat{T}(q) := \sum_{j=j_1}^{j_2} a_j \log_2 \widehat{S}_j(q) \,, \tag{4.6}$$

where the regression weights a_j must verify the two conditions $\sum_j a_j = 0$ and $\sum_j j\, a_j = 1$ (see [2] and [5]). Thus, we can estimate $f(\alpha)$ through a local slope of $\widehat{T}(q)$ at values

$$\widehat{\alpha}(q_l) \,=\, \big[\widehat{T}(q_{l+1}) - \widehat{T}(q_l)\big] / q_0 \,, \qquad q_l = l\, q_0 \,,$$

as

$$\widehat{f}\big(\alpha(q_l)\big) \,=\, q_l\, \alpha(q_l) - \widehat{T}(q_l) \,.$$

Multifractal spectra can even be found for monofractal processes, where the spectra generated from such processes are ramp-like with a dominant (modal) irregularity corresponding to the theoretical Hurst exponent (see [23]). The MFS can be easily generalized to higher dimensions (see [6, 22]).

4.1. Multifractal descriptors

The multifractal spectrum can be approximately described by three canonical descriptors, which are:

(**1**) Spectral Mode (Hurst exponent, SM);
(**2**) left slope (LS) or left tangent (LT);
(**3**) width spread (Broadness, B) or right slope (LS) or right tangent (RT).

A typical multifractal spectrum can be quantitatively described as shown in Figure 7. In particular, SM represents the apex of spectrum or most common Hölder regularity index α found within the signal, and LS (or LT) represents the slope of the distribution produced by the collection of Hölder regularity index α with smaller values of the mode (SM). However, broadness (B) is a more intricate descriptor of the multifractal spectrum. Broadness (B) is believed to be more meaningful than right slope (RS) or right tangent (RT), because it is a compound measure representing the overall nature of the multifractal spectra, taking into account the overall variability among the Hölder regularity index α. In addition, broadness (B) partially accounts for right slope (RS) or right tangent (RT) in calculation, as the resulting value of B is based on the relative values of RS and LS. Both slopes (or both tangents) can be easily obtained using the interpolation technique, while it is not straightforward to define the broadness (B) automatically. There are many ways to define the broadness (B). In this work, we select the method proposed by [27]. The overall multifractal descriptors are also graphically presented in Figure 7.

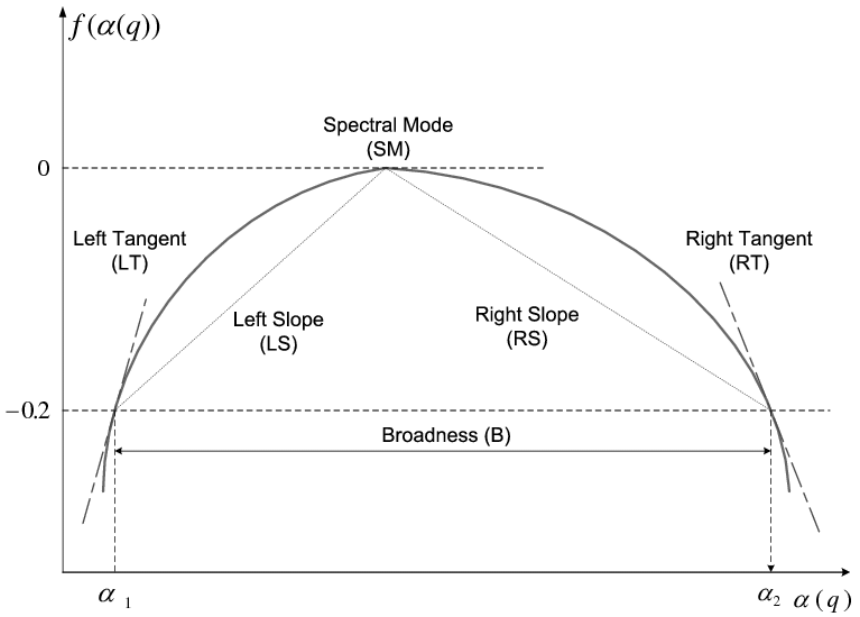

Figure 7: Illustration of geometric descriptors of multifractal spectra. Note that the horizontal axis represents values of Hölder regularity index $\alpha(q)$, while the vertical axis represents values proportional to the relative frequency of these indices, $f(\alpha(q))$.

4.2. Application to mammographic tissue images

In this section, we apply the wavelet-based multifractal spectra to two digital mammogram images (shown in Figure 8) of size 512×512 representing mastopathic and cancerous tissues. We refer to the paper of [12] for a detailed description of the images.

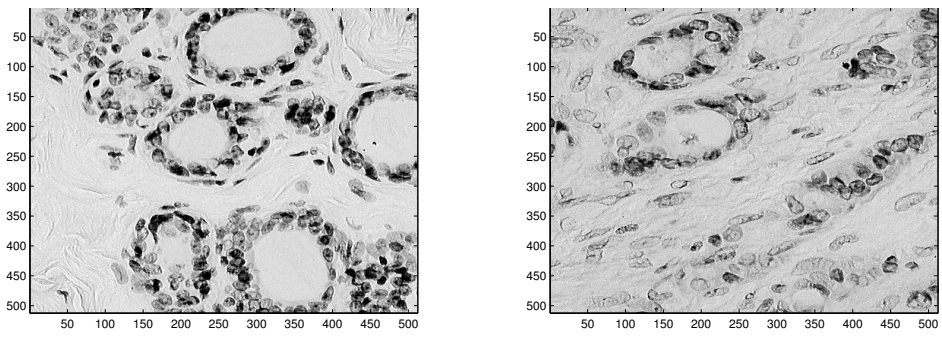

Figure 8: (a) Mastopathic tissue;
(b) Mammary cancer (invasive ductal mammary carcinoma).

First, we perform the 2D discrete complex wavelet transform for each image of size 512×512 by using complex Daubechies 6-tap filter (see [12, 15]), then we evaluate the wavelet multifractal spectra by extending (4.1) and (4.3) to 2D. Figure 9 compares the multifractal spectrum of the mastopathic tissue with the cancerous mammogram image. Although they seem to have a similar behavior it is evident that the Hurst exponents representing the local regularity are different for the two images.

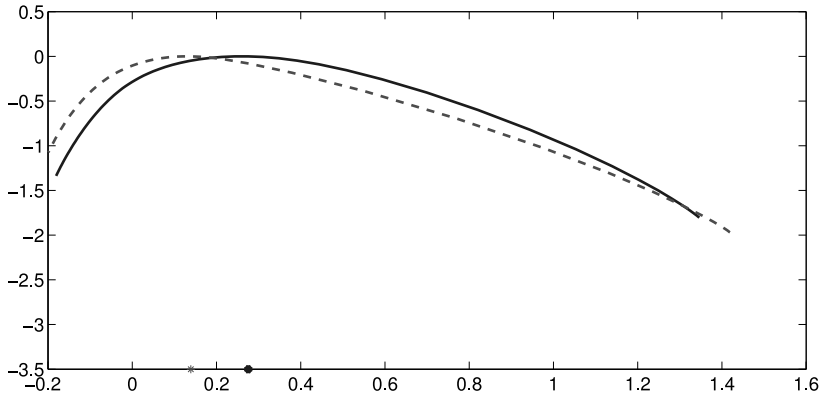

Figure 9: Wavelet multifractal spectrum for the mastopathic tissue (solid line) and cancerous tissue (dashed line). The filled dot and the asterisk on the horizontal axis represent the spectral mode for the mastopathic and cancerous tissue, respectively.

The different fractality is also confirmed by the calculation of the multifractal descriptors shown in Table 3. The mastopathic tissue seems to be more regular than the cancerous one (the regularity is represented by the SM or Hurst exponent) and the range (or broadness) of the local Hölder index is larger than for the cancerous tissue.

Table 3: Wavelet multifractal descriptors.

Tissue	**H**	**L1**	**L2**	**R1**	**R2**	**B**
masto	0.26	2.2	−1.2	0.89	−0.70	0.51
mamca	0.13	2.8	−1.2	1.15	−0.76	0.43

Hence, we conclude that the multifractal spectrum and its descriptors could be used in classification algorithms for discriminating between mastopathic and cancerous tissue. This could provide an automatic tool to support medical decisions.

5. DISCUSSION AND CONCLUSION

Due to its prevalence and mortality a cancer diagnosis is one of the main fears of the general public. Certainly due to modern diagnostic tools as well as improvements in therapy, cancer can be seen as a chronic disease, where some of the patients will be living for several years after diagnosis. Earlier studies have proven that a positive attitude will lead to a significant increase in life expectancy of cancer patients. The risk of suicide or a burn-out is rapidly increasing within the first weeks after cancer diagnosis due to the very stressful first period. Utilizing psycho-oncologic care gives assistance in these situations. However, this option is quite unknown to most patients such that only 1% uses this support [25]. All these facts support the necessity to find quick, semi or full-automated methods for tissue discrimination. The discrimination between the groups in terms of distributional fit allows the interpretation that more abnormal tissue follows normal distribution. Moreover, it has been shown that gamma as well as Weibull distributions are proper distributions for fitting mammary as well as mastopathic box-counting dimensions. Combining several instruments for cancer testing is of major importance, because e.g. deciding for mammary cancer or mastopathy just on the basis of box-counting dimension may lead to many miss-specifications. Medical staff can be supported in the decision process by these fractal measures, nevertheless, other supporting tools as shape analysis of the cancer (see [14] among others) or alternative cancer therapies in cases of high risks for cancer (see [25]) are desired to have the highest possible medical attendance for patients. Additionally the impact of environmental factors on developing cancer or as preventive strategy have to be taken into account.

Criticism on the use of screening mammography due to over-diagnosis led some researchers to show that one in three breast cancers identified by mammography would not cause symptoms in a patient's lifetime (see [16]). Therefore, alternative and accurate screening technologies must be developed. The functional and technical background of dynamic infrared (IR) imaging has the potential for early detection of breast cancer and treatment response evaluation if optimal diagnostic algorithms are developed. We have shown that the wavelet-based multifractal analysis of dynamic IR thermograms is able to discriminate between cancerous breasts with monofractal (cumulative) temperature temporal fluctuations characterized by a unique singularity exponent ($h = c_1$), and healthy breasts with multifractal temperature fluctuations requiring a wide range of singularity exponents as quantified by the intermittency coefficient $c_2 \gg 0$.

ACKNOWLEDGMENTS

M. Stehlík acknowledges the support of Fondecyt Proyecto Regular No. 1151441 and Proyecto Interno 2015, REGUL. MAT 12.15.33, Modelacion del crecimiento de tejidos con aplicaciones a la Investigacion del cancer. P. Hermann thanks to the support of ANR project Desire FWF I 833-N18. Last but not least, the authors are very grateful to the Editor and the Reviewers for their valuable comments.

REFERENCES

[1] Abràmoff, M.D.; Magalhães, P.J. and Ram, S.J. (2004). Image processing with ImageJ, *Biophotonics international*, **11**(7), 36–43.

[2] Abry, P.; Roux, S.; Vedel, B. and Wendt, H. (2010). *The contribution of wavelets in multifractal analysis*. In "Wavelet Methods in Mathematical Analysis and Engineering, Series in contemporary applied mathematics" (A. Damlamian and S. Jaffard, Eds.), 55–98.

[3] Baish, J.W. and Jain, R.K. (2000). Fractals and cancer, *Cancer Res.*, **60**, 3683–3688.

[4] Chen, D.R. *et al.* (2005). Classification of breast ultrasound images using fractal feature, *Clinical Imaging*, **29**(4), 235–245.

[5] Delbeke, L. and Abry, P. (2000). Stochastic integral representation and properties of the wavelet coefficients of linear fractional stable motion, *Stochastic Processes and their Applications*, **86**, 177–182.

[6] Derado, G.; Lee, K.; Nicolis, O.; Bowman, F.D.; Newell, M.; Ruggeri F. and Vidakovic, B. (2008). *Wavelet-based 3-D Multifractal Spectrum with Applications in Breast MRI Images*. In "Bioinformatics Research and Applications" (Mandoiu, Sunderraman and Zelikovsky, Eds.), Lecture Notes in Computer Science, Springer, 281–292.

[7] Doubeni, C.A. *et al.* (2012). Contribution of Behavioral Risk Factors and Obesity to Socioeconomic Differences in Colorectal Cancer Incidence, *JNCI – J. Natl. Cancer Inst.*, **104**(18), 1353–1362.

[8] Ellis, R.S. (1984). Large Deviations for a General Class of Random Vectors, *The Annals of Probability*, **12**(1), 1–12.

[9] Frisch, U. and Parisi, G. (1985). *On the singularity structure of fully developed turbulence*. In "Turbulence and Predictability in Geophysical Fluid Dynamics and Climate Dynamics" (M. Gil, R. Benzi and G. Parisi, Eds.), 84–88, Amsterdam, North-Holland, Elsevier.

[10] George, L.E. and Kamal, H.S. (2007). *Breast cancer diagnosis using multifractal dimension spectra*. In "2007 IEEE International Conference on Signal Processing and Communications (ICSPC 2007)".

[11] Gonçalves, P.; Riedi, H. and Baraniuk, R. (1998). *Simple statistical analysis of wavelet-based multifractal spectrum estimation*. In "Proceedings 32nd Asilomar Conference on Signals, Systems and Computers", Pacific Grove, CA, Nov. 1998.

[12] Hermann, P.; Mrkvička, T.; Mattfeldt, T.; Minárová, M.; Helisová, K.; Nicolis, O.; Wartner, F. and Stehlík, M. (2015). Fractal and stochastic geometry inference for breast cancer: a case study with random fractal models and Quermass-interaction process, *Statistics in Medicine*, **34**(18), 2636–2661. doi:10.1002/sim.6497

[13] Hermann, P.; Piza, S.; Ruderstorfer, S.; Spreitzer-Gröbner, S. and Stehlík, M. (2014). Fractal Case Study for Mammary Cancer, *Biometrie und Medizinische Informatik – Greifswalder Seminarberichte*, **23**, 149–166.

[14] Hermann, P. and Stehlík, M. (2015). *On some issues on statistical analysis for Wilms tumor*. In "Proceedings of the 30th International Workshop on Statistical Modelling" (H. Friedl and H. Wagner, Eds.), Volume 2, Springer, 115–118.

[15] Jeon, S.; Nicolis, O. and Vidakovic, B. (2014). Mammogram diagnostics via 2-D complex wavelet based self-similarity measures, *The São Paulo Journal of Mathematical Sciences*, **8**(2), 265–284.

[16] Jørgensen, K.J. and Gøtzsche, P.C. (2009). Overdiagnosis in publicly organised mammography screening programmes: systematic review of incidence trends, *BMJ Publishing Group Ltd*, **339**, b2587.

[17] Mandelbrot, B. (1969). Long-run linearity, locally Gaussian process, H-spectra, and infinite variances, *International Economic Review*, **10**, 82–111.

[18] Mandelbrot, B. (1972). A statistical methodology for non-periodic cycles: From the covariance to R/S analysis, *Annals of Economic and Social Measurement*, **1**, 259–290.

[19] Mrkvička, T. and Mattfeldt, T. (2011). Testing histological images of mammary tissues on compatibility with the boolean model of random sets, *Image Anal. Stereol.*, **30**, 11–18.

[20] Paseka, J.; Solovyov, S.A. and Stehlík, M. (2015). Lattice-valued bornological systems, *Fuzzy Sets and Systems*, **259**(15), 68–88.

[21] Paseka, J.; Solovyov, S.A. and Stehlík, M. (2014). On the category of lattice-valued bornological vector spaces, *Journal of Mathematical Analysis and Applications*, **419**, 138–155.

[22] Ramírez-Cobo, P. and Vidakovic, B. (2013). A 2D wavelet-based multiscale approach with applications to the analysis of digital mammograms, *Computational Statistics & Data Analysis*, **58**, 71–81.

[23] Riedi, R.H. (1999). Multifractal Processes, *Technical Report* 99-06.

[24] Riedi, R.H. (1998). *Multifractals and Wavelets: A potential tool in Geophysics*. In "Proceedings of the SEG Meeting", New Orleans, Louisiana/USA.

[25] Runow, K.D. (2013). *Krebs – Eine Umweltkrankheit? Risiko minimieren, Therapie optimieren*, Südwest Verlag, München.

[26] Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*, Wiley.

[27] Shi, B.; Moloney, K.P.; Pan, Y.; Leonard, V.K.; Vidakovic, B.; Jacko, J.A. and Sainfort, F. (2006). Wavelet classification of high frequency pupillary responses, *Journal of Statistical Computation and Simulation*, **76**(5), 431–445.

[28] Stehlík, M. (2016). On convergence of topological aggregation functions, *Fuzzy Sets and Systems*, **287**, 48–56.

REVSTAT – Statistical Journal
Volume 14, Number 2, April 2016, 157–169

RISK ANALYSIS AND RETROSPECTIVE UNBALANCED DATA

Authors: Francesca Pierri
– Department of Economics,
Statistical Section University of Perugia, Italy
francesca.pierri@unipg.it

Elena Stanghellini
– Department of Economics,
Statistical Section University of Perugia, Italy
elena.stanghellini@unipg.it

Nicoló Bistoni
– Graduate of Department of Economics,
University of Perugia, Italy
nicobistoni@yahoo.it

Received: October 2015 Revised: February 2016 Accepted: February 2016

Abstract:

- This paper considers three different techniques applicable in the context of credit scoring when the event under study is rare and therefore we have to cope with unbalanced data. Logistic regression for matched case-control studies, logistic regression for a random balanced data sample and logistic regression for a sample balanced by ROSE (Random OverSampling Examples, Lunardon, Menardi and Torelli, 2014) are tested. We applied the methods to real data: balance sheets indicators of small and medium-sized enterprises and their legal status are considered. The event of interest is the opening of insolvency proceedings of bankruptcy.

Key-Words:

- *bankruptcy; case-control studies; data augmentation; logistic regression; ROSE method; unbalanced data.*

AMS Subject Classification:

- 62J05, 62M20, 62P20, 91G40.

1. INTRODUCTION

In recent years, mainly because of the economic crisis that involves several European countries, the measurement of credit risk plays an important role; it simply concerns classifying out-of-sample units into two categories, bad and good, but it is crucial for its implications. The different classes of risk such as Probability of Default, Loss Given Default, Exposure at Default, Expected or Unexpected Loss are subjects of special attention from financial institutions which are making more and more frequently use of quantitative tools in decision-making. Credit quality is in fact crucial to the profitability and stability of banking systems.

An approach to estimate the probability of default is represented by statistical models, known as *credit scoring* techniques, and logistic regression is widely used in this context (Stanghellini, 2009). Frequently we have data where one of the two events is rare, so even in the case of all categorical explanatory variables, contingency tables will have very low or zero frequencies in the cells related to this event. Things get more extreme when there is at least one continuous variable in the set of the explanatory variables. In this situation, estimation of the logistic regression model may lead to high classification errors of rare units (King and Zeng, 2001). The aim of this paper is to compare different techniques which allow accurate estimation under these conditions.

The study is carried out on data selected from the *AIDA* database, concerning balance sheet indicators of companies in the Tuscany region of Italy which contains a large number of small and medium-sized enterprises. The event of interest is the opening of insolvency proceedings for bankruptcy which, luckily from an economic point of view, can be considered rare.

In order to face this problem we applied logistic regression to a retrospective data collection, using different sampling techniques: case-control sampling, balanced random sampling and random oversampling (ROSE method). From the full dataset we built a training and a hold-out sample: the first one forms the basis of data for the implementation of the different methodologies, and the second is used to compare the three classification methods on the basis of the Receiver Operating Characteristic (ROC) Curve (Fawcett, 2006). The theoretical illustration of the three methodologies (Section 2, 3 and 4) is followed by a brief description of the data (Section 5) and their application. The three models are then compared, based on the area underlying the ROC Curve.

2. THE LOGISTIC MODEL WITH BALANCED DATA

A prospective study often involves a long follow-up period and a large sample and therefore many investigations relay on a retrospective technique. The default status is regarded as a fixed variable, while variables specifying risk factors are viewed as random conditional on the default status. A retrospective study draws separate samples of cases (the bankruptcy event occurred) and controls (good firms) and therefore a smaller total sample size is usually required in comparison to a prospective study. Mantel and Haenszel (1959) and Mantel (1973) provide discussions of retrospective studies and their relationship with prospective ones. The logistic model is widely used in the analysis of retrospective studies, but it is necessary to ensure that the retrospective sample includes a representative sample of cases and controls from the population.

Let Y be the Bernoulli random variable taking value 1 when the event occurs (bankruptcy) and 0 otherwise (good firm). Let $\mathbf{x}' = (x_1, x_2, ..., x_p)$ be a covariate vector representing risk factors thought to be related to event under study. Assume the suitability of the retrospective sample and that $P(y|\mathbf{x})$ is represented by the logistic model

$$P\big(Y=1|\mathbf{x}\big) \;=\; \frac{e^{\alpha+\boldsymbol{\beta}'\mathbf{x}}}{1+e^{\alpha+\boldsymbol{\beta}'\mathbf{x}}}\,, \tag{2.1}$$

where α is an unknown scalar parameter and $\boldsymbol{\beta}' = (\beta_1, \beta_2, ..., \beta_p)$ is an unknown vector of coefficients. Now consider the hypothetical population to which (2.1) refers and let the marginal distribution of the covariates be denoted by $P(\mathbf{x})$. We draw a random retrospective sample of size n, with n_1 cases $(Y=1)$ and n_0 controls $(Y=0)$, in such a way that the marginal distribution of Y in the retrospective sample has M good cases for each bad one.

Let Z be a binary variable which takes the value 1 if a unit is included in the sample and 0 otherwise; moreover define $K_1 = P(Z=1|Y=1)$ the probability to extract a default unit and $K_0 = P(Z=1|Y=0)$ the complementary probability, both independent of the p dimensional vector of $\mathbf{x}$ covariates.

Let $P(\cdot|\mathbf{x}, Z=1) = P^*(\cdot|\mathbf{x})$ represent the distribution which is conditional on being observed in the retrospective sample. The probability distribution of Y given $\mathbf{x}$, conditional on being observed, is the following:

$$P^*\big(Y=1|\mathbf{x}\big) \;=\; \frac{K_1 P\big(Y=1|\mathbf{x}\big)}{K_1 P\big(Y=1|\mathbf{x}\big) \,+\, K_0 P\big(Y=0|\mathbf{x}\big)}\,, \tag{2.2}$$

$$\log \frac{P^*\big(Y=1|\mathbf{x}\big)}{P^*\big(Y=0|\mathbf{x}\big)} \;=\; \log \frac{K_1}{K_0} + \log \frac{P\big(Y=1|\mathbf{x}\big)}{P\big(Y=0|\mathbf{x}\big)}\,,$$

and substituting in the second term the logistic model expression, we have:

$$\log \frac{P^*\big(Y=1\,|\,\mathbf{x}\big)}{P^*\big(Y=0\,|\,\mathbf{x}\big)} \;=\; \log \frac{K_1}{K_0} + \alpha + \beta_1 x_1 + \cdots + \beta_p\, x_p \;, \tag{2.3}$$

where α and $\beta_j,\ j = 1, 2, 3, ..., p$, are the unknown parameters. It then follows that the logistic model for the retrospective balanced data has different intercepts but equal slopes and inference about α would require knowledge of $\frac{K_1}{K_0}$.

If $\alpha^\star$ denotes the intercept in the logistic model in the population with $Z = 1$, it follows that:

$$\alpha \;=\; \alpha^\star - \log \frac{K_1}{K_0} \;.$$

We indicate with the distribution $P(\mathbf{x}\,|\,y)$ the conditional distribution of the covariates given the response and with $P^*(\mathbf{x}\,|\,y)$ the same distribution conditional on being in the sample. As the sampling is independent of the covariates, the two distributions should be the same. The likelihood function for the retrospective sample can then be written:

$$\prod_{i=1}^{n} P\big(\mathbf{x}_i\,|\,y_i\big) \;. \tag{2.4}$$

Let $P^*(y_i\,|\,\mathbf{x}_i)$ be the conditional distribution of the response given that the covariates in a unit i are observed in the sample. Furthermore, let $P^*(y)$ denote the distribution of y and $P^*(\mathbf{x})$ represents the distribution of $\mathbf{x}$ conditional on being in the sample. Then from Bayes's rule (2.4) can be written as:

$$\prod_{i=1}^{n} \frac{P^*\big(y_i\,|\,\mathbf{x}_i\big)\, P^*(\mathbf{x}_i)}{P^*(y_i)} \;. \tag{2.5}$$

By the sampling scheme we know $P^*(Y=1)$ and $P^*(Y=0)$ are respectively equal to $\frac{n_1}{n}$ and $\frac{n_0}{n}$. For maximum likelihood inference (V.T. Farewell, 1979) we maximize (2.5) subject to the constraint

$$\sum_{\mathbf{x}} P^*\big(Y=1\,|\,\mathbf{x}\big)\, P^*(\mathbf{x}) \;=\; \frac{n_1}{n} \;, \tag{2.6}$$

where we have assumed that $\mathbf{x}$ is discrete. Anderson (1972) shows that the constrained maximum likelihood estimates of α^* and $\boldsymbol{\beta}$ are algebraically equivalent to the unconstrained estimates which maximize

$$\prod_{i=1}^{n} P^*\big(y_i\,|\,\mathbf{x}_i\big) \;, \tag{2.7}$$

while R.L. Prentice and R. Pyke (1979) show that the constrained estimation of (2.5) is a reparametrization of a likelihood based on the population model, where the constraints are defined in terms of the population value of $P(Y=1)$.

3. THE LOGISTIC MODEL FOR MATCHED CASE-CONTROL STUDIES

The logistic regression model for matched case-control studies, developed and widely used in epidemiology, may be considered as a refinement of the logistic modelling for balanced data. This method stratifies subjects on the basis of variables believed to be associated with the outcome. Again, we assume that the population model is logistic, as in (2.1). Within each stratum, samples of cases ($Y=1$) and controls ($Y=0$) are chosen, according to a $1-1$ design or $1-M$ design, where M is usually no more than five (Hosmer, Lemenshow and Sturdivant, 2013, p. 243). Let K be the number of strata, n_{1k} and n_{0k} respectively the cases and the controls within the k-th stratum, where $k = 1, 2, ..., K$. The stratum-specific logistic regression model for a unit in the sample, is

$$P\big(Y=1 \,|\, \mathbf{x},\, K=k\big) \;=\; \pi_k(\mathbf{x}) \;=\; \frac{e^{\alpha_k+\boldsymbol{\beta}'\mathbf{x}}}{1+e^{\alpha_k+\boldsymbol{\beta}'\mathbf{x}}}\,, \tag{3.1}$$

where α_k represents the contribution of all constant terms within the k stratum (i.e. stratification variable or variables) and $\boldsymbol{\beta}$ the vector of the p coefficients $\boldsymbol{\beta}' = (\beta_1, \beta_2, ..., \beta_p)$. From (2.3), it follows that the relationship between α and the stratum-specific parameters α_k varies among strata. Therefore, α_k are nuisance terms and should be eliminated from the set of parameters on which we want to make inference. The conditional likelihood method gives consistent and asymptotically normally distributed estimates of the β_j slope coefficients (Prentice and Pyke, 1979). The conditional likelihood for the k-th stratum is the probability that the observed case and control configuration is verified, conditioned on the stratum total and total number of observed case. Denoting $n_k = n_{1k}+n_{0k}$ as the number of subjects, the conditional likelihood for each stratum gives the probability to observe the data, conditioned on all possible assignment of cases n_{1k} and controls n_{0k}. The number of possible assignments of case status to n_{1k} among the n_k subjects is given by:

$$c_k \;=\; \binom{n_k}{n_{1k}} \;=\; \frac{n_k!}{n_{1k}!\,(n_k-n_{1k})!}\,.$$

Let the subscript j denote any one of these c_k assignments; moreover let subjects from 1 to n_{1k} correspond to the cases and subjects $n_{1k}+1$ to n_k to the controls. Any assignment is indexed by i for the observed data and by i_j for the j^{th} possible assignment. The conditional likelihood for the k-stratum is

$$l_k(\boldsymbol{\beta}) \;=\; \frac{\displaystyle\prod_{i=1}^{n_{1k}} P\big(\mathbf{x}_i \,|\, Y_i=1\big) \prod_{i=n_{1k}+1}^{n_k} P\big(\mathbf{x}_i \,|\, Y_i=0\big)}{\displaystyle\sum_{j=1}^{c_k} \left\{ \prod_{i_j=1}^{n_{1k}} P\big(\mathbf{x}_{ji_j} \,|\, Y_{i_j}=1\big) \prod_{i_j=n_{1k}+1}^{n_k} P\big(\mathbf{x}_{ji_j} \,|\, Y_{i_j}=0\big) \right\}} \tag{3.2}$$

and the full conditional likelihood over the K strata would be given by the product:

$$L(\boldsymbol{\beta}) = \prod_{k=1}^{K} l_k(\boldsymbol{\beta}) \,. \tag{3.3}$$

Assuming (3.1) is true and applying Bayes's rule to each $P(\mathbf{x}|y)$ term, we can write equation (3.2) as follows:

(3.4)

$$l_k(\boldsymbol{\beta}) = \frac{\displaystyle\prod_{i=1}^{n_{1k}} \left[\frac{P\big(Y_i=1|\mathbf{x}_i\big)\, P(\mathbf{x}_i)}{P\big(Y_i=1\big)} \right] \prod_{i=n_{1k}+1}^{n_k} \left[\frac{P\big(Y_i=0|\mathbf{x}_i\big)\, P(\mathbf{x}_i)}{P\big(Y_i=0\big)} \right]}{\displaystyle\sum_{j=1}^{c_k} \left\{ \prod_{i_j=1}^{n_{1k}} \left[\frac{P\big(Y_{i_j}=1|\mathbf{x}_{ji_j}\big)\, P(\mathbf{x}_{ji_j})}{P\big(Y_{i_j}=1\big)} \right] \prod_{i_j=n_{1k}+1}^{n_k} \left[\frac{P\big(Y_{i_j}=0|\mathbf{x}_{ji_j}\big)\, P(\mathbf{x}_{ji_j})}{P\big(Y_{i_j}=0\big)} \right] \right\}} \,.$$

Remembering that $P(Y_i=1|\mathbf{x}_i) = \pi(\mathbf{x}_i)$ and $P(Y_i=0|\mathbf{x}_i) = 1-\pi(\mathbf{x}_i)$ we can write:

$$l_k(\boldsymbol{\beta}) = \frac{\displaystyle\prod_{i=1}^{n_{1k}} \left[\frac{\pi(\mathbf{x}_i)\, P(\mathbf{x}_i)}{P\big(Y_i=1\big)} \right] \prod_{i=n_{1k}+1}^{n_k} \left[\frac{\big[1-\pi(\mathbf{x}_i)\big]\, P(\mathbf{x}_i)}{P\big(Y_i=0\big)} \right]}{\displaystyle\sum_{j=1}^{c_k} \left\{ \prod_{i_j=1}^{n_{1k}} \left[\frac{\pi(\mathbf{x}_{ji_j})\, P(\mathbf{x}_{ji_j})}{P\big(Y_{i_j}=1\big)} \right] \prod_{i_j=n_{1k}+1}^{n_k} \left[\frac{\big[1-\pi(\mathbf{x}_{ji_j})\big]\, P(\mathbf{x}_{ji_j})}{P\big(Y_{i_j}=0\big)} \right] \right\}}$$

and also

(3.5)

$$l_k(\boldsymbol{\beta}) = \frac{\displaystyle\prod_{i=1}^{n_{1k}} \left[\frac{\frac{e^{\alpha_k+\beta'\mathbf{x}_i}}{1+e^{\alpha_k+\beta'\mathbf{x}_i}}\, P(\mathbf{x}_i)}{P\big(Y_i=1\big)} \right] \prod_{i=n_{1k}+1}^{n_k} \left[\frac{\frac{1}{1+e^{\alpha_k+\beta'\mathbf{x}_i}}\, P(\mathbf{x}_i)}{P\big(Y_i=0\big)} \right]}{\displaystyle\sum_{j=1}^{c_k} \left\{ \prod_{i_j=1}^{n_{1k}} \left[\frac{\frac{e^{\alpha_k+\beta'\mathbf{x}_{ji_j}}}{1+e^{\alpha_k+\beta'\mathbf{x}_{ji_j}}}\, P(\mathbf{x}_{ji_j})}{P\big(Y_{i_j}=1\big)} \right] \prod_{i_j=n_{1k}+1}^{n_k} \left[\frac{\frac{1}{1+e^{\alpha_k+\beta'\mathbf{x}_{ji_j}}}\, P(\mathbf{x}_{ji_j})}{P\big(Y_{ij}=0\big)} \right] \right\}} \,.$$

Moreover, collecting common terms of the form

$$\frac{1}{1+e^{\alpha_k+\boldsymbol{\beta}'\mathbf{x}}}$$

we can write (3.5) as:

$$l_k(\boldsymbol{\beta}) = \frac{\displaystyle\prod_{i=1}^{n_{1k}} \Big[e^{\alpha_k+\beta'\mathbf{x}_i} \Big] \prod_{i=1}^{n} \left[\frac{1}{1+e^{\alpha_k+\beta'\mathbf{x}_i}} \right] \prod_{i=1}^{n_k} \left[\frac{P(\mathbf{x}_i)}{P(Y_i)} \right]}{\displaystyle\sum_{j=1}^{c_k} \left\{ \prod_{i_j=1}^{n_{1k}} \Big[e^{a_k+\beta'\mathbf{x}_{ji_j}} \Big] \prod_{i_j=1}^{n} \left[\frac{1}{1+e^{\alpha_k+\beta'\mathbf{x}_{ji_j}}} \right] \prod_{i_j=1}^{n_k} \left[\frac{P(\mathbf{x}_{ji_j})}{P(Y_{i_j})} \right] \right\}} \,. \tag{3.6}$$

Further algebraic simplification leads to the following:

$$l_k(\boldsymbol{\beta}) = \frac{\prod_{i=1}^{n_{1k}} e^{\boldsymbol{\beta}'\mathbf{x}_i}}{\sum_{j=1}^{c_k}\prod_{i_j=1}^{n_{1k}} e^{\boldsymbol{\beta}'\mathbf{x}_{ji_j}}}, \tag{3.7}$$

where $\boldsymbol{\beta}$ is the only unknown parameter.

In a $1-1$ matched design each case is matched to a single control. Let $\mathbf{x}_{1k}$ and $\mathbf{x}_{0k}$ respectively denote the data vector for the case and the control in the k-th stratum, the conditional likelihood for the k-th stratum is

$$l_k(\boldsymbol{\beta}) = \frac{e^{\boldsymbol{\beta}'\mathbf{x}_{1k}}}{e^{\boldsymbol{\beta}'\mathbf{x}_{1k}} + e^{\boldsymbol{\beta}'\mathbf{x}_{0k}}} \tag{3.8}$$

given specific value for $\boldsymbol{\beta}$, $\mathbf{x}_{1k}$ and $\mathbf{x}_{0k}$, equation (3.8) is the probability that the unit identified as the case is in fact the case. If data for case and control are identical, $\mathbf{x}_{1k}= \mathbf{x}_{0k}$, it follows from equation (3.8) that $l_k(\boldsymbol{\beta}) = 0.5$ for every value of $\boldsymbol{\beta}$ and the stratum will be considered as *uninformative* meaning that the covariates do not discriminate cases from controls.

In a $1-1$ matched data design with a binary explanatory variable X, the conditional maximum likelihood estimator is the log of the ratio of discordant pairs (see Breslow and Day, 1980). It follows that it is advisable to classify (in a 2×2 table) cases versus controls for each dichotomous variable to verify the presence of discordant pairs: the absence of both types of pairs ($x_{1k} = 1$, $x_{0k} = 0$) and ($x_{1k} = 0$, $x_{0k} = 1$) gives rise to an undefined estimator.

In a $1-M$ matched design each case is matched to M controls, so there are $M + 1$ units in each stratum. Letting $M = 4$ and denoting by $\mathbf{x}_{k1}$ the case and by $\mathbf{x}_{k2}$, $\mathbf{x}_{k3}$, $\mathbf{x}_{4k}$, $\mathbf{x}_{5k}$ the controls in the k^{th} stratum, the contribution to the likelihood for this stratum of matched subjects from equation (3.7) is

$$l_k(\boldsymbol{\beta}) = \frac{e^{\beta'\mathbf{x}_{k1}}}{e^{\boldsymbol{\beta}'\mathbf{x}_{k1}} + e^{\boldsymbol{\beta}'\mathbf{x}_{k2}} + e^{\boldsymbol{\beta}'\mathbf{x}_{k3}} + e^{\boldsymbol{\beta}'\mathbf{x}_{k4}} + e^{\boldsymbol{\beta}'\mathbf{x}_{k5}}}. \tag{3.9}$$

Given the coefficients' values (3.9) gives the probability that the unit with the observed data $\boldsymbol{x}_{k1}$ is the case relative to four controls with data $\mathbf{x}_{k2}$, $\mathbf{x}_{k3}$, $\mathbf{x}_{k4}$, and $\mathbf{x}_{k5}$. If the four covariates have the same value, then $l_k(\boldsymbol{\beta}) = 0.20$ for each $\boldsymbol{\beta}$ value. Hence, for each covariate at least one control should have a value different from the case, otherwise the stratum would be considered *uninformative*.

4. THE LOGISTIC MODEL FOR "ROSE" DATA

Random OverSampling Examples (Lunardon, Menardi and Torelli, 2014) is a new procedure developed in the R language (R Development Core Team, 2015), based on the generation of new artificial data according to a smoothed bootstrap approach (Efron and Tibshirani, 1993). Let $P(\mathbf{x}) = f(\mathbf{x})$ be the probability density function on X. Let $n_j < n$ be the size of Y_j, $j = 0, 1$. A new sample is generated by the following three steps:

1. select $y = Y_j$, $j \in \{0, 1\}$, with probability $\frac{1}{2}$;
2. select $(\mathbf{x}_i, y_i)$ in the sample such that $y_i = y$ with probability $p_i = \frac{1}{n_j}$;
3. sample the vector of covariates $\mathbf{x}$ from the kernel probability distribution $K_{\boldsymbol{H}_j}(\cdot, \mathbf{x}_i)$, centered on $\mathbf{x}_i$ and depending on the matrix of smoothing parameters $\boldsymbol{H}_j$.

According to the ROSE method, in the training sample you extract a unit belonging to one of the two classes with the same probability. Then a new sample is generated in its neighborhood of width determined by $\boldsymbol{H}_j$. Generally $K_{\boldsymbol{H}_j}$, is chosen as symmetric and unimodal. Therefore the generation of new samples for the class Y_i according to ROSE corresponds to the generation of data from the kernel density estimate of $f(\mathbf{x}, Y_j)$, with matrix of smoothing parameters $\boldsymbol{H}_j$:

$$\hat{f}\big(\mathbf{x}|y=Y_j\big) = \sum_{i=1}^{n_j} p_i \, P\big(\mathbf{x}|\mathbf{x}_i\big) = \sum_{i=1}^{n_j} \frac{1}{n_j} \, P\big(\mathbf{x}|\mathbf{x}_i\big) = \sum_{i=1}^{n_j} \frac{1}{n_j} \, K_{\boldsymbol{H}_j}(\mathbf{x} - \mathbf{x}_i) \,, \tag{4.1}$$

see Menardi and Torelli (2014).

5. DATA ANALYSIS

Data are drawn from the *AIDA* database[1], one of the most important Italian databases containing historical balance sheets as well as financial, commercial and demographic information on more than one million Italian firms. We selected all the firms in the Tuscany region having positive revenues of sales in the year 2006 and for these we extracted revenues, profits, fixed assets, financial indicators, indexes of resultant profits and current management. On May 2010, we verified their legal status: the database provides the legal status of each firm, which is periodically updated without indicating the reference date; therefore, we do not know the exact time at which a firm was declared bankrupt. The selected time interval of four years is due to the delay between the bankruptcy event and the availability of the company balance sheet. The distribution according to the legal status is shown in Table 1.

[1]The database is distributed by Bureau Van Dijk s.p.a.; `https://aida.bvdinfo.com/`

Table 1: Companies' distribution by state law in May 2010.

Legal status	**Frequency**	**Percentage**
Active	33798	89.23
Bankruptcy	537	1.42
Liquidation	2800	7.39
Not active	744	1.96

Due to the lack of information on the causes of inactivity and liquidation, we included in the analysis only the firms that are active (33798) and bankrupt (537). Data clearly shows the rareness of the default event (1.56%) and therefore the inadequacy of a logistic regression model due to the unbalanced data. We built a training sample, to implement the methods, and a randomly selected hold-out sample consisting of 10% of the whole sample. Since the aim of the study is to compare three different methodologies, from among the balance sheet indicators we selected those that were found to be most informative in a previous case-control study (Pierri, 2013) on the same data.

Logistic regression is estimated on three different data sets: a balanced sample with 2505 observations where the frequencies of Y_0 and Y_1 are respectively 501 and 2004; a stratified sample (2440 observations) with strata formed on Legal Form and the first two numbers of the ATECO code (industry sector) jointly considered, where the frequencies of Y_0 and Y_1 are 488 and 1952, respectively; and a ROSE data set of 68000 observations where the frequencies of Y_0 and Y_1 are respectively 33671 and 34329. We used the ROSE routine included in R software to generate data based on the ROSE method.

In the multivariable logistic model we considered as explanatory variables Net Profit (NP), Asset Coverage Index (AC), Liabilities index (L), Quick Ratio (QR), Debt Ratio (DR), Asset Turnover (AT) and EBITDA value. The linearity in the odds of these variables was checked following the methodology proposed by Hosmer, Lemenshow and Sturdivant (2013, Ch. 4): transformation of variables, applied where necessary, led to a final model including two quadratic forms. For a detailed implementation, see also Pierri, Burchi and Stanghellini (2013). We refer to Table 2 for a summary of our main results. The logistic model on the balanced sample indicates that for Asset Turnover the quadratic form is not statistically significant. The same holds for the ROSE sample where Net Profit is also not significant. The same table also displays the estimated coefficients for the models considering only the significant ($p < 0.05$) covariates (Balanced2 and ROSE2). Balanced and case-control methods produce close estimates, while ROSE method gives smaller values. This may due to the use of artificial data. The economic interpretation of the model is also consistent with the expected results: Tuscany region is characterized by small and medium-sized enterprises as the 98% of the Italian firms. In this context the financial structure plays an

important role, because they are often under-capitalized. From data in Table 2 we can notice that the forms of debt that a company chooses is of great importance in determining the probability of default: the negative value of the Debt Ratio combined with a positive Asset Coverage Index, show that the probability of having a healthy company increases if you prefer forms of internal financing. Moreover companies with a positive Quick Ratio and a negative Liabilities Index are less exposed to the risk of default as more able to obtain long-term funding, while short-term debt may compromise the health of a company.

Table 2: Estimates of the coefficients applying the three different methods.

Explanatory Variables	**Balanced**	**Balanced2**	**ROSE**	**ROSE2**	**Case Control**
NP	0.00063	0.00064	2.22e−07*	—	0.00241
AC	0.12323	0.12541	0.05821	0.05813	0.10371
AC^2	−0.01069	−0.01082	−0.00636	−0.00636	−0.01176
L	−1.05589	−0.98879	−0.61510	−0.61631	−0.88630
QR	0.59219	0.60493	0.44750	0.44740	0.72559
DR	−0.00583	−0.00590	−0.00174	−0.00174	−0.00472
AT	0.47878	0.3058	0.26130	0.26927	1.02132
AT^2	−0.05038*	—	−0.00289*	—	−0.17471
EBITDAV	0.01522	0.01520	0.00747	0.00747	0.00574

(* *p-value* > 0.1)

We compared the predictive and discriminatory ability of the three methods looking at the ROC curves built with the hold-out sample. In Figure 1 we notice that the logistic model on balanced data (AUC = 0.7955) has the greatest capability to discriminate between good and bad firms while ROSE (AUC = 0.7645)

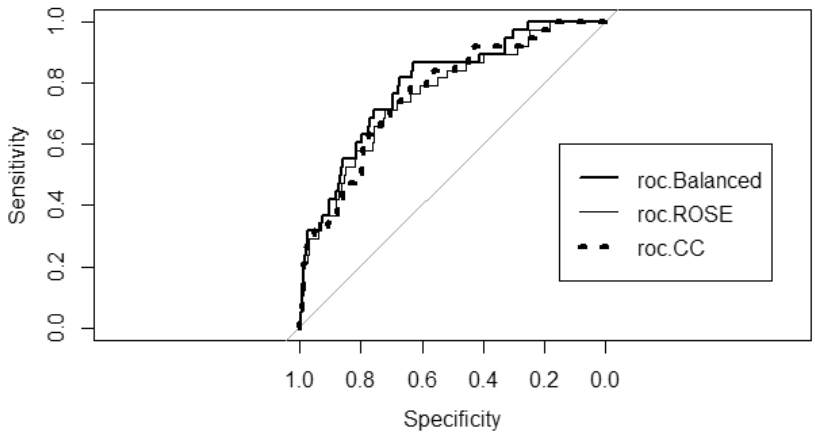

Figure 1: Estimated ROC curve in the three models using hold-out sample: Balanced (AUC = 0.7955); ROSE (AUC = 0.7645); Case Control (AUC = 0.7686).

and Case-Control (AUC = 0.7686) methods exhibit very similar results to each other. Testing the difference between their AUC, we find a significant difference ($p\text{-}value < 0.05$) only between Balanced and both Case Control and ROSE areas. We achieve similar results if we consider ROC curves built for Balance2 (AUC = 0.7911) and ROSE2 (AUC = 0.7685) models.

6. DISCUSSION

Three different methodologies have been compared. On the basis of the data and the model applied, the oversampling (ROSE) and case-control studies methods seem to give very similar results, on the other hand logistic regression on balanced data show the best predictive capabilities. We underline some particularities: ROSE allows only for continuous covariates; in case-control studies, confidence intervals are generally narrower than in standard logistic regression, but does not produce the predicted probability of bankruptcy; standard logistic regression, applied over a random balanced sample, is very easy and quick to implement. Future developments of this study will test whether stepwise model selection procedures applied to the different datasets will lead to different models.

REFERENCES

[1] Anderson, J.A. (1972). Separate sample logistic discrimination, *Biometrika*, **59**(1), 19–35.

[2] Breslow, N.E. and Day, N.E. (1980). *Statistical Methods in Cancer Research, Vol. 1: The Analysis of Case-Control Studies*, International Agency of Cancer, Lyon, France.

[3] Efron, B. and Tibshirani, R. (1994). *An Introduction to the Bootstrap*, Chapman & Hall/CRC, Boca Raton London New York Washington, D.C..

[4] Farewell, V.T. (1979). Some results on the estimation of logistic models based on retrospective data, *Biometrika*, **66**(1), 27–32.

[5] Fawcett, T. (2006). An introduction to ROC analysis, *Pattern Recognition Letter*, **27**(8), 861–874.

[6] Hosmer, D.W.; Lemeshow, S. and Sturdivant, R.X. (2013). *Applied Logistic Regression*, Wiley, New Jersey.

[7] King, G. and Zeng, L. (2001). Logistic Regression in Rare Events Data, *Political Analysis*, **9**(2), 137–163.

[8] Lunardon, N.; Menardi, G. and Torelli, N. (2014). ROSE: a package for Binary Imbalanced Learning, *R Journal*, **6**(1), 79–89.

[9] Mantel, N. (1973). Synthetic retrospective studies and related topics, *Biometrics*, **29**(3), 479–486.

[10] Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Biometrics*, **29**(3), 479–486.

[11] Menardi, G. and Torelli, N. (2014). Training and Assessing Classification Rules with Imbalanced Data, *Data Mining and Knowledge Discovery*, **28**(1), 92–122.

[12] Pierri, F. (2013). *Valutazione del rischio di default nelle piccole e medie imprese attraverso uno studio caso-controllo*. In "Gli Analytics come motore per i big data, la ricerca ed il sistema paese" (A. Di Ciaccio and W. Lanzani, Eds.), Aracne Editrice, Roma (Italy), 57–68.

[13] Pierri, F.; Burchi, A. and Stanghellini, E. (2013). La capacitá predittiva deglii indicatori di bilancio delle PMI, *Piccola Impresa Small Business*, **1**, 85–106.

[14] Prentice, R.L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies, *Biometrika*, **66**(3), 403–411.

[15] R Development Core Team, (2015). R: A Language and Environment for Statistical Computing, *R Foundation for Statistical Computing*, Vienna, Austria, `http://www.R-project.org`.

[16] Stanghellini, E. (2009). *Introduzione ai metodi statistici per il credit scoring*, Springer Verlag (in Italian).

[17] Zhang, B. (2006). Prospective and retrospective analyses under logistic regression models, *Journal of Multivariate Analysis*, **97**, 211–230.

REVSTAT – Statistical Journal
Volume 14, Number 2, April 2016, 171–192

MODELING NON-LIFE INSURANCE PRICE FOR RISK WITHOUT HISTORICAL INFORMATION

Authors: Filipe Charters de Azevedo
– Department of Sciences and Technology, Universidade Aberta, Portugal
fchartersazevedo@hotmail.com

Teresa A. Oliveira
– Center of Statistics and Applications (CEAUL), University of Lisbon, and Department of Sciences and Technology, Universidade Aberta, Portugal
teresa.oliveira@uab.pt

Amilcar Oliveira
– Center of Statistics and Applications (CEAUL), University of Lisbon, and Department of Sciences and Technology, Universidade Aberta, Portugal
amilcar.oliveira@uab.pt

Received: October 2015 Revised: February 2016 Accepted: February 2016

Abstract:

- How should an insurer price a risk for which there is no history? This work intends to show, step by step, which main mechanisms are needed to capture the tariff model of another insurance company minimizing the risk involved. The document generally deals with the price-making mechanisms in non-life insurance through the GLM regression models — Generalized Linear Model, more precisely the Poisson, Gamma and Tweedie models. Given the complexity of the application of these models in experimental design, it is studied a simpler way to characterize the rate, namely considering the Box–Cox transformation with SUR — Seemingly Unrelated Regression. An orthogonal experimental design to collect information is also presented as well as an application of these methods in the motor industry considering different companies.

Key-Words:

- *pricing (non-life insurance); GLM; Box–Cox; optimal designs; SUR — Seemingly Unrelated Regression.*

AMS Subject Classification:

- 62J12, 62K05, 91B24, 91B30.

1. INTRODUCTION

An insurance company bases its production model in the value of a commodity with unknown cost by the time of production. Furthermore, the company "purchase" claims and "sell" safety, if a company buys the claims at a low price then it makes money; if it buys the claims at an expensive price then it loses money. In the value chain, a company can rely on the law of large numbers that mitigates volatility and market uncertainty — provides security on average.

The bottom line of a company is then how to evaluate the purchase price of claim: What is the cost of a risk (pure premium)? Usually an insurer has historical data that allow to estimate this value: based on the behavior of their customers it is reasonable to offer a premium, that is identical to the liabilities assumed (adding administrative costs, distribution and shareholder remuneration). But how should an insurer do to price a risk, for which there is no history? Should a company to "pay to view" — risking prices and their future sustainability? What should an insurer do if both — market and risk — are unknown?

From a practical point of view these questions are extremely important once the market has a strong barrier to overcome — the knowledge of the cost of raw materials. However there are solutions available in the literature. Some companies:

- Hire experienced technicians that heuristically define a charging table. Many investors are attracted to base their decisions on the information "currently available in their minds" see (Nocetti [5] and [6]). Thus, many times even when company has some historical data, experts opinions can be more plausible than the detailed analysis.
- Adopt reinsurance for (almost) 100% of the costs, transferring the risk for more experienced companies (which will draw a tariff) and that have financial muscle (to support higher risks).

In both cases (hiring experienced technicians or reinsurance) there is risk, and/or potential revenue loss. Are the companies locked to this reality? In any case, the insurer will always bear the costs of administration and distribution.

The challenge assumes more interesting contours since it is known that the *player* who first entered the market, or which has a higher market, has a strong competitive advantage: its historical references provide knowledge, which in this industry means the ability to determine more accurately the cost of the raw material. The *player* with no experience, only will get an interesting share if he gets a similar competitive advantage over the incumbent.

The aim of this work is thus to present a minimization method of pricing risk by capturing the tariff model, enabling a comparative advantage in the market to

smaller *players* (in terms of market share), with no relevant history and without financial padding to buy knowledge in a significantly way that is assuming risk. This capture method is based on the assumption that the smallest company can access a reasonable number of simulations with surgically chosen risk profiles. This collection can be performed, for example, by a mystery shopper or using their own mediators.

The methodological approach for answering to this challenge, follows the classic process of experimental design:

- Step 1: Identify the factors that define the product;
- Step 2: Identify levels that define the product;
- Step 3: Optimal Design;
- Step 4: Gathering information;
- Step 5: Analysis.

Considering the particular case of motor insurance an application will be performed in the sequence.

This work is organized in the following chapters:

- General Linear Models. In chapter 2 attention is given to changing pricing methodologies, particularly with regard to the GLM model associated to Tweedie distribution.
- Experimental design in context of a Tweedie population. The purpose of chapter 3 is to build a sample design which minimizes the field of endeavor, by using an Optimal Design and Box–Cox transformation. This is a practical solution once considering Tweedie populations, the variation component is not easily determined in an experimental design.
- Optimal Design. In chapter 4 the orthogonality concept is presented in order to gather information and allows discussing their suitability to the main objectives of the project: reducing the volume of information to be collected in order to obtain a manageable model and efficient estimates to facilitate the risk modeling. In this chapter special emphasis will be given to Seemingly Unrelated Regression — SUR — in order to maximize the predictability capacity.
- Applications. The methodologies explored in previous chapters are applied in chapter 6. We are working on a confidential real database, considering motor insurance data from a Portuguese insurance company in 2011.
- Conclusion and remarks. In the last chapter 7 emphasis should be given to the widespread conditions.

2. GLM — GENERALIZED LINEAR MODELS

How does a company know that it is expensive or cheap to pay for a policy? Going to market a company subscribes a policy and accept risks for which the real cost is unknown.

The well known expression of what usually is known as pure premium, which supports the rational of an insurance rate construction is:

$$\text{(2.1)} \qquad \textit{Pure Premium} \;=\; \textit{sinister frequency} \times \textit{average claim cost}\,(+\textit{error})\;.$$

Usually an insurer apply statistical models to estimate the frequency of an accident and their average claim cost. This problem can be seen isolated (frequency and average claim cost) or estimated jointly (Pure Premium). The concept of regression tackles this problem successfully, whatever its formulation. It should be noted that in a generalized regression model there are two components:

- **i)** A random vector $Y = (Y_1, ..., Y_n)'$ is following a distribution with unknown parameters vector $\mu = (\mu_1, ..., \mu_n)'$;
- **ii)** A function relation between μ and the involved parameters' vector $\beta = (\beta_1, ..., \beta_k)'$, such that $\mu = f(\beta)$, considering $f(\cdot)$ a continuous and univocal function.

Following the terminology of Jørgensen ([15]), these two components are referred respectively as the random component and the systematic component of the model. The random vector Y is designated as the response, while the random component can assume any stochastic process, including errors measurement. The function $f(\cdot)$ is designated as the regression function and the β parameters represent the regression parameters. This whole system of vectors and distributions is defined though the average for each Y_i on the conditions of μ, $E(Y_i\,|\,\mu)$. The variation associated with $E(Y_i\,|\,\mu)$ provide a measure of the adjustment quality.

An important class of regression models can be expressed as:

$$\begin{aligned} g(\mu_i) \;&=\; \eta_i \;, && i = 1, ..., n \;; \\ \eta_i \;&=\; \sum_{j=1}^{k} x_{ij}\,\beta_j \;, && i = 1, ..., n \;. \end{aligned}$$

The function $g(\cdot)$ is continuous and unequivocal and is designated as *link function*. The matrix $X = \{x_{ij}\}$ is the design matrix model and x_{ij} are the covariates or explanatory variables. A model of this form is said to be linear.

When $g(\cdot)$ is an identity function and Y distribution is homoscedastic or even normal, the simple linear regression model is considered. Usually in this

simple case parameters are estimated by the least squares error minimization or by the maximum likelihood.

In some cases it is possible to question the type of function $g(\cdot)$ assumes as well as the distribution associated with Y, which is a very convenient way to determine the heterogeneity of the data. It is typical to assume that Y distribution is defined by a Poisson probability function and that Gamma distribution is used to compute the average cost. When a claim frequency is the goal, to establish the terms in (2.1) together, the most traditional mechanism is through GLM composite models. In this case, a very convenient way to determine the heterogeneity of the data is assuming that the Y distribution is defined by an exponential distribution model (ED):

$$p(y,\theta,\lambda) \;=\; \alpha(\lambda,y)\, e^{\left[\lambda\left\{y\theta-k(\theta)\right\}\right]}\,, \qquad \text{with} \;\; y\in\mathbb{R}\,.$$

Note that $\alpha(\cdot)$ and $k(\cdot)$ represent functions, and $\lambda>0$ and θ belongs to a real domain.

Thus, let $Y\sim ED(\mu,\sigma^2)$ where $\mu=k'(\theta)$ represent the expected value of Y and $\sigma^2=\frac{1}{\lambda}$ represent the variance where ED refers to the family of exponential distributions and in Jørgensen ([15]) there is a particular case of this family distributions, characterized by $V(Y)=\sigma^2=\phi\,V(\mu)$.

The particular cases of $V(Y)=\sigma^2=\phi\,\mu^p$, to diverse p assume an important class usually associated to the Tweedie distribution model. This class can be:

- a normal data generator when $p=0$;
- a Poisson data generator when $p=1$;
- a Gamma data generator when $p=2$;
- an Inverse Gaussian when $p=3$.

Considering $1<p<2$ the Tweedie exponential distribution assumes the expression:

$$p(y,\theta,\lambda) \;=\; \sum_{n=1}^{\infty} \frac{\left\{(\lambda\omega)^{1-\alpha}\,k_\alpha\!\left(-\frac{1}{y}\right)\right\}^n}{\Gamma(-n\,\alpha)\,n!\,y}\; e^{\left[\lambda\left\{y\theta-k_\alpha(\theta)\right\}\right]}\,, \qquad \text{to} \;\; y>0\,.$$

Let $P(Y{=}0)=e^{\{\lambda\omega k_\alpha(\theta_0)\}}$ where $k_\alpha(\theta)=\frac{\alpha-1}{\alpha}\left(\frac{\theta}{\alpha-1}\right)^\alpha$, $\theta_0=\theta\,\lambda^{\frac{1}{(1-\alpha)}}$, and ω represents the weight associated to the observation exposition. As we can observe, for the Tweedie distribution the density function depends on the parameter p which relates to the variance $V(Y)=\sigma^2=\phi\,\mu^p$. This parameter p is thus defined exogenously before the estimation process, usually due to the analyst experience. These GLM models are widely recognized in the industry. Anderson *et al.* ([1]) presents it as the standard method to define motor and other lines of commercial branches tariffs. It is also indicate that these models are used by companies

in the UK, Ireland, France, Italy, the Netherlands, Scandinavia, Spain, Portugal, Belgium, Switzerland, South Africa, Israel and Australia. The referred paper also states that this model has gained popularity in Canada, Japan, Korea, Brazil, Singapore, Malaysia and Eastern European countries. More details on this distribution applied to the actuarial context can be obtained in Jørgensen and de Souza ([17]).

The type of function $g(\cdot)$ follows some rationals, as the context of analysis and distribution of Y.

3. EXPERIMENTAL DESIGN CONSIDERING TWEEDIE POPULATIONS

It has been noted that insurance companies are usually using a way of charging based on GLM that combine Poisson|Gamma model or on a composite model (known as Tweedie model). For the sample design in Tweedie regression, see Jørgensen and de Souza ([17]), it should be noted the following approaches, which are well known in the literature:

(**i**) **Sequential design**: The sequential design for binary responses has a rich history, which dates back to the 1940s, see Wald ([23]), on trying to find designs which results lead to asymptotic properties, see also Haines *et al.* ([12]), Ivanova and Wang ([14]) and Karvanen *et al.* ([18]). These authors concentrate their work on one factor designs and the challenge in our work is to extend this research to multifactorial designs. In Woods *et al.* ([24]) and Dror and Steinberg ([8]), solutions to this multifactoriality problematic are presented. However, such solutions are computationally complex, and the associated methodology is based on "ebb and flow" and trial and error, making the process complex and nonintuitive.

(**ii**) **Design based on clusters**: The Tweedie regression is based on the estimation of three parameters vectors: φ, θ and p, where p conditions affects the other two parameters computation. The design by clusters seek to find homogeneous groups of observations in order to determine p in an exogenously way. This idea is conceptually interesting, and is computationally easy to perform.
In Dror and Steinberg ([7]) is suggested an approach based on K-means cluster — since this process allows rapid exploration of various designs outperform the existing alternatives. The authors mention "given the set of location D-optimal designs, the core of the proposed method is to combine them into the set of vectors location and use K-means clustering to derive a robust design".

The possibility of finding an optimum location with this method has, however, a serious problem with respect to the other model coefficients: how to evaluate the estimated degree of accuracy? The question arises once the clusters were defined exogenously and a sample experimental design is always reduced to allow "good" experiences performance. As an avenue for improvement one can explore the use of computational simulations to ensure the best model based on different levels of p. Algorithms based on *random forest* may be an important issue to consider. However this is not the main goal of this project.

3.1. Experimental design 1: A pragmatic solution in Tweedie populations

As the GLM models Tweedie are not easily applicable in the experimental context it is necessary to find a pragmatic solution. The main problem is to have an experimental model analysis under heterocedasticity conditions or dispersion models, where Tweedie distribution fits well.

Trying to stabilize the variance, see Box and Cox ([4]), the usual method is to determine empirically or theoretically the ratio between the variance and the mean. The empirical relationship can be found by the logarithm and the average graph, or to make a transformation in the dependent variable.

For positive expected value, the well-known Box–Cox Transformation is frequently used:

$$y^{(\lambda)} = \begin{cases} \dfrac{x^{\lambda}-1}{\lambda} & \text{to } \lambda \neq 0\,, \\ \log(x) & \text{to } \lambda = 0\,. \end{cases} \tag{3.1}$$

The choice of λ however, in our days, is usually done automatically, while Osborne ([20]) and Harrison and McCabe ([13]) propose the following algorithm:

1. Splitting the key variable in 10 (or more) intervals;
2. Calculate the mean and standard deviation for each interval;
3. Design a graphic with $\log(\sigma)$ vs. $\log(\mu)$ for each of the regions;
4. Estimate the average slope of the graphic, and use the $1-\lambda$ as initial value of λ.

It is important to refer that this algorithm is not an unanimous choice for researchers and usually, as in Ripley *et al.* ([22]) it is assumed that the best way to estimate λ is the one that guarantees the maximum likelihood. Drawing the evolution of the maximum likelihood function can be useful in this case.

However this method is not prudent, since any $|\lambda|$ too high will reduce the variability of the variable goal. Therefore when you re-build the target variable of $y^{(\lambda)}$ to y the result may be an estimated variable without any variability. Some software and statistical packages (MASS in R, STATA) maintain this approach, but impose limits to $y^{(\lambda)}$, usually in $y^{(\lambda)} < 1$, $y^{(\lambda)} < 2$.

Finally, another alternative is to look for a λ value that makes sense to the analyst. A careful reading of Box and Cox ([4]) points in that direction. So it is noteworthy that the way loglin $\Longleftrightarrow \lambda = 0$ is theoretically the one that makes more sense to use for the premium model, and is easier to interpret:

1. The distribution of total / pure premium costs (i.e. Tweedie with reliable parameters) is visually close to a log normal or gamma;

2. The log-lin model has the advantage that the coefficients represent elasticities; a very meaningful concept in terms of premiums.

A very simple way to find and test the data transformation — at least between the linear form and log — is presented in Mackinnon and Davidson ([19]). Although the estimation process may seem a little complex, the test logic is very simple: if the linear model is in fact correct, the formula $e^{(\widehat{\log(y_{\text{based on log model}})})}$ will be related to the model under evaluation (so, it will be enough to use the regression and a t-test).

In short, to determine the tariff model competition in experimental design context, the Box–Cox methodology is preferable to the Tweedie regression. In addition, after the analysis of the best functional form, subsequently a Tweedie regression may be applied but using Box–Cox, for a rough indication of what value p may assume (i.e. the shape of the Tweedie), so that to overcome the already mentioned difficulties. That is why we propose this strategy to overcome the existing computational difficulties.

3.2. Experimental design 2: Box–Cox regression correction

The key variable to estimate in this case is $y = \textit{commercial premium}$ and not $y^{(\lambda)}$. When the estimation process is integrated there is the need to decompose the Box–Cox formulation in the correct formulation:

$$\widehat{y} = \begin{cases} \widehat{y^{(\lambda)}}\,\lambda + 1 & \text{to } \lambda \neq 0\,, \\ e^{(\widehat{y^{(\lambda)}})} & \text{to } \lambda = 0\,. \end{cases} \tag{3.2}$$

However, this formulation is not the most statistically efficient. In fact the application of Box–Cox expression underestimates the y expected value.

The mass point of the linear Box–Cox regression (the average of x and y on average) is not identical in both equations, in order to $\widehat{y^{(\lambda)}}$ and in order to y. In Wooldridge ([25]) is presented a solution for the case $\lambda = 0$, which can be generalized to any variant of Box–Cox regression. It is possible to obtain a corrected model by regressing y to $\widehat{y}$ without the constant component. The coefficient associated to $\widehat{y}$ gives the correction factor of the mass points. So, taking into account the correction, the final prediction for two stages estimation is:

$$\widehat{\widehat{y}} = \widehat{y} \times \textit{correction coefficient} \, . \tag{3.3}$$

4. EXPERIMENTAL DESIGN: IDENTIFICATION OF FACTORS AND LEVELS TO COLLECT

Regarding the determination of the factors, the experimental work is easy: Each customer must fill out a quotation document so that a quote can be issued. Usually, there are no data to work beyond the required (although it is known that some insurers in bancassurance partnerships use the bank behavior data, and in other countries it is known that the profile on social networks can be used). The work on validation factors in brainstorming sessions and interviews with different experts, see Barker ([2]), is dispensed as companies in the quotation indicate which factors are supposed to be investigated.

In the case of motor insurance, the factors usually considered are:

1. **Characteristics of the insured**
   - Gender: The rational of this variable is associated mainly to a different frequency of accidents according to gender. It should be noted, however, that in March 1, 2011 the European Court of Justice ruled that insurance companies which use gender as a risk factor were to disregard EU equality laws. However, in Portugal, in February 2015 (Law No. 9/2015), it became amicable to have same gender discrimination if premiums and benefits "are proportionate and justified by a risk assessment based on actuarial and statistically relevant and accurate data". The possibility of using this variable from 2015 thus became a reality.
   - Age: The rational of this variable is to measure the inexperience and risk trend of the insured. It is a variable impacting the accident frequency and, depending on the coverage, the average cost.
   - Claims History: The claim record can be consulted by the Portuguese insurers in SegurNet (a managed platform for the sector's association with the claim record).

- Age when the driver got driver license: When combined with age, it attain an instrumental variable of the driver's experience.
- Status: Rarely used in Portugal, although there is some sensitivity to point to the fact that married drivers have fewer accidents than the rest of the population.
- Usual path: Variable indicating the accident frequency — the greater the distance house-work, the greater the likelihood of an accident.
- Payment: The payment method reveals the financial pressure that the driver is subject; is a factor that correlates with the driving profile — frequency. In addition, the payment method is correlated with the insurer capital consumption and therefore to effect on the commercial premium via the administrative burdens and profit loading.

2. **Features of insurance risks**
   - Vehicle rating: The power to weight ratio increases the sinister frequency.
   - Brand and classification of vehicle: There are brands whose parts cost more than others, so this variable has an impact on the average cost. The type of construction, security features also has its influence on the average cost.
   - Using the insured object: If the object is essential for day-to-day, or for professional use, accident frequency increases while the frequency per unit of exposure (measured in km driving) decreases. Thus, having a profession and any instrumental variable of the insured object use is relevant.

3. **Regional and general contexts**
   - Weather: The loss context has been the least considered issue in the construction of a tariff. For example, if it rains, there are more accidents, but companies have rates for a given country in which implicitly rainfall rates do not vary. If there is a crisis, people use less the car, so there are fewer accidents. These context variables are linked to the evolution of times and this issue should be carefully analyzed.
   - Region: Regional variables are often neglected since everything is placed in large commercial areas and not with sufficient granularity.
   - Sector: For professional cases in certain sectors, for example transport and distribution, there is a greater exposure.

4. **Company**
   - To the mentioned factors another one should de added: the company. Presumably this factor has strong impact on the relationship among all the others: each company defines their particular pricing model.

This information should be used to determine the true market risk, since in theory all companies are measuring the same risk: frequency and average cost. Otherwise, since there is no exchange of tariff models between insurance companies, there will be as many models as the number of companies: the model is statistically different from company to company and that it is not controlled with a simple randomization. The inclusion of this information in the estimation process will be detailed in the next section. A mathematized way, and considering that x has the usual reading of exogenous variables, the model assumes the form:

$$y_i = f_{company_j}(x_i) \tag{4.1}$$

and not

$$y_i = f(company_i, x_i)\ . \tag{4.2}$$

Regarding the levels the question is different since the collection is performed continuously on some key variables. More, there may be some ratios and values derived (the calculation of the power weight is perhaps the most obvious case). It should be noted, however, that the choice of levels must be such as to minimize the information or the variance, leading to an efficient and feasible project. At this point it must be assumed that using a panel of experts, see Barker ([2]), it is possible to minimize/aggregate the number of levels, where it is emphasized that is best to arrange an experiment as a team effort and use the brainstorming technique to scope the entire problem.

5. OPTIMAL DESIGNS

The Completely Randomized Design (CRD) is the simplest form of statistical experimental design. In a CRD the treatments are randomly assigned and the model is linear. It is necessary to check a set of hypotheses, often called classical hypotheses and to estimate the generating process of tariffs by maximum likelihood, in order to obtain a centered model. With classic conditions (linearity in the parameters, random sample, absence of perfect multicollinearity) it is possible to ensure the centering of the maximum likelihood estimators. Thus, in a random sample, as is customary in regression work context, there is a random mechanism which selects the sample within all possible samples.

In experimental design context, adopting the usual notation, as in Graybill ([9]), it is possible to interpret the problem differently. If there is a data generating process (and not a random selection mechanism of samples), the data

structure remains the same, the estimator formula will be the same and centering is guaranteed (strict conditions).

It is thus possible to collect any data to ensure the $\widehat{\boldsymbol{\beta}}$ estimator centering. But centering is not the only prerequisite. Under this assumption of data generating mechanism it is also possible to go further in terms of variance test.

The second condition for a good sample design is to maximize the power of the tests. To ensure that the process is efficient, i.e. that the standard deviation associated with each of the estimates of the betas is minimal, there is the need to examine the beta estimator formulation (to see its derivation see, e.g., Gujarati ([11])):

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{y}\,. \tag{5.1}$$

The $\widehat{\boldsymbol{\beta}}$ variance is given by:

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}\,. \tag{5.2}$$

For an efficient estimator there is the need to have enough number of cases in order to estimate regression and that each of the elements in the diagonal of matrix $(\mathbf{X}'\mathbf{X})^{-1}$ is minimum (assuming no interaction effects). The best way is to guarantee that the matrix $(\mathbf{X}'\mathbf{X})^{-1}$ is only filled in the diagonal — there is no correlation between the different x, (cf. Cramer Rao, see [3]). In such case the matrix $(\mathbf{X}'\mathbf{X})^{-1}$ is orthogonal. So, we have:

$$\widehat{\boldsymbol{\beta}} = \mathbf{I}\mathbf{X}'\mathbf{y}\,. \tag{5.3}$$

In such case the estimator values are easily obtained and will have minimum variance. Note that if $(\mathbf{X}'\mathbf{X})^{-1} \neq \mathbf{I}$ there will be *confounding* and it will not be possible to estimate without an high error associated to the estimated coefficients.

5.1. Optimal Design — functional adjustments

When the sample design and the data analysis are performed, it is possible to obtain a simple linear regression model to determine the importance of factors (through a t-test) and the degree of criticality of their levels (again with t-test, assuming levels as dummy variables), possibly setting the best functional form.

But it is worth exploring the meaning of (4.2) and the need to have a function for company seen in the previous section. Relation (4.2) indicates that the option is to collect and model data for a single company, assuming that each company should have autonomous pricing models. If one considers only two

companies, the model can be described as:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}, \qquad \text{or moreover} \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}\,, \tag{5.4}$$

where

$$\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \mathbf{e} \sim N\left[\begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} \sigma_1 I_T & 0 \\ 0 & \sigma_2 I_T \end{bmatrix} = \mathbf{W} \right], \qquad \mathbf{e} \sim N(\mathbf{0}, \mathbf{W})\,.$$

Another issue arises: Companies are operating in the same market, is it all the information available to capture the tariff model on the market being used? Or even more directly: "The tariff models are autonomous, but are they independent?" In a more mathematized form (applying the same rational Griffiths *et al.* ([10])): And if the mistakes of the different equations, e_1 and e_2, are correlated?

Thus, consider

$$\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \mathbf{e} \sim N\left[\begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} \sigma_{11} I_T & \sigma_{12} I_T \\ \sigma_{21} I_T & \sigma_{22} I_T \end{bmatrix} \right]. \tag{5.5}$$

The idea is that we can estimate the (4.2) per blocks in order to take different functional forms, and perhaps different explanatory variables; but considering (5.5) we can have greater accuracy in forecasting and more power in the tests. The demonstration of these statements follows below, considering just the case of two companies, although the generalization is directly (and it can be confirmed in the Annex to Sec. 17, see Griffiths *et al.* ([10]).

Considering maximum likelihood, it follows that:

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \left[\begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}' \begin{bmatrix} \sigma_{11} I_T & \sigma_{12} I_T \\ \sigma_{21} I_T & \sigma_{22} I_T \end{bmatrix}^{-1} \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \right]^{-1} \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}' \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}. \end{aligned}$$

$E(\mathbf{e}\mathbf{e}') \neq \boldsymbol{\sigma}\,\mathbf{I}$, the usual case does not apply, so: $E(\mathbf{e}\mathbf{e}') = \mathbf{W}$ and $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}'\,\mathbf{W}^{-1}\mathbf{X})^{-1}$.

However, this estimator is not likely to be calculated, since the matrix W is not known; then it must be estimated. In other words, it is necessary to establish the following relationship: $\widehat{\sigma} = \widehat{\mathbf{e}}\,\widehat{\mathbf{e}}'$. Thus:

$$\begin{aligned} \widehat{\widehat{\boldsymbol{\beta}}} &= (\mathbf{X}'\,\widehat{\mathbf{W}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \left[\begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}' \begin{bmatrix} \widehat{\sigma_{11}} I_T & \widehat{\sigma_{12}} I_T \\ \widehat{\sigma_{21}} I_T & \widehat{\sigma_{22}} I_T \end{bmatrix}^{-1} \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \right]^{-1} \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}' \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \end{aligned} \tag{5.6}$$

$$= \left[\begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}' \begin{bmatrix} \widehat{e_1'}\widehat{e_1} I_T & \widehat{e_2'}\widehat{e_1} I_T \\ \widehat{e_1'}\widehat{e_2} I_T & \widehat{e_2'}\widehat{e_2} I_T \end{bmatrix}^{-1} \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \right]^{-1} \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}' \begin{bmatrix} y_1 \\ y_2 \end{bmatrix},$$

with $\mathrm{Cov}(\widehat{\widehat{\beta}}) = (\mathbf{X}' \widehat{\mathbf{W}}^{-1} \mathbf{X})^{-1}$.

Since without other companies the model is not homoscedastic (i.e., $E(\mathbf{e}\,\mathbf{e}') \neq \boldsymbol{\sigma}\,\mathbf{I}$), the different $\widehat{\boldsymbol{\beta}}$ estimators are not centered with minimum efficiency. However $\widehat{\widehat{\boldsymbol{\beta}}}$ is homoscedastic in case of good parametrization. The analysis of t and F tests will be more precise and the experimental design gains more consistency. And in this way it can be captured which are the mechanisms that generate pure data/premiums from different companies.

Evidently the SUR adjustment should be applied before the deconstruction of objective variable and accordingly the correction of the mass point determined already indicated.

6. APPLICATIONS

To better understand the application of this concept, a database that portrays the conditions of the insurance market in 2011 for the Portuguese aggregate car liability coverage and travel assistance was studied. On the computation procedures R software ([21]) was used. The data presented were slightly retouched in order to guarantee the anonymity of the companies under review. The variable "form of payment" was excluded from the analysis in order to create one more element of non-identification of the insurers, impairing pure orthogonality.

The following subsections accompany the classic stages of experimental design.

Stage 1: Identify the factors and levels that define the product

The variables considered were not defined within this article nor by its authors. The discussion of these variables is indicated in Table 1. Each broker must deliver a quotation under a specific scenario for the minimum legal capital requirement plus a minimum coverage for travel assistance.

Stage 2: Optimal Design

The sample was drawn by mystery shopping imposing a minimum number of observations: A standard case was set up for the main factors, and variable levels were changed in five subsamples. Therefore, orthogonality was not guaranteed. The estimated model is well centered, but is not necessarily statistically efficient.

Table 1: Description of factors and levels.

Variables	**Number of Levels**	**Identification of levels**
Gender	2	Male Female
Age	7	19 23 28 35 45 57 67
Claims history	21	0 injury/10 years/15 years 0 injury/15 years/15 years 0 injury/ 2 years/ 2 years 0 injury/ 4 years/ 4 years 0 injury/ 5 years/10 years 0 injury/ 5 years/12 years 0 injury/ 5 years/ 5 years 0 injury/ 5 years/ 6 years 0 injury/ 5 years/ 7 years 0 injury/ 5 years/ 8 years 0 injury/ 7 years/ 9 years 1 injury/ 0 years/ 1 years 1 injury/ 0 years/ 2 years 1 injury/ 0 years/ 3 years 1 injury/ 0 years/ 4 years 1 injury/ 0 years/ 5 years 1 injury/ 1 years/ 5 years 1 injury/ 2 years/ 9 years 1 injury/ 3 years/ 9 years 1 injury/ 4 years/10 years no experience
motor classification	6	Picup truck Light vehicle Commercial vehicle Multi-purpose vehicle Pickup Off-road vehicle
Automobile age	13	{0,1,...,10}; 15 and 20
Region	58	58 Municipalities
Companies	7	Confidential

Stage 3: Collection of information

This database was collected at the beginning of the decade for a specific consulting project. A detailed technical specifications could allow customer identification, project objectives and market conditions. Thus, all data were carefully calibrated so as not to allow companies to determine the target or operating results.

It should be noted that for the realization of information collection resorted in non-exclusive agents. These brokers collected quotations, according to a scenario/profile structured to be simulated beforehand and subsequent recording of information on observation grids.

Care was taken to ensure that the brokers were gathering in municipalities belonging to the working regions.

Stage 4: Analysis

First of all, it should be noted that the model to estimate will run with the following steps:

1. Building a model for insurance and calculation of λ;

2. Estimation of the model by SUR;

3. Mass point correction;

4. Obtaining the best β estimator;

5. Critical analysis of the results and, optionally, repetition of the cycle.

S4.1. Building a model for insurance and calculation of λ

As discussed, the model should include all the variables collected for each company.

For choosing λ it was decided initially to obtain a graphic with the evolution of the likelihood function between -10 and 1, and the limits around zero and in order to include the maximum of each of the functions. The results can be observed in Figure 1 and indicate that the λ which maximize the objective functions are: -3.504, -1.723, -4.055, -3.469, -1.380, -3.276 and -3.024.

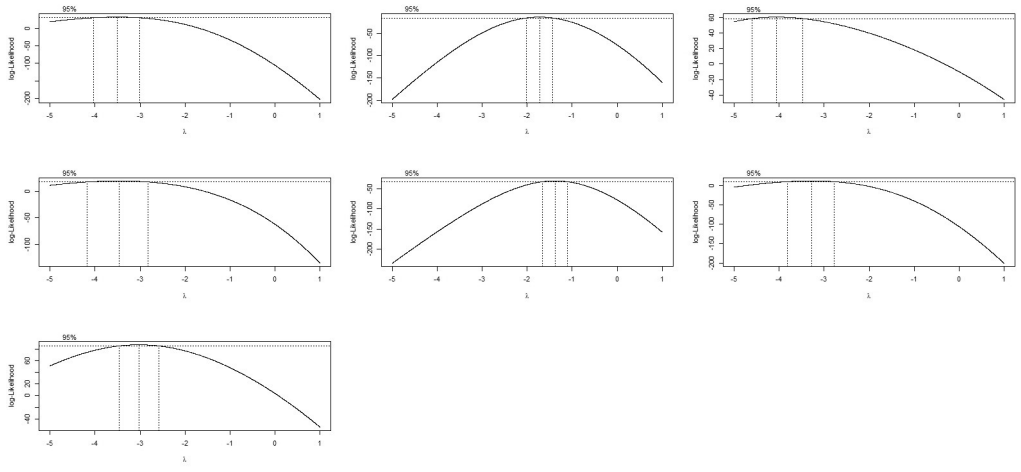

Figure 1: Evolution of the likelihood functions for each of the companies and obtaining λ.

All optimal points move away from intervals with easy interpretation. It should also be noted that apparently the companies use different tariff models reinforcing the idea of estimating each model separately, according to equation (4.2).

When using these lambda values, and estimates the data generating process by ordinary regression and reconstructs the variable ultimate goal, the result is bleak. The choice as accurate lambda ultimately eliminate all variability in the model. It's worth mention in Figure 2 the results with $\lambda < \sim -3.2$ the result is a parallel to the x-axis. The model is therefore estimated using $\lambda = 0$.

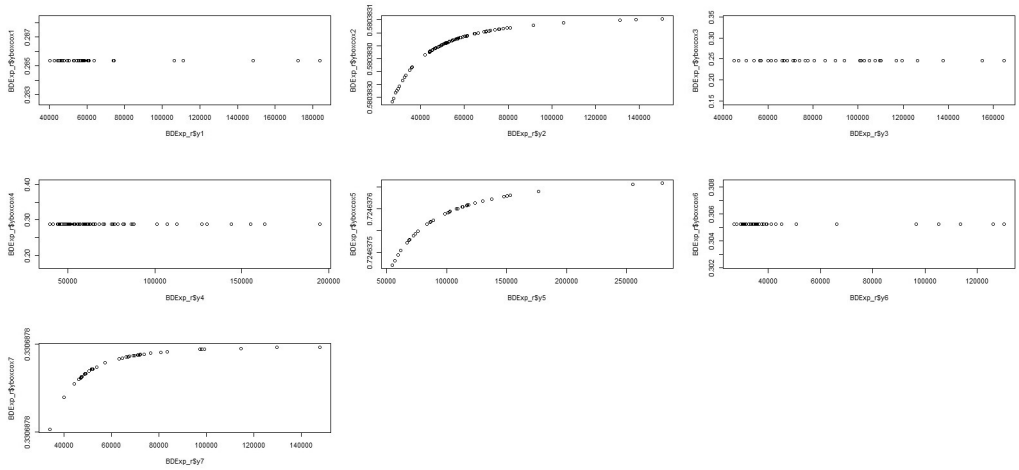

Figure 2: Function of $\widehat{y}$ with Box–Cox that maximizes the likelihood function.

S4.2. Estimation with model SUR

The model estimated by SUR, even with the transformation Box–Cox, shows a strong correlation matrix (qualitative assessment) between models of different companies. Indeed, the correlation between the estimated rates varies between 49% and 93%.

Table 2: Estimated models per enterprise — the error correlation matrix.

	eq1	eq2	eq3	eq4	eq5	eq6	eq7
eq1	1.00	0.28	0.40	0.53	0.64	0.51	0.34
eq2	0.28	1.00	0.60	0.52	0.40	0.32	0.56
eq3	0.40	0.60	1.00	0.56	0.48	0.54	0.38
eq4	0.53	0.52	0.56	1.00	0.52	0.29	0.58
eq5	0.64	0.40	0.48	0.52	1.00	0.49	0.24
eq6	0.51	0.32	0.54	0.29	0.40	1.00	0.01
eq7	0.34	0.56	0.38	0.58	0.24	0.01	1.00

S4.3. Mass point correction

Correction of the mass point has different impacts: in some cases is almost negligible, in others may require a correction of $> 4\%$. In fact, per company it is possible to find the following correction factors: 1.0427, 1.0202, 1.0053, 1.0177, 1.0263, 1.0398 and 1.0079 .

S4.4. Critical analysis of the results and, optionally, repetition of the cycle

In this work only the $R2$ between the final estimated variable and the variable goal is analyzed — indeed to the general objective of this work the main interest is in evaluating the predictive power of the models. So we have the following coefficients $R2$: 0.88021, 0.81354, 0.91175, 0.87484, 0.89378, 0.90491, and 0.8674 — very high values indicating excellent adjustment capacity.

The remaining quality indicators usually calculated on a regression analysis may also be applied. In any case it is interesting to compare the estimated model with the observed pattern. As can be observed in Figure 3, the largest deviation holds mainly thanks to the existence of outliers that were not treated/corrected. Therefore the determination of the final model will be made using the matrix W.

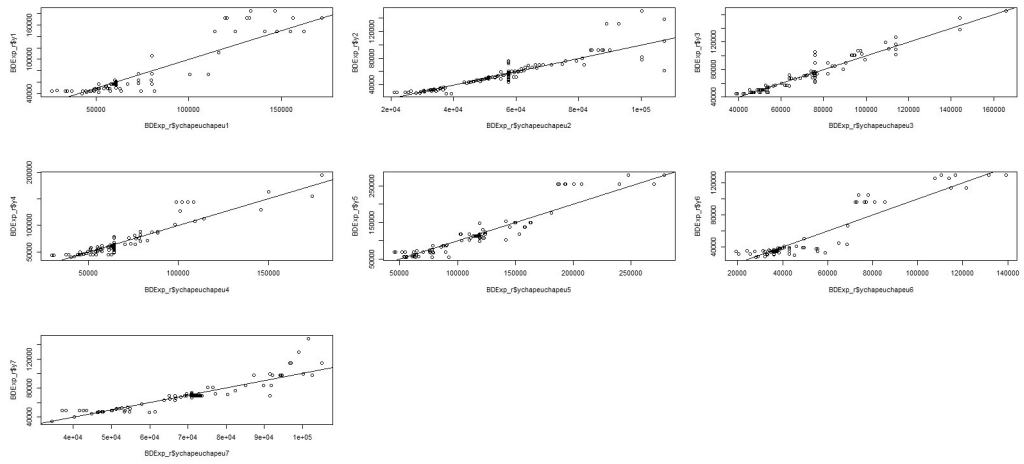

Figure 3: Comparison of SUR model with original data.

7. CONCLUSIONS AND FUTURE RESEARCH

The main purpose of this work was to present a method for collecting and capturing the tariff model for an insurance company and this goal was achieved. An approach to sample design based on the principles of orthogonality was presented as well as the linear regression model, with Box–Cox transformation and point correction for a first analysis. We presented a methodology to integrate information from more than one company and therefore increasing the efficiency of the estimators through a SUR model.

For a future work is the possibility of designing a more complex experimental design model with GLM — Tweedie. This would potential provide a greater adherence to data , specially if one can indicate how to get a rough estimate for the dispersion factor p. It will be interesting to investigate how the Box–Cox model may contribute for an efficient estimation of Tweedie on the determination of p. It will be also interesting to assess the prevalence of the SUR approach in the case of GLM.

ACKNOWLEDGMENTS

The authors acknowledge the valuable suggestions from the referees.

Teresa A. Oliveira and Amílcar Oliveira were partially sponsored by national funds through the Fundação Nacional para a Ciência e Tecnologia, Portugal — FCT under the project UID/MAT/00006/2013 and by the sabbatical scholarships SFRH/BSAB/113669/2015 and SFRH/BSAB/113590/2015, respectively.

REFERENCES

[1] Anderson, D.; Feldblum, S.; Modlin, C.; Schirmacher, D.; Schirmacher, E. and Thandi, N. (2007). *A Practitioner's Guide to Generalized Linear Models — a foundation for theory, interpretation and application*, Third edition, CAS Discussion Paper Program.

[2] Barker, T.B. (1994). *Quality By Experimental Design*, Second edition, Chapman and Hall/CRC.

[3] Bhapkar, V.P. (2014). *Cramer-Rao Inequality*, Wiley StatsRef: Statistics Reference Online, John Wiley & Sons, Ltd.

[4] Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations (with Discussion), *J. R. Statist. Soc. B*, **26**, 211–256.

[5] Nocetti, D. (2005). A model of mental effort and endogenous estimation, *Econ. Bull*, **4**(14), 1–10.

[6] Nocetti, D. (2006). Portfolio Selection with endogenous estimation risk, *Econ. Bull*, **7**(6), 1–9.

[7] Dror, H. and Steinberg, D.M. (2006). Robust Experimental Design for Multivariate Generalized Linear Models, *Technometrics*, **48**(4), 520–529.

[8] Dror, H. and Steinberg, D.M. (2008). Sequential Experimental Designs for Generalized Linear Models, *Journal of the American Statistical Association*, **103**, 288–298.

[9] Graybill, F.A. (1976). Theory and Application of the Linear Model, First edition, *Duxbury Classic Series*.

[10] Griffiths, W.; Hill, R.C. and Judge, G. (1993). *Learning and Practicing Econometrics*, John Wiley and Sons.

[11] Gujarati, D.N. (1995). *Basic Econometrics*, Third edition, McGraw Hill.

[12] Haines, L.M.; Perevozskaya, I. and Rosenberger, W.F. (2003). Bayesian-Optimal Designs for Phase I Clinical Trials, *Biometrics*, **59**, 591–600.

[13] Harrison, M.J. and McCabe, B.P.M. (1979). A Test for Heteroscedasticity based on Ordinary Least Squares Residuals, *Journal of the American Statistical Association*, **74**, Issue 366a, 494–499.

[14] Ivanova, A. and Wang, K. (2004). A Non-Parametric Approach to the Design and Analysis of Two-Dimensional Dose-Finding Trials, *Statistics in Medicine*, **23**, 1861–1870.

[15] Jørgensen, B. (1989). *The Theory of Exponential Dispersion Models and Analysis of Deviance*. In "Lecture notes for short course, School of Linear Models", University of São Paulo, 129 pages. Second Edition, 1992: *Monografias de Matemática*, **51**, IMPA, Rio de Janeiro.

[16] Jørgensen, B. (1987). Exponential dispersion models, *Journal of the Royal Statistical Society*, Series B, **49**(2), 127–162.

[17] Jørgensen, B. and de Souza, M.P. (1994). Fitting Tweedie's compound Poisson model to insurance claims data, *Scand. Actuarial J.*, **1994**, 69–93.

[18] Karvanen, J.; Vartiainen, J.J.; Timofeev, A. and Pekola, J. (2007). Experimental Designs for Binary Data in Switching Measurements on Superconducting Josephson Junctions, *Applied Statistics*, **56**, 167–181.

[19] Mackinnon, H.W. and Davidson, R. (1983). Tests for Model Specification in the Presence of Alternative Hypothesis; Some Further Results, *Journal of Econometrics*, **21**, 53–70.

[20] Osborne, J. (2010). Improving your data transformations: Applying the Box–Cox transformation, *Practical Assessment, Research & Evaluation*, **15**(12). `http://pareonline.net/getvn.asp?v=15&n=12`.

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

[22] Ripley, B.; Venables, B.; Bates, D.M.; Hornik, K.; Gebhardt, A. and Firth, D. (2015). *Package MASS*. `http://www.stats.ox.ac.uk/pub/MASS4`.

[23] Wald, A. (1947). *Sequential Analysis*, John Wiley and Sons, New York.

[24] Woods, D.C.; Lewis, S.M.; Eccleston, J.A. and Russel, K.G. (2006). Designs for Generalized Linear Models With Several Variables and Model Uncertainty, *Technometrics*, **48**, 284–292.

[25] Wooldridge, J. (2003). *Introductory Econometrics: A Modern Approach*, Second edition, Thomson South-Western.

[26] Zellner (1962). Exponential dispersion models, *Journal of the Royal Statistical Society*, Series B, **49**(2), 127–162.

REVSTAT – Statistical Journal
Volume 14, Number 2, April 2016, 193–215

EXTREME VALUE ANALYSIS – A BRIEF OVERVIEW WITH AN APPLICATION TO FLOW DISCHARGE RATE DATA IN A HYDROMETRIC STATION IN THE NORTH OF PORTUGAL

Authors: Helena Penalva
– Department of Economics and Management,
School of Business Administration, and CEAUL,
Polytechnic Institute of Setúbal, Portugal
helena.penalva@esce.ips.pt

Sandra Nunes
– Department of Economics and Management,
School of Business Administration, CEAUL and CMA/FCT/UNL,
Polytechnic Institute of Setúbal, Portugal
sandra.nunes@esce.ips.pt

M. Manuela Neves
– Department of Sciences and Engineering of Biosystems,
Institute of Agronomy, and CEAUL, University of Lisbon, Portugal
manela@isa.ulisboa.pt

Received: October 2015 Revised: February 2016 Accepted: February 2016

Abstract:

- Extreme value theory is dedicated to characterise the behaviour of the extreme observations. The interest is then focused in the tails of the underlying distribution. It is important to test for the adequate shape of the tail, because it influences the estimation of parameters of extreme or even rare events. The aim of this work is to present a brief overview on several tests and parameter estimation procedures available in the literature. They will be applied to daily mean flow discharge rate values in the hydrometric station of Fragas da Torre in the river Paiva, collected from 1946/47 to 2005/06.

Key-Words:

- *extreme value theory; parametric and semi-parametric estimation; statistical choice of the tail; statistical tests.*

AMS Subject Classification:

- 62G32, 62G10, 62E20, 62P12.

1. INTRODUCTION

Extreme value theory (EVT) is concerned with the stochastic behaviour of extremes values. In EVT we need to deal with events that are more extreme than any that have already been observed. The question is how to make inference beyond the sample data. Obviously, statistical inference can be deduced only from those observations which are extreme in some sense.

There are a few parameters whose estimation is of major importance. The extreme value index (EVI), which is directly related with the heaviness of the right tail of the underlying distribution of the data, is a crucial parameter. It influences the estimation of other parameters of extreme values, such as, high quantiles of probability $1-p$, with p "small", i.e., the *high levels* usually designed by the *return levels* associated with the *return periods*, $1/p$, i.e. the expected waiting time between independent exceedances of a specific high level.

In all areas of application it is of major importance to use adequate and accurate statistical methods. The R software (R Development Core Team, [55]) is an open source environment that incorporates in its base a huge amount of statistical packages built and made freely available by the scientific community.

Penalva *et al.* ([52], [53]) have illustrated the application of some procedures of modelling and estimating in EVT, under a parametric framework. Some R packages were explained and some data sets were considered. The *Block Maxima* (BM), the *Peaks Over Threshold* (POT) and the *k Largest Observations* (k-LO) methods were described and applied. Different methodologies for parameter estimation were also considered. In Neves *et al.* ([50]) R procedures for the semi-parametric estimation in EVT have been presented and discussed. A real data set of daily mean flow discharge rate values from the hydrometric station of Fragas da Torre in the river Paiva during the years from 1946/47 to 1996/97 was considered.

In this paper parametric and semi-parametric frameworks are briefly reviewed. In both cases EVT theory relies on certain assumptions that should be validated when dealing with an application. Regardless the framework followed statistical inference will be improved if one makes the choice of the most adequate tail previously. A brief overview of some testing procedures for the so-called *extreme value condition* and for the *statistical choice of the tail* will be given. An application to a larger data set than the one mentioned above will be performed, now considering the years from 1946/47 to 2005/06.

Section 2 provides a brief review on the basic notions in EVT. In Section 3 parametric and semi-parametric statistical approaches in EVT are summarized and the main statistical methods for the estimation of parameters are described.

Section 4 and 5 are dedicated to a brief reference to testing issues and finally Section 6 presents a case study and the application of some of the methods described in the previous sections, still giving some attention to the main packages available in R software for the extreme value analysis.

2. PRELIMINARIES IN EVT LIMITING LAWS

Classic theory of extremes is concerned with the limiting behaviour of the maximum $M_n := \max\big(X_1, ..., X_n\big)$ or the minimum $m_n := \min\big(X_1, ..., X_n\big)$, as $n \to \infty$, of a sample $(X_1, ..., X_n)$ of independent and identically distributed (i.i.d.) or possibly stationary, weakly dependent, random variables with unknown distribution function (d.f.) F. It is well known that in those conditions the distribution of the maximum M_n is $F^n(\cdot)$, and also for the minimum m_n, i.e., $1 - [1 - F(\cdot)]^n$. However the d.f. F^n is of little help in practice since F is itself unknown and should F be misspecified, this can lead to large errors in the distribution of the maximum.

First results in EVT date back to Fréchet ([27]), Fisher and Tippet ([22]), Gumbel ([39]) and von Mises ([60]), but Gnedenko ([30]) and de Haan ([41]) have solved the problems related with the asymptotic behaviour of statistical extremes, giving conditions for the existence of sequences $\{a_n\} \in \mathbb{R}^+$ and $\{b_n\} \in \mathbb{R}$ such that

$$\lim_{n\to\infty} P\left(\frac{M_n - b_n}{a_n} \le x\right) = \lim_{n\to\infty} F^n(a_n x + b_n) \;=\; EV_\xi(x) \qquad \forall x \in \mathbb{R}\,, \tag{2.1}$$

where EV_ξ is a nondegenerate distribution function.

This function, known as the Extreme Value d.f., is usually denoted by EV_ξ and is given by

$$EV_\xi(x) = \begin{cases} \exp\big\{-[1+\xi x]^{-1/\xi}\big\}\,, & 1+\xi x > 0\,, \quad \text{if} \;\; \xi \neq 0\,, \\ \exp\big\{-\exp[-x]\big\}\,, & x \in \mathbb{R}\,, \quad \text{if} \;\; \xi = 0\,, \end{cases} \tag{2.2}$$

where $\xi \in \mathbb{R}$ is the shape parameter.

Definition 2.1. We say that F is in the domain of attraction (for maxima) of EV_ξ and write $F \in \mathcal{D}_{\mathcal{M}}(EV_\xi)$, whenever (2.1) holds.

As a consequence of the existence of that limit, when $n \to \infty$ we may consider the approximation, $P[M_n \le x] = F^n(x) \approx EV_\xi\big((x - b_n)/a_n\big)$.

The EV_ξ incorporates the three (Fisher–Tippett) families: the *Gumbel family*, that is the limit for exponential tailed distributions, $\Lambda(x) = EV_0(x) =$

$\exp\big(-\exp(-x)\big)$, $x \in \mathbb{R}$, $\xi = 0$; the *Fréchet family*, that is the limit for heavy tailed distributions, $\Phi_\alpha(x) = EV_{1/\alpha}\big(\alpha(x-1)\big) = \exp(-x^{-\alpha})$, $x > 0$, $\xi = 1/\alpha > 0$ and the *Weibull family*, that is the limit for short tailed distributions, $\Psi_\alpha(x) = EV_{-1/\alpha}\big(\alpha(x+1)\big) = \exp\big(-(-x)^\alpha\big)$, $x < 0$, $\xi = -1/\alpha < 0$.

The shape parameter, ξ, is the so-called *extreme value index* (EVI), it is the primary parameter in EVT and it measures the heaviness of the right-tail, $\overline{F} := 1 - F$. If $\xi = 0$, the right tail is of an exponential type; if $\xi > 0$, the right tail is heavy, it is of a negative polynomial type and if $\xi < 0$, the right tail is short and F has a finite right endpoint.

The limit distribution family, EV_ξ in (2.2), seems to present some difficulties due to the normalizing constants, $\{a_n\}$ and $\{b_n\}$ be unknown. However that limit can be interpreted, for sufficiently large n, as

$$P\left(\frac{M_n - b_n}{a_n} \le x\right) \approx EV_\xi(x) \iff P\big(M_n \le x\big) \approx EV_\xi\left(\frac{x - b_n}{a_n}\right). \tag{2.3}$$

We can further consider location and scale parameters, $\lambda \in \mathbb{R}$ and $\delta \in \mathbb{R}^+$, respectively, in the EV_ξ d.f., denoting it by $EV_\xi(x; \lambda, \delta) \equiv EV_\xi\big((x-\lambda)/\delta\big)$, so the constants in (2.3) can incorporate this location/scale version.

Instead of just considering the maximum value of a sample as an extreme value, we may consider all the observations, X_i, above a high level or threshold, u, established previously, as extremes. The differences $X_i - u$, are called exceedances over that threshold. Balkema and de Haan ([1]) and Pickands ([54]) proved that if $F \in \mathcal{D}_{\mathcal{M}}(EV_\xi)$, see Definition (2.1), then for large enough u, $Y = \big((X-u)\,|\,X > u\big)$ is approximately the generalized Pareto (GP) d.f.,

$$H_\xi(y) = 1 - \big(1 + \xi\, y/\widetilde{\delta}\big)^{-1/\xi}, \quad \text{for } y > 0 \text{ and } \big(1 + \xi\, y/\widetilde{\delta}\big) > 0\,, \tag{2.4}$$

where ξ is the shape parameter, equal to that of the corresponding EV distribution, and the scale parameter $\widetilde{\delta} = \delta + \xi(u - \lambda)$, where λ is the location parameter in the EV d.f.. The reciprocal of the stated above is also true.

We can also consider the joint distribution of the k top order statistics. More specifically, if X is a random variable with d.f. F belonging to the domain of attraction of an EV d.f. then, for fixed k, the limiting distribution, as $n \to \infty$, of the k-dimensional random vector, suitably normalized by constants $\{a_n\} \in \mathbb{R}^+$ and $\{b_n\} \in \mathbb{R}$, $\left(\frac{M_n^{(1)} - b_n}{a_n}, ..., \frac{M_n^{(k)} - b_n}{a_n}\right)$, where $M_n^{(k)} \equiv X_{n-k+1:n} :=$ k largest of $\{X_1, ..., X_n\}$ and the joint probability density function is given by

$$g(w_1, ..., w_k) = EV_\xi(w_k) \prod_{i=1}^{k} \frac{ev_\xi(w_i)}{EV_\xi(w_i)}\,, \qquad w_1 > \cdots > w_k\,, \tag{2.5}$$

with $EV_\xi(w)$ defined in (2.2) and where $ev_\xi(w) = \frac{\partial EV_\xi(w)}{\partial w}$ is the probability density function of the EV model. This model is known as the Multivariate-EV_ξ model, also known as the extremal process, Dwass ([20]).

3. MODELLING AND ESTIMATING IN EVT

Statistical inference in EVT is based on extreme observations, however there are different ways of defining such observations leading to the application of different models. Classical parametric approaches for modelling and estimation were the first to appear, based on limiting distributions defined in the previous section. In the late seventies, estimation procedures in EVT began to be performed on a semi-parametric approach based on probabilistic asymptotic results in the tail of the unknown distribution.

3.1. Parametric statistical approaches and estimation

The first approach for modelling extremes is the so-called *Block Maxima* (BM), *Annual Maxima* or *Gumbel's* approach, Gumbel ([40]). 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_ξ 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.

However, in many applications there is no natural way of defining blocks of observations. Besides it may occur that the maximum within a block has a lower value than some values in another block. Thus, some extreme values contained in a block may not be included in data for the analysis. So, BM methodology may not be the best method for studying the behaviour of extreme values.

Another methodology consists of setting a high level or threshold, u, and defining as extremes all the observations above that value. The idea is then to fit the model referred to in (2.4) to the *excesses* over such a high level, u. This method, known as *Peaks Over Thresholds* (POT) method, uses relevant information that can be lost by the BM method. Details of this procedure can be seen in Davison ([15]), Davison and Smith ([16]) and Smith ([57]).

Another approach, in some sense parallel to the previous one, consists of considering the k top order statistics of the sample. In this methodology, usually denoted as the *k-Largest Observations* (k-LO), inference can be done when the size n of the sample is large and k fixed, based on the multivariate structure of the k top order statistics, referred to in (2.5). This model was developed and studied by Weissman ([61]) and Gomes ([31]).

Note that the use of POT method needs the choice of a suitable threshold, u, what is equivalent to the choice of the number, k, of upper order statistics to be taken on the k-LO approach.

We can also think of combining the BM and the k-LO approaches. In each of the m sub-samples, we can collect a few top order statistics and, in this case, inference is based on the m random k-dimensional vectors. These m random k-dimensional vectors, after being suitably normalized by constants $\{a_n\} \in \mathbb{R}^+$ and $\{b_n\} \in \mathbb{R}$, are well modelled by the Multivariate-EV_ξ defined in (2.5). This methodology is known as *Multidimensional-EV_ξ* approach.

For estimating extreme value parameters several procedures have been proposed: (i) graphical methods; (ii) moment-based methods and (iii) likelihood methods. All these procedures have been extensively studied and applied in classical parametric modelling. In this work we will review parameter estimation using the maximum likelihood (ML) method , the profile likelihood (PL) method and the probability weighted moments (PWM) method.

Difficulties that arose with the "regularity conditions" for the maximum likelihood estimation were solved by Smith ([58]), who showed that the usual property of asymptotic normality holds provided the extreme value parameter ξ is larger than -0.5. Recently, Zhou ([62], [63]) showed that the ML estimators verify the property of asymptotic normality for $\xi > -1$. This condition, that is not verified for very light tailed distributions, is satisfied for most environmental applications.

The asymptotic normality, that would allow to obtain confidence intervals, is not very accurate because the normal approximation to the true sampling distribution of the estimator is rather poor. An alternative, and usually more accurate method of estimation is based on the profile likelihood function. Given a parameter vector $\boldsymbol{\theta}$ the *profile log-likelihood* function of the component θ_i is defined as $\log L_p(\theta_i) := \max_{\boldsymbol{\theta}_{-i}} \log L(\theta_i, \boldsymbol{\theta}_{-i})$ where $\boldsymbol{\theta}_{-i}$ denotes a vector with all components of vector $\boldsymbol{\theta}$ excluding θ_i. For each value of θ_i, the profile log-likelihood is defined as the maximized log-likelihood with respect to the other components of the parameter vector $\boldsymbol{\theta}$.

So, for example, for the estimation of ξ in the EV model,

$$\log L_p(\xi) := \max_{\lambda,\delta\,|\,\xi} \log L(\lambda, \delta, \xi) \; .$$

Under suitable regularity conditions, see Beirlant *et al.* ([3]), for large n, the *deviance function* is:

$$D_p(\xi) := 2\left\{\log L(\widehat{\lambda}, \widehat{\delta}, \widehat{\xi}) - \log L_p(\xi)\right\} \; \dot{\sim} \; \chi^2_{(1)} \; ,$$

where $\widehat{\lambda}, \widehat{\delta}$ and $\widehat{\xi}$ are the maximum likelihood estimators of λ, δ and ξ, respectively. This property is used to obtain the $(1-\alpha)\times 100\%$ confidence interval for the parameters of the underlying distribution. Particularly, for a singular component, for example ξ, the $(1-\alpha)\times 100\%$ confidence interval is $\left\{\xi\colon D_p(\xi) \le q_{1-\alpha}\right\} =$

$\left\{\xi\colon \log L_p(\xi) \geq \log L(\widehat{\lambda},\widehat{\delta},\widehat{\xi}) - \frac{q_{1-\alpha}}{2}\right\}$, where $q_{1-\alpha}$ is the $(1-\alpha)$ quantile of $\chi^2_{(1)}$. Therefore the profile log-likelihood ratio statistic

$$-2\log\Lambda \;=\; -2\log\left\{\frac{L_p(\xi_0)}{L_p(\widehat{\xi})}\right\} \;=\; 2\left\{\log L_p(\widehat{\xi}) - \log L_p(\xi_0)\right\},$$

to test $H_0\colon \xi = \xi_0$ versus $H_1\colon \xi \neq \xi_0$ has, under the hypothesis H_0, asymptotic distribution $\chi^2_{(1)}$, when $n \to \infty$. H_0 is rejected at a level of significance α if $-2\log\Lambda > q_{1-\alpha}$, see Coles ([13]) and Beirlant *et al.* ([3]) for more details.

The probability-weighted moments (PWM) (Greenwood *et al.*, [38]) of a random variable X, with d.f. F are defined as

$$M_{p,r,s} \;:=\; E\left\{X^p\left[F(X)\right]^r\left[1-F(X)\right]^s\right\}, \qquad p,r,s \in \mathbb{R}\,.$$

For the *EV* d.f., these moments were extensively studied by Hosking *et al.* ([44]). Considering a random sample $(X_1,...,X_m)$ from a *EV* population, the PWM estimator, $(\widehat{\lambda},\widehat{\delta})$, when $\xi = 0$, is the solution of the system of equations:

$$\begin{cases} \widehat{M}_{1,0,0} = \lambda + \delta\,\Gamma'(1) \\ 2\,\widehat{M}_{1,1,0} - \widehat{M}_{1,0,0} = \log 2\delta \end{cases} \qquad \text{where} \quad \widehat{M}_{1,r,0} = \frac{1}{m}\sum_{i=1}^{m}\left(\prod_{l=1}^{r}\frac{i-l}{m-l}\right)X_{i:m}\,,$$

with $X_{1:m} \leq X_{2:m} \leq \cdots \leq X_{m:m}$ the ascending order statistics associated with the random sample $(X_1, X_2, ..., X_m)$.

For $0 < \xi < 1$, we can obtain the PWM estimator, $(\widehat{\lambda},\widehat{\delta},\widehat{\xi})$ solving the equations system,

$$\begin{cases} \widehat{M}_{1,0,0} \;=\; \lambda - \dfrac{\delta}{\xi}\big(1-\xi(1-\xi)\big) \\ 2\,\widehat{M}_{1,1,0} - \widehat{M}_{1,0,0} \;=\; \dfrac{\delta}{\xi}\,\xi(1-\xi)\,(2^{\xi}-1) \\ \dfrac{3\,\widehat{M}_{1,2,0} - \widehat{M}_{1,0,0}}{2\,\widehat{M}_{1,1,0} - \widehat{M}_{1,0,0}} \;=\; \dfrac{3^{\xi}-1}{2^{\xi}-1} \end{cases}.$$

Also in this method the asymptotic normality for the PWM estimator (λ,δ,ξ) holds provided that $\xi < 0.5$ and $m \to \infty$ (see Beirlant *et al.*, [3]).

3.2. Semi-parametric statistical framework and EVI estimation

In the late seventies estimation in EVT began to be performed in a semi-parametric approach. Here it is not necessary to fit a specific parametric model, dependent upon a location, scale and shape parameters, but only assume that the underlying distribution function F belongs to $\mathcal{D}_{\mathcal{M}}(EV_\xi)$, for an appropriate value of ξ in specific sub-domain of $\mathcal{D}_{\mathcal{M}}(EV_\xi)$, being ξ the primordial parameter to be

estimated. Estimates are based on the k top order statistics in the sample, or on the excesses over a high random threshold, u. For the consistence of the estimators we need to work with an intermediate sequence $k \equiv k_n$, i.e., $k \equiv k_n \to \infty$ and $k/n \to 0$ as $n \to \infty$.

In this framework several EVI estimators have been proposed. We will refer to the classical ones, such as, the Hill estimator ([43]), the Moment estimator, Dekkers *et al.* ([17]), the Generalized Hill estimator, introduced in Beirlant *et al.* ([4]) and studied later in Beirlant *et al.* ([2]), the Mixed Moment estimator, Fraga Alves *et al.* ([26]) and also a recent estimator of reduced bias and minimum variance, (MVRB), Caeiro *et al.* ([10]). A family of estimators based on the logarithm of the *mean of order p* (MOP) of $X_{n-i-1:n}/X_{n-k:n}$, $1 \le i \le k < n$, has been very recently proposed by Brilhante *et al.* ([7]). See also other related estimators such as the harmonic mean estimator introduced in Beran *et al.* ([5]) and a family of estimators introduced in Paulauskas and Vaiciulis ([51]).

Let $X_{1:n} \le X_{2:n} \le \cdots \le X_{n:n}$ the ascending order statistics associated with the random sample $(X_1, X_2, ..., X_n)$, and for $r \ge 1$ let us define

$$L_{k,n}^{(r)} := \frac{1}{k} \sum_{i=1}^{k} \left[1 - \frac{X_{n-k:n}}{X_{n-i+1:n}} \right]^r \quad \text{and} \quad M_{k,n}^{(r)} := \frac{1}{k} \sum_{i=1}^{k} \left[\ln \frac{X_{n-i+1:n}}{X_{n-k:n}} \right]^r . \tag{3.1}$$

Among the aforementioned estimators we will consider:

The Hill estimator ($\xi > 0$)

$$\widehat{\xi}_{k,n}^{H} := M_{k,n}^{(1)} , \qquad k = 1, 2, ..., n-1 ; \tag{3.2}$$

The Moments estimator ($\xi \in \mathbb{R}$)

$$\widehat{\xi}_{k,n}^{M} := M_{k,n}^{(1)} + 1 - \frac{1}{2} \left(1 - \frac{\left(M_{k,n}^{(1)}\right)^2}{M_{k,n}^{(2)}} \right)^{-1} , \qquad k = 1, 2, ..., n-1 ; \tag{3.3}$$

The Generalized Hill estimator ($\xi \in \mathbb{R}$)

$$\widehat{\xi}_{k,n}^{GH} := M_{k,n}^{(1)} + \frac{1}{k} \sum_{i=1}^{k} \left[\ln \frac{M_{i,n}^{(1)}}{M_{k,n}^{(1)}} \right] , \qquad k = 1, 2, ..., n-1 ; \tag{3.4}$$

The Mixed Moment estimator ($\xi \in \mathbb{R}$)

$$\widehat{\xi}_{k,n}^{MM} := \frac{\widehat{\varphi}_{k,n} - 1}{1 + 2 \min(\widehat{\varphi}_{k,n} - 1, 0)} , \qquad k = 1, 2, ..., n-1 , \tag{3.5}$$

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

The **MVRB** estimators, Caeiro *et al.* ([10]) have revealed a better performance than the classical estimators in the context of heavy tails ($\xi > 0$). This class of estimators has the functional form

$$\overline{\widehat{\xi}^{H}_{k,n}}(\widehat{\beta},\widehat{\rho}) := \widehat{\xi}^{H}_{k,n}\Big(1-\widehat{\beta}(n/k)^{\widehat{\rho}}/(1-\widehat{\rho})\Big)\,, \tag{3.6}$$

with $\widehat{\xi}^{H}_{k,n}$ the Hill estimator and $(\widehat{\beta},\widehat{\rho})$ consistent estimators of second order parameters $(\beta,\rho) \in (\mathbb{R},\mathbb{R}^{-})$. About reduced bias estimation, we may also refer to Gomes *et al.* ([36]), Gomes *et al.* ([33]) and Caeiro *et al.* ([9]), among others.

For the estimation of ρ we consider a particular member of a class of estimators introduced in Fraga Alves *et al.* ([24]). This class, parametrized in a control parameter $\tau \in \mathbb{R}$, which here we will take as $\tau = 0$, see Gomes *et al.* ([37]), is defined as: $\widehat{\rho}(k) \equiv \widehat{\rho}_0(k) := \min\Big(0,\, \frac{3\left(T^{(0)}_n(k)-1\right)}{T^{(0)}_n(k)-3}\Big)$, being $T^{(0)}_n(k)$ defined as

$$T^{(0)}_n(k) := \left[\ln\Big(M^{(1)}_{k,n}\Big) - \frac{1}{2}\ln\Big(M^{(2)}_{k,n}/2\Big)\right] \Big/ \left[\frac{1}{2}\ln\Big(M^{(2)}_{k,n}/2\Big) - \frac{1}{3}\ln\Big(M^{(3)}_{k,n}/6\Big)\right],$$

with $M^{(j)}_{k,n}(k),\ j = 1,2,3$, defined above.

For the estimation of the second order scale parameter, β, we will consider

$$\widehat{\beta}_{\widehat{\rho}}(k) := \left(\frac{k}{n}\right)^{\widehat{\rho}} \Big[d_{\widehat{\rho}}(k)\, D_0(k) - D_{\widehat{\rho}}(k)\Big] \Big/ \Big[d_{\widehat{\rho}}(k)\, D_{\widehat{\rho}}(k) - D_{2\widehat{\rho}}(k)\Big]\,,$$

with $\widehat{\rho} = \widehat{\rho}_0(k)$, $d_\alpha(k) := \frac{1}{k}\sum_{i=1}^{k}\big(\frac{i}{k}\big)^{-\alpha}$ and $D_\alpha(k) := \frac{1}{k}\sum_{i=1}^{k}\big(\frac{i}{k}\big)^{-\alpha} U_i$, for $\alpha \le 0$, with $U_i := i\left[\ln(X_{n-i+1:n}/X_{n-i:n})\right],\ 1 \le i \le k$.

In order not to have an increase in the variance of the estimator $\overline{\widehat{\xi}^{H}_{k,n}}$, estimators $\widehat{\rho}_0(k)$ and $\widehat{\beta}_{\widehat{\rho}}(k)$ must be calculated at $k = k_1$, with $k_1 = \lfloor n^{1-\epsilon} \rfloor$, $\epsilon = 0.001$, see Gomes and Martins ([35]), Gomes *et al.* ([33]) and Caeiro *et al.* ([9]), for more details. Alternative estimators for β can be seen in Caeiro and Gomes ([8]) and Gomes *et al.* ([34]).

4. TESTING EXTREME VALUE CONDITIONS

In any of the aforementioned procedures it is assumed that the underlying d.f. F belongs to $\mathcal{D}_{\mathcal{M}}(EV_\xi)$, for an appropriate value of ξ, or it is in a specific sub-domain of $\mathcal{D}_{\mathcal{M}}(EV_\xi)$. This condition is called the *extreme value condition* and is not always fulfilled. So, before performing an application, it is important to check whether the extreme value condition is reasonable for a data set or not. So, we must test the hypothesis:

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

Dietrich *et al.* ([18]) proposed the E, PE tests (if we assume $\xi \geq 0$) and Drees *et al.* ([19]) proposed the T test (assuming $\xi > -1/2$).

Let $X_1, X_2, ..., X_n$ be independent random variables with d.f. F and suppose that some additional second order conditions hold, then for $\eta > 0$ the corresponding test statistics are:

$$E_n := k \int_0^1 \left(\frac{\log X_{n-\lfloor kt \rfloor,n} - \log X_{n-k,n}}{\widehat{\xi}_+} - \frac{t^{-\widehat{\xi}_-} - 1}{\widehat{\xi}_-} \left(1 - \widehat{\xi}_-\right) \right)^2 t^\eta \, dt \,, \tag{4.1}$$

$$PE_n := k \int_0^1 \left(\frac{\log X_{n-\lfloor kt \rfloor,n} - \log X_{n-k,n}}{\widehat{\xi}_+} + \log t \right)^2 t^\eta \, dt \,, \tag{4.2}$$

$$T_n := k \int_0^1 \left(\frac{n}{k} \overline{F}_n \left(\widehat{a}_{n/k} \, \frac{x^{-\widehat{\xi}} - 1}{\widehat{\xi}} + \widehat{b}_{n/k} \right) - x \right)^2 x^{\eta - 2} \, dx \,, \tag{4.3}$$

where the estimates for ξ_+ and ξ_- are obtained through the moment estimators in Dekkers *et al.* ([17]), and k is again an intermediate sequence, $k = k_n \to \infty$, $k/n \to 0$ and $k^{1/2} A(n/k) \to 0$ as $n \to \infty$. A is related to the second order condition. Hüsler and Li ([45]) present an algorithm for testing H_0 using the test statistic E_n in (4.1). They have carried out an extensive simulation study with guidelines for obtaining the value of η and have provided tables of critical values. See also Neves and Fraga Alves ([48]) for a description of those tests.

5. STATISTICAL CHOICE OF EXTREME DOMAINS OF ATTRACTION — SEMI-PARAMETRIC APPROACH

In a semi-parametric framework, ξ is the primordial parameter since determines the shape of the tail of the underlying distribution function F. A negative value for ξ is associated to the Weibull domain of attraction in which all the d.f.'s are short tailed with finite right endpoint. If $\xi > 0$ we have the Fréchet domain of attraction to which the heavy tailed d.f.'s with polynomially decaying tail belong. The case of $\xi = 0$ is particularly important, due to the simplicity of inference, within the Gumbel domain which contains a great variety of d.f.'s with an exponential tail having finite right end point or not. Whenever we intend to perform a statistical inference in extreme values we should look for the most adequate procedures according to the domain of attraction selected. Therefore, it is of great benefit to test the Gumbel domain against the Fréchet or Weibull domains. The hypothesis to test is:

$$H_0 \colon F \in \mathcal{D}_{\mathcal{M}}(EV_0) \quad \textit{vs.} \quad H_1 \colon F \in \mathcal{D}_{\mathcal{M}}(EV_\xi)_{\xi \neq 0} \;; \tag{5.1}$$

or *versus* the one-sided alternatives $F \in \mathcal{D}_{\mathcal{M}}(EV_\xi)_{\xi<0}$ or $F \in \mathcal{D}_{\mathcal{M}}(EV_\xi)_{\xi>0}$.

Several tests have been proposed in the literature, among which we can mention Galambos ([28]), Castilho *et al.* ([12]), Hasofer and Wang ([42]), Falk ([21]), Fraga Alves and Gomes ([23]) and Correia and Neves ([14]), that have proposed a slight modification of the Hasofer and Wang statistic, Marohn ([46, 47]), Fraga Alves ([25]) and Segers and Teugels ([56]). More recently Brilhante ([6]) derived a resistant and robust test for the exponential *versus* the generalized Pareto, Neves and Fraga Alves ([48]) introduced three tests statistics based on the reformulation of the Hasofer and Wang statistic. Those tests were later studied in Neves and Fraga Alves ([49]). Castillo *et al.* ([11]) provided a test based on the properties of the *coefficient of variation*.

In this work the tests introduced in Neves and Fraga Alves ([48]) will be considered. The statistics for testing (5.1) are based on the k excesses over the $(n-k)$-th ascending intermediate order statistic $X_{n-k:n}$, Thus, under the null hypothesis of Gumbel domain of attraction and further assuming: (i) second order conditions on the upper tail of F and (ii) the intermediate sequence $k \equiv k_n$, such that $k^{1/2}A(n/k) \to 0$ as $n \to \infty$ where A is related to the second order condition, Neves and Fraga Alves ([48]) have defined the following tests:

The ratio-test

$$R_n^* := \frac{X_{n:n} - X_{n-k:n}}{\frac{1}{k}\sum_{i=1}^{k}\big(X_{n-i+1:n} - X_{n-k:n}\big)} - \log k \xrightarrow[n\to\infty]{d} \Lambda \ ; \tag{5.2}$$

The GT-test

$$G_n(k) := \frac{\frac{1}{k}\sum_{i=1}\big(X_{n-i+1:n} - X_{n-k:n}\big)^2}{\left(\frac{1}{k}\sum_{i=1}^{k} X_{n-i+1:n} - X_{n-k:n}\right)^2} \ , \tag{5.3}$$

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

The HW-test

$$W_n(k) := \frac{1}{k}\left[1 - \frac{G_n(k) - 2}{1 + \big(G_n(k) - 2\big)}\right], \tag{5.4}$$

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

where Λ is a Gumbel random variable.

The null hypothesis in (5.1) is rejected if $T_n^* < \chi_{\alpha/2}$ or $T_n^* > \chi_{1-\alpha/2}$, where T^* has to be replaced by R^*, G^* or W^* and χ_p is the p probability quantile of the corresponding distribution.

If we are interested in the one-sided tests, and being χ_p the p probability quantile of the corresponding distribution, the critical regions for:

Gumbel vs Weibull domain of attraction are:

$$R_n^*(k) < \chi_\alpha \ , \qquad G_n^*(k) < \chi_\alpha \ , \qquad W_n^*(k) > \chi_{1-\alpha} \ ; \tag{5.5}$$

Gumbel vs Fréchet domain of attraction are:

$$R_n^*(k) > \chi_{1-\alpha} \ , \qquad G_n^*(k) > \chi_{1-\alpha} \ , \qquad W_n^*(k) < \chi_\alpha \ . \tag{5.6}$$

As an illustration of the methodologies reviewed in the previous sections and also for showing some functions available in the R software, a real data set will be studied in the next section.

6. A CASE STUDY: DAILY MEAN FLOW DISCHARGE RATE

Here we will focus our attention on the estimation of the EVI. Packages and/or functions available in the R environment will be used and mentioned. R software contains already a large number of packages with several functions for modelling extreme data, such as `evd`, `ismev`, `evir`, `POT`, `fExtremes`, `evdbayes`, `copula`, `SpatialExtremes`, among others. Gilleland *et al.* ([29]) give an excellent software review for extreme value analysis. They describe and compare packages available in R with other software.

6.1. A preliminary data analysis

Our data set consists of daily mean flow discharge rate in the hydrometric station of Fragas da Torre in the river Paiva. The source of this river is in the Serra de Leomil, in the north of Portugal, it is an effluent of the river Douro, with a watershed area of approximately 700 Km. More precisely the data set studied is the daily mean flow discharge rate values (m^3/s) from 1 October 1946 to 30 September 2006, collected from the "SNIRH: Sistema Nacional de Informação dos Recursos Hídricos" and the interest is to analyse the extreme values.

After some previous graphical analyses on the empirical tail behaviour of the different months showing the occurrence of the maximum values, advices of hydrologists and taking into account a previous work that considered a few initial years of these data, Gomes ([32]), only the months from November until April were used in each year. We had then a total of 10860 daily mean flow discharge rate values. The results of a preliminary graphical and descriptive analysis are shown in Figure 1 and Table 1.

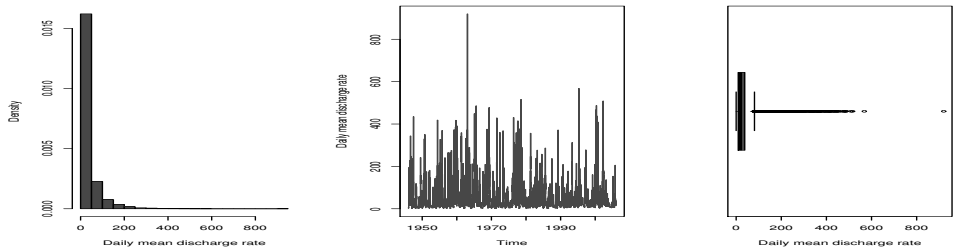

Figure 1: Histogram (left); chronogram (center) and boxplot (right).

Table 1: Basic descriptive statistics for the data.

n	Min	1st Quart.	Median	Mean	3rd Quart.	Max	St Dev.	Skew.	Kurt.
10860	0	9.20	17.30	34.83	38.00	920.00	50.92	4.15	27.31

The stationarity was also studied by the Augmented Dickey–Fuller Test through the function `adf.test()`, available in the *package* `tseries`. The boxplot, the histogram and the descriptives statistics, in particular the skewness = 4.15 and the kurtosis = 27.31 indicate a tail heavier than the normal one.

6.2. Testing extreme value conditions

Following the brief introduction given in section 4, we will use here the test E, Dietrich *et al.* ([18]) and Hüsler and Li ([45]). The function `MTestEVC1d()` in the *package* `TestEVC1d` gave the results shown in Figure 2. We observe that the values

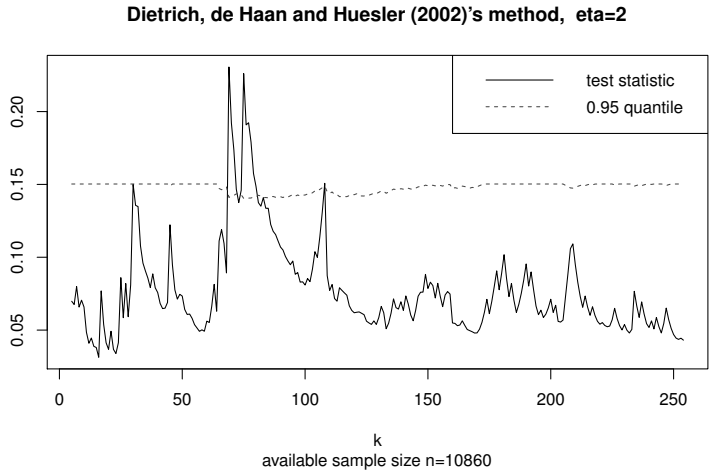

Figure 2: Sample path of E statistic.

of the test statistic E are smaller than the corresponding asymptotic 0.95-quantile for a large range of k-values. So, since the sample path of the 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.

6.3. Parametric framework

The BM methodology

In this framework, we have considered the years as blocks of observations and have picked the maximum values up in each block. So, we will use the maximum values of each of 60 years — these are all the years available in SNIRH: "Sistema Nacional de Informação dos Recursos Hídricos" for the hydrometric station of Fragas da Torre in river Paiva.

We have now obtained the skewness $= 0.998$ and the kurtosis $= 2.265$. Graphical analyses are shown in Figure 3.

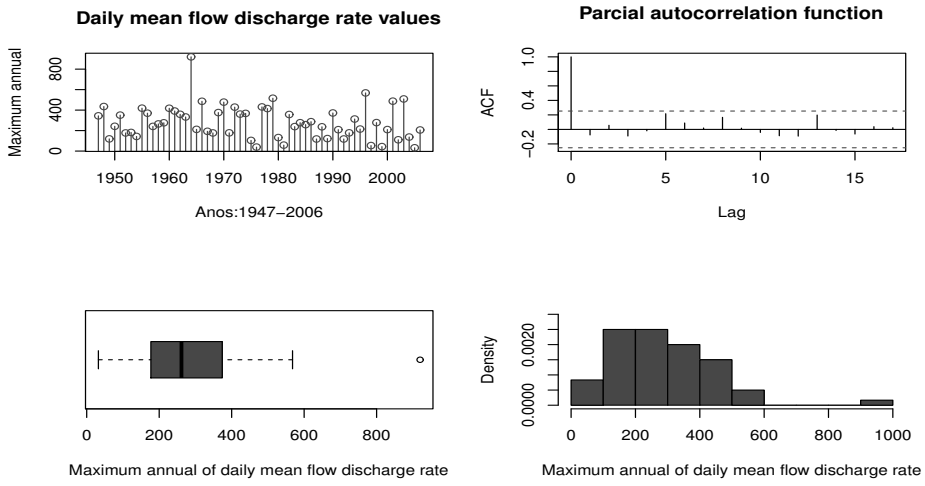

Figure 3: Chronogram (top left); ACF (top right); boxplot (bottom left) and histogram (bottom right).

The histogram, the boxplot and the skewness indicate a moderate positive asymmetry. From the autocorrelation partial function (ACF), it seems reasonable to assume that these data are not correlated. So an EV_ξ was fitted to the maxima in each year.

The ML fitting for the *EV* distribution for all the parameters can be obtained through the *package* `evd` and the function `fgev()`, see Table 2. The parameter estimates by the PWM method can be obtained using the *package* `fExtreme` and running `gevFit( ,type="pwm")`[1]. See results in Table 2.

Table 2: Parameters estimates (standard errors in parenthesis) and the profile Log-Likelihood (pLog-L) 95% confidence intervals.

	$\widehat{\lambda}$	$\widehat{\delta}$	$\widehat{\xi}$
ML	210.08 (18.77)	129.81 (13.62)	-0.03 (0.09)
PWM	213.65	137.37	-0.09

	λ	δ	ξ
pLog-L	(174.15; 248.23)	(106.41; 160.79)	$(-0.16; 0.19)$

Using the same *package*, Wald confidence intervals of level $1-\alpha$, can be obtained through `confint(fgvev(), level=`$1-\alpha$`)`. Greater accuracy for the confidence intervals is usual attained by the profile log-likelihood. Plots for the profile log-likelihood for all parameters can be obtained by `plot(profile(fgev()),ci = c(0.95, 0.99))`. The confidence interval limits can be obtained through `confint(profile(fgvev()),level=`$1-\alpha$`)` and are given in Table 2.

Notice that the confidence intervals for ξ include zero, so lead to not reject the null hypothesis, $\xi = 0$.

The POT methodology

The POT method is based on fitting the statistical model in (2.4) to the *excesses* over a given threshold u. A challenge here is the choice of u. Choosing a value too high can lead to a very small number of observations in the tail resulting in estimators with high variance, but a small threshold may lead to the violation on the Pickands Theorem.

The most traditional methods for the choice of u are graphical procedures. A graph widely used is the *mean residual life* (mrl) plot, based on the mean value of the GP distribution, which is a linear function of u. If the GP model is valid for the excesses above u_0 then will also be valid for all $u > u_0$. So, this graph should show a linear behaviour above a suitable choice of the threshold u. Another graphical method is based on the *threshold choice* (tc) plot, which represents the estimated values of the GP model over a set of thresholds. The threshold u will be a "good" choice if the parameter estimates appear approximately constant

[1] `gevFit()` function can also determine the maximum likelihood estimates, setting type="mle".

above u. The function `mrlplot()` in the *package* `evd` plots the mean excess plot, and the function `tcplot()` plots two graphs for both parameters, see Figure 4.

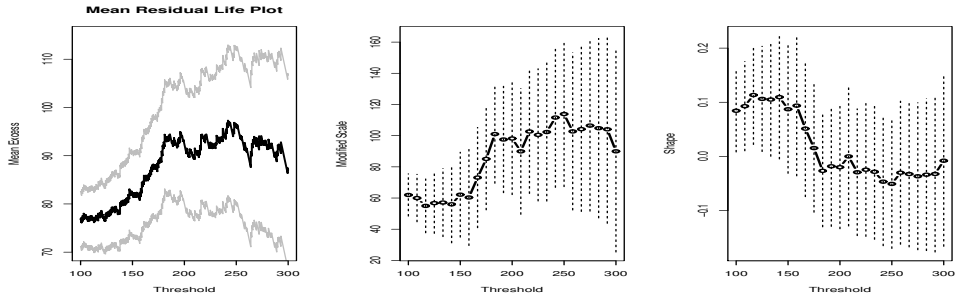

Figure 4: The mean residual life plot (left) and the tc plots (centre and right).

A threshold around 200 is suggested. We have chosen $u = 180$, corresponding to a number of 254 exceedances. Figure 5 shows those exceedances, no correlation of the exceedances and the asymmetry of the data. Using the function `gpd()` in *package* `evir`, we got similar results to those by the BM method, see Table 3.

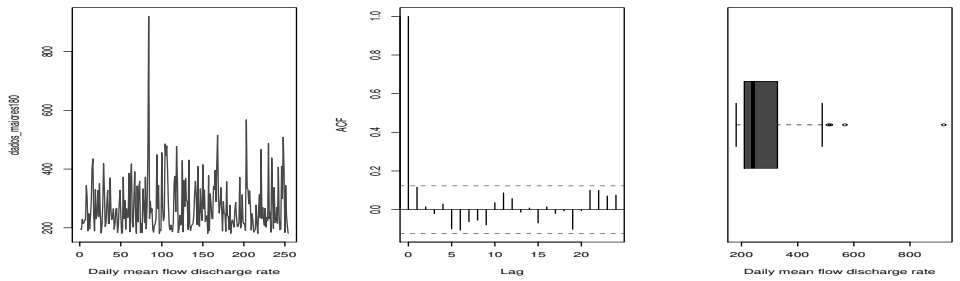

Figure 5: Chronogram (left) with u=180, partial autocorrelation function (center) and boxplot (right).

Table 3: Parameters Estimates (standards errors in parenthesis) and the profile Log-Likelihood (pLog-L) 95% Confidence Intervals.

	$\widehat{\widetilde{\delta}}$	$\widehat{\xi}$
ML	94.69 (7.83)	-0.02 (0.05)
PWM	95.47(9.47)	-0.03 (0.07)

	δ	ξ
pLog-L	(80.01;110.80)	$(-0.10;0.11)$

Note again that the results obtained indicate a value for ξ close to zero.

6.4. Semi-parametric framework

In this approach ξ is the primordial parameter to be estimated. As we referred to in section 3.2, estimates are based on the k top order statistics in the sample, with k an intermediate sequence, assuming that the underlying distribution function F belongs to $\mathcal{D}_{\mathcal{M}}(EV_\xi)$, for an appropriate value of ξ. Since, there are specific estimation procedures according to the signal of ξ, we should start this framework by testing the Gumbel max-domain against Fréchet or Weibull max-domains.

The choice of the tail

To test $H_0 \colon F \in \mathcal{D}_{\mathcal{M}}(EV_0)$ *vs.* $H_1 \colon F \in \mathcal{D}_{\mathcal{M}}(EV_\xi)_{\xi \neq 0}$ or against the one-sided alternatives $F \in \mathcal{D}_{\mathcal{M}}(EV_\xi)_{\xi<0}$ or $F \in \mathcal{D}_{\mathcal{M}}(EV_\xi)_{\xi>0}$ we will consider the *Ratio-test*, the *Gt-test* and the *HW-test*, mentioned in section 5. Figure 6 presents the sample paths of G^*, R^* and W^* for several values of k. As we can see in Figure 6, for a large range of k-values, the three tests statistics present values that belong to the corresponding region of no rejection. So we find no evidence to reject the null hypothesis, $F \in D_M(EV_0)$.

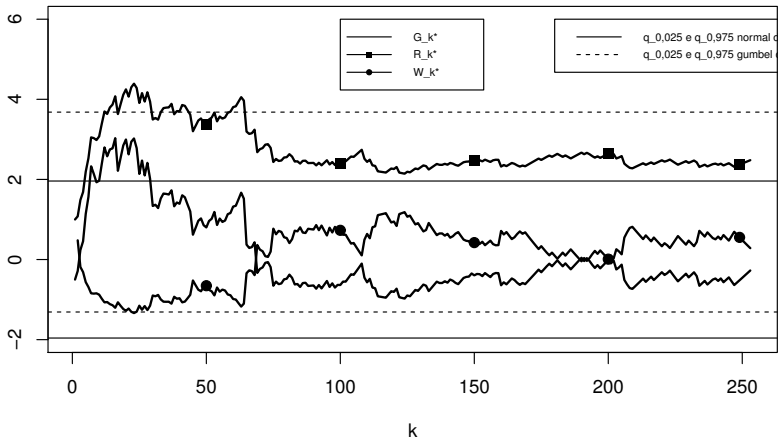

Figure 6: Sample paths of G^*, R^* and W^* statistics.

Some semi-parametric estimates

As specified in Section 3.2, we will consider here the Hill estimator, the Moment estimator, the Generalized Hill estimator, the Mixed Moment estimator and the MVRB estimator. Although having been led above to the non rejection of the Gumbel domain of attraction we present here the results of application of all those estimators.

Figure 7 shows the sample paths of the estimates obtained for each k. It is worthwhile to mention that the Hill estimator and the MVRB estimator, specifically built for $\xi > 0$ show results that are far from those previously obtained (notice that the MVRB estimates show a very stable path, but around positive values of $\widehat{\xi}$). The other estimators present sample paths near $\xi = 0$.

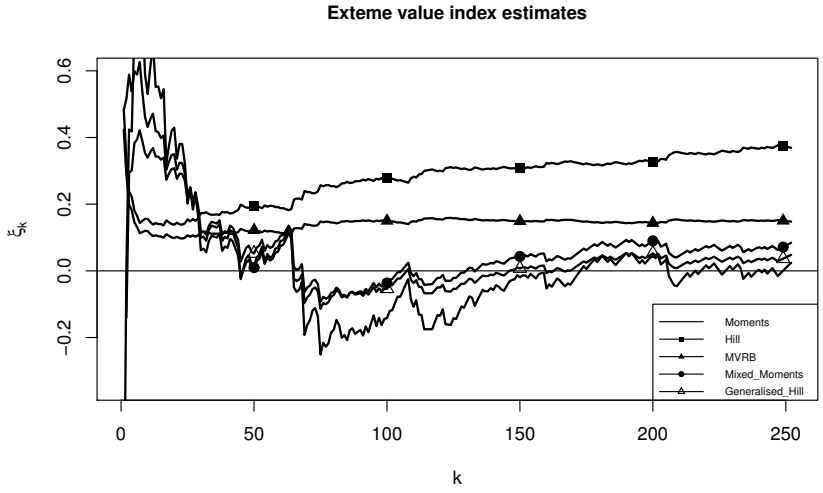

Figure 7: Sample paths of ξ estimates.

7. A FEW OVERALL COMMENTS

Testing whether $F \in \mathcal{D}_{\mathcal{M}}(EV_\xi)$, for a certain ξ, is a crucial topic when an application of extreme values procedures is needed to be considered. This subject has been dealt in several articles mentioned along this paper. However, several times a real problem in the area of EVT is studied without that previous analysis.

With the study of this application we intended to motivate the discussion regarding the need of a previous analysis on the choice of the tail before applying the well theoretically studied estimators. The influence of the estimate of the tail index parameter in the estimation of high quantiles, parameters of major interest for preventing catastrophes that can occur in this domain of application, is also another important issue, however out of the scope of this study.

ACKNOWLEDGMENTS

Research partially supported by National Funds through **FCT** — Fundação para a Ciência e a Tecnologia, projects PEst-OE/MAT/UI0006/2014 (CEAUL).

REFERENCES

[1] Balkema, A.A. and de Haan, L. (1974). Residual life time at great age, *Ann. Probab.*, **2**, 792–804.

[2] Beirlant, J.; Dierckx, G. and Guillou, A. (2005). Estimation of the extreme-value index and generalized quantile plots, *Bernoulli*, **11**(6), 949–970.

[3] Beirlant, J.; Goegebeur, Y.; Teugels, J. and Segers, J. (2004). *Statistics of Extremes: Theory and Applications*, Wiley, England.

[4] Beirlant, J.; Vynckier, P. and Teugels, J. (1996). Excess functions and estimation of the extreme-value index, *Bernoulli*, **2**, 293–318.

[5] Beran, J.; Schell, D. and Stehlik, M. (2014). The harmonic moment tail index estimator: asymptotic distribution and robustness, *Annals of the Institute of Statistical Mathematics*, **66**(1), 193–220.

[6] Brilhante, M.F. (2004). Exponentiality versus generalized Pareto — a resistant and robust test, *RevStat*, **2**(1), 1–13.

[7] Brilhante, M.F.; Gomes, M.I. and Pestana, D. (2013). A simple generalization of the Hill estimator, *Comput. Stat. Data Anal.*, **57**(1), 518–535.

[8] Caeiro, F. and Gomes, M.I. (2006). A new class of estimators of a "scale" second order parameter, *Extremes*, **9**, 193–211.

[9] Caeiro, F.; Gomes, M.I. and Henriques-Rodrigues, L. (2009). Reduced-bias tail index estimators under a third order framework, *Comm. Statist. Th. Meth.*, **38**(7), 1019–1040.

[10] Caeiro, F.; Gomes, M.I. and Pestana, D.D. (2005). Direct reduction of bias of the classical Hill estimator, *Revstat*, **3**, 111–136.

[11] Castillo, J. del; Daoudi, J. and Lockhart, R. (2014). Methods to distinguish between polynomial and exponential tails, *Scandinavian Journal of Statistics*, **41**, 382–393.

[12] 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. Häusler and R.-D. Reiss, Eds.), Springer, Berlin-Heidelberg, **51**, 181–190.

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

[14] Correia, A.L. and Neves, M. (1996). *Escolha estatística em modelos extremais — testes de ajustamento.* In "Bom Senso e Sensibilidade" (J. Branco, P. Gomes e J. Prata, Eds.), Actas do III Congresso Anual da Sociedade Portuguesa de Estatística, Edições Salamandra, 223–236.

[15] Davison, A.C. (1984). *Modelling excesses over high threshold, with an application.* In "Statistical Extremes and Applications" (J. Tiago de Oliveira, Ed.), Reidel, Dordrecht, 389–410.

[16] Davison, A.C. and Smith, R.L. (1990). Models for exceedances over high thresholds (with discussions), *J. Royal Statistical Society*, **52**(3), 393–442.

[17] 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.

[18] Dietrich, D.; De Haan, L. and Hüsler, J. (2002). Testing extreme value conditions, *Extremes*, **5**(1), 71–85.

[19] Drees, H.; De Haan, L. and Li, D. (2006). Approximation to the tail empirical distribution function with application to testing extreme value conditions, *J. Stat. Plan. Inference*, **136**, 3498–3538.

[20] Dwass, M. (1964). Extremal processes, *Ann. Math. Stat.*, **35**, 1718–1725.

[21] Falk, M. (1995). On testing the extreme value index via the POT-method, *Annals of Statistics*, **23**, 2013–2035.

[22] Fisher, R.A. and Tippett, L.H.C. (1928). On the estimation of the frequency distributions of the largest or smallest member of a sample, *Proceedings of the Cambridge Philosophical Society*, **24**, 180–190.

[23] Fraga Alves, M.I. and Gomes, M.I. (1996). Statistical choice of extreme value domains of attraction — a comparative analysis, *Comm. in Stat. Th. and Meth.*, **25**(4), 789–811.

[24] Fraga Alves, M.I.; Gomes M.I. and de Haan, L. (2003). A new class of semi-parametric estimators of the second order parameter, *Portugaliae Mathematica*, **60**, 194–213.

[25] Fraga Alves, M.I. (1999). Asymptotic distribution of Gumbel statistic in a semi-parametric approach, *Portugaliae Mathematica*, **56**(3), 282–298.

[26] 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.

[27] Fréchet, M. (1927). Sur la loi de probabilité de l'écart maximum, *Ann. Soc. Polon. Math. (Cracovie)*, **6**, 93–116.

[28] 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.

[29] Gilleland, E.; Ribatet, M. and Stephenson, A.G. (2012). A software review for extreme value analysis, *Springer (Springerlink.com). Published online: 20 July 2012)*, DOI 10.1007/s10687-012-0155-0.

[30] Gnedenko, B.V. (1943). Sur la distribution limite d'une série aléatoire, *Annals of Mathematics*, **44**, 423–453.

[31] Gomes, M.I. (1981). *An i-dimensional limiting distribution function of largest values and its relevance to the statistical theory of extremes.* In "Statistical Distributions in Scientific Work" (C. Taillie *et al.*, Eds.), D. Reidel, **6**, 389–410.

[32] Gomes, M.I. (1993). On the estimation parameters of rare events in environmental times series, *Statistics for the Environment*, 226–241.

[33] Gomes, M.I.; de Haan, L. and Henriques-Rodrigues, L. (2008). Tail index estimation for heavy-tailed models: accommodation of bias in weighted log-excesses, *J. Royal Statist. Soc. B*, **70**, 31–52.

[34] Gomes, M.I.; Henriques-Rodrigues, L.; Pereira, H. and Pestana, D. (2010). Tail index and second order parameters' semi-parametric estimation based on the log-excesses, *J. Statist. Comput. and Simul.*, **80**, 653–666.

[35] Gomes, M.I. and Martins, M.J. (2002). "Asymptotically unbiased" estimators of the tail index based on external estimation of the second order parameter, *Extremes*, **5**, 5–31.

[36] Gomes, M.I.; Martins, M.J. and Neves, M.M. (2007). Improving second order reduced bias extreme value index estimation, *Revstat*, **5**(2), 177–207.

[37] Gomes, M.I.; Martins, M.J. and Neves, M.M. (2013). Generalized Jackknife-Based Estimators for Univariate Extreme-Value Modeling, *Comm. Statist. Th. Meth.*, **42**(7), 1227–1245.

[38] Greenwood, J.A.; Landwehr, J.M.; Matalas, N.C. and Wallis, J.R. (1979). Probability-weighted moments: definition and relation to parameters of several distributions expressable in inverse form, *Water Resources Research*, **15**, 1049–1054.

[39] Gumbel, E.J. (1935). Les valeurs extrêmes des distributions statistiques, *Ann. Inst. Henri Poincaré*, **5**(2), 115–158.

[40] Gumbel, E.J. (1958). *Statistics of Extremes*, Columbia University Press, New York.

[41] 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.

[42] Hasofer, A. and Wang, J.Z. (1992). A test for extreme value domain of attraction, *Journal of the American Statistical Association*, **87**, 171–177.

[43] Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution, *Annals of Statistics*, **3**, 1163–1174.

[44] Hosking, J.R.M.; Wallis, J.R. and Wood, E.F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments, *Technometrics*, **27**, 251–261.

[45] Hüsler, J. and Li, D. (2006). On testing extreme value conditions, *Extremes*, **9**, 69–86.

[46] Marohn, F. (1998a). An adaptive efficient test for Gumbel domain of attraction, *Scandinavian Journal of Statistics*, **25**, 311–324.

[47] Marohn, F. (1998b). Testing the Gumbel hypothesis via the POT-method, *Extremes*, **1**(2), 191–213.

[48] Neves, C. and Fraga Alves, M.I. (2007). Semi-parametric approach to the Hasofer-Wang and Greenwood statistics in extremes, *Test*, **16**, 297–313.

[49] Neves, C. and Fraga Alves, M.I. (2008). Testing extreme value conditions — an overview and recent approaches, *Revstat*, **6**(1), 83–100.

[50] 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, Eds.), 533–538.

[51] Paulauskas, V. and Vaiciulis, M. (2013). On the improvement of Hill and some others estimators, *Lith. Math. J.*, **53**, 336–355.

[52] Penalva, H.; Neves, M. and Nunes, S. (2013). Topics in data analysis using R in extreme value theory, *Advances in Methodology & Statistics / Metodoloski zvezki*, **10**(1), 17–29.

[53] Penalva, H.; Neves, M. and Nunes, S. (2014). *Statistical Modeling and Inference in Extremes: Applications with R.* In "Statistical and Biometrical Challenges: Theory and Applications, Biometrie und Medizinische Informatik — Greifswalder Seminarberichte" (T. Oliveira, K.-E. Biebler, A. Oliveira and B. Jager, Eds.), Shaker-Verlag, Aachen, 281–309.

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

[55] R Core Team (2013). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org/`

[56] Segers, J. and Teugels, J. (2000). Testing the Gumbel hypothesis by Galton's ratio, *Extremes*, **3**(3), 291–303.

[57] Smith, R.L. (1984). *Threshold methods for sample extremes.* In "Statistical Extremes and Applications" (J. Tiago de Oliveira, Ed.), Reidel, Dordrecht, 621–638.

[58] Smith, R.L. (1985). Maximum likelihood estimation in a class of nonregular cases, *Biometrika*, **72**, 67–90.

[59] Smith, R.L. (1986). Extreme value theory based on their largest annual events, *J. Hydrology*, **86**, 27–43.

[60] 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.

[61] Weissman, I. (1978). Estimation of parameters and large quantiles based on the largest observations, *J. Amer. Statist. Ass.*, **73**, 812–815.

[62] Zhou, C. (2009). Existence and consistency of the maximum likelihood estimator for the extreme value index, *J. Multivariate Anal.*, **100**(4), 794–815.

[63] Zhou, C. (2010). The extent of the maximum likelihood estimator for the extreme value index, *J. Multivariate Anal.*, **101**(4), 971–983.

REVSTAT – Statistical Journal
Volume 14, Number 2, April 2016, 217–228

NON-STATIONARY MODELLING OF EXTREME TEMPERATURES IN A MOUNTAINOUS AREA OF GREECE

Authors: Chrys Caroni
– Department of Mathematics,
School of Applied Mathematical and Physical Sciences,
National Technical University of Athens, Greece
ccar@math.ntua.gr

Dionysia Panagoulia
– Department of Water Resources and Environmental Engineering,
School of Civil Engineering,
National Technical University of Athens, Greece
dpanag@hydro.ntua.gr

Received: October 2015 Revised: February 2016 Accepted: February 2016

Abstract:

- The generalised extreme value (GEV) distribution is often fitted to environmental time series of extreme values such as annual maxima and minima of temperatures. It is often necessary to allow the distribution's parameters to depend on time or other covariates (non-stationary GEV). Increasingly, model fitting within the GAMLSS framework is being used as an alternative approach. A case study is presented of temperature extremes in a mountainous area of Greece divided into nine zones by altitude. Model fitting supported non-stationary GEV models for temperature with the location parameter depending linearly on year and zone, showing the expected dependence on altitude along with an increasing trend in annual maxima and declining trend in annual minima. The scale parameter for maxima depended on zone, with greater variability at higher altitudes. The scale parameter for minima increased over time. Fitting non-stationary Inverse Gaussian, Lognormal and Gamma distributions within the GAMLSS framework identified the same dependence on zone and year. There was little difference in goodness of fit of the various distributions.

Key-Words:

- *modelling extremes; GEV distribution; GAMLSS; non-stationary models; extreme temperatures.*

AMS Subject Classification:

- 60G70, 62P12.

1. INTRODUCTION

The study of extreme values in climatological time series is an area of intense scientific activity. Examples of this type of data are series of annual or monthly maxima of precipitation or temperature. This is the block maximum approach to defining the extremes of time series; the alternative peaks-over-threshold approach will not be considered in this paper (Beirlant *et al.*, [1]; Chavez-Demoulin and Davison, [2]; Gomes and Guillou, [9]; Scarrott and Macdonald, [22]). An early method of analysis that subsequently became well-established was to fit the Generalised Extreme Value distribution, assuming a series of independently and identically distributed values over time (Jenkinson, [11]). However, it has become increasingly clear that many series do not possess this property of stationarity, because of natural climate variability or anthropogenic climate change (Jain and Lall, [10]; Milly *et al.*, [18]; Serinaldi and Kilsby, [23]). Consequently, it becomes necessary to move from stationary to non-stationary models.

Introducing non-stationarity within the framework of standard statistical distributions requires extended models with covariate-dependent changes in one or more of a distribution's parameters (Coles, [4]). For example, a trend towards higher temperatures could be represented by the time-dependence of the parameter that represents the distribution's mean, or increased variability in a rainfall series by time-dependence of the parameter that is associated with the distribution's variance. Spatial trends and dependence on any other available covariates can be represented in a similar way.

2. STATISTICAL MODELLING

2.1. Generalised extreme value (GEV) distribution

The GEV distribution is widely employed in the environmental sciences and elsewhere for modelling extremes (Reiss and Thomas, [20]). It depends on three parameters: location μ, scale σ and shape ξ. In the non-stationary GEV distribution (El Adlouni *et al.*, [7]; Leclerc and Ouarda, [14]), these parameters are expressed as a function of time t and possibly other covariates (Coles, [4]). If, as is usually done, we allow non-stationarity of the location and scale parameters but not of the shape parameter, this non-stationary GEV$(\mu(t),\, \sigma(t),\, \xi)$ distribution has distribution function

$$F\big(y; \mu(t), \sigma(t), \xi\big) \;=\; \exp\left\{ -\left[1 + \xi\, \frac{y - \mu(t)}{\sigma(t)}\right]^{-1/\xi} \right\}. \tag{2.1}$$

In the simplest case, the following regression structures could be considered for the location and scale parameters

$$\begin{aligned} \mu(t) &= \mu_0 + \mu_1 t + \mu_2 t^2 + \mu_3 t^3 \,, \\ \sigma(t) &= \exp\bigl(\sigma_0 + \sigma_1 t + \sigma_2 t^2 + \sigma_3 t^3\bigr) \,, \end{aligned} \tag{2.2}$$

allowing up to cubic dependence on time t. We denote by GEVjk the model with time dependence of order j in the location parameter and order k in the scale parameter. A convenient tool for fitting either stationary or non-stationary GEV distributions is the *gev.fit* function in the R package 'ismev' (available from `http://cran.r-project.org/package=ismev`), which employs the maximum likelihood method. Other relevant R packages are listed by Gomes and Guillou ([9]). Bayesian and other estimation methods are discussed by, for example, Beirlant *et al.* ([1]), Chavez-Demoulin and Davison ([2]), and Gomes and Guillou ([9]). The estimation of the shape parameter ξ may sometimes cause difficulty, as observed by Coles and Dixon ([5]) who proposed using a penalized likelihood function to avoid this problem. Similarly, Martins and Stedinger ([17]) proposed restricting the estimate of ξ to fall within the range $[-0.5, +0.5]$ by using a suitable prior distribution. However, we have not encountered any difficulty in the estimation of ξ in the practical problems that we have investigated.

2.2. GAMLSS

Generalised additive models for location, scale and shape (GAMLSS; Rigby and Stasinopoulos, [21]) represent a very wide class of non-stationary distributions. GAMLSS provide a highly flexible framework for modelling, because as many as four parameters of a distribution chosen from an extensive family are allowed to depend on covariates. The first applications of GAMLSS to meteorological data appear to have been by Villarini and colleagues, who examined the fit of Gumbel, Weibull, Gamma, Lognormal and Logistic distributions to data on rainfall and temperature in Rome (Villarini *et al.*, [25]), and the first four of these to flood peaks in the United States (Villarini *et al.*, [26]). Further examples of its application are now quite common; recent examples include Lopez and Frances ([15]), who fitted the Gumbel, Lognormal, Weibull, Gamma and Generalized Gamma distributions, Garcia Galiano *et al.* ([8]) (fitting the Lognormal, Weibull and Gamma distributions) and Machado *et al.* ([16]) (Lognormal, GEV and two-component extreme value distributions).

The general format of the model for parameter θ_k is

$$g_k(\theta_k) = X_k \beta_k + \sum_{j=1}^{J_k} Z_{jk}\, \gamma_{jk} \,, \tag{2.3}$$

where g_k is a link function, X_k is a design matrix containing the values of J_k covariates for each of n independent observations, β_k is a parameter vector of length J_k, Z_{jk} is another known design matrix of dimension $n \times q_{jk}$ and γ_{jk} is a q_{jk}-dimensional random vector. In the absence of random effects, the first term on the left-hand side of (2.3) gives a parametric linear model; in this case, the advantages of GAMLSS over generalised linear models or generalised additive models are its not being restricted to exponential family distributions and its ability to model several parameters of the distribution, not just the mean. Furthermore, Rigby and Stasinopoulos ([21]) and Stasinopoulos and Rigby ([24]) demonstrate how the second term of (2.3) can be used to construct a wide variety of models, although this generality will not be required in the present paper.

GAMLSS modelling is implemented in the R package 'gamlss' (`http://cran.r-project.org/package=gamlss`; Stasinopoulos and Rigby, [24]), which makes it easy to include features such as random effects or non-polynomial dependence on covariates by means of splines. The method of fitting is penalized maximum likelihood. A recent extension to 'gamlss.spatial' (`http://cran.r-project.org/package=gamlss.spatial`) offers a facility for spatial modelling by including Markov Random Field additive terms.

2.3. Model selection

When searching for the best fitting model among many alternatives, it is important to have objective procedures for making the selection from the various candidates. The likelihood ratio test can be used if the models are hierarchically nested. The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are also widely employed for model selection. If $\hat{\ell}$ is the maximized value of the likelihood from a model that contains p parameters, and n is the sample size, these criteria are defined as

$$AIC_C = -2\hat{\ell} + 2p + \frac{2p(p+1)}{n-p-1} \tag{2.4}$$

(this is the corrected AIC — the third term is a small-sample adjustment) and

$$BIC = -2\hat{\ell} + p \ln n \,. \tag{2.5}$$

The preferred model minimizes the chosen criterion although alternative models with values close to the minimum should not be ignored. More details of model selection procedures can be found in Claeskens and Hjort ([3]), for example.

Panagoulia *et al.* ([19]) carried out a simulation study in order to evaluate empirically the performance of the AICc and BIC in identifying the true model among the set of models GEVjk ($j = 0, 1, 2, 3$; $k = 0, 1, 2, 3$), for samples of sizes $n = 20, 50$ or 100. Both criteria had high success rates in detecting non-stationarity.

The BIC was the more successful in identifying the correct model: over 80% of the time for $n = 50$ and over 90% for $n = 100$, although these percentages obviously depend on the parameter values selected for the study. AICc was better for $n = 20$, although this is a small sample in relation to the number of parameters in some of these models and neither selection criterion performed very well.

2.4. Uncertainty

Apart from obtaining a description of the phenomenon, one of the major objectives of fitting a distribution to climate data is to obtain estimates of its quantiles, especially those related to the return periods of extreme events: for example, the upper 1% point of the distribution of annual maxima corresponds to a $1/.01 = 100$-year return period. Good estimation of the uncertainty in extreme levels can be as important as the estimate of the level itself (Coles, [4]; Khaliq *et al.*, [12]). Parametric confidence intervals based on a normal distribution approximation cannot be expected to be accurate for extreme quantiles; that is, their actual coverage probabilities will not be close to the nominal values. As a result, confidence interval construction by bootstrap methods has been examined for GEV models, first by Kysely ([13]) in the stationary case and subsequently by Panagoulia *et al.* ([19]) in the non-stationary case. Amongst several methods compared, the best was found to be the parametric bootstrap with confidence intervals constructed by the bias corrected and accelerated (BCa) technique. Serinaldi and Kilsby ([23]) expressed a preference for percentile parametric bootstrap confidence intervals, although this appears to be based on general considerations rather than detailed studies. However, they warned that the estimation of extreme quantiles is inherently so uncertain that the discrepancy between different types of confidence intervals is not of major relevance.

Uncertainty in predictions also stems from model selection. The above confidence intervals are based on the assumption that the correct model has been selected and take no account of the alternatives that were considered. Model averaging procedures exist and are used in many contexts to overcome this objection, especially in the Bayesian framework, but will not be considered here.

3. CASE STUDY

3.1. Data

Our analysis concerns time series of meteorological data from one catchment area in the mountains of Central Greece. Further description of this location can be found in Panagoulia *et al.* ([19]), where analyses of annual maxima of rainfall

over the whole catchment area are carried out for historical data and for data simulated under climate change scenaria. In the present paper, analyses are carried out for annual maxima and minima of temperature over the period 1972–1992, over the whole area and in nine zones corresponding to a partition of the area by elevation.

3.2. GEV modelling

The series of minima can be analysed by GEV modelling after taking the negative of its values and fitting the same models as to the series of maxima (Chavez-Demoulin and Davison, [2]). The series of annual extremes for the entire area do not appear to be stationary, as the GEV*10* model offers significantly improved fit over the GEV*00* model (comparing minus twice the change in log-likelihood to the chi-squared distribution with one degree of freedom, $p = 0.05$ for maximum temperatures, $p = 0.01$ for minima). The smooth curves fitted to the annual minima in Figure 1 and annual maxima in Figure 2 demonstrate a decreasing and an increasing trend, respectively. The suggestion in Figure 1 of greater variance of the minima in the later years is not borne out by statistical

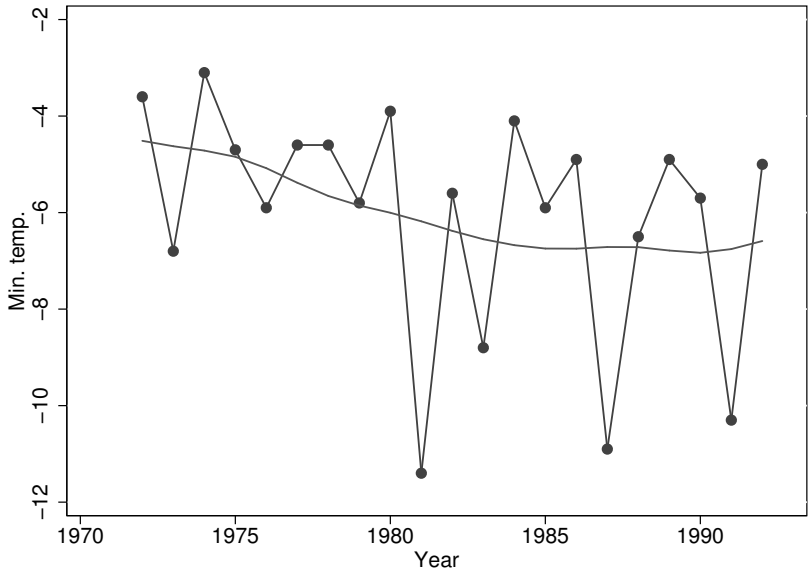

Figure 1: Annual minimum temperatures in °C over the whole study area, with trend fitted by locally weighted scatterplot smoothing.

tests for a linear trend in the scale parameter ($p = 0.40$ for GEV*11* versus GEV*10*; $p = 0.91$ for the corresponding test for the maxima). In contrast to the results of the analysis of rainfall data in Panagoulia *et al.* ([19]), the GEV model for

temperatures did not reduce to the Gumbel ($\xi = 0$) as the former had much better fit than the latter (AIC 96.8 compared to 102.1).

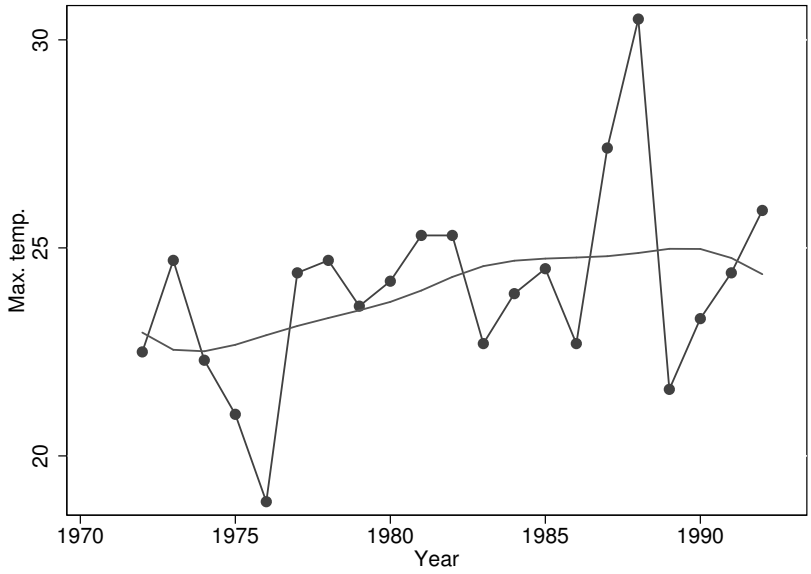

Figure 2: Annual maximum temperatures in ℃ over the whole study area, with trend fitted by locally weighted scatterplot smoothing.

Parameter estimates obtained from fitting the stationary GEV*00* model to the data from each zone separately are shown in Table 1. There appear to be trends with zone, that is, with altitude. In particular, as would be expected,

Table 1: Fitting stationary GEV to annual maximum and minimum temperatures in each zone separately: estimates of location μ, scale σ and shape ξ.

Zone	**Maxima**			**Minima**		
	μ	σ	ξ	μ	σ	ξ
1	25.2	2.05	−0.23	2.16	1.96	−0.310
2	24.4	2.05	−0.19	2.63	1.91	−0.230
3	23.9	2.08	−0.18	3.15	1.78	−0.080
4	23.5	2.12	−0.17	3.97	1.58	0.080
5	23.0	2.22	−0.17	5.07	1.43	0.210
6	22.8	2.49	−0.19	6.24	1.56	0.170
7	22.5	2.73	−0.11	7.70	1.85	0.090
8	22.3	2.85	−0.02	8.76	2.05	0.040
9	22.0	3.06	0.06	10.24	2.28	0.002
Standard error:	0.53	0.35	0.11	0.46	0.34	0.18
median (range)	(0.48–0.75)	(0.32–0.56)	(0.10–0.16)	(0.36–0.57)	(0.29–0.41)	(0.17–0.20)

the location parameter appears to decline as the altitude increases from Zone 1 to Zone 9 for the maxima and also — apparently to a much greater degree — for minus the minima. Furthermore, the relationship with zone is very close to linear. Also, fitting the GEV*10* model separately in each zone (results not shown) suggests time dependence of the scale in many zones, as was noted above in the analysis of the entire area. The next step was to fit models to the annual maxima and minima by year and zone, allowing various forms of dependence on both of these covariates.

The best fitting model for annual maxima, selected using likelihood ratio tests and the AICc and BIC criteria, included linear dependence of μ on zone and year, and log-linear dependence of σ on zone. The shape parameter ξ did not depend on either covariate, despite the indication of a trend in Table 1 which may have been due to correlations between the estimates of the three parameters. The fitted model was:

$$\begin{aligned}
\hat{\mu} &= \underset{(0.36)}{25.00} - \underset{(0.073)}{0.285}\, Zone + \underset{(0.030)}{0.118}\,(Year - 1982)\,, \\
\ln\hat{\sigma} &= \underset{(0.109)}{0.554} + \underset{(0.021)}{0.068}\, Zone\,, \\
\hat{\xi} &= \underset{(0.045)}{-0.142}\,.
\end{aligned}$$

(Standard errors are shown in parentheses below the parameter estimates to which they refer.) We note that the estimate of ξ is clearly significantly different from zero, meaning that the Gumbel distribution is not suitable here. This is different from the finding for rainfall over the same catchment area in Panagoulia *et al.* ([19]), although that analysis was for the total area not broken down by zone. The corresponding analysis for (minus) the annual minima, produced a slightly different model, with σ depending on year instead of zone. The fitted model was:

$$\begin{aligned}
\hat{\mu} &= \underset{(0.314)}{0.438} + \underset{(0.060)}{1.043}\, Zone + \underset{(0.024)}{0.121}\,(Year - 1982)\,, \\
\ln\hat{\sigma} &= \underset{(0.063)}{0.579} + \underset{(0.011)}{0.026}\,(Year - 1982)\,, \\
\hat{\xi} &= \underset{(0.063)}{0.004}\,.
\end{aligned}$$

In this case, the estimate of ξ is clearly not significantly different from zero, implying that the Gumbel distribution could be employed. Comparing the two sets of equations, it is noticeable that annual maxima are increasing and annual minima are decreasing, meaning that the temperature range is increasing. In fact,

the coefficients representing the time dependence of the location μ are almost equal: annual maxima are increasing at the same rate as annual minima are decreasing. However, the coefficient of the dependence of μ on altitude is much bigger for minima than for maxima. This gives an expected result, that minimum temperatures fall more steeply than maximum temperatures with altitude. The results for scale show increasing variability of minima with time — which is what Figure 1 indicated, but did not emerge from the analysis for the entire area aggregated across zones. The variability of maxima increases at higher altitude but is not changing with time.

3.3. GAMLSS modelling

There is no theory to guide the choice of which distribution to fit from among the many available in GAMLSS. We carried out the modelling using the Inverse Gaussian, Gamma and Lognormal distributions, allowing non-stationarity in the form of polynomial dependence of the parameters on year and zone just as we did for the GEV distribution. The preferred models coincided with those chosen for the GEV. We found that the fits of these non-stationary distributions were almost identical. For maxima, AIC values were 911.2 for the Inverse Gaussian distribution, 910.6 for the Lognormal and 912.0 for the Gamma distribution. Graphs demonstrating goodness-of-fit are not presented because the lines showing each distribution are virtually indistinguishable. Furthermore, estimated percentiles were very close.

Close similarity of fits between different models is probably a usual feature of modelling data of this kind. For example, Villarini *et al.* ([26]) analysed annual flood peaks from many stations using GAMLSS and found (see their Table 7) that the Lognormal distribution provided the best fit in 16 sets of data, the Gamma in seven, the Gumbel in 5 and the Weibull in one. In the absence of theory to guide the choice, the preference for one or the other may well just be a matter of sampling variability.

4. CONCLUSION AND COMMENTS

When the underlying distribution is stationary, the choice of the GEV distribution for modelling extremes is well supported on theoretical grounds, precisely because it is an extreme value distribution. That is, it is a form that necessarily arises in the limit to describe the distribution of the maxima of a series of independent and identically distributed random variables (Cox *et al.*, [6]; Gomes and Guillou, [9]). In the non-stationary case, however, the original se-

quence does not consist of identically distributed variables. We are not aware that any limiting form necessarily arises in this case. This suggests that there is no compelling reason to use the non-stationary GEV in preference to many other distributions that are available. The choice of distribution then becomes entirely empirical. This is the approach that seems to have been taken in the various papers that have appeared in the literature so far on the application of GAMLSS to meteorological and related data. These papers tend to demonstrate the possibilities that this flexible approach to modelling offers but not to go on to draw conclusions about which models are the most appropriate on general grounds. Searching through alternative distributions — which the GAMLSS framework tends to encourage — also adds an extra layer of uncertainty to the model selection procedure which ought to be accounted for in predictions.

REFERENCES

[1] Beirlant, J.; Caeiro, F. and Gomes, M.I. (2012). An overview and open research topics in statistics of univariate extremes, *Revstat*, **10**, 1–31.

[2] Chavez-Demoulin, V. and Davison, A.C. (2012). Modelling time series extremes, *Revstat*, **10**, 109–133.

[3] Claeskens, G. and Hjort, N.L. (2008). *Model Selection and Model Averaging*, Cambridge University Press, Cambridge.

[4] Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*, Springer, New York.

[5] Coles, S.G. and Dixon, M.J. (1999). Likelihood-based inference for extreme value models, *Extremes*, **2**, 5–23.

[6] Cox, D.R.; Isham, V.S. and Northrop, P.J. (2002). Floods: some probabilistic and statistical approaches, *Philosophical Transactions of the Royal Society of London Series A*, **360**, 1389–1408.

[7] El Adlouni, S.; Ouarda, T.B.M.J.; Zhang, X.; Roy, R. and Bobee, B. (2007). Generalized maximum likelihood estimators for the nonstationary generalized extreme value distribution, *Water Resources Research*, **43**, W03410.

[8] Garcia Galiano, S.G.; Gimenez, P.O. and Giraldo-Osorio, J.D. (2015). Assessing nonstationary spatial patterns of extreme droughts from long-term high-resolution observational dataset on a semiarid basin (Spain), *Water*, **7**, 5458–5473.

[9] Gomes, M.I. and Guillou, A. (2014). Extreme value theory and statistics of univariate extremes: A review, *International Statistical Review*, **83**, 263–292.

[10] Jain, S. and Lall, U. (2001). Floods in a changing climate: does the past represent the future? *Water Resources Research*, **37**, 3193–3205.

[11] Jenkinson, A.F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements, *Quarterly Journal of the Royal Meteorological Society*, **81**, 158–171.

[12] Khaliq, M.N.; Ouarda, T.B.M.J.; Ondo, J.-C.; Gachon, P. and Bobee, B. (2006). Frequency analysis of a sequence of dependent and/or non-stationary hydro-meteorological observations: A review, *Journal of Hydrology*, **329**, 534–552.

[13] Kysely, J. (2008). A cautionary note on the use of nonparametric bootstrap for estimating uncertainties in extreme-value models, *Journal of Applied Meteorology and Climatology*, **47**, 3236–3251.

[14] Leclerc, M. and Ouarda, T. (2007). Non-stationary regional flood frequency analysis at ungauged sites, *Journal of Hydrology*, **343**, 254–265.

[15] Lopez, J. and Frances, F. (2013). Non-stationary flood frequency analysis in continental Spanish rivers, using climate and reservoir indices as external covariates, *Hydrology and Earth System Sciences*, **17**, 3189–3203.

[16] Machado, M.J.; Botero, B.A.; Lopez, J.; Frances, F.; Diez-Herrero, A. and Benito, G. (2015). Flood frequency analysis of historical flood data under stationary and non-stationary modelling, *Hydrology and Earth System Sciences*, **19**, 2561–2576.

[17] Martins, E.S. and Stedinger, J.R. (2000). Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data, *Water Resources Research*, **36**, 737–744.

[18] Milly, P.C.D.; Betancourt, J.; Falkenmark, M.; Hirsch, R.M.; Kundzewicz, Z.W.; Lettenmaier, D.P. and Stouffer, R.J. (2008). Stationarity is dead: whither water management? *Science*, **319**, 573–574.

[19] Panagoulia, D.; Economou, P. and Caroni, C. (2014). Stationary and non-stationary GEV modelling of extreme precipitation over a mountainous area under climate change, *Environmetrics*, **25**, 29–43.

[20] Reiss, R.D. and Thomas, M. (2007). *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields*, 3rd ed., Birkhäuser, Basel.

[21] Rigby, R.A. and Stasinopoulos, D.M. (2005). Generalized additive models for location scale and shape, *Applied Statistics*, **54**, 507–554.

[22] Scarrott, C. and Macdonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification, *Revstat*, **10**, 33–60.

[23] Serinaldi, F. and Kilsby, C.G. (2015). Stationarity is undead: Uncertainty dominates the distribution of extremes, *Advances in Water Research*, **77**, 17–36.

[24] Stasinopoulos, D.M. and Rigby, R.A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R, *Journal of Statistical Software*, **23**, 1–46.

[25] Villarini, G.; Smith, J.A. and Napolitano, F. (2010). Nonstationary modelling of a long record of rainfall and temperature over Rome, *Advances in Water Resources*, **33**, 1256–1267.

[26] Villarini, G.; Serinaldi, F.; Smith, J.A. and Krajewski, W.F. (2009). On the stationarity of annual flood peaks in the continental United States during the 20th century, *Water Resources Research*, **45**, W08417.

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 February, one in April, one in June and the other in October.

— On average, five articles will be published per issue

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.