ISSN 1645-6726

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

REVSTAT
Statistical Journal

Volume 9, No.2
June 2011

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 21 842 6364
    Web site: http://www.ine.pt
    Customer Support Service
    (National network) : 808 201 808
    Other networks: + 351 22 605 07 48
- **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**
  - *350 copies*
- **LEGAL DEPOSIT REGISTRATION**
  - *N.º 191915/03*

PRICE

[VAT included]

- Single issue …….. **€ 8**
- Annual subscription (No. 1 Special Issue, No. 2 and No.3)……. **€ 19**
- Annual subscription (No. 2, No. 3) …….. **€ 13**

INDEX

A Measure of Departure from Average Marginal Homogeneity for Square Contingency Tables with Ordered Categories

Kouji Yamamoto, Shuji Ando and *Sadao Tomizawa* 115

Statistics of Extremes in Athletics

Lígia Henriques-Rodrigues, M. Ivette Gomes and *Dinis Pestana* 127

A Spatial Unit Level Model for Small Area Estimation

Pedro S. Coelho and *Luís N. Pereira* 155

Generalized Sum Plots

J. Beirlant, E. Boniphace and *G. Dierckx* 181

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 9, Number 2, June 2011, 115–126

A MEASURE OF DEPARTURE FROM AVERAGE MARGINAL HOMOGENEITY FOR SQUARE CONTINGENCY TABLES WITH ORDERED CATEGORIES

Authors: Kouji Yamamoto
– Center for Clinical Investigation and Research, Osaka University Hospital, 2-15, Yamadaoka, Suita, Osaka, 565-0871, Japan
yamamoto-k@hp-crc.med.osaka-u.ac.jp

Shuji Ando
– Graduate School of Sciences and Technology, Tokyo University of Science, Noda City, Chiba, 278-8510, Japan

Sadao Tomizawa
– Department of Information Sciences, Faculty of Science and Technology, Tokyo University of Science, Noda City, Chiba, 278-8510, Japan
tomizawa@is.noda.tus.ac.jp

Received: February 2010 Revised: November 2010 Accepted: December 2010

Abstract:

- For the analysis of square contingency tables, Tomizawa, Miyamoto and Ashihara (2003) considered a measure to represent the degree of departure from marginal homogeneity. However, the maximum value of this measure cannot distinguish two kinds of marginal inhomogeneity. This paper proposes a measure which can distinguish two kinds of marginal inhomogeneity for square tables with ordered categories. The measure is constructed using the arc-cosine function of symmetric cumulative probabilities. Especially the proposed measure is useful for representing the degree of departure from marginal homogeneity when the extended marginal homogeneity model holds. Examples are given.

Key-Words:

- *average marginal homogeneity; extended marginal homogeneity; measure; ordinal data.*

AMS Subject Classification:

- 62H17.

1. INTRODUCTION

Consider an $R \times R$ square contingency table with the same row and column classifications. Let p_{ij} denote the probability that an observation will fall in the i-th row and j-th column of the table ($i = 1, ..., R$; $j = 1, ..., R$), and let X and Y denote the row and column variables, respectively. The marginal homogeneity model is defined by

$$\Pr(X = i) = \Pr(Y = i) \qquad \text{for} \quad i = 1, ..., R \ ,$$

namely

$$p_{i\cdot} = p_{\cdot i} \qquad \text{for} \quad i = 1, ..., R \ ,$$

where $p_{i\cdot} = \sum_{k=1}^{R} p_{ik}$ and $p_{\cdot i} = \sum_{k=1}^{R} p_{ki}$. See, for example, Stuart (1955), Bishop, Fienberg and Holland (1975, p. 294).

Let

$$G_{1(i)} = \sum_{s=1}^{i} \sum_{t=i+1}^{R} p_{st} \qquad \left[= \Pr(X \le i, \, Y \ge i+1)\right] ,$$

and

$$G_{2(i)} = \sum_{s=i+1}^{R} \sum_{t=1}^{i} p_{st} \qquad \left[= \Pr(X \ge i+1, \, Y \le i)\right] ,$$

for $i = 1, ..., R-1$. Then, by considering the difference between the cumulative marginal probabilities, $F_i^X - F_i^Y$ for $i = 1, ..., R-1$, where $F_i^X = \Pr(X \le i)$ and $F_i^Y = \Pr(Y \le i)$, we see that the marginal homogeneity model may also be expressed as

$$G_{1(i)} = G_{2(i)} \qquad \text{for} \quad i = 1, ..., R-1 \ .$$

Namely, this model also states that the cumulative probability that an observation will fall in row category i or below and column category $i+1$ or above is equal to the cumulative probability that the observation falls in column category i or below and row category $i+1$ or above for $i = 1, ..., R-1$.

When the marginal homogeneity model does not hold, we are interested in measuring the degree of departure from the marginal homogeneity model.

For square contingency tables with ordered categories, Tomizawa, Miyamoto and Ashihara (2003) proposed the following measure $\Gamma^{(\lambda)}$ to represent the degree of departure from marginal homogeneity: assuming that $\{G_{1(i)} + G_{2(i)} \ne 0\}$, for $\lambda > -1$,

$$\Gamma^{(\lambda)} = \frac{\lambda(\lambda+1)}{2^{\lambda}-1} \sum_{i=1}^{R-1} \left(G_{1(i)}^{*} + G_{2(i)}^{*}\right) I_i^{(\lambda)}\left(\left\{G_{1(i)}^{c}, G_{2(i)}^{c}\right\}; \left\{\frac{1}{2}, \frac{1}{2}\right\}\right) ,$$

where

$$\Delta = \sum_{i=1}^{R-1} \left(G_{1(i)} + G_{2(i)} \right),$$

$$G^*_{1(i)} = \frac{G_{1(i)}}{\Delta}, \qquad G^*_{2(i)} = \frac{G_{2(i)}}{\Delta},$$

$$G^c_{1(i)} = \frac{G_{1(i)}}{G_{1(i)} + G_{2(i)}}, \qquad G^c_{2(i)} = \frac{G_{2(i)}}{G_{1(i)} + G_{2(i)}},$$

with

$$I_i^{(\lambda)}(\cdot;\cdot) = \frac{1}{\lambda(\lambda+1)} \left[G^c_{1(i)} \left\{ \left(\frac{G^c_{1(i)}}{1/2} \right)^{\lambda} - 1 \right\} + G^c_{2(i)} \left\{ \left(\frac{G^c_{2(i)}}{1/2} \right)^{\lambda} - 1 \right\} \right],$$

and the value at $\lambda = 0$ is taken to be the limit as $\lambda \to 0$. Note that $I_i^{(\lambda)}(\cdot;\cdot)$ is the Cressie and Read (1984) power-divergence between two distributions (also see Read and Cressie, 1988, p. 15).

The measure $\Gamma^{(\lambda)}$ has characteristics that:

(**i**) it lies between 0 and 1;

(**ii**) $\Gamma^{(\lambda)} = 0$ if and only if the marginal homogeneity model holds;

(**iii**) $\Gamma^{(\lambda)} = 1$ if and only if the degree of departure from marginal homogeneity is maximum (that is, $G_{1(i)} = 0$ (then $G_{2(i)} > 0$) or $G_{2(i)} = 0$ (then $G_{1(i)} > 0$) for all $i = 1, ..., R-1$).

However, using the measure $\Gamma^{(\lambda)}$, we cannot distinguish two kinds of marginal inhomogeneity, namely, that the marginal inhomogeneity is which of

(**i**) $G_{1(i)} = 0$ (then $F_i^X < F_i^Y$) for all $i = 1, ..., R-1$,

or

(**ii**) $G_{2(i)} = 0$ (then $F_i^X > F_i^Y$) for all $i = 1, ..., R-1$.

Since these two kinds of marginal inhomogeneity indicate the opposite different maximum departures from marginal homogeneity, we are now interested in proposing a measure which can take the different values for them.

The purpose of this paper is to propose such a measure which can distinguish two kinds of marginal inhomogeneity for square contingency tables with ordered categories. We note that Tahata, Yamamoto, Nagatani and Tomizawa (2009) investigated average symmetry. In the present paper, we consider the average marginal homogeneity using a similar ideas to Tahata *et al.* (2009) and using as example the same data.

2. A MEASURE FOR MARGINAL HOMOGENEITY

Consider an $R\times R$ table with ordered categories. Let

$$\Delta = \sum_{i=1}^{R-1} \left(G_{1(i)} + G_{2(i)}\right) ,$$

and

$$G^*_{1(i)} = \frac{G_{1(i)}}{\Delta} , \quad G^*_{2(i)} = \frac{G_{2(i)}}{\Delta} , \qquad \text{for} \quad i = 1, ..., R-1 \ .$$

Assuming that $\left\{G_{1(i)} + G_{2(i)} \neq 0\right\}$, consider a measure defined by

$$\Psi \ = \ \frac{4}{\pi} \sum_{i=1}^{R-1} \left(G^*_{1(i)} + G^*_{2(i)}\right) \left(\theta_i - \frac{\pi}{4}\right) ,$$

where

$$\theta_i \ = \ \cos^{-1}\left(\frac{G_{1(i)}}{\sqrt{G^2_{1(i)} + G^2_{2(i)}}} \right) .$$

Noting that the range of θ_i is $0 \leq \theta_i \leq \pi/2$, we see that the measure Ψ lies between -1 and 1. The measure Ψ has characteristics that:

- (**i**) $\Psi = -1$ if and only if $G_{2(i)} = 0$ (then $F^X_i > F^Y_i$) for all $i = 1, ..., R-1$, [marginal inhomogeneity with all probabilities zero of lower left triangle (say, L-marginal inhomogeneity)];
- (**ii**) $\Psi = 1$ if and only if $G_{1(i)} = 0$ (then $F^X_i < F^Y_i$) for all $i = 1, ..., R-1$, [marginal inhomogeneity with all probabilities zero of upper right triangle (say, U-marginal inhomogeneity)].

In addition, $\Psi = 0$ indicates that the weighted average of $\{\theta_i - \frac{\pi}{4}\}$ equals zero. Thus when $\Psi = 0$, we shall refer to this structure as the average marginal homogeneity. We note that if the marginal homogeneity holds then the average marginal homogeneity holds, but the converse does not hold.

Therefore, using the measure Ψ, we can see whether the average marginal homogeneity departs toward the L-marginal inhomogeneity or toward the U-marginal inhomogeneity. As the measure Ψ approaches -1, the departure from the average marginal homogeneity becomes greater toward the L-marginal inhomogeneity. While as the Ψ approaches 1, it becomes greater toward the U-marginal inhomogeneity.

3. RELATIONSHIPS BETWEEN THE MEASURE AND SOME MODELS

First, we consider the relationship between the measure Ψ and the extended marginal homogeneity model. The extended marginal homogeneity model considered by Tomizawa (1984), is defined by

$$G_{1(i)} = \tau\, G_{2(i)} \qquad \text{for} \quad i = 1, ..., R-1 \ .$$

A special case of this model obtained by putting $\tau = 1$ is the marginal homogeneity model. If the extended marginal homogeneity model holds true, then the measure Ψ can be expressed as

$$\Psi \ = \ \frac{4}{\pi} \cos^{-1}\left(\frac{\tau}{\sqrt{\tau^2+1}}\right) - 1 \ . \tag{1}$$

Therefore, $\Psi = 0$ if and only if the marginal homogeneity model holds, i.e., $\tau = 1$, thus $G_{1(i)} = G_{2(i)}$ for $i = 1, ..., R-1$. As the value of τ approaches the infinity, the measure Ψ approaches –1. While as the value of τ approaches zero, Ψ approaches 1. Thus when the extended marginal homogeneity model holds in a table, the measure Ψ represents the degree of departure from marginal homogeneity toward the L-marginal inhomogeneity or toward the U-marginal inhomogeneity.

Next, consider the conditional symmetry model (McCullagh, 1978) defined by

$$p_{ij} = \tau\, p_{ji} \qquad \text{for} \quad i < j \ .$$

This model implies the extended marginal homogeneity model. Therefore, if the conditional symmetry model holds true, then the measure Ψ can also be expressed as (1).

Therefore for comparisons in several tables, if it can be estimated that there is a structure of extended marginal homogeneity or conditional symmetry in each table, then the measure Ψ would be adequate for representing and comparing the degree of departure from the marginal homogeneity toward the L-marginal inhomogeneity and U-marginal inhomogeneity.

The measure Ψ should be applied to the ordinal data of square tables with the same row and column classifications because the Ψ is not invariant under arbitrary similar permutations of row and column categories.

4. APPROXIMATE CONFIDENCE INTERVAL FOR THE MEASURE

Let n_{ij} denote the observed frequency in the i-th row and j-th column of the table ($i = 1, ..., R$; $j = 1, ..., R$). Assuming that a multinomial distribution applies to the $R \times R$ table, we shall consider the approximate variance for estimated measure and large-sample confidence interval for the measure Ψ using delta method, the descriptions of which are given by, e.g., Bishop *et al.* (1975, Sec. 14.6). The sample version of Ψ, i.e., $\hat{\Psi}$, is given by Ψ with $\{p_{ij}\}$ replaced by $\{\hat{p}_{ij}\}$, where $\hat{p}_{ij} = n_{ij}/n$ and $n = \sum\sum n_{ij}$. Using delta method, $\sqrt{n}(\hat{\Psi} - \Psi)$ has asymptotically (as $n \to \infty$) a normal distribution with mean zero and variance,

$$\sigma^2[\Psi] \;=\; \sum_{k<l}\sum \left(p_{kl}\, D_{kl}^2 + p_{lk}\, D_{lk}^2\right),$$

where for $k < l$,

$$D_{kl} = \frac{4}{\pi\Delta}\sum_{i=k}^{l-1}\left[\cos^{-1}\left(\frac{G_{1(i)}}{\sqrt{G_{1(i)}^2+G_{2(i)}^2}}\right) - \frac{G_{2(i)}\left(G_{1(i)}+G_{2(i)}\right)}{G_{1(i)}^2+G_{2(i)}^2}\right] - \frac{(l-k)\,(\Psi+1)}{\Delta},$$

$$D_{lk} = \frac{4}{\pi\Delta}\sum_{i=k}^{l-1}\left[\cos^{-1}\left(\frac{G_{1(i)}}{\sqrt{G_{1(i)}^2+G_{2(i)}^2}}\right) + \frac{G_{1(i)}\left(G_{1(i)}+G_{2(i)}\right)}{G_{1(i)}^2+G_{2(i)}^2}\right] - \frac{(l-k)\,(\Psi+1)}{\Delta}.$$

Let $\hat{\sigma}^2[\Psi]$ denote $\sigma^2[\Psi]$ with $\{p_{ij}\}$ replaced by $\{\hat{p}_{ij}\}$. $\hat{\sigma}[\Psi]/\sqrt{n}$ is an estimated standard error for $\hat{\Psi}$, $\hat{\Psi} \pm z_{p/2}\,\hat{\sigma}[\Psi]/\sqrt{n}$ is an approximate $100(1-p)\%$ confidence interval for Ψ, where $z_{p/2}$ is the percentage point from the standard normal distribution that corresponds to a two-tail probability equal to p.

The maximum likelihood estimates of expected frequencies under each of the marginal homogeneity, extended marginal homogeneity and average marginal homogeneity models can be obtained using the Newton–Raphson methods to the log-likelihood equations. The marginal homogeneity, extended marginal homogeneity and average marginal homogeneity models can be tested for goodness-of-fit by, e.g., the likelihood ratio chi-squared statistic with $R-1$, $R-2$, and 1 degrees of freedom, respectively.

5. ANALYSIS OF DATA

5.1. Analysis of Table 1(a)

Consider the data in Table 1(a) taken from Stuart (1955). These are data on unaided distance vision of 7477 women aged 30 to 39 employed in Royal Ordnance factories in Britain from 1943 to 1946. These data have been analyzed by many statisticians, e.g., including Stuart (1955), Caussinus (1965), Bishop *et al.* (1975, p. 284), McCullagh (1978), Goodman (1979), Agresti (1983), Tomizawa (1993), and Tomizawa and Tahata (2007), etc.

Table 1: The unaided vision data of
(a) 7477 women in Britain (from Stuart, 1955),
(b) 3242 men in Britain (from Stuart, 1953),
(c) 4746 students in Japan (from Tomizawa, 1984).

(a) Women in Britain

Right eye grade	Left eye grade				Total
	Best (1)	Second (2)	Third (3)	Worst (4)	
Best (1)	1520	266	124	66	1976
Second (2)	234	1512	432	78	2256
Third (3)	117	362	1772	205	2456
Worst (4)	36	82	179	492	789
Total	1907	2222	2507	841	7477

(b) Men in Britain

Right eye grade	Left eye grade				Total
	Best (1)	Second (2)	Third (3)	Worst (4)	
Best (1)	821	112	85	35	1053
Second (2)	116	494	145	27	782
Third (3)	72	151	583	87	893
Worst (4)	43	34	106	331	514
Total	1052	791	919	480	3242

(c) Students in Japan

Right eye grade	Left eye grade				Total
	Best (1)	Second (2)	Third (3)	Worst (4)	
Best (1)	1291	130	40	22	1483
Second (2)	149	221	114	23	507
Third (3)	64	124	660	185	1033
Worst (4)	20	25	249	1429	1723
Total	1524	500	1063	1659	4746

We see from Table 2 that for the data in Table 1(a), the value of estimated measure $\hat{\Psi}$ is -0.102 and all values in the confidence interval for Ψ are negative. Therefore, the average marginal homogeneity for the women's right and left eyes departs toward the L-marginal inhomogeneity. Table 3 gives the values of likelihood ratio chi-squared statistic for testing goodness-of-fit of each model.

Table 2: The estimates of Ψ, estimated approximate standard errors for $\hat{\Psi}$, and approximate 95% confidence intervals for Ψ, applied to Tables 1(a), 1(b) and 1(c).

Applied data	Estimated measure	Standard error	Confidence interval
Table 1(a)	-0.102	0.029	$(-0.160, -0.045)$
Table 1(b)	0.038	0.044	$(-0.048, +0.123)$
Table 1(c)	0.128	0.040	$(+0.049, +0.206)$

We see from Table 3 that each model of marginal homogeneity and average marginal homogeneity fits the data in Table 1(a) poorly, but the extended marginal homogeneity model fits these data well. So we can see from the estimated measure that the degree of departure from marginal homogeneity for the vision data in Table 1(a) is estimated to be 10.2 percent of the maximum departure toward the L-marginal inhomogeneity. This indicates that the right eye is better than her left eye for all women.

Table 3: The values of likelihood ratio chi-squared statistic for the models of marginal homogeneity, average marginal homogeneity and extended marginal homogeneity, applied to Tables 1(a), 1(b) and 1(c).

Table 1(a)

Applied models	degrees of freedom	Likelihood ratio chi-square
Marginal homogeneity	3	11.99^*
Average marginal homogeneity	1	11.98^*
Extended marginal homogeneity	2	0.005

Table 1(b)

Applied models	degrees of freedom	Likelihood ratio chi-square
Marginal homogeneity	3	3.68
Average marginal homogeneity	1	0.73
Extended marginal homogeneity	2	2.94

Table 1(c)

Applied models	degrees of freedom	Likelihood ratio chi-square
Marginal homogeneity	3	11.18^*
Average marginal homogeneity	1	9.94^*
Extended marginal homogeneity	2	0.56

$*$ means significant at the 0.05 level.

5.2. Analysis of Table 1(b)

Consider the data in Table 1(b) taken from Stuart (1953). These are data on unaided distance vision of 3242 men in Britain.

We see from Table 2 that for the data in Table 1(b), the value of measure $\hat{\Psi}$ is 0.038 and the confidence interval for Ψ includes zero. So this would indicate that there is a structure of average marginal homogeneity in the data in Table 1(b). Also we see from Table 3 that the marginal homogeneity model fits these data well, and each model of average marginal homogeneity and extended marginal homogeneity also fits these data well. Therefore, it is estimated that there is a structure of marginal homogeneity for the data in Table 1(b), and also the estimated measure $\hat{\Psi}$ would indicate it.

5.3. Analysis of Table 1(c)

Consider the data in Table 1(c) taken from Tomizawa (1984). These are data on unaided distance vision of 4746 students aged 18 to about 25 including about 10% women in Faculty of Science and Technology, Science University of Tokyo in Japan examined in April 1982.

For the data in Table 1(c), we see from Table 2 that the value of $\hat{\Psi}$ is 0.128 and all values in the confidence interval for Ψ are positive. Therefore, the average marginal homogeneity for the students' right and left eyes departs toward the U-marginal inhomogeneity. This is a contrast to the women's vision data in Table 1(a). We see from Table 3 that each model of marginal homogeneity and average marginal homogeneity fits the data in Table 1(c) poorly, but the extended marginal homogeneity model fits these data well. So we can see from the estimated measure that the degree of departure from marginal homogeneity for the vision data in Table 1(c) is estimated to be 12.8 percent of the maximum departure toward the U-marginal inhomogeneity. This indicates that the left eye is better than his/her right eye for all students.

In addition, when we compare the data in Tables 1(a) and 1(c) using the estimated measure $\hat{\Psi}$, the degree of departure from the marginal homogeneity for the right and left eyes is greater in the students data in Table 1(c) than in the women data in Table 1(a) (see Table 2). Since the $\hat{\Psi}$ is negative for the women vision data and positive for the students vision data, a woman's right eye tends to be greater than her left eye, and a student's left eye tends to be greater than his/her right eye.

ACKNOWLEDGMENTS

The authors would like to express their thanks to anonymous referee for his/her helpful suggestions and comments.

REFERENCES

[1] Agresti, A. (1983). A simple diagonals-parameter symmetry and quasi-symmetry model, *Statistics and Probability Letters*, **1**, 313–316.

[2] Bishop, Y.M.M.; Fienberg, S.E. and Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*, The MIT Press, Cambridge, Massachusetts.

[3] Caussinus, H. (1965). Contribution à l'analyse statistique des tableaux de corrélation, *Annales de la Faculté des Sciences de l'Université de Toulouse*, **29**, 77–182.

[4] Cressie, N. and Read, T.R.C. (1984). Multinomial goodness-of-fit tests, *Journal of the Royal Statistical Society, Series B*, **46**, 440–464.

[5] Goodman, L.A. (1979). Multiplicative models for square contingency tables with ordered categories, *Biometrika*, **66**, 413–418.

[6] McCullagh, P. (1978). A class of parametric models for the analysis of square contingency tables with ordered categories, *Biometrika*, **65**, 413–418.

[7] Read, T.R.C. and Cressie, N. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*, Springer, New York.

[8] Stuart, A. (1953). The estimation and comparison of strengths of association in contingency tables, *Biometrika*, **40**, 105–110.

[9] Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification, *Biometrika*, **42**, 412–416.

[10] Tahata, K.; Yamamoto, K.; Nagatani, N. and Tomizawa, S. (2009). A measure of departure from average symmetry for square contingency tables with ordered categories, *Austrian Journal of Statistics*, **38**, 101–108.

[11] Tomizawa, S. (1984). Three kinds of decompositions for the conditional symmetry model in a square contingency table, *Journal of the Japan Statistical Society*, **14**, 35–42.

[12] Tomizawa, S. (1993). Diagonals-parameter symmetry model for cumulative probabilities in square contingency tables with ordered categories, *Biometrics*, **49**, 883–887.

[13] Tomizawa, S.; Miyamoto, N. and Ashihara, N. (2003). Measure of departure from marginal homogeneity for square contingency tables having ordered categories, *Behaviormetrika*, **30**, 173–193.

[14] Tomizawa, S. and Tahata, K. (2007). The analysis of symmetry and asymmetry: orthogonality of decomposition of symmetry into quasi-symmetry and marginal symmetry for multi-way tables, *Journal de la Société Française de Statistique*, **148**, 3–36.

REVSTAT – Statistical Journal
Volume 9, Number 2, June 2011, 127–153

STATISTICS OF EXTREMES IN ATHLETICS

Authors: Lígia Henriques-Rodrigues
– Instituto Politécnico de Tomar and C.E.A.U.L., Portugal
Ligia.Rodrigues@aim.estt.ipt.pt

M. Ivette Gomes
– Universidade de Lisboa, F.C.U.L. (D.E.I.O.) and C.E.A.U.L., Portugal
ivette.gomes@fc.ul.pt

Dinis Pestana
– Universidade de Lisboa, F.C.U.L. (D.E.I.O.) and C.E.A.U.L., Portugal
dinis.pestana@fc.ul.pt

Received: July 2010 Revised: December 2010 Accepted: February 2011

Abstract:

- TV shows on any athletic event make clear that those who want *gold medals* cannot dispense *statistics*. And the statistics more appealing to champions and coachers are the *extreme order statistics*, and in particular *maximum* (or *minimum*) *values* and *records*. The models in *statistics of extremes* are usually semi-parametric or even non-parametric in nature, with the imposition of a few *regularity conditions* in the appropriate tail of the unknown model underlying the available data. The primordial parameter is the *extreme value index*, the shape parameter in the (unified) *extreme value* distribution. The estimation of the *extreme value index* is one of the basis for the estimation of other parameters of rare events, like the *right endpoint* of the model underlying the data, a *high quantile*, the *return period* and the *probability of exceedance* of a high level. In this paper, we are interested in an application of *statistics of extremes* to the best personal marks in a few athletic events. Due to the way data are collected, we begin with a parametric data analysis, but we pay special attention to the semi-parametric estimation of the extreme value index and the right endpoint whenever finite, the possible *world record*, given the actual conditions. In order to achieve a better decision we consider a few alternative semi-parametric estimators available in the literature, and heuristic rules for the choice of thresholds.

Key-Words:

- *statistics of extremes; athletics; semi-parametric estimation; extreme value index; right endpoint; excellence indicators; Monte Carlo methodology.*

AMS Subject Classification:

- 62G32, 62E20, 65C05.

1. INTRODUCTION AND OUTLINE OF THE PAPER

Statistical facts are quite commonly used by sports commentators. We all have listened to programs on different athletic events, showing that *statistics* is an instrument that champions instructors need to use. It is without doubt a subject which cannot be dispensed by those who want *gold medals*, and the statistics more appealing to the champions are the *extreme order statistics* (o.s.'s), and in particular *maximum* (or *minimum*) *values* and *records*.

The models in *statistics of extremes* are usually semi-parametric or even non-parametric in nature, with the imposition of a few "regularity conditions" in the right-tail $\big($or left-tail$\big)$, $\overline{F}(x) := 1 - F(x)$, as $x \to +\infty$ $\big($or $F(x)$, as $x \to -\infty\big)$, of an unknown model F underlying the available data, whenever we are interested in large (or small) values. The primordial parameter is the *extreme value index*. For large values, the extreme value index is the shape parameter γ in the distribution function (d.f.)

$$G_\gamma(x) = \begin{cases} \exp\big(-(1+\gamma x)^{-1/\gamma}\big), & 1+\gamma x > 0\ , \qquad \text{if } \gamma \neq 0\ , \\ \exp\big(-\exp(-x)\big), & x \in \mathbb{R}\ , \qquad\qquad\ \ \text{if } \gamma = 0\ , \end{cases} \tag{1.1}$$

the (unified) *extreme value* distribution. The extreme value index needs to be estimated in a "precise" way, because such an estimation plays a major role in the estimation of other parameters of extreme and large events, like the *right endpoint* of the model F underlying the data,

$$x^* := \sup\Big\{x\colon F(x) < 1\Big\}\ , \tag{1.2}$$

a *high quantile* with probability $1-p$, p small, i.e., $\chi_{1-p} := \inf\big\{x\colon F(x) \geq 1-p\big\}$, $p < 1/n$, with n the available sample size, the *return period* and the *probability of exceedance* of a *high level*.

In this paper, we shall be interested in an application of *statistics of extremes* to the best personal marks attained at a few athletic events, in a context similar to the one used in Einmahl and Magnus (2008). We shall pay special attention to the estimation of γ, in (1.1), as well as of the right endpoint x^*, in (1.2), whenever finite, and of an indicator of the "excellence" of the level $x_{n:n}$, the maximum of the n available observations. The right endpoint provides an estimate of the possible "world record" given the actual conditions, and the closer to one the "excellence" indicator of the level $x_{n:n}$ is, the better is the actual world record. In Section 2, we present some preliminary results in *extreme value theory*. In Section 3, we refer a few details on the semi-parametric estimation of a few parameters of extreme events. In Section 4, we provide heuristic choices of the thresholds for an adaptive semi-parametric estimation of the parameters of interest. Such heuristic choices take essentially into account the similarities of a few simple and alternative estimators of those parameters. In Section 4, we analyze data related to six athletic events and draw some final comments.

2. PRELIMINARY RESULTS IN EXTREME VALUE THEORY

Let us think on any athletic event, like for instance the women marathon. Let us denote the best personal marks of n different athletes by $X_1, X_2, ..., X_n$ and by

$$X_{1:n} \le X_{2:n} \le \cdots \le X_{n:n}$$

the associated ascending o.s.'s. Under this set-up, $(X_1, X_2, ..., X_n)$ can be considered as independent, identically distributed (i.i.d.) observations from an underlying model F, obviously unknown. Let us also assume that, if necessary, data are transformed so that we can speak of maximum values (and not of minimum values). Indeed, any result for maxima can be easily reformulated for minima, due to the fact that $\min(X_1, X_2, ..., X_n) = -\max(-X_1, -X_2, ..., -X_n)$. However, this is not the transformation used later on in this paper for the analysis of data in athletics, where we shall convert running times in speeds, so that the higher the speed, the better. We shall thus work with upper o.s.'s.

One of the main results in extreme value theory is related to the possible limiting laws of the sequence $X_{n:n} := \max(X_1, X_2, ..., X_n)$, of maximum values, as $n \to \infty$. Since

$$\mathbb{P}\big(X_{n:n} \le x\big) \,=\, \mathbb{P}\left(\bigcap_{i=1}^{n} \{X_i \le x\}\right) \,=\, F^n(x) \,\underset{n\to\infty}{\longrightarrow}\, \begin{cases} 0 & \text{if } F(x) < 1 \,, \\ 1 & \text{if } F(x) = 1 \,, \end{cases}$$

we obviously have

$$X_{n:n} \,\underset{n\to\infty}{\xrightarrow{p}}\, x^* \,,$$

with x^* given in (1.2).

In order to obtain a possible non-degenerate behaviour for $X_{n:n}$, we thus need to normalize it. Similarly to the *central limit theorem* for sums or means, we know that if the maximum $X_{n:n}$, linearly normalized, converges to a non-degenerate random variable (r.v.), then there exist real constants $\{a_n\}_{n\ge1}$ $(a_n > 0)$ and $\{b_n\}_{n\ge1}$, the so-called *attraction coefficients* of F to G_γ, in (1.1), such that

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{X_{n:n} - b_n}{a_n} \le x\right) = \lim_{n\to\infty} F^n(a_n\, x + b_n) \,=\, G_\gamma(x) \,, \tag{2.1}$$

for some $\gamma \in \mathbb{R}$ (Gnedenko, 1943; de Haan, 1970). We then say that F is in the *domain of attraction* (for maxima) of G_γ and we use the notation $F \in \mathcal{D}_{\mathcal{M}}(G_\gamma)$.

The extreme value index γ, in (1.1), measures essentially the weight of the right-tail $\overline{F} = 1 - F$. If $\gamma < 0$, the right-tail is *light*, i.e., F has a finite right endpoint $(x^* < +\infty)$. If $\gamma > 0$, the right-tail is *heavy*, of a negative polynomial type, i.e., F has an infinite right endpoint. If $\gamma = 0$, the right-tail is of an *exponential* type and the right endpoint can be either finite or infinite. In Figure 1,

we represent graphically the probability density function (p.d.f.) associated with the extreme value d.f., in (1.1), i.e. $g_\gamma(x) = dG_\gamma(x)/dx$, for $\gamma = -0.5$, 0 and 0.5. We also picture the standard normal p.d.f., $\varphi(x) = \exp(-x^2/2)/\sqrt{2\pi}$, $x \in \mathbb{R}$, as well as a "zoom" of the right-tails of these four models. It is clear the lightness of the right-tail of G_γ for $\gamma < 0$ (finite right endpoint), followed by the normal tail and next the Gumbel tail ($\gamma = 0$). It is also clear the heaviness of the right-tail of G_γ for $\gamma > 0$.

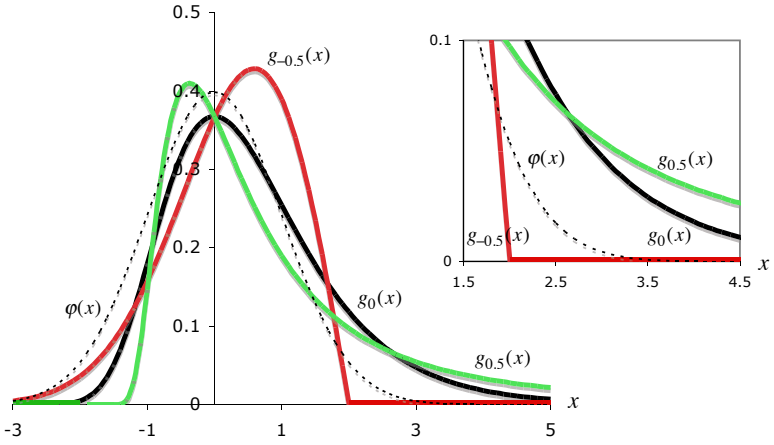

Figure 1: Extreme value p.d.f.'s, $g_\gamma(x)$, with $\gamma = -0.5, 0$ and 0.5, and normal p.d.f., $\varphi(\cdot)$.

Remark 2.1. Note that to say that $F \in \mathcal{D}_{\mathcal{M}}(G_\gamma)$ is equivalent to saying that for all x real and such that $0 < G_\gamma(x) < 1$, $\lim_{n\to\infty} n \ln F(a_n x + b_n) = \ln G_\gamma(x) = -(1+\gamma x)^{-1/\gamma}$. Consequently, $F(a_n x + b_n) \to 1$ for those values of x. Since $\lim_{n\to\infty} \big(-\ln F(a_n x + b_n)\big)\big/\big(1 - F(a_n x + b_n)\big) = 1$, we equivalently have

$$\lim_{n\to\infty} n\big(1 - F(a_n x + b_n)\big) \;=\; -\ln G_\gamma(x) \;=\; (1+\gamma x)^{-1/\gamma}\,. \tag{2.2}$$

Let us define

$$U(t) := F^{\leftarrow}(1 - 1/t) \;\; (t > 1)\,, \qquad F^{\leftarrow}(x) := \inf\big\{y\colon F(y) \ge x\big\}\,, \tag{2.3}$$

with $F^{\leftarrow}$ denoting thus the generalized inverse function of F. It is reasonably easy to prove (de Haan, 1984) that, with G_γ^{-1} denoting the inverse function of the extreme value d.f. G_γ in (1.1),

$$\lim_{t\to\infty} \frac{U(tx) - b_t}{a_t} \;=\; G_\gamma^{-1}\big(\exp(-1/x)\big) \;=\; \frac{x^\gamma - 1}{\gamma}\,,$$

for all $x > 0$, with $a_t \equiv a(t) \equiv a_{[t]}$, $[t]$ = integer part of t and a_n the scale attraction coefficient in (2.1). Also $b_t \equiv b(t) \equiv b_{[t]}$, with b_n the location attraction

coefficient, also in (2.1). Moreover, we can choose $b_t = U(t)$, with $U(\cdot)$ defined in (2.3) (see Theorem 1.1.2 of de Haan and Ferreira, 2006). More generally,

$$F \in \mathcal{D}_{\mathcal{M}}(G_\gamma) \iff \lim_{t\to\infty} \frac{U(tx)-U(t)}{a(t)} = \frac{x^\gamma - 1}{\gamma}\,, \tag{2.4}$$

for all $x > 0$, with $U(\cdot)$ defined in (2.3).

Remark 2.2. When $\gamma = 0$, and by continuity arguments, the functions $-\ln G_\gamma(x) = (1+\gamma x)^{-1/\gamma}$ and $G_\gamma^{-1}\big(\exp(-1/x)\big) = (x^\gamma - 1)/\gamma$ should be interpreted as $\exp(-x)$ and $\ln x$, respectively.

3. SEMI-PARAMETRIC ESTIMATION OF A FEW RELEVANT PARAMETERS OF EXTREME EVENTS

On the basis of the available random sample, $(X_1, X_2, ..., X_n)$, let us see how to estimate the extreme value index γ, the primordial parameter in statistics of extremes, the scale a, the location b, the right endpoint x^* and the return period of a high level $x_{\rm H}$, usually defined as the expected number of exceedances of such a level.

3.1. Estimation of the extreme value index

For any integer $j \geq 1$, let us denote

$$L_{k,n}^{(j)} := \frac{1}{k}\sum_{i=1}^{k}\left\{1 - \frac{X_{n-k:n}}{X_{n-i+1:n}}\right\}^j \tag{3.1}$$

and

$$M_{k,n}^{(j)} := \frac{1}{k}\sum_{i=1}^{k}\Big\{\ln X_{n-i+1:n} - \ln X_{n-k:n}\Big\}^j\,. \tag{3.2}$$

These statistics have revealed to be fundamental in *statistics of extremes*. For the estimation of γ, we shall first refer three estimators, valid, i.e. consistent, for all $\gamma \in \mathbb{R}$:

1. The *moment* (M) estimator (Dekkers *et al.*, 1989), with the functional form

$$\hat{\gamma}_{k,n}^{M} \equiv M_{k,n} := M_{k,n}^{(1)} + \frac{1}{2}\left\{1 - \left(\frac{M_{k,n}^{(2)}}{\big[M_{k,n}^{(1)}\big]^2} - 1\right)^{-1}\right\}, \tag{3.3}$$

$M_{k,n}^{(j)}$, $j = 1, 2$, defined in (3.2).

2. The *generalized Hill* (*GH*) estimator introduced in Beirlant *et al.* (1996), further studied in Beirlant *et al.* (2005), and based on the Hill estimator (Hill, 1975), the statistic $M_{k,n}^{(1)}$, in (3.2), also denoted

$$\hat{\gamma}_{k,n}^{H} \equiv H_{k,n} := \frac{1}{k} \sum_{i=1}^{k} \Big\{ \ln X_{n-i+1:n} - \ln X_{n-k:n} \Big\} , \tag{3.4}$$

and valid only for $\gamma \geq 0$. The *GH*-estimator, valid for all $\gamma \in \mathbb{R}$, and with $\hat{\gamma}_{k,n}^{H}$ given in (3.4), has the functional form

$$\hat{\gamma}_{k,n}^{GH} \equiv GH_{k,n} := \hat{\gamma}_{k,n}^{H} + \frac{1}{k} \sum_{i=1}^{k} \Big\{ \ln \hat{\gamma}_{i,n}^{H} - \ln \hat{\gamma}_{k,n}^{H} \Big\} . \tag{3.5}$$

3. The *mixed moment* (*MM*) estimator (Fraga Alves *et al.*, 2009), with the functional form

$$\hat{\gamma}_{k,n}^{MM} \equiv MM_{k,n} := \frac{\hat{\varphi}_{k,n} - 1}{1 + 2 \min(\hat{\varphi}_{k,n} - 1, 0)} , \qquad \hat{\varphi}_{k,n} := \frac{M_{k,n}^{(1)} - L_{k,n}^{(1)}}{\big(L_{k,n}^{(1)}\big)^2} , \tag{3.6}$$

$L_{k,n}^{(1)}$ and $M_{k,n}^{(1)}$ defined in (3.1) and (3.2), respectively.

The three estimators in (3.3), (3.5) and (3.6) are consistent in $\mathcal{D}_{\mathcal{M}}(G_\gamma)$, $\gamma \in \mathbb{R}$, if $k = k_n$ is an intermediate sequence, i.e., a sequence of integers such that

$$k = k_n \to \infty \quad \text{and} \quad k_n = o(n) , \qquad \text{as} \quad n \to \infty . \tag{3.7}$$

Due to the specificity of the data, we shall also consider another simple estimator:

4. The *location invariant* estimator (*F*) introduced in Falk (1995),

$$\hat{\gamma}_{k,n}^{F} \equiv F_{k,n} := \frac{1}{k} \sum_{i=1}^{k-1} \ln \frac{X_{n:n} - X_{n-i:n}}{X_{n:n} - X_{n-k:n}} , \tag{3.8}$$

valid only for a negative extreme value index smaller than -0.5.

We still would like to refer the so-called "*maximum likelihood*" (*ML*) estimator, introduced in Smith (1987) and further studied in Drees *et al.* (2004). Such an estimator is valid and asymptotically normal for all $\gamma > -1$ (see Zhou, 2009, 2010, for details in the region $-1 < \gamma \leq -1/2$). The extreme value index *ML*-estimator is based on the application of the maximum likelihood methodology to the excesses $X_{n-i+1:n} - X_{n-k:n}$, $1 \leq i \leq k$. These excesses are approximately the k top o.s.'s of a sample of size k from a *generalized Pareto* model, strongly related to the *extreme value* d.f. G_γ in (1.1), through the relation

$$GP(x; \gamma, \alpha) = 1 + \ln G_\gamma(\alpha x/\gamma) = 1 - (1 + \alpha x)^{-1/\gamma}, \quad 1 + \alpha x > 0, \; x > 0 , \tag{3.9}$$

with $\alpha, \gamma \in \mathbb{R}$. This is a re-parametrization due to Davison (Davison, 1984).

Then, with such a re-parametrization, the *ML*-estimator of γ has an explicit expression as a function of the *ML*-estimator $\widehat{\alpha} = \widehat{\alpha}_{ML}$ of α and the sample of the excesses. We have

$$\hat{\gamma}_{k,n}^{ML} \;=\; \widehat{\gamma}_{k,n,\hat{\alpha}}^{ML} \;\equiv\; ML_{k,n} \;:=\; \frac{1}{k}\sum_{i=1}^{k} \ln\big(1+\hat{\alpha}(X_{n-i+1:n}-X_{n-k:n})\big)\;. \tag{3.10}$$

The estimates $\widehat{\alpha}=\widehat{\alpha}_{ML}$ are obtained through numerical iterative methods, usually computationally time-consuming. This is the reason why we shall not consider these estimators in the Monte Carlo simulation in Section 4, related to heuristic choices of thresholds. We shall however consider the *ML*-estimators in the data analysis provided in Section 5.3.2, due to their nice asymptotic properties for $-1/2<\gamma<0$ (see, for instance, Gomes and Neves, 2008, among others).

For a large variety of models, and under mild second-order conditions, we can guarantee the asymptotic normality of all the above mentioned estimators and can build approximate confidence intervals (CI's) for γ, as well as for all other parameters of extreme events, like the ones discussed next in Section 3.2. We merely need to assume the existence of a function $A(t)$, converging to 0, as $t\to\infty$, which measures the rate of convergence of the sequence of maximum values to a non-degenerate limit r.v. and that "measures" also the bias of the estimators in a great variety of situations (see de Haan and Ferreira, 2006, for details). Such a second-order condition can be written as

$$\lim_{t\to\infty} \frac{\frac{U(tx)-U(t)}{a(t)} - \frac{x^{\gamma}-1}{\gamma}}{A(t)} \;=\; H_{\gamma,\rho}(x) \;:=\; \frac{1}{\rho}\left(\frac{x^{\gamma+\rho}-1}{\gamma+\rho} - \frac{x^{\gamma}-1}{\gamma}\right), \tag{3.11}$$

for all $x>0$, where $\rho\le 0$ is a second order parameter controlling the speed of convergence in the first-order condition, (2.4), and $|A(t)|\in RV_{\rho}$, with RV_a standing for the class of regularly varying functions at infinity with an index of regular variation a, i.e. positive measurable functions g such that $\lim_{t\to\infty} g(tx)/g(t)=x^a$, for all $x>0$. Note that for the *extreme value* d.f., in (1.1), condition (3.11) holds, with $\rho=-1$ if $\gamma\neq 1$ and $\rho=-2$ if $\gamma=1$. For the Generalized Pareto d.f., in (3.9), $U(t)=(t^{\gamma}-1)/\gamma$, and we can say that condition (3.11) holds with $A(t)\equiv 0\;\;(\rho=-\infty)$.

3.2. Semi-parametric estimation of other parameters of interest

3.2.1. Estimation of location and scale

As mentioned before, we have $b_t=U(t)$, with $U(\cdot)$ defined in (2.3). On another side, the universal uniform transformation enables us to guarantee that $\forall\, F$, unknown and underlying the r.v. X, $X\overset{d}{=}U(Y)$, with Y a unit Pareto r.v.,

i.e. a r.v. with d.f. $F_Y(y) = 1 - 1/y,\ y \geq 1$. Consequently,

$$X_{n-k:n} \stackrel{d}{=} U(Y_{n-k:n})\,, \quad \text{and since} \quad Y_{n-k:n} \stackrel{p}{\sim} n/k\,, \ \text{ as } n \to \infty\,,$$

where $X_n \stackrel{p}{\sim} Y_n$ means that X_n/Y_n converges in probability to one, as $n \to \infty$, it is sensible to consider

$$\hat{b} = \hat{b}_{k,n} = \widehat{U}(n/k) = X_{n-k:n}\,.$$

And for any extreme value index estimator, $\hat{\gamma}^\bullet \equiv \hat{\gamma}^\bullet_{k,n}$, we can consider (de Haan and Ferreira, 2006)

$$\hat{a}^\bullet = \hat{a}^\bullet_{k,n} = X_{n-k:n} M^{(1)}_{k,n}\big(1 - \min(0, \hat{\gamma}^\bullet)\big)\,,$$

with $M^{(1)}_{k,n}$ given in (3.2).

3.2.2. Estimation of the right endpoint for $\gamma < 0$

For large values of t and $\gamma \neq 0$, if we take into account the validity of condition (2.4), we can write the approximation $U(tx) \approx U(t) + a(t)\,(x^\gamma - 1)/\gamma$. But $x^* = U(\infty)$ and for all $\gamma < 0$, $(x^\gamma - 1)/\gamma \to -1/\gamma$, as $x \to \infty$. If we consider $t = n/k$, with k intermediate, we can thus guarantee that, whenever $\hat{\gamma}^\bullet < 0$,

$$x^* \approx U(n/k) - a(n/k)/\gamma \quad \Longrightarrow \quad \hat{x}^*_\bullet := \hat{b} - \hat{a}^\bullet/\hat{\gamma}^\bullet\,.$$

As we have the obvious restriction $x_{n:n} \leq x^*$, we shall instead consider the right endpoint estimator

$$\hat{x}^*_{k,n|\bullet} = \max\left(X_{n:n},\, X_{n-k:n}\left(1 - M^{(1)}_{k,n}\big(1 - \min(0, \hat{\gamma}^\bullet_{k,n})\big)/\hat{\gamma}^\bullet_{k,n}\right)\right). \tag{3.12}$$

3.2.3. Estimation of the return period of a high level $x_{\rm H}$ and similar indicators

In a pure framework of i.i.d. observations, if we think on the number of observations $N_{\rm H}$ needed to reach a value higher than $x_{\rm H}$, such a r.v. has support $\{1, 2, ...\}$ and $\mathbb{P}(N_{\rm H} = r) = p_{\rm H}(1 - p_{\rm H})^{r-1}$, $r \geq 1$, with $p_{\rm H} = \mathbb{P}(X > x_{\rm H}) = 1 - F(x_{\rm H})$, i.e. $N_{\rm H}$ is a *geometric* r.v.. The *return period* of the high level $x_{\rm H}$ is usually defined as the mean value of $N_{\rm H}$, being given by

$$R(x_{\rm H}) := \frac{1}{p_{\rm H}} = \frac{1}{1 - F(x_{\rm H})}\,.$$

In the framework of this paper, it is perhaps sensible to think on the n athletes under consideration, and to define an indicator associated with a high level $x_{\rm H}$

as the mean number of athletes, among the n, who will have in the future a personal mark larger than $x_{\rm H}$. We thus have the mean value of a Binomial$\big(n, p_{\rm H} = 1-F(x_{\rm H})\big)$ r.v., given by

$$MN(x_{\rm H}) := n\big(1-F(x_{\rm H})\big)\ ,$$

with *MN* standing for *mean number*.

On the basis of the limiting relation in (2.2), we can then consider the estimators

$$\widehat{R}^{\bullet}(x_{\rm H}) := \frac{n}{k}\left(\min\left(+\infty,\ 1+\hat{\gamma}^{\bullet}\left(\frac{x_{\rm H}-\hat{b}}{\hat{a}^{\bullet}}\right)\right)\right)^{1/\hat{\gamma}^{\bullet}}$$

and

$$\widehat{MN}^{\bullet}(x_{\rm H}) \equiv n\,\widehat{p}^{\bullet}_{\rm H} := k\left(\max\left(0,\ 1+\hat{\gamma}^{\bullet}\left(\frac{x_{\rm H}-\hat{b}}{\hat{a}^{\bullet}}\right)\right)\right)^{-1/\hat{\gamma}^{\bullet}}$$

of $R(x_{\rm H})$ and $MN(x_{\rm H})$, respectively.

Note that for $x_{\rm H} = x_{n:n}$, F absolutely continuous, and denoting $(U_1, ..., U_n)$ a random sample from a uniform d.f. in $(0,1)$, we have

$$MN(X_{n:n}) = n\big(1-F(X_{n:n})\big) \stackrel{d}{=} n\,U_{1:n}\ ,$$

which converges weakly towards a unit exponential r.v., as $n\to\infty$. Consequently, the sequence of r.v.'s $\exp\big(-MN(X_{n:n})\big)$ converges weakly towards a uniform r.v. in $(0,1)$. In the data analysis provided in Section, 5.3.2 we shall thus consider

$$\widehat{E}^{\bullet}_{n} \equiv \widehat{E}^{\bullet}_{k,n} := \exp\big(-\widehat{MN}^{\bullet}(X_{n:n})\big) \tag{3.13}$$

as an estimator of an indicator of the "excellence" of the world record $X_{n:n}$, given by $E_n := \exp\big(-MN(X_{n:n})\big)$. Note that the E-indicator was chosen merely because it lies in the finite support [0,1]. The closer to 1 this indicator is, the better is the actual world record. Such an indicator is strongly related to the *quality of the current world record*'s indicator $Q := -\ln E_n = n\big(1-F(X_{n:n})\big)$ of Einmahl and Magnus (2008), the expected number of exceedances of the current world record, $X_{n:n}$, conditional on this world record.

For further details on most of the subjects of this Section, see Chapters 1 and 4 of de Haan and Ferreira (2006).

4. HEURISTIC CHOICES OF THRESHOLDS IN THE SEMI-PARAMETRIC EXTREME VALUE INDEX, RIGHT END-POINT AND EXCEEDANCE PROBABILITY ESTIMATION: A MONTE-CARLO STUDY

For any arbitrary estimator, $\hat{\gamma}^{\bullet}_{k,n}$, of γ, like the ones in (3.3), (3.4), (3.5), (3.6), (3.8) and (3.10), and under the validity of a second-order condition like the one in (3.11), we get an asymptotic distributional representation of the type

$$\hat{\gamma}^{\bullet}_{k,n} \stackrel{d}{=} \gamma + \frac{\sigma_{\bullet}\, P^{\bullet}_{k}}{\sqrt{k}} + v_{\bullet}\, A(n/k)\,\big(1+o_p(1)\big)\,, \tag{4.1}$$

with $P^{\bullet}_{k} \stackrel{a}{\sim} \text{Normal}(0,1)$. Consequently, for intermediate levels k, i.e., levels such that (3.7) holds, and also such that $\sqrt{k}\,A(n/k) \to \lambda$, finite, $\exists\, v_{\bullet} \in \mathbb{R}$ and $\sigma_{\bullet} \in \mathbb{R}^{+}$ such that

$$\sqrt{k}\,(\hat{\gamma}^{\bullet}_{k,n} - \gamma) \xrightarrow[n\to\infty]{d} \text{Normal}(\lambda v_{\bullet},\, \sigma^2_{\bullet})\,. \tag{4.2}$$

The "*asymptotic mean squared error*" (*AMSE*) is defined as

$$AMSE\big(\hat{\gamma}^{\bullet}_{k,n}\big) := \frac{\sigma^2_{\bullet}}{k} + v^2_{\bullet}\, A^2(n/k)\,,$$

i.e. we get asymptotic bias and variance given by $BIAS_{\infty}\big(\hat{\gamma}^{\bullet}_{k,n}\big) := v_{\bullet}\, A(n/k)$ and $\text{Var}_{\infty}\big(\hat{\gamma}^{\bullet}_{k,n}\big) := \sigma^2_{\bullet}/k$, respectively. If $\lambda = 0$, the mean value of the limiting normal law in (4.2) is equal to zero.

Let us define $k^{\bullet}_0 = k^{\bullet}_0(n) := \arg\min_k MSE\big(\hat{\gamma}^{\bullet}_{k,n}\big) \sim \arg\min_k AMSE\big(\hat{\gamma}^{\bullet}_{k,n}\big)$, the level associated with a minimal *AMSE*, as the optimal level for the estimation of γ through $\hat{\gamma}^{\bullet}_{k,n}$, and let us denote $\hat{\gamma}^{\bullet}_{n0} := \hat{\gamma}^{\bullet}_{k^{\bullet}_0,n}$, the estimator computed at its optimal level. With the notation $A(t) = \beta t^{\rho}$, $\rho < 0$, the value $\sigma_{\bullet}$ is a function of γ and $v_{\bullet}$ is usually a function of β and ρ (possibly also of γ). We then get

$$k^{\bullet}_0 \;=\; \big(\sigma^2_{\bullet}/(-2\,\rho\, v^2_{\bullet}\beta^2)\big)^{1/(1-2\rho)}\, n^{-2\rho/(1-2\rho)}\,. \tag{4.3}$$

In order to estimate $k^{\bullet}_0$ in (4.3), in a simple and precise way, we thus need to have "nice" estimates of the second-order parameters (β, ρ). However, whereas such an estimation is reliable for $\gamma > 0$ (see, for instance, Caeiro *et al.*, 2005; Gomes and Pestana, 2007; Gomes *et al.*, 2007, 2008, among others), this is not the case for $\gamma \le 0$. Notice however that we can estimate ρ, for a general $\gamma \in \mathbb{R}$, through the estimators in Fraga Alves *et al.* (2003). Even so, the optimal level, in (4.3), depends often not only on β but also on γ. The estimation of k_0 can then be made recursively, but it induces a high volatility in the estimates and a drastic loss of efficiency. Alternatively, we could also use, for instance, bootstrap

methods (Draisma *et al.*, 1999; Danielson *et al.*, 2001; Gomes and Oliveira, 2001) for an optimal adaptive choice of k. Here, after deciding on a negative value for γ, as will be the case in Section 5, we propose the following heuristic choice of the threshold k. Let us denote $\hat{\gamma}_{k,n}^{(i)}$, $i \in \mathcal{K} = \{1,2,3,4\}$, the set of alternative (and computationally simple to obtain) *EVI*-estimators in (3.3), (3.5), (3.6) and (3.8). Then, consider

$$k^*_{\min} := \arg\min_k \sum_{(i,j)\in\mathcal{K},\, i\neq j} \big(\hat{\gamma}_{k,n}^{(i)} - \hat{\gamma}_{k,n}^{(j)}\big)^2 \tag{4.4}$$

and

$$T^* := T_{k^*_{min},n}\,, \qquad \text{with} \quad T = M \text{ or } GH \text{ or } MM \text{ or } F\,, \tag{4.5}$$

$M_{k,n}$, $GH_{k,n}$, $MM_{k,n}$, $F_{k,n}$ and $k^*_{\min}$ given in (3.3), (3.5), (3.6), (3.8) and (4.4), respectively. We cannot claim any kind of asymptotic optimality for the choice $k^*_{\min}$, in (4.4), in the sense that we would like to have $k^*_{\min}/k_0^\bullet \to 1$, as $n \to \infty$. However, if $b_\bullet \neq 0$, we can guarantee that the value $k^*_{\min}$ in (4.4) is of the order of $n^{-2\rho/(1-2\rho)}$, i.e., of the same order of the optimal value $k_0^\bullet$ in (4.3). Consequently, (4.2) holds whenever we there replace k by $k^*_{\min}$. Moreover, the value in (4.4) seems to be heuristically appealing whenever we want to take into account a set of alternative semi-parametric estimators of a parameter of extreme events. It is expected that there will be a region where all estimators work, and in such a region we surely get close values for all estimates and the smallest possible value for the indicator in (4.4). If we enlarge the set $\mathcal{K}$, in order to include the extreme value index *ML*-estimator, in (3.10), as we shall do in the data analysis performed in Section 5.3.2, we shall use the notations $k^{**}_{\min}$ and T^{**} for the entities equivalent to the ones in (4.4) and (4.5), respectively.

We shall also consider the same type of heuristic procedure for the estimation of the right endpoint x^*, in (1.2), done through similar adaptive right endpoint estimators,

$$\hat{x}^*_T := \hat{x}^*_{k^{*x}_{\min},n|T}\,, \qquad \text{again with} \quad T = M \text{ or } GH \text{ or } MM \text{ or } F\,, \tag{4.6}$$

$\hat{x}^*_{k,n|\bullet}$ given in (3.12), and where, for the same set $\mathcal{K}$ and the same notation as before,

$$k^{*x}_{\min} := \arg\min_k \sum_{(i,j)\in\mathcal{K},\, i\neq j} \big(\hat{x}^*_{k,n|(i)} - \hat{x}^*_{k,n|(j)}\big)^2 =: k^{*x}\,. \tag{4.7}$$

Similarly, we shall use the notations $k^{**x}_{\min}$ and $\hat{x}^{**}_T$, whenever we include in $\mathcal{K}$ the *ML*-estimator, in (3.10), for the estimation of the right endpoint. A similar method was also applied to the estimators of the "excellence" indicators, in (3.13) (or equivalently to the exceedance probability of $X_{n:n}$). We shall use the obvious similar notations $\widehat{E}^*_\bullet, \widehat{E}^{**}_\bullet$ for those adaptive estimators and $k^{*E}_{\min} \equiv k^{*E}$, $k^{**E}_{\min} \equiv k^{**E}$ for the adaptive choices of k.

In order to obtain distributional properties of the adaptive estimators under consideration, we have performed simulation studies of size 5000×10 for sample sizes $n = 100$, 200, 300, 400, 500, 1000, 2000 and 5000, from a reasonably large variety of models. Due to characteristics of the data, which are maxima of a certain number of marks, and should consequently be associated with an underlying d.f. quite close to the *extreme value* (EV) model, we shall uniquely present, as an illustration, the results associated with an underlying model $F(x) = G_\gamma(x)$, with $G_\gamma(x)$ given in (1.1), $\gamma = -0.1$ and -0.3.

For each value of n, we have simulated not only the mean values and root mean squared errors of the four estimators in (4.5), but also of the similar adaptive right endpoint estimators in (4.6). A similar method was also applied to the estimators of the "excellence" indicators, in (3.13) (or equivalently to the exceedance probability of $X_{n:n}$). As mentioned before, we shall use the obvious notation $\widehat{E}^*_\bullet$ for those adaptive estimators and $k^{*E}_{\min} \equiv k^{*E}$ for the adaptive choice of k. Due to the stability of the sample paths of the estimators in (3.8), even when we cannot guarantee their consistency, the results do not depend on either the inclusion or the non-inclusion of such an estimator.

For underlying EV models, with $\gamma = -0.1$ and -0.3, the estimates of the absolute bias ($|BIAS|$) and root mean squared error ($RMSE$) of the adaptive EVI-estimators are presented in Figures 2 and 3, respectively. We also present in these figures the corresponding values of at least one of the estimators at its simulated optimal level, denoted T_0, with $T = M$, GH, MM or F. For the bias structure, we present only one T_0, the one with the lowest absolute bias for large values of n. The introduction of the ML-estimator, in (3.10), does not lead to very different conclusions, but increases drastically the time of computation and consequently the loss of precision associated with the $REFF$ indicators. Similar patterns have been obtained for underlying GP and reversed-Burr parents, and we see no need to present those extra results.

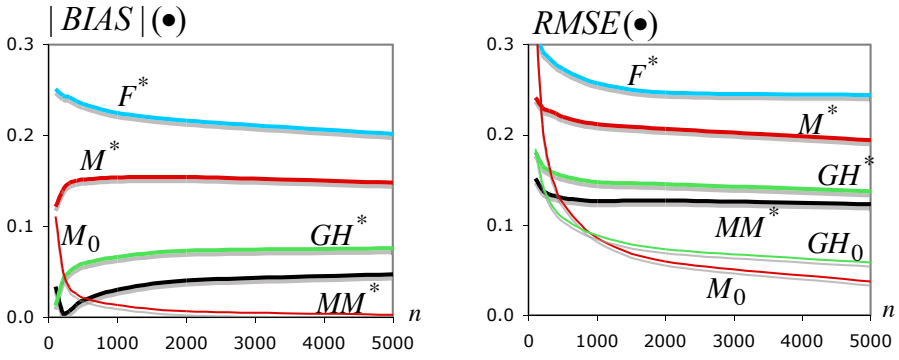

Figure 2: Absolute values of bias (*left*) and root mean squared errors (*right*) of the adaptive extreme value index estimators in (4.5), for an *extreme value* model with $\gamma = -0.1$.

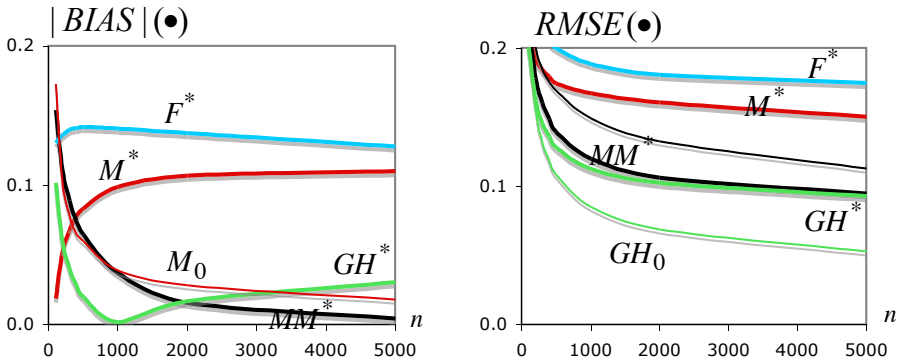

Figure 3: Absolute values of bias (*left*) and root mean squared errors (*right*) of the adaptive extreme value index estimators in (4.5), for an *extreme value* model with $\gamma = -0.3$.

A few remarks related with the adaptive estimators in (4.5) for underlying *EV* models:

- For $\gamma = -0.1$, the absolute bias of MM^* is the smallest one, except for $n = 100$. For this sample size, and regarding absolute bias, GH^* beats MM^*. Regarding *MSE*, the best of the adaptive estimates is MM^*, for all n. As γ decreases, and regarding bias, MM^* is replaced by GH^* for moderate n and by M^* for small n.
- For $\gamma = -0.3$, the absolute bias of MM^* is the smallest one, only for $n \geq 2000$. For $300 \leq n \leq 1000$, GH^* beats the other estimators. For $n \leq 200$, the smallest absolute bias is the one of M^*. Regarding *MSE*, the best of the estimates is GH^*, quite close to MM^* for all n.
- Notice the overall worst performance of the estimator F^*, essentially due to the region of γ-values under consideration.

Regarding the "potential" estimators T_0 at simulated optimal levels, with $T = M, GH, MM$ or F, we draw the following comments:

- At simulated optimal levels, GH_0 achieves the minimum *MSE* for all n, if $\gamma = -0.3$. For the other values of γ, GH_0 is the best one for small n, but M_0 becomes the best for large n ($n \geq 1000$ for $\gamma = -0.1$).
- Regarding smallest absolute bias at simulated optimal levels, M_0 is the best for all n, if $\gamma = -0.1$. For the other values of γ, M_0 is the best for $n \geq 200$. For $n = 100$, GH_0 overpasses all other ones.

In Table 1, for *EV* underlying parents, for a few values of n, and for $T = M, GH, MM$ and F, we present two relative efficiency indicators of T^*, in (4.5), relatively to T_0, and to S_0, the best T_0-estimator, i.e. the one with smallest *MSE*

at optimal level. With the notation $MSE_s(S_0) = \min\big(MSE_s(M_0), MSE_s(GH_0), MSE_s(MM_0), MSE_s(F_0)\big)$, we have simulated

$$REFF_1 := \sqrt{\frac{MSE_s(T_0)}{MSE_s(T^*)}} \qquad \text{and} \qquad REFF_2 := \sqrt{\frac{MSE_s(S_0)}{MSE_s(T^*)}}\ ,$$

the values placed in the first and second line, respectively, of any entrance T^*. Note that the higher than one these indicators are, the better T^* performs. Moreover, we obviously have $REFF_1 \geq REFF_2$. For all n, the highest $REFF_1$-indicator is written in **bold** and the highest $REFF_2$ indicator is written in *italic*. The results obtained are consistent with the remarks made above.

Table 1: Simulated *REFF*'s of the adaptive *EVI*-estimators under study, together with associated 95% CI's, for *extreme value* underlying parents.

***EV* parent**, $\gamma = -0.1$					
n	100	200	500	1000	2000
M^*	**1.403** $\pm$ 0.015 0.762 $\pm$ 0.008	0.921 $\pm$ 0.011 0.632 $\pm$ 0.006	0.508 $\pm$ 0.130 0.491 $\pm$ 0.006	0.404 $\pm$ 0.002 0.404 $\pm$ 0.002	0.289 $\pm$ 0.003 0.289 $\pm$ 0.003
MM^*	1.212 $\pm$ 0.009 *1.212* $\pm$ 0.009	**1.116** $\pm$ 0.016 *1.050* $\pm$ 0.014	**0.973** $\pm$ 0.016 *0.833* $\pm$ 0.014	**0.874** $\pm$ 0.013 *0.677* $\pm$ 0.010	**0.771** $\pm$ 0.014 *0.469* $\pm$ 0.007
GH^*	1.021 $\pm$ 0.010 1.015 $\pm$ 0.010	0.877 $\pm$ 0.010 0.877 $\pm$ 0.010	0.700 $\pm$ 0.009 0.700 $\pm$ 0.009	0.599 $\pm$ 0.009 0.579 $\pm$ 0.009	0.506 $\pm$ 0.006 0.410 $\pm$ 0.004
F^*	1.023 $\pm$ 0.009 0.587 $\pm$ 0.006	0.898 $\pm$ 0.002 0.496 $\pm$ 0.004	0.768 $\pm$ 0.003 0.395 $\pm$ 0.002	0.695 $\pm$ 0.004 0.332 $\pm$ 0.003	0.628 $\pm$ 0.003 0.243 $\pm$ 0.001

***EV* parent**, $\gamma = -0.3$					
n	100	200	500	1000	2000
M^*	**1.932** $\pm$ 0.020 0.978 $\pm$ 0.011	**1.355** $\pm$ 0.020 0.807 $\pm$ 0.005	0.896 $\pm$ 0.013 0.621 $\pm$ 0.006	0.657 $\pm$ 0.006 0.509 $\pm$ 0.008	0.489 $\pm$ 0.003 0.428 $\pm$ 0.003
MM^*	1.060 $\pm$ 0.006 0.962 $\pm$ 0.008	1.121 $\pm$ 0.009 0.871 $\pm$ 0.005	**1.214** $\pm$ 0.008 0.780 $\pm$ 0.006	**1.245** $\pm$ 0.014 0.711 $\pm$ 0.007	**1.246** $\pm$ 0.012 0.645 $\pm$ 0.006
GH^*	1.038 $\pm$ 0.010 *1.038* $\pm$ 0.010	0.959 $\pm$ 0.006 *0.959* $\pm$ 0.006	0.847 $\pm$ 0.010 *0.847* $\pm$ 0.010	0.757 $\pm$ 0.007 *0.757* $\pm$ 0.007	0.670 $\pm$ 0.008 *0.670* $\pm$ 0.008
F^*	1.097 $\pm$ 0.010 0.895 $\pm$ 0.009	0.959 $\pm$ 0.005 0.718 $\pm$ 0.006	0.788 $\pm$ 0.007 0.544 $\pm$ 0.005	0.693 $\pm$ 0.145 0.452 $\pm$ 0.005	0.597 $\pm$ 0.006 0.388 $\pm$ 0.004

The behaviour of the right endpoint semi-parametric estimators is quite erratic, even when we consider equation (3.12), to make them coherent with the data. Such a behaviour is even more catastrophic when we do not make them coherent with the data, and the most usual estimators in the literature are in fact "raw", in the sense that they have not been modified in order to be larger than the maximum in the sample, as needed. Indeed, alternative semi-parametric estimators of the right endpoint are urgently needed. The bias and the *RMSE* of the estimators in (4.6) almost overlap, and we see no reason to present figures similar to the ones drawn for the adaptive *EVI*-estimators in (4.5). A similar comment applies to the adaptive estimators of the "excellence indicator".

Table 2 is equivalent to Table 1, but for the adaptive right endpoint estimation. Similarly, Table 3 is equivalent to Table 1, now for the exceedance probability estimation (or equivalently, for the "excellence" indicator).

Table 2: Simulated *REFF*'s of the adaptive right endpoint estimators under study, together with associated 95% CI's, for *extreme value* underlying parents.

***EV* parent**, $\gamma = -0.1$					
n	100	200	500	1000	2000
M^*	3.272 ± 0.771 *0.877* ± 0.006	**2.786** ± 0.569 *0.865* ± 0.005	**1.660** ± 0.293 *0.866* ± 0.007	**1.001** ± 0.002 *0.863* ± 0.008	0.998 ± 0.000 *0.858* ± 0.005
MM^*	1.819 ± 0.483 *0.877* ± 0.006	1.000 ± 0.000 *0.865* ± 0.005	1.000 ± 0.294 *0.866* ± 0.007	1.000 ± 0.000 *0.863* ± 0.008	**1.000** ± 0.000 *0.858* ± 0.005
GH^*	**5.454** ± 1.139 *0.877* ± 0.006	2.106 ± 0.698 *0.865* ± 0.005	1.000 ± 0.000 *0.866* ± 0.007	1.000 ± 0.000 *0.863* ± 0.008	**1.000** ± 0.000 *0.858* ± 0.005
F^*	0.877 ± 0.006 *0.877* ± 0.006	0.865 ± 0.005 *0.865* ± 0.005	0.866 ± 0.007 *0.866* ± 0.007	0.863 ± 0.008 *0.863* ± 0.008	0.858 ± 0.005 *0.858* ± 0.005

***EV* parent**, $\gamma = -0.3$					
n	100	200	500	1000	2000
M^*	8.116 ± 1.817 0.810 ± 0.003	**3.586** ± 0.728 0.793 ± 0.005	1.001 ± 0.001 0.779 ± 0.006	0.988 ± 0.007 *0.773* ± 0.005	0.947 ± 0.004 *0.765* ± 0.005
MM^*	**8.118** ± 1.926 0.844 ± 0.005	1.429 ± 0.440 0.809 ± 0.006	**1.002** ± 0.003 0.779 ± 0.006	**1.000** $\pm$ **0.000** *0.773* ± 0.005	**1.000** $\pm$ **0.000** *0.765* ± 0.005
GH^*	0.875 ± 0.007 0.828 ± 0.003	0.822 ± 0.009 0.804 ± 0.006	0.779 ± 0.006 0.779 ± 0.006	0.773 ± 0.005 *0.773* ± 0.005	0.765 ± 0.005 *0.765* ± 0.005
F^*	0.875 ± 0.007 *0.875* ± 0.007	0.822 ± 0.009 *0.822* ± 0.009	0.779 ± 0.006 *0.779* ± 0.006	0.773 ± 0.005 *0.773* ± 0.005	0.768 ± 0.003 *0.765* ± 0.005

Table 3: Simulated *REFF*'s of the adaptive estimators of exceedance probabilities of $x_{n:n}$, and associated 95% CI's, for *extreme value* underlying parents.

***EV* parent**, $\gamma = -0.1$					
n	100	200	500	1000	2000
M^*	4.149 ± 0.899 0.698 ± 0.005	**5.265** ± 0.923 0.700 ± 0.009	**2.257** ± 0.293 0.657 ± 0.015	1.259 ± 0.021 0.590 ± 0.011	1.048 ± 0.021 0.537 ± 0.004
MM^*	3.343 ± 1.369 0.733 ± 0.008	1.518 ± 0.043 0.702 ± 0.014	1.453 ± 0.294 0.641 ± 0.015	1.394 ± 0.016 0.579 ± 0.011	1.371 ± 0.011 0.533 ± 0.004
GH^*	**7.690** ± 1.944 0.702 ± 0.007	3.349 ± 1.020 0.680 ± 0.013	1.498 ± 0.036 0.629 ± 0.015	**1.422** ± 0.017 0.570 ± 0.011	**1.387** ± 0.012 0.525 ± 0.004
F^*	0.858 ± 0.010 *0.858* ± 0.010	0.793 ± 0.016 *0.793* ± 0.016	0.710 ± 0.017 *0.710* ± 0.017	0.637 ± 0.013 *0.637* ± 0.013	0.580 ± 0.005 *0.580* ± 0.005

***EV* parent**, $\gamma = -0.3$					
n	100	200	500	1000	2000
M^*	12.177 ± 3.181 0.587 ± 0.006	5.746 ± 1.475 0.601 ± 0.018	1.774 ± 0.149 0.580 ± 0.022	1.111 ± 0.040 0.545 ± 0.029	0.920 ± 0.039 0.540 ± 0.022
MM^*	**24.493** ± 6.833 0.702 ± 0.015	8.523 ± 3.839 0.722 ± 0.038	**6.527** ± 0.426 0.692 ± 0.039	**6.928** ± 0.343 0.633 ± 0.048	**4.865** ± 0.305 0.623 ± 0.034
GH^*	8.535 ± 3.339 0.518 ± 0.008	**9.516** ± 5.754 0.583 ± 0.024	1.840 ± 0.095 0.614 ± 0.033	0.812 ± 0.047 0.586 ± 0.041	0.601 ± 0.034 0.592 ± 0.029
F^*	0.865 ± 0.012 *0.865* ± 0.012	0.787 ± 0.029 *0.787* ± 0.029	0.712 ± 0.031 *0.712* ± 0.031	0.641 ± 0.040 *0.641* ± 0.040	0.618 ± 0.029 *0.618* ± 0.029

On the basis of the simulated results, the adaptive estimation procedure seems to provide interesting results, in the sense that we have obtained *REFF* indicators reasonably high for small n and all parameters of interest. Regarding *EVI*-estimation, and despite of the fact that it is not possible to claim that MM^* has, for all models in $\mathcal{D}_{\mathcal{M}}(G_\gamma)$, $\gamma < 0$, the best performance among the four adaptive estimators in (4.5), it is clear that if we have to elect one of these four adaptive estimators, we are inclined to the choice of MM^*. This is particularly if the model is not a long way from an *EV* model, and we have a light indication for this underlying parent, not only on the basis of the undertaken parametric data analysis in Section 5.1, but also due to the nature of the data. This is the reason why in Section 5.3.2, we shall compute the final estimates of γ on the basis of MM^*. Note however that, for small n, GH^* is also a serious alternative. For the right endpoint estimation all adaptive estimators in (4.6) are almost equivalent, and we thus see no reason not to use also x^*_{MM}. A similar comment applies to the estimators of the exceedance probability (or equivalently, of the excellence indicator).

5. DATA ANALYSIS OF INDOOR ATHLETIC EVENTS

The data under analysis are related to three running and three jumping events, all for men, the 60 Metres Hurdles (60MH), 400 (400M) and 1500 Metres (1500M), as well as the high jump (HJ), long jump (LJ) and pole vault (PV). The *sources* were `http://www.iaaf.org/statistics/toplists/index.htmx` and `http://hem.bredband.net/athletics/athletics_all-time_best.htm`. Data was collected until the end of 2007 and for any athlete only the best mark was taken into account. As mentioned before, we are dealing with right-tails. Consequently, for all running events we have converted *running times* into *speeds*, i.e., 10.00 seconds in the 60MH (equal to 0.06 kilometers) is transformed to a speed of $3600 \times 0.06/10 = 21.6$ km/h. Like this, the higher the speed, the better, just as the higher the jump, the better. Contrarily to what has been done in Einmahl and Smeets (2009), we have not paid attention to doping related times, and we are conscious that slightly different estimates could then be obtained, despite of the usual robustness of the methods to a few outliers in the data.

5.1. Parametric data analysis

Prior to a semi-parametric analysis of the data, the most common framework of *statistics of extremes*, we shall proceed to a parametric data analysis, in the lines of Robinson and Tawn (1995) and Barão and Tawn (1999), who considered the annual best times in the women's 3000m event. Also Smith (1988) has

proposed a maximum likelihood method of fitting models to a series of records, and applied his method to athletics records for the mile and the marathon. The attempts made in these papers to predict an ultimate world record are based on the development of top performances over time. This is not the case in this paper. Here, as in Einmahl and Magnus (2008), as well as in Einmahl and Smeets (2009), we are not interested in predicting the world record in the future. We are using only the top performances associated with a set of n athletes, and consequently, our estimated ultimate record tells us what, in principle, is possible at this moment, given today's knowledge and material.

We first illustrate in Figures 4 and 5, the Gumbel QQ-plots associated with all data sets under analysis. In all figures we have thus plotted the points $\big(x_{i:n},\, p_i^{\Lambda} = -\ln(-\ln(i/(n+1)))\big)$, $1 \le i \le n$, and proceeded to the fitting of a least-squares line.

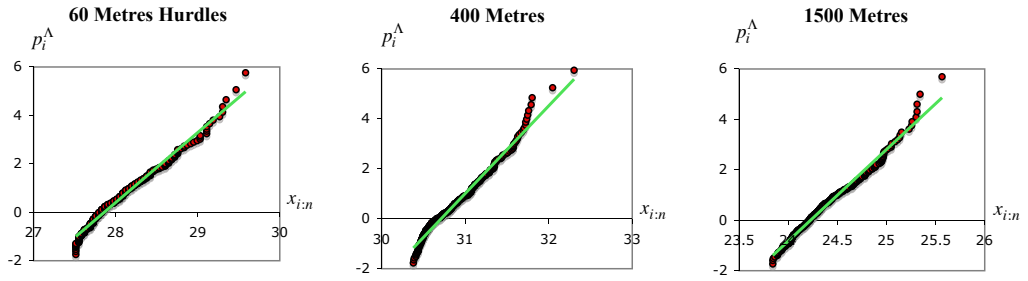

Figure 4: Gumbel QQ-plot related to the *running events* under analysis — 60 Metres Hurdles, 400 and 1500 Metres.

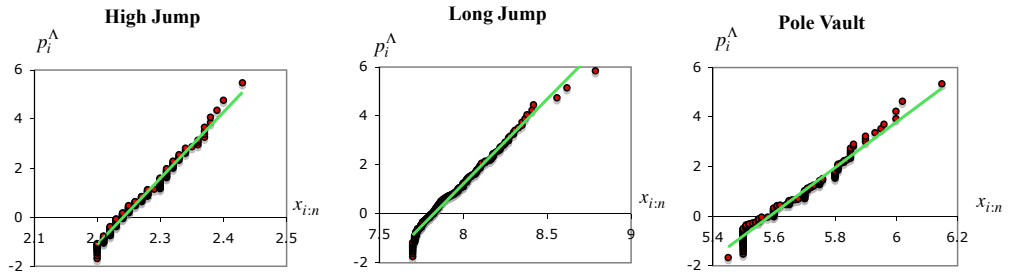

Figure 5: Gumbel QQ-plot related to the *jumping events* under analysis — High Jump, Long Jump and Pole Vault.

Apart from the Long Jump event, where $\gamma = 0$ can perhaps provide a reasonable fit to the right-tail, despite of a slight deviation of top o.s.'s smaller than the third largest value, all other events exhibit a light right-tail, i.e. a negative extreme value index, and consequently a finite right endpoint x^*.

Due to the fact that the observed data considered are already maxima, possibly of a small and dependent number of marks associated with any of the n athletes, but the *extreme value* limiting law, in (1.1), is "robust" to changes of the i.i.d. assumption, we have first tried the fitting, through maximum likelihood, of an extreme value model $F(x;\lambda,\delta,\gamma) = G_\gamma\big((x-\lambda)/\delta\big)$, with $G_\gamma(x)$ given in (1.1). We have used the EVIR package in the *R*-software. The estimate of the right endpoint is then provided by $\hat{x}^* = \max(x_{n:n}, \hat{\lambda}-\hat{\delta}/\hat{\gamma})$, with $(\hat{\lambda},\hat{\delta},\hat{\gamma})$ the maximum likelihood estimates of the unknown parameters, (λ,δ,γ). The results obtained are presented in Table 4.

Table 4: Maximum likelihood estimates of $(\lambda,\delta,\gamma,x^*)$ for an underlying model $G_\gamma\big((x-\lambda)/\delta\big)$, with $G_\gamma(x)$ given in (1.1): $'$ – Km/h, $''$ – metres.

Event	n	$(x_{1:n}, x_{n:n})$	$\hat{\lambda}$	$\hat{\delta}$	$\hat{\gamma}$ (95% CI)	$\hat{x}^*$
60 MH	312	$(27.52, 29.59)'$	27.84	0.28	-0.21 $(-0.328, -0.090)$	29.59
400 M	380	$(30.38, 32.31)'$	30.70	0.25	-0.15 $(-0.277, -0.024)$	32.36
1500 M	296	$(23.84, 25.57)'$	24.23	0.26	-0.06 $(-0.166, +0.042)$	28.33
HJ	235	$(2.20, 2.43)''$	2.24	0.03	-0.09 $(-0.223, +0.040)$	2.61
LJ	340	$(7.70, 8.79)''$	7.81	0.11	-0.26 $(-0.392, -0.130)$	8.79
PV	205	$(5.45, 6.15)''$	5.58	0.09	-0.15 $(-0.342, +0.041)$	6.21

As expected, all estimates of γ are negative. But for the 1500 Metres, High Jump and Pole Vault, the upper limits of the associated 95% CI's are positive, suggesting that the value $\gamma = 0$ could possibly be adequate. The estimation of the right endpoint, which provides estimates equal to the maximum value in the data, the value $x_{n:n}$, for two of the athletic events, 60 Metres Hurdles and Long Jump, can be considered slightly problematic.

5.2. Fitting the extreme value model

In Figure 6, we picture in *light grey* the asymptotic 95% critical values (CV), $1.36/\sqrt{n}$, of the Kolmogorov–Smirnov statistic for testing a model without unknown parameters. The observed values of the Kolmogorov–Smirnov statistic, $KS_n := \max_{1\le i\le n}\Big(\big|G_{\hat{\gamma}}\big((x_{i:n}-\hat{\lambda})/\hat{\delta}\big) - i/n\big|,\ \big|G_{\hat{\gamma}}\big((x_{i:n}-\hat{\lambda})/\hat{\delta}\big) - (i-1)/n\big|\Big)$ are pictured in *black*. The simulated 95% critical points of the Kolmogorov–Smirnov statistic, for testing an extreme value model $G_{\hat{\gamma}}\big((x-\hat{\lambda})/\hat{\delta}\big)$, have been based on 1000 runs, and are pictured in *grey*, showing again the "conservative property" of the Kolmogorov–Smirnov test — if we are led to rejection of a model without taking into account the maximum likelihood estimation of the parameters, we are *a fortiori* led to a rejection of the same model whenever we appropriately estimate the unknown parameters through the maximum likelihood approach.

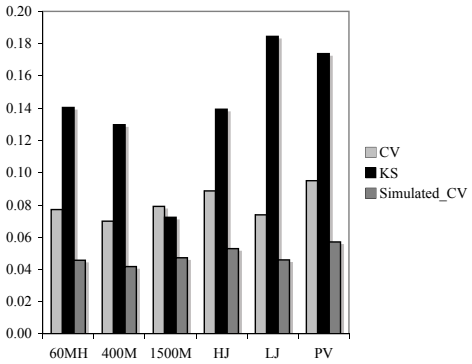

Figure 6: Asymptotic critical values CV= $1.36/\sqrt{n}$ (*light grey*), simulated critical values (*grey*) and observed values (*black*) of the Kolmogorov–Smirnov statistic, for all athletic events under analysis.

At the significance level $\alpha = 0.05$, the hypothesis of a (unified) *extreme value* model has thus been rejected by the Kolmogorov–Smirnov test for all data sets, as could also have been inferred graphically from Figure 7 and Figure 8, where we picture the empirical d.f., in grey, and the fitted extreme value d.f., in black.

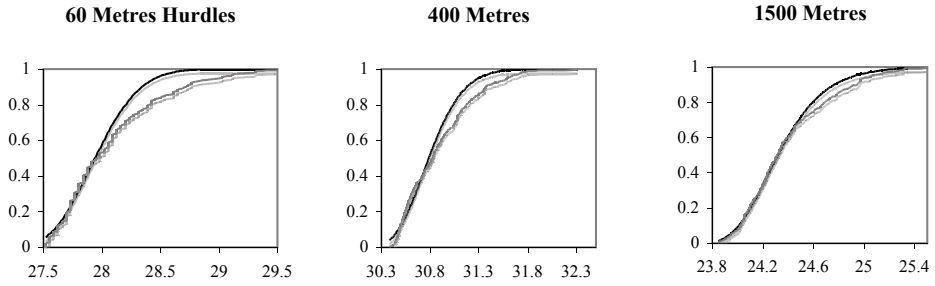

Figure 7: Empirical d.f. (*grey*) and fitted extreme value d.f. (*black*) for the running events under analysis.

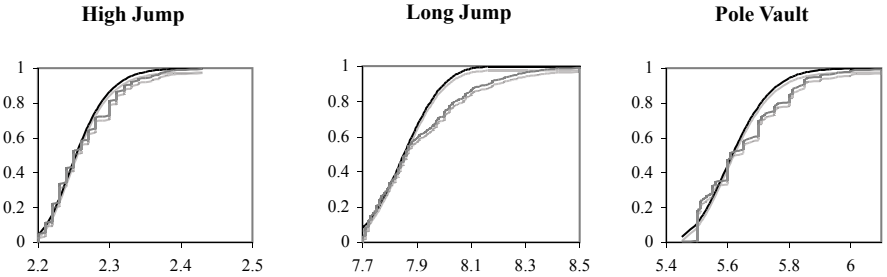

Figure 8: Empirical d.f. (*grey*) and fitted extreme value d.f. (*black*) for the jumping events under analysis.

Alternative parametric models have even provided worse fitting results. There is thus a claim for the need of a semi-parametric data analysis, to be developed next, in Section 5.3.

5.3. A semi-parametric data analysis

5.3.1. Testing the extreme value index sign

As mentioned before, whenever we place ourselves under a semi-parametric framework, we assume only that (2.4) holds, or equivalently, that $F \in \mathcal{D}_{\mathcal{M}}(G_\gamma)$, for a certain γ, being γ the primordial parameter of extreme events.

In many areas where extremes are relevant, the simplest case $\gamma = 0$ is often considered. Moreover, if we clearly think that $\gamma < 0$ or that $\gamma > 0$, we have specific procedures for the estimation of γ, possibly more reliable than the procedures valid for a general $\gamma \in \mathbb{R}$. Prior to a deeper semi-parametric analysis of the tail associated with this type of data, it thus seems sensible to test

$$
\begin{array}{ll}
H_0: & F \in \mathcal{D}_{\mathcal{M}}(G_\gamma)_{\gamma=0} \quad \big(\text{or } \ F \in \mathcal{D}_{\mathcal{M}}(G_\gamma)_{\gamma\geq 0}\big) \\
& \text{versus} \\
H_1: & F \in \mathcal{D}_{\mathcal{M}}(G_\gamma)_{\gamma<0} \ ,
\end{array}
\tag{5.1}
$$

through the use of any semi-parametric test statistic.

We shall consider here two test statistics of a similar type, i.e. both based on the excesses over a high random threshold $X_{n-k:n}$, with k satisfying (3.7). The first one was introduced by Greenwood (1946) and the second one by Hasofer and Wang (1992). These two statistics were further studied, under a semi-parametric framework, by Neves and Fraga Alves (2007). They are given by

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

and

$$
W_{k,n} := \frac{1}{k(G_{k,n}-1)} \ .
$$

Under the null hypothesis H_0 in (5.1) and extra mild conditions on the right-tail of F and on the growth of $k = k_n$, they both have an asymptotic normal behaviour. More specifically,

$$
G^*_{k,n} := \sqrt{k/4}\,\big(G_{k,n}-2\big)\,\big|_{F\in\mathcal{D}_{\mathcal{M}}(G_0)} \ \xrightarrow[n\to\infty]{d} \ N(0,1) \tag{5.2}
$$

and

$$
W^*_{k,n} := \sqrt{k/4}\,\big(k\,W_{k,n}-1\big)\,\big|_{F\in\mathcal{D}_{\mathcal{M}}(G_0)} \ \xrightarrow[n\to\infty]{d} \ N(0,1) \ . \tag{5.3}
$$

Motivated by the important contribution of the maximum to the sum of the k excesses, $X_{n-i+1:n} - X_{n-k:n}$, $1 \le i \le k$, Neves *et al.* (2006) introduced the following complimentary statistic,

$$R_{k,n} := \frac{X_{n:n} - X_{n-k:n}}{\frac{1}{k}\sum_{i=1}^{k} X_{n-i+1:n} - X_{n-k:n}} ,$$

also considered in the analysis of the data under study. The asymptotic behaviour of $R_{k,n}$ is provided by the Gumbel d.f., $\Lambda = G_0$, with G_γ given in (1.1). More specifically,

$$R^*_{k,n} := R_{k,n} - \ln k \,\big|_{F \in \mathcal{D}_{\mathcal{M}}(G_0)} \xrightarrow[n\to\infty]{d} Z \frown G_0 \ . \tag{5.4}$$

As a function of k both $G^*_{k,n}$ and $R^*_{k,n}$ tend to have a slope with the sign of γ. The statistic $W^*_{k,n}$ works the other way round.

As an illustration, we present, in Figure 9, the sample paths of the three test statistics $G^*_{k,n}$, $W^*_{k,n}$ and $R^*_{k,n}$ in (5.2), (5.3) and (5.4), respectively, associated with the Long Jump. In this figure we also picture the quantiles $\left(\chi^{\bullet}_{0.025}, \chi^{\bullet}_{0.975}\right)$ of the standard normal Φ, equal to $(-1.96, +1.96)$, and of the standard Gumbel $\Lambda \equiv G_0$, equal to $(-1.31, +3.68)$.

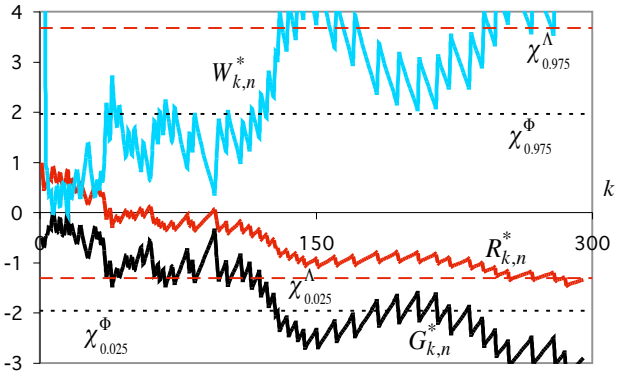

Figure 9: Sample paths of the test statistics for the Long Jump event.

For all other data sets under analysis the graphs are quite similar, showing clearly a decreasing trend of $R^*_{k,n}$ and $G^*_{k,n}$ (with $G^*_{k,n}$ below $\chi^{\Phi}_{0.025}$ for a large number of k-values), as well as an increasing trend of $W^*_{k,n}$ (above $\chi^{\Phi}_{0.975}$ for moderate up to large values of k). Such a trend is mainly related to bias, but bias is strongly related to the extreme value index sign. These results provide a strong suggestion of a negative extreme value index, as expected. Despite of that, notice that, in Figure 9, the sample path of $R^*_{k,n}$ is within the 95% CI for almost all k-values. This was also expected, because it is well known (see, for instance, Neves and Fraga Alves, 2008) that $R^*_{k,n}$ tends to be a conservative test and the true value of γ is for sure close to zero.

5.3.2. Semi-parametric estimates of the extreme value index and the right endpoint

In Table 5 we present a summary of the performed data analysis, with estimates and 95% CI's for the extreme value index γ. These estimates of γ were obtained through the mixed moment (MM) estimates, computed at the value $k^*_{\min}$, in (4.4), i.e. they are the adaptive estimate MM^* in (4.5).

Table 5: Estimates of the extreme value index, based on M^*: $'$ – Km/h, $''$ – metres.

Event	n	$(x_{1:n}, x_{n:n})$	MM^* (95% CI)	$k^*_{\min}$
60 MH	312	$(27.52, 29.59)'$	-0.34 $(-0.469, -0.214)$	305
400 M	380	$(30.38, 32.31)'$	-0.26 $(-0.445, -0.080)$	128
1500 M	296	$(23.84, 25.57)'$	-0.38 $(-0.520, -0.241)$	275
HJ	235	$(2.20, 2.43)''$	-0.32 $(-0.468, -0.173)$	219
LJ	340	$(7.70, 8.79)''$	-0.20 $(-0.315, -0.087)$	296
PV	205	$(5.45, 6.15)''$	-0.31 $(-0.472, -0.151)$	182

In this semi-parametric data analysis, we have also considered the adaptive estimators MM^{**} and ML^{**}, the estimators in (3.6) and (3.10), respectively, computed at the value $k^{**}_{\min}$, obtained through a minimization procedure of the type of the one in (4.4), but including also the ML-estimator. The reason for the consideration of the ML-estimator lies on the fact that in the region $-1/2 < \gamma < 0$, where the estimates lie, $\sigma^2_{ML} = (1+\gamma)^2$ is smaller than $\sigma^2_{MM} = \sigma^2_M = (1-\gamma)^2\,(1-2\,\gamma)\,(1-\gamma+6\,\gamma^2)/\big((1-3\,\gamma)\,(1-4\,\gamma)\big)$ for all γ, with $\sigma_\bullet$ the asymptotic standard deviation in the asymptotic representation (4.1) (see Gomes and Neves, 2008, for further details). These estimates are presented in Table 6. For the LJ athletic event $k^*_{\min} = k^{**}_{\min}$. Then, the estimates $MM^* = MM^{**}$ and associated CI's are written in *italic*.

Table 6: Estimates of the extreme value index, based on ML^{**} and MM^{**}.

Event	n	ML^{**} (95% CI)	MM^{**} (95% CI)	$k^{**}_{\min}$
60 MH	312	-0.30 $(-0.377, -0.216)$	-0.31 $(-0.438, -0.186)$	294
400 M	380	-0.22 $(-0.351, -0.085)$	-0.26 $(-0.434, -0.077)$	133
1500 M	296	-0.32 $(-0.400, -0.239)$	-0.31 $(-0.440, -0.180)$	273
HJ	235	-0.29 $(-0.387, -0.201)$	-0.31 $(-0.456, -0.165)$	220
LJ	340	-0.17 $(-0.262, -0.073)$	*-0.20 $(-0.315, -0.087)$*	296
PV	205	-0.29 $(-0.396, -0.191)$	-0.30 $(-0.458, -0.142)$	183

As it can be seen from Tables 5 and 6, there is only a small difference between $k^*_{\min}$ and $k^{**}_{\min}$, as expected. All semi-parametric γ-estimates at $k = k^{**}_{\min}$ are within the CI's provided in Table 5 and based on MM^*. Similarly, all estimates in Table 5 are within the CI's provided in Table 6. However, apart from the parametric estimates of γ associated with the 400 Metres and Long Jump events, the parametric estimates in Table 4 are outside the CI's provided in Table 5, as well as the other way round. The parametric estimates are above the semi-parametric estimates for the six events considered. Note also that, contrarily to what generally happens, the values $k^*_{\min}$ and $k^{**}_{\min}$ are quite large, comparatively with the sample size n. This is essentially due to the fact that, for large k, the samples paths of the different estimators are reasonably stable as functions of k and close to each other (a small bias, contrarily to the most common situations in practice) and volatile for small k (large variance for small k, as usual).

Also as an illustration, we present, in Figure 10, the estimates $M \equiv M_{k,n}$, $GH \equiv GH_{k,n}$, $MM \equiv MM_{k,n}$, and $F \equiv F_{k,n}$ of γ, defined in (3.3), (3.5), (3.6) and (3.8), respectively, again for the Long Jump athletic event. We also picture the sample paths of the γ-estimator $ML \equiv ML_{k,n}$, in (3.10).

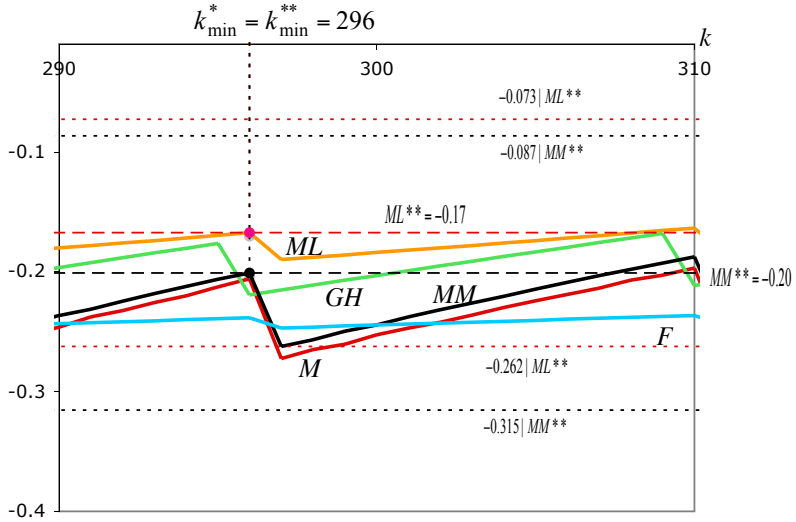

Figure 10: Sample paths of the extreme value index estimates under consideration, for the Long Jump event.

Analogously, and for the estimation of the right endpoint, apart from the adaptive estimators $\hat{x}^*_{MM}$, the estimators $\hat{x}^*_{k,n|\bullet}$, in (3.12), for $\bullet \equiv MM$, computed at the value k^{*x}, in (4.7), we have also considered the adaptive estimators $\hat{x}^{**}_{MM}$ and $\hat{x}^{**}_{ML}$, the estimators $\hat{x}^*_{k,n|\bullet}$ in (3.12) for $\bullet \equiv MM$ and ML, computed at the value $k^{**x}_{\min} \equiv k^{***x}$, obtained through a minimization procedure of the type of the one used for the adaptive endpoint estimators in (4.6), but including also the

ML-estimator. Similarly, and as mentioned before, for the estimation of the excellence indicator, we use the notation $\widehat{E}_{\bullet}^{**}$ for the estimator $\widehat{E}_{k,n}^{\bullet}$, in (3.13), computed at the value $k_{\min}^{**E} \equiv k^{**E}$. Next, in Table 7 we present the estimates of the right endpoints of the models underlying the different data sets under study.

Table 7: Estimates of the right endpoint: $'$ – km/h, $^\bullet$ – minutes, $''$ – metres.

Event	$x_{n:n}$	k^{*x}	$\hat{x}_{MM}^{*}$	k^{**x}	$\hat{x}_{ML}^{**}$
60 MH	$29.59'$ $(00:07.30)^\bullet$	53	$29.81'$ $(00:07.25)^\bullet$	53	$29.71'$ $(00:07.27)^\bullet$
400 M	$32.31'$ $(00:44.57)^\bullet$	32	$32.45'$ $(00:44.37)^\bullet$	128	$32.68'$ $(00:44.06)^\bullet$
1500 M	$25.57'$ $(03:31.18)^\bullet$	119	$25.63'$ $(03:30.69)^\bullet$	119	$25.70'$ $(03:3012)^\bullet$
HJ	$2.43''$	219	$2.44''$	219	$2.46''$
LJ	$8.79''$	144	$8.84''$	281	$9.12''$
PV	$6.15''$	82	$6.16''$	182	$6.22''$

In Table 8 we present the estimates of the associated "excellence" indicators of the levels $x_{\rm H} = x_{n:n}$, provided in (3.13). Note that for all data sets we got $k^{*E} = k^{**E}$, smaller than expected for some of the data sets (60 MH, 1500 M and LJ).

Table 8: Estimates of an "excellence" indicator of the level $x_{n:n}$.

Event	k^{*x}	$\widehat{E}_{k^{*x},n}^{MM}$	k^{**x}	$\widehat{E}_{k^{**x},n}^{ML}$	$k_{\min}^{*E} = k_{\min}^{**E}$	$\widehat{E}_{MM}^{*} \mid \widehat{E}_{ML}^{*}$
60 MH	53	0.66	53	0.67	11	0.62 | 0.72
400 M	32	0.98	128	0.88	148	0.99 | 0.90
1500 M	119	0.95	119	0.82	36	0.89 | 0.81
HJ	219	0.98	219	0.90	222	0.91 | 0.89
LJ	144	0.99	281	0.92	39	0.80 | 0.78
PV	82	0.98	182	0.94	132	0.99 | 0.94

Despite of slight discrepancies of the different estimates of the relevant parameters of extreme events, the results in Tables 5, 6, 7 and 8 mean that, under the present conditions, there are finite upper limits for all jumping events under analysis, as well as finite lower limits in the times associated with all running events under analysis. From the "excellence" indicators of the world records, we can say that the current 400 Metres, High Jump and Pole Vault world records are very good (indicators above 89%). The lowest "excellence" indicator, around 65%, corresponds to the 60 Metres Hurdles.

ACKNOWLEDGMENTS

Research partially supported by FCT/OE and PTDC/FEDER.

REFERENCES

[1] Barão, M.I. and Tawn, J. (1999). Extremal analysis of short series with outliers: sea-levels and athletic records, *Applied Statistics*, **48**, 469–487.

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

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

[4] Caeiro, C.; Gomes, M.I. and Pestana, D. (2005). Direct reduction of bias of the classical Hill estimator, *Revstat*, **3**(2), 113–136.

[5] Danielsson, J.; Haan, L. de; Peng, L. and de Vries, C.G. (2001). Using a bootstrap method to choose the sample fraction in the tail index estimation, *J. Multivariate Analysis*, **76**, 226–248.

[6] Davison, A.C. (1984). *Modelling excesses over high thresholds.* In "Statistical Extremes and Applications" (J. Tiago de Oliveira, Ed.), D. Reidel, Dordrecht, Holland, 461–482.

[7] Dekkers, A.; Einmahl, J. and de Haan, L. (1989). A moment estimator for the index of an extreme-value distribution, *Annals of Statistics*, **17**, 1833–1855.

[8] Draisma, G.; Haan, L. de; Peng, L. and Pereira, T.T. (1999). A bootstrap-based method to achieve optimality in estimating the extreme-value index, *Extremes*, **2**, 367–404.

[9] Drees, H.; Ferreira, A. and de Haan, L. (2004). On maximum likelihood estimation of the extreme value index, *Annals of Applied Probability*, **14**, 1179–1201.

[10] Einmahl, J. and Magnus, J.R. (2008). Records in athletics through extreme-value theory, *J. American Statistical Association*, **103**, 1382–1391.

[11] Einmahl, J. and Smeets, S.G.W.R. (2011). Ultimate 100-m world records through extreme-value theory, *Statistica Neerlandica*, **65**(1), 32–42.

[12] Falk, M. (1995). Some best parameter estimates for distributions with finite endpoint, *Statistics*, **27**(1)–(2), 115–125.

[13] Fraga Alves, M.I.; de Haan, L. and Lin, T. (2003). Estimation of the parameter controlling the speed of convergence in extreme value theory, *Mathematical Methods of Statistics*, **12**(2), 155–176.

[14] Fraga Alves, M.I.; Gomes, M.I.; de Haan, L. and Neves, C. (2009). The mixed moment estimator and location invariant alternatives, *Extremes*, **12**, 149–185.

[15] Gnedenko, B.V. (1943). Sur la distribution limite du terme maximum d'une série aléatoire, *Annals of Mathematics*, **44**(6), 423–453.

[16] Gomes, M.I. and Neves, C. (2008). Asymptotic comparison of the mixed moment and classical extreme value index estimators, *Statistics and Probability Letters*, **78**(6), 643–653.

[17] Gomes, M.I. and Oliveira, O. (2001). The bootstrap methodology in Statistics of Extremes: choice of the optimal sample fraction, *Extremes*, **4**(4), 331–358.

[18] Gomes, M.I. and Pestana, D. (2007). A sturdy reduced bias extreme quantile (VaR) estimator, *J. American Statistical Association*, **102**(477), 280–292.

[19] Gomes, M.I.; Martins, M.J. and Neves, M. (2007). Improving second order reduced bias extreme value index estimation, *Revstat*, **5**(2), 177–207.

[20] 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 Statistical Society*, **B70**(1), 31–52.

[21] Greenwood, M. (1946). The statistical study of infectious diseases, *J. Royal Statistical Society*, **A109**, 85–109.

[22] Haan, L. de (1970). *On Regular Variation and its Application to the Weak Convergence of Sample Extremes*, Mathematical Centre Tract 32, Amsterdam.

[23] Haan, L. de (1984). *Slow variation and characterization of domains of attraction.* In "Statistical Extremes and Applications" (J. Tiago de Oliveira, Ed.), D. Reidel, Dordrecht, Holland, 31–48.

[24] Haan, L. de and Ferreira, A. (2006). *Extreme Value Theory: an Introduction*, Springer, USA.

[25] Hasofer, A. and Wang, J.Z. (1992). A test for extreme value domain of attraction, *J. American Statistical Association*, **87**, 171–177.

[26] Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution, *Annals of Statistics*, **3**(5), 1163–1174.

[27] Neves, C. and Fraga Alves, M.I. (2007). Semi-parametric approach to Hasofer-Wang and Greenwood statistics in extremes, *Test*, **16**, 297–313.

[28] Neves, C. and Fraga Alves, M.I. (2008). Testing extreme value conditions — an overview and recent approaches, *Revstat*, **6**(1), 83–100.

[29] Neves, C.; Picek, J. and Fraga Alves, M.I. (2006). The contribution of the maximum to the sum of excesses for testing max-domains of attraction, *J. Statistical Planning and Inference*, **136**(4), 1281–1301.

[30] Robinson, M.E. and Tawn, J. (1995). Statistics for exceptional athletic records, *Applied Statistics*, **44**, 499–511.

[31] Smith, R.L. (1987). Estimating tails of probability distributions, *Annals of Statistics*, **15**(3), 1174–1207.

[32] Smith, R.L. (1988). Forecasting records by maximum likelihood, *J. American Statistical Association*, **83**, 331–338.

[33] Zhou, C. (2009). Existence and consistency of the maximum likelihood estimator for the extreme value index, *J. Multivariate Analysis*, **100**(4), 794–815.

[34] Zhou, C. (2010). The extent of the maximum likelihood estimator for the extreme value index, *J. Multivariate Analysis*, **101**(4), 971–983.

REVSTAT – Statistical Journal
Volume 9, Number 2, June 2011, 155–180

A SPATIAL UNIT LEVEL MODEL FOR SMALL AREA ESTIMATION

Authors: Pedro S. Coelho
– ISEGI – Universidade Nova de Lisboa, Portugal
Faculty of Economics, Ljubljana University, Slovenia
psc@isegi.unl.pt

Luís N. Pereira
– Escola Superior de Gestão, Hotelaria e Turismo,
Centro de Investigação sobre o Espaço e as Organizações,
Universidade do Algarve, Portugal
lmper@ualg.pt

Received: May 2010 Revised: April 2011 Accepted: May 2011

Abstract:

- This paper approaches the problem of small area estimation in the framework of spatially correlated data. We propose a class of estimators allowing the integration of sample information of a spatial nature. Those estimators are based on linear models with spatially correlated small area effects where the neighbourhood structure is a function of the distance between small areas. Within a Monte Carlo simulation study we analyze the merits of the proposed estimators in comparison to several traditional estimators. We conclude that the proposed estimators can compete in precision with competitive estimators, while allowing significant reductions in bias. Their merits are particularly conspicuous when analyzing their conditional properties.

Key-Words:

- *combined estimator; empirical best linear unbiased prediction; small area estimation; spatial models; unit level models.*

AMS Subject Classification:

- 62D05, 62F40, 62J05.

1. INTRODUCTION

Sample survey data are extensively used to provide reliable direct estimates of parameters of interest for the whole population and for domains of different kinds and sizes. When the domains were not originally planned, they usually are poorly represented in the sample or even not represent at all. These domains are called small areas and they usually correspond to small geographical areas, such as a municipality or a census division, or a small subpopulation like a particular economic activity or a subgroup of people obtained by cross-classification of demographic characteristics. Traditionally, sample sizes are chosen to provide reliable estimates for large domains and the lack of sample data from the target small area seriously affects the precision of estimates obtained from area-specific direct estimators. This fact has given rise to the development of various types of estimators that combine both the survey data for the target small areas and auxiliary information from sources outside the survey, often related to recent censuses and current administrative data, in order to increase precision. Under this context, the use of indirect estimators has been extensively applied. Such indirect estimators are based on either implicit or explicit models that provide a link to related small areas through auxiliary data.

Although traditional indirect estimators based on implicit models, which include synthetic and composite estimators, are easy to apply, they usually present undesirable properties. For that reason, other model based methods of small area estimation have been suggested in the literature. These methods can make specific allowance for local variation through complex error structures in the models that link the small areas, can be validated from the sample data and can handle complex cases such as cross-sectional, time series and spatial data. Such methods are often based on explicit Linear Mixed Models. The Best Linear Unbiased Prediction (BLUP) approach, using Henderson's method ([13]), is the most popular technique for estimating small area parameters of interest (usually the mean or the total). Under this approach and from the model point of view the small area parameters of interest are functions of fixed ($\boldsymbol{\beta}$) and random ($\mathbf{u}$) effects. Consequently, the prediction of small area parameters of interest is based on the estimation/prediction of these model effects. In practice this type of models always involves unknown variance components in the variance-covariance structure of random effects. When these unknown components are substituted by consistent estimates the resulting estimator is usually named as Empirical Best Linear Unbiased Predictor (EBLUP).

In the context of unit level spatial data, little work has been done on model-based methods of small area estimation. [25], [26] and [5] proposed a spatial unit level random effects model with spatial dependence incorporated in the error structure through a simultaneous autoregressive (SAR) error process.

Other findings are due to [4], [36], [30], [31], [32] and [40]. All these approaches consider a contiguity matrix to describe the neighbourhood structure between small areas. Nevertheless, there has been a lack of work regarding the explicit modeling of spatial correlation as a function of the distance between observations or small areas.

The main aim of this paper is to propose an approach to the problem of small area estimation in circumstances in which the sample data are of a spatial nature (or in other contexts in which it is possible to establish some kind of proximity between the domains of study), using an estimator that explicitly consider spatial correlation as a function of distance between small areas of study. This estimator, applicable to unit level data, exploits both auxiliary information relating to other known variables on the population and structures of spatial correlation between the sample data through the specification of an adequate non-diagonal structure for the variance-covariance matrices of random effects. It is based on a general class of models that includes some of the existing models as special cases and can be understood as an EBLUP of the small area totals. Consequently, it does not require the specification of a specific prior distribution for model random effects. We also aim to evaluate this estimator in comparison with traditional synthetic and composite estimators that do not explicitly consider spatial variability. The paper is organized under five sections. Section 1 introduces the context of the small area estimation and the goals of the paper. Section 2 reviews some traditional indirect estimators. Section 3 proposes an EBLUP estimator for small area totals based on spatial unit level data. The estimator is assisted by a class of models that fits into the general linear mixed theory. Section 4 describes the design of the Monte Carlo simulation study and presents empirical results. This study analyzes the performance of the proposed estimator over the direct and indirect estimators using a real data set from an agricultural survey conducted by the Portuguese Statistical Office. Discussion of the main findings of this study, along with some of its limitations and possible future developments are the subject of Section 5.

2. INDIRECT ESTIMATORS

One possible approach for "borrow information" in the context of small area estimation based on implicit models is to use direct modified estimators. These estimators maintain certain design-based properties such as approximately unbiased. This is the case of the regression estimator ([37])

$$\hat{\tau}_{d,\text{reg}} = \hat{\tau}_d + (\boldsymbol{\tau}_{\mathbf{x}d} - \hat{\boldsymbol{\tau}}_{\mathbf{x}d})' \hat{\boldsymbol{\beta}} , \qquad d = 1, ..., D , \tag{2.1}$$

where $\hat{\tau}_d$ is an estimator of the d^{th} domain total of the interest variable, usually the Horvitz–Thompson or a post-stratified estimator and $\hat{\boldsymbol{\tau}}_{\mathbf{x}d}$ have the same

meaning in relation to the vector of auxiliary variables $\mathbf{x}_i = (x_{i1}, ..., x_{ip})'$, $\hat{\boldsymbol{\beta}} = \left[\sum_{d\in U}\sum_{i\in s_d} \upsilon_i^{-2}\pi_i^{-1}\mathbf{x}_i\mathbf{x}_i'\right]^{-1}\sum_{d\in U}\sum_{i\in s_d}\upsilon_i^{-2}\pi_i^{-1}\mathbf{x}_i\, y_i$ are estimators of the regression coefficients obtained using data from the whole sample, υ_i^2 are regression weights and π_i the inclusion probabilities resulting from the sampling design. Estimator (2.1) is approximately unbiased, since $E(\hat{\tau}_{d,\mathrm{reg}}) = \tau_d + \boldsymbol{\tau}_{\mathbf{x}d}' E(\hat{\boldsymbol{\beta}}) - E(\hat{\boldsymbol{\tau}}_{\mathbf{x}d}'\boldsymbol{\beta}) \approx \tau_d$ (supposing $E\left[(\hat{\boldsymbol{\tau}}_{\mathbf{x}d} - \boldsymbol{\tau}_{\mathbf{x}d})'(\hat{\boldsymbol{\beta}} - \mathbf{b})\right] \ll \boldsymbol{\tau}_{\mathbf{x}d}\,\mathbf{b}$, where $\mathbf{b} = E(\hat{\boldsymbol{\beta}})$ from the design-based perspective[1]). Although using information from outside the domain for estimating the regression coefficients, usually these estimators still show low precision.

An alternative is the synthetic estimation (whose properties depend on the assumptions of a postulated model). From the design-based point of view these estimators can be biased and inconsistent. A synthetic regression estimator can be presented as:

$$\hat{\tau}_{d,\mathrm{sreg}} = \boldsymbol{\tau}_{\mathbf{x}d}'\,\hat{\boldsymbol{\beta}}\ , \qquad d = 1, ..., D\ , \tag{2.2}$$

where $\hat{\boldsymbol{\beta}}$ is obtained as before. A more extreme attitude under a pure model-based approach would ignore the inclusion probabilities in estimating the regression parameters. The design-based bias of estimator (2.2) is $B(\hat{\tau}_{d,\mathrm{sreg}}) \approx \boldsymbol{\tau}_{\mathbf{x}d}'(\mathbf{b} - \mathbf{b}_d)$, assuming the regression weights are such that $\upsilon_i^2 \propto \sum_{j=1}^{p} a_j\, x_{ij},\ i \in U_d$, where $a_j,\ j = 1, ..., p$, are arbitrary constants. This condition is always assured in the most typical situation of a non-weighted regression where the parameters υ_i^2 are assumed constant. Further, typically estimator (2.2) has smaller variance than the direct modified regression estimator but it is biased from the design-based point of view. When dealing with small areas the reduction in variance associated with the synthetic estimator can be such that it will assure a mean square error (MSE) lower than the one obtained through the use of the direct modified estimator. There will always be a risk of a high bias and consequently the invalidity of any confidence intervals obtained under repeated sampling. A significant advantage of synthetic estimation lies in the fact that it is always possible to obtain domain estimates, even in situations where the sample is very small or even zero. In order to prevent the quality of the estimator being totally dependent on the postulated model, some combined or composite estimators have been proposed. A combined estimator typically presents the form of the weighted average of a design-based estimator (approximately unbiased but with high variance) and a synthetic estimator (biased but with low variance):

$$\hat{\tau}_{d,\mathrm{com}} \ = \ \lambda_d\,\hat{\tau}_{d,\mathrm{des}} + (1 - \lambda_d)\,\hat{\tau}_{d,\mathrm{syn}}\ , \qquad d = 1, ..., D\ , \tag{2.3}$$

with $0 \le \lambda_d \le 1$. These estimators can be classified in two main types (according to the way the weights λ_d are chosen): sample-size dependent weights and data dependent weights. It is also possible to assume that the weights are chosen in

[1] This condition supposes there is a sufficiently week correlation between $\hat{\boldsymbol{\tau}}_{\mathbf{x}d}$ and $\hat{\boldsymbol{\beta}}$ what is usually easily achieved as $\hat{\boldsymbol{\tau}}_{\mathbf{x}d}$ and $\hat{\boldsymbol{\beta}}$ are estimated at different aggregation levels.

a deterministic way using, for example, some previous knowledge or an informed guess. This would result in what [39] call weights fixed in advance. A good example of a combined regression estimator where the weights depend on sample size is the dampened regression estimator ([38]):

$$\hat{\tau}_{d,\text{dreg}} = \lambda_d \hat{\tau}_{d,\text{reg}} + (1-\lambda_d)\hat{\tau}_{d,\text{sreg}} , \qquad d = 1, ..., D , \tag{2.4}$$

where $\lambda_d = 1$ if $\hat{N}_d \geq N_d$ and $\lambda_d = 0$ otherwise, where N_d is the d^{th} domain population size and h is a positive constant. The authors suggested to use $h = 2$. The basic idea for choosing h is to assure that the bias contribution from the synthetic component of the estimator is kept within acceptable limits. Another possible approach for "borrow information" in the context of small area estimation is to use a data-dependent combined estimator, through the modeling of the bias of the synthetic part of the estimator, thus producing indirect estimates for the weights. Many of the models that have been proposed include random area effects and can be seen as particular cases of linear mixed models. One of the best known models applicable at unit level is the nested error regression model ([8], [3]). All these approaches implicitly consider some kind of sectional correlation and the domain estimators are obtained through EBLUP, empirical Bayes or hierarchical Bayes approaches. The well-known nested error regression model ([3]) has the form $y_{di} = \mathbf{x}'_{di}\boldsymbol{\beta} + u_d + \epsilon_{di},\ d = 1, ..., D,\ i = 1, ..., N_d,$ where u_d and ϵ_d are assumed to be iid with zero means. It is also assumed that u_d and ϵ_{di} are mutually independent, $V_m(u_d) = \sigma_u^2$ and $V_m(\epsilon_{di}) = \sigma^2 \upsilon_{di}^2$ where υ_{di} are known constants. Here a common covariance between any two observations in the same small area is assumed, $\text{Cov}_m(y_{di}, y_{dj}) = \sigma_u^2\ (i \neq j)$. In this kind of models it is assumed that there is no sample selection bias, resulting that they are assumed to hold both for the population and for the sample. This may be a very limiting assumption since small domain estimation is frequently needed in the context of informative sampling designs.

An alternative to the EBLUP in the context of informative sampling designs is the pseudo-EBLUP estimator ([29]). This estimator, based on the nested error regression model, depends on survey weights and it is design-consistent.

3. A COMBINED ESTIMATOR FOR SPATIAL DATA

3.1. A class of models

Let y_{di} be the value of the interest variable for unit i $(i = 1, ..., n_d)$ in small area d $(d = 1, ..., D)$ and let $\mathbf{x}'_{di} = (x_{di1}, ..., x_{dip})$ a vector of p unit level explana-

tory variables referring to the same unit. Consider the following class of models:

$$(3.1) \quad y_{di} = \mathbf{x}'_{di}\boldsymbol{\beta} + \sum_{h=1}^{H} \mathbf{x}'_{(1)di}\, q_{h,di}\, \mathbf{u}_h^{(1)} + \mathbf{x}'_{(2)di}\, \mathbf{u}_d^{(2)} + \epsilon_{di}\,, \quad d = 1, ..., D, \;\; i = 1, ..., n_d\,,$$

where $\boldsymbol{\beta}$ is a vector of p fixed effects; $\mathbf{x}'_{(j)di}$ is a vector of p_j explanatory variables (typically a subvector of $\mathbf{x}'_{di}$) for the i^{th} unit in small area d; $q_{h,di}$ are design variables used to take into account the sampling design and indicate that unit di belongs to a stratum or a sampling unit h ($h = 1, ..., H$); $\mathbf{u}_h^{(1)} = \text{col}_{1 \le j \le p_1}(u_{hj})$ is a vector of p_1 random (or fixed) design effects associated with stratum (or sampling unit) h; $\mathbf{u}_d^{(2)} = \text{col}_{1 \le j \le p_2}(u_{dj})$ is a vector of p_2 random effects associated with domain d; l_d represents the geographical location associated to the centroid of domain d; $f(l_d - l_e)$ is a function of the vector $l_d - l_e$ and ϵ_{di} is the residual term associated with unit di. We assume that $E_m\big(\mathbf{u}_h^{(1)}\big) = \mathbf{0}$, $E_m\big(\mathbf{u}_d^{(2)}\big) = \mathbf{0}$, $E_m(\epsilon_{di}) = 0$, $E_m\big(\mathbf{u}_h^{(1)}\mathbf{u}_g^{\prime(1)}\big) = \begin{cases} \boldsymbol{\Sigma}^{(1)}, & h = g \\ 0, & \text{otherwise} \end{cases}$, $E_m(\epsilon_{di}\,\epsilon_{ej}) = \begin{cases} \sigma^2_{di}, & d = e,\; i = j \\ 0, & \text{otherwise} \end{cases}$, $E_m\big(\mathbf{u}_d^{(2)}\mathbf{u}_e^{\prime(2)}\big) = \boldsymbol{\Sigma}^{(2)} f(l_d - l_e)$, with $\boldsymbol{\Sigma}^{(1)} = \left\{\sigma^2_{U_{jk}^{(1)}}\right\}$ $(j, k = 1, ..., p_1)$, $\sigma^2_{U_{jk}^{(1)}} = E(u_{hj}\, u_{hk})$, $\boldsymbol{\Sigma}^{(2)} = \left\{\sigma^2_{U_{jk}^{(2)}}\right\}$ $(j, k = 1, ..., p_2)$, and $\sigma^2_{U_{jk}^{(2)}} = E(u_{dj}\, u_{dk})$. The model is applied to data from a sample of total size $n = \sum_{d=1}^{D} n_d$, where n_d is the number of sampling units in area d. It is also assumed that random effects associated with different aggregation levels are not correlated, $E\big(\mathbf{u}^{(j)}\,\mathbf{u}^{\prime(k)}\big) = \mathbf{0}$ for $j \neq k$, and that the errors are non-correlated with random effects, $E\big(\mathbf{u}^{(j)}\boldsymbol{\epsilon}'\big) = \mathbf{0}$, $\forall j$.

In the proposed model, domain effects show a structure of spatial variability. The covariance between the random effects associated with domains d and e depends on the vector defined by their geographical coordinates $l_d - l_e$. Some functions that can be applied to this context are presented in [28]. When the spatial covariance only depends on the distance $|l_d - l_e|$, then the function $f(|l_d - l_e|)$ is said to be isotropic ([6]) and is typically such that $\lim_{|l_d - l_e| \to 0} f = 1$. As the domains are not points in space, but areas, these coordinates are defined by their centroids. The assumption is that the lowest level of aggregation for which the georeferencing is available is the domain. Situations where the level of aggregation for which the referencing is available does not coincide with the domains of study can generate special cases of model (3.1). In particular, when georeferencing is possible at unit level, spatial variation can be modeled through variances-covariances of the errors vector $\boldsymbol{\epsilon}$.

Domain random effects represent the characteristics specific to the domain of study that affect the values of the interest variable and are not represented by the fixed effects at a higher level of aggregation. They can be thought of as modeling the bias of the synthetic part of the model. Moreover, these domain random effects will now have the additional role of bringing information from other domains, to explain the values of the interest variable in each domain.

Design effects are used to take into account the sampling design. The goal is to allow the model to be applied to contexts with informative sampling designs, overcoming the limitations ([27], [20]) of other data-dependent combined estimators that implicitly assume that the sampling design is ignorable.

The methodology proposed can therefore be seen as model assisted. The sample s is the result of a two-step procedure. First it is supposed that the finite population can be approximately described by a superpopulation model. In the second step it is assumed that a sample is drawn from the finite population through a specific sampling design. It is assumed that the sample can be approximately described by model (3.1), which has taken into account the existence of these two steps.

3.2. Estimation of model parameters

The model 3.1 can be presented as a special case of the general linear mixed model, grouping the unit-specific models over the population:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon} \ , \tag{3.2}$$

where $\mathbf{y}$ is a vector of the target variable, $\mathbf{X}$ is a design matrix of explanatory variables with rows given by $\mathbf{x}'_{di}$, $\mathbf{Z} = \left[\mathbf{Z}_{(1)}\, \mathbf{Z}_{(2)}\right]$ is a design matrix, $\mathbf{u} = \mathrm{col}_{1\leq j\leq 2}\left(\mathbf{u}^{(j)}\right)$ is a vector of random effects and $\boldsymbol{\epsilon}$ is a vector of errors. The covariance matrix of $\mathbf{u}$ is given by $\mathbf{G} = V_m(\mathbf{u}) = \mathrm{blockdiag}_{1\leq j\leq 2}\left[\mathbf{G}^{(j)}\right]$, where $\mathbf{G}^{(1)} = \mathrm{blockdiag}_{1\leq h\leq H}\left\{\boldsymbol{\Sigma}^{(1)}\right\}$ and $\mathbf{G}^{(2)} = \mathbf{F}\otimes\boldsymbol{\Sigma}^{(2)}$ with $\mathbf{F} = \left\{f(l_d - l_e)\right\},\ d,e = 1,...,D$. Further, $\mathbf{R} = V_m(\boldsymbol{\epsilon}) = \mathrm{diag}_{\substack{1\leq i\leq n_d \\ 1\leq d\leq D}}\{\sigma^2_{di}\}$, $E_m\left(\mathbf{u}^{(1)}\right) = \mathbf{0}$, $E_m\left(\mathbf{u}^{(2)}\right) = \mathbf{0}$ and $E_m(\boldsymbol{\epsilon}) = \mathbf{0}$. Both covariance matrices $\mathbf{G}$ and $\mathbf{R}$ involve unknown variance components, represented by $\boldsymbol{\theta}$. Also the flexibility of the proposed class of models recommends proceeding in each application to the selection of a specific model, i.e. to the choice of the explanatory variables and appropriate variance-covariance structures to $\mathbf{u}$ and $\boldsymbol{\epsilon}$. This step in model selection and diagnosis is crucially important to obtaining a model that can adequately describe the behavior of the target population and can be performed as a systematic procedure like those proposed by [7] or [43].

Once a specific model has been selected, the variance components $\boldsymbol{\theta}$ need to be estimated in order to assess the variability of estimators or to predict the fixed and random effects. Several methods are available for estimating variance components, such as the analysis of variance (ANOVA) method ([12]), the minimum norm quadratic unbiased estimation (MINQUE) method ([33], [34], [35]) and the likelihood methods. Some references about the maximum likelihood estimation (MLE) method due to Fisher may be found in [9], [1], [23], [21], [15] and [18]. On the other hand, references about the residual maximum likelihood estimation

(RMLE) method proposed by [41] and its extensions can be found in [24], [10], [11], [2], [42], [14], [16], [17], among others. For details about estimation general linear mixed models see [19].

Now consider the decomposition of all matrices into sample and non-sample components, where the subscript s is associated with the n sample units and r is associated with the $(N-n)$ non-sample units. The omission of the subscript indicates that the respective matrices allude to the whole population $U \equiv s \cup r$. Assuming model 3.2 holds and variance components are known, the best linear unbiased estimator of $\boldsymbol{\beta}$ and the best linear unbiased predictor of $\boldsymbol{\theta}$ are given by

$$\widetilde{\boldsymbol{\beta}} = \left(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s\right)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}\,\mathbf{y}_s\ , \tag{3.3}$$

$$\widetilde{\mathbf{u}} = \mathbf{G}\,\mathbf{Z}_s'\mathbf{V}_s^{-1}\left(\mathbf{y}_s - \mathbf{X}_s\,\widetilde{\boldsymbol{\beta}}\right)\ , \tag{3.4}$$

where $\mathbf{V}_s = E\left[(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta})\,(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta})'\right] = \mathbf{Z}_s\,\mathbf{G}\,\mathbf{Z}_s' + R_{ss}$. The vectors $\mathbf{u}^{(j)}$ may be predicted using $\widetilde{\mathbf{u}}^{(j)} = \mathbf{G}^{(j)}\,\mathbf{Z}_{(j)s}'\mathbf{V}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}})$, while the predictors of the errors $\boldsymbol{\epsilon}$, can be obtained as $\boldsymbol{\epsilon} = \mathbf{R}_{.s}\mathbf{V}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}})$, where $\mathbf{R}_{.s} = [\mathbf{R}_{ss}'\,\mathbf{R}_{rs}']'$.

When the covariance matrix $\mathbf{R} = \begin{bmatrix} \mathbf{R}_{ss} & \mathbf{R}_{sr} \\ \mathbf{R}_{rs} & \mathbf{R}_{rr} \end{bmatrix}$ is block-diagonal, i.e. when there is no correlation between errors associated with the observations inside and outside the sample, then $\mathbf{R}_{rs} = 0$ and $\widetilde{\boldsymbol{\epsilon}}_r = \mathbf{0}$. This is the case for model (3.1). Nevertheless, it should be noted that some situations can be devised, particularly when the spatial correlation can be established at unit level, where there is a correlation between model errors that can be used in the prediction of $\boldsymbol{\epsilon}_r$.

3.3. Estimation of domain totals

The objective of the inference can be seen as to predict the total of an interest variable, τ_d, that under the model corresponds to the summation of the realizations of the variable of interest over all the elements in the small area d:

$$\tau_d\ = \sum_{i\in U_d} y_{di}\ =\ \boldsymbol{\tau}_{\mathbf{x},d}'\boldsymbol{\beta} + \sum_{h=1}^{H}\boldsymbol{\tau}_{\mathbf{x}(1),hd}'\,\mathbf{u}_h^{(1)} + \tau_{\mathbf{x}(2),d}'\,\mathbf{u}_{ad}^{(2)} + \tau_{\epsilon,d}\ , \tag{3.5}$$

where $\tau_{\epsilon,d} = \sum_{i\in U_d}\epsilon_{di}$. It should be noted that, from the model-based point of view, (3.5) is a predictable function producing inference in the narrow inference space [22]. An estimator for the small area total, τ_d, can be obtained as

$$\widetilde{\tau}_d\ =\ \mathbf{1}_{N_d}'\widetilde{\mathbf{y}}_d\ =\ \sum_{i\in U_d}\widetilde{y}_{di}\ =\ \boldsymbol{\tau}_{\mathbf{x}d}'\widetilde{\boldsymbol{\beta}} + \mathbf{v}_{\boldsymbol{\tau}s}'\mathbf{V}_s^{-1}\left(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}}\right)\ , \tag{3.6}$$

where $\widetilde{\mathbf{y}}_d$ is the EBLUP of the vector $\mathbf{y}_d$ and $\mathbf{v}_{\boldsymbol{\tau}s}' = \boldsymbol{\tau}_{\mathbf{z}d}'\,\mathbf{G}\,\mathbf{Z}_s' + \mathbf{1}_{N_d}'\mathbf{R}_{d\cdot,s}$ is the line vector of the model-based covariances between the small area total τ_d and

the observable vector $\mathbf{y}_s$, $\mathbf{R}_{d\cdot,s} = E(\boldsymbol{\epsilon}_{ad}\boldsymbol{\epsilon}_s)$, and $\mathbf{1}'_{N_d}$ is a unit vector of size N_d. It should be noted that the estimator $\widetilde{\tau}_d$ is the EBLUP of τ_d, given the observable random vector $\mathbf{y}_s$ (cf. Appendix 1).

When $\mathbf{R}_{d,rs}$ is a null matrix, then the EBLUP of the total is τ_d is given by a simplified expression:

$$\widetilde{\tau}_d \;=\; \widetilde{E}(\tau_{d,r}|\mathbf{u}) \;=\; \tau_{\mathbf{y},d,s} + \boldsymbol{\tau}'_{\mathbf{x},d,r}\widetilde{\boldsymbol{\beta}} + \boldsymbol{\tau}'_{\mathbf{z},d,r}\,\mathbf{G}\,\mathbf{Z}'_s\mathbf{V}_s^{-1}\big(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}}\big)\;, \tag{3.7}$$

where $\tau_{\mathbf{y},d,s}$ is the observed sample total in small area d (cf. Appendix 2). It should also be noted that many of the regression estimators that have been proposed for small area estimation may be viewed as EBLUP of domain totals for particular cases of the class of models (3.1). For instance, the form of the nested error regression model and the random coefficient model presented in Section 2 accord with class (3.1), with $\mathbf{u}_d^{(2)}$ scalar, $\mathbf{G}^{(1)} = \mathbf{0}$, $\mathbf{G}^{(2)} = \sigma_u^2\,\mathbf{I}_D$ and $\mathbf{R} = \sigma^2\,\mathbf{I}_n$. Also, the model underlying the synthetic regression estimator (2.2) is equivalent to considering (3.1) with $\mathbf{u}_h^{(1)}$ scalar and taken as a fixed effect, $\mathbf{G}^{(2)} = \mathbf{0}$ and $\mathbf{R} = \sigma^2\,\mathbf{I}_n$. Moreover, the direct modified regression estimator (2.1), can be obtained considering $\mathbf{u}_h^{(1)}$ and $\mathbf{u}_d^{(2)}$ scalars and taken as a fixed effects and $\mathbf{R} = \sigma^2\,\mathbf{I}_n$.

3.4. Domains not represented in the sample

Situations may arise where some domains are not represented in the sample. If no sample falls into small area d, then the respective random effects $\mathbf{u}_d^{(2)}$ may still be predicted if there is covariance between $\mathbf{u}_d^{(2)}$ and at least one of the small area random effects represented in the sample $\mathbf{u}_e^{(2)}$ $(e = 1, ..., D;\; e \neq d)$. We have then

$$\widetilde{\mathbf{u}}_d^{(2)} \;=\; \mathbf{G}_{d,\cdot}^{(2)}\,\mathbf{Z}'_{(2)s}\mathbf{V}_s^{-1}\big(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}}\big)\;, \tag{3.8}$$

where $\mathbf{G}_{d,\cdot}^{\prime(2)} = \underset{\substack{n_e\neq 0\\ 1\le e\le D}}{\mathrm{col}}\big[\mathbf{G}_{d,e}^{\prime(2)}\big] = E\big[\mathbf{u}^{(2)}\mathbf{u}_d^{\prime(2)}\big]$ and $\mathbf{G}_{d,e}^{(2)} = \boldsymbol{\Sigma}^{(2)} f(l_d - l_e)$. In an extreme situation where the small area effects $\mathbf{u}_d^{(2)}$ are not correlated with any other small area effect for a domain represented in the sample, i.e. when $\mathbf{G}_{d,e}^{(2)} = 0$, $\forall\, e \neq d\colon n_e \neq 0$, then $\widetilde{\mathbf{u}}_d^{(2)} = \mathbf{0}$. The estimator $\widetilde{\tau}_d$ is then reduced to a form similar to a following synthetic estimator:

$$\widetilde{\tau}_d \;=\; \tau_{y,d,s} + \boldsymbol{\tau}'_{\mathbf{x},d,r}\widetilde{\boldsymbol{\beta}} + \sum_{h=1}^{H}\boldsymbol{\tau}'_{\mathbf{x}(1),dh,r}\,\widetilde{\mathbf{u}}_h^{(1)}\;. \tag{3.9}$$

It may be noted that this estimator may be written in the same generic form

$$\hat{\tau}_d \;=\; \boldsymbol{\tau}'_{\mathbf{x},d}\hat{\boldsymbol{\beta}} + \mathbf{f}'\big(\mathbf{y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}\big)\;, \tag{3.10}$$

where $\mathbf{f}$ is used to weight the regression residuals. This form puts in evidence that estimator (3.9) can be seen as a combined estimator where the weights in $\mathbf{f}'$ allow a correction of the synthetic part of the estimator $\boldsymbol{\tau}'_{\mathbf{x},d}\widetilde{\boldsymbol{\beta}}$ through the prediction errors in the domain that is the target of inference, but also in other domains spatially correlated. When no correlation between domains is specified, the correction factor depends only on the prediction errors in the target small area and the estimator is reduced to a similar form to the data-dependent combined estimators presented in Section 2.

These characteristics seem to be particularly interesting when estimating in small domains, where the available sample size is small, since it borrows information from outside the domain of study in order to assist the estimation. Moreover, taking advantage of the potential spatial correlation of data it is possible to avoid the reduction of the proposed estimators to pure synthetic estimators even when the sample size in the domain is null.

4. MONTE CARLO SIMULATION STUDY

4.1. Generation of the pseudo-population

For the simulation a pseudo-population is used. This population is obtained from a real data set containing the responses to the 1993 wave of the Agricultural Structure Survey. It is an agricultural survey conducted by the Portuguese Statistical Office in the period between agricultural censuses. The responses for the variable total production of cereals were extracted and circumscribed to the NUTSII of *Alentejo*. The total sample size in this region is 7,060 and the population size 47,049. The design for the Agricultural Structure Survey is based on stratified sampling. The sample is first stratified using the *região agrária* as the level for geographic stratification. A *região agrária* is an administrative division used for agricultural purposes. In each *região agrária* a new stratification is established based on Used Agricultural Surface (UAS) classes. In the same *região agrária* some other strata are defined based on the value of other variables considered weakly correlated with UAS. In *Alentejo* there are 19 strata.

For simulation purposes a pseudo-population is generated by replicating the agricultural establishments in the sample proportionally to the inverse of their inclusion probabilities. The sampling frame resulting from this replication includes the value of production of cereals for each establishment in 1993, and the same value reported for 1989 (year of the agricultural census). The production in 1989 is used as an auxiliary variable in the models used in the simulation. Also, geographical coordinates associated with the centroids of *freguesias* were recorded.

This was the lowest level of aggregation for which geographical referencing was available. This means that geographical differentiation between establishments included in the same *freguesia* is not available. A *freguesia* is an administrative division that segments the *Alentejo* into 284 sub-regions.

4.2. Description of the simulations

Using the sampling frame corresponding to the pseudo-population of agricultural establishments we have run a Monte Carlo simulation. The goal is to evaluate the design-based properties of a set of alternative estimators. Note that the approach followed in this paper is to evaluate the properties and relative merits of the proposed estimators through simulation. In fact, due to the complexity of these estimators their design-based properties (e.g. bias, variance) are impossible to obtain through analytical methods. Also, their model-based properties would be of limited interest from the point of view of a benchmark with alternative direct estimators used in this simulation whose properties only make sense to evaluate from the design-based perspective. The target parameter is the total of the variable production of cereals at *freguesia* level. The number of simulations performed is 560. In each simulation a sample is drawn from the pseudo-population U^*, using a stratified design similar to the one used in the Agricultural Structures Survey. The only difference in relation to that survey design is that the sample size by stratum was reduced to 30% of the original size (2,118 establishments). The goal is to simulate a framework similar to that survey, but with a smaller sample size, enabling the evaluation of the estimators' behavior in "critical" situations where the domain sample size is very small (sometimes only a few units or even none). This sampling design leads to a relative precision of 7.5% (for a 95% confidence-level) in the estimation of the total of the interest variable at the population level, using the Horvitz–Thompson estimator. The expected sample sizes for the 284 domains of interest vary from 0.3 to 45.8 units.

4.3. Estimators

The estimators analyzed in the simulation are presented in this section. They are mainly implementations of the direct, synthetic and combined regression estimators presented in Sections 2 and 3. It should be noted that all the regression estimators include the same auxiliary variables (associated with the fixed effects), allowing a fair comparison of their relative merits. In what follows, the notation ad is used to represent the small area d of region a, where the regions correspond to the level of aggregation of NUTSIII and the small area of interest to *freguesia*. Table 1 summarizes the estimators used in the simulation.

Table 1: Estimators used in the simulation study.

Estimator	Description
$\hat{\tau}_{ad1}$	Horvitz–Thompson estimator
$\hat{\tau}_{ad2}$	Direct modified regression estimator (2.1)
$\hat{\tau}_{ad3}$	Dampened regression estimator (2.4)
$\hat{\tau}_{ad4}$	Pure synthetic regression estimator (2.2) with fixed effects estimated ignoring the sampling design
$\hat{\tau}_{ad5}$	Synthetic regression estimator where the sampling design is explicitly considered through the inclusion of a vector of design variables E_{hi} indicating the belonging of each establishment i to the strata $h=1,...,H$
$\widetilde{\tau}_{ad6}$	Data-dependent combined regression estimator based on the nested error unit level regression model
$\widetilde{\tau}_{ad7}$	Data-dependent combined regression estimator, similar to $\widetilde{\tau}_{ad6}$ but including fixed strata effects β_{h0}, $h=1,...,H$
$\widetilde{\tau}_{ad8}$	Data-dependent combined regression estimator, based on a model included in the proposed class of models (3.1), with random small area effects presenting a spatial covariance structure following an isotropic exponential model.

The isotropic exponential model used to represent spatial variability in $\hat{\tau}_{ad8}$ was suggested in the model diagnosis phase. We have tested several structures (exponential, spherical, linear, log-linear and Gaussian), through the evaluation the significance of covariance parameters (using Wald tests) and information criteria (such as AIC and BIC). Among the structures that showed statistical significance we retained the one that minimized the several information criteria. Although we have chosen the exponential model, some of the other structures resulted in very similar adjustments. Also note that for the data-dependent regression estimators the variance components are estimated through REML method. The only exception regards estimator $\widetilde{\tau}_{ad8}$, where the parameter c_e was estimated *a priori* through the adjustment of an exponential semivariogram to an empirical semivariogram.

Note that the estimators included in the simulation vary in nature: $\hat{\tau}_{ad1}$ and $\hat{\tau}_{ad2}$ are design-based estimators, $\hat{\tau}_{ad4}$ and $\hat{\tau}_{ad5}$ are synthetic estimators, while the others can be classified as combined estimators as described in previous sections. The included estimators also differentiate in the way the sampling design is (or not) taking into account: in $\hat{\tau}_{ad1}$, $\hat{\tau}_{ad2}$ and $\hat{\tau}_{ad3}$ the sampling design information is taking into account using sampling weights, $\hat{\tau}_{ad5}$, $\hat{\tau}_{ad7}$ and $\hat{\tau}_{ad8}$ include fixed strata effects, while in $\hat{\tau}_{ad4}$ and $\hat{\tau}_{ad6}$ the sampling design information is ignored.

4.4. Precision and bias measures

The estimators under consideration are evaluated using a set of precision and bias measures. In what follows K represents the number of simulations, and $\hat{\tau}_{kd}$ the d^{th} small area estimate of the total obtained from the simulation k $(k = 1, ..., K)$.

4.4.1. Unconditional analysis

Taking into account the high number of small areas in the population (284) and in order to facilitate the presentation of the simulation results, the small areas are divided into six groups. Each group g contains D_g small areas. Table 2 presents the definition of each group and the number of small areas involved.

Table 2: Small area groups in the simulation study.

Group	Expected sample size	Number of small areas
0	—	20
1	$[0; 2]$	20
2	$[2; 3.5]$	43
3	$[3.5; 5]$	49
4	$[5; 10]$	87
5	$[10; +\infty]$	65

Groups 1 to 5 were defined according to the expected sample size of the small areas. Group 0 includes small areas for which the total of the interest variable is zero (*freguesias* where there is no cereal production) regardless their size. The goal is to separate these small areas from the other groups to prevent them from changing the conclusions regarding the relative merits of the estimators. The Monte Carlo relative error for the estimators' expected value is on average 8.0% in group 1 and varies between 3.4% and 4.1% in groups 2 to 5.

For the unconditional analysis the following measures were considered for each group g:

Average absolute bias: $AAB_g = D_g^{-1} \sum_{d=1}^{D_g} AB_d$, where $AB_d = K^{-1} \sum_{j=1}^{K} |\hat{\tau}_{jd} - \tau_d|$;

Average MSE: $AMSE_g = D_g^{-1} \sum_{d=1}^{D_g} MSE_d$, where $MSE_d = K^{-1} \sum_{j=1}^{K} (\hat{\tau}_{jd} - \tau_d)^2$;

Average variance: $AV_g = D_g^{-1} \sum_{d=1}^{D_g} V_d$, where $V_d = K^{-1} \sum_{j=1}^{K} (\hat{\tau}_{jd} - \bar{\hat{\tau}}_d)^2$;

Average absolute bias ratio: $AABR_g = D_g^{-1} \sum_{d=1}^{D_g} ABR_d$, where $ABR_d = AB_d / \sqrt{V_d}$;

Average coverage rate for a design-based $100(1-\alpha)$ confidence interval: $ACR_g = D_g^{-1} \sum_{d=1}^{D_g} TC_d$,

where $TC_d = 100 \times R_d / K$ and R_d represents the number of simulations for which the confidence interval $\hat{\tau}_{jd} \pm t_{\alpha/2} \sqrt{V_d}$ contains the true parameter τ_d .

4.4.2. Conditional analysis

A conditional analysis was also conducted using a set of precision and bias measures for each small area d. The superscript (n_d) indicates that the respective measure is conditioned to the realized sample size, n_d, in small area d. They are:

Conditional relative bias: $CRB_d^{(n_d)} = K_{n_d}^{-1} \sum_{j=1}^{K_{n_d}} (\hat{\tau}_{jd} - \tau_d) / \tau_d$;

Conditional relative standard error: $CRSE_d^{(n_d)} = \sqrt{K_{n_d}^{-1} \sum_{j=1}^{K_{n_d}} (\hat{\tau}_{jd} - \tau_d)^2} / \tau_d$;

Conditional variation coefficient: $CVC_d^{(n_d)} = \sqrt{V_d^{(n_d)}} / \tau_d$,

where $V_d^{(n_d)} = K_{n_d}^{-1} \sum_{j=1}^{K_{n_d}} (\hat{\tau}_{jd} - \bar{\hat{\tau}}_d)^2$ is the conditional variance ;

Conditional bias ratio: $CBR_d^{(n_d)} = B_d^{(n_d)} / \sqrt{V_d^{n_d}}$,

where $B_d^{(n_d)} = K_{n_d}^{-1} \sum_{j=1}^{K_{n_d}} (\hat{\tau}_{jd} - \tau_d)$ is the conditional bias ;

Coverage rate of the conditional design-based confidence interval: $CR_d^{(n_d)} = 100 \times R_d^{(n_d)} / K_{n_d}$,

where $R_d^{(n_d)}$ represents the number of simulations for which the confidence interval $\hat{\tau}_{jd} \pm t_{\alpha/2} \sqrt{V_d^{(n_d)}}$ contains the true parameter τ_d .

4.5. Results

4.5.1. Unconditional analysis

Table 3 summarizes the unconditional results of the simulation study. The values for absolute bias, variance and MSE are presented relatively to the respective value associated with $\hat{\tau}_{ad2}$.

Table 3: Unconditional results.

Group	$\hat{\tau}_{ad1}$	$\hat{\tau}_{ad2}$	$\hat{\tau}_{ad3}$	$\hat{\tau}_{ad4}$	$\hat{\tau}_{ad5}$	$\widetilde{\tau}_{ad6}$	$\widetilde{\tau}_{ad7}$	$\widetilde{\tau}_{ad8}$
Absolute bias								
0	0.00	1.00	2.39	96.54	19.67	91.80	15.50	11.07
1	0.60	1.00	3.44	23.64	10.32	22.57	9.63	9.46
2	1.13	1.00	4.15	22.72	15.14	24.30	12.07	12.59
3	1.33	1.00	4.65	45.75	18.54	44.16	13.72	13.55
4	1.46	1.00	4.40	51.13	20.76	49.14	13.60	14.47
5	1.09	1.00	3.36	77.28	22.51	68.04	14.14	18.14
Variance								
0	0.00	1.00	1.07	0.67	0.03	0.60	0.21	0.29
1	1.10	1.00	0.79	0.03	0.01	0.09	0.06	0.09
2	2.06	1.00	0.76	0.06	0.02	0.26	0.18	0.25
3	1.77	1.00	0.80	0.10	0.03	0.45	0.34	0.40
4	2.01	1.00	0.79	0.19	0.05	1.18	0.94	1.10
5	1.84	1.00	0.85	0.45	0.10	2.29	1.91	2.13
MSE								
0	0.00	1.00	1.08	14.74	0.68	13.18	0.63	0.61
1	1.09	1.00	0.85	0.93	0.40	0.91	0.37	0.39
2	2.06	1.00	0.79	0.92	0.45	1.21	0.47	0.57
3	1.77	1.00	0.82	1.67	0.45	1.89	0.58	0.63
4	2.01	1.00	0.82	3.36	0.63	3.99	1.20	1.46
5	1.84	1.00	0.87	7.88	0.76	7.39	2.27	2.76
Bias ratio								
0	0.01	0.04	0.09	5.39	4.55	5.15	2.29	1.68
1	0.03	0.04	0.19	5.36	6.40	3.86	2.58	1.63
2	0.03	0.03	0.18	3.64	3.70	2.18	1.44	1.26
3	0.03	0.03	0.14	3.45	3.05	2.13	0.85	0.71
4	0.04	0.03	0.15	3.74	3.58	2.11	0.75	0.62
5	0.03	0.03	0.12	3.70	2.60	1.76	0.40	0.42
Coverage Rate								
0	1.00	0.96	0.96	0.01	0.16	0.00	0.59	0.71
1	0.97	0.97	0.95	0.07	0.32	0.21	0.51	0.69
2	0.96	0.96	0.94	0.28	0.37	0.48	0.65	0.70
3	0.96	0.95	0.94	0.23	0.46	0.50	0.82	0.84
4	0.96	0.95	0.95	0.23	0.35	0.48	0.84	0.86
5	0.96	0.95	0.95	0.18	0.44	0.60	0.91	0.91

As expected the two design-based estimators $\hat{\tau}_{ad1}$ and $\hat{\tau}_{ad2}$ are approximately unbiased, even for very small areas. For these estimators the average coverage rate is very near the nominal confidence level (95%). Nevertheless, especially in the smallest domains, they show variance and MSE substantially higher than those observed for the synthetic and combined estimators. It is also to be noted that $\hat{\tau}_{ad2}$ brings significant precision gains when compared with the Horvitz–Thompson estimator.

On the other hand, the two synthetic estimators show very different behavior. $\hat{\tau}_{ad4}$, which can be viewed as a pure synthetic estimator shows disastrous behavior both in terms of bias and precision. These results are clear evidence of the effects of ignoring an informative sampling design. On the other hand, the synthetic estimator with fixed strata effects, $\hat{\tau}_{ad5}$, shows significant precision gains when compared to the direct regression estimator. These precision gains show a tendency to decrease as the expected sample size in the small areas increases. Its major drawback is related to the high bias, which originates average bias ratios that are always above 2.6, compromising the construction of design-based confidence intervals. In fact the average coverage rate for this estimator is always below 0.46, and in many cases near 0.30.

The combined regression estimator with sample-size dependent weights, $\hat{\tau}_{ad3}$, shows a systematic precision gain when compared to the direct regression estimator (with the exception of group 0, the ratio between the average MSE of the two estimators is between 0.79 and 0.87). Nevertheless, these gains are always moderate and substantially lower than the ones observed for the synthetic regression estimator. This estimator also shows a very good behavior in what regards bias. In fact, although having an absolute bias higher than those observed for the direct estimators, the average bias ratio is always lower than 0.2, which originates an average coverage rate very near the nominal confidence level. With regard to the combined estimators with data-dependent weights it can once again be observed that disregard of the sampling design, as in estimator $\widetilde{\tau}_{ad6}$, produces undesirable properties both in terms of bias and precision. In fact, $\widetilde{\tau}_{ad6}$ systematically shows higher MSE than the direct regression estimator (with the exception of group 1). It also exhibits dramatic biases (of the same magnitude as the synthetic estimator $\hat{\tau}_{ad4}$) and bias ratios that on average are situated between 1.76 and 5.15.

The combined estimators $\widetilde{\tau}_{ad7}$ and $\widetilde{\tau}_{ad8}$, which explicitly consider strata effects, show very different behavior. These estimators show average MSE that for the smallest small areas (groups 0 to 3) is near those observed for the best synthetic estimator, while allowing significant reduction in bias and bias ratio.

In particular, $\widetilde{\tau}_{ad7}$ based on a nested error model with fixed strata effects reveals important precision gains when compared to the direct regression estimator, $\hat{\tau}_{ad2}$, or the sample-size dependent regression estimator, $\hat{\tau}_{ad3}$, in groups 0 to 3.

It is to be noted that for groups 4 and 5 (corresponding to expected sample sizes higher than 5 units) $\widetilde{\tau}_{ad7}$ shows some precision loss regarding the direct regression estimator, which is particularly important in group 5. It is also significant that, in the groups 0 to 3, $\widetilde{\tau}_{ad7}$ exhibits a MSE that is similar or even smaller than that associated with the synthetic estimator $\hat{\tau}_{ad5}$. In these groups the increase in variance in $\widetilde{\tau}_{ad7}$ is more than compensated by the reduction in bias. In what regards bias measures, the estimator $\widetilde{\tau}_{ad7}$ shows a behavior that is situated between those recorded for the direct and the synthetic estimators. The average absolute bias ratios vary from 0.40 to 2.58 (increasing with the reduction of the expected sample size) and are strikingly lower than the ones associated with the synthetic estimators (varying between 15% and 50% of those obtained for the best synthetic estimator). This results in average coverage rates for a design-based confidence interval that are substantially higher than those observed for the synthetic estimators.

The estimator $\widetilde{\tau}_{ad8}$ considers a spatial covariance based on an isotropic exponential structure at the small area level. For all small area groups (with the exception of group 0) it shows a small loss of precision when compared to $\widetilde{\tau}_{ad7}$, but still allows important precision gains regarding the direct estimators and the sample-size dependent regression estimator, $\hat{\tau}_{ad3}$, in groups 0 to 3. It is worth noting that this decline in precision is mainly induced by an increase in variance, since $\widetilde{\tau}_{ad8}$ shows average absolute bias that is very near or even smaller than for $\widetilde{\tau}_{ad7}$ (mainly in the smaller areas). The bias ratios for $\widetilde{\tau}_{ad8}$ are, for groups 0 to 4, substantially smaller than those observed for $\widetilde{\tau}_{ad7}$, varying now from 0.42 to 1.68. The reduction in the bias ratio tends to diminish with the increase in the expected sample size, resulting that in group 5 the bias ratio of the two estimators is similar. In fact it is in smaller areas that $\widetilde{\tau}_{ad7}$ approximates more closely a synthetic estimator, allowing the additional sample information used in $\widetilde{\tau}_{ad8}$ (from other spatially correlated small areas) to contribute to bias reduction. Between the combined estimators that allow precision gains in small areas groups 0 to 3, $\widetilde{\tau}_{ad8}$ is the one that shows the best behavior in terms of average bias ratio, which varies from 16% to 37% of those obtained for the best synthetic estimator.

4.5.2. Conditional analysis

Figure 1 and Table 4 summarize the simulation's conditional results for one small area in the study. Considering the large number of small areas in the study (284), these data are only intended to illustrate typical results associated with one of the smallest areas. The results refer to a small area with expected sample size of 4.2 units, thus belonging to group 3.

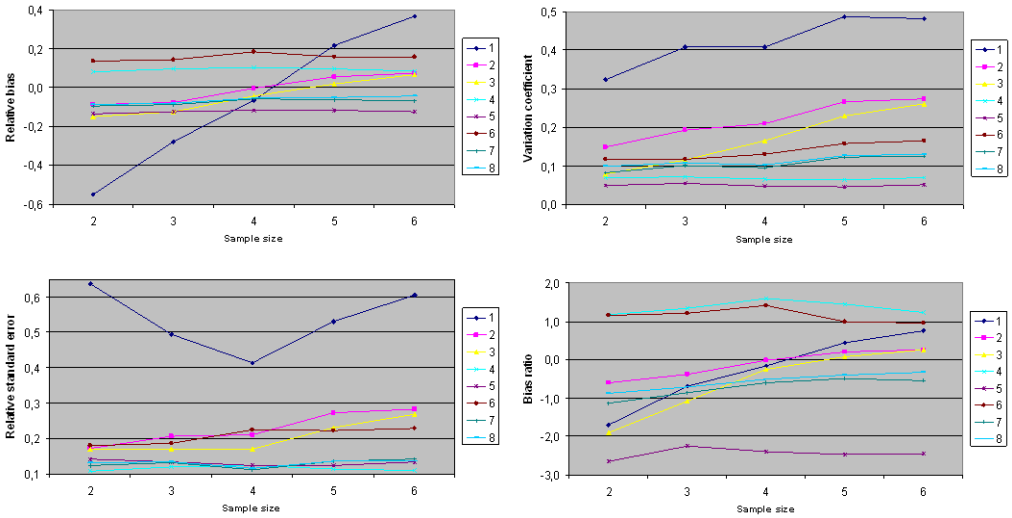

Figure 1: Conditional results.

Table 4: Unconditional coverage rates.

Sample size	$\hat{\tau}_{ad1}$	$\hat{\tau}_{ad2}$	$\hat{\tau}_{ad3}$	$\hat{\tau}_{ad4}$	$\hat{\tau}_{ad5}$	$\widetilde{\tau}_{ad6}$	$\widetilde{\tau}_{ad7}$	$\widetilde{\tau}_{ad8}$
2	0.52	0.90	0.51	0.80	0.31	0.83	0.81	0.84
3	0.88	0.89	0.81	0.74	0.32	0.77	0.90	0.88
4	0.98	0.95	0.95	0.61	0.29	0.72	0.91	0.90
5	0.91	0.96	0.96	0.66	0.35	0.89	0.94	0.94
6	0.88	0.97	0.96	0.78	0.27	0.89	0.92	0.92

The two design-based estimators $\hat{\tau}_{ad1}$ and $\hat{\tau}_{ad2}$ show bad conditional properties, namely in what regards bias. In fact, both estimators show important conditional biases and bias ratios when the effective sample size departs from the expected sample size. This phenomenon is particularly notable in the Horvitz–Thompson estimator. This bias tends to be negative for effective sample sizes that are smaller than expected and positive for expected sample sizes that are larger than expected. Also, when effective sample size is smaller than the expected sample size, the conditional variation coefficients tend to show a rising pattern with the increase in the effective sample size. The combined result of this bias and variance behavior is a conditional relative standard error that for both estimators tends to increase with the effective sample size (for sample sizes above the expected). When the effective sample size is significantly smaller or greater than the expected sample size, these estimators (and mainly $\hat{\tau}_{ad1}$) show a significant degradation in precision. These patterns are particularly notable in the small areas with a very small sample size.

On the other hand, although the synthetic estimators show very high bias and bias ratios (resulting in very low conditional coverage rates for a design-based confidence interval), they are seen to be approximately constant and therefore independent of the effective sample size for each small area. The conditional variation coefficient is clearly constant showing independence from the sample size in the small area. The combined result of these patterns is a relative standard error which is also approximately invariant with the effective sample size. From this conditional point of view, the synthetic regression estimator can still be considered one of the most precise estimators.

For effective sample sizes that are smaller than the expected sample size the combined regression estimator with sample-size dependent weights, $\widetilde{\tau}_{ad3}$, shows important conditional biases and bias ratios that for significant departures are similar to those observed for the synthetic estimators. For sample sizes that are greater than expected $\widetilde{\tau}_{ad3}$ tends to show a behavior similar to the direct regression estimator $\hat{\tau}_{ad2}$. Therefore, the resulting conditional coverage rates for a design-based confidence interval also show a behavior similar to a synthetic estimator for sample sizes that are smaller than expected and similar to the direct regression estimator when they are higher than expected. The relative standard error also tends to show the bad property observed for the direct estimator, characterized by an increase with the effective sample size, mainly for the smallest areas.

The combined estimators $\widetilde{\tau}_{ad7}$ and $\widetilde{\tau}_{ad8}$ show interesting conditional properties as they show a mixed behavior between the direct regression estimator $\hat{\tau}_{ad2}$ and the synthetic regression estimator $\hat{\tau}_{ad5}$. This behavior is characterized by a significant resistance of precision and bias to departures between the effective sample size and the expected sample size.

It can be observed that $\widetilde{\tau}_{ad7}$ shows a conditional bias and bias ratio that are approximately constant, although with a slight tendency to increase with the reduction of the effective sample size. As to bias, this estimator shows a clear advantage when compared to the synthetic estimators and even when compared to the direct estimators and the sample-size dependent combined estimator, particularly when the sample size departs from the expected sample size. This results in conditional coverage rates which in extreme situations are closer to the confidence level than those associated with some design-based estimators. The conditional variance only shows a very slight tendency to rise with an increase in the effective sample size, resulting in a conditional relative standard error that is approximately constant. From the precision point of view it can be seen that this estimator is still competitive with the best synthetic estimator and maintains important precision gains when compared to the direct estimators and the sample-size dependent combined estimator (especially when the effective and expected sample sizes are different).

The estimator $\widetilde{\tau}_{ad8}$ magnifies these bias and bias ratio reductions as it continues to show smaller conditional bias and bias ratios than $\widetilde{\tau}_{ad7}$. Although not seen in this illustrative small area, global results showed that these bias reductions are particularly notable for effective sample sizes smaller than the expected sample sizes. In fact, it is in the smaller areas and particularly when the effective sample sizes are smaller than expected that there is an opportunity to reduce bias by borrowing information from other domains through the use of spatial correlations. The conditional variation coefficient still shows significant resistance to departures from the expected sample size. For sample sizes below the expected the variation coefficient tends to be slightly higher than the one obtained for $\widetilde{\tau}_{ad7}$, and still shows a pattern of a slight increase with the growth of the effective sample size. This increase is now lessened since a part of the variance is due to data provided by small areas in the neighborhood of the target small area. The conditional relative standard error continues to be reasonably constant and not substantially higher than the one obtained for $\widetilde{\tau}_{ad7}$. Overall, it can be concluded that among the combined estimators analyzed $\widetilde{\tau}_{ad8}$ is the estimator that exhibits the best conditional properties for bias and coverage rates for a design-based confidence interval.

5. MAIN FINDINGS AND DISCUSSION

The results of the empirical study show that the combined estimators obtained from the model classes proposed can compete in precision with the best synthetic estimators analyzed, while also allowing large reductions at the level of bias and, particularly, the bias ratio. They manage to show better precision than synthetic estimators for very small domains, and thus provide an important alternative to such estimators. The results attained seem to confirm that the combination of a synthetic and a direct component manages to take into account a significant part of the bias in the purely synthetic estimator, trading it for an increase in variance.

It should be noted that for this population the proposed estimators only prove interesting for inference related to domains of a small expected sample size (up to 5–10 units for the population analyzed). For larger sample sizes they cease to show precision gains in comparison with the best direct estimators (particularly with some direct modified regression estimators).

When the adjusted data displays spatial variability, the estimators that take advantage of the spatial correlation between observations tend to present reductions in bias (and mainly bias ratio) when compared with estimators that ignore this variability. These reductions are usually accompanied by a modest

loss of precision, resulting in bias ratios that are generally substantially lower than those obtained for these other estimators. This fact is easy to understand if we take into account that the consideration of spatial information implies the use of observations that are exogenous to each small area when estimating its random effect. It is natural that the inclusion of such information will also introduce some additional variability in the resulting estimator. The spatial information permits a repositioning of the estimator, which will display behavior that is further away from that presented by a synthetic estimator and gain the characteristics of a direct estimator. It should be pointed out that when the sample size in the inference target domain is very small or even non-existent, the introduction of spatial information relating to other domains can prevent the estimators being reduced to 'pure' synthetic estimators and maintain mixed characteristics between a synthetic and a design-based estimator. This fact helps to explain the good behavior of these estimators in domains with a very small sample size.

The proposed estimators clearly show interesting conditional characteristics, as they tend to behave in a way that is typified by strong robustness, both in precision and bias, to differences in effective and expected sample size. Their remarkable conditional behavior is clearly demonstrated by the fact that their conditional bias ratios are in many cases lower than those registered for direct estimators, specially when there are significant discrepancies between the effective and expected sample size. In particular, estimators that exploit spatial correlation continued to show reductions in conditional bias and conditional bias ratios when compared with estimators that ignore this variability. In fact, we can conclude that while the proposed estimator shows interesting unconditional properties, it is within a conditional point of view that its advantage over competitive estimators strikes.

One of the main limitations of this study lies on the fact that only the specification of isotropic spatial covariance structures was considered. In fact, in a context where the differences between the coast and the hinterland are presumably very different from those between the north and south, resort to anisotropic spatial models can allow the reality to be more satisfactorily represented. However, the sheer complexity of calculation presented by these structures, arising from the need to process a considerable amount of data, rendered the estimation of these models unviable. Although the proposed estimator is thought to be applicable to the context where data of spatial nature is present, it would be interesting to test its application to other contexts, whenever is possible to establish some kind of proximity between the small areas of study.

It should be also stressed that the conclusions presented are depended of the used data set. Although we have used a realistic data set based on real data from a National Statistical Office, the use of different data sets, for example exhibiting different spatial correlation, can lead to different results and possibly different conclusions. Therefore, the proposed estimators should be tested with

other sets of real and artificial data before they are selected for application in other contexts. In fact, empirical studies have revealed to be a fundamental stage in the process of choosing an estimator. The results of such studies can, moreover, help to create greater confidence on the part of potential users of these kinds of estimators.

A. APPENDIX 1

The estimation of τ_d is performed thought the prediction of the realizations of the vector $\mathbf{y}_d$. Under the model (3.2) the EBLUP is:

$$\begin{aligned}\widetilde{\mathbf{y}}_d &= \mathbf{X}_d\widetilde{\boldsymbol{\beta}} + \mathbf{Z}_d\widetilde{\mathbf{u}} + \widetilde{\boldsymbol{\epsilon}}_d \\ &= \mathbf{X}_d\widetilde{\boldsymbol{\beta}} + \mathbf{Z}_d\mathbf{G}\,\mathbf{Z}_s'\mathbf{V}_s^{-1}\big(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}}\big) + \mathbf{R}_{d,s}\mathbf{V}_s^{-1}\big(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}}\big) \\ &= \mathbf{X}_{ad}\widetilde{\boldsymbol{\beta}} + \mathbf{V}_{ad,s}\mathbf{V}_s^{-1}\big(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}}\big)\,,\end{aligned}$$

where the subscript d indicates that the respective matrices only include observations from the small area d, $\mathbf{R}_{d,s} = E(\boldsymbol{\epsilon}_d\boldsymbol{\epsilon}_s)$ and $\mathbf{V}_{ad,s} = E\big[(\mathbf{y}_d - \mathbf{X}_d\boldsymbol{\beta})\,(\mathbf{y}_d - \mathbf{X}_d\boldsymbol{\beta})'\big]$ $= E(\mathbf{Z}_d\,\mathbf{u}\,\mathbf{u}'\,\mathbf{Z}_s') + E(\boldsymbol{\epsilon}_d\boldsymbol{\epsilon}_s) = \mathbf{Z}_d\mathbf{G}\,\mathbf{Z}_s' + \mathbf{R}_{d,s}$. With the EBLUP $\widetilde{\mathbf{y}}_d$, the estimator of τ_d may be obtained as:

$$\begin{aligned}\widetilde{\tau}_d = \sum_{i\in U_d}\widetilde{y}_{di} &= \boldsymbol{\tau}_{\mathbf{x}d}'\widetilde{\boldsymbol{\beta}} + \boldsymbol{\tau}_{\mathbf{z},d}'\mathbf{G}\,\mathbf{Z}_s'\mathbf{V}_s^{-1}\big(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}}\big) + \mathbf{1}_{N_d}'\mathbf{R}_{d\cdot,s}\mathbf{V}_s^{-1}\big(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}}\big) \\ &= \boldsymbol{\tau}_{\mathbf{x}d}'\widetilde{\boldsymbol{\beta}} + \mathbf{v}_{\boldsymbol{\tau}s}'\mathbf{V}_s^{-1}\big(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}}\big)\,,\end{aligned}$$

where $\mathbf{R}_{d\cdot,s} = E(\boldsymbol{\epsilon}_{ad}\boldsymbol{\epsilon}_s)$ and $\mathbf{v}_{\boldsymbol{\tau}s}' = E\big[(\tau_d - \boldsymbol{\tau}_{\mathbf{x},d}'\boldsymbol{\beta})\,(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta})'\big] = \boldsymbol{\tau}_{\mathbf{z},d}'\mathbf{G}\,\mathbf{Z}_s' + \mathbf{1}_{N_d}'\mathbf{R}_{d\cdot,s}$.

B. APPENDIX 2

The vector $\widetilde{\mathbf{y}}_d$ may be decomposed into $\widetilde{\mathbf{y}}_d' = (\widetilde{\mathbf{y}}_{d,s}', \widetilde{\mathbf{y}}_{d,r}')'$. From mixed model theory it is straightforward that $\widetilde{\mathbf{y}}_{d,s} = \mathbf{y}_{d,s}$, with the unobservable part of $\mathbf{y}_d$ predicted by

$$\widetilde{\mathbf{y}}_{d,r} = \mathbf{X}_{d,r}\widetilde{\boldsymbol{\beta}} + \mathbf{Z}_{d,r}\widetilde{\mathbf{u}} + \widetilde{\boldsymbol{\epsilon}}_{d,r} = \mathbf{X}_{d,r}\widetilde{\boldsymbol{\beta}} + \mathbf{V}_{dr,s}\mathbf{V}_s^{-1}\big(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}}\big)\,,$$

where $\mathbf{R}_{dr,s} = E(\boldsymbol{\epsilon}_{d,r}\,\boldsymbol{\epsilon}_s)$ and $\mathbf{V}_{dr,s} = E\big[(\mathbf{y}_{d,r} - \mathbf{X}_{d,r}\boldsymbol{\beta})\,(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta})'\big] = E(\mathbf{Z}_{d,r}\,\mathbf{u}\,\mathbf{u}'\,\mathbf{Z}_s')$ $+ E(\boldsymbol{\epsilon}_{d,r}\,\boldsymbol{\epsilon}_s) = \mathbf{Z}_{d,r}\,\mathbf{G}\,\mathbf{Z}_s' + \mathbf{R}_{dr,s}$. When $\mathbf{R}_{dr,s}$ is a null matrix, then the covariances between the unobservable vector $\mathbf{y}_{d,r}$ and the observable vector $\mathbf{y}_s$ are uniquely determined by the random effects $\mathbf{u}$. Consequently the EBLUP of $\mathbf{y}_{d,r}$ coincides with the EBLUP of $E(\mathbf{y}_{d,r}|\mathbf{u})$:

$$\widetilde{\mathbf{y}}_{d,r} = \widetilde{E}(\mathbf{y}_{d,r}|\mathbf{u}) = \mathbf{X}_{d,r}\widetilde{\boldsymbol{\beta}} + \mathbf{Z}_{d,r}\mathbf{G}\,\mathbf{Z}_s'\mathbf{V}_s^{-1}\big(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}}\big)\,.$$

The EBLUP total for the unobservable part of the small area $\tau_{d,r}$ is now equal to the EBLUP of $E(\tau_{d,r}|\mathbf{u})$, with

$$\begin{aligned}\widetilde{\tau}_{d,r} \;=\; \widetilde{E}(\tau_{d,r}|\mathbf{u}) \;&=\; \boldsymbol{\tau}'_{\mathbf{x},d,r}\widetilde{\boldsymbol{\beta}} + \sum_{h=1}^{H} \boldsymbol{\tau}'_{\mathbf{x}(1),dh,r}\widetilde{\mathbf{u}}_h^{(1)} + \boldsymbol{\tau}'_{\mathbf{x}(2),d,r}\widetilde{\mathbf{u}}_d^{(2)} \\ &=\; \boldsymbol{\tau}'_{\mathbf{x},d,r}\widetilde{\boldsymbol{\beta}} + \boldsymbol{\tau}'_{\mathbf{z},d,r}\,\mathbf{G}\,\mathbf{Z}'_s\mathbf{V}_s^{-1}\big(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}}\big)\;,\end{aligned}$$

and the estimator of τ_d is given in a simplified expression by

$$\widetilde{\tau}_d \;=\; \tau_{\mathbf{y},d,s} + \boldsymbol{\tau}'_{\mathbf{x},d,r}\widetilde{\boldsymbol{\beta}} + \boldsymbol{\tau}'_{\mathbf{z},d,r}\,\mathbf{G}\,\mathbf{Z}'_s\mathbf{V}_s^{-1}\big(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}}\big)\;,$$

where $\tau_{\mathbf{y},d,s}$ is the observed sample total in small area d.

REFERENCES

[1] Amemiya, T. (1971). The Estimation of the Variances in a Variance-Components Model, *International Economic Review*, **12**, 1–13.

[2] Barndorff-Nielsen, O. (1983). On a formula for the distribution of the maximum likelihood estimator, *Biometrika*, **70**(2), 343–365.

[3] Battese, G.E.; Harter, R.M. and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data, *Journal of the American Statistical Association*, **83**, 28–36.

[4] Bernardinelli, L. and Montomoli, C. (1992). Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk, *Statistics in Medicine*, **11**, 983–1007.

[5] Chandra, H.; Salvati, N. and Chambers, R. (2007). Small area estimation for spatially correlated populations — a comparison of direct and indirect model-based methods, *Statistics in Transition – new series*, **8**(2), 331–350.

[6] Cressie, N. (1991). *Statistics for Spatial Data*, John Wiley & Sons, New York.

[7] Diggle, P. (1988). An approach to the analysis of repeated measurements, *Biometrics*, **44**, 959–971.

[8] Fuller, W.A. and Battese, G.E. (1973). Transformations for estimation of linear models with nested-error structures, *Journal of the American Statistical Association*, **68**, 625–632.

[9] Hartley, H. and Rao, J.N.K. (1967). Maximum-likelihood estimation for the mixed analysis of variance model, *Biometrika*, **54**, 93–108.

[10] Harville, D.A. (1974). Bayesian inference for variance components using only error contrasts, *Biometrika*, **61**(2), 383–385.

[11] Harville, D.A. (1977). Maximum likelihood approaches to variance components estimation and to related problems, *Journal of the American Statistical Association*, **72**, 320–337.

[12] Henderson, C.R. (1953). Estimation of variance and covariance components, *Biometrics*, **9**, 226–252.

[13] Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model, *Biometrics*, **31**, 423–447.

[14] Heyde, C.C. (1997). *Quasi-Likelihood and Its Application*, Springer-Verlag, New York.

[15] Jennrich, R. and Schluchter, M. (1986). Unbalanced repeated-measures models with structured covariance matrices, *Biometrics*, **42**, 805–820.

[16] Jiang, J. (1996). REML estimation: Asymptotic behaviour and related topics, *The Annals of Statistics*, **24**(1), 255–286.

[17] Jiang, J. (1997). A derivation of BLUP — best linear unbiased predictor, *Statistics and Probability Letters*, **25**, 321–324.

[18] Jiang, J. (1998). Consistent estimators in generalized linear mixed models, *Journal of the American Statistical Association*, **93**, 720–729.

[19] Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*, Springer-Verlag, New York.

[20] Kott, P. (1989). Robust small domain estimation using random effects modeling, *Survey Methodology*, **15**(1), 3–12.

[21] Laird, N.M. and Ware, J.H. (1982). Random-effects models for longitudinal data, *Biometrics*, **38**, 963–974.

[22] McLean, R.; Sanders, W. and Stroup, W. (1991). A unified approach to mixed linear models, *The American Statistician*, **45**, 54–64.

[23] Miller, J.J. (1977). Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance, *The Annals of Statistics*, **5**(4), 146–762.

[24] Patterson, H.D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal, *Biometrika*, **58**(3), 545–554.

[25] Petrucci, A. and Salvati, N. (2004). *Small Area Estimation considering spatially correlated errors: the unit level random effects model*, Working Paper No. 2004/10, Dipartimento di Statistica "G. Parenti", Firenze.

[26] Petrucci, A. and Salvati, N. (2006). Small area estimation for spatial correlation in watershed erosion assessment, *Journal of Agricultural, Biological and Environmental Statistics*, **11**(2), 169–182.

[27] Pfeffermann, D. (1993). The role of sampling weights when modeling survey data, *International Statistical Review*, **61**(2), 317–337.

[28] Pinheiro, J. and Coelho, P. (2004). Spatial variability in the general mixed model, *Revista de Estatística*, **2**, 35–80.

[29] Prasad, N.G.N. and Rao, J.N.K. (1999). On robust small area estimation using a simple random effects model, *Survey Methodology*, **25**(1), 67–72.

[30] Pratesi, M. and Salvati, N. (2004). *Spatial EBLUP in agricultural survey. An application based on census data*, Report No. 256, Department of Statistics and Mathematics, University of Pisa, Pisa.

[31] Pratesi, M. and Salvati, N. (2004). *Small area estimation: the EBLUP estimator with autoregressive random area effects*, Report No. 261, Department of Statistics and Mathematics, University of Pisa, Pisa.

[32] Pratesi, M. and Salvati, N. (2008). Small area estimation: the EBLUP estimator based on spatially correlated random area effects, *Statistical Methods and Applications*, **17**(1), 113–141.

[33] Rao, C.R. (1970). Estimation of heteroscedastic variances in linear models, *Journal of the American Statistical Association*, **65**, 161–172.

[34] Rao, C.R. (1971). Estimation of variance and covariance components — MINQUE theory, *Journal of Multivariate Analysis*, **1**, 257–275.

[35] Rao, C.R. (1972). Estimation of variance and covariance components in linear models, *Journal of the American Statistical Association*, **67**, 112–115.

[36] Salvati, N. (2004). *Small area estimation by spatial models: the spatial empirical best linear unbiased prediction (spatial EBLUP)*, Working Paper No. 2004/3, Dipartimento di Statistica "G. Parenti", Firenze.

[37] Särndal, C.E. (1984). Design-consistent versus model-dependent estimators for small domains, *Journal of the American Statistical Association*, **79**, 624–631.

[38] Särndal, C.E. and Hidiroglou, M.A. (1989). Small domain estimation: a conditional analysis, *Journal of the American Statistical Association*, **84**, 266–275.

[39] Singh, M.B.; Gambino, J. and Mantel, H.J. (1994). Issues and strategies for small area data, *Survey Methodology*, **20**, 3–22.

[40] Singh, B.B.; Shukla, G.K. and Kundu, D. (2005). Spatio-temporal models in small area estimation, *Survey Methodology*, **31**(2), 183–195.

[41] Thompson, W.A. (1962). The problem of negative estimates of variance components, *The Annals of Mathematical Statistics*, **33**(1), 273–289.

[42] Verbyla, A.P. (1990). A conditional derivation of residual maximum likelihood, *Australian Journal of Statistics*, **32**(2), 227–230.

[43] Wolfinger, R. (1993). Covariance structures selection in general mixed models, *Communications in Statistics, Simulation and Computing*, **22**(4), 1079–1106.

REVSTAT – Statistical Journal
Volume 9, Number 2, June 2011, 181–198

GENERALIZED SUM PLOTS

Authors: J. Beirlant
– Department of Mathematics, Campus Kortrijk and Leuven Statistics Research Center, Katholieke Universiteit Leuven, Belgium
Jan.Beirlant@wet.kuleuven.be

E. Boniphace
– Department of Mathematics and Leuven Statistics Research Center, Katholieke Universiteit Leuven, Belgium
Edwin.Boniphace@wis.kuleuven.be

G. Dierckx
– Department of Mathematics and Statistics, Hogeschool-Universiteit Brussel, Belgium
Goedele.Dierckx@hubrussel.be

Received: September 2010 Revised: May 2011 Accepted: May 2011

Abstract:

- Sousa and Michailidis (2004) developed the sum plot based on the Hill (1975) estimator as a diagnostic tool for selecting the optimal k when the distribution is heavy tailed. We generalize their method to any consistent estimator with any tail type (heavy, normal and light tail). We illustrate the method associated to the generalized Hill estimator and the moment estimator.

  As an attempt to reduce the bias of the generalized Hill estimator, we propose new estimators based on the regression model which are based on the estimates of the generalized Hill estimator. Here weighted least squares and weighted trimmed least squares is proposed. The bias and the mean squared error (MSE) of the estimators is studied using a simulation study. A few practical examples are proposed.

Key-Words:

- *sum plot; generalized sum plot; extreme value analysis; generalized quantile plot; weighted regression model.*

AMS Subject Classification:

- 62G32, 62J05, 62F35, 62P05, 62P12.

1. INTRODUCTION

In order to estimate a tail index using k upper order statistics, one needs to determine an appropriate value of k. There exist a variety of diagnostic plots and adaptive estimation methods that assist in threshold selection. The list of plots includes Zipf, Hill, empirical mean-excess and sum plots. Adaptive selection procedures are listed for instance in Beirlant *et al.* (2005). The aim of this paper is to generalize the graphical tool developed by Sousa and Michailidis (2004) assisting in choosing a sensible estimate or a value of k. Their sum plot is based on the assumption that the distribution is heavy tailed. We extend the approach to all estimators which use a set of extreme order statistics in the estimation of a real valued extreme value index. Here we illustrate the approach using the generalized Hill estimator introduced in Beirlant *et al.* (1996b) and the moment estimator proposed by Dekkers *et al.* (1989).

In this paper we also propose new estimators of the extreme value index based on the regression associated to the estimates of the (generalized) Hill estimator for various $k = 1, ..., K$ for some K.

The article is organized as follows. In section 2, we first specify the original sum plot in subsection 2.1. Then we generalize it using the generalized Hill estimator in subsection 2.2 and using the moment estimator in subsection 2.3. In subsection 2.4 we illustrate the method with some simulation results. The new estimators based on regression models are introduced in section 3, first for the original Hill sum plot in subsection 3.1 and then for the generalized Hill sum plot in 3.2. Finally, some simulations and practical examples are presented in subsections 3.3 and 3.4.

2. SUM PLOTS

The sum plot by Sousa and Michailidis (2004) and Henry III (2009), are examples of the following principle. Let $\hat{\gamma}_{k,n}$ (which uses k upper order statistics from the total sample of size n) be a consistent estimator of γ as $k, n \to \infty$ and $k/n \to 0$. Assume first that $\hat{\gamma}_{k,n}$ is an unbiased estimator i.e. $\mathbf{E}\hat{\gamma}_{k,n} = \gamma$. Define the random variables S_k, for $k = 1, 2, ..., n-1$, by

$$S_k := k\,\hat{\gamma}_{k,n} \tag{2.1}$$

then $\mathbf{E}S_k = k\,\gamma$. Therefore the plot (k, S_k) is approximately linear for the range of k where $\hat{\gamma}_{k,n} \approx \gamma$, i.e. $\hat{\gamma}_{k,n}$ is constant in k. The slope of the linear part of the graph (k, S_k) can then be used as an estimator of γ. Assume now that $\hat{\gamma}_{k,n}$ is a consistent estimator but biased, that is $\mathbf{E}\hat{\gamma}_{k,n} = \gamma + (\text{bias})$, then $\mathbf{E}S_k =$

$k\,\gamma + k(\text{bias})$. If the bias is constant in k then (k, S_k) is again linear with the slope equal to $\gamma + (\text{bias})$. Typically though the bias is not constant in k and hence the path of (k, S_k) will depend on the non constant function in k defining the bias.

The sum plot introduced in Sousa and Michailidis (2004) is based on the Hill estimator (Hill, 1975). The sum plot by Henry III (2009) is based on a harmonic moment estimator. Both proposals were limited to the family of Pareto-type distributions.

So given $\hat{\gamma}_{k,n}$, any consistent estimator of γ based on k top order statistics, we propose a sum plot (k, S_k) based on $\hat{\gamma}_{k,n}$ with S_k defined in (2.1). The only strong assumption on $\hat{\gamma}_{k,n}$ is consistency which is a natural requirement on any estimator. This plot could be helpful in identifying an appropriate region of k, the number of order statistics to be used in $\hat{\gamma}_{k,n}$. One could argue that the plots (k, S_k) and $(k, \hat{\gamma}_{k,n})$ are statistically equivalent. The sum plot naturally leads to the estimation of the slope whereas $(k, \hat{\gamma}_{k,n})$ leads to horizontal plots and hence estimation of the intercept. Here we consider the case of a real-valued γ and hence increasing or decreasing sum plots allow to assess the sign of γ.

Since each estimator will have its own sum plot, we hereafter name the associated sum plot along the name of the estimator. For example the sum plot based on the Hill estimator is named the Hill sum plot.

In the following subsections we illustrate the proposed sum plot principle using the Hill, the generalized Hill and the moment estimator. We also illustrate the performance of these sum plots on simulated data and on some real data sets.

2.1. The Hill sum plot

Let $X_{1,n} < X_{2,n} < \cdots < X_{n,n}$ denote the order statistics of a random sample $(X_1, X_2, ..., X_n)$ from a heavy tailed distribution F with

$$1 - F(x) = x^{-1/\gamma}\, l_F(x)\ , \qquad x > 0\ , \tag{2.2}$$

where l_F is a slowly varying function at infinity satisfying

$$l_F(\lambda x)/l_F(x) \to 1 \quad \text{when} \quad x \to \infty\,, \qquad \text{for all} \quad \lambda > 0\ .$$

Let the random variables $S^H_{k,n}$ $(k = 1, ..., n)$ be defined as

$$S^H_{k,n} = \sum_{j=1}^{k} Z_j := \sum_{j=1}^{k} j \log \frac{X_{n-j+1,n}}{X_{n-j,n}}\ . \tag{2.3}$$

Sousa and Michailidis (2004) introduced the diagnostic plot $(k, S^H_{k,n})$, the sum plot for estimating the tail index γ. This plot is called the Hill sum plot since

the Hill (1975) estimator $H_{k,n}$ satisfies

$$H_{k,n} \,=\, \frac{1}{k}\sum_{j=1}^{k} \log X_{n-j+1,n} - \log X_{n-k,n} \,=\, \frac{1}{k}\, S^H_{k,n}\ . \tag{2.4}$$

To understand the behavior of the Hill sum plot we rely on a representation of the variables Z_j from (2.3) ($j = 1, ..., n$) provided in Beirlant *et al.* (2001). We remind that the model (2.2) is well-known to be equivalent to

$$U(x) \,=\, x^{\gamma}\, l_U(x)\ , \tag{2.5}$$

where $U(x) = \inf\{y\colon F(y) \geq 1 - 1/x\}$ ($x > 1$) and with l_U again a slowly varying function. Often, the following second order condition on l_U is assumed

$$\frac{l_U(tx)}{l_U(x)} \,=\, 1 + b(x)\,\frac{t^{\rho}-1}{\rho}\,\big(1+o(1)\big)\ ,$$

where b is a rate function satisfying $b(x) \to 0$ as $x \to \infty$ and $\rho < 0$. Under this second order condition, Beirlant *et al.* (2001) have shown that

$$\left| Z_j \,-\, \left(\gamma + b_{n,k}\Big(\frac{j}{k+1}\Big)^{-\rho}\right) E_j \,+\, \beta_j \right| \,=\, o_P(b_{n,k})\ , \tag{2.6}$$

uniformly in $j \in \{1, ..., k\}$, as $k, n \to \infty$ with $k/n \to 0$, where $(E_1, ..., E_k)$ is a vector of independent and standard exponentially distributed random variables, $b_{n,k} := b\big((n+1)/(k+1)\big)$, $2 \leq k \leq n-1$ and $\frac{1}{k}\sum_{j=1}^{k} \beta_j = o_P(b_{n,k})$.

Hence, for $S^H_k = \sum_{j=1}^{k} Z_j$,

$$\left| S^H_k \,-\, \left(k\,\gamma \,+\, k\, b_{n,k}/(1-\rho) \,+\, \gamma \sum_{j=1}^{k} (E_j - 1) \,+\, o_P(k\, b_{n,k})\right)\right| \,=\, o_P(k\, b_{n,k})$$

since $\frac{1}{k}\sum_{j=1}^{k}\big(\frac{j}{k+1}\big)^{-\rho} \sim \frac{1}{1-\rho}$ as $k, n \to \infty$ with $k/n \to 0$, where as usual, $a_n \sim b_n$ is equivalent to $a_n/b_n \to 1$ as $n \to \infty$.

In the specific case where $b(x) = C x^{\rho}\big(1 + o(1)\big)$ for some real constant C (Hall, 1982), then we obtain

$$\left| S^H_k \,-\, \left(k\,\gamma + C\, n^{\rho} k^{1-\rho} + \gamma \sum_{j=1}^{k} (E_j - 1)\right) + o_P(k\, b_{n,k})\right| \,=\, o_P(k\, b_{n,k})\ . \tag{2.7}$$

Sousa and Michailidis (2004) only considered the case $C = 0$.

The Hill sum plot is a graphical tool in which one is searching for a range of k where the sum plot is linear, or equivalently where $H_{k,n}$ is constant in k, if such a behaviour becomes apparent. Whereas the Hill estimator can be seen as

an estimator of the slope in a Pareto quantile plot (see for instance Beirlant *et al.* (1996a) and Kratz and Resnick (1996)), the sum plot now can be viewed as a regression plot from which new estimators can be constructed by regression of $S^H_{k,n}$ on k, as suggested by (2.7). As the regression error will turn out smaller on the sums of noise variables $\gamma(E_j - 1)$ rather than on extreme log-data, regression on the sum plot appears to be an interesting alternative approach. In practice we put $\rho = -1$ so that in that case we will fit a quadratic regression model as discussed in Section 3. The second order parameter could be replaced by estimators such as discussed in Fraga Alves *et al.* (2003). In the simulation study the case of the Burr distribution with $\rho = -0.5$ gives an idea of the loss of accuracy by setting $\rho = -1$.

2.2. The generalized Hill sum plot

Using a similar approach we derive the generalized sum plot for $\gamma \in \mathbb{R}$ based on the generalized Hill estimator by Beirlant *et al.* (1996b). Here the underlying model is that the distribution belongs to a maximum domain of attraction: there exist sequences of constants $(a_n;\, a_n > 0)$ and (b_n) such that

$$\lim_{n\to\infty} P\left(\frac{X_{n,n} - b_n}{a_n} \le x\right) = \exp\bigl(-(1+\gamma x)^{-1/\gamma}\bigr)\,, \qquad 1+\gamma x > 0\,.$$

Define the function UH as follows

$$UH := U(x)\,\mathbf{E}\Bigl(\log X - \log U(x) \;\Big|\; X > U(x)\Bigr)\,. \tag{2.8}$$

This function possesses the regular variation property for the full range of γ. The empirical counterpart of UH at $x = n/k$ is given by

$$UH_{k,n} := X_{n-k,n}\left(\frac{1}{k}\sum_{i=1}^{k} \log X_{n-i+1,n} - \log X_{n-k,n}\right) = X_{n-k,n}\, H_{k,n}\,. \tag{2.9}$$

Using the property of regular variation of UH, Beirlant *et al.* (1996b) proposed an estimator of $\gamma \in \mathbb{R}$ by fitting a constrained least-squares line to the points with coordinates $(-\log(j/n), \log UH_{j,n})$ $(j = 1, ..., k)$ to obtain the generalized Hill estimator $H^*_{k,n}$. Similarly as in (2.4), $H^*_{k,n}$ is given by

$$H^*_{k,n} = \frac{1}{k}\sum_{i=1}^{k}\left((i+1)\,\log\frac{UH_{i,n}}{UH_{i+1,n}} + \frac{i+1}{i} - (i+1)\log\frac{i+1}{i}\right). \tag{2.10}$$

Define random variables S^{UH}_k, for $k = 1, ..., n-2$, as

$$S^{UH}_k := \sum_{i=1}^{k}(i+1)\left(\log\left(\frac{X_{n-i,n}}{X_{n-i-1,n}}\,\frac{H_{i,n}}{H_{i+1,n}}\right) + \frac{1}{i} + \log\frac{i+1}{i}\right). \tag{2.11}$$

Since $S_k^{UH} = k\,H_{k,n}^*$, we obtain the generalized Hill sum plot (k, S_k^{UH}), and we expect that for the range of k where $H_{k,n}^*$ is constant (or stable) the plot will be linear. Note that the range of k where the Hill estimator is constant, the term $H_{j,n}/H_{j+1,n} \to 1$, and (2.11) is almost reduced to S_k^H, except that the term including the largest observation is deleted.

Under general second order regular variation conditions, in Dierckx (2000) it is shown that for $1 \le j \le k$, $2 \le k \le n-2$, it holds for

$$Z_j^* := (j+1)\left(\left(\log UH_{j,n} - \log UH_{j+1,n}\right) + \frac{1}{j} + \log\frac{j+1}{j}\right)$$

that

$$\left| Z_j^* - \left(\left(\gamma + \tilde{b}_{n,k}\left(\frac{j+1}{k+1}\right)^{-\rho}\right)E_{j+1} + \gamma\,(E_{j+1}-1) + (j+1)\left(\log\frac{\bar{E}_j}{\bar{E}_{j+1}} - \log\frac{j+1}{j} + \frac{1}{j}\right)\right) + \tilde{\beta}_j \right| = o_P\big(\tilde{b}_{n,k}\big) \tag{2.12}$$

as $k, n \to \infty$ with $k/n \to 0$, where $(E_1, ..., E_k)$ is a vector of independent and standard exponentially distributed random variables, $\bar{E}_j$ denotes the sample mean of $(E_1, ..., E_j)$, $\tilde{b}_{n,k}$ is some generic notation for a function decreasing to zero, $\rho < 0$ and $\frac{1}{k}\sum_{j=1}^{k}\tilde{\beta}_j = o_P(b_{n,k})$. Note also that for $\gamma < 0$, the above expression only holds for $j \to \infty$.

Let us denote $e_j := \gamma\,(E_{j+1}-1) + (j+1)\left(\log\frac{\bar{E}_j}{\bar{E}_{j+1}} - \log\frac{j+1}{j} + \frac{1}{j}\right)$. In Dierckx (2000) it is shown that

$$\begin{aligned} &E\,e_i = 0\ , \\ &\mathrm{Cov}(e_i, e_j) = \frac{\gamma}{j}\ , \qquad i < j\ , \\ &\mathrm{Var}(e_i) = (\gamma-1)^2 + \frac{1+2\,i}{i^2}\ . \end{aligned} \tag{2.13}$$

Model (2.12) is a direct generalization of the regression model (6) used in the Hill sum plot, leading to a generalized Hill sum plot regression approach. In practice we fit the regression model

$$S_k^{UH} = k\,\gamma + C_\rho k^{1-\rho} + \sum_{j=1}^{k} e_j\ . \tag{2.14}$$

In the simulations below we will replace ρ by the canonical choice -1 so that later on we fit a quadratic regression model to the responses S_k^{UH}, $k = 1, ..., K$ for some $K > 0$.

2.3. The moment sum plot

Let $H_{k,n}^{(2)}$ be defined as follows

$$H_{k,n}^{(2)} := \frac{1}{k}\sum_{i=1}^{k}\Big(\log X_{n-i+1,n} - \log X_{n-k,n}\Big)^2 .$$

The moment estimator $M_{k,n}$ (Dekkers *et al.* (1989)) is given by

$$M_{k,n} := H_{k,n} + 1 - \frac{1}{2}\left(1 - \frac{H_{k,n}^2}{H_{k,n}^{(2)}}\right)^{-1} \tag{2.15}$$

where $H_{k,n}$ is the Hill estimator from (2.4).

Let the random variables S_k^M, for $k = 1, ..., n-1$, be defined as

$$S_k^M := \left(\sum_{i=1}^{k}\log X_{n-i+1,n} - \log X_{n-k,n}\right) + k - \frac{k}{2}\left(1 - \frac{H_{k,n}^2}{H_{k,n}^{(2)}}\right)^{-1} . \tag{2.16}$$

Definition (2.16) is equivalent to writing $S_k^M = k\, M_{k,n}$, hence (k, S_k^M) is the moment sum plot. However here a regression model has not been established to the best of our knowledge.

2.4. Simulation results

The different sum plots have been applied to some simulated data sets. Six distributions are considered:

- The strict Pareto distribution given by $F(x) = 1 - x^{-1/\gamma}$, $x > 1$, $\gamma > 0$. We have chosen $\gamma = 1$. Here $b(x) = 0$.
- The standard Fréchet distribution given by $F(x) = \exp(-x^{-1/\gamma})$, $x > 0$, $\gamma > 0$. We have chosen $\gamma = 1$. Here $\rho = -1$.
- The Burr distribution $F(x) = 1 - \Big(\frac{\eta}{\eta + x^{-\tau}}\Big)^{\lambda}$, $x > 0$, $\eta, \tau, \lambda > 0$. We have chosen $\eta = 1$, $\tau = 0.5$, $\lambda = 2$, such that $\gamma = 1$. Here $\rho = -1/\lambda = -0.5$.
- The gamma distribution $F(x) = \frac{1}{b^a\,\Gamma(a)}\int_0^x t^{a-1}\exp(-t/b)\,dt$, $x > 0$, $a, b > 0$. Here we have chosen $a = 2$, $b = 1$. Always, $\gamma = 0$.
- The uniform distribution $F(x) = x$ $(0 < x < 1)$. Here $\gamma = -1$.
- The reversed Burr distribution $F(x) = 1 - \Big(\frac{\beta}{\beta + (x_+ - x)^{-\tau}}\Big)^{\lambda}$, $x > 0$, $\eta, \tau, \lambda > 0$, x_+ denotes the right endpoint of the distribution. We have chosen $\eta = 1$, $\tau = 0.5$, $\lambda = 2$, $x_+ = 2$, such that $\gamma = -1$. Here $\rho = -1/\lambda = -0.5$.

The Hill sum plots, the generalized Hill sum plots and the moment sum plots of these distributions are shown in Figure 1.

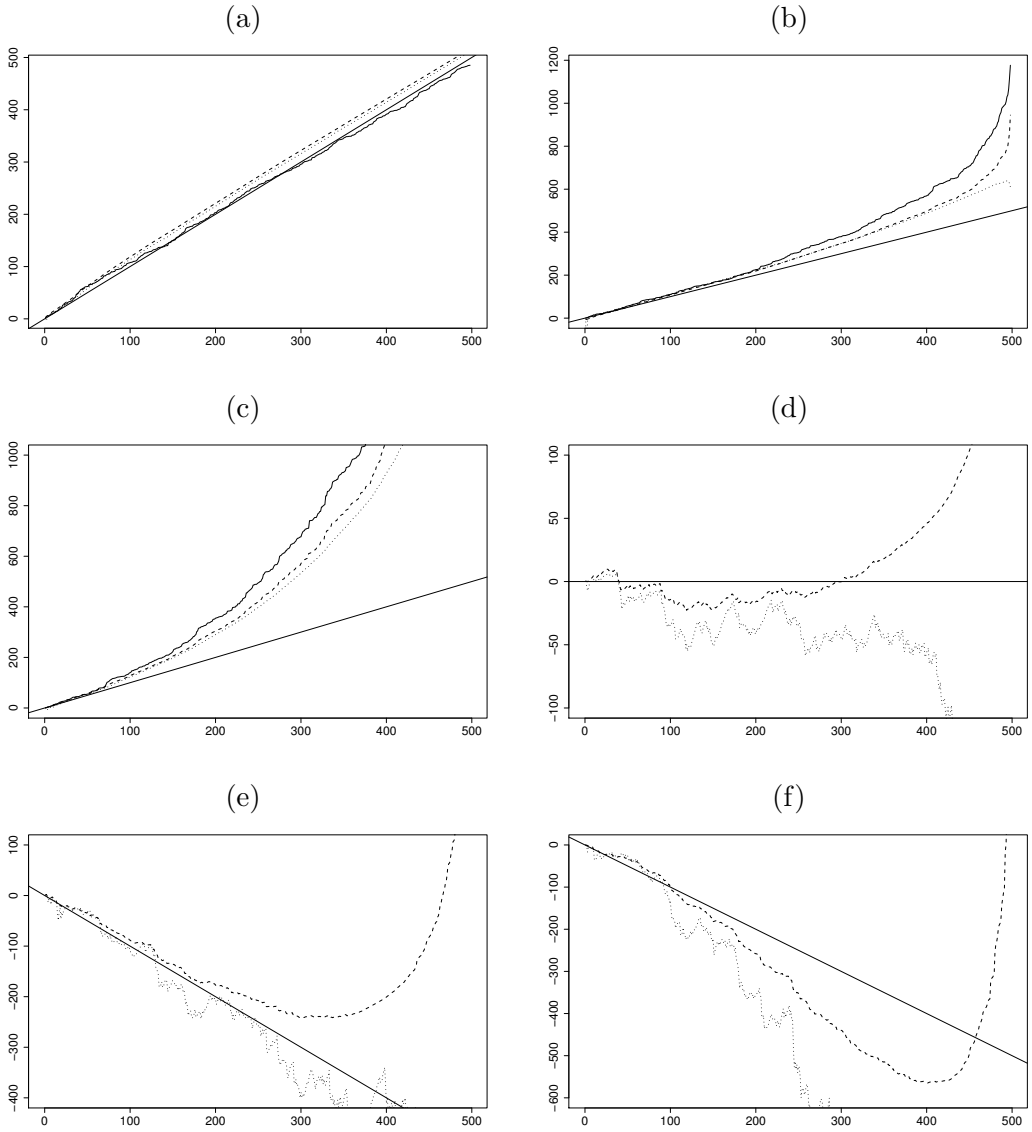

Figure 1: The Hill sum plot (full line); generalized Hill sum plot (dashed line) and the moment sum plot (dotted line) are plotted for simulated data sets of size n=500 from the: (a) strict Pareto; (b) Fréchet; (c) Burr; (d) gamma; (e) uniform; (f) reversed Burr distribution.

For $\gamma > 0$, the three sum plots are comparable for the linear parts of the plots. For $\gamma = 0$ and $\gamma < 0$, the generalized Hill sum plot and the moment sum plot are comparable on these particular data sets. However the generalized Hill sum plot seems to be less volatile. The sum plots can be used to identify the sign of γ

for a given data set: increase in k indicates $\gamma > 0$, a horizontal pattern indicates $\gamma = 0$ and decrease in k indicates $\gamma < 0$. In future work this could be used to test the domain of attraction condition. For an overview of this problem we can refer to Neves and Fraga Alves (2008). Moreover, the linear part of these generalized sum plots can be used to estimate the value of tail index.

3. REGRESSION ESTIMATORS

As indicated before, we propose regression estimators for the extreme value index γ based on the slope of the Hill and the Generalized Hill sum plots.

3.1. Hill sum plot estimators

Huisman *et al.* (2001) introduced a new estimator for $\gamma > 0$ based on the Hill sum plot which can be understood from (6). It indeed follows from (6) that for some constant D

$$\left| H_{k,n} - \left(\gamma + Dk^{-\rho} + \epsilon_k\right) \right| = o_P(b_{n,k}) \tag{3.1}$$

where $\epsilon_k = \gamma/k \sum_{j=1}^{k}(E_j - 1)$, leading to the regression model

$$H_{k,n} = \gamma + Dk^{-\rho} + \epsilon_k \ , \qquad k = 1, ..., K \ . \tag{3.2}$$

Since the variance of the error term $\mathrm{Var}(\epsilon_k) = \gamma^2/k$ is not constant, a weighted least squares regression is applied with a $K \times K$ diagonal weight matrix $W = \mathrm{diag}\sqrt{1}, ..., \sqrt{K}$. Note that in this way, Huisman *et al.* (2001), did not take into account that the error terms are not independent.

In practice, we put $\rho = -1$. Huisman *et al.* (2001) assumed that $\rho = -1/\gamma$ which is the case for an extreme value distribution.

Remark that when deleting the second order term $Dk^{-\rho}$ in the regression model, one obtains a simple average of K Hill estimators. Due to the volatile behaviour of Hill estimators $H_{k,n}$ as a function of k it is known that a robust average of Hill estimators provides better estimators. This will be discussed in more detail in case of the generalized Hill sum plot where we apply weighted trimmed least squares regression.

3.2. Generalized Hill sum plot estimators

In a similar way, a new estimator can be introduced for real valued γ based on the generalized Hill sum plot. Indeed from (2.14)

$$H^*_{k,n} \;=\; \gamma + Dk^{-\rho} + \tilde{\epsilon}_k \;, \qquad k = 1, ..., K \;. \tag{3.3}$$

with $\tilde{\epsilon}_k = \sum_{j=1}^{k} e_j/k$. The variance of $\tilde{\epsilon_k}$ is asymptotically equal to the asymptotical variance $\mathrm{AVar}(H^*_{k,n})$ which, according to Beirlant *et al.* (2005) is equal to

$$\begin{aligned} \mathrm{AVar}(H^*_{k,n}) \;&=\; \frac{1+\gamma^2}{k}; \quad \gamma \geqslant 0 \\ &=\; \frac{(1-\gamma)\,(1+\gamma+2\,\gamma^2)}{(1-2\,\gamma)\,k}; \quad \gamma < 0 \;. \end{aligned}$$

Since the variance of the error term $\mathrm{Var}(\epsilon_k) = C_\gamma/k$ is not constant, a weighted least square regression is applied with the same $K \times K$ weight matrix W as in case of the Hill sum plot. Here again we ignore the fact that the error terms are not independent. We also put $\rho = -1$.

We also apply weighted trimmed least squares regression minimizing the sum of the $\lfloor n/2 \rfloor + 1$ smallest squared residuals. For more information we refer to Rousseeuw and Leroy (1987).

3.3. Simulation results

In Figures 2 till 5 we show the simulation results we obtained concerning weighted least squares regression estimators, trimmed and non trimmed, for some of the distributions considered in Section 2.4. For each distribution 100 repetitions of samples of size $n = 500$ were performed.

Weighted trimmed least squares yields less bias but somewhat higher mean squared error compared with the non robust regression algorithm. In case $\gamma > 0$ we also show the results for the weighted least squares estimators based on the Hill sum plot. Hill sum plots then yield better results than the generalized Hill sum plot. Also the trimmed regression algorithm is typically better than the non-robust version in case of the generalized Hill sum plot.

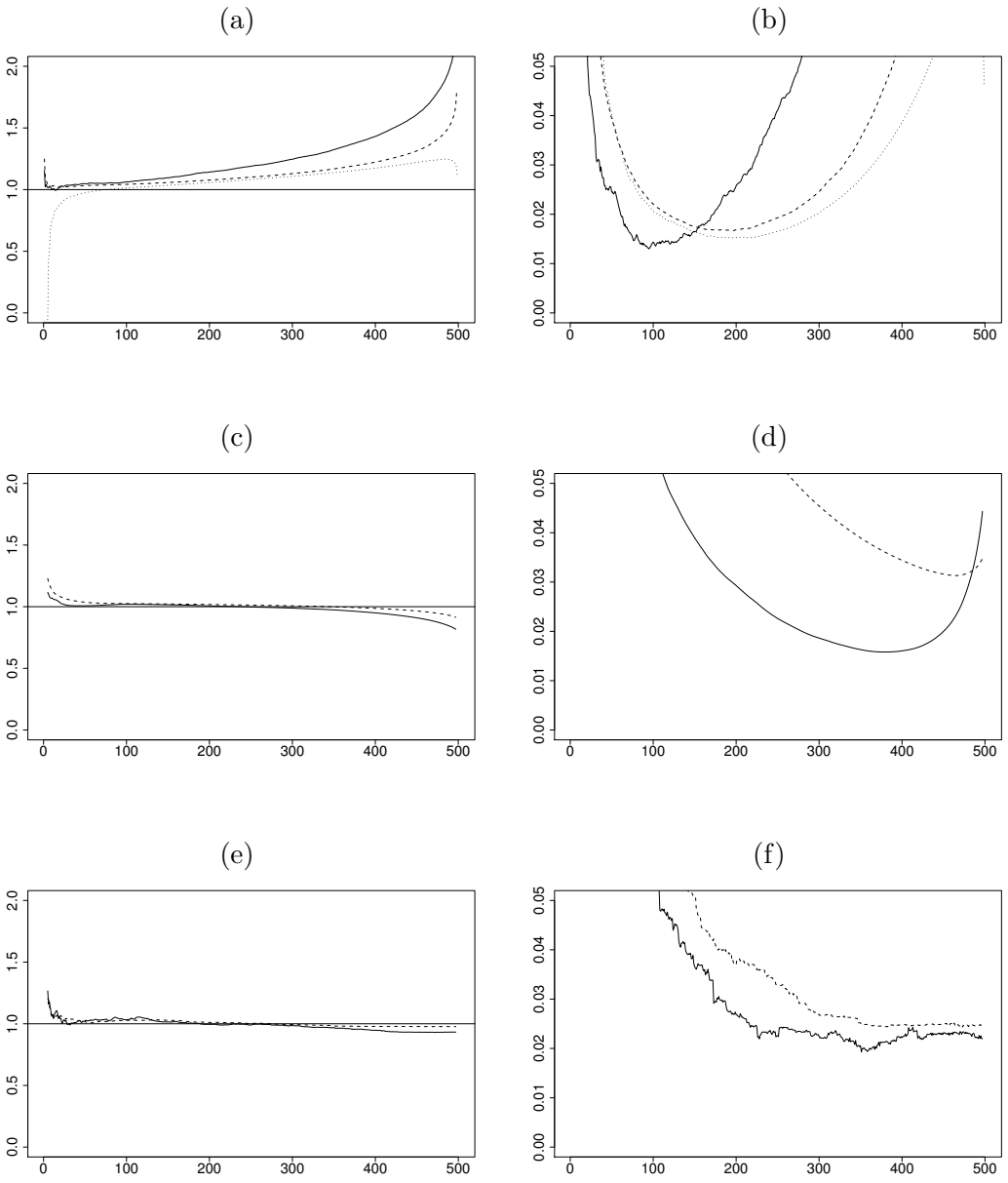

Figure 2: Fréchet distribution: (a) means of $H_{k,n}$ (full line), $H^*_{k,n}$ (dashed line) and $M_{k,n}$ (dotted line) as a function of k; (b) MSE of the estimators in (a); (c) means of weighted least squares estimators based on the regression model of the Hill sum plot (full line) and generalized Hill sum plot (dashed line); (d) MSE of the estimators in (c); (e) same as in (c), but now weighted trimmed least squares is used; (f) MSE of the estimators in (e).

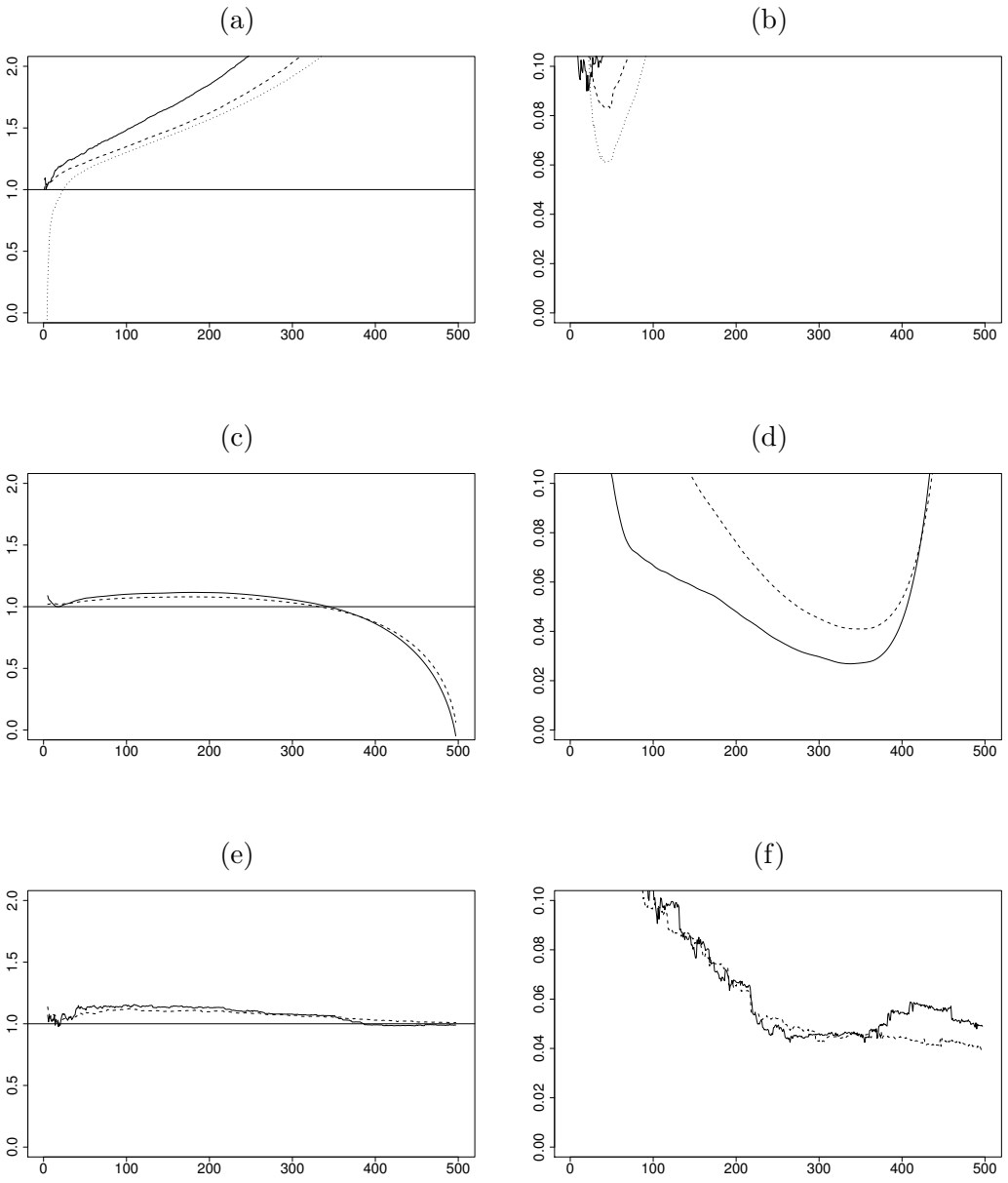

Figure 3: Burr distribution with $\rho = -0.5$: (a) means of $H_{k,n}$ (full line), $H^*_{k,n}$ (dashed line) and $M_{k,n}$ (dotted line) as a function of k. (b) MSE of the estimators in (a). (c) means of weighted least squares estimators based on the regression model of the Hill sum plot (full line) and generalized Hill sum plot (dashed line); (d) MSE of the estimators in (c); (e) same as in (c), but now weighted trimmed least squares is used; (f) MSE of the estimators in (e).

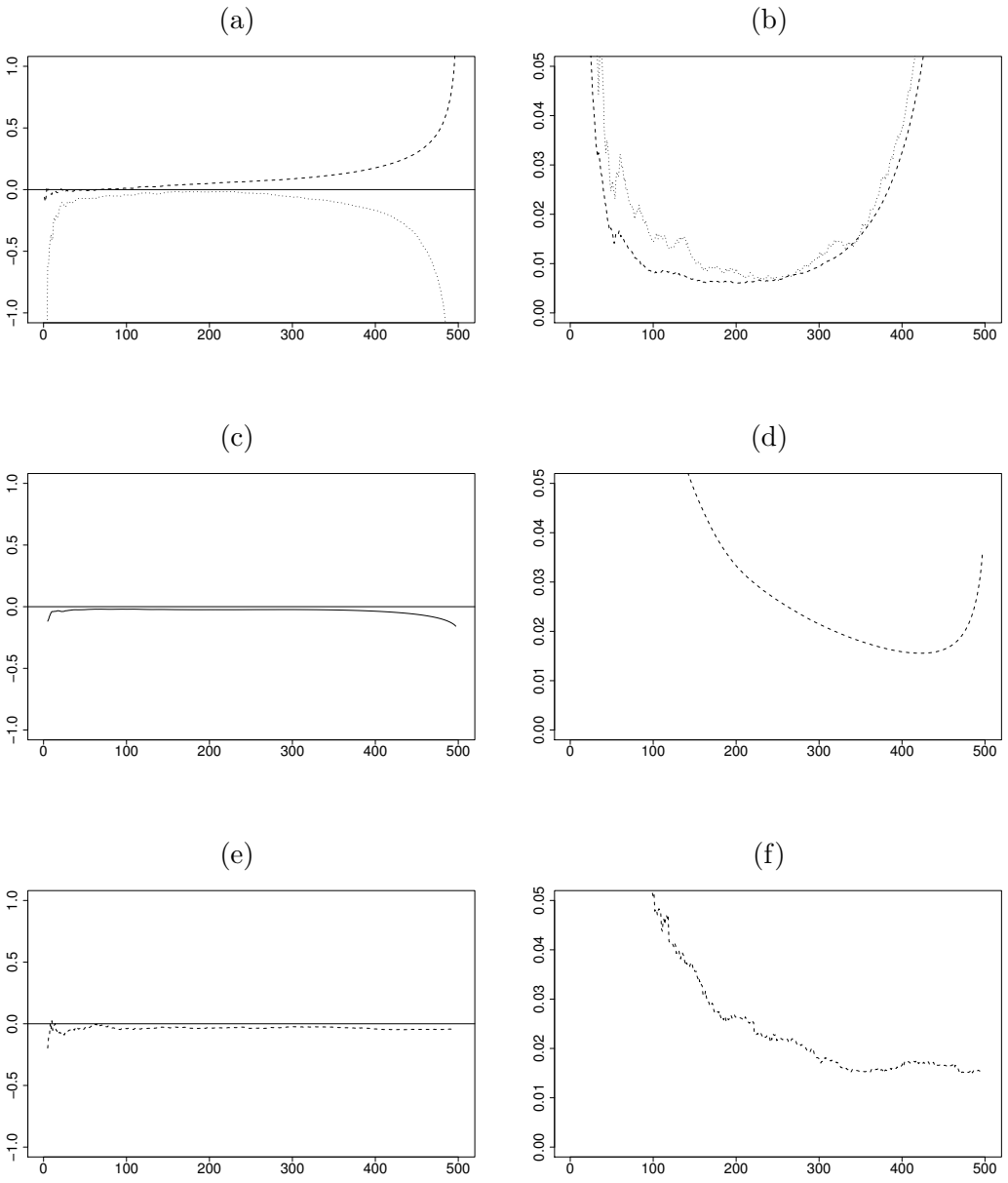

Figure 4: Gamma distribution: (a) means of $H^*_{k,n}$ (dashed line) and $M_{k,n}$ (dotted line) as a function of k; (b) MSE of the estimators in (a); (c) means of weighted least squares estimators based on generalized Hill sum plot (dashed line); (d) MSE of the estimators in (c); (e) same as in (c), but now weighted trimmed least squares is used; (f) MSE of the estimators in (e).

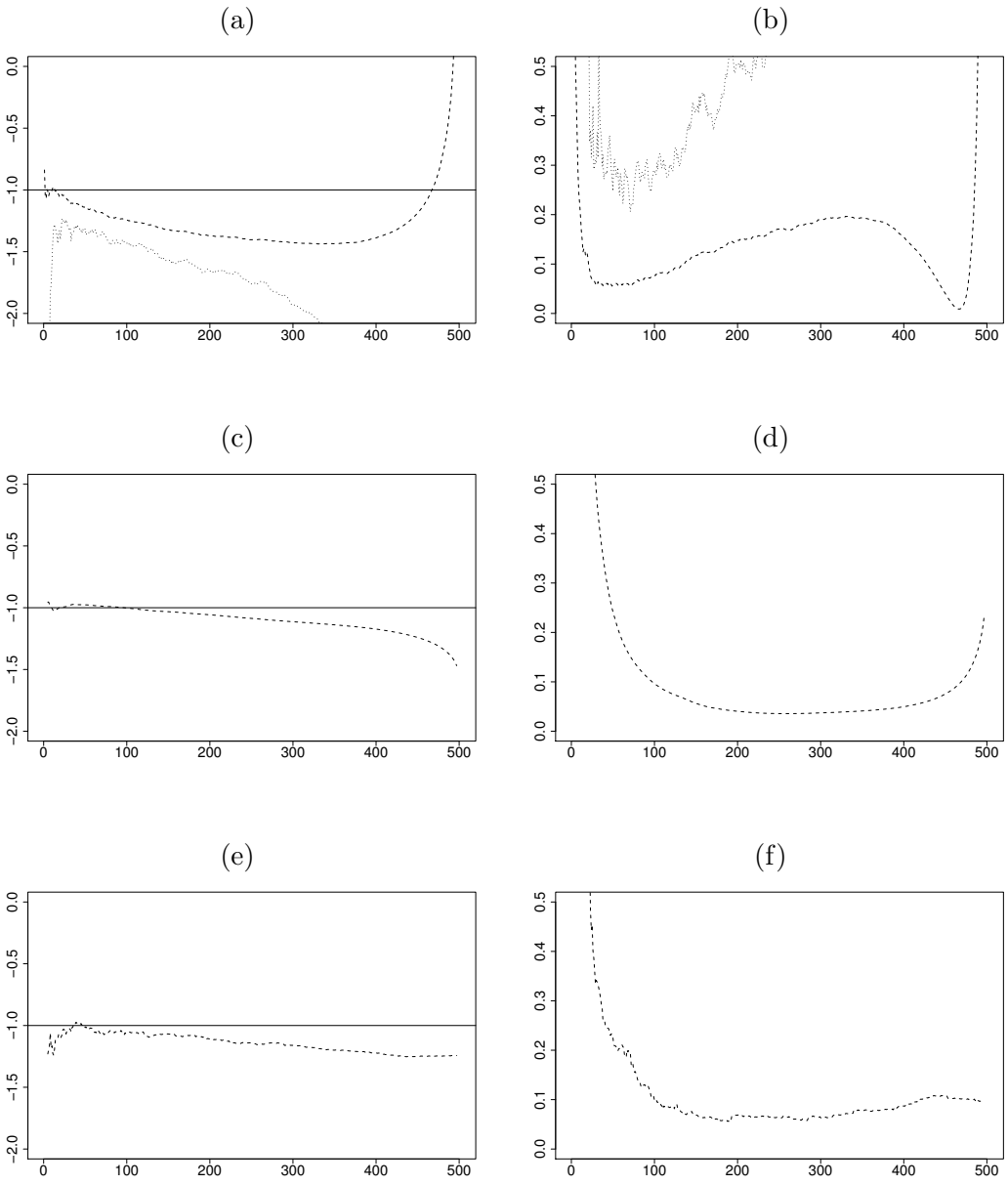

Figure 5: Reversed Burr: (a) means of $H^*_{k,n}$ (dashed line) and $M_{k,n}$ (dotted line) as a function of k; (b) MSE of the estimators in (a); (c) weighted least squares estimators based on the generalized Hill sum plot (dashed line); (d) MSE of the estimators in (c); (e) same as in (c), but now weighted trimmed least squares is used; (f) MSE of the estimators in (e).

3.4. Some practical examples

We end this paper showing the proposed methods into action. We apply the methods to two data sets proposed earlier in Beirlant *et al.* (2004). The data sets themselves can be found on `http://lstat.kuleuven.be/Wiley`.

The first data set contains daily maximal wind speeds at Brussels airport (Zaventem) from 1985 till 1992.

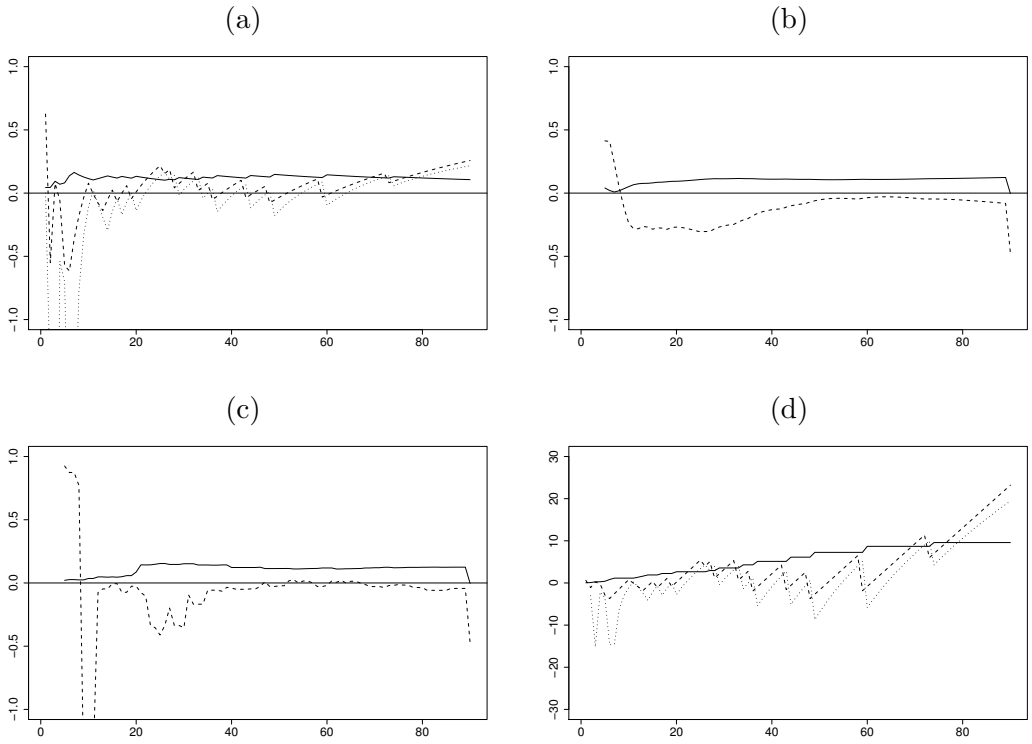

Figure 6: Zaventem daily maximum wind speed data: (a) $H_{k,n}$ (full line), $H^*_{k,n}$ (dashed line) and $M_{k,n}$ (dotted) as a function of k; (b) weighted least squares estimators based on the regression model of the Hill sum plot (full line) and generalized Hill sum plot (dashed line); (c) same as in (b), but now weighted trimmed least squares is used; (d) sum plots.

In Example 1.1 in Beirlant *et al.* (2004) the authors come to the conclusion that the data follow a simple exponential tail beyond 80 km/hr, and hence γ equals 0. The weighted (trimmed) least squares estimates based on the generalized Hill sum plot indeed indicates a zero valued extreme value index. Also the moment and generalized Hill sum plots indeed exhibit an overall horizontal behaviour for $K = 1, ..., 80$. The generalized Hill sum plot is less volatile however.

Finally we consider the AoN Re Belgium fire portfolio data introduced in section 1.3.3 in Beirlant *et al.* (2004). Here we omit the covariate information concerning sum insured and type of building. Here the estimate $\hat{\gamma} = 1$ follows from the weighted trimmed least squares regression analysis. These estimates are indeed quite stable over K-values compared to the non-robust version. The sum plots in Figure 7(d) are quite comparable for $K = 1, ..., 80$.

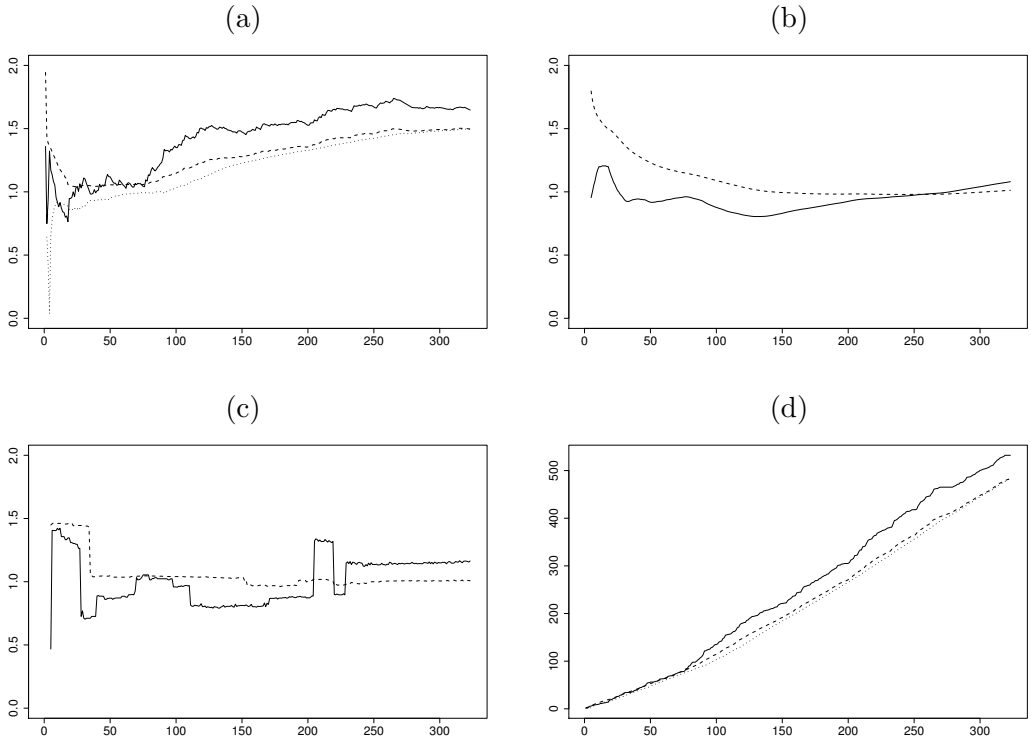

Figure 7: AoN claim size data: (a) $H_{k,n}$ (full line), $H^*_{k,n}$ (dashed line) and $M_{k,n}$ (dotted line) as a function of k; (b) weighted least squares estimators based on the regression model of the Hill sum plot (full line) and generalized Hill sum plot (dashed line); (c) same as in (b), but now weighted trimmed least squares is used; (d) sum plots.

ACKNOWLEDGMENTS

This research is sponsored by FWO grant G.0436.08N.

REFERENCES

[1] Beirlant, J.; Teugels, J.L. and Vynckier, P. (1996a). *Practical Analysis of Extreme Values*, Leuven University Press, Leuven.

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

[3] Beirlant, J.; Dierckx, G.; Guillou, A. and Starica, C. (2001). On exponential Representations of Log-Spacings of Extreme Order Statistics, *Extremes*, **5**, 157–180.

[4] Beirlant, J.; Goegebeur, Y.; Segers, J. and Teugels, J.L. (2004). *Statistics of Extremes*, Wiley.

[5] Beirlant, J.; Dierckx, G. and Guillou, A. (2005). Estimation of the extreme-value index and generalized quantile plots, *Bernoulli*, **5**(6), 949–970.

[6] Dekkers, A.L.M.; Einmahl, J.H.J. and de Haan, L. (1989). A moment estimator for the index of an extreme-value distribution, *Ann. Statist.*, **17**, 1833–1855.

[7] Dierckx, G. (2000). *Estimation of the Extreme Value Index*, Doctoral thesis, Katholieke Universiteit Leuven.

[8] 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**, 193–213.

[9] Hall, P. (1982). On some simple estimates of an exponent of regular variation, *Journal of the Royal Statistical Society B*, **44**, 37–42.

[10] Huisman, R.; Koedijk, K.; Kool, C. and Palm, F. (2001). Tail-index estimates in small samples, *Journal of Business and Economic Statistics*, **19**(2), 208–216.

[11] Henry III, J.B. (2009). A harmonic moment tail index estimator, *Journal of Statistical Theory and Applications*, **8**(2), 141–162.

[12] Hill, B. (1975). A simple general approach to inference about the tail of a distribution, *Ann. Statist.*, **3**, 1163–1174.

[13] Kratz, M. and Resnick, S. (1996). The qq-estimator of the index of regular variation, *Communications in Statistics: Stochastic Models*, **12**, 699–724.

[14] Neves, C. and Fraga Alves, M.I. (2008). Testing extreme value conditions: an overview and recent approaches, *REVSTAT – Statistical Journal*, **6**, 83–100.

[15] Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*, Wiley.

[16] Sousa, B. de and Michailidis, G. (2004). A diagnostic plot for estimating the tail index of a distribution, *J. Comput. Graph. Statist.*, **13**(4), 974–1001.

REVSTAT – Statistical Journal

Background

Statistical Institute of 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.

— Three volumes are scheduled for publication, one in April, one in June and the other in November.

— On average, four 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 / indexed in

REVSTAT is expected to be abstracted / indexed at least in *Current Index to Statistics*, *Statistical Theory and Method Abstracts* and *Zentralblatt für Mathematik*.

Instructions to Authors, special-issue editors and publishers

Papers may be submitted in two different ways:

— By sending a paper copy to the Executive Editor and one copy to one of the two Editors or Associate Editors whose opinion the author(s) would like to be taken into account, together with a postscript or a PDF file of the paper to the e-mail: revstat@fc.ul.pt.

— By sending a paper copy to the Executive Editor, together with a postscript or a PDF file of the paper to the e-mail: 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.

Submitted 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 not have more than 30 pages.

The first page should include the name, affiliation and address of the author(s) and a short abstract with the maximum of 100 words, followed by the key words up to the limit of 6, 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), *Mackintosh*, *Linux* and *Solaris Systems* (StuffIt format), and *Mackintosh System* (BinHex Format), are available in the REVSTAT link of the National Statistical Institute's 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: liliana.martins@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.

Also send always the final paper version to:

Maria José Carrilho
Executive Editor, REVSTAT – STATISTICAL JOURNAL
Instituto Nacional de Estatística, I.P.
Av. António José de Almeida
1000-043 LISBOA
PORTUGAL

Copyright