HURST EXPONENTS FOR NON-PRECISE DATA

M. ALVO AND F. THÉBERGE

Abstract. We provide a framework for the study of statistical quantities related to the Hurst phenomenon when the data are non-precise with bounded support.

1. Introduction

We consider non-precise data as defined by Viertl in [7]. We concentrate on a special type of non-precise data, where the characterizing function for each data point is symmetric around some central value x_i and has bounded support. Starting from a sample $x_1, ..., x_n$, which we assume to be normalized so that $\overline{x} = 0$, we consider each point as non-precise. At some λ-level, we get intervals $[x_i - \epsilon_\lambda, x_i + \epsilon_\lambda]$ called λ-cuts, where ϵ_λ is a non-increasing function of $0 \leq \lambda \leq 1$. With such a non-precise sample, we study the variance s^2, the partial sums $S_j = \sum_{i=1}^{j} x_i$, the range R of partial sums, and finally the ratio R/s and the related *Hurst exponent* H.

Analysis of the ratio R/s is motivated by the work of Hurst [4]. In this work, several natural phenomena were empirically shown to exhibit long-term dependence, which is quantified by the Hurst exponent. In this paper, we provide a framework by which we can take the non-precision of the data into account, so we can quantify the uncertainty associated with the statistics related to the Hurst phenomenon.

The rest of the paper is organized as follows. In Section 2, we review the non-precise formulation and introduce the notion of symmetric characterizing functions. In Section 3, we present a small example studied by Hurst [4]: the annual water discharges from Lake Albert. We use this example to describe the *Hurst exponent*, which is used to quantify the long-term behaviour of several natural phenomena. In Section 4, we study the functions described above for non-precise data. We bring our conclusion in Section 5. The Lake Albert data set as well as some computational considerations are presented in the Appendix.

2. Non-precise Data

We give a brief overview of the notation and concepts used in defining non-precise quantities. More details can be found in [7] and [1].

We define a non-precise quantity x via its *characterizing function*. This concept is similar to the notion of membership function used in fuzzy data analysis.

Received: November 2011; Accepted: November 2012
Key words and phrases: Hurst phenomenon, Non-precise data.

Definition 2.1. A characterizing function $\xi(.)$ of a non-precise number is a real function of a real variable such that:
(i) $\xi : \mathbb{R} \to [0,1]$
(ii) $\exists\ x_0 \in \mathbb{R} : \xi(x_0) = 1$
(iii) $\forall\ \lambda \in (0,1]$, the set $B_\lambda = \{x \in \mathbb{R} : \xi(x) \geq \lambda\} = [a_\lambda, b_\lambda]$ is a finite closed interval, a λ-cut of ξ.

Viertl [7] has shown that a characterizing function can be uniquely determined by the family of λ-cuts $\{B_\lambda : \lambda \in (0,1]\}$ and moreover

$$\xi(x) = \max_{\lambda \in (0,1]} \lambda I_{B_\lambda}(x),\ \forall x \in \mathbb{R}. \tag{1}$$

Characterizing functions can also be defined for a non-precise n-dimensional vector x^*.

Definition 2.2. A characterizing function $\xi(.)$ of a non-precise vector x^* is a real function of n variables such that:
(i) $\xi(.) : \mathbb{R}^n \to [0,1]$
(ii) $\exists\ \mathbf{x}_0 \in \mathbb{R}^n : \xi(\mathbf{x}_0) = 1$
(iii) $\forall\ \lambda \in (0,1]$, the set $B_\lambda(x^*) = \{x \in \mathbb{R}^n : \xi(x) \geq \lambda\}$ is a star-domain, by which we mean that the line segment joining any two points in the set lies entirely in the set.

An example of a non-precise vector is the location of an object on a radar screen: the object appears as a cloud in two-dimensional space and the characterizing function may be constructed in terms of the light intensity function. Given n non-precise observations, $x_1^*, x_2^*, ..., x_n^*$, each taking values in a space M with corresponding characterizing functions $\xi_1, ..., \xi_n$, it is possible to define a characterizing function $\xi : M^n \to [0,1]$ for the combined sample via the product rule or the minimum rule respectively as:

$$\xi(x_1, x_2, ..., x_n) = \Pi_{i=1}^n \xi_i(x_i) \tag{2}$$

$$\xi(x_1, x_2, ..., x_n) = \min_i \xi_i(x_i). \tag{3}$$

We use the minimum rule (3). Any statistical function $f(x_1, x_2, ..., x_n)$ which is the basis of inference for precise data $x = (x_1, x_2, ..., x_n)$ is then adapted for non-precise data by computing its characterizing function in accordance with the rule:

$$\psi(y) = \left\{ \begin{array}{ll} \sup\{\xi(x) : x \in \mathbb{R}^n, f(x_1, x_2, ..., x_n) = y\} & \text{for } f^{-1}(\{y\}) \neq \varnothing \\ 0 & \text{for } f^{-1}(\{y\}) = \varnothing \end{array} \right\} \forall y \in \mathbb{R} \tag{4}$$

We consider symmetric characterizing functions, as in [6], for the non-precise version of our data so that every λ-cut is of the form $B_\lambda(x_i) = [x_i - \epsilon_\lambda, x_i + \epsilon_\lambda]$. Some important families of symmetric characterizing functions with bounded support are listed next.

(1) The *truncated Gaussian* characterizing functions are of the form $\xi(x) = e^{-(x-\mu)^2/(2\sigma^2)}$ for $|x - \mu| \leq T$ for some threshold T and $\xi(x) = 0$ otherwise.

(2) The *uniform* characterizing functions: $\xi(x) = 1$ if and only if $|x - \mu| \leq \sigma$ and 0 otherwise.
(3) The *triangular* characterizing functions: $\xi(x) = 1 - \frac{|x-\mu|}{\sigma}$ for $|x - \mu| \leq \sigma$ and 0 otherwise.
(4) The *trapezoidal* characterizing functions: $\xi(x) = 1$ when $|x-\mu| \leq \sigma$, $\xi(x) = 1 - \frac{|x-\mu|-\sigma}{\beta}$ for some $\beta > 0$ when $|x - \mu| > \sigma$ and $|x - \mu| \leq \sigma + \beta$, and 0 otherwise.

The functions are illustrated in Figure 1, centered at $\mu = 0$. We set $\sigma = 1$, $T = 2$ for the truncated Gaussian, and $\beta = 1$ for the trapezoidal. In each case, a λ-cut yields an interval of values symmetric around 0, as illustrated with the horizontal dashed lines at $\lambda = 0.5$.

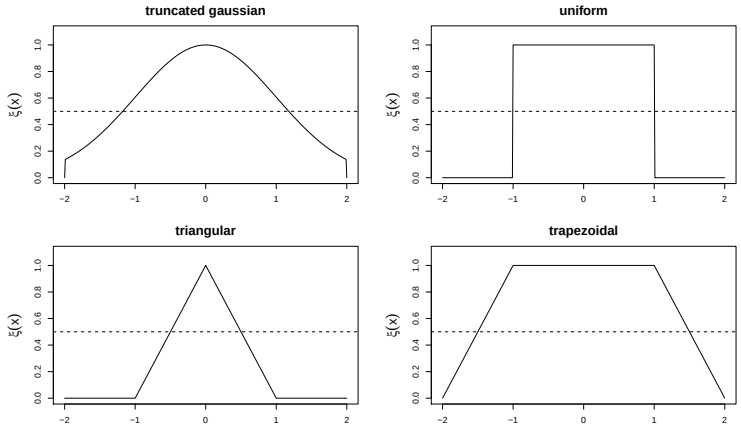

FIGURE 1. Some Symmetric Characterizing Functions

3. **Lake Albert Example**

We consider the data from Table 1 on page 7 of [4], which lists the annual total discharge from Lake Albert from 1904 to 1957 in billion of cubic meters. We let $q_1, q_2, ..., q_n$ represent the annual discharge, with $n = 54$, and $x_i = q_i - \overline{q}$. The data are listed in Appendix A. We define $S_k = \sum_{i=i}^{k} x_i$, the partial sums, R the range of the $\{S_1, ..., S_n\}$ and s^2 the sample variance. If the q_i are independent and identically distributed, Feller [3] has shown that for any sequence of i.i.d. random variables with finite variance, $E(R/s) = (n\pi/2)^{1/2}$ for large n.

On the other hand, for many seemingly unrelated natural phenomena, Hurst observed that $R/s = (n/2)^H$ with the Hurst exponent $H > 1/2$. In fact, it is reported in [4] that based on 690 experiments, the average value for H was 0.73. In order to allow for easy comparison with the random hypothesis, we use the *generalized* Hurst coefficient, as detailed in [5], namely

$$R/s = (n\pi/2)^H. \tag{5}$$

With this definition, for a process which does not exhibit dependency, we expect $H = 0.5$ for any value of n. The average value for H as defined in (5) for the 690 experiments reported by Hurst in [4] is 0.57. We will re-visit this example in the context of non-precise data.

4. Non-precise Data Analysis

Consider a sample $\hat{x}_1, \hat{x}_2, ..., \hat{x}_n$ of precise quantities, which we assume to be the central values for n identically shaped, symmetric characterizing functions $\xi_i(\cdot)$ with bounded support. Let $\text{Supp}(\hat{x}_i)$ be the support of the characterizing function centered at $\hat{x}_i$. Then for all $1 \leq i \leq n$, we have $\xi_i(\hat{x}_i) = 1$ and $\xi_i(t) = 0,\ t \notin \text{Supp}(\hat{x}_i)$. Without loss of generality, we assume that the data is centered around 0, namely $\sum_{i=1}^n \hat{x}_i = 0$.

Given $0 < \lambda \leq 1$, there exists a $\epsilon_\lambda \geq 0$ such that $x_i \in [\hat{x}_i - \epsilon_\lambda, \hat{x}_i + \epsilon_\lambda]$ for $1 \leq i \leq n$, the λ-cuts. We define the set of *feasible configurations* at λ as:

$$F_\lambda = \{x = (x_1, x_2, ..., x_n) :\ x_i = \hat{x}_i + \alpha_i,\ \alpha_i \in [-\epsilon_\lambda, \epsilon_\lambda]\ \forall i,\ \sum_{i=1}^n \alpha_i = 0\}. \quad (6)$$

The last constraint ensures that the sample mean of the x_i remains at 0. Given λ and some $x \in F_\lambda$, we define the following quantities related to the Hurst phenomenon:

- the *sample variance* $s^2 = \sum_{i=1}^n x_i^2/(n-1)$,
- the *partial sums* $S_j = \sum_{i=1}^j x_i,\ 1 \leq j \leq n$,
- the *range* of partial sums:

$$R = \max_{1 \leq j \leq n} S_j - \min_{1 \leq k \leq n} S_k = \max_{1 \leq i < j \leq n} |x_i + ... + x_j|,$$

- the *ratio* R/s, and the *Hurst exponent* $H = \log(R/s)/\log(n\pi/2)$.

In order to compute the characterizing functions for the quantities related to the Hurst phenomenon, we use the construction given in equation (4), along with the minimum rule given in (3).

(1) For the sample variance s^2, we define

$$f(x_1, ..., x_n) = \begin{cases} \sum_{i=1}^n x_i^2/(n-1); & \sum_{i=1}^n x_i = 0 \\ 0;\ \text{otherwise.} \end{cases}$$

(2) For the range of partial sums R, we define

$$f(x_1, ..., x_n) = \begin{cases} \max_{1 \leq i < j \leq n} |x_i + ... + x_j|; & \sum_{i=1}^n x_i = 0 \\ 0;\ \text{otherwise.} \end{cases}$$

(3) For the ratio R/s, we define

$$f(x_1, ..., x_n) = \begin{cases} \left(\max_{1 \leq i < j \leq n} |x_i + ... + x_j|\right) / \left(\sum_{i=1}^n x_i^2/(n-1)\right); & \sum_{i=1}^n x_i = 0 \\ 0;\ \text{otherwise.} \end{cases}$$

Finally, we let $H = \log(R/s)/\log(n\pi/2)$ to build the characterizing function for the Hurst exponent. The computation of the characterizing functions above as given in expression (4) uses the fact that each point has bounded support, thus $\xi(x) = 0$ when at least one $x_i \notin \text{Supp}(\hat{x}_i)$. The characterizing functions are built

by considering various levels of λ in turn, and the corresponding points $x \in F_\lambda$. For the computation of the sample variance s^2 and the range of partial sums R, some results given in Appendix B can be used to speed up the process.

Water Discharge Examples. In Figure 2, we plot the characterizing functions for the Hurst exponent H, respectively with *truncated Gaussian* and *triangular* characterizing functions (with $\mu = 0$, $\sigma = 1$), for the Lake Albert dataset described earlier. The dashed horizontal line indicates the range of possible values when $\lambda = 0.5$. For triangular characterizing function, $\lambda = 0.5$ corresponds to $\epsilon_\lambda = 0.5$, the range of truncation error for integer-valued data. These plots further support Hurst's observations since the value 0.5 is further on the left and is thus less plausible.

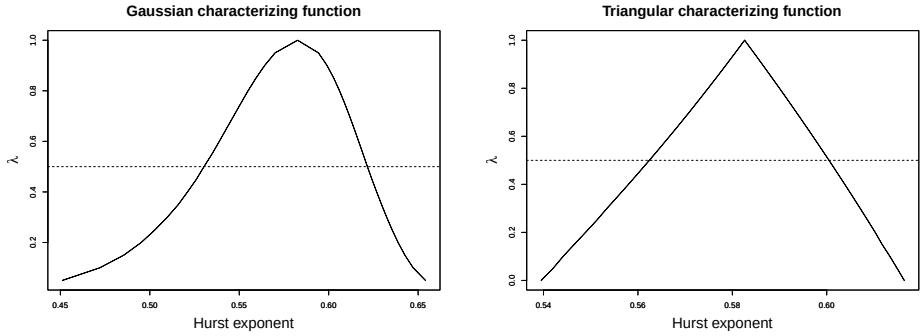

FIGURE 2. Characterizing Functions for H for the Lake Albert dataset, Respectively with Gaussian and Triangular Characterizing Functions for the Data ($\mu = 0, \sigma = 1$)

We ran a similar analysis on another set of hydrometric data from [2], which lists the average monthly and annual water discharge of the Ottawa river from 1961 to 2007. We analyze the average flows in m^3/s rather than the total flow to avoid the issue of months having unequal number of days. For the monthly data, equation (5) yields $H \approx .554$ while with the annual data, we get $H \approx .509$. With such results, it is not a-priori clear if the Hurst phenomenon is present in this data. Considering non-precision is thus useful to quantify our conclusion.

In Table 1, we list the range of values taken by H under various ϵ_λ ranging from 0 (precise data) to ± 5 m^3/s. A given ϵ_λ corresponds, for example, to triangular characterizing functions as defined in section 2, with $\sigma = 2\epsilon_\lambda$ and $\lambda = 0.5$. We see that the Hurst phenomenon is supported for the monthly data. However with wider intervals on non-precision, the Hurst phenomenon is unclear for the annual data. In Figure 3, we plot the characterizing functions for H respectively for the annual and monthly data. We used triangular characterizing functions for the measurements with $\mu = 0$ and $\sigma = 5$.

For the monthly data, we then considered a deseasonalized version where we subtracted the average for each month throughout the years. The adjusted rates

ϵ_λ	monthly data	annual data
0	[.554,.554]	[.509,.509]
.5	[.553,.555]	[.508,.511]
1	[.553,.555]	[.507,.512]
2.5	[.550,.557]	[.503,.516]
5	[.547,.561]	[.496,.522]

TABLE 1. Range for H for the Ottawa River Data at Various Levels of ϵ_λ

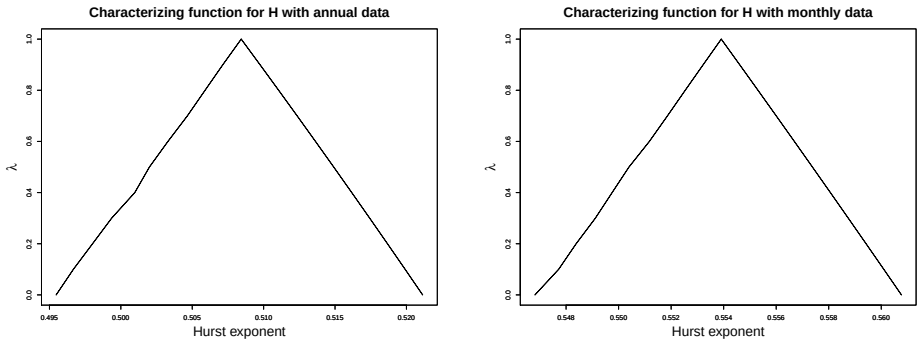

FIGURE 3. Characterizing Functions for H for the Ottawa River Datasets, Respectively Annual and Monthly, with Triangular Characterizing Functions ($\mu = 0, \sigma = 5$)

as well as the characterizing function for H are shown in Figure 4. The evidence for the Hurst phenomenon is even stronger in this case.

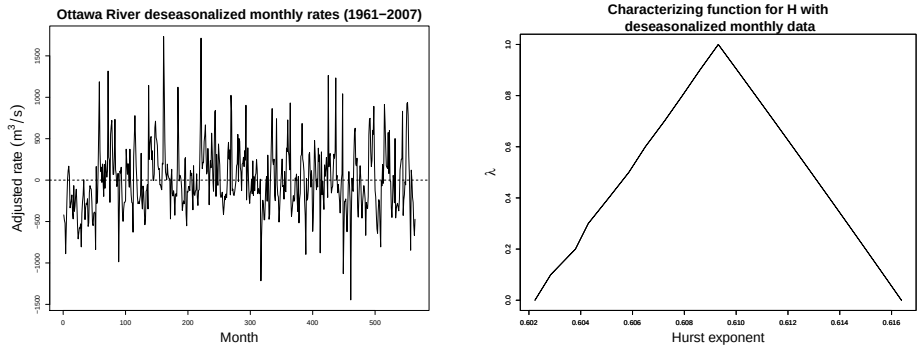

FIGURE 4. Adjusted Monthly Rates and Characterizing Functions for H for the Ottawa River Datasets, with Triangular Characterizing Functions ($\mu = 0, \sigma = 5$)

Asymptotic Considerations. The Hurst exponent depends on the ratio R/s, and both values carry some non-precision, which makes the evaluation of the characterizing function for H difficult. For large values of n, can we fix s and concentrate on the non-precision due to R? We ran a series of experiments based on the Lake Albert dataset with the q_i as i.i.d. Gaussian random variables, with $\mu = 23.72$ and $s = 6.88$. The characterizing function for the Hurst exponent is given by

$$\Psi(H) = \sup\left\{\xi(x_1, ..., x_n)\ ;\ H = \frac{\log(R)}{\log(n\pi/2)} - \frac{\log(s)}{\log(n\pi/2)}\right\}.$$

In this example, the second term in the expression for H is about .15 with $n = 10^6$ and .07 with $n = 10^{12}$, which is non-trivial considering that $H \approx .5$, so it can hardly be ignored in practice.

5. **Conclusion**

Hurst observed that for several environmental series of data there was a persistent long term effect which he quantified through the Hurst coefficient. For such series the value coefficient was greater than 0.5. When data exhibit a persistent long term effect, they can in general be suitably modeled. In this article we re-considered the calculation of the Hurst coefficient when the data are treated as non-precise. Our analysis provides strong evidence to support Hurst's findings for the Lake Albert annual discharge data set. We then applied our methodology on the Ottawa river data for the period 1961-2007. We conclude that there is evidence to support the existence of the Hurst phenomenon when the measurements are made monthly, in particular with deseasonalized data, but not when they are made annually. By treating the data as non-precise, we develop confidence in our conclusions.

Appendix A. **Lake Albert Dataset**

year	discharge (q_i)	year	discharge (q_i)	year	discharge (q_i)
1904	35	1922	13	1940	20
1905	31	1923	14	1941	19
1906	34	1924	18	1942	29
1907	33	1925	16	1943	26
1908	26	1926	19	1944	18
1909	29	1927	25	1945	15
1910	26	1928	21	1946	16
1911	22	1929	19	1947	25
1912	19	1930	21	1948	28
1913	20	1931	26	1949	24
1914	21	1932	28	1950	18
1915	24	1933	29	1951	17
1916	27	1934	23	1952	25
1917	47	1935	20	1953	21
1918	48	1936	20	1954	19
1919	29	1937	24	1955	20
1920	23	1938	26	1956	21
1921	17	1939	24	1957	23

The annual discharges q_i are in billions of cubic meters. We define the partial sums $S_k = \sum_{i=1}^{k} q_i$. We get $\max S_k = 91.44$ at $k = 16$ (year 1919), $\min S_k = 0$, so the range is $R = 91.44$. We compute the sample standard deviation $s = 6.88$, so $R/s = 13.29$ and $H \approx 0.57$, the Hurst exponent as defined in (3).

Appendix B. **Computational Considerations**

Computing the characterizing functions for the statistics described in Section 4 is done by fixing λ and empirically estimating the range of values taken by this statistic over all $x \in F_\lambda$. For the variance s^2 and the partial sum range R, some results can be used to speed up this process.

Consider all $x \in F_\lambda$ with ordering of the indices such that $x_1 \leq x_2 \leq ... \leq x_n$. We first show that the maximum value taken by the sample variance over all $x \in F_\lambda$ is obtained with $\alpha_i = -\epsilon_\lambda,\ i \leq n/2$ when n is even, with all other $\alpha_i = \epsilon_\lambda$. With n odd, we do the same and leave the middle point as it is.

Lemma B.1. *If we have a set of values $x = (x_1, ..., x_n) \in F_\lambda$ with $x_1 \leq .. \leq x_n$ such that $\exists i < j,\ \alpha_i > -\epsilon_\lambda,\ \alpha_j < \epsilon_\lambda$, then there exists another $x \in F_\lambda$ with larger variance.*

Proof. Start with $S_1 = \sum_{i=1}^{n} x_i^2$, and let $\overline{x} = 0$ w.l.g. Let $\alpha_i \leftarrow \alpha_i - \delta,\ \alpha_j \leftarrow \alpha_j + \delta$ such that $\alpha_i \geq -\epsilon_\lambda,\ \alpha_j \leq \epsilon_\lambda,\ \delta > 0$, then we compute:
$S_2 = \sum_{k \neq i,j} x_k^2 + (x_i - \delta)^2 + (x_j + \delta)^2 = S_1 + 2\delta(\delta + x_j - x_i) > S_1$. □

To find the configuration with smallest variance, we use the following result.

Lemma B.2. *If we have a set of values $x = (x_1, ..., x_n) \in F_\lambda$ with $x_1 \leq .. \leq x_n$ such that $\exists x_i < x_j,\ \alpha_i < \epsilon_\lambda$ and $\alpha_j > -\epsilon_\lambda$, then there exists another $x \in F_\lambda$ with smaller variance.*

Proof. Start with $S_1 = \sum_{i=1}^{n} x_i^2$, and let $\overline{x} = 0$ w.l.g. Let $\alpha_i \leftarrow \alpha_i + \delta,\ \alpha_j \leftarrow \alpha_j - \delta$ such that $\alpha_i \leq \epsilon_\lambda,\ \alpha_j \geq -\epsilon_\lambda,\ (x_j - x_i)/2 > \delta > 0$, then we compute:
$S_2 = \sum_{k \neq i,j} x_k^2 + (x_i + \delta)^2 + (x_j - \delta)^2 = S_1 + 2\delta(\delta + x_i - x_j) < S_1$. □

We turn our attention to the partial sums $S_j = \sum_{i=1}^{j} x_i$ for all $x \in F_\lambda$. We assume that $S_n = 0$, thus $\overline{x} = 0$. For any $x \in F_\lambda$, let $m = \text{argmin}_{1 \leq i \leq n} S_i$ and $M = \text{argmax}_{1 \leq i \leq n} S_i$. Thus the overall range of the partial sums corresponding to all $x \in F_\lambda$ is given by $R = S_M - S_m = \max_{1 \leq i < j \leq n} |x_i + x_{i+1} + ... + x_j|$. Assume first that $M < m$, so the maximum partial sum is reached before the minimum. We know that $S_m \leq 0$ and $S_M \geq 0$ with $R = S_M - S_m$. We note that, with m and M known and fixed:

- changing a value in $x_1, ..., x_M$ affects *both* S_m and S_M in the same way, leaving R unchanged;
- changing a value in $x_{M+1}, ..., x_m$ affects only S_m, thus also R;
- changing a value in $x_{m+1}, ..., x_n$ affects neither S_m nor S_M.

We do not know the values of m and M for the optimal configuration, but we can test the assumptions $M = i, m = j, s = (j - i) \geq 1$ in turn. In order to maximize

R, we need:

$$\alpha_{i+1} + ... + \alpha_j = \begin{cases} -s\epsilon; & s \leq n/2 \\ -(n-s)\epsilon; & s \geq n/2 \end{cases}$$

therefore:

$$\alpha_1 + ... + \alpha_i + \alpha_{j+1} + ... + \alpha_n = \begin{cases} +s\epsilon; & s \leq n/2 \\ +(n-s)\epsilon; & s \geq n/2. \end{cases}$$

This leads to the following algorithm to compute $\max(R)$ over all $x \in F_\lambda$.

(1) Pre-compute $\Delta = (\epsilon_\lambda, 2\epsilon_\lambda, ..., (n/2)\epsilon_\lambda, (n/2-1)\epsilon_\lambda, ..., \epsilon_\lambda)$ of length $n-1$; let Δ_i be the i^{th} element in vector Δ.

(2) For $M = i$, the maximum range is given by $\max_{i+1 \leq j \leq n}(x_i - (x_j - \Delta_{j-i}))$; try all $M = i$ and take maximum value

$$\max_{1 \leq i < n} \left(\max_{i+1 \leq j \leq n} (x_i - (x_j - \Delta_{j-i})) \right).$$

(3) Verify all cases where $m = i, M = j, i < j$ in a similar way and take the overall maximum.

References

[1] M. Alvo and F. Théberge, *The problem of classification when the data are non-precise*, Austrian Journal of Statistics, **34** (2005), 375-390.

[2] Environment Canada, *Archived data for station 02KF005, Ottawa river at Britannia*, `www.ec.gc.ca/rhc-wsc`.

[3] W. Feller, *The asymptotic distribution of the range of sums of independent random variables*, The Annals of Mathematical Statistics, **22** (1951), 427-432.

[4] H. E. Hurst, R. P. Black and Y. M. Simaika, *Long-term storage: an experimental study*, Constable (London), 1965.

[5] A. I. McLeod and K. W. Hipel, *Preservation of the rescaled adjusted range: A reassessment of the Hurst phenomenon*, Water Resources Research, **14(3)** (1978), 491-508.

[6] J. Valente de Oliveira and W. Pedrycz, *Advances in Fuzzy Clustering and its Applications*, Chapter 8, Wiley, 2007.

[7] R. Viertl, *On statistical inference for non-precise data*, Environmetrics, **8** (1997), 541-568.

Mayer Alvo*, Department of Mathematics & Statistics, University of Ottawa, 585 King Edward, Ottawa, ON (K1N 5N1), Canada

E-mail address: `malvo@uottawa.ca`

Francois Théberge, Department of Mathematics & Statistics, University of Ottawa, 585 King Edward, Ottawa, ON (K1N 5N1), Canada

E-mail address: `ftheberg@uottawa.ca`

*Corresponding author