Malaysian Journal of Mathematical Sciences 10(1): 1–21 (2016)

MALAYSIAN JOURNAL OF MATHEMATICAL SCIENCES

Journal homepage: http://einspem.upm.edu.my/journal

Modelling Record Times in Sport with Extreme Value Methods

Adam, M. B. *[1] and Tawn, J. A. [2]

[1] *Department of Mathematics, Universiti Putra Malaysia*
[2] *Department of Mathematics and Statistics, Lancaster University*

E-mail: bakri@upm.edu.my
** Corresponding author*

ABSTRACT

We exploit connections between extreme value theory and record processes to develop inference methods for record processes. Record processes have trends in them even when the underlying process is stationary. We study the problem of estimating the underlying trend in times achieved in sports from record data. We develop new methods for inference, simulating record series in non-stationary contexts, and assessing fit which account for the censored characteristic of record data and we apply these methods to athletics and swimming data.

Keywords: Athletics, extreme value theory, Poisson process, records, swimming.

1. Introduction

In many fields records are of particular interest such as in weather (highest flood, hottest temperature, strongest wind speed) and in finance (largest insurance claim, biggest stock market crash). In sporting events, such as athletics and swimming, records are the vital events in characterising the development of the sport, they are the target for competitors, and they receive most publicity. Records are however simply the most extreme values in a series at the time of their occurrence and so there are natural connections between the theory for records and the theory for extreme values, this has been exploited in probabilistic results for records by Ahsanullah (2004), Arnold et al. (1998), Ahsanullah and Bhoj (1996), Bairamov (1996), David and Nagaraja (2003), Glick (1978), Sibuya and Nishimura (1997), Smith (1988) and Benested (2004). In this paper we apply and adapt our knowledge of extreme value methods to the analysis of record data for sporting events, focusing on athletics track events and swimming competitions.

To illustrate the issues in this paper first consider the minimum record process $\{Y_t\}$ which arises from a stationary sequence $\{X_t\}$, i.e. $Y_t = \min(X_1, \ldots, X_t)$. We assume that $\{X_t\}$ are continuous random variables. Let $R_t = I(X_t < Y_{t-1})$ be an indicator variable for the occurrence of a record at time t and let N_n be the number of records until time n, with $N_1 = 1$ by definition. Then $R_t \sim \text{Bernoulli}(t^{-1})$ and $E(N_n) = \sum_{t=1}^{n} E(R_t) = \sum_{t=1}^{n} \frac{1}{t}$, so for large n, $E(N_n) \approx \log n + \gamma$, with γ is Euler's constant 0.5772... . Arnold et al. (1998) point out that about 7 records are expected to occur in a sequence of 100,000 observations of a stationary sequence, so when studying records of stationary processes we need to recognise that records are broken rarely. Furthermore, there is a decreasing trend in the $\{Y_t\}$ process as $Y_{t+1} \leq Y_t$ for all $t \geq 1$.

If an analysis is to be undertaken from a series of record data $\{Y_t\}$ alone then at first sight it may appear that the information in the series is limited as many of the values in the series will be identical. The few values in the series where the record changes, i.e. $Y_t < Y_{t-1}$ tell us directly that $X_t = Y_t$. However, the record remaining unchanged does also provide information about the underlying $\{X_t\}$ process as $Y_t = Y_{t-1}$ tells us that $X_t > Y_{t-1}$, so this gives censored information about the underlying $\{X_t\}$ process. Despite the information from censored data, for stationary series the information about the record process is gathered very slowly because when the probability of not breaking the record is high there is very limited information in the censored data.

As sports have increasingly become professional, their training techniques

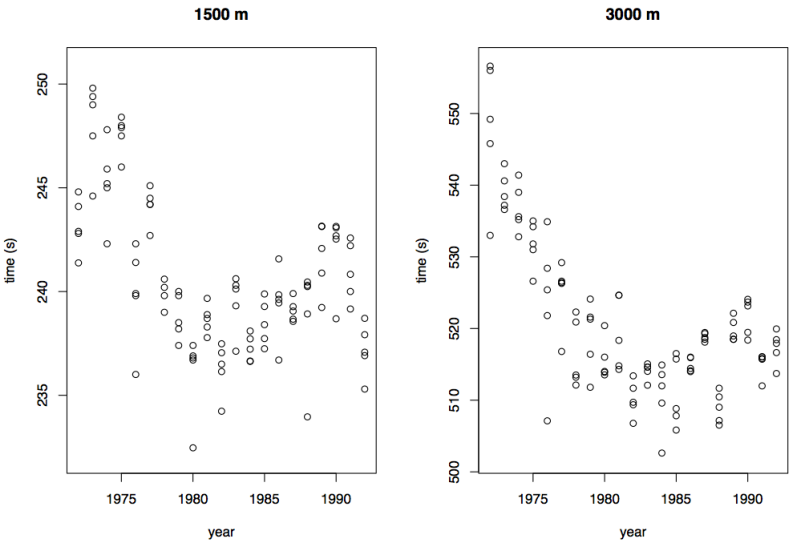

Figure 1: Plot A - women's athletics 3000 m with annual minima (circle) and annual record data (dot) from 1972 to 1992; Plot B - women's athletics 1500 m with annual minima (circle) and annual record data (dot) from 1972 to 1992.

have been refined, dietary knowledge has been improved, more people are training to high standard in every sport, and it is clear that the underlying marginal distribution of elite performance in sport has changed. This is seen in the athletics and swimming data we study. Figure 1 shows the annual best times achieved in recognised international competitions over the period of 1972-1992 for the women's athletics 1500 m and 3000 m track events, see Robinson and Tawn (1995). The record process, derived from these annual data, is also on the figure. It shows that the number of records being broken is greater than would be expected if the annual best performance was stationary over time.

Similarly, Figure 2 shows the FINA world record series of the men's 400 m freestyle swimming event plotted against the actual times in the year when the records are achieved. Unlike the athletics data, this type of record data shows when the record has been broken more than once in a year. Again the number of records is greater than would be expected if the series was stationary, indicating a trend in elite performances. We call the two types of data annual record data and actual record data respectively.

As indicated earlier, the record process exhibits a trend even when the

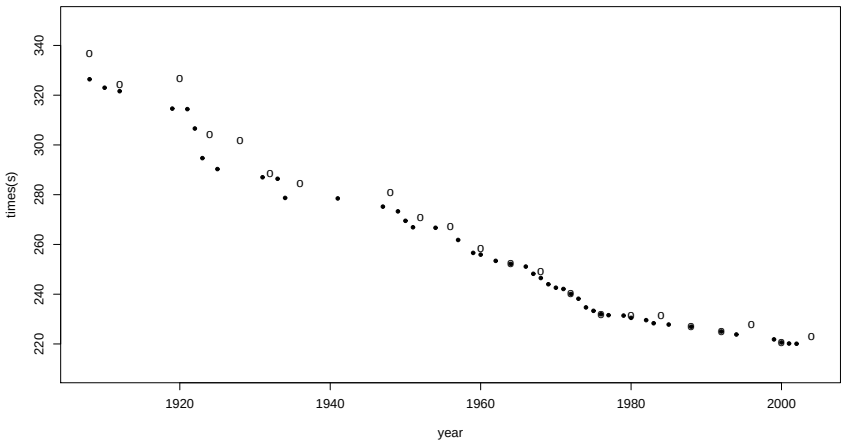

Figure 2: The men's Olympic 400 m freestyle gold medallist's times (empty dot) and the men's actual world record data for 400 m freestyle (dark dot) from 1912 to 2004.

underlying process is stationary. We can tell from the number of records in the actual record data that there must be non-stationarity in the underlying process but it is clear that separating the component of the trend due to records from that due to non-stationarity is not obvious from information on the record process alone. For the annual records data in athletics we also have annual best data which allows us to identify the separation as the records decrease over time faster than the mean of the annual best data.

Most often data on records takes the form of actual record data as annual record data exclude records if the record is broken more than once in a year. Furthermore, annual best data (from which annual record data are typically derived) are difficult to find for early periods in the history of an event whereas all actual records are well documented. As annual best data, and their associated annual record data, provide information on both the underlying process and the record process we first study these data, this helps in identifying the level of loss of information in analysing annual records in comparison to annual best data and identifies the need for some independent information on the trend in the underlying series.

As the data are minima times the appropriate distribution for the annual best is the generalised extreme value distribution for minima, GEVM(μ, σ, ξ), which has a distribution function

$$H(x) = 1 - \exp\left\{ - \left[1 - \xi \left(\frac{x - \mu}{\sigma} \right) \right]_+^{-1/\xi} \right\} \tag{1}$$

with a location parameter $\mu \in \mathbb{R}$, a scale parameter $\sigma > 0$, a shape parameter $\xi \in \mathbb{R}$ and $[y]_+ = \max(y, 0)$, see Coles (2001). The negative Gumbel is given by the GEVM distribution with $\xi \to 0$, the negative Fréchet by $\xi > 0$ and the Weibull by $\xi < 0$. The justification for this distribution is that it is the only possible non-degenerate distribution for the limiting distribution of the minima of stationary sequences after linear normalisation, see Leadbetter et al. (1983).

Stationarity over a year may be a reasonable approximation to justify the choice of the GEVM distribution, but the obvious non-stationarity in the annual best times suggest that at least some of the parameters of the GEVM need to change over time, as in the modelling framework of Davison and Smith (1990) and Coles (2001). Here we assume that the GEVM parameter $\mu(t)$ varies smoothly over time t, with σ and ξ being a constant, i.e. the non-stationarity is purely a location shift in the elite performance in a year. This is justified by findings in Adam and Tawn (2011) and Adam and Tawn (2012) more generally for data of this type but also given the limited data in our applications it is also unlikely that we can find evidence that the scale and shape parameter changes over time. For the athletics example we take $\mu(t)$ as a simple exponential decay but for the swimming data no simple parametric model seems appropriate so we use non-parametric methods for estimating $\mu(t)$.

We study annual records using the information that they are simply a partially censored series of annual best data. Following the study of annual records we then develop a general Poisson process limiting characterisation for modelling the actual record data. Specifically, point process results for extremes, Pickands (1971) and Smith (1989), are adapted to record values recognising that records are a form of censored extreme value data.

Again modelling the non-stationarity of the underlying data needs to be addressed. Evidence from the analysis of annual records shows the need for information on the trend in elite performance (in the form of annual best data or something equivalent) to supplement information from the annual record to get any form of useful inference. In order to make inferences from the actual record data we need additional data that provides information about the trend of the underlying series. Annual best data are not available generally for other record series so some form of proxy data are required. For the actual record

data for swimming we use the Olympic gold medallists' times as proxy data for inference on the trend in elite performances.

The structure of this paper is as follows. In Sections 2 and 3 we derive models and inference methods for annual record data and actual record data respectively. As a large amount of data are censored this information needs to be accounted for not just in the fitting but also in methods of goodness-of-fit for both methods. Then we apply the theory to athletics data in Section 5 and swimmming data in Section 6.

2. Models and Inference for Annual Records

If $Z_1, \ldots, Z_n$ is the sequence of annual minima from $\{X_t\}$ with Z_t the annual minima in year t then we define $Y_t = \min(Z_1, Z_2, \ldots, Z_t)$ for $t \geq 1$ so that $\{Y_t\}$ is the annual record process. We assume that $\{Z_t\}$ are independent with $Z_t \sim \text{GEVM}(\mu(t), \sigma, \xi)$ and so $\mu(t)$, $t = 1, \ldots, n$, σ and ξ determine the distribution of the annual record process. Suppose that we have data for $\{Y_t\}$ but not $\{Z_t\}$. When $Y_t < Y_{t-1}$, i.e. a record occurs, then $Y_t = Z_t$ is an observation from a $\text{GEVM}(\mu(t), \sigma, \xi)$ variable. When $Y_t = Y_{t-1}$, i.e. not a record, then $Z_t > Y_t$ so Z_t is a $\text{GEVM}(\mu(t), \sigma, \xi)$ variable censored below at Y_t. Consequently, the log-likelihood function for the record sequence $\boldsymbol{y} = (y_1, \ldots, y_n)$ is

$$\begin{aligned} \ell(\boldsymbol{y}; \boldsymbol{\mu}, \sigma, \xi) &= -\sum_{t=1}^{n} I_t \log \sigma - \sum_{t=1}^{n} I_t \left(1 + \frac{1}{\xi}\right) \log \Psi_t \\ &\quad - \sum_{t=1}^{n} I_t \Psi_t^{-1/\xi} - \sum_{t=1}^{n} (1 - I_t) \Psi_t^{-1/\xi} \end{aligned} \tag{2}$$

where $\Psi_t = \left[1 - \xi \left(\frac{y_t - \mu(t)}{\sigma}\right)\right]_+$, with $I_t = I(Y_t < Y_{t-1})$ and $\boldsymbol{\mu} = (\mu(1), \ldots, \mu(n))$.

If $\mu(t)$ can be parametrically specified then maximum likelihood estimates are obtained by maximizing (2) directly. However if $\mu(t)$ is specified only to be a smooth function we use a penalized log-likelihood function for the GEVM distribution of the form

$$\ell(\boldsymbol{y}; \boldsymbol{\mu}, \sigma, \xi) - \frac{\lambda}{2} \boldsymbol{\mu}^T K \boldsymbol{\mu}, \tag{3}$$

where K is a symmetric $n \times n$ matrix of rank $n - 2$, defined in Green and Silverman (1994) which depends only on $(1, 2, \ldots, n)$ and λ is a smoothing parameter we select using the AICc criteria. The second term in the penalized

log-likelihood (3) imposes a penalty for the smooth $\boldsymbol{\mu}$ values. We maximize Equation (3) to get the best estimate values of $\hat{\boldsymbol{\mu}}$, $\hat{\sigma}$ and $\hat{\xi}$ using the Fisher's scoring method explained in Adam (2007).

As the estimated parameters determine the distribution of $\{Y_t\}$ through $\{Z_t\}$ we need to derive the distribution of $\{Y_t\}$. When $\{Z_t\}$ are independent and identically distributed (IID) then

$$\begin{aligned}\Pr(Y_t > y) &= \Pr(Z_1 > y)\Pr(Z_2 > y)\ldots\Pr(Z_t > y)\\ &= \exp\left\{-\left[1-\xi\left(\frac{y-\mu_t^*}{\sigma_t^*}\right)\right]_+^{-1/\xi}\right\}, \end{aligned} \tag{4}$$

where $\mu_t^* = \mu - \frac{\sigma}{\xi}(t^\xi - 1)$ and $\sigma_t^* = \sigma t^\xi$, so $Y_t \sim \text{GEVM}(\mu_t^*, \sigma_t^*, \xi)$, so the expected value for Y_t is $E(Y_t) = \mu_t^* + \frac{\sigma_t^*}{\xi}[1-\Gamma(1-\xi)]$. For the non-stationary case as $\Pr(Y_t > y)$ is complex we resort to deriving the distribution of $\{Y_t\}$ by simulation from $\{Z_t\}$.

We can assess the fit of the model through comparison of Y_m and its estimated expected value $\hat{E}(Y_m)$. This comparison is complicated by the strong dependence in $\{Y_t\}$. Instead we prefer to exploit the property that if $Z_t \sim \text{GEVM}(\mu(t), \sigma, \xi)$ then $Z_t = \mu(t) + E_t$, where $\mu(t)$ is the trend in Z_t and $E_t \sim \text{GEVM}(0, \sigma, \xi)$. Then the estimated sequence of $\{E_t\} = \{Z_t - \hat{\mu}(t)\}$ for $t = 1, \ldots, n$ are IID. We use estimated values of $\{E_t\}$ to construct the P-P and Q-Q plots for assessing the goodness of fit of the record progression data.

The initial step is to get the E_t value. If Y_t is a record then $Z_t = Y_t$, we get $E_t = Y_t - \mu(t)$. Otherwise, if Y_t is not a record then $Z_t > Y_{t-1}$ so $E_t > Y_t - \mu(t)$. Data on $\{E_t\}$ are therefore censored when no record is broken. To estimate the distribution of E we need to account for censoring, with the standard approach being the Kaplan-Meier estimator. Let $e_{(1)} < \ldots < e_{(m)}$ denote the residuals of the observed records and d_i be the number of $\{E_t\} > e_{(i)}$. The Kaplan-Meier estimate for a survival function is given by $\hat{H}_{KM}(z) = \prod_{i:e_{(i)}<z}\left(1-\frac{1}{d_i}\right)$. The model-based estimate of the distribution function for E is $\hat{H}(z) = 1 - \exp\left\{-\left[1-\hat{\xi}\left(\frac{z}{\hat{\sigma}}\right)\right]_+^{-1/\hat{\xi}}\right\}$. By comparing the model-based survival probability and empirical Kaplan Meier the P-P and Q-Q plots can be constructed. Let $e_{(j)}^*$ be the jth largest of $e_1, \ldots, e_n$ (i.e. observed and censored values together) then the P-P plot is constructed by plotting

$$\hat{H}(e_{(i)}^*) \quad \text{against} \quad \hat{H}_{KM}(e_{(i)}^*) \quad \text{for} \quad i = 1, \ldots, n.$$

and the Q-Q plot, using the exponential quantile plot, for the $\{e_i\}$ is

$$-\log\left[1-\hat{H}(e^*_{(i)})\right] \quad \text{versus} \quad -\log\left[1-\hat{H}_{KM}(e^*_{(i)})\right] \quad \text{for} \quad i=1,\ldots,n.$$

The transformation exponential quantile scale changes the lower tail into the upper tail.

3. Models and Inference for Actual Records

3.1 Point process model

First consider a stationary sequence $\{X_t\}$. We define the point process $P_n = \{[i/(n+1), (X_i - b_n)/a_n] : i = 1,\ldots,n\}$, where the sequence of constants $\{a_n > 0\}$ and $\{b_n\}$ are such that

$$\Pr\left[\frac{\min(X_1,\ldots,X_n)-b_n}{a_n} > x\right] \xrightarrow{n\to\infty} 1-H(x)$$

with $H(x)$ a non-degenerate limit distribution. Then $H(x)$ is GEVM(μ,σ,ξ) given by (1.1), see Leadbetter et al. (1983). Let P be the limit as $n\to\infty$ of P_n. Let $x_F = \min\{x : F_X(x) > 0\}$ where F_X is the distribution function of $\{X_t\}$, and $b_l = \lim_{n\to\infty}(x_F - b_n)/a_n$. For $A \subseteq [0,1]\times(-\infty, b_l)$ let $N_n(A)$ be the number of points of P_n that fall in A and $N(A)$ be the number of points of P in A. Following Pickands (1971) and Smith (1989) it follows that P is non-homogeneous Poisson processes with intensity

$$\lambda(t,r) = \frac{1}{\sigma}\left[1-\xi\left(\frac{r-\mu}{\sigma}\right)\right]_+^{-1-1/\xi}. \tag{5}$$

Now, if we let $A = [t_0,t_1]\times(-\infty,r)$, with $0 < t_0 < t_1 < 1$ and $r < b_l$, the limiting distribution of $N_n(A)$ is Poisson$\big(\Lambda(A)\big)$ with $\Lambda(A) = (t_1 - t_0)\left[1-\xi\left(\frac{r-\mu}{\sigma}\right)\right]_+^{-1/\xi}$. Now, we consider the non-stationary case with location parameter $\mu(t)$. The intensity measure for this non-homogeneous Poisson processes is then $\lambda(t,r) = \frac{1}{\sigma}\left[1-\xi\left(\frac{r-\mu(t)}{\sigma}\right)\right]_+^{-1-1/\xi}$ where here time is scaled onto $[0,1]$ each for t in $\mu(t)$.

Suppose that between times t_0 and t_{m+1} the times and values of a sequence of m consecutive actual records are $(t_1,r_1),\ldots,(t_m,r_m)$, with r_i the record value obtained at time t_i with $r_1 > \cdots > r_{m-1} > r_m$. It is assumed that the current record at time t_0 is r_0. We derive the likelihood for this actual

record sequence in stages. Using the Poisson processes the probability of no new record in the time interval (t_0, t_1) is

$$\exp\left[-\int_{t_0}^{t_1}\int_{-\infty}^{r_0}\lambda(t,r)drdt\right] = \exp\left\{-\int_{t_0}^{t_1}\left[1-\xi\left(\frac{r_0-\mu(t)}{\sigma}\right)\right]_+^{-1/\xi}dt\right\}.$$

The probability of one record in the interval $(r_i-\delta r, r_i)$ in the time interval $(t_i, t_i+\delta t)$,

$$\approx \lambda(t_i,r_i)\delta r\delta t\exp\left\{-\int_{t_i}^{t_i+\delta t}\int_{r_i-\delta r}^{r_i}\frac{1}{\sigma}\left[1-\xi\left(\frac{r-\mu(t)}{\sigma}\right)\right]_+^{-1-1/\xi}drdt\right\}$$

for small $\delta r > 0$ and $\delta t > 0$.

The probability of no record in the time interval $(t_i+\delta t, t_{i+1})$ is

$$\exp\left\{-\int_{t_i+\delta t}^{t_{i+1}}\int_{-\infty}^{r_i}\frac{1}{\sigma}\left[1-\xi\left(\frac{r-\mu(t)}{\sigma}\right)\right]_+^{-1-1/\xi}drdt\right\}.$$

So the joint probability of a record r_i at time t_i and then no record until t_{i+1} is proportional to

$$\lambda(t_i,r_i)\exp\left\{-\int_{t_i}^{t_{i+1}}\int_{-\infty}^{r_i}\frac{1}{\sigma}\left[1-\xi\left(\frac{r-\mu(t)}{\sigma}\right)\right]_+^{-1-1/\xi}drdt\right\}.$$

The likelihood for the record progression from t_0 to t_{m+1} is proportional to

$$\left[\prod_{i=1}^{m}\lambda(t_i,r_i)\right]\exp\left\{-\sum_{i=0}^{m}\int_{t_i}^{t_{i+1}}\left[1-\xi\left(\frac{r_i-\mu(t)}{\sigma}\right)\right]_+^{-1/\xi}dt\right\}. \tag{6}$$

As $\mu(t)$ changes slowly we expect the integral in Equation (6) to also change smoothly with t, and so for numerical simplification we approximate the sum of integrals in the log-likelihood function. In order to get an accurate value of the approximation of the integral, we break the time from t_j to t_{j+1} with $t_{j+1} = t_j + \Delta k_j$, where k_j is the number of partitions from t_j to t_{j+1} with equal distance of Δ and $s_{j,i} = t_j + (t_{j+1}-t_j)(i-1)/k_j, i = 1,\ldots,k_j+1$ to give

$$\sum_{i=0}^{m}\sum_{j=1}^{k_i}\Delta\left[1-\xi\left(\frac{r_i-\mu(s_{j,i})}{\sigma}\right)\right]_+^{-1/\xi}. \tag{7}$$

As in Section 2 non-parametric inference for $\mu(t)$ can be obtained by penalized likelihood with a penalty term identical to that in Equation (2).

3.2 Constructing tolerance bands

In order to develop new methods for diagnosing whether the model is acceptable or not for the actual record data, we need to deal with two aspects which are the time when the record occurred, t, and the value of the new record, r, given that a record occurs at time t. Our first approach is to generate tolerance intervals for the record progression under the assumption that the fitted model is correct. In constructing such tolerance intervals, the fitted model is used to simulate replicates of the record progression data. The resulting pointwise tolerance intervals produced provide a region of values of record progressions which can be used as a criteria for judging whether our fitted model is acceptable when compared to the observed record progression data. The observed record progression data should be within the region to indicate that the model is good.

Critical to this approach is the need to be able to simulate a record progression for the fitted model. For efficiency reasons we only want to simulate record times and values. To generate the record progression there are two stages given a current record time t_j and a current record value r_j. The first stage is to generate the next record time $t_{j+1}\,(t_{j+1} > t_j)$ and the second stage is to generate the next record value $r_{j+1}\,(r_{j+1} < r_j)$. Furthermore, we need an initialisation stage given (t_0, r_0) and a way for terminating the simulation. We give details of all these stages below. Fundamental to our approach in each case is the use of the limiting non-homogeneous Poisson process P. Refer to Equation (5).

First let us derive the waiting time distribution for the next record given that a record value of r_j occurred at time t_j. Let this waiting time random variable be W_{j+1}. Then $\Pr(W_{j+1} > w) = \Pr(\text{ no points of } P \text{ in } A_j)$ where $A_j = [t_j, t_j + w] \times (-\infty, r_j)$.

$$\Pr(\text{ no points of } P \text{ in } A_j) \quad = \quad \exp\{-\Lambda(A_j)\} = \exp\left[-\int_{t_j}^{t_j+w}\int_{-\infty}^{r_j} \lambda(s,r)drds.\right]$$

To simulate T_{j+1}, the time of the $(j+1)$th record value, we use the probability integral transformation for the survivor function of W_{j+1}. Specifically if u is a realisation of $U \sim U(0,1)$, then solving $u = \Pr(W_{j+1} > w)$ gives a realisation w from the waiting time distribution. The next record time is then generated as $t_{j+1} = t_j + w$, this is a realisation of T_{j+1}. Hence we solve for

t_{j+1} in $u = \exp\left[-\int_{t_j}^{t_{j+1}}\int_{-\infty}^{r_j}\lambda(s,r)drds\right]$, we get

$$-\log u = \int_{t_j}^{t_{j+1}}\int_{-\infty}^{r_j}\lambda(s,r)drds \approx \sum_{i=1}^{k_j+1}\Delta\left\{1-\frac{\xi}{\sigma}[r_j-\mu(s_i)]\right\}_+^{-1/\xi}. \quad (8)$$

The second stage is to derive the distribution of the decrease V_j in the record value given that the $(j+1)$th record value occurs at time t_{j+1}. The previous record value r_j is now a threshold for any further data to be the next new record value $r_{j+1} = r_j - v$ for $v > 0$. The probability of getting a decrease in record level greater than v at time t_{j+1} is $\Pr(V_{j+1} < v | T_{j+1} = t_{j+1}) = \Pr(R_{j+1} < r_j - v | R_{j+1} < r_j, T_{j+1} = t_{j+1})$ where R_{j+1} is the new record progression at t_{j+1}. However,

$$\Pr(R_{j+1} < r_{j+1} | R_{j+1} < r_j, T_{j+1} = t_{j+1}) = \left\{1-\xi\left[\frac{v}{\sigma-\xi(r_j-\mu(t_{j+1}))}\right]\right\}_+^{-1/\xi},$$

is the Generalised Pareto type distribution, $\mathrm{GPD}(\tilde{\sigma},\xi)$ with parameter $\tilde{\sigma} = \sigma - \xi(r_j - \mu(t_{i+1}))$. We simulate a new record value from the decrease random variable be V_i. To simulate r_{j+1} we use the probability integral transform for the distribution function of V_i. Using the realisation u of $U \sim U(0,1)$. Then solving $u = \Pr(V_j < v)$ gives a realisation from getting new record value distribution given time of new record occur. The next record value is then generated as $r_{j+1} = r_j - v$. Hence solve for v from $u = \left\{1-\xi\left[\frac{v}{\sigma-\xi(r_j-\mu(t_{j+1}))}\right]\right\}_+^{-1/\xi}$, we get,

$$r_{j+1} = r_j + \left[\frac{\sigma}{\xi} - (r_j - \mu(t_{j+1}))\right](1-u^{-\xi}). \quad (9)$$

The algorithm for constructing the tolerance bands is as follows:

1. Initiate $\sigma = \hat{\sigma}_{rec}$, $\xi = \hat{\xi}_{rec}$ and $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}} = \{\hat{\mu}_1, \ldots, \hat{\mu}_n\}$.
2. Selection of $r_0 (\leq r_1)$ and $t_0 (\leq t_1)$.
3. First stage: Simulate t_{j+1} from Equation (8) using t_j and r_j.
4. Second stage: Simulate r_{j+1} from Equation (9) using r_j and t_{j+1}.
5. If $t_{j+1} < T$ repeat steps 3 and 4 with T is the maximum time of interest. Stop if $t_{j+1} \geq T$.

We replicate the algorithm above k times to get k realisations of the record progression process of fitted model. We used the pointwise quantiles as the upper and lower bound of possible record value for the observed record progression. If the real record progression falls reasonably within the pointwise acceptance bounds, it indicates than the estimated value of the parameters are appropriate to the selected model and the model is reasonable description of the data.

As already we knew that $Z_i \sim \text{GEVM}(\mu_i, \sigma, \xi)$ then $Z_i = \mu_i + E_i$, where μ_i is the trend in Z_i and $E_i \sim \text{GEVM}(0, \sigma, \xi)$. Then the sequence of $\{E_i\}$ for $i = 1, \ldots, n$ are independent and identically distributed by separating the trend from the data. The E_i is used to construct the P-P and Q-Q plots for assessing the goodness of fit of the record progression data.

The initial step is to get the E_i value. If Y_i is a record then $y_i < y_{i-1}$ where $Z_i = Y_i$, we get $E_i = Y_i - \mu_i$. Otherwise, if Y_j is not a record then $y_j = y_{j-1}$ as $Z_j > Y_{j-1}$. We get $E_j > y_{j-1} - \mu_{j-1}$.

Data on E_j are therefore censored when no record is broken. To estimate the distribution of E we need to account for censoring, with the standard approach being the Kaplan-Meier estimator. Initially we need to plot the Kaplan-Meier estimate, $KM_j = \prod_j (1 - \frac{d_j}{t_j})$, to estimate the empirical survivor function of E, where d_j is the number of records occur until time j and t_j is the number of events left from time j. The model-based estimate of the survivor function for E_i is $\hat{S}(z) = \exp\left\{ -\left[1 - \hat{\xi}\left(\frac{z}{\hat{\sigma}}\right)\right]_+^{-1/\hat{\xi}} \right\} = 1 - \hat{H}(z)$. By comparing the model-based survival probability and empirical Kaplan Meier estimate from graph, the P-P and Q-Q plots can be constructed. When we plotting KM_j against $S(e_j^*)$, where e_j^* is the jth largest of $e_1, \ldots, e_n$ value, or transformations of these quantities give P-P and Q-Q plots. The P-P plot is constructed by plotting

$$\hat{H}(e_i^*) = 1 - \hat{S}(e_i^*) \quad \text{versus} \quad 1 - KM_i \quad \text{for} \quad i = 1, \ldots, n.$$

When studying the minima data, the fit of the upper tails is needed to relocate to the right hand of the graph. The Q-Q plot using the exponential quantile plot for the $\{e_i\}$ is

$$-\log\left[\hat{S}(e_i^*)\right] \quad \text{versus} \quad -\log\left[KM_i\right] \quad \text{for} \quad i = 1, \ldots, n.$$

In summary, to overcome the non-iid difficulties, we separate the trend from the data to get the iid residual data. Then construct the P-P and Q-Q plots from the residual data.

4. Estimating Trend from Additional Data

As the record data do not provided enough information to fully separate the record progression from a trend we need another different set of data from similar type of process which is able to give additional information to enable better separation. This additional source of data should be able to represent whole data including the record data.

Adam (2007) shown the difficulties of analysing extreme value data using parametric methods compared to non-parametric methods. The non-parametric approach allows the data to "speak for themselves" with less assumptions being made. Complex trends can be handled better using non-parametric approach compared to parametric approach.

We assume that $Z_1, Z_2, \ldots$ are the extreme additional data which can provide an information of the model trend in location for the record progression data. In this case we will using the annual minima type of data.

4.1 Non-parametric approach for GEVM

To allow for non-stationarity we replace μ with the smooth function g as we are estimating the trend in location using non-parametric method in this case of study. Then the density function of the GEVM at time t with smooth function $g(t)$ is defined as

$$h_t(z) = \frac{1}{\sigma}\left[1-\xi\left(\frac{z-g(t)}{\sigma}\right)\right]_+^{-1-1/\xi} \exp\left\{-\left[1-\xi\left(\frac{z-g(t)}{\sigma}\right)\right]_+^{-1/\xi}\right\}.$$

The penalized log-likelihood function for the GEVM distribution is

$$\ell_{\lambda_g}(\boldsymbol{z};\boldsymbol{g},\sigma,\xi) - \frac{\lambda_g}{2}\boldsymbol{g}^T K \boldsymbol{g} \tag{10}$$

with $\boldsymbol{g} = \big(g(t_1),\ldots,g(t_n)\big) = (g_1,\ldots,g_n)$ and $\ell_{\lambda_g}(z_i;g_i,\sigma,\xi)$ is the log-likelihood contribution for $\boldsymbol{z} = \{z_1,\ldots,z_n\}$ where

$$\begin{aligned}\ell_{\lambda_g}(\boldsymbol{z};\boldsymbol{g},\sigma,\xi) = &- n\log\sigma - (1+1/\xi)\sum_{i=1}^{n}\log\left[1-\xi\left(\frac{z_i-g_i}{\sigma}\right)\right]_+ \\ &- \sum_{i=1}^{n}\left[1-\xi\left(\frac{z_i-g_i}{\sigma}\right)\right]_+^{-1/\xi}\end{aligned} \tag{11}$$

and λ_g, the smoothing parameter value, is selected using the AICc criteria. We smooth g through the penalty function of (10) and maximize (11), to get the

best estimate values of $\hat{\boldsymbol{g}}$, $\hat{\sigma}$ and $\hat{\xi}$ using the Fisher's scoring method explained in Adam (2007). As our priority is to model and to analyse the record progression data, the only information relevant from this inference is the estimated $\hat{\boldsymbol{g}}$.

4.2 Non-parametric approach to estimate μ_t in a Poisson process model of record progression

To model μ_t Adam (2007) shown that using a non-parametric approach is preferable. When a non-parametric is implemented in the analysis, the μ_t values in Equation (7) are replaced by the smooth function g_t at time t with $t = 1, \ldots, m$. The log-likelihood function with g_t is

$$\ell(\boldsymbol{r}; g_t, \sigma, \xi) = \quad + \sum_{i=1}^{m} \log \lambda(t_i, r_i) - \sum_{i=0}^{m} \sum_{j=1}^{k_i} \Delta \left[1 - \xi \left(\frac{r_i - g_j}{\sigma}\right)\right]_+^{-1/\xi}. \quad (12)$$

In order to get an accurate value of approximation of the integration, we break the time from t_i to t_{i+1} to a small equal distance of Δ from partitions $[s_0, s_1]$ to $[s_{k_i-1}, s_{k_i}]$ we get the modified log-likelihood as

$$\ell(\boldsymbol{r}; g_t, \sigma, \xi) \quad = \quad + \sum_{i=1}^{m} \log \lambda(t_i, r_i) - \sum_{i=0}^{m} \sum_{j=1}^{k_i} \Delta \left[1 - \xi \left(\frac{r_i - g_{s_j}}{\sigma}\right)\right]_+^{-1/\xi} \quad (13)$$

where the distance from s_0 to s_{k_i} is the distance from t_i to t_{i+1} for $i = 1, \ldots, n$ with the distances between s_0 to s_{k_i} has been partitions to $[s_0, s_1], \ldots, [s_{k_i-1}, s_{k_i}]$.

5. Application to Athletics Annual Records

The five fastest annual order statistics for women's athletics 1500 m and 3000 m track events from recognized international events over the period of 1972-1992 are used in this study. See Robinson and Tawn (1995).

For this analysis we assume times run by every athlete in all events are independent. There is a trend for both events, it is based on the exponential decay of the annual minima over time. We use the trend, $\mu_t = \alpha - \beta[1 - \exp(-\gamma t)]$ where $\beta > 0$, $\gamma > 0$ and t is the year, taking 1971 as the base year. This model was used by Robinson and Tawn (1995).

We redefined our record as $Y_i = \min(Z_1, \ldots, Z_i), i = 1972, \ldots, 1992$ as in Figure 1. Let $Z_i \sim GEVM(\hat{\mu}_i, \hat{\sigma}, \hat{\xi})$ for $i = 72, \ldots, 92$, where this is independent but not identically distributed random variables. The random variable for

the number of record broken over the 21 years is $N_{21} = \sum_{i=72}^{92} I_i$, where $I_i = 1$ if $Z_i < \min(Z_{72}, \ldots, Z_{i-1})$ and is zero otherwise.

In Figure 3 the distribution of the number of records in annual minimum is shown for three situations: firstly if the annual minima are iid; second for the annual minima with trend and year to year variation as estimated for the women's athletics 1500 m data; finally the trend and year to year variation as estimated for the women's athletics 3000 m data. For iid case, it is possible for the record not to be broken for long time intervals. The distribution of the number of records is skewed to the right. For non-iid case, at least six records broken is predicted for the women's athletics 3000 m (5 times for the women's athletics 1500 m). This is because for the athletics, which is not identically distributed as $\hat{\mu}_i$ gives some obvious trend early in the period, we predict that the record tends to be broken frequently. The box-plot in Figure 3 summarise the discussion of a numbers of breaking record for the women's athletics 1500 m and 3000 m. In 1500 m track event the records have been broken 3 times

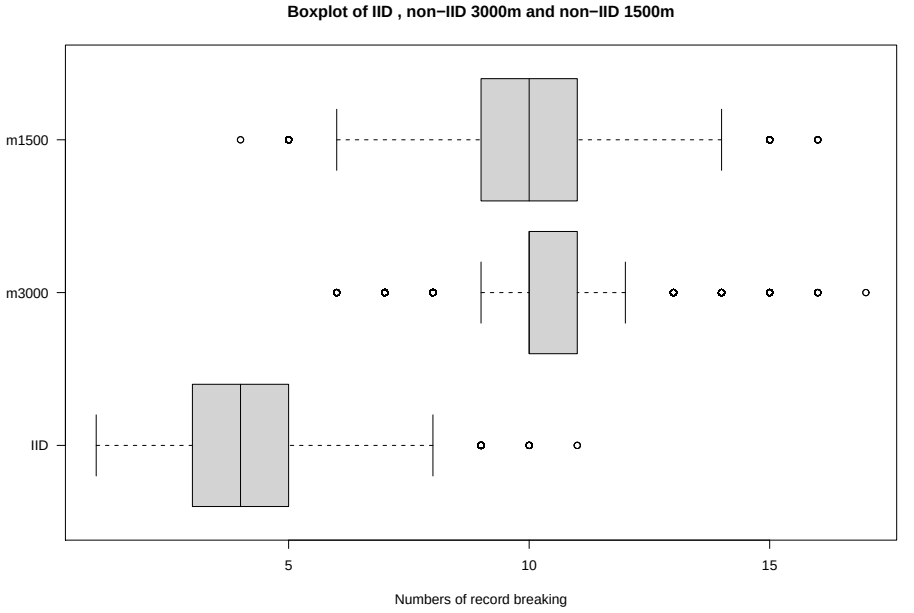

Figure 3: The box-plot of numbers of record breaking for annual minima for iid case and non-iid case for the women's athletics 3000 m and 1500 m, evaluated by Monte Carlo with 100000 repetitions.

since 1972 and some breaking records is predicted to happen. The women's athletics 3000 m at 1992, current number of breaking record is 6 times and more record also will be predicted to happen.

We now consider the women's athletics 1500 m and 3000 m events of Figure 1, focusing on the record data only. This led to the maximum likelihood estimate from Equation (2), we get the parameter estimates in Table 1. The 95 % CI for ξ for 1500 m and 3000 m events are (-5.585,5.018) and (-2.196,2.156). This is are unacceptably large confidence intervals. When maximizing the log likelihood with fixed α, β and γ values at their estimated values (using all annual minima data), the standard error for the shape are lowered, this is because separating the trend from the record process is hard if only the record are observed.

Table 1: The parameter estimation for record for the women's athletics 1500 m and 3000 m track events from 1972 to 1992.

Parameter	α	β	γ	σ	ξ
1500 m	248(29)	8.9(31.1)	0.251(0.609)	4.79(14.03)	-0.567(2.546)
3000 m	548(9)	37.7(13.9)	0.212(0.212)	4.12(7.23)	-0.020(1.110)

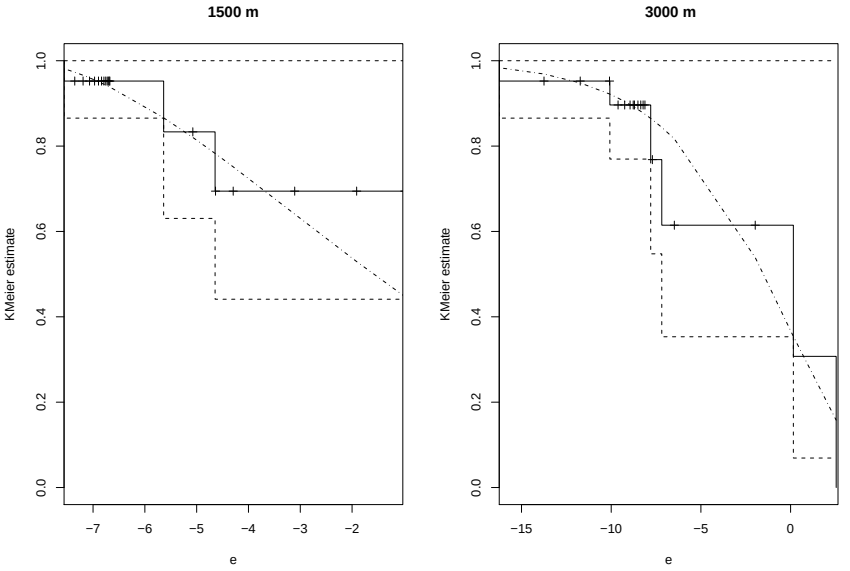

Figure 4: Kaplan-Meier estimates for the women's athletics 1500 m and 3000 m of e_i with the dotted line is survival function, $\hat{S}(e^*)$.

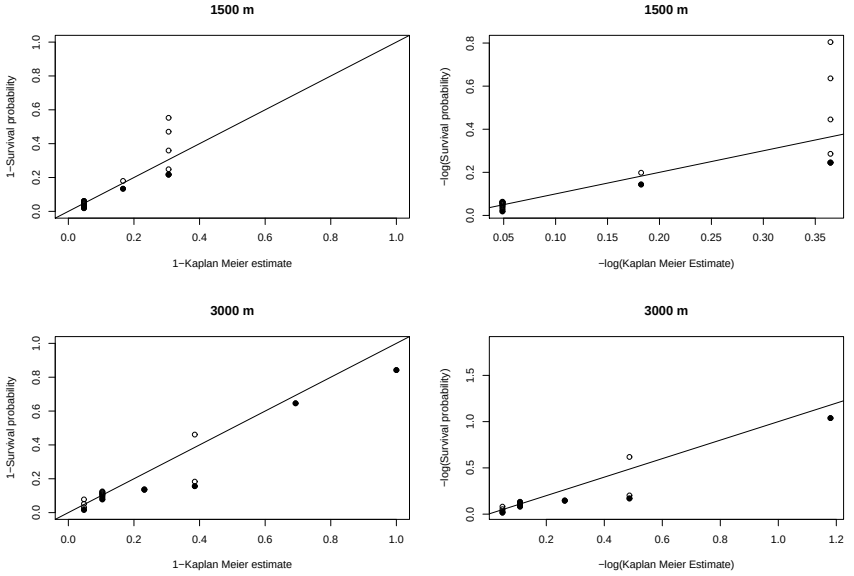

Figure 5: The P-P and Q-Q plots for the women's athletics 1500 m and 3000 m, with record points marked with black circle.

In order to fit the model, as the record data involving the censor data, we replace the empirical density function with a Kaplan-Meier estimate in P-P and Q-Q plots, see Figure 4. The P-P and Q-Q plots in Figure 5 shows that the model's fit is acceptable for the women's athletics 1500 m and 3000 m events as there are signs of linear pattern for in plots for censored data and the record (solid circle). We noted an important point, i.e. a number of the record depend wholly on the form of the trend of the mean.

6. Application to Actual Records Swimming

We applied the methods of analysis developed in this chapter using data on the records progression of the men's 400 m Freestyle swimming event with data which are recognised all around the world by FINA. Additional information is given by the men's Olympic 400 m Freestyle data.

We implement the theory and fitting the model propose in Sections 2-3. We used the men's annual record progression for 400 m Freestyle swimming event within 97 years time from 1908 to 2004. In the early stage of the competition, the Freestyle event is also included the Breaststroke and the Butterfly until

1953, when these two styles where separated from the Freestyle event.

Dealing with record data always led us to loose information regarding the trend in mean of true performance of the men's 400 m Freestyle. In order to overcome this missing information from record progression data, we will used the gold medallist of the men's Olympic 400 m Freestyle swimming events from 1908 to 2004 to obtain the trend on the mean.

We initially minimize the Equation (10) for the Olympic event to obtain the non-parametric trend in mean $\hat{g}_t$ (not shown here). This Fisher scoring method is capable in dealing with complexes trend in mean compared to parametric method, see Adam (2007).

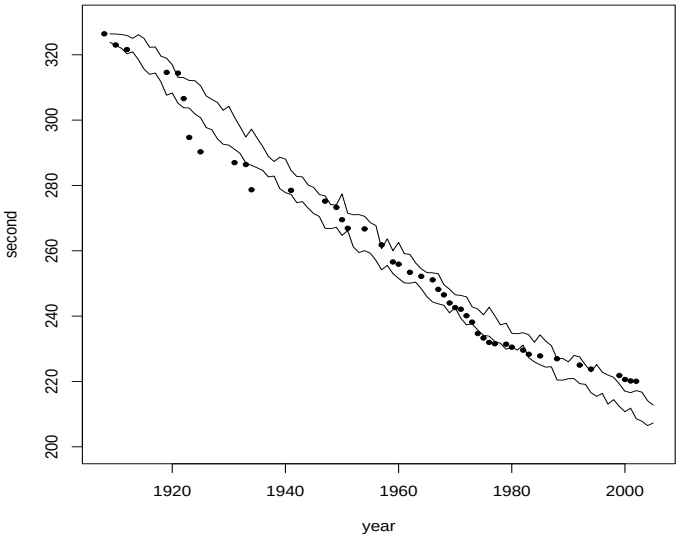

Figure 6: The maximum (upper line) and the minimum (lower line) value of the simulation record progression with $m = 33$ where m is the number of record have been simulated using $\hat{g}_t$, $\hat{\sigma}_{rec} = 5.32$ and $\hat{\xi}_{rec} = -0.202$.

Figure 6 shows that using the parameter estimates for the record progression using information from the men's Olympic events gives a fairly good fit compared to the observed records. More than 80% of the observed records are within the acceptance area (between the maximum and the minimum of simulated records). In Figure 7, the proposed P-P and Q-Q plots showed a good fit for the model.

We treated the event in 1908 as r_0, which means that it is not a considered as a record. Then there are 45 record have been recorded until 2004 for the

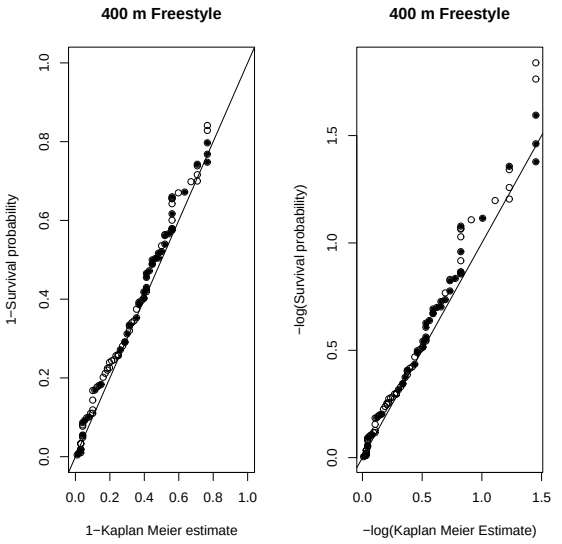

Figure 7: The P-P (left) and Q-Q (right) plots using Kaplan Meier estimation for record progression and censored data using method II in Section 3.

recognised men's 400 m Freestyle swimming events. From simulation of $m = 33$ times of record progression, the record have been broken between 34 to 50 times tabulated in Table 2, inclusively with the average record breaking is 43 times (further indicate that the simulated model is approximately similar with the real model). Noted that we only considered the annual progression of the best performance by men's 400 m Freestyle swimmer each years.

Table 2: Summary result of the number of records broken from $m = 33$ of simulation data using $\hat{g}_t$, $\hat{\sigma}_{rec}$ and $\hat{\xi}_{rec}$.

m	true	minima	median	mean	maximum
33	45	34	43	43.15	50

From the minimizing the log-likelihood of Equation (13), the parameter estimates for $\hat{\sigma}_{rec}$ is 5.32(0.79) and for $\hat{\xi}_{rec}$ is -0.202(0.089). The estimate of σ_{rec} values is quite similar with $\sigma_{Olympic}$ (5.625) indicate that the model proposed from Equation (6) is good.

7. Concluding Remarks

We have introduced a new method to analysis the progression of record data. We use the Poisson processes in order to build the joint density function for the record data. We have also introduced a new test for goodness-o-fit for the records. To account for trends in the underlying data the records are drawn from we used information from another existing events e.g. for the swimming record analysis we used the performance of the swimmers from Olympic event to estimate the non-parametric trend in mean. We limited the study to annually data which mean that if in a year few record broken events occurred we will stick to the latter record in our study.

We have also introduced two goodness-of-fit methods how to test the fitting model for the record progression data i.e. the tolerance regions of the model using some simulation data from estimated parameters and if we treated the record data as a censored type of extreme data, we used the P-P and Q-Q plots as alternative model diagnostics. To proceed the tolerance regions and plots we still need an information from another existing events.

The only drawback of the methods proposed here is that they require a higher capability of computer power. Two main aspects consume most computer ability: Fisher scoring method in estimating $\hat{g}_t$ and when constructing the tolerance region for the new proposed goodness of fit procedure.

References

Adam, M. B. (2007). *Extreme Value Modelling for Sports Data.* PhD thesis, Lancaster University, Lancaster, UK.

Adam, M. B. and Tawn, J. A. (2011). Modification of Pickand's dependent for ordered bivariate extreme distribution. *Communications in Statistics: Theory and Methods*, 40(9):1687–1700.

Adam, M. B. and Tawn, J. A. (2012). Bivariate extreme analysis of Olympic swimming data. *Journal of Statistical Theory and Practice*, 6(3):510–523.

Ahsanullah, M. (2004). *Record Values -Theory and Application.* University Press of America, Oxford.

Ahsanullah, M. and Bhoj, D. S. (1996). Record values of extreme value distributions and a test for domain of attraction of type i extreme value distribution. *Sankhyâ: The Indian Journal of Statistics*, 58:151–158. Series B.

Arnold, B. C., Balakrishnan, N., and Nagaraja, H. N. (1998). *Records*. Wiley, Canada.

Bairamov, I. G. (1996). Some distribution free properties of statistics based on record values and characterizations of the distributions through a record. *Journal of Applied Statistical Science*.

Benested, R. E. (2004). Record-values, non-stationary tests and extreme value distributions. *Global and Planetary Change*, pages 11–26.

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

David, H. A. and Nagaraja, H. N. (2003). *Order Statistics*. Wiley, New Jersey, third edition.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over hight thresholds. *Journal of the Royal Statistical Society Series B (Methodological)*, 52:393–442.

Glick, N. (1978). Breaking records and breaking boards. *The American Mathematical Monthly*, 85(1):2–26.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall, London.

Leadbetter, M. R., Lindgren, G., and Rootzen, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer, Berlin.

Pickands, J. (1971). The two-dimensional poisson process and extremal processes. *Journal of Applied Probability*, 8:745–756.

Robinson, M. E. and Tawn, J. A. (1995). Statistics for exceptional athletics records. *Applied Statistics*, 44(4):499–511.

Sibuya, M. and Nishimura, K. (1997). Prediction of record breaking. *Statistica Sinica*, 7:893–906.

Smith, R. L. (1988). Forecasting records by maximum likelihood. *Journal of the American Statistical Association*, 83:331–338.

Smith, R. L. (1989). Extreme value analysis of environmental time series: An example based on ozone data (with discussion). *Statistical Sciences*, 4:367–393.