ERASMUS UNIVERSITEIT ROTTERDAM

Ultimate Records in Athletics using Extreme Value Theory

Author:
Alexander R. J. Gruisen (416038)

Supervisor:
Xuan Leng
Second assessor:
dr. Alex J. Koning

Bachelor Thesis
Econometrics & Operations Research

Erasmus School of Economics
Erasmus University Rotterdam

July 2, 2017

Abstract

This paper addresses two questions: What is the ultimate record in a specific athletic event? and, how good is a current athletic world record? Data consist of all the personal bests of athletes in 28 events, 14 for both men and women. As an ultimate record is considered as a finite endpoint of an underlaying distribution, Extreme Value Theory and corresponding estimators are used to determine those endpoints. A simulation is performed to show the accuracy and consistency of the different estimators. After checking the finiteness of the endpoints, the ultimate records are calculated. The quality is measured by deriving its limiting distribution. Results indicate that some events have enough room for improvement (men's marathon, 3 minutes), while other current world records lay close to the ultimate record (women's shot put, 7 cm). Moreover, some current records have a high quality (women's marathon), whereas other records are likely to be sharpened in the near future (men's 110m hurdles). As extension, it is tested if the outdoor long jump data are contaminated by measurement errors, as the data do not satisfy initial assumptions. Results indicate that this is indeed the case, with the wind as possible factor that contributes to the measurement errors.

1 Introduction

On May 1st 2017 the European Athletics Council (EAC) endorsed a revolutionary plan to approach doping in athletics in a rigorous way. Doping is, from the beginning of athletics, a huge smear on the sport. Even with the modern drug tests it is still believed that many athletes are using these performance-enhancing substances. This came to an outburst when the International Association of Athletics Federations (IAAF) proved that the Russian Athletics Federation had been systematically forced doping upon their athletes, whom in consequence were expelled from participation to the Olympic Games in 2016. Ever since, people started questioning the records made by athletes. Subsequently, records are meaningless if people do not believe in the human capabilities. This doubt raises questions concerning the capabilities of humans in athletics. What can possibly be achieved by an athlete, given the today's state of art? In this paper two research question will be investigated:

"What is the ultimate world record in a specific athletic event?"
"How good is a current athletic world record?"

The first question measures the possible world record in the near future, given the present knowledge, materials and drug laws. The second research question can be interpreted as the difficulty of improving the standing world record. This enables to make the comparison between the quality of world records in the different athletic disciplines.

An ultimate world record, meaning no further improvement is possible, can be considered as the endpoint, an extreme-value, of the underlying distribution for achieved records. In the literature, estimates for endpoints can be obtained using Extreme Value Theory (EVT). This theory offers an semi-parametric approach to account for the tail region of the underlying distribution. Specifically, the tail region can be characterized by the Extreme Value Index (EVI), the shape parameter that displays the heaviness of the tail. Therefore, in the framework of EVT, the finite endpoint is estimated by the expression of the EVI. Since the EVI comes as a parameter of a limit distribution, after normalizing the sample maximum for the underlaying distribution, we can estimate the EVI in several approaches. Traditional methods to estimate the EVI include the Moment approach (Dekkers, Einmahl, & de Haan, 1989), the Maximum Likelihood Estimation (Smith, 1987), and the Probability Weighted Moment (PWM) estimator Hosking and Wallis (1987).

Despite the asymptotic normality of the PWM estimator, the estimator also contains an asymptotic bias. Recently, Cai, de Haan, and Zhou (2013) succeeded to correct the asymptotic bias of the PWM estimator, and made it applicable for estimating the EVI. They also showed that the choice of k, the number of high order statistics used for estimation, is far more flexible when one uses the bias-corrected PWM estimator. Specifically, the estimator allows a higher value of k which decreases the asymptotic variance of the estimator. These estimation methods, along with the framework of EVT, have been thoroughly studied and summarized in Fereira and de Haan (2006).

With the presented theory on estimating the EVI, applications with the focus on endpoints were ready to be performed. So did Einmahl and Magnus (2008), who studied the same two research questions as those of this paper. They used the Moment estimator as well as the Maximum Likelihood (ML) estimator for the EVI. They argued and showed that in sports the extreme values should be finite, and therefore considered endpoint estimation. Their findings were fairly accurate with the available data and techniques.

Extending Einmahl and Magnus (2008), most recent and ongoing research, performed by Leng et al. (2017), has proposed a method to estimate endpoints when the observations are contaminated by normally distributed errors. To put this in the context of this paper, it is also possible to get the endpoint estimator in some sport events where the measurement error plays a role, e.g. the outdoor long jump.

The remainder of the paper has the following structure. Section 2 introduces the dataset and how it will be made suitable to this application. Also it reports their descriptive statistics. Section 3 regards the methodology of the EVT and the different estimation methods used to estimate the EVI and endpoints. Section 4 describes how Monte Carlo Simulation will be used to investigate the finite sample performance of the estimation methods mentioned in Section 3. Section 5 reports and discusses the results of the estimation process applied on the sports dataset, as well as the results of an extension to the research. Section 6 concludes the paper. Afterwards, the used references are summarized.

2 Data

As mentioned before, Einmahl and Magnus (2008) already studied the two research questions proposed in this paper. Because this paper extends the aforementioned study, the dataset of Einmahl and Magnus (2008) will be used again. Therefore, we also focus on 28 sport events in the athletics (14 for both men and women). The selected events are as follows:

Running disciplines: 100m, 110/100m hurdles, 200m, 400m, 800m, 1500m, 10km, marathon
Throwing disciplines: Discus throw, javelin throw, shot put
Jumping disciplines: High jump, (outdoor) long jump, pole vault

With 110m hurdles for men and 100m hurdles for women. As the dataset only has data coverage up to April 30th 2005, new data will be supplemented to get a representative dataset up to now, May 1st 2017. The data will be obtained from the official IAAF site (*www.iaaf.org/statistics/top lists/index.html*) and the Swedish website compiled by Peter Larsson (*http://www.alltime-athletics. com/index.html*).

The data consist of all the personal best of the athletes in the events. Notice, as the interest of this paper is the records and not its development, an athlete can only appear once in the dataset. Hence only the personal best of an athlete is noted, even when this athlete has broken the world record multiple times. This data-filtering process has to be handled with caution and care, because several names of athletes are spelled differently or even misspelled over time. Specifically, many female athletes took on their husband's family name after marriage. A summary of the resulted dataset with the number of athletes and the worst and best record for each sport event can be seen in Table 1.

A preliminary look at the data indicates that the data occur in clusters, i.e. some records are the same between the athletes. For instance, in the running disciplines, equal records are likely caused by imperfect timing. Therefore the dataset will be smoothened using the method suggested by Einmahl and Magnus (2008), which is constructed as follows. For example, if c athletes share the same record, $d = 9.28$, it will be smoothened by

$$d_i = 9.275 + .01\frac{2i-1}{2c} \quad i = 1, \ldots, c$$

Note that the running records are expressed in time, which is inconvenient when considering finite right endpoints. Therefore the times will be transformed to the speeds of the athletes. For example, a time of 10.00 seconds in the 100m sprint is transformed to a speed of 36.00 km/h.

Table 1: Data summary

Event	Men			Women		
	Depth	Worst	Best	Depth	Worst	Best
Running						
100m	1217	10.3	9.58	715	11.38	10.49
110/100m hurdles	982	13.83	12.80	487	13.20	12.20
200m	1029	20.66	19.19	654	23.14	21.34
400m	826	45.74	43.03	646	52.02	47.60
800m	1026	1:46.61	1:40.91	663	2:01.07	1:53.28
1500m	1075	3:38.79	3:26.00	705	4:09.03	3:50.07
10km	2109	28:30.17	26:17.53	1175	33:04.00	29:17.45
Marathon	2223	2:13:36	2:02:57	1532	2:36:06	2:15:25
Throwing						
Shot put	405	19.80	23.12	232	18.42	22.63
Javelin throw	457	77.00	98.48	324	54.08	72.28
Discus throw	364	62.84	74.08	260	62.52	76.80
Jumping						
Long jump	804	7.80	8.95	836	6.30	7.52
High jump	527	2.26	2.45	433	1.90	2.09
Pole vault	625	5.50	6.16	632	4.00	5.06

3 Methodology

The methodology used in this paper can be split up into two distinct parts. The first part regards the EVT. The second part discusses the different estimators used in the EVT framework.

3.1 Extreme-Value Theory

The n records in an event are denoted as i.i.d. observations $X_1, \dots, X_n$, ordered by $X_{1,n} \leq \cdots \leq X_{n,n}$, with world record $X_{n,n}$, and are considered to come from some underlaying distribution F. Using EVT, if for the maximum $X_{n,n}$ a scale $a_n > 0$ and shift $b_n \in \mathbb{R}$ exists, such that

$$a_n^{-1}\left(X_{n,n} - b_n\right) \xrightarrow{d} G_\gamma(x) = \exp\left(-(1+\gamma x)^{-1/\gamma}\right) \tag{1}$$

as $n \to \infty$ and $1 + \gamma x > 0$, then F is in the max-domain of attraction (MDA) of G_γ. The γ is then called the EVI. For convenience, one can take logarithms of Equation (1), resulting in

$$\lim_{t\to\infty} t\left(1 - F(a_t x + b_t)\right) = -\log G_\gamma(x) = (1+\gamma x)^{-1/\gamma} \tag{2}$$

Taking the inverse in both sides of Equation (2), it can be equivalently written as

$$\lim_{t\to\infty} \frac{U(tx) - U(t)}{a(t)} = \frac{x^\gamma - 1}{\gamma} \tag{3}$$

with $b(t) = U(t)$, the quantile function that equals the left continuous inverse of F, i.e. $U(t) = F^{-1}\left(1 - \frac{1}{t}\right)$. In literature, Equation (3) is referred to as the first order condition. To ensure the asymptotic normalities of the estimators, a second order condition also has to be imposed. Assume a scale function $A(t)$ that satisfies $\lim_{t\to\infty} A(t) = 0$, such that for $x > 0$

$$\lim_{t\to\infty} \left(\frac{U(tx) - U(t)}{a(t)} - \frac{x^\gamma - 1}{\gamma}\right) / A(t) = \frac{x^{\gamma+\rho} - 1}{\rho(\gamma+\rho)} \tag{4}$$

Coming back to (3), one can observe that when t becomes large it can be written heuristically that

$$U(tx) \approx U(t) + a(t)\frac{x^{\gamma} - 1}{\gamma} \tag{5}$$

As mentioned, this paper's application is estimating the ultimate records, which are considered as right finite endpoints $x_F = \sup\{x|F(x) < 1\}$. To ensure finiteness, the EVI must be negative. If $\gamma < 0$, then for large x holds

$$x_F \approx U(t) - \frac{a(t)}{\gamma} \tag{6}$$

By setting $t = n/k$, with $k \to \infty$ and $k/n \to 0$ as $n \to \infty$, the ultimate record will be estimated by

$$\hat{x}_F = X_{n-k,n} - \frac{\hat{a}(n/k)}{\hat{\gamma}} \tag{7}$$

Where $U(t) = X_{n-k,n}$, the empirical analog being a high threshold, n the amount of athletes in an event and k the number of upper order statistics used for estimation.

The quality of a current world record will be measured by $n\left(1 - F(X_{n,n})\right)$, which is the expected amount of times that the current record will be improved. The lower, the more difficult to improve the standing record, hence higher quality. One can argue to use $x_F - X_{n,n}$ to measure the expected amount hence the quality of a record. However, as this quantity can be infinite, and moreover, it does not incorporate the tail behavior of underlaying distribution F, it is not convenient to this paper's application. From Equation (2), the quality $n\left(1 - F(X_{n,n})\right)$ can be approximated by $k\left(1 + \gamma\frac{X_{n,n} - b(n/k)}{a(n/k)}\right)^{-1/\gamma}$. Therefore, the expected amount will be estimated by

$$Q = k\left[\max\left(0, 1 + \hat{\gamma}\frac{X_{n,n} - \hat{b}}{\hat{a}}\right)\right]^{-1/\hat{\gamma}} \tag{8}$$

Assuming that the second order condition (4) holds, and combining it with Theorem 1 in Einmahl and Magnus (2008), it can be concluded that Q indeed coincides in probability with $n\left(1 - F(X_{n,n})\right)$, i.e. $\frac{Q}{n(1-F(X_{n,n}))} \xrightarrow{p} 1$. Hence, $X_{n,n}$ will cover all asymptotic randomness in Q, and not the estimation of F.

e^{-Q} will present the quality, as it serves a relative and absolute measurement due to its uniform (0,1) distribution in the limit. Thus, a higher value of e^{-Q} implies a better record, whereas a low e^{-Q} value indicates a relatively easy to improve record in the near future.

To make use of EVT and to estimate endpoints, the γ and $a(n/k)$ have to be estimated. The parameters will be estimated simultaneously, using different estimators for $(\hat{\gamma}, \hat{a}(n/k))$. As replication of Einmahl and Magnus (2008), the Moment estimator will be used. Instead of using the Maximum Likelihood estimator, the PWM estimator will be used in our paper. This allows for the comparison with the bias-corrected PWM estimator proposed by Cai et al. (2013), which will be implemented as extension. Moreover, to deal with potential measurement errors, the endpoint estimator of Leng et al. (2017) will be implemented, also as an extension.

3.2 The Moment estimator

The Moment estimator (called the M estimator) divides the EVI in two distinct parts, when negative or when positive. This can be formally written as $\gamma = \gamma_- + \gamma_+$, with $\gamma- = \min(\gamma, 0)$ and $\gamma_+ = \max(\gamma, 0)$. Define for $j = 1, 2$

$$M^{(j)} = M^{(j)}(k) = \frac{1}{k}\sum_{i=1}^{k}\left(\log X_{n-i+1,n} - \log X_{n-k,n}\right)^j$$

As can be seen, $M^{(j)}$ does not depend on the sample size n. Next, set $\hat{\gamma}_- = 1 - \frac{1}{2}\left(1 - \frac{\left(M^{(1)}\right)^2}{M^{(2)}}\right)^{-1}$ and $\hat{\gamma}_+ = M^{(1)}$. Subsequently, this results in

$$\hat{\gamma}_M = \hat{\gamma}_- + \hat{\gamma}_+ \quad \text{and} \quad \hat{a}_M(n/k) = X_{n-k,n}\hat{\gamma}_+(1 - \hat{\gamma}_-) \tag{9}$$

Dekkers et al. (1989) showed that $\left(\hat{\gamma}_M, \frac{\hat{a}_M(n/k)}{a(n/k)}\right)$ converges in probability to their true value $(\gamma, 1)$, stating the consistency. The asymptotic variance of the estimators and endpoint can be also found in their paper.

3.3 The PWM estimator

First assume a random variable $V \sim H_\gamma(x/a(n/k))$ i.e. V is generalized Pareto distributed, with $\mathrm{E}\{V\} = \frac{a(n/k)}{1-\gamma}$ and $\mathrm{E}\{V(1 - H_\gamma(x/a(n/k)))\} = \frac{a(n/k)}{2(2-\gamma)}$. Hosking and Wallis (1987) showed that if one approximates V by $X_{n-i+1,n} - X_{n-k,n}$, one can consistently estimate γ and $a(n/k)$ as follows. Define for $j = 1, 2$

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

Then $\mathrm{E}\{V\} \approx I_1$ and $\mathrm{E}\{V(1 - H_\gamma(x/a(n/k)))\} \approx I_2$, neglecting a possible scale, and get estimators

$$\hat{\gamma}_{PWM} = \frac{I_1 - 4I_2}{I_1 - 2I_2} \quad \text{and} \quad \hat{a}_{PWM}(n/k) = \frac{2I_1I_2}{I_1 - 2I_2} \tag{10}$$

The asymptotic variance of the estimators and endpoint is shown in Fereira and de Haan (2006).

3.4 The bias-corrected PWM estimator

The bias-corrected PWM estimator (referred to as the UB estimator) is estimated sequentially. The $\hat{\gamma}_{UB}$ is formed by subtracting the bias term from the PWM estimator, mentioned in Section 3.3. As the bias term depends on parameters from the second order condition in Equation (4), additional parameters need to be created. Firstly, a general estimator for γ is constructed as

$$\hat{\gamma}_{q,r} = \frac{q^2 I_q - r^2 I_r}{qI_q - rI_r}$$

Notice that $\hat{\gamma}_{2,1}$ is equal to $\hat{\gamma}_{PWM}$. Secondly, the second order index ρ and scale function $A\,(n/k)$ are constructed as

$$\hat{\rho} = 1 - \hat{\gamma}_{2,1} - \frac{1}{\frac{2-\hat{\gamma}_{2,1}}{1-\hat{\gamma}_{2,1}} \cdot \frac{\hat{\gamma}_{3,1}-\hat{\gamma}_{4,1}}{\hat{\gamma}_{3,2}-\hat{\gamma}_{4,2}} - 1} \quad \text{and} \quad \hat{A}\,(n/k) = (\hat{\gamma}_{2,1} - \hat{\gamma}_{3,1}) \cdot \frac{(1 - \hat{\gamma}_{2,1} - \hat{\rho})(2 - \hat{\gamma}_{2,1} - \hat{\rho})(3 - \hat{\gamma}_{2,1} - \hat{\rho})}{\hat{\rho}(1 - \hat{\gamma}_{2,1})}$$

Notice, the ρ depends on the choice of k (through $\hat{\gamma}_{q,r}$). Subsequently, the bias term can be constructed and the EVI can be estimated by

$$\hat{\gamma}_{UB} = \hat{\gamma}_{2,1} - \hat{A}\,(n/k)\,\frac{(1 - \hat{\gamma}_{2,1})(2 - \hat{\gamma}_{2,1})}{(1 - \hat{\gamma}_{2,1} - \hat{\rho})(2 - \hat{\gamma}_{2,1} - \hat{\rho})} \tag{11}$$

$\hat{\gamma}_{UB}$ depends on a different choice of k than the k corresponding to ρ. Therefore, one should first establish an accurate estimator for ρ, which will be used as a fixed parameter for the other estimates. In addition, the chosen k associated with ρ should be at a higher level compared to the k belonging to γ_{UB}, i.e. $\frac{k_\gamma}{k_\rho} \to 0$.

Cai et al. (2013) argued that the asymptotic bias of $\hat{a}_{PWM}(n/k)$ is incorrect, and therefore must be corrected. The resulting estimator for the scale function is given by

$$\hat{a}_{UB}(n/k) = \hat{a}(n/k) \cdot \exp\left(-\hat{A}\,(n/k)\,\frac{(1-\hat{\gamma}_{UB})(2-\hat{\gamma}_{UB}) - \hat{\rho}(3-2\hat{\gamma}_{UB})}{\hat{\rho}(1-\hat{\gamma}_{UB}-\hat{\rho})(2-\hat{\gamma}_{UB}-\hat{\rho})} \right) \tag{12}$$

The bias correction in $(\hat{\gamma}_{UB}, \hat{a}_{UB}(n/k))$ has also its effect on the estimation of the endpoint x_F. So, if the UB estimation method is used, not the standard endpoint estimator but the following bias-corrected endpoint estimator should be used:

$$\hat{x}_{UB} = X_{n-k,n} - \frac{\hat{a}_{UB}(n/k)}{\hat{\gamma}_{UB}} - \frac{\hat{A}\,(n/k)\,\hat{a}_{UB}(n/k)}{\hat{\rho}(\hat{\gamma}_{UB}+\hat{\rho})} \tag{13}$$

Following, the asymptotic variance of the estimators and endpoint is also altered. A thorough description can be found in Cai et al. (2013).

3.5 Endpoint estimator with measurement errors

In contrast to what is assumed before, data are often contaminated by measurement errors. Instead of observing the i.i.d. observations $X_i, \ldots, X_n$, a variable Y_i is observed, with $Y_i = X_i + \varepsilon_i$, $i = 1, \ldots, n$, where ε_i are i.i.d. random errors with mean 0, unknown variance σ^2 and are independent of X_i. The idea of the estimator (henceforth the G estimator) is to estimate the endpoint x_F by correcting the sample maximum directly with a shift quantity that approximates the error maximum. This may result in $\hat{x}_F = \max_{1\le i\le n} Y_i - \sigma\sqrt{2\log n}$. Ordering the Y_i gives the sample maximum $Y_{n,n}$, hence $x_F = Y_{n,n} - \sigma\sqrt{2\log n}$. Leng et al. (2017) showed that if one estimates the error standard deviation by

$$\hat{\sigma}_G = \frac{\sqrt{\log(n/k)}}{\sqrt{2}k}\sum_{i=1}^{k-1} g(i/k)\,(Y_{n-i,n} - Y_{n-k,n})$$

then an estimator of the endpoint is given as

$$\hat{x}_G = Y_{n,n} - \hat{\sigma}_G\sqrt{2\log n}$$

Notice, the G estimator depends on the sample size n, which is in contrast with the other estimators. To ensure the asymptotic normality of $\hat{\sigma}_G$, some conditions on $g(s), s \in (0,1]$ have to hold, which can be found in Leng et al. (2017). Using $g(s) = -\log s$ satisfies all conditions, hence it will be implemented in this estimator. The resulting asymptotic variance of the endpoint can be found in the proposed paper. As the G estimator only tries to filter out the measurement error in the endpoint, conditions on the EVI, such as $\gamma < 0$, will be omitted from the analysis.

4 Simulation

To assess the accuracy and consistency of the first three estimators (M, PWM, UB), a simulation will be performed by generating data from the Reversed Burr distribution

$$F(x) = 1 - \left(1 + (x_F - x)^{-\tau}\right)^{-\lambda}$$

with $x < x_F$, $x_F = 4$, $\tau = 4$, $\lambda = 1.25$, such that $\gamma = -1/(\tau\lambda) = -0.20$. As mentioned before, the UB estimator depends on a fixed ρ, which is set to -0.80. 100 samples are drawn with sample size $n = 5000$. The generated samples are sorted in ascending order and used as data to estimate the parameters $(\hat{\gamma}, \hat{a}(n/k))$, as well as the endpoint $\hat{x}_F$, using the different estimators. The estimated parameters and endpoints depend on the upper order statistic k. Therefore, the average estimates over the 100 samples will be calculated and plotted against k, with k varying from 1 to 1000, where $k/n < 20\%$.

The $\hat{\gamma}$ will be determined by balancing its bias and variance. Specifically, the optimal k will be chosen that minimizes the Root Mean Squared Error (RMSE). The accuracy of the corresponding $\hat{\gamma}$ will be tested by checking if the true γ value is included in the 95% Confidence Interval (CI), which will be constructed as follows

$$\left[\hat{\gamma} - z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{k}}, \hat{\gamma} + z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{k}}\right]$$

with $z_{\alpha/2}$ the 1-$\alpha/2$ quantile of the standard normal distribution with significance level $\alpha = 0.05$.

The UB estimator works only in the simulation when $\rho < 0$. Therefore the k_ρ and corresponding estimated $\hat{\rho}$ will be determined by searching amongst the negative values for the first stable plot. With the $\hat{\rho}$ fixed, the γ_{UB} can be estimated.

A similar procedure will be performed on the endpoint estimation. As alternative evaluation criteria, a box plot of the estimated endpoints will be made to show the accuracy and dispersion of the different estimators compared to the true endpoint value.

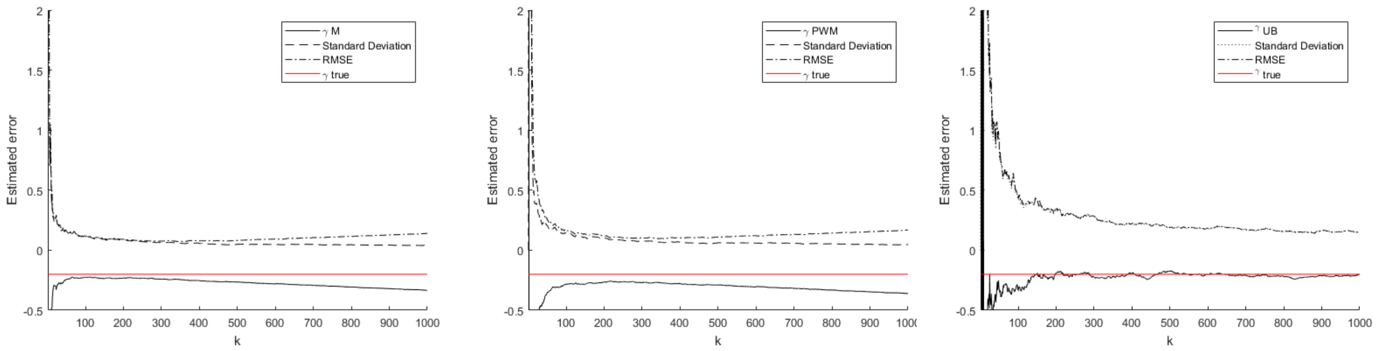
As the empirical analysis also considers endpoints with possible contamination of measurement errors, the proposed G estimator will be implemented and tested in a simulation to show its accuracy and consistency. To do this, data will be generated from $Y_i = X_i + \varepsilon_i, i = 1, \ldots, n$, where the X_i will come from the aforementioned Reversed Burr distribution and the parameter as stated in begin of this section. The ε_i will be drawn from a normal distribution $N(0, \sigma^2)$, with $\sigma = 2$ and $\sigma = 3$. Again, 100 samples with sample size $n = 5000$ will be generated, which will be used to estimate the G estimator.

For $\sigma = 2$ and $\sigma = 3$, choosing the k that minimizes the RMSE, the accuracy of the corresponding $\hat{x}_G$ will be assessed by testing if $x_F = 4$ is in its 95% CI. To analyze the consistency, a box plot of the estimated endpoints will be constructed. The performance of the G estimator will be compared with that of the PWM and UB estimator for x_F, which are the adequate estimators only when ignoring the presence of measurement errors in the data.

The simulation procedure of the estimators on fictional data will be performed in steps. Firstly, the accuracy and consistency of the EVI and endpoint estimation, in case of no measurement errors, will be tested by simulating the Reversed Burr distribution. In addition, the G estimator will be applied by simulating the Reversed Burr distribution with normally distributed errors, which should verify the accuracy and consistency of the estimator for $\hat{x}_F$.

Figure 1 shows the estimated γ versus k and its standard deviation and RMSE for the M, PWM and UB estimator.

Figure 1: EVI estimation for Reversed Burr Distribution



Note: Estimated EVI, standard deviation and RMSE for different values of k using the M, PWM and UB estimator. Data generated from the Reversed Burr Distribution without measurement errors, using 100 samples with sample size $n = 5000$.

One can observe that both the M estimator and the PWM estimator from the stable regions for the EVI slightly underestimate the true value. Preliminary to the EVI estimation of the UB estimator, the $\hat{\rho}$ has to be determined. The first stable plot is found when $k = 970$, which yields $\hat{\rho} = -0.93$. This differs from the true value -0.80, but as stated in Cai et al. (2013), the inaccuracy in estimating ρ does not impose a significant error when estimating γ. Therefore the $\hat{\rho}$ is kept fixed at -0.93. Looking at Figure 1, it can be concluded that the UB estimator honors its name, as it corrects the bias when k becomes large, and moreover, the asymptotic variance drops when k increases.

To assess the accuracy of the estimated EVI, the 95% CI is build for the chosen k that minimizes the RMSE. The k value, the γ point estimate, as well as its 95% CI can be seen in Table 2.

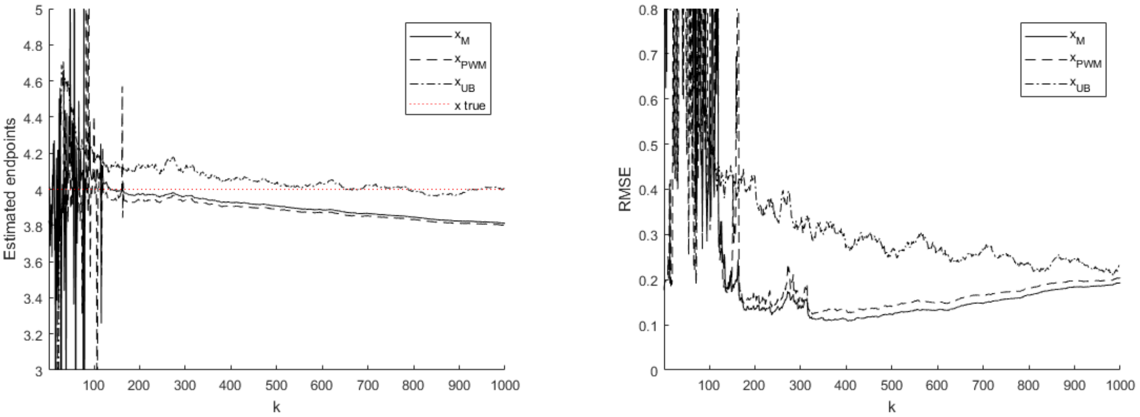
Table 2: 95% CI of $\hat{\gamma}$ for corresponding k

	k	$\hat{\gamma}$	95% CI
$\hat{\gamma}_M$	362	-0.247	[-0.355, -0.139]
$\hat{\gamma}_{PWM}$	347	-0.270	[-0.400, -0.139]
$\hat{\gamma}_{UB}$	770	-0.183	[-0.484, 0.117]

It becomes evident that the optimal k for the UB estimator is twice the value of the k for the M and PWM estimator. This is no coincidence, as the UB estimator performs well for a wider range of k. Moreover, as $k_{\gamma_{UB}} = 770$, the condition for the UB estimator that $\frac{k_\gamma}{k_\rho} \to 0$ when $k \to \infty$ is satisfied. As the true $\gamma = -0.20$ is in each of estimators' 95% CI, it can not be rejected that the estimated EVI equals the true one. From this it can be concluded that the different estimators are accurate regarding the EVI estimation.

Next, the accuracy and consistency of the endpoints are investigated. Figure 2 shows the estimated endpoints and RMSE versus k, using the different estimators.

Figure 2: Endpoint estimation for Reversed Burr Distribution



Note: Estimated endpoints x_F and RMSE for different values of k using the M, PWM and UB estimator. Data generated from the Reversed Burr Distribution without measurement errors, using 100 samples with sample size $n = 5000$.

When k is small, the M and PWM estimator have trouble to estimate the endpoints accurately. When k starts to increase, they become more stable, resulting in their RMSE to drop. However, the M and PWM estimator start to develop a bias when k becomes large, which consequently leads to an increase in the RMSE. The course of the UB estimator is more stable, as it shows no outliers. Moreover, it converges to the true endpoint. The RMSE drops gradually, but for $k < 1000$ it stays above the RMSE of the M and PWM estimator. As done with the EVI, the k that minimizes the RMSE is chosen to determine the endpoint and corresponding 95% CI, which are shown in Table 3.

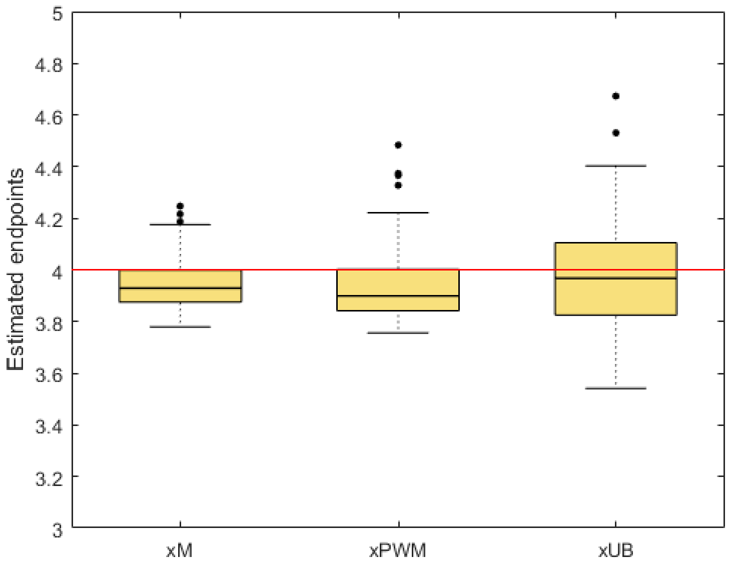
Table 3: 95% CI of $\hat{x}_F$ for corresponding k

	k	$\hat{x}_F$	95% CI
$\hat{x}_M$	352	3.946	[3.749, 4.143]
$\hat{x}_{PWM}$	356	3.918	[3.722, 4.113]
$\hat{x}_{UB}$	781	3.987	[3.574, 4.400]

Again, the optimal k for the UB estimator doubles the k chosen for the M and PWM estimator, which share similar k values. The UB estimator estimates the endpoint the most accurate, but has also the highest variance. As can be seen in Table 3, every 95% CI includes the true endpoint value 4, hence it can not be rejected that the estimated endpoints are similar to the true endpoint. This shows that the different estimators estimate the endpoint accurate. As mentioned before, in addition, a box plot of the estimated endpoints is shown in Figure 3.

Despite that the M and PWM estimator slightly underestimate the endpoints, the box plot shows that the estimators perform well and are fairly similar to each other; the medians are close to the true endpoint and variations are low. This concludes that the estimators are accurate as well as consistent estimators for the EVI and endpoints, when considering data without measurement errors.

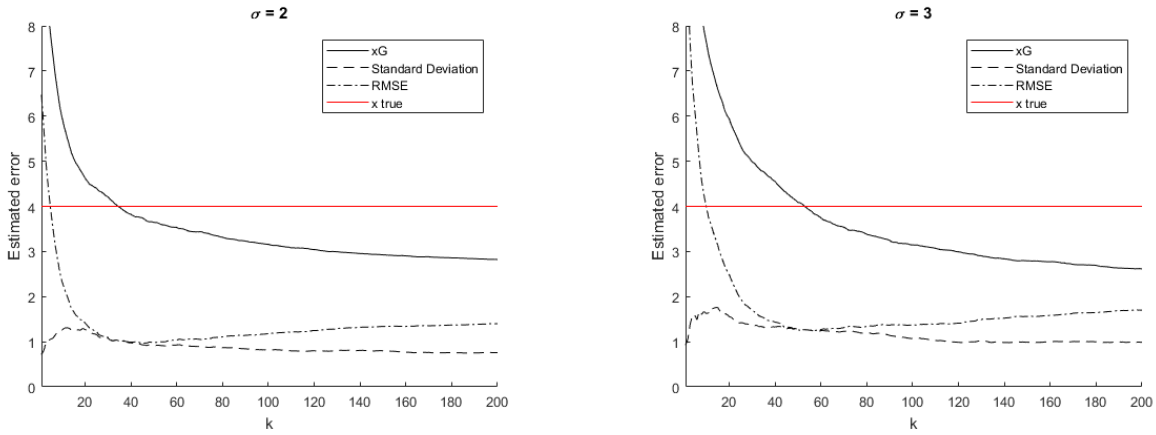
Figure 3: Box plots of estimated endpoints



Note: Box plots of estimated endpoints x_F in optimal k, using the M, PWM and UB estimator. Data generated from the Reversed Burr Distribution without measurement errors, using 100 samples with sample size $n = 5000$.

Next, the accuracy and consistency of the G estimator is tested on the data with measurement errors. Figure 4 shows the estimated endpoints x_F of the G estimator versus k. The standard deviation and RMSE are also plotted.

Figure 4: Endpoint estimation for Reversed Burr Distribution



Note: Estimated endpoints x_F, standard deviation and RMSE for different values of k using the G estimator. Data generated from the Reversed Burr Distribution with normally distributed measurement errors, using 100 samples with sample size $n = 5000$.

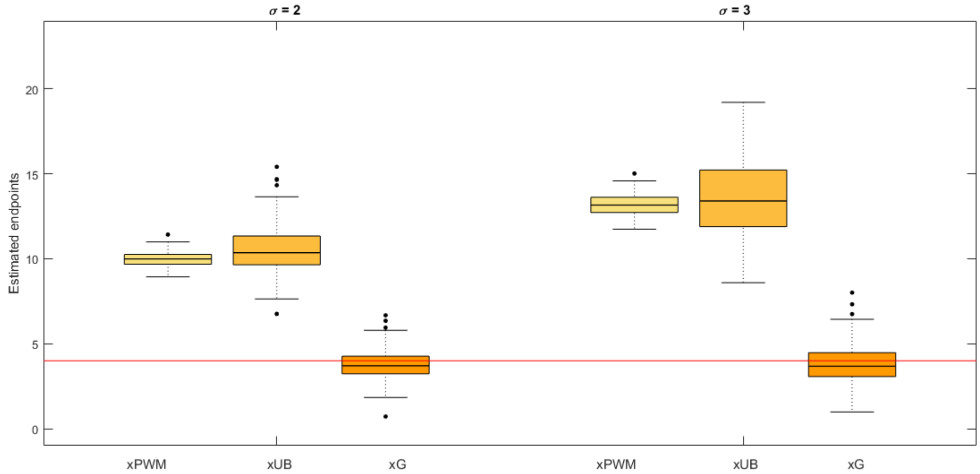
It becomes visible that the estimated endpoints create a bias which underestimate the true endpoint, denoted as the red line. Again, the k is chosen that minimizes the RMSE. The optimal k for data with $\sigma = 2$ and 3 is located at 45 and 57 respectively. Compared to other estimators, this is relatively early in the plot. Afterwards, the RMSE increases rapidly to a constant level of 1.5 and 1.8 for $\sigma = 2$ and 3 respectively. To still show the point where the RMSE is minimized, both plots are created with k up to 200. The estimated x_F and its 95% CI, belonging to the optimal k, are constructed to evaluate the accuracy. Table 4 presents the 95% CI for the different σ levels.

Table 4: 95% CI of $\hat{x}_F$ for corresponding k

	$\sigma = 2$			$\sigma = 3$		
	k	$\hat{x}_F$	95% CI	k	$\hat{x}_F$	95% CI
$\hat{x}_G$	45	3.745	[3.477, 4.017]	57	3.832	[3.512, 4.152]

The true endpoint $x_F = 4$ is included in the 95% CI, for both $\sigma = 2$ and 3. Hence it can not be rejected that the estimated endpoints are similar to the true endpoint. This confirms that the G estimator is an accurate estimator for the endpoint when the data are contaminated by normally distributed measurement errors. The performances of the PWM, UB and G estimator in case of measurement errors are compared using a box plot, which is shown in Figure 5. The chosen k for the PWM and UB estimator is also based on the minimization of the RMSE.

Figure 5: Box plots of estimated endpoints



Note: Box plots of estimated endpoints x_F in optimal k, using the M, PWM and UB estimator. Data generated from the Reversed Burr Distribution with normally distributed measurement errors, using 100 samples with sample size $n = 5000$.

The PWM and UB estimator largely overestimate the true endpoint $x_F = 4$. They become even more inaccurate when the σ increases from 2 to 3. In contrast, the G estimator is accurate for both levels of σ. Moreover, the median is close to the true endpoint and the dispersions remain low. Hence, it can be concluded that the G estimator is accurate and consistent in estimating endpoints with data being contaminated by measurement errors.

With the accuracy and consistency of the different estimators been shown, the estimation process on the sports dataset can be performed in parts. Firstly, the assumptions that the EVI exists must be tested, i.e. first order condition (3) must hold for an arbitrary $\gamma \in \mathbb{R}$. Einmahl and Magnus (2008) showed that the EVI exists for all events except the men's and women's pole vault. Therefore, these two events will be omitted from further analysis, which leaves a 13 x 2 dataset. It is true that this paper uses new data, which potentially causes different results. However, to make the comparison and for convenience, this paper will use the aforementioned results concerning existence of the EVI. Nevertheless, the assumption that $F \in MDA(G_\gamma)$ for $\gamma < 0$ holds for each sport event, ensuring finite right endpoints, will be tested using the 95% CI of the different estimators for $\hat{\gamma}$. Because $\hat{\gamma}$ depends on k, a plot of $(k, \hat{\gamma})$ will be made, with k starting from 1 to where $k/n < 20\%$. Determining the first coinciding stable region of the different estimators, and taking averages over the region and estimators, will result in the estimate for γ.

Secondly, for events that have $\hat{\gamma} < 0$, ensuring finite endpoints, the ultimate world records x_F will be estimated using the different estimators for $(\hat{\gamma}, \hat{a}(n/k))$, where k is determined from the plot of $(k, \hat{\gamma})$. In addition, the $\hat{\gamma}$ found in the previous step will be used as a fixed EVI across k, such that the endpoint x_{fixed} only depends on k through $X_{n-k,n}$ and $M^{(1)}$, which produces a much more stable plot of x_F, as shown in Einmahl and Magnus (2008). Subsequently, the quality of a current world record e^{-Q} will be determined using the previous procedure, likewise with the fixed estimator Q_{fixed} included.

Lastly, as extension, the possible presence of measurement errors in records in the long jump discipline will be investigated. This will be done by making the distinction between an outdoor and indoor long jump data. First the hypothesis that $\gamma \geq 0$ will be tested for the indoor and outdoor data. If the tests result in different conclusions, contamination of measurement errors could be present in the outdoor data and consequently the endpoint. Therefore, for outdoor long jump, the endpoints will be estimated using the G estimator, as it accounts for possible measurement errors. Notice, the G estimator requires a predetermined sample size n. As n should reflect the total number of top athletes in an event, the data could possibly not represent all those athletes, i.e. n of the data could be lower than the actual n. Therefore an arbitrary value $n = 3000$ is taken to address this caveat. After estimation, the estimated endpoint for outdoor, using G estimator, and that of indoor, using the PWM and UB estimator, will be compared. Notice, for indoor long jump the M estimator will not be used. If the endpoints mutually agree with each other, the G estimator for outdoor long jump data can be verified. In case of different values, the estimated endpoints are compared with the sport records, to find a possible cause for the measurement errors.

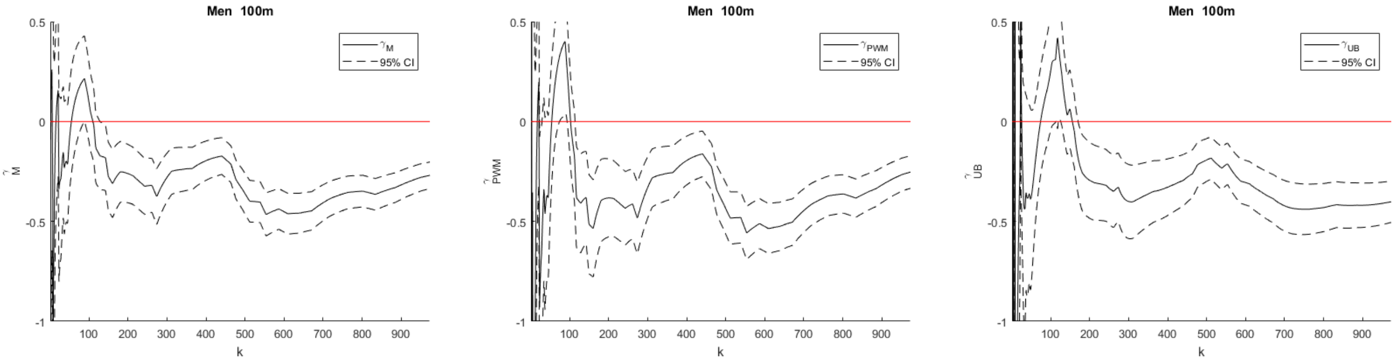
5 Results

5.1 Extreme Value Index

As mentioned before, a finite endpoint can only be ensured if the EVI is negative. Testing if $\gamma \geq 0$ is performed for each of the 26 sport events, using the same procedure. As example, the men's 100m is used to show this testing-procedure graphically. Figure 6 shows the plots of the estimated γ with their 95% CI versus k, using the M, PWM, and UB estimator. After inspection, the $\hat{\rho}$ of the UB estimator is set to -8.80.

As can be seen from the graphs, the three estimators behave quite similar. For $k > 100$, The M and PWM estimator show some dispersion, whereas the UB estimator is more stable. To make conclusive statements regarding $\gamma \geq 0$, the significance level of the estimated γ is needed. For this the estimated γ has to be determined using stable regions. The first stable region of the M estimator can be located between $k = 140$ and 475. For the PWM estimator, this can be found from $k = 120$ to 475. Lastly, the first stable region for the UB estimator runs from $k = 175$ to 575, for which it holds that $\frac{k_\gamma}{k_\rho} \to 0$ when $k \to \infty$. Notice that the UB estimator is stable for a wider span of k, as proposed. Consequently, the region where all three estimators are stable is from $k = 175$ to 475. Taking averages over the different k and estimators, results in $\hat{\gamma} = -0.29$.

Figure 6: Estimated EVI



Note: Estimated EVI with corresponding 95% confidence intervals for different values of k, using the M, PWM and UB estimator.

Next, the 95% CI of the estimated γ is assessed and tested if $\gamma \geq 0$. As can be seen in Figure 6, for the k included in the stable region, all estimates' 95% CI do not contain 0 or a positive value. Statistically, the average 95% CI for the average estimate $\hat{\gamma} = -0.29$ is $[-0.45, -0.15]$, which also does not include positive values. Based on the 95% CI, one can reject that $\gamma \geq 0$ for men's 100m, and therefore the finite endpoint is ensured in this particular event.
This procedure is performed for every event, both for men and women. Table 5 shows the average estimated EVI of the estimators. To compare, the $\hat{\gamma}$ from Einmahl and Magnus (2008) are included.

Table 5: EVI $\hat{\gamma}$

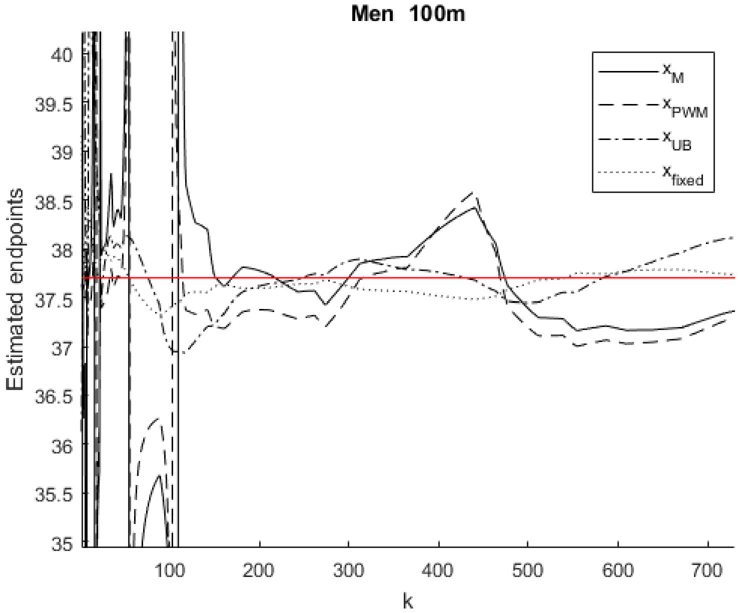
	Men		Women	
Event	$\hat{\gamma}$	Einmahl's $\hat{\gamma}$	$\hat{\gamma}$	Einmahl's $\hat{\gamma}$
Running				
100 m	-0.29	-0.11	-0.43	-0.14
110/100-m hurdles	-0.27	-0.16	-0.39	-0.25
200 m	-0.12	-0.11	-0.07*	-0.18
400 m	-0.06*	-0.07*	-0.25	-0.15
800 m	-0.33	-0.20	-0.27	-0.26
1500 m	-0.26	-0.20	-0.15	-0.29
10 km	-0.16	-0.04*	-0.11	-0.08*
Marathon	-0.17	-0.27	-0.27	-0.11
Throwing				
Shot put	-0.28	-0.18	-0.49	-0.30
Javelin throw	-0.20	-0.15	-0.18	-0.30
Discus throw	-0.24	-0.23	-0.22	-0.16
Jumping				
High jump	0.03*	-0.20	-0.34	-0.22
Long jump	-0.03*	0.06*	-0.11*	-0.07*

Results indicate that some events have a positive EVI value, e.g. high jump for men, whereas other events have indeed a negative EVI value which can not statistically be rejected that $\gamma \geq 0$, e.g. 400m for men, 200m for women and the outdoor long jump for men and women. These events, and those rendered insignificant by Einmahl, are marked with a * in Table 5. The differences between this paper's and Einmahl's $\hat{\gamma}$ are relatively big. Specifically, for 20 of the 26 estimated EVIs, Einmahl's estimate is much higher. A possible explanation is the way of estimating the EVI. Einmahl and Magnus (2008) use the M and ML estimator to determine the $\hat{\gamma}$, whereas this paper considers the M, PWM and UB estimator. Moreover, in both papers the EVI is based on determining the first stable region, which is a semi-subjective view. Therefore different k values could be used for the estimates which can cause the variations in estimates. A more obvious answer is the fact that the datasets are different, hence different results. Nevertheless, the different values of $\hat{\gamma}$ will have its effect on the estimated endpoint, which will be discussed next.

5.2 Ultimate World Records

With the estimated EVI, the first research question, the ultimate world records, can be addressed. For events that have an insignificant negative EVI, a finite endpoint can not be guaranteed. Therefore only for the events that have a significant negative γ, the endpoint is estimated. As example, Figure 7 shows the estimated endpoints versus k for the men's 100m.

Figure 7: Estimated endpoints



Note: Estimated endpoints x_F for different values of k using the M, PWM, UB and fixed estimator, resulting in endpoint estimator $\hat{x}_F$, indicated as the red line.

As can be seen in Figure 7, the x_{fixed} (dotted line) gives indeed a more stable plot of the endpoints than the other estimators. Similar to the EVI plot, the first coinciding stable region runs from $k = 175$ to 475. Taking averages over the region and different estimators, including x_{fixed}, results in $x_F = 37.70$, which is represented as the red line. This procedure is repeated for the remaining 22 of 26 events which have a significant $\gamma < 0$. Table 6 shows the estimated ultimate records and its 95% CI, for each discipline. For the running events the speeds are transformed back to times. The choice for the 95% CI rather than the standard errors is because it is unreasonable to convert the errors from speeds to time, as ought with the running disciplines. For comparison, the current world records, as well as the estimated ultimate records of Einmahl and Magnus (2008) are presented.

Table 6: Ultimate world records $\hat{x}_F$

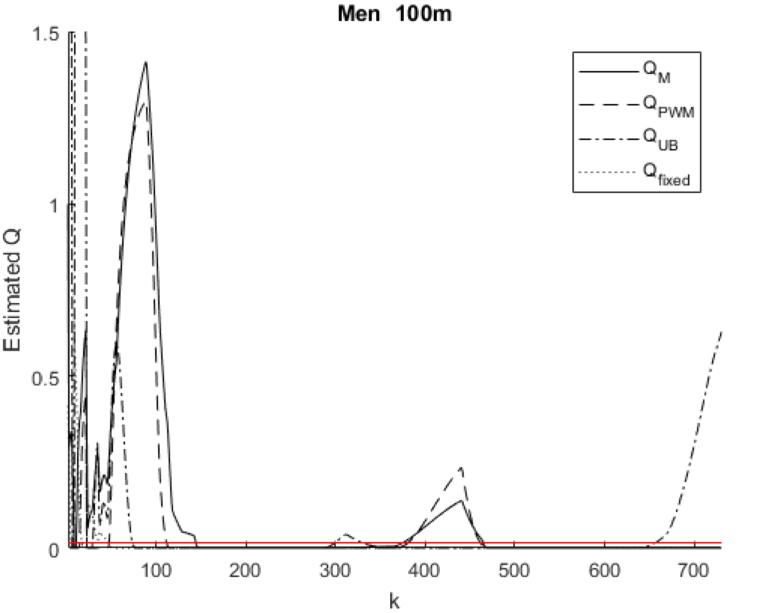
	Men				Women			
Event	Ultimate record	95% CI	Current record	Einmahl's record	Ultimate record	95% CI	Current record	Einmahl's record
Running								
100m	9.55	[9.50, 9.58]	9.58	9.29	10.29	[10.20, 10.38]	10.49	10.11
110/100m h	12.58	[12.36, 12.79]	12.80	12.38	12.09	[11.87, 12.28]	12.20	11.98
200m	18.68	[18.38, 19.08]	19.19	18.63	—	—	21.34	20.75
400m	—	—	43.03	—	47.01	[45.81, 48.00]	47.60	45.79
800m	1:40.29	[1:39.28, 1:41.28]	1:40.91	1:39:65	1:52.37	[1:50.77, 1:53.92]	1:53.28	1:52.28
1500m	3:24.16	[3:21.56, 3:26.56]	3:26.00	3:22.63	3:33.80	[3:33.26, 3:34.43]	3:50.07	3:48.33
10km	25:20.16	[24:34.79, 26:03.03]	26:17.53	—	26:49.84	[25:50.43, 27:42.83]	29:17.45	—
Marathon	1:59:55	[1:57:00, 2:02:09]	2:02:57	2:04:06	2:15:17	[2:12:57, 2:17:38]	2:15:25	2:06:35
Throwing								
Shot put	23.75	[22.78, 24.78]	23.12	24.80	22.69	[21.60, 23.84]	22.63	23.70
Javelin throw	100.97	[95.31,105.95]	98.48	106.50	79.11	[71.60, 87.52]	72.28	72.50
Discus throw	77.59	[71.22, 84.60]	74.08	77.00	80.17	[74.66, 85.26]	76.80	85.00
Jumping								
High jump	—	—	2.45	2.50	2.11	[2.06, 2.15]	2.09	2.15
Long jump	—	—	8.95	—	—	—	7.52	—

The records of events that could not guarantee a finite endpoint, are left blank. Looking at the results, differences between men and women become visible. The men's marathon record has enough time to improve (3 minutes), while for women it is already close to the boundary (8 seconds). With the 10km run this is the other way around; men can only improve the record by 57 seconds, but women can sharpen the current world record by almost 2.5 minutes. Big differences occur when comparing the ultimate records with those of Einmahl and Magnus (2008). For 15 of the 19 comparable ultimate records, Einmahl's endpoints are higher, i.e. greater distance and fewer time. Only for the men's marathon, men's discus throw, women's 1500m and women's javelin throw, this paper estimated a larger frontier. This is most likely caused by the fact that each of the 4 disciplines had their world record broken after April 30, 2005, the day that the data collection of Einmahl and Magnus (2008) ended. Moreover, the world records are sharpened after October 25, 2007, the data used by Einmahl and Magnus (2008) to establish the current world records in their paper. Because the records are improved, i.e. the maximum values of the dataset are higher, the estimators could consider higher endpoints. However, for 7 of the 15 events, where Einmahl's endpoints are estimated higher, their world records were also improved after April 2007. So recent set records can not causally be accountable for the contrast between the ultimate records. More thorough explanation lays in the nature of the endpoint. In Section 5.1 it became obvious that this paper's EVI estimates are lower than those in Einmahl and Magnus (2008). When the EVI is relatively low, the estimated underlaying distribution of the endpoint is relatively steep near the right endpoint. Hence, there is not much space for improvement of the record, resulting in ultimate records close to the current record. This is also the reason why the relatively high EVI values of Einmahl and Magnus (2008) give such high endpoints. This leaves the question why Einmahl and Magnus (2008) estimate the EVI higher. An explanation can be found in the dataset differences. As can be seen in the data section, in each event the number of athletes has increased. Despite the sharpening of the world records, most of the new athletes have a record that is, compared to the existing data, a rather weak record. In other words, new records are mainly distributed in the bottom of the dataset rather than uniformly spread. Consequently, the top records are more isolated and considered to be more 'ultimate' than those in Einmahl and Magnus (2008). Therefore the estimators will estimate an endpoint that is closer to the current record, as the data suggest that the world record is already a sort of ultimate. This does not imply that the estimated endpoints in Einmahl and Magnus (2008) are wrong, but given new data, their ultimate records are a little overestimated. The estimated endpoints of this paper can be considered more realistic to the actual but unknown frontier.

5.3 Quality of Current Records

With the estimated $(\hat{\gamma}, \hat{a}(n/k))$ using the different estimators, it is possible to assess the second research question, the quality of the current world records. Determining the Q depends, again, on the upper order statistic k. So, as done with the EVI and endpoint estimation, the Q is plotted versus k for the men's 100m, which can be seen in Figure 8.

Figure 8: estimated Q



Note: Estimated Q versus k using the M, PWM, UB and fixed Q estimator, resulting in $\hat{Q}$, indicated as the red line.

As the first coinciding stable region of the EVI and endpoint estimation is the same, the same region will be applied for the Q plot, i.e. k runs from $k = 175$ to 475. Taking averages over the region and different estimators, results in $Q = 0.02$. Transforming this to the quality yields $e^{-Q} = 0.98$. As this quantity is 'uniformly' distributed over [0,1], it can be concluded that the current world record of men's 100m is almost unbeatable.

For the other 25 events, the $\hat{Q}$ is estimated using the k values of the coinciding stable region of the EVI and endpoint estimation procedure, as performed in previous steps. The resulting quality of the world records, with corresponding athlete, record and year, are ordered and outlined in Table 7. For comparison, the estimated qualities of Einmahl and Magnus (2008) are also shown, with a * if it concerns the same world record.

Table 7: Quality of world records with corresponding athlete, record and year

Athlete	Event	Record	Year	e^{-Q}	Einmahl's e^{-Q}
Paula Radcliffe	Marathon (W)	2:15:25	2003	0.99	0.86*
Florence Griffith-Joyner	100m (W)	10.49	1988	0.99	0.86*
Javier Sotomayor	High jump (M)	2.45	1993	0.98	0.86*
Usain Bolt	100m (M)	9.58	2009	0.98	0.47
Jan Zelezny	Javelin throw (M)	98.48	1996	0.97	0.93*
David Rudisha	800m (M)	1:40.91	2012	0.96	0.74
Usain Bolt	200m (M)	19.19	2009	0.94	0.92
Marita Koch	400m (W)	47.60	1985	0.85	0.78*
Natalya Lisovskaya	Shot put (W)	22.63	1987	0.84	0.50*
Gabriele Reinsch	Discuss throw (W)	76.80	1988	0.80	0.55*
Jarmila Kratochvilova	800m (W)	1:53.28	1983	0.79	0.78*
Wayde van Niekerk	400m (M)	43.03	2016	0.72	0.67
Stefka Kostadinova	High jump (W)	2.09	1987	0.72	0.64*
Barbora Spotakova	Javelin throw (W)	72.28	2008	0.69	0.98
Hicham El Guerrouj	1500m (M)	3:26.00	1998	0.69	0.74*
Genzebe Dibaba	1500m (W)	3:50.07	2015	0.65	0.86
Jurgen Schult	Discuss throw (M)	74.08	1986	0.59	0.74*
Randy Barnes	Shot put (M)	23.12	1990	0.54	0.45*
Mike Powell	Long jump (M)	8.95	1991	0.47	0.27*
Florence Griffith-Joyner	200m (W)	21.34	1988	0.45	0.74*
Almaz Ayana	10km (W)	29:17.45	2016	0.41	0.50
Kendra Harrison	100m h (W)	12.20	2016	0.27	0.33
Dennis Kimetto	Marathon (M)	2:02:57	2014	0.25	0.95
Kenenisa Bekele	10km (M)	26:17.53	2005	0.24	0.33*
Galina Chistyakova	Long jump (W)	7.52	1998	0.20	0.30*
Aries Merritt	110m h (M)	12.80	2012	0.16	0.20

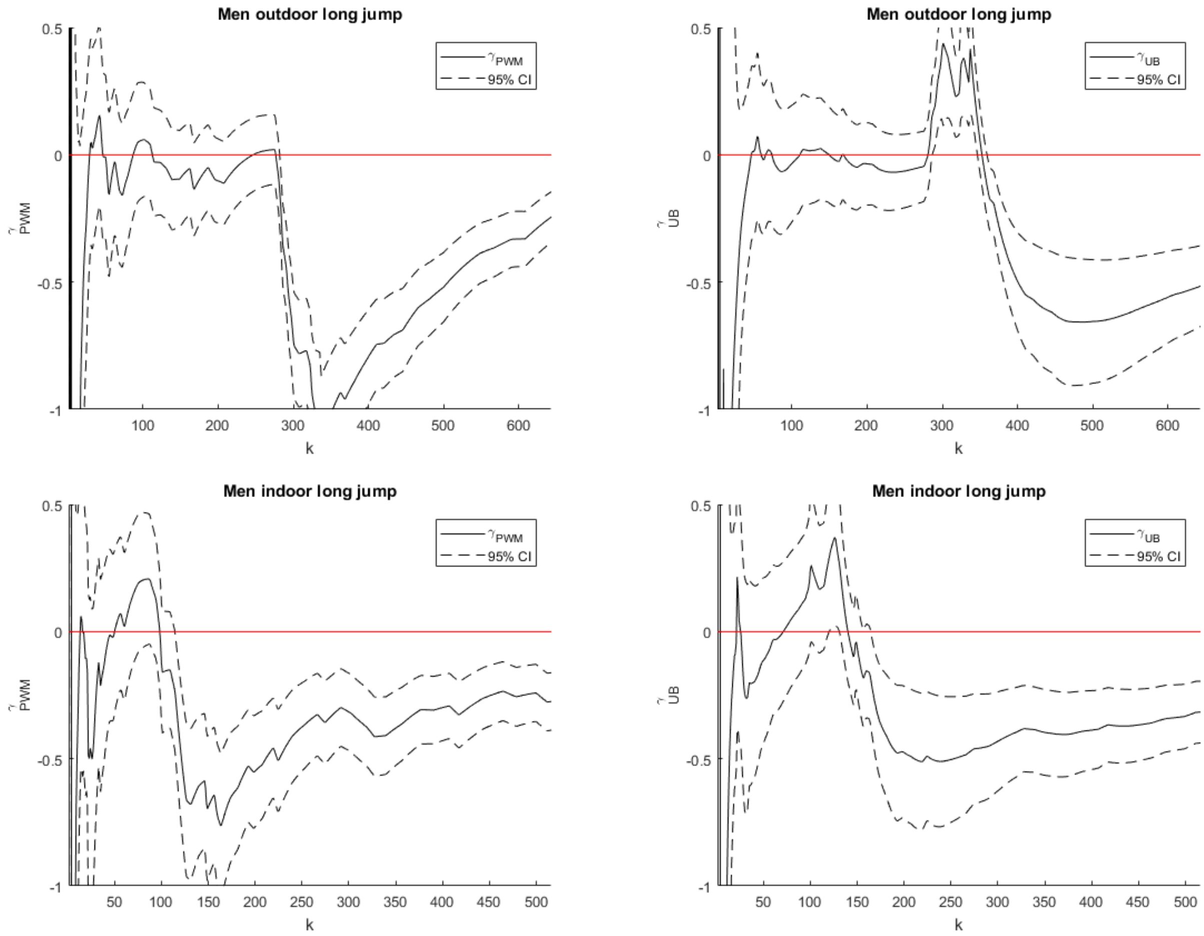
When combining Table 6 and Table 7, some interesting differences become evident. It can be possible that a current world record is of high quality but still has much room for improvement (men's 200m). In contrast, a high quality record can also lay very close to its endpoint (women's marathon). This can also be concluded for records of low quality, where very little progress is still possible (men's 110m hurdles) and where there is more to improve (men's marathon). The main cause of these differences are the estimated EVI. For events that have high (low) quality and much (little) room to improve, the $\hat{\gamma}$ is relatively high (low), and vice versa ($\hat{\gamma} = -0.12$ for men's 200m against $\hat{\gamma} = -0.27$ for men's hurdles and $\hat{\gamma} = -0.17$ for women's marathon against $\hat{\gamma} = -0.27$ for men's marathon).

In 16 of the 26 events the world records has not been improved after October 2007. Except four events, all those qualities have been improved. This may seem logical, as unsharpened records generate higher qualities over time. However, it does not apply to all 16 events. An explanation for this can be boiled down to the data, which is similar to the one for the differences in endpoints, see Section 5.2. The world records that have been sharpened, also show some variation in the qualities. Some events had a high quality but after improvement have low quality (men's marathon, from 0.95 to 0.25), whereas other qualities increase from low to high (men's 100m, from 0.47 to 0.98). The first can be attributed to the fact that the old world record was outlying to the other records, which generated an ultimate record close to the world record. However, these records have been improved multiple times, which led to a fatter tail region of the estimated distribution. Consequently, due to the corresponding low EVI, the quality of the new world record dropped. Reversing this explanation gives the cause for the shift in quality of the men's 100m.

5.4 Extension

As discussed in Section 5.1 and seen in Figure 9, the first coinciding stable region of the outdoor long jump for men renders the EVI not significantly different from 0 or a positive value, i.e. it can not be rejected that $\gamma \geq 0$. This conclusion also applies to the women's outdoor long jump data. A possible explanation is that the outdoor data are contaminated by measurement errors, which affects the estimated endpoint. To further investigate this assumption, first, the EVI of the indoor long jump for men and women is estimated using the PWM and UB estimator. As example, the estimated EVI for men's indoor long jump is plotted against k in Figure 9.

Figure 9: Estimated EVI

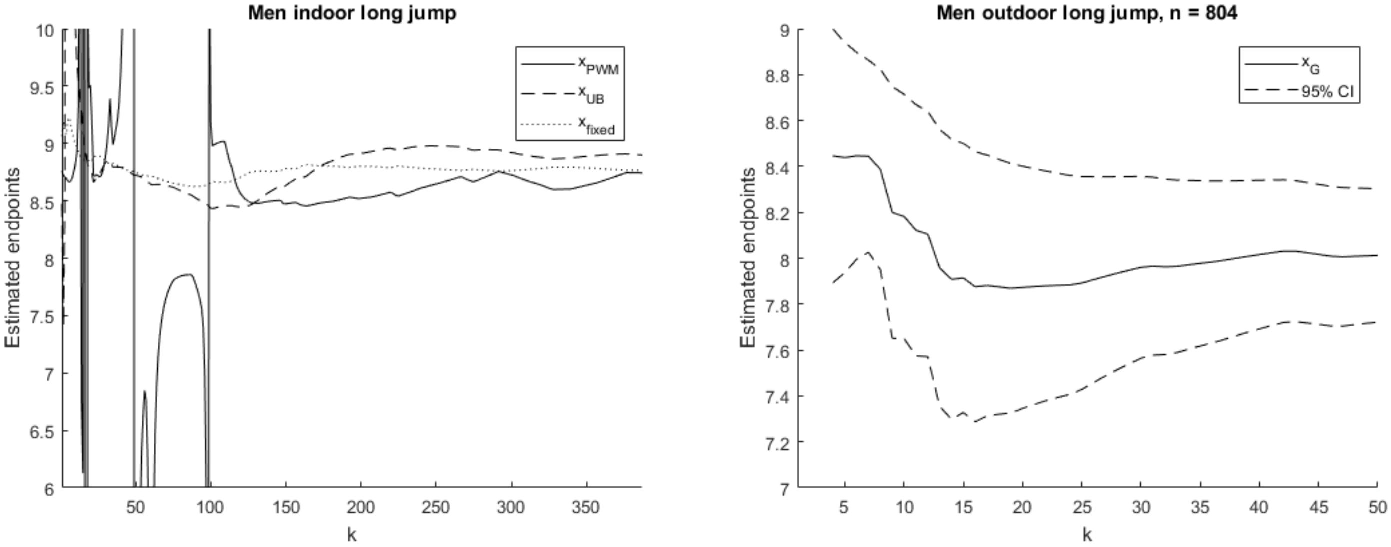


Note: Estimated EVI γ of indoor and outdoor long jump for men, with corresponding 95% confidence intervals for different values of k, using the PWM and UB estimator.

The first stable region of the PWM estimator can be found from $k = 120$ to 320. For the UB estimator, the region runs from $k = 180$ to 350. Hence, the first coinciding stable region is from $k = 180$ to 320. Taking averages over the region and estimators results in $\hat{\gamma}$. The same procedure is performed for the women's indoor long jump. Table 8 reports the $\hat{\gamma}$ with its 95% CI for men and women. Graphically, the 95% CI of the $\hat{\gamma}$ over the stable region does not contain 0 or a positive value, for both estimators. Likewise, the statistical 95% does not include 0 or positive values. Therefore, $\gamma \geq 0$ is rejected for both men and women, hence the finite endpoint exists. The different conclusions regarding $\gamma \geq 0$ between outdoor and indoor data, support the assumption that the outdoor long jump data are contaminated by measurement errors.

Based on these findings, the G estimator is used to estimate the endpoints for the outdoor long jump data, while the PWM and UB estimator are used for the indoor long jump data. Figure 10 shows the estimated endpoints versus k for men's indoor and outdoor long jump. For convenience, the estimated x_{fixed} is also plotted.

Figure 10: Estimated endpoints



Note: Estimated endpoints x_F of indoor long jump data using the PWM and UB estimator, and estimated endpoints x_F of outdoor long jump data using the G estimator with corresponding 95% confidence intervals.

For men's indoor long jump, the coinciding stable region can be found from $k = 180$ to 320, which is identical to the region of the EVI. In the simulation results in Section 4, it became evident that the G estimator performs well when k is small. Hence, the k of the G estimator is plotted from 1 to 50. The first stable region can be found from $k = 5$ to 10. To follow the procedure of Leng et al. (2017), instead of taking averages over the region, a single k value is chosen to determine the estimated endpoint. Therefore, the k with smallest RMSE in the region is chosen, which yields $k = 8$. The resulting endpoints for the indoor and outdoor long jump for both men and women are reported in Table 8.

Table 8: estimated EVI and endpoints: Long jump data

	Men				Women			
	n	k	Point	95% CI	n	k	Point	95% CI
Outdoor								
$\hat{\gamma}$	-	100-275	-0.03	[-0.19, 0.13]	-	50-150	-0.11	[-0.36, 0.13]
$\hat{x}_G$	804	8	8.39	[7.95, 8.82]	836	6	7.27	[7.05, 7.49]
$\hat{x}_G$	3000	8	8.25	[7.71, 8.79]	3000	6	7.21	[6.94, 7.49]
Indoor								
$\hat{\gamma}$	-	180-320	-0.36	[-0.63, -0.23]	-	150-200	-0.34	[-0.56, -0.12]
$\hat{x}_F$	-	180-320	8.78	[8.37, 9.19]	-	150-200	7.44	[7.05, 7.84]

It is observed that the value of $\hat{x}_G$ of the outdoor data drops a little when the sample size n increases. Moreover, the 95% CI of $\hat{x}_G$ becomes wider for larger n. However, this is a minor difference, hence $\hat{x}_G$ is stated to be insensitive to the sample size n. It becomes also evident that the estimated endpoints using G estimator for outdoor long jump are all lower than those from applying the PWM and UB estimator on indoor long jump data. Nevertheless, the 95% CI of the outdoor endpoint $\hat{x}_G$ includes the indoor endpoint $\hat{x}_F$ and vice versa, hence the the estimated endpoints still coincide in a certain extent with each other.

The main interest of this extension is if the outdoor long jump data are contaminated by measurement errors. The aforementioned results indicated indeed that this can be the case. To make conclusive remarks regarding this assumption, the results of Table 8 are compared to the actual observations in the data. The PWM and UB estimator based on the indoor long jump data suggest that the endpoints for men's and women's long jump are 8.78 and 7.44 respectively. However, when assessing the outdoor long jump data, multiple observations exceed those estimated endpoints. Specifically, for men's long jump there are four better records than 8.78, while for women's long jump five records are higher than 7.44. The only explanation for the difference must be due to a positive measurement error value, ergo the outdoor long jump data are contaminated by measurement errors. In the context of the long jump, one can think of the wind that assisted the records positively. As the dataset provides the wind speed for each record, this claim can be checked. Assessing the data, in eight of the nine aforementioned occasions the wind speeds were indeed positive. Therefore, the data support the claim that the wind can be a possible factor that contributes to measurement errors in the outdoor long jump data.

6 Conclusion

This paper's aim is to answer two questions: (1) What is the ultimate world record in a specific athletic event? and (2) How good is a current athletic world record? In total 28 sport events in athletics are investigated, 14 for both men and women. To get a representative impression of the records, only the personal bests of the top athletes are considered in the dataset. An ultimate world record can be seen as the endpoint of an underlaying distribution based on the data. Therefore, Extreme Value Theory is used for this framework. After checking if the Extreme Value Index exists and is negative, which ensures finite endpoints, several semi-parametric estimators are implemented to estimate the existing endpoints. Using simulation on fictional data with and without normally distributed errors, the accuracy and consistency of the different estimators are shown. Subsequently, the ultimate records are estimated, as shown in Table 6. Some ultimate records are close to the current world record (men's 100m), while other events have enough room for improvement (men's marathon). Concerning the quality of current records, shown in Table 7, some events have records that are of high quality (women's marathon), whereas other records are likely to be improved in near future (women's long jump). Note, results indicate that records close to the current record do not always imply that the record is of high quality (men's 200m), or vice versa (men's 110m hurdles). The events that did not satisfy a negative EVI, could have data that are contaminated by measurement errors. One event, the outdoor long jump for men and women, is chosen to further investigate this assumption. Using different estimators on outdoor and indoor data, results show that outdoor long jump data are indeed contaminated by measurement errors. Moreover, outdoor long jump data support the claim that the wind potentially contributes to the measurement errors in this particular event.

As this paper replicates the study of Einmahl and Magnus (2008), comparisons concerning the ultimate records and quality of current world records are made. The majority of the ultimate records of this paper lay closer to current records than those of Einmahl and Magnus (2008). One has to be cautious when interpreting these differences, as the estimates are determined semi-subjectively. Therefore, we suggest further investigation in how to determine the optimal upper statistic k. Nevertheless, this paper's results should be considered for further research, as the estimates are a better approximation of the true ultimate records and qualities due to the recently updated dataset and implementation of a more accurate and consistent estimator. Furthermore, the proposed methods can not only be used to assess the quality of records in multi-event sports, such as decathlon and heptathlon, but can also be applied in other scientific fields of interest, such as human life spawn and still water level.

References

Cai, J.-J., de Haan, L., & Zhou, C. (2013). Bias correction in extreme value statistics with index around zero. *Extremes*, *16*(2), 173–201.

Dekkers, A. L., Einmahl, J. H., & de Haan, L. (1989). A moment estimator for the index of an extreme-value distribution. *The Annals of Statistics*, 1833–1855.

Einmahl, J. H., & Magnus, J. R. (2008). Records in athletics through extreme-value theory. *Journal of the American Statistical Association*, *103*(484), 1382–1391.

Fereira, A., & de Haan, L. (2006). *Extreme value theory. an introduction.* Springer Series in Operations Research and Financial Engineering. NY: Springer.

Hosking, J. R., & Wallis, J. R. (1987). Parameter and quantile estimation for the generalized pareto distribution. *Technometrics*, *29*(3), 339–349.

Leng et al., X. (2017). Endpoint estimation for observations with normal measurement errors. *Submitted*.

Smith, R. L. (1987). Estimating tails of probability distributions. *The annals of Statistics*, 1174–1207.