An Extreme Value Approach to Age Limit Analysis

Christ-Roi Musonere

Master's thesis
2020:E19

Lund University

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

Master's Theses in Mathematical Sciences 2020:E19
ISSN 1404-6342

LUNFMS-3088-2020

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lth.se/

Abstract

A fundamental question in aging research is if the human lifespan has reached its maximum, as it has practical implications and affects the sustainability of our societies and health care systems. This thesis uses methods and techniques from extreme value theory to study human lifespan at extreme age, in particular the age at death of the world's oldest person (WOP) titleholders. Our study is a contribution to aging research for understanding whether extreme value analysis is useful in modeling life at extreme age and if the models used to fit the data predict the existence or lack of a limit to human life length.

Both the stationary and non-stationary models were considered for modeling the data. The stationary generalized extreme value (GEV) distribution and stationary generalized Pareto (GP) distribution which were fitted to the data indicated existence of a limit to human lifespan with an upper bound ranging between 123 and 126 years. A better fit was obtained with the non-stationary GEV model with trend in the location parameter, which took into account the trend in the individual's ages over the years. The non-stationary model does not however indicate an existence of a limit to human life length. Assuming that the linear trend in the location parameter holds in the future, a forecast probability of beating the current world's age record (122.45 years) for individuals born between 1902 and 1965 indicated more than 60% chance of observing a new age record for individuals born after 1940.

Acknowledgements

I would like to express my deepest gratitude to my supervisor Nader Tajvidi at the division of Mathematical Statistics, Lund University, for introducing me to this subject. He has been a great support through the time of research and writing of this thesis. I would also like to thank the Gerontology Research Group (GRG) for providing the data that was used in this thesis.

Contents

1 Introduction

1.1 Background

The maximum human life span has increased considerably over the past century, mainly due to factors such as the technological progress, the hygiene and biomedicine. Demographic studies in Sweden, which gathered data on the recorded maximum ages at death for both men and women, have shown that the maximum age at death has increased remarkably between the 1860s and 1990s from 101 to 108 years, respectively. The rate of increase was observed to accelerate from 0.44 years per decade in the 1960s to 1.11 years per decade the following years. Studies also showed that women tend to live 1.7 years longer than men. The results obtained in Sweden could be generally applied to other highly industrialized countries in the world (Western Europe and North America) and attest that more than 70% of the increase in maximum age at death, is due to considerable reductions in death rate above age 70.

The maximum lifespan, together with the average lifespan, are a stable characteristic of a species. However the existence or lack of a limit to lifespan has divided researchers over the years and convincing answers to this question have not been given yet. Some of the supporters of a limit for human lifespan claim that "the maximum lifespan of humans is fixed and subject to natural constraints" (Dong et al. (2016)) [1], others claim that "a biologic barrier forecast" limit further lifespan progress (Antero-Jacquemin et al (2014)) [2]. Some researchers provide arguments both for and against the existence of a limit for human lifespan as seen in papers by Aarssen and de Haan (1994) [3], Wilmoth et al. (2000) [4] and (Weon and Je(2009)) [5]. The supporters of a non-limit for human lifespan such as Gampe (2010) [6] claim that "human mortality after age 110 is flat at a level corresponding to an annual probability of death of 50%", Rootzén and Zholud. (2017) [7] give a similar conclusion as Gampe [6] that "after 110 the risk of dying is constant and is about 47% per year". Medford and Vaupel (2019) [8] investigate how likely it is that the current age record has held until now and how long it might stand in the future by studying life length records via the (inter-) record times and using methods from statistical Records Theory. Their findings indicate "a 25% chance that the record would have survived until now and around a one in five chance that it will survive until 2050". Studies [9] in aging research confirm factors that help slowing down aging such as: healthy diet, exercise, social engagement, meaning and purpose. Researchers agree that 20 to 30% of longevity is based upon genetics, which means that 70% to 80% is due to the factors named above.

The paper by Rootzén and Zholud (2017) [7], which inspired this thesis project, studies the mortality of supercentenarians (someone who has reached 110 years) using Extreme Value Theory. The Generalized Pareto (GP) distribution is used to model excess lifespan of individuals with threshold $u = 110$. Due to the method of data sampling used, the data is truncated and the GP distribution gives three different behaviours of force of mortality:

- The first case, the force of mortality tends to infinity and life is limited.
- The second case, the force of mortality is constant, life has no limit but it is short.
- The third case, the force of mortality tends to zero with age, life is unlimited.

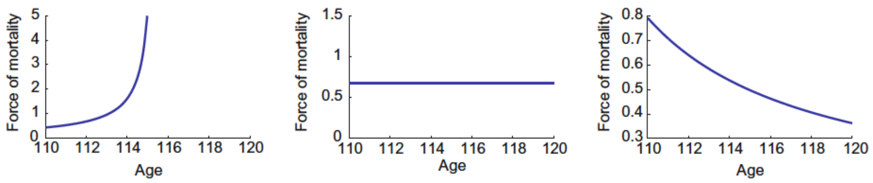

Figure 1: Force of mortality against age [7]

The paper considers the GP distribution and exponential distribution (this corresponds to the GP distribution with the shape parameter $\gamma = 0$), to be the appropriate models for the data and studies the force of mortality for supercentenarians, also the difference in survival based on various factors such as gender, birthday and region. No difference in survival based on gender, birthday and region is observed and a prediction for the next 25 years shows that maximum life length will fall between 119 and 128 years.

1.2 Objective and aims

As discussed in the previous section, the complexity of modeling extreme life length has led to many controversies and debates between researchers. A lack of adequate mathematical model and the lack of available and reliable data for extreme old age are the main challenges for this subject. The main goal of this project is to build a mathematical model for extreme life length and make predictions on maximum life length in the future. The results will contribute to earlier studies on understanding the extreme life length modeling, in particular for previous researches based on extreme value theory.

Extreme value theory provides statistical techniques and models for analyzing the distribution of the extreme parts of data such as extreme floods, the amount of large insurance losses, road safety analysis or extreme life length in this case. The thesis will apply extreme value theory models used by Rootzén and Zholud (2019) [7] such as the Generalized Pareto distribution and the Exponential distribution but also models not used by Rootzén and Zholud (2019) [7] such as the Generalized Extreme Value (GEV) model, and the non-stationary GEV model. The parameter estimates of various distributions and the inference for return levels and upper bound will be analyzed in an attempt to understand future predictions on extreme life length.

This thesis aims to model the life length of the world's oldest person (WOP) titleholders. One of the obstacles faced in the study of extreme life length is the lack of reliable data on extreme age, the data is obtained from the Gerontology Research Group (GRG) database [10], which contains validated ages of the WOP titleholders from 1955 to 2018. The last section will take into account the current living person using the survival analysis method: censoring [11].

This project will try to bring answers to the following questions:

- Can extreme life length (in this case the world's oldest person titleholders) be modeled by extreme value theory models such as the Generalized Extreme Value distribution and the Generalized Pareto distribution?
- If so, do the models used to fit the data predict the existence or lack of a limit to human life length?

The results of this thesis will be compared to the results and conclusions of papers written on this topic by Rootzén and Zholud.(2017) [7] and Medford and Vaupel (2019)) [8].

2 Extreme Value Theory

The theory on extreme value analysis presented in the sections below is based on the book [12] by Coles (2011), which provides a more detailed review on the theory of statistical modeling of extreme values.

Let $X_1, ..., X_n$ be a sequence of independent random variables with a common distribution function F. The goal is to model the statistical behaviour of $M_n = max\{X_1, ..., X_n\}$, where M_n is the maximum of the process over n time units of observation.

In case F is known, the distribution of M_n is given for all values of n:

$$\begin{aligned} Pr\{M_n \leq z\} &= Pr\{X_1 \leq z, ..., X_n \leq z\} \\ &= Pr\{X_1 \leq z\} \times ... \times Pr\{X_n \leq z\} \\ &= \{F(z)\}^n, \end{aligned}$$

In practice, F is unknown therefore we consider F^n, estimated using extreme data. For $z < z_+$ such that $F(z) = 1$, $F^n(z)$ converges to 0 as $n \to \infty$. The consequence will be that the distribution of M_n degenerates to a point mass z_+. This problem is solved by letting a linear re-normalization of the variable M_n:

$$M_n^* = \frac{M_n - b_n}{a_n}$$

where $\{a_n > 0\}$ and $\{b_n\}$ are constant such that, when well chosen, stabilize the location and scale of M_n^* as n increases. The limiting distribution of M_n^* always belongs to one of the following distributions:

1. $G(z) = \exp\left\{-\exp\left[-\left(\frac{z-b}{a}\right)\right]\right\} \quad -\infty < z < \infty;$
2. $G(z) = \begin{cases} 0, & \text{z} \leq b \\ \exp\left\{-\left(\frac{z-b}{a}\right)^{-\alpha}\right\} & \text{z} > b \end{cases}$
3. $G(z) = \begin{cases} \exp\left\{-\left[-\left(\frac{z-b}{a}\right)\right]^{\alpha}\right\} & \text{z} < b \\ 1, & \text{z} \geq b \end{cases}$

These extreme value distributions are named Gumbel, Fréchet and Weibull distributions, respectively with location and scale parameters b and $a > 0$, respectively and for distributions 2 and 3, $\alpha > 0$.

2.1 Generalized Extreme Value Distribution

The family of distributions considered in the previous section can be combined into a single distribution family: the generalized extreme value distribution (GEV) defined by Coles (2001) [12] as below:

If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$Pr\left\{(M_n - b_n)/a_n \leq z\right\} \to G(z)$$

as n→ ∞ for a non-degenerate distribution function G, then G is a member of the GEV family.

$$G(z) = \exp\left\{-\left[1 + \gamma\left(\frac{z-\mu}{\sigma}\right)\right]^{\frac{-1}{\gamma}}\right\} \tag{1}$$

defined on $\{z : 1 + \gamma(\frac{z-\mu}{\sigma}) > 0\}$, where $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \gamma < \infty$.

When $\gamma = 0$, the GEV family is considered as the limit of eq.(1) as $\gamma \to 0$. This is the Gumbel family with distribution function:

$$G(z) = \exp\left\{-\exp\left[-\left(\frac{z-\mu}{\sigma}\right)\right]\right\} \quad -\infty < z < \infty;$$

Let $X_1, X_2..., X_n$ be independent observations divided into m blocks of equal length, then the maximum values M_n in each block have a distribution, which can be approximated asymptotically by the GEV distribution.

2.1.1 Inference for the GEV distribution

There are different techniques used to estimate parameters in extreme value model, the likelihood-based technique will be used in this project because of its all-round utility and its ability to adapt to complex model-building. However, the use of likelihood methods requires certain conditions to be satisfied, otherwise the results won't be applicable: in case $\gamma > -0.5$, the estimators have the usual asymptotic properties, for for $-1 < \gamma < -0.5$, the estimators can be obtained but will not have the standard asymptotic properties, and when $\gamma < -1$, the maximum likelihood estimators are unlikely to be obtainable.

The log-likelihood for the parameters considering $Z_1, ..., Z_m$ as independent variables (block maxima) with GEV distribution and $\gamma \neq 0$ is:

$$l(\mu,\sigma,\gamma) = -m\log(\sigma) - \left(1+\frac{1}{\gamma}\right)\sum_{i=1}^{m}\log\left[1+\gamma\left(\frac{z_i-\mu}{\sigma}\right)\right] - \sum_{i=1}^{m}\left[1+\gamma\left(\frac{z_i-\mu}{\sigma}\right)\right]^{\frac{-1}{\gamma}} \tag{2}$$

given that:

$$1+\gamma\left(\frac{z_i-\mu}{\sigma}\right) > 0 \tag{3}$$

for $i = 1, ..., m$. In case eq.(3) is not satisfied, the likelihood will be zero and log-likelihood equal to ∞.

The case $\gamma = 0$ corresponds to the Gumbel distribution and the log likelihood is:

$$l(\mu,\sigma) = -m\log(\sigma) - \sum_{i=1}^{m}\left(\frac{z_i-\mu}{\sigma}\right) - \sum_{i=1}^{m}\exp\left\{-\left(\frac{z_i-\mu}{\sigma}\right)\right\}.$$

The maximum likelihoods estimates $(\hat{\mu}, \hat{\sigma}, \hat{\gamma})$ follow a multivariate normal distribution with mean (μ, σ, γ) and variance-covariance matrix is obtained from the inverse of the observed information matrix evaluated at the maximum likelihood estimate.

2.1.2 Return level

The extreme quantiles of the annual maximum distribution are estimated by inverting eq.(1) and obtain:

$$\hat{z}_p = \begin{cases} \hat{\mu} - \frac{\hat{\sigma}}{\hat{\gamma}}[1 - y_p^{-\hat{\gamma}}], & \hat{\gamma} \neq 0 \\ \hat{\mu} - \hat{\sigma} \log y_p, & \hat{\gamma} = 0 \end{cases}$$

where $y_p = -\log(1 - p)$

$$Var(\hat{z}_p) \approx \bigtriangledown z_p^T V \bigtriangledown z_p, \tag{4}$$

V is the variance-covariance matrix of $(\hat{\mu}, \hat{\sigma}, \hat{\gamma})$ and

$$\begin{aligned} \bigtriangledown z_p^T &= \left[\frac{\partial z_p}{\partial \mu}, \frac{\partial z_p}{\partial \sigma}, \frac{\partial z_p}{\partial \gamma}\right] \\ &= \left[1, -\gamma^{-1}(1 - y_p^{-\gamma}), \sigma\gamma^{-2}(1 - y_p^{-\gamma}) - \sigma\gamma^{-1}y_p^{-\gamma}\log y_p\right] \end{aligned}$$

evaluated at $(\hat{\mu}, \hat{\sigma}, \hat{\gamma})$.

Right end-point

In case $\hat{\gamma} < 0$, the upper bound of the distribution (infinite observation return period) can be estimated. The maximum likelihood estimate is :

$$\hat{z}_0 = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\gamma}}$$

The variance is obtained as in eq.(4) with

$$\bigtriangledown z_0^T = [1, -\gamma^{-1}, \sigma\gamma^{-2}]$$

evaluated at $(\hat{\mu}, \hat{\sigma}, \hat{\gamma})$. The right end-point is equal to infinity if $\hat{\gamma} \geq 0$.

The results for return level inferences should be considered with caution because the assumption of normal distribution of the maximum likelihood estimates might not give the best results.

2.1.3 Profile likelihood

The evaluation of the parameters by profile likelihood consists of keeping one parameter fixed and maximize the log-likelihood eq.(2) with respect to the other parameters. This process is repeated for a range of values of the fixed parameter.

Profile likelihood is also used for estimating the right end-point:

$$\hat{z}_0 = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\gamma}}$$

by reparameterizing the GEV distribution (eq.1) and write z_0 as a parameter of the model. This can be done by rewriting the location parameter:

$$\hat{\mu} = \hat{z}_0 + \frac{\hat{\sigma}}{\hat{\gamma}}$$

We replace $\hat{\mu}$ in the log-likelihood function and find the maximum likelihood estimate of $\hat{z}_0$. The confidence interval is evaluated by choosing the points where the log-likelihood function $l(\hat{z}_0, \sigma, \gamma)$ intercepts the function

$$l(\hat{z}_0, \sigma, \gamma) - \frac{1}{2}\chi_1^2(\alpha),$$

where $\chi_1^2(\alpha)$ is the α quantile of the χ_1^2 distribution.

2.2 Generalized Pareto distribution

The fact that the GEV model only considers the block maxima is less effective and overlooks other extreme data available. The Peak over Threshold (POT) approach includes the extreme events over a threshold u. The number of exceedances over threshold follows a Poisson distribution asymptotically.

If $X_1, X_2, ...$ is a sequence of independent and identically distributed random variables with a common distribution F, then it makes sense to consider even those observations x_i that exceed a predefined threshold. Let X represents an arbitrary term in the X_i sequence, then the following conditional probability describes the stochastic behaviour of extreme events:

$$Pr\{X > u + y | X > u\} = \frac{1 - F(u + y)}{1 - F(u)}, \quad y > 0$$

Similar to the GEV model, the distribution of threshold exceedances has an unknown parent distribution F and other approximation methods will be applied.

Let $X_1, X_2, ...$ be a sequence of independent random variables with common distribution function F, and let

$$M_n = \max\{X_1, ..., X_n\}.$$

Denote an arbitrary term in the X_i sequence by X, and suppose that F satisfies the GEV distribution, so that for large n,

$$Pr\{M_n \leq z\} \approx G(z),$$

where

$$G(z) = \exp\left\{ - \left[1 + \gamma \left(\frac{z - \mu}{\sigma}\right)\right]^{\frac{-1}{\gamma}} \right\}$$

for some μ, $\sigma > 0$, and γ. Then, for large enough u, the distribution function of $(X - u)$, conditional on $X > u$, is approximately

$$H(y) = 1 - \left(1 + \frac{\gamma y}{\tilde{\sigma}}\right)^{\frac{-1}{\gamma}} \tag{5}$$

defined on $\{y : y > 0 \text{ and } \left(1 + \frac{\gamma y}{\tilde{\sigma}}\right) > 0\}$, where

$$\tilde{\sigma} = \sigma + \gamma(u - \mu).$$

Eq.(5) defines the family of Generalized Pareto distributions. The parameters of GPD are uniquely determined by those of GEV; in particular, the shape parameter γ is the same for both distributions. The relationship between the GEV and generalized Pareto families means that γ is important in explaining the behaviour of the GPD. If $\gamma < 0$, the distribution has an upper bound of $\frac{\tilde{\sigma}}{|\gamma|}$; in case $\gamma > 0$ the distribution does not have an upper bound. When $\gamma = 0$, the distribution has no upper limit and is written as:

$$H(y) = 1 - \exp\left(-\frac{y}{\tilde{\sigma}}\right), \quad y > 0,$$

$H(y)$ is an exponential distribution with scale parameter $\tilde{\sigma}$.

2.2.1 Modeling Threshold Excesses

Just as for the block maxima approach, the choice of the right threshold can be an issue because too low a threshold will lead to bias and too high a threshold gives a high variance. Two methods will be used to choose the appropriate threshold: the first method is based on the mean of GPD and the second method applies a range of thresholds and choose the value above which the shape parameter estimate γ is approximately constant.

Mean Residual Life Plot

Let $X_1, ..., X_n$ be a sequence that generates excesses of a threshold u_0 modeled by GPD and X an arbitrary term of the sequence such that:

$$E\left(X - u_0 | X > u_0\right) = \frac{\sigma_{u0}}{1-\gamma},$$

given that $\gamma < 1$, σ_{u0} represents the scale parameter that corresponds to the excesses of u_0. If GPD holds for u_0, it should also hold for all thresholds $u > u_0$, with an appropriate change of scale parameter to σ_u. For $u > u_0$:

$$\begin{aligned} E\left(X - u | X > u\right) &= \frac{\sigma_u}{1-\gamma} \\ &= \frac{\sigma_{u0} + \gamma u}{1-\gamma} \end{aligned}$$

Then, the conditional mean is a linear function of u. The sample mean of the threshold excesses of u gives an empirical estimate of $E\left(X - u | X > u\right)$. For values of u for which GPD model is appropriate, the estimates will change linearly with u. Let $x_{(1)}, ..., x_{(n_u)}$ be observations that exceed u, the locus of points

$$\left\{ \left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} (x_{(i)} - u) \right) : u < x_{max} \right\},$$

where x_{max} is the largest of the X_i and n_u is the number of observations exceeding u, is the mean residual life plot. This plot should be linear in u above the threshold u_0 at which the GPD provides an appropriate approximation. The confidence intervals are included in the plot and the sample means follow a normal distribution.

Model Based Approach

Another method of choosing the appropriate threshold is the model based approach, which looks for the stability of parameter estimates. As described in the previous section, if the generalized Pareto distribution can model the excesses of a threshold u_0, it should also be able to model the excesses of a threshold u such that $u > u_0$. Even if the shape parameter of the two distributions are identical, the scale parameter for u, σ_u changes with u unless $\gamma = 0$:

$$\sigma_u = \sigma_{u0} + \gamma(u - u_0),$$

To solve this problem, the generalized Pareto scale parameter is reparameterized to be constant with respect to u.

$$\sigma^* = \sigma_u - \gamma u,$$

If u_0 is the right threshold for excesses to follow a GPD, the estimates of both σ^* and γ will be constant above u_0. Hence, the valid threshold is chosen by plotting both $\hat{\sigma}^*$ and $\hat{\gamma}$ against u, together with confidence intervals for each of these quantities and selecting u_0 as the lowest value of u for which the estimates remain near constant.

2.2.2 Inference for generalized Pareto distribution

Maximum likelihood is used to estimate the parameters of GPD. Let $y_1, ..., y_k$ be the sequence of threshold excesses, when $\gamma \neq 0$ the log-likelihood is:

$$l(\sigma, \gamma) = -k \log \sigma - \left(1 + \frac{1}{\gamma}\right) \sum_{i=1}^{k} \log\left(1 + \frac{\gamma y_i}{\sigma}\right),$$

where $\left(1 + \frac{\gamma y_i}{\sigma}\right) > 0$ for $= 1, ..., k$.

When $\gamma = 0$, the log-likelihood is:

$$l(\sigma) = -k \log \sigma - \sigma^{-1} \sum_{i=1}^{k} y_i.$$

Numerical methods are used for parameter estimation and the choice of parameters requires extra care otherwise the algorithm will fail.

2.2.3 Return Level

In extreme value modeling, it is usually more interesting to study the behaviour of quantiles than individual parameter estimates. Let the GPD with parameters σ and γ be a convenient model for the threshold u exceedances given by X. Given $x > u$,

$$Pr\left\{X > x | X > u\right\} = \left[1 + \gamma\left(\frac{x-u}{\sigma}\right)\right]^{\frac{-1}{\gamma}},$$

Then, we have

$$Pr\{X > x\} = \zeta_u \left[1 + \gamma\left(\frac{x-u}{\sigma}\right)\right]^{\frac{-1}{\gamma}},$$

where $\zeta_u = Pr\{X > u\}$. Let m be the period of observation, the return level x_m that is exceeded on average once every m observations is given by rearranging

$$\zeta_u \left[1 + \gamma\left(\frac{x-u}{\sigma}\right)\right]^{\frac{-1}{\gamma}} = \frac{1}{m}$$

it follows that

$$x_m = u + \frac{\sigma}{\gamma}\left[(m\zeta_u)^{\gamma} - 1\right],$$

given that m is large enough such that $x_m > u$.

Sometimes there are $n_y > 1$ observations per year, then the number of observations m is given by $m = N \times n_y$ and the N-year return level is:

$$z_N = u + \frac{\sigma}{\gamma}\left[(Nn_y\zeta_u)^{\gamma} - 1\right],$$

Right end-point

Similar the GEV distribution, the GPD has upper end-point when the shape parameter $\gamma < 0$. Here the return level estimate $\hat{z}_0$, also know as the infinite period observation, is given by

$$\hat{z}_0 = \frac{\hat{\sigma}}{|\gamma|}$$

2.3 Model Diagnostics

It is of great interest to check the accuracy of the fitted model since the reliability of the results about some aspect of the population from which the data were drawn depend on it. Due to a lack of additional sources of data, the accuracy of the model will be judged using the data from which it was derived.

2.3.1 Probability Plot

Given an ordered sample of independent observations

$$x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$$

from a population with estimated distribution function $\hat{F}$, a probability plot consists of the points

$$\left\{\left(\hat{F}\left(x_{(i)}\right), \frac{i}{n+1}\right) : i = 1, ..., n\right\}$$

The estimated distribution function $\hat{F}$ is accepted if the points of the probability plot lie close the unit diagonal.

2.3.2 Quantile plot

Given an ordered sample of independent observations

$$x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$$

from a population with estimated distribution function $\hat{F}$, a quantile plot consists of the points

$$\left\{ \left(\hat{F}^{-1} \left(\frac{i}{n+1} \right), x_{(i)} \right) : i = 1, ..., n \right\}.$$

The estimate distribution function $\hat{F}$ is an appropriate estimate of F if the quantile plot lies close to the unit diagonal.

2.4 Non-Stationary model

Non-stationary processes have characteristics that change systematically with time. For example in the case of environmental processes, the seasonal effects create different climate patterns, which raise doubts of applying a constant distribution through time. This requires the non-stationary processes to have a different treatment.

When the annual maximum observations increase linearly with time, but in other aspects the distribution (GEV distribution in this case) does not change, then it's plausible to model Z_t, the annual maximum observation as

$$Z_t \sim GEV(\mu(t), \sigma, \gamma),$$

where

$$\mu(t) = \beta_0 + \beta_1 t$$

for parameters β_0 and β_1. The latter represents the annual rate of change in annual maximum observation.The location parameter μ may be expressed in other forms, for example as a quadratic model, change-point model, etc. The non-stationary model applies to the other extreme value parameters as well. For example, the scale parameter σ can be represented as an exponential function as follows:

$$\sigma(t) = \exp\left(\beta_0 + \beta_1 t\right),$$

the use of the exponential function keeps σ positive for all values of t. Given the complexity of modeling extreme value shape parameters, they are usually kept constant.

2.4.1 Parameter inference

Maximum likelihood is preferred to other methods for its adaptability to changes in model structure. Let the annual maximum observation Z_t for $t = 1, ..., m$ have a GEV distribution:

$$Z_t \sim GEV\left(\mu(t), \sigma(t), \gamma(t)\right),$$

For $\gamma(t) \neq 0$, then the log-likelihood is:

$$l(\mu, \sigma, \gamma) = -\sum_{t=1}^{m} \left\{ \log \sigma(t) + \left(1 + \frac{1}{\gamma(t)}\right) log \left[1 + \gamma(t)\left(\frac{z_t - \mu(t)}{\sigma(t)}\right)\right] + \left[1 + \gamma(t)\left(\frac{z_t - \mu(t)}{\sigma(t)}\right)\right]^{\frac{-1}{\gamma(t)}} \right\}$$

given that

$$1 + \gamma(t)\left(\frac{z_t - \mu(t)}{\sigma(t)}\right) > 0$$

for $t = 1, ..., m$

The parameter estimates are obtained by maximizing the log-likelihood function using numerical techniques.

2.4.2 Autocorrelation Function (ACF)

The autocorrelation function evaluates the dependency of a process according to different time periods. Let's consider a process X such that the stochastic variable $X_t, X_s \in X$, where t and s are different time-points, then the autocorrelation function R, is defined by:

$$R(s,t) = \frac{E[X_t - \mu_t]E[X_s - \mu_s]}{\sigma_t \sigma_s}$$

where μ_t, μ_s and σ_t, σ_s is the mean and standard deviation of the stochastic variables, respectively. The range of the autocorrelation function lies between $[-1, 1]$, where the value of 1 indicates a perfect correlation and -1 indicates perfect anti-correlation.

2.4.3 Likelihood Ratio Test

The likelihood ratio test is used to compare two nested models say $\mathcal{M}_0$ (with parameter $\theta^{(2)}$) and $\mathcal{M}_1$ (with parameter $\theta_0 = (\theta^{(1)}, \theta^{(2)})$) with condition that the k-dimensional subvector $\theta^{(1)} = 0$. Let $l_0(\mathcal{M}_0)$ and $l_1(\mathcal{M}_1)$ be the maximized values of the log-likelihood for models $\mathcal{M}_0$ and $\mathcal{M}_1$ respectively, a test of the validity of model $\mathcal{M}_0$ relative to $\mathcal{M}_1$ at the α level of significance is to reject $\mathcal{M}_0$ in favor of $\mathcal{M}_1$ if $D = 2\{l_1(\mathcal{M}_1) - l_0(\mathcal{M}_0)\} > c_\alpha$, where c_α is the $(1-\alpha)$ quantile of the χ^2_k distribution.

2.5 Censoring

Survival analysis studies time-to-event data in diverse fields such as economics, medicine, biology, public health, etc. This analysis might present some problems depending on how the data is featured. One of those features is known as censoring and this means that for some data, the event is known to have happened in some period interval, whereas for the rest of the data, the exact lifetime is known. There are different types of censoring: right censoring is when the individual is known to be alive at a given time, for left censoring, the individual has experienced the event of interest before the start of the study and interval censoring is when the event of interest is known to occur within some interval. Only right censoring will be used in this project. Theory in this section is based on the book [11] by Klein and Moeschberger (1997), which is recommended for a more detailed review of the theory on survival analysis.

2.5.1 Right censoring

For a given individual in the experiment, there is a lifetime X and a fixed censoring time C_r. When $X \leq C_r$, the exact lifetime is known otherwise the individual is a survivor and the event time is censored at C_r. A pair of random variables (T, δ) is used to represent the data from the study. When $\delta = 1$, the lifetime X is observed and equals T, when $\delta = 0$, the lifetime is censored and C_r equals T.

In this project, individuals enter the study at different time and the censoring time is fixed by the investigator. The data will be represented by a Lexis diagram where the calendar time is on the horizontal axis and lifetime is given by 45° line.

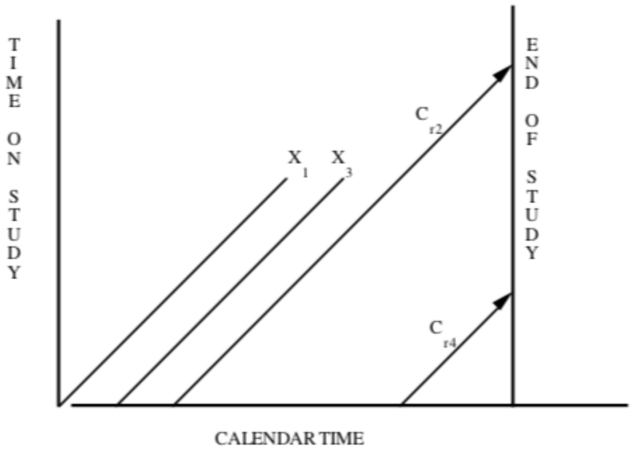

Figure 2: Lexis diagram for right censoring [11]

In the figure above, the life length of patients 1 and 3 are known ($\delta = 1$), whereas patients 2 and 4 are censored given that they are still alive at the end of the experiment ($\delta = 0$).

2.5.2 Likelihood construction

Let's assume that the lifetimes and censoring times are independent, likelihood construction requires extra care on what information each observation provides. If an observation has an exact event time, it provides information that the event occurs at this time, which can be approximated to the density function of X at this time. In case of right censored observation, the lifetime is larger than the time of study so the information is the survival function evaluated at the on study time.

$$L \propto \prod_{i \in D} f(x_i) \prod_{i \in R} S(C_r)$$

D represents the set of death times, R the set of right censored observations, $S(C_r)$ is the survival function.

A detailed construction of the likelihood for right censoring is given below:

For $\delta = 0$,

$$\begin{aligned} Pr[T, \delta = 0] &= Pr[T = C_r | \delta = 0] Pr[\delta = 0] = Pr(\delta = 0) \\ &= Pr(X > C_r) = S(C_r). \end{aligned}$$

For $\delta = 1$,

$$\begin{aligned} Pr[T, \delta = 1] &= Pr[T = X | \delta = 1] Pr[\delta = 1] \\ &= Pr(X = T | X \leq C_r) Pr(X \leq C_r) \\ &= \left[\frac{f(t)}{1 - S(C_r)} \right] [1 - S(C_r)] = f(t). \end{aligned}$$

Combining the two expressions into one gives:

$$P_r(t, \delta) = [f(t)]^{\delta} [S(t)]^{1-\delta}.$$

With the random sample of pairs (T_i, δ_i), $i = 1, ..., n$, the likelihood function is

$$L = \prod_{i=1}^{n} Pr[t_i, \delta_i] = \prod_{i=1}^{n} [f(t_i)]^{\delta_i} [S(t_i)]^{1-\delta_i} \tag{6}$$

The GEV distribution will be considered in this project and the parameters are estimated by maximizing the log-likelihood function of eq.(6) using numerical techniques.

3 Data and Methods

The first part of this section describes how the data was collected and which parts were used in the study. The second part contains the results obtained from different analyses, given in this order: Generalized extreme value distribution, Gumbel distribution, Peak over Threshold approach, Exponential distribution, non-stationary model, the probability of observing a new age record and the non-stationary model with censoring.

3.1 Data collection

The world's oldest person (WOP) titleholders was obtained from Gerontology Research Group (GRG) database[1], which is a global group of international researchers specialized in gerontology (study of the social, cultural, psychological, cognitive and biological aspects of ageing). The GRG tracks and verifies supercentenarians or individuals who are at least 110 years old. The information about a person recorded by the WOP titleholders database consists of their name, gender, race, birthplace, birthday, age at death, year of death, when oldest, age at accession, length of reign and the death place. The record contains information about individuals since 1955 and the last update was done on July 31,2018.

The data used in this project for the extreme values modeling includes the individual's age (age+days, since given), year of death and later in the project the year of birth. From 1955 to 2018, there have been 65 WOP titleholders, the current titleholder included. For data modeling, only deceased former titleholders are considered and the one who is still alive will be included in survival analysis using censoring. Data modeling and analysis was done using MATLAB and *RStudio* in particular the package *in2extRemes.*

[1] https://grg.org/Adams/C.HTM

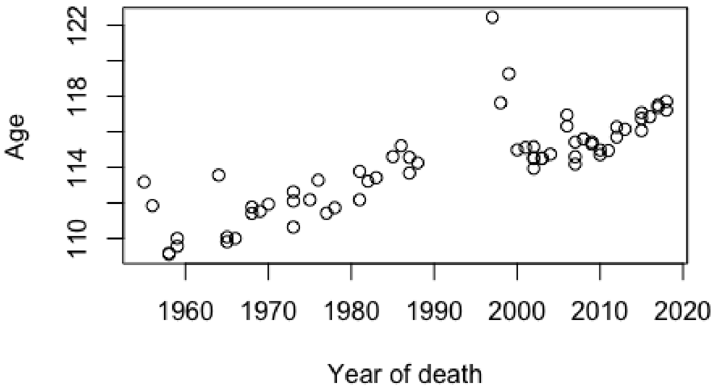

Figure 3: Age against Year of death

The plot of age against year of death (Figure 3) shows clearly that age at death of the WOP titleholders increased with time. The first step in the data analysis is to check for dependency in the time series data using the autocorrelation function. The size of the lag between elements of the data is represented on the x-axis. Lag 0 is equal to 1 since it indicates the correlation of an element with itself. When the spikes are higher than the dashed horizontal lines, that means that the correlation is high.

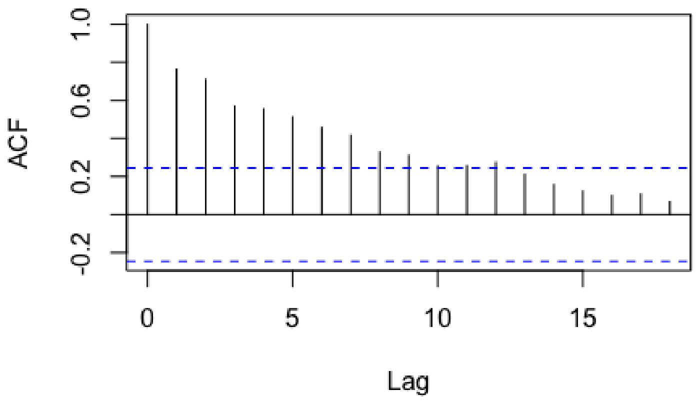

Figure 4: Autocorrelation function of age of WOP titleholders

As it can be seen on Figure 4, the autocorrelation function indicates that there exists dependency between the ages at death of individuals. This dependency is observed up to lag 10. However, it is hard to imagine dependency between the ages at death of the world's oldest person titleholders, this could be explained by a trend possibly created by a long-term change in ages at death of individuals. This issue will be considered later when the non-stationary model is used. The autocorrelation function will be plotted again after removing the possible trend in the data.

3.2 Extreme Value Analysis

In this section, three extreme value distributions are fit to the data: Generalized extreme value (GEV) distribution, Generalized Pareto distribution and the non-stationary GEV distribution.

3.2.1 Generalized Extreme Value distribution

The WOP titleholders data was fit to the GEV distribution and the first step is to check the goodness-of-fit using the diagnostic plots. Figure 5 shows that both the probability plot and the quantile plot confirm the validity of the fitted model, each set of plotted points is near-linear. Also, the density plot agrees with the histogram of the data.

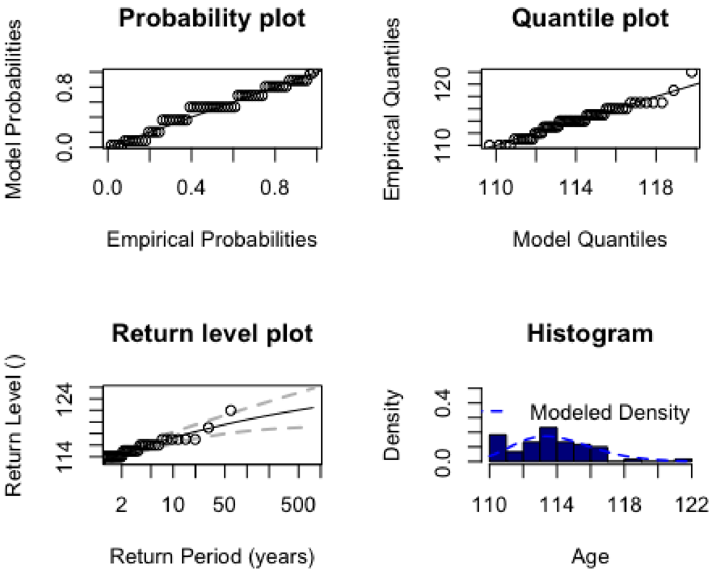

Figure 5: Diagnostic plots for GEV fit to the WOP data

Parameter estimates

The GEV parameter estimates $(\hat{\mu}, \hat{\sigma}, \hat{\gamma})$ are obtained by the maximum likelihood method and the profile log-likelihood is used to find the 95% confidence intervals of the parameter estimates.

Table 1: Parameter estimates with confidence interval

	95% Lower Bound	Point Estimate	95% Upper Bound
$\hat{\mu}$	112.5	113.141	113.8
$\hat{\sigma}$	2.116	2.5127	3.056
$\hat{\gamma}$	-0.2934	-0.1936	-0.0421

The shape parameter is of most importance in this part as it indicates the behaviour of the tail of distribution. We observe that $\hat{\gamma}$ is negative, which implies the existence of a finite upper bound of the distribution. The confidence interval of the shape parameter confirms the existence of an upper end-point of the distribution.

Right end-point

Given that the shape parameter estimate is negative, the distribution has an upper-bound. Table 2 and Figure 6 show the maximum likelihood estimate and the 95% confidence intervals for the right end-point of the distribution based on profile log-likelihood.

Table 2: Right end-point estimate with confidence interval

95% Lower Bound	Point estimate	95% Upper Bound
125.485	126.120	126.745

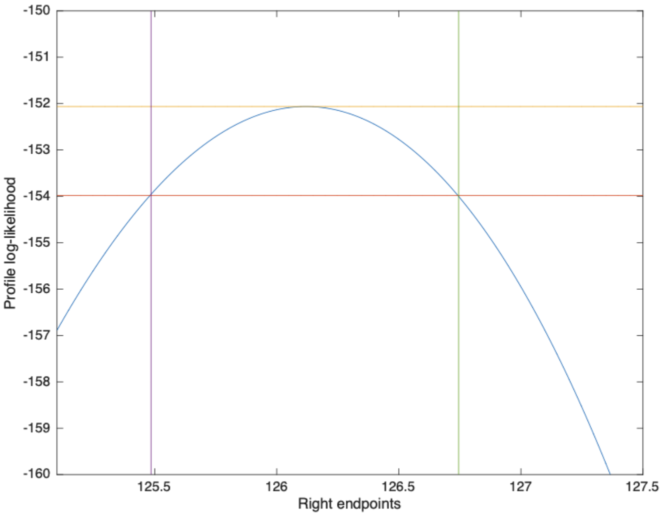

Figure 6: Right end-point of the distribution based on profile likelihood

Reducing the confidence intervals to 99% does not change significantly the results as it can be seen in Figure 7.

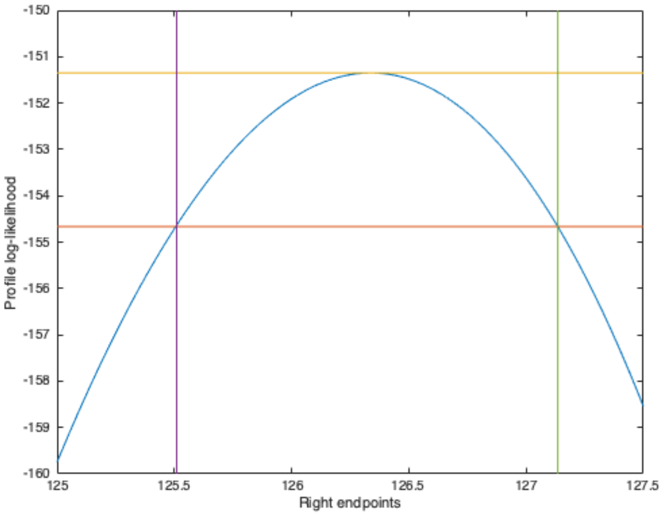

Figure 7: Right end-point of the distribution based on profile likelihood

3.2.2 Gumbel distribution

The Gumbel distribution considers the shape parameter $\gamma = 0$ and has the distribution as follows:

$$G(z) = \exp\left\{-\exp\left[-\left(\frac{z-\mu}{\sigma}\right)\right]\right\} \qquad -\infty < z < \infty;$$

The diagnostic plots in Figure 8 produce a set of plotted points that lie close to the unit diagonal, even if the quantile plot deviates for higher values.

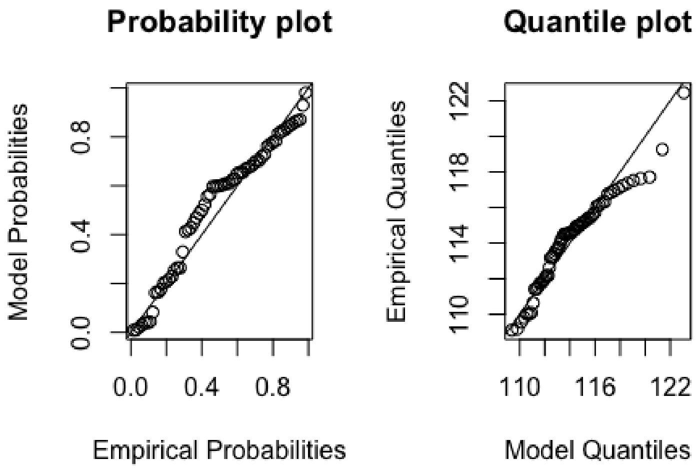

Figure 8: Diagnostic plots for the Gumbel distribution fit to the age of WOP titleholders

Parameter estimates

The parameters estimated by maximum likelihood, together with the 95% confidence interval are represented in Table 3. These estimates are not very different from GEV parameter estimates.

Table 3: Parameter estimates with confidence interval

	95% Lower Bound	Point estimate	95% Upper Bound
$\hat{\mu}$	112.243	112.879	113.515
$\hat{\sigma}$	2.003	2.449	2.894

Likelihood Ratio Test

The likelihood ratio test is used to compare the Gumbel distribution (null hypothesis) to the GEV distribution (alternative hypothesis).

$$H_0 : \gamma = 0$$
$$H_1 : \gamma \neq 0$$

The deviance statistic is defined as:

$$D = 2(l_1(M_1) - l_0(M_0)) > c_\alpha$$

where $l_0(M_0)$ is the maximized log-likelihood of the Gumbel distribution and $l_1(M_1)$ is the maximized log-likelihood of the GEV distribution with $\gamma \neq 0$. The result follows below:

$$D = 2(-152.0628 + 154.8771) = 5.6286 > 3.841$$

The likelihood ratio test statistic for the reduction to the Gumbel distribution is big compared to the χ_1^2 distribution. Thus, the GEV distribution is adequate for these data and the shape parameter, $\gamma \neq 0$.

3.2.3 Peak Over Threshold (POT)

The POT model fits the threshold excesses data to the Generalized Pareto distribution. In this section, the appropriate threshold is found based on the Mean Residual Life Plot and Model Based Approach, then the GPD parameters and right end-point are estimated.

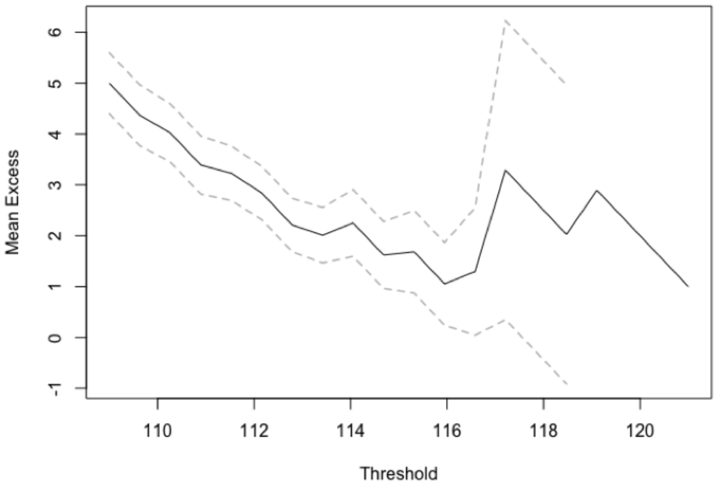

Figure 9: Mean residual life plot for the age of WOP titleholders

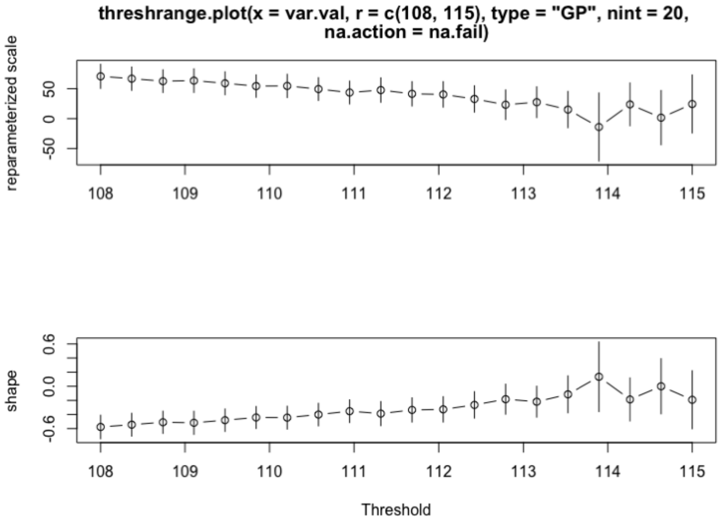

Figure 10: Parameter estimates against threshold for the age of WOP titleholders

The mean residual life plot in Figure 9 is not easy to interpret for threshold selection therefore, we will refer to the model based approach. After examining Figure 10, the threshold $u = 110$ is chosen because it is where the parameters appear to be near constant.

After choosing the appropriate threshold, the next step is to look at the goodness-of-fit of the distribution using the diagnostic plots. Both the probability and quantile plots are near-linear, the quantile plot slightly deviates for higher ages but in general, the threshold excesses can be fitted to the model.

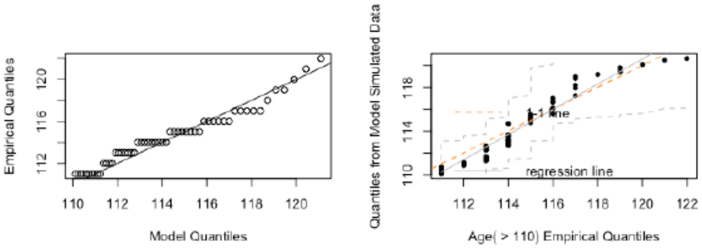

Figure 11: Diagnostic plots for the GPD fit to the WOP data

Parameter estimates

Table 4: Parameter estimates with confidence interval

	95% Lower Bound	Point estimate	95% Upper Bound
$\hat{\sigma}$	6.0370	6.2565	6.6770
$\hat{\gamma}$	-0.490	-0.4783	-0.3832

Similar to GEV model, the shape parameter estimate $\hat{\gamma}$ is of most importance for the Generalized Pareto distribution. Table 4 shows that $\hat{\gamma}$ and its confidence intervals are negative, which supports the existence of an upper bound point of the distribution.

Right end-point

Now knowing that the $\hat{\gamma}$ is negative, it is of great interest to make inference on the right end-point of the distribution. This is the observation that will be reached with $p = 0$, also known as the infinite observation return period. The estimate and 95% confidence interval based on profile likelihood are presented in Table 5 and Figure 12 below.

Table 5: Right end-point estimate with confidence interval

95% Lower Bound	Point estimate	95% Upper Bound
122.50	123.08	124.85

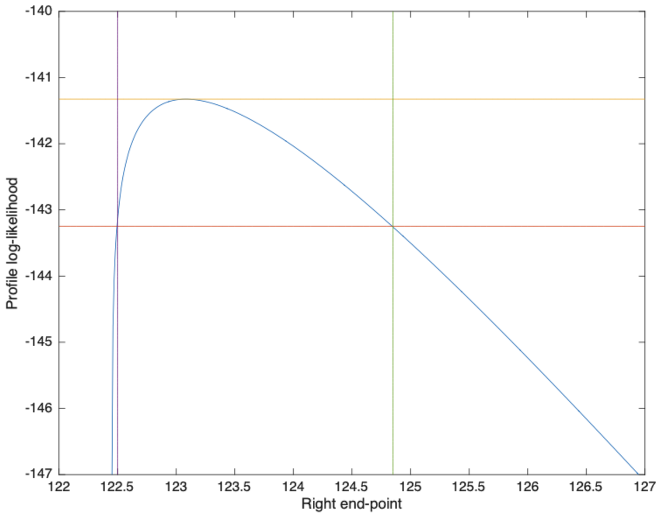

Figure 12: Right end-point of the distribution based on profile likelihood

The upper-bound estimate $\hat{z}_0 = 123.08$ is surprisingly low compared to the GEV right end-point estimate and the current validated age record, 122.45.

3.2.4 Exponential distribution

The exponential distribution is a special case of the Generalized Pareto (GP) distribution, it corresponds to the shape parameter $\gamma = 0$. To study this distribution, the diagnostic plots and the parameter estimates with 95% confidence interval are analyzed, the threshold is kept equal to $u = 110$. The GP distribution takes the form below:

$$H(y) = 1 - \exp\left(-\frac{y}{\sigma}\right), \quad y > 0,$$

where σ is the scale parameter of the distribution.

The diagnostic plots in Figure 13 show that both the probability plot and the quantile plot fail to confirm the validity of the fitted model since the set of plotted points does not lie close to the unit diagonal.

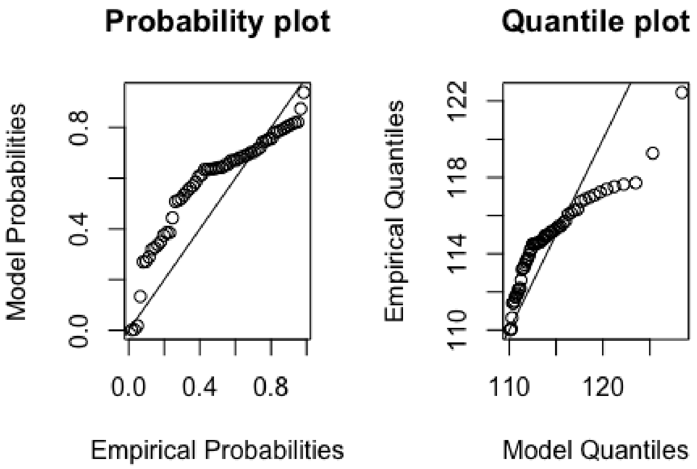

Figure 13: Diagnostic plots for the Exponential distribution fit to the WOP data

The scale parameter estimate with 95% confidence intervals are presented in Table 6:

Table 6: Parameter estimates with confidence interval

	95% Lower Bound	Point estimate	95% Upper Bound
$\hat{\sigma}$	3.346	4.479	5.613

Likelihood Ratio Test

The log-likelihood ratio test compares two nested models against one another. The exponential distribution is the null hypothesis and the GP distribution with $\gamma \neq 0$ is the alternative hypothesis.

$$H_0 : \gamma = 0$$
$$H_1 : \gamma \neq 0$$

The deviance statistic for comparing the two models is:

$$D = 2(l_1(M_1) - l_0(M_0)) > c_\alpha$$

where $l_0(M_0)$ is the log-likelihood maximum of exponential distribution and $l_1(M_1)$ is the log-likelihood maximum of the GP distribution with $\gamma \neq 0$. The result follows below:

$$D = 2(-141.3213 + 149.9664) = 17.2902 > 3.841$$

The deviance statistic value is greater than chi-square distribution, therefore the null hypothesis is rejected at the 0.05 significance level.

Looking at the GP shape parameter estimate and confidence intervals in table 4, it is not likely that the shape parameter will take the value of zero since the estimated values are not close to zero. It can also be seen by comparing Figure 11 and Figure 13 that the better fit for the probability plot and quantile plot is obtained when $\gamma \neq 0$ (Figure 11).

3.2.5 Non-Stationary Model

Returning to Figure 3, there seems to be visual evidence of a trend in the age data series and the strength of this evidence needs to be tested using the non-stationary modeling. This time the year of birth of the WOP titleholders will be used instead of year of death. In the data used for the project, the year of birth ranges between 1842 and 1901 and there appears to be a trend here as well (Figure 14).

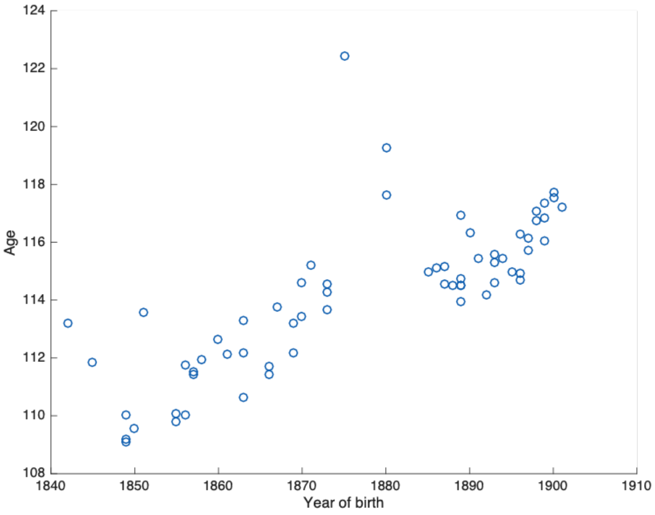

Figure 14: Age against Year of birth

The GEV distribution will be used in this section with trend in the location parameter μ. The diagnostic plots in Figure 15 show that the probability plot and quantile plot confirm the validity of the fitted model. The fitted density plot agrees with the empirical density of the observed data.

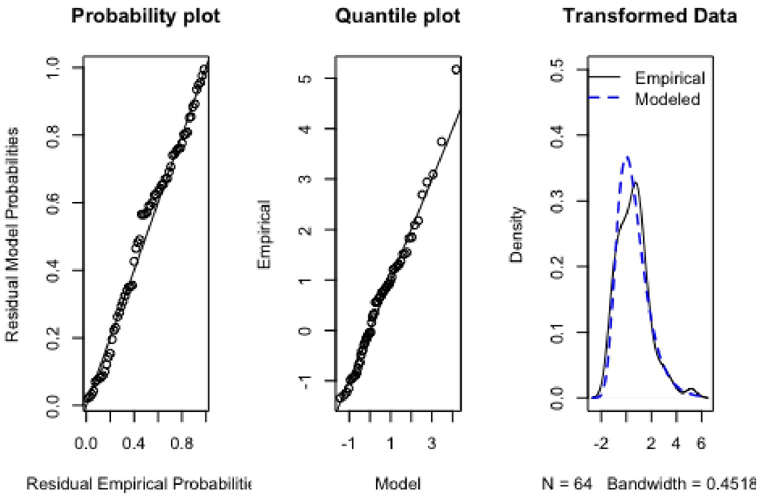

Figure 15: Diagnostic plots in linear trend GEV model of the WOP data

Parameter estimates

Let Z_t represent the age of the WOP titleholders . Z_t increases with time t (year) and is modeled as:

$$Z_t \sim GEV(\mu(t), \sigma, \gamma),$$

where

$$\mu(t) = \beta_0 + \beta_1 t$$

The location parameter $\mu(t)$ is represented as a linear function where the parameter β_1 corresponds to the annual increase in age of the WOP titleholders. t represents the transformed year of birth of the individuals, it is transformed for numerical stability reasons such that $t \in [0, 1]$ in the following way:

$$t = \frac{Y_i - Y_1}{Y_{end} - Y_1}, \quad i = 1, ..., 64$$

where Y_i represents the year of birth of individual i, Y_1 is the earliest year of birth in the data, corresponding to 1842 and Y_{end} is the latest year of birth in the data, corresponding to 1901.

Table 7: Parameter estimates with standard errors

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}$	$\hat{\gamma}$
Parameter	108.9922	0.1247	0.9123	0.2397
estimates	(0.2330)	(0.0053)	(0.1089)	(0.1209)

Table 7 shows that the estimated increase in age at death, $\hat{\beta}_1$ for an individual is $45.55 \approx 46$ days per year. It is important to notice that the shape parameter estimate $\hat{\gamma}$ is now positive, which changes the assumption on the distribution of the age of the WOP titleholders. Indeed, when $\hat{\gamma}$ is positive, the upper-bound is equal to infinity and this has a major impact on the conclusion of life length analysis.

In Figure 16 the linear trend in μ is plotted relative to the original data. The quality of the plot shows that the model fits well the data.

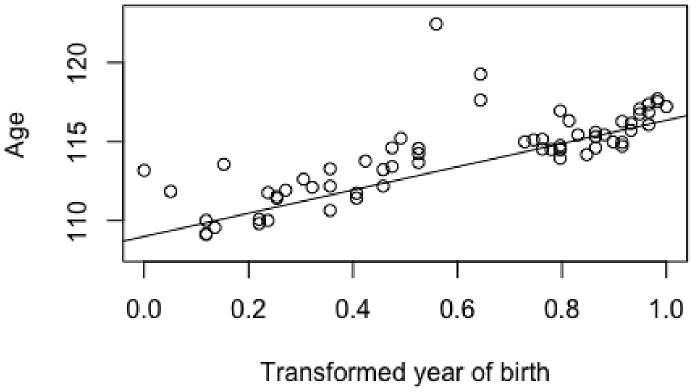

Figure 16: Fitted estimates for μ in linear trend GEV model of the WOP data

Likelihood Ratio Test

The log-likelihood ratio test compares two nested models against one another. The stationary GEV model is the null hypothesis and the non-stationary GEV model is the alternative hypothesis.

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

The deviance statistic for comparing the two models is:

$$D = 2(l_1(M_1) - l_0(M_0)) > c_\alpha$$

where $l_0(M_0)$ is the log-likelihood maximum of the stationary GEV model and $l_1(M_1)$ is the log-likelihood maximum of the non-stationary GEV model. The result is shown below:

$$D = 2(-103.7968 + 152.0628) = 96.532 > 3.841$$

The deviance statistic value is highly relevant compared with a chi-square distribution, hence providing strong evidence in favor of the non-stationary GEV model (alternative hypothesis). The result supports the existence of a linear trend in the location parameter μ.

Different models are tested to examine if there is a better choice than the linear trend in μ of GEV model. The first model considers a quadratic trend in μ, defined as:

$$\mu(t) = \beta_0 + \beta_1 t + \beta_2 t^2,$$

The second model considers a trend in the scale parameter of the GEV distribution:

$$\sigma(t) = \exp(\beta_0 + \beta_1 t),$$

the exponential function is used to make sure that the values of σ are positive. The maximum log-likelihoods and parameter estimates, with standard errors in brackets, of different models for μ and σ in GEV model are shown in Table 8

Table 8: Maximum log-likelihood, Parameter estimates with standard errors

Model	Log-likelihood	$\hat{\beta}$	$\hat{\sigma}$	$\hat{\gamma}$
Constant	-152.06	113.03 (0.31)	2.15 (0.21)	-0.14 (0.07)
Linear	-103.8	(108.99,0.12) (0.23,0.0053)	0.91(0.10)	0.24 (0.12)
Quadratic	-103.47	(108.65,0.16,-0.028)	0.90(0.11)	0.23 (0.12)
Linear (location and scale)	-103.21	(109.2,0.12) (0.33,0.008)	(1.19,-0.0073)(0.3,0.007)	0.193 (0.12)

The deviance statistic for comparing the linear model and quadratic model is:

$$D = 2(-103.4657 + 103.7968) = 0.662 < 3.841$$

The null hypothesis can not be rejected, the linear trend in the location parameter for GEV model fits better the data. The deviance statistic for comparing the linear model and the model with trend in location and scale parameter is:

$$D = 2(-103.2117 + 103.7968) = 0.279 < 3.841$$

The null hypothesis can not be rejected, hence the GEV model with linear trend in the location parameter is chosen as the appropriate model that fits better the data.

Autocorrelation function after trend removal

Having determined that the GEV model with trend in location parameter is the more adequate model that describes well the data, the autocorrelation function from section 3.1 is considered again, this time after removing the trend in location parameter. Figure 17 shows the plot of the data when trend is removed.

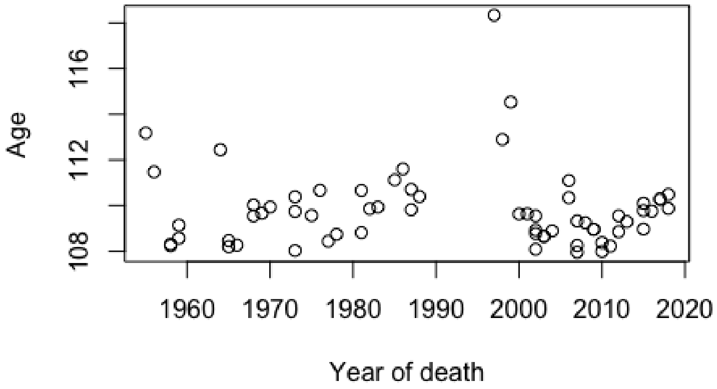

Figure 17: Age against year of death after removing trend in location parameter

The autocorrelation function after removing the trend in the location parameter (Figure 18) is much more improved compared to Figure 4 and indicates that data is nearly independent (lag 2 and 3 slightly above the dashed horizontal lines).

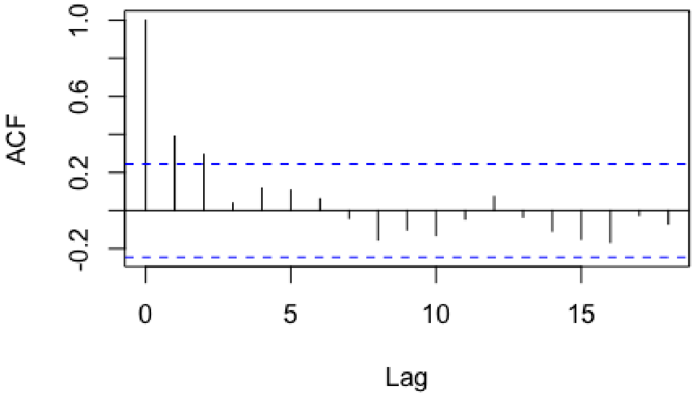

Figure 18: Autocorrelation function after removing trend in location parameter

Return Level

The GEV distribution with trend in location parameter μ was chosen as the model that fits well the data, so from now on it will be the model used for the rest of the project.

Given that the shape parameter estimate $\hat{\mu}$ is positive, the upper bound point is equal to infinity. Return levels, which are values that are exceeded on average once every m years observations, are estimated below. The equation for estimating return levels for GEV as seen earlier is:

$$\hat{z}_p = \begin{cases} \hat{\mu} - \frac{\hat{\sigma}}{\hat{\gamma}}[1 - y_p^{-\hat{\gamma}}], & \hat{\gamma} \neq 0 \\ \hat{\mu} - \hat{\sigma} \log y_p, & \hat{\gamma} = 0 \end{cases}$$

where

$$y_p = -\log(1-p)$$

This time since the location parameter μ is time-varying $\hat{\mu}(t) = \hat{\beta}_0 + \hat{\beta}_1 t$, there will be time-varying return levels $\hat{z}_p(t)$ as well. Figure 19 shows the 10,50 and 100-year return levels, the index corresponds to the year of birth of different individuals (0 and 64 correspond to years 1842 and 1901 respectively). For the next 10, 50 and 100 years, age is expected to increase with time, however only in the next 100 years, the return level is expected to reach new records (above 122.45 years). For the 10-year return level, age ranges between 111.71 and 119.07 years, for 50-year return level, age ranges between 114.88 and 122.24 years, for the 100-year return level, age ranges between 116.65 and 124.01 years. Looking at Figure 19, the 50 and 100-year return levels lie above the current observed ages for the WOP titleholders, which indicates that Jeanne Calment, the oldest validated person, could be treated as an outlier.

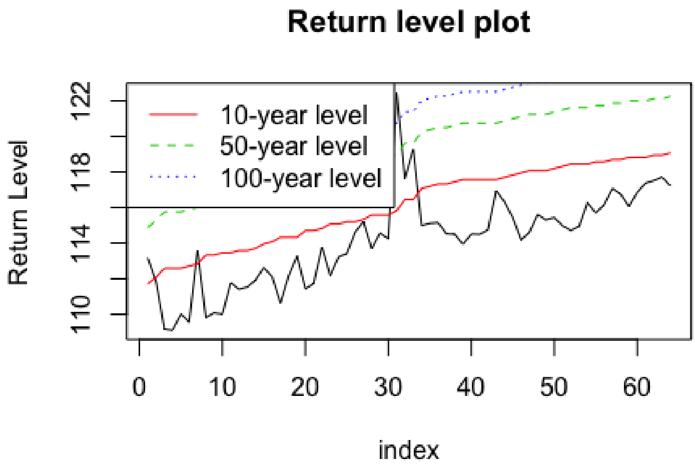

Figure 19: 10-, 50- and 100-year return level plot of the WOP data

Probability of beating Jeanne Calment's record

The return level estimate is assumed to vary for individuals born between 1902 and 1965. This forecast assumes that the linear trend in the location parameter still holds as we move beyond the range of the data. Therefore the probability of observing a new record from individuals in this group is:

$$P(X \geq x_{rec}) = 1 - \left(\exp\left\{ - \left[1 + \gamma \left(\frac{x_{rec} - \mu(t)}{\sigma} \right) \right]^{\frac{-1}{\gamma}} \right\} \right)$$

where $\mu(t) = \beta_0 + \beta_1 t$ and x_{rec} is the record age (122.45).

From Figure 20, as it could be expected the probability increases with time but it is very low for individuals born earlier in the 20th century. A significant increase is observed for the individuals born after 1940 with probabilities ranging between 60% and 99%. This means that it is expected to see individuals alive now with a new age record if the linear trend in μ of GEV model still holds for this group.

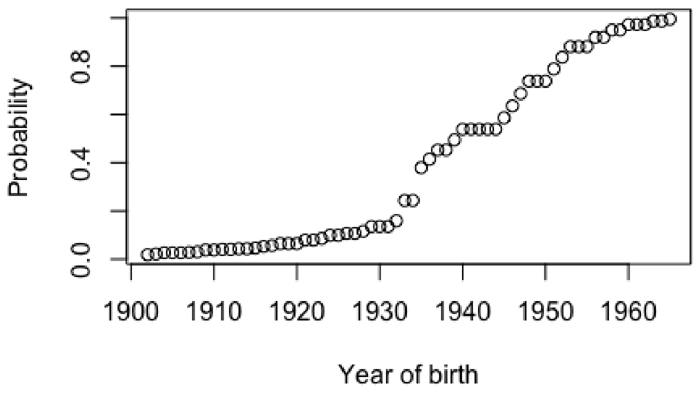

Figure 20: Probability for new age record for individuals born between 1902 and 1965

3.2.6 Non-stationary model with censoring

In this section, the current world's oldest person is included in the data and survival analysis is used. This last observation is treated differently because the individual, on the opposite from the others, has not died yet as of July 2018, date of the last update. The deceased individuals in the study will have the lifetime X and the living individual will have the censored time C_r, which is the time she has on July 2018. The GEV with linear trend in the location parameter μ model, chosen as the most appropriate model for the data, is used. The likelihood function for this model is given below:

$$L = \prod_{i=1}^{n} Pr[t_i, \delta_i] = \prod_{i=1}^{n} [f(t_i)]^{\delta_i} [S(t_i)]^{1-\delta_i}$$

$[T, \delta]$ is used to represent the data from the study. When $\delta = 1$, the lifetime X is observed and equals T, when $\delta = 0$, the lifetime is censored and C_r equals T. $f(t_i)$ is the density function of the GEV distribution, $S(t_i)$ is the survival function defined as:

$$S(C_r) = Pr(X > C_r)$$

The log-likelihood is maximized using numerical techniques.

Parameter estimates

Table 9: Parameter estimates with standard errors

Parameter estimates	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}$	$\hat{\gamma}$
	109.071 (0.0042)	0.126 (0.0063)	0.847 (0.0025)	0.276 (0.0013)

In Table 9, it is seen that the shape parameter estimate $\hat{\gamma}$ is still positive, confirming the existence of an infinite upper bound of the distribution. $\hat{\beta}_1$ is equal to 0.126, which corresponds to an estimated annual increase of $45.91 \approx 46$ days for the WOP titleholders.

4 Conclusion and Discussion

This project investigates different models of the extreme value theory for describing human life length at extreme age, in this case the age at death of the world's oldest person titleholders. The use of validated data (from GRG database) is useful in order to produce reliable results. The data evaluated by the autocorrelation function (ACF) indicated correlation between the data (Figure 4) due to the trend in the age of the individuals created over the years, but the data was independent after the removal of the linear trend in the location parameter (Figure 18).

The GEV distribution was a good fit for the data as the probability plot and quantile plot followed a unit diagonal (Figure 5). The obtained shape parameter estimate was negative indicating the existence of an upper bound to the distribution and this was confirmed by the confidence intervals, which were all negative. The right-end point of the distribution, which is the highest age that will be exceeded with probability, $p = 0$, was estimated by profile log-likelihood and was about 126.12 years (Figure 2), this value is around four years greater than the highest documented human age, 122.45. The GEV distribution was also a better fit for the data compared to the Gumbel distribution ($\gamma = 0$). This was confirmed by the standard diagnostic graphical checks, with a deviation for higher values for the Gumbel distribution (Figure 8) and the likelihood-ratio test. The GP distribution produced a good fit for the data (Figure 11) both with the probability plot and quantile plot (with a small deviation for higher values). Just like the GEV model, the GP distribution produced a negative shape parameter estimate and negative confidence intervals, which indicates the existence of a right end-point. The estimated upper bound is 123.08 with confidence interval $(122.5, 124.85)$. This estimate compared to the GEV model estimate is surprisingly small, also given that the highest documented human age is of 122.45. The exponential distribution ($\gamma = 0$) proved not to be a better model than the GP distribution ($\gamma \neq 0$) by the likelihood ratio test but also as it can be seen on the diagnostic plots (Figure 13 and Figure 11), the GP distribution produced a better fit.

The non-stationary GEV model with trend in the location parameter provided the best fit for the data. The diagnostic plots were better than the stationary GEV distribution and the GP distribution, with set of plotted points close the unit diagonal. The likelihood ratio-test between the two models provided strong evidence in favor of the non-stationary GEV model. Also when the linear trend in μ was plotted relative to the original data, the plot was of good quality except for the abnormally high age of Jeanne Calment (Figure 16). Larger non-stationary GEV models (quadratic trend in the location parameter, trend in scale parameter and both linear trend in location parameter and scale parameter) were not an improvement compared to the GEV model with linear trend in the location parameter. Conclusively the world's oldest person titleholders age is best described by the non-stationary model with linear trend in the location parameter. Surprisingly the shape parameter of the model and its confidence intervals are positive indicating that age distribution has no upper bound, this goes against earlier results found by stationary GEV and GP distributions where the distribution has an upper bound. For individuals born between 1842 and 1901, age has increased linearly with an increase of approximately 46 days per year (Table 7). The return level plot is expected to increase with time for the 10,50 and 100-year levels (Figure 19). Also assuming that linearity in the location parameter holds for individuals born between 1902 and 1965, the probability of observing a new age record increases significantly for those born in 1940 and later. This means that it is expected to see individuals alive now with a new age record if the linear trend in the location parameter of GEV model still holds for this group. Finally the non-stationary model with censoring was used in order to add the current living world's oldest person, the results were very similar to the non-censored model with a slight increase in the shape parameter and an estimated annual increase of 45.91 days for the WOP titleholders.

As seen above, extreme value analysis of the age at death of the world's oldest person titleholders does not indicate an existence of a limit for human life length at extreme age. These findings agree with the studies by Rootzén and Zholud (2017) [7] that there is no evidence of a finite upper limit to the human lifespan. The difference between the two studies is that the age of the WOP titleholders is best modeled by the non-stationary GEV model ($\gamma > 0$), whereas the age of validated supercentenarians was best modeled by the Exponential distribution ($\gamma = 0$). In this project, the Exponential distribution was not a good model for the data. Hence, the return level plot in this project is concave with no evidence of a finite bound, whereas the return level in the paper by Rootzén and Zholud (2017) [7] is linear with no evidence of a finite bound as well. The results agree with the conclusions made by Medford and Vaupel (2019) [8] that there is a small chance (around one in five chance for the paper by Medford and Vaupel (2019) [8]), that the current world's age record will survive until 2050, compared to the results obtained in this project that the probability of beating the record for individuals born in 1940 and later, is above 60% (Figure 20). Results from this study bring support to efforts being invested into finding the "cure for aging" in order to extend human life length, Häggström (2016) [13], [9].

References

[1] Dong, X., Milholland, B., Vijg, J.: Evidence for a limit to human lifespan. *Nature* 538, 257–259 (2016)

[2] Antero-Jacquemin, J., Berthelot, G., Marck, A., Noirez, P., Latouche, A., Toussaint, J.-F.: Learning from leaders: Life-span trends in olympians and supercentenarians. *J. Gerontol. Biol. Sci. Med. Sci.* 70, 1–6 (2014)

[3] Aarssen, K., de Haan, L.: On the maximal life span of humans. *Math. Popul. Stud.* 4, 259–281 (1994)

[4] Wilmoth, J.R., Deegan, L., Lundstrom, H., Horiuch, S.: Increase of maximum life-span in Sweden, 1861- 1999. *Science* 289, 2366–2368 (2000)

[5] Weon, B.M., Je, J.H.: Theoretical estimation of maximum human lifespan. *Biogerontology* 10, 65–71 (2009)

[6] Gampe, J.: Human mortality beyond age 110. In: H. Maier et al. (eds.) *Supercentenarians*, Chapter 13, pp. 219–230. Springer-Verlag, Heidelberg (2010)

[7] Rootzén H, Zholud D.: Human life is unlimited-but short. *Extremes.* (2017); p. 1-16.

[8] Medford A, Vaupel J.W (2019): Human lifespan records are not remarkable but their durations are. PLoS ONE 14(3): e0212345. https://doi.org/10.1371/journal.pone.0212345

[9] Pigg På Ålderns Höst: Hälsans Hemilgheter. UR (2018)
https://urplay.se/program/214749-halsans-hemligheter-pigg-pa-alderns-host

[10] GRG: Gereontology Research Group List of Validated Deceased Supercentenarians. www.grg.org (2016)

[11] Klein, J.P., Moeschberger, M.L.: Survival Analysis Techniques for Censored and Truncated Data. Springer (1997)

[12] Coles, S.G.: An Introduction to Statistical Modeling of Extreme Values. *Springer*, London (2001)

[13] Häggström, O.: Here be Dragons, Science, Technology and the Future of Humanity. Oxford University press (2016)

A Right end-point estimate with profile likelihood

In this section, the MATLAB program used to estimate the right end-point and confidence intervals for GEV distribution is given (Figure 2). The details are explained in section 2.1.3, the transformed log-likelihood function is presented as below:

$$l(\mu,\sigma,\gamma) = -m\log(\sigma) - \left(1+\frac{1}{\gamma}\right)\sum_{i=1}^{m}\log\left[1+\gamma\left(\frac{z_i-\left(\hat{z}_0+\frac{\hat{\sigma}}{\hat{\gamma}}\right)}{\sigma}\right)\right]$$
$$-\sum_{i=1}^{m}\left[1+\gamma\left(\frac{z_i-\left(\hat{z}_0+\frac{\hat{\sigma}}{\hat{\gamma}}\right)}{\sigma}\right)\right]^{\frac{-1}{\gamma}}$$

The MATLAB code is presented below:

```
%% Right End-points estimation using Profile log-likelihood (GEV)
% Age_d= age+days of individuals
% sigma= scale parameter estimate
% gamma= shape parameter estimate
% endpoints= range of right end-points of the distribution
% N= number of individuals
N=length(Age_d);
endpoints=(125:0.001:128);
m=zeros(length(Age_d),length(endpoints));
n=zeros(length(Age_d),length(endpoints));
l=zeros(length(endpoints),1);
    for j=1:length(endpoints)
        for i=1:length(Age_d)
            m(i,j)=log(1+(gamma*(Age_d(i)-(endpoints(j)+(sigma/gamma)))/sigma));
            n(i,j)= (1+(gamma*(Age_d(i)-(endpoints(j)+(sigma/gamma)))/sigma)).^(-1/gamma);
        end
        l(j)=(-N*log(sigma))-((1+(1/gamma))*sum(m(:,j)))-sum(n(:,j)); % log-likelihood function
    end
```

```
xrange=(125:128);
yrange=(-160:-150);
chisqe=max(l)-(chi2inv(0.95,1)/2);
plot(endpoints,l)
hold on
plot(xrange,repmat(chisqe,1,length(xrange)))
hold on
plot(xrange,repmat(max(l),1,length(xrange)))
hold on
plot(repmat(125.485,1,length(yrange)),yrange) % Lower bound
hold on
plot(repmat(126.745,1,length(yrange)),yrange) % Upper bound
xlim([125.1,127.5])
ylim([-160,-150])
xlabel('Right endpoints')
ylabel('Profile log-likelihood')
```