Online Supplement to Statistical Features of Persistence and Long Memory in Mortality Data

Gareth W. Peters [1]
Department of Actuarial Mathematics and Statistics, Heriot-Watt University
Hongxuan Yan [2]
Academy of Mathematics and Systems Science, Chinese Academy of Sciences
Jennifer Chan [3]
School of Mathematics and Statistics, The University of Sydney

Contents

1 Survival Curves for England and Wales over time.

It was observed that for the senior age groups ranging between 80-99, there is a consistent downward trend of estimated H for both genders. This feature is due to the fact that the mortality persistence is more difficult to detect in a smaller population of seniors. Furthermore, such population sizes in this age bracket have rapidly changed throughout the period of analysis of this time series for such age groups, which is evident if one observes the change in life expectancy over time. As an illustration of this point, one may refer to Figure 1 which shows the change in life expectancy as captured by survival curves of the population as it changes through time from 1851 to 2031 plotted over decades.

[1]Email: garethpeters78@gmail.com
[2]Email: yhx19901122@gmail.com
[3]Email: jchan@maths.usyd.edu.au

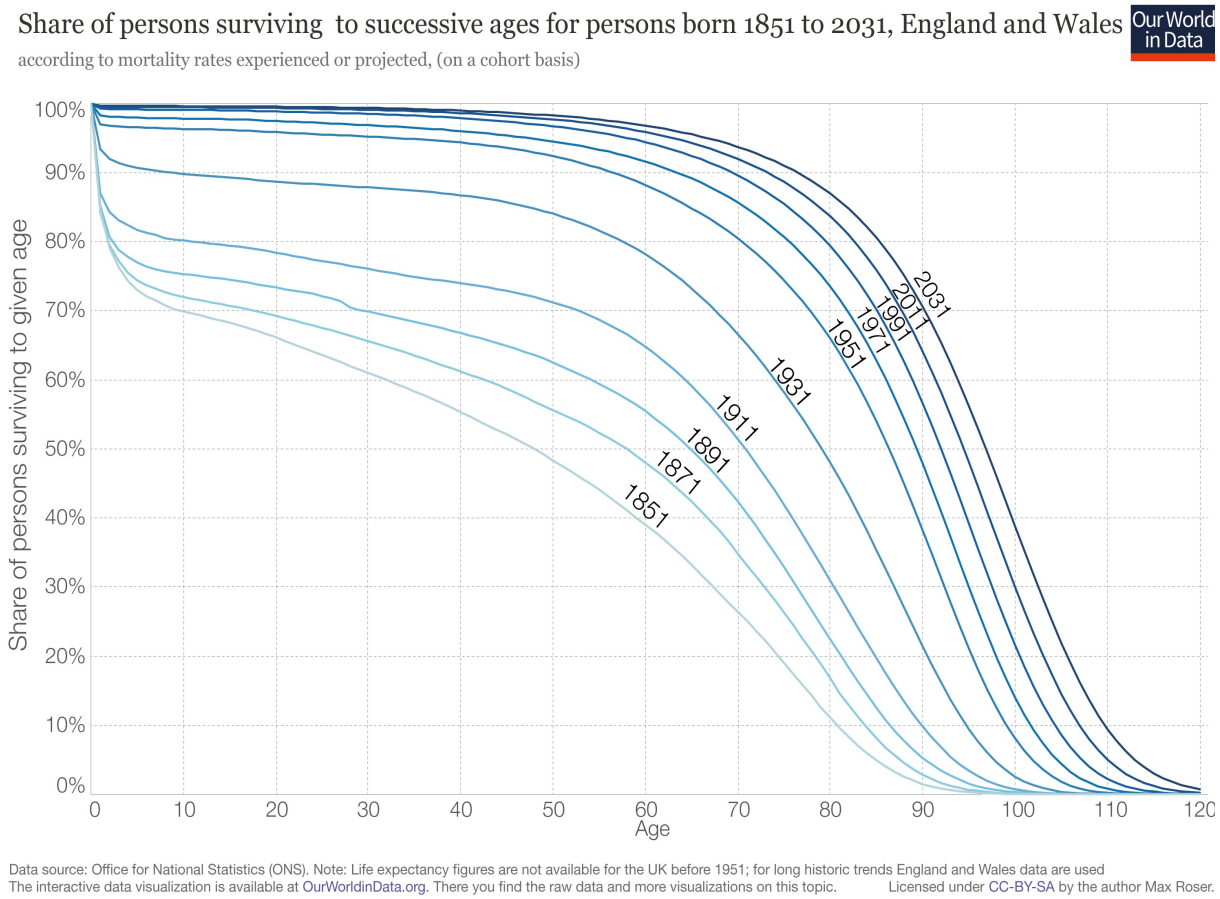

Figure 1: Source: Max Roser and Esteban Ortiz-Ospina (2019) – "A different view on mortality by age – survival curves". Published online at OurWorldInData.org. Retrieved from: https://ourworldindata.org [Online Resource]

2 Long memory pattern before and after world war II

In this study, we seek to determine the strength of a population mortality shock that alters the persistence features of the mortality processes in a range of countries. In order to facilitate such a study, we focus on two population regimes where the post-shock structural properties of the time series may be altered. We show that such a population shock due to the effect of world war two (WW2) influences the mortality processes in different countries in various ways. In particular, countries heavily involved in the war seem to have a different structural response in the mortality processes than those less involved. For instance, we observe evidence similar in nature to the results obtained in Fung et al. (2017), which indicates that as a consequence of mortality shock, there is a creation of a cohort effect in European populations. However, despite this, the long memory structure prevails as a statistically significant feature. The features of long memory structures of 21 age groups and 15 countries before and after WW2 are investigated, excluding Japan because there is no fully available record of mortality data before 1947 in Japan. Heat-maps in Figure 2 and Figure 3 show that the H estimates for all age groups after WW2 increases. Such an increase of mortality persistence indicates a more stable living environment and better living standards, including food quality and medical services after WW2. However, the persistence for youth and senior age groups is higher than the age groups 25-40, consistent with our previous observations.

To study the age group effect and country effect separately, we aggregate the H estimates in the Heat-map over countries and age groups respectively in Figure 4 and Figure 5. Figure 4 shows that

the distribution of H estimates before WW2 spreads more widely than the H after WW2 for all age groups. This observation indicates that the persistence of mortality differs more across countries before WW2. Before WW2, the median of H for the male is mostly lower than the median for female age during age 15-45. However, after WW2, such difference disappears or even reverses. This implies that the mortality persistence for male increases relatively to female after WW2. Figure 5 illustrates that the pre-war and post-war periods for these 16 countries are totally different in their mortality persistence. For Australia, Canada, Italy, Spain, U.K. and U.S.A, the median of H before WW2 is smaller than H after WW2. This agrees with the previous observation. For Belgium, Denmark, Finland, France, Netherlands and Norway, the range of estimated H for male before WW2 is much wider than male after WW2. The mortality persistence for male before WW2 varies more across age groups among these European countries than after WW2.

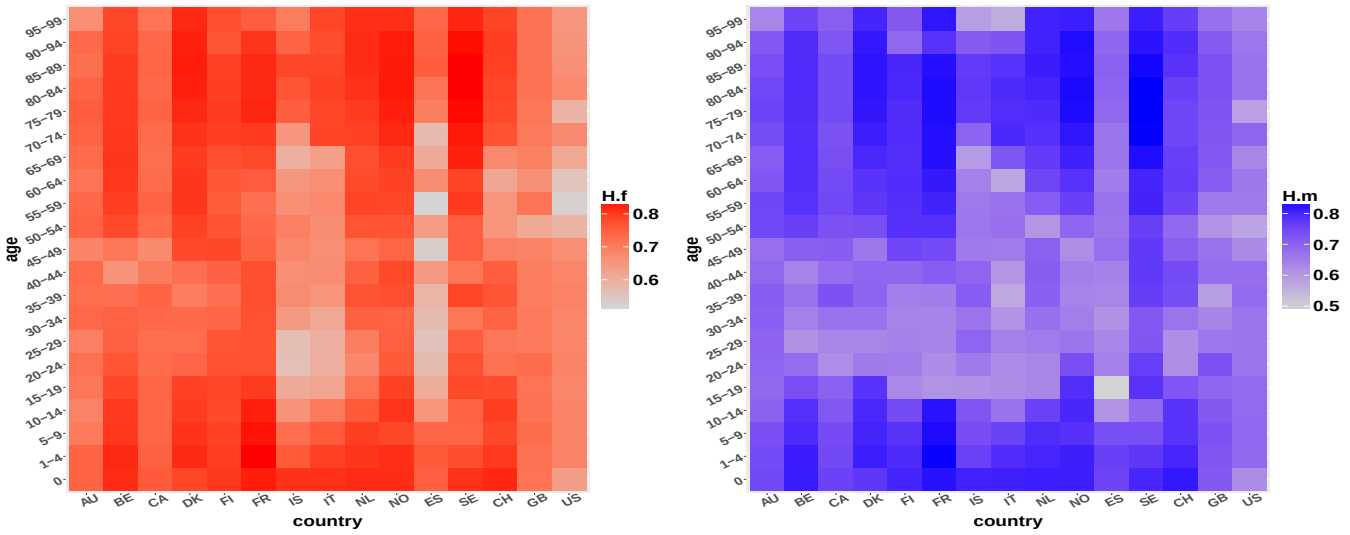

Figure 2: Heat map of estimated Hurst exponent across gender, age groups and countries before WW2 for female (left) and male (right).

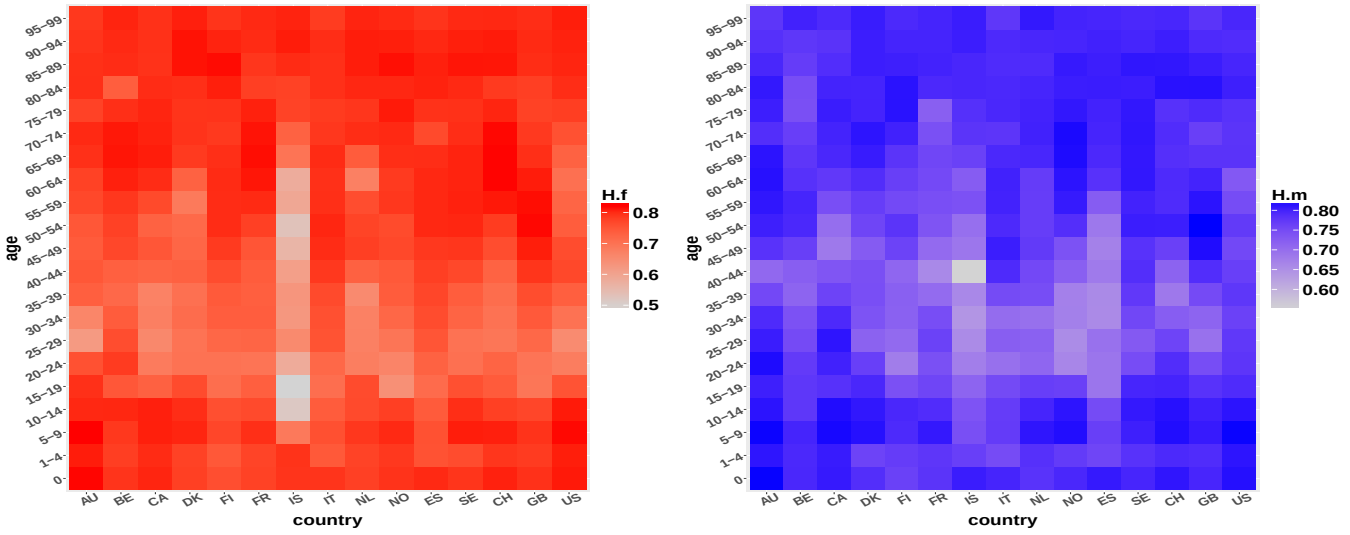

Figure 3: Heat map of estimated Hurst exponent across gender, age groups and countries after WW2 for female (left) and male (right).

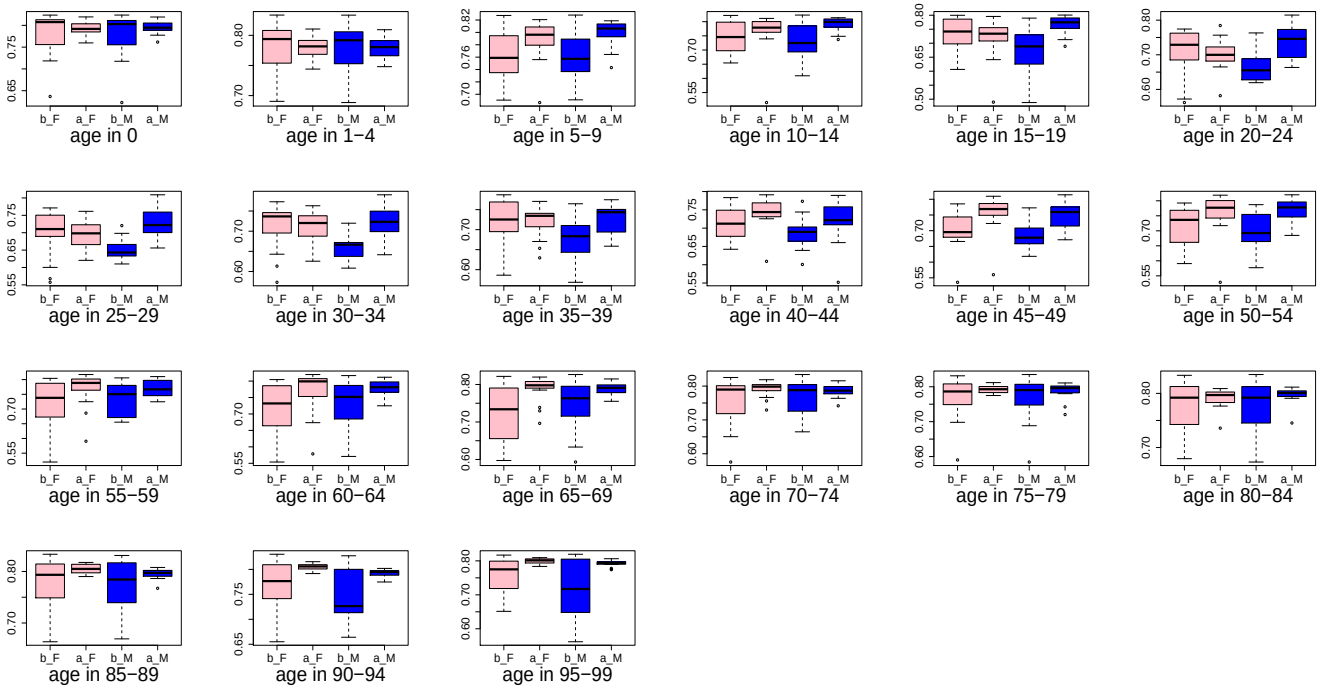

Figure 4: Boxplot of estimated H before/after world war two across age groups aggregated over countries with b_F, a_F b_M, and a_M representing before WW2 female (red), after WW2 female, before WW2 male (blue), and after WW2 male respectively.

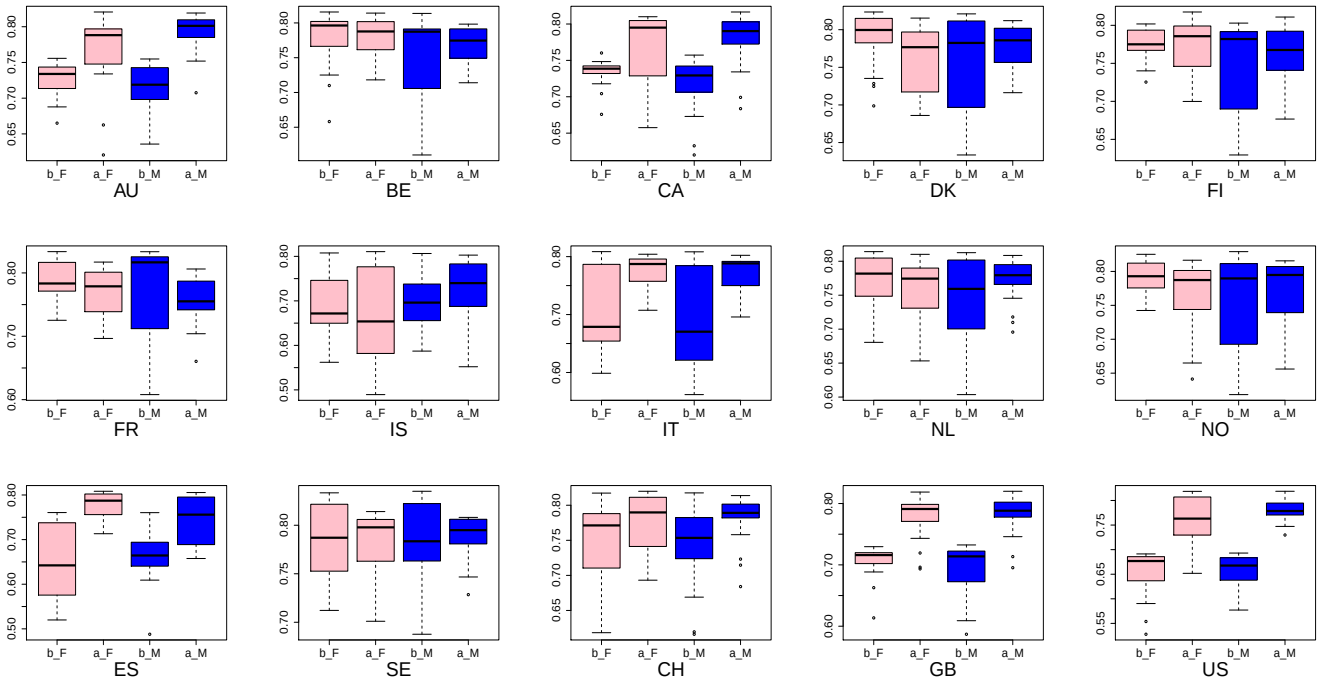

Figure 5: Boxplot of estimated H by ages before/after world war two across countries aggregated over age groups for female (red) and male (blue).

3 Brief overview of Kernel K-Means clustering

In this section, we will briefly explain the generic machine learning method and concept of kernel clustering in the weighted k-means variant. Kernel K-means uses an implicit typically non-linear feature mapping $\varphi(\cdot) : \mathbb{R}^d \mapsto \mathcal{H}$ embedding input features $\boldsymbol{f}_p \in \mathbb{R}^d$ as points $\varphi_p \equiv \varphi(\boldsymbol{f}_p)$ in a reproducing kernel Hilbert space (RKHS). In the examples in this paper, the feature vectors will be the multi-fractal Hurst exponents $\boldsymbol{f}_i = (H(q_1), H(q_2), \ldots, H(q_d))$ for d-values of q the time resolution or scales selected in the multi-fractal feature extraction estimation stage for the i-th population data (country, gender and age)-group death count time series.

The objective of the clustering method on these embedded or kernelised features is then to minimize the squared errors in the embedded space corresponding to the objective function

$$O(\mathcal{S}, \boldsymbol{\mu}) = \sum_{k}^{K} \sum_{p \in S^k} w_p ||\varphi(\boldsymbol{f}_p) - \boldsymbol{\mu}_k||^2_{\mathcal{H}} \tag{1}$$

where $\mathcal{S} = \left(S^1, S^2, \ldots, S^K\right)$ is a partitioning (clustering) of Ω into K clusters and $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K)$ is a set of parameters for the clusters with $|| \cdot ||_{\mathcal{H}}$ the Hilbert norm and w_p are positive weights for each data point.

As detailed in the machine learning literature, see Dhillon et al. (2004), the objective function $O(\mathcal{S}, \boldsymbol{\mu})$ can be optimized with respect to parameters $\boldsymbol{\mu}$ for each cluster. The resulting solution is closed form and corresponds to cluster means in the embedded feature space:

$$\boldsymbol{\mu}^*_k = \frac{\sum_{q \in S^k} w_q \varphi(\boldsymbol{f}_q)}{\sum_{q \in S^k} w_q} \tag{2}$$

The Euclidean distance from φ_p to the cluster center $\boldsymbol{\mu}_k$ is given by

$$\begin{aligned} &\varphi(\boldsymbol{f}_p) \cdot \varphi(\boldsymbol{f}_p) - \frac{2 \sum_{q \in S^k} w_q \varphi(\boldsymbol{f}_p) \cdot \varphi(\boldsymbol{f}_q)}{\sum_{q \in S^k} w_q} + \frac{\sum_{q,l \in S^k} w_q w_l \varphi(\boldsymbol{f}_q) \cdot \varphi(\boldsymbol{f}_l)}{\left(\sum_{q \in S^k} w_q\right)^2} \\ &= \kappa\left(\boldsymbol{f}_p, \boldsymbol{f}_p\right) - \frac{2 \sum_{q \in S^k} w_q \kappa\left(\boldsymbol{f}_p, \boldsymbol{f}_q\right)}{\sum_{q \in S^k} w_q} + \frac{\sum_{q,l \in S^k} w_q w_l \kappa\left(\boldsymbol{f}_q, \boldsymbol{f}_l\right)}{\left(\sum_{q \in S^k} w_q\right)^2} \end{aligned} \tag{3}$$

where $\kappa\left(\boldsymbol{f}_p, \boldsymbol{f}_p\right)$ is a Mercer kernel, for instance the Gaussian Radial Basis kernel can be selected for k-means and is defined as

$$\kappa\left(\boldsymbol{f}_p, \boldsymbol{f}_p\right) = \exp\left(\frac{-\left\|\boldsymbol{f}_p - \boldsymbol{f}_p\right\|^2}{2\alpha^2}\right),$$

and the dot products $\varphi(\boldsymbol{f}_q) \cdot \varphi(\boldsymbol{f}_l)$ are computed using kernel functions.

The key aspect of kernel k-means is that the clustering in the feature mapped function space is performed without ever having to know explicitly the map $\varphi(\cdot)$, instead of one simply selects a kernel family to characterize a space of functions $\varphi(\cdot)$. The natural choice of the kernel for k-means based kernel clustering is the p-kernel widely used and known as the radial basis function kernel, presented above.

4 Words of caution when utilising open source software for long memory estimation.

One needs to be careful when using open source free software packages. The following is simple guidance on modifications that we had to perform, when using some of the free R packages in the CRAN for estimation of the Hurst exponent, in order to make sure that the Hurst exponent estimators for the RSA, DFA and PR were working correctly and matched the mathematical expressions presented in this paper. In several cases the implementations did not reflect the formulas presented in this paper and consequently produced some erroneous and biased estimation on synthetic controlled experiments.

To assist the data scientist seeking to use this method, some common errors we found in public R code making these estimations:

- the most serious of these errors was the incorrect calculations of the fluctuations used in the DFA estimator code and therefore input of the wrong regression variable to the linear estimator for d or H. For the reader, we advise that you carefully check free packages to ensure that the average root mean square fluctuation is calculated and not something else which may be approximately this. We identified this error in a few DFA functions in R that are publicly available.;
- units of log-scale used didn't match the expected output in the linear regression, causing a units bias offset on output parameters so check the log-base used and we advise to convert all to natural logarithm units as presented in our paper;
- incorrect parameter reporting of H or d. We advise the reader to check if they are really being reported d and not accidentally H as some functions presented;
- spacing and scaling issues with the periodogram method in the DFT spacing of frequencies not correctly transformed.

These were a few of the main issues one should be alert to, but simple changes to match the formulas provided in this paper then yield correctly behaving estimators as we demonstrate in our results in Appendix 2.

5 Comparison of R/S, DFA and PR estimators

This section compares the performance of the R packages in estimating Hurst exponent using R/S, DFA and PR three different methods. In order to evaluate the performance of these packages in estimating Hurst exponent with different long memory strength d and sample size n, we choose true values for $d \in \{0.15, 0.25, 0.35, 0.45\}$ and $n \in \{50, 100, 250, 500, 1000, 2000\}$ which we used to generate from an AFRIMA$(1, d, 1, 0)$ model. For each parameter combination we generated, 1,000 samples of time series data and for each generated time series sample, we performed estimation with each of the estimators and compared the estimate to the true parameters. The results are demonstrated in a box-plot to show the distribution of the estimated Hurst exponent for each true value of d as sample size increases.

Figures 6, 7 and 8 show the estimated Hurst exponent using R/S, DFA and PR three different methods, after corrections to the R codes were made to match correct statistical estimators. The x-axis represents the true value $d = H - 0.5$ and y-axis show the estimated H values. With the increasing of sample size, these box-plots present less variability.

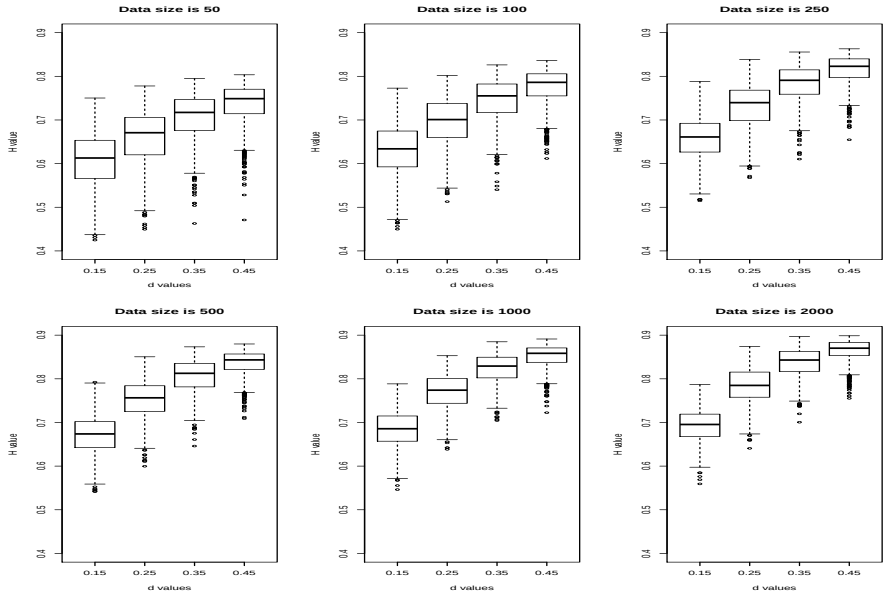

Figure 6: Boxplots of estimated H using RS method

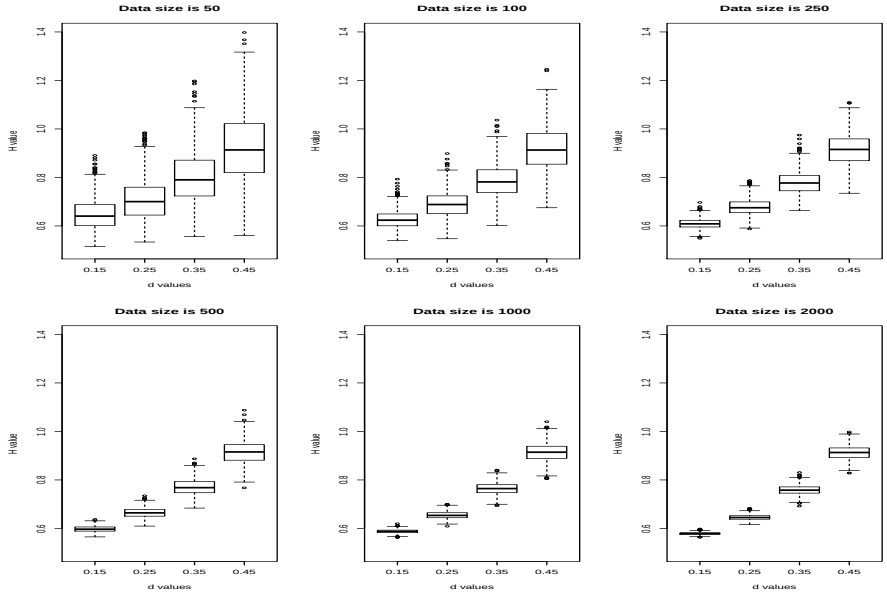

Figure 7: Boxplots of estimated H using modified DFA method

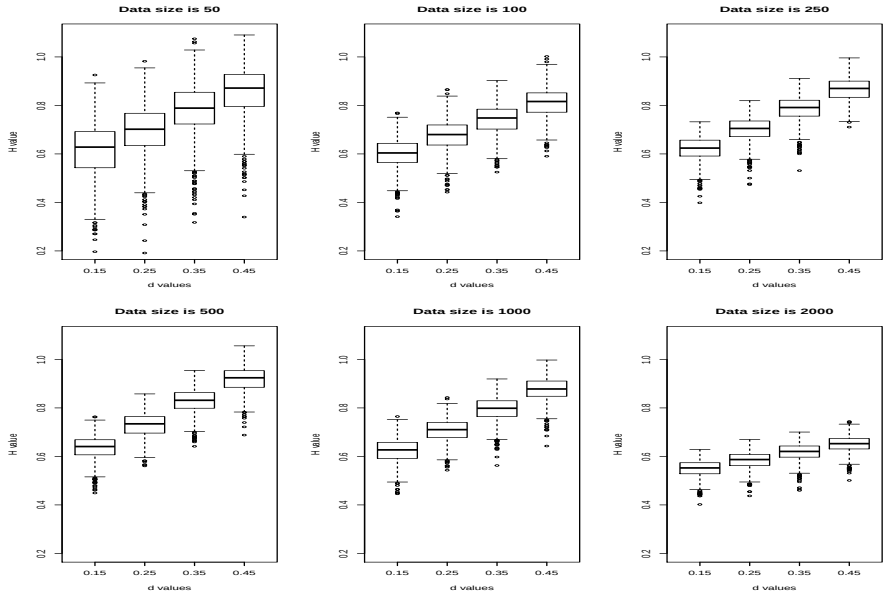

Figure 8: Boxplots of estimated H using modified PR method

6 Synthetic study: assessing the properties of long memory estimation

We consider a study of time series with mono-fractal properties, for varying parameter d and we display the box plots of the estimated function $H(q)$ over a grid of values of q. We know that at $q = 2$ one would expect to obtain a close estimate of the mono-fractal long-memory parameter. We choose $d \in \{0, 0.15, 0.25, 0.35, 0.45\}$ and $n = 500$. The vector of scales was chosen to be $5 : 20$ because our data length is around 90. The polynomial order for the detrending is 1. q-order of the moment are from -9 to 9. We further modify the package by apply divide the estimate of H by $\log 2(\exp(1))$ since this package returns base 2 logarithm rather than natural logarithm. Figure 9 shows the estimated Hurst exponent using MFDFA. The x-axis represents different q values and y-axis show the estimated H values. With increases in the value of d, these box-plots present less variability and the estimated H get closer to true value for positive q. There exist more bias for negative q, which is perhaps expected as mortality data may not vary significantly over small time scales, except perhaps for seasonal flu effects or holiday/festive seasons, which will be at most annual or bi-annual in general for most countries under study. Therefore, this will not be a big problem for the analysis to be presented, and we will exclude the negative values, using only $q > 0$ for real data studies.

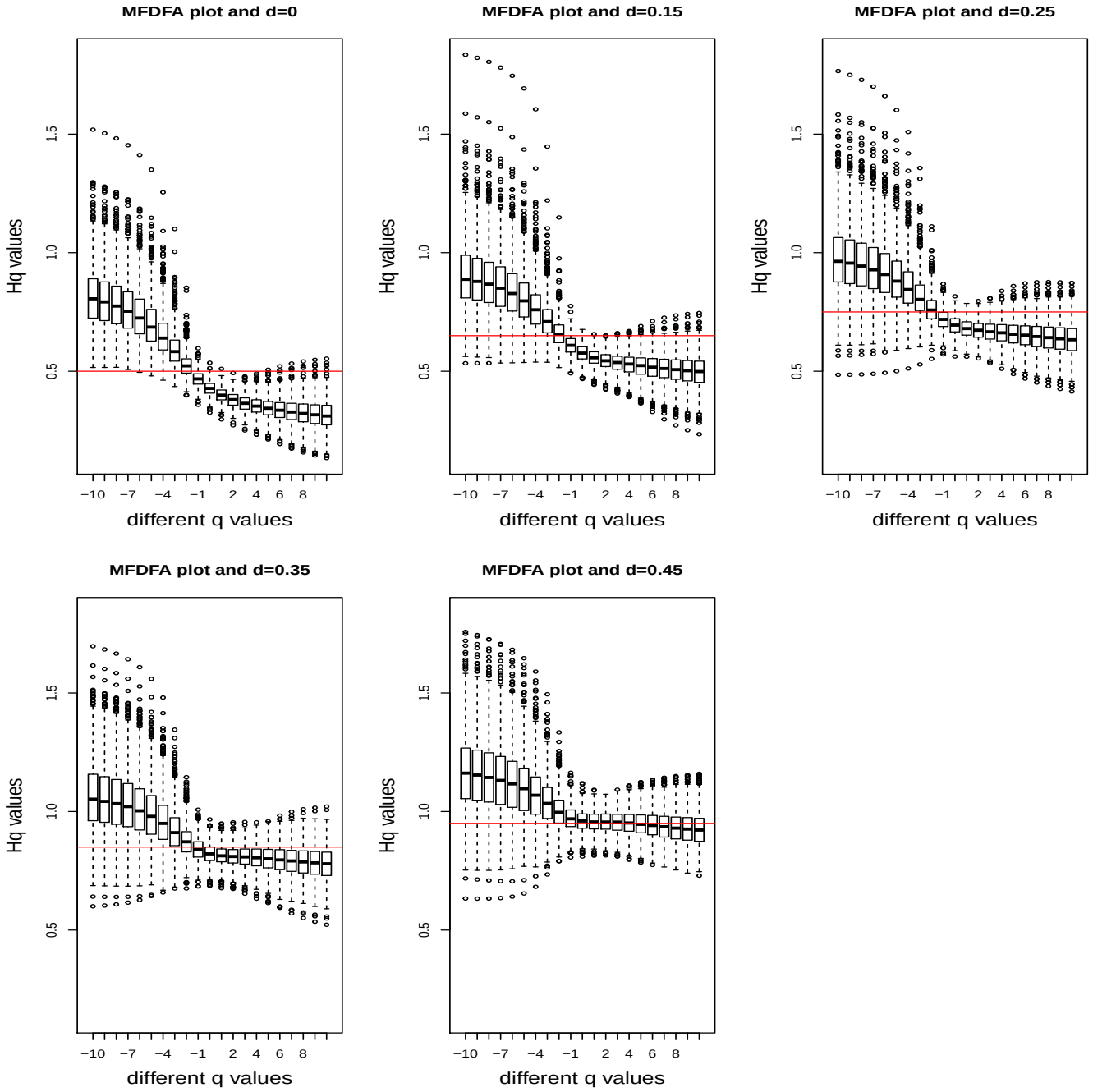

Figure 9: Boxplots of estimated H for different d values

References

Dhillon, I. S., Guan, Y., and Kulis, B. (2004). Kernel k-means: spectral clustering and normalized cuts. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 551–556.

Fung, M. C., Peters, G. W., and Shevchenko, P. V. (2017). Cohort effects in mortality modelling: a Bayesian state-space approach. SSRN: https://ssrn.com/abstract=2907868.