Smoothing Mortality Rates Using Random Forests

Torsten Sauer*and Roland Rau†

November 1, 2019

Abstract

To smooth sparse mortality age schedules, existing approaches use parametric, penalized non-parametric, relative and Bayesian approaches. Statistical learning has become important in various scientific disciplines to solve complex problems. The question arises if it can also be used to smooth mortality age schedules. We introduce a new algorithm based on Random Forests, a non-parametric statistical learning approach. The algorithm is trained on data from the Human Mortality Database about the shape of smooth mortality schedules. The trained models are then used to predict smooth age trajectories on sparse data. To test the method we apply cross validation with simulations of sparse data. It is found that the approach can smooth moderate sparse patterns comparably well to TOPALS. The algorithm introduces a new way of smoothing and its promising results makes it also considerable for age-specific fertility rates.

*University of Rostock & Max Planck Institute for Demographic Research
†University of Rostock & Max Planck Institute for Demographic Research

1 Introduction

Age-specific mortality rates of populations are of great interest in demographic research. They are the basis of life tables to calculate mortality measures like life expectancy or lifespan disparity.
In small populations the number of age-specific events can become very small, or even zero, that these patterns exhibit large fluctuations.
Existing approaches to smooth out the random fluctuations can be broadly classified as traditional parametric models (e.g. Gompertz [1], Kannisto-Tathcher [2]), relational models (e.g. Brass, TOPALS [3, 4]), non-parametric models (e.g. P-Splines [5, 6]) and Bayesian models [7].
What has not been used so far in demography are methods of statistical learning, despite by successfully employment in other disciplines, such as economics [8, 9] and medicine [10, 11], to solve complex problems. Therefore the question arises if statistical learning can be used to smooth sparse mortality patterns. In this paper we introduce a new algorithm to smooth mortality schedules with large fluctuations using Random Forests [12], an non-parametric statistical learning approach.

2 Methods and Data

The Algorithm

Lets imagine a smooth mortality schedule with n age-specific death rates (m_x). If now one age-specific death rate m_i is zero than it is relatively simple to guess the level of m_i just by hand, using the information of all other $m_{x \neq i}$ death rates. This intuitive replacement is based on our knowledge about human mortality schedules. We know mortality is decreasing from age 0 to around age 15 and then nearly exponentially increasing until old ages. We also know that in a smooth schedule death rate m_i is close to its neighbors m_{i-1} and m_{i+1}, which are close to their neighbors, which are close to their neighbors, and so on. As a consequence this means that in a smooth mortality schedule every m_i-th death rate provides information about all other death rates and vice versa. These information have different weights, depending on the relationships between the single death rates.
Our algorithm tries to learn these relationships in single regressions models, where every age-specific death rate is once the dependent variable and all others the predictors. Because our example uses single ages from 0 to 100, this leads to 101 models which are fitted within the algorithm. These trained models are then used to predict smooth trajectories on sparse mortality schedules.
We choose Random Forests [12] as the algorithms regression method. A Random Forest is a so called assembling method, because it consists out of hundreds of decision trees, each based on a unique bootstrap sample. The introduction of randomness in multiple steps, its capability to handle a huge number of predictors and its ability to model non-linear relationships, makes a Random Forest to one of the most accurate learning algorithms [13].

Data

To train and test we use death counts (D_x) and exposures (N_x) from the Human Mortality Database [14], by single year and single age to calculate age-specific death rates (m_x).
For the training and testing procedure, only complete ($D_x \neq 0$) and sufficiently smooth mortality schedules with a life expectancy of 50+ years are selected. This leads to a total of 2893 mortality schedules for 49 countries from the years 1850 to 2017.
To test the algorithm in a cross validation setting, the dataset is randomly divided into 80% (n=2314) training and 20% (n=579) test data.

Experimental Design

The mortality schedules of the test data are used to simulate age-specific death rates. To simulate sparse mortality data the simulations are done for six populations sizes decreasing by an order of magnitude from 10,000,000 to 1000. Age-specific population weights (c_x) are drawn randomly for every schedule. All simulations are repeated 100 times. This leads to 579x6x100 simulated schedules which are fitted using the algorithm. The same schedules are also fitted with TOPALS, to compare our algorithm with one of the state of the art approaches. The mean absolute error for all schedules by all six population sizes is calculated, to see how well the algorithm is smoothing and compare its precision to the TOPALS method.

3 Results

Algorithm Training

Since Random Forest don't provide coefficients and p-values to asses the impact of each variable for the prediction, a different measurement is needed to get insights into the models. For this purpose Random Forest provide the so called Variable Importance (VIMP). Variable Importance shows the impact of a variable by displaying how much the prediction accuracy would suffer if this variable would be excluded from the process. The bigger the reduction in accuracy, when a variable is left out, the bigger is its assumed impact on the prediction of the outcome.
Like previously described, the trained algorithm consist of 101 trained Random Forests, one for each age, where all other ages are predictors. Therefore, the algorithm contains the Variable Importance measurements of all ages in all models. Figure 1 displays these measurements, stacked row by row. The models for the response ages are depicted on the y-axis, where as the predictors are depicted on the x-axis. Consequently, every row displays the importance measurements for one prediction of a age-specific death rate. The diagonal is blank, because a death rate can't be outcome and predictor at the same time. The darker the color (white to blue scale), the higher the predictability of the age in the respective model. Note that the boundaries are shown on a logarithmic scale, because the decrease in accuracy is normally relatively small when excluding one variable from 101.
Even if all measurements come from single models, stacked they form an interesting pattern. To give a

short reading example for Figure 1: a) If we are interested in which ages are informative in predicting the death rate of age 25, we follow the horizontal line of this age. It reveals that from age 0 on the predictability of the ages is decreasing until age 15 and than start to rise again to peak around 25 itself, followed by another reduction in predictability. From age 60 the relative importance of the ages goes up again until the old ages. b) If we are interested in how much information age 25 contains about all other ages, we follow the vertical line of it. This shows that age 25 is important for its +-10 neighbor ages and that its impact is relatively small when it comes to predict death rates from 45 to 100.

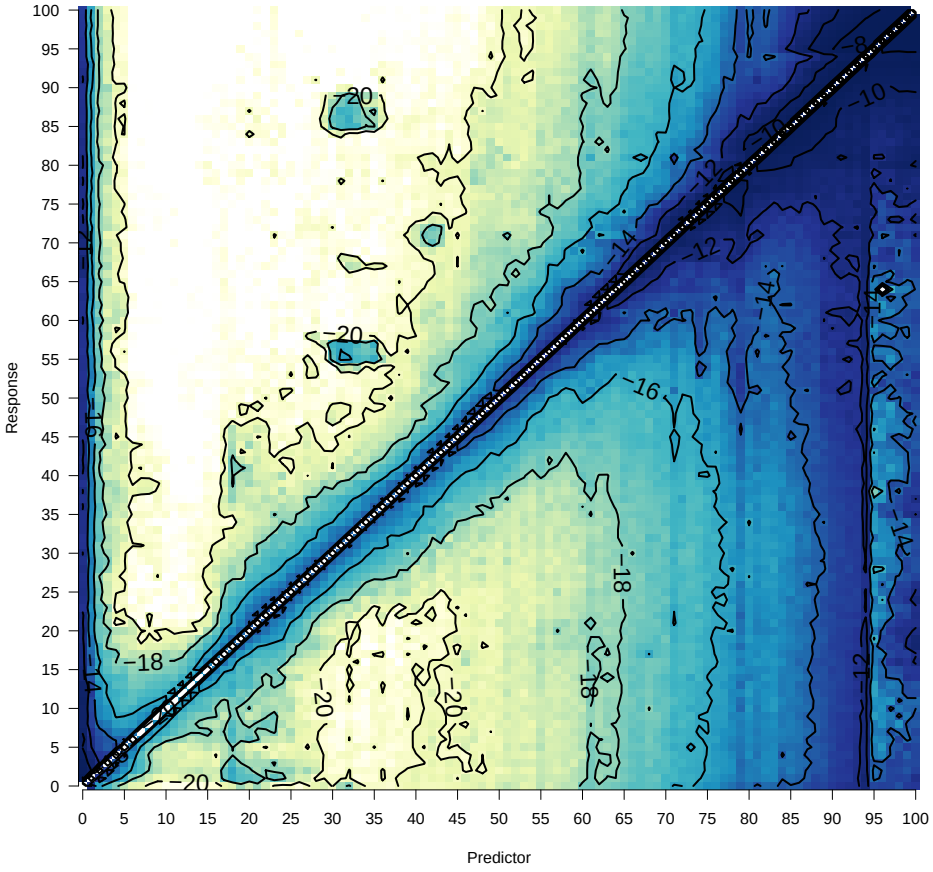

Figure 1: Importance Matrix for all age specific Random Forest Models. predictors are depict on the x-axis and response variables on the y-axis. The counter boundaries are set on a logarithmic scale. The darker the color the more important is the age specific death rate as a predictor for the depending age specific death rate.

Overall Figure 1 reveals three main findings. First, from the dark coat around the diagonal it can derived that the neighboring death rates are always important when it comes to predict the death rate in their middle. Second, from the dark vertical line at ages 0 to 5, it is visible that these ages always contain important information about all other ages in the age schedule of mortality. Third, the predictability for all ages, is rising from 40 until the highest ages.

When it comes to use all models to predict death rates of all ages these pattern can be interpreted as

impact weights of each age, which are derived from the training process and therefore were "intuitively" found by the algorithm.

Prediction Results

To evaluate the algorithm a cross validation procedure is used, where training and testing rely on different datasets. Figure 2 shows the estimation of the algorithm for the example of the male mortality pattern in England and Wales of the year 1972 (randomly picked). It is visible that even when data is becoming scattered and sparse the algorithm still suggest a form which can be identified as a schedule of human mortality. In Figure 2 it is also shown that the algorithm is capable of fitting the underlying mortality schedule (gray) very well, even when there are increasingly no realizations of death cases in a lot of ages.

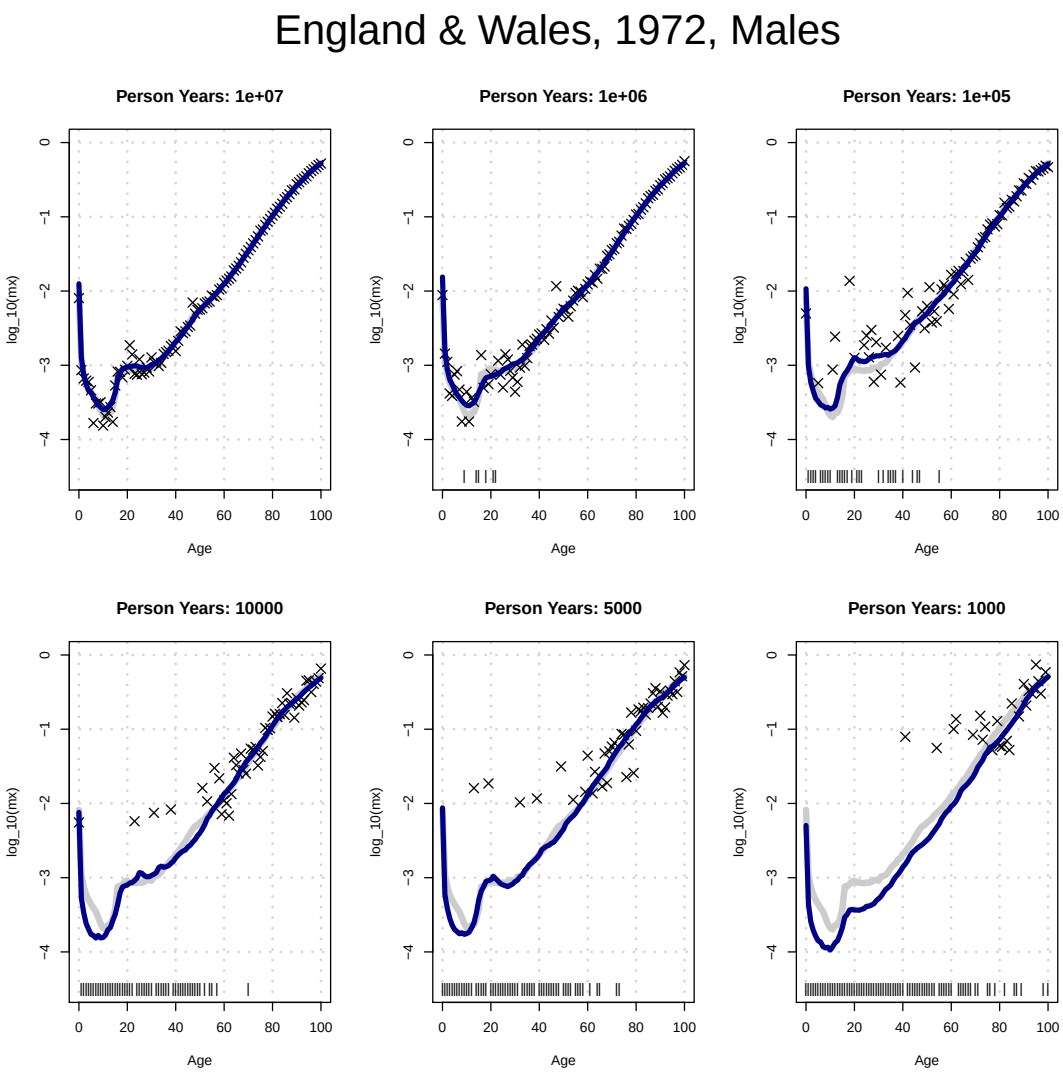

Figure 2: Example of the Random Forest fit (blue) on simulated death rates (black crosses) by deceasing population size. The underlying mortality schedule is showed in grey. Ages with zero mortality are indicated by black lines on the bottom.

To evaluate the accuracy and compare the algorithm estimations with TOPALS fits, the mean of the absolute error for every schedule by all six population sizes is calculated. The results are shown in Figure 3. It is visible that with decreasing population size the error is increasing. The algorithm seems to perform slightly better when the population is relatively large. From a population size of 100,000 its error is faster increasing than of TOPALS's. Nevertheless even at a population size of 5000, the algorithm still performs relatively well.

Because every age-specific death rate is a single estimate, for very sparse input patterns, the prediction can still contain fluctuation. Using this prediction results in an second iteration of prediction can smooth out theses fluctuations and lead to lower error.

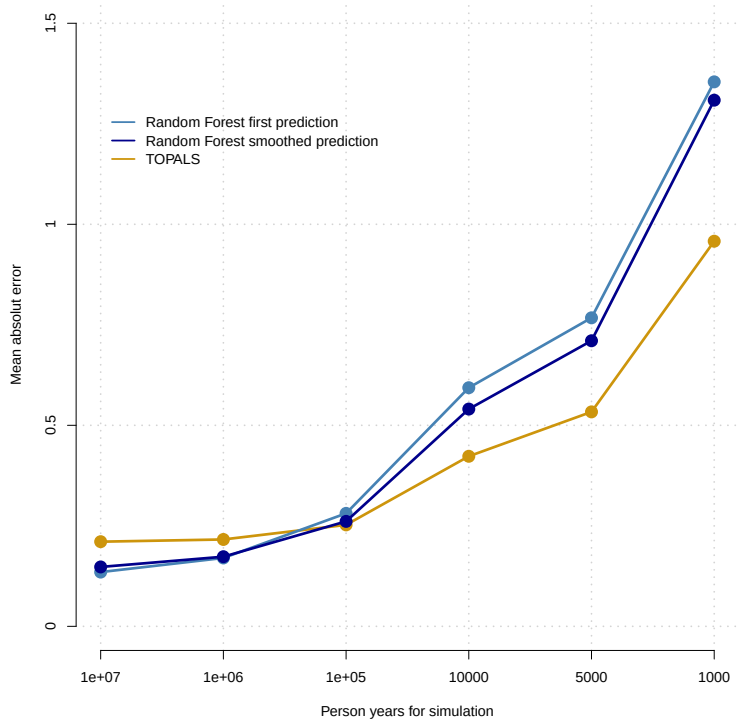

Figure 3: Mean absolute error over all test schedules (579), comparison of first Random Forest and smooth Random Forest prediction with TOPALS

4 Conclusion

The algorithm is a new approach to smooth age schedules of mortality. It shows promising results, which still have the potential to improve, by using feature reduction (the reducing of unnecessary predictors) and a finer tuning of hyper parameters of all models. Given its non-parametric nature and the lack of assumptions one may consider to use the algorithm also to smooth other rates, like age-specific fertility.

References

[1] Gompertz Benjamin. XXIV. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. In a letter to Francis Baily, Esq. F. R. S. &c *Philosophical Transactions of the Royal Society of London.* 1825;115:513–583.

[2] Thatcher A. R.. The long-term pattern of adult mortality and the highest attained age *Journal of the Royal Statistical Society: Series A (Statistics in Society).* 1999;162:5–43.

[3] Beer Joop. Smoothing and projecting age-specific probabilities of death by TOPALS *Demographic Research.* 2012;27:543–592.

[4] Gonzaga Marcos Roberto, Schmertmann Carl Paul. Estimating age- and sex-specific mortality rates for small areas with TOPALS regression: an application to Brazil in 2010 *Revista Brasileira de Estudos de População.* 2016;33:629–652.

[5] Currie Iain D, Durban Maria, Eilers Paul HC. Smoothing and forecasting mortality rates *Statistical Modelling: An International Journal.* 2004;4:279–298.

[6] Camarda Carlo G, others . MortalitySmooth: An R package for smoothing Poisson counts with P-splines *Journal of Statistical Software.* 2012;50:1–24.

[7] Schmertmann Carl P., Gonzaga Marcos R.. Bayesian Estimation of Age-Specific Mortality and Life Expectancy for Small Areas With Defective Vital Records *Demography.* 2018;55:1363–1388.

[8] Einav L., Levin J.. Economics in the age of big data *Science.* 2014;346:1243089–1243089.

[9] Brenner Thomas. , ed. *Computational Techniques for Modelling Learning in Economics.* Springer US 1999.

[10] Ascent of machine learning in medicine *Nature Materials.* 2019;18:407–407.

[11] BELLAZZI R, ZUPAN B. Predictive data mining in clinical medicine: Current issues and guidelines *International Journal of Medical Informatics.* 2008;77:81–97.

[12] Breiman Leo. Random Forests *Machine Learning.* 2001;45:5–32.

[13] James Gareth, Witten Daniela, Hastie Trevor, Tibshirani Robert. *An Introduction to Statistical Learning.* Springer New York 2013.

[14] University of California, Berkeley (USA), and Max Planck Institute for Demographic Research, Rostock, (Germany) . Human Mortality Database Available at `http://www.mortality.org` 2019.