*Cliodynamics: The Journal of Quantitative History and Cultural Evolution*

# A Bayesian Approach to Survivorship Bias in Historical Data Analysis

Tobias Wand

*University of Münster, Institute of Theoretical Physics and Rissho University, Faculty of Data Science*

Datasets such as Seshat have allowed researchers to quantitatively test hypotheses about premodern societies and states with great success. Nevertheless, one has to consider potential sources of bias in the data such as a survivorship bias favouring the inclusion of long-lived over short-lived states. Bayesian methods can be used to complement standard modelling procedures to take this issue into account as is demonstrated by analysing the longevity distribution of premodern states.

## Introduction

Cliodynamics has paved the way for the scientific analysis of human societies on long time scales via modelling and comparison to empirical data (Turchin 2003). Complemented by the efforts of the Seshat data bank (Turchin, Brennan et al. 2015), this methodology allows for detailed quantitative analyses of how the sampled states and societies developed (Turchin, Currie et al. 2018; Shin et al. 2020). In a recent article, Scheffer et al. have used the Seshat data in combination with their own Moros database to analyse the longevity of historical states by modelling the states' vulnerability or resilience against existential threats (Scheffer et al. 2023). They conclude that states' resilience decreases during their first 200 years of existence until it reaches a plateau and efforts are underway to find a microscopic model that can reproduce such

*Corresponding author's e-mail:* t_wand01@uni-muenster.de

*Citation:* Wand, Tobias. 2024. "A Bayesian Approach to Survivorship Bias in Historical Data Analysis." *Cliodynamics* 15(1): Report 1, 99–106.

behaviour (Schunck et al. 2024). However, the data gathering might suggest an alternative explanation: What if the states that collapsed within their first 200 years simply did not leave enough traces behind to be included in the dataset, i.e. what if a survivorship bias makes us believe that young states were more resilient than older states? As an example, the present paper explores this possibility for the resilience analysis of Scheffer et al., but its main idea should be kept in mind whenever historical data is analysed.

# Modelling the Data

### Resilience Analysis by Scheffer et al.

The resilience of a system can be defined as its tendency to return to its original state after a shock whereas vulnerability describes conversely the inability to restore the original state (Scheffer et al. 2023). Observing historical societies as the systems of interest, one can use their longevity distribution as a measure of how resilient they have been because a non-resilient state would collapse after it faces the first major crisis. Results from complex system science show that if the resilience of a state is constant throughout its lifetime, the resulting ensemble of recorded lifetimes will resemble an exponential survival distribution

$$f_s(t) = \exp(-\lambda_s t)$$

which can be compared to the empirical data. This shows that the naive $f_s$ overpredicts the number of states that should have collapsed at a fairly young age whereas the Seshat and Moros data have a well-defined maximum in their histogram distribution at around 200 years. Hence, Scheffer et al. develop the Saturating Risk model which assumes that the resilience is high for young states (i.e. just very few of them collapse) and decreases towards a constant plateau after about 200 years. Compared to other

alternative mechanisms that lead to different survival functions, Scheffer et al. show that the Saturating Risk model has the best fit to the data by using the AIC model selection criterion where a lower AIC indicates a better fit (Akaike 1998).

## Bayesian Modelling of Survivorship Bias

What if the number of states which died young is not *over*estimated by the naive model but rather *under*sampled by the data gathering? If a premodern state collapsed within the first few years after its foundation, it may well have left so few artefacts behind that archaeologists have not been able to recover any traces of its existence, thereby causing this state to be lost to time. Additionally, short-lived sociopolitical structures may have been classified as revolutionary periods or uprisings rather than full-fledged states (consider e.g. the eleven years of the republican Commonwealth of England). And because the Seshat data was mostly sampled at a frequency of 100 years, short-lived states which did not overlap with the turn of a century either must be deliberately included by researchers or simply fall by the wayside. Both mechanisms could be responsible for an undersampling of such states and therefore reflect a survivorship bias in databanks like Seshat or Moros.

Such concerns can be taken into account by using Bayesian statistics and incorporating those beliefs into a prior distribution (Von der Linden, Dose and von Toussaint 2014). As a simple example, the chance of a political formation with age $t$ to be considered a state by researchers and included in the data may be modelled naively by an exponential discovery rate via the prior distribution

$$f_p(t) = 1 - \exp\left(-\lambda_p t\right).$$

Hence, young states with age close to zero have a very low chance to be included in the datasets whereas long-lived states converge to a

probability 1 to almost surely be included. Assuming that the states now follow the naive longevity distribution for constant resilience (irrespective of whether they have been discovered by researchers or not), the observed longevity distribution is proportional to the product of these equations

$$f_B(t) \sim f_p(t) f_s(t) = \exp(-\lambda_s t) - \exp\left(-\left(\lambda_p + \lambda_s\right)t\right)$$

according to the Bayesian methodology. Note that because the prior distribution for the discovery rate converges to 1, $f_p$ cannot be normalised, but the posterior distribution $f_B$ can be normalised as a probability density. One could formally solve the problem of $\int_0^\infty \mathrm{f_p(t)dt} = \infty$ by replacing the infinite upper limit of the integral with the highest possible age of states τ, e.g. as the time since human settlements and agriculture were first established. Nevertheless, the normalisation constant is an irrelevant scaling factor for the problem presented here.

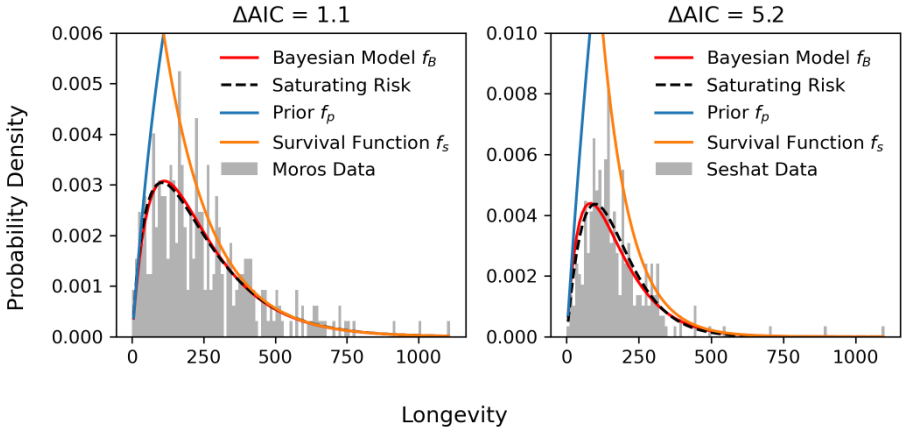


**Figure 1** Scheffer et al.'s Saturating Risk model is compared to the Bayesian model for the Moros (left) and Seshat (right) data. The ΔAIC indicates how much better Scheffer et al.'s model is compared

to the two-stage model. The Bayesian prior and the naive survival function are also depicted.

By using the likelihood optimisation scheme from (Scheffer et al. 2023) and the function *optimize.minimize* from scipy (Virtanen et al. 2020), the Seshat and Moros data is fitted to the Bayesian model $f_B$. The results in Figure 1 show that this model captures the general trend of the data extremely well. The AIC is used to evaluate its goodness of fit as in (Scheffer et al. 2023) by comparing the differences in AIC

$$\Delta AIC = AIC_B - AIC_{SR}$$

between the Bayesian and the Saturating Risk model. More information about the interpretation of $\Delta AIC$ and characteristic threshold values can be found in (Burnham, Anderson and Huyvaert 2010).

For the Moros data, the Bayesian model is essentially as accurate as the Saturating Risk model because the $\Delta AIC \approx 1.1$ is negligibly small. For the Seshat sample, $\Delta AIC$ is slightly larger, but the model is still reasonably accurate as discussed in (Burnham, Anderson and Huyvaert 2010). Hence, even though both datasets are fitted slightly better by the Saturating Risk model, the Bayesian model is essentially an equally good description within the fluctuation expected by statistical noise. Because both compared models have the same number of parameters, likelihood-ratio tests and $\Delta AIC$ evaluation imply the same results. Note that for short and high longevity, $f_p(t)$ and $f_s(t)$ approximate their respective tail of the data.

# Discussion

While this brief study is by no means exhaustive, it indicates that the Bayesian model is at least viable for the Moros data and perhaps, if coupled with a more refined survival function from

(Scheffer et al. 2023), could also become viable for the Seshat data and even surpass the accuracy of the Saturating Risk model. The exact form of the prior distribution could certainly be varied, too. Considering that a long-lasting state did not only have more time to produce its own set of archaeological records, but also to be mentioned in contemporary sources from other geographic areas, it might be reasonable to use a super-exponential discovery rate instead. Fine-tuning the Bayesian estimation might be an interesting endeavour, but is ultimately beyond the scope of this paper as the author's intention is to merely illustrate the potential effects of survivorship bias with an example from recent research. This article is not supposed to falsify the findings in (Scheffer et al. 2023), but rather to highlight issues that need to be considered during the analysis of historical data and offer the first attempt at a solution.

The product in the Bayesian model equation for $f_B$ might superficially resemble the product of hazard function $h$ and survival function $S$ in (Scheffer et al. 2023), but there is an important difference in the methodologies behind these approaches. In (Scheffer et al. 2023), both factors of the product stem from the same underlying process and are therefore not independent of each other. Moreover, the analysis assumes that the maximum of the longevity distribution is caused by a decrease in resilience and therefore originates from the underlying dynamics. On the contrary, the Bayesian approach of the present article highlights potential problems with the data gathering as the source of this feature in the observed distribution.

Of course, Scheffer et al. were aware of the imperfections found in historical datasets such as Seshat and Moros and described some of these issues in their Supplementary Information to (Scheffer et al. 2023) as well as listed some problematic sources and polities. Moreover, they explicitly stated that their polity selection “does pose a bias against areas with less well documented states”, thereby motivating the present article. Hence, it seems natural to posit an

alternative model that takes into account the probability of a given polity to actually be considered as a state and to be included in the datasets via a Bayesian prior distribution in order to account for this potential source of bias. Additionally, because historical polities often had less direct control over their territories than modern states and often originated as a subpolity of a larger empire, it is challenging to assign a clear beginning and end to their lifetime, which is reflected by the efforts to construct the CrisisDB database within the Seshat project (Turchin 2023; Hoyer, Bennett et al. 2023; Hoyer, Holder et al. 2024). "Crafting better databases" is included in their research agenda in (Scheffer et al. 2023) to alleviate those issues, but if the accuracy of such datasets is nevertheless limited by the sparsity of archaeological records, Bayesian methods might provide a suitable alternative to incorporate the underlying uncertainty.

# References

Akaike, Hirotogu. 1998. "Information Theory and an Extension of the Maximum Likelihood Principle". In *Selected Papers of Hirotugu Akaike*, edited by Emanuel Parzen, Tanabe Kunio and Kitagawa Genshiro, 199-213. New York: Springer.

Burnham, Kenneth P., David R. Anderson, and Kathryn P. Huyvaert. 2010. "AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons." *Behavioral Ecology and Sociobiology* 65(1): 23-35.

Hoyer, Daniel, et al. 2024. All Crises are Unhappy in their Own Way: The role of societal instability in shaping the past. *SocArXiv Preprint.* doi:10.31235/osf.io/rk4gd

Hoyer, Daniel, et al. 2023. "Navigating polycrisis: long-run socio-cultural factors shape response to changing climate." *Philosophical Transactions of the Royal Society B* 378: 20220402.

Scheffer, Marten, Egbert H. van Nes, Luke Kemp, Timothy A. Kohler, Timothy M. Lenton, and Chi Xu. 2023. "The vulnerability of aging

states: A survival analysis across premodern societies." *PNAS* 120(48): e2218834120. doi: 10.1073/pnas.2218834120.

Schunck, Florian, Marc Wiedermann, Jobst Heitzig, and Jonathan F. Donges. 2024. "A Dynamic Network Model of Societal Complexity and Resilience Inspired by Tainter's Theory of Collapse." *Entropy* 26(2): 98.

Shin, Jaeweon, Michael Holton Price, David H. Wolpert, Hajime Shimao, Brendan Tracey, and Timothy A. Kohler. 2020. "Scale and information-processing thresholds in Holocene social evolution." *Nature Communications* 11(1): 2394.

Turchin, Peter. 2023*. End Times: Elites, Counter-Elites, and the Path of Political Disintegration*. New York: Penguin Press.

— 2003. *Historical Dynamics: Why States Rise and Fall. Princeton*: Princeton University Press.

Turchin, Peter, et al. 2018. "Quantitative historical analysis uncovers a single dimension of complexity that structures global variation in human social organization." *PNAS* 115(2): E144-E151.

Turchin, Peter, et al. 2015. "Seshat: The Global History Databank." *Cliodynamics* 6(1).

Virtanen, Pauli, et al. 2020. "SciPy 1.0: Fundamental Algorithms for Scientific Python." *Nature Methods* 17: 261-272.

Von der Linden, Wolfgang, Volker Dose, and Udo von Toussaint. 2014. *Bayesian probability theory: applications in the physical sciences*. Cambridge: Cambridge University Press.