Article:
Fagan, Brennen, Knight, Marina I, MacKay, Niall et al. (2020) Change point analysis of historical battle deaths. Journal of the Royal Statistical Society: Series A (Statistics in Society). ISSN: 1467-985X

https://doi.org/10.1111/rssa.12578

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2020)

Change point analysis of historical battle deaths

Brennen T. Fagan, Marina I. Knight, Niall J. MacKay and A. Jamie Wood

University of York, UK

[Received April 2019. Revised March 2020]

Summary. It has been claimed and disputed that World War II has been followed by a 'long peace': an unprecedented decline of war. We conduct a full change point analysis of well-documented, publicly available battle deaths data sets, using new techniques that enable the robust detection of changes in the statistical properties of such heavy-tailed data. We first test and calibrate these techniques. We then demonstrate the existence of changes, independent of data presentation, in the early to mid-19th century, as the Congress of Vienna system moved towards its collapse, in the early to mid-20th century, bracketing the World Wars, and in the late 20th century, as the world reconfigured around the end of the Cold War. Our analysis provides a methodology for future investigations and an empirical basis for political and historical discussions.

Keywords: Battle deaths; Change point analysis; Correlates of war; Heavy-tailed data; Long peace; Power law distribution

1. Introduction

Is war declining? The record of historical battle deaths surely embodies more human value than any other conceivable data set, for every unit in every data point is a human life violently taken, yet its structure remains poorly understood. Pioneering work was done in the *Journals of the Royal Statistical Society* (Richardson, 1944, 1946, 1952; Moyal, 1949) by the Quaker pacifist Lewis Fry Richardson. Richardson discovered one of the few robust quantitative results in political science (Richardson (1960), pages 143–167), that deaths in deadly quarrels are well described by two power law distributions (Clauset *et al.*, 2009), with powers of approximately 2.4 from murders up to events with about 1000 dead, and 1.5 for events of more than 1000 dead ('wars') (Richardson (1960), Fig. 4). On the question of whether humanity's propensity for deadly violence has fundamentally altered, Richardson's final conclusion was that

> 'the observed variations [in battle deaths] might be merely random, and not evidence of any general trend towards more or fewer fatal quarrels'

(Richardson (1960), page 141). The newly apparent phenomenon of the 60 years since Richardson's book is the post World War II 'long peace', although one might just as well characterize the 20th century by the 'great violence' (Clauset (2018), page 4) or 'hemoclysm' (Pinker (2011), page 229, originally due to Matthew White) of its first half.

Every point of these data takes place in a web of human society, culture and politics. To analyse this requires a broad sweep of multidisciplinary qualitative analysis, and an astonishing book by Pinker—suffused with individual statistics, but not overtly a statistical work—concludes that an individual's likelihood of violent death has greatly declined over the centuries (Pinker, 2011),

Address for correspondence: Brennen T. Fagan, Department of Mathematics, University of York, Heslington, York, YO10 5DD, UK.
E-mail: btf500@york.ac.uk

especially with the advent of the 'long peace' and a later, more tentative 'new peace' after 1989. Goldstein (2011) reached similar conclusions, giving much credit to the United Nations. The idea of an invariant human tendency towards violence retains its proponents (Huntington, 1989; Gray, 2012), although others who accept the violence of precivilized societies (e.g. Gat (2013)) nevertheless stress its amelioration by the continuing development of the Hobbesian state and superstate Leviathan. A classic work by Gaddis (1986) lays out multiple possible explanations for the post World War II absence of large-scale war.

The question has become hugely controversial in the last few years, playing out rather publicly in the pages of *Significance* between Michael Spagat and Stephen Pinker on the one hand and Pasquale Cirillo and Nassim Nicholas Taleb on the other (Spagat, 2015; Cirillo and Taleb, 2016a; Spagat and Pinker, 2016). Cirillo and Taleb (2016b) applied techniques from extreme value theory to an unpublished data set covering the years from 60 until 2015 and failed to find evidence for any change in arrival time or distribution. Clauset (2018) arrived at a similar conclusion by applying standard statistical techniques to the publicly available 'Correlates of war' (COW) data set as used in this paper. Spagat and Pinker considered it erroneous to conclude that there was no change in the distribution of violence since World War II without explicit comparison and testing of the periods immediately before and after. Indeed, they identified several qualitative changes that suggest that the world has become more peaceful, in line with results from Pinker (2011) who identified possible changes after 1945 and 1989. In the same vein Spagat and van Weezel (2018) tested the null hypothesis of no change in the magnitude of large wars before and after 1945 or 1950, using the same Gleditsch (2004) data set as in this paper, and found sufficient evidence to reject it for some definitions of large wars. Hjort (2018) performed a restrictive change point analysis limited to a single change point and requiring parametric assumptions, using the same COW data set, and subsequently found 1965 to be the most likely candidate for a change in the sizes of wars.

What has not been done so far, and is the subject of this paper, is a full and comprehensive change point analysis on a fully documented and freely available historical battle death data set. To conduct a full change point analysis on heavy-tailed data, in which extreme data points are 'common', is a difficult task for which the methodology has until recently been inadequate. Our contributions are

(a) to calibrate the components of the flexible methodology of Killick *et al.* (2012) and Haynes *et al.* (2017a, b) through simulation studies on generated data with traits akin to the historical data and
(b) to employ the proposed algorithm to infer in a data-driven manner whether there is sufficient historical evidence to support distributional changes.

We do not posit the existence of any fixed change point(s). To do so, after all, might cause us to miss other interesting phenomena in the data, and introduces human bias—we shall not impose a 2019 view of which moments may have been epochal. In a historical sense, should one or more change points be detected, this provides candidates for approximate times at which something changed in the distribution of wars. If, for example, a change point near World War II were detected, following which the distribution yields fewer deadly events, this would lend credence to the 'long peace'.

The paper is structured as follows. In Section 2 we introduce the historical battle deaths data sets. In Section 3 we calibrate the relevant methodology, focusing on simulated data and showing that there does indeed exist a change point methodology that is successful in identifying statistical changes in power law distributions. In Section 4 we use this methodology to analyse the historical data sets. We conclude in Section 5 with an interpretation and discussion.

The programs that were used to analyse the data can be obtained from

`https://rss.onlinelibrary.wiley.com/hub/journal/1467985x/series-a-datasets`.

2. Battle deaths data sets

Since the pioneering work of Richardson there have been many attempts to create data sets quantifying violence. The construction of these data sets raises some important questions, first of definition and then also of incomplete or biased knowledge. Richardson (1960), pages xxxvi and 4–12, was acutely aware of these issues, which is why he chose to focus on 'deadly quarrels' of all sizes and types. More recent approaches to data collection often focus on subtypes of deadly quarrels, such as battle deaths above a set threshold, as in the COW data sets (Sarkees and Wayman, 2010), or terrorism, as in the global terrorism database (National Consortium for the Study of Terrorism and Responses to Terrorism (2016), pages 9–10). For recent reviews see the works of Bernauer and Gleditsch (2012) and Clauset and Gleditsch (2018).

Even if we do settle on an appropriate subset of violence, there are still several issues to be decided. There are complex questions regarding the inclusion of non-combatants, particularly in asymmetric (typically, insurgent) warfare. An extreme example is the Taiping rebellion in 19th century China. There is no question that this tragic campaign led to enormous loss of life, but how many of the dead were combatants? How many civilian deaths have been accounted for? How do we separate battle deaths from those caused by famine and disease and those caused in other simultaneous rebellions? Estimates for this particular event vary over at least an order of magnitude. It is commonly stated that approximately 20 million died in total in the Taiping rebellion (Spence, 1996; Reilly, 2004; Fenby, 2013). Sivard (1991) indicated 5 million military deaths with 10 million total (in comparison with 300000 due to simultaneous rebellions) by using data due to Eckhardt. Worden *et al.* (1988) reported that 30 million were reported killed over 14 years. Platt (2012) reported in the epilogue 70 million dead, along with the standard 20 million–30 million figure and criticisms of both of these numbers. Deng (2003) indicated similar numbers from Chinese sources but noted their interrelationship with famine. However, the COW data set reports only 26000 (Chinese), 85000 (Taipings) and 25 (UK) battle deaths—albeit only for the second, interstate phase of the war. Battle deaths for the initial phase are listed as unknown. The Gleditsch (2004) data set is consistent with the COW values. Particular difficulty arises where there is disagreement between contemporary (or even political descendants of) participants, and especially where one or the other side has a different level of control or vested interest in the interpretation of the event (Sarkees, 2010).

A further issue emerges regarding granularity and data aggregation (Cirillo and Taleb, 2016b). What constitutes an individual event, and to what extent should individual actions be distinguished within a larger conflict? For example, should the different fronts in World War II be considered separate? Should World Wars I and II be considered merely as more active periods within a global conflagration which encompasses both? This might seem more natural from a Russian or Chinese perspective than from the Anglosphere—for example, how should we handle the Japanese invasion of Manchuria and the Sino-Japanese War of 1931–1945, or the Russian Civil War of 1917–1922? And, since such events (and related combinations thereof) happen over an extended period, to which point in time should we assign the combined event? Both inappropriate aggregation and inappropriate disaggregation can lead to artefacts (Cristelli *et al.*, 2012). Such problems cannot be wholly avoided but certainly require that we work only with well-known, publicly available data sets that handle the data consistently and use clearly stated assumptions on data gathering and aggregation.

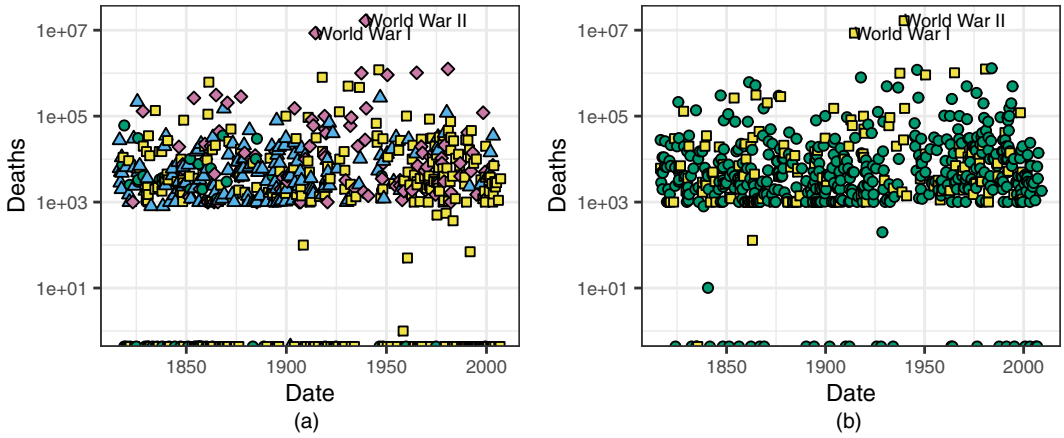


Fig. 1. Data sets on a logarithmic axis (the World Wars are labelled for reference): (a) COW data set (△, extrastate; □, intrastate; ◇, interstate; ○, non-state); (b) Gleditsch data set (□, inter; ○, civil)

We acknowledge that none of the available data sets is ideal, as each has varying criteria for inclusion of events; and indeed the available historical data themselves are not ideal, because of, for instance, biases in the record. The two data sets that we use are the COW data set (Sarkees and Wayman, 2010) and a data set due to Gleditsch (2004). The former has been used by, for example, Clauset (2018) and Hjort (2018), whereas the latter has been used by Spagat and van Weezel (2018). We note that the Gleditsch data set was originally based on the COW data set, although divergent evolution has occurred since. The COW data set has four subsets (interstate, intrastate, extrastate and non-state), whereas the Gleditsch data set identifies civil and interstate wars. For both data sets, each data point is a war, including its combatants (nation states or other organizations) and each combatant's battle deaths, date(s) of entry, date(s) of exit, alliance, outcome and status as initiator (Gleditsch, 2004) and additionally within COW the location and transition status (Sarkees and Wayman, 2010). We select these data sets for their availability, continuous maintenance, reputability and, most importantly for our work, their dedication to consistent application of their definitions. For example, by focusing exclusively and consistently on battle deaths, instead of more generally war deaths, uncertainties about their definitions and their definitions' influence on numbers in the data sets are kept to a minimum.

In our analysis, for simplicity, we consider each event to have occurred at its start date for the purposes of ordering. Indeed, *not* allocating to each war a single point in time has been shown in the literature to induce auto-correlation (Beard, 2018). Intuitively, if battle deaths are assigned per combatant to the date that a combatant joins the war, then auto-correlation arises due to the correlations between battle deaths of the combatants. Similarly, per-year battle deaths are correlated within a single war (Beard, 2018). Consistent event-based disaggregation is thus increasingly preferred in political and peace data (Gleditsch *et al.*, 2014; Clayton *et al.*, 2017). A natural future research direction then would be to develop a more disaggregated data set using better-resolved events—battles, perhaps—and to perform a change point analysis there.

In Fig. 1, we show the COW data set (Fig. 1(a)) and the Gleditsch data set (Fig. 1(b)), on a logarithmic scale for better visual representation of the data. For events that are listed but have no value recorded, we present the events on the bottom of the plots at their listed time, but we do not include them in the analysis.

A controversial question is whether we should consider the absolute number of deaths that are caused in a conflict or the number relative to global population; see for example the work

of Spagat and van Weezel (2018). There are good arguments for each choice reflecting two important questions about human value. The relative number, which was favoured by Pinker (2011) and Spagat and van Weezel (2018), approximates the probability of being killed in a particular event, and thus the significance of the event to the average person alive at the time. However, each unit within the data is a human life so we must acknowledge the criticisms of Epstein (2011) and Cirillo and Taleb (2016b), page 16, that we should not be satisfied merely with a decreasing proportion of battle deaths if the raw values stay high or increase. We therefore conduct our analyses on both raw data and data normalized by world population, computed by using the HYDE 3.2.1 data set (Klein Goldewijk *et al.*, 2017). Of course any change points in the (normalizing) population will interfere with change point detection in the battle deaths data set, and the two analyses need to be considered separately.

3. Methodology for detecting change points for power law distributions

3.1. Brief review of change-point-detection methodology

Recall that our aim is to identify in the battle deaths data sets the existence and locations, if any, at which we observe a change in the statistical properties. It has long been understood that battle deaths data are characterized by heavy tails and are typically modelled by using a power law distribution (Richardson, 1944, 1960; Clauset *et al.*, 2009; González-Val, 2016; Chatterjee and Chakrabarti, 2017; Clauset, 2018; Martelloni *et al.*, 2018). This yields some complex issues, as we explain next in the change-point-detection context.

Simply put, a typical change-point-search method consists of three components: an algorithm, a cost function and a penalty. The combination of cost function and penalty balances explanatory worth against model complexity, valuing the ability of the change points to describe the data while penalizing the additional complexity that they introduce (usually in proportion to their number). Often, the cost function is parametric in nature; it assumes some knowledge about how the distribution is parameterized. This may range from a simple assumption, for example, that the mean or variance exists (e.g. cumulative sums of squares CSS; Inclán and Tiao (1994)) to something more specific, such as that the data follow a normal distribution.

Formally, we denote the time-ordered observations by $y_1, \ldots, y_n$ with potentially m change points at (integer) ordered locations $\boldsymbol{\tau}_{1:m} \equiv (\tau_1, \tau_2, \ldots, \tau_m)$, with $1 \leqslant m \leqslant n-1$. Also denote $\tau_0 = 0$ and $\tau_{m+1} = n$. Within subscripts, we write intervals with the notation $(\tau_{i-1}, \tau_i]$ to indicate that the left point τ_{i-1} is always excluded (open) but the right τ_i is always included (closed) in the current segment. The change points thus split the data into $m+1$ segments $\{\mathbf{y}_{(\tau_{i-1},\tau_i]} \equiv (y_{(\tau_{i-1}+1)}, \ldots, y_{\tau_i})\}_{i=1}^{m+1}$ and a cost is associated with each segment, denoted $\mathcal{C}(\mathbf{y}_{(\tau_{i-1},\tau_i]})$ (see for example Haynes *et al.* (2017a)). The penalty function, which is denoted f, aims to control the segmentation size m and contributes to formulating a penalized minimization problem:

$$\min_{m,\boldsymbol{\tau}_{1:m}} \left\{ \sum_{i=1}^{m+1} \mathcal{C}(\mathbf{y}_{(\tau_{i-1},\tau_i]}) + f(m) \right\}.$$

Common cost choices are the negative log-likelihood (Chen and Gupta, 2000) and quadratic loss (Rigaill, 2015). The penalty is often chosen to be a linear function $f(m) = (m+1)\beta$, with for example $\beta = 2p$ (Akaike's information criterion AIC (Akaike, 1974)), $\beta = p\log(n)$ (Bayesian information criterion BIC, which is also known as Schwarz's information criterion SIC (Schwarz, 1978)), or $\beta = 2p\log\{\log(n)\}$ (Hannan and Quinn, 1979) where p denotes the additional number of parameters that are introduced by adding a change point. We provide a brief overview on the use of different penalties in appendix A (Table A1) of the on-line supplemental material.

The heavy-tailed nature of battle deaths data usually manifests in sampling extremely large values that will dominate the sample statistics, so we expect the standard methods and cost functions of change point detection to fail if they depend on properties that heavy tails do not have. At minimum, then, change point detection on data distributed according to (multiple) power law distributions has ill-defined behaviour if the change point detection is according to whether there is a change in the mean or variance of the empirical distribution.

To cope with the heavy tails of battle deaths data we explore the utility of a non-parametric change point analysis which preferentially considers the tail of the distribution by weighted subsampling when making choices regarding the inclusion of change points (Haynes *et al.* (2017b), section 3.1). A non-parametric approach was first proposed by Zou *et al.* (2014) and then incorporated by Haynes *et al.* (2017b) by means of the empirical distribution (ED) into the dynamic programming algorithm for optimal segmentation search of Killick *et al.* (2012), PELT, thus referred to as ED-PELT. We explore (ED-)PELT with the classical penalty choices that were introduced above, but we also consider the modified Bayesian information criterion mBIC of Zhang and Siegmund (2007) and the change points for a range of penalties algorithm CROPS of Haynes *et al.* (2017a) that explores optimal segmentations across a range of penalties to bypass the disadvantage of ED-PELT of having to supply a value for p. Although ED-PELT (Haynes *et al.*, 2017b) has been shown to outperform competitor methods when mildly deviating from the usual normal distribution assumption for the observed data, to the best of our knowledge none of the standard methods for change point detection (for a recent review see Truong *et al.* (2018)) has been specifically tested on data obeying power law distributions.

3.2. Simulation study

This section performs a detailed exploration of the performance of existing segmentation methods for simulated data that are specifically chosen to mimic the properties that have been documented for historical battle deaths. We concentrate on simulating from (multiple) power law distributions with powers selected to be consistent with those reported in the historical literature (e.g. Richardson (1960) and Clauset *et al.* (2009)).

The wide pool of candidate methods is first narrowed down in Section 3.2.1, and the thorough testing in the subsequent sections leads us to propose a change-point-detection algorithm (algorithm 1 in Section 3.2.3) that is suitable for our context. Furthermore, we carry out a simulation study to investigate the effects of data aggregation on the change points identified (Section 3.2.4) and perform a sensitivity analysis to assess the effects of data normalization (Section 3.2.5).

To compare methods, we consider three metrics: the Hausdorff metric, the adjusted Rand index (ARI) and the true detection rate (TDR). The first measures segmentation by reporting the largest (worst) minimum distance between two points in the true and discovered change point sets (Truong *et al.*, 2018). The Rand index measures (cluster) accuracy by comparing the relationships of data in each cluster in the discovered change point set to the true (Truong *et al.*, 2018). We use the ARI, implemented in `mclust`, to account for clustering due to chance (Scrucca *et al.*, 2017). Total agreement between clusters results in an ARI of 1, whereas the expected value of a random partition of the set is 0. Finally, the TDR gives us an understanding of how many change points detected are true or false by checking to see whether a true change point happened near a detected change point (Haynes *et al.*, 2017b). A TDR of 1 indicates that every change point detected is within a given distance of at least one true change point, whereas a TDR of 0 indicates that every change point is outside such a distance. First, for direct comparison, we consider a radius of acceptance of 0 (Haynes *et al.*, 2017b). To choose appropriate further radii, we consider the historical context—for example, World War I might easily have begun a year or two earlier because of conflict in the Balkans, and the start date of

World War II might easily have varied by a year or two, depending on when the western allies reached the limit of their willingness to accommodate Hitler. On one side of 65.9%, 78.2%, and 77.6% of wars, three, five and eight new wars will have occurred within 1, 2 or 3 years respectively. Hence, we use radii of 3, 5 and 8 to represent roughly 1, 2 or 3 years in the historical data set. We also do not include the end points of the data as change points for this calculation.

Note that the metrics above are effectively trying to measure the quality of our fit by whether there are too many change points (overfitting) or too few change points (underfitting) and whether the change points are in the right location (low bias) or not. We obtain a low Hausdorff metric if there is a detected change point near every true change point, and hence a single change point near a cluster of true change points or vice versa is not highly punished. A low ARI suggests that the segments detected are improperly placed (high bias) or sized (overfitted or underfitted) and is thus a good all-round measurement. A high TDR rewards fits that are very close to correct, whereas a low TDR indicates either overfitting (too many positive discoveries) or incorrect placement (not a high number of true positive discoveries). Of course, these metrics are not all encompassing, which is why we present both the placement of detected change points, above, and the percentage of times that a particular number of change points was identified, below, in the figures that follow.

All simulation tests were carried out in R. In particular, data generation was performed by using the `poweRlaw` R package (Gillespie, 2015), whereas change point analyses were carried out by using the `changepoint` (Killick *et al.*, 2016) and `changepoint.np` (Haynes and Killick, 2019) R packages. As the name suggests, the extension ⋆`.np` in the package name and associated function stands for the non-parametric approach of Haynes *et al.* (2017b). Visuals were compiled by using the `ggplot2` R package (Wickham, 2016).

3.2.1. Initial method screening

To benchmark the various candidate methods, we first screened the possible combinations of cost and penalty corresponding to different data modelling distributions. Table 1 summarizes

Table 1. Function options for changepoint and changepoint.np R-packages†

Function	*Penalty*	*Method*	*test.stat*
`cpt.mean`	SIC or BIC	AMOC	Normal
`cpt.var`	mBIC	PELT	CUSUM (`cpt.mean` only)
`cpt.meanvar`	AIC	SegNeigh	CSS (`cpt.var` only)
	Hannan–Quinn	BinSeg	Exponential (`cpt.meanvar` only)
	Asymptotic		Poisson (`cpt.meanvar` only)
	CROPS		
`cpt.np`	SIC or BIC	PELT	Empirical distribution
	mBIC		
	AIC		
	Hannan–Quinn		
	CROPS		

†The first column corresponds to the R function used, whereas the other three correspond to arguments that determine how the analysis is performed. Not every combination of options within a function is valid: SegNeigh (Auger and Lawrence, 1989) cannot be used with mBIC; PELT, mBIC and Asymptotic cannot be used with CUSUM; PELT and mBIC cannot be used with CSS (Inclán and Tiao, 1994); Asymptotic cannot be used with Poisson; CROPS was designed for use in conjunction with PELT. In particular, `cpt.np` is particularly restricted.

the available functions and options, as implemented in the change point packages above, while noting restrictions on combinations of methods. Some of the arguments that are provided require additional information which we set to be the same across all tests. Specifically, the type I error probability when using the Asymptotic penalty was set to 0.05, the penalty range for CROPS was set to 10^0–10^6, the maximum number of segments in SegNeigh (Auger and Lawrence, 1989) was set to 61 and the maximum number of change points required by BinSeg (Scott and Knott, 1974; Sen and Srivastava, 1975) was set to 60.

We assessed segmentation outcomes across $N = 1000$ trials with data of length $n = 600$ featuring a single change point ($m = 1$) at $\tau_1 = 300$. The first segment consisted of data simulated from a power law distribution with parameter $\alpha = 2.05$, whereas for the second segment we chose $\alpha = 2.55$ (in the range of powers akin to those documented for historical battle deaths). Across our simulations we set the minimum value that is attainable by the power law distribution to be 10 or 1000. We present the results in the former case; the latter case is nearly equivalent and its results appear in appendix B of the on-line supplemental material.

Figs 2–5 give illustrative examples of the types of behaviour of the analyses that were conducted. The bottom subplot of each plot indicates the percentage of trials in which a given number of change points was detected by the analysis. The top subplots are arranged by the number of change points that were found and use boxplots to show the location of each change point so found. The middle broken line is placed along the change point. Across the combinations tested, most failed to identify that there was only a single change point, let alone to pinpoint its precise location.

We also note that Figs 2–5 do not showcase all possible outcomes. For example, some combinations result in approximately correct numbers of change points but incorrect locations. Even when using `cpt.np` overfitting is still common with penalties such as AIC or BIC. PELT and CROPS are also no guarantee of success; `cpt.mean` with PELT, CROPS and a normal distribution results in preferential selection for even numbers of change points, overfitting and placement in the middle of the $\alpha = 2.05$ segment. Of the change point methods, the 'at most one change point' method AMOC (Killick *et al.*, 2016) was naturally most successful. It was tied with itself for second-lowest median Hausdorff measure (39), third-highest median ARI (0.76) and second-highest TDR (0, 0.03; 3, 0.11; 5, 0.16; 8, 0.20). However, it had to be discarded because of its obvious intrinsic restriction. On the basis of these findings from our simulations, we therefore select ED-PELT with CROPS and mBIC to use with the real data (implemented under function `cpt.np` in the `changepoint.np` package). We find appealing not only their strong behaviour but also the lack of parametric assumptions, suitable for our context. ED-PELT's preferential sampling of the tail of the distribution explains its better performance in our power-law-distributed simulated data, i.e. if a change point is detected by ED-PELT with CROPS or mBIC, then there is statistically significant evidence that the segment before the change point has a different power law exponent than the segment after the change point (Haynes *et al.*, 2017b).

3.2.2. Investigation in the presence of at most one change point

For our explorations to be relevant to the real battle deaths data, we choose power law exponents α that are close in value to Richardson's law, and we test the segmentation robustness against numerical proximity, order and false positive detection, as detailed in Table 2. In general, we found that ED-PELT performs well with both CROPS and mBIC penalties, but with CROPS outperforming mBIC in most cases. Both benefit from increased segment lengths with increased precision of the number of change points that are detected and increased ARI (in contrast

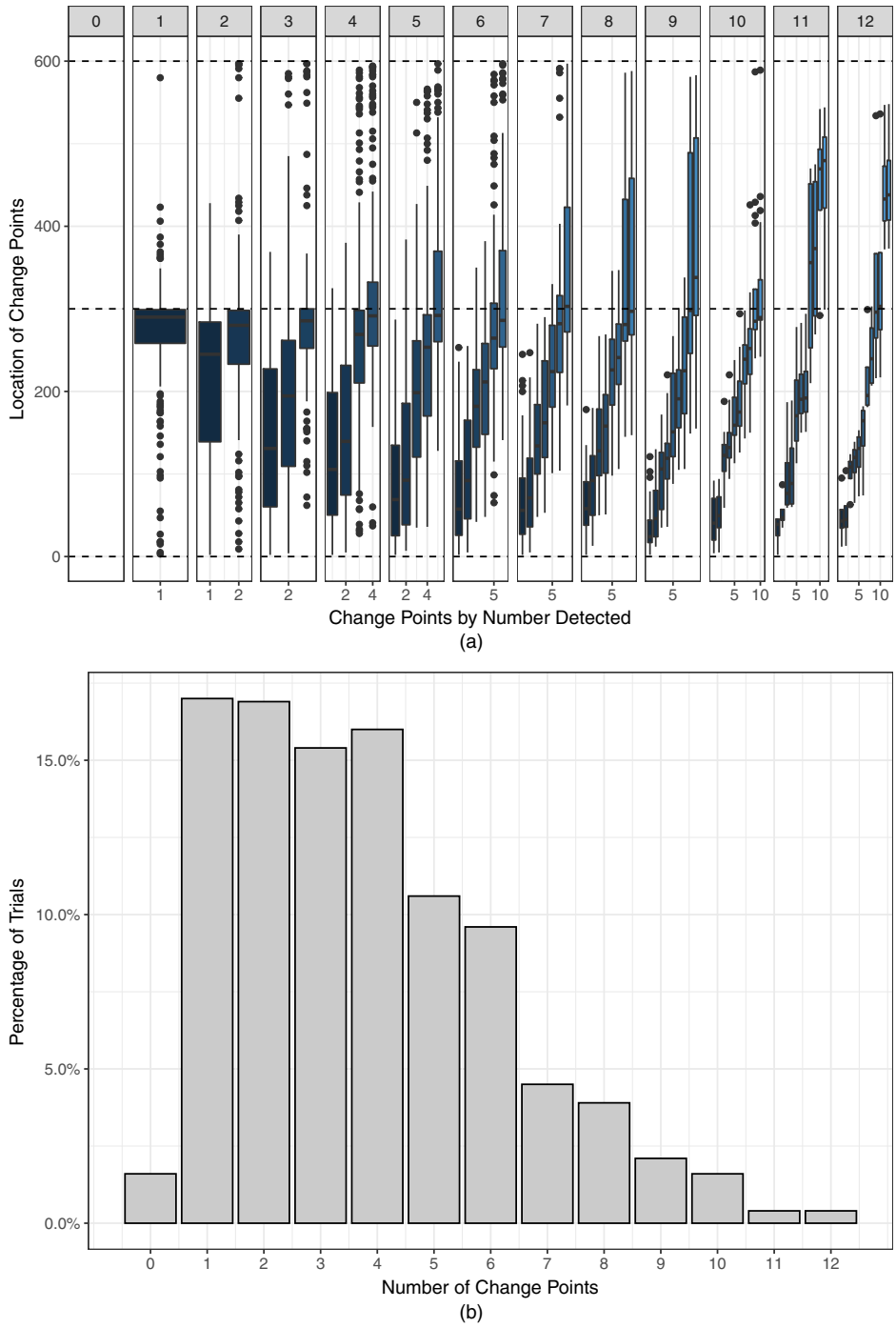


Fig. 2. Test case with better than average behaviour: the simulation is of two power-law-distributed segments of length 300 with exponents 2.05 and 2.55; segmentation is generated by using cpt.meanvar with BinSeg, mBIC and an exponential distribution; although there are good aspects to this finding, the method commonly overfits and tends to assume that change points happen in the $\alpha = 2.05$ segment; this combination has a median Hausdorff of 189, median ARI of 0.64 and TDR 0, 0.01, 3, 0.04, 5, 0.05, and 8, 0.07

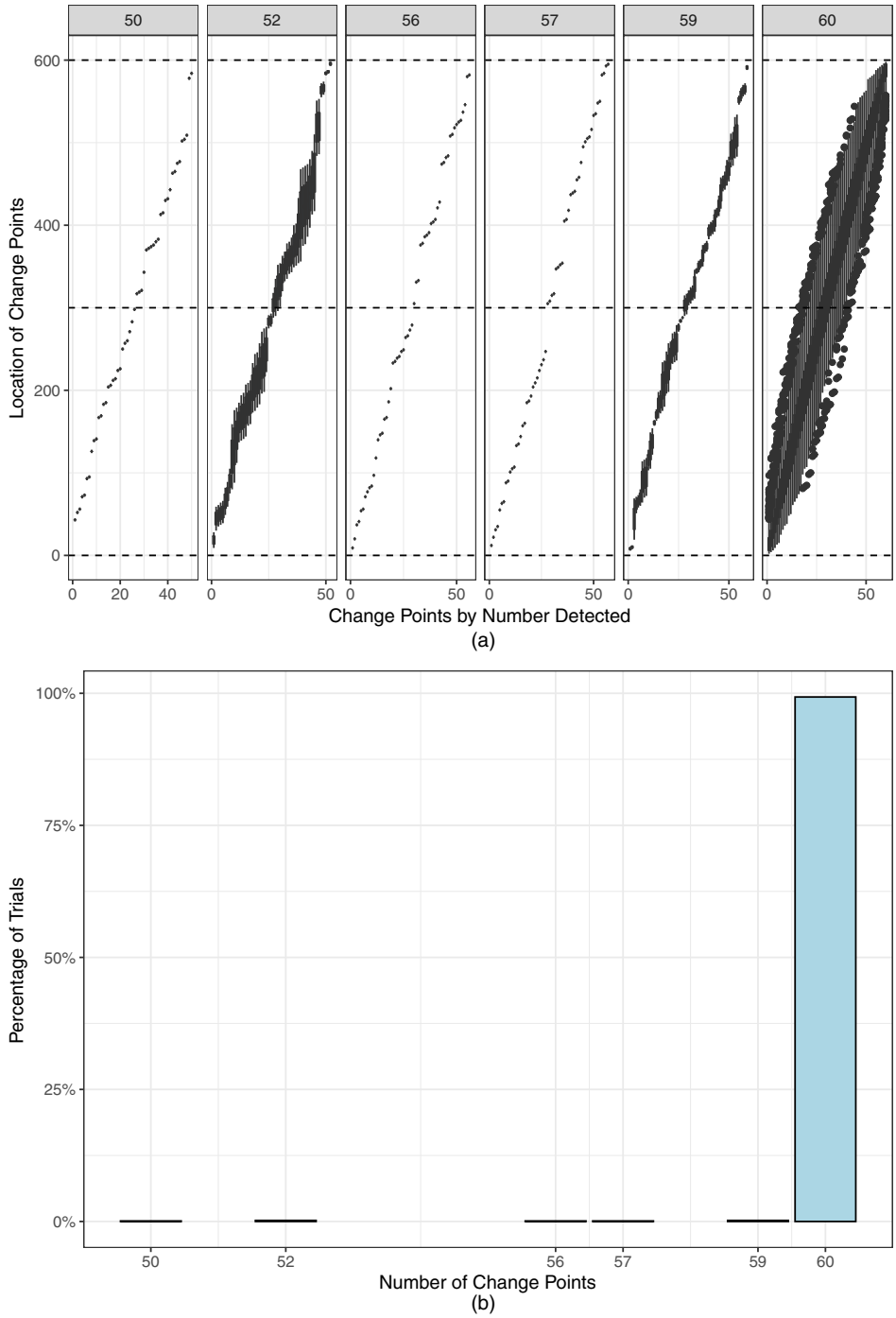


Fig. 3. Test case with worst behaviour: the simulation is as in Fig. 2; segmentation is generated by using cpt.meanvar with SegNeigh, an asymptotic penalty and a normal distribution; results such as this occur with many combinations and can be regarded as failures; many combinations result in more than 10 false positive results and are only stopped by the maximums provided; this combination has median Hausdorff 294, median ARI 0.07 and TDR 0, 0.00, 3, 0.01, 5, 0.01, and 8, 0.01

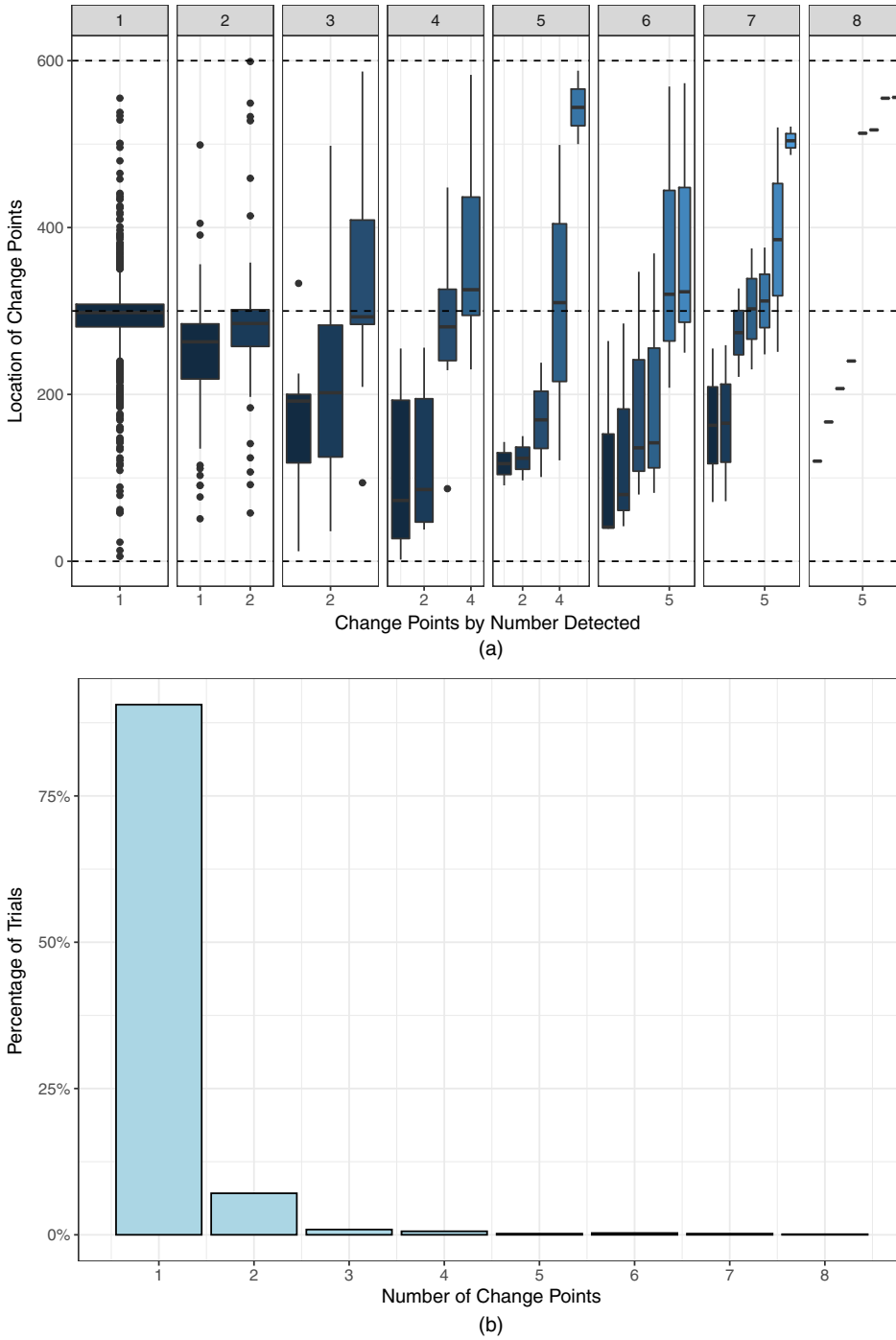


Fig. 4. Test case with best behaviour: the simulation is as in Fig. 2; segmentation is generated by using cpt.np with ED-PELT and CROPS and shows some of the best achievable behaviour; although qualitatively similar to Fig. 2(a), there is improved accuracy in the positioning of the change points and improved precision and accuracy in the number of points so detected; this combination has lowest median Hausdorff 15.50 highest median ARI 0.90 and highest TDR 0, 0.03, 3, 0.17, 5, 0.24, and 8, 0.32

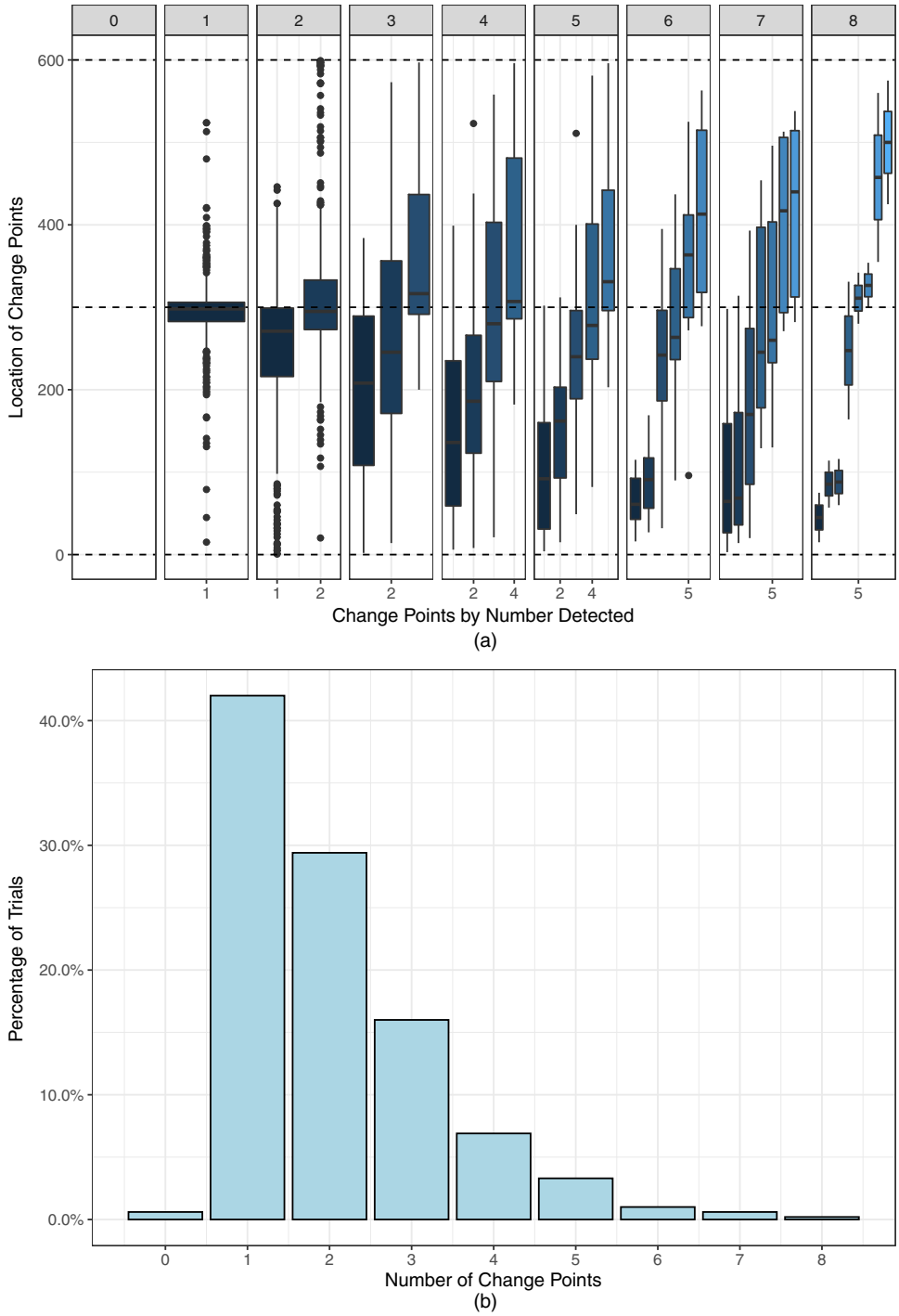


Fig. 5. Test case with second-best behaviour: the simulation is as in Fig. 2; segmentation is generated by using cpt.np with ED-PELT and mBIC; although not as good at detecting change points as CROPS, cpt.np with ED-PELT and mBIC still shows strong potential; this combination has a median Hausdorff 50.5, second-highest median ARI 0.79 and TDR 0, 0.02, 3, 0.09, 5, 0.13, and 8, 0.17

Table 2. Test set parameters†

Exponent α	*Exponent modifier* α_{mod}	*Order*	*Segment length s*
1.7	0	Low–high	30
2.3	± 0.05	High–low	100
	± 0.15		300
	± 0.25		1000
	± 0.5		

†Each column represents a parameter and its options for the simulated data in Section 3.2.2. Each test was performed with $N = 1000$ trials with $m = 1$ change point(s) at $\tau_1 = s$, where $n = 2s$ with generated data with power law parameters $\alpha \pm \alpha_{\text{mod}}$ and $\alpha \mp \alpha_{\text{mod}}$ on each segment.

with the other penalty options, which claim that more change points occur as segment lengths increase). The performance for both is consistent regardless of exponent and order.

However, mBIC does outperform CROPS in one notable situation: when the two distributions are very close, such as $\alpha_{\text{mod}} \leqslant 0.05$, or coincide (no change point). When this occurs, CROPS has a tendency to overfit the number of change points dramatically whereas mBIC is more likely to report correctly no change points.

Fig. 6 shows examples where there is no change point in the data; mBIC can detect this reasonably well, but CROPS dramatically overfits. This unique failure mechanism of dramatic overfitting occurs often: three or more change points are identified in about 63.9% of the trials when the exponent is 1.7, and in about 63.3% of the trials when the exponent is 2.3. In contrast mBIC correctly detects 0 change points 56.1% of the time when the exponent is 1.7 and 54.6% of the time when the exponent is 2.3. This important case means that we cannot rely solely on CROPS to determine our change points. Subsequent results should be viewed through the lens of these limitations.

3.2.3. Investigation in the presence of several change points

We now expand our investigations beyond the presence of at most one change point and explore the outcomes that are obtained when the data feature several (specifically, two, four or eight) change points controlled for variable segment length and data granularity. Fig. 7 shows some representative results of each procedure.

In general, our previous findings extend to the case of multiple change points, as can be seen in Figs 7(a) and 7(c), where CROPS proves to be more precise and accurate in its identification of change points (higher median ARI and TDR; lower Hausdorff distance), although sometimes too conservative. Figs 7(b) and 7(b) illustrate an uncommon case in which the change across one particular change point is so drastic that CROPS identifies it as the only change, missing the less pronounced changes. In contrast, mBIC mostly successfully identifies these change points, showcased in higher ARI and lower Hausdorff distances, albeit with a lower TDR. This uncommon case is more likely to occur when there are a large number of true change points (e.g. 8) and small segment sizes. We thus conclude that one cannot rely solely on one penalty but must use the joint findings of CROPS and mBIC to assess the presence of change points. The combined use of the two methods gives good confidence in accurately detecting the correct number of change points, as well as their location. CROPS findings are accurate when small numbers of change points are found, whereas change points that are found only by mBIC should

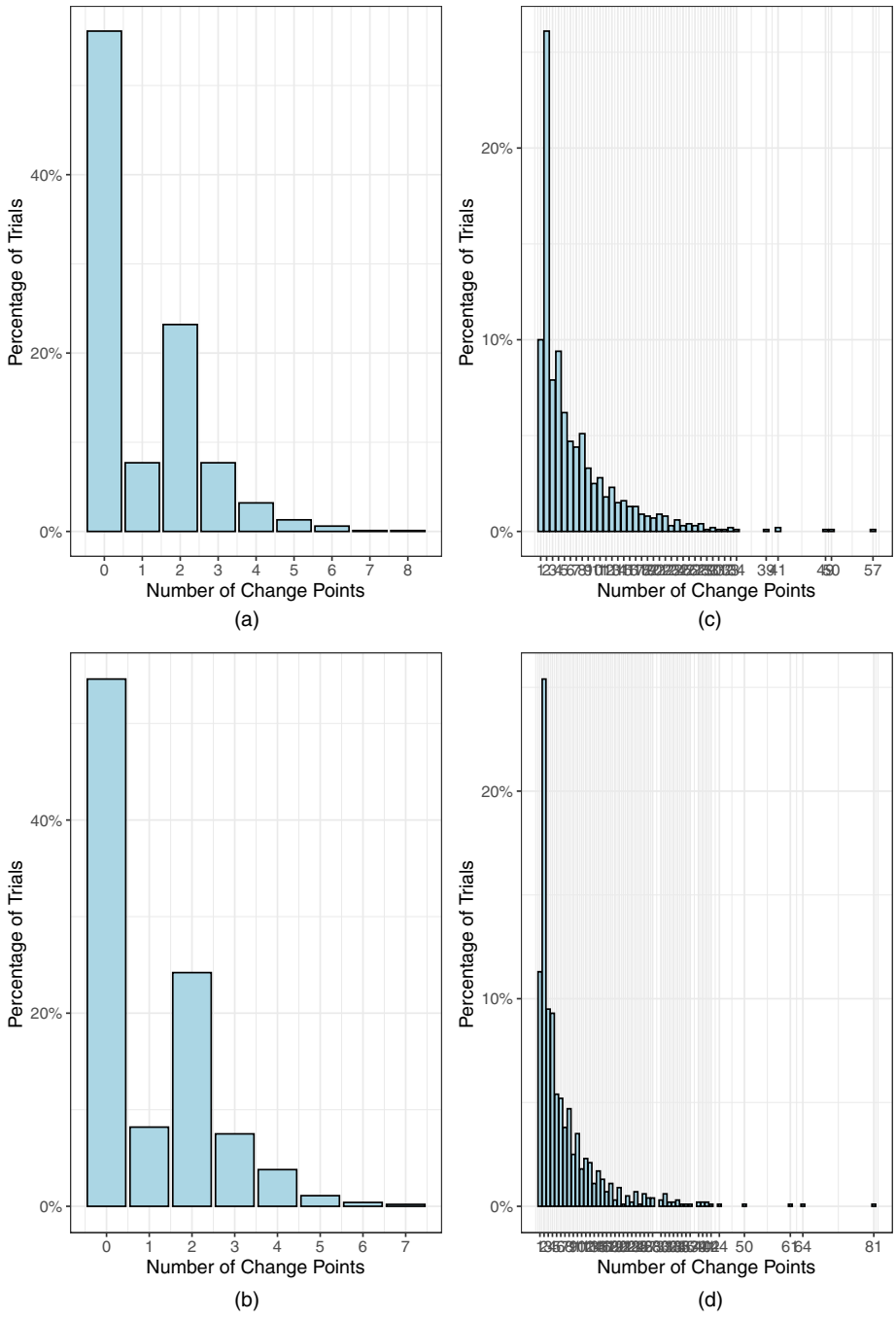


Fig. 6. Examples comparing behaviour of (a), (b) mBIC and (c), (d) CROPS with no change points present: (a), (c) power law exponent 1.7 and (b), (d) 2.3; the sequence length is 600; note that CROPS has pathological behaviour, whereas mBIC succeeds with reasonable precision and accuracy; the behaviour of CROPS is due to a known feature; Haynes *et al.* (2017b) recommended choosing the optimal number of change points for CROPS such that it maximizes the estimated curvature of the penalty as a function of the number of change points; this naturally truncates the data over which the curvature is estimated, removing the possibility of obtaining zero change points on a potentially flat line

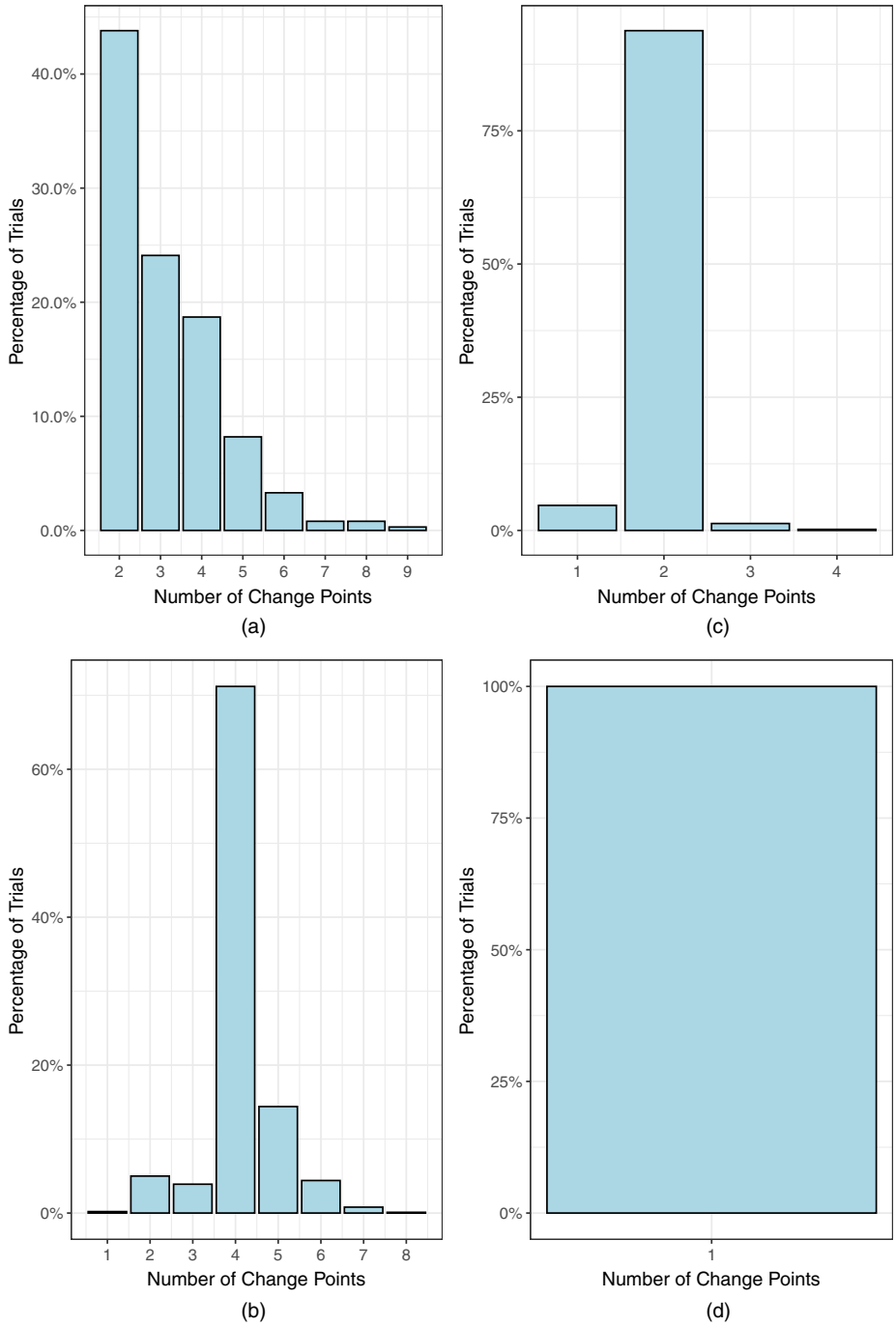


Fig. 7. Examples comparing behaviour of (a), (b) mBIC and (c), (d) CROPS with multiple change points: (a), (c) sequence length $n = 1000$; (b), (d) sequence length 575; in (a) and (c), two change points, marking the power law exponent change from 2.3 to 1.7 to 2.1, are present, and CROPS gives a more accurate result; simulations show that this is the common pattern; in the second case four change points are present, transitioning across exponents 2.87, 1.83, 2.49, 1.67 and 1.06; CROPS detects only a single change point with high precision; mBIC outperforms CROPS in this uncommon case

Table 3. Algorithm 1: proposed change-point-detection algorithm for power law distributions†

Given the time-ordered observations $\mathbf{y} = \{y_1, \ldots, y_n\}$, segment $\mathbf{y}$ by applying ED-PELT with penalty

(a) *mBIC*—denote the estimated set of change points as τ_{BIC};
(b) *CROPS*—denote the estimated set of change points as τ_{CROPS}

If $|\tau_{\mathrm{mBIC}}| = 0$ and $|\tau_{\mathrm{CROPS}}| > 2$, then $m = 0$ and the change point set is $\tau = \emptyset$
Else

(a) set $\tau = \tau_{\mathrm{mBIC}} \cap \tau_{\mathrm{CROPS}}$ and $m = |\tau|$;
(b) for $(\tau_{\mathrm{mBIC}} \cup \tau_{\mathrm{CROPS}}) \setminus \tau$, interpretation is required

†'$|\cdot|$' denotes cardinality.

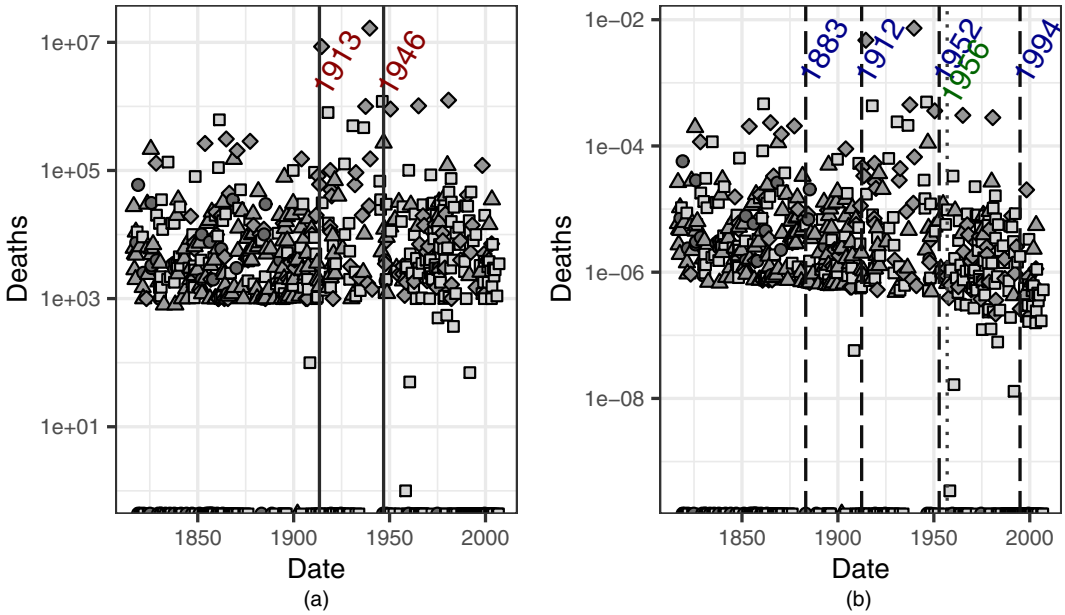


Fig. 8. Results from applying algorithm 1 to the COW data set for all data subsets (△, extrastate war; □, intrastate war; ◇, interstate war; ○, non-state war; ⋮, CROPS change points; ⋮, MBIC change point; |, both change point types): (a) raw data; (b) data rescaled by world population at the time of conflict

be viewed with caution. Of particular note is that mBIC appears to have an extremely low false negative rate: if mBIC does not find a clear break in the data, then we may be confident that no change point is present. Where mBIC and CROPS agree on identified change points, we have a high degree of confidence that this marks a real change of distribution in the data.

In the light of the results above, we propose the following change-point-detection algorithm (algorithm 1 in Table 3) to employ on the real battle deaths data sets. This protects against the pathological CROPS case, resulting in increased TDR when considering the change-point-intersection set, while also allowing for a more liberal interpretation of the union of detected change points.

3.2.4. Aggregation effects

As pointed out in Section 2, data aggregation has often been debated in the historical literature

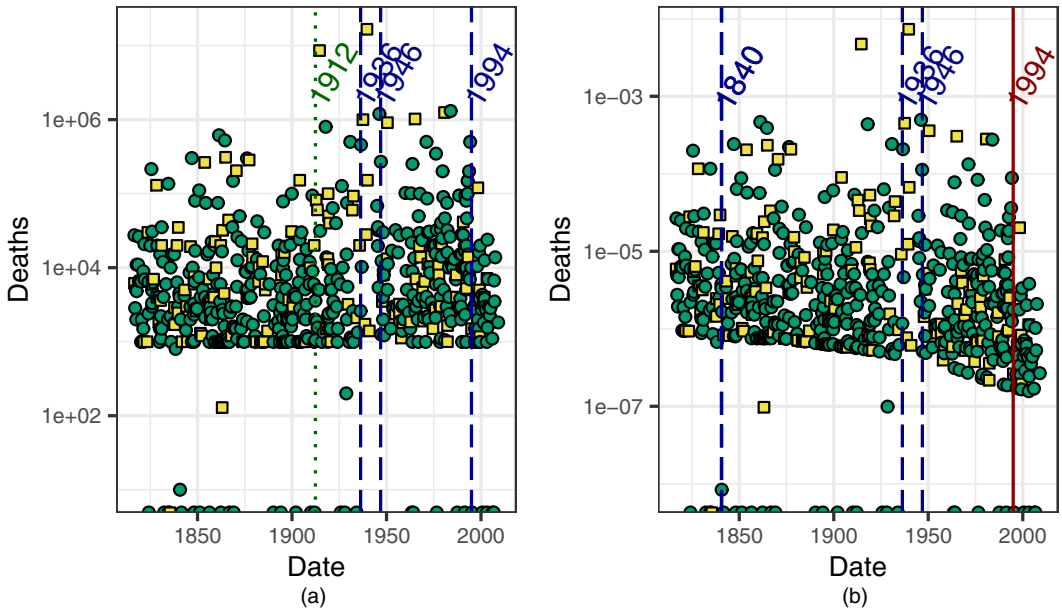


Fig. 9. Results from applying algorithm 1 to Gleditsch's combined data sets (□, interstate war; ●, civil war; ⋮, CROPS change point; ⋮, mBIC change points; |, both types of change point): (a) raw data; (b) data rescaled by world population at the time of the conflict

(Cirillo and Taleb, 2016b; Gleditsch *et al.*, 2014; Clayton *et al.*, 2017). Binning data into time periods affects the size of the tail, which is potentially important in our context as the methods that we use are sensitive to changes in the tail of the distribution (Haynes *et al.* (2017b), section 3.1). Thus for completeness we now empirically address the effects of aggregation of distinct events.

For this, we performed simulations in which we aggregate the observations in consecutive windows of lengths 2 and 4 (the precise details can be found in appendix C of the on-line supplemental material). Our results (e.g. Figs C7–C10 in the supplemental material) indicate an accuracy–precision trade-off, much like one might experience with confidence intervals: larger aggregations appear to make the methods more likely to detect the correct number of change points but yield less information about the precise change location in the original data set. Set in its historical context we might regard this as an intuitive result: the start dates of many wars might easily have varied by a year or two—recall the World Wars—so there is natural imprecision in the historical realization of the underlying transition to war. However, we do not on such grounds advocate aggregation, as we identified some associated disadvantages, e.g. no prior knowledge of the segment size that is appropriate for meaningful aggregation (should it exist), lack of clarity on the use of metrics for quantifying method performance or the potential collapse of legitimate consecutive change points.

3.2.5. Sensitivity analysis: effects of normalizing by world population

To assess fully the effect of data presentation on the change points discovered, we also carried out a sensitivity analysis that highlights how changes in the (normalizing) population interfere with the changes identified in the battle data sets. An outline of the world population data set traits and full simulation details appear in appendix D of the on-line supplemental material. We might naturally expect that the normalization could induce the appearance of additional change points if there is a sudden change in the world population, but our simulations demonstrate

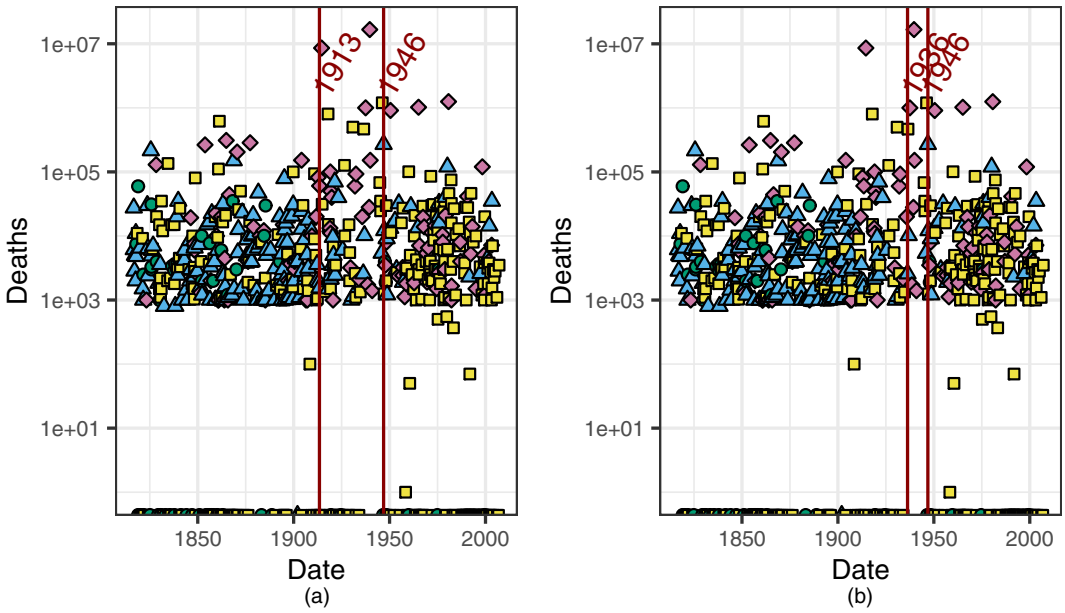


Fig. 10. Results from applying algorithm 1 to the combined COW data set, rescaled as recommended by Cirillo and Taleb (2016b), pages 30 and 32 (△, extrastate war; ◇, interstate war; □, intrastate war; ○, non-state war; ⋮, CROPS change point; ⋮, mBIC change point; |, both types of change point): (a) data rescaled by using the current (2018) world population; (b) data rescaled by using the world population at the time of the conflict

more complex effects (e.g. Fig. D12–D18 in the supplemental material). If a population change is induced near a (simulated) change point, then highly contrasting effects are possible—the change point may dominate, or it may disappear and possibly induce another change point elsewhere. The resulting behaviour depends on several factors such as power law intensities and sharpness of population change. This is not necessarily a problem—if changes in the population balance change in war, so that the probability of dying does not change even as the battle deaths do, this is certainly meaningful. But our analysis demonstrates that the sources of a change point, in the numerator or denominator, cannot be effectively disentangled. Thus the search for change points in raw battle deaths and the corresponding search in normalized battle deaths must be conducted separately and subjected to comparative assessment of the findings.

4. Change point analysis of historical battle deaths

Using the insights that were gained in the simulation study above, we now apply the proposed algorithm to the data sets that were described in Section 2. The results indicate with confidence the existence of change points in the data. In the raw COW data set, which is shown in Fig. 8, there are two changes: just before World War I and just after World War II. When scaled by population two more candidate change points emerge, in the late 19th century (1883) and in 1994 (and the post World War II point shifts slightly), but there is less confidence in the change points overall since the results are not identical across CROPS and mBIC. The further detected post World War II change point in particular is likely to be the result of a sharp change in population in the 1950s. A similar effect can be observed in Fig. D12 in appendix D of the online supplemental material. This supports the proponents of the long peace hypothesis, albeit via an argument for the 'great violence'.

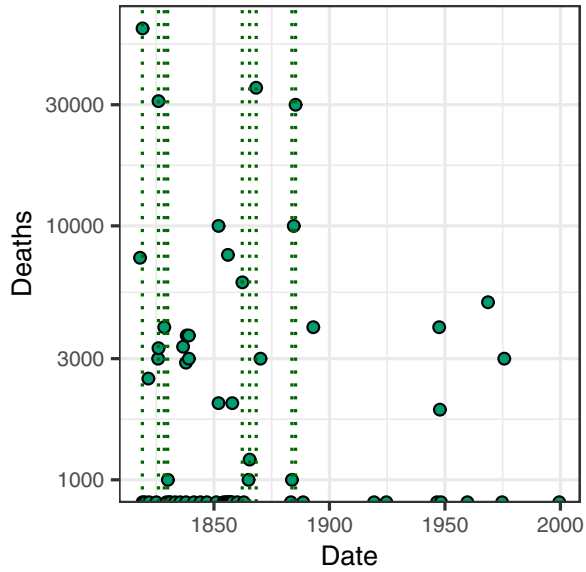
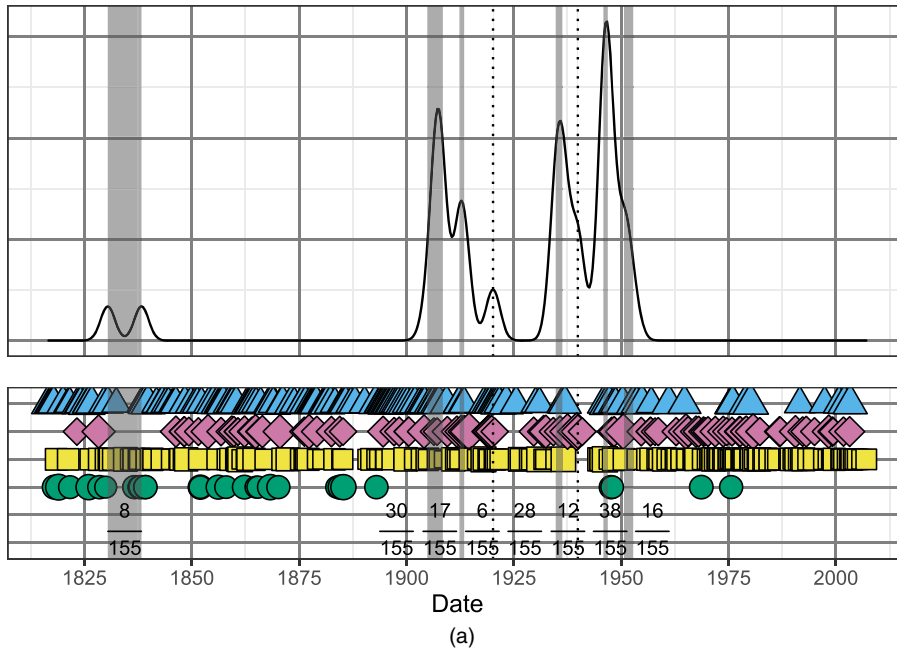


Fig. 11. Results from applying algorithm 1 to the non-state COW data set (△, non-state war; ┆, CROPS change point; ┆, mBIC change point; ┆, both types of change point): no change points are found by the mBIC penalty and CROPS finds a large number of tightly clustered points; note that this result is extremely indicative of no change points; for comparison, see Fig. 6

It is less clear to assign change points in the Gleditsch raw data set, but the emerging 1994 change point in data scaled by population size is now conclusively found (Fig. 9). The broad message is similar, with candidate change points found pre World War I, post World War II and 1994. In contrast with the raw COW data set, we find evidence for change in the 1840s. In addition, the Gleditsch analysis suggests a change in the mid-1930s.

The suggestions that were made by Cirillo and Taleb (2016b), pages 30 and 32, to transform the data to account for the finite upper bound appear to have little effect (Fig. 10). Neither transforming the data to impose a size limit of the 2018 world population on any single war, nor doing so with each event bounded by population at the time of the war, typically changes the number or location of change points, especially in the Gleditsch data set. Among the COW data set and its various subsets, an exception is the combined COW data set, as shown in Fig. 10. The limited sensitivity to such transformations is probably due to the lack of data points that are sufficiently far in the tail of the distribution—no single war results in the death of a high proportion of world population through battle. These results do suggest some sensitivity within the COW combined data set, in that the 1913 change point in the raw data has a similar likelihood of being identified to that of the 1936 change point in the transformed data.

As noted in Section 3.2.5, normalization of power law distributed data by world population can obscure or even eliminate what would be considered change points in the raw data. As such, we must consider the analyses as separate, but we can exploit their relationship to understand what we might reasonably expect if we assumed that one or the other analysis was true. To do so, we carry out *a posteriori* robustness checks by assuming that the raw segmentation is true, fitting a power law to each segment by using the methods of Clauset *et al.* (2009, implemented by Gillespie (2015)), simulating from the non-parametric model, and assessing the change points identified.



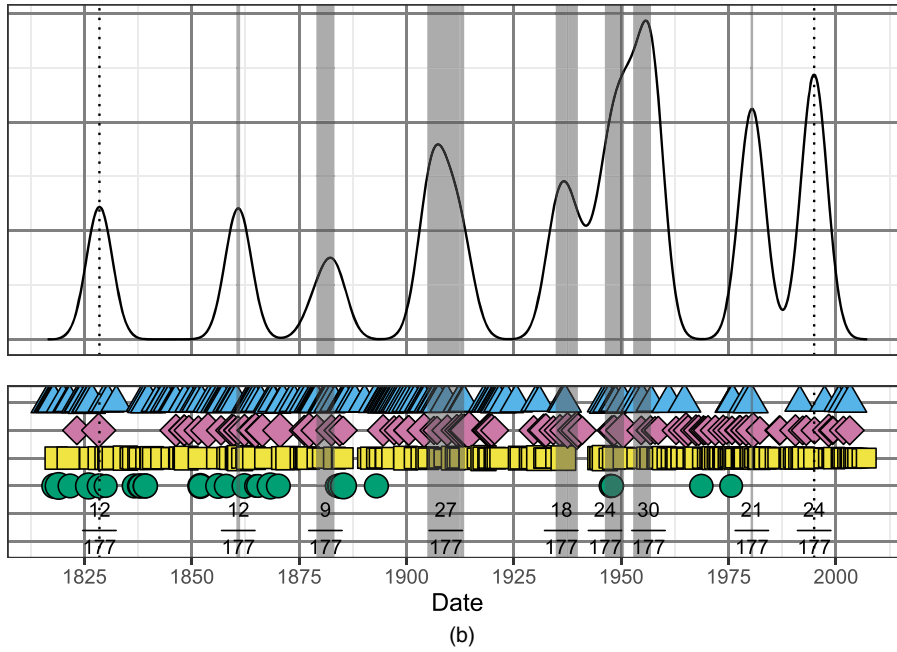
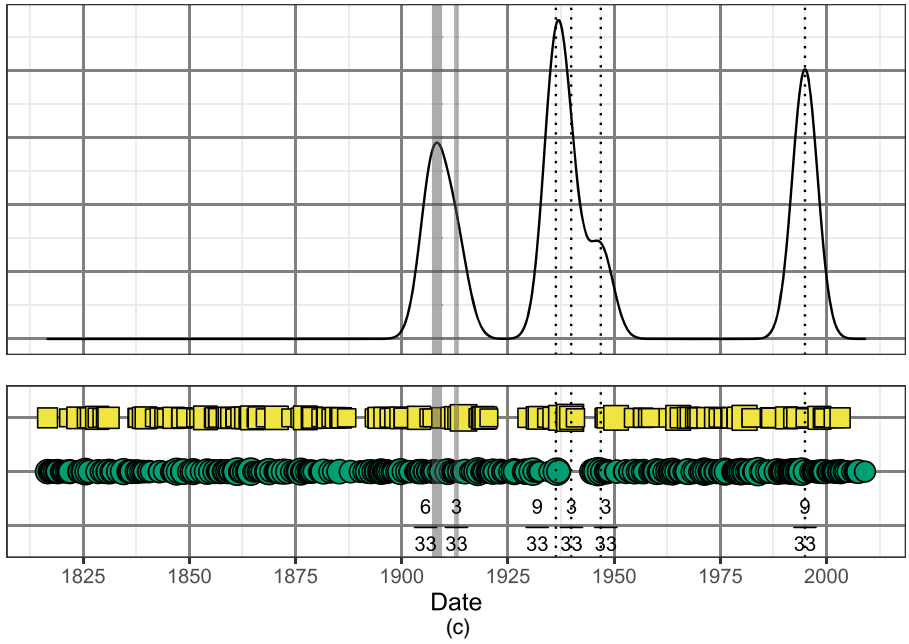


Fig. 12. Results for internal meta-analyses performed on all change points found in any combination of subsets within the data sets (in each plot, there are two images; the lower of each pair of images is a time line of events occurring, sorted by subset; above it is a density estimate of the locations of change points detected; the area under the curve of the estimate is proportional to the probability of finding a change point within that part of the data set; |, ¦, change points, in different locations or the same location respectively, that have been clustered; numbers below the time lines indicate the fraction of identified change points so clustered): (a) COW normalized data (△, extrastate war; ◇, interstate war; □, intrastate war; ●, non-state war; ○, 10^1; ○, 10^3; ○, 10^5; ○, 10^7); (b) COW normalized data (△, extrastate war; ◇, interstate war; □, intrastate war; ●, non-state war; ○, 10^{-8}; ○, 10^{-6}; ○, 10^{-4}) (c) Gleditsch raw data (□, interstate war; ●, civil war; ○, 10^1 deaths; ○, 10^3 deaths; ○, 10^5 deaths; ○, 10^7 deaths); (d) Gleditsch normalized data (□, interstate war; ●, civil war; ○, 10^{-8} deaths; ○, 10^{-6} deaths; ○, 10^{-4} deaths)



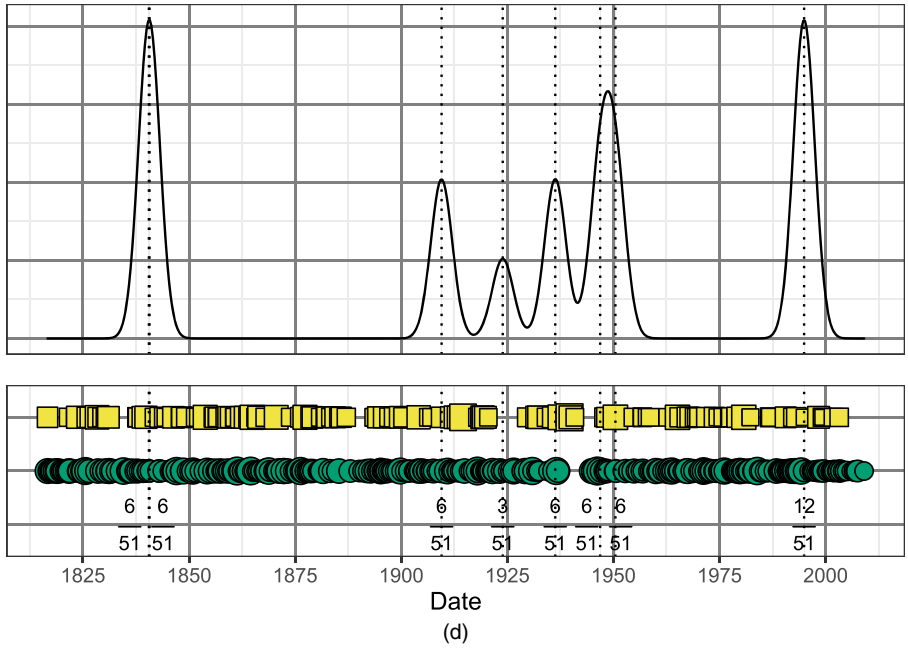


Fig. 12 *(continued)*

This analysis reveals that we can be confident in the results for the COW data set: if the estimated raw change points were true, we would indeed detect an approximate 1910–1916 change point by CROPS and mBIC, as well as an approximate 1944–1950 change point. Similar robustness findings hold for the normalized data; hence if the raw COW results were indeed true, we should expect to detect change points that are similar to those we have discovered.

For the raw Gleditsch data, algorithm 1 does not identify a specific change point because of the failure of CROPS and mBIC to support each other. Instead, we require further analysis and interpretation, and indeed our *a posteriori* robustness checks indicate that our observed results would be extremely unlikely in the absence of a true change point. This behaviour is unsurprising, given that the algorithm aims to protect against false positive discoveries. For the normalized Gleditsch data, we are confident that the 1994 change point has been identified.

Another important check of consistency is whether any real data sets exhibit no change points. We recall that we have already demonstrated that the no-change-points case for our methodology is evidenced by a particular combination of a large number of (false) positive discoveries from CROPS and few or no (false) positive discoveries from mBIC for a wide range of data points, seen in Fig. 6. Although we have established the robustness of the methods against artificial data, the existence of a data set with no change points would clearly help to validate our methods while also identifying a setting that is consistent with the null hypothesis of no change in the statistical properties. It is therefore worthy of comment that such a data set within the COW data does exist: the COW non-state data set, which is shown in Fig. 11, has a response that is clearly of the same type as in Fig. 6, indicating a potential unchanging underlying mechanistic reason for this phenomenon.

To obtain a sense of the robustness of our approach and to represent the overall prevalence of change points, in Fig. 12 we present an internal meta-analysis across all the analyses that we have performed on the COW and Gleditsch data sets. Fig. 12 shows where change points are found in all composing internal data subsets by the proposed algorithm identified in Section 3.2. In the top panel of each subfigure, we place a kernel density estimate of the locations of change points; the subfigures and density estimates were created by using a one-fifth adjustment to the default bandwidth to sharpen the location of change points. In the bottom panel of each subfigure, we present the data subsets as a time line. Shaded regions, for change points over a period of time, and dotted lines, for change points at a single time, indicate the location of clusters of change points that are clustered by using the k-means algorithm (Wang and Song, 2011). The area under the density estimation curve is therefore a rough aggregate measure of the likelihood of a change point during the period, independent of the magnitude of the change. This panel gives a clear sense of the robustness of the 1994 change point, less strongly the robustness of the 1830s change points and the variations that exist in the period 1910–1950 of the change points. The graph shows the location of individual points but also more finely grained variation where multiple methods and data sets produce change points at approximately the same point in time. The R-package `Ckmeans.1d.dp` by Wang and Song (2011) was used for clustering in this context.

5. Discussion

We have shown that recent advances in non-parametric change-point analysis now allow for analysis of heavy-tailed data: an important class of data with unusual properties. Previous methods are prone to detecting too many change points by comparison. Our simulation study demonstrates that no single method fully captures the behaviour of heavy-tailed data, and we

concluded that a combination of analyses more fully addressed the task of detecting change points. In particular, we showed evidence for obtaining the best segmentation results when combining ED-PELT (Haynes *et al.*, 2017b) with CROPS (Haynes *et al.*, 2017a) and mBIC (Zhang and Siegmund, 2007) penalties; moreover, this approach has the notable advantage of carrying no model-specific assumptions.

We emphasize that our approach is purely data driven and we are explicitly not attempting to prove or disprove a particular thesis with our work. The active and important debate about historical battle casualties has been hampered by disagreement over the existence and position of change points and the entanglement of the two strands of argument. In particular, the tendency within the literature to require that any putative change point be supported by an argument for its cause, and even in some cases to go looking for change points to support a hypothesis, creates a real danger of bias. This leads to several issues, not least the potential for skewing the literature towards studies that find no change points. In this context, it is nonetheless appropriate for us to speculate on possible reasons for the change points that we have detected.

Applying our findings to historical battle deaths data, long considered power law distributed (Richardson, 1944, 1960; Clauset *et al.*, 2009; González-Val, 2016; Chatterjee and Chakrabarti, 2017; Clauset, 2018; Martelloni *et al.*, 2018), revealed both new and old insights into how the data may have changed in time. We detected the approximate beginning and end of the 'great violence' 1910–1950 as change points, which are consistent with the idea that the World Wars marked a particularly violent period in human history. A change point in the Gleditsch data in the 1930s might also reflect the complex civil and proxy wars that took place around this time. We also observed possible change points in the 1830s and the 1990s across data sets and data presentations. The former might indicate the gradual change away from the so-called congress era, and the beginnings of the events that led to the revolutions of 1848. The latter change point, around the end of the Cold War, supports the hypothesis that was put forward by Gurr (2000) (see also the work of Cederman *et al.* (2017)).

Our study provides a demonstration of a practical methodology, leveraging recent techniques to provide the best possible answer to whether change points exist in battle deaths data. Additional rigour would require the development of change-point-detection techniques specifically that are designed for power law distributions while retaining the ability to detect multiple change points. Such distributions are of significant potential interest, including diverse areas such as blackouts, book sales and terrorism (Clauset *et al.*, 2009). Furthermore we have not considered the possibility of continuous changes in underlying distributions over time beyond the world population such as those postulated by Pinker (2011, 2018). Our analysis takes an important step forward in answering whether changes exist but stops short of integrating analysis of both continuous and discrete changes. Nonetheless our study provides an essential statistical benchmark: driven by only the features of the data, we have demonstrated that the latest techniques show the existence of change points in well-documented and publicly available data sets of battle deaths.

Acknowledgements

We thank M. Spagat for a seminar at York which seeded the project, and A. Clauset, K. Gleditsch, N. P. Gleditsch, S. Pinker, M. Spagat and the reviewers for comments and discussions. Additionally, parts of this project were performed on the Viking Cluster, which is a high performance computing facility provided by the University of York. We are grateful for computational support from the University of York High Performance Computing Service, Viking and the Research Computing team.

References

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Trans. Autom. Control*, **19**, 716–723.

Auger, I. E. and Lawrence, C. E. (1989) Algorithms for the optimal identification of segment neighborhoods. *Bull. Math. Biol.*, **51**, 39–54.

Beard, S. (2018) Is there really evidence for a decline of war? One Earth Future, Broomfield. (Available from `https://oefresearch.org/think-peace/evidence-decline-war`.)

Bernauer, T. and Gleditsch, N. P. (2012) New event data in conflict research. *Int. Interacns*, **38**, 375–381.

Cederman, L.-E., Gleditsch, K. S. and Wucherpfennig, J. (2017) Predicting the decline of ethnic civil war: was Gurr right and for the right reasons? *J. Peace Res.*, **54**, 262–274.

Chatterjee, A. and Chakrabarti, B. K. (2017) Fat tailed distributions for deaths in conflicts and disasters. *Rep. Adv. Phys. Sci.*, **1**, article 1740007.

Chen, J. and Gupta, A. K. (2000) *Parametric Statistical Change Point Analysis*. Boston: Birkhäuser.

Cirillo, P. and Taleb, N. N. (2016a) What are the chances of war? *Significance*, **13**, no, 2, 44–45.

Cirillo, P. and Taleb, N. N. (2016b) On the statistical properties and tail risk of violent conflicts. *Physica* A, **452**, 29–45.

Clauset, A. (2018) Trends and fluctuations in the severity of interstate wars. *Sci. Adv.*, **4**, no. 2, article eaa03580.

Clauset, A. and Gleditsch, K. S. (2018) Trends in conflict: what do we know and what can we know? In *The Oxford Handbook of International Security* (eds A. Gheciu and W. C. Wohlforth). Oxford: Oxford University Press.

Clauset, A., Shalizi, C. R. and Newman, M. E. J. (2009) Power-law distributions in empirical data. *SIAM Rev.*, **51**, 661–703.

Clayton, G., Kathman, J., Beardsley, K., Gizelis, T.-I., Olsson, L., Bove, V., Ruggeri, A., Zwetsloot, R., van der Lijn, J., Smit, T., Hultman, L., Dorussen, H., Ruggeri, A., Diehl, P., Bosco, L. and Goodness, C. (2017) The known knowns and known unknowns of peacekeeping data. *Int. Peacekeep.*, **24**, 1–62.

Cristelli, M., Batty, M. and Pietronero, L. (2012) There is more than a power law in Zipf. *Scient. Rep.*, **2**, article 812.

Deng, K. (2003) Fact or fiction?: Re-examination of Chinese premodern population statistics. *Working Paper 76/03*. Department of History, London School of Economics and Political Science, London. (Available from `http://www.lse.ac.uk/Economic-History/Assets/Documents/WorkingPapers/Economic-History/2003/wp7603.pdf`.)

Epstein, R. (2011) Review of *The Better Angels of our Nature: Why Violence has Declined*. *Scient. Am.*, **305**, no. 4.

Fenby, J. (2013) *The Penguin History of Modern China*, 2nd edn. London: Penguin Books.

Gaddis, J. L. (1986) The long peace: elements of stability in the postwar international system. *Int. Secur.*, **10**, 99–142.

Gat, A. (2013) Is war declining—and why? *J. Peace Res.*, **50**, 149–157.

Gillespie, C. (2015) Fitting heavy tailed distributions: the powerlaw package. *J. Statist. Softwr.*, **64**, 1–16.

Gleditsch, K. S. (2004) A revised list of wars between and within independent states, 1816-2002. *Int. Interactns*, **30**, 231–262.

Gleditsch, K. S., Metternich, N. W. and Ruggeri, A. (2014) Data and progress in peace and conflict research. *J. Peace Res.*, **51**, 301–314.

Goldstein, J. S. (2011) *Winning the War on War: the Decline of Armed Conflict Worldwide*. London: Penguin.

González-Val, R. (2016) War size distribution: empirical regularities behind conflicts. *Def. Peace Econ.*, **27**, 838–853.

Gray, C. S. (2012) *Another Bloody Century: Future Warfare*. London: Weidenfeld and Nicholson.

Gurr, T. R. (2000) Ethnic warfare on the wane. *For. Aff.*, **79**, 52–64.

Hannan, E. J. and Quinn, B. G. (1979) The determination of the order of an autoregression. *J. R. Statist. Soc.* B, **41**, 190–195.

Haynes, K., Eckley, I. A. and Fearnhead, P. (2017a) Computationally efficient changepoint detection for a range of penalties. *J. Computnl Graph. Statist.*, **26**, 134–143.

Haynes, K., Fearnhead, P. and Eckley, I. A. (2017b) A computationally efficient nonparametric approach for changepoint detection. *Statist. Comput.*, **27**, 1293–1305.

Haynes, K. and Killick, R. (2019) changepoint.np: methods for nonparametric changepoint detection. *R Package Version 1.0.1*.

Hjort, N. L. (2018) Towards a more peaceful world [insert '!' or '?' here]. Department of Mathematics, University of Oslo, Oslo. (Available from `http://www.mn.uio.no/math/english/research/projects/focustat/the-focustat-blog!/krigogfred.html`.)

Huntington, S. P. (1989) No exit: the errors of endism. *Natn. Intrst*, no. 17, 3–11.

Inclán, C. and Tiao, G. C. (1994) Use of cumulative sums of squares for retrospective detection of changes of variance. *J. Am. Statist. Ass.*, **89**, 913–923.

Killick, R., Fearnhead, P. and Eckley, I. A. (2012) Optimal detection of changepoints with a linear computational cost. *J. Am. Statist. Ass.*, **107**, 1590–1598.

Killick, R., Haynes, K. and Eckley, I. A. (2016) changepoint: an R package for changepoint analysis. *R Package Version 2.2.2*.

Klein Goldewijk, K., Beusen, A., Doelman, J. and Stehfest, E. (2017) Anthropogenic land use estimates for the Holocene–HYDE 3.2. *Earth Syst. Sci. Data*, **9**, 927–953.
Martelloni, G., Di Patti, F. and Bardi, U. (2018) Pattern analysis of world conflicts over the past 600 years. *Preprint arXiv:1812.08071*. Interuniversity Consortium for Science and Technology of Materials.
Moyal, J. E. (1949) The distribution of wars in time. *J. R. Statist. Soc.* A, **112**, 446–449.
National Consortium for the Study of Terrorism and Responses to Terrorism (2016) *Global Terrorism Database Codebook: Inclusion Criteria and Variables*.
Pinker, S. (2011) *The Better Angels of Our Nature: the Decline of Violence in History and Its Causes*. London: Penguin.
Pinker, S. (2018) *Enlightenment Now: the Case for Reason, Science, Humanism, and Progress*. London: Allen Lane.
Platt, S. R. (2012) *Autumn in the Heavenly Kingdom*. London: Atlantic Books.
Reilly, T. H. (2004) *The Taiping Heavenly Kingdom*. Seattle: University of Washington Press.
Richardson, L. F. (1944) The distribution of wars in time. *J. R. Statist. Soc.*, **107**, 242–250.
Richardson, L. F. (1946) The number of nations on each side of a war. *J. R. Statist. Soc.*, **109**, 130–156.
Richardson, L. F. (1952) Contiguity and deadly quarrels: the local pacifying influence. *J. R. Statist. Soc.* A, **115**, 219–231.
Richardson, L. F. (1960) *Statistics of Deadly Quarrels*. London: Stevenson.
Rigaill, G. (2015) A pruned dynamic programming algorithm to recover the best segmentations with 1 to k-max change-points. *J. Soc. Fr. Statist.*, **156**, 180–205.
Sarkees, M. R. (2010) The COW typology of war: defining and categorizing wars. (Available from `http://www.correlatesofwar.org/data-sets/COW-war`.)
Sarkees, M. R. and Wayman, F. W. (2010) *Resort to War: a Data Guide to Inter-state, Extra-state, Intra-state, and Non-state Wars, 1816-2007*. Washington DC: CQ.
Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.
Scott, A. J. and Knott, M. (1974) A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, **30**, 507–512.
Scrucca, L., Fop, M., Murphy, T. B. and Raftery, A. E. (2017) mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *R J.*, **8**, 205–233.
Sen, A. and Srivastava, M. S. (1975) On tests for detecting change in mean. *Ann. Statist.*, **3**, 98–108.
Sivard, R. L. (1991) World military and social expenditures, 1991. World Priorities.
Spagat, M. (2015) World War III—what are the chances? *Significance*, **12**, no. 6, 10.
Spagat, M. and Pinker, S. (2016) Warfare. *Significance*, **13**, no. 3, 44.
Spagat, M. and van Weezel, S. (2018) On the decline of war. *Working Paper Series 18*. Centre for Economic Research, University of California at Davis, Davis. (Available from `http://www.ucd.ie/t4cms/WP18.15.pdf`.)
Spence, J. D. (1996) *God's Chinese Son*. London: Harper Collins.
Truong, C., Oudre, L. and Vayatis, N. (2018) Selective review of offline change point detection methods. *Preprint*. École Normale Supérieure Paris-Saclay, Paris. (Available from `http://arxiv.org/abs/1801.00718`.)
Wang, H. and Song, M. (2011) Ckmeans.1d.dp: optimal *k*-means clustering in one dimension by dynamic programming. *R J.*, **3**, 29–33.
Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.
Worden, R. L., Savada, A. M. and Dolan, R. E. (1988) *China: a Country Study*. Washington DC: Library of Congress. (Available from `https://lccn.loc.gov/87600493`.)
Zhang, N. R. and Siegmund, D. O. (2007) A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, **63**, 22–32.
Zou, C., Yin, G., Feng, L. and Wang, Z. (2014) Nonparametric maximum likelihood approach to multiple change-point problems. *Ann. Statist.*, **42**, 970–1002.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Changepoint analysis of historical battle deaths: Supplemental material'.