Quantifying War;

From the Battle of Britain to Terrorism

Brennen Taylor Fagan

PhD

UNIVERSITY OF YORK

MATHEMATICS

DECEMBER 2020.

Abstract

Conflicts are central events in history, defining eras and the lives of their peoples. In this thesis, we demonstrate the use of three quantitative methods to three case studies from historical and conflict modelling. We begin with the application of the bootstrap to the Battle of Britain. The bootstrap allows us to answer counterfactual questions about the Battle of Britain, including the importance of targeting and tactical decisions on the final outcome of the battle. We quantify the final outcome using theoretical prior distributions associated with historical viewpoints. We next conduct a changepoint analysis of historical battle deaths. This requires the adaptation of changepoint analysis methods to heavy-tailed data, for which we formulate an algorithm before applying the algorithm to the case study. We find evidence for changes in the distribution of battle deaths through time. We finish with a case study of coalescence and fragmentation modelling, which has been proposed for insurgent-counter-insurgent conflict. We demonstrate that gel-shatter cycles are a previously unrecognised yet ubiquitous feature of such systems and discuss the robustness of these systems to perturbations in the underlying rules. Together, these case studies demonstrate the ability of modern methods to refine and deepen our understanding of historic conflicts.

Contents

List of tables

List of figures

Preface

After accepting this PhD but before beginning it, I made the usual rounds of saying goodbye to family and friends. This led to perhaps the first academic question from someone about the PhD: why study conflict? At the time, I was confused as the answer had seemed obvious to me: we study conflict to better understand how it happens with the hopes of reducing its danger, much like we study disease. They too were confused for different reasons: 2016 had been a year where many people looked at their friends, families, and neighbours and saw them and their beliefs in a different and more polarised light. On reflection, their question was more akin to the ethical questions we require of many disciplines: are you doing good and how do you make sure you are doing good without (or with minimal) harm?

The material in this thesis covers conflicts at very different levels. We study a prolonged aerial conflict in preparation for an invasion, the sequence of conflicts and their deaths due to battle as a dataset, and a recurring theme and model of modern conflicts. In each case we investigated, I was vaguely surprised by the meta-conflicts that arose. The Battle of Britain is still surrounded by a national myth that led to a very large, mixed public response and the techniques used therein are situated in the middle of the debate on (the validity of) counterfactual history. The changepoint analysis of historical battle deaths was conceived of from the antagonistic tensions amongst those arguing about trends in the historical data and those arguing whether said data has discrete changes. Even the more strictly mathematical chapter on coalescence and fragmentation is in tension with some existing literature, although this tension is perhaps the least dramatic as a natural result of the scientific process.

In each case, we sought to identify truths about the world to the best of our abilities. By the nature of each work, we did no harm directly, nor is harm an obvious result of our studies. We did good by providing tools or evaluating the quality of tools used in scientific study. Our objects of study are not things that can be used to guide a weapon or take a life. Their primary purpose is to aid in decision making about models, implicit or explicit. In studying these tools, we may not have solved conflict, but hopefully we have helped our fellow scientists and decision makers to better understand their own models and how to interact with them.

My co-authors and I would like to acknowledge the support of several academics relating to the chapters of this thesis. We thank T. Lucas of the US Naval Postgraduate School; he provided the initial suggestion for the project discussed in Chapter 3. We thank M. Spagat for a seminar at York which seeded Chapter 4, and A. Clauset, K. Gleditsch, N. P. Gleditsch, S. Pinker, and M. Spagat for comments and discussions. Similarly, we thank S. Connor and G. Delius for comments and discussions regarding Chapter 5. We are also grateful for computational support from the University of York High Performance Computing service, Viking, and the Research Computing team. We also acknowledge the anonymous referees for their feedback on the papers that we have published. Finally, I would like to thank G. Constable and K. Glazebrook for examining the thesis and providing many helpful comments for refining the chapters.

On a more personal note, I would like to thank my advisors for their help over the last four years. Niall MacKay and A. Jamie Wood have been extremely supportive in a rather difficult time and have always made a point of trying their utmost to engage with varied material on a weekly basis despite their own extraordinarily busy lives. Similarly, I should like to thank Stephen Connor and Gustav Delius for their feedback as a part of my advisory panel.

I also would like to thank my fellow students. They have entertained many silly queries and trivial thoughts about their own research and have been a vital sounding board for my own inane thoughts and approaches to problem solving. Thank you again for the additional mathematical grounding, knowledge, and insanity that is far outside my own area of expertise.

I would like to thank my family and friends. You have put up with mathematics that I am not certain that you ever wanted to hear, many days and nights (or nights and days for those in the United States) when I was unable to attend to you, and long stretches of time where I was irritable when I could pay attention to you.

Finally, I would like to thank everyone in York for the patience and welcome you have given the American in his and his family's time here. It is naturally terrifying to come such a far way from the place one is born, especially in these trying times, but you all have made every effort to help us acclimate.

Declaration

I declare that this thesis is a presentation of original work and is a mixture of published and unpublished material. Chapter 1 is drawn from a series of unpublished book chapters as well as various drafts of papers and conversations with advisors and collaborators Prof. Niall MacKay and Dr A. Jamie Wood, with whom the works following were all co-authored, as well as Drs Horwood and Price of York St. John. No claims of originality are held for Chapter 2; some of that material is also drawn as necessary from the papers used in the following chapters. Chapter 3 draws heavily from work originally published in 2020a in the *Journal of Military History* with authorship shared by Fagan, Horwood, MacKay, Price, Richards, and Wood. The primary contribution by the author of this thesis was the construction and running of the bootstraps, as well as the summarisation of the results of the bootstraps. Material in Chapter 4 was originally published in 2020b in the *Journal of the Royal Statistical Society, Series A* with authorship shared by Fagan, Knight, MacKay, and Wood. The primary contribution by the author of this thesis was the simulation study, the creation of the algorithm, the investigation of the datasets, and the application of the algorithm to the data. Chapter 5 includes work currently under review; authorship is shared amongst Fagan, MacKay, Pushkin, and Wood. The author of this thesis participated in the mathematical analysis of the model and is primarily responsible for varying simulations of the model, conceptualising and implementing rules variations to study robustness, and statistically analysing the resulting data. This work has not previously been presented for an award at this, or any other, university. All sources are acknowledged as references.

1 Introduction

We all know that art is not truth. Art is a lie that makes us realize truth, at least the truth that is given us to understand. *– Picasso, 1923*

Human history is filled with conflict, whether one merely examines the last two hundred years or whether one stretches back to the beginnings of our species. Even in just the last two hundred years, battles in wars alone, neglecting those due to poverty, famine, disease and the like, resulted in approximately 40.5 million deaths (*e.g.* Sarkees and Wayman, 2010). At the same time whenever two or more nation -states wish it, the human species could be dealt destruction that dwarfs the losses of our past. Yet clearly, despite such threats, the world continues to survive humanity's conflicts with the human species (more or less) intact. Many authors have tried to grapple with the nature of conflict and its place within both history and human nature. Philosophers aplenty have debated whether mankind was some brutal savage whose only salvation can be found in the Hobbesian state or whether instead the state was responsible for all the brutalities that the innocent suffered. Such debates continue in a variety of forms to this day of course; it would be remiss for a discussion of conflict to not mention the conflict of the discussions, even if by pen rather than sword. If one needs an example, one need only consider the controversy surrounding the writings of popular scientist Steven Pinker on how humanity is improving itself to be, among other things, less violent [2011; 2018]. Despite his data-driven approach, critics have attacked Pinker from a variety of directions with more or less vitriol (*e.g.* Epstein, 2011; Cirillo and Taleb, 2016a).

Pinker has written down a (largely statistical but qualitative in mechanism) model for why violence has declined, one which can be discussed, replicated, and disputed with attention to data, methods, and conclusions. Regardless of criticisms of his model, such as cohesiveness, predictive power, or qualitative nature, the model is written down for all to examine, compare and contrast (*e.g.* with the military horizon model of Turney-High (1949)), and debate (*e.g.* Cirillo and Taleb (2016b)). Whenever someone tries to determine how something does or might work, they construct a model, knowingly or not. The challenge is to write down the most explicit model possible so that each relationship, such as cause and effect, can be examined. As Epstein (2008) points out, explicit models enable some of the best scientific behaviours that we can ask for. Explicit models allow us to better explain and convey our beliefs about what is going on, better enabling collaboration with others and questioning of the models. Such models can also guide us to better understand how the phenomenon happens, “illuminat[ing] core dynamics”. The mere fact that Pinker has written down his model has enriched the debate about the decline of conflict (or lack thereof). Much like what has occurred with other datasets (*e.g.* the debate about the possible decline of battle deaths when comparing and contrasting the Correlates of War and Peace Research Institute Oslo Battle Deaths datasets: Harrison and Wolf, 2012; Gleditsch and Pickering, 2014; Harrison and Wolf, 2014), differing researchers can examine the model or dataset and come to differing conclusions and discuss the results (Spagat, 2015; Cirillo and Taleb, 2016a; Spagat and Pinker, 2016).

Of course, this thesis is also about modelling various conflicts, rather than just about the possible decline of conflict. Perhaps the most natural place to start is immediately prior to World War I. At the time, the soon-to-be belligerents were engaged in a build-up of arms for the ensuing war and were equally interested in how best to apply their arms, knowing that every advantage would be needed. In parallel in the United States (Chase, 1902; Fiske, 1916), France (Baudry, 1914), Russia (Osipov, 1915), and Britain (Lanchester, 1916), a simple model of combat was worked out for when two forces came into conflict with each other. This family of models is now conventionally known as Lanchester’s Laws. The central idea is that when two forces come into contact with each other, they instantaneously begin attacking and deal damage to the other side in proportion to their weapons’ effectiveness and some power of their numbers. For modern warfare, the power is taken to be 1, reflecting the ability to aim and thus concentrate modern firepower to eliminate the enemy. For ancient warfare, the power is taken to be 0, reflecting instead a set of duels where combatants cannot readily exploit their numbers. This naturally leads to a set of differential equations with a conserved quantity, for which the former case is known as the square law and the latter as the linear law. Due to the model’s simplicity, it has been readily extended and adapted to numerous contexts, especially within warfare (*e.g.* spatial models: Spradlin and Spradlin, 2007; González and Villena, 2011), but also without (*e.g.* markets: Campbell and Roberts, 1986).

Once one has a model, generally one wants to see how well it matches reality and investigate it to determine what lessons one can learn from the model. Perhaps ironically given Lanchester’s original goal of understanding aerial combat (Lanchester, 1916), Lanchester’s Laws do not actually perform well for modelling aerial campaigns (Johnson and MacKay, 2011; MacKay, 2011; Horwood *et al.*, 2014). This is a problem we must return to in Chapter 3 where we discuss the Battle of Britain. Nevertheless, Lanchester’s Laws still offer lessons about the asymmetries between attacking and defending air space and the importance of the defender’s ability to choose how to respond to the attacker. The parallel inventors, on the other hand, sought to apply the model to naval combat (Chase, 1902; Baudry, 1914; Fiske, 1916) or battles on the ground with mixed unit types (Osipov, 1915). These pursuits have in general been more fruitful for adaptation and usage. For example, MacKay *et al.* (2016) applied a stochastic interpretation of Lanchester’s Laws with approximate Bayesian computation to calibrate and than analyse the likelihood of historical outcomes at the Battle of Dogger Bank in World War I. A classic extension of the ground battle by Deitchman (1962) combines the aimed fire case (square law) with the unaimed fire case (which instead has casualties proportional to both forces, but also produces a linear law as in duels (Lanchester, 1916)). In Deitchman’s work, the insurgents are able to aim effectively, while the counter-insurgents have to guess their enemy’s location, leading to a mixed conserved quantity. This allowed Deitchman to explain the dynamics between the forces and advise how to tilt the system in favour of one side or the other. He compared his model’s parameters to those of then recent insurgencies, including Vietnam, and observed the relationship between the force ratio and the success of the insurgents. This is not the first time that Lanchester's Laws were used to inform the military[1], MacKay *et al.* (2016) noted that Admiral Jellicoe, the commander of the British Grand Fleet, had observed the popularity of the Lanchester's Laws to Lanchester in 1916 and the model was also used to debate and justify the sizes of navies in the Washington Conference of 1921-22 (Kahn, 2004).

Whereas Lanchester's Laws tell us about the application of force, the works of Lewis Fry Richardson, FRS [1960a; 1960b] tell us more about how force accrues and the consequences of its application. Richardson was a pacifist polymath who investigated weather, numerical analysis, geography, fractals, psychology, and war. He was inspired and horrified by the atrocities of World War II and sought to model how such a terrible war could occur and if this represented some new normal in humankind's propensity for "deadly quarrels", his preferred term. He believed that these questions needed to be rationally investigated for patterns and explanations, rather than governed by the horrors of the recent wars. To investigate how the wars might occur, Richardson used the concept of an arms race: two (or more) countries are in tension with each other and wish to equip themselves to defend against a first strike. Both countries have some estimate of the amount of arms the other has and build up their arms in proportion to their would-be opponent's, but their current level of arms have costs to maintain and replace. These are taken to be linear terms in a system of differential equations augmented by a constant level of friendliness (if inclined to disarm) or grievances (if inclined to build up arms). Standard techniques in dynamical systems can be used to analyse how the system then progresses for given initial conditions and parameters, leading to stability, disarmament, or an ever expanding arms race which could break out into a large scale war (Richardson, 1960a).

Richardson's work on deadly quarrels is a much more statistical work, in comparison to his work on arms races, but no less grand in scale. For deadly quarrels, Richardson (1960b) put together an impressive collection of conflicts from 1820 to 1949, ranging from murders to large-scale wars, sorted by the magnitude of deaths caused for each event. This allowed him to consider statistical models and tests of humanity's violent output. For example, he considered simple changepoint analyses, in which one looks for the presence of statistical changes in the behaviour of the data. He conducted an analysis for 1885, which divided the data into two consecutive 65 year intervals (p. 141), and found that there was not significant evidence for a change in frequency. Similarly, he checked if there might be a change in frequency, controlling for magnitude and time, after 1920 which was the point of largest contrast (p. 142) and found that wars appeared to be more frequent afterwards. He also is the first to identify the distribution of deadly quarrels as power-law, a type of distribution which we discuss more thoroughly in Section 2.1, but which motivated the study of the logarithm of the magnitude of deaths, rather than the deaths themselves. His dataset is an important, albeit vastly more general, ancestor of datasets like the Correlates of War dataset (Sarkees and Wayman, 2010).

Due to his explicit treatment of his models and assumptions within his work, retained from his background as a physicist, Richardson's theory can be re-examined (e.g. Wilkinson, 1980) or extended. Recent letters in *Significance* (Spagat, 2015; Cirillo and Taleb, 2016a; Spagat and Pinker, 2016) inspired work to extend Richardson's naive changepoint analysis. Clauset (2018), Spagat and van Weezel (2018) and Hjort (2018) all tried to investigate whether large scale conflicts

[1] Deitchman discussed his life history and advising of the United States military with Sheldon *et al.* (2010).

had changed in magnitude. All three proceeded with acknowledgement to Richardson's power-law distributions but each applied different methods and came to differing conclusions. Clauset (2018) relied on simulations, Spagat and van Weezel (2018) applied probability, and Hjort (2018) used parametric changepoint models[2]. The first of these failed to reject the null hypothesis for no change in battle deaths, but the latter two found significant evidence against this hypothesis, stoking the recent debate. It is thus in the spirit of Richardson that we will present Chapter 4, in which we contribute to the debate of whether conflicts have changed since the early 1800s using modern non-parametric methods which only require us to assume that the tail of the distribution is an important trait of the battle deaths.

Lanchester and Richardson are not the only researchers to try modelling conflicts with techniques from mathematics and physics. Many other models have been proposed to deal with the asymmetries of modern warfare. We concentrate on the abilities of insurgents to create terrorist attacks. It has been shown that such events are consistent with power-law distributions with an exponent parameter of around 2.5 (Clauset *et al.*, 2007, 2009b). This observation alone is suggestive of various models for generating power-law distributions; classic references include Bak (1996), Mitzenmacher (2004), and Newman (2005). For example, Clauset *et al.* (2007) propose a model, repurposed from the work of Reed and Hughes (2002), in which first a terrorist (group) plans an attack which can then be stopped by a state. Their central assumptions are that the severity of an event is proportional to the exponential of the amount of time required to plan the event as well as that the probability of the event's success is proportional to the amount of time required to plan it. The assumption of two exponentials in tension is sufficient to yield a power-law whose exponent depends on the specific rate constants for the growth of events and their detection and subsequent failure. Indeed, Clauset *et al.* (2010) propose another alternative pair of competing exponentials where population density at a location and the "attractiveness" of the location to terrorists are supposed to have roughly exponential distributions. These types of explanations are particularly attractive as models due to their simplicity as well as their agreement with statistical analysis suggesting that the size of a terrorist organisation and the severity of events that it produces are not related to each other (Clauset and Gleditsch, 2012).

In contrast to the models proposed by Clauset and collaborators, we examine, especially in Chapter 5, a model of coalescence and fragmentation that has been used by Johnson and collaborators to model the size of terrorist attacks (e.g. Johnson *et al.*, 2005; Bohorquez *et al.*, 2009; Johnson *et al.*, 2016). The general concept is quite simple: a population of insurgents seek to combine forces to make the largest events possible, but exist in a hostile environment that might cause them to break-up and go into hiding. This model again produces an appropriate power-law distribution, and has been compared to event data from multiple conflicts as a form of validation (Bohorquez *et al.*, 2009). Johnson and collaborators adapted this model from the work of D'Hulst and Rodgers (2000)[3], who used it to model herding in financial markets. For comparison, financial agents are proposed to coordinate by trading information and break-up when the shared information is used in a transaction.

Putting aside the appeal of this model for "dynamical analogies", *i.e.* its variety of applications to other contexts (Epstein, 2008), this model is of particular interest for its relationship to graph

[2] Hjort's work was later published with co-authors with more Richardsonian parallels (Cunen *et al.*, 2020).

[3] This chain of citations does not appear to include von Smoluchowski (1916) who first conceived of this type of model. Instead, citations stem back to the random graph theory of Erdös and Rényi. Aldous (1999) explains the connection from coalescence and fragmentation to random graphs.

(equivalently, network) theory. The connections between individuals in society can be thought of as a graph (in the sense of nodes and edges) along which information, goods, diseases, *etc.* can spread. In the context of graph theory, the coalescence and fragmentation model of well-mixed individuals can be thought of as a graph where each individual is connected to every other individual and coalescence represents agreements to coordinate amongst individuals or the spread of some other state throughout the graph. A natural generalisation is then to consider inhomogeneous graphs, but, in order to do so, we must have a full and complete understanding of the base model's behaviour and how this behaviour changes in the presence of perturbations. Ideally, such a model is robust to small perturbations in the underlying rules, similar to how the models of Reed and Hughes (2002) used by Clauset and collaborators depend only on exponential forms (Clauset *et al.*, 2007, 2010). Despite the apparent disconnect between the data and the size-severity (Clauset and Gleditsch, 2012), we expect the analysis of the coalescence and fragmentation model to be a key step in understanding networks of terrorist support, such as those studied by Johnson *et al.* (2016). A natural result of checking for robustness will determine if Bohorquez *et al.* (2009) suffered from a street-light effect in which one looks only in places one knows well. Bohorquez *et al.* claim that the observed data matches with a specific form of coalescence and fragmentation model, but the data are drawn from various conflicts with different underlying rules for engagement. Is this just good fortune in that Bohorquez *et al.* chose precisely the right model and parameters that match the data, or is the model robust to variations in its parameters and such data could reasonably emerge from many such variations?

In each of the example models given so far, the models are necessarily applied, calibrated, and tested against data, but the nature of data in conflicts and, more generally, history is quite different from the reproducibility of a laboratory setting. No event happens exactly the same way twice in a historical setting. Each event is contingent on those that came before it and there is very little chance for, for example, combat to be truly independently and identically distributed (IID) as one finds commonly assumed in statistics. It would be hard to claim that terrorists the world over were uninfluenced by the effects of the September 11 attacks on the World Trade Center, nor does it seem reasonable to assume that World War I and World War II were unrelated. How then does one surmount such a problem?

One simple way is to suppose the data is sufficiently IID to model them as such anyway. This is not such a naive approach: a factory machine undergoes wear and tear as it produces its products and eventually can break down without proper maintenance, yet it is still natural to attempt to model its products as produced in an IID way. We take this approach in Chapter 4 with each war's battle deaths assigned to the start date of the war. This approach is taken to prevent improper aggregation or disaggregation. As we discuss in the chapter, we must rely on consistent data handling in order to maintain this assumption. If the data handling is inconsistent, *e.g.* if two wars are separated or combined inappropriately, then we risk introducing artefacts. For example, if wars are divided up by year instead of being considered discrete events, autocorrelation would be induced in the dataset (Beard, 2018). Beard (2018) also determines that standard controls for autocorrelation are not particularly influential in a regression analysis of battle deaths. This suggests that using an IID assumption for battle deaths of wars, despite historical justification otherwise, can be appropriate for modelling. In contrast to our work, one could consider the use of secular cycles in war and peace (Turchin, 2007) similar to the use of Kondratieff waves (Goldstein, 1985). These cycles help alleviate the need for an IID assumption. One possibly interesting direction for future research might be to reconcile these two ideas. For example, are there sufficiently many of these cycles between state actors that the resultant wars give the appearance of independence over time? This is a parallel long timescale problem to the short timescale result which we discuss in Chapter 5, where we find natural cycles arising from the stochasticity in coalescence and fragmentation systems that are described in terms of power-law distributions. This might be one interesting place to look for a dynamical analogy as well as to check if dynamics on the short time scale might drive those of the long.

For Chapter 3, we use similar arguments, but we run afoul of a related issue. Whereas we argue that battle deaths in war and the days in an aerial campaign are sufficiently IID, such an assumption is far harder to justify for the actions of leaders which are only taken once for each scenario, especially when we seek to examine how these actions could have been different. There is no obvious factory analogy, nor conditions that might imply the usage of an IID assumption. Instead there is but a single datum from some unknown distribution. This is actually the subject of acrimonious debate amongst modern historians under the term counterfactual. The concept is simple enough. From our perspective, in our present, history appears to be a single continuous, if complicated, narrative in which events naturally lead up to the present. Looking into our future, many possible paths can unfold however. Both small changes and large can occur: one could choose to eat a different meal for breakfast or take a different route to work, or one could choose a different career or different partner. But of course, historical actors make precisely these sorts of decisions within their own presents, for which the results are only now known in the futures that the actors created. One might wish to analyse the results of the decisions the historical actors made by considering the repercussions of the decisions not taken, from which comes the term counterfactual. Determining the ripples resulting from going forward is where the debate begins.

Evans (2014) presents the standard historiographic view: counterfactual analysis is a limited tool for historical analysis akin to speculation that at best merely reproduces work that can be done within the standard historical paradigm. At worst, counterfactual analysis instead results not just in the introduction of the bias of the author, but the complete capture of the work by said bias. Evans contends that the result of the analysis performed is merely whatever utopia (or dystopia) the author wishes. Counterfactual analysis should only ever be used in the most minimal capacity possible as a result with the acknowledgement that a small amount of counterfactual analysis must occur to describe why history did not take another course.

Despite the warnings of Evans, counterfactual analysis still has a tantalizing allure. Part of the allure is due to ease and human nature, for who among us has not contemplated how life might have been different if only we had taken the other course of action. A larger part stems from the fact that "small" counterfactuals feel as if they should have large repercussions: what if Archduke Ferdinand's driver had taken a different turn and prevented the Archduke's happenstance assassination prior to World War I, what if German leader Adolf Hitler had gotten into art school, or, for an even more modern example, what if Democratic presidential candidate Gore had continued to contest Florida in the 2000 United States presidential election? How can we safely evaluate the importance of events, without assuming that all events merely reduce to variations around the observed reality to which these variations must surely converge? It certainly is true that not all decisions matter, *e.g.* whether one takes the tuna mayo sandwich over the ploughman's, but some seemingly inconspicuous events, *e.g.* taking a wrong turn or getting rejected by a school, certainly could. How can we safely apply counterfactual analysis to determine whether events are consequential without being consumed by our biases?

Part of an answer to the question of the usability of counterfactual analysis is provided by Megill (2004). Megill draws a line between "exuberant" counterfactuals and "restrained" counterfactuals. The former is akin to "the world of historically-based game-playing", in which one begins right before a critical event and begins to speculate as to what if the factual decisions were replaced with counterfactual decisions, compounding forward until a new "virtual history" has been written. On the other hand, restrained counterfactuals can be thought of as moving backwards, rather than forwards, in time. One begins with a fixed outcome and then explores how such an event might have not come to pass, usually by examining each candidate change and how it influences the event under consideration. This is meant to be a measured thought experiment, similar to adjusting experimental conditions in a lab setting. For Megill, the question is not the perhaps philosophical one of agency and freedom or convergence and determinism. It is about determining the necessary causes for an event, for which the event would not have happened if one such predecessor event (or sufficiently many of a set) had not occurred as well.

How do we investigate counterfactual decision making, which we are interested in in Chapter 3, in a restrained manner? These thought experiments must be structured to obey rules, and scholars such as Tetlock and Belkin (1996) have studied how to do so. One such way is to combine existing theoretical models with critical junctures, events which decisively determined the proceeding course of history (Capoccia and Kelemen, 2007); we adopt this style in Chapter 3 using the concept of bootstrapping that we discuss in Section 2.2, but other models can be used such as game theory. The six key attributes of Tetlock and Belkin (1996) for maintaining a consistent and rational counterfactual connect the why and how of modelling with the processes of history (Epstein, 2008). First, they list clarity, by which they mean that we must be explicit in identifying our model, its assumptions, its inputs, and its outputs. Second, there must be a logical consistency, *e.g.* one must not assume that an actor is simultaneously weak and strong. Third, there must be historical consistency: in examining the counterfactual, we must also determine how much of the past must be rewritten to arrive at the counterfactual. The more of history that must be rewritten, the more tenuous the counterfactual itself. Fourth and fifth, there should be theoretical and statistical consistency. For example, we would expect actors to behave rationally and not to rely on having won the lottery or other low probability events. Finally, the lessons taken from the counterfactual should be projectable and ideally falsifiable. It is inevitably hard to maintain all six of these principles, which may end up in tension with another (*e.g.* statistical consistency can come into conflict with the other principles if the unlikely did actually happen as in the work of MacKay *et al.*, 2016), but it is necessary to do our utmost to maintain them in order to properly consider counterfactual scenarios.

A critical juncture to a mathematician is perhaps best described in terms of dynamical systems. Consider a system, like a Richardson's arms race selected such that either disarmament or exponential growth of arms may occur for different initial conditions. We hold the parameters constant, the level of grievances, the cost of arms, and our reactions to our opponents arms, as they represent the state we are in on a long time scale. Such things would have needed to be changed long in our past and prepared for far in advance of the situation itself as they operate on far longer time scales than our immediate day-to-day decision to arm or disarm. Without stochasticity, or foresight for that matter, we simply continue on our course, choosing to arm or disarm in reaction to our perception of our opponent's decision. If we are near the edge between the arming and disarming regimes, however, a critical juncture will arise. In the dynamical systems view, this juncture looks like a saddle node, a point that attracts trajectories in at least one direction, but also repels in a different direction. Around this point, many trajectories, with possibly quite varying starting conditions, will become very near to each other, before rapidly separating as they leave the area around the point. Precisely one trajectory will approach the point itself: this trajectory separates those that end in disarmament from those that end in arms race. This is a critical juncture in that a slight change in our decision making can allow us to move to a nearby curve, of which there are many, and end up in varying end states. Most changes will not affect the qualitative outcome of arms race or disarmament, but they can affect the final numbers of arms on both sides. On the other hand, some decisions at the crucial moment could send us over that critical trajectory allowing us to choose a different outcome. The presence of saddle nodes, and other regions where trajectories become tightly packed, are analogous to critical junctures in that they naturally balance the ideas of historical forces creating a single history against the idea that every event mattered. Most events, *i.e.* those far from those regions of dense trajectories due to saddle nodes or similar features, do not have a significant earth-shattering impact on the final outcome, but some do.

Where do we work with critical junctures in this thesis? The answer is different in each chapter. Their influence is most directly felt in Chapter 3. In this chapter, we must naturally handle the problem of finding the Battle's critical junctures. There are two types that are implicitly investigated therein. The first is aspects of the preparations for the battle. These include fairly explicit "great man" counterfactuals, where the personality of Hitler is made substantially more aggressive and decisive than he was in the actual battle or where Luftwaffe leader Reichsmarschall Hermann Goering has different beliefs about how best to defeat the British forces, as well as counterfactuals regarding the amount of information available to the German forces. Great man theory lived and died in the 19th century (Carlyle, 1840), and it is worth noting criticisms such as those of Spencer (1896, p. 31): "Before [the great man] can remake his society, his society must make him." and that "he is powerless in the absence of... his society." Nonetheless, it would be a mistake to discard the personalities of those involved completely from the analysis, for they clearly had a role (on both sides!) of the Battle of Britain.

The second type of critical juncture investigated in Chapter 3 is the condition for launching the invasion by which we judge the success or failure of the German forces.[4] This critical juncture is one that is documented as having not occurred. It is conceivable that such a meeting and decision could have happened, but its absence is a conspicuous point that needs addressing. Due to its absence, we need to investigate the counterfactual causes of its occurrence and settle on the conditions that would trigger an invasion. While doing so gives our theoretical model a historically plausible method with which to trigger invasion, we do not have any good information as to how far we historically were from our counterfactual invasion. As such, we instead ask the reader to supply their own beliefs which we can then feed into the model. Three theoretical historians are then used to finish the model, with which we can then simulate counterfactual aerial campaigns.

In Chapter 4, we look for changepoints, which are conceptually similar to the *effects* of a critical juncture. Indeed, one might even claim that changepoint analysis within historical data is merely the data-driven search for critical junctures. Statistically, changepoints are the point at

[4] The criticality of this decision could be debated, as does Forester (1971). He has a counterfactual historian state "[I]t is hard to escape the conclusion that Hitler's decision to attempt the invasion [of Britain] was most important in shortening the war and hastening his own destruction." On the other hand, if it had shortened the war, it could be argued that it was critical for the (quantitative) consequences of the war (*e.g.* lives or economics) but not the qualitative result.

which there is a measurable change in the distribution, comparing data before the point to data after the point. Here, we use a statistic that emphasises the influence of the tail of the battle deaths distribution. Whereas the critical junctures of Chapter 3 are those which could change the conclusion of a campaign, Chapter 4 is about the evolution of wars over time. Given the time span and the global dataset, it is hard to know whether broad historical forces make the changepoints found inevitable or if a critical juncture is directly responsible, akin to whether the decline of slavery in the United States was actually prevented by the invention of the cotton gin or not (and thus influencing the subsequent civil war or not) or whether the fall of the Soviet Union was guaranteed by historical forces or a result of United States President Ronald Reagan. This inexactness means that we have no sense nor control for possible delays: we know when the effects of the proposed critical juncture emerge in the dataset, but we have no idea as to the what or when of the critical juncture that caused the effects. While we cannot be absolutely certain about the exact cause of a changepoint, the detection thereof provides historians with lampposts to guide their searches for factors, whether they be historical forces, major events, or critical junctures that lead to changes on a global scale.

Finally, we turn to Chapter 5, which has the most theoretical relationship to critical junctures. Here the model we consider is theoretical in that we are not applying it to real world data, but instead examining properties of the model through mathematical analysis and simulation. The critical junctures are the decisions which determine the rules of the model. Do the counter-insurgents choose to pursue clusters of small size? Do the counter-insurgents prioritise clusters of larger size? How often do the counter-insurgents attempt to eliminate clusters of interest? Similar questions exist for the insurgents: *e.g.* do the insurgents try to aggressively recruit the population? In a reality where the model is true, one would have a military deciding how to engage with insurgent forces, with each rule's variation representing the differing approaches. This is analogous to the usage of war-gaming to make sure that the right decisions are made on the field.

Modelling power-law distributed or historical data is not an easy task, and combining them does not make the matter simpler. As we shall see in Chapter 2, we will need specific tools chosen for the their applicability to the tasks at hand. Power-law distributed data has quirks to its behaviour that need to be incorporated into the model. Not only are large events common in power-law distributions, but combining samples from such distributions is not a trivial matter. At the same time, we rarely have a full day-by-day knowledge of distributions of armies and their interactions. The usage of the bootstrap allows us to circumvent this issue, so long as the data does not change enough to fail to be considered IID, as we shall discover. This technique also allows us to, in a restrained manner, investigate counterfactuals, so long as they are sufficiently similar to the data we factually observed. We also prepare for the interaction of power-law distributions and historical data with a discussion of changepoint analysis; the presence of power-law distributions within the historical data necessitates the adoption of modern methods which we discuss. We finish with a discussion of the development of coalescence and fragmentation models, which prepares us to study their repercussions on the results of Bohorquez *et al.* (2009).

The most natural starting point for our contributions to the literature is the most self-contained and earliest conflict that we study: the Battle of Britain (1940) in Chapter 3. The Battle of Britain is an excellent example of the application of an explicit model to counterfactual historical analysis. In this chapter, we present a simple statistical model called the bootstrap in order to create counterfactual campaigns, but the bootstrap alone is insufficient to move very far from factual history. The counterfactuals it makes are necessarily explorations of the variance in the distribution of combat. Supplemented with historical analysis, we are able to motivate an innovative change in the model which enables the discussion of counterfactuals further from reality and steeped in the decisions of the German military machine. We use these counterfactuals to better understand whether the Germans could have won the Battle of Britain. To our knowledge, this is the first usage of the bootstrap in historical or counterfactual analysis, and we are excited for its possible application to other datasets.

In Chapter 4, we turn to a much wider temporal scope and instead study the time series of battle deaths from wars. Here, an all-encompassing model of conflict would be a heroic, but likely foolhardy endeavour. Instead, we choose an approach that is motivated solely by qualities that we know about the data. Richardson (1960b) has already shown that we should expect battle deaths to be power-law distributed, and recent literature suggests that there is a possible discrete change involving the World Wars, after which a period called the "long peace" emerged (Gaddis, 1986). These pre-existing models and debates motivate us to study the data using changepoint analysis calibrated to power-law distributed simulated data. This provides us with a tool that we are confident can detect changepoints with accuracy, and which we then use to analyse whether there are changepoints in historical battle deaths. We find that these do exist in most, but not in all, subsets of the datasets under consideration and conclude that there is evidence that supports the long peace hypothesis.

Whereas the temporal scope of Chapter 4 was wider, the operational scope of Chapter 5 is significantly larger, covering a model that has recently been used for modelling insurgent dynamics but has historically been used for a variety of more physical purposes. This model of coalescence and fragmentation is assumed to be a model for how insurgents cooperate and share resources with each other in order to attack larger targets. Questions naturally arise: if this model is true, how do the insurgent dynamics evolve over time? How best can we intervene to slow or stop the insurgent organisations? We address these questions and find that, contrary to expectations in the literature, stochastic cycles or steady-state size distributions form depending on parameters and that the system is not quite as robust to interventions as it initially might seem.

We conclude this thesis with a discussion of the relationships between the various chapters and how they weave together to tell us about the various conflicts that we have studied. These relationships naturally provide ideas for new projects to follow. While we have not closed the proverbial book on any of the problems we have considered, the introduction of new techniques and mathematical perspectives will be useful to our fellow researchers.

2 Methods

> *The craftsman there, the smith with that metal of his, with these tools, with these cunning methods, – how little of all he does is properly* his *work! All past inventive men work there with him; – as indeed with all of us, in all things.* *– Carlyle, 1840*

2.1 Power-law distributions

Most distributions are fairly well-behaved. They possess means, variances, and obey the Central Limit Theorem, converging to a normal distribution if there are enough samples. Not all distributions are so well-behaved. In this section, we discuss power-law and related distributions, which are well known for not behaving like “common” distributions. This ill-behaviour has repercussions for the fitting and modelling of such distributions. For most of this section, we use the continuous power-law distribution, but most of this discussion holds equally well for the discrete power-law distribution.

The best way to contrast common distributions with exponential tails, *e.g.* normal distributions, with those that are heavy-tailed, *i.e.* distributions that decay slower than an exponential function such as power-law distributions, is by example. Crook (2006) and Clauset and Wiegel (2010) provide one such example: human height is approximately normally distribution and, as such, knowing the mean and standard deviation well-characterises the limits of human height. On the other hand, income is heavy-tailed. If human height were to follow a power-law with approximately the same average as the factual normal distribution, then we would have “nearly $60{,}000$ individuals... at 2.72 meters”, the tallest human on record, “10,000... as tall as an adult male giraffe, 1 as tall as the Empire State Building (381 meters), and 180 million... a mere 17 centimeters tall.” Distributions with exponential tails are unlikely to result in data far from the well-defined mean, while distributions with heavier tails will often have comparatively huge values. In the case where there is no well-defined mean, these large values will dominate the sample mean as sampling continues until the next more extreme value is sampled. Comparing cumulative distribution functions using the `poweRlaw` R package (Gillespie, 2017; R Core Team, 2018), a standard normal distribution produces values larger than 1.64 with probability approximately 0.05, but a

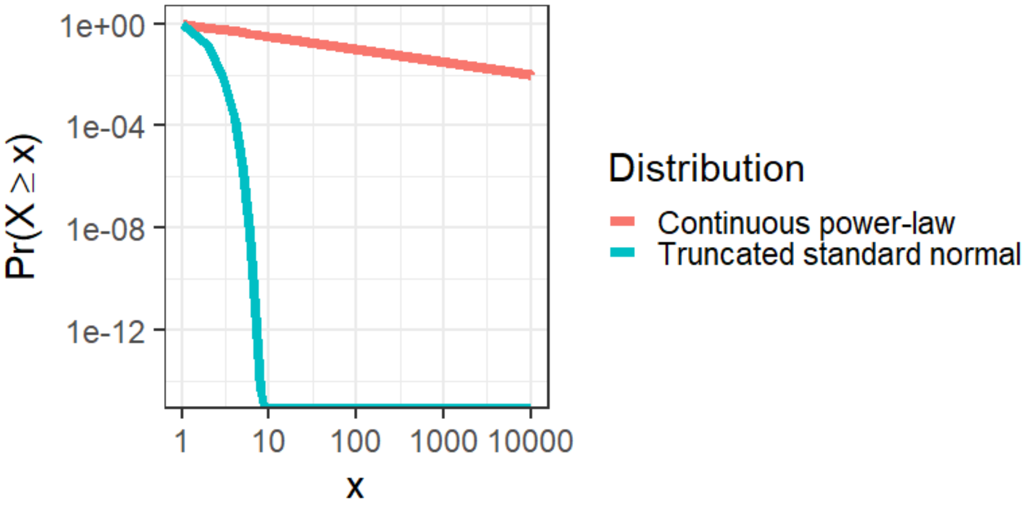


Figure 2.1: A comparison of the heavy-tailed continuous power-law distribution with $\alpha = 1.5$, $x_{\text{min}} = 1$ and a truncated standard normal distribution beginning at $x_{\text{min}} = 1$. We plot the complementary cumulative distribution on log-log axes. Exponentially-tailed distributions show a characteristic drop-off towards 0, while the power-law distribution instead follows a straight line with slope $\alpha - 1 = 0.5$ on log-log axes.

continuous power-law with exponent 1.5 beginning at 1 produces values larger than 400 with the same probability. The same standard normal produces values larger than 5 with probability of 3×10^{-7}, but the aforementioned power-law would be producing values larger than 10^{11} with that probability.

Power-law distributions have the scaling $\Pr(X \geq x) \propto x^{-\alpha+1}$ for $x \geq x_{\text{min}}$, where α is the scale parameter and x_{min} is the location parameter that determines the start of the distribution. The tail of a power-law distribution, *i.e.* the behaviour as $x \to \infty$, dominates exponential tails. This is easiest to see on log-log axes, as in Figure 2.1, where we contrast a truncated standard normal distribution beginning at $x_{\text{min}} = 1$ with a continuous power-law with $\alpha = 1.5$, $x_{\text{min}} = 1$. Note that Figure 2.1 shows the complementary cumulative distributions $\Pr(X \geq x)$. The power-law distribution follows a straight line with slope $\alpha - 1 = 0.5$ on the log-log axes, while the truncated normal quickly falls off to infinitesimal values.

The probability density functions for power-law distributions are fairly straightforward to express. For $x \in \mathbb{R}$ and $x \geq x_{\text{min}} \geq 0$,

$$p(x|\alpha, x_{\text{min}}) = \frac{(\alpha - 1)}{x_{\text{min}}} \left(\frac{x}{x_{\text{min}}} \right)^{-\alpha}, \tag{2.1}$$

while for $x \in \mathbb{N}_{\geq 1}$ and $x \geq x_{\text{min}} \geq 1$,

$$p(x|\alpha, x_{\text{min}}) = \frac{x^{-\alpha}}{\sum_{i \geq x_{\text{min}}} i^{-\alpha}}. \tag{2.2}$$

Unless the distribution is truncated from above, it is required that $\alpha > 1$. Otherwise, the functions will fail to have the standard probability mass property (*e.g.* $\alpha = 1$ for the discrete distribution invokes a truncated harmonic series). Similarly, higher moments require higher values of α: the nth moment requires $\alpha > n + 1$ in order to exist. The mean thus requires $\alpha > 2$ while a well-defined variance requires $\alpha > 3$. These properties are precisely why the sampling behaviour discussed at the beginning of this section is so strange in comparison to distributions with exponential tails. While this behaviour is not limited to power-law distributions, other distributions are heavy-tailed as well, power-law distributions are useful exemplars and are used throughout this text. While we usually use the full distribution, sometimes we use only a power-law distributed tail when fitting, as discussed below. Finally, we note that $\log \frac{x}{x_{\min}}$ follows an exponential distribution, which is far better behaved.

Many authors have been inspired by graphs like Figure 2.1 to fit power-law distribution to data by fitting a straight line on log-log axes. Unfortunately, the appearance of a straight line on log-log axes is necessary but not sufficient to conclude the tail of a dataset is power-law distributed (Clauset *et al.*, 2009b). Maximum likelihood estimation (MLE) can, of course, be used. Given a realised data set $\mathbf{x} = (x_1, \ldots, x_i, \ldots, x_n)$,

$$\hat{x}_{\min} = \min_i x_i \qquad\qquad \hat{\alpha} = 1 + n \left(\sum_{i=1}^{n} \ln \frac{x_i}{\hat{x}_{\min}} \right)^{-1}. \tag{2.3}$$

The discrete equivalent requires the solution of a transcendental equation, although numerical maximisation of the likelihood can be used (Clauset *et al.*, 2009b). Unfortunately, we are often interested in where the tail of the distribution begins, while the MLE $\hat{x}_{\min}$ simply uses the minimum of the dataset to begin fitting the exponent in the power-law distribution. (Similarly, if a distribution is truncated above, the maximum of the dataset is the MLE maximum.) When $x_{\min}$ is a known quantity, the MLE $\hat{\alpha}$ has useful properties. It is, for instance asymptotically normal and approaches the true α as n grows, and its standard error can be calculated and is of order $n^{-1/2}$ (Clauset *et al.*, 2009a).

A more rigorous approach for estimating the value of $x_{\min}$ when interested in a power-law distributed tail, which we refer to as Kolmogorov-Smirnov maximum likelihood estimation (KSMLE) for reasons that will become rapidly apparent, is supplied by Clauset *et al.* (2009b). Broadly speaking, their approach is as follows:

1. For each candidate $x_{\min}$, calculate $\hat{\alpha}$.

2. Choose the $(x_{\min}, \hat{\alpha})$ pair that minimises the Kolmogorov-Smirnov (KS) distance between the cumulative distribution functions of the data and the model beginning at $x_{\min}$. (Note that the KS distance is the maximum distance between two cumulative distribution functions (Clauset *et al.*, 2009b).)

3. Test the best power-law model by comparing the obtained KS distance against those of artificial data sets. The artificial data sets are drawn using the fitted model's parameters and the data below the tail and then refitted using steps 1 and 2. The original power-law model is rejected if its KS distance is less than 10% of the KS distances of the artificial data sets. In terms of statistical tests, the null hypothesis is that the power-law is consistent with the data, and this is rejected, *i.e.* statistically significantly different, at the 0.10 level.

4. If the power-law model was not rejected, compare it to alternative hypotheses using likelihood ratios.

There are other, subtle ways of encountering problems when working with power-laws. Power-law distributed sampled sets cannot be trivially combined or sub-sampled even when samples are drawn from the same distribution without distorting some of the properties within the set. Cristelli *et al.* (2012) demonstrated problems of this type using sets that are expected to follow Zipf's law, power-laws with $\alpha = 2$. If one has a power-law distributed set of size N, then one can order the objects by their size to create a ranking. For a theoretical power-law distributed set that follows Zipf's law (rather than a sample), we write down some of these relationships as follows. For a rank r and size k, the ratio of successive sizes is

$$\frac{k(r+1)}{k(r)} = \frac{r}{r+1}, \tag{2.4}$$

but, as a trivial example, if one samples only those objects below some rank r^*, then the new ratio for the re-ranked objects r' is

$$\frac{k(r'+1)}{k(r')} = \frac{r^* + r'}{r^* + r' + 1}. \tag{2.5}$$

Similarly, the function $k(r)$ is usually

$$k(r) = \frac{k(1)}{r}, \tag{2.6}$$

but, for the re-ranked sub-sample, it is

$$k(r') = \frac{k(r^*)}{(r'/r^*) + 1 + (1/r^*)}. \tag{2.7}$$

Sub-sampling can distort properties of a power-law distributed set and make the sample appear to have been derived from a different distribution. Similar issues arise when aggregating two sets; including two copies of the largest size item is expected to strongly distort these internal structures. Sampling and aggregation need to be conducted carefully in order to preserve the innate structure of these sets. Part of this problem is that we are dealing with power-law distributed sets, rather than an independently and identically distributed (IID) sample, which implicitly introduces a dependence on N (Cristelli *et al.*, 2012). As such, when we create power-law distributed sample sets, we will analyse them separately rather than combine them to create averaged statistics or sub-sample them to create bootstrap statistics in order to reduce the likelihood of introducing artefacts due to structural problems.

2.2 Bootstrapping

For such a simple technique, bootstrapping is surprisingly recent in the history of statistics. Introduced by Efron, Efron described the bootstrap as "a more primitive method" than some of its contemporaries, although 'primitive' is not bad here due to the bootstrap's wide applicability. Efron (1979a) has been cited over 19,000 times, and *An Introduction to the Bootstrap* by Efron and Tibshirani (1993) has been cited over 40,000 times. The recency of the bootstrap stems from its dependence on modern computing power; explicit formulae are fairly rare in practice. The core concept of the bootstrap is fairly simple and easily modified: consider a sample from the population as a population itself. One can then sample (again) from, or otherwise compute, the original sample's distribution without resampling from the initial population. Hence, one can estimate properties of the sampling distribution, such as the variation in the mean or standard deviation. By assuming that the original sample is representative of the population, one can then infer properties of the original population. In what follows, we write as if we are working exclusively with continuous distributions, but the ideas hold with minimal modification for discrete distributions as well.

Informally, the bootstrap proceeds simply. Assume that there is some well-behaved population from which we have drawn a sample in order to ask some question about the population (that is, crucially, not about (extremal) rank statistics). The classical question to ask is "what is the population mean", but one could equally ask about the proportion above a fixed value, or to compare with a different population. The answer can be approximated by asking the question of the sample, *e.g.* computing the sample mean. More can be said by exploiting the variation inherent in the sample. We might want to construct not just an estimate, but also an interval around the estimate to enable hypothesis testing. The bootstrap solution to constructing such an interval, or answering other related problems, is to construct a large number of secondary samples by sampling with replacement from the original sample. Since sampling is done with replacement, secondary samples are different from the original sample with high probability. One can ask the same question of each secondary sample. This will form a distribution of answers which can then be used to construct the required interval. Crucially, this procedure is non-parametric, meaning there is no required assumptions of the distribution or its parameters.

To discuss the bootstrap in more detail, we break up our discussion according to various developments of the bootstrap. First, we discuss the classic bootstrap and some considerations. Then, we tackle building confidence intervals; these are crucial to how we discuss our results when applied to the Battle of Britain in Chapter 3. In order to understand some of the theory of the weighted bootstrap we use, we then present the Bayesian bootstrap. We finally discuss some recent developments in bootstrap theory by addressing a paper on causality that was developed in parallel with our work.

Formally, consider n IID random variables $\mathbf{X} = (X_1, \ldots, X_i, \ldots, X_n)$ distributed according to some distribution $\mathcal{P}$. Then our empirical distribution $\tilde{\mathcal{P}}$ is the distribution which assigns weight uniformly to each element of $\mathbf{X}$. Suppose we are trying to obtain an understanding of some parameter, typically notated $\theta = t(\mathcal{P})$, that has some estimator $\tilde{\theta}$. Normally, one would obtain an estimate by observing a realization of $\mathbf{X}$, denoted $\mathbf{x}$, and calculating $\tilde{\theta} = \tilde{t}(\mathbf{x})$. The original application of the bootstrap is to non-parametrically estimate the standard error of $\tilde{\theta}$ as follows (Efron, 1979a; Efron and Tibshirani, 1993, Ch. 6).

1. Create B resamples of $\mathbf{x}$, $\mathbf{x}^{*1}, \ldots, \mathbf{x}^{*B}$, each of the same size as the original sample by sampling uniformly at random with replacement from $\mathbf{x}$ or, equivalently, by sampling from the empirical distribution $\tilde{\mathcal{P}}$.

2. For each resample $i \in [1, B]$, calculate $\tilde{\theta}^{*i} = \tilde{t}\left(\mathbf{x}^{*i}\right)$.

3. Calculate the standard deviation of the resamples as an estimator for the standard error of $\tilde{\theta}$:

$$\tilde{\mathrm{se}} = \sqrt{\frac{1}{B-1} \sum_{i=1}^{B} \left(\tilde{\theta}^{*i} - \sum_{j=1}^{B} \frac{\tilde{\theta}^{*j}}{B} \right)^2}. \tag{2.8}$$

Note that we are not estimating θ or its properties, but $\tilde{\theta}$ or its properties instead. As such, as $B \to \infty$, the standard error described above tends towards the standard error of $\tilde{\theta}$ over the empirical distribution $\tilde{\mathcal{P}}$ created by $\mathbf{x}$, as opposed to that of θ (Efron and Tibshirani, 1993, Ch. 6). When $\tilde{t}\,(\mathbf{x}) = t\,(\mathcal{P})$, the bootstrap estimate is (approximately) unbiased by construction (Efron and Tibshirani, 1993, Ch. 19). If need-be, the bootstrap can be used again to estimate the bias of $\tilde{\theta}$ (Efron and Tibshirani, 1993, Ch. 10).

There is error induced from the variance of the (non-infinite) resampling process that is only eliminated in the aforementioned $B \to \infty$ limit. Unfortunately, the threshold for reducing this error is not so cut-and-dry as to have a singular recommendation. Efron (1987, Sec. 9) and Efron and Tibshirani (1993, Ch. 19) show an example of how this differs, but the crucial point is that the more $\tilde{\theta}$ depends on the tails of $\tilde{\mathcal{P}}$ and the larger n is, the larger B must be to compensate. Fortunately, Efron and Tibshirani indicate that small values of B are quite effective, with $B = 50$ satisfactory for estimating the standard error, and $B = 1{,}000$ satisfactory for estimating confidence intervals. In practice, the time taken to perform 'extra bootstraps' is usually trivial unless operating in real time. For example, we show that the results of taking $B = 100{,}000$ or $B = 10{,}000$ in our use case are clearly quite similar, see Figures 3.1 and 3.2 in Section 3.5. Alongside this primary work on the Battle of Britain, however, we have constructed a tool to aid in producing similar analyses. For this tool, $B = 1{,}000$; empirically, this is the smallest order of magnitude for which the data appears to be consistently well-behaved. Values larger than this impede the user experience (30 seconds on our local desktop to perform a bootstrap), while values lower can look irregular and unsmooth to the layperson.

One standard example of when the bootstrap fails is estimation of the maximum of a uniform distribution, for which the empirical distribution does not converge smoothly to the original distribution (Bickel and Freedman, 1981; Efron and Tibshirani, 1993). Another example, due to Athreya (1987), is the bootstrap of the sample mean when the original distribution lacks a finite variance: in this case, the bootstrapped mean does not converge to the sample mean.

With these constraints in mind, a natural question is to ask how we can use bootstrapping to estimate confidence intervals. In particular, we need to know how to estimate one sided confidence intervals in Chapter 3 to summarise the counterfactuals scenarios. From the standard Central Limit Theorem results, it is trivial to estimate a confidence interval when the limiting distribution of the estimator $\tilde{\theta}$ is normal. In this case, one can construct a confidence interval by appeal to a normal distribution (Efron and Tibshirani, 1993, Ch. 12): given an estimate for $\tilde{\theta}$ and the bootstrapped estimated standard error $\tilde{\mathrm{se}}$ of $\tilde{\theta}$ and assuming the standard normal distribution relationship

$$\frac{\tilde{\theta} - \theta}{\tilde{\mathrm{se}}} \sim N(0,1),$$

write the confidence interval for θ as

$$\left(\tilde{\theta} - \tilde{\mathrm{se}} \cdot z^{(1-\frac{c}{2})}, \tilde{\theta} - \tilde{\mathrm{se}} \cdot z^{(\frac{c}{2})}\right), \tag{2.9}$$

where $z^{(c)}$ (as in z-score) denotes the $100c$th percentile of a standard normal distribution. To relax the normality assumption we can instead use a t-statistic (and associated t distribution with $n-1$ degrees of freedom). While this helps account for the error from estimating the variance, this does not account for problems due to higher moments of the distribution. Errors due to estimating the variance or any higher moments can be handled by creating a case-specific t-distribution through bootstrapping.

To construct the bootstrapped t-distribution, one calculates the bootstrapped t-statistic for the

ith bootstrap as

$$\frac{\tilde{\theta}^{*i} - \tilde{\theta}}{\tilde{\text{se}}^{*i}},$$

where $\tilde{\text{se}}^{*i}$ is the standard error associated to the ith boostrap. We write $t^{(c)}$ as the value below which there are cB bootstrapped t- statistics. The confidence interval then becomes

$$\left(\tilde{\theta} - \tilde{\text{se}} \cdot t^{(1-\frac{c}{2})}, \tilde{\theta} - \tilde{\text{se}} \cdot t^{(\frac{c}{2})}\right) \tag{2.10}$$

Unfortunately, this fix for the confidence interval is unwieldy at best. Efron and Tibshirani (1993, Ch. 12) note that this procedure can be unstable in practice and dependent on the (poorly behaved) outliers. Furthermore, one needs to estimate $\tilde{\text{se}}^{*i}$, the standard error of the θ^{*i}, which can require a new bootstrap in turn for each bootstrap estimate.

Bootstrap procedures can yield distributions that are sometimes drastically different from a normal distribution however. One trivial solution is to construct the confidence interval directly from the bootstrapped data (Efron, 1979b). In this case, we construct the interval to contain an appropriate amount of the corresponding empirical (bootstrapped) data. For example, suppose that we have already performed a bootstrap with $B = 100$ independent estimates of $\tilde{\theta}$. In order to write a 90% confidence interval (and thus $c = 0.1$), we first order the bootstrap estimates $\tilde{\theta}^{*i}$, $i = 1 \dots B$ from smallest to largest and denote them as $\tilde{\theta}^{*(i)}$. Assuming a central, equally-tailed interval, we then look for the bootstrap estimates that will contain 90% of the estimated values. Here, we choose $i = Bc/2 = 5$ for the lower bound so that the interval begins at the 5th empirical percentile and $i = B(1 - c/2)$ for the upper bound so that the interval ends at the 95th empirical percentile.

$$\left[\tilde{\theta}^{*(B\frac{c}{2})}, \tilde{\theta}^{*(B(1-\frac{c}{2}))}\right]. \tag{2.11}$$

This is perhaps the most straightforward way to construct a confidence interval and this is how we construct our intervals in Section 3.5. It is not the best that one can theoretically do, as its coverage is less than what is desired (a 95% confidence interval might, for example, only contain the true value 90% of the time) and it can struggle with bias inherent to the distribution of the estimator (Efron and Tibshirani, 1993, Ch. 13).

Both construction of the t-distribution from the bootstrapped data and calculation directly from the bootstrapped data have their flaws. A general improvement over both can be found in the "Bias Corrected and Accelerated" percentile method (BC$_a$) and approximations (*e.g.* Approximate Bias Corrected or ABC) to it (Efron and Tibshirani, 1993, Ch. 14). BC$_a$ is so named due to its dependence on a bias-correction parameter, $\hat{z}_0$, and on an acceleration parameter, $\hat{a}$. The bias-correction forces the distribution to be centred on the original (estimated) median, while the acceleration accounts for the rate of change of $\tilde{\text{se}}$ with respect to the estimator. Technical details are presented in the original publication by Efron (1987).

The Bayesian bootstrap, first proposed by Rubin (1981), homes in on the relationship between bootstrapping and sampling from a multinomial model, which proves useful in our reweighting procedure in Chapter 3. As originally stated by Rubin, when calculating the $\tilde{\theta}^*$, weights $\mathbf{p}^* = (p_1, \dots, p_i, \dots, p_n)$ for each datum x_i are formulated by first realizing $n - 1$ uniform random variables over $[0, 1]$ as $u_1, \dots, u_{n-1}$ and taking $u_0 = 0$ and $u_n = 1$. The differences of the ordered $\mathbf{u}$ is $\mathbf{p}^*$: $p_i = u_{(i)} - u_{(i-1)}$ and $\tilde{\theta}^* = \tilde{t}(\mathbf{p}^*)$. This makes the standard (frequentist) bootstrap a discrete approximation of the Bayesian bootstrap. While the Bayesian weights may vary across $[0,1]$, the frequentist weights may only take values in $\left\{\frac{i}{n} : i \in \mathbb{N}_{\geq 0}\right\}$. The frequentist "approximation" becomes better as n increases.

For our purposes, it is more helpful to note the Bayesian bootstrap's prior is explicitly a non-informative Dirichlet distribution, as formulated by Hastie *et al.* (2009, Ch. 8.4). A Dirichlet distribution can be thought of as a generalization of the beta distribution to the case of multiple variables n. Instead of taking values on the line segment $[0,1]$ (a 1-simplex), it takes values on an $(n-1)$-simplex whose vertices lie at the end of unit vectors. Let ε be an information parameter that we intend to take to 0 and $\mathbf{1}$ be a vector of ones of appropriate size. The prior for the weights $\mathbf{P}^* = (P_1, \ldots, P_n)$ before observing the data is

$$\mathbf{P}^* \sim \text{Dirichlet}\left(\varepsilon \mathbf{1}\right).$$

If the data has empirical weights $\mathbf{P}^0$ such that, for each i, $n\tilde{\mathcal{P}}_i$ is the number of occurrences of each datum x_i, then the posterior is

$$\mathbf{P}^* \sim \text{Dirichlet}\left(\varepsilon \mathbf{1} + n\tilde{\mathcal{P}}\right).$$

Applying the proposed limit $\varepsilon \to 0$ yields

$$\mathbf{P}^* \sim \text{Dirichlet}\left(n\tilde{\mathcal{P}}\right).$$

(This is analogous to Haldane's prior in the beta distribution.) In comparison, the frequentist approach would analogously be sampling from the aforementioned multinomial distribution with parameters n and $\tilde{\mathcal{P}}$ to directly obtain the bootstrapped counts (Hastie *et al.*, 2009).

There are natural critiques of the Bayesian, and thus the frequentist, bootstrap. Rubin (1981) puts the most important quite simply: "Is it reasonable to use a model specification that effectively assumes all possible distinct values of X have been observed?" This very criticism is used by Rubin to predict the issue identified by Athreya (1987): bootstrapped moments are sensitive to the model's tail probabilities. This is a recurring criticism of the non-parametric bootstrap and forces us to assume that everything that could have happened did so as well as that there are no (unobserved) extraordinary events.[5] Additionally, Rubin indicates concerns about smoothness and independence as well (*i.e.* that values near each other should be near in probability as well). Unfortunately, to deal with these problems requires some parametric assumptions on the bootstrap, rather than using strictly the data from the original sample.

In parallel to the publication of Bootstrapping the Battle of Britain (Fagan *et al.*, 2020a), several papers on combining causality with bootstrapping were produced. We briefly discuss a bootstrap for causal parameters using copula to impute missing experimental or observational outcomes due to Imbens and Menzel (2018). This method is based on the causal framework advocated by Imbens and Rubin (2015).

Imbens and Menzel (2018) are concerned with uncertainty arising from the differences between sample construction (*e.g.* what happens if instead of observing Alice's and Bob's reactions in an experiment, we were to observe Airy's and Barry's) and experimental design (*e.g.* what happens if instead of assigning Alice to the control group and Bob to the treatment, vice versa had occurred). The problem is that the bootstrap only considers sampling uncertainty. Imbens

[5] The popular term for such an event is a 'black swan', attributed to Taleb (2007).

and Menzel view this as a data imputation problem: the bootstrap is imputing to the population values from the sample, but this cannot trivially be done for the treatment assignment due to the additional design uncertainty. They propose to proceed as follows.

1. Extend the realised sample (of size n) to generate a population of the same size as the actual population. Care is taken to obtain a generated population with empirical distributions approximately equal to the sample's empirical distributions with error of order n^{-1}. (The distributions are separated by treatment.)

2. Choose a copula iteratively from the set of all copulas that fit the data by fixing moments to the least favourable values (*i.e.* values that minimise the information, usually in the sense of Fisher) until either a unique copula remains or there is a sufficiently precise copula. The resulting copula then is conservative. (A copula is a method of coupling marginal cumulative distribution functions into a joint distribution that share the same correlational structures.) This copula is then used to impute the missing values in the population, although the specific copula will determine exactly how the imputation is performed. For a simple treatment effect model, one can use the rank of the individual's factual outcome to assign the same rank for the counterfactual outcome(s).

3. Construct bootstrap replications by repeatedly sampling without replacement from the generated population. Assign treatment to each sample by Bernoulli trials with probability as in the realised sample. Sample statistics can be computed on the bootstrap replicates treated sample.

In practice, Imbens and Menzel (2018, Sec. 4) observe more overcoverage for the standard bootstrap in contrast to the causal version when constructing confidence intervals. This is due to the association between treatments. As the causal relationship becomes more complicated, the bootstrap does see decreasing performance, but consistently via overcoverage, while the causal bootstrap remains reasonably close to the correct level of coverage.

2.3 Changepoint analysis

Changepoint analysis is a family of statistical techniques for detecting abrupt changes in the distributions of data and thus breaking a dataset into segments with common properties. We restrict ourselves to offline changepoint analysis, *i.e.* analysis of data after the data is finished recording, as opposed to online changepoint analysis, which tries to estimate the locations of changepoints as new data is introduced to the dataset. Briefly, the standard tactic is to collect a set of summary statistics about the data set, then look for places where the summary statistics suddenly change. The weight of the change then needs to be compared to some form of tolerance. In this section, we will explore how various forms of changepoint analysis work in practice in preparation for Chapter 4. First, we will talk about a general changepoint analysis, before exploring in more detail the components that we will use, implemented in the `changepoint` and `changepoint.np` R packages (Killick *et al.*, 2016; Haynes *et al.*, 2016; R Core Team, 2018)

A changepoint analysis is a minimisation problem that can be broken down into three interacting components: the algorithm, the the cost function corresponding to a segment's model, and the cost due to segmentation (typically referred to in the literature as a per-segment penalty). We first give a bottom-up conceptual discussion of changepoint analysis before presenting a top-down discussion in the context of the packages used. Consider first a simple model in which we fit a line to a dataset. We can compute measures of goodness-of-fit for our line, regardless of how linear the dataset actually is. This component is comparable to what we will refer to as the cost function. Adding more parameters makes the fit better, with some form of diminishing returns as we approach the true model. We can use the measures of goodness-of-fit to compare between models with the same number of parameters, but it is not apparent how we should compare between models with differing numbers of parameters. To make this comparison, we will rely on what is typically called the penalty. Finally, the method with which we expand our model is important. For example, our line has two parameters, a slope and intercept, but adding another parameter gives us access to a quadratic as well as a step function: the first and second levels, and the point at which the levels change. A fourth parameter grants access to cubics, step and line, piecewise continuous lines, and so on. The specific choices as to how to proceed through these models, not allowing discontinuities or only using piecewise continuous lines for example, can be thought of as our fitting algorithm. The main difference here is that, instead of fitting curves to the data of interest, we are instead interested in fitting distributions.

Proceeding now from the top, the algorithm can be thought of as the strategy for the placement of changepoints and thus the governing strategy for minimising the costs of modelling the data. Trivial algorithms include changepoints between every datum and changepoints between no datum. A more commonly encountered and sensible algorithm is to place "at most one changepoint" (AMOC) (Killick *et al.*, 2016). Such a simple algorithm is a natural first step to a recursive algorithm in which a changepoint is placed and then the algorithm repeats to check if there should be any new changepoints to the left or right of the one just placed (Scott and Knott, 1974; Sen and Srivastava, 1975). The algorithm corresponds to the `method` argument in the `changepoint` and `changepoint.np` R packages.

The cost function, denoted $\mathcal{C}$, assigns to each segment between changepoints a cost to model the segment. The cost function can be thought of as characterising how well explained or internally consistent a segment is in that a model that very well explains a segment is quite cheap, while a model that works poorly is expensive to maintain. The cost function might be thought of as a "negative benefit" function instead if one considers values like the R^2 of a regression or likelihood of a distribution to be measurements of benefit. Cost functions are commonly parametric, with assumptions ranging from general, *e.g.* lower order moments exist (*e.g.* CSS, Inclán and Tiao, 1994), to specific, such as assuming that each segment is normally distributed. When an explicit model is expected, such as a normal distribution, the cost function might take the form of the negative log-likelihood, since this is a minimisation problem, associated with the maximum likelihood estimated distribution for a given segment (Chen and Gupta, 2000). Another example might be to measure the quadratic loss as the sum of squared distances to the mean of a segment (Rigaill, 2015). Fully non-parametric options exist as well; one simple example would be as quadratic loss but to the median of the segment instead. One is not required to use only a single cost function, but to do otherwise would likely require a specialised knowledge of how to model each segment individually. As such, we do not consider it here. The cost function corresponds to the `test.stat` argument in the `changepoint` and `changepoint.np` R packages.

Given that we have already included costs for modelling inside of the cost function, why do we need to include an additional cost in the form of a per-segment penalty? The penalty exists to prevent overly complex segmentations by penalising high numbers of segments, while still allowing complexity when the cost function warrants it. In our example of fitting curves, the penalty serves to truncate the process, rather than trying to find an extremely complex model that fits all points. After some point, diminishing returns will be dominated by the penalties accrued for each parameter used. Usually, penalties are of a linear form, *e.g.* $f(m) = (m+1)\beta$ where m is the number of changepoints and β is the per parameter scaling. With p as the number of parameters and n as length of the original data, common choices of β include $\beta = 2p$ (Akaike information criterion or AIC (Akaike, 1974)), $\beta = p\log(n)$ (Bayesian information criterion or BIC, also known as Schwarz's information criterion or SIC (Schwarz, 1978)), or $\beta = 2p\log\log(n)$ (Hannan and Quinn, 1979). The penalty corresponds to the `penalty` and `pen.value` arguments in the `changepoint` and `changepoint.np` R packages.

Formally, take $\mathbf{x} = (x_1, \ldots, x_i, \ldots, x_n)$ to be a time-ordered dataset. Suppose further that $1 \leq m \leq n-1$ changepoints have been placed at integer time-ordered locations $\boldsymbol{\tau}_{1:m} \equiv (\tau_1,\tau_2,\ldots,\tau_m)$, with $1 \leq m \leq n-1$. Take $\tau_0 = 0$ and $\tau_{m+1} = n$. This segments the data into $m+1$ segments, with indices written as the intervals $(\tau_{i-1},\tau_i]$ with the usual convention for bracketing intervals. These correspond to the data segments: $\{\mathbf{x}_{(\tau_{i-1},\tau_i]} \equiv (x_{(\tau_{i-1}+1)}, \ldots, x_{\tau_i})\}_{i=1}^{m+1}$. Each segment has a cost associated to it, and the penalty is applied over the result, which yields minimisation Problem (2.12).

$$\min_{m,\boldsymbol{\tau}_{1:m}} \left(\sum_{i=1}^{m+1} \mathcal{C}\left(\mathbf{x}_{(\tau_{i-1},\tau_i]}\right) + f(m) \right). \tag{2.12}$$

Problem (2.12) is usually solved by one of two approaches: exactly or approximately. Approximate algorithms are generally simpler to implement, but are not guaranteed to come to the true minimum, in contrast to exact approaches. The algorithm employed in Chapter 4 is a modification of pruned exact linear time (PELT) due originally to Killick *et al.* (2012). As its name suggests, PELT is intended to solve Problem (2.12) exactly in linear time with the length of the data via pruning candidate changepoints during minimisation. PELT is in turn a modification of an exact searching method for optimal partitioning by Jackson *et al.* (2005). Killick *et al.* (2012) take the optimal partitioning algorithm and add to it a pruning condition, which removes candidate changepoints that cannot lower the cost of segments. This retains the exactness of the original algorithm, while reducing the cost to linear time.

For the purposes of Chapter 4, we choose to use a non-parametric model. As discussed above, many cost functions are intrinsically parametric, *i.e.* require some level of model assumption, but for Chapter 4, we cannot even necessarily assume that the mean of the underlying distribution exists. The solution we use is a cost function proposed by Haynes *et al.* (2017b), extending the work of Zou *et al.* (2014). The cost function is based upon integrating the log-likelihood of the empirical distribution $\hat{\mathcal{P}}$ for the theoretical distribution $\mathcal{P}$. Given cumulative distribution function (CDF) $G(t)$ and empirical CDF $\hat{G}(t)$,

$$n\hat{G}(t) \sim \text{Binomial}(n, G(t)), \tag{2.13}$$

with log-likelihood

$$\mathcal{L}(\mathbf{x}|t) = n\left(\hat{G}(t)\log G(t) + (1-\hat{G}(t))\log(1-G(t))\right) \tag{2.14}$$

Maximum likelihood estimation is used to assign a theoretical cost to each segment, with n replaced by the segment length and each G and $\hat{G}$ replaced by the segment's $\hat{G}$:

$$\mathcal{L}(\mathbf{x}_{(\tau_{i-1},\tau_i]}|t) = (\tau_i - \tau_{i-1})\left(\hat{G}_i(t)\log\hat{G}_i(t) + (1-\hat{G}_i(t))\log(1-\hat{G}_i(t))\right). \tag{2.15}$$

This is weighted and integrated over the entire dataset to form a cost for each segment:

$$\mathcal{C}(\mathbf{x}_{(\tau_{i-1},\tau_i]}) = \int_{-\infty}^{\infty} -\mathcal{L}(\mathbf{x}_{(\tau_{i-1},\tau_i]}|t)\,(G(t)(1-G(t)))^{-1}\,\mathrm{d}G(t). \tag{2.16}$$

Even if the integral were to be tractable, $G(t)$ is not known. The cost function devised by Haynes *et al.* (2017b) evaluates Equation (2.16) through the use of empirical quantiles which are preferentially selected from the tail of the distribution. If Q quantiles are to be used, set the qth evaluation point to

$$t_q = \left((1+(2n-1)\exp\left(\frac{-\log(2n-1)}{Q}(2q-1)\right)\right)^{-1} \tag{2.17}$$

and define the final cost function by

$$\mathcal{C}(\mathbf{x}_{(\tau_{i-1},\tau_i]}) = \frac{2\log(2n-1)}{Q}\sum_{q=1}^{Q}\mathcal{L}(\mathbf{x}_{(\tau_{i-1},\tau_i]}|t_q). \tag{2.18}$$

This cost function is implemented in the `changepoint.np` R package under the `test.stat` argument as the `empirical_distribution` option. When combined with PELT, we refer to this as ED-PELT.

Finally, we address the segment penalty arguments we will primarily use. We use two penalties: the modified Bayesian information criterion (mBIC) (Zhang and Siegmund, 2007) and changepoints over a range of penalties (CROPS) (Haynes *et al.*, 2017a). The former of these is, as its name implies, a modification of the BIC $\beta = p\log(n)$ (Schwarz, 1978) for the changepoints setting from the regression setting, which has differing regularity conditions. Approximately, mBIC sets $\beta = \frac{3}{2}p\log n$, but the actual computation incorporates the relative positions of the changepoints (Zhang and Siegmund, 2007). The mBIC penalty is derived theoretically from the asymptotics of the Bayes factor for Brownian motion with changing drift, whereas the BIC penalty is derived without dependence on priors or parameters using asymptotics of a Taylor expansion. Technically, β becomes nontrivially dependent on m through the (relative) locations of changepoints placed in the datasets. The incorporation of the placement of the changepoints is guaranteed to add the extra (relative to BIC) factor of $\frac{1}{2}$ in the approximate value of β.

CROPS, on the other hand, is an algorithm for evaluating the effects of varying the penalty over a continuous range, *i.e.* varying $\beta \in [\beta_{\min}, \beta_{\max}] \subset \mathbb{R}_{\geq 0}$. Penalty values are recursively chosen to identify when the optimal number of changepoints changes. When no or trivial changes are all that is left in an interval, the algorithm terminates. This is more computationally complex, partially due to needing to rerun the algorithm (with its cost function) for each penalty that is chosen. The result is optimal changepoints over ranges of penalties, but this leaves the user to decide which set of changepoints is best for their purposes. We adopt the recommendation of Haynes *et al.* (2017a, who cite Lavielle (2005)): we select the "elbow" in a plot of the (unpenalised) cost against the number of changepoints. We use the location of the numerical maximum of the second derivative, maximising the curvature. Identifying the elbow in the plot is intended to identify the point of diminishing returns: if the penalty is too weak, we are in danger of false positives, but if the penalty is too strong, we are in danger of false negatives. As we shall see in Chapter 4, this elbow does not always properly exist, for which we will need a solution. These penalty functions are implemented in the `penalty` and `pen.value` arguments in the `changepoint` and `changepoint.np` R packages as `MBIC` and `CROPS`.

We finish with a note about a limitation of changepoint analysis. Changepoint analysis looks for *discrete* changes in the underlying generative processes. Naturally, not all changes are discrete. It is not surprising that changepoint analysis can then fail in the presence of trends in the underlying processes. One solution has been to attempt to detrend a dataset, but this requires some knowledge of the trend (*e.g.* Serinaldi and Kilsby, 2016). Recent attention has begun to try to work on changepoint analysis in the presence of trends (*e.g.* Gallagher *et al.*, 2013), but we are unaware of publications resulting in a combined changepoint and trend analysis algorithm suitable for Chapter 4. Due to its applicability, this is a natural direction for future work.

2.4 Coalescence and fragmentation models

As a mathematical field, coalescence (alternatively: aggregation, coagulation) and fragmentation is just over a century old, having begun with von Smoluchowski (1916). Many developments have occurred in the past 100 years. We begin our discussions of coalescence and fragmentation models with a general description of some of these developments in coalescence theory; the reader is encouraged to consult classical references Dubovski (1994) for a collection of theoretical results and Aldous (1999) for a discussion of applications and interpretations, especially in probability theory. We highlight the appearance of gelation, the formation of a particularly large (*i.e.* of size comparable to the size of the entire system) or infinite sized particle. We then discuss models with both coalescence and fragmentation.

Von Smoluchowski sought to better understand the dynamics of colloidal chemistry: chemistry involving particles suspended in a fluid, which is usually taken to be well-mixed in coalescence and fragmentation modelling. Each particle is a discrete object that can clump, or coalesce, with other particles to make bigger particles; when a particle is a single atomic object, we refer to it as a monomer while particles in general will be referred to as clusters. Note that mass is conserved and only binary collisions are considered. An example of such can be seen in Figure 2.2 (left). Under assumptions of attraction and Brownian motion leading to mixing, von Smoluchowski was able to derive coalescence mean-field equations of the form

$$\dot{n}_k(t) = \frac{1}{2}\sum_{i=1}^{k-1} n_i(t)n_{k-i}(t) - n_k(t)\sum_{i=1}^{\infty} n_i(t) \tag{2.19}$$

where n_k is the density of clusters of size k ($k \geq 1$), $t \geq 0$ is a rescaling of time, and we denote the time derivative with a dot. Henceforth, we neglect the time argument when the meaning is clear. The first term represents gains due to coalescence: clusters of size i interact with clusters of size $k-i$ to create clusters of size k. The $1/2$ prevents double counting (*i.e.* index i can be both 1 and $k-1$ *etc.* and when two choices are made from the candidate clusters for $i = k - i$). The second term helps to preserve mass by recording the losses due to coalescence, which can be trivially seen by taking $\sum_{k=1}^{\infty} k\dot{n}_k$ and re-indexing the sums. We might naturally expect the system size M, the amount of mass (per the fixed system volume) in the system, to be constant, whether the system has begun coalescing or is strictly monomers at complete disaggregation. Equation (2.19)

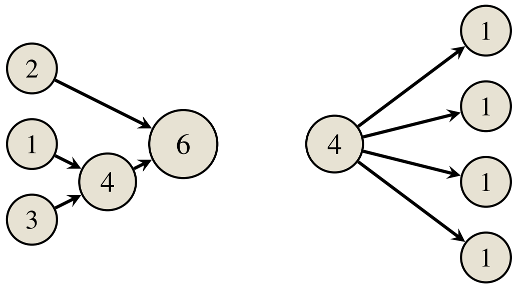


Figure 2.2: Example of coalescences (left) and shattering fragmentation (right).

has solutions of the form

$$n_k = n_0 \frac{(\frac{1}{2}n_0 t)^{k-1}}{(1+\frac{1}{2}n_0 t)^{k+1}}, \tag{2.20}$$

where n_0 is the number of monomers per unit volume when the initial condition constrains all mass to monomers and $k \geq 1$ (von Smoluchowski, 1916, Eq. 24). These equations are in general agreement with experiments conducted at the time (von Smoluchowski, 1916).

Equation (2.19) would come to be known as a form of the Smoluchowski coagulation equation and can be considered an infinite-volume mean-field theory (Aldous, 1999). More general formulations introduce a coalescence rate kernel $K(i, j)$ and can take discrete or continuous (referring exclusively to cluster sizes, time is regarded usually as continuous) form:

$$\dot{n}_k = \frac{1}{2}\sum_{i=1}^{k-1} K(i, k-i) n_i n_{k-i} - n_k \sum_{i=1}^{\infty} K(i, k) n_i, \tag{2.21}$$

$$\dot{n}_k = \frac{1}{2}\int_0^k K(i, k-i) n_i n_{k-i}\, \mathrm{d}i - n_k \int_0^{\infty} K(i, k) n_i\, \mathrm{d}i. \tag{2.22}$$

Many variant kernels have been considered (examples available in, *e.g.*, Dubovski, 1994; Aldous, 1999), but among them the constant $K(i,j) = 1$, additive $K(i,j) = i + j$, and multiplicative $K(i, j) = ij$ kernels are particularly tractable. A valid kernel can always be written as symmetric and is a function of the amount of mass in each cluster. Additionally, we neglect the (strictly positive and time rescaleable) constant when discussing theory (Dubovski, 1994; Aldous, 1999).

The behaviours of mean-field Systems (2.21) and (2.22) depend strongly on the kernel chosen. Solutions for each of the aforementioned kernels are listed for both discrete and continuous systems by Aldous (1999, Table 2) with citations therein. We call particular attention to von Smoluchowski's solution for the constant kernel for the discrete system, Equation (2.20), as well as to McLeod's solution for the multiplicative kernel for the discrete system:

$$n_k = \frac{t^{k-1} k^{k-2} \exp(-kt)}{k!}, \tag{2.23}$$

for $t \leq 1$ (McLeod, 1962, Eq. 2.5)[6]. Aside from the obviously different forms, there is a crucial distinction between Equation (2.20) and Equation (2.23): in the former the solution is valid for all non-negative t, but for the latter the phenomenon of gelation occurs at $t = 1$.

Gelation is the formation of a particle of infinite mass (or of size comparable to the size of the system in a discrete setting), called a gel, within finite time and represents a phase transition. In contrast, the remainder of the finite system is typically called the sol. Despite beginning with a

[6] Aldous (1999) identifies Equation (2.23) as a Borel distribution divided by the cluster size.

finite system, such a cluster can emerge in finite time when $K(i,j) = ij$, and more generally when the degree of the homogeneity of the kernel[7] $\alpha : \forall c > 0 : K(ci,cj) = c^{\alpha} K(i,j)$ has $\alpha > 1$ (Ziff *et al.*, 1982; Aldous, 1999; Pushkin and Aref, 2002). A useful argument to illustrate the problem is to analyse the flux of mass through the system (Davies *et al.*, 1999; Wattis, 2006). With flux

$$J_k = \sum_{i=1}^{k} \sum_{j=k+1-i}^{\infty} iK(i,j) n_i n_j \tag{2.24}$$

for $k \geq 1$ and $J_0 = 0$, then the change in mass amongst clusters of size k is

$$\dot{(kn_k)} = J_{k-1} - J_k \tag{2.25}$$

and so the change in the total mass, $M = \sum_k kn_k$, is

$$\dot{M} = - \lim_{i \to \infty} J_i. \tag{2.26}$$

If equation 2.26 has the limit decay sufficiently fast, then we expect mass to be conserved, but for gelling kernels, this limit is non-zero (Davies *et al.*, 1999; Wattis, 2006). The loss of mass is attributed to and used to measure the size of the gel (Dubovski, 1994).

Different strategies have been used to explicitly model the formation of the gel and its interaction (or lack thereof) with the sol. In approaches usually attributed to Stockmayer and Flory, a gel (inert in the approach of Stockmayer, reactive in the approach of Flory or Ziff and Stell (Ziff and Stell, 1980)) is measured by its effects on the sol (Ziff and Stell, 1980; Lushnikov, 2006). A different approach is via truncation of the Smoluchowski coagulation Equations (2.21) and (2.22). The gel is then assumed to absorb any clusters whose size exceeds the truncation (Lushnikov and Piskunov, 1983; Lushnikov, 2006).

Does a gel still form if we are tracking each cluster, as opposed to their densities? Such a microscopic system was considered by Marcus (1968) and Lushnikov (1978) and might be considered the finite-volume mean-field theory in contrast to the differential equations above (Aldous, 1999). Mean-field approaches provide a natural description for the sizes of the clusters, but are usually ignorant of finite system size, which is more important in microscopic systems where the explicit number of clusters of each size is tracked, and the probabilities of obtaining specific configurations are examined. The three kernels mentioned above correspond to probabilistic structures: $K(i,j) = 1$ to Kingman's coalescent, $K(i,j) = i + j$ to the continuum random tree, and $K(i,j) = ij$ to random graph models of Erdös and Rényi (Aldous, 1999). In the last context, the gel is the giant component that manifests around the critical time $t = 1$. The gel is then of the same scale as the system size[8] and the remainder of the system is fairly disaggregated sol with which the gel reacts in the sense of Flory (Aldous, 1999; Lushnikov, 2006).

So far, we have discussed purely coalescent systems. There are many ways to introduce fragmentation, an example of which can be seen in Figure 2.2 (right), both in terms of when it occurs and its mechanics. Fragmentation can be a part of the same process as coalescence such as when two asteroids impact each other and may remain together or break each other apart (Tanaka *et al.*, 1996), fragmentation may occur whenever a cluster gets too large or unstable (Birnstiel *et al.*,

[7] Note that homogeneity is not necessarily required, but is a generally useful assumption. Asymptotics for non-homogeneous kernels are expected to agree (Aldous, 1999).

[8] In Θ notation, the size of the gel is $\Theta(M)$.

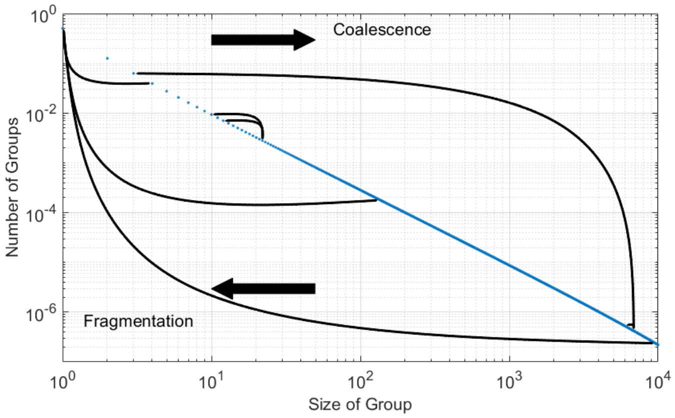


Figure 2.3: Example of (binary) coalescence and (shattering) fragmentation (black lines) moving mass through the system. Coalescence moves mass to the right towards larger clusters, while fragmentation moves mass to the left towards smaller clusters. This is expected to form a steady-state, here depicted as a line of blue dots on log-log axes.

2011), or fragmentation might be an exogenous process such as a state spotting and breaking up a social group of individuals (Bohorquez *et al.*, 2009; Clauset and Wiegel, 2010). The distribution of fragments might be a part of the distribution of coalescents, independent and random, or even fixed. Due to the variety of ways to define fragmentation, the form of the kernel is often ambiguous. In principle, fragmentation inputs some number of clusters and outputs a (usually different) number of clusters. We will focus on spontaneous or exogenous fragmentation, in which a cluster breaks apart without interacting with another cluster. Writing the fragmentation kernel as F, a general form of spontaneous fragmentation (in the discrete case) is then

$$\dot{n}_k = \frac{1}{2}\sum_{i=1}^{k-1} K(i,(k-i))n_i n_{k-i} - n_k \sum_{i=1}^{\infty} K(i,k)n_i + \sum_{i=k}^{\infty} \frac{i}{k} F(i,k) n_i - n_k \sum_{i=1}^{k} F(k,i). \quad (2.27)$$

The fragmentation kernel $F(i,k)$ here determines how frequently a cluster of size i fragments into a cluster of size k. The ik^{-1} on the third term accounts for the difference in sizes between the cluster fragmented and the cluster formed. In practice, various authors write the arguments to reflect their desired emphasis and we are no different. Usually, we will write a single argument fragmentation kernel $F(k)$ which encodes how a cluster of size k fragments. This is because we will usually adopt spontaneous and atomising or shattering[9] forms of fragmentation, where a cluster of size k is reduced to k monomers and the destination argument is then implicitly obvious.

Fragmentation is usually seen as providing some form of steady-state to the system although it is not the only way to achieve equilibrium; for example, one might continuously introduce new small clusters (Pushkin and Aref, 2002). The reason for why fragmentation might provide such a steady-state is that fragmentation sends mass to smaller clusters while coalescence sends mass to larger clusters, as shown in Figure 2.3. In practice, how obvious the steady-state is varies. In the Becker-Döring model, within which clusters grow and shrink by monomers only, steady-states are well-known when the rate kernels are sub-linear and the total system mass M is finite (Ball *et al.*, 1986). When a coalescence and fragmentation system's reactions are reversible, *i.e.* fragmentation

[9] Our usage follows Birnstiel *et al.* (2011), but shattering has been used to describe the creation of size zero particles, 'dust', instead (e.g. McGrady and Ziff, 1987).

exactly undoes coalescence[10], and if the fragmentation and coalescence rates satisfy the detailed balance condition, then the solutions adopt the same form as those of the Becker-Döring model (Wattis, 2006).

When fragmentation results in shattering , alternative tools are used to establish a steady state. We present an abbreviated argument for a steady-state solution which has been presented in various forms and with various extensions and that holds for gelling coalescence kernels (e.g. D'Hulst and Rodgers, 2000; Brilliantov *et al.*, 2015; Johnson *et al.*, 2016). Clauset and Wiegel (2010) showed that this steady-state was approximately correct regardless of the fragmentation kernel F so long as spontaneous shattering is required. Taking for $k \geq 2$ (and $k = 1$ for completeness)

$$\begin{aligned} \dot{n}_k &= \frac{1}{2}\sum_{i=1}^{k-1} K(i, k-i) n_i n_{k-i} - n_k \sum_{i=1}^{\infty} K(i,k) n_i - F(k) n_k, \\ \dot{n}_1 &= \sum_{i=1}^{\infty} iF(i) n_i - n_1 \sum_{i=1}^{\infty} K(i,1) n_i \end{aligned} \tag{2.28}$$

with $K(i,j) = c(ij)^\alpha$ for some $c, \alpha \geq 0$, in steady-state $\dot{n}_k = 0$ for all k and using generating functions[11]

$$f(z) \equiv \sum_{k=1}^{\infty} k^\alpha n_k z^k \qquad\qquad g(z) \equiv \sum_{k=1}^{\infty} (k) n_k z^k, \tag{2.29}$$

we can write

$$0 = \frac{1}{2} c f^2(z) - c f(1)(f(z) - n_1 z) + g(z). \tag{2.30}$$

Clauset and Wiegel (2010) argue that the leading orders of magnitude with the system size allow us to neglect $g(z)$, as the other terms are proportional to M^2 rather than just M. Solving for $f(z)$ yields

$$f(z) = 2n_1(1 - \sqrt{1-z}), \tag{2.31}$$

noting that $f(1) = 2n_1$. Power series expansion or Cauchy's theorem may be used to obtain coefficients, with the latter yielding the asymptotic result

$$k^\alpha n_k \approx \frac{1}{\sqrt{\pi}} n_1 (k+1)^{-3/2}. \tag{2.32}$$

For the case $b(k) \propto k, \alpha = 1$, this argument can proceed exactly (D'Hulst and Rodgers, 2000; Johnson *et al.*, 2016). Furthermore, D'Hulst and Rodgers (2000) and Kyprianou *et al.* (2018) showed there exist steady-state solutions when coalescence involves more than two particles assuming shattering, a single time scale (by scaling the kernels by M), and multiplicative (in the former's work) or constant (in the latter's) rate kernels. These steady-state solutions are zero for impossible cluster sizes (*e.g.* no clusters of size two if only triples coalesce). The variant of Kyprianou *et al.* (2018) allows multiple types of coalescence and finds only the smallest type (rather than the fastest) dictates which cluster sizes are non-zero in the steady-state.

While this answers whether there is a theoretical steady-state, it does not tell us about the inter-

[10] This is usually written as $S_a + S_b \rightleftharpoons S_{a+b}$ where $S_\cdot$ is a type, or species, of cluster and a and b are indices usually indicating size.

[11] Note that generating functions are not necessarily concerned with questions of convergence as they are usually considered formal power series (Stanley, 2011).

action between the gel and fragmentation. Can fragmentation inhibit or even prohibit gel formation? As reported by da Costa (1995), the answer is yes, but fragmentation needs to be sufficiently strong. Da Costa (1995) showed for binary coalescence and binary spontaneous fragmentation, *i.e.* a fragmentation kernel $F(i, j)$ that breaks a single $(i + j)$-sized cluster into an i-sized cluster and a j-sized cluster, that fragmentation can prevent gel formation if, given coalescence kernel

$$K(i,j) \leq c(ij)^{\alpha}, \alpha \in \left(\frac{1}{2}, 1\right], \tag{2.33}$$

the fragmentation kernel has the property that $\exists\gamma > \alpha : \forall\mu \geq 1, \exists C(\mu) > 0 :$

$$\sum_{j=1}^{\lfloor (r-1)/2 \rfloor} j^{\mu} F(j, r - j) \geq C(\mu) r^{\gamma+\mu} \tag{2.34}$$

for all $r \geq 3$. As examples, $F(i, j) = (i + j)^{\beta}$ and $F(i, j) = (ij)^{\beta}$ satisfy the above property for $\beta > -1$, dependent on α of course. While not necessarily for our primary use case, this provides theoretical evidence that fragmentation can act as a balancing force as one might expect.

3 The Battle of Britain

> *[The great man] is powerless in the absence of the co-existing population, character, intelligence, and social arrangements [of his society].* *– Spencer, 1896*

3.1 Introducing the Battle

Air campaigns in general, and the Battle of Britain (1940) in particular, have a stationarity to them that is thoroughly lacking in a traditional land campaign. If longbowmen on one side annihilate infantry or knights bogged down in marsh, the next day those infantry and knights cannot be redeployed, that same territory is likely lost, and the battle has shifted its position and terrain. On the other hand, if opposing aerial squadrons begin to dogfight, both sides can have their pilots evacuate from the field[12], new planes can be ready to be deployed[13], and the same airspace is likely to see multiple dogfights. Over a short enough period of time, even the weather is unchanged.

At its simplest, the Battle of Britain was a World War II air campaign in the summer and autumn of 1940. It was fought between the Luftwaffe, the German air force, and (British) Royal Air Force (RAF) Fighter Command in the skies of the English Channel, south-eastern England, and London. The Luftwaffe sought to deplete Fighter Command forces, while Fighter Command conversely needed to remain a force in being. Factually, the Luftwaffe had demonstrably failed in this regard by October, while the RAF was growing stronger.

It has long been folklore that this victory was won on a "narrow margin", with many books and articles published on both sides of the debate (e.g. Dempster and Wood, 1961; Bungay, 2000). Similarly, it is a popular event to consider counterfactually (Forester, 1971; Cox, 1974; Messenger, 2002; Forczyk, 2016); that is, to ask why Britain won, and the German war machine failed to invade. We too seek to answer this question, but we must acknowledge that "[t]here was no

[13] Consider the following back of the envelope: the average number of pilots removed from battle per sortie per day is about 0.01 (from data), while 33% of novice pilots are lost within a month in 501 Squadron during the Battle of Britain (Bungay, 2000, p. 373). A geometric distribution with a probability of success of about 0.013 on a given trial would need 30 trials to have about a 33% probability of having a single success. Taking 0.013 as the probability that a given pilot is removed on a given day and dividing it by the average number of pilots removed per sortie and day gives about 1.25 sorties which we interpret as the number of sorties flown by a (novice) pilot in a day.

[13] For example, Dempster and Wood (1961) indicate that, per day, about 400 monoplanes were being produced compared to about 260 pilots by the British during the Battle of Britain.

obvious step-change for Fighter Command to make" (Fagan *et al.*, 2020a). As a result, we instead focus on the decisions of the Germans, and adopt the approach of Forester (1971, p. 131): "[German leader Adolf Hitler] must be given in this narrative every possible chance, but none of the impossible ones".

For the purposes of this chapter, we consider only the air campaign over England with the counterfactual invasion campaign and its other prerequisites firmly outside of our scope. There were, of course, other aspects to the preparations for the invasion of Britain: simultaneously RAF Bomber Command and the Royal Navy were attacking German ports and craft to prevent the build-up of an invasion fleet. If an invasion were to be attempted, then the Royal Navy would have been critical to preventing the arrival of landing craft (Grinnell-Milne, 1958). This does not mean that the air battle was meaningless, as has been claimed (James, 2006; Cumming, 2010). Indeed, such a portrayal has been considered 'a silly season story par excellence' (Goulter *et al.*, 2006). The destruction of Fighter Command and the resulting ability of the Luftwaffe to do as they wished in the skies of England would surely have influenced the remainder of the war.

A defeat of Fighter Command in the air campaign would entail, at a minimum, a gradual retreat from south-eastern England. RAF 11 Group commander Keith Park had no obvious abrupt change to make, nor event that would have resulted in such a change. Instead, his fighters would be more restricted in their appearances before disappearing altogether, granting the Germans a brief window of opportunity. If the Germans took this opportunity, the RAF would have rallied in parallel with the ensuing naval action.[14]

We quantitatively investigate the Battle of Britain with a novel application of the method of bootstrapping, for which standard methods are discussed in Section 2.2. The novelty stems from a reweighting scheme in which we analyse the effect on counterfactual aerial campaigns by biasing the bootstrap to select from days that are representative of the counterfactual under consideration. To investigate the Battle of Britain quantitatively though, we will need to engineer an invasion criterion. We begin this investigation by providing more background on the campaign in Section 3.2 before discussing modelling considerations and the bootstrap in Section 3.3. With these factors in mind, we discuss the data, the invasion criterion, and the counterfactual scenarios in Section 3.4. This finally allows us to apply the bootstrap in Section 3.5. We present refinements to our analysis in Section 3.6 before concluding in Section 3.7 and place the chapter in the context of the thesis in Section 3.8.

3.2 Going into the Battle

Hitler was a victim of his own success in May 1940. His invasion of France had gone exceedingly well and forced the Allies into a corner. Hitler found himself faced with subjugating the remainders of the French, attempting to push his advantage against the British, and trying to recover from his troops' losses. In late May, he hesitated, frustrating his own troops; Heinz Guderian, the famously thrusting tank commander, noted that it had "never occurred" to him that Hitler "would now be the one to be frightened by his own temerity and would order our advance to be stopped at once" (Guderian, 2009, p. 109). This hesitation, a "halt order" on 24 May, helped stop the German advance, aiding the Allies in evacuating the French port of Dunkirk.

[14] Such an invasion has been wargamed by the 1974 Royal Military Academy and subsequently fictionalized by Cox (1974). Schenk (1990) provides details of German preparations and a summary of the strategic planning is given in the German Plans (2017).

Hitler's hesitations with Britain did not stop there. Nazi Germany needed the British out of the war, but its leadership was unsure of how to do so. Peace was one possibility: the British could unite with Germany in their shared Germanic heritage. British leader Winston Churchill may have had no intention of peace with Nazi Germany, but a broken nation might force the matter. Invasion was another possibility, but a costly one. To do so would require the construction of transports for the army, subduing the Royal Navy, and weakening the RAF before even beginning the invasion proper. All of this would need to be prepared at the same time as the second phase of the Battle of France against some 200,000 soldiers in June (Alexander, 2007).

Factually the Germans failed in their objectives. There was no cowing of the populace, no capitulation of the government, and not even a proper attempt at invasion by the German military. This renders the quantification of such an invasion immensely difficult and in turn making it difficult to come to an objective understanding of how close the Germans came to success. Even whether the Germans truly tried has been debated; some German veterans and historians deny that the "Battle of Britain" was a singular campaign, suggesting instead that invasion was never the true intention and that the battle was merely another front of the ongoing struggle on the mainland (Deighton, 1977, p. 51; Bungay, 2000, p. 33; Kershaw, 2008). Regardless of their "true" intentions, over the next month German high command and Hitler studied the situation before coming to conclusion. On 16 July, Hitler ordered that preparations for Operation Sea Lion, the invasion of the United Kingdom, begin.

Two counterfactuals are commonly derived from this approximate point in time, albeit usually by popular writers as opposed to academic historians. The first, less frequent and more exuberant counterfactual requires a singularly minded and aggressive Hitler who desires the immediate destruction or subjugation of the British. This entails a quick and decisive strike at British defences in order to exploit the losses and confusion after the evacuation from Dunkirk (Forester, 1971; Macksey, 1980; Messenger, 2002). This counterfactual is untenable for a few reasons beyond Hitler's desire to defeat Britain: the Luftwaffe would need to make French resources and bases available earlier as well as have suffered fewer losses than they factually did in the first phase of the Battle of France. Indeed, the Luftwaffe lost "about half its operational strength" in that campaign and needed to rebuild both its fighters and bombers (Bungay, 2000, p. 104 – 5).

The more restrained counterfactual instead supposes that after a long but ultimately successful campaign to weaken the British, the Germans find themselves in position to attempt invasion in September of 1940 (Cox, 1974; Kieser, 1997; Evans and McGeoch, 2004; Forczyk, 2016). To do so requires that the Germans have control of air space that is defended by "11 Group", the RAF forces in charge of the defence of south-eastern England. This counterfactual more closely matches the German plans for invasion.

Whether Nazi Germany truly intended to invade or not, the Luftwaffe was tasked with achieving air supremacy over the English Channel, south-eastern England, and London. The Luftwaffe's primary goal then was to deplete the forces of RAF's Fighter Command in the region before too late in October, when the weather alone would be enough to prevent achieving air supremacy. In the mind of Luftwaffe leader Reichsmarschall Hermann Goering, this meant drawing up, engaging with, and destroying the enemy fighters in the air (Townsend, 1970). To accomplish this, the Luftwaffe would employ both fighters and bombers: the bombers would destroy targets on the ground which the RAF was obliged to defend. The RAF fighters that were scrambled to intercept the bombers would then be targets for the Luftwaffe's own fighters.

Just as the Germans were preparing for the invasion, the RAF was preparing to repel them. The country was divided into regions, the aforementioned groups, that would be defended by portions of Fighter Command. The RAF also recognised German air supremacy as necessary for an invasion and that Fighter Command needed to hold off the Luftwaffe while Bomber Command and the Royal Navy prevented the build up of German landing craft.

The RAF had a few key advantages that it would leverage in its defence. While the Luftwaffe could not know how the RAF would respond to its forces due to the asymmetry of attacking versus defending over foreign soil, the RAF had a strong information-gathering infrastructure in place to rapidly inform them of and respond to the Luftwaffe's movements (Holwell and Checkland, 1998; Bungay, 2000). Due to the asymmetry in objectives (the Luftwaffe needed to defeat the RAF, while the RAF needed to merely not be defeated), the British also could rescue their pilots, while the Luftwaffe had no accurate way of knowing how much damage their tactics caused. Furthermore, the British had a high fighter production and could rapidly replace lost aircraft, while the Luftwaffe were spread thin from repeated conflicts (Deighton, 1977; Bungay, 2000). Park could not only rapidly respond to threats, but he could also decide exactly how he wanted to respond, although his available choices in practice were to not respond, respond with small forces, or to assemble a large force while the Luftwaffe bombed its targets.

The main constraint Park faced instead appears to be pilot training. Data on pilot training suggests that newly trained pilots were matching losses (due to injury or death) during the battle (Dempster and Wood, 1961). On the other hand, RAF Fighter Command's leader Air Chief Marshal Hugh Dowding strongly believed that neither a novice monoplane fighter pilot, nor an experienced pilot of other types, was worth a combat-experienced fast monoplane fighter pilot (Dowding, 2015). Furthermore, such training could not be easily ramped up, as might aircraft production lines.

Faced with the need to protect his resources, Park chose to use his men sparingly, emulating guerrilla warfare in the air: Park would receive reports of German numbers and respond with as few forces as he could to disrupt the enemy bombers and fighters. Perhaps unsurprisingly, Park's decision was controversial within the RAF; RAF Air Vice-Marshal and 12 Group commander Trafford Leigh-Mallory insisted that aircraft be gathered together in a massive flight to engage the enemy. This tactic, the usage of "Big Wings", mirrored the belief of Goering that aircraft should be destroyed in the air (Townsend, 1970). This debate would end up sullying the reputation of Park, resulting in lack of recognition for leading the bulk of the defence against the Luftwaffe (Bungay, 2000).

3.3 Modelling the Battle

One major question that can be asked is if Park's general approach was correct or if Leigh-Mallory's approach would have been more effective against the Germans. Questions of this nature were already being asked before World War II. A particularly important set of models are Lanchester's Laws, which were formulated specifically to model aerial combat by Lanchester (1916) and help us understand the scalings of attrition.[15] This question was addressed by Johnson and MacKay (2011) and MacKay (2011), who found that Lanchester's equations did not fit the data

[15] Although they bear Lanchester's name they were not *only* developed by Lanchester, as discussed in the introduction. Most of the formulations of the equations are actually for naval warfare (Chase, 1902; Baudry, 1914; Fiske, 1916). Of particular note is Chase (1902), who developed the idea over a decade earlier and Osipov (1915), who attempted it for land combat with inhomogeneous forces.

well. Their results indicate that the aerial combat more closely resembles duels, instead of a situation where aircraft can combine their fire to quickly destroy the smaller enemy.

The situation is actually even graver for Leigh-Mallory's approach than we would expect from the conclusion that aerial combat resembled duels. While combat does more closely resembled duels than square-law aimed fire, days with more sorties *favoured the Luftwaffe*. MacKay (2011) suggests that Park was leveraging the aerial equivalent of guerrilla warfare's cover and concealment. Leigh-Mallory would eventually be able to apply his approach over France, which, for a variety of reasons including lessons not learned by Leigh-Mallory in the Battle of Britain, result ed in large RAF losses (Bungay, 2000).

More generally, MacKay (2011) and Horwood *et al.* (2014) argue that, not only is the Battle of Britain not effectively modelled by Lanchester's Laws, but air combat in general does not accept such a model. Johnson and MacKay (2011) highlight a few points that make the Battle of Britain tractable for non-parametric analysis, however. There is very little obvious change in the data, although they stop short of a full changepoint analysis and do identify a possible change on September 15th in the loss-rate per total sortie number. Johnson and MacKay (2011) additionally observe a strong degree of spatial homogeneity, and the Battle is homogeneous temporally with respect to the types of units employed (Dempster and Wood, 1961).

The non-parametric tool we use here is the bootstrap, discussed in Section 2.2 (Efron, 1979a; Efron and Tibshirani, 1993). The core of the bootstrap is to create artificial samples to statistically analyse by re-sampling with replacement from the original sample, which is viewed as drawn from a single distribution. This allows one to explore the effects of natural variance on the data if the data is independently and identically distributed (IID). When the data is days of a campaign, each artificial sample can be thought of as a counterfactual campaign, in which, more than merely rearranging the days, we have sampled some days more than once and some days not at all.

This naivety seems unwarranted; in a typical campaign, each day depends steadfastly on the previous. Each soldier or pilot who is lost is not there the next day and, even worse, some of these losses would appear to occur multiple times with the lost potentially harming their enemy in the meantime. We can reconcile the naive approach and historicity however. Instead of viewing each loss as occurring to a specific individual, we instead view each loss as an event affecting the strength of a side. In this way, we assume that if a day is drawn twice, its events happen to a different set of people in subtly different but numerically equivalent ways.

Each day is factually unique and no surprise attack should be able to happen twice. We must turn to the historical context: the Battle of Britain was fought with consistent equipment, personnel, weather, and tactics (Dempster and Wood, 1961). There were no surprise attacks in the Battle of Britain due to the British information systems, and very little damage was actually done to this infrastructure. The closest case, argued by Bungay (2000, p. 368–9), is the closure of Biggin Hill for a few hours. What about the days where the Luftwaffe did very well when bombing infrastructure? First, there were many targets for infrastructure, especially given that the Luftwaffe were unsure as to what infrastructure was being used by the RAF (Bungay, 2000). Hence, the Luftwaffe could reasonably choose a different target to attack in the same manner with ease. Second, the RAF's damages were due less to the bombing that occurred and more to the dogfighting that ensued (Ramsey, 1989; Honour Roll). Together, these imply that attacks on airfields were reproducible in the sense of being draws from the same distribution.

There is still the problem of the possible existence of changepoints in the data. Changepoints are a problem in that they mark places where the underlying distribution before and after a point in time is (statistically significantly) different. To circumvent this problem, we will need to identify any changepoints, preferably with historical justifications, and treat each segment as distinct. This motivates a natural order to the data; all data within the first segment should arrive before the second segment's data and so on. Note that we do not order the data within a segment and that the data within each segment then should be IID. Our result then is a counterfactual which will follow the logic of the original campaign with some variation in the dates of transition between each phase of the campaign.

Merely bootstrapping a campaign helps us to understand the variations that may have arisen naturally from the historical actors' decisions, but it does not help us understand the decisions themselves. In order to better understand, for example, the costs of Hitler's hesitance to attack the British, we need to do something more. We modify the bootstrap by reweighting it: changing from picking days uniformly at random to picking days with a bias determined by the details of the day and the details of the counterfactual. These reweightings allow us to bias the bootstrap to sample more or less data than what factually occurred based on the counterfactual changes to the historical actors and the orders given. Crucially, we cannot answer our counterfactual questions directly due to never having observed the true victory condition. We will need to rely on our prior beliefs in conjunction with the weighted bootstrap and historical context in order to analyse the counterfactuals we propose.

Due to the constraints of the data and historical context, we cannot, for example, have Leigh-Mallory lead 11 Group. We do not have robust or extensive data for how he would have led 11 Group and, furthermore, Park's approach was exclusively defensive and minimal. We can, on the other hand, examine German decision making. One famous example of such a change is the switch to targeting London, which has been declared a crucial error by many authors including contemporary actors (Spaatz, 1946; Churchill, 1950; Townsend, 1970), airpower theorists (*e.g.* Warden, 1989, p. 103), and historians (*e.g.* Macksey, 1987, p. 45). The Luftwaffe clearly had no idea of how to achieve air supremacy (Bungay, 2000, Ch. 30). Each of its targets, Channel shipping, airfields and aircraft factories, and London, was an attempt by the Luftwaffe to guess at how best to achieve air supremacy using bombers and fighters. They knew they needed to destroy RAF Fighter Command's aircraft, but the Germans had no way of truly knowing which targets Fighter Command felt it needed to defend and no way of knowing how much damage their bombers were actually achieving. As such, modifying targets chosen is an obvious counterfactual.

A different alternative is to alter the beliefs of the German leadership. For example, what would happen if Goering had not shared Leigh-Mallory's opinion that airframes are to be destroyed in the air in large-scale confrontation? Such a Goering might focus more heavily on destroying airframes where ever they might be found, in turn concentrating on the airfields. The success of such a tactic has been disputed in the literature. Bungay (2000) maintains that even attacks on airfields were ineffective, but Warden (1989) holds it to be the correct concentration and application of airpower. Goering is not the only German leader whose belief we can alter. We also consider a more aggressive Hitler, one who clearly prioritises the removal of the British from the war above the now diminished French forces.

No matter the modifications that we use for the bootstrap, we are still bound by the historical and physical context. Our aggressive Hitler still must coordinate, plan, and prepare with his army, navy, and air force. In particular, the Luftwaffe was only able to use appropriate French bases during June, so we can bring the campaign forward at most three weeks. Coordination with the army and navy is also costly, regardless of the counterfactual scenario, due to the need to cross the Channel. To do so, the invasion force would need neap tides to optimise their odds of surviving the crossing. Such tides allow the invasion barges freedom of movement, reduce tidal races, and assist barges with beaching just before dawn with some moonlight (Grinnell-Milne, 1958, Ch.7: "Time and the Tides"; German Plans, 2017). This constraint results in a fortnightly cycle of viable invasion dates: within 3 days of the quarter-moons on 26 August, 8 September, 24 September, and 8 October, for which a decision to invade must be made ten days prior. The first of these invasion dates is only available for counterfactuals where the campaign begins earlier as well, since the factual first feasible date was mid-September.

3.4 Methods for the Battle

In order to better understand the Battle of Britain, we gathered a variety of data, presented in Tables B.1 and B.2. Table B.1 consists of British and German total airframe losses (Hurricanes, Spitfires, or otherwise; and fighters, bombers or otherwise; respectively), British pilot casualties, and primary target type (C = docks, shipping and coastal, R = reconnaissance merely, A = aerodromes, L = London, Kent and Thames estuary). Table B.2 additionally provides the number of British sorties and (British estimates of) German sorties as well as regional weather (W = rain (W for wet), C = clear, O = overcast).[16]

We also divide the campaign into four phases in Table B.1, following the official history (James, 2000). These are (P1): 10 July - 7 August, principally of coastal attacks and armed reconnaissance; (P2): 8 August - 18 August, of heavy attacks on mostly coastal targets; (P3): 24 August - 6 September, of sustained attacks gradually concentrating on airfields; and finally (P4): 7 September - 31 October, following the Luftwaffe's switch to London as its principal target. Note the five-day lull between (P2) and (P3) and denoted 0 in Table B.1; we treat this lull as a separate phase in our scenarios. We do not include 25 September and 16 October when reweighting based on target; these days correspond respectively to an attack on Filton and to general German air sweeps.

In Appendix A, we present an analysis of the IID assumption for this data. We find that we cannot reject stationarity within each phase for British pilot casualties except for in Phase 4, which is instead consistent with the presence of a trend. Autocorrelations identified that are significant appear to be false positives when controlling for the false discovery rate. Hence, the pilot casualty data appears to be mostly consistent with the IID assumption within phases.

We now need a "victory" criterion, by which we mean a trigger for invasion. Within RAF and War Cabinet policy there is no one obvious change which might have occurred and would have constituted the defeat of Fighter Command; the need to retain a force capable of contesting invasion, combined with the ability to withdraw Fighter Command northwards, ensure this. Further, we find no evidence of a plan for such a single collective withdrawal, which would probably rather have been gradual, and not immediately obvious to the Germans, who sometimes erroneously assumed that unrelated airfields belonged to Fighter Command (Deighton, 1977, p. 216; Bungay, 2000, p. 208, 211).

[16] Airframe losses were compiled from Ramsey (1989). British pilot casualties were also compiled from Ramsey (1989), and were cross-checked with the RAF Battle of Britain Honour Roll and with Wynn (2015). Targets are from Dowding (1941-1947). Weather is from the RAF Campaign Diaries. Sortie numbers are from James (2000). German sortie numbers are not generally available before 1st August.

In terms of the constraints on Fighter Command, it was pilot supply that approached criticality. British monoplane fighter airframe production, running at about 400 per month (Dempster and Wood, 1961, p. 104), would always have been sufficient to provide a modern aircraft for every available pilot. In contrast the supply of newly-trained pilots was running at around 260 per month, supplemented by non-U.K. volunteers and refugees, and pilots re-allocated from other types. As noted above, Dowding considered a novice fighter pilot worth less than a combat-experienced one, and this is consistent with, for example, the 501 Squadron figures (Bungay, 2000, p. 373), with 33% of novice pilots lost within a month compared with 22% of experienced ones. If we therefore assign a novice pilot a value of $67\%/78\% = 0.85$ of a pilot lost, a reasonable estimate of the strength of Fighter Command, $BS(t)$, might therefore be $BS(t) := B_0 + sbt - BL(t)$. Alternative methods of assigning value to novice pilots are, of course, possible. For example, one might consider ratios of daily survival probabilities or ratios of mean lifetimes (measured in days or months). We choose to use the 501 figures as presented in the source material so as to minimise assumptions used. We anticipate, however, that the way we are using the bootstrap should be insensitive to such choices due to the way we calibrate the bootstrap to prior knowledge.

Here B_0 is the initial Fighter Command pilot strength, $B_0 = 1259$ on 6 July 1940 or $B_0 = 1094$ on 15 June, all assumed to be trained and experienced (Dempster and Wood, 1961, Appendix 11). The new pilot contribution is $s = 0.85$, $b = 11$ is the average daily complement of new pilots, and $BL(t)$ is the total number of British pilots lost so far, so that $BS(t) = B_0 + sbt - BL(t)$ is the total Fighter Command pilot strength at time t. Pilot losses $BL(t)$ are computed using data from a number of sources, given in Table B.1: essentially they include all pilots killed, seriously wounded or missing in action, and exclude the slightly wounded, who typically returned to action within a few days if not within the same day. For reference, the total number of (possibly non-unique) pilot losses, wounded, and slightly wounded over the course of the Battle are 487, 236, and 165 respectively in our data. In contrast, the number of pilots available in October (dependent on week) is listed by Dempster and Wood (1961, Appendix 11) as about 1,700. Assuming that pilots available does not include pilots lost, this suggests around 2,100 pilots were employed over the Battle of Britain.

We now need to decide how the values of $BS(t)$ might be used to trigger invasion. Let T (for "T-Tag") be the planned date of invasion, which we recall must lie within three days either side of the quarter-moon Q. Recall further that an initial decision for invasion needs to be taken on $T-10$. For invasion to be triggered, the Luftwaffe must appear to have been gaining air superiority during (let us say) the five days before this, and for invasion not to be cancelled or postponed the same must apply for five days after the initial decision – beyond this, too much effort would have been put into preparations such as the sowing of minefields for poor air combat reports to cause cancellation. So we choose some critical value BS_C according to the procedure outlined in the next paragraph, and say that invasion is triggered if $BS < BS_C$ throughout the period from $T-15$ to $T-5$ for any T between $Q-3$ and $Q+3$. Thus the latest date for which we simulate air combat is 6 October. Naturally, some of these assumptions are difficult, if not impossible, to validate. We again will rely on the calibration of our bootstrap to the prior beliefs of its users in order to provide robustness to changes in the assumptions.

What critical threshold value BS_C of BS would constitute the defeat of Fighter Command? As noted earlier, the crux of our method is not to attempt to answer this directly, but rather to calibrate it to prior beliefs using bootstrap methods. Imagine three historians of differing views.

Left	0.495	0.140	0.832	3.15	0.433
Right	0.186	0.649	1.699	0.967	1.523

Table 3.1: Example sample from a two armed bandit. The true distribution for the left arm is the absolute value of a normal distribution with zero-mean and a standard deviation of 1.22. The true distribution for the right arm is an exponential distribution with a rate parameter of 5. Note that, in truth, the values obtained randomly for the right arm are fairly unusual and underscore the importance of online learning methods in bandit problems.

One of them believes that the British margin of victory was nil – that the battle was won on a coin toss – and thus that the Germans had a 50.0% probability of victory. A second believes that the British had a modest margin of victory, that it would have taken a moderate amount of deviation from the mean for the Germans to win, and thus that the British probability of victory was 84.0%, corresponding to one standard deviation σ from the mean in a normal distribution. A third believes that a German victory was very unlikely, and would have taken a 2σ event or change and thus that the British probability of victory was 97.7%. We then run a simple bootstrap on the Battle of Britain as actually fought, which results in a bell curve of outcomes centred on the actual outcome, and choose the three values of BS_C which generate the three historians' British victory probabilities specified above. We return to the normality assumption in Section 3.6, but find that the assumption suffices for our purposes.

We then use these three values of BS_C in our counterfactual scenarios, resulting for each scenario in three new probabilities. These are robust to small changes in the form of the victory criterion, since this merely mediates between the figures of interest, which are each historian's belief (expressed as a victory probability estimate) about the actual battle, and the belief which it would then be rational for them to assign, on the basis only of the evidence from the actual fighting, to each counterfactual scenario.[17]

We next need to map our counterfactual scenarios to the bootstrap, which we will do by reweighting sampling of days from uniform sampling. This is particularly simple to accomplish within the Bayesian framework discussed in Section 2.2. Instead of taking a prior which in the limit is uninformative, we instead take an informative prior. One simple example is the following: suppose that we play a two-armed bandit ten times, pulling each arm five times and obtain the results in Table 3.1 (see, *e.g.*, Slivkins, 2019, for a review of bandit problems and algorithms). Assuming that each pull is independent, we might be curious about how mixtures affect the number of draws d to reach a certain score, say 10.

Using n for the number of draws and $\tilde{\mathcal{P}}$ as the empirical distribution, the Bayesian unweighted bootstrap then proceeds by sampling, say, $B = 100$ times from Dirichlet $\left(n\tilde{\mathcal{P}}\right)$. The weighted version instead uses Dirichlet $\left(n\tilde{\mathcal{P}} + \mathbf{Be}\right)$, where $\mathbf{Be}$ (as in prior belief) is our vector of prior weights, analogous to $\epsilon\mathbf{1}$ in Section 2.2. Note that each component of $\mathbf{Be}$ must be strictly greater than the negative of its respective component in $n\tilde{\mathcal{P}}$. We then consider five $\mathbf{Be}$ corresponding to an unweighted 50-50 split, a 25-75 split, and a 0-100 split and vice versa. To determine our number of draws d for each sample to reach 10, we then use the data $\mathbf{x}$ and bootstrapped weights

[17] That is, to the extent to which they believe that additional days' fighting would have followed the pattern of actual days' results. Since it uses only the actual days of fighting, what the bootstrap cannot do is to include any of the unlikely things that one could argue might have happened, but did not. There are no "black swans" in bootstrapped counterfactuals (Taleb, 2007).

Left-Right	Min.	1st Qu.	Median	Mean	3rd Qu.	Max
100-0	4.854	8.191	10.532	10.991	13.036	27.212
75-25	4.815	8.598	10.239	10.589	12.179	21.565
50-50	5.066	8.845	10.314	10.325	11.567	17.232
25-75	5.732	8.798	9.989	10.354	11.576	16.544
0-100	7.576	9.294	10.131	10.276	11.184	14.785

Table 3.2: Example summary statistics for Bayesian bootstrapping a two armed bandit. Bootstrapping is conducted over the results from Table 3.1 using different weightings. Values reported correspond to the bootstrapped d^* from Equation (3.1).

Left-Right	Min.	1st Qu.	Median	Mean	3rd Qu.	Max
100-0	4	9	10.5	11.57	14	23
75-25	4	9	11	10.90	13	18
50-50	4	9	11	11.21	13	18
25-75	6	9	10	10.38	12	16
0-100	8	9	10	10.43	11	14

Table 3.3: Example summary statistics for frequentist bootstrapping a two armed bandit. Bootstrapping is conducted over the results from Table 3.1 using different weightings. Values reported correspond to the bootstrapped d^* from Equation (3.2). Compare with the results in Table 3.2.

$\mathbf{P}^*$ to calculate

$$d^* = \frac{10}{\mathbf{P}^* \cdot \mathbf{x}^\intercal}. \tag{3.1}$$

We present some simple summary statistics in Table 3.2. Due to the better median (note that the means are essentially the same at 1.01 for the left arm and 1.00 for the right arm), we see nicer tail behaviour if we focus on the right arm, but can drastically decrease our minimum by 'getting lucky' from a few pulls of the left arm.

As the bootstrap becomes more complicated, these complexities can make the system seem particularly unphysical. For example, why should one be able to get only some proportion of a draw from an arm, as opposed to whole numbers? A more physical approach is to, analogously to how we changed $\mathbf{Be}$ to a non-zero vector, introduce such a vector in our multinomial model. The simplest way to do this is to take our multinomial probabilities to be the normalised Dirichlet parameter vector. This simplification to the frequentist bootstrap has another benefit however. Since we will be drawing events, it is easier to measure a more physical d^* that depends on the order of the draws:

$$\text{choose } d^* \in \mathbb{N} : \sum_{i=1}^{d^*-1} x_i^* < 10 \leq \sum_{i=1}^{d^*} x_i^*. \tag{3.2}$$

This conforms more with our physical intuition of the system as well, while retaining the ability to modify our strategy from the Bayesian bootstrap, albeit at the loss of not being able to sample from the entire space of possible values. (To accomplish such a "physical" interpretation for the Bayesian bootstrap would require an additional step of using the sampled distribution as a prior for a multinomial distribution.) The results from the frequentist approximation are shown in Table 3.3.

Henceforth, we use the frequentist approximation of the weighted bootstrap. Our re-weightings will come from our counterfactual scenarios, and will be used to preferentially select certain days, analogous to choosing to pull the left or right arm of the bandit. Here, the arms of our bandits

Scenario	Summary	Starting Date	Bootstrap Reweighting (Days of Target or Phase)			
			P1	P2	P3	P4
Real	Actual Battle	10th July	29	11	19	30
CF1	No switch to London	10th July	29	11	49	0
CF2	Early Start	16th June	32	18.7	23.8	30
CF3	No London and early start	16th June	32	29.7	37.8	0
Scenario	Summary	Starting Date	Airfield	Coast	London	Recon
Real	Actual Battle	10th July	16	47	36	13
CF4	Target airfields	10th July	43	33	0	13
CF5	Target airfields and early start	16th June	54	42	0	17

Table 3.4: Summary of counterfactual scenarios.

correspond to the phases or targets used in the actual battle, and each draw is the number of RAF pilot casualties. Next, we need to determine how each counterfactual affects the strategies used for pulling the arms of our bandits.

In the most radical counterfactual (CF) fiction the Luftwaffe's initial hopes of swift achievement of (at least) air superiority are realized, followed by an early invasion (Forester, 1971; Macksey, 1980; Messenger, 2002). We cannot and do not pursue such ideas: rather our counterfactuals are air campaigns which depart from the actual campaign in their dates or targeting but which are built up using data from it. Instead we consider five counterfactuals that can be well-posed in terms of our bootstrapping method. These are summarized in Table 3.4.

CF1: What if the switch to bombing London had not occurred?

That the Luftwaffe switch to bombing London was an error is a standard argument, as noted earlier. To capture it here we simply extend P3 to 6 October, eliminating P4 entirely.

CF2: What if Hitler had been fundamentally in favor of invasion from the outset?

In this case we assume that planning would be brought forward: German navy commander Grand Admiral Erich Raeder's visit to Hitler on 21 May would, in its effects, have taken the place of that of 20 June; air campaign planning would have been initiated much earlier than the actual 30 June (Macksey, 1980, p. 13). We take the net result as bringing forward the air campaign by three weeks – as much as seems reasonable given the Luftwaffe's need to make the Channel-littoral airbases operational. Thus we bring forward P1 to 16 June - 17 July, and spread P2 and P3 proportionally over 18 July - 6 September, with P4 thereafter. Since the battle begins early, this also gives time for the Germans to take advantage of the 26 August neap tides. We anticipate that this strengthens the German position considerably due to cutting into the number of pilots the RAF can field and granting the usage of the extra and early neap tides. As we will see in the day-to-day distributions of the number of pilots (looking ahead to Figure 3.3), the early days of the Battle are important in building up the strength of the RAF forces.

CF3 combines CF1 (no fourth phase) and CF2 (early onset):

We take CF2, but with no switch to London: bring forward P1 to 16 June-17 July, and spread proportionally P2 and P3 over 18 July-6 October.

For our next counterfactual we switch from contracting or prolonging phases to alterations of targeting. Recall that in the actual battle the numbers of days for the principal target types were $(A, C, L, R) = (16, 47, 36, 13)$.

CF4: What if Goering and his staff had believed that Fighter Command could be more easily destroyed on the ground than in the air?

Townsend (1970, p. 325)[18] notes the belief of both Goering and staff officer Paul Deichmann that Fighter Command would be more easily destroyed in the air than on the ground (paralleling the beliefs of Big Wing advocate Trafford Leigh-Mallory in the RAF). Indeed, Townsend (1970, p. 360) records Deichmann's view that the Luftwaffe should not destroy radar stations, whose work would simply bring the RAF's fighters to the Luftwaffe's, facilitating their destruction. Thus for this counterfactual we take an eighty-nine-day battle terminating on 6 October, with R unchanged, L untargeted, and A exceeding C, with $(A, C, L, R) = (43, 33, 0, 13)$.

CF5 combines CF2 (an early start) with CF4 (targeting of Fighter Command on the ground):

We take $(A, C, L, R) = (54, 42, 0, 17)$ to combine a commitment for invasion with a firm belief in the destruction of Fighter Command on the ground as a prerequisite.

In Section 3.6, we additionally investigate some of the problems caused by trying to determine the impact of the weather on the Battle of Britain. We use our attempt to create counterfactual weather for the Battle primarily as a cautionary tale. We also present some work that might be able to handle these problems with future development.

3.5 Bootstrapping the Battle

Before tackling the counterfactuals, we first need to apply the unweighted bootstrap, creating many new samples (henceforth we call these re-samples "trials") of the Battle of Britain from the actual days' fighting. We begin with the results of a bootstrap with 100,000 trials. Compared to a standard $10{,}000$ trials this will allow us to better fit a normal distribution and thereby obtain suitable critical values BS_C. We then compare the effects of bootstrapping to the battle as actually fought with a standard 10,000 trials. With these baselines in mind, we can then proceed to address the genuine counterfactuals.

As discussed earlier, to begin exploring the counterfactuals we must first obtain the threshold values which match various prior beliefs as to the probability of the Luftwaffe's obtaining air superiority. To do so, we will impose a normal distribution on the results of a large number of trials without any reweighting.[19] The results of such a calibration run are shown in Figure 3.1. We observe immediately that the normal distribution provides a good fit (validated further in Section 3.6). Thus we can take the expected value (the average or mean) of the normal distribution as the 50.0% threshold (median). The standard deviation and second standard deviation to the left of the average then agree well with our desired 84.0% and 97.7% probabilities of British victory. The corresponding thresholds are $1{,}437.5$, $1{,}383$, and $1{,}328.5$ pilots respectively. To summarize: the historian who believes that the German invasion decision was evenly-balanced would use a threshold pilot strength of $1{,}437.5$ in our victory criterion, while the historians who believe in moderate and large British margins of victory would use $1{,}383$ and $1{,}328.5$ respectively.[20]

[18] Townsend himself is quite clear: "Goering made a crucial error. For Fighter Command was more vulnerable on the ground than in the air" (p.379).

[19] To be clear: we are not imposing a normal distribution on the number of pilots $BS(t)$ evaluated on a single day, nor on one neap cycle's possible launch dates. Instead, we are imposing a normal distribution on the lowest number of pilots during any of the three possible neap windows. In this way, we ensure that the probabilities correspond to whether or not Germany launches an invasion at all during a given trial, and prevent over-counting trials where the number of pilots remains low multiple times.

[20] Note that fractional values are possible due to our assigning less value to new pilots in $BS(t)$.

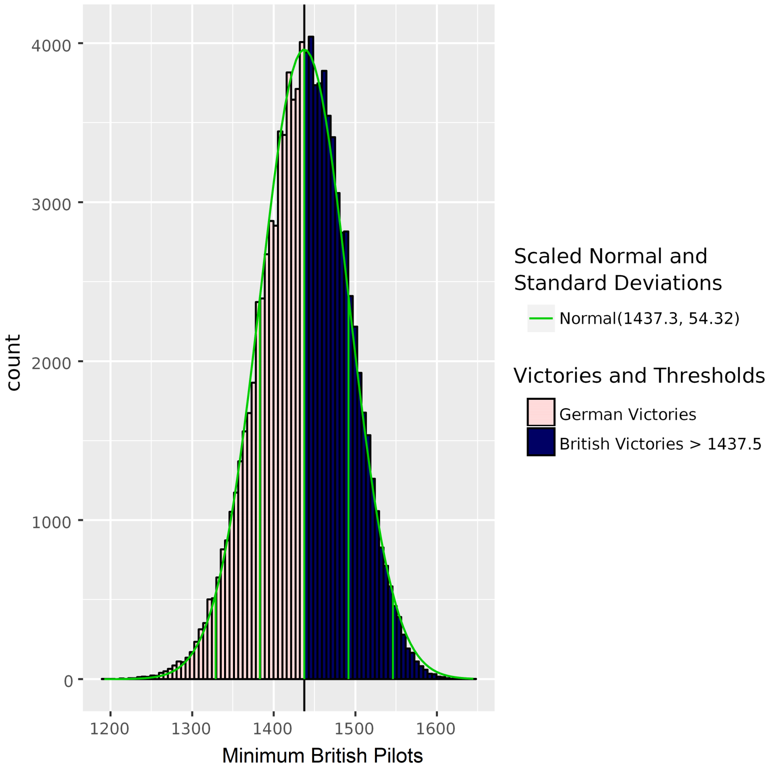


Figure 3.1: A calibration run in which $100{,}000$ trials are run and then fitted with a normal distribution (superimposed green curve). We use this run to inform our choices of thresholds, here shown as the leftmost two superimposed vertical green lines, representing two and one standard deviations below the mean, and the solid black line dividing German victories from British victories, representing the mean.

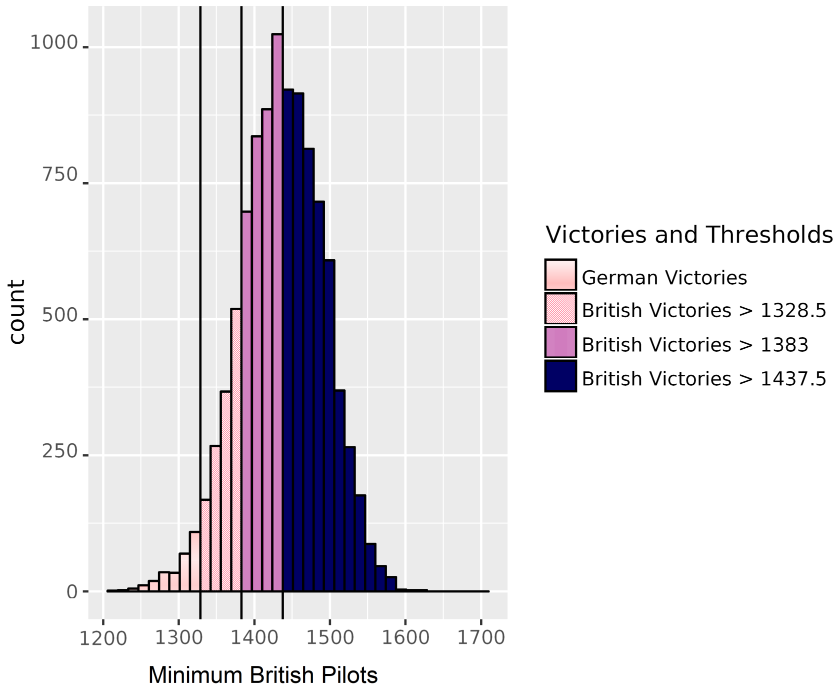


Figure 3.2: Bootstrapping the Battle of Britain with sampling in proportion to the phases as actually fought. The thresholds correspond to 97.2%, 83.9%, and 49.5% probabilities of British victories.

Next, we compare our now-calibrated bootstrap to the battle as actually fought. We use the same number of trials as for our counterfactuals: 10,000. Owing to the lower number of trials, we should expect more variation, meaning that the results will be less precise and may deviate from our ideal values. The results of this run can be seen in Figures 3.2 and 3.3. Figure 3.2 is equivalent to Figure 3.1, but now for the smaller number of trials. Figure 3.3 shows plots of the day-to-day number of British pilots, with the actual number of pilots (by day) superimposed.[21]

CF1: What if the switch to bombing London had not occurred?

The results of our first counterfactual, where the Luftwaffe does not switch target to London (that is, enter into P4), are shown in Figure 3.4. It is immediately clear that the probability of British victory has significantly decreased. If one believed that the British won the real battle with probability 50%, the implied threshold now yields a British victory probability of just 9.1%. Our lowest threshold, which had given the British a victory probability of 97.7% in the real battle, brings this down to 63.7%. This reinforces the common narrative that the switch to targeting London was a mistake.

CF2: What if Hitler had been fundamentally in favour of invasion from the outset?

Our second counterfactual is grimmer still for the British: an eager Hitler pushes for an earlier beginning to the campaign, catching Fighter Command with approximately 165 fewer pilots ini-

[21] A box-and-whisker plot shows the spread of data by way of five values – the median, quartiles and 5th and 95th centiles – supplemented by outliers. The middle value is the median, above and below which 50% of the data lie. The lower (respectively, upper) edges of the boxes correspond to the 1st (resp. 3rd) quartiles, which divide the data into the lowest 25% (resp. 75%) and highest 75% (resp. 25%). The distance between the 1st and 3rd quartiles is then used to compute the locations of the whiskers, beyond which all points are considered outliers. As we should expect, the number of pilots from day to day in the actual battle matches well with the trends of the bootstrap, due to ordering the days of our trials in order of phase.

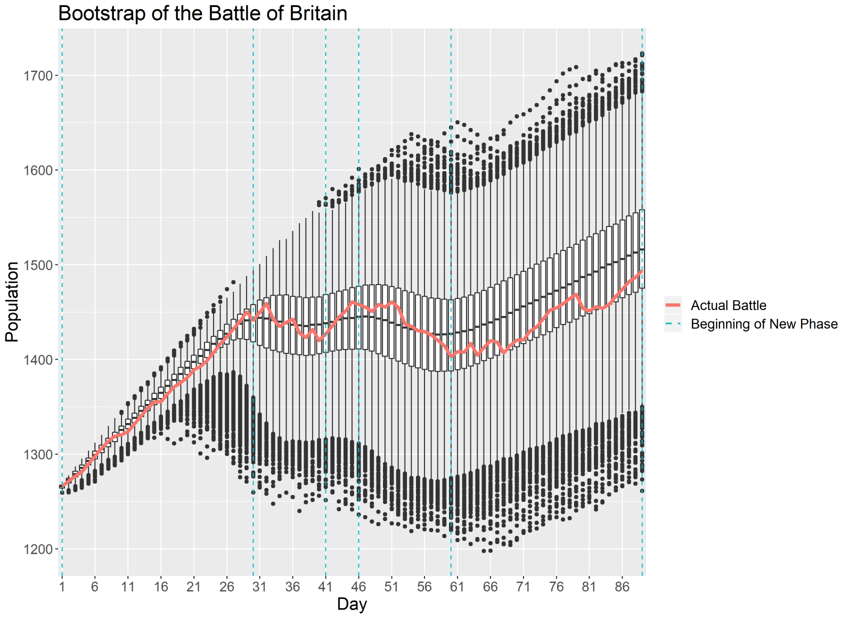


Figure 3.3: Bootstrapping the Battle of Britain with sampling in proportion to the phases as actually fought. The box-and-whisker plots show the day-to-day distributions of the number of pilots, the dashed vertical lines show boundaries between the phases of the battle as actually fought, and the solid curve is the number of pilots in the battle as actually fought.

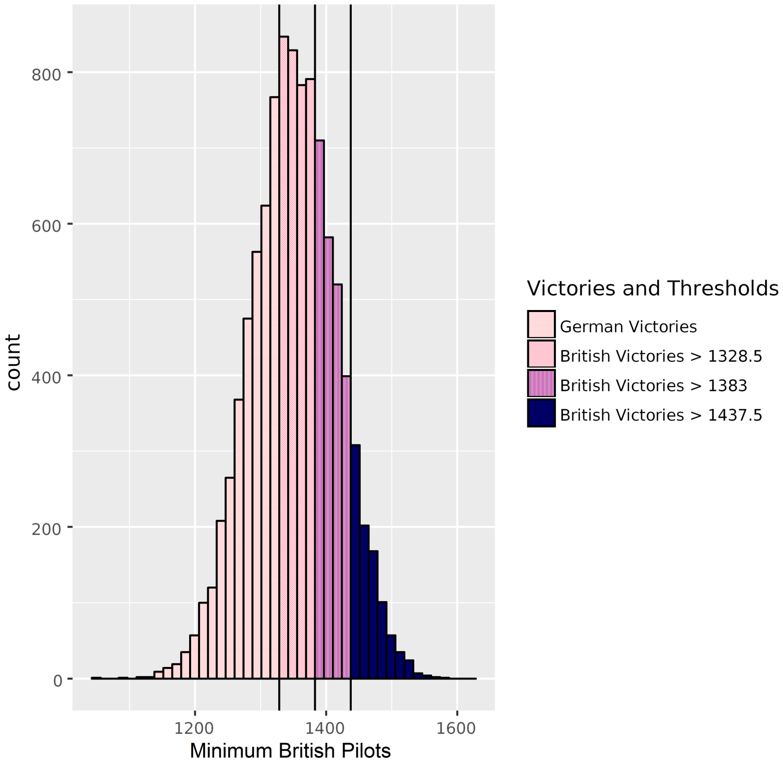


Figure 3.4: Bootstrapped results of the scenario where the Luftwaffe had not switched to targeting London by entering the fourth phase. Our thresholds now correspond to 63.7%, 31.2%, and 9.1% probabilities of British victory.

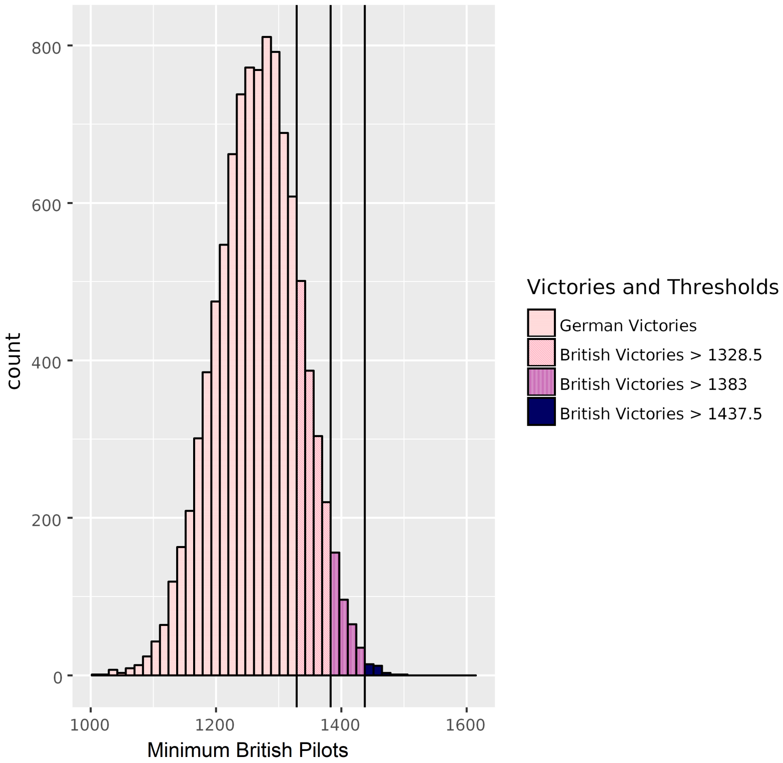


Figure 3.5: Bootstrapped results of the scenario where the Luftwaffe began their assault earlier. Our thresholds now correspond to 18.0%, 3.8%, and 0.3% probabilities of British victory.

tially available. The constant threat allows the Luftwaffe to engage earlier RAF pilots who in the real battle would have had more training and combat experience. Additionally, it gives the German forces access to the earliest possible invasion date on 24th August. Figure 3.5 shows the damage this does to the probability of British victory: the 50% victory possibility has now become 0.3%, and the most optimistic 97.7% threshold is now just 18.0%.

CF3 combines CF1 (no fourth phase) and CF2 (early onset):

As one might now expect, combining an early attack with no switch to London decreases further the viability of the British defense. This counterfactual also helps remind us of the prospect of diminishing returns: British chances are not utterly destroyed by the combined changes. The probability of British victory is simply reduced further: the 50% victory probability is now 0.1% (CF2: 0.3%), while the 97.7% is now 8.3% (CF2: 18.0%).

CF4: What if Goering and his staff had believed that Fighter Command could be more easily destroyed on the ground than in the air?

Instead of a change in the phasing of the battle, we now investigate the effects of not attacking London at all. We have already mentioned the German belief that it was easiest to destroy the RAF in the air, lured up by bombers, especially over London. The effects of this belief are brought out by our fourth counterfactual, in which the Luftwaffe focuses much more on the airfields. Results are shown in Figure 3.7. The results of CF4 are a more drastic variation of CF1 (no fourth phase), but are inevitably closely aligned with it. The 50% victory probability is reduced to 1.1% (CF1: 9.1%) and 97.7% to 33.6% (CF1: 63.7%). Not only should the Germans not have made their early-September switch; they paid dearly by choosing to attack London at all.

CF5 combines CF2 (an early start) with CF4 (targeting of Fighter Command on the

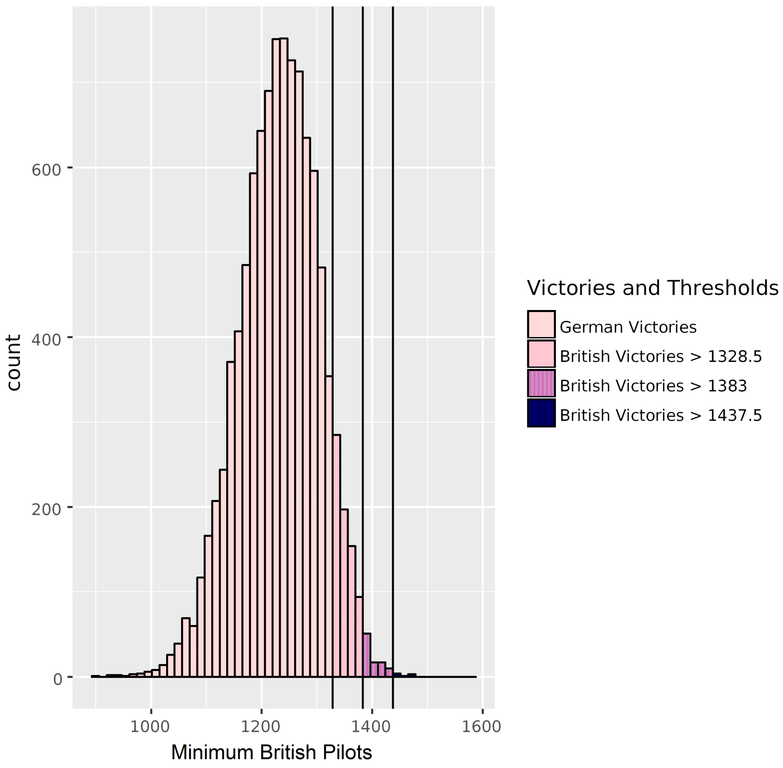


Figure 3.6: Bootstrapped results of the scenario where the Luftwaffe began their assault early and did not choose to switch to primarily targeting London by entering the fourth phase. Our thresholds now correspond to 8.3%, 1.0%, and 0.1% probabilities of British victory.

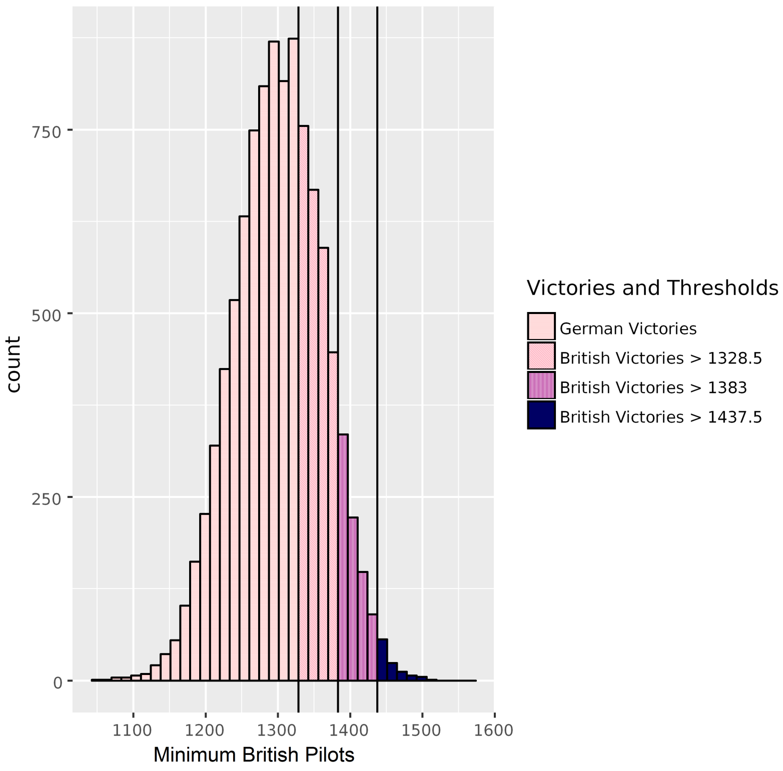


Figure 3.7: Bootstrapped results of the scenario where the Luftwaffe neglected to attack London entirely and had instead focused on the airfields. Our thresholds now correspond to 33.6%, 9.0%, and 1.1% probabilities of British victory.

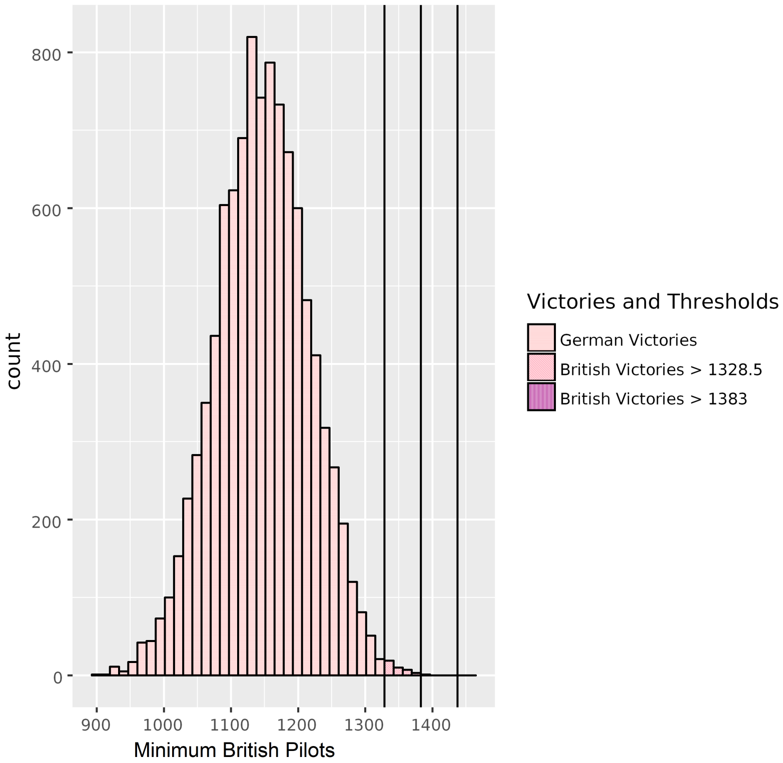


Figure 3.8: Bootstrapped results of the scenario where the Luftwaffe began their assault early as well as not attacked London and instead focused on the airfields. Our thresholds now correspond to 0.4%, 0.01%, and almost-nil probabilities of British victory.

ground):

We now come to our last counterfactual: if Hitler had been eager for invasion, giving the Luftwaffe an early start, and if the Luftwaffe had been dedicated to targets associated with destroying the RAF on the ground. This is our most negative counterfactual for Britain, as shown in Figure 3.8: if one believes that the probability of British victory in the battle as actually fought was 50%, or even 84%, then this alternative yields fewer than 1 in $10{,}000$ victories for Britain. If one holds that the probability of British victory was 97.7%, this situation yields just 0.4% for the British. However much one might not believe in a "narrow margin" of British victory in the battle actually fought, the British aerial victory, it seems, depended very strongly on poor German choices. We summarize the results of all our counterfactuals in Table 3.5.

Scenario	Summary	Probabilities of British Victory by Threshold		
		1,437.5	1,383	1,328.5
Real	Actual Battle	49.5%	83.9%	97.2%
CF1	No switch to London	9.1%	31.2%	63.7%
CF2	Early Start	0.3%	3.8%	18.0%
CF3	No London and early start	0.1%	1.0%	8.3%
CF4	Target airfields	1.1%	9.0%	33.6%
CF5	Target airfields and early start	0%	0.01%	0.4%

Table 3.5: Summary of counterfactual scenario outcomes.

3.6 Refinements

The bootstrap is a flexible tool and can be used to explore more than counterfactuals. We can also explicitly examine the effects of each of the strategies employed by the Luftwaffe alone. While these are not counterfactuals per say, they do help us understand the extent to which each strategy was effective in and of itself. With this knowledge, we can better understand why the counterfactuals in Section 3.5 have the results that they do.

We begin by noting that each of the histograms in Figures 3.1 – 3.8 presented looks very normal. One might wonder if this is statistically the case; we consider quantile-quantile plots to check for normality. Such a plot will appear linear if the empirical distribution is effectively a rescaling and translation of the reference distribution (here, normal). We apply this method to each of the seven datasets generated in Figures 3.1 – 3.8 in Figure 3.9. We note that the normality of the tails in each case is subject to questioning, but that in each case, the center of the distribution (roughly quantiles $[-2, 2]$ or approximately the central 95% interval) is fairly close to normal.

Next, we consider how the various strategies employed by the Luftwaffe related to its performance in battle. Each strategy can be seen in Figures 3.10 – 3.13. As much as London was touted as a critical failure for the Luftwaffe, it would appear that the Luftwaffe's worst performance was actually at the Channel targets, which the Luftwaffe targeted 47 times. Strangely, the Luftwaffe performed worse when attacking Channel targets than they did when performing mere reconnaissance. London was only targeted 36 times, 22 of which were in October, putting many of these days as out of the range of the final quarter-moon. It is not so much that the switch to London was an absolutely bad move (the shift indicates it was actually slightly better than the average), but instead that it was relatively better for the British than attacks on Aerodromes by a wide margin. It seems that it would have been better for the Luftwaffe to have not attacked the Channel as opposed to not attacking London.

Furthermore, Figures 3.10 – 3.13 suggest that normality is not guaranteed and should not be taken for granted with the bootstrap. In this case, normality fails drastically for targeting only the Channel as can be seen from Figure 3.11 or from Figure 3.14. In this case, we would need to rely on BC$_a$, see Section 2.2 and discussed shortly in this section, in order to obtain accurate estimates, although the counterfactuals that these represent are obviously implausible for historical reasons (*e.g.* requiring the Germans to commit to performing only reconnaissance in an attempt to obtain aerial supremacy or otherwise committing to a single tactic when they historically varied their tactics due to a lack of information).

While the methods used previously are sufficient to give a general idea as to how the bootstraps behave, there do exist refinements, as discussed in Section 2.2. We construct alternative 95% confidence intervals with the emphasis on the lower confidence point (since we are interested in the right confidence interval). For comparison, the intervals are for the unweighted case and are measured over the lowest number of pilots available during any invasion opportunity across a single run of the bootstrap. (We use 95% since this is a standard confidence interval size and the lower confidence point is at 2.5%, which is approximately the 2.3% from two standard deviations that we used in practice.)

For the bootstrap-t method, we calculate the bootstrap standard errors by a second round of bootstraps. When calculating the BC$_a$ method, we encounter a problem: since our statistic of interest is based on the data order, performing the jackknife, the usual method of calculating the acceleration of the standard error, does not make sense. We note that our model generates a single

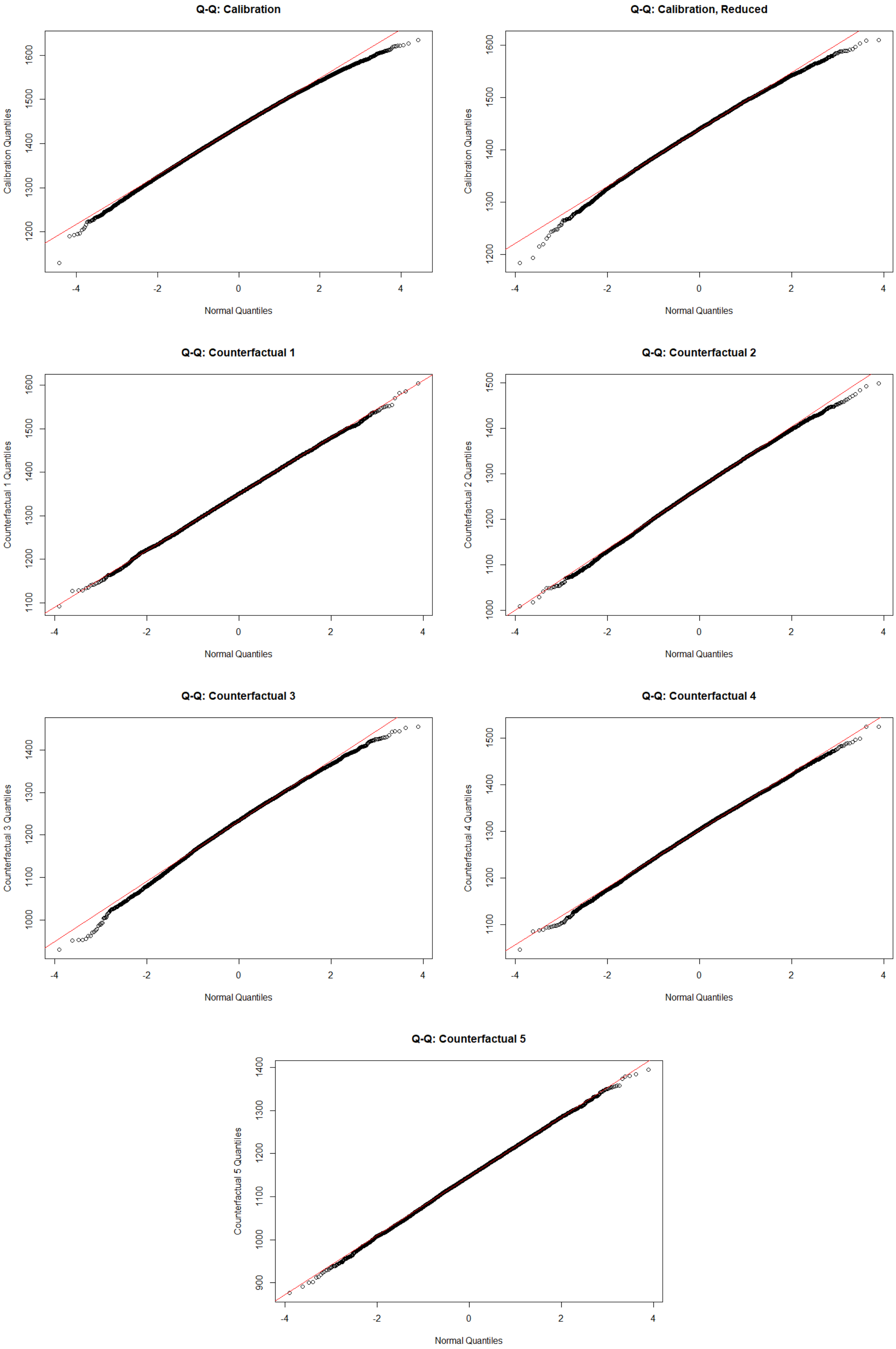


Figure 3.9: Quantile plots comparing the counterfactual bootstraps to normal distributions. If the bootstraps were identical to normal distributions, we would expect them to follow the red line; deviations from the line are deviations from normality. We see such deviations commonly in the tails, but the bulk of the distributions are well-behaved.

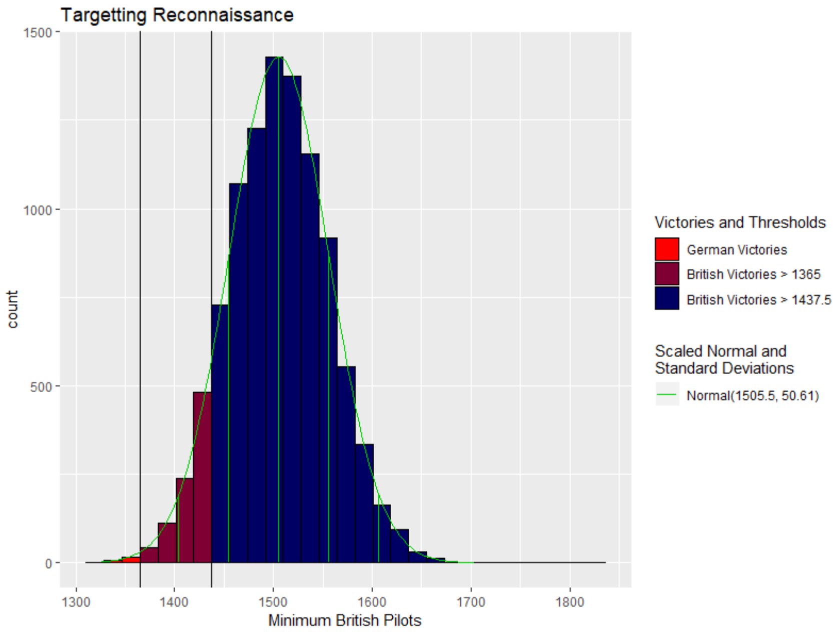


Figure 3.10: Histograms of bootstraps conducted using only reconnaissance targets. The vertical black guiding lines correspond to 1365 (approximately 90% chance of British victory in the original Figure 3.1) and 1437.5 (50%); the bars with at least 1437.5 pilots are presented in blue, bars with at least 1365 are presented in purple, and the remaining bars are presented in red. Smooth overlaid green lines show a fitted normal distribution. Note the positive shift corresponding to approximately a 1.2σ shift in the mean.

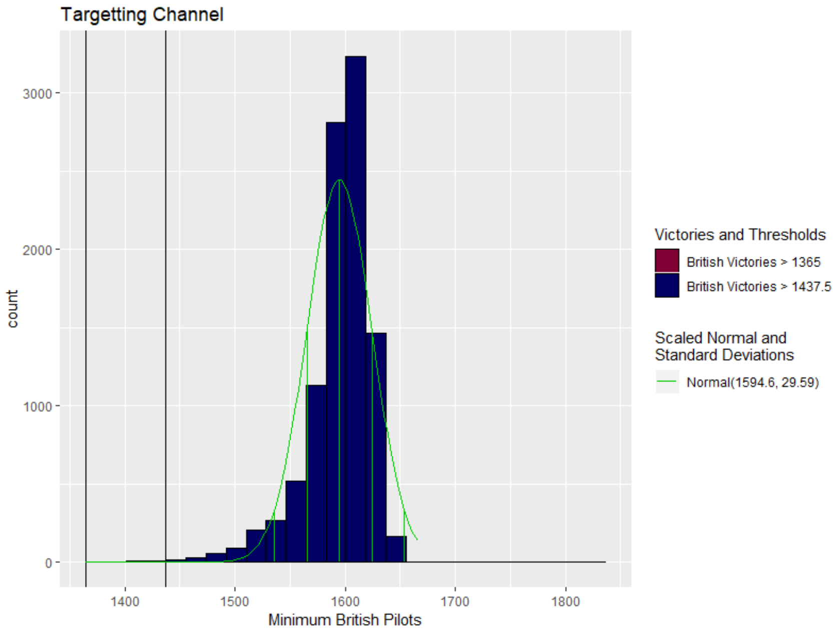


Figure 3.11: Histograms of bootstraps conducted using only Channel targets. The colors are as in Figure 3.10. Note the extreme behaviour. The magnitude of the shift in the mean is approximately 2.9σ.

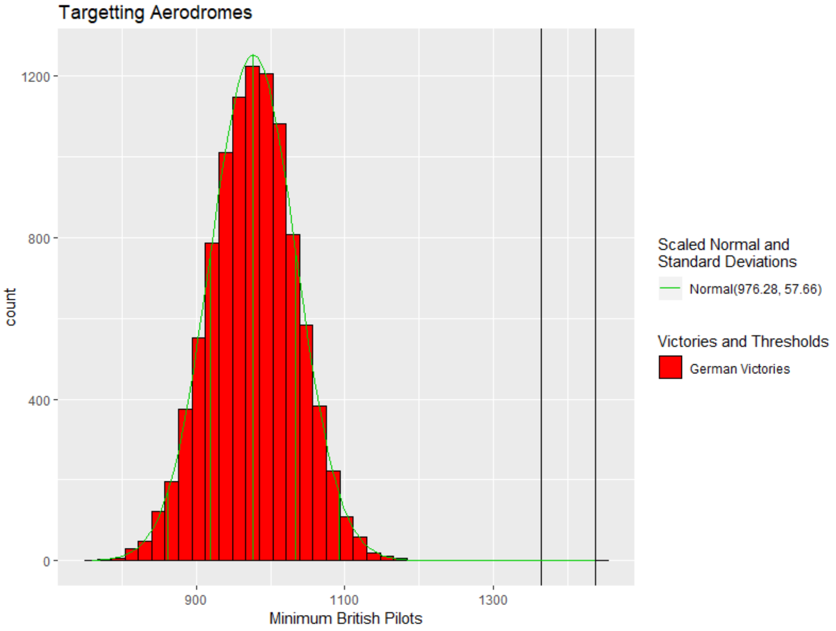


Figure 3.12: Histograms of bootstraps conducted using only aerodrome targets. The colors are as in Figure 3.10. Targeting the aerodromes clearly has the deadliest results of the lot for the British. The magnitude of the shift is approximately 8.5σ.

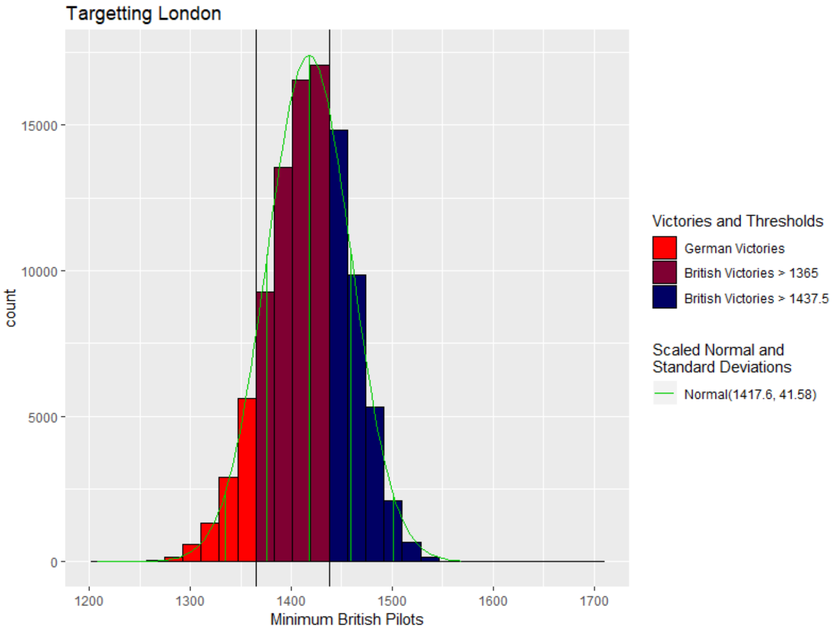


Figure 3.13: Histograms of bootstraps conducted using only London targets. The colors are as in Figure 3.10. In contrast to Figures 3.10 – 3.12, London has nearly average behaviour. The magnitude of the shift is approximately 0.37σ.

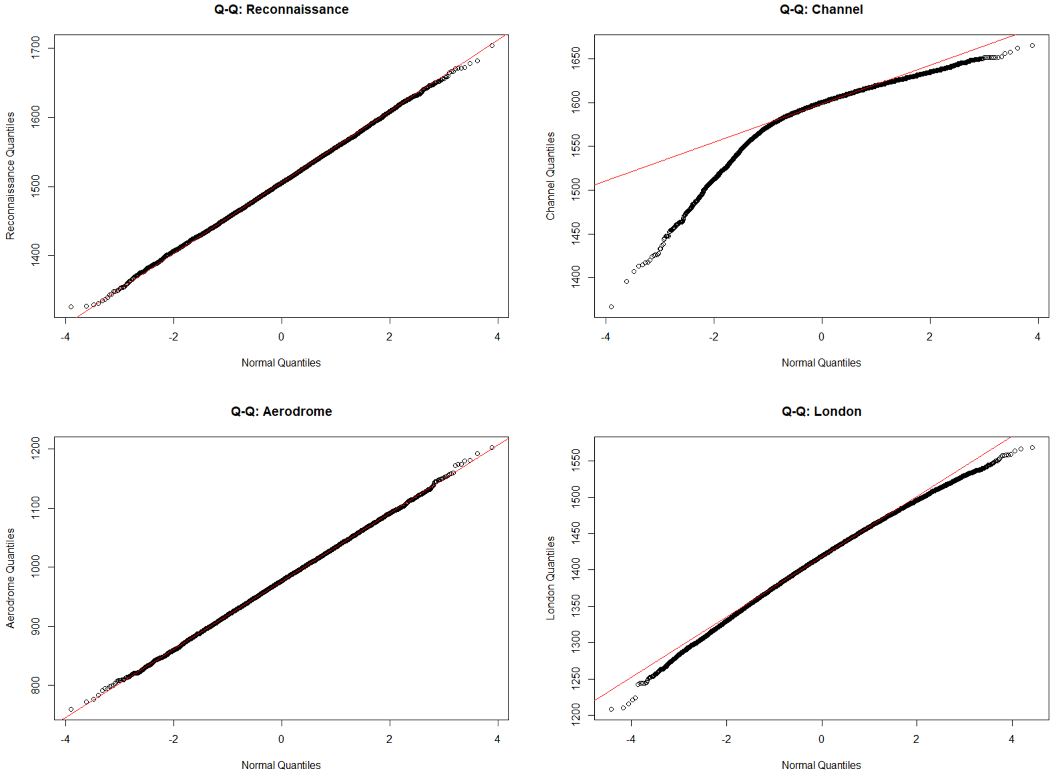


Figure 3.14: Quantile plots comparing the targeting bootstraps to normal distributions. If the bootstraps were identical to normal distributions, we would expect them to follow the red line; deviations from the line are deviations from normality. The targeting of only the Channel leads to large deviations from normality, while reconnaissance and aerodrome targets are extremely normal.

Method	Left (2.5%)	Right (97.5%)
Original	≈ 1331	≈ 1544
Percentile	1328.15	1539.91
Normal	1314.36	1525.84
Student's t	1314.35	1525.85
Bootstrap t	1292.56	1506.18
BC$_a$	1195.15	1477.65
Causal-Overall	1384.05	1490.25
Causal-Phase	1364.50	1500.65

Table 3.6: Results for applying various confidence interval methods for bootstrapping the Battle as actually fought. The original values are computed using a normal distribution with mean 1437.5, standard deviation 54.32. Methods are discussed in Section 2.2. Causal bootstrapping receives additional focus in the text. To calculate the standard error for the bootstrap t, we performed a second bootstrap for each of the bootstrapped samples.

value though, which means that the acceleration should be approximately the bias-correction $\hat{a} \approx \hat{z}_0$ for large n (Efron and Tibshirani, 1993, Ch. 22.5). We must be cautious about this source of additional error, especially considering BC$_a$ is expected to be more accurate than the other methods employed (Efron and Tibshirani, 1993, Ch. 14).

The results are available in Table 3.6. Compared to our original lower confidence point, which we can estimate using a normal distribution as 1331, we see that there is a fair amount of variation in the exact value that each method reports. Most striking is the BC$_a$ method's large shift in values, which suggests that more than 100 less pilots are needed to to achieve 97.5% confidence in British victories. Indeed, its right confidence point 1477.65 is just above the 50% confidence point in our original bootstrap 1437.5, indicating the sheer scale of the difference. This may be a result of the $\hat{a} \approx \hat{z}_0$ approximation, but it also might be a result of the properties of BC$_a$. The bootstrap t and BC$_a$ methods are both second-order accurate, but the percentile and BC$_a$ methods are both transformation respecting, *i.e.* commute with transformations on the bootstrapped data. If both of these properties result in shifts in the same direction, this might cause the shift we see in the BC$_a$ results.

Despite the theoretical advantages to the alternative thresholds proposed, we maintain that the results in Section 3.5 are our current standard. If the alternative bootstraps are better, they would imply our chosen standard is too conservative and that the Germans had less chance of achieving aerial supremacy than stated before, which is consistent with the standard historiographical views regarding the margin of the Battle of Britain. On the other hand, the other methods have their own flaws. The normal and Student's t distributions both require parametric assumptions that may not be in line with the asymptotic distribution actually reached (*e.g.* see Figure 3.11). As we noted in Section 2.2, the bootstrap-t method can overcome problems due to mismatches in higher order moments, but, as noted by Efron and Tibshirani (1993, Ch. 12), this method can be unstable and poorly behaved. The BC$_a$ would normally be the best of these methods, but, as noted above, the standard method for using the BC$_a$ method does not appear to be applicable in our situation. Additionally, as we shall see, it is not naively obvious how to utilise the causality in the data appropriately to use the causal bootstraps. We thus prefer the percentile method that we used in Section 3.5, but the theoretically superior BC$_a$ and causal bootstraps are excellent avenues for future research.

In contrast to the other results in Table 3.6, the causal bootstrap appears to have much narrower boundaries. We end our discussion of refinements to the bootstrap with the causal bootstrap, as discussed in Section 2.2. We consider each combination of weather and target as a treatment and we investigate the difference in treatment effects for each factual day. For simplicity, we use the simple minimisation copula, although we note that further work could be done on identifying the most appropriate copula. In contrast to the classic bootstrap, causal bootstrapping proceeds by randomising the assigned treatment, rather than randomising the days in the sample. One simple issue that we run into is whether the data imputation should be run over the whole of the dataset or on each subset phase; for completeness we perform both.

A standard, not reweighted run of the causal bootstrap can be seen in Figure 3.15, comparable to Figures 3.1 and 3.3. The means are quite similar: our calibration run, Figure 3.1, had a mean of approximately 1437.3, while the within phase imputation yields 1432.9 and the overall imputation yields 1437.2. Much of the change we observe is within the estimate of the standard deviation: originally we had 54.32, but now we have 34.46 and 27.45. This is suggestive of the over-coverage of the classic bootstrap claimed by Imbens and Menzel (2018): compared to the 1331 number of pilots required for 97.5% confidence of a British victory, the within phase bootstrap yields 1364.50, while the overall bootstrap has 1384.05. The difference between the within phase and the overall figures is the amount of data used for the imputation: when imputing by phase, we are much more likely to have only one or two events of a given type compared to imputing over the entire data set. It is not clear that the extra information is accurate for each phase though, given the possibility of changepoints, suggested by the official history by James (2000) and discussed in Appendix A. The within phase imputation instead respects these proposed distributional changes.

In contrast to the methods in Section 3.6 which are all refinements to the frequentist and Bayesian bootstraps, the causal bootstrap works under a different statistical paradigm. The central conceptual difference is whether we are exploring the distribution of potential days or whether we are exploring the distribution of the effects of treatments. The former requires a less informed decision as to what is being modelled; the latter requires one to identify the variables that are to be considered treatments as well as specify a model as to how the treatments relate (in the form of the copula). For example, it is not obvious that weather should be one of the variables that is considered a treatment, or merely a variable to control for when applying the imputation. Furthermore, the causal bootstrap requires a deeper statistical knowledge than the standard bootstrap does. While the causal bootstrap, like the other refinements, is useful future research, we again consider the standard bootstrap more appropriate, especially given the contrast with traditional improvements on the bootstrap, as seen in Table 3.6.

In particular, the causal bootstrap might have a particularly useful application when dealing with confounding variables. The bootstrap relies on the idea that the data is IID. Conversely, hidden common factors or changes in the distribution can greatly affect the results of a bootstrap. Weather, despite seeming to be an obvious variable to consider in our analysis, exhibits problems of this sort. We will attempt to make the weather “more typical” in our bootstrap, but we will run into problems regarding data quality and mutual dependence on time that make this bootstrap ill-posed.

We begin with retrieving data, but obtaining properly detailed data has proven difficult. The data we use here is the 1929-1979 record of sunlight hours from the Meteorological Office’s Oxford weather station, which provides typical ranges of sunlight during July to October, in conjunction with the more qualitative Meteorological Office 1940 Daily Weather Reports. This procedure

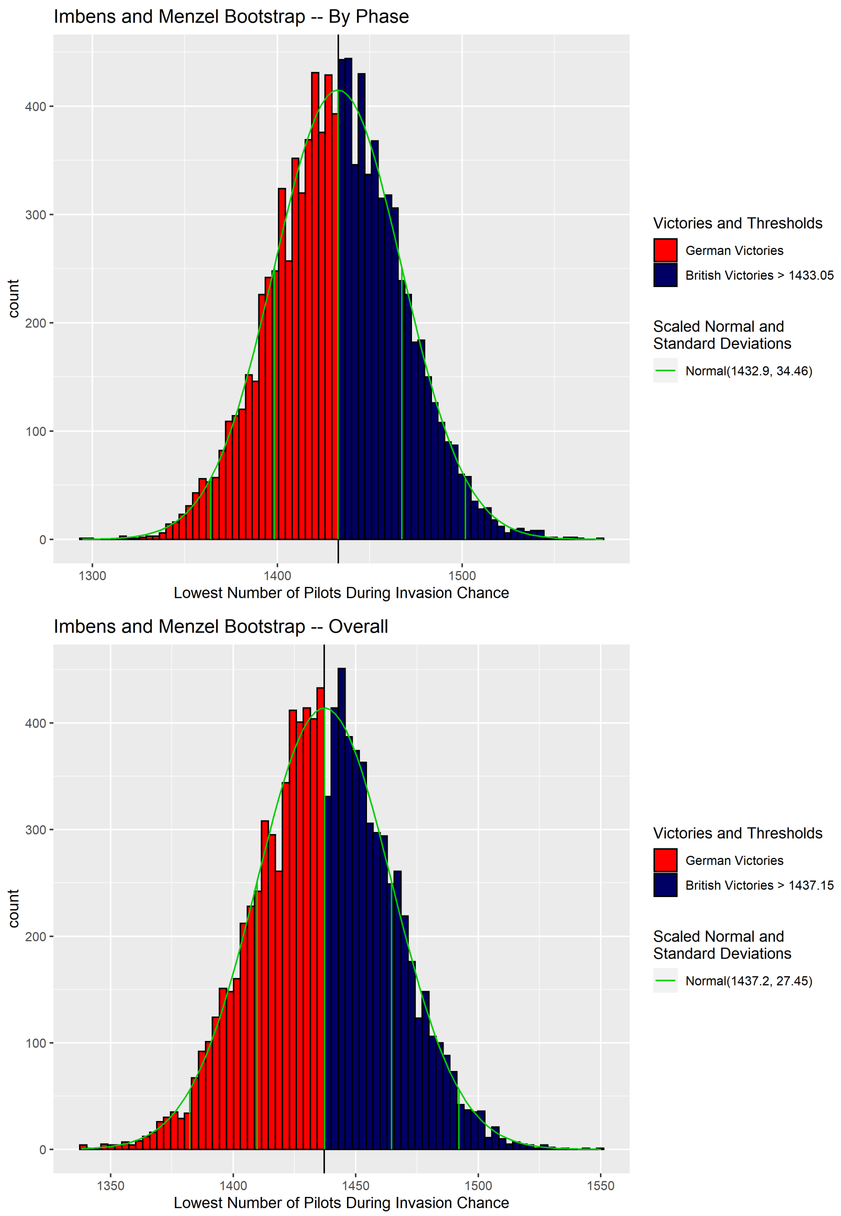


Figure 3.15: We apply the causal bootstrap to the Battle of Britain data, treating each weather and target combination as a treatment and using the simple model posed in their work. Above, we apply this model within each phase, while below we apply it to the overall data set. In each case, we also fit a normal distribution to the bootstraps. Note the subtle difference in scalings of the horizontal axis, indicative of the change in scale of the distributions. Compare Figures 3.1 and 3.3. Figure 3.1 has a mean of approximately 1437.3 and standard deviation of 54.32.

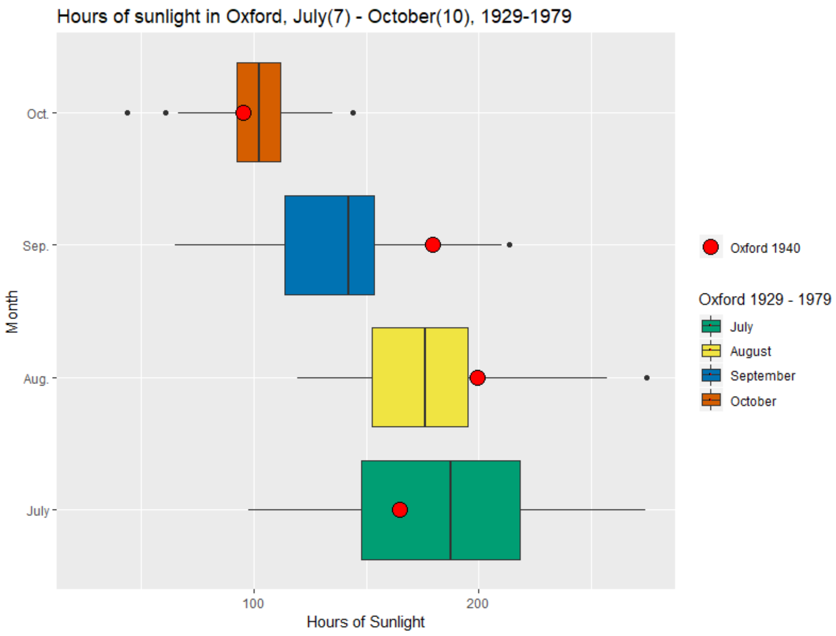


Figure 3.16: Box plots of the monthly hours of sunlight from July through October from 1929 – 1979. We indicate 1940 data with large red dots. Data was obtained from the Meteorological Office's Oxford weather station.

	Target's Weather		
Phase	C	O	W
1	9	17	3
2	6	5	0
0	0	5	0
3	9	1	4
4	13	25	17

Table 3.7: Contingency table counting the weather of the primary target in each phase.

results in qualitative and geographically restricted data. We then can compare this data with the weather data from each target attacked during the Battle of Britain. First, we note that the amount of sunlight, which we will use to help rescale the number of clear days, is not actually that unusual on a month to month basis for (Oxford in) 1940, seen in Figure 3.16. Furthermore, there is no natural distinction between months; to bootstrap weather properly, we should instead use *daily* estimates, although we were unable to find such.

Next, we can illustrate the problem with time dependence. One might naturally expect that weather does have an effect on the ability of airframes to do combat, but there is a problem testing this assumption. We naturally expect seasonal variation, but this seasonal variation takes place at the same time that the changes in phases occur. We demonstrate this with Table 3.7, which is a contingency table of phase and target weather. Even ignoring the shortest phase P0, there are differences in the proportions of, say, rainy days between each phase. As such, it should be unsurprising that a Pearson's Chi-squared test rejects the null hypothesis (with $p = 0.0015$) that the phase and weather are independent. As for the actual values of each type of weather, we used the median hours of sunlight for each month to adjust the number of clear days and proportionately rescale the number of cloudy and rainy days within that month. This rescaling's effects are shown in Table 3.8; the change has little obvious effect.

	Original Weather			Adjusted Weather		
Month	Clear	Overcast	Rainy	Clear	Overcast	Rainy
July	7	12	3	8.0	11.2	2.8
August	13	16	2	11.5	17.3	2.2
September	13	7	10	10.3	8.1	11.6
October	0.8	3.5	1.7	0.8	3.4	1.7

Table 3.8: Attempt at making weather more like the median using hours of sunlight. The July weather represents the number of days of July actually in the battle. The October values are averaged over the month and then scaled down to the number of days included in the bootstrap.

Proceeding to bootstrap by trying to make weather more like the median creates only small changes in the probability of a British victory: the 50% victory probability increases to 58% and our 97.7% increases to 98.5%. This is likely due to the change in phase sampling however: for example, the Channel in the middle of August during Phase 0 (P0) was exclusively overcast, and P0 saw few losses. Increasing the number of overcast days in August would make P0 probabilistically longer, which aids the British. We present this cautionary tale because there is no obvious way for the bootstrap to inform its user that something has gone wrong. The user is responsible for verifying the bootstrap's assumptions.

This is not to say that there is potentially no recourse to handle the weather and its entanglement with phase. The causal bootstrap views the problem as one of data imputation. We have already imputed the effects of the weather on a specfic day's results and bootstrapped on the imputed data by choosing amongst the weather and target combinations for each day. Now we apply a reweighting scheme by changing the ratio of treatments to favour those involving better weather, as suggested by Table 3.8.

In the original reweighting, we observed that the 50% victory probability increased to 58% and the 97.7% increased to 98.5%, but that this was likely due to the correlation between weather and phase. Here, phase has not changed, only the treatments by change of weather. The results can be seen now in Figure 3.17. In comparison to our previous weather attempt, which saw our 50% threshold increase to 58%, the within phase model has 50% increase to 58.6% and the overall model has 50% to 56.3%. We therefore conclude that, for this simplistic model of weather, the relationship between phase and weather is not the driver of the increase, as the days and their phases remain fixed in the bootstraps. Instead, this suggests that the relationship between target (for which phase would have acted as a proxy) and weather results in the change. Performing a Pearson's Chi-squared test sees rejection of the null hypothesis (with $p = 0.041$) that the target and its weather are independent, and this result strengthens if the targeting of Filton and the air sweeps are excluded from the data ($p = 0.019$). One could, in principle, fix the targets and phase and bootstrap only on the weather, but we stop here due to the weakness of our weather data. Improved weather data would make this a natural avenue for future research.

3.7 Discussion

In this chapter, we have discussed applications of the bootstrap and its variations to the Battle of Britain. At the surface level, our results confirm the standard view that the German Luftwaffe made a serious mistake when they chose to target London in Phase 4, rather than continue their assault on aerodromes as in Phase 3. The Battle of Britain appears to have been won on a narrow

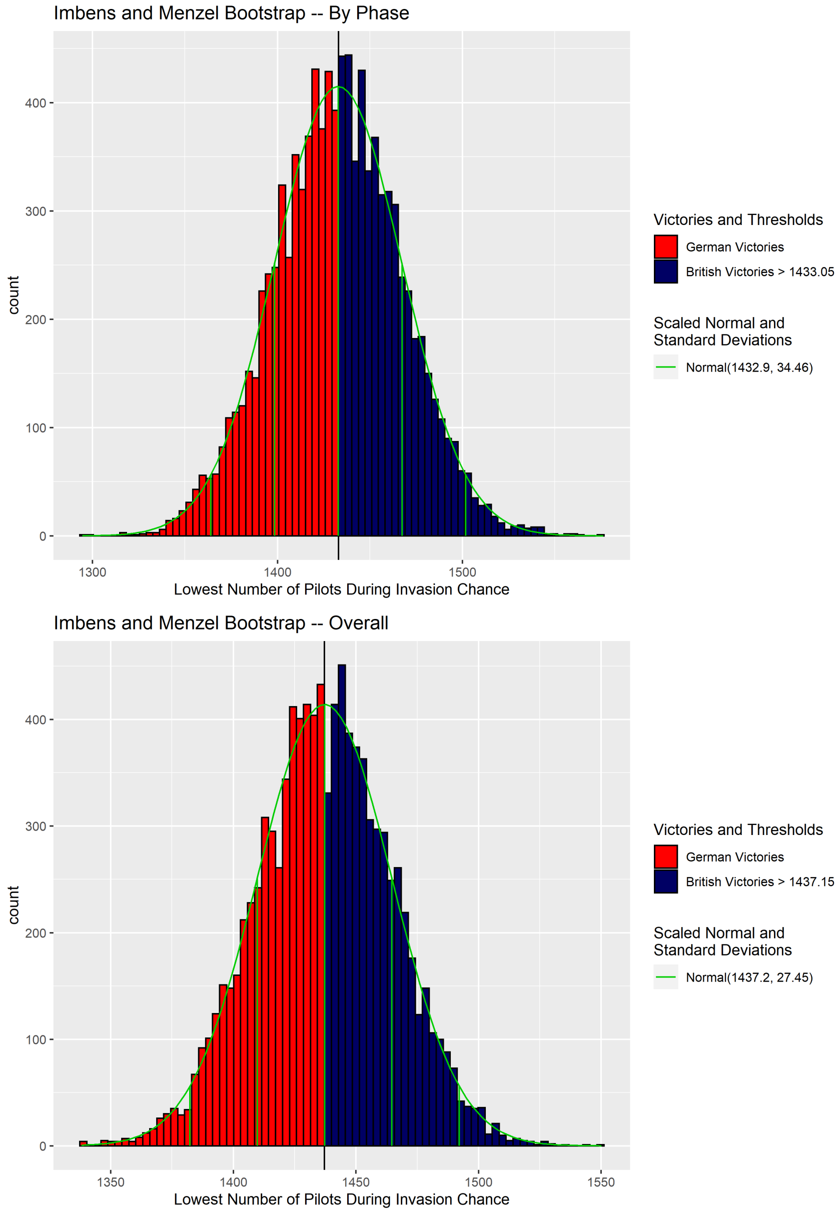


Figure 3.17: Causal bootstrapping the weather of the Battle of Britain. Treatment is as in Figure 3.15, except we counterfactually change the weather to make it more "average", as suggested by Table 3.8. The vertical line marks the mean of the normal distribution fitted to the bootstraps from Figure 3.15. When applying the classic bootstrap counterfactually by changing the proprotions of days, we had observed that our 50% threshold had increased to 58%. When applying the model within each phase, we instead see an increase from 50% to 58.6%. When applying the model to the overall data set, we see an increase from 50% to 56.3%.

victory, although not perhaps in the sense of barely having superior forces. Rather, the victory appears to have been narrow in that the Germans had the capability to win, but did not know how to achieve this victory.

Our results are more nuanced then this, however. While the standard view is that switching to targeting London was a critical error (Spaatz, 1946; Churchill, 1950; Townsend, 1970; Macksey, 1987; Warden, 1989; Bungay, 2000), this was not the largest mistake of the Germans. Eliminating London (**CF1**, Figure 3.5) results in a change of about one standard deviation σ, while targeting London alone would have been comparable to the battle as actually fought with a shift of about 0.37σ. Two other mistakes seem far more important factors in the Germans' loss. The Luftwaffe had its poorest performance when it was instead attacking Channel shipping, which, comparing strategy to strategy and ignoring confounding relationships with weather (*i.e.* attacks on the Channel were far more likely to take place when weather was overcast or clear in comparison to attacks on London, for which rainy days were more common), had an almost 3σ shift in favour of the British. Despite the criticism of the switching to London, which had been targeted only 36 times and many of these too late to matter for an invasion attempt, perhaps criticism should be levied instead at the continued attacking of Channel targets, which were targeted 47 times early in the campaign. Of course, more important still is the lack of initiative amongst the German leadership: an early campaign (CF2, Figure 3.6) yields an almost 3σ effect in favour of the Germans.

Historically speaking, these problems naturally emerge from the Germans' position. After the fall of France in 1940, the Germans were surprised to find themselves fighting to subjugate the French at the same time as forcing the British to capitulate. With their attention split and their resources diminished, it is unsurprising that the Germans had difficulty focusing on one enemy. The French were particularly demanding: the Germans were initially repelled and required the employment of "Blitzkrieg" tactics which in turn divided the Luftwaffe's attention (Alexander, 2007). At the same time, the Luftwaffe had lost approximately half of its forces, dividing its strength even further (Bungay, 2000). With German attention so split and forces so weakened, Macksey (1987) notes, "The basic reason for Germany's failure to invade Britain in 1940 is... the lack of any preconceived will or intention to do so."

While we have quantified the effects of each of our counterfactuals, we cannot quantify how likely the counterfactual scenarios themselves would have been. On the other hand, we can qualify the scenarios' chances. In each counterfactual, we have notably given much more information and much more clarity to German leadership. CF2 requires a Hitler who is not just aggressively enthusiastic for the defeat of Britain, but who also dedicates significant resources to this battle even as the battle with France is ongoing despite his anxieties. Similarly, CF1 and CF4 require a Goering with a great deal more insight as to the effects that Luftwaffe attacks were having on Britain. Such a Goering must know that the British had not already moved north of the Thames (Dempster and Wood, 1961, p. 212). Furthermore, Goering must realise that assaults on airbases are extremely effective for defeating the RAF in comparison to attacks on London, which might theoretically yield political victories, or attacks on the Channel, coast, and shipping, which theoretically support invasion. It requires a highly knowledgeable Goering, who has ready access to knowledge of what the RAF is doing and how it responds to Luftwaffe attacks, despite factually having not prepared at all to discover such information. Merely realising that destroying fighters on the ground and the supporting infrastructure is advantageous is already counterfactual: we have already noted that staff officer Paul Deichmann believed the Luftwaffe should not destroy the radar stations that brought the Luftwaffe's targets before them.

Given the details of what our counterfactuals entail, it is not perhaps so surprising that such great turnarounds can be observed (CF3 and CF5). While it is certainly possible for the Germans to have won the Battle of Britain, to do so would have required far greater resources, knowledge, and strategic acumen than was available to the Germans. Future research might benefit from trying to find a crossover between losing and winning the Battle of Britain as a function of the counterfactual advantages given to the Germans. Presumably, this would start with far less radical counterfactual advantages and progress to more radical ones to explore the crossover. The British had acquired each of these with their production of fighters, their information system, and Park's usage of his fighters. While the Germans could have won if they had fought the battle more proficiently, our results support standard historiographical views: the Battle of Britain as actually fought was not won on a narrow margin.

3.8 Conclusion

Beyond the exploration of the Battle of Britain, this chapter also focuses on applications of the bootstrap to historical modelling. The bootstrap naturally has some constraints that are antithetical to historical analysis. It ignores any possibility of developing tactical responses, improving training, or improving technology due to either side's changed position, as the non-parametric bootstrap can only draw from the battle as actually fought. It also requires a large amount of stationary data, a rarity in history. On the other hand, the (weighted) bootstrap provides a powerful tool for exploring counterfactual realities in an open and transparent way. Both the method of the alteration and its resultant effects can be presented in intuitive ways to better understand a range of factual and counterfactual outcomes.

Future work can proceed in a variety of directions. Standard bootstrap techniques, such as the bootstrap-t method, can be profitably used to explore a historical setting using more computing power. Implementing the BC$_a$ method as a standard tool for exploring data with a time dependent statistic would also be useful for historical analysis. Moving away from the non-parametric setting, integrating historical models with bootstraps could yield interesting results to, for example, examine the effects of tactical changes in real time models. A simpler but more widely applicable direction would be to investigate appropriate copula for various historical environments to make better usage of the causal bootstrapping technique of Imbens and Menzel (2018). On the other hand, one could also seek to identify other historical contexts in which the bootstrap could be applied, such as the Battle of the Atlantic.

Amongst the chapters of this thesis, the bootstrap is perhaps unique in its simplicity, which greatly facilitates applying it to a historical context, as we have done with the Battle of Britain. While the bootstrap is useful for its ability to address counterfactual claims, other methods are more useful when examining different properties of a dataset. As we shall see in the next chapter, there is a great deal of debate as to how conflicts have changed over time, let alone as to by what measure a "unit" of conflict even should be measured. Instead of looking towards counterfactual analysis, we instead must begin by looking for whether there even are changes across these conflicts, to which we turn to changepoint analysis.

4 Changepoint analysis of historical battle deaths

> *The occurrence of two world wars in the present century is apt to leave us with the vague belief that the world has become more warlike. But this belief needs logical scrutiny. A long future may perhaps be coming without a third world war in it.* – *Richardson, 1960b*

4.1 History of conflict

Conflict is a central aspect of human nature, whether discussing disputes of opinion, murder, or war. Humans have been fighting with one another for the entirety of the history of our species, whether with sticks and stones or guns and grenades (Keeley, 1996). What is not as clear is whether violence is an invariant quality of human nature and if our ability and tendency to cause harm to each other, especially on massive scales, has similarly stayed the same. Some maintain that humans have always been violent and that such superficial structures as the League of Nations did not stop or change our nature (Huntington, 1989; Gray, 2012). This is the subject of great debate: others such as Gat (2013) maintain that, despite our violent heritage, civilisation, with its states and super-states, has managed to curb our tendency towards conflict. Others still point to modern developments like bipolarity or the United Nations as evidence of mankind pulling itself out of its violent past, such as Gaddis (1986) or Goldstein (2011). This sort of question has been brought to the public by popular books, particularly those of Steven Pinker (Pinker, 2011, 2018). Pinker contributes a variety of quantitative analysis, including statistics of commerce and conflicts, and qualitative analysis, including discussions of democracy and rationalism, to conclude that the likelihood of violent death has greatly declined over the centuries at the very least at the level of the individual.

Pinker's contribution, due to its public nature, has been particularly controversial. For example, Michael Spagat and Stephen Pinker engaged in a series of letters with Pasquale Cirillo and Nassim Nicholas Taleb in *Significance*, the joint US/UK magazine for professional statisticians, on the merits of the theoretical decline in conflict (Spagat, 2015; Cirillo and Taleb, 2016a; Spagat and Pinker, 2016). The statistical discussion is far older, however. Quaker pacifist Lewis Fry Richardson, FRS contributed formative work in this arena (Richardson, 1944, 1946, 1952). He collected one of the first data sets on what he called “deadly quarrels”, events with violent deaths. Using power-law distributions, for which we write their density’s exponent parameter as α, he found that deaths in deadly quarrels are well described by two such distributions, with powers of approximately 2.4 for murders up to events with about 1,000 dead, and 1.5 for events of more than 1,000 dead (‘wars’) (Richardson, 1960b, Figure 4). His motivation stemmed from the recent World Wars and whether they were evidence of some shift in human violence. He concluded 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, 1960b, p. 141). The more recent controversy since Richardson’s time is due to the post-World War II “long peace” (Gaddis, 1986; Pinker, 2011), the idea that recent time has been relatively peaceful and devoid of major conflict, such as what one might have seen had the United States and Union of Soviet Socialist Republics gone to war. This too is controversial: it has been argued that the early 20th century was unusual instead due to its “great violence” (Clauset, 2018, p. 4) or ‘hemoclysm’ (Pinker, 2011, p. 229, attributed to Matthew White).

Recent statistical discussions have centred on whether there has been a decline and how to test for the differences supposed of recent human history with different approaches reaching different conclusions. Regarding whether there had been a decline, Sarkees *et al.* (2003) use the publicly available Correlates of War dataset, which we also use, and found a ‘disquieting constancy in warfare’ when studying both inter-state and intra-state (civil) wars. Lacina *et al.* (2006) dispute this claim, finding instead that battle deaths ‘declined significantly after World War II and again after the end of the Cold War’ when examining the Peace Research Institute Oslo (PRIO) Battle Deaths, based on the UCDP/PRIO Armed Conflict data set (Gleditsch *et al.*, 2002). This debate has continued between differing groups of authors (Lacina and Gleditsch, 2013; Gohdes and Price, 2013; Harrison and Wolf, 2012; Gleditsch and Pickering, 2014; Harrison and Wolf, 2014).

Regarding how to test for differences in recent human history, techniques from extreme value theory, when applied to an unpublished dataset of deaths due to war from 60 CE until 2015 CE, failed to reject a null hypothesis of no change in arrival time or distribution (Cirillo and Taleb, 2016b). More standard statistical techniques using bootstrap testing, applied to the Correlates of War dataset, have similarly failed to reject a no change null hypothesis (Clauset, 2018). On the other hand, Spagat and Pinker identify several qualitative changes in line with Pinker (2011) that indicate possible changes after 1945 and 1989. Additionally, a null hypothesis of no change in the magnitudes of large wars was rejected by Spagat and van Weezel (2018) for some definitions of large wars within the Gleditsch data set, also used herein. A parametric single changepoint analysis with a variety of models was conducted by Cunen *et al.* (published in parallel with our work in 2020, see also Hjort (2018)) using the aforementioned Correlates of War dataset and identified 1965 as the most likely changepoint for a change in the magnitudes of wars.

We apply a full nonparametric changepoint analysis on reputable and freely available historical battle death data sets. To do so requires adopting modern methods to cope with the well-known presence of heavy-tails in the battle deaths data sets. The long-standing consensus is that battle deaths data are power-law distributed (Richardson, 1944, 1960b; Clauset *et al.*, 2009b; González-Val, 2016; Chatterjee and Chakrabarti, 2017; Clauset, 2018; Martelloni *et al.*, 2018). We demonstrate recent methods by Killick *et al.* (2012) and Haynes *et al.* (2017a,b), discussed in Section 2.3, are up to the task using simulation studies that mimic historical data. We then use an algorithm synthesised from these methods to examine the case for changepoints in the distribution of the magnitudes of wars using only the data. We do not make use of an informed prior, such as that World War II is a natural changepoint. One might argue that this ignores modern data, beliefs, and the value of hindsight, but we would rather our approach be purely data-driven. The algorithm is a neutral observer, looking for features that seem obvious as well as those that do not using the strength of the signal in the data alone. This procedure has repercussions on the debate on whether conflict has changed and if the world has become more peaceful: if a changepoint near World War II was detected following which we observed less intense deadly events, this would support the long peace hypothesis.

We begin by discussing the datasets we intend to use in Section 4.2. We establish and conduct our simulation studies in Section 4.3. This allows us to establish Algorithm 1, which we then apply in Section 4.4. We end with a discussion of our results and future directions in Section 4.5 and 4.6.

4.2 Battle deaths datasets

Since the pioneering work of Richardson there have been many attempts to create datasets quantifying violence. The construction of these datasets raises a number of important questions, first of definition and then also of incomplete or biased knowledge. Richardson (1960b, p. xxxvi, 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 sub-types of deadly quarrels, such as battle deaths above a set threshold, as in the Correlates of War datasets (Sarkees and Wayman, 2010), or terrorism, as in the Global Terrorism Database (National Consortium for the Study of Terrorism and Responses to Terrorism (START), 2016, p. 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 a number of 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 CE 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 does one 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) indicates 5 million military deaths with 10 million total (in comparison to 300,000 due to simultaneous rebellions) using data due to Eckhardt. Worden *et al.* (1988) report that 30 million were reported killed over 14 years. Platt (2012) reports in the epilogue 70 million dead, along with the standard 20 – 30 million figure and criticisms of both of these numbers. Deng (2003) indicates similar numbers from Chinese sources, but notes their interrelation with famine. However, the Correlates of War dataset reports only 26,000 (Chinese), 85,000 (Taipings) and 25 (U.K.) battle deaths – albeit only for the second, inter-state phase of the war. Battle deaths for the initial phase are listed as unknown. The Gleditsch dataset is consistent with the Correlates of War values. Particular difficulty arises where there is disagreement between contemporary (or even political descendants of) participants, and especially where one or 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 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 datasets that handle the data consistently and use clearly stated assumptions on data gathering and aggregation.

We acknowledge that none of the available datasets is ideal, as each has varying criteria for inclusion of events; and indeed the available historical data themselves are not ideal, due to, for instance, biases in the record. The two datasets we use are the *Correlates of War* (Sarkees and Wayman, 2010, first version published in 1982, hereafter, *CoW*) and a dataset due to Gleditsch (2004, hereafter the *Gleditsch dataset*). The former has been used by, for example, Clauset (2018) and Hjort (2018), while the latter has been used by Spagat and van Weezel (2018). We note that the Gleditsch dataset was originally based upon the CoW dataset, although divergent evolution has occurred since. The CoW dataset has four different subsets (inter-state, intra-state, extra-state and non-state), whereas the Gleditsch dataset identifies civil and inter-state wars. For both datasets, each data point is a war, including its combatants (nation-states or other organisations) 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 datasets 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 datasets 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 autocorrelation (Beard, 2018). Intuitively, if battle deaths are assigned per-combatant to the date a combatant joins the war, then autocorrelation 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 dataset using better-resolved events – battles, perhaps – and perform a changepoint analysis there.

In Figure 4.1, we show the CoW dataset, on the left, and the Gleditsch dataset, right, 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 do not include them in the analysis.

A controversial question is whether we should consider the absolute number of deaths caused in a conflict or the number relative to global population, see *e.g.* the work of Spagat and van

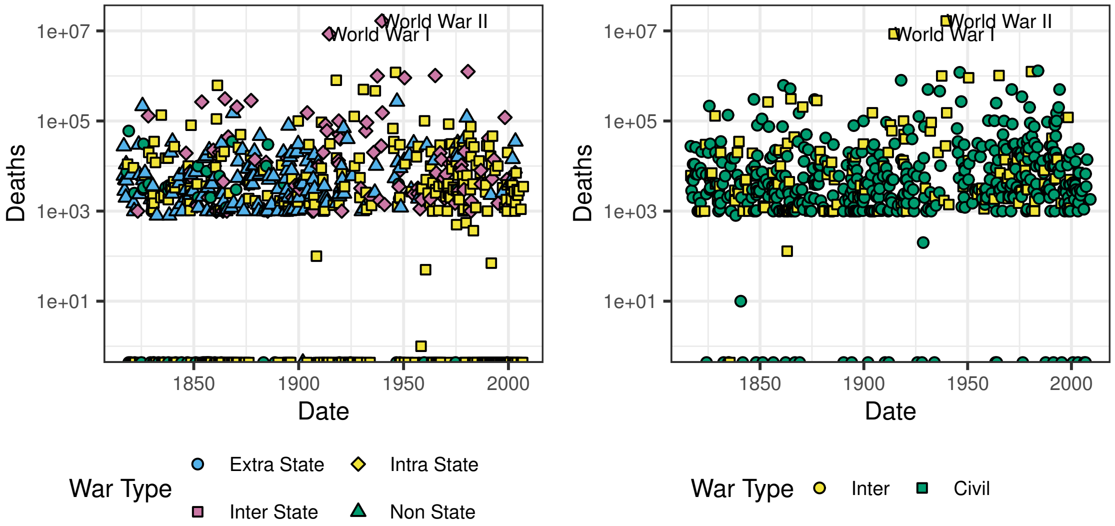


Figure 4.1: Datasets on logarithmic axis. Left: CoW dataset. Right: Gleditsch dataset. The World Wars are labelled for reference. Colours indicate the different subsets defined for each dataset and are indicated in the respective legends.

Weezel (2018). There are good arguments for each choice reflecting two important questions about human value. The relative number, 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. On the other hand, each unit within the data is a human life so we must acknowledge the criticisms of Epstein (2011) and Cirillo and Taleb (2016b, p. 16) that one 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 using the HYDE 3.2.1 dataset (Klein Goldewijk *et al.*, 2017). Of course any changepoints in the (normalising) population will interfere with changepoint detection in the battle deaths dataset, and the two analyses need to be considered separately.

4.3 Simulation study

This section performs an in-depth exploration of the performance of existing segmentation methods for simulated data specifically chosen to mimic the properties 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, 1960b; Clauset *et al.*, 2009b).

The wide pool of candidate methods is first narrowed down and thorough testing leads us to propose a changepoint detection algorithm (Algorithm 1) suitable for our context. Furthermore, we carry out a simulation study to investigate the effects of data aggregation on the identified changepoints and perform a sensitivity analysis in order to assess the effects of data normalisation.

In order 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 changepoint sets (Truong *et al.*, 2018). The Rand index measures (cluster) accuracy by comparing the relationships of data in each cluster in the discovered changepoint set to the true (Truong *et al.*, 2018). We use the adjusted Rand index, implemented in `mclust`, to account for clustering due to chance

(Scrucca *et al.*, 2017). Total agreement between clusters results in an ARI of 1, while the expected value of a random partition of the set is 0. Finally, the true detection rate gives us an understanding of how many changepoints detected are true or false by checking to see if a true changepoint happened near a detected one (Haynes *et al.*, 2017b). A TDR of 1 indicates that every changepoint detected is within a given distance of at least one true changepoint, while a TDR of 0 indicates that every changepoint is outside such a distance. First, for direct comparison, we consider a radius of acceptance of 0, *i.e.* exactly identifying the changepoints where a positive radius h would report a detected changepoint as accepted if it within h data points of a true changepoint (Haynes *et al.*, 2017b). In order to choose appropriate further radii, we consider the historical context – for example, World War I might easily have begun a year or two earlier due to 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, 3, 5, and 8 new wars will have occurred within 1, 2, or 3 years respectively. Hence, we use radii of 3, 5, and 8 to roughly represent 1, 2, or 3 years in the historical dataset. We also do not include the endpoints of the data as changepoints for this calculation.

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

All simulation tests were carried out in R (R Core Team, 2018). In particular, data generation was performed using the `poweRlaw` R package (Gillespie, 2017), while changepoint analyses were carried out using the `changepoint` (Killick *et al.*, 2016) and `changepoint.np` (Haynes *et al.*, 2016) R packages. As the name suggests, the extension `⋆.np` in the package name and associated function stands for the nonparametric approach of Haynes *et al.* (2017b). Visuals were compiled using the `ggplot2` R package (Wickham, 2016).

To benchmark the various candidate methods, we first screened the possible combinations of cost and penalty corresponding to different data modelling distributions. Table 4.1 summarises the available functions and options, as implemented in the changepoint packages above, while noting restrictions on combinations of methods. Some of the arguments provided require additional information which we set to be the same across all tests. Specifically: the type I error probability 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 changepoints required by BinSeg (Scott and Knott, 1974; Sen and Srivastava, 1975) was set to 60.

We assessed segmentation outcomes across $N = 1{,}000$ trials with data of length $n = 600$ featuring a single changepoint ($m = 1$) located at $\tau_1 = 300$. The first segment consisted of data simulated from a power-law distribution with parameter $\alpha = 2.05$, while for the second segment we chose $\alpha = 2.55$ (in the range of powers akin to those documented for historical battle deaths).

Table 4.1: Function options for `changepoint` and `changepoint.np` R packages. The first column corresponds to the R function used, while the other three correspond to arguments that determine how the analysis is performed. Note that not every combination of options within a function are 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. Additional descriptions of the function options are available in Table C.1.

Function	`penalty`	`method`	`test.stat`

`cpt.mean`
`cpt.var`
`cpt.meanvar`
	SIC/BIC
	mBIC
	AIC
	Hannan-Quinn
	Asymptotic
	CROPS
		AMOC
		PELT
		SegNeigh
		BinSeg
			Normal
			CUSUM (`cpt.mean` only)
			CSS (`cpt.var` only)
			Exponential (`cpt.meanvar` only)
			Poisson (`cpt.meanvar` only)

`cpt.np`	SIC/BIC mBIC AIC Hannan-Quinn CROPS	PELT	Empirical Distribution

Across our simulations we set the minimum value attainable by the power-law distribution to be 10 or $1{,}000$. We present the results in the former case; the latter case is nearly equivalent and its results appear in Appendix D. We note at this time that power-law distributions with $\alpha > 2$ have finite means, but $\alpha < 3$ leaves us with infinite higher moments. As such, we expect occasional very large events, similar to how the World Wars are considered very large relative to other wars.

Figures 4.2 – 4.5 give illustrative examples of the types of behaviour of the analyses conducted. The bottom subplot of each plot indicates the percentage of trials in which a given number of changepoints was detected by the analysis. The top subplots are arranged by the number of changepoints found and use boxplots to show the location of each changepoint so found. The middle dashed line is placed along the changepoint. Across the tested combinations, most failed to identify that there was only a single changepoint, let alone to pinpoint its precise location.

We also note that Figures 4.2 – 4.5 do not showcase all possible outcomes. For example, some combinations result in approximately correct numbers of changepoints 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 changepoints, overfitting, and placement in the middle of the $\alpha = 2.05$ segment. Of the `changepoint` methods, 'at most one changepoint' (henceforth, AMOC, Killick *et al.*, 2016) was naturally most successful. It was tied with itself for second lowest median Hausdorff metric 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. Based on 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. That is, if there is a changepoint detected by ED-PELT with CROPS or

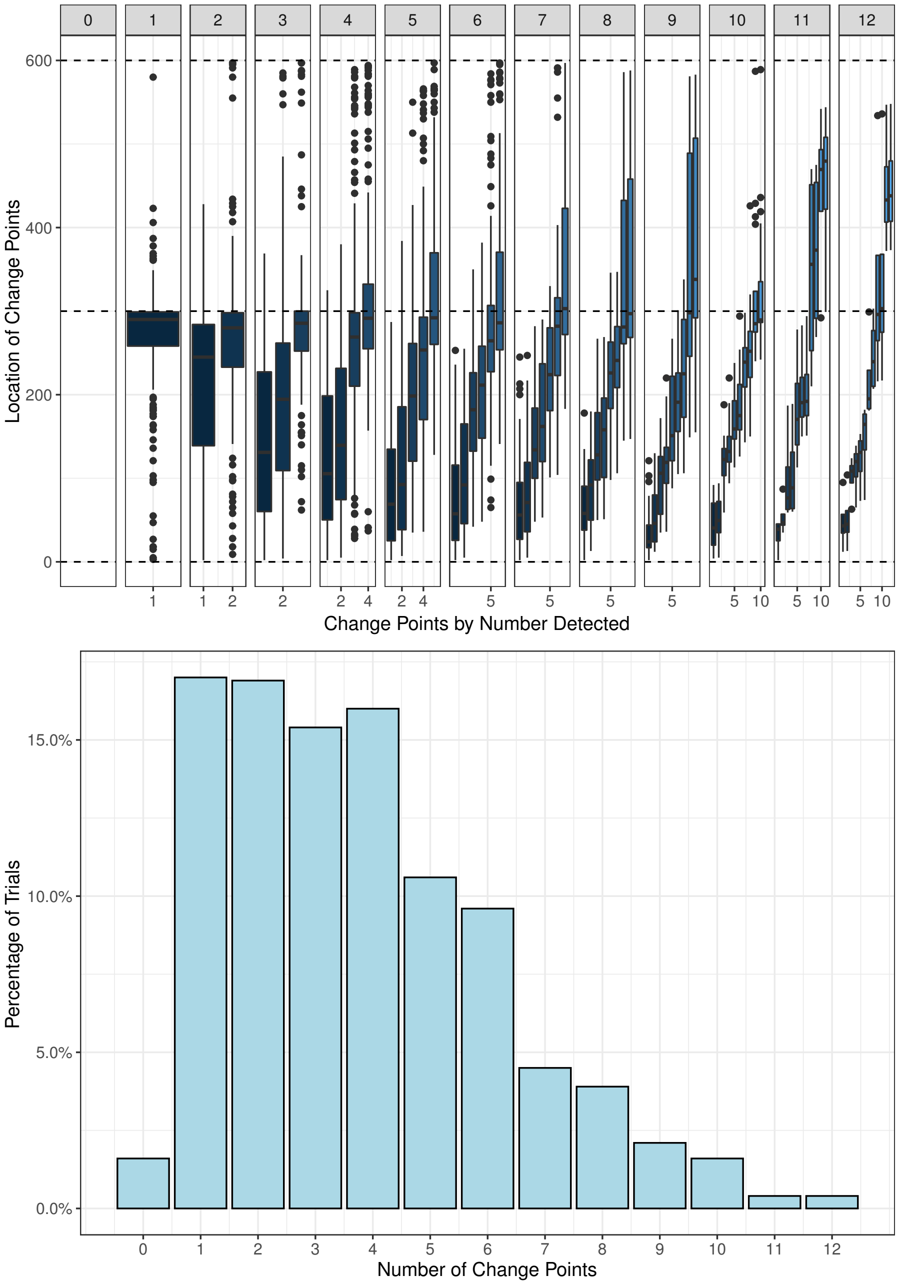


Figure 4.2: Test case with better than average behaviour. The simulation is of two power-law distributed segments of length 300 with exponents 2.05 and 2.55 respectively. Segmentation generated using `cpt.meanvar` with BinSeg, mBIC and an exponential distribution. Whilst there are good aspects to this finding, the method commonly overfits and tends to assume changepoints happen in the $\alpha = 2.05$ segment. This combination has median Hausdorff metric of 189, median ARI of 0.64, and TDR 0: 0.01, 3: 0.04, 5: 0.05, 8: 0.07.

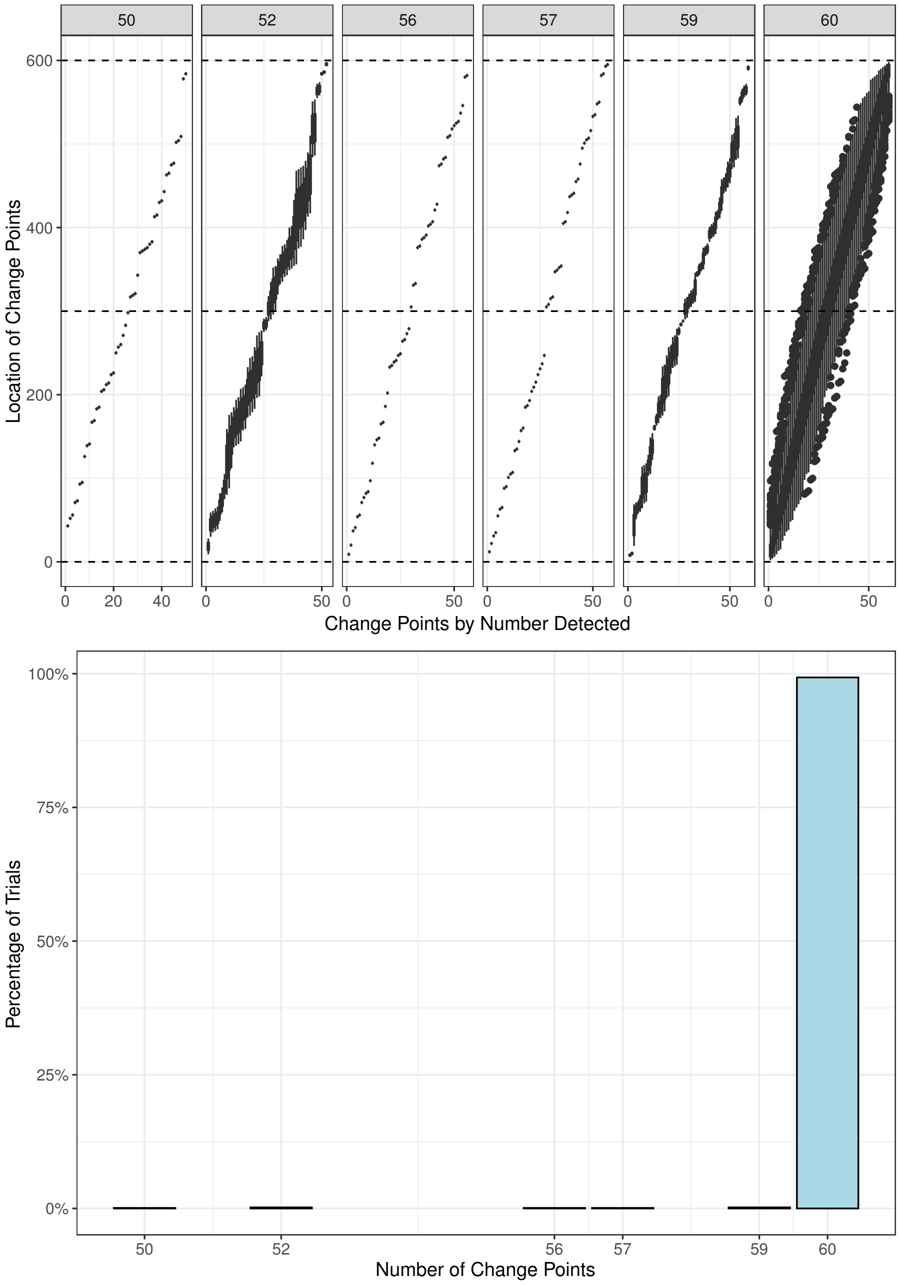


Figure 4.3: Test case with worst behaviour. The simulation is as in Figure 4.2. Segmentation generated using `cpt.meanvar` with SegNeigh, an asymptotic penalty and a normal distribution. Results such as this occur with many combinations, and can be regarded as failures. Many combinations result in more than 10 false positives and are only stopped by the maximums provided. This combination has median Hausdorff metric of 294, median ARI of 0.07, and TDR 0: 0.00, 3: 0.01, 5: 0.01, 8: 0.01.

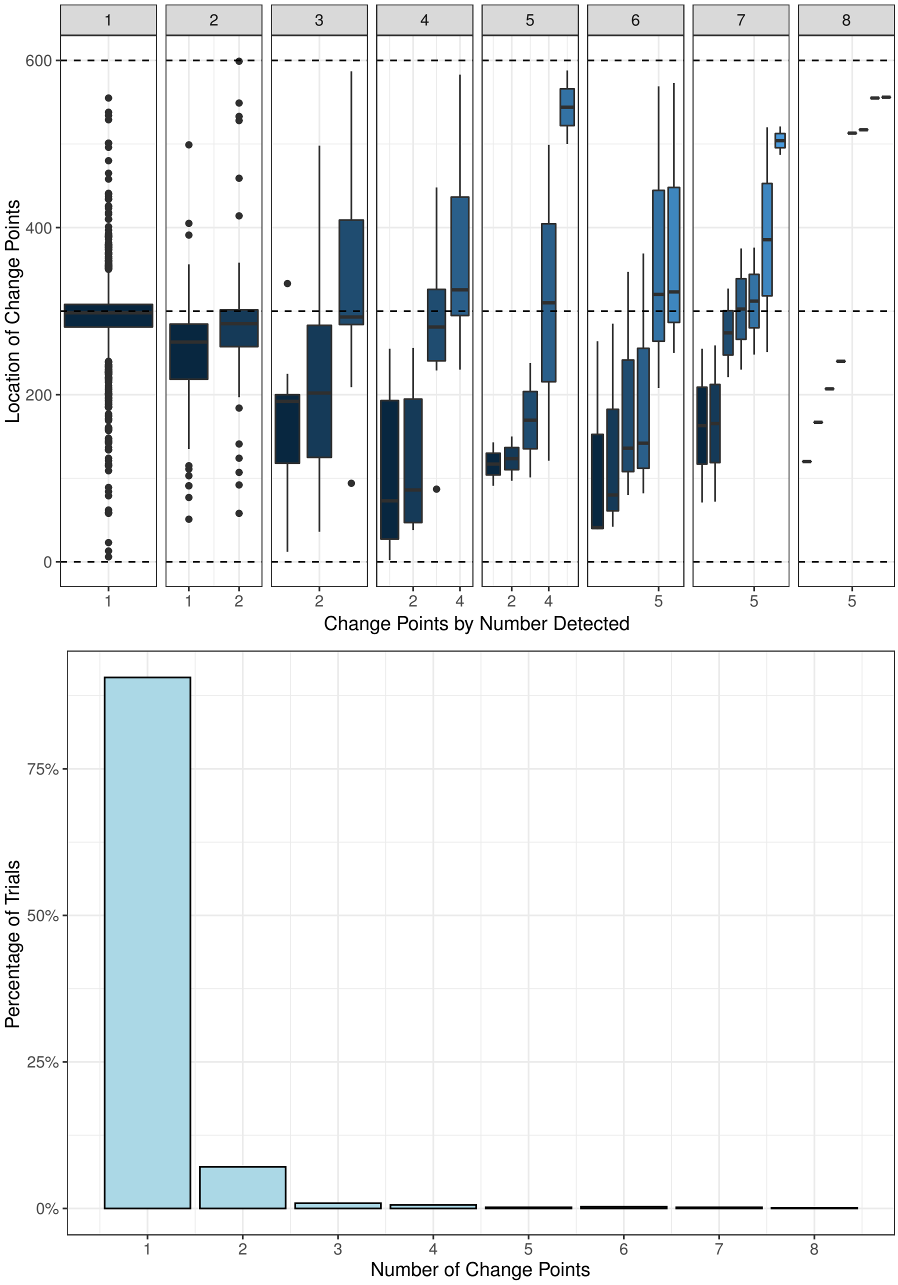


Figure 4.4: Test case with best behaviour. The simulation is as in Figure 4.2. Segmentation generated using `cpt.np` with ED-PELT and CROPS and shows some of the best achievable behaviour. Although qualitatively similar to the top sub-plot of Figure 4.2, there is improved accuracy in the positioning of the changepoints and improved precision and accuracy in the number of points so detected. This combination has the lowest median Hausdorff metric of 15.50, highest median ARI of 0.90, and highest TDR 0: 0.03, 3: 0.17, 5: 0.24, 8: 0.32.

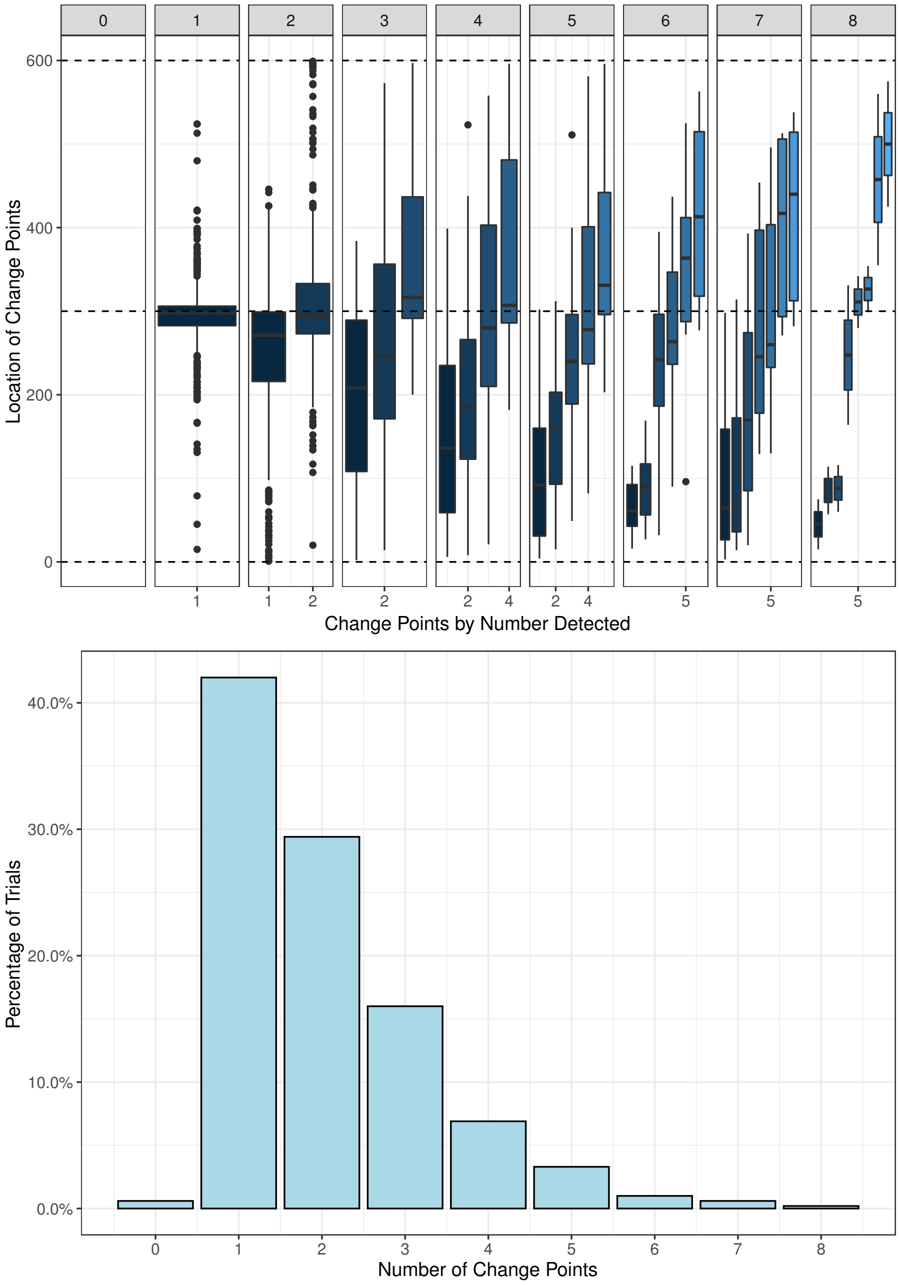


Figure 4.5: Test case with second best behaviour. The simulation is as in Figure 4.2. Segmentation generated using `cpt.np` with ED-PELT and mBIC. While not as good at detecting changepoints as CROPS, `cpt.np` with ED-PELT and mBIC still shows strong potential. This combination has median Hausdorff metric of 50.5, second highest median ARI of 0.79, and TDR 0: 0.02, 3: 0.09, 5: 0.13, 8: 0.17.

Table 4.2: Test set parameters. Each column represents a parameter and its options for the simulated data in Section 4.3. Each test was performed with $N = 1000$ trials with $m = 1$ changepoint(s) located 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.

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		

mBIC, then there is statistically significant evidence that the segment before the changepoint has a different power-law exponent than the segment after the changepoint (Haynes *et al.*, 2017b).

We next examine the performance of our remaining candidates in the presence of at most one changepoint. In order for our explorations to be relevant to the real battle deaths data, we choose power-law exponents (α) close in value to Richardson's law, and test the segmentation robustness against numerical proximity, order and false positive detection, as detailed in Table 4.2. Some of our tests have one or both $\alpha < 2$, which will result in more extreme tail and sampling behaviour. 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 number of changepoints detected and increased ARI (in contrast to the other penalty options, which claim more changepoints occur as segment lengths increase). 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 $\alpha_{\text{mod}} \leq 0.05$, or coincide (no changepoint). When this occurs, CROPS has a tendency to dramatically overfit the number of changepoints whereas mBIC is more likely to correctly report no changepoints.

Figure 4.6 shows examples where there is no changepoint in the data; mBIC can detect this reasonably well, but CROPS dramatically overfits. This unique failure mechanism of dramatic overfitting occurs often: 3 or more changepoints 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. On the other hand, mBIC correctly detects 0 changepoints 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 changepoints. Subsequent results should be viewed through the lens of these limitations.

We now expand our investigations beyond the presence of at most one changepoint and explore the outcomes obtained when the data feature several (specifically, two, four, or eight) changepoints controlled for variable segment length and data granularity. Figure 4.7 shows some representative results of each procedure.

In general, our previous findings extend to the case of multiple changepoints, as can be seen in the first row of Figure 4.7, where CROPS proves to be more precise and accurate in its identification of changepoints (higher median ARI and TDR, lower Hausdorff metric distance), although sometimes too conservative. The second row of Figure 4.7 illustrates an uncommon case in which the change across one particular changepoint is so drastic that CROPS identifies it as the only change, missing the less pronounced changes. In contrast, mBIC mostly successfully identifies

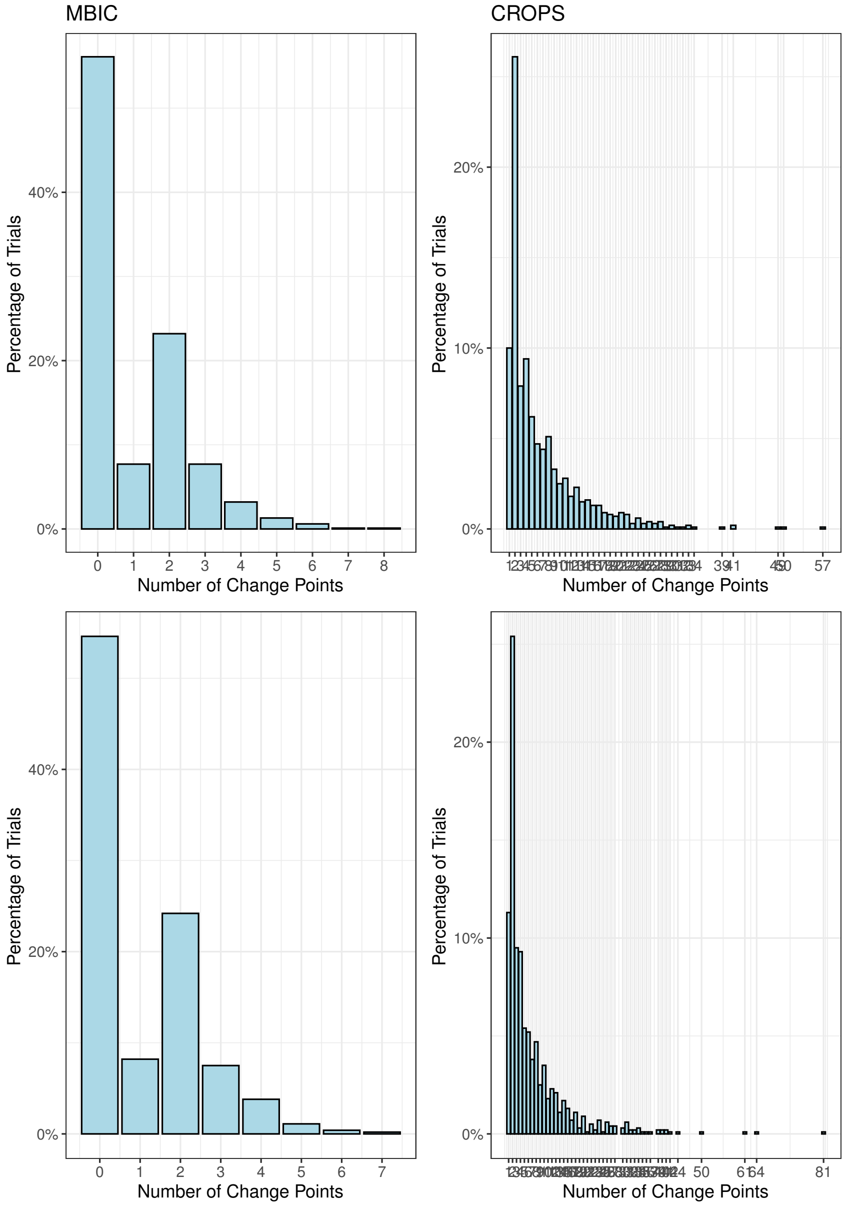


Figure 4.6: Examples comparing behaviour of CROPS and mBIC with no changepoints present. Each row is a different scenario using different power-law exponents: 1.7 and 2.3 respectively. The sequence length is 600. Note that CROPS has pathological behaviour, while mBIC succeeds with reasonable precision and accuracy. The behaviour of CROPS is due to a known feature: Haynes *et al.* (2017b) recommend choosing the optimal number of changepoints for CROPS such that it maximises the estimated curvature of the penalty as a function of the number of changepoints. This naturally truncates the data over which the curvature is estimated, removing the possibility of obtaining 0 changepoints on a potentially flat line.

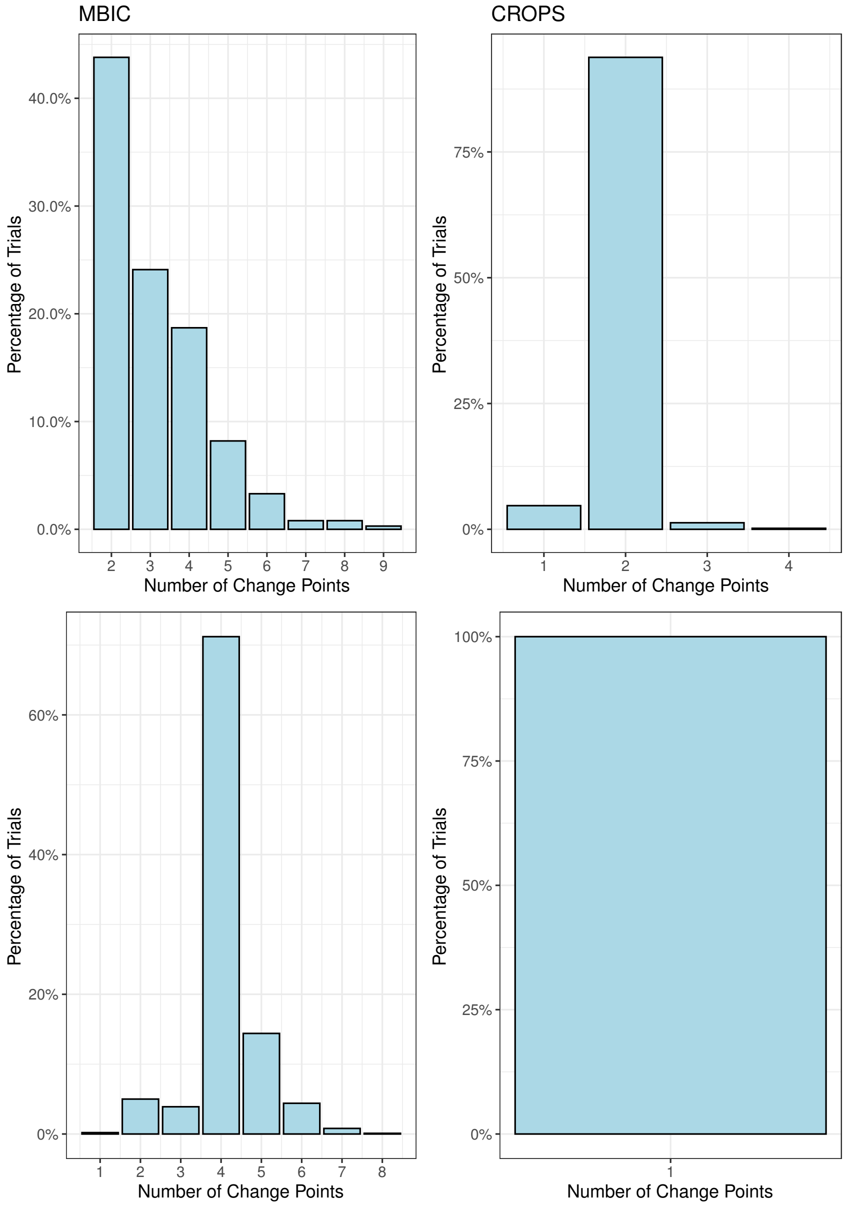


Figure 4.7: Examples comparing behaviour of CROPS and mBIC with multiple changepoints. Each row is a different scenario using power-law distributions. The sequence length is $n = 1000$ and 575 respectively. In the first row, two changepoints, marking the power-law exponent change from 2.3 to 1.7 to 2.1, are present, and CROPS gives a more accurate result. Simulations show this is the common pattern. In the second case four changepoints are present, transitioning across exponents 2.87, 1.83, 2.49, 1.67, and 1.06. CROPS detects only a single changepoint with high precision. mBIC outperforms CROPS in this uncommon case.

these changepoints, showcased in higher ARI and lower Hausdorff metric distances, albeit with a lower TDR. This uncommon case is more likely to occur when there are a large number of true changepoints (*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 changepoints. The combined use of the two methods gives good confidence in accurately detecting the correct number of changepoints, as well as their location. When using both methods in tandem, we preserve the ability of mBIC to identify systems consistent with no changepoints (58.2% of the time for $\alpha = 1.7$ and 54.8% for $\alpha = 2.3$ when used in conjunction with CROPS). Furthermore, we know from our simulation study that CROPS has the best ability to correctly detect changepoints, with the almost universally best median ARI, TDR, and Hausdorff metric distances. Indeed, as the distance between the exponents increases and as the amount of data increases, CROPS detects increasingly well, with, for example, a 50.7% probability of detecting exactly the changepoint and a 99.0% probability of being with 8 data points of the changepoint for $\alpha = 1.7 \pm 0.5$ and $s = 300$. The combination still has weaknesses that can be improved upon, including improving performance as α increases, but we are confident that, supplemented with additional testing, the combination performs well as a tool for detecting changepoints.

In summary, CROPS findings are accurate when small numbers of changepoints are found, whereas changepoints found only by mBIC should 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 changepoint is present. Where mBIC and CROPS agree on identified changepoints, 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 changepoint detection algorithm (Algorithm 1) to employ on the real battle deaths datasets. This protects against the pathological CROPS case, *i.e.* by setting $\underline{\tau} = \emptyset$ when $|\underline{\tau}_{\text{CROPS}}| > 2$ while $|\underline{\tau}_{\text{mBIC}}| = 0$, resulting in increased TDR when considering the changepoint intersection set, while also allowing for a more liberal interpretation of the union of detected changepoints, such as when CROPS detects only the strongest changepoints and mBIC detects additional weaker changepoints. As stated, Algorithm 1 is not fully specified, leaving part of the procedure to the user. This reflects that the patterns that are found by CROPS and mBIC can be complicated but similar, *e.g.* if CROPS detects a changepoint while mBIC detects one on each side and nearby to the one detected by CROPS. Hence, when presenting results, we will highlight $\underline{\tau}$ as well as $\underline{\tau} \setminus \underline{\tau}_{\text{mBIC}}$ and $\underline{\tau} \setminus \underline{\tau}_{\text{CROPS}}$. When using simple decision metrics in simulation studies, behaviour is similar to our existing results: accepting all results looks similar to using mBIC alone, while accepting only changepoints recommended by both CROPS and mBIC looks similar to using CROPS alone except in anomalous cases. Our algorithm could benefit from future research investigating how one might explicitly build in a tolerance for (dis)agreement in changepoint detection.

As pointed out in Section 4.2, data aggregation has often been debated in the historical literature (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 we use are sensitive to changes in the tail of the distribution (Haynes *et al.*, 2017b, Section 3.1). Furthermore, the beginnings and ends of conflicts can be hard to distinguish – were the ‘great violence’ (Clauset, 2018, p. 4) or the Chinese civil wars of the 1860s each coherent singular conflicts or two sets of discrete wars? Thus for completeness we now empirically address the effects of ag-

Algorithm 1: Proposed changepoint detection algorithm for power-law distributions. (Note $|\cdot|$ denotes cardinality.)

```
Given the time-ordered observations $\underline{y} = \{y_1, \ldots, y_n\}$, segment $\underline{y}$ by applying ED-PELT
with penalty
    1. mBIC; denote the estimated set of changepoints as $\underline{\tau}_{\text{mBIC}}$;
    2. CROPS; denote the estimated set of changepoints as $\underline{\tau}_{\text{CROPS}}$.
    3. Set $\underline{\tau} = \underline{\tau}_{\text{mBIC}} \cap \underline{\tau}_{\text{CROPS}}$ and $m = |\underline{\tau}|$.
    4. For $(\underline{\tau}_{\text{mBIC}} \cup \underline{\tau}_{\text{CROPS}}) \setminus \underline{\tau}$, interpretation is required.
```

gregation of distinct events. In particular, we show that additional, but not an extreme amount of, aggregation induces an accuracy-precision trade-off, in that aggregation increases the probability of correctly identifying a changepoint, but decreases information as to its exact location.

Our procedure is a straightforward extension of our previous simulation studies. We again simulate data with power-law exponents in the range described by historical literature, but once the data have been generated, we then aggregate adjacent pairs or quartets to replicate the idea of binning data into time periods or other systematic aggregations. For example, if our data are $X_1, X_2, X_3, X_4, \ldots$, we would instead use the sequence $X_1+X_2, X_3+X_4, \ldots$ or $X_1+X_2+X_3+X_4, \ldots$ respectively. For brevity, we present one example analysed by CROPS and mBIC with aggregation of pairs ("aggregation 2") and aggregation of quartets ("aggregation 4") in Figures 4.8 – 4.11. These images are directly comparable to Figures 4.4 and 4.5, with which they share the same underlying distribution(s).

One immediate disadvantage to aggregation is that it makes it harder to use our chosen metrics. For example, is an exact changepoint achievable if the true changepoint is in the middle of an aggregation segment? Another example: if the distance in the aggregated data is 6; in the original data this would be outside our chosen range of 8, so should we include it in the TDR with radius 8? For the ARI, should there be some form of penalty if the changepoint occurs in the middle of an aggregation segment? In general, these questions can be summarised as: should we disaggregate (or further aggregate) the data for comparison? We choose not to do so here, and thus rely more heavily on the information from the figures.

Overall, we note a common trend amongst Figures 4.8 – 4.11 by comparison to Figures 4.4 and 4.5. In each case, the probability of identifying a single (true) changepoint has somewhat increased, and other simulations largely yield similar results. However, this increase in the probability of a correct detection is counterbalanced by the decreased amount of information as to the exact position of the changepoint. The former point seems to indicate that the summation of power-law distributed samples exaggerates the differences between the samples. For instance, it might be that, due to the aforementioned problems with power-law samples, the tail behaviour will dominate any sums in which it is relevant. An interesting future direction might be to see if the increased probability of detection is maintained if the aggregation has a random component.

In order to fully assess the impact of data presentation on the discovered changepoints, we also carried out a sensitivity analysis that highlights how changes in the (normalising) population interfere with the identified changes in the battle datasets. First, we present in Figure 4.12 the population and its inverse as recorded by Klein Goldewijk *et al.* (2017). (We only present years that are present in the CoW dataset.) Both plots make it clear that (i) the population is changing

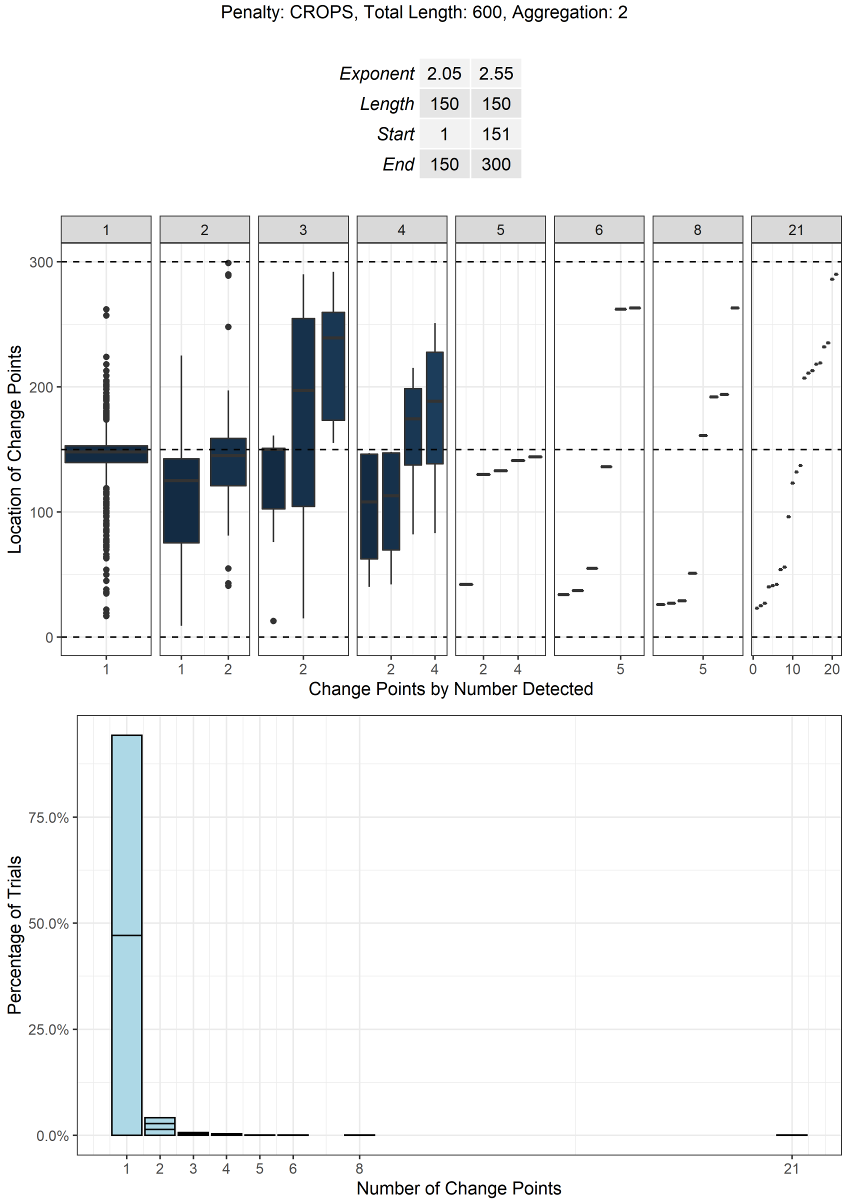
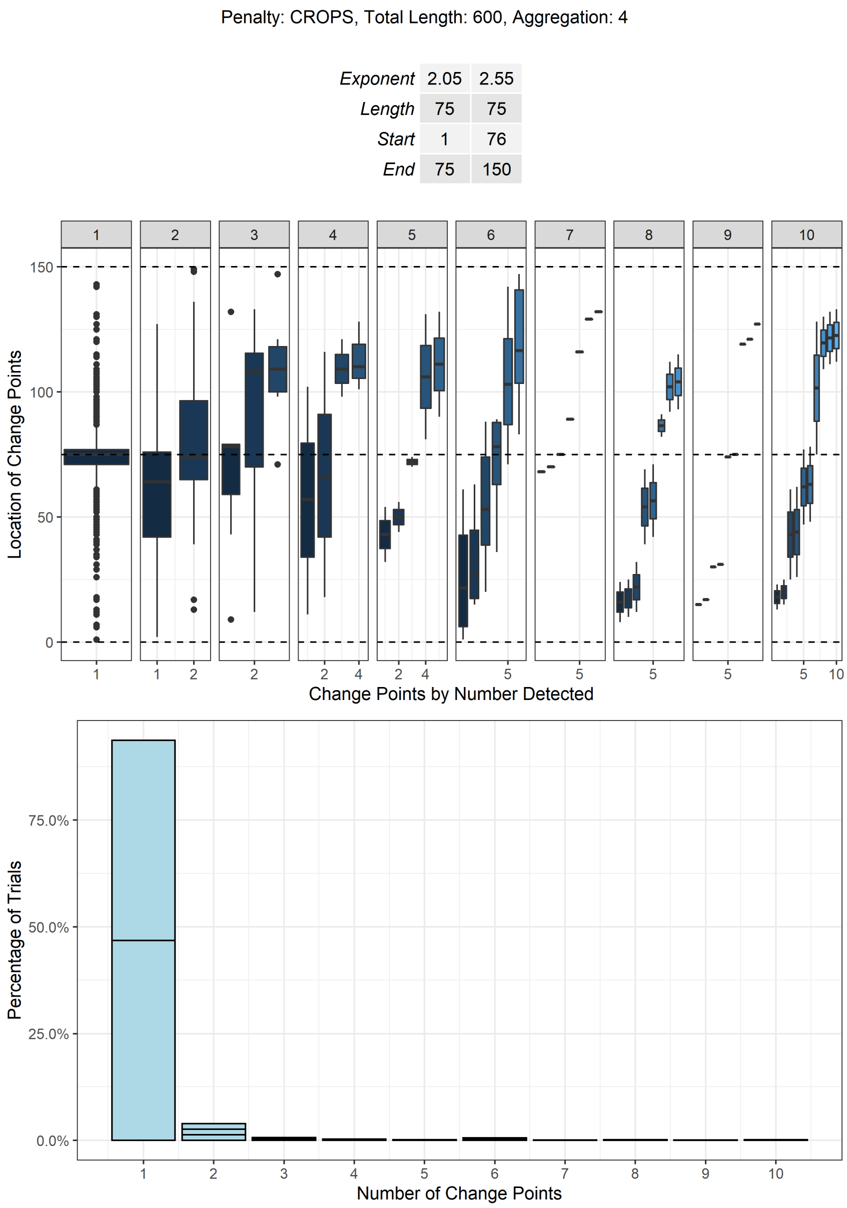
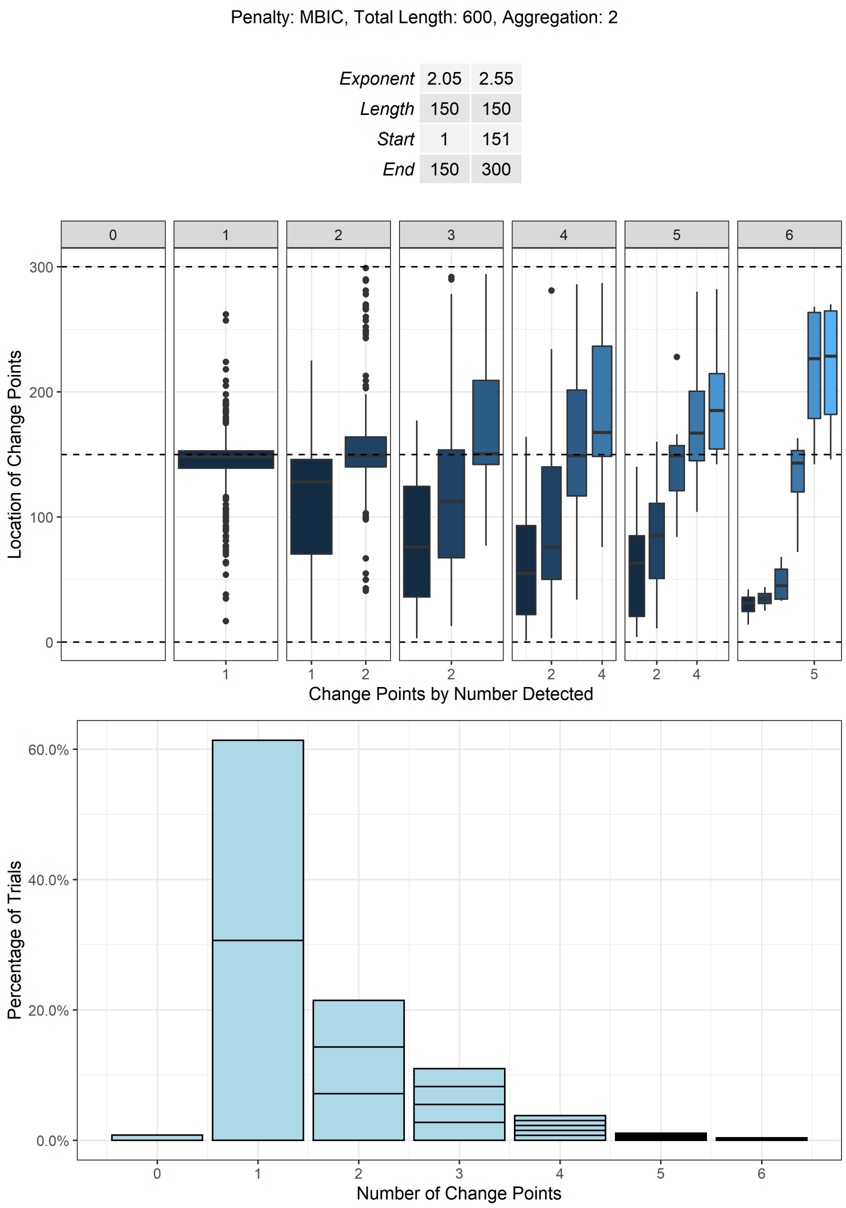


Figure 4.8: Test case with CROPS and “aggregation 2”. The simulation is as in Figure 4.2, but with an aggregation level of 2. Segmentation generated using `cpt.np` with ED-PELT and CROPS. Compare with Figure 4.4 and Figure 4.9. The segments in the lower image correspond to the number of changepoints detected and are guides for the eye only. The probability of correctly identifying that there is only one changepoint in this case is 94.3% compared to 91.9% in Figure 4.4.



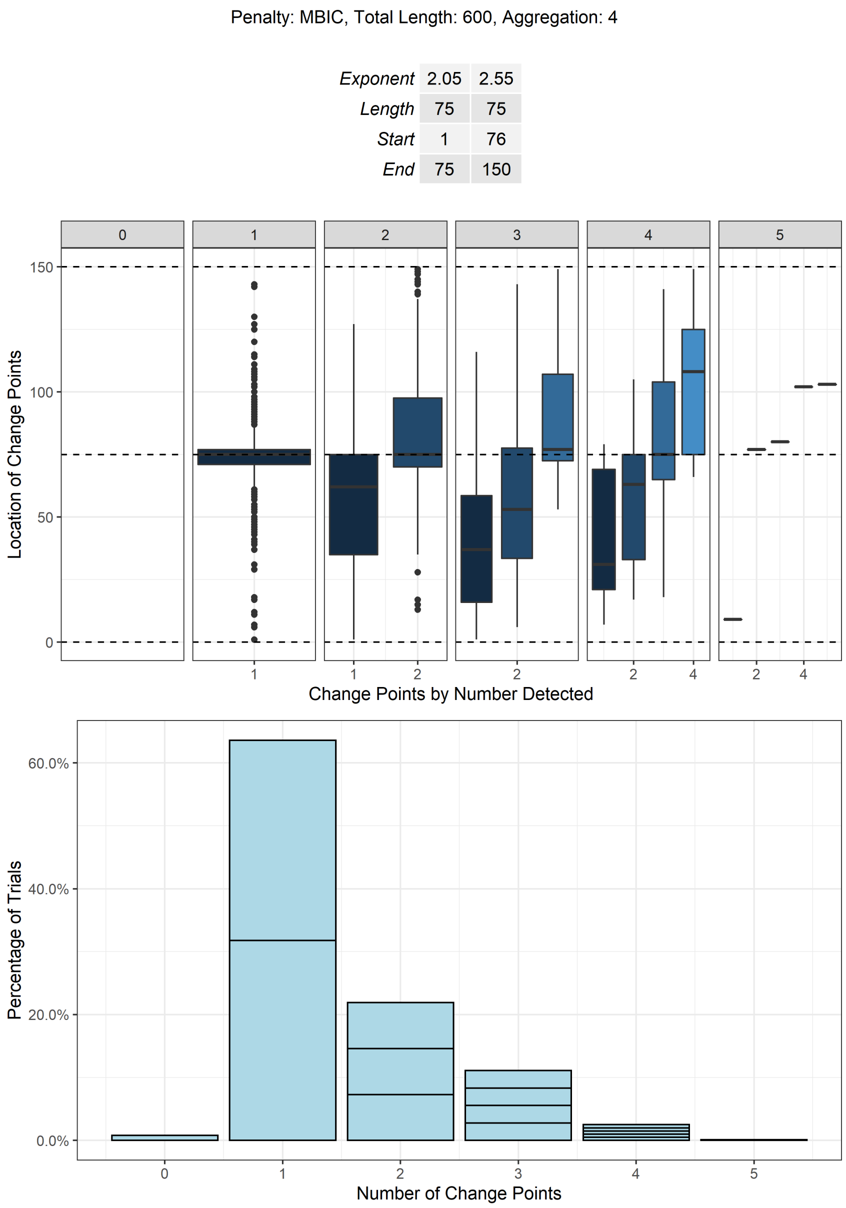
Exponent	2.05	2.55
Length	75	75
Start	1	76
End	75	150

Figure 4.9: Test case with CROPS and “aggregation 4”. The simulation is as in Figure 4.2, but with an aggregation level of 4. Segmentation generated using `cpt.np` with ED-PELT and CROPS. Compare with Figure 4.4 and Figure 4.8. The segments in the lower image correspond to the number of changepoints detected and are guides for the eye only. The probability of correctly identifying that there is only one changepoint in this case is 93.7% compared to 91.9% in Figure 4.4.



Exponent	2.05	2.55
Length	150	150
Start	1	151
End	150	300

Figure 4.10: Test case with mBIC and "aggregation 2". The simulation is as in Figure 4.2, but with an aggregation level of 2. Segmentation generated using `cpt.np` with ED-PELT and mBIC. Compare with Figure 4.5 and Figure 4.11. The segments in the lower image correspond to the number of changepoints detected and are guides for the eye only. The probability of correctly identifying that there is only one changepoint in this case is 61.4% compared to 40.0% in Figure 4.5.



Penalty: MBIC, Total Length: 600, Aggregation: 4

Exponent	2.05	2.55
Length	75	75
Start	1	76
End	75	150

Figure 4.11: Test case with mBIC and “aggregation 4”. The simulation is as in Figure 4.2, but with an aggregation level of 4. Segmentation generated using `cpt.np` with ED-PELT and mBIC. Compare with Figure 4.5 and Figure 4.10. The segments in the lower image correspond to the number of changepoints detected and are guides for the eye only. The probability of correctly identifying that there is only one changepoint in this case is 63.6% compared to 40.0% in Figure 4.5.

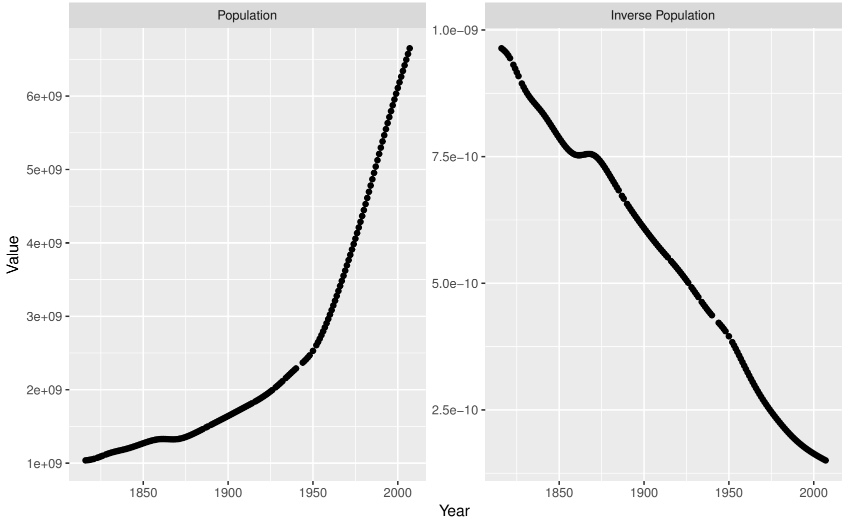


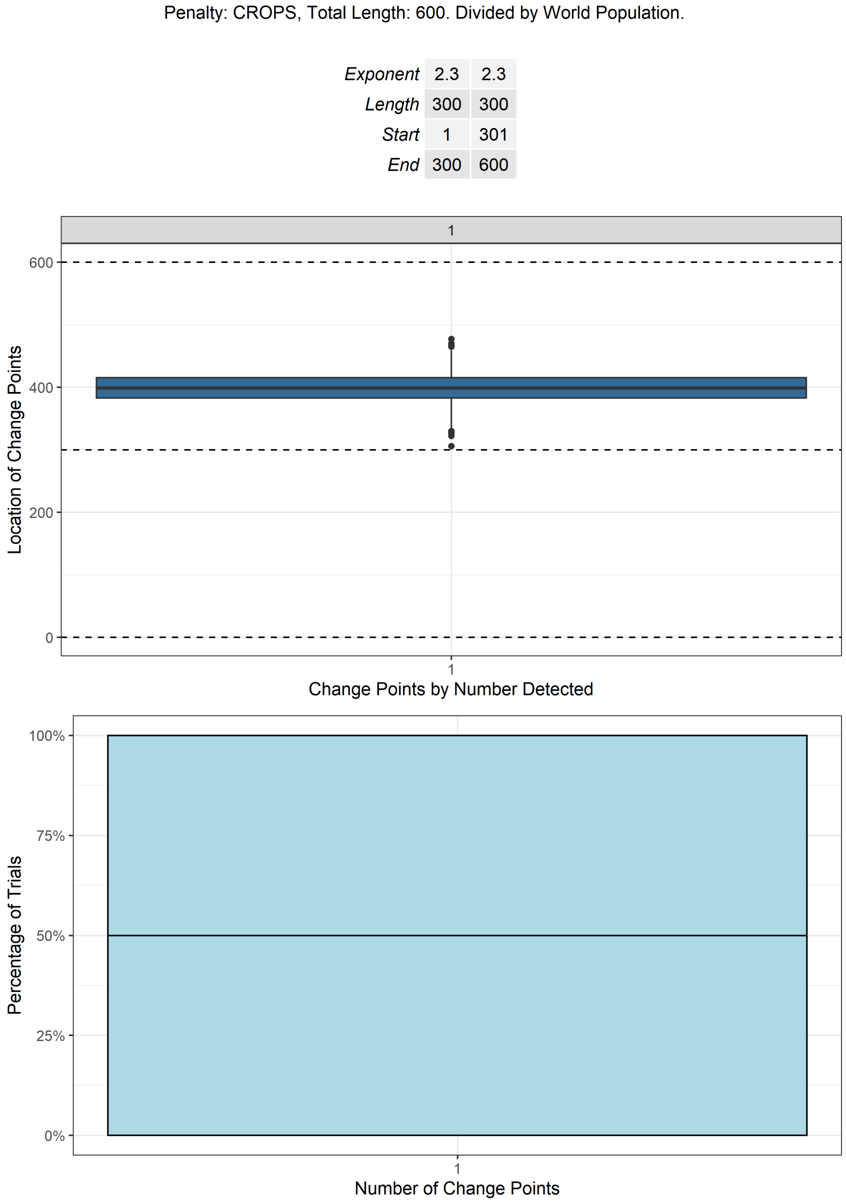
Figure 4.12: World population and its inverse. Only dates in the CoW dataset have been plotted. The inverse is taken to be the multiplicative inverse.

by a factor of approximately 6.5 over the time period of interest and (ii) there are several "kinks" in the population data, notably those around 1870 and 1950. One might naturally expect that the normalisation then could induce the appearance of additional changepoints if there is a sudden change in the world population, as in Figure 4.13, but our simulations demonstrate more complex effects.

For each simulation, we randomly sample with replacement years from the CoW dataset after removing ineligible entries (those without a number of deaths specified). The number of years sampled corresponds exactly to the size of the power-law samples. We then order these years and convert them to their corresponding populations. The data sampled from the power-law distributions are then divided by the population samples, making sure to preserve order throughout.

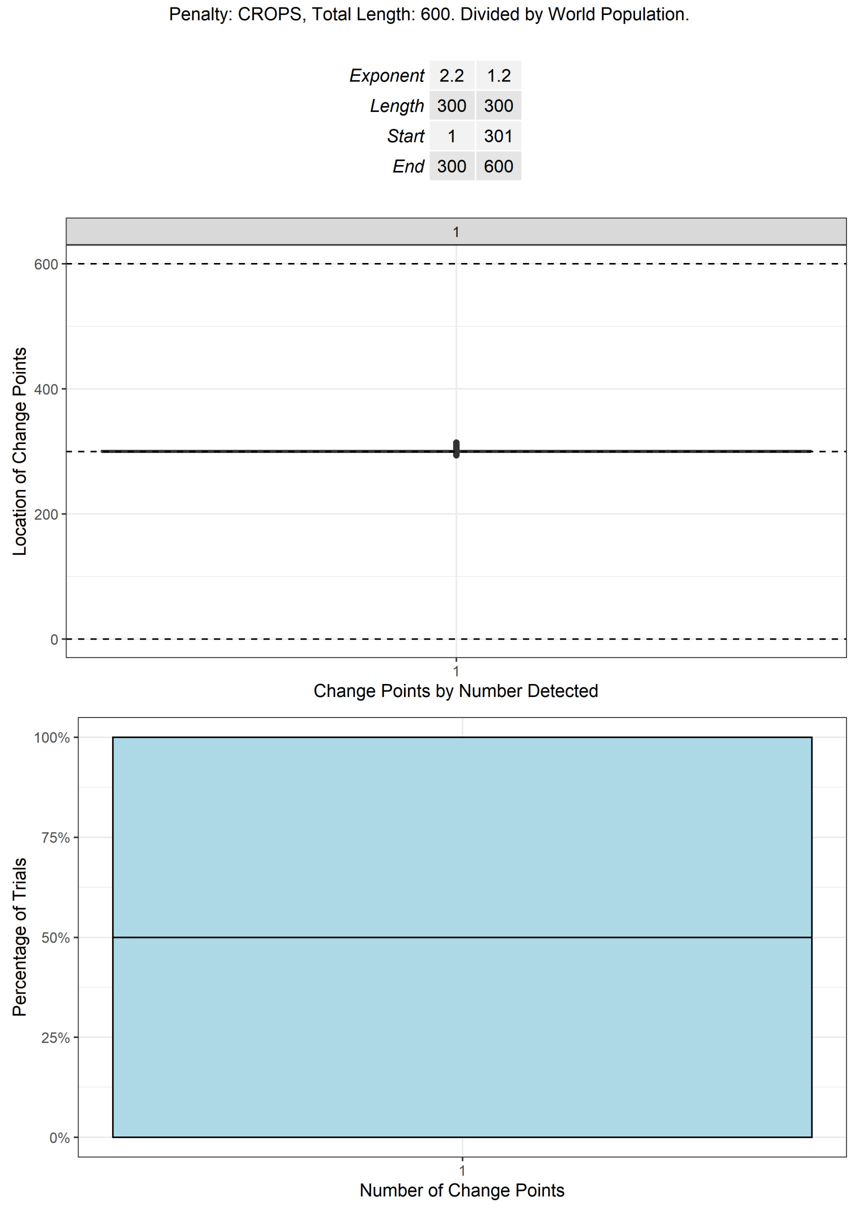
If a population change is induced near a (simulated) changepoint, then highly contrasting effects are possible – the changepoint may dominate, or it may disappear and possibly induce another changepoint elsewhere. We show in Figures 4.14 – 4.17 corresponding outcomes. Naturally, since CROPS and mBIC show somewhat different behaviour, with mBIC producing more changepoints (see Figure 4.17) while CROPS emphasises one detected changepoint (see the other figures), the repercussions are also complex for Algorithm 1. The resulting behaviour appears to depend on a number of factors such as power-law intensities and sharpness of population change. We also note a disagreement on the changepoint locations when combining CROPS and mBIC via their intersection (see Figures 4.18 and 4.19). This is unlike their behaviour on the raw data, but unsurprising since the distributional properties of the raw and normalised data are different.

Altogether, this is not necessarily a problem – if changes in the population balance changes 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 changepoint, in numerator or denominator, cannot be effectively disentangled. Thus the search for changepoints in raw battle deaths and the corresponding search in normalised battle deaths must be conducted separately and subjected to comparative assessment of the findings.



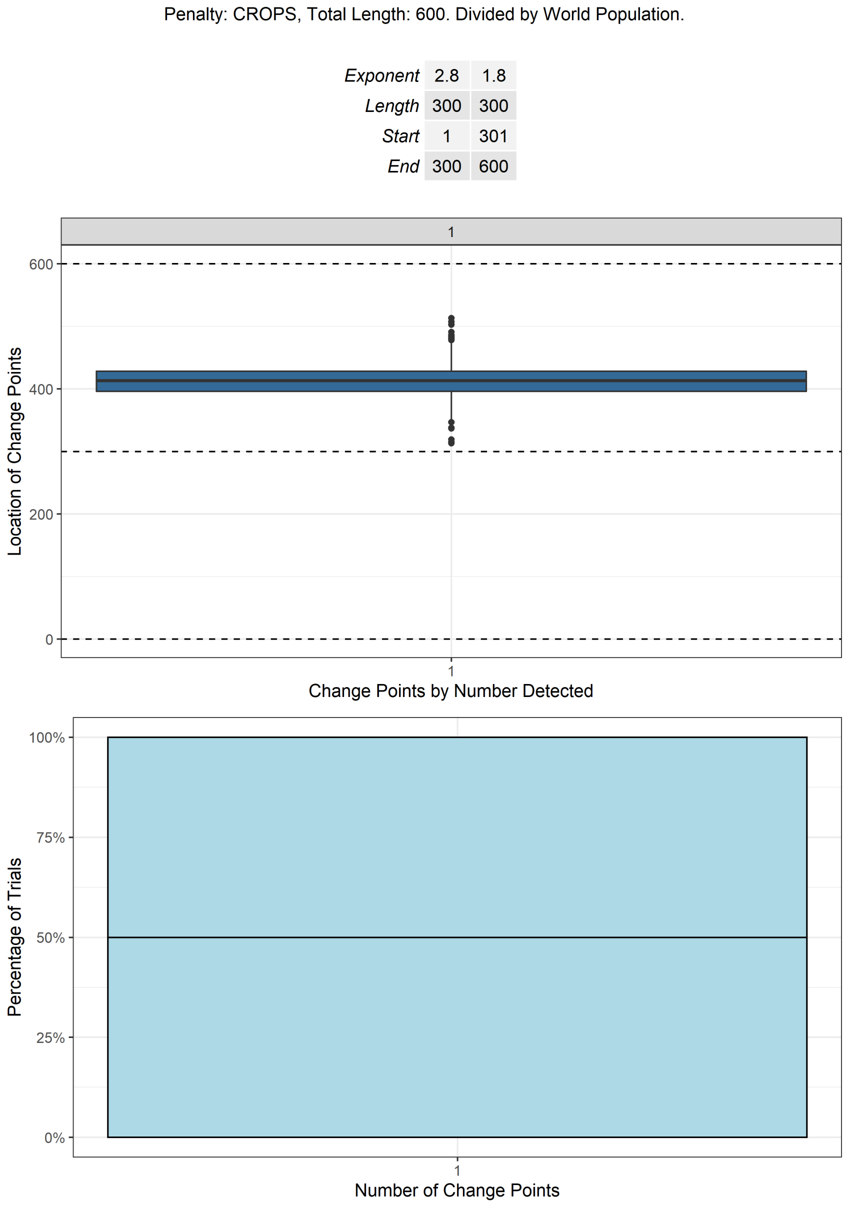
Exponent	2.3	2.3
Length	300	300
Start	1	301
End	300	600

Figure 4.13: A “no changepoints” simulation divided by population. Note that the algorithm identifies a median changepoint that corresponds to the years 1948 – 1950, when the world population experienced a sharp rise (see Figure 4.12).



Exponent	2.2	1.2
Length	300	300
Start	1	301
End	300	600

Figure 4.14: A one changepoint simulation divided by population (I). In this case, the first 300 data are sampled from a power-law with exponent 2.2, and the next 300 are sampled from a power-law with exponent 1.2. Both use minimum values of 10. Despite division by the world population by year, only the power-law changepoint is detected by CROPS.



Penalty: CROPS, Total Length: 600. Divided by World Population.

Exponent	2.8	1.8
Length	300	300
Start	1	301
End	300	600

Figure 4.15: A one changepoint simulation divided by population (II). In this case, the first 300 data are sampled from a power-law with exponent 2.8, and the next 300 are sampled from a power-law with exponent 1.8. Both use minimum values of 10. In contrast to Figure 4.14, in this case only the world population related changepoint is detected by CROPS due to the division of the power-law data world population by year. Also compare Figure 4.16, which makes it clear that the order of the power-laws now matters.

Penalty: CROPS, Total Length: 600. Divided by World Population.

Exponent	1.8	2.8
Length	300	300
Start	1	301
End	300	600

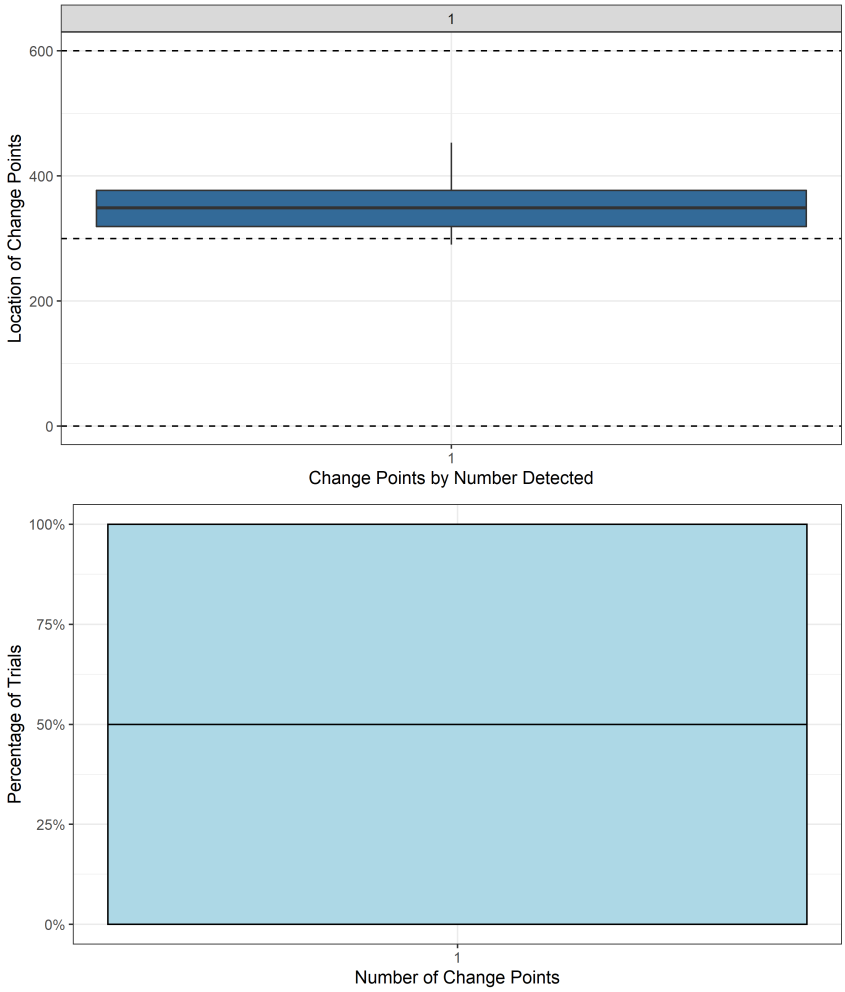
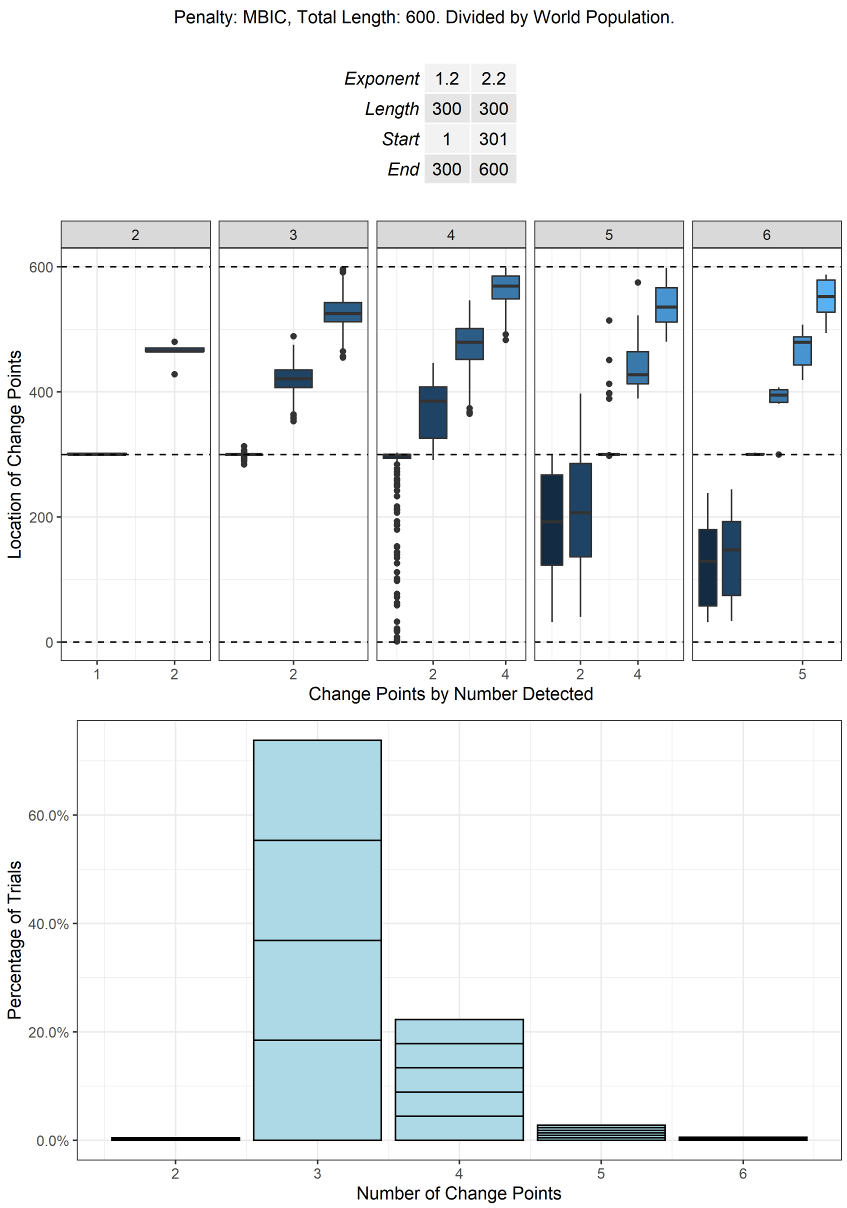


Figure 4.16: A one changepoint simulation divided by population (III). In this case, the first 300 data are sampled from a power-law with exponent 1.8, and the next 300 are sampled from a power-law with exponent 2.8. Both use minimum values of 10 and the simulated power-law data is divided by the world population. In contrast to Figure 4.15 in which the power-law exponents are switched, in this case CROPS detects a single changepoint in between the power-law changepoint and the changepoint induced by the world population.



Exponent	1.2	2.2
Length	300	300
Start	1	301
End	300	600

Figure 4.17: A one changepoint simulation divided by population (IV). In this case, the first 300 data are sampled from a power-law with exponent 1.2, and the next 300 are sampled from a power-law with exponent 2.2. Both use minimum values of 10 and the simulated power-law data is divided by the world population. In contrast to Figures 4.14 – 4.16, mBIC detects multiple changepoints corresponding to the power-law changepoint, one induced by the world population, and a third changepoint not seen in the previous figures.

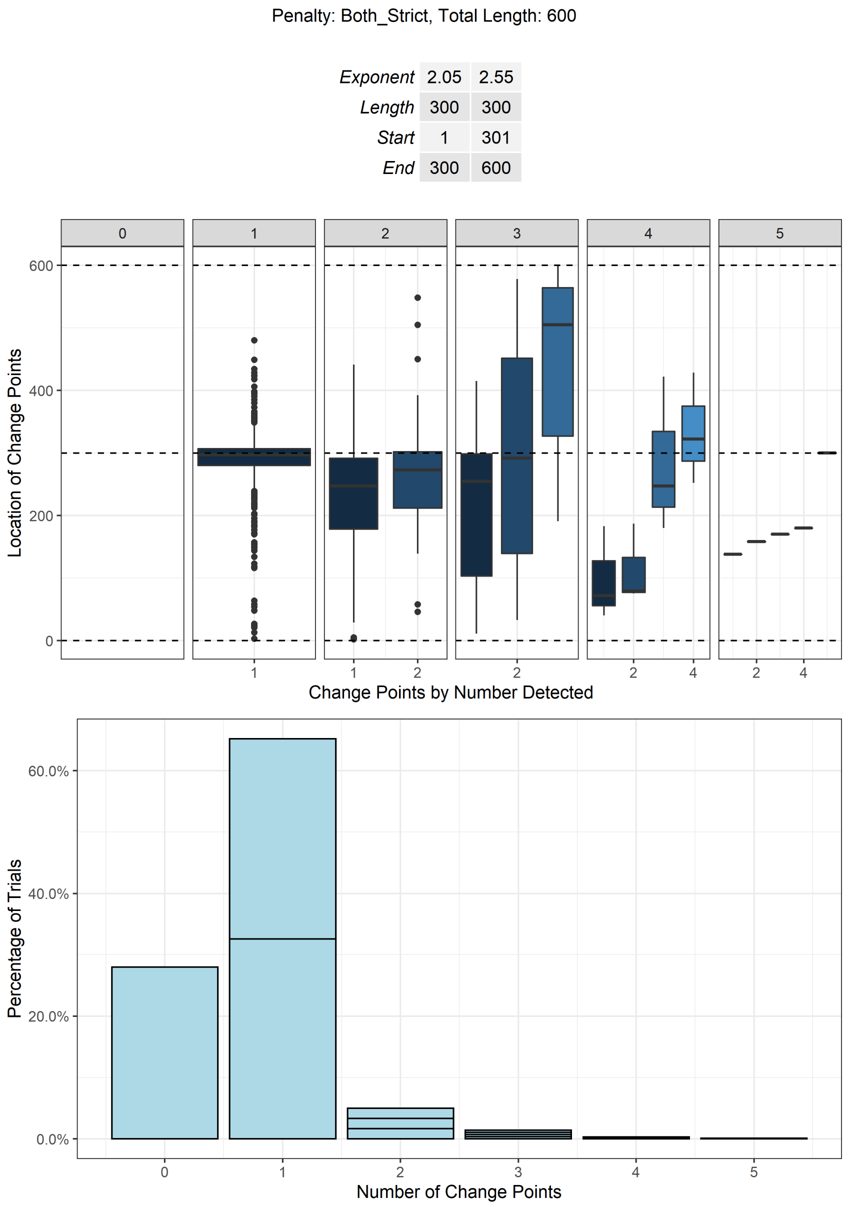
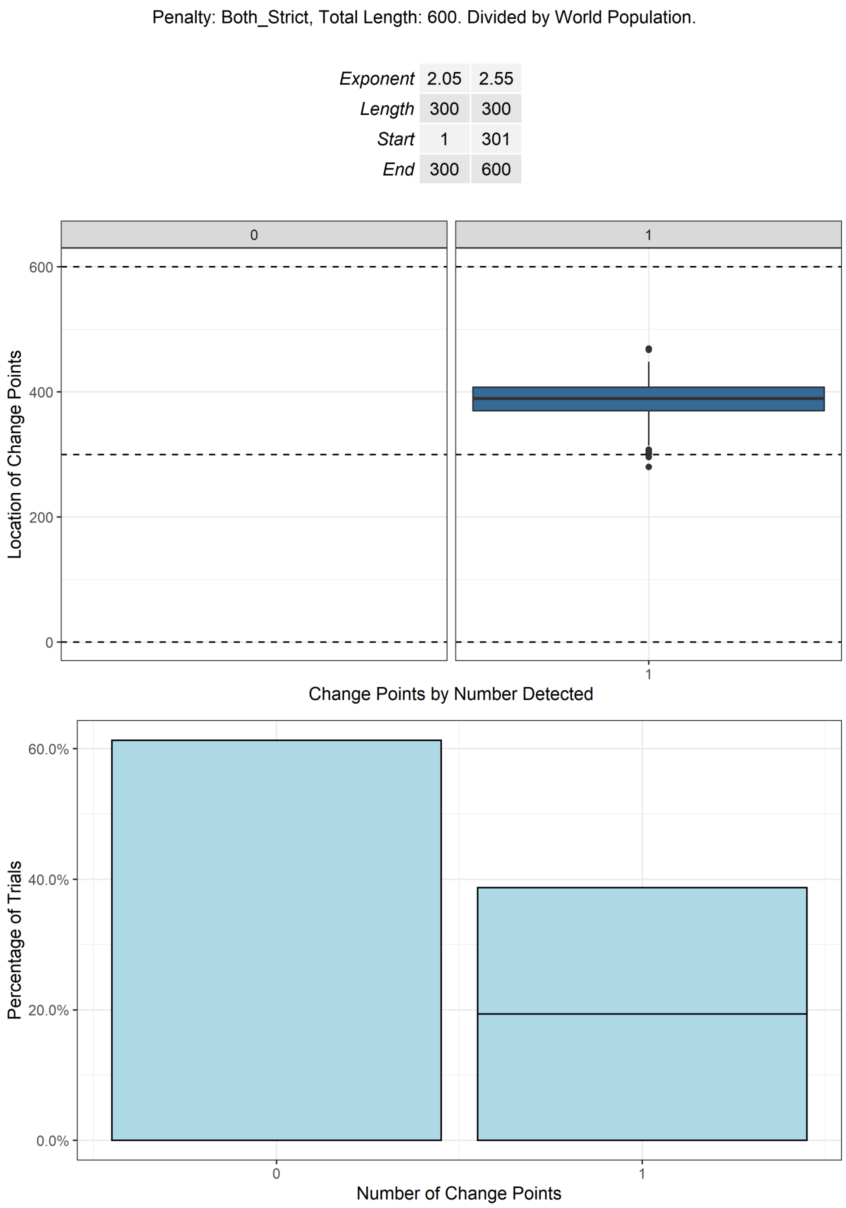


Figure 4.18: The intersection of CROPS and mBIC on raw data. The simulation is of two power-law distributed segments of length 300 with exponents 2.05 and 2.55 respectively and minimum value of 10. For each simulation, both CROPS and mBIC are run on the data as in Algorithm 1, but with only the intersection reported. Here, CROPS and mBIC mostly agree on the presence and location of a single changepoint. Compare to Figure 4.19, which features normalised data.



Penalty: Both_Strict, Total Length: 600. Divided by World Population.

Exponent	2.05	2.55
Length	300	300
Start	1	301
End	300	600

Figure 4.19: The intersection of CROPS and mBIC on normalised data. The simulation is of two power-law distributed segments of length 300 with exponents 2.05 and 2.55 respectively and minimum value of 10, divided by the world population by year. For each simulation, both CROPS and mBIC are run on the data as in Algorithm 1 , but with only the intersection reported. In contrast to Figure 4.18, CROPS and mBIC are substantially more likely to disagree about the exact placement of changepoints and thus tend to report that no changepoints were found.

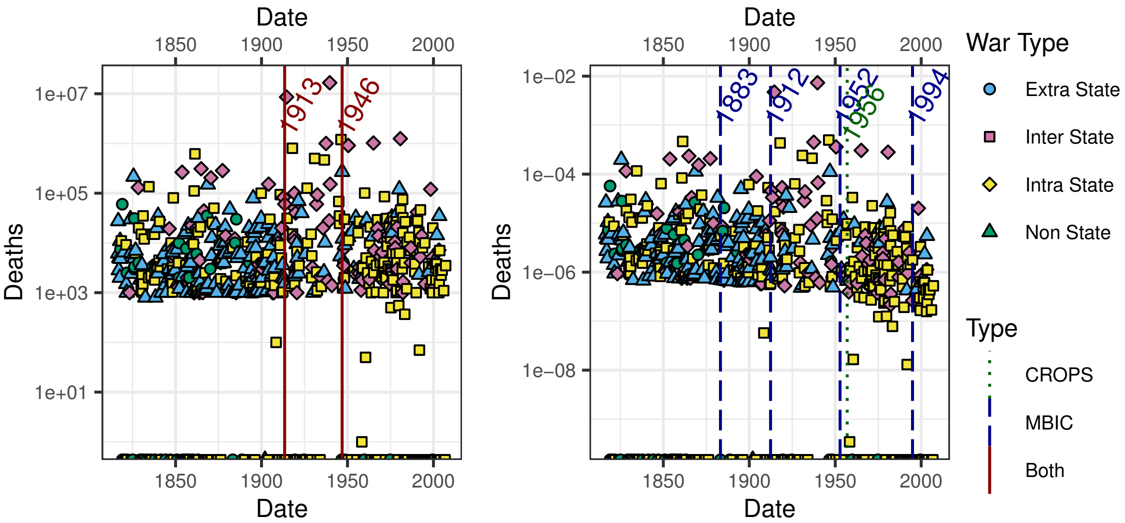


Figure 4.20: Results from applying Algorithm 1 to CoW for all data subsets. On the left we use raw data; on the right, data rescaled by world population at the time of the conflict. Vertical bars indicate detected changepoints annotated by exact years for clarity. *A-posteriori* power-law distribution fits suggest that α moves from 1.77 to 1.51 to 1.65 in the raw data.

4.4 Analysis of historical battle deaths

Using the insights gained in the simulation study above, we now apply the proposed algorithm to the datasets described in Section 4.2. The results indicate with confidence the existence of changepoints in the data. In the entire raw CoW dataset, shown in Figure 4.20, there are two changes, just prior to World War I and just after World War II. When scaled by population two more candidate changepoints 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 changepoints overall since the results are not identical across CROPS and mBIC. The further detected post-World War II changepoint in particular is likely the result of a sharp change in population in the 1950s. A similar effect can be observed in Figure 4.13. This supports the proponents of the long peace hypothesis, albeit via an argument for the great violence.

It is less clear to assign changepoints in the entire Gleditsch raw dataset, but the emerging 1994 changepoint in data scaled by population size is now conclusively found (Figure 4.21). The broad message is similar, with candidate changepoints found pre-World War I, post-World War II, and 1994. In contrast to the raw CoW dataset, we find evidence for change in the 1840s. In addition, the Gleditsch analysis suggests a change in the mid-1930s.

The suggestions made by Cirillo and Taleb (2016b, p. 30, 32) to transform the data in order to account for the finite upper bound on battle deaths due to having a finite world population appear to have little impact (see Figure 4.22). 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 changepoints, especially in the Gleditsch dataset. Among CoW and its various subsets, an exception is the combined CoW dataset, as shown in Figure 4.22. The limited sensitivity to such transformations is probably due both the inherent scale invariance of power-law distributions as well as to the lack of data points located 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

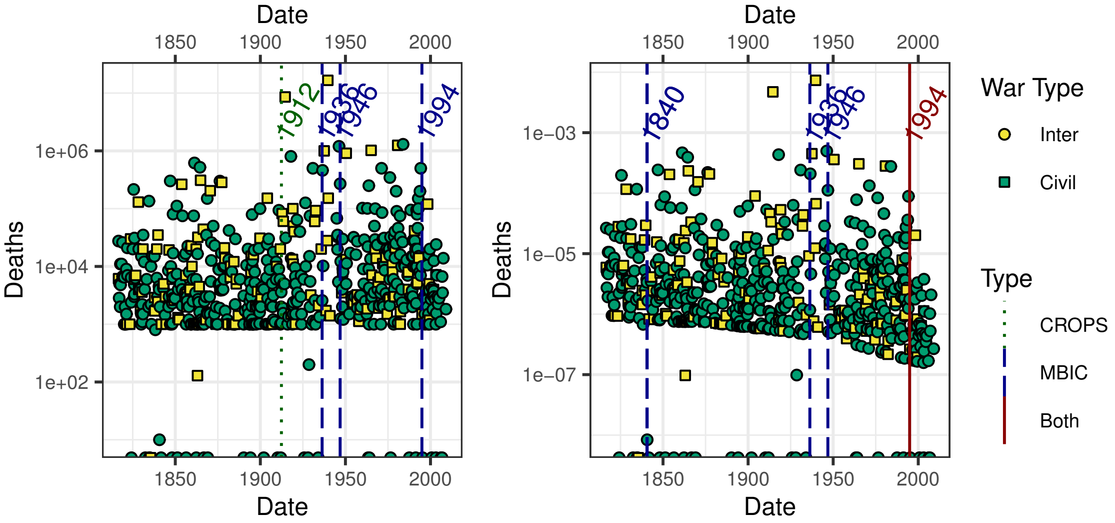


Figure 4.21: Results from applying Algorithm 1 to Gleditsch's combined datasets. On the left we use raw data; on the right, data rescaled by world population at the time of the conflict. Vertical bars indicate detected changepoints annotated by exact years for clarity. *A-posteriori* power-law distribution fits suggest that α moves from 1.74 to 1.65 when the raw data is partitioned using CROPS. When the data is partitioned using mBIC, α instead moves between $1.71, 1.44, 1.73$, and 2.08.

the CoW combined dataset, in that the 1913 changepoint in the raw data has a similar likelihood of being identified as the 1936 changepoint in the transformed data.

As noted in Section 4.3, normalisation of power-law distributed data by world population can obscure or even eliminate what would be considered changepoints 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 one or the other analysis was true. To do so, we carry out *a-posteriori* robustness checks by assuming the raw segmentation to be true, fitting a power-law to each segment using the methods of Clauset *et al.* (2009b, implemented by Gillespie (2015)), simulating from the now-parametric model, and assessing the identified changepoints.

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

For the raw Gleditsch data, Algorithm 1 does not identify a specific changepoint due to 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 changepoint. This behaviour is unsurprising, given the algorithm aims to protect against false positives. For the normalised Gleditsch data, we are confident that the 1994 changepoint is identified.

Another important consistency check is whether any real datasets exhibit no changepoints. We recall that we have already demonstrated that the no-changepoints case for our methodology is evidenced by a particular combination of a large number of (false) positives from CROPS and few or no (false) positives from mBIC for a wide range of data points, seen in Figure 4.6. Whilst we have established the robustness of the methods against artificial data, the existence of a dataset

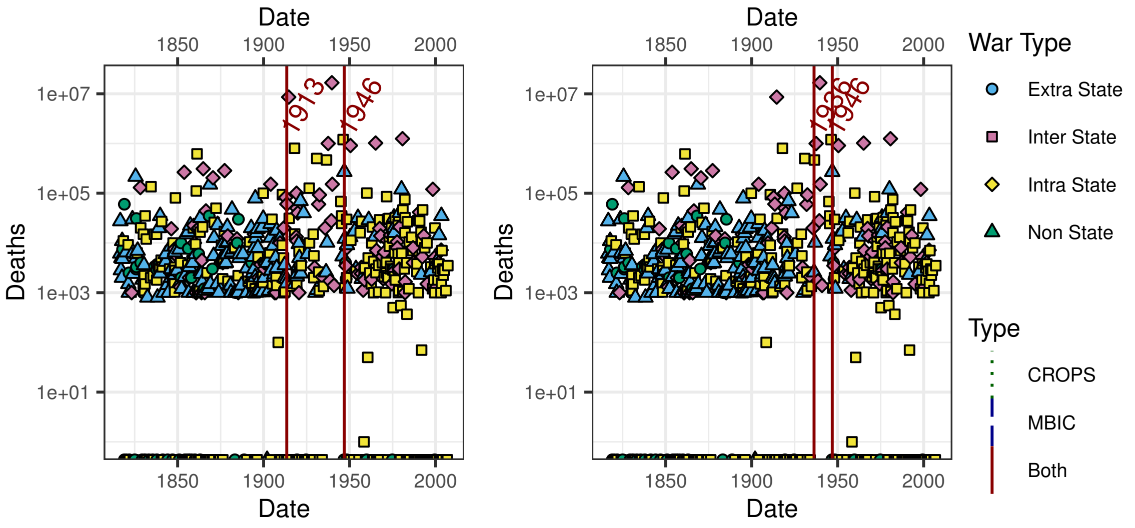


Figure 4.22: Results from applying Algorithm 1 to the combined CoW dataset, rescaled as recommended by Cirillo and Taleb (2016b, p. 30, 32). On the left, the data is rescaled using the current (2018) world population. On the right, data is rescaled using the world population at the time of the conflict. Vertical bars indicate detected changepoints annotated by exact years for clarity.

with no changepoints would clearly help validate our methods while also identifying a setting consistent with the null hypothesis of no change in the statistical properties. It is therefore worthy of comment that such a dataset within CoW does exist: the CoW non-state dataset, shown in Figure 4.23, has a response clearly of the same type as in the bottom row of Figure 4.6, indicating a potential unchanging underlying mechanistic reason for this phenomenon.

To get a sense of the robustness of our approach and to represent the overall prevalence of changepoints, in Figure 4.24 we present an internal meta-analysis across all the analyses we have performed on the CoW and Gleditsch datasets. This figure shows where changepoints are found in all composing internal data subsets by the proposed algorithm identified in Section 4.3. In the top panel of each sub-figure, we place a kernel density estimate of the locations of changepoints; the sub-figures and density estimates were created using a $1/5$th adjustment to the default bandwidth to sharpen the location of changepoints. In the bottom panel of each sub-figure, we present the data subsets as a time line. Shaded regions, for changepoints over a period of time, and dotted lines, for changepoints located at a single time, indicate the location of clusters of changepoints clustered 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 changepoint during the period, independent of the magnitude of the change. This panel gives a clear sense of the robustness of the 1994 changepoint, less strongly the robustness of the 1830s changepoints, and the variations that exist in the period 1910-1950 of the changepoints. The graph shows the location of individual points but also more finely-grained variation where multiple methods and datasets produce changepoints 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.

4.5 Discussion

In this chapter, we have used simulation studies to highlight a key problem with changepoint analysis of heavy-tailed data, as well as to come to a solution using a synthesis of modern techniques.

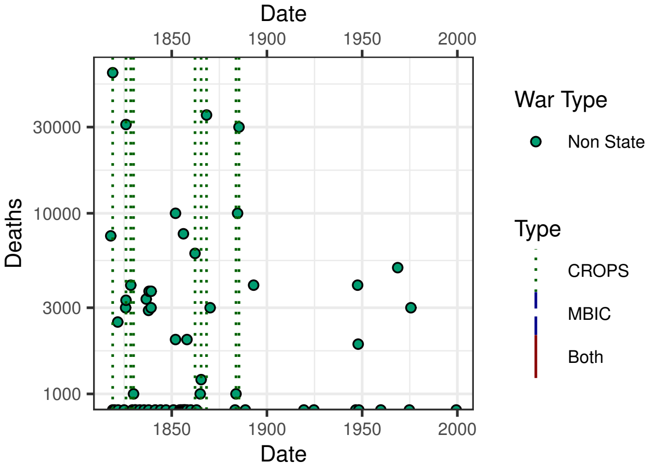
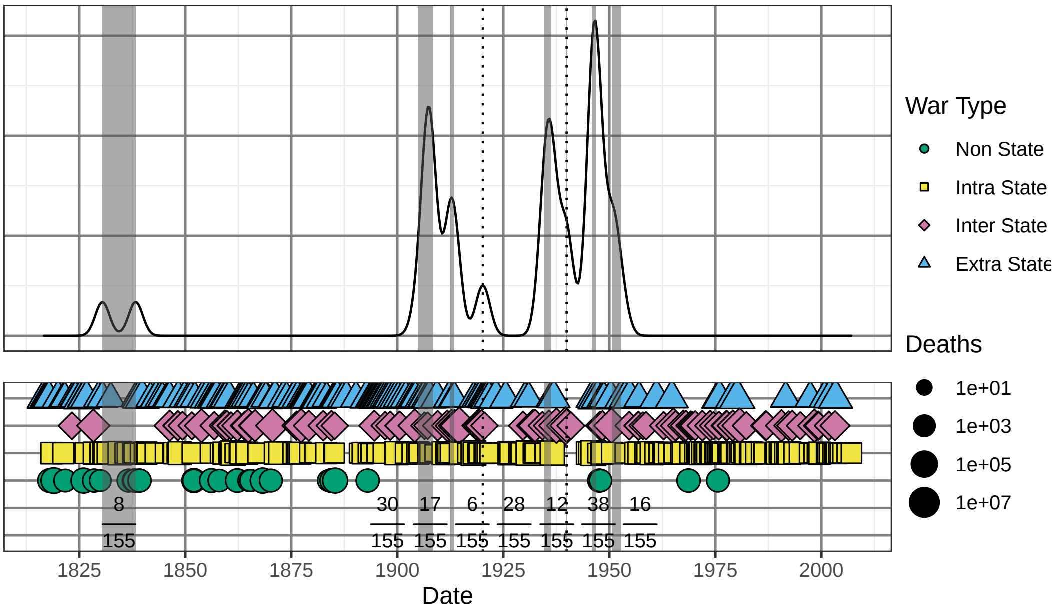


Figure 4.23: Results from applying Algorithm 1 to the non-state CoW dataset. No changepoints are found by the mBIC penalty and CROPS finds a large number of tightly clustered points. Note that this result is extremely indicative of no changepoints. For comparison, see Figure 4.6.

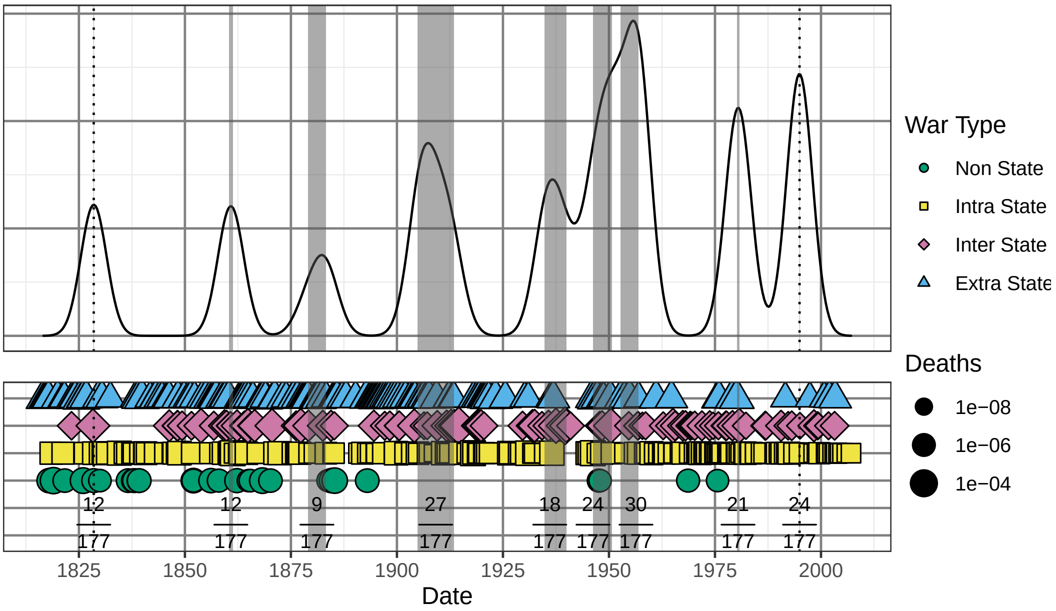
As shown in Section 4.3, many previous techniques result in over-segmentation of power-law distributed data. Using Algorithm 1, a combination of PELT (Killick *et al.*, 2012) with a cost function based upon the tail of the empirical distribution (Haynes *et al.*, 2017b) and both the CROPS (Haynes *et al.*, 2017a) and mBIC (Zhang and Siegmund, 2007), allowed us to arrive at vastly more accurate segmentations. While we have tested this algorithm on power-law distributed data, it is theoretically sensitive to when the tails of the empirical distribution sufficiently change due to its non-parametric approach, which would theoretically make it a general tool for heavy-tailed data.

Algorithm 1 was designed with the distributions of battle deaths in mind, which have long been known to be power-law distributed (Richardson, 1944, 1960b; Clauset *et al.*, 2009b; González-Val, 2016; Chatterjee and Chakrabarti, 2017; Clauset, 2018; Martelloni *et al.*, 2018). Due to its non-parametric nature, Algorithm 1 is purely data-driven and provides an absolutely neutral attempt at looking for changepoints in a field where significant ink has been spilt over the biases of both authors and their methods. Algorithm 1 does have some natural limitations: it is designed for detection of abrupt changes in distribution, but continuous changes need to be handled explicitly, as we have done with the world population. Continuous changes as postulated by Pinker (2011, 2018) are beyond the scope of our work and would be an excellent future research direction.

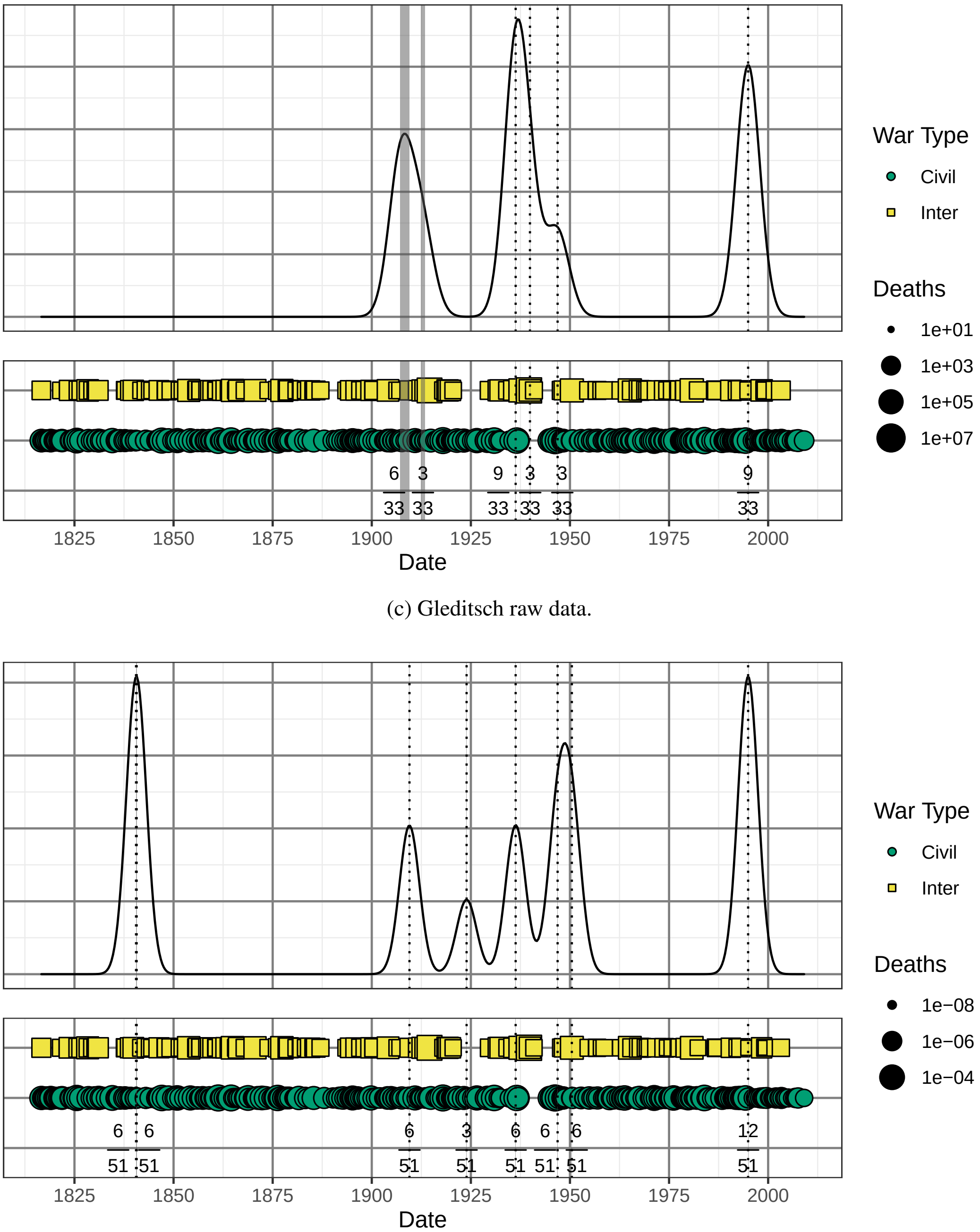
Applying Algorithm 1 to the CoW and Gleditsch data sets indicated the existence of changepoints. Consistently, 1910-1950 are identified as changepoints, but significant variations exist, including whether the World Wars are both included, whether they form a pair of changepoints, and whether the post-World War II period is included. This provides evidence for some form of long peace, albeit sometimes via argument for the great violence. In particular, the possibility of a changepoint in the 1930s in the Gleditsch data set might reflect the complex civil wars of that period. Our results are more vague about the existence of a changepoint in the 19th century in the CoW dataset. Possible landmark events include revolts around the 1830s, the American Civil War and Taiping Rebellion of the 1860s, and colonial wars in the 1880s. More precise is the identification, when normalised by world population, of a segment containing the single point of the Second Syrian Civil War Phase 2, September 1840. This point is likely an anomaly within the Gleditsch data set given it has only 10 recorded deaths. The algorithm detects that this point is far outside of the normal behaviour of either section before and after this point. On the other hand,



(a) CoW raw data



(b) CoW normalised data



(c) Gleditsch raw data.

(d) Gleditsch normalised data.

Figure 4.24: Results for internal meta-analyses performed on all changepoints found in any combination of subsets within the datasets. In order, these plots correspond to (i) CoW raw data, (ii) CoW normalised data, (iii) Gleditsch raw data, and (iv) Gleditsch normalised data. In each plot, there are two images. The lower of each pair of images is a timeline of events occurring, sorted by subset. Above it is a density estimation of the locations of changepoints detected. The area under the curve of the estimate is proportional to the probability of finding a changepoint within that part of the dataset. Grey bars and dotted lines represent changepoints, in different locations or the same location respectively, that have been clustered. Numbers below the timelines indicate the fraction of identified changepoints so clustered.

there is a much more robust changepoint in 1994, consistently identified by CROPS and mBIC in the normalised CoW and both normalised and unnormalised Gleditsch data sets. This changepoint supports the hypothesis of Gurr (2000, see also the work of Cederman *et al.* (2017)) who identified a decline in ethnic conflict since the mid-1990s.

4.6 Conclusion

In this chapter, we have discussed the pairing of changepoint analysis with the long peace hypothesis for the decline of war. Our work is a natural extension of existing work in both fields, but required the development and extension of methods to the case of power-law distributed data. There are some weaknesses in our technique: we do not explicitly handle every case and we can only detect discrete changepoints. Nonetheless, our methods are the most up-to-date and have been shown in a variety of simulation studies to robustly handle power-law distributions, which have unusual properties that normally hinder changepoint analysis (see Section 2.1 for some such properties).

While our analysis addresses the existence of changepoints in a controversial and important debate using modern techniques, more work can still be done. Leveraging the design of Algorithm 1 for power-law distributed data also suggests that it could be applied to many other datasets such as blackouts, book sales, and terrorism (Clauset *et al.*, 2009b). Changepoint analysis could also be developed specifically for the heavy-tailed multiple changepoint setting. One future possibility is to examine whether any transformations can make the problem more tractable. For example, as mentioned in Chapter 2, the logarithm of power-law distributed data normalised by its minimum would make the data theoretically exponentially distributed. Additional adaptation of our algorithm to the multivariate setting would allow better exploration of the distributions of wars, given their multi-faceted nature. Duration, arrival time, and war type are all relevant parameters that we have not explored in this analysis. Our analysis also necessarily does not include both continuous and discrete changes outside of the world population. Development could also continue in the direction of data collection: producing and investigating a more disaggregated dataset using better-resolved events.

While changepoint analysis highlights when the effects of a discrete change in a distribution are felt, it does not and cannot tell us the true nature of the change. We can try to characterise these changes using various test statistics in our analysis, but this is no guarantee that we have identified the best test statistic, let alone characterise the effects and driving factors of of the change. To do so requires an understanding of the generative process of the distribution and how that generative process changes over time, something that we have avoided in this chapter to be purely data-driven. In our next chapter, we exchange roles and instead look at one model that has been proposed as a generative process for terrorism data. This model is coalescence and fragmentation and we seek to understand how apt and robust this model is for its many applications.

5 Coalescence and fragmentation

[A]nalogies are more than beautiful testaments to the unifying power of models: they are headlights in dark unexplored territory... Models can surprise us, make us curious, and lead to new questions. *– Epstein, 2008*

5.1 Introduction

Coalescence and fragmentation models have always had an applied nature, going back even to their first use by von Smoluchowski (1916). Such is their nature as a collection of reasonably simple, but extremely flexible and deep models. It is perhaps little surprise that 'particles', be they probabilistic (Aldous, 1999), chemical (von Smoluchowski, 1916; Spicer and Pratsinis, 1996), fish (Datta *et al.*, 2011), people (D'Hulst and Rodgers, 2000; Bohorquez *et al.*, 2009), banks (Pushkin and Aref, 2004), or asteroids and dust (Tanaka *et al.*, 1996; Birnstiel *et al.*, 2011), obey similar rules regarding how they collide and stick together or how they break apart. As discussed in Section 2.4, these models have interesting emergent properties, such as the presence of gelation, the formation of a superparticle of infinite size in finite time or even instantaneously (Carr and da Costa, 1992). Many variations exist that we must neglect in this chapter due to space and time, such as the spontaneous formation of particles of infinitesimal size (McGrady and Ziff, 1987). We will limit our discussion to gelation, its consequences, and the system's robustness for their repercussions on applications.

Our motivation stems from work done by Bohorquez *et al.* (2009) and Johnson *et al.* (2016), in which a simple well-mixed coalescence and fragmentation model was used to model first terrorist events and then the evolution of pro-ISIS (Islamic State) networks within the Russia-based social network VKontakte. In the work of Bohorquez *et al.* (2009), the overall model is formed of two parts: the first is coalescence and fragmentation to model the various groups within the (approximately) fixed homogeneous (terrorist) population, while the second is a group consensus model for when to launch an attack (with sizes proportional to the groups initiating the attack). Coalescence corresponds to individuals in the groups agreeing to coordinate resources; each individual is equally capable of doing so using arbitrary methods with arbitrary ranges and so a multiplicative coalescence kernel is assumed. This reflects that any individual can coordinate with any other individual, which suggests that for two groups of sizes i and j that there are ij ways for two groups to begin coordinating. Fragmentation is a rare event (Bohorquez *et al.* (2009) choose their fragmentation rate such that there is approximately 1 fragmentation for every 99 coalescence events) in which an individual in a group perceives danger to cause the entire group to cease co-operation immediately. As such, we will also work with processes where coalescence is more frequent than fragmentation, although we will vary the exact probabilities. They claim robustness of the model with respect to changes in the underlying rules with citation to D'Hulst and Rodgers (2000) , Hui *et al.* (2003), and Ruszczycki *et al.* (2009) as well as by considering the cases of two interacting populations (*e.g.* insurgents and counter-insurgents) and two interacting populations with one non-interacting population (*e.g.* civilians) in their supplemental materials[22].

In the latter work by Johnson *et al.* (2016), single particles are considered to be total number of 'follows', while the groups are composed of numbers of follows by followers, due to the bipartite nature of the networks[23] under consideration (Johnson *et al.*, 2016). Despite the bipartite nature, Johnson *et al.* (2016) observe that groups on occasion completely absorb other groups. Fragmentation is an exogenous process: an outside force/predator/moderator identifies the pro-ISIS organisations and shuts them down. Johnson *et al.* (2016) also observe 'shark-fin' dynamics where there is growth to apparent plateaus before sudden collapse driven by the combined coalescence and fragmentation dynamics, a critical threshold for moderation below which the pro-ISIS organisations 'grow exponentially fast into one super-aggregate', and an intervention strategy based on targeting (only) medium sized, as opposed to (only) large, organisations. In their supplemental materials, they provide a table of model variants to which they claim the steady-state and shark-fin dynamics are robust. This table includes higher order coalescences, fragmentation to fragments of a single size larger than 1, population fluctuations, a spatial grid, exponents on the number of groups, and heterogeneous followers.

In both usage cases, we want to understand the appropriateness of applying coalescence and fragmentation models. While we noticeably are lacking in data on real time terrorist group formations, we use simulations to investigate how robust the model is to changes in the modelling assumptions. For example, it seemed unlikely to us that an entire group would scatter upon sensing danger or being targeted by a predator and it thus is important to know if making such an assumption is fundamental to how the model works, or if the system is robust to weakening the form of the fragmentation used. Similarly, the observation, and prevention, of a 'super-aggregate' or gel is a critical feature of a broad class of coalescence models which have fundamentally different behaviour than systems lacking gelation. Ideally, we would be able to establish the effects of changes in the assumptions and robustness, or lack thereof, in the homogeneous population model before continuing on to consider a networked population, such as the one investigated by Johnson *et al.* (2016). From there, many different types of complex networks could be investigated, with similarly many potential applications.

There are competitor models (e.g. Clauset and Gleditsch, 2012) and these and the group consensus model are outside of the scope of this chapter. We do not cover the whole breadth of the disciplines of coalescence, fragmentation, their combination, and neighbouring topics; to do so would require discussions including current analysis, semigroup theory, graph theory, probabil-

[22] Curiously, the supplemental materials of Bohorquez *et al.* (2009) also suggest the model is robust to 'more rigid' 'larger attack groups', but this text is invisible and without citation (Figure 2).

[23] Networks are composed of individuals and organisations. Individuals can follow organisations, but other interactions (*e.g.* individual-to-individual private messaging) are not observed.

ity, and other 'pure' disciplines alongside discussions of a wide variety of applications. In this chapter we rely on the overview given in Section 2.4 and instead focus on a few aspects of von Smoluchowski's coagulation equation with fragmentation. We discuss in Section 5.2 findings pertaining to the common steady-state analysis of, *e.g.*, Clauset and Wiegel (2010). In Section 5.3, we find that the proposed steady-state is not always arrived at in simulations. We discover a new phenomenon in which the system is dominated alternatively by coalescence and fragmentation leading to *gel-shatter cycles*. A natural question then arises as to the ubiquity of the cycles we find; we address this in Section 5.4 before finishing this chapter with a discussion of future directions in Section 5.5 and placing the chapter in the context of the thesis in Section 5.6.

5.2 Steady-state

In Section 2.4, we presented a now standard argument for the steady-state for systems of the form

$$\begin{aligned} \dot{n}_k &= \frac{1}{2}\sum_{i=1}^{k-1} K(i,k-i)n_i n_{k-i} - n_k \sum_{i=1}^{\infty} K(i,k)n_i - F(k)n_k, \quad k \geq 2, \\ \dot{n}_1 &= \sum_{i=2}^{\infty} iF(i)n_i - n_1 \sum_{i=1}^{\infty} K(i,1)n_i, \end{aligned} \tag{5.1}$$

where n_k is the number of clusters of size k per unit volume, $K(i,j)$ is the (binary) coalescence kernel, and $F(i)$ is the shattering fragmentation kernel. For the special case $K(i,j) = \hat{K}ij$ and $F(i) = \hat{F}i$, where $\hat{K}$ and $\hat{F}$ are positive constants, we can easily observe a link from the steady-state, $\forall k, \dot{n}_k = 0$, to a combinatoric construct, the Catalan numbers (The On-Line Encyclopedia of Integer Sequences). The Catalan numbers have many applications: the easiest might be the number of ways to (correctly in the standard sense) insert i pairs of brackets in a sequence of length $i + 1$. They have the form

$$C_i = \sum_{k=0}^{i-1} C_k C_{i-1-k} = \frac{1}{i+1}\binom{2i}{i} \tag{5.2}$$

with $C_0 = 1$. To our knowledge this link has not been directly observed in the literature, but methods used are often similar to the methods used to analyse the Catalan numbers.

At the steady-state with a multiplicative kernel, rearrange for $k \geq 2$ and set $\rho_k = kn_k$ so that

$$\rho_k = \frac{\frac{1}{2}\hat{K}}{\hat{F} + \hat{K}\sum_{i=1}^{\infty}\rho_i} \sum_{i=1}^{k-1} \rho_i \rho_{k-i}. \tag{5.3}$$

Provided that $\sum_{i=1}^{\infty}\rho_i$ is well-behaved, this is a (offset) standard recurrence relation for the Catalan numbers except with a prefactor (Hilton and Pedersen, 1991). Labelling the prefactor in recurrence relation 5.3, say, γ, and using the generating function

$$f(z) \equiv \sum_{k=1}^{\infty} \rho_k z^k \tag{5.4}$$

yields

$$f(z) - \rho_1 z = \gamma f^2(z) \implies f(z) = \frac{1 \pm \sqrt{1 - 4\gamma\rho_1 z}}{2\gamma}. \tag{5.5}$$

Binomial series expansion of the physical negative branch[24] yields

$$\sum_{k=1}^{\infty} \rho_k z^k = \sum_{k=1}^{\infty} \frac{(2k-3)!!}{(2k)!!} 2^{2k-1} \gamma^{k-1} \rho_1^k z^k, \tag{5.6}$$

but the terms can be rewritten using the Catalan numbers:

$$\rho_k = C_{k-1} \gamma^{k-1} \rho_1^k. \tag{5.7}$$

Application of Stirling's law provides the expected truncated power-law exponent $-3/2$. For the special case we are considering, the total amount of mass (over the closed and fixed unit system volume) in the system is expected to be fixed, $M = \sum_k \rho_k = \sum_k k n_k = f(1)$, allowing for straightforward computation, since

$$\gamma = \frac{\frac{1}{2}\hat{K}}{\hat{F} + \hat{K} f(1)}, \qquad \rho_1 = f(1) - \gamma f^2(1) = \frac{\hat{F} + \frac{1}{2}\hat{K} M}{\hat{F} + \hat{K} M} M.$$

Relating the coalescence and fragmentation equations to Catalan numbers naturally relates multiple results together. While our results above make the connection clearly for the multiplicative kernel, similar techniques can be employed for other kernels. For the constant kernel, the result still holds, although $f(1)$ is no longer the mass of the system, but rather the number of clusters in the steady-state. Other generalisations are also possible. For example, consider the generalised Catalan numbers (Hilton and Pedersen, 1991)

$$_p d_k = \frac{1}{(p-1)k+1} \binom{pk}{k}, \tag{5.8}$$

for which $C_i = {}_2 d_i$. We hypothesise that the generalised Catalan numbers could be used to reframe the results for generalised coalescence kernels of D'Hulst and Rodgers (2000) and Kyprianou *et al.* (2018), given that these calculations have final forms that are reminiscent of Equation (5.8). This helps to establish some of the robustness of coalescence and fragmentation systems in applications. Just as the Catalan numbers are generalisable, we would expect coalescence and fragmentation systems of the form of System 5.1.

These results about the expected steady-state do not tell us about how much we might expect to deviate from the steady-state at any given timea critical factor in understanding how much we might deviate from the analytical solution in practice. Some work for the coagulation equation has already been done on deviations by van Dongen and Ernst (1987) and van Dongen (1987) using the system size expansion methods of van Kampen (1992) for the Smoluchowski coagulation equations with multiplicative kernel and constant, additive, and general homogeneous kernels respectively. Van Dongen and Ernst (1987) identify some problems with application of the system size expansion to binary fragmentation. For the case of shattering fragmentation, problems can also arise due to the presence of large (system sized) fluctuations when shattering large clusters. If the largest cluster size is sufficiently small, such approaches are more tractable. We apply Fourier

[24] The positive branch has $f(0) = \gamma^{-1}$, which conflicts with our definition of $f(z)$. Additionally, we are not worried about the radius of convergence in terms of z as we will not be evaluating $f(z)$ in the expanded form. It may be reassuring to note that we will be working with $\hat{K} \gg \hat{F}$, so $\gamma \approx \frac{1}{2f(1)}$. If we do not have a gel, which empirical steady-state systems support, then $\rho_1 \approx \frac{f(1)}{2}$, altogether suggesting that the radius of convergence in terms of z is $|z| < 1$. Hence we should expect agreement at $f(0)$.

analysis to the Langevin equation derived from the system size expansion for small systems, here with maximum cluster size of 3, following the extension of van Kampen (1992) by McKane *et al.* (2007). This approach allows us to look for oscillations in the number of clusters of each size, which are to be expected in dynamical systems representing interacting chemical or biological species (McKane and Newman, 2005; McKane *et al.*, 2007). Problems emerge analytically when the maximum cluster size is allowed to reach the system size, as the higher order fluctuations then grow, rather than shrink, as the system size increases. Future work may be able to reconcile raising the maximum cluster size while controlling the higher order fluctuations, but for now we restrict the maximum cluster size to enable numeric approaches.

Consider a slight variant coalescence and fragmentation system for clusters that can only create clusters of up to three or disassociate into monomers even for large populations

$$\begin{aligned}
\dot{n}_3 &= 2K(1,2)n_1n_2 - F(3)n_3 \\
\dot{n}_2 &= K(1,1)n_1(n_1-1) - 2K(1,2)n_1n_2 - F(2)n_2 \\
\dot{n}_1 &= -2K(1,2)n_1n_2 - 2K(1,1)n_1(n_1-1) + 2F(2)n_2 + 3F(3)n_3.
\end{aligned} \tag{5.9}$$

The variation is absorbing the factor of $1/2$ into the coalescence kernel and explicitly forbidding self-coalescence, as we will be examining what happens with a limited population. In the simulation, we set $K(i,j) = (1 - \text{Pr}\,(\text{Frag.}))i(j-\delta_{i,j})/(M(M-i))$, noting that we can take the average of $K(i,j)$ and $K(j,i)$ in calculations. We take $M = 100$ and size-biased kernels with $\hat{K} \approx M^{-2}(1 - \text{Pr}\,(\text{Frag.})), \hat{F} = M^{-1}\text{Pr}\,(\text{Frag.})$. We set the probability of fragmentation $\text{Pr}\,(\text{Frag.}) = 0.1$. An example plot of the (n_1, n_2) plane can be seen in Figure 5.1 (left). Clearly, no limit cycle like behaviour appears. On the other hand, sample simulations of this system for the initial condition $(n_1, n_2) = (50, 25)$ (and $n_3 = 0$) can be seen in Figure 5.1 (right), in which oscillations emerge.

This system can also be thought of microscopically as a continuous time Markov process. The state is the number of clusters of each size allowed in the system. Then we generally speaking have two types of transitions between states: coalescence between clusters of sizes i and j to form a cluster of size $i+j$, whose rate we will write as $\text{Coal.}_{i,j,i+j}$, and shattering fragmentation of a cluster i into i monomers, whose rate we will write as $\text{Frag.}_{i,1}$. We can write the 'reactions' and their transition rates then as

$$\begin{gathered}
n_i + n_j \overset{K(i,j)}{\rightarrow} n_{i+j}, \\
\text{Coal.}_{i,j,i+j} = K(i,j)n_in_j = \left(\frac{1-\text{Pr}\,(\text{Frag.})}{2}\right)\left(\frac{in_i}{M}\frac{j-\delta_{i,j}}{M-i} + \frac{jn_j}{M}\frac{i-\delta_{i,j}}{M-j}\right)
\end{gathered} \tag{5.10}$$

and

$$n_i \overset{F(i)}{\rightarrow} in_1, \quad \text{Frag.}_{i,1} = F(i)n_i = \text{Pr}\,(\text{Frag.})\,\frac{in_i}{M}. \tag{5.11}$$

Each of the rates incorporates the probability of an event of the appropriate type and the probability of choosing clusters of the appropriate sizes (in_iM^{-1}). There is a (small for most situations we are concerned with) correction to account for not being able to choose the same cluster twice from the population in the coalescence terms as well. Technically, this represents a deviation from a strict multiplicative kernel, but the assumption of a steady-state (about which we are studying the oscillations) enforces that $i,j \ll M$, making the results comparable.

The oscillations can be averaged out or analysed by using the transition rates to write a prob-

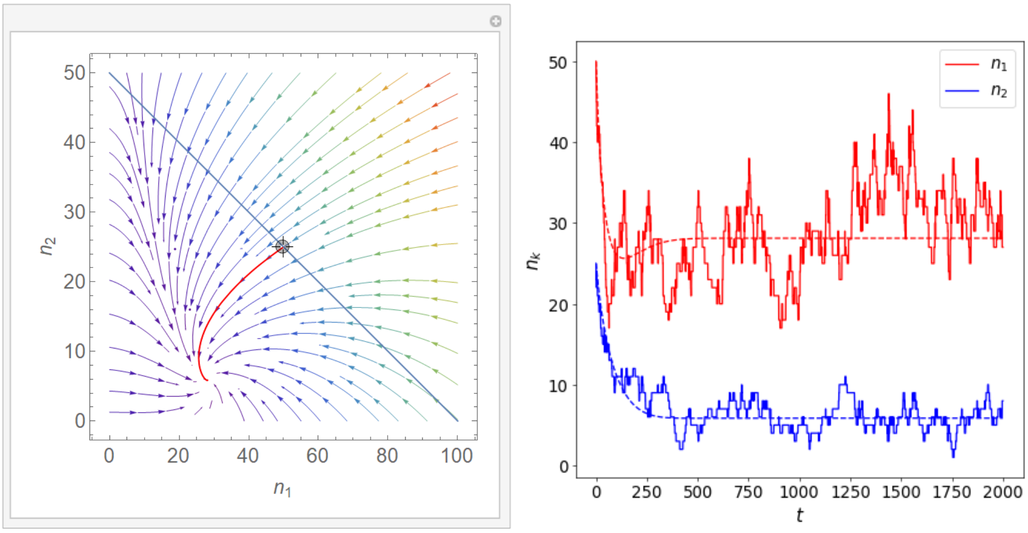


Figure 5.1: Comparison of the dynamical system to stochastic realisation for $\kappa = 3$. The dynamical system is as in Equations (5.9). (Left) The system is evaluated in Mathematica (Wolfram Research, Inc., 2019) in a stream plot. The diagonal blue line is the border of the physical region, while the dot is centred at the initial condition, with the solid red curve representing its trajectory over time. (Right) We present a stochastic realisation, created with `GillesPy2`, a Python implementation (Python Software Foundation, 2020) of Gillespie's stochastic simulation algorithm (Gillespie, 1977; Abel *et al.*, 2016). Dashed lines correspond to the trajectory in the left plot, while solid lines represent the values actually attained by a simulation. We take $\Pr(\text{Frag.}) = 0.1$ and $M = 100$ with initial condition $(n_1, n_2) = (50, 25)$.

abilistic master equation and utilising the van Kampen system size expansion (McKane and Newman, 2005). Due to space, we do not present the full equations here for either System (5.9) or general coalescence and fragmentation. The method is theoretically straightforward and can be thought of as a power series expansion of the master equation in terms of the size of the system. The system proceeds under the assumption that changes in the state of the system are small for large values of the system size (or other appropriate parameter) and that fluctuations are centred around the (assumed to exist) steady-state. These fluctuations also are assumed to be of a magnitude proportional to the square root of the system size parameter. This ansatz on the relationship between the fluctuations, the steady-state, and the system size is "the essential step". Additionally, the problem becomes more difficult if the steady-state is not (globally) stable, which we do not consider here. If fluctuations grow in time, the analysis naturally fails when they are of the same order as the steady-state (van Kampen, 1992).

Define the maximum allowed cluster size to be κ. Following van Kampen (1992), introduce first $P(\mathbf{n}, t)$, the probability of being in state $\mathbf{n}$ at time t, governed by the probabilistic master equation

$$\dot{P} = \sum_{i=1}^{\kappa-1} \sum_{j=1}^{\kappa-i} \left(\mathcal{E}^{i,1}\mathcal{E}^{j,1}\mathcal{E}^{i+j,-1} - 1\right) \text{Coal.}_{i,j,i+j} P + \sum_{i=1}^{\kappa} \left(\mathcal{E}^{i,1}\mathcal{E}^{1,-i} - 1\right) \text{Frag.}_{i,1} P \tag{5.12}$$

where step operators $\mathcal{E}^{i,j}$ indicates a change in index i by amount j in the value of $\mathbf{n}$ in the argument of P, $\text{Coal.}_{i,j,i+j}$ represents the rate of coalescences of clusters of sizes i and j into clusters of size $i + j$, and $\text{Frag.}_{i,1}$ represents the rate of shattering of clusters of size i into monomers. The groups of step operators represent entering the current state, while the subtraction of 1 represents leaving the current state with the same type of event. The central ansatz is that the system state $\mathbf{n}$ can be written in terms of the normalised mean-field solution $\boldsymbol{\phi}$ and fluctuations $\boldsymbol{\epsilon}$:

$$\mathbf{n} = M\boldsymbol{\phi} + \sqrt{M}\boldsymbol{\epsilon}. \tag{5.13}$$

The step operators can then be expanded in terms of the system size M:

$$\mathcal{E}^{i,\pm 1} = 1 \pm M^{-1/2}\partial_{\epsilon_i} + M^{-1}(2!)^{-1}\partial^2_{\epsilon_i} + \dots \tag{5.14}$$

Since the $\boldsymbol{\phi}$ solve the deterministic system, we are more interested in studying the fluctuations $\boldsymbol{\epsilon}$ for which we write $P(\mathbf{n}, t) = P(M\boldsymbol{\phi} + \sqrt{M}\boldsymbol{\epsilon}, t)$ =$\Pi(\boldsymbol{\epsilon}, t)$, whose time derivative is related to P:

$$\dot{P} = \partial_t \Pi - M^{1/2} \sum_{i=1}^{\kappa} \frac{\mathrm{d}\phi_i}{\mathrm{d}t} \partial_{\epsilon_i} \Pi. \tag{5.15}$$

Upon performing these substitutions and expansions, terms associated with $M^{1/2}$ solve the deterministic system.

Terms associated with M^0 can then be collected and govern the (highest-order) perturbations to the deterministic system (assuming $\kappa \ll M$). They constitute a Fokker-Planck equation of the form

$$\partial_t \Pi = -\sum_{i,j} A_{i,j}(t)\partial_{\epsilon_i}\{\epsilon_j \Pi\} + \frac{1}{2}\sum_{i,j} B_{i,j}(t)\partial_{\epsilon_i}\partial_{\epsilon_j}\{\Pi\} \tag{5.16}$$

where A corresponds to the Jacobian of the deterministic system and B is symmetric and non-negative semi-definite (van Kampen, 1992). To study the oscillations using Fourier analysis, we follow McKane *et al.* (2007) in converting the Fokker-Planck equation to a Langevin equation

$$\dot{\boldsymbol{\epsilon}} = A\boldsymbol{\epsilon} + \boldsymbol{\eta}(t) \tag{5.17}$$

where $\boldsymbol{\eta}$ is multivariate zero-mean Gaussian noise with covariances

$$\langle \boldsymbol{\eta}(t)\boldsymbol{\eta}^\dagger(t')\rangle = B\delta(t - t'). \tag{5.18}$$

We use † to denote the conjugate transpose. Neglecting proportionality constants, this allows us to apply the Fourier transform, with frequency power spectrum

$$W(\omega) = \mathrm{Tr}\left((-\mathrm{i}\omega\mathbb{1} - A)^{-1}B(-\mathrm{i}\omega\mathbb{1} - A)^{\dagger^{-1}}\right). \tag{5.19}$$

(Here, $\mathbb{1}$ indicates the identity matrix and i, the imaginary unit.)

We note a few details before discussing the results. Due to the de-meaning in Fourier analysis, the zero frequency signal should have (near) 0 power and we exclude it in what follows to not distort the plots. Additionally, conducting the above analysis for System (5.9) induces a singular matrix A, since $\phi_1 + 2\phi_2 + 3\phi_3 = 1$. Using this relationship to reduce the size of the matrix removes not only the singularity, but also the direct evaluation of fluctuations in ϵ_3, so we do not do so here. There is also a proportionality factor missing, so when presenting the power spectrum directly we scale both the empirical data and theoretical values by the area under the curve (in the former case, integration between $(-0.5, 0.5)$, but $(-\infty, \infty)$ in the latter). Each of our simulations is created by tracking the number of clusters for each size and applying one of coalescence or

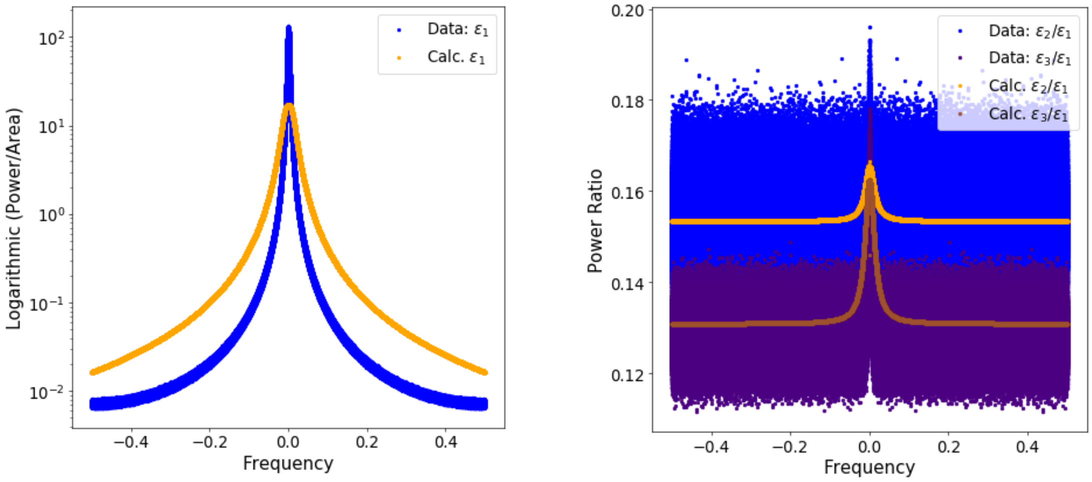


Figure 5.2: Fourier analysis of empirical and theoretical coalescence and fragmentation for $\kappa = 3$, $M = 100$, and $\Pr(\text{Frag.}) = 0.1$, using the average of 1,000 simulations. (Left) Power spectra for ϵ_1, empirical ('data') in blue with theoretical ('calculated') in orange. Both spectra are rescaled by their area and plotted on a logarithmic power (y) axis. The shape is broadly correct, but the shoulders are higher for the theoretical spectra, possibly due to simulation constraints and truncation in the theoretical calculation. (Right) Ratios of power spectra for ϵ_2 and ϵ_3 to that of ϵ_1. Good agreement is observed over the majority of frequencies, with divergences occurring for very low frequencies. Variations in the empirical results are over a narrow range and can be reduced with more trials.

fragmentation each discrete time-step using $M = 100, \Pr(\text{Frag.}) = 0.1$; impossible coalescences fail to occur. This is in contrast to the stochastic simulation algorithm used in Figure 5.1, which attempts to simulate continuous time directly. For a single simulation the results of Fourier analysis are quite noisy, so we have averaged the power spectra of 1,000 simulations in order to better show the results.

In Figure 5.2 (left), we contrast the theoretical power spectrum with the average power spectrum for the monomer fluctuations on a logarithmic power axis. Other power spectra are similar in appearance. Broadly speaking, the calculated fluctuations in ϵ_1 have broadly the same shape, but different shoulders (at high frequencies) and different maximum powers (at low frequencies). On the other hand, we also examine the relationships between the (unnormalised) fluctuations in Figure 5.2 (right) in which we divide the fluctuations of clusters of sizes 2 and 3 by those of monomers. There is rough agreement between the data and the calculated ratios of the fluctuations. The median result for the data is 0.154 for the ratio of the power spectra of ϵ_2 to that of ϵ_1 and 0.128 for ϵ_3 to ϵ_1 with 95% intervals of $[0.141, 0.167]$ and $[0.121, 0.137]$ respectively. The theoretical ratio evaluated at approximately the same frequencies has median ratios of 0.153 and 0.131 respectively.

Generally the results of Figure 5.2 indicate that there is modest agreement between the empirical and theoretical spectra, but there is strong room for improvement. There are two obvious sources of error. First, there is the truncation of the Fourier analysis, since time and sampling are not truly continuous compared to in the theoretical situation. Second, there is the linear approximation in the theoretical calculation, *i.e.* the truncation of the powers of M. It is not obvious how these errors necessarily interact, but we conjecture that expanding the theoretical calculation would improve the peak near the zero frequency as well as the shoulders of the distribution (due to the area normalisation), while more sampling would improve the shoulders in the data. There is fairly good agreement between the ratios of the power spectra of the fluctuations to each other, which supports the broad assertion that the theory and data match. If the ratios had not agreed, that would have suggested that there were systemic deviations that were not accounted for in the bulk of the theoretical calculation. Instead, deviations occur in the frequencies that additional powers of M would be expected to influence. We would expect improvement in Figure 5.2 as M is increased, since errors from both forms of truncation would be mitigated. Of course, increasing M would also reduce the stochasticity in the system and cause the system to deviate less from the mean, since it would also reduce the effects of the higher powers of M in the theoretical calculation. This in turn risks potentially eliminating the oscillations that we are trying to study here.

While we have presented the proposed steady-state, the analysis in this section assumes that the steady-state exists and is stable in the sense of persistence and attraction of other states. While there are deviations when simulating a discrete system, those deviations should fall off rapidly as M is increased, leaving the system to be absorbed into the proposed steady-state. As discussed in Section 2.4, fragmentation is used to create a steady-state by balancing coalescence, but such a balance has not been guaranteed. In the next section, we examine the existence of cycles that emerge when fragmentation does not balance coalescence, which we term gel-shatter cycles.

5.3 Gel-shatter cycles

Despite this wealth of individual results, a clear understanding of the dynamics underlying the macroscopic power-law steady-state is lacking. As we shall show, cyclicity, beyond what we would normally expect from stochasticity, is a ubiquitous feature of the system, despite the presumed existence of a steady-state. This cyclicity appears in the form of the repetition of steady accumulation followed by sudden shattering.

It is commonly presumed that in the absence of external changes to the system the cluster size distributions emerging from the tension between coalescence and fragmentation necessarily tend to a steady-state. However, it has recently been discovered that this does not need to be the case: perpetual oscillations were observed in the mean-field simulations of a system with a special choice of coalescence and fragmentation (C-F) rules (Matveev *et al.*, 2017). Most recently, it has been shown that temporal oscillations can arise via a Hopf bifurcation in a class of C-F processes where clusters grow or shrink by addition or deletion of monomers (Pego and Velázquez, 2020). Both findings predicted deterministic oscillations and relied on mean-field descriptions.

In this section, we use simulations to show that stochastic cyclical dynamics, dependent on system size, emerge for a broad class of C-F processes. We focus on a particularly important class of multiplicative coalescence and multiplicative fragmentation processes which are widely used in applications. We provide simple scaling arguments to explain the observed dependence of the recurrence time on the system size.

Recall Smoluchowski's equations for C-F, Equation (5.1). In these equations, we have both a coalescence kernel K and a fragmentation kernel F. In this section, we take the multiplicative coalescence kernel, which naturally emerges in describing random network growth: when two nodes are randomly connected by an edge per unit time, two clusters of connected nodes coalesce at the rate $K(i,j) = \hat{K}(i/M)(j/M)$, where $\hat{K}$ is constant and the system size M is the total number of nodes. Similarly, we take the multiplicative fragmentation rate to be proportional to the cluster size, given by $F(i) = \hat{F}(i/M)$, where $\hat{F}$ is constant. This particular case is that of shattering, discussed in Section 2.4. These kernels can be conveniently simulated numerically by picking a random node and, if coalescing, picking another random node and connecting every node in the corresponding clusters or, if fragmenting, removing every connection within the cluster. In essence, this is an alternative formulation of the stochastic system that we posed in Section 5.2, which we can then simulate in either discrete or continuous time.

The typical stochastic dynamics of four such simulations and the emerging cycles are seen in Figure 5.3(a), which shows a sample of the trajectory in the $(k_{\max},N)$ plane, where N is the total number of clusters and $k_{\max}$ is the size of the largest cluster. Each distinct counterclockwise cycle begins in the top-left with a predominance of monomers. A period of gradual growth of $k_{\max}$ is accompanied by a decrease in N, taking the trajectory slowly down and to the right before accelerating rightward through gelation until growth ends in an abrupt random jump back to the top-left with $\Delta N \approx -\Delta k_{\max}$. This clearly evidences that 1) coalescence leads to the formation of very large clusters during the periods of gradual cluster size growth and 2) such periods end in shattering of the largest cluster. However, the cycling dynamics is not limited to the growth and shattering of the largest cluster but involves the whole cluster size distribution. Indeed, our simulations show that the cluster size distribution is very broad and can be fitted to (truncated) power laws with the time-dependent exponent $\alpha(t)$ cycling in the range $2.7 - 2.9$. During the period of coalescence, the cluster size distribution is broadening and α decreases appreciably. To the best of our knowledge, these gel-shatter cycles have not been previously described.

The gel-shatter cycles are strongly dependent on the system size M. Figure 5.4 presents a 'heat map'[25] of the times the simulation spends in different regions of the $(k_{\max},N)$ plane (normalised

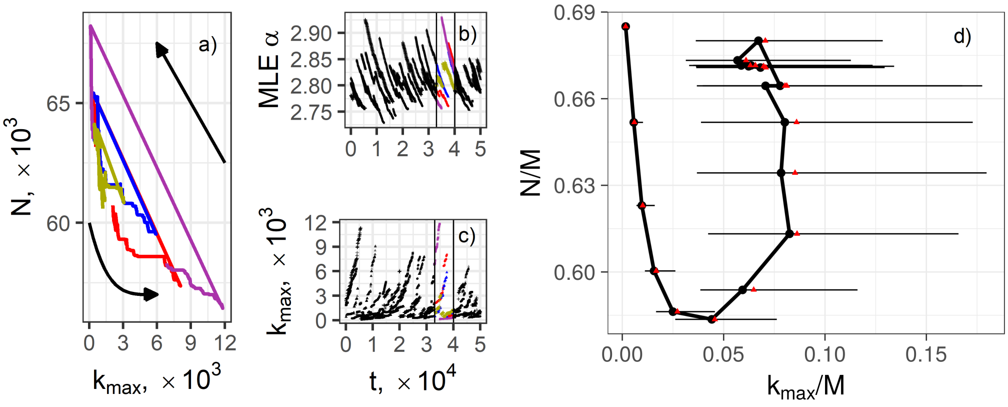


Figure 5.3: Stochastic coalescence-fragmentation cycles. (a) Number of clusters N versus maximum cluster size $k_{\max}$. The counterclockwise trajectory shows periods of relatively gradual growth of the largest cluster followed by its abrupt shattering. (b) The time-dependent power law exponent α obtained by fitting the instantaneous cluster-size distribution to a (truncated) power law using maximum likelihood estimation. The coloured region corresponds to the trajectories shown in (a). (c) Maximum cluster size as a function of time. (d) For comparison, a deterministic numerical evaluation of $M = 3 \times 10^3$ where we plot the normalised N calculated directly from the equations against simulated $k_{\max}$. To simulate $k_{\max}$, we treat the mass as a probability distribution and repeatedly draw cluster sizes until we the total cluster sizes is greater than M. We then discard the last cluster drawn and replace it with the remaining mass needed to reach M. The points are plotted at the median and red triangles are plotted at the mean. The bars represent 95% of results. Here $\hat{K} = 0.99$, $\hat{F} = 0.01$ and the total populations is $M = 10^5$ in plots (a)-(c).

[25] A heat map displays the number of times a value is observed in a two dimensional region, analogous to a two dimensional histogram. Here, darker colours signal more occurrences on a logarithmic scale.

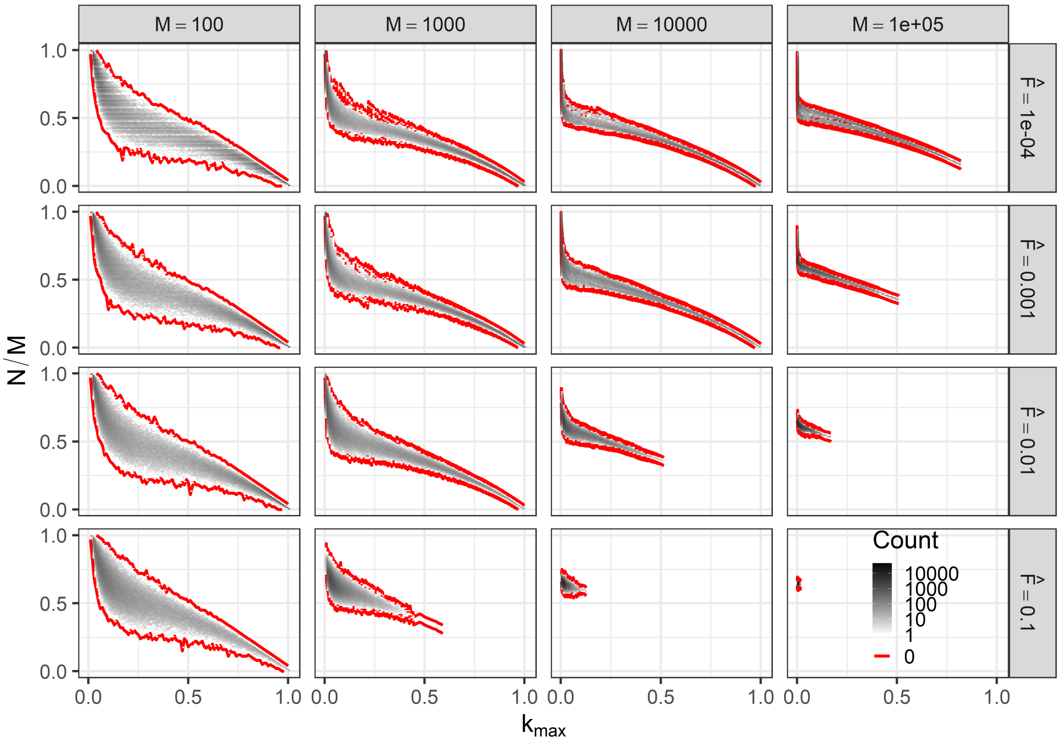


Figure 5.4: Cluster summary statistics. System sizes are $M = 10^2, 10^3, 10^4$, and 10^5 (left to right), and fragmentation rates $\hat{F} = 10^{-4}, 10^{-3}, 10^{-2}$, and 10^{-1} (top to bottom, with coalescence rates $\hat{K} = 1 - \hat{F}$). The number of clusters N and the maximum cluster size k_{max} are normalised by M. Darker regions represent states that emerge more often. The red border line denotes the boundary of the region visited by the system.

by M for comparison between different populations). For low fragmentation rates and small system sizes (top-left corner) the system visits a broad shoulder-like region. The 'elbow', the change in gradient of its lower boundary, corresponds roughly to gelation. As fragmentation rate and system size increase, the broad distribution collapses to a very small region within which the stochasticity is not visible as distinct cycles. This shows that the assumption of a steady-state should not be considered given for all C-F systems. We do not always see the dissipation of dynamics as the system evolves with time due to the rarity and magnitude of fragmentation events, which are averaged over in the mean-field system. Furthermore, the necessary components appear to only be the existence of gelation and sufficiently strong fragmentation, as we shall observe. Note that, due to the natural coupling between the system size and the maximum cluster size as well as the lack of the apparent steady-state and persistence with large system size, the methods employed in Section 5.2 cannot analyse these cycles. (That is, such methods might be applicable when $M = 10^5$ and $\hat{F} = 0.1$, but not when $\hat{F} = 10^{-4}$, compare and contrast in Figure 5.4.)

The mean recurrence time $\langle t_{\text{rec}} \rangle$ is defined to be the number of computational steps between successive shatterings of the largest cluster, averaged across simulations with the same parameters. This can be thought of as the average duration of a cycle in the gel-shatter regime. The dependence of the mean recurrence time $\langle t_{\text{rec}} \rangle$ on system size M for varying fragmentation rates $\hat{F}$ shows crossovers between three distinct regimes: for small systems and small fragmentation rates, $\langle t_{\text{rec}} \rangle \sim M^0$ while modestly increasing M or $\hat{F}$ yields $\langle t_{\text{rec}} \rangle \sim M^{1/2}$, before crossing over into a regime with superlinear growth of $\langle t_{\text{rec}} \rangle$ with M. Remarkably, these curves, *i.e.* plotting functions of $\langle t_{\text{rec}} \rangle$ against functions of M, show data collapse when $\hat{F}\langle t_{\text{rec}} \rangle$ is plotted as a function

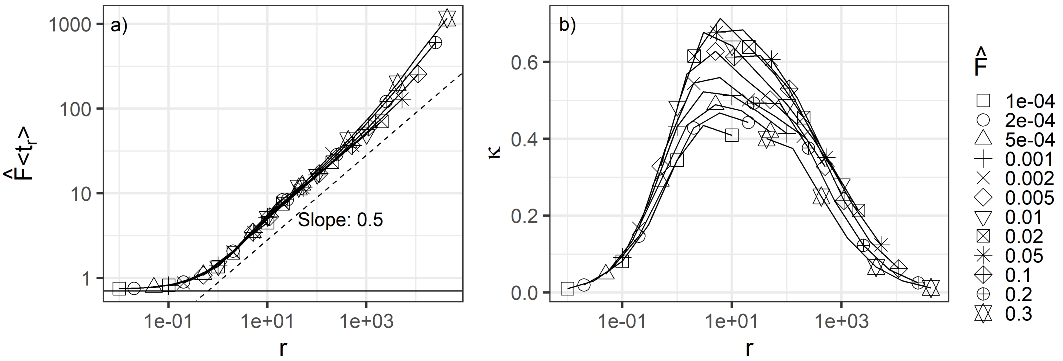


Figure 5.5: Observed data collapse in terms of r. Points are plotted for $M = 10^2, 10^3, 10^4$ and 10^5, but data were additionally gathered at $M = 3 \times 10^2, 3 \times 10^3$ and 3×10^4. (a) When plotting the mean recurrence time $\langle t_{\text{rec}} \rangle$ multiplied by the fragmentation rate $\hat{F}$, we see that the middle region of unforced gel-shatter cycles has strong data collapse following a nearly linear trend on log-log axes. (b) When plotting the order parameter $\mathcal{K}$, we observe a plateau of $\mathcal{K}$ in the region of unforced gel-shatter cycles.

of the dimensionless parameter

$$r = \frac{\hat{F}M}{\hat{K}}. \tag{5.20}$$

The physical meaning of r is the ratio of the characteristic times of gelation $T_g \sim M/\hat{K}$ and shattering $T_f \sim 1/\hat{F}$. Figure 5.5(a) demonstrates that

$$\langle t_r \rangle = \hat{F}^{-1} g(r), \tag{5.21}$$

where the scaling function $g(r) \sim 1$ for $r \ll 0.1$. For larger values of r, the scaling function crosses over to $g(r) \sim r^{1/2}$ and this behaviour persists for almost four orders of magnitude of r. Finally, for $r \gtrsim 10^3$, the scaling breaks down.

In order to explore further the nature of the stochastic gel-shatter cycles and distinguish cyclical dynamics from acyclical stochastic fluctuations, we introduce the cyclicity order parameter $\mathcal{K}$ as the ratio of the number of computational steps that result in growth of $k_{\max}$ minus those that result in loss to the total number of computational steps. Large $\mathcal{K}$ indicates many steps of growth followed by abrupt collapse, characterizing gel-shatter cycles, and in contrast to generic stochastic fluctuations which have no preferred sense. The order parameter $\mathcal{K}$ depends non-trivially on r, see Figure 5.5(b). Here $\mathcal{K}(r)$ forms hump-shaped curves with the maximum around $r = 10$. The order parameter is appreciably large, $\mathcal{K} > 0.1$, for $0.1 < r < 10^3$, which constitutes the middle regime with $\langle t_r \rangle \sim M^{1/2}$.

We can now identify the different scalings as three physical regimes:

(i) **Weak fragmentation or forced cycles.** Here $T_g \ll T_f$ and coalescence quickly results in a single gel cluster. The gel is shattered at the rate $\hat{F}$ and hence the emerging cycles have the mean recurrence time $\langle t_r \rangle \sim \hat{F}^{-1}$. This regime is physically unsurprising and we are not the first to note it (Ruszczycki *et al.*, 2009, p. 301).

(ii) **Unforced gel-shatter cycles.** Here $T_g \sim T_f$ and the cycles arise from a non-trivial interplay between the dynamics of gelation and shattering. The cyclicity $\mathcal{K}$ reaches a maximum in this regime. The presence and extent of this regime in gelling C-F systems is our main finding.

(iii) **Fragmentation-dominance.** Here $T_g \gg T_f$ and fragmentation is so strong as to preclude formation of large clusters. This annihilates the gel-shattering cycles, leaving only stochastic variation in a small region.

We now provide a simple heuristic scaling argument to explain the behaviour of the mean recurrence times and gain further insight into the gel-shatter cycles.

Assume the dynamics are dominated by continuous growth and shattering of the largest cluster. Let the growth of the largest cluster due to coalescence be approximated by $v(t)$. The probability $P(t)$ that the largest cluster is not shattered by time t is governed by

$$\frac{dP(t)}{dt} = -\hat{F}M^{-1}v(t)P. \tag{5.22}$$

Hence the probability density for shattering at time t is

$$p(t) = -\frac{dP(t)}{dt} = \hat{F}M^{-1}v(t)e^{-\hat{F}M^{-1}\int_0^t v(\tau)d\tau}. \tag{5.23}$$

In the forced-cycles regime (i), coalescence quickly leads to a single large cluster of size M, hence $v(t) = M$ and

$$p(t) = \hat{F}\exp\left(-\hat{F}t\right).$$

This is the exponential distribution with mean $\langle t_{\text{rec}}\rangle = \hat{F}^{-1}$; thus $g(r) \sim 1,\ r \ll 1$.

In the unforced gel-shattering cycles regime (ii), assume linear growth $v(t) = c\hat{K}t$, where $c > 0$. Then

$$p(t) = c\hat{F}\hat{K}M^{-1}te^{-c\hat{F}\hat{K}M^{-1}t^2/2}.$$

This is the Rayleigh distribution with scale parameter $\left(\sqrt{c\hat{F}\hat{K}M^{-1}}\right)^{-1}$ and mean

$$\langle t_{\text{rec}}\rangle = \sqrt{\pi M/(2c\hat{F}\hat{K})}. \tag{5.24}$$

Hence, $g(r) \sim r^{1/2}$ in regime (ii). Also, it is straightforward to show that the largest realisable cluster size scales as

$$\frac{v_{max}}{M} \sim r^{-1/2}.$$

This scaling explains the shortening of the 'elbow' on Figure 5.4 for larger M and $\hat{F}$.

In this section, we have described the distinctive new phenomenon of gel-shatter cycles. We have studied how they emerge as a result of imperfect balance of otherwise independent coalescence and fragmentation processes and provided statistics to characterise the transition from forced cycles through stochastic unforced gel-shatter cycles to the steady-state anticipated in the literature. Crucially for applications, these cycles do not emerge in the mean-field limit. In the next section, we examine the robustness of the steady-state and gel-shatter cycles to explore their ubiquity in the presence of alterations to the underlying rules of the system.

5.4 Robustness

In the previous section, we characterised different regimes of coalescence and shattering fragmentation systems with multiplicative kernels. We write the binary coalescence kernel as $K(i,j) = \hat{K}(i/M)(j/M)$ and the shattering fragmentation kernel as $F(i,k) = \hat{F}(i/M)\delta_{k,1}$, where $\hat{K}$ and

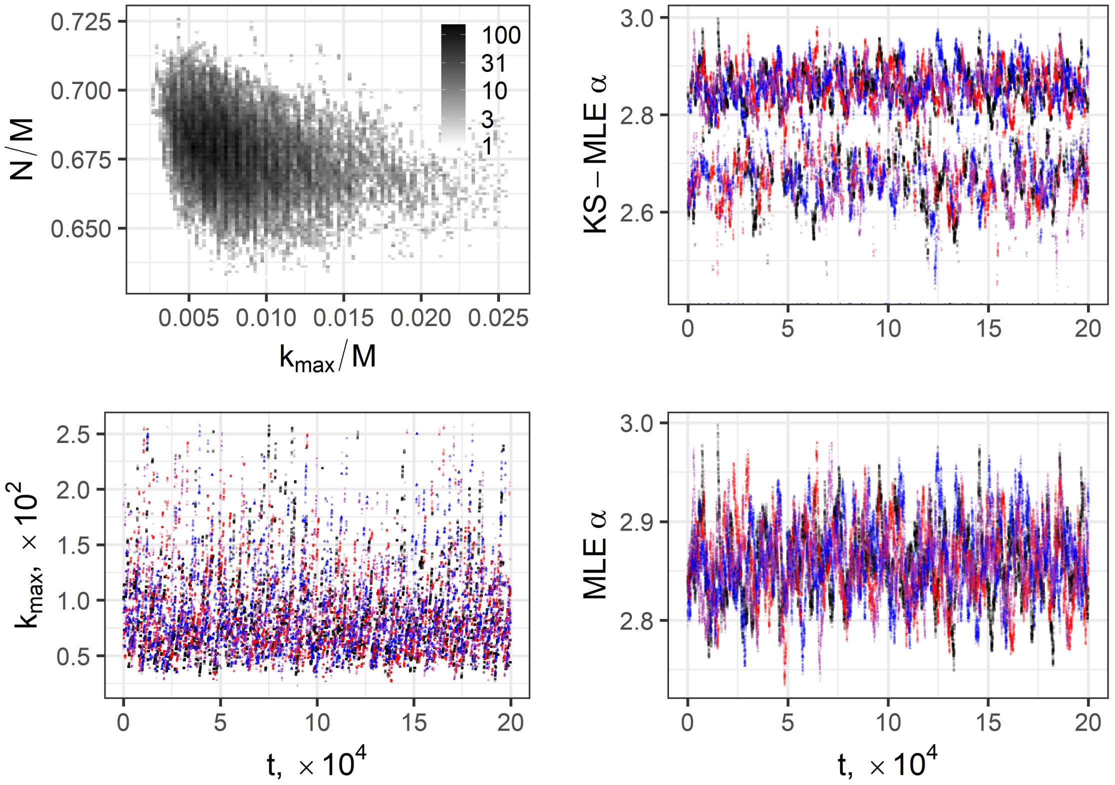
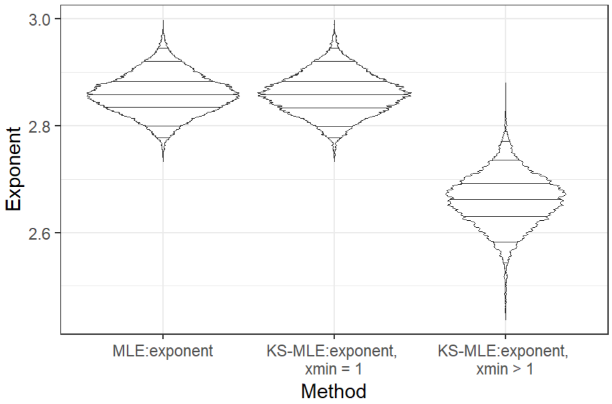


Figure 5.6: Coalescence and fragmentation summary statistics, $M = 10^4$, $\Pr(\text{Frag.}) = 0.20$. Four simulations were conducted (black, blue, red, and purple). Beginning at the top-left and proceeding counter-clockwise, we present a heatmap of the locations of the simulations in the $(\frac{k_{\max}}{M}, \frac{N}{M})$ plane, the time series of $k_{\max}$, the time series of the MLE α estimate, and the time series of the KSMLE α estimate. Note we have trimmed very large (> 5) and very small (< 1) estimates of the KSMLE α; these correspond to failures to fit the system.

$\hat{F}$ are reaction rates, i, j, and k are cluster sizes, and M is the total system size. We were able to characterise the different regimes using a dimensionless parameter $r = \hat{F}M/\hat{K}$, corresponding to the ratio of time to gelation to time to shatter, and cyclicity order parameter $\mathcal{K}$, which is the average of the sign of the change in the largest cluster size $k_{\max}$ over time. The three regimes we characterised were steady-state, in which the system exhibited slight stochasticity in a narrow region, stochastic gel-shatter cycles, in which the system alternated between growth over a long period that randomly ends in shattering before repeating, and forced gel-shatter cycles, in which the system is dominated by only growth until no more growth can occur, at which point the system shatters and begins anew. In the previous section, these regimes are demonstrated and explored by varying M and $\hat{F} = 1 - \hat{K} \in [0, 1]$, while the underlying rules of the system were untouched. In this section, we take the opposite approach: we fix $M = 10^4$ and take values of $\hat{F}$ that result in two model systems, one that is approximately in steady-state and one that experiences stochastic gel-shatter cycles. We then perturb these model systems by changing the types of coalescence and fragmentation processes used.

Multiple results from the literature suggest that the steady-state size distribution of a system of coalescence and fragmentation is robust to perturbations (*e.g.* D'Hulst and Rodgers, 2000; Ruszczycki *et al.*, 2009; Clauset and Wiegel, 2010; Kyprianou *et al.*, 2018). As such, aside from the impacts on cyclicity, which we will measure using $\mathcal{K}$, we also will attempt to measure the impacts on the proposed power-law steady-state. As discussed in Section 2.1, there are a few



Stat.	MLE	KSMLE	KSMLE
		$x_{\text{min}} = 1$	$x_{\text{min}} > 1$
0.99	2.95	2.95	2.77
0.95	2.92	2.92	2.73
0.75	2.88	2.88	2.69
Mean	2.86	2.86	2.66
0.50	2.86	2.86	2.66
0.25	2.83	2.83	2.63
0.05	2.80	2.80	2.58
0.01	2.78	2.78	2.54

Figure 5.7: Coalescence and fragmentation exponent summary statistics. We present the distribution of estimated exponents for Figure 5.6. Together, these simulations constitute $80{,}000$ samples. (Left) Violin plots compare how the fitted exponents are distributed. (Right) A table showing quantiles and the means for the fitted exponents. The quantiles correspond to the horizontal lines within the violin plots. MLE α and KSMLE α agree when the KSMLE $x_{\text{min}} = 1$. When $x_{\text{min}} > 1$, the KSMLE α decreases by approximately 0.2. The theoretical result indicates a truncated power-law distribution with exponent 2.5.

different ways to fit power-law distributions to data. We focus on the fitted exponent α from two methods, MLE and KSMLE, as well as the empirical largest cluster k_{max}. (We find that the fitted minimum x_{min} is uninformative for discussions of the perturbations.)

In order to characterise the perturbations, we begin by characterising the effectiveness of the MLE and KSMLE estimates of α. To do so, we will identify our sample steady-state system and its parameters and the behaviour of the fitting methods. We then can proceed to describe the influence of various rules perturbations on the steady-state system. After exploring the results for the steady-state system, we can perform our analysis again on a system selected for its gel-shatter cycles.

Multiplicative coalescence and fragmentation kernels are together expected to generate a steady-state that is a truncated power-law distribution with an exponent of (approximately) $\alpha = 2.5$, see Section 2.4. As shown in Section 5.3, we need to be careful about selecting an appropriate C-F system to study, since not all such systems will tend towards a steady-state size distribution. We take $M = 10^4$ and $\Pr(\text{Frag.}) = \hat{F} = 0.2 = 1 - \hat{K}$, which, with size-biased and system size normalised kernels, yields $r = 2.5 \times 10^3 > 10^3$. Hence, this system is expected to have a steady-state due to the frequent fragmentation. The results of four simulations of this system can be seen in Figure 5.6, where we show, top-to-bottom, left-to-right, a heat map of the locations the system visits (after burn-in) in $(k_{\text{max}}/M, N/M)$ space, k_{max} over time, the KSMLE α values over time, and the MLE α values over time. In practice, this system has a small amount of cyclicity $\mathcal{K} = 0.123$ on average across the simulations, which can be seen upon close examination of the time series. In comparison to Figure 5.4, the region explored is far narrower with k_{max}/M within $[0.0023, 0.0554]$ and N/M within $[0.6318, 0.7261]$ across all simulations. Of these simulations, 99% of time steps sampled remained within $[0.0035, 0.0276]$ and $[0.6427, 0.7106]$ respectively, demonstrating the far steadier results. For comparison, $\hat{F}\langle t_{\text{rec}}\rangle \approx 120$. The average cyclicity and 95% intervals for KSMLE and MLE α are collected in the first row of Table 5.1, within which we will collect analogous values for each system and perturbation used.

The first question is whether the deviations in the exponent of our model system are issues

Perturbation	Cyclicity $\mathcal{K}$	KSMLE ($x_{\min} > 1$)	MLE
Steady-state	0.123	$[2.568, 2.754]$	$[2.788, 2.932]$
Accretion (3, Uniq.)	0.550	$[2.246, 2.685]$	$[1.907, 2.135]$
Attrition (3, Uniq.)	0.010	$[2.849, 3.047]$	$[3.122, 3.229]$
Accr. and Attr. (3, Uniq.)	0.251	$[2.341, 2.769]$	$[2.105, 2.256]$
Accretion (3)	0.557	$[2.248, 2.682]$	$[1.906, 2.136]$
Attrition (3)	-0.010	$[2.118, 4.848]$	$[3.528, 3.635]$
Accr. and Attr. (3)	0.168	$[2.468, 2.806]$	$[2.178, 2.315]$
Coal. Stick-breaking	0.094	$[2.534, 2.768]$	$[2.725, 2.829]$
Frag. Stick-breaking	0.412	$[1.890, 2.166]$	$[2.241, 2.402]$
Coal. and Frag. Stick-breaking	0.236	$[2.208, 2.393]$	$[2.298, 2.369]$
CRP $\theta = 1.20$	0.436	$[2.075, 2.257]$	$[2.274, 2.376]$
CRP $\theta = 1.50$	0.246	$[2.299, 2.426]$	$[2.434, 2.511]$
CRP $\theta = 1.80$	0.156	$[2.459, 2.620]$	$[2.640, 2.734]$

Perturbation	Cyclicity $\mathcal{K}$	KSMLE ($x_{\min} > 1$)	MLE
Gel-shatter	0.489	$[2.474, 2.829]$	$[2.656, 3.163]$
Accretion (3, Uniq.)	0.725	$[2.248, 2.957]$	$[1.559, 2.800]$
Attrition (3, Uniq.)	0.169	$[2.615, 2.991]$	$[2.970, 3.316]$
Accr. and Attr. (3, Uniq.)	0.567	$[2.370, 2.884]$	$[1.960, 2.710]$
Accretion (3)	0.718	$[1.603, 2.849]$	$[2.261, 3.043]$
Attrition (3)	-0.064	$[2.904, 3.191]$	$[3.276, 3.435]$
Accr. and Attr. (3)	0.496	$[2.420, 2.883]$	$[2.062, 2.673]$
Coal. Stick-breaking	0.497	$[2.387, 2.626]$	$[2.495, 2.850]$
Frag. Stick-breaking	-0.104	$[0.686, 1.563]$	$[0.902, 1.440]$
Coal. and Frag. Stick-breaking	0.902	$[2.216, 2.984]$	$[2.225, 2.650]$
CRP $\theta = 1.20$	0.396	$[0.979, 1.838]$	$[1.027, 1.956]$
CRP $\theta = 1.50$	0.888	$[1.875, 2.519]$	$[2.357, 2.964]$
CRP $\theta = 1.80$	0.806	$[2.362, 2.609]$	$[2.588, 2.799]$

Table 5.1: Summary of robustness results in coalescence and fragmentation. In all cases, $M = 10^4$ and four simulations were used. For steady-state cases, $\Pr(\text{Frag.}) = 0.20$. For gel-shatter cases, $\Pr(\text{Frag.}) = 0.01$. To measure cyclicity, we use the order parameter $\mathcal{K}$ from Section 5.3 averaged over the simulations. The ranges given for KSMLE and MLE estimates of power-law α contain 95% of empirical results. Perturbations used are explained in the text. $\mathcal{K} > 0.2$ is characteristic of gel-shatter cycling.

For steady-state, without failures to fit (KSMLE $\alpha < 0.5$ or $\alpha > 5$) removed, for KSMLE we instead have $[0, 2.759]$. Attrition (3, unique) and attrition (3) instead take $[0, 18.21]$ and $[0, 14.545]$ respectively. Accr. and attr. (3) instead takes $[0, 2.807]$. Coal. stick-breaking instead takes $[0, 2.735]$ and coal. and frag. stick-breaking takes $[0, 2.390]$. CRP $\theta = 1.50$ instead takes $[0, 2.425]$ and $\theta = 1.80$ takes $[0, 2.623]$.

For gel-shatter cycles, the KSMLE range is $[0, 2.840]$. Attrition (3, unique) and attrition (3) KSMLE instead take $[0, 3.005]$ and $[0, 20.446]$. Coal. stick-breaking KSMLE takes $[0, 2.628]$. Frag. stick-breaking KSMLE takes $[0, 1.492]$ while MLE takes $[0.909, \infty)$. CRP $\theta = 1.20$ KSMLE takes $[0, 1.810]$.

with the underlying system, problems with fitting, or finite size effects. To begin, we make a note about the presence of two regions in the KSMLE α plot in contrast to the MLE α plot. This is due to the sensitivity of the KSMLE α to the calculated $x_{\min}$ for which the power-law distribution begins. The higher values arise when $x_{\min} = 1$ is preferred by the algorithm[26], while lower values arise when $x_{\min} \geq 2$. All cases where KSMLE technique has $x_{\min} = 1$ also have the MLE α equal to the KSMLE α. There is not an obvious relationship between the MLE α and the KSMLE α when the KSMLE $x_{\min}$ is otherwise, although there is still an expected correlation (0.454). The KSMLE α and MLE α are then equal 61.6% of the time. The empirical distributions and summary statistics of the KSMLE α and MLE α can be seen in the violin plots[27] and table of Figure 5.7. Divergence from a normal distribution is only evident in the tails of the distribution.[28]

In the theoretical solution to the coalescence and fragmentation equation, discussed in Sections 2.4 and 5.2, we observed a truncated power-law with exponent 2.5 for multiplicative kernels. Here we have values that are mostly higher: only 3.92% of cases have KSMLE below 2.5 and, of those, 97.0% have values below 1, suggesting that there was a failure to fit a power-law distribution instead of a normal result. The minimum value of the MLE, on the other hand, is 2.73. This is not an uncommon occurrence, as can be seen in Figure 5.8 where we reproduce the violin plots of Figure 5.7 for varying M. The empirical distribution appears to collapse to a narrow set of values as M increases, but these values are distinctly above the theoretical 2.5 even for quite large populations of $M = 10^5$. The only time these values really appear to coincide is for $M = 10^3$, for which $r = 250$ and we thus have gel-shatter cycles, suggesting that the model is not, in fact, in steady-state.

This sort of difference between the fitted exponent and the theoretical exponent is actually not surprising. A simple experiment to demonstrate this is to truncate and normalise the solution to the theoretical coalescence and fragmentation equation and then simulate partitions from it. Here, we do so by repeatedly

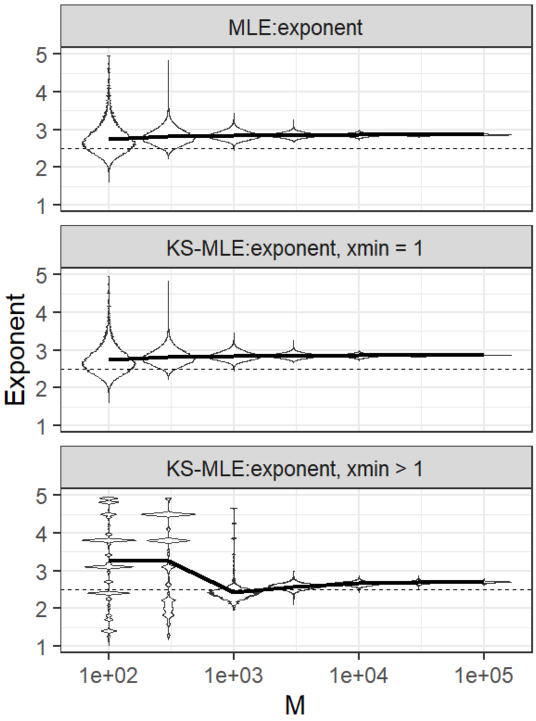
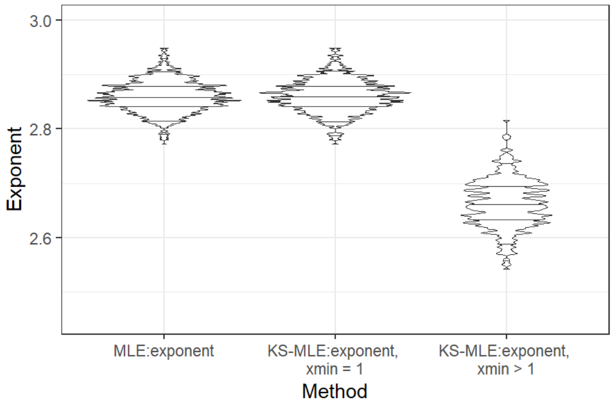


Figure 5.8: Coalescence and fragmentation exponent summary statistics by system size. We fix the probability of fragmentation to $\Pr(\text{Frag.}) = 0.20$ and vary the system size M. For each system size, we simulate the system four times and fit exponents using the MLE and KSMLE methods before displaying the distributions of the fitted exponents using violin plots. The thick black line additionally represents the mean in each case. In each case, as M increases, we see a reduction of variance and a plateau to estimates above the theoretical 2.5. Note that $r = \frac{\hat{F}M}{\hat{K}}$ indicates that we expect gel-shatter cycles for the simulations $M \leq 3 \times 10^3$ as opposed to steady-states.

[26] Note that, as discussed in Section 2.1, that the MLE for the minimum of a distribution is merely the minimum of the dataset and so we do not see split behaviour when applying the MLE method.

[27] In a violin plot, empirical densities of (subsets of) the data are fitted and then plotted adjacent to each other. The densities fitted are usually mirrored about the axis they are fitted on, giving a shape similar to a violin. The intention is to allow rapid comparison of empirical densities (*e.g.* between dataset subsets).

[28] Determined using quantile-quantile plots, not shown. Comparison is against a standard normal distribution, and divergence is visible to the naked eye beginning two standard deviations away from the mean.



Stat.	MLE	KSMLE	KSMLE
		$x_{\min} = 1$	$x_{\min} > 1$
0.99	2.93	2.93	2.77
0.95	2.91	2.91	2.74
0.75	2.88	2.88	2.70
Mean	2.86	2.86	2.66
0.50	2.86	2.86	2.66
0.25	2.84	2.84	2.63
0.05	2.81	2.81	2.59
0.01	2.79	2.79	2.55

Figure 5.9: Coalescence and fragmentation exponent simulated summary statistics. We present the distribution of summary statistics of $1{,}000$ samples from the theoretical coalescence and fragmentation solution with $M = 10^4$ and Pr (Frag.) $= 0.20$. Compare with Figure 5.7 for which we observe good agreement, suggesting deviations from the expected exponent $\alpha = 2.5$ are systematic.

drawing from the normalised solution until the sum of the cluster sizes drawn would exceed the amount of mass in the system. We discard the last cluster drawn and add the missing mass as an additional cluster. While inexact, for large enough systems the difference should be small. When this procedure is performed $1{,}000$ times, as in Figure 5.9, we observe good agreement with Figure 5.7. A natural consequence is that previous studies that rely on the power-law implementation of the methods of Clauset *et al.* (2009b) without taking account of the truncation in the theoretical solution can expect a positive bias in their result unless they happen to use a population of $M = 10^3$.

There is a naturally large variety of possible variations in the microscopic rules that can be implemented. We will confine our discussion to variations that have interesting effects on the population and that are plausible for a social group to experience. We begin our variations with the addition of Becker-Döring type mechanics in conjunction with the existing size-biased coalescence and shattering (Ball *et al.*, 1986). For a social group, this might represent natural turn-over in which individuals leave autonomously (*e.g.* due to changing beliefs) or join a group without being actively recruited. We refer to these as attrition, the loss of monomers from clusters, and accretion, the gain of monomers to clusters, beyond the existing coalescence and fragmentation structure. On a given time step, attrition and/or accretion can occur to multiple clusters (which we may require to be unique) at a fixed rate. This adds additional transitions to the underlying process:

$$n_i \overset{\mathrm{at}i}{\to} n_{i-1} + n_1 \tag{5.25}$$

and

$$n_i \overset{\mathrm{ac}i}{\to} n_{i+1} \tag{5.26}$$

for attrition and accretion respectively. Note that we do not have the rates at and ac dependent on n_1, but to conserve mass we will deduct from n_1 when accretion occurs. We will assume that both attrition and accretion operate on individuals within groups or clusters such that the processes are size-biased. This can be thought of as a process where the clusters are stable or unstable in contrast to neutrally reacting with their surroundings. Expanding System 5.1 to account for attrition and accretion, we have

$$
\begin{aligned}
\dot{n}_k =& \frac{1}{2}\sum_{i=1}^{k-1} K(i,k-i)n_i n_{k-i} - n_k \sum_{i=1}^{\infty} K(i,k)n_i - F(k)n_k \\
&+ \text{at}((k+1)n_{k+1} - kn_k) + \text{ac}((k-1)n_{k-1} - kn_k), \qquad\qquad k \geq 2, \\
\dot{n}_1 =& \sum_{i=2}^{\infty} iF(i)n_i - n_1 \sum_{i=1}^{\infty} K(i,1)n_i + (\text{at} - \text{ac})\sum_{i=1}^{\infty} in_i - \text{ac}n_1 + 2\text{at}n_2,
\end{aligned}
\tag{5.27}
$$

where we use at and ac to represent the attrition and accretion rates. Due to the dependence of MLE α on $x_{\min} = 1$, fluctuations in n_1 can significantly affect the effectiveness of MLE α, while we expect KSMLE α to be more robust to the effects of accretion or attrition. We have six cases we wish to consider: accretion alone, attrition alone, and both in combination, as well as the requirement that any accretion or attrition be conducted on unique clusters within a time step. As we shall see, requiring that unique clusters experience accretion or attrition will slow down these processes in extreme cases resulting in different final distributions[29]. To discuss whether these perturbations force the system out of steady-state and into gel-shatter cycles, we recall the cyclicity order parameter $\mathcal{K}$. As in Section 5.3, $\mathcal{K}$ is the ratio of the number of computational steps that result in growth of $k_{\max}$ minus those that result in loss to the total number of steps. For our standard system in Figure 5.6, $\mathcal{K} = 0.123$. Values higher than 0.2 are strongly suggestive of unforced cyclicity, while values nearer 0.1 are near the border of cyclic behaviour and steady-state behaviour.

Accretion alone does not depend strongly on whether we require unique clusters or not, but the magnitude of the accretion does have an effect. As predicted, the MLE is more strongly affected than the KSMLE, but both show declines in the exponent predicted. The parts of state space encountered are lower in N/M, but higher in $k_{\max}/M$. Accretion also increases $\mathcal{K}$ to 0.550 for accretion to three unique clusters, suggesting cyclic behaviour, albeit of a different character than that found in Section 5.3. The primary difference is that the accretion system does not as fully disaggregate as observed in gel-shatter cycles. In contrast, attrition does depend strongly on the uniqueness requirement; if uniqueness of clusters chosen to attrit is not required then the largest clusters can more rapidly deteriorate than they can recuperate. This manifests in the complete lack of any cycles and the inability of the system to grow a cluster beyond the size of 30 when attriting 3 monomers per time step without a uniqueness requirement. Requiring uniqueness instead allows dynamics more reminiscent of the original system, but with clusters of size approximately 120 or less when attriting 3 monomers per time step and $\mathcal{K} = 0.010$. Exponents are also larger than in the standard system, suggesting a less aggregated system overall.

Combining accretion and attrition, set to the same rate, does not generate results similar to those found in the original simulations, despite the symmetry in their effects. The number of clusters (monomers included) N decreases substantially as the rates increase, more-so if uniqueness is required, and recorded $k_{\max}$ are larger. While the effects are not as strong as in the pure accretion case, the effects of accretion and attrition at a rate of 9 per time step are similar to those of pure accretion at a rate of 3 per time step. Additionally, some cyclicity appears to be present, *e.g.* $\mathcal{K} = 0.251$ for a rate of 3 per time step with uniqueness. Example violin plots are available in

[29] Not requiring uniqueness would be more similar to a standard Gillespie style implementation. The additional complexity produces variations worth discussing and might reflect reactions of a social group to being spontaneously joined or spontaneously losing members.

Figure 5.10 (a). Altogether, this demonstrates the sensitivity of the model to changes in the sizes of the smallest clusters. For our next set of rules, we consider weakening the coalescence and fragmentation, instead of strengthening them as we did with accretion and attrition.

In the astrophysics literature, collisions can perform both roles of coalescence and fragmentation dependent on the interaction between the particles colliding (Tanaka *et al.*, 1996). While we do not adopt that approach here, we do consider a somewhat similar idea. Instead of coalescence resulting in the larger cluster completely absorbing the smaller one or fragmentation resulting in the complete shattering of a cluster, we instead consider partial variations in which a cluster can break into multiple clusters. In a social group context, this represents a group that does not get completely broken up, *e.g.* if some people come as pairs or if a group otherwise scatters into still organised subunits. In order to represent these variations, we rewrite System 5.1 as

$$\dot{n}_k = \frac{1}{2}\sum_{i=1}^{k-1} K(i,(k-i))n_i n_{k-i} - n_k \sum_{i=1}^{\infty} K(i,k)n_i - F(k)n_k + \sum_{i=k}^{\infty} F(i)b(i,k)\frac{i}{k}n_i, \quad (5.28)$$

where we now allow $k \geq 1$, relying upon $\sum_{i=1}^{0} = 0$, and we use $b(i,k)\frac{i}{k}$ to represent how a fragmented cluster of size i contributes to clusters of size k. The function $b(i,k)$ is normalised so that $\sum_{k=1}^{i} b(i,k) = 1$ for mass conservation during fragmentation. (The factor of $\frac{i}{k}$ assists in converting between the different cluster sizes.) We recover the original system for $b(i,k) = \delta_{1,k}$. The first variation we concentrate on is (discrete) stick-breaking, the process of repeated random division (Pitman, 2006). The implementation is simple: consider a stick, randomly choose a location to break it, and store the left half. Repeat on the right half, finishing when the right half is indivisible. More formally, one might write this process as follows. Take n to be the length of the stick to be broken up and X_i to be the value of the i-th break. Then we sample $X_i \sim \text{Unif}(1 + \sum_{j=1}^{i-1} X_j, n)$ where we write Unif to refer to a discrete uniform distribution on natural numbers. The lengths of the X_i, *i.e.* the differences $X_i - X_{i-1}$ with $X_0 = 0$, then correspond to the sizes of the clusters that were fragmented. We then interpret $b(i,k)i/k$ as the expected number of clusters of size k formed from the fragmentation of a cluster of size i, *i.e.* the expected number of parts of size k in a random partition of i. The result is $b(i,k) = 1/i$. We will show that the system is particularly sensitive to using stick-breaking as a fragmentation variant, but the system is also sensitive to stick-breaking the smaller of the pair coalescing.

The distribution of exponents can be seen in Figure 5.10 (b), but these distributions do not convey the full effect on the system. For the case of stick-breaking fragmentation, we not only have a drop in exponent (the mean drops to 2.03 from 2.66 in our unmodified case for KSMLE α with $x_{\min} > 1$ and to 2.32 from 2.86 for MLE α), but the system has a cluster that contains at least one third of the system size 6.58% of the time (and reaches a high of 5,438). While stick-breaking fragmentation can completely change the outputted distribution, the effects of stick-breaking coalescence are more subtle. There is a slight drop in fitted exponents (mean KSMLE α with $x_{\min} > 1$: 2.43, mean MLE $\alpha = 2.77$) and a drop in the largest $k_{\max}$ (down to 291 from 554). Unlike with stick-breaking fragmentation, where there were obvious visual differences when comparing to the unmodified system's output, only the scales appear to change. Stick-breaking coalescence is also less cyclic, $\mathcal{K} = 0.094$, while stick-breaking fragmentation appears to have transitioned to a cyclic regime, $\mathcal{K} = 0.412$.

Combining both variants in a single system results in an intermediate: exponents are depressed by a lesser amount than for stick-breaking fragmentation alone (mean KSMLE α with $x_{\min} > 1$: 2.10, mean MLE $\alpha = 2.33$), but $\max(k_{\max}) = 703$, more closely resembling (visually and in value) the stick-breaking coalescence case. On the other hand, $\mathcal{K} = 0.236$, indicating more cyclicity than the unmodified and stick-breaking coalescence cases, but substantially less than stick-breaking fragmentation. In both the combined and solely stick-breaking fragmentation cases, the increased $\mathcal{K}$ suggests that more large clusters that would attract fragmentation (which, as a reminder, is size-biased) instead of the strictly largest cluster. Cycling does not, as in the standard case, result in a fully disaggregated system, similar to the observed problem with accretion.

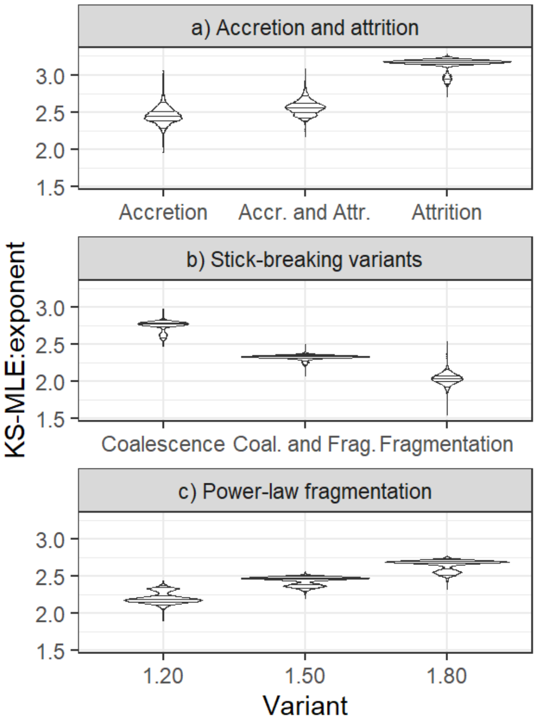


Figure 5.10: Coalescence and fragmentation exponent summary statistics across system variations. We set Pr (Frag.) = 0.20 and $M = 10^4$. For each variant, we simulate the system four times and analyse the KSMLE α for each time step, displayed in violin plots with the same area. Accretion and attrition variants select 3 unique clusters to add or remove a monomer. For stick-breaking coalescence, the smaller cluster chosen to coalesce has the stick-breaking procedure applied. For power-law fragmentation, the exponent θ used in the fragmentation's power-law distribution is listed.

What makes the stick-breaking fragmentation variant so potent? There are a few variants we have used to explore the issue. We have considered cases where a fixed proportion of the cluster chosen to fragment is shattered, ranging from one-tenth to nine-tenths. This does not reproduce such large changes in the distribution however. An alternative way to generalise stick-breaking is to instead look at the uniformity with which the stick is broken. The uniform distribution is a beta distribution with parameters $(1, 1)$, which allows consideration of other parameter combinations. The system is slightly sensitive to this variation. A third generalisation is more rewarding however. It is straightforward to prove that stick-breaking a discrete quantity will provide a bounded power-law distribution with exponent 1 in expectation[30].

In order to generalise the usage of stick-breaking to generate partitions with power-law probabilities, we turn to the Chinese Restaurant Process (CRP)[31] (Pitman, 2006; Goldwater *et al.*, 2011). The CRP allows us to explore the influence of the exponent used in power-law fragmentation, call

[30] Proceed by induction: $n = 1$ yields 1 with probability 1. For $n = 2$, $\Pr(X_1 = 2) = 0.5, \Pr(X_1 = 1) = 0.5$. In the $X_1 = 1$ case, proceed to $n = 1$ and one thus has two 1's. Hence, one expects 0.5 2's and one expects 1 1's. The induction proceeds in the obvious way.

[31] Consider a queue at restaurant of length M. Inside the restaurant are (round) tables which can seat any number of people. The process begins by seating the first person and offering each person in turn an opportunity to sit at any occupied table with probability proportional to the number already sitting at the table or to occupy a new table with probability proportional to a parameter which we fix to 1. The parameter $\theta, 0 < \theta < 1$, is then introduced as a penalty on each occupied table. The penalised mass is then redistributed to a new table (Goldwater *et al.*, 2011). The returned value is the partition of customers grouped by table. As such, if X_i is the assignment of the i-th customer to table X_i and $\mathbf{z}_i$ is the partition of i customers amongst the tables $z_1 \ldots z_{|\mathbf{z}_i|}$, the assembly begins with $X_1 = 1$ and proceeds

it θ, on the exponent of the fitted power-law distribution, α. It is limited to exponents between 1 and 2 however[32]. A natural question is what then is the expected number of clusters of size k generated by CRP fragmentation of a cluster of size i, $b(i,k)i/k$. Adopting the notation of Pitman (2006), we define

$$(x)_{n\uparrow a} := x(x+a)\cdots(x+(n-1)a) = \prod_{i=0}^{n-1}(x+ia).$$

As the size i of the cluster fragmented increases, the expectation of the number of clusters of size k asymptotically scales as

$$b(i,k)\frac{i}{k} \sim \left(\frac{(1+\theta)_{i\uparrow 1}}{\theta(2)_{i-1\uparrow 1}} - \frac{1}{\theta}\right)\frac{\theta(1-\theta)_{k-1\uparrow 1}}{k!}$$

(Pitman, 2006, Eq. 3.13 and Lemma 3.11). Results can be seen in Figure 5.10 (c). In combination with (b) and Figure 5.7, we see a smooth transition as θ increases from 1 to 2 in α from a KSMLE ($x_{\text{min}} > 1$) mean of 2.03 in the stick-breaking procedure to a mean of 2.66 in the original unaltered simulation. Similarly, the cyclicity varies as well: $\mathcal{K} = 0.412$ at $\theta = 1$, 0.436 at 1.2, 0.246 at 1.5, and 0.156 at 1.8.

To summarise, we began by noting and characterising problems in fitting the proposed theoretical power-law and demonstrated that they are endemic to the model considered. Then, while the process of coalescence and fragmentation is considered to be quite robust in the literature (e.g. D'Hulst and Rodgers, 2000; Ruszczycki *et al.*, 2009; Clauset and Wiegel, 2010; Kyprianou *et al.*, 2018), we have identified various perturbations that can substantially change the fitted power-law distribution. Despite the robustness of the KSMLE α of a power-law distribution to perturbations in the minimum value of the fitted distribution x_{min}, adding attrition or accretion of monomers to clusters can drastically change the exponent of the steady-state and even move the system out of the steady-state. We also identified a particular form of rules perturbation, power-law distributed fragmentation, that the system is particularly sensitive to and responds smoothly to. Notably, the perturbed system appears to coincide with the original system when the fragmentation exponent $\theta > 2$, which contributes to the argument for robustness in the literature.

Next, we turn to determining how a gel-shatter system responds to the same type of perturbations that we applied to the steady-state system. We again set a common set of parameters, here $\hat{F} = 0.01$ and $M = 10^4$ so that $r \approx 101$ and we have gel-shatter cycles with $\mathcal{K} = 0.489$. This system is shown in Figure 5.11 and the violin plots of exponents can be seen in Figure 5.12. We have seen that we can induce cyclicity with perturbations, seen from the increase in $\mathcal{K}$ although not necessarily resulting in gel-*shatter* cycles due to the removal of strict shattering and the lack of access to the most disaggregated regions. We also note that using constant and multiplicative kernels together creates imbalances that are reminiscent of extreme forced cycling or steady-states[33]

with

$$\Pr\left(X_i = z_j | \mathbf{z}_{i-1}, \theta\right) = \begin{cases} \frac{|z_j|-\theta}{i} & 1 \leq j \leq |\mathbf{z}_{i-1}| \\ \frac{|\mathbf{z}_{i-1}\theta+1}{i} & j = |\mathbf{z}_{i-1}| + 1. \end{cases}$$

[32] Theoretically, one could use Price's network models to access power-law distributed partitions with exponent greater than 2 (Newman, 2010). In practice, the results did not converge to the expected exponent when partitioning small numbers ($<$ 1,000).

[33] Multiplicative fragmentation without multiplicative coalescence does not result in gelation and prevents the system from growing large enough to begin cycling. Multiplicative coalescence without multiplicative fragmentation results in the system gelling into a single cluster in a manner similar to the forced cycles, with fragmentation only serving

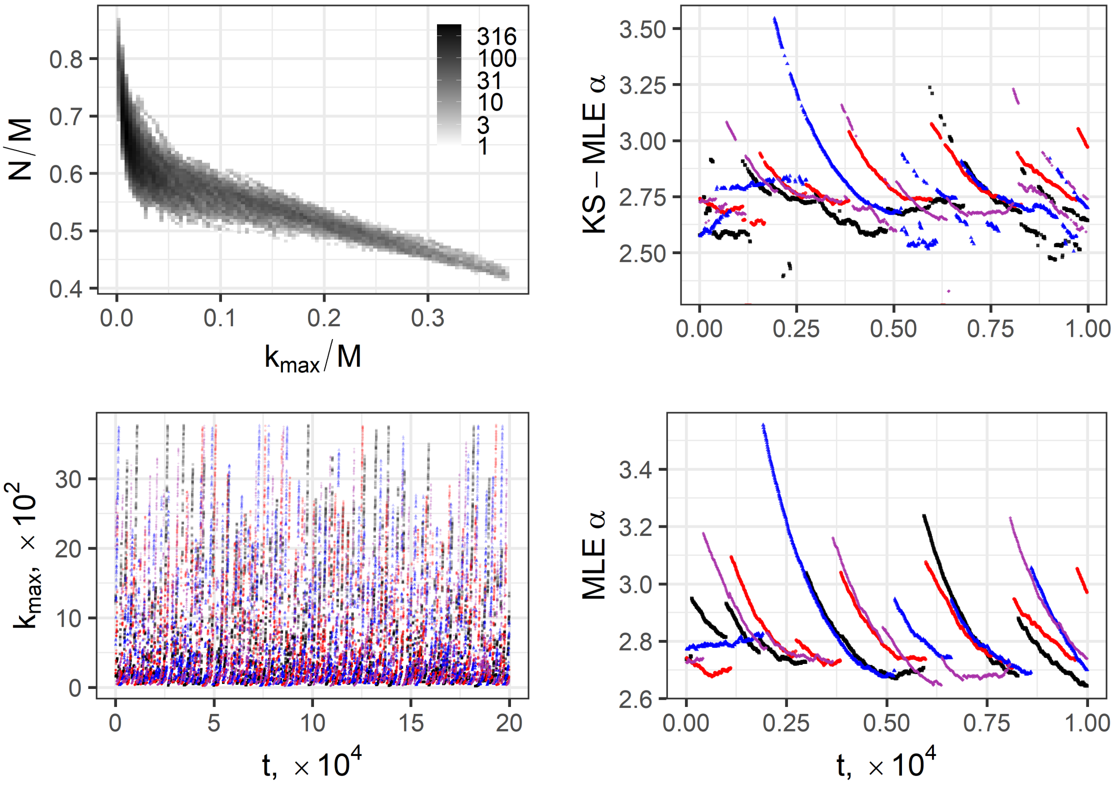
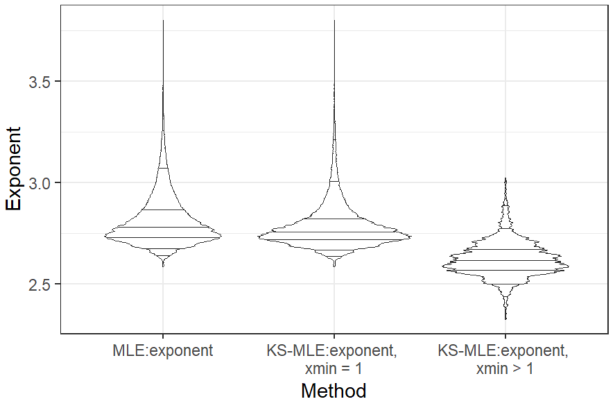


Figure 5.11: Coalescence and fragmentation summary statistics, $M = 10^4$, $\Pr(\text{Frag.}) = 0.01$. Plot types are the same as in Figure 5.6. The time series shown for estimates of the exponent are shorter than those for $k_{\max}$ or those in Figure 5.6 to better show details.



Stat.	MLE	KSMLE $x_{\min} = 1$	KSMLE $x_{\min} > 1$
0.99	3.28	3.21	2.89
0.95	3.08	3.01	2.77
0.75	2.87	2.82	2.67
Mean	2.82	2.79	2.62
0.50	2.78	2.76	2.62
0.25	2.73	2.72	2.57
0.05	2.68	2.68	2.50
0.01	2.64	2.64	2.44

Figure 5.12: Coalescence and fragmentation exponent summary statistics. Conditions are as in Figure 5.7, except $\Pr(\text{Frag.}) = 0.01$ with simulations shown in Figure 5.11. The theoretical result, on the other hand, indicates a truncated power law distribution with exponent 2.5.

while a barrier can be used to reduce the amount of stochasticity in the system. As such, we focus on the same combinations as in the first part of this section: we begin with a discussion of the effects of accretion and attrition, before turning to stick-breaking and power-law fragmentation.

As expected, accretion increases the size of the largest cluster in a cycle, while attrition decreases the same. As accretion increases, cycles become longer and cyclicity increases, from $\mathcal{K} = 0.489$ with no accretion to 0.798 with accretion at 9 monomers per time-step. Attrition with uniqueness heavily reduces cycles, while removing uniqueness outright eliminates them with $\mathcal{K} = 0.154$ and -0.004 respectively at 9 monomers per time-step. In both accretion and attrition, a value of 1 monomer per time-step causes notable changes in the cyclicity parameter (a rise or drop of about 0.2 respectively), but effects on the MLE exponents fitted are more subtle. Divergence from the norm is more obvious in the KSMLE fitted exponents; violin plots for the case with uniqueness and 3 individuals per time-step are presented in Figure 5.13 (a). Accretion and attrition combined results in no change to slight increases in cyclicity without uniqueness with $\mathcal{K}$ ranging from 0.491 (1 monomer per time-step) to 0.506 (9). With uniqueness, more sizeable increases compared to that of pure accretion are observed, with $\mathcal{K}$ now ranging from 0.485 (1) to 0.725 (9) respectively. In contrast to the reduction in the number of clusters N observed before, there is a much wider range of N and k_{max} visited stemming from the system's ability to still gel and shatter. Fitted exponents are similar to those of the accretion case, which are somewhat lower than the standard system. The general observation we make about the MLE fits is that accretion and attrition appear to be able to exaggerate or diminish the observed cycles without making substantial visual changes to the KSMLE fits.

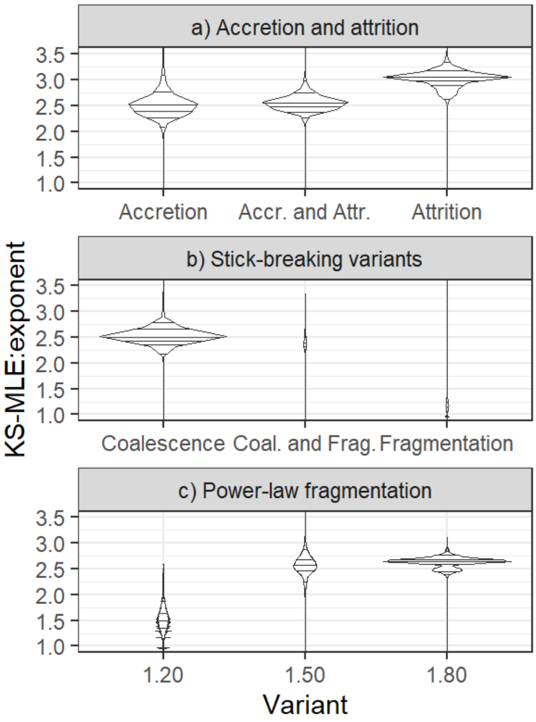


Figure 5.13: Coalescence and fragmentation exponent summary statistics system for variants of a gel-shatter system. The system is as in Figure 5.10, but with $\Pr(\text{Frag.}) = 0.01$. We present only $\alpha \leq 3.5$, but some tails extend beyond this point. Attrition is the most affected (0.41% have $\alpha \geq 3.5$) with a maximum of 57.9. Similarly, we truncate the minimum to 0.5, for which stick-breaking fragmentation is the most affected (8.67% have $\alpha < 0.5$ with a minimum non-zero value of 0.0008). The violins in the stick-breaking cases are barely visible, but present.

Stick-breaking coalescence has very little in the way of obvious effects above and beyond the model system we are studying. There is a decrease in the largest observed cluster size, as well as a decrease in the exponents observed (approximately 0.2 for the center and left tail and 0.3 for the right tail when using MLE and KSMLE with $x_{\text{min}} = 1$ and 0.2 throughout when using KSMLE with $x_{\text{min}} > 1$), but cyclicity is nearly unaffected $\mathcal{K} = 0.497$. Stick-breaking fragmentation is

to reset the system from this state. With both kernels constant, we instead observe a slow build-up of medium size clusters that then rapidly combine together in abrupt competition with fragmentation. This last version could be said to be cyclic due to the kernels' balance and behaviour, but is of a different nature than the *gel*-shatter cycles. Examples for comparison are given at the end of this section in Figure 5.16.

much more potent, even more so than in the steady-state case. The largest cluster sizes observed are those of the entire system as the system struggles to escape the fully gelled state. Whether fitting via KSMLE or MLE, the fitted exponents are strongly different and fairly uniformly sampled[34]. The system clearly no longer experiences gel-shatter cycles: $\mathcal{K} = -0.104$ signifying that the system spends more computational time fragmenting than growing. The apparent paradox of having a low probability of fragmentation yet having fragmentation as the dominant feature of the system is due to the sampling occurring after computational steps.

Combining these two features, stick-breaking coalescence and stick-breaking fragmentation, results in a hybrid system, Figure 5.14, despite the lack of an impact by coalescence beforehand. For example, the largest cluster size now contains approximately 80% of the system size at its largest. This largest cluster is not sufficiently fragmented to cease acting like a gel. It rapidly begins coalescing with the other clusters in the system. Stick-breaking coalescence prevents the immediate return to a fully coalesced state: with each coalescence, the smaller cluster is first subjected to stick-breaking before the larger cluster receives any portion of the smaller. This slows the system down and provides the state-space's unique form compared to those of previous figures. This combination of rare fragmentation, common large clusters, and ineffective coalescence also gives rise to a high cyclicity value $\mathcal{K} = 0.902$, reflecting a very consistent process of growth and disaggregation that is nonetheless distinct from the gel-shatter cycle.

There is also a shift in the values of the fitted exponents compared to stick-breaking coalescence. MLE α fits are reduced by an additional 0.3 for the left tail to 0.2 for the center and right tail. KSMLE α fits, on the other hand, feature a much wider range. For the standard case, the 1st and 99th quantiles were 2.68 and 3.34, but for stick-breaking coalescence they were 2.37 and 2.69 and for stick-breaking both coalescence and fragmentation we instead have 2.12 and 3.09. Combined with our knowledge of the violin plots, Figure 5.13 (b), we conclude that KSMLE α fits are not concentrated around a single value as with stick-breaking coalescence alone, but are more likely to appear in a broader range. The fitted exponents have another deviation from the original process's fits however. Whereas Figure 5.11 has fits that decrease with time before fragmentation forces them upwards, Figure 5.14 shows generally opposite behaviour.

To better understand this behaviour we again turn to other variations of fragmentation. Fixed proportions fragmented do not result in the change in orientation, nor do they appear to escape the gel-shatter cycle so much as remove the disaggregated portions ('top' tail) of the cycle in state-space. Varying the beta distribution used in stick-breaking does not make substantive changes compared to stick-breaking fragmentation. Power-law distributed fragmentation, on the other hand, again interpolates quite well between the observed results for stick-breaking fragmentation (θ near 1) through stick-breaking coalescence and fragmentation ($\theta \approx 1.5$) to the original system (θ above 2). As can be seen by the values fitted in Figure 5.13, (c) in comparison to (b), this is an approximate comparison. Nonetheless, studying the power-law distributed fragmentations for varying θ can give us insight into how the underlying process is behaving. Recall that θ controls the steepness of the power-law distribution of fragments; higher θ results in fewer large fragments and more small fragments. For very low θ, the fragmentation can be undone in very few steps. Performing the CRP $10{,}000$ times for a power-law with $\theta = 1.2$ on a cluster of size $1{,}000$, 52%

[34] Comparison via quantile-quantile plot against a uniform distribution with the same minimum, *e.g.* 0.501 for MLE, and maximum, 1.809, values. Deviations are present in the tails, but a uniform distribution matches the bulk quite well from 0.8 to 1.6. Unfortunately a Kolmorgorov-Smirnov test should not be used due to the presence of ties which presumably arise due to the limited configurations possible.

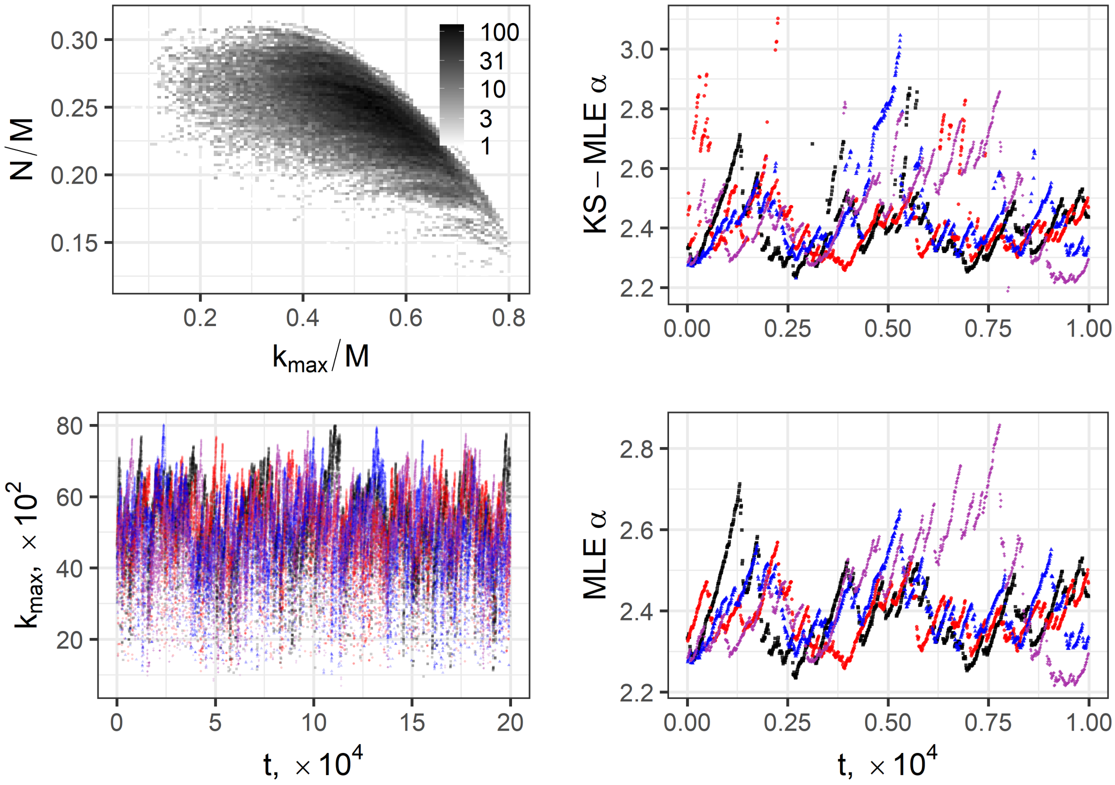


Figure 5.14: Coalescence and fragmentation summary statistics combining gel-shatter cycles with stick-breaking coalescence and fragmentation. The system parameters are as in Figure 5.11, but we have replaced both the coalescence and fragmentation rules with stick-breaking variants, as described in the text.

have a fragment of size 500 or larger and 95% have less than 30 fragments. Increasing θ to 1.5 drops these to 27% and 7.2% respectively. Hence, the CRP for middle values of θ is similar to the case of stick-breaking coalescence and stick-breaking fragmentation together in that there are simultaneously large clusters after fragmentation that require many coalescence events to return to a fully coalesced system. The key difference between the two is reduced to whether large numbers of small clusters are formed during the rare fragmentation event or the common coalescence events.

How then does the transition from power-law distributed fragmentation to standard coalescence and fragmentation gel-shatter cycles occur as θ increases? At $\theta = 1.80$, the KSMLE α fits with $x_{\text{min}} > 1$ are similar to those of stick-breaking coalescence: the 1st and 99th quantiles have 2.35 and 2.66 respectively compared to stick-breaking coalescence's 2.37 and 2.69 with most major divergence happening above the 99th quantile. The tails of the distributions are then fairly similar, but the MLE values are substantially different and there is the aforementioned difference in whether the fitted exponents increase or decrease leading up to the fragmentation. Power-law distributed fragmentation has the 'wrong' direction for the change in the exponents. Increasing θ through to 1.95 reveals the transition that occurs. Thinking about the shape of the exponent curves as 'U' shaped, as θ increases the left arm increases in size, while the right arm decreases in size, with balance appearing to occur around $\theta \approx 1.85$. By 1.90 the fits appear to have been mirrored and at 1.95 the continuity of the fits appears to improve. We provide examples for the MLE fitted exponents at $\theta = 1.80$ and $\theta = 1.90$ in Figure 5.15. KSMLE α fits obey a similar transition, but

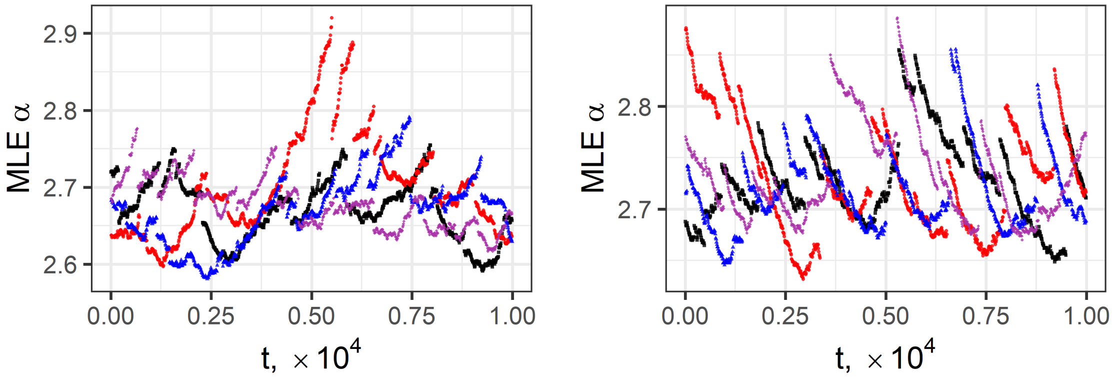


Figure 5.15: Coalescence and fragmentation MLE exponent summary statistics combining gel-shatter cycles with power-law distributed fragmentation, $M = 10^4$, $\Pr(\text{Frag.}) = 0.01$, and $\theta = 1.80$, left, and $\theta = 1.90$, right. Four simulations were conducted (black, blue, red, and purple). Compare with Figures 5.11 and 5.14.

the end results are more damped.

We conjecture that this changeover in proceeding from low-to-high and high-to-low exponents reflects the sensitivity of the fitting procedure to the center of the distribution. For lower θ and low-to-high values of α, more mass will be sent to the middle of the distribution, while for high θ and high-to-low values of α the mass is sent towards the far left tail. In the low-to-high case, that leads to a larger reservoir of medium sized clusters. As coalescence occurs, there is a slight drain on the medium sized clusters, but larger clusters experience a 'large' (relative) growth. As these larger clusters grow without impeding the smaller clusters, the exponent falls. In the high-to-low case though, the largest cluster rapidly depletes the middle, which lacks a suitable reservoir. The largest cluster grows, acting as an outlier to a distribution that increasingly is formed of only small clusters. The exponent then increases to accommodate the small clusters.

In this section, we have discussed the robustness of both the steady-state expected in the literature and the phenomenon of gel-shatter cycles. We began by acknowledging some difficulties from fitting the distribution using the MLE and KSMLE methods of estimating the exponent α and used these difficulties to understand potential bias in the estimators. We then discussed the effects of some simple perturbations to coalescence and fragmentation that can have drastic effects on the resultant detected distribution, including accretion, attrition, stick-breaking, and variant fragmentations. We noted power-law fragmentations for their robust ability to transition between the standard system and stick-breaking fragmentation. We repeated this analysis for a coalescence and fragmentation process in the gel-shatter regime, again observing that each variant can have strong effects. Stick-breaking variants and power-law fragmentation were particularly powerful and could substantially modify the system and its behaviour. In both steady-state and gel-shatter regimes, we saw that $\theta \gtrsim 2$ resulted in a return to the standard systems, indicative of the robustness discussed in the literature. On the other hand, behaviour was drastically different below these points, despite a smooth transition. To conclude this section, we present Figure 5.16, which showcases the changing behaviour in $k_{\max}$ across the variations we have discussed in this section alongside some mentioned elsewhere in this chapter for gel-shatter cycles. This figure highlights some of the robustness and sensitivity of the cycles to the exact rules of the simulation.

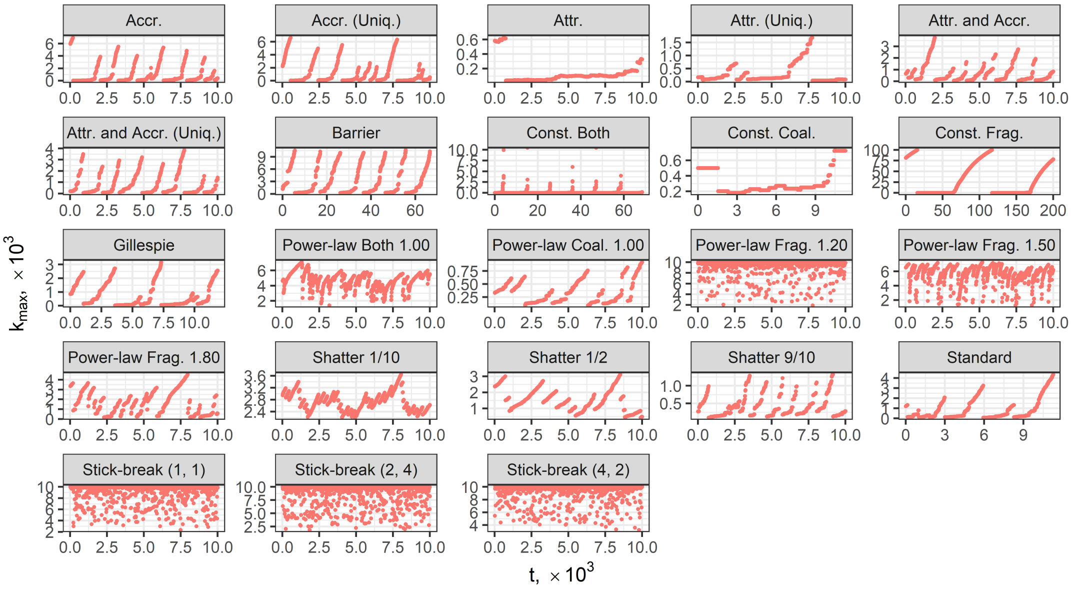


Figure 5.16: k_{max} of variations of the gel-shatter cycles. Each panel represents a single simulation of a single variation. Each simulation has $M = 10^4$ and $\hat{F} = 0.01$ ($\hat{K} = 1 - \hat{F}$) except constant fragmentation with $M = 10^5$. Accretion and attrition act on three monomers per time step. Gillespie refers to the computational steps of the stochastic simulation algorithm. Stick-breaking refers to stick-breaking fragmentation with a beta distribution and power-law coalescence is equivalent to stick-breaking coalescence. Most panels show cyclic behaviour. The exceptions, some of the power-law distributed fragmentations and stick-breaking, instead have fragmentation that removes so little of the cluster fragmented that the system almost immediately returns to a single cluster. Note that scales vary between panels to better show differing cycles.

5.5 Discussion

In this chapter, we have examined a well-known model of coalescence and fragmentation which has been used in a variety of works to model particles coming together and breaking apart in a well-mixed volume. We began Section 5.2 by noting a previously unrecognised relationship between different kernels and methods used to analyse them. In particular, we observed that the steady-state can be expressed in terms of the Catalan numbers, a well studied sequence with many results that have been effectively reproduced in the coalescence and fragmentation literature. We also observed that there are natural stochastic oscillations in the predicted steady-state of the system. Normally, these stochastic oscillations are computationally complex, but can be computed for small systems.

In practice, these stochastic oscillations are fairly small in comparison to system-wide gel-shatter cycles, in which the competition between coalescence and fragmentation is unbalanced due to rare shattering and frequent gelation and does not lead to the steady-state expected in the literature. As this lack of balance is exaggerated, the system becomes almost deterministic in forming gels before shattering to reset the entire system, as seen in Section 5.3 with the forced cycles. We studied these cycles and discovered that they are characterised by a dimensionless parameter r representing the ratio of the characteristic times of gelation and shattering and by a cyclicity order parameter $\mathcal{K}$. The critical regime for gel-shatter cycles to exist empirically appears to be $0.1 < r < 10^3$ and $\mathcal{K} \gtrsim 0.1$, with cycles as the dominant feature when $\mathcal{K} > 0.2$. This system is robust to some simple perturbations in the rules, but it also does not take much to disturb the system, as discussed in Section 5.4. We recognised limitations in the standard methods of fitting power-law distributions, either by MLE or by KSMLE, to data from the coalescence and fragmentation model, likely due to the presence of a truncated power-law tail. Even with this limited characterisation, we were able to observe large changes in the fitted exponent α when allowing simultaneous Becker-Döring type dynamics of accretion and attrition. The system, whether with steady-state parameters or gel-shatter parameters, was also particularly sensitive to power-law fragmentation, even though it was robust to other types of fragmentation. The power-law fragmentation exponent θ could smoothly, but non-linearly, interpolate between the standard case $\theta > 2$ and more exotic variations.

These results have broad consequences for the usage of coalescence and fragmentation models in terrorist modelling. Due to the observed sensitivity of the resultant distribution on the underlying rules, it quickly becomes important to ascertain the exact rules for how the monomers interact. For coalescence, this is less of a problem in fields like chemistry where chemical pathways are well-studied and understood, but is critical for social clusters where the equivalent pathways can easily be unknown. For fragmentation, the system is strongly sensitive to rules like power-law fragmentation for $\theta < 2$, and it cannot be assumed that a system automatically shatters upon encountering fragmentation. For example, for a government wanting to reduce the size of social clusters within a population that follows a multiplicative coalescence kernel, that government needs to make sure that it fragments clusters more than a power-law with $\theta = 2$ to replicate the effectiveness of shattering. Otherwise, the results become complex and depend heavily on how frequent fragmentation is.

Even beyond questions of robustness, the mere emergence of gel-shatter cycles in place of a steady-state has profound repercussions for future studies. For example, Johnson *et al.* (2016) indicates the presence of reincarnation of clusters of followers in a social network. It is not obvious that these are necessarily different from gel-shatter cycles, and analysis of the social network data needs to be done with the knowledge that gel-shatter cycles may be the natural state of the system as opposed to a steady-state. As such the presence of this non-trivial dynamic instead of the presumed steady-state and the sensitivity to perturbations in the rules needs to be considered when models of this type are used in applications and these directions represent useful avenues for future theoretical research.

5.6 Conclusion

There is still much work to be done in modelling coalescence and fragmentation, which is a highly divided field with many workers unaware of each other's work (understandably so, given the field's diversity). We have contributed to this vast body of literature first a series of connections between existing bodies of knowledge, with a discussion of the Catalan numbers. Next, we described both the expected oscillations, arising from stochasticity, and the unexpected, arising from mismatches of fundamental rules. Finally, we characterised the robustness of both the expected steady-state and the new gel-shatter cycles in terms of their responses to simple changes in the rules. While robust to some changes, there are many simple changes with drastic effects on the cluster size distribution.

Obvious extensions to our findings would be to better analytically characterise the gel-shatter cycles or the various perturbations and their effects on the underlying system. The former might explicitly consider the importance of the (effective) homogeneity of the coalescence kernel in the formation of the gel-shatter cycle. The homogeneity of the coalescence kernel should also be contrasted with the effects of power-law fragmentation, in addition to the scaling of the fragmentation rate (*e.g.* size-biased or constant kernel). Formal analytical results beyond scaling arguments for the regimes in Section 5.3 would also be welcome, especially if it would allow for understanding of the inherent variation in the system (*e.g.* recurrence time). Similarly, formal analytical results and metrics for perturbations would better characterise the robustness of the systems, which is especially important given the broad variety of physical systems that coalescence and fragmentation are used to model.

Having discussed three different systems at three different scales and with three different sets of techniques, we place coalescence and fragmentation modelling in context as well. Coalescence and fragmentation modelling is the broadest project in this thesis in terms of its operational scope due to its microscopic scale. Its simple rules suggest that it might be appropriate to model many quite different generative processes that are cast as dynamical analogies to each other. Its appropriateness as such a model is constrained and informed by our knowledge of what happens at other scales. We should not desire to use a coalescence and fragmentation model because we observe a power-law distribution, but instead because the observed behaviour of the system at other scales resembles coalescence and fragmentation. With these thoughts in mind, we proceed to our conclusion, in which we consider relationships both between the chapters of the thesis and stretching beyond the topics already discussed.

6 Conclusion

The future will be better tomorrow. – *Dan Quayle*

The last century has suffered many different conflicts, from classical clashes between armies and navies, to trenches, to aerial campaigns, to bombing campaigns, and to asymmetrical guerrilla warfare and to 'hearts and minds'. These conflicts have been terrifying for its populace, with some of the deadliest wars, both for those fighting the war, and for those whose homes have become the battlefront. In this thesis, we have taken case studies representing a cross section of the conflicts of the last century and applied different and new quantitative methods to them. In Chapter 3, we began with the Battle of Britain, an aerial campaign which the Germans might have won, if only they had better intelligence, initiative, and the central desire to defeat the British. We continued in Chapter 4 with a recent debate over whether the last century has been different from the previous: has there been a "long peace" or a "great violence"? We answered that, consistently, changepoint analysis indicated yes, something has changed in the time series of battle deaths data, supporting the affirmative claim. We finished with Chapter 5, in which we detailed a model that has been proposed for modern conflict, *i.e.* how insurgents (or other coalescing units) might come together, and how counter-insurgent forces (or some other fragmenting force) might be applied in order to prevent the population from organising itself.

Aside from the most general of relationships, "it is all vaguely conflict modelling!", how then do these topics relate? What binds this thesis together? Techniques are one obvious riposte. Each chapter builds on the same foundational belief of mathematical modelling that we can study the world around us by studying simple mathematical models. The main tool we used to do so is stochastic simulation, *i.e.* that we can prescribe some distribution of events and draw from the distribution to create a time series that we wish to analyse. Bootstrapping the Battle of Britain is exactly this, where our distribution is the empirical distribution, while in our changepoint analysis of historical battle deaths, we simulated time series that reflected Richardson's law for deadly quarrels to create alternative 'histories' (if one is permitted to call a time series such) with known changepoints to test modern methods upon. The simulations of coalescence and fragmentation models are similar, but instead of drawing static events, as in the changepoint analysis, or drawing from a changing pool of events and interpreting their cumulative value, as in the Battle of Britain, events are drawn using a plausible mechanistic process on an underlying and dynamic population. In each case, random sampling is the simple part of the problem; understanding what the simulations mean for the system as a whole is where the value is truly drawn.

There are other relationships between the models that could be exploited, as suggested in the introduction. Consider the interactions between the changepoint analysis and the other two chapters. The changepoint analysis was conducted on historical battle deaths, which were supposed to be the same regardless of the type of underlying war that was fought, but this is obviously at odds with the other materials in this thesis. One way to relate these chapters is to consider if battle deaths datasets have a unifying theory centred on generative processes. For example, we can employ Lanchester's Laws and related techniques to understand the scalings of battle deaths in battles. Battles are then simulated according to war types. For clashes between classical armies such as in inter-state wars or some civil wars, *e.g.* the United States civil war, we might combine Lanchestrian scalings with bootstrapping. When dealing with non-state populations and/or insurgents, coalescence and fragmentation modelling can be used to model how an insurgency might organise itself before battle, such as the War in Afghanistan or the Australian Aboriginal wars fought between settlers and Aboriginal Australians. When dealing with forces using guerrilla tactics, Bohorquez *et al.* (2009, supplemental materials) considered using encounters between populations as fragmentation, and one could then resolve such an encounter using Lanchester's Laws as well, *e.g.* following Deitchman (1962). Lanchester's Laws might very well be inappropriate, but this unifying model would be testable and is a good first step for future work.

The methods and ideas of changepoint analysis also feed back into bootstrapping the Battle of Britain and coalescence and fragmentation modelling. Consider Appendix A, in which we used changepoint analysis to largely support the official historical phases (James, 2000) that we employed in Chapter 3. Such work validates our decision to split up the bootstrap in turn by phase, which is clearly a more natural way of splitting the time series than, say, month. Coalescence and fragmentation modelling also has a natural time series component. There are multiple systems where such a component is important: *e.g.* detecting the introduction of a catalyst in a chemical system, a change in tactics among an insurgent population, or a change in social grouping behaviour due to changes in lockdown rules. Additional complexity arises in that there is not necessarily an obvious instantaneous change in summary statistics for all rules. In Figure 5.15, for example, one would not necessarily switch orientations immediately, even though the range of estimated power-law exponents is nearly identical. Rather, it seems more likely that the system would reorganise internally around the new rule and we would want to detect the beginning of this 'burn-in' period, to borrow the language of Markov chain Monte Carlo. Such a changepoint analysis would require an understanding of the possible variations induced by the rules as well as the ability to perform a changepoint analysis of the trends in the population or summary statistics.

What then of the Battle of Britain and insurgency modelling? Are these merely linked via being conflicts? Recall the discussion of trying to model an aerial campaign with Lanchester's Laws, in which Johnson and MacKay (2011) identified that the Battle of Britain did not match theoretical expectations, but more closely resembled duels. Unaimed fire also resembles duels, in that they both produce linear laws, but unaimed fire is also used to model the inability of forces to aim on insurgent forces within hiding. In both cases, the defender has the ability to choose to engage or retreat more readily than the attacker. RAF 11 Group commander Keith Park was essentially employing guerrilla tactics: the Luftwaffe, the German air force, was unaware of where the RAF forces were located, which aerodromes the RAF were using, and which targets the RAF felt obliged to defend, while the RAF in turn could choose when to engage, how to engage, and with how many troops to engage. Given that coalescence and fragmentation modelling has been used for guerilla tactics, one might seek to combine the two.

As forces approach each other, they have a tactically chosen size distribution (of bombers and fighters) which then is broken up when the forces come into contact. As the dogfights resolve, the remaining forces coalesce with each other to try to pursue each other until combat is finished. If coalescence is sufficiently weak or slow, then the system easily remains fragmented (and thus, resulting primarily in duels) until the end of the battle, which is consistent with the discussion by Johnson and MacKay (2011). Such a combined model might also explain the weak benefit that the Luftwaffe obtained from greater sortie numbers. Luftwaffe fighters were under some obligation (even if the exact amount is up for debate) to defend their allied bombers, which the RAF were attempting to prevent from bombing Britain. As such, the Luftwaffe might be more obliged to remain coalesced. One could imagine using footage of dogfights to determine clustering and the interactions between the different particle types. Perhaps an agent-based model could be implemented to replicate the dogfighting itself in a three dimensional space. Questions that this approach might be able to analyse include how to optimise the size distribution preferred by the forces before entering the dogfight, as well as why higher sortie numbers might have favoured the Luftwaffe: was it due to the bombers, the status as attacker, or something else entirely? On the other hand, this approach might be rapidly falsified with questions like ‘when dogfights resolve, do victors immediately seek out their allies to assist them or return to formation?’ This model might also struggle with distances that might result from the dogfights, where the repulsion might make it extremely unlikely that coalescence might reoccur without directly coding it into the system. The ability to test the model, that it is falsifiable, is important to prevent the model from being used simply because it seems obvious (the street-light effect).

There are other things we can learn by applying the theory of one project to the others. Consider the work of Clauset (2018), in which bootstrapping was used to simulate counterfactual time series to which probabilistic arguments could be applied to estimate the probability of obtaining something like the long peace. While our method of changepoint analysis is an alternative technique, the two ideas could be conceptually combined. This is similar to how we conducted our *a-posteriori* robustness checks; we supposed that a certain model existed from our changepoint analysis and drew from it to see how likely such a model would be identified if it were the true model. One could instead have used the bootstrap again, as opposed to power-law distributions, in order to verify the results instead. Further analyses similar to those conducted by Clauset could also be conducted, but with changepoint analysis replacing probabilistic methods to answer questions such as do the World Wars necessarily induce a segment between them, or can they exist within the same segment?

On the other hand, using an informed model can improve the bootstrap for a specific application. Consider the problem of applying the bootstrap to time series with a constraint much like coalescence and fragmentation, where only certain transitions are allowed. For a concrete example, consider a coalescing and fragmenting population which we know the size distribution of at multiple points in time and we want to bootstrap the data. Just using a bootstrap alone would ignore any correlations in the time series (a well-known problem (Efron and Tibshirani, 1993)) so one solution would be to create a bootstrap and then perform a “guided” simulation in between each point in the time series bootstrap. For models that do not fit the process generating the data well, the process will have to make many low likelihood decisions to interpolate, while a model that fits the process quite well (such as the empirical events in the original data for very well-sampled processes) would proceed without many such decisions.

We have thus far described some broad interactions between the various chapters in this thesis, which contain a wide variety of techniques and ideas for modelling, each of which could naturally be expanded further. Many variants of the bootstrap exist, and determining a best variant for historical data is perhaps its own reward. Finding datasets and ways to apply the bootstrap to them is an equally viable direction to go in. Much the same could be said of changepoint analysis or coalescence and fragmentation modelling; technical developments and datasets to apply the developments to are always in high demand in applied mathematics. Instead, we close this chapter and thesis with a discussion of some more specific ideas and avenues for future development. Some developments lay at the intersection between two projects, while others are continuations.

We begin with the Battle of Britain. A natural first idea that was proposed in the original paper is the investigation of the Battle of the Atlantic. Instead of thinking about pilot deaths, one would think about the movements of the convoys as they tried to deliver resources to the United Kingdom and Soviet Union from Canada and the United States while evading the presence of German and Italian submarines (U-boats). Some care would need to be taken to understand the appropriate hierarchy to model with, given possible complications from aerial cover, equipment such as radar, convoy composition, and tactics. The weighted bootstrap would then be employed within the hierarchy in much the same way as it was in the Battle of Britain to resolve each convoy's fate and determine the importance of the details and decision making of both sides. An obvious omission in our analysis of the Battle of Britain is our lack of usage of causal techniques. This includes a well-developed usage of the causal bootstrap (Imbens and Menzel, 2018), but also more generally includes the omission of techniques advocated by Pearl (2000) or Imbens and Rubin (2015). Such techniques could be profitably explored both for the Battle of Britain as well as in the Battle of the Atlantic. They are likely important to properly characterise and utilise a proposed hierarchy.

Focussing on the Battle itself, one intriguing detail that is nevertheless hard to resolve is what would have happened if the Germans had decided to invade. The temptation is to form a triptych, beginning with bootstrapping the Battle of Britain's aerial campaign, followed by modelling the subsequent naval encounter (conditioned on the success of the Germans in the previous project), and finishing with how well the subsequent landing would have gone (again conditioned on the previous project). The difficulties that arise are related to determining appropriate rules for wargaming each part of the triptych and then implementing the wargame in an appropriate simulation such that a broad understanding could be gathered. This is obviously not impossible to do (Cox, 1974), but would be fairly labour intensive and would require more than just a bootstrap. The bootstrap could be used anywhere where empirical data existed, *e.g.* if a dataset exists for the effects of an armament against a type of armour, but it is more likely that distributions would need to be conjectured from physical and historical observations instead. Understanding the sensitivity of the model to these parameter assumptions is likely critical, whatever model is used. Implementation would likely be agent-based, but there is a danger there in estimating the personalities of the captains of various vessels. One possibility is to look for historical personalities who would have been likely to be present at the counterfactual Sea Lion and incorporate them as the agents. A different route would be to crowd source the captaining of the larger vessels, turning the model into a multiplayer game. A large enough crowd could then grant many realisations of the simulation, at the cost of longer development time and a greater difficulty of historical consistency, especially if the crowd begins to coordinate (*e.g.* using tools outside of the simulation to communicate faster than would have been historically feasible).

Between the results from the simulated invasion and the historical documentation, one could then compile what a landing scenario would look like. This is probably the most exuberant proposal, in that it is contingent on the Germans succeeding in both the aerial campaign and the naval campaign. It is hard to know exactly the best way to model the land invasion, especially as the results would likely be contingent on just how many troops would be able to land and how much support they would receive from the air and sea, both of which would likely still be contested as the land invasion was occurring. Hence, this would rely most on the historical context, such as the German plans for invasion (German Plans, 2017), and counterfactual, rather than mathematical modelling, principles. Modelling could be used to elucidate specific points, but would be hard to use to cover the general strategy. A slightly different tactic might be to use the causal methods (Pearl, 2000; Imbens and Rubin, 2015) in order to organise how the different parts of the counterfactual might come together.

Regarding the changepoint analysis, again multiple avenues are possible. We anticipate that the datasets will be refined over time, so adapting to a per battle battle deaths dataset (as opposed to the per war datasets we used) would be one obvious refinement. As discussed above, one could also perform additional robustness checks following the lead of Clauset (2018). More interesting for the mathematical audience would be to incorporate additional information from the battle deaths datasets to complement the death totals themselves. The battle deaths datasets have fields such as the type of war, combatants, or location (Gleditsch, 2004; Sarkees and Wayman, 2010). Performing a (now multivariate) changepoint analysis that incorporates these elements could better guide us towards looking for spatial changepoints. This in turn addresses questions such as if the period after the World Wars was characterised by different types of wars, such as proxy wars, or if the long peace was a particularly western phenomenon. A different but related problem might be to consider the problem of changepoint analysis of hybrid data, where wars have multiple types that change over time such as the Taiping Rebellion, or fuzzy data, where the properties of a datum are not necessarily fully known such as if the number of deaths for each war have a minimum, maximum, and mode rather than a single point estimate.

More difficult still would be to try to adapt changepoint analysis to more fully address the long term trends posited by Pinker (2011, 2018). This would require more datasets which cover the trends postulated and a multivariate changepoint analysis technique that is intended to identify changepoints in the raw data or in trends throughout the data, which is an ongoing topic of discussion in changepoint literature. A starting point might be to consider the initial approach of Richardson (1960b) in which deadly quarrels of all sizes are analysed, rather than just wars. The more complicated the model, the harder it is to determine which of the moving parts was responsible for a given outcome, which is perhaps the most damning problem for this direction.

In the opposite direction of Pinker's work on better angels, one could try to model the generative process(es) for conflict. Richardson's law, that deadly quarrels are power-law distributed, is suggestive here, in that power-law distributions are scale-free. On the other hand, Richardson's law actually required the use of two power-law distributions, which might require multiple processes before one even begins to consider problems of arms races or differing war types, both in the sense of aimed fire and duels or unaimed fire as well as inter-state, *etc.* Many processes could be tested. For example, one might consider a large number of exponentials that are paired at random following the approach of Clauset *et al.* (2010, using the model of Reed and Hughes (2002)). Such an approach might naturally reconcile with the ideas of Kondratieff wave-like cycles in war and peace (Goldstein, 1985; Turchin, 2007) if there is a cooling off period following the pairing of exponentials. One possible problem is that this is a well-mixed two-body problem, but defensive pacts and the like arise and space must be considered (as noted by Richardson (1960a) who considered whether the lengths of shared borders influenced the propensity for war between two nations). Additionally, there is the question of what drives the exponential dynamics, unlike with power-law dynamics which appear to be the same 'all the way down' due to the scale-free nature. Instead of beginning with a top-down approach, a bottom up approach, such as coalescence and fragmentation modelling could be postulated. For example, a spatial coalescence and fragmentation approach with some sense of dynamic identity amongst the population could be used, where similar identities prefer to coalesce while differing identities prefer to fragment each other. In this case, the quandary becomes how to produce multiple power-law results and whether the definition of the population is *de facto* changing over time (*e.g.* do tribal villagers and modern civilians coalesce in the same way? Do they produce warriors and soldiers in the same way?). In either case, identifying the proper model of combat would be important.

Focusing on coalescence and fragmentation modelling, a straightforward task would be to explore the existence of cycles in coalescence and fragmentation datasets such as the one obtained by Johnson *et al.* (2016) or other social media data. The goal would be to fit the coalescence and fragmentation model to the dataset to determine which regime the dataset belongs to. If the dataset behaves as expected of the regime, this would help validate the model, especially if cycles are observed and these cycles are of the same stochastic gel-shatter type as observed in Chapter 5. Comparing with multiple datasets would obviously be ideal for validation. If the cycles observed by Johnson *et al.* (2016) are different in character than those of gel-shatter cycles, characterising the difference and if they should be expected from the base model would be a good step forward. A particular challenge can be found in the form of external driving forces. If one investigates cycles in insurgencies for instance, one might need to rule out political cycles, especially deterministic ones, amongst either the insurgent or counter-insurgent forces.

One obvious direction is to pursue stronger analytic results, similar to results by Pego and Velázquez (2020). Topics of interest include analytically describing the three regimes of forced gel-shatter cycles, stochastic unforced gel-shatter cycles, and a steady-state, establishing the stability of the cycles, and determining the necessary and sufficient conditions, such as the degree of homogeneity required. There is a great deal of literature that has already been done on the steady-state and parallel work currently being done on cyclicity, so this should be a highly competitive direction to pursue research in. We are excited to see what results will be achieved to better understand cyclicity within coalescence and fragmentation models.

It might also be interesting to examine what happens when varying models are combined. For example, consider taking the multiple populations coalescence and fragmentation model and combine it with Lanchester's Laws to create a toy conflict model. In this case, fragmentation would occur when two different populations encounter each other (Bohorquez *et al.*, 2009, supplemental materials) and would take place using (a stochastic or deterministic implementation of) Lanchester's Laws. One can then ask much the same types of questions that we have asked in Chapter 5.

Does a stable steady-state emerge, or do we instead see cycles? How much imbalance is necessary to force one population to gel while the other population(s) are forced to remain disaggregated? What are the effects of using linear law, square law, or asymmetric laws on the system? What are the effects of using different kernels for different population interactions? A natural starting point would likely be just to consider the multiple populations case before changing the fragmentation to something more complicated. Regardless, proposals of this sort seem unlikely to be easily tractable, and would likely require simulation or other techniques.

Implicit in many of these ideas and interactions is again the nature of a valid restrained counterfactual. For example, the idea of not choosing low likelihood models is a variation of the fifth rule for counterfactual analysis due to Tetlock and Belkin (1996): we should choose statistically likely results over those that would rely on the unlikely having occurred. This is a guiding principle in most statistical work and a principle well-used both within this thesis, *e.g.* our *a-posteriori* robustness checks to see how likely the proposed model would occur if it were true in Chapter 4, and without, *e.g.* the work on analysing the long peace by Clauset (2018). The other principles are no less important though. We have strived throughout this thesis to make clear our assumptions regarding the systems examined and methods used. Where necessary, we have discussed the logical consistency of using our methods with the data, crucial especially for Chapters 3 and 4 where *e.g.* independently and identically distributed (IID) is not necessarily obvious. The third principle of historical consistency pairs naturally with statistical consistency in analysing the results of our re-weighted bootstraps for the Battle of Britain, helping us understand how the Germans squandered the weaknesses of the British before the Battle even began. Where we are able, we have appealed to theoretical consistency regarding knowledge of past datasets as well as Lanchester's Laws and Richardson's works (*e.g.* Richardson, 1960a,b). Finally, our results are also falsifiable in theory, although achieving the correct underlying conditions might prove difficult. One might need a multi-front war including two balanced aerial powers with which the aerial campaign is segregated from the other ongoing parts of the conflict to try to project and test lessons from the Battle of Britain. The ongoing debate on the decline of war might more readily test our results, just as explicit 'insider' knowledge of how insurgents operate and respond to force might readily test the assumptions and results of the coalescence and fragmentation model. Our gel-shatter cycles are perhaps particularly falsifiable: one would need to find a model system where coalescence occurs much more rapidly than shattering and study its size distribution over time to see if the pattern of gelation followed by shattering repeats.

One broad criticism of our methods is that we have extended a 'small' amount of data far beyond its original domain. For example, in the Battle of Britain, we extended airfield targeting from 16 days in the Battle as actually fought to 54 in CF5. Much of this thesis is about how far we can safely extend what we know about, or how robust is, a system. Our answers are consistently quite far, with caveats. For the more than tripling the number of times a type of target was chosen, we again consider the conditions of the Battle: much of the Battle did not change. There was not adequate time or ability to change the planes and tools used, nor were there substantial changes in leadership or pilots or underlying conditions. So long as we believe that each day was drawn from and is representative of the same distribution, these days represent all that we can know. Naturally we cannot factor in learning or the development of new tactics, but we also do not have evidence that those really occurred in the Battle proper. More importantly, the chapters in this thesis represent jumping off points for thinking about problems. The lessons of this thesis are meant to detail how we can think about modelling counterfactual history, or addressing whether war has declined, or determining if a model is appropriate for its application. We have learned how to reconcile assumptions with mathematical techniques like the bootstrap. We have learned how to explore the interrelationships between tools and datasets. Furthermore, we have learned how to stress-test a model to verify it as a description of a process.

In learning these lessons, this thesis has covered a broad swath of history. We have discussed the Battle of Britain, and how Hitler could have fared better if only he was not Hitler, how Goering could have fared better if only he was not Goering, and how the Luftwaffe could have won the Battle of Britain if only they had understood the battle they were fighting. We have discussed the time series of battle deaths data, created an algorithm to investigate the data, and found evidence that there have been changes that coincide with those proposed by the long peace hypothesis. We have discussed a modern model of coalescence and fragmentation that has been used to describe the interactions of insurgents and investigated how this model differs in behaviour from predictions in the literature, both with and without perturbations.

More important than our individual applications is the lessons about what tools are available and how we can safely use them. We clearly have not closed the book on conflict regardless of whether Hobbes or Rousseau or both in varying proportions were right about the causes of conflict. Instead, we have shown how to apply mathematical tools in contexts to which they were foreign. We have shown the usefulness of the statistical technique of bootstrapping to our colleagues in history. We have helped introduce changepoint analysis to peace studies where the primary arguments were made using regressions and probabilistic models. We have brought to the forefront previously undocumented behaviour of coalescence and fragmentation, which might drive cycles that are observed elsewhere in the literature without requiring novel explanation. Each of these applications enriches the literature around it and has laid the groundwork for future work, both for ourselves as well as our colleagues.

A Changepoint analysis of the Battle of Britain

The primary assumption of the bootstrap is that the (original) sample must be independently and identically distributed (IID). Previous analysis of data from the Battle of Britain, however, suggested the existence of a changepoint on September 15th due to the fiercer combat before that date, with twice as many deaths per total number of sorties (Johnson and MacKay, 2011, Fig. 1). While we can work around changepoints by considering each segment separately, each segment needs to be identified and verified as IID. Due to the historical constraints, we are most interested in the properties of the time series of British pilot casualties, but we also present other variations for completeness, including airframes and sorties for each side and in total, the difference in number of sorties, and the airframes and British casualties per total or British number of sorties.

Our initial battery of tests are tests of autocorrelation and stationarity. If data is autocorrelated, then each datum appears to depend directly on one or more previous data points. If data is non-stationary, then there would appear to be some form of dependence on time. While autocorrelation is available in the base implementation of R (R Core Team, 2018), stationarity tests have been implemented in the `tseries` R package (Trapletti and Hornik, 2019).

As with all statistical tests, the general results depend on the level of significance a, but broadly speaking nearly all of our data has significant autocorrelations at the $a = 0.05$ level, with a significant lag of 3 amongst all variations considered except the difference in number of sorties flown. The time series of British pilot casualties has significant autocorrelation for lags of 1, 3, 4, and 20, with the lags of 3 and 20 significant at the $a = 0.01$ level.

Controlling the false discovery rate (under the assumption of independence) to correct for the large number of tests using the procedure due to Benjamini and Hochberg (1995, implemented in R) indicates that only a lag of 1 of the number of airframes per British sortie and a lag of 3 of the number of German airframes are significant, and only at the $a = 0.05$ level. The lags of 3 and 20 for British pilot casualties have an adjusted level of significance of $a \approx 0.15$. As this procedure controls the expected number of false discoveries, it does not necessarily discount the significant results, but instead suggests that a high proportion of them are false positives. We need to follow up with additional tests as a result.

We next check for (non-)stationarity using the augmented Dickey-Fuller (ADF) test (Trapletti and Hornik, 2019). This test adopts the null hypothesis that the data is not stationary and alternative that it is. Using this test, most of the data can be regarded as consistent with the null hypothesis that the data is not stationary. Exceptions include a very weak rejection for British pilot casualties ($p = 0.0926$) and a more significant rejection for British sortie numbers ($p = 0.021$).

We can compare the results with the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test (Trapletti and Hornik, 2019). This test has the opposite convention: the null hypothesis is that the data are (level or trend) stationary and the alternative is that they are not. (A process is trend stationary if it tends to revert to a trend.) Trend stationarity is rejected in most cases except for British sorties and British pilot casualties per total number of sorties. In the level case, most data sets are consistent with the null hypothesis of stationarity, except for British pilot casualties per British sortie ($p = 0.028$) and airframes per British sortie ($p < 0.01$).

Together this suggests that the majority of (combinations of) data sets contain insufficient information to conclude whether the data is (level) stationary. The exceptions are that British sortie numbers are significantly consistent with stationarity, and the British pilot casualties are more

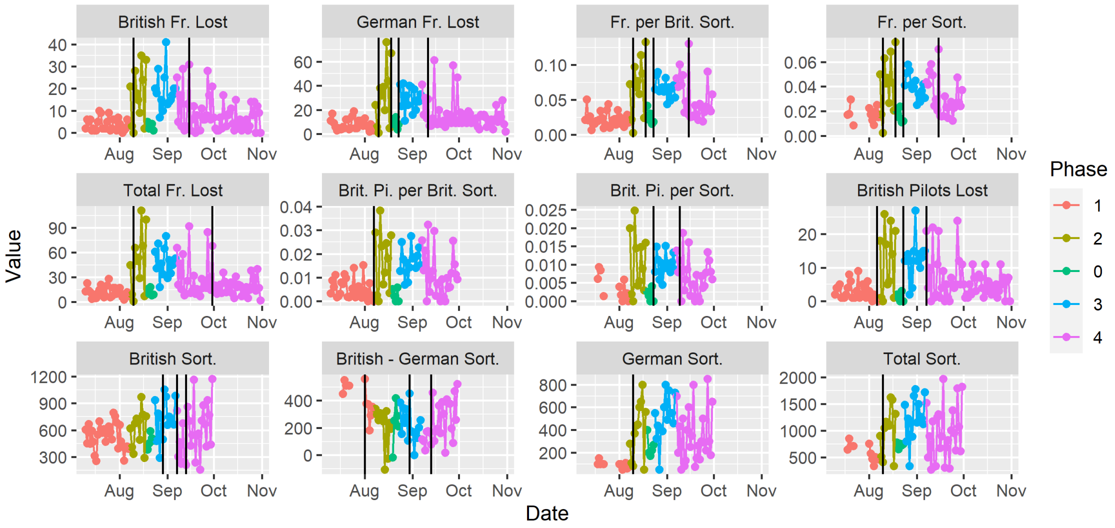


Figure A.1: Changepoint analysis of the Battle of Britain variables. We abbreviate British as Brit., Frames as Fr., Pilots as Pi., and Sorties as Sort. The changepoints found agree excellently with the historical phases when considering British pilots lost. Phases are those recorded by James (2000). Agreement is less good, but still roughly present in the other plots.

weakly consistent with stationarity, while British pilot casualties per British sorties and airframes per British sortie are significantly consistent with non-stationarity.

Obviously, we would like to be more certain about the properties of the data, especially the British pilot casualties. Instead of viewing the whole of the data as a single sample, we can instead look for changepoints to create segments to investigate. We borrow a method from Section 2.3: we apply a non-parametric changepoint analysis using ED-PELT with mBIC as a penalty. Performing this method for each variable combination of interest, we see that the phases from the official history (James, 2000) line up quite well with the British pilot casualties, seen in Figure A.1.

Dividing up data according to phases leads to additional problems due to segment lengths. The second and third phases have only 11 and 14 days associated to them, while Phase 0 is only 5 days long. We note that there has been some limited work discussing the applicability of unit root and stationarity tests that suggests that the ADF and KPSS tests maintain statistical power (low probability of false negatives) for short (length 25) time series (Arltová and Fedorová, 2016).

Reapplying the earlier analysis, Phase 1 is broadly uninformative as to stationarity, with failures to reject on both tests except for a dual rejection for German Sortie numbers and consistency with stationarity for the British casualties per British sortie and total airframes lost per British sortie. Phase 2 is more informed: British pilot casualties and lost airframes are both consistent with stationarity, as are British pilot losses per total or British sortie numbers. On the other hand, German airframe losses, total airframe losses, and airframe losses per British sortie are all not consistent with the stationarity hypothesis. We note that Phase 0 lacks sufficient data to come to a conclusion using the ADF test, and that the KPSS indicates that the data sets generally fail to reject the stationarity null hypothesis. Phase 3 similarly is unable to reject either null hypothesis. Finally, Phase 4 is largely consistent with rejection of stationarity: British pilot casualties, airframe losses, sorties, as well as total airframes lost and total airframes lost per either total or British sorties are all consistent with rejection of stationarity. Both German airframes lost and the difference in the number of sorties reject both null hypotheses. In nearly all of these cases, rejections occur only at the $a = 0.10$ level.

In summary, we have found that British pilot casualties, when divided according to the official historical phases (James, 2000), has insufficient evidence to conclude whether the data is stationary or not (Phases 1, 0, and 3), is consistent with stationarity in Phase 2, and is consistent with non-stationarity in Phase 4. The last phase is somewhat concerning for application of the bootstrap. Further analysis of Phase 4 using the KPSS test fails to reject the trend-stationary null hypothesis. While our focus in the main text is not on Phase 4, we could control for this further by attempting to detrend the Phase 4 data. Alternatively, it is possible that the problem stems from the September 15th changepoint detected by Johnson and MacKay (2011) and that there may be another effective phase that emerges at around that date.

Returning to the autocorrelations, dividing the data according to phases drastically reduces the autocorrelations present, possibly due to the lack of data. Autocorrelation is still present even to $a = 0.01$ level in Phase 1 for British pilots lost per total sortie, the difference in sortie numbers, and the total sortie numbers for lags of 1 day. Curiously German sorties have a similarly significant lag of 3 days with an (partial) autocorrelation of 0.504 and airframes per total number of sorties has significant lags of 2, 4, and 6 days (values -0.648, -0.544, and -0.907). Some other lags are significant at the 0.10 level, but are larger than 1 day. Phases 2, 0, and 3 have no lags significant at the 0.01 level, while Phase 4 only has a 5 day lag in German sorties and a 9 day lag in total airframes per total number of sorties significant at the 0.01 level. Overall then, the data is now more consistent with an independence assumption, although some weak evidence to the contrary remains.

British pilot casualties have only a lag of 6 significant at the $a = 0.10$ level in Phase 1, a lag of 2 significant at the $a = 0.05$ level in Phase 2, and a lag of 4 and a lag of 6 significant at the $a = 0.10$ and $a = 0.05$ level in Phase 4. Attempting to control the false discovery rate suggests that nearly all lags are false discoveries with high probability. As such, we conclude that the data, divided up by official phase and equipped by historical context appears to largely be consistent with the IID assumption in terms of stationarity and autocorrelation.

B Battle of Britain tables

Table B.1: Combat loss and target data. Phases are 1: 10 Jul - 7 Aug, principally of coastal attacks and armed reconnaissance, 2: 8 Aug - 18 Aug, of heavy attacks on mostly coastal targets, 0: 19th Aug - 23 Aug, of little interaction between the forces, 3: 24th Aug - 6th Sept, of attacks gradually concentrating on aerodromes and 4: 7th Sept - 31st Oct, following the switch to principally bombing London. Primary targets are A: aerodromes, C: docks, shipping and coastal or L: London, Kent, and Thames estuary. Note that pilots lost, wounded or slightly wounded are measured by incident. Thus some values represent pilots receiving a slight wound, flying again, and then receiving another wound.

Day	Date	British Airframe Losses	German Airframe Losses	British Pilots Lost	British Pilots Wounded	British Pilots Slightly Wounded	Phase	Primary Target
1	10/07/1940	2	11	2	0	0	1	C
2	11/07/1940	6	17	3	1	0	1	C
3	12/07/1940	5	9	4	0	1	1	C
4	13/07/1940	6	6	5	0	1	1	C
5	14/07/1940	1	3	1	0	0	1	C
6	15/07/1940	2	5	0	1	0	1	C
7	16/07/1940	1	4	1	0	0	1	C
8	17/07/1940	1	4	1	1	0	1	C
9	18/07/1940	5	6	4	0	0	1	C
10	19/07/1940	10	5	4	4	0	1	C
11	20/07/1940	9	12	6	0	2	1	C
12	21/07/1940	2	12	1	0	0	1	C
13	22/07/1940	2	4	1	0	0	1	C
14	23/07/1940	2	5	0	1	0	1	C
15	24/07/1940	5	15	3	0	1	1	C
16	25/07/1940	9	19	7	2	2	1	C
17	26/07/1940	1	5	1	0	0	1	C
18	27/07/1940	2	5	2	0	0	1	C
19	28/07/1940	6	11	1	4	1	1	C
20	29/07/1940	6	11	3	1	0	1	C
21	30/07/1940	1	9	0	1	0	1	C
22	31/07/1940	7	7	4	2	0	1	A
23	01/08/1940	4	13	3	0	0	1	C
24	02/08/1940	3	7	1	0	0	1	C
25	03/08/1940	0	6	0	0	0	1	C
26	04/08/1940	1	2	2	0	0	1	C
27	05/08/1940	2	8	1	0	1	1	C
28	06/08/1940	6	6	1	0	1	1	R
29	07/08/1940	4	3	0	0	1	1	R
30	08/08/1940	21	24	17	1	3	2	C
31	09/08/1940	3	6	1	0	0	2	R

Day	Date	British Airframe Losses	German Airframe Losses	British Pilots Lost	British Pilots Wounded	British Pilots Slightly Wounded	Phase	Primary Target
32	10/08/1940	0	1	0	0	0	2	C
33	11/08/1940	28	38	25	1	3	2	C
34	12/08/1940	18	32	11	6	2	2	C
35	13/08/1940	15	39	4	1	5	2	C
36	14/08/1940	9	20	4	2	1	2	C
37	15/08/1940	35	76	16	8	5	2	C
38	16/08/1940	24	44	9	5	6	2	A
39	17/08/1940	2	5	0	1	0	2	R
40	18/08/1940	33	67	10	11	6	2	A
41	19/08/1940	5	11	2	0	1	0	C
42	20/08/1940	2	8	1	0	0	0	C
43	21/08/1940	4	14	0	0	1	0	C
44	22/08/1940	4	4	2	1	0	0	C
45	23/08/1940	1	8	0	0	0	0	C
46	24/08/1940	20	41	5	7	6	3	R
47	25/08/1940	18	23	11	1	3	3	C
48	26/08/1940	29	42	4	10	9	3	R
49	27/08/1940	7	11	2	0	0	3	A
50	28/08/1940	15	32	6	6	2	3	R
51	29/08/1940	10	24	2	2	4	3	A
52	30/08/1940	25	40	11	4	0	3	A
53	31/08/1940	41	39	9	18	7	3	A
54	01/09/1940	13	16	6	6	0	3	A
55	02/09/1940	14	37	4	10	2	3	A
56	03/09/1940	15	20	6	4	7	3	A
57	04/09/1940	17	28	11	2	4	3	A
58	05/09/1940	20	27	8	7	3	3	A
59	06/09/1940	20	33	8	7	8	3	A
60	07/09/1940	25	41	15	6	7	4	A
61	08/09/1940	5	16	2	2	0	4	A
62	09/09/1940	17	30	6	4	7	4	A
63	10/09/1940	3	13	0	0	0	4	R
64	11/09/1940	29	29	14	8	7	4	L
65	12/09/1940	1	7	1	0	0	4	R
66	13/09/1940	3	7	2	0	1	4	L
67	14/09/1940	13	13	4	7	0	4	L
68	15/09/1940	31	61	16	5	6	4	L
69	16/09/1940	1	10	0	1	0	4	L
70	17/09/1940	6	8	3	1	1	4	L
71	18/09/1940	12	20	4	5	3	4	L

Day	Date	British Airframe Losses	German Airframe Losses	British Pilots Lost	British Pilots Wounded	British Pilots Slightly Wounded	Phase	Primary Target
72	19/09/1940	0	10	0	0	0	4	L
73	20/09/1940	8	8	5	0	2	4	L
74	21/09/1940	1	11	1	0	0	4	C
75	22/09/1940	1	6	0	0	0	4	L
76	23/09/1940	11	17	3	4	0	4	C
77	24/09/1940	6	11	2	3	3	4	L
78	25/09/1940	6	16	3	1	0	4	(Filton)
79	26/09/1940	8	9	3	2	1	4	C
80	27/09/1940	28	57	20	4	2	4	L
81	28/09/1940	17	12	10	2	2	4	L
82	29/09/1940	6	9	2	3	0	4	C
83	30/09/1940	21	47	6	5	5	4	L
84	01/10/1940	7	9	4	1	0	4	L
85	02/10/1940	2	18	0	1	1	4	L
86	03/10/1940	1	9	1	1	0	4	C
87	04/10/1940	1	15	1	0	0	4	L
88	05/10/1940	7	14	2	3	1	4	L
89	06/10/1940	2	9	2	0	0	4	L
90	07/10/1940	17	19	9	2	4	4	C
91	08/10/1940	8	17	7	0	0	4	L
92	09/10/1940	3	9	3	0	0	4	L
93	10/10/1940	8	12	6	1	0	4	L
94	11/10/1940	9	10	4	4	1	4	C
95	12/10/1940	11	13	5	2	3	4	C
96	13/10/1940	4	6	1	3	0	4	L
97	14/10/1940	1	4	1	0	0	4	R
98	15/10/1940	15	16	6	5	2	4	L
99	16/10/1940	3	15	3	0	0	4	(sweeps)
100	17/10/1940	5	16	5	1	0	4	L
101	18/10/1940	6	14	5	0	0	4	R
102	19/10/1940	1	6	2	0	1	4	L
103	20/10/1940	5	11	3	1	0	4	L
104	21/10/1940	2	7	2	0	0	4	L
105	22/10/1940	6	12	4	1	0	4	L
106	23/10/1940	1	4	1	0	0	4	R
107	24/10/1940	3	12	3	0	0	4	L
108	25/10/1940	14	24	6	5	1	4	L
109	26/10/1940	8	10	5	0	0	4	L
110	27/10/1940	14	16	6	1	1	4	L
111	28/10/1940	0	14	0	0	0	4	L

Day	Date	British Airframe Losses	German Airframe Losses	British Pilots Lost	British Pilots Wounded	British Pilots Slightly Wounded	Phase	Primary Target
112	29/10/1940	12	28	5	1	2	4	L
113	30/10/1940	9	8	6	1	2	4	L
114	31/10/1940	0	2	0	0	0	4	R

Table B.2: Sortie and weather data. Entries are NA: not available, for which no estimate exists, C: clear, O: overcast or cloudy and R: rainy.

Day	Date	British Sortie	German Sortie	Channel & Coast Weather	London, Kent, and Estuary Weather	Midlands & North Weather
1	10/07/1940	609	NA	R	R	R
2	11/07/1940	452	NA	O	O	R
3	12/07/1940	670	NA	O	R	R
4	13/07/1940	449	NA	O	O	O
5	14/07/1940	593	NA	C	C	C
6	15/07/1940	470	NA	O	O	O
7	16/07/1940	313	NA	O	O	O
8	17/07/1940	253	NA	O	O	O
9	18/07/1940	549	100	O	R	O
10	19/07/1940	701	150	O	O	O
11	20/07/1940	611	100	O	R	R
12	21/07/1940	571	NA	C	C	C
13	22/07/1940	611	100	O	O	R
14	23/07/1940	470	NA	O	O	R
15	24/07/1940	561	NA	O	O	R
16	25/07/1940	641	NA	C	C	O
17	26/07/1940	581	NA	R	R	R
18	27/07/1940	496	NA	C	O	R
19	28/07/1940	794	NA	C	C	C
20	29/07/1940	758	NA	C	C	C
21	30/07/1940	688	NA	R	R	R
22	31/07/1940	395	NA	O	C	C
23	01/08/1940	659	100	O	O	C
24	02/08/1940	477	100	O	R	C
25	03/08/1940	425	50	O	O	O
26	04/08/1940	261	80	C	C	C
27	05/08/1940	402	110	O	C	C
28	06/08/1940	416	60	O	O	O
29	07/08/1940	393	70	C	O	C
30	08/08/1940	621	280	C	R	R
31	09/08/1940	409	110	O	O	O
32	10/08/1940	336	80	O	R	R
33	11/08/1940	679	370	O	O	O
34	12/08/1940	732	440	C	C	C
35	13/08/1940	700	450	O	C	C
36	14/08/1940	494	600	C	O	O
37	15/08/1940	974	650	C	C	C
38	16/08/1940	776	800	O	C	C

Day	Date	British Sortie	German Sortie	Channel & Coast Weather	London, Kent, and Estuary Weather	Midlands & North Weather
39	17/08/1940	288	50	C	C	C
40	18/08/1940	755	560	O	O	O
41	19/08/1940	383	400	O	O	O
42	20/08/1940	453	200	O	O	R
43	21/08/1940	589	170	O	O	O
44	22/08/1940	509	220	O	O	O
45	23/08/1940	482	270	O	O	R
46	24/08/1940	936	550	C	C	R
47	25/08/1940	480	325	O	O	O
48	26/08/1940	787	440	C	C	O
49	27/08/1940	288	50	O	O	R
50	28/08/1940	739	400	C	C	C
51	29/08/1940	498	390	O	R	R
52	30/08/1940	1054	600	C	C	C
53	31/08/1940	978	800	O	C	C
54	01/09/1940	661	490	C	C	C
55	02/09/1940	751	750	C	C	C
56	03/09/1940	711	550	O	C	R
57	04/09/1940	678	550	O	C	R
58	05/09/1940	662	460	O	C	C
59	06/09/1940	987	730	C	C	C
60	07/09/1940	817	700	O	O	O
61	08/09/1940	305	200	O	O	O
62	09/09/1940	466	430	C	R	R
63	10/09/1940	224	50	O	O	O
64	11/09/1940	678	500	O	C	C
65	12/09/1940	247	80	R	R	R
66	13/09/1940	209	130	R	R	R
67	14/09/1940	860	400	O	R	R
68	15/09/1940	705	600	C	C	C
69	16/09/1940	428	250	R	R	R
70	17/09/1940	544	350	R	R	R
71	18/09/1940	1165	800	C	C	C
72	19/09/1940	237	75	R	R	R
73	20/09/1940	540	150	O	O	O
74	21/09/1940	563	260	C	C	C
75	22/09/1940	158	140	O	O	O
76	23/09/1940	710	300	C	C	C
77	24/09/1940	880	500	O	O	O
78	25/09/1940	668	290	C	C	C

Day	Date	British Sortie	German Sortie	Channel & Coast Weather	London, Kent, and Estuary Weather	Midlands & North Weather
79	26/09/1940	417	220	C	C	C
80	27/09/1940	939	850	C	R	R
81	28/09/1940	770	300	O	O	C
82	29/09/1940	441	180	C	C	C
83	30/09/1940	1173	650	C	C	C
84	01/10/1940	NA	NA	O	O	O
85	02/10/1940	NA	NA	C	C	C
86	03/10/1940	NA	NA	R	R	R
87	04/10/1940	NA	NA	R	R	R
88	05/10/1940	NA	NA	R	R	R
89	06/10/1940	NA	NA	R	R	R
90	07/10/1940	NA	NA	O	O	O
91	08/10/1940	NA	NA	O	O	O
92	09/10/1940	NA	NA	R	O	O
93	10/10/1940	NA	NA	R	R	R
94	11/10/1940	NA	NA	C	R	C
95	12/10/1940	NA	NA	O	O	O
96	13/10/1940	NA	NA	O	O	O
97	14/10/1940	NA	NA	O	R	O
98	15/10/1940	NA	NA	O	C	C
99	16/10/1940	NA	NA	O	O	O
100	17/10/1940	NA	NA	R	R	R
101	18/10/1940	NA	NA	O	O	O
102	19/10/1940	NA	NA	O	O	O
103	20/10/1940	NA	NA	O	O	O
104	21/10/1940	NA	NA	O	O	O
105	22/10/1940	NA	NA	O	O	O
106	23/10/1940	NA	NA	R	R	R
107	24/10/1940	NA	NA	O	C	C
108	25/10/1940	NA	NA	O	O	O
109	26/10/1940	NA	NA	O	O	R
110	27/10/1940	NA	NA	O	O	O
111	28/10/1940	NA	NA	O	O	O
112	29/10/1940	NA	NA	O	O	O
113	30/10/1940	NA	NA	R	R	R
114	31/10/1940	NA	NA	R	O	O

C Other `changepoint` package arguments

Table C.1: Explanation of function options for `changepoint` and `changepoint.np` R-packages. We use p to indicate the number of parameters in a model and n to indicate the length of the data when calculating a penalty. Note that mBIC, CROPS, PELT, and ED-PELT are discussed further in Section 2.3.

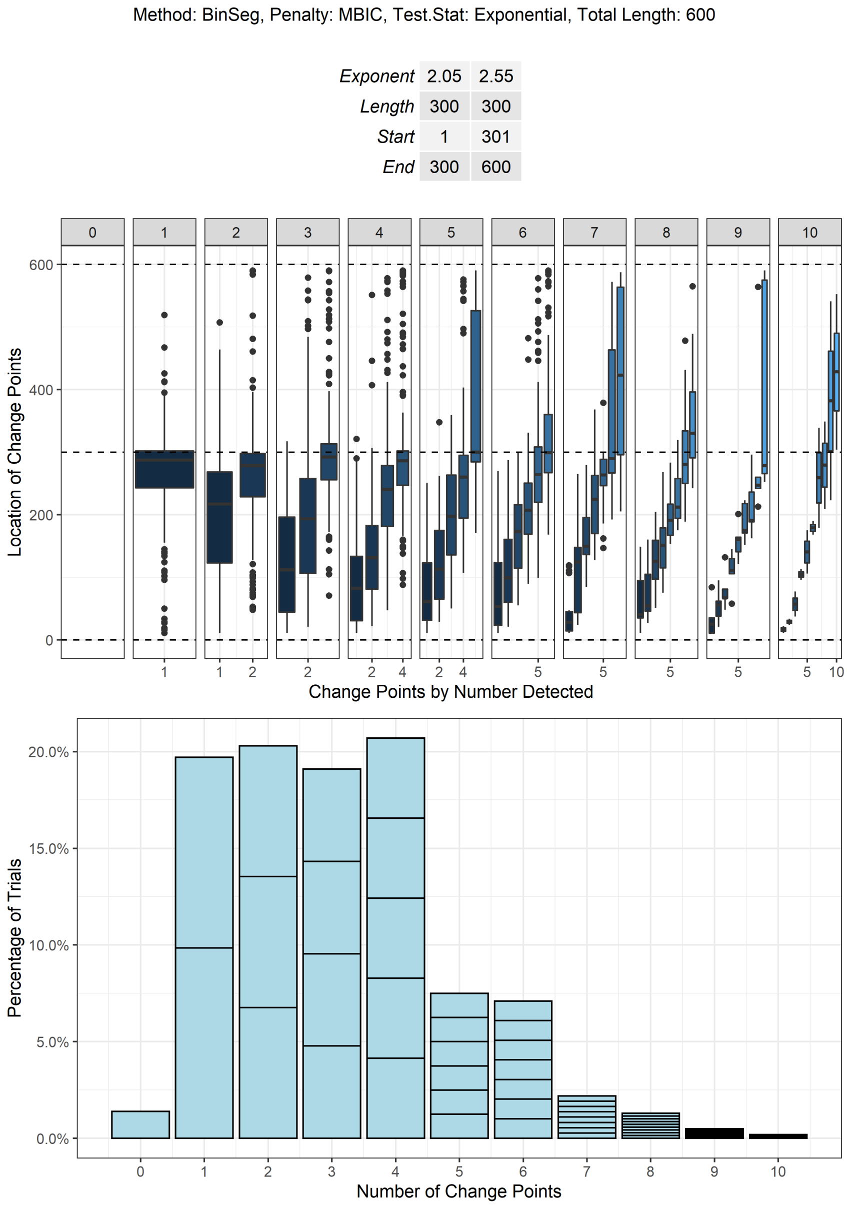
Type	Name	Calculation and Usage
Penalty	AIC	$2p$, intended for model selection problems (Akaike, 1974)
	SIC/BIC	$p\log(n)$, response to Akaike (1974) using Bayesian context (Schwarz, 1978)
	Hannan-Quinn	$2p\log\log(n)$, intended for auto-regression problems (Hannan and Quinn, 1979)
	Asymptotic	For single changepoint detection, use the false positive rate and asymptotic distribution to approximate a penalty (Hinkley, 1970)
	CROPS	Changepoints for a range of penalties (Haynes *et al.*, 2017a)
Method	AMOC	At most one changepoint (Killick *et al.*, 2016)
	SegNeigh	Use dynamic programming to exactly calculate an additional changepoint by calculating optimal models and reusing the previous calculation, for protein segmentation (Auger and Lawrence, 1989)
	BinSeg	Recursively calculate changepoints for detected segments (Scott and Knott, 1974; Sen and Srivastava, 1975)
	PELT	Optimal partitioning with pruning (Killick *et al.*, 2012)
Test Statistic	Normal	The distribution.
	Exponential	The distribution.
	Poisson	The distribution.
	CUSUM	Detect a change in a parameter (e.g. mean) if its cumulative sum passes a threshold (Page, 1954)
	CSS	Detect changes in variance if the normalised cumulative sum of squares passes a threshold (repeatedly) (Inclán and Tiao, 1994)
	Empirical Distribution	Tail biased non-parametric quantile PELT (Haynes *et al.*, 2017b)

D Changepoint analysis and higher minimum values

In the main text in Section 4.3, we note that we present the simulations with minimum value attainable by the power-law distribution of 10, as opposed to a minimum value of 1,000 that is more in-line with the datasets examined. Here, we note some of the differences that one can detect that are due solely to this change in minimum value.

First, we compare and contrast Figures 4.2 – 4.5 in the main text with Figures D.1 – D.4. These figures use the same methods, but the simulations are now two power-laws with exponents 2.05 and 2.55 respectively and shared minimum value 1,000. The results are mostly comparable to the main paper with the exception of Figure D.2. By comparison to Figure 4.3, this combination greatly benefits from the increased minimum value in that overfitting is drastically reduced, albeit not to a usable level. The other combinations show very little change.

Figures 4.3 and D.2 illustrate the most dramatic change, but the change can be abruptly seen in the other direction as well (see Figures D.5 and D.6), perhaps due in both cases to the parametric assumptions of the relevant methods. These changes never modify our existing judgement of what are considered good methods, as the changes never give sufficiently large improvements to make a combination serviceable in our case.



Method: BinSeg, Penalty: MBIC, Test.Stat: Exponential, Total Length: 600

Exponent	2.05	2.55
Length	300	300
Start	1	301
End	300	600

Figure D.1: 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 respectively and minimum value of 1,000. Segmentation generated using `cpt.meanvar` with BinSeg, mBIC and an exponential distribution. Whilst there are good aspects to this finding, the method commonly overfits and tends to assume changepoints happen in the $\alpha = 2.05$ segment. This combination performs slightly better than its counterpart in Figure 4.2, with median Hausdorff of 186 (compared to 189), median ARI of 0.63 (0.64), and TDR 0: 0.01 (0.01), 3: 0.05 (0.04), 5: 0.06 (0.05), 8: 0.09 (0.07). The segments in the lower image correspond to the number of changepoints detected and are guides for the eye only.

Method: SegNeigh, Penalty: Asymptotic, Test.Stat: Normal, Total Length: 600

Exponent	2.05	2.55
Length	300	300
Start	1	301
End	300	600

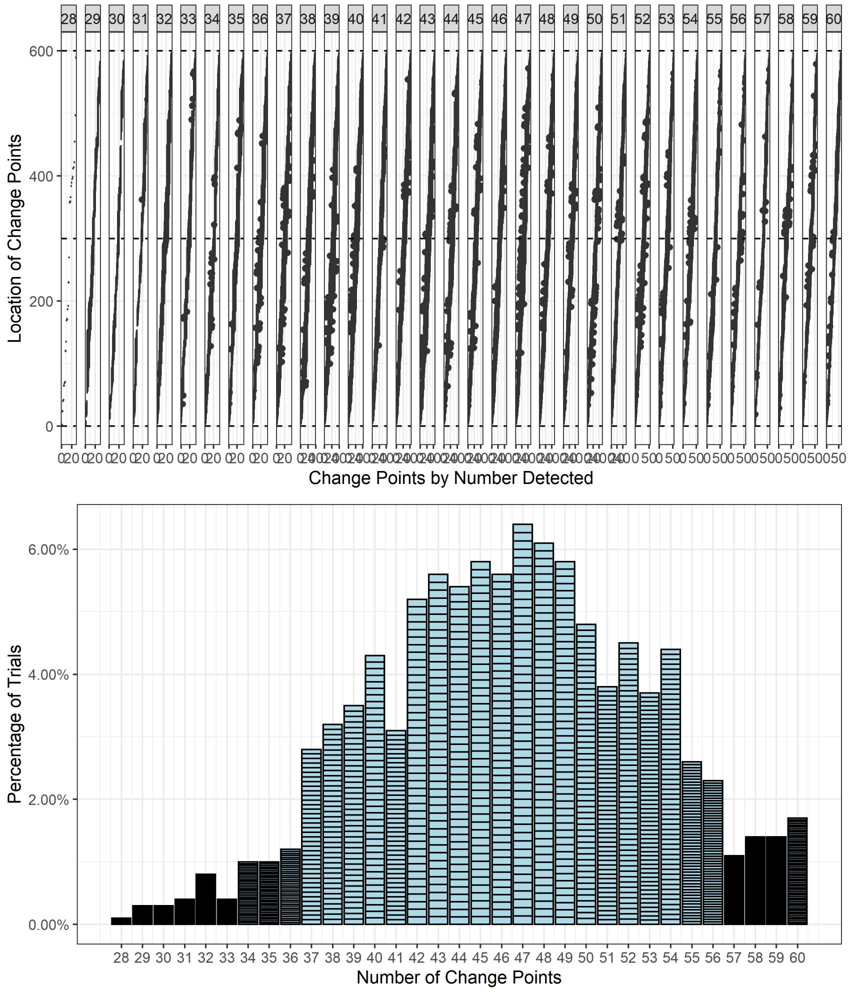
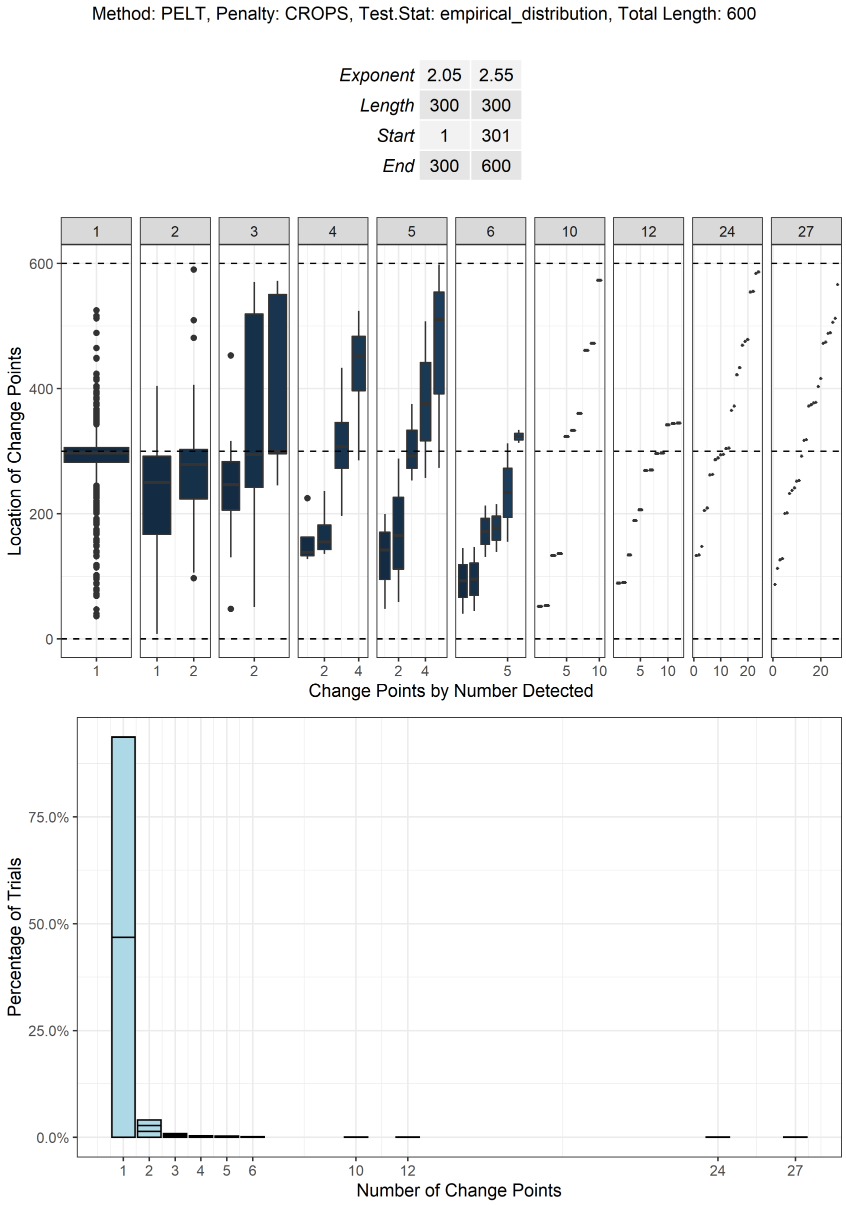
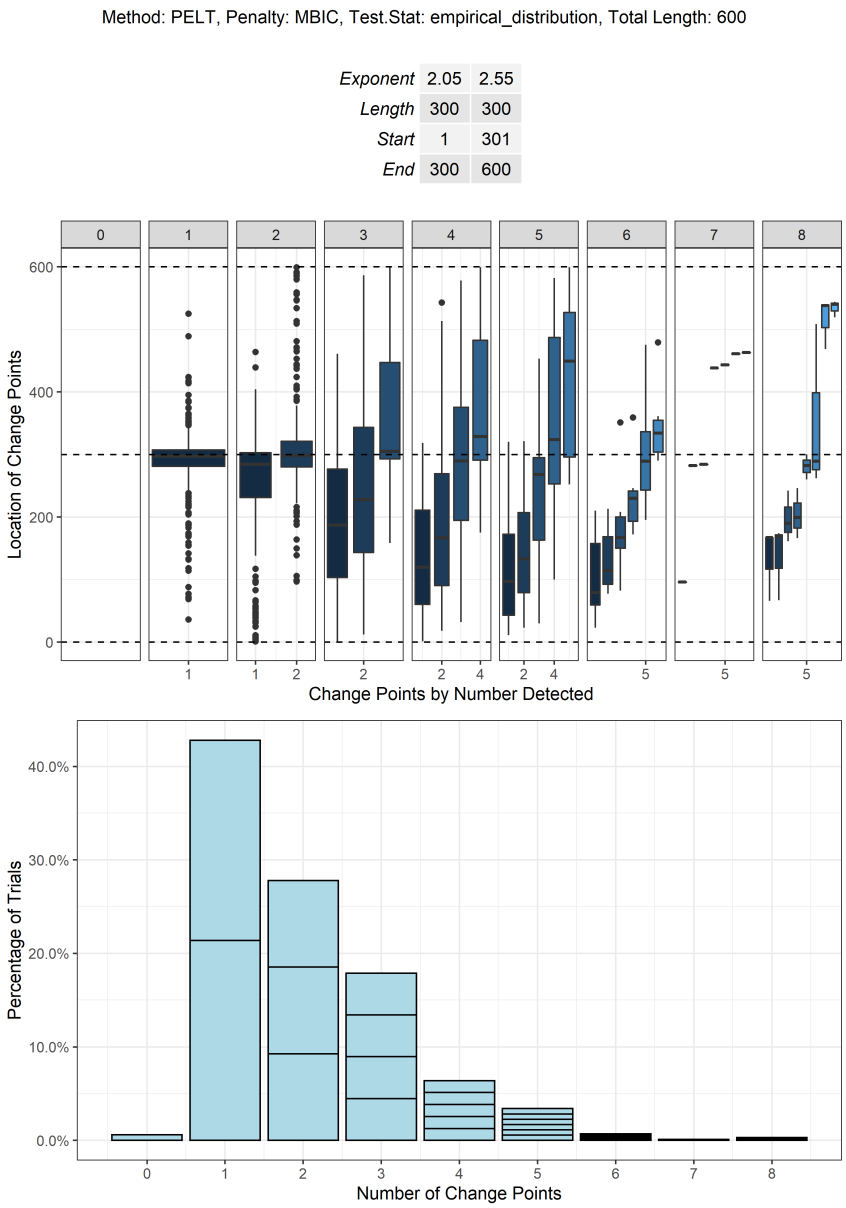


Figure D.2: Test case with worst behaviour. The simulation is as in Figure D.1. Segmentation generated using `cpt.meanvar` with SegNeigh, an asymptotic penalty and a normal distribution. Results such as this occur with many combinations, and can be regarded as failures. Many combinations still result in more than 10 false positives and are only stopped by the maximums provided. This combination performs significantly better than its counterpart Figure 4.3, although this is only noticeable from the number of changepoints detected; its median Hausdorff is 294 (compared to 294), median ARI is 0.08 (0.07), and TDRs are 0: 0.00 (0.00), 3: 0.01 (0.01), 5: 0.01 (0.01), 8: 0.02 (0.01).



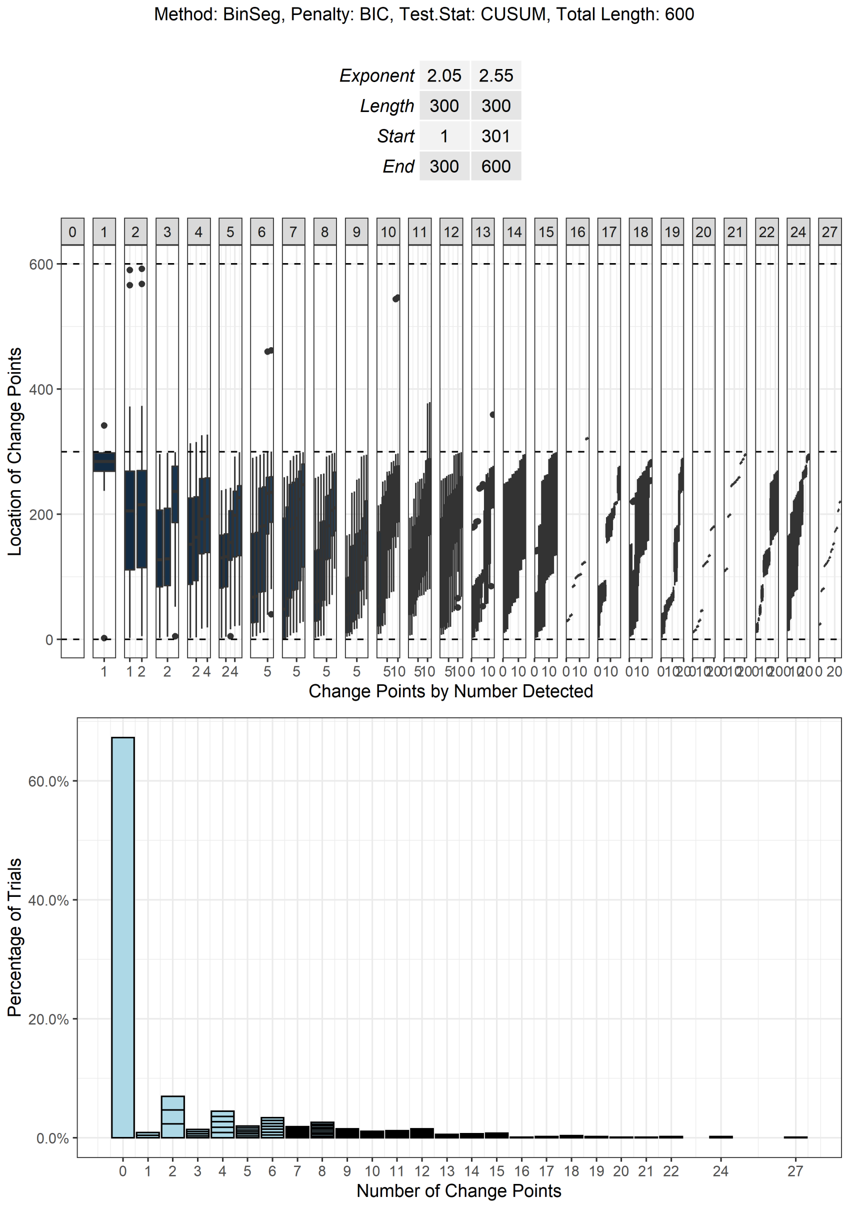
Exponent	2.05	2.55
Length	300	300
Start	1	301
End	300	600

Figure D.3: Test case with best behaviour. The simulation is as in Figure D.1. Segmentation generated using `cpt.np` with ED-PELT and CROPS and shows some of the best achievable behaviour. Although qualitatively similar to the top sub-plot of Figure D.1, there is improved accuracy in the positioning of the changepoints and improved precision and accuracy in the number of points so detected. Increasing the minimum value improves this combination leading to median Hausdorff of 13 (compared to 15 in Figure 4.4), median ARI of 0.92 (0.90), and TDR 0: 0.04 (0.04), 3: 0.19 (0.18), 5: 0.28 (0.25), 8: 0.36 (0.33).



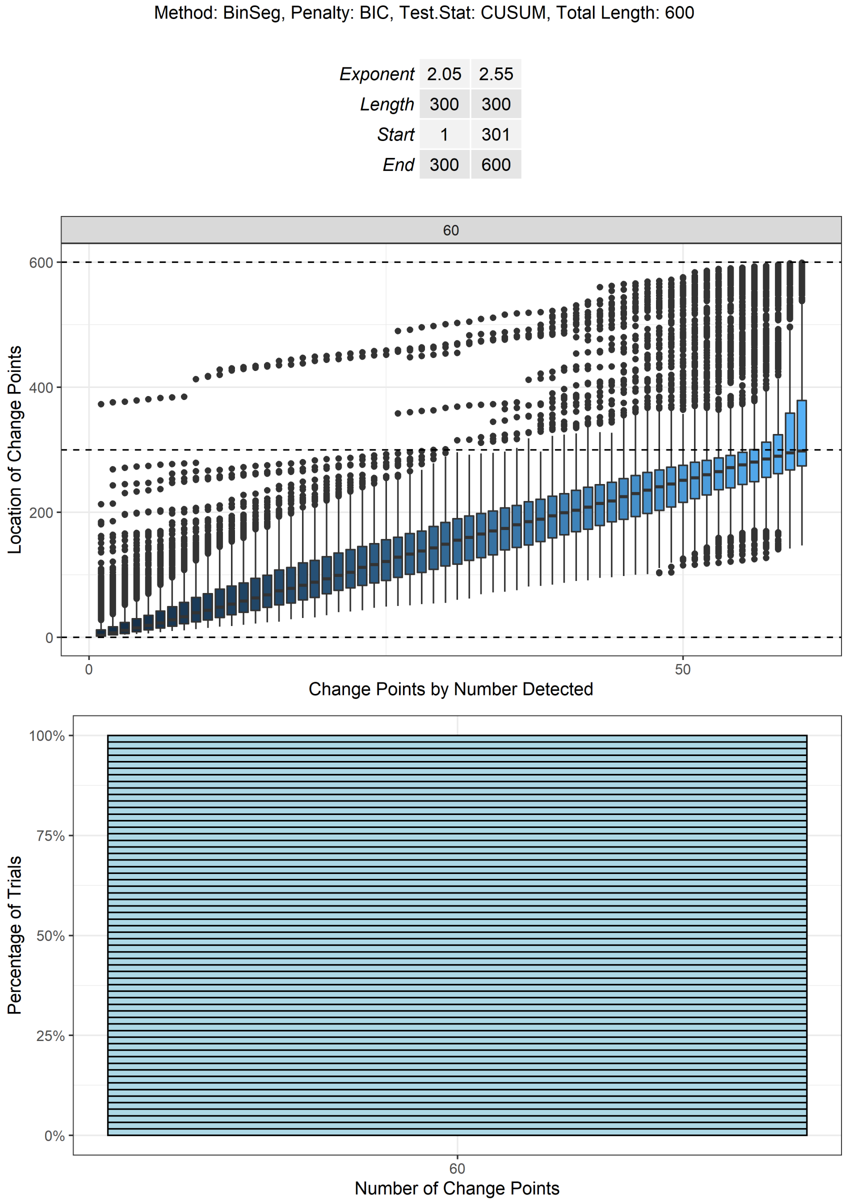
Exponent	2.05	2.55
Length	300	300
Start	1	301
End	300	600

Figure D.4: Test case with second best behaviour. The simulation is as in Figure D.1. Segmentation generated using `cpt.np` with ED-PELT and mBIC. While not as good at detecting changepoints as CROPS, `cpt.np` with ED-PELT and mBIC still shows strong potential. Increasing the minimum value has again improved the statistics slightly compared to Figure 4.5, with median Hausdorff of 50 (compared to 51.5), median ARI of 0.81 (0.80), and TDR 0: 0.02 (0.02), 3: 0.10 (0.09), 5: 0.15 (0.13), 8: 0.19 (0.17).



Exponent	2.05	2.55
Length	300	300
Start	1	301
End	300	600

Figure D.5: Example where increasing the minimum value decreases performance (I). The simulation is as in Figure D.1, but with a minimum value of 10, as in the main text. Segmentation generated using `cpt.mean` with binary segmentation and CUSUM. In this case we observe clear underfitting, but with some occasional overfitting.



Exponent	2.05	2.55
Length	300	300
Start	1	301
End	300	600

Figure D.6: Example where increasing the minimum value decreases performance (II). The simulation is as in Figure D.1. Segmentation generated using `cpt.mean` with binary segmentation and CUSUM. In this case, we observe only extreme overfitting.

List of acronyms

Notation	Description	Page List
ADF	Augmented Dickey-Fuller stationarity test	149, 150
AIC	Akaike Information Criterion	35, 161
ARI	Adjusted Rand Index, measurement of clustering that accounts for chance	79–86, 89, 90, 164–167
BC$_a$	Bias Corrected and Accelerated bootstrap method	31, 61, 66, 73
BIC	Bayesian (or Schwarz's) Information Criterion	35, 36, 161
C-F	Coalescence and Fragmentation	119, 121, 122, 125
CDF	Cumulative Distribution Function	35
CF	CounterFactual scenario	53, 54, 56, 58, 60, 72, 73
CoW	Correlates of War (Sarkees and Wayman, 2010)	78, 79, 90, 95, 103–109
CROPS	Changepoints over a Range Of PenaltieS	36, 80, 81, 84–92, 95, 97–99, 103, 104, 106, 109, 161, 166, 167
CRP	Chinese Restaurant Process, generates power-law distributed partitions (Pitman, 2006; Goldwater *et al.*, 2011)	126, 131, 132, 135, 136
ED-PELT	Algorithm to solve changepoint optimisation for Empirical Distributions using Pruning, Exactly, and in Linear Time. See also PELT	36, 81, 84–86, 90–94, 150, 161, 166, 167
KPSS	Kwiatkowski-Phillips-Schmidt-Shin non-stationarity test	149–151
KSMLE	Kolmogorov-Smirnov Maximum Likelihood Estimation (Clauset *et al.*, 2009b)	27, 124–137, 139
mBIC	Modified Bayesian Information Criterion	36, 81, 85–90, 93–95, 100, 103, 104, 106, 109, 150, 161, 164, 167
MLE	Maximum Likelihood Estimation	10, 27, 124–131, 133–137, 139
PELT	Algorithm to solve changepoint optimisation using Pruning, Exactly, and in Linear Time	35, 36, 81, 106, 161, 171

Notation	Description	Page List
RAF	(British) Royal Air Force	43–49, 53, 58, 60, 72, 143
TDR	True Detection Rate, how many points in one set are near those of another.	79–86, 89, 90, 164–167

List of notation

Notation	Description	Page List
α	Theoretical exponent of a power-law distribution	26–28, 80–82, 86, 89, 103, 104, 120, 124–132, 134, 135, 137, 139
B	Number of bootstrap resamples	29, 30
BS	Effective British (pilot) strength	50, 54, 173
BS_C	Critical value of BS, below which invasion can occur	50, 51, 54
$\dot{}$	Dot. Time derivative. Usually the time argument is dropped	37, 38, 40, 41, 113, 115, 116, 129, 130
$\mathcal{E}^{i,j}$	Step operator: changes argument index i by amount j in the subsequent function	116, 117
$F(i)$	Fragmentation rate kernel for clusters of size i	40–42, 113, 115, 119, 123, 129, 130
$\hat{}$	Hat. Likelihood related estimator	27, 35, 36
IID	Independently and identically distributed	19, 20, 23, 28, 29, 47–49, 67, 147, 149, 151
$\mathcal{K}$	Cyclicity order parameter, ratio of computational steps that result in increases to $k_{\max}$ minus those that result in decreases, all divided by the total number of computational steps	122, 124–126, 129–132, 134, 135, 138
κ	Maximum attainable cluster size	10, 116–118
$K(i,j)$	Coalescence rate kernel for clusters of sizes i and j	38–42, 113, 115, 119, 123, 129, 130
$k_{\max}$	Empirical maximum of a coalescence and fragmentation process	120–122, 124, 125, 129–131, 133, 134, 137, 138, 173
M	System size; amount of mass or total number of possible monomers in simulations, which we consider to have unit volume. In equations, this is more appropriately a density.	10, 37, 39–41, 114–129, 131–133, 137, 138

Notation	Description	Page List
N	Number of discrete objects or clusters	28, 120, 121, 125, 129, 134
n_k	Number of objects or clusters of size k	37, 38, 40, 41, 113–116, 129, 130
Pr	Probability	10, 26, 115, 116, 118, 124–128, 131–134, 137
r	Dimensionless parameter formed from the ratio of characteristic times of gelation and shattering	122–125, 127, 132, 138
se	Standard error	29–31
τ	Location of a changepoint	35, 36, 80, 86, 89, 90
$\tilde{\cdot}$	Tilde. Bootstrapping related estimator	29–32
$x_{\min}$	Theoretical minimum of a power-law distribution	10, 26, 27, 125–134, 136

References

J. H. Abel, B. Drawert, A. Hellander, and L. R. Petzold. GillesPy: A Python package for stochastic model building and simulation. *IEEE Life Sciences Letters*, 2(3):35–38, 2016. doi: 10.1109/LLS.2017.2652448.

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716 – 723, December 1974. ISSN 0018-9286. doi: 10.1109/TAC.1974.1100705.

D. J. Aldous. Deterministic and stochastic models for coalescence (aggregation and coagulation): A review of the mean-field theory for probabilists. *Bernoulli*, 5(1):3–48, Feb. 1999. URL http://www.jstor.org/stable/3318611.

M. S. Alexander. After Dunkirk: The French army's performance against Case Red, 25 May to 25 June 1940. *War in History*, 14(2):219 – 264, 2007.

M. Arltová and D. Fedorová. Selection of unit root test on the basis of time series length and value of AR(1) parameter. (3), 2016. URL https://www.czso.cz/csu/czso/statistika-statistics-and-economy-journal-no-32016.

K. B. Athreya. Bootstrap of the mean in the infinite variance case. *Ann. Statist.*, 15(2):724 – 731, 06 1987. doi: 10.1214/aos/1176350371. URL https://doi.org/10.1214/aos/1176350371.

I. E. Auger and C. E. Lawrence. Algorithms for the optimal identification of segment neighborhoods. *Bulletin of Mathematical Biology*, 51(1):39 – 54, Jan 1989. ISSN 1522-9602. doi: 10.1007/BF02458835. URL https://doi.org/10.1007/BF02458835.

P. Bak. *How Nature Works: the science of self-organized criticality*. Copernicus, 1996.

J. M. Ball, J. Carr, and O. Penrose. The Becker-Döring cluster equations: Basic properties and asymptotic behaviour of solutions. *Communications in Mathematical Physics*, 104:657 – 692, 12 1986. ISSN 1432-0916. doi: 10.1007/BF01211070. URL https://doi.org/10.1007/BF01211070.

A. Baudry. *The Naval Battle, studies of the tactical factors*. 1914.

S. Beard. Is there really evidence for a decline of war?, May 2018. URL https://oefresearch.org/think-peace/evidence-decline-war.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289 – 300, 1995. ISSN 00359246. URL http://www.jstor.org/stable/2346101.

T. Bernauer and N. P. Gleditsch. New event data in conflict research. *International Interactions*, 38(4):375 – 381, 2012. doi: 10.1080/03050629.2012.696966. URL https://doi.org/10.1080/03050629.2012.696966.

P. J. Bickel and D. A. Freedman. Some asymptotic theory for the bootstrap. *The Annals of Statistics*, 9(6):1196 – 1217, 11 1981. ISSN 00905364. doi: 10.1214/aos/1176345637. URL http://www.jstor.org/stable/2240410.

T. Birnstiel, C. W. Ormel, and C. P. Dullemond. Dust size distributions in coagulation/fragmentation equilibrium: numerical solutions and analytical fits. *Astronomy & Astrophysics*, 525:A11, 2011. doi: 10.1051/0004-6361/201015228. URL https://doi.org/10.1051/0004-6361/201015228.

J. C. Bohorquez, S. Gourley, A. R. Dixon, M. Spagat, and N. F. Johnson. Common ecology quantifies human insurgency. *Nature*, (462):911–914, 2009. doi: 10.1038/nature08631. URL http://www.nature.com/nature/journal/v462/n7275/full/nature08631.html.

N. Brilliantov, P. L. Krapivsky, A. Bodrova, F. Spahn, H. Hayakawa, V. Stadnichuk, and J. Schmidt. Size distribution of particles in Saturn's rings from aggregation and fragmentation. *Proceedings of the National Academy of Sciences*, 112(31):9536–9541, 2015. ISSN 0027-8424. doi: 10.1073/pnas.1503957112. URL https://www.pnas.org/content/112/31/9536.

S. Bungay. *The Most Dangerous Enemy: a History of the Battle of Britain*. Aurum, 2000.

Campaign Diaries. RAF campaign diaries. URL raf.mod.uk/history/campaigndiaries.cfm. See also https://web.archive.org/web/20170313205334/http://www.raf.mod.uk:80/campaign/battle-of-britain-75th/bob-campaign-diary/.

N. C. G. Campbell and K. J. Roberts. Lanchester market structures: A Japanese approach to the analysis of business competition. *Strategic Management Journal*, 7(3):189 – 200, 1986. ISSN 01432095, 10970266. URL http://www.jstor.org/stable/2486072.

G. Capoccia and R. D. Kelemen. The study of critical junctures: Theory, narrative, and counterfactuals in historical institutionalism. *World Politics*, 59(3):341 – 369, 2007. doi: 10.1017/S0043887100020852.

T. Carlyle. *On Heroes, Hero-worship, and the Heroic in History*. Project Gutenberg, 1840. URL https://www.gutenberg.org/files/1091/1091-h/1091-h.htm. Published online 2012.

J. Carr and F. P. da Costa. Instantaneous gelation in coagulation dynamics. *Zeitschrift für angewandte Mathematik und Physik ZAMP*, 43, 11 1992. ISSN 1420-9039. doi: 10.1007/BF00916423. URL https://doi.org/10.1007/BF00916423.

L.-E. Cederman, K. S. Gleditsch, and J. Wucherpfennig. Predicting the decline of ethnic civil war: Was Gurr right and for the right reasons? *Journal of Peace Research*, 54(2):262 – 274, 2017. doi: 10.1177/0022343316684191. URL https://doi.org/10.1177/0022343316684191.

J. V. Chase. *Sea Fights: A Mathematical Investigation of the Effect of Superiority of Force in Combats upon the Sea*. US Naval War College, 1902.

A. Chatterjee and B. K. Chakrabarti. Fat tailed distributions for deaths in conflicts and disasters. *Reports in Advances of Physical Sciences*, 01(01):1740007, 2017. doi: 10.1142/S2424942417400072. URL https://doi.org/10.1142/S2424942417400072.

J. Chen and A. K. Gupta. *Parametric Statistical Change Point Analysis*. Birkhauser, 2000. doi: 10.1007/978-1-4757-3131-6.

W. Churchill. *Their Finest Hour: The Second World War*. Cassell, 2nd edition, 1950.

P. Cirillo and N. N. Taleb. What are the chances of war? *Significance*, 13(2):44 – 45, 2016a. doi: 10.1111/j.1740-9713.2016.00903.x. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1740-9713.2016.00903.x`.

P. Cirillo and N. N. Taleb. On the statistical properties and tail risk of violent conflicts. *Physica A: Statistical Mechanics and its Applications*, 452:29 – 45, 2016b. ISSN 0378-4371. doi: 10.1016/j.physa.2016.01.050. URL `http://www.sciencedirect.com/science/article/pii/S0378437116000923`.

A. Clauset. Trends and fluctuations in the severity of interstate wars. *Science Advances*, 4(2), 2018. doi: 10.1126/sciadv.aao3580. URL `http://advances.sciencemag.org/content/4/2/eaao3580`.

A. Clauset and K. S. Gleditsch. The developmental dynamics of terrorist organizations. *PLOS ONE*, 7(11):1–11, 11 2012. doi: 10.1371/journal.pone.0048633. URL `https://doi.org/10.1371/journal.pone.0048633`.

A. Clauset and K. S. Gleditsch. Trends in conflict: What do we know and what can we know? In A. Gheciu and W. C. Wohlforth, editors, *The Oxford Handbook of International Security*. Oxford University Press, 2018. ISBN 9780198777854.

A. Clauset and F. W. Wiegel. A generalized aggregation-disintegration model for the frequency of severe terrorist attacks. *Journal of Conflict Resolution*, 54(1):179 – 197, 2010. doi: 10.1177/0022002709352452. URL `http://dx.doi.org/10.1177/0022002709352452`.

A. Clauset, M. Young, and K. S. Gleditsch. On the frequency of severe terrorist events. *Journal of Conflict Resolution*, 51(1):58–87, 2007. doi: 10.1177/0022002706296157. URL `https://doi.org/10.1177/0022002706296157`.

A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661 – 703, 2009a. ISSN 00361445, 10957200. URL `http://www.jstor.org/stable/25662336`.

A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661 – 703, 2009b. ISSN 00361445, 10957200. URL `https://www.jstor.org/stable/25662336`.

A. Clauset, M. Young, and K. S. Gleditsch. A novel explanation of the power-law form of the frequency of severe terrorist events: Reply to Saperstein. *Peace Economics, Peace Science and Public Policy*, 16(1), December 2010. doi: 10.2202/1554-8597.1213. URL `https://www.degruyter.com/view/journals/peps/16/1/article-peps.2010.16.1.1213.xml.xml`.

G. Clayton, J. Kathman, K. Beardsley, T.-I. Gizelis, L. Olsson, V. Bove, A. Ruggeri, R. Zwetsloot, J. van der Lijn, T. Smit, L. Hultman, H. Dorussen, A. Ruggeri, P. Diehl, L. Bosco, and C. Goodness. The known knowns and known unknowns of peacekeeping data. *International Peacekeeping*, 24(1):1 – 62, 2017. doi: 10.1080/13533312.2016.1226768. URL https://doi.org/10.1080/13533312.2016.1226768.

R. Cox. *Operation Sealion*. Thornton Cox, 1974.

M. Cristelli, M. Batty, and L. Pietronero. There is more than a power law in Zipf. *Scientific Reports*, 2(812), 11 2012. doi: 10.1038/srep00812. URL http://dx.doi.org/10.1038/srep00812.

C. Crook. The height of inequality. *The Atlantic Monthly*, September 2006. URL https://www.theatlantic.com/magazine/archive/2006/09/the-height-of-inequality/305089/.

A. J. Cumming. *The Royal Navy and the Battle of Britain*. Naval Institute Press, 2010.

C. Cunen, N. L. Hjort, and H. M. Nygård. Statistical sightings of better angels: Analysing the distribution of battle-deaths in interstate conflict over time. *Journal of Peace Research*, 57(2): 221–234, 2020. doi: 10.1177/0022343319896843. URL https://doi.org/10.1177/0022343319896843.

F. P. da Costa. Existence and uniqueness of density conserving solutions to the coagulation-fragmentation equations with strong fragmentation. *Journal of Mathematical Analysis and Applications*, 192(3):892 – 914, 1995. ISSN 0022-247X. doi: 10.1006/jmaa.1995.1210. URL http://www.sciencedirect.com/science/article/pii/S0022247X85712103.

Daily Weather Reports. Daily weather reports, 1940. URL digital.nmla.metoffice.gov.uk/archive/sdb%3AdeliverableUnit%7Ceda9f47f-c326-4991-ada4-a3e4c8bd11d9/.

S. Datta, G. W. Delius, R. Law, and M. J. Plank. A stability analysis of the power-law steady state of marine size spectra. *Journal of Mathematical Biology*, 63(4):779–799, Oct 2011. ISSN 1432-1416. doi: 10.1007/s00285-010-0387-z. URL https://doi.org/10.1007/s00285-010-0387-z.

S. C. Davies, J. R. King, and J. A. D. Wattis. Self-similar behaviour in the coagulation equations. *Journal of Engineering Mathematics*, 36:57 – 88, 1999. ISSN 1573-2703. doi: 10.1023/A:1004589822425. URL https://doi.org/10.1023/A:1004589822425.

L. Deighton. *Fighter: The True Story of the Battle of Britain*. Jonathan Cape, 1977.

S. J. Deitchman. A Lanchester model of guerrilla warfare. *Operations Research*, 10(6):818 – 827, 1962. ISSN 0030364X, 15265463. URL http://www.jstor.org/stable/168104.

D. D. Dempster and D. Wood. *The Narrow Margin: the Battle of Britain and the rise of air power 1930-1940*. Hutchinson, 1961.

K. Deng. Fact or fiction? *Working Paper*, 76(3), July 2003.

R. D'Hulst and G. J. Rodgers. Exact solution of a model for crowding and information transmission in financial markets. *International journal of theoretical and applied finance*, 3:609 – 616, 2000.

H. Dowding. *Enemy Air Offensive Against Great Britain*, 1941-1947. Held as AIR 2/7771.

H. Dowding. Battle of Britain despatch. *Royal Air Force Air Power Review*, 18(2), 2015.

P. B. Dubovski. *Mathematical Theory of Coagulation*. 1994. URL https://www.researchgate.net/publication/265038723_Mathematical_Theory_of_Coagulation_1.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1 – 26, 01 1979a. doi: 10.1214/aos/1176344552. URL http://www.jstor.org/stable/2958830.

B. Efron. Computers and the theory of statistics: Thinking the unthinkable. *SIAM Review*, 21(4): 460 – 480, 1979b. ISSN 00361445. URL http://www.jstor.org/stable/2030104.

B. Efron. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171 – 185, 1987. ISSN 01621459. URL http://www.jstor.org/stable/2289144.

B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. 1993.

J. M. Epstein. Why model? *Journal of Artificial Societies and Social Simulation*, 11(4):12, 2008. ISSN 1460-7425. URL http://jasss.soc.surrey.ac.uk/11/4/12.html.

R. Epstein. Book review: The better angels of our nature: Why violence has declined. *Scientific American*, October 2011.

M. M. Evans and A. McGeoch. *Invasion! Operation Sea Lion, 1940*. Routledge, 2004.

R. J. Evans. *Altered pasts*. Brandeis University Press, 2014.

B. Fagan, I. Horwood, N. MacKay, C. Price, E. Richards, and A. J. Wood. Bootstrapping the Battle of Britain. *Journal of Military History*, 84(1):151 – 86, January 2020a.

B. T. Fagan, M. I. Knight, N. J. MacKay, and A. J. Wood. Change point analysis of historical battle deaths. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(3):909 – 933, 2020b. doi: 10.1111/rssa.12578. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssa.12578.

B. T. Fagan, N. J. MacKay, D. O. Pushkin, and A. J. Wood. Stochastic gel-shatter cycles in coalescence-fragmentation models. *Europhysics Letters*, Under Review.

J. Fenby. *The Penguin history of modern China*. Penguin Books, second edition, 2013. ISBN 9780141975153.

B. A. Fiske. *The Navy as a Fighting Machine*. Charles Scribner's Sons, 1916.

R. Forczyk. *We March Against England: Operation Sea Lion, 1940–41*. Bloomsbury, 2016.

C. S. Forester. If Hitler had invaded England. In *Gold from Crete*. Pan, 1971.

J. L. Gaddis. The long peace: Elements of stability in the postwar international system. *International Security*, 10(4):99 – 142, 1986. ISSN 01622889, 15314804. URL `http://www.jstor.org/stable/2538951`.

C. Gallagher, R. Lund, and M. Robbins. Changepoint detection in climate time series with long-term trends. *Journal of climate*, 26(14):4994?5006, Jul 2013. ISSN 0894-8755. doi: 10.1175/JCLI-D-12-00704.1. URL `https://journals.ametsoc.org/view/journals/clim/26/14/jcli-d-12-00704.1.xml`.

A. Gat. Is war declining - and why? *Journal of Peace Research*, 50(2):149 – 157, 2013. doi: 10.1177/0022343312461023. URL `https://doi.org/10.1177/0022343312461023`.

German Plans. *German Plans for the Invasion of England: Operation Sealion, 1940*, 2017.

C. Gillespie. Fitting heavy tailed distributions: The poweRlaw package. *Journal of Statistical Software, Articles*, 64(2):1 – 16, 2015. ISSN 1548-7660. doi: 10.18637/jss.v064.i02. URL `https://www.jstatsoft.org/v064/i02`.

C. Gillespie. *poweRlaw: Analysis of Heavy Tailed Distributions*, 2017.

D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977. doi: 10.1021/j100540a008. URL `http://pubs.acs.org/doi/abs/10.1021/j100540a008`.

K. S. Gleditsch. A revised list of wars between and within independent states, 1816-2002. *International Interactions*, 30(3):231 – 262, 2004. doi: 10.1080/03050620490492150. URL `https://doi.org/10.1080/03050620490492150`.

K. S. Gleditsch and S. Pickering. Wars are becoming less frequent: a response to Harrison and Wolf. *The Economic History Review*, 67(1):214 – 230, 2014. doi: 10.1111/1468-0289.12002. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0289.12002`.

K. S. Gleditsch, N. W. Metternich, and A. Ruggeri. Data and progress in peace and conflict research. *Journal of Peace Research*, 51(2):301–314, 2014. doi: 10.1177/0022343313496803. URL `https://doi.org/10.1177/0022343313496803`.

N. P. Gleditsch, P. Wallensteen, M. Eriksson, M. Sollenberg, and H. Strand. Armed conflict 1946-2001: A new dataset. *Journal of Peace Research*, 39(5):615 – 637, 2002. doi: 10.1177/0022343302039005007. URL `https://doi.org/10.1177/0022343302039005007`.

A. Gohdes and M. Price. First things first: Assessing data quality before model quality. *The Journal of Conflict Resolution*, 57(6):1090 – 1108, 2013. ISSN 00220027, 15528766. URL `http://www.jstor.org/stable/24545604`.

J. S. Goldstein. Kondratieff waves as war cycles. *International Studies Quarterly*, 29(4):411 – 444, 1985. ISSN 00208833, 14682478. URL http://www.jstor.org/stable/2600380.

J. S. Goldstein. *Winning the War on War: The Decline of Armed Conflict Worldwide*. Penguin Publishing Group, 2011. ISBN 9781101549087.

S. Goldwater, T. L. Griffiths, and M. Johnson. Producing power-law distributions and damping word frequencies with two-stage language models. *Journal of Machine Learning Research*, 12:2335 – 2382, July 2011. URL http://www.jmlr.org/papers/v12/goldwater11a.html.

E. González and M. Villena. Spatial Lanchester models. *European Journal of Operational Research*, 210(3):706 – 715, 2011. ISSN 0377-2217. doi: 10.1016/j.ejor.2010.11.009. URL http://www.sciencedirect.com/science/article/pii/S0377221710007666.

R. González-Val. War size distribution: Empirical regularities behind conflicts. *Defence and Peace Economics*, 27(6):838 – 853, 2016. doi: 10.1080/10242694.2015.1025486. URL https://doi.org/10.1080/10242694.2015.1025486.

C. Goulter, A. Gordon, and G. Sheffield. The Royal Navy did not win the 'Battle of Britain'. *Journal of the Royal United Services Institute*, 151(5):66 – 67, 2006.

C. S. Gray. *Another Bloody Century: Future Warfare*. Hachette UK, 2012. ISBN 9781780223919.

D. W. Grinnell-Milne. *The Silent Victory: September 1940*. Bodley Head, 1958.

H. Guderian. *Panzer Leader*. Penguin, 2009.

T. R. Gurr. Ethnic warfare on the wane. *Foreign Affairs*, 79(3):52 – 64, 2000.

E. J. Hannan and B. G. Quinn. The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):190 – 195, 1979. ISSN 00359246. URL http://www.jstor.org/stable/2985032.

M. Harrison and N. Wolf. The frequency of wars. *The Economic History Review*, 65(3):1055 – 1076, 2012. doi: 10.1111/j.1468-0289.2011.00615.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0289.2011.00615.x.

M. Harrison and N. Wolf. The frequency of wars: reply to Gleditsch and Pickering. *The Economic History Review*, 67(1):231 – 239, 2014. doi: 10.1111/1468-0289.12008. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0289.12008.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2nd edition, 2009. URL https://web.stanford.edu/~hastie/ElemStatLearn/.

K. Haynes, R. Killick, and P. Fearnhead. *changepoint.np: Methods for Nonparametric Changepoint Detection*, 2016.

K. Haynes, I. A. Eckley, and P. Fearnhead. Computationally efficient changepoint detection for a range of penalties. *Journal of Computational and Graphical Statistics*, 26(1):134 – 143, 2017a.

doi: 10.1080/10618600.2015.1116445. URL https://doi.org/10.1080/10618600.2015.1116445.

K. Haynes, P. Fearnhead, and I. A. Eckley. A computationally efficient nonparametric approach for changepoint detection. *Statistics and Computing*, 27(5):1293 – 1305, Sep 2017b. ISSN 1573-1375. doi: 10.1007/s11222-016-9687-5. URL https://doi.org/10.1007/s11222-016-9687-5.

P. Hilton and J. Pedersen. Catalan numbers, their generalization, and their uses. *The Mathematical Intelligencer*, 13:64 – 75, 1991. URL https://doi.org/10.1007/BF03024089.

D. V. Hinkley. Inference about the change-point in a sequence of random variables. *Biometrika*, 57 (1):1 – 17, 1970. ISSN 00063444. URL http://www.jstor.org/stable/2334932.

N. L. Hjort. Towards a more peaceful world [insert '!' or '?' here], January 2018. URL http://www.mn.uio.no/math/english/research/projects/focustat/the-focustat-blog!/krigogfred.html.

S. Holwell and P. Checkland. An information system won the war. In *IEE Proceedings - Software*, volume 145, pages 95 – 99, 1998.

Honour Roll. RAF Battle of Britain honour roll. URL raf.mod.uk/campaign/battle-of-britain-75th/the-few/battle-of-britain-roll-of-honour/.

I. Horwood, N. MacKay, and C. Price. Concentration and asymmetry in air combat: Lessons for the defensive employment of air power. *RAF Air Power Review*, 17(2):68 – 91, 2014.

P. M. Hui, N. F. Johnson, and P. Jefferies. *Financial Market Complexity*. OUP Catalogue. Oxford University Press, New York, 2003. ISBN 9780198526650.

S. P. Huntington. No exit: The errors of endism. *The National Interest*, (17):3 – 11, Fall 1989.

G. Imbens and K. Menzel. A causal bootstrap. *arXiv e-prints*, art. arXiv:1807.02737, July 2018. URL https://arxiv.org/abs/1807.02737.

G. Imbens and D. B. Rubin. *Causal inference for statistics, social, and biomedical sciences : an introduction*. New York, NY : Cambridge University Press, 2015. ISBN 9780521885881.

C. Inclán and G. C. Tiao. Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association*, 89(427):913 – 923, 1994. ISSN 01621459. URL http://www.jstor.org/stable/2290916.

B. Jackson, J. D. Scargle, D. Barnes, S. Arabhi, A. Alt, P. Gioumousis, E. Gwin, P. Sangtrakulcharoen, L. Tan, and Tun Tao Tsai. An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters*, 12(2):105–108, 2005.

B. James. Pie in the sky? *History Today*, September 2006.

T. C. G. James. *Royal Air Force Official Histories: Air Defence of Great Britain*. Routledge, new edition, 2000.

I. R. Johnson and N. J. MacKay. Lanchester models and the Battle of Britain. *Naval Research Logistics*, 58:210 – 222, 2011. URL https://doi.org/10.1002/nav.20328.

N. Johnson, M. Spagat, J. Restrepo, J. Bohorquez, N. Suarez, E. Restrepo, and R. Zarama. From old wars to new wars and global terrorism. *ArXiv Physics e-prints*, June 2005.

N. F. Johnson, M. Zheng, Y. Vorobyeva, A. Gabriel, H. Qi, N. Velasquez, P. Manrique, D. Johnson, E. Restrepo, C. Song, and S. Wuchty. New online ecology of adversarial aggregates: ISIS and beyond. *Science*, 352(6292):1459–1463, 2016. ISSN 0036-8075. doi: 10.1126/science.aaf0675. URL http://science.sciencemag.org/content/352/6292/1459.

D. Kahn. *The Reader of Gentlemen's Mail*. Yale University Press, 2004. ISBN 9780300129885. URL http://ebookcentral.proquest.com/lib/york-ebooks/detail.action?docID=3420152.

L. H. Keeley. *War Before Civilization : The Myth of the Peaceful Savage*. Oxford University Press USA - OSO, 1996. ISBN 9780199761531. URL http://ebookcentral.proquest.com/lib/york-ebooks/detail.action?docID=694019.

I. Kershaw. *Fateful choices: ten decisions that changed the world, 1940 – 1941*. Penguin, 2008.

E. Kieser. *Hitler on the Doorstep: Operation "Sea Lion": the German Plan to Invade Britain, 1940*. Naval Institute Press, 1997.

R. Killick, P. Fearnhead, and I. A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590 – 1598, 2012. doi: 10.1080/01621459.2012.737745. URL https://doi.org/10.1080/01621459.2012.737745.

R. Killick, K. Haynes, and I. A. Eckley. *changepoint: An R Package for Changepoint Analysis*, 2016.

K. Klein Goldewijk, A. Beusen, J. Doelman, and E. Stehfest. Anthropogenic land use estimates for the Holocene–HYDE 3.2. *Earth System Science Data*, 9(2):927 – 953, 2017. doi: 10.5194/essd-9-927-2017. URL https://www.earth-syst-sci-data.net/9/927/2017/essd-9-927-2017.html.

A. E. Kyprianou, S. W. Pagett, and T. Rogers. Universality in a class of fragmentation-coalescence processes. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 54(2):1134–1151, 2018.

B. Lacina and N. P. Gleditsch. The waning of war is real: A response to Gohdes and Price. *Journal of Conflict Resolution*, 57(6):1109 – 1127, 2013. doi: 10.1177/0022002712459709. URL https://doi.org/10.1177/0022002712459709.

B. Lacina, N. P. Gleditsch, and B. Russett. The declining risk of death in battle. *International Studies Quarterly*, 50(3):673 – 680, 2006. ISSN 00208833, 14682478. URL http://www.jstor.org/stable/4092798.

F. W. Lanchester. *Aircraft in Warfare*. Constable and Company, 1916. URL https://archive.org/details/aircraftinwarfar00lancrich.

M. Lavielle. Using penalized contrasts for the change-point problem. *Signal Processing*, 85 (8):1501 – 1510, 2005. ISSN 0165-1684. doi: 10.1016/j.sigpro.2005.01.012. URL `http://www.sciencedirect.com/science/article/pii/S0165168405000381`.

A. Lushnikov. Gelation in coagulating systems. *Physica D: Nonlinear Phenomena*, 222(1 - 2):37 – 53, 2006.

A. Lushnikov and V. Piskunov. Analytic solutions in the theory of coagulating systems with sinks. *Journal of Applied Mathematics and Mechanics*, 47(6):743 – 750, 1983. ISSN 0021-8928. doi: 10.1016/0021-8928(83)90110-7. URL `http://www.sciencedirect.com/science/article/pii/0021892883901107`.

A. A. Lushnikov. Coagulation in finite systems. *Journal of Colloid and Interface Science*, 65 (2):276 – 285, 1978. ISSN 0021-9797. doi: 10.1016/0021-9797(78)90158-3. URL `http://www.sciencedirect.com/science/article/pii/0021979778901583`.

N. MacKay, C. Price, and A. J. Wood. Weighing the fog of war: Illustrating the power of Bayesian methods for historical analysis through the Battle of the Dogger Bank. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 49(2):80–91, 2016. doi: 10.1080/01615440.2015.1072071. URL `https://doi.org/10.1080/01615440.2015.1072071`.

N. J. MacKay. Is air combat Lanchestrian? *Phalanx: the Bulletin of Military Operations Research*, 44:12 – 14, 2011.

K. Macksey. *Invasion: The German Invasion of England, July 1940*. Greenhill, 1980.

K. Macksey. *Military Errors of World War Two*. Arms & Armour, 1987.

A. H. Marcus. Stochastic coalescence. *Technometrics*, 10(1):133–143, 1968. ISSN 00401706. URL `http://www.jstor.org/stable/1266230`.

G. Martelloni, F. Di Patti, and U. Bardi. Pattern analysis of world conflicts over the past 600 years. *arXiv e-prints*, art. arXiv:1812.08071, Dec 2018.

S. A. Matveev, P. L. Krapivsky, A. P. Smirnov, E. E. Tyrtyshnikov, and N. V. Brilliantov. Oscillations in aggregation-shattering processes. *Phys. Rev. Lett.*, 119:260601, Dec 2017. doi: 10.1103/PhysRevLett.119.260601. URL `https://link.aps.org/doi/10.1103/PhysRevLett.119.260601`.

E. D. McGrady and R. M. Ziff. “Shattering” transition in fragmentation. *Phys. Rev. Lett.*, 58: 892–895, Mar 1987. doi: 10.1103/PhysRevLett.58.892. URL `https://link.aps.org/doi/10.1103/PhysRevLett.58.892`.

A. J. McKane and T. J. Newman. Predator-prey cycles from resonant amplification of demographic stochasticity. *Phys. Rev. Lett.*, 94:218102, Jun 2005. doi: 10.1103/PhysRevLett.94.218102. URL `https://link.aps.org/doi/10.1103/PhysRevLett.94.218102`.

A. J. McKane, J. D. Nagy, T. J. Newman, and M. O. Stefanini. Amplified biochemical oscillations in cellular systems. *Journal of Statistical Physics*, 128(1):165–191, Jul 2007. ISSN

1572-9613. doi: 10.1007/s10955-006-9221-9. URL https://doi.org/10.1007/s10955-006-9221-9.

J. B. McLeod. On an infinite set of non-linear differential equations. *The Quarterly Journal of Mathematics*, 13(1):119–128, 01 1962. ISSN 0033-5606. doi: 10.1093/qmath/13.1.119. URL https://doi.org/10.1093/qmath/13.1.119.

A. Megill. The new counterfactualists. *Historically Speaking*, 5(4):17 – 18, March 2004. URL https://muse.jhu.edu/article/423436/pdf.

C. Messenger. The Battle of Britain 1940: Triumph of the Luftwaffe. In P. G. Tsouras, editor, *Third Reich Victorious: Alternative Histories of World War II*, pages 65 – 96. Greenhill, 2002.

M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226–251, 2004. doi: 10.1080/15427951.2004.10129088. URL https://doi.org/10.1080/15427951.2004.10129088.

National Consortium for the Study of Terrorism and Responses to Terrorism (START). *Global terrorism database codebook: Inclusion criteria and variables*, 2016.

M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5): 323 – 351, 2005. doi: 10.1080/00107510500052444. URL https://doi.org/10.1080/00107510500052444.

M. E. J. Newman. *Networks: An Introduction*, chapter IV: Network Models. Oxford University Press, 2010.

M. Osipov. The influence of the numerical strength of engaged forces on their casualties. *Tzarist Russian Journal: Military Collection*, 1915. URL https://apps.dtic.mil/dtic/tr/fulltext/u2/a241534.pdf. Translated September 30th, 1991.

Oxford weather station. Oxford weather station. URL www.metoffice.gov.uk/pub/data/weather/uk/climate/stationdata/oxforddata.txt.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100 – 115, 1954. ISSN 00063444. URL http://www.jstor.org/stable/2333009.

J. Pearl. *Causality: models, reasoning, and inference*. Cambridge University Press, Cambridge, 2000. ISBN 0521773628.

R. L. Pego and J. J. L. Velázquez. Temporal oscillations in Becker–Döring equations with atomization. *Nonlinearity*, 33(4):1812–1846, feb 2020. doi: 10.1088/1361-6544/ab6815. URL https://doi.org/10.1088%2F1361-6544%2Fab6815.

P. Picasso. Picasso speaks. In A. H. J. Barr, editor, *Picasso*, pages 9 – 12. The Museum of Modern Art, 1923. URL https://monoskop.org/images/0/0a/Picasso_Forty_Years_of_His_Art_MoMA_1939.pdf. Originally published in The Arts, 3: 315–326.

S. Pinker. *The Better Angels of Our Nature: The Decline of Violence In History And Its Causes*. Penguin Books Limited, 2011. ISBN 9780141959740.

S. Pinker. *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress*. Allen Lane, 2018. ISBN 9780241004319.

J. Pitman. *Combinatorial stochastic processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2006. ISBN 978-3-540-30990-1; 3-540-30990-X. doi: 10.1007/b11601500. URL http://bibserver.berkeley.edu/csp/april05/bookcsp.pdf.

S. R. Platt. *Autumn in the Heavenly Kingdom*. Atlantic Books, 2012. ISBN 9780857897695. URL https://books.google.co.uk/books?id=Oa_RfS6D1zkC.

D. O. Pushkin and H. Aref. Self-similarity theory of stationary coagulation. *Physics of Fluids*, 14(2):694–703, 2002. doi: 10.1063/1.1430440. URL https://doi.org/10.1063/1.1430440.

D. O. Pushkin and H. Aref. Bank mergers as scale-free coagulation. *Physica A: Statistical Mechanics and its Applications*, 336(3):571 – 584, 2004. ISSN 0378-4371. doi: 10.1016/j.physa.2003.12.056. URL http://www.sciencedirect.com/science/article/pii/S0378437103012391.

Python Software Foundation. Python documentation, 2020. URL https://docs.python.org/3/.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL https://www.R-project.org/.

W. G. Ramsey. *The Battle of Britain: Then and Now*. Battle of Britain prints international, 5th edition, 1989.

W. J. Reed and B. D. Hughes. From gene families and genera to incomes and internet file sizes: Why power laws are so common in nature. *Phys. Rev. E*, 66:067103, Dec 2002. doi: 10.1103/PhysRevE.66.067103. URL https://link.aps.org/doi/10.1103/PhysRevE.66.067103.

T. H. Reilly. *The Taiping Heavenly Kingdom*. China Program Bks. University of Washington Press, 2004. ISBN 9780295801926.

L. F. Richardson. The distribution of wars in time. *Journal of the Royal Statistical Society*, 107(3/4):242 – 250, 1944. ISSN 09528385. URL http://www.jstor.org/stable/2981216.

L. F. Richardson. The number of nations on each side of a war. *Journal of the Royal Statistical Society*, 109(2):130 – 156, 1946. ISSN 09528385. URL http://www.jstor.org/stable/2981178.

L. F. Richardson. Contiguity and deadly quarrels: The local pacifying influence. *Journal of the Royal Statistical Society. Series A (General)*, 115(2):219 – 231, 1952. ISSN 00359238. URL http://www.jstor.org/stable/2981156.

L. F. Richardson. *Arms and Insecurity*. The Boxwood Press and Quadrangle Books, 1960a.

L. F. Richardson. *Statistics of deadly quarrels*. Stevenson & Sons, 1960b.

G. Rigaill. A pruned dynamic programming algorithm to recover the best segmentations with 1 to k_max change-points. *Journal de la Société Française de Statistique*, 156(4):180 – 205, 2015. URL `http://journal-sfds.fr/article/view/485/`.

D. B. Rubin. The Bayesian bootstrap. *The Annals of Statistics*, 9(1):130 – 134, 01 1981. doi: 10.1214/aos/1176345338. URL `https://www.jstor.org/stable/2240875`.

B. Ruszczycki, B. Burnett, Z. Zhao, and N. F. Johnson. Relating the microscopic rules in coalescence-fragmentation models to the cluster-size distribution. *The European Physical Journal B*, 72(2):289, 2009. ISSN 1434-6036. doi: 10.1140/epjb/e2009-00354-5. URL `http://dx.doi.org/10.1140/epjb/e2009-00354-5`.

M. R. Sarkees. The COW typology of war: Defining and categorizing wars. (Available at http://www.correlatesofwar.org/data-sets/COW-war [accessed: 16 May 2018]), 2010.

M. R. Sarkees and F. W. Wayman. *Resort to war: a data guide to inter-state, extra-state, intra-state, and non-state wars, 1816-2007*. Correlates of war series. CQ Press, 2010. ISBN 9780872894341.

M. R. Sarkees, F. W. Wayman, and J. D. Singer. Inter-state, intra-state, and extra-state wars: A comprehensive look at their distribution over time, 1816-1997. *International Studies Quarterly*, 47(1):49 – 70, 2003. ISSN 00208833, 14682478. URL `http://www.jstor.org/stable/3096076`.

P. Schenk. *Invasion of England 1940: The Planning of Operation Sealion*. Conway, 1990.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461 – 464, 1978. ISSN 00905364. URL `http://www.jstor.org/stable/2958889`.

A. J. Scott and M. Knott. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, 30(3):507 – 512, 1974. ISSN 0006341X, 15410420. URL `http://www.jstor.org/stable/2529204`.

L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):205 – 233, 2017. URL `https://journal.r-project.org/archive/2017/RJ-2017-008/RJ-2017-008.pdf`.

A. Sen and M. S. Srivastava. On tests for detecting change in mean. *The Annals of Statistics*, 3(1): 98 – 108, 1975. ISSN 00905364. URL `http://www.jstor.org/stable/2958081`.

F. Serinaldi and C. G. Kilsby. The importance of prewhitening in change point analysis under persistence. *Stochastic environmental research and risk assessment: research journal*, 30(2): 763?777, Feb 2016. ISSN 1436-3240. doi: 10.1007/s00477-015-1041-5. URL `https://doi.org/10.1007/s00477-015-1041-5`.

B. Sheldon, Y. Wong, and S. J. Deitchman. Military Operations Research Society (MORS) Oral history project interview of Mr. Seymour J. Deitchman. *Military Operations Research*, 15(2):

61 – 106, 2010. ISSN 10825983, 21632758. URL http://www.jstor.org/stable/43941227.

R. L. Sivard. *World military and social expenditures, 1991.* World Priorities, 14th edition, 1991. ISBN 0918281075.

A. Slivkins. Introduction to multi-armed bandits. *arXiv e-prints*, art. arXiv:1904.07272, Apr. 2019. URL https://arxiv.org/abs/1904.07272.

C. Spaatz. Strategic air power: Fulfillment of a concept. *Foreign Affairs*, 24(3):385 – 396, 1946.

M. Spagat. World war three: what are the chances? *Significance*, 12(6):10 – 11, 2015. doi: 10.1111/j.1740-9713.2015.00860.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1740-9713.2015.00860.x.

M. Spagat and S. Pinker. Letters/puzzle. *Significance*, 13(3):44 – 46, 2016. doi: 10.1111/j.1740-9713.2016.00922.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1740-9713.2016.00922.x.

M. Spagat and S. van Weezel. On the decline of war. *UCD Centre for Economic Research Working Paper Series*, 18(15), August 2018. URL http://www.ucd.ie/t4cms/WP18_15.pdf.

J. D. Spence. *God's Chinese Son*. Harper Collins, 1996.

H. Spencer. *The Study of Sociology*. D. Appleton, 1896.

P. T. Spicer and S. E. Pratsinis. Coagulation and fragmentation: Universal steady-state particle-size distribution. *AIChE Journal*, 42(6):1612–1620, 1996. doi: 10.1002/aic.690420612. URL https://aiche.onlinelibrary.wiley.com/doi/abs/10.1002/aic.690420612.

C. Spradlin and G. Spradlin. Lanchester's equations in three dimensions. *Computers & Mathematics with Applications*, 53(7):999 – 1011, 2007. ISSN 0898-1221. doi: 10.1016/j.camwa.2007.01.013. URL http://www.sciencedirect.com/science/article/pii/S0898122107000557.

R. P. Stanley. *Enumerative Combinatorics*, volume 1. Springer, second edition, July 2011.

N. N. Taleb. *The Black Swan: The impact of the highly improbable*. Random House, 2007.

H. Tanaka, S. Inaba, and K. Nakazawa. Steady-state size distribution for the self-similar collision cascade. *Icarus*, 123(2):450 – 455, 1996. ISSN 0019-1035. doi: 10.1006/icar.1996.0170. URL http://www.sciencedirect.com/science/article/pii/S0019103596901700.

P. E. Tetlock and A. Belkin. Counterfactual thought experiments in world politics: Logical, methodological, and psychological perspectives, 1996.

The On-Line Encyclopedia of Integer Sequences. Catalan numbers. URL https://oeis.org/A000108.

P. Townsend. *Duel of Eagles*. Cassell, 1970.

A. Trapletti and K. Hornik. *tseries: Time Series Analysis and Computational Finance*, 2019. URL https://CRAN.R-project.org/package=tseries. R package version 0.10-47.

C. Truong, L. Oudre, and N. Vayatis. Selective review of offline change point detection methods. *arXiv e-prints*, January 2018. URL http://arxiv.org/abs/1801.00718.

P. Turchin. *War and Peace and War: The Rise and Fall of Empires*. Penguin Publishing Group, 2007. ISBN 9781101126912.

H. H. Turney-High. *Primitive War*. 1949. ISBN 9780872491960.

P. G. J. van Dongen. Fluctuations in coagulating systems. ii. *Journal of Statistical Physics*, 49:927 – 975, 12 1987. ISSN 1572-9613. doi: 10.1007/BF01017554. URL https://doi.org/10.1007/BF01017554.

P. G. J. van Dongen and M. H. Ernst. Fluctuations in coagulating systems. *Journal of Statistical Physics*, 49:879 – 926, 12 1987. ISSN 1572-9613. doi: 10.1007/BF01017553. URL https://doi.org/10.1007/BF01017553.

N. van Kampen. *Stochastic Processes in Physics and Chemistry*. North-Holland Personal Library. Elsevier Science, 1992. ISBN 0444893490.

M. von Smoluchowski. Versuch einer mathematischen theorie der koagulationskinetik kolloider lösungen. *Zeitschrift fuer physikalische Chemie*, 92:129 – 168, 1916. URL http://publikationen.ub.uni-frankfurt.de/frontdoor/index/index/docId/13699.

H. Wang and M. Song. Ckmeans.1d.dp: Optimal k-means clustering in one dimension by dynamic programming. *The R Journal*, 3(2):29 – 33, 2011. URL https://journal.r-project.org/archive/2011-2/RJournal_2011-2_Wang+Song.pdf.

J. A. Warden. *The Air Campaign: planning for combat*. Brassey's, 1989.

J. A. D. Wattis. An introduction to mathematical models of coagulation–fragmentation processes: a discrete deterministic mean-field approach. *Physica D: Nonlinear Phenomena*, 222(1-2):1–20, 2006.

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL http://ggplot2.org.

D. Wilkinson. *Deadly Quarrels*. University of California Press, 1980.

Wolfram Research, Inc. Mathematica, Version 12.0, 2019. URL https://www.wolfram.com/mathematica. Champaign, IL.

R. L. Worden, A. M. Savada, R. E. Dolan, and Library of Congress. *China: A Country Study*. Federal Research Divison, Library of Congress, 1988. URL https://lccn.loc.gov/87600493.

K. G. Wynn. *Men of the Battle of Britain: A Biographical Dictionary of The Few*. Frontline, 2015.

N. R. Zhang and D. O. Siegmund. A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22 – 32, 2007. doi: 10.1111/j.1541-0420.2006.00662.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2006.00662.x`.

R. M. Ziff and G. Stell. Kinetics of polymer gelation. *The Journal of Chemical Physics*, 73 (7):3492–3499, 1980. doi: 10.1063/1.440502. URL `https://doi.org/10.1063/1.440502`.

R. M. Ziff, E. M. Hendriks, and M. H. Ernst. Critical properties for gelation: A kinetic approach. *Phys. Rev. Lett.*, 49:593–595, Aug 1982. doi: 10.1103/PhysRevLett.49.593. URL `https://link.aps.org/doi/10.1103/PhysRevLett.49.593`.

C. Zou, G. Yin, L. Feng, and Z. Wang. Nonparametric maximum likelihood approach to multiple change-point problems. *The Annals of Statistics*, 42(3):970 – 1002, 2014. ISSN 00905364. URL `http://www.jstor.org/stable/43556312`.

Index

Notation	Description	Page List
Shattering	Fragmentation in which a cluster is reduced solely to monomers.	38, 40, 41, 113, 114, 116, 119, 120, 123, 128, 130, 138, 174
Sol	The parts of a system that are not a part of the gel.	38, 39