

# **Epidemic and contagion processes in self-organizing systems**

Sumit Kumar Ram

This dissertation is submitted for the degree of  
*Doctor of Philosophy*



## Acknowledgements

First and foremost, I would like to convey my heartfelt appreciation to my advisor, Prof. Didier Sornette, for his unwavering support of my Ph.D. studies and research, as well as his patience, drive, passion, and vast expertise. His advice and instructions on recognizing, crafting, and answering the right questions will be essential in the future. I couldn't have asked for a better advisor and mentor for my PhD studies.

In addition to my advisor, I'd like to thank the other members of my thesis committee: Prof. Claudio Tessone and Dr. Shyam Nandan for co-refereeing my PhD thesis.

My heartfelt gratitude also goes to my mentors, coauthors, and colleagues Dr. Shyam Nandan, Dr. Guy Ouillon, and Dr. Yavor Kamer for their guidance and support along this journey.

I would like to thank my current and former colleagues at the Chair of Entrepreneurial Risks, Jan-Christian, Giuseppe Ferro, Ali Ayoub, Sandro Lera, Dongshuai Zhao, Chahat Bhatia, Ming Chen, Michael Schatz, Ran Wei, Graciela Rojo Limón, Rebecca Westphal, Spencer Wheatley, Richard Senner, Ke Wu, and former master and bachelor students Ferdinand Wittman, Devendra Shintre, Niklas Leppe, and Daniel Partida. I am grateful for all of the stimulating debates and thoughts exchanged. In addition, I'd like to thank Adriana Schellenbaum-Lenner, Isabella Bieri, and Judith Holzheimer for their assistance in providing entertaining and easy administrative support.

Last but not least, I am grateful to my friends, Ananya, my parents, and other family members for their unwavering support during this journey.

## Abstract

Unlike equilibrium systems, where it is possible to study the impact of endogenous and exogenous perturbations using fluctuation-dissipation theorems, it is difficult to understand and predict the impact of shocks in out-of-equilibrium systems. This thesis tries to uncover and quantify the epidemic processes in out-of-equilibrium self-organizing systems of both endogenous and exogenous origin. We address the questions using massive datasets from social, biological, and physical systems and quantify the endogenous and exogenous origin of contagion. We further extend our discussion on how to build robust data aggregation systems to obtain and process large datasets to better understand the out of equilibrium systems.

In the first part of the thesis, we study the active systems of YouTube with  $\sim 2$  billion active users watching billions of videos, and a Scholarly network, with  $\sim 200$  Million articles with  $\sim 1.6$  Billion citation links. We mine these massive datasets to uncover epidemic processes that impact individual productivity and success. We quantify the epidemic processes in these self-organizing systems to validate their endogenous origin. We further show how the endogenously fueled exuberance within the network participants tend to synchronize both the individual productivity and success patterns.

The second part of the thesis examines hundreds of thousands of performances in international cricket to show that the performance sequence can be efficiently modeled as an epidemic process. With a number of statistical analyses, we provide convincing evidence for the presence of hot-hand effect that is success breeds success, implying large value of endogeneity in the game of cricket. Again we show the presence of a predictable performance sequence, in individual careers. However, we further uncover that the game itself aggregates and digests the information in such a way that the overall team performance and game outcome becomes unpredictable.

The third part of the thesis investigates the Earth's crust as an out-of-equilibrium system. With the help of an augmented Epidemic-Type Aftershock Sequence (ETAS) model that offers a direct path to thoroughly investigate the response of this self-organizing system to the exogenous (background earthquakes) and endogenous (aftershock earthquake sequence) shocks, we quantify the distance of the earthquake catalogs from the criticality. We show that our model that accounts specifically for spatial variation of background rate ( $\mu(x,y)$ ) of events outperforms the standard ETAS model with uniform  $\mu$ . Interestingly, the relatively low value of branching ratio ( $n < 1$ ) in the superior ETAS model suggests the crust is not be operating at a critical point, as it has been believed for a long time.

The fourth part of the thesis considers the society we live in as an out-of-equilibrium self-organizing system, where we are constantly exposed to infectious pathogens. The pathogens trigger epidemic of infectious diseases at all magnitudes. The pandemic due to the SARS-CoV-



2 virus is one such examples. The virus causing the disease is very efficient in transmitting from human to human. We thus model the epidemic process with the help of a probabilistic contagion model with inhomogeneous source terms that takes into account both endogenous (human to human) and exogenous (contaminated surfaces) modes of transmission. We use a number of Bayesian tools to effectively estimate the impact of governmental interventions (namely exogenous shocks) on epidemic progression during this period.

## Kurzfassung

Im Gegensatz zu Gleichgewichtssystemen, bei denen die Auswirkungen endogener und exogener Schocks mit Hilfe der sogenannten Fluktuations-Dissipations-Theoreme untersucht werden können, ist es schwierig, die Auswirkungen von Schocks in Systemen, die sich außerhalb des Gleichgewichts befinden (englisch: „Out-of-Equilibrium-Systems“, hier kurz OOES), zu verstehen und vorherzusagen. Der wissenschaftliche Beitrag dieser Arbeit ist es, epidemische Prozesse in selbstorganisierenden OOES aufzudecken und zu quantifizieren. Die Thematik wird anhand verschiedenster Beispiele untersucht, mit Datensätzen aus der Soziologie, Biologie und Physik. Ziel ist es, epidemische Prozesse und Ansteckungseffekte innerhalb von Systemen zu untersuchen, und ihre exogenen und endogenen Anteile zu quantifizieren. Da die Erforschung solcher Systemeffekte typischerweise die Nutzung massiver Datensets erfordert, wird in dieser Arbeit ebenfalls thematisiert, wie robuste Strukturen zur Datenspeicherung, -aggregation und -nutzung aufgebaut werden können.

Im ersten Teil der Arbeit untersuchen wir zum einen YouTube-Daten von über  $\sim 2$  Milliarden aktiven Nutzern, sowie zum anderen wissenschaftliche Publikationsdaten, basierend auf ca.  $\sim 200$  Millionen Artikeln und  $\sim 1,6$  Milliarden Zitationslinks. Wir untersuchen diese riesigen Datensätze, um epidemische Prozesse aufzudecken, die sich auf die Produktivität und den Erfolg des Einzelnen auswirken und verdeutlichen so die Zusammenhänge zwischen individueller Produktivität und Erfolg.

Im zweiten Teil der Arbeit untersuchen wir Daten von internationalen Cricket-Matches, um zu zeigen, dass der Erfolg einzelner Spieler effizient als epidemischer Prozess modelliert werden kann. Mit einer Reihe von statistischen Analysen demonstrieren wir das Vorhandensein eines Hot-Hand-Effekts, d. h., dass „Erfolg Erfolg erzeugt“. Dennoch decken wir auch auf, dass das Spiel selbst so unterschiedlichen Einflüssen unterliegt, dass die Gesamtleistung der Mannschaft und das Spielergebnis aus der Performance einzelner Spieler nicht allgemeingültig vorhergesagt werden können.

Der dritte Teil der Arbeit untersucht die Erdkruste als ein OOES. Sogenannte Epidemic-Type Aftershock Sequence (ETAS)-Modelle bieten einen direkten Weg, die Reaktion von selbstorganisierenden Systemen wie der Erdkruste auf exogenen und endogene Schocks (respektive Hintergrundbeben und Nachbeben) mathematisch zu untersuchen. Mithilfe eines solchen erweiterten Modells quantifizieren die sogenannte „Kritikalität“ des Systems Erdkruste. Wir zeigen, dass unser erweitertes Modell, welches die sogenannte Hintergrundrate ( $\mu(x,y)$ ) der von Ereignissen als eine räumlich varrierende Einheit darstellt, das ETAS-Standardmodell mit einer einheitlichen (nicht ortsabhängigen) Hintergrundrate  $\mu$  übertrifft. Interessanterweise deutet der relativ niedrige Wert des Verzweigungsverhältnisses (englisch

„Branching Ratio“) ( $n < 1$ ) im erweiterten ETAS-Modell darauf hin, dass die Kruste der Erde kein System ist, welches am kritischen Punkt operiert, wie lange Zeit angenommen wurde.

Der vierte Teil der Arbeit betrachtet die Übertragung von Krankheiten. Wir modellieren die Gesellschaft in der wir leben als ein selbstorganisierendes OOES, in dem wir ständig infektiösen Krankheitserregern ausgesetzt sind. Diese Übertragung dieser Erreger löst Epidemien von Infektionskrankheiten in allen Größenordnungen aus. Die Pandemie aufgrund des SARS-CoV-2-Virus ist ein solches Beispiel. Wir modellieren den epidemischen Prozess mit Hilfe eines probabilistischen Ansteckungsmodells mit inhomogenen Quelltermen, das sowohl endogene (von Mensch zu Mensch) als auch exogene (z.B. kontaminierte Oberflächen) Übertragungswege berücksichtigt. Wir verwenden eine Reihe von Bayes'schen Methoden, um die Auswirkungen staatlicher Eingriffe auf den Verlauf von Epidemien zu untersuchen.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	1
1.2	Self-excited point process model and its applications . . . . .	2
1.3	Detailed overview of the thesis . . . . .	4
1.3.1	Synchronized bursts of productivity and success in individual career	4
1.3.2	Uncovering predictability of individual and team success: Significant Hot Hand Effect in International Cricket . . . . .	4
1.3.3	Is Seismicity Operating at a Critical Point? . . . . .	5
1.3.4	Impact of Governmental interventions on epidemic progression and workplace activity during the COVID-19 outbreak . . . . .	5
<b>2</b>	<b>Synchronized bursts of productivity and success in individual career</b>	<b>7</b>
2.1	Introduction . . . . .	7
2.2	Data . . . . .	7
2.3	Productivity in Scientific Career . . . . .	8
2.4	Success in YouTube . . . . .	8
2.5	Scaling in individual productivity and success . . . . .	9
2.6	Synchronization of productivity and success . . . . .	9
2.7	Discussion . . . . .	10
2.8	Conclusion . . . . .	11
<b>3</b>	<b>Uncovering predictability of individual and team success: Significant Hot Hand Effect in International Cricket</b>	<b>15</b>
3.1	Introduction . . . . .	15
3.2	Background . . . . .	16
3.3	Methods . . . . .	17
3.3.1	Data . . . . .	17
3.3.2	Clustering point process representation . . . . .	18

3.3.3	Hawkes point process along the “performance time” . . . . .	19
3.4	Results . . . . .	20
3.4.1	Distributions of temporal locations of best performances . . . . .	20
3.4.2	Hot individual hands . . . . .	21
3.4.3	Hot team hands . . . . .	23
3.4.4	Hot winning hands . . . . .	23
3.5	Discussion . . . . .	24
<b>4</b>	<b>Is seismicity operating at a critical point?</b>	<b>31</b>
4.1	Introduction . . . . .	31
4.2	Methods . . . . .	33
4.2.1	Self-consistent estimation of spatially variable background rate using an extended <i>EM</i> algorithm . . . . .	33
4.2.2	Dataset . . . . .	35
4.2.3	Parameter calibration . . . . .	35
4.3	Results . . . . .	36
4.3.1	Synthetic tests of the bias in branching ratio $n$ due to uniform back- ground rate . . . . .	36
4.3.2	Pseudo-prospective forecasting experiments as a validation step for $ETAS_{\mu(x,y)}$ . . . . .	36
4.4	Discussion and Conclusions . . . . .	37
<b>5</b>	<b>Impact of Governmental interventions on epidemic progression and workplace activity during the COVID-19 outbreak</b>	<b>41</b>
5.1	Introduction . . . . .	41
5.2	Materials and Methods . . . . .	43
5.2.1	Data . . . . .	43
5.2.2	Models and methodology . . . . .	44
5.3	Results . . . . .	49
5.4	Discussion . . . . .	52
<b>6</b>	<b>Conclusion</b>	<b>63</b>
	<b>Bibliography</b>	<b>67</b>
	<b>Appendix A Supplementary Materials to Synchronized Bursts of Productivity and Success in individual career</b>	<b>79</b>
A.1	Preparation of citation data . . . . .	79

A.2	Preparation of YouTube data . . . . .	79
A.2.1	Quasi-Monte Carlo search for finding related YouTube channels . . . . .	80
A.2.2	Analyzing the contents of the videos . . . . .	80
A.2.3	Reconstructing the YouTube channel’s career . . . . .	80
A.3	Sample Data . . . . .	80
A.4	Validation of model prediction . . . . .	81
 <b>Appendix B Supplementary Materials to uncovering predictability of individual and team success: Significant Hot Hand Effect in International Cricket</b>		<b>83</b>
B.1	Data Preparation . . . . .	83
B.2	About game of cricket . . . . .	84
B.3	Supporting Results . . . . .	85
B.3.1	Model comparison . . . . .	86
B.4	Hot team hands . . . . .	88
B.5	Hot winning hands . . . . .	89
B.5.1	Multiple Hypothesis Testing . . . . .	89
B.5.2	Winning Streaks in ODI Teams . . . . .	91
B.5.3	Winning Streaks in Test Teams . . . . .	97
 <b>Appendix C Supplementary material for “Is the Earth crust operating at a critical point?”</b>		<b>101</b>
C.1	Extended Expectation-Maximization Algorithm for joint estimation of ETAS parameters and space varying background rate . . . . .	101
C.2	The three data sets and their magnitude of completeness . . . . .	102
C.3	Setup of pseudo-prospective experiments conducted on the three regions . . . . .	103
 <b>Appendix D Appendix 1</b>		<b>113</b>
D.1	Supporting figures . . . . .	113
D.2	Supporting Tables . . . . .	119
D.2.1	Impact of interventions on effective reproduction number and growth of epidemic . . . . .	119
D.3	Supplemental figures . . . . .	125
D.3.1	Mobility Impact . . . . .	125
D.3.2	Epidemic Impact . . . . .	134





# Chapter 1

## Introduction

### 1.1 Motivation

Both equilibrium and out-of-equilibrium systems, when subjected to external driving forces or exogenous shocks endogenously organize themselves in order to respond to the external changes [1–5]. While it is possible to exactly quantify the response to external influences through the theorem of fluctuation-dissipation in equilibrium systems [1], it is difficult to quantify the same in out-of-equilibrium self-organizing complex systems. Unlike equilibrium systems, out-of-equilibrium systems even in the absence of external influences constantly evolve under the influence of multiple sources of endogenous fluctuations [1]. Thus, quantifying the response is particularly challenging, as it amounts to distinguishing between endogeneity and exogeneity of the system [6]. This is difficult in most physical systems because externally imposed perturbations may lie outside the complex attractor, which itself may exhibit bifurcations, making observable perturbations difficult to classify [1].

Though there is no standard protocol to exactly characterize and quantify the response of the self-organizing systems to the exogenous and endogenous perturbations, the universal behaviour of their response to different shocks can be exploited for prediction [5]. Depending on their position with respect to the critical point, the components of the self-organizing out-of-equilibrium systems becomes highly interactive and also relax through a hierarchy of avalanches of all sizes [7]. This leads to a chain reaction or an epidemic process within the system [4]. Even though it is difficult to exactly decluster the exogenous, and endogenous fluctuations within the system, it is still possible to characterize the system by quantifying the response of the system [1, 6, 8, 9]. These systems exhibit different responses to endogenous and exogenous shocks that further depends on the distance from their critical points [8, 9]. Thus, with effective modeling of the underlying contagion process, it is possible to characterize and predict the behaviour of the systems [10–12].

The fundamental law of contagion, defined broadly as the transmission of influence from one entity to another, occupies an important place in epidemiology [13, 14], physics [15], sociology [16–19], finance and economics [20–22], geosciences [23, 12], to name a few. It manifests itself in sociology in a variety of ways, including the diffusion of innovations [24, 25], the propagation of cultural fads [26–28], and the emergence of political or social unrest [29, 30]. It's used in finance to describe events like flash crashes [11, 31], financial crises [1, 3], and financial bubbles [32, 33]. The bursts of earthquake and magnitude-dependent earthquake cascades [34–36, 12, 23] are commonly studied as contagion processes. Lastly, the spread of infectious diseases [37, 38] is the most studied example of a contagion process.

During a contagion process, the propagation of influence between a "infectious entity" and a "susceptible entity" is often studied in one of two ways: (i) the infectious entity influences the susceptible entity with a certain probability, resulting in contagion, or (ii) the probability of infection from susceptible to infectious entity changes rapidly from low to high as a critical threshold, thus, the effect of any single exposure depends strongly on the number of other exposures. Without loss of generality, it is thus sufficient to study the time evolution of infectious entities to efficiently quantify the underlying epidemic process [39, 40]. In this thesis, we generalize the notion of contagion to consistently model the contagion processes in a number of out-of-equilibrium systems.

We model the contagion process as self-excited conditional point process [39, 40]. This allows us to explicitly quantify the rate at which infectious entities are generated at any given time. We adapt and design the functional form of contagion intensity for different systems that best explains the underlying mechanisms. We use this contagion mechanism to quantify the endogenous and exogenous sources of influence that propagate through self-organizing out-of-equilibrium social, biological, and physical processes, as well as to comprehend their implications.

The Hawkes process and some of its applications in social, financial, biological, and geophysical systems are briefly discussed in the following sections. Following that, we offer a summary of the topics covered in this thesis.

## 1.2 Self-excited point process model and its applications

In the scientific literature, point processes are commonly used to model discrete random events in space and time, such as arrival of market orders [41, 11], arrival of emails [42], and software development [43]. The simplest of these processes is the Poisson point process, in which events occur independently of one another with a constant arrival rate  $\lambda$ . Since the

Poisson point process lacks a correlation structure, it is unable to explain the stylized facts found in real-world datasets, such as (i) clustering of events, (ii) long memory in inter-event times, (iii) fat-tailed decay of the inter-event distribution, and so on. A more consistent model that incorporates the self-excitations between the consecutive points, also known as the Hawkes process [39, 40], describes a large set of empirical observations.

The Hawkes point process is a generalization of the non-homogeneous Poisson process, whose intensity  $\lambda(t)$  [defined such that  $\lambda(t)dt$  is the expected value of the number of events in the time interval  $[t, t + dt)$ ] depends not only on time  $t$  but also on the history of the process according to

$$\lambda(t) = \mu(t) + \sum_{t_i < t} \phi(t - t_i) \quad (1.1)$$

In expression eq. (1.1), the first term  $\mu(t)$  on the right-hand side quantifies the exogenous background events that are uninfluenced by the past events. The second term describes how past events can trigger future events. This is a convenient and elegant way to account for the possibility of triggering, since each next point is function of the whole history, with a weight quantified by the memory or kernel function  $\phi(t - t_i) > 0$ , which decays as a function of its argument (points further in the past have a weaker influence). Thus, the sum  $\sum_{t_i < t} \phi(t - t_i)$  quantifies the influence of the history of past events on the present events.

Thus the self-excited Hawkes conditional point process is one of the simplest models to account for how the past can influence the future, while keeping a very convenient dependence of the past onto the future. The most important parameter of the Hawkes model is its branching ratio defined by eq. (1.2).

$$n = \int_0^{\infty} \phi(t) dt. \quad (1.2)$$

The branching ratio  $n$  represents the average number of events (or points) that a given event triggers. Additionally, it is the fraction of points (events) that have been caused by previous events. A value of  $n$  near the critical value 1 thus indicates a high degree of triggering (strong endogeneity). Hawkes process has received considerable attention in recent years as a consequence of its simplicity and application within a variety of fields, most notably the description of social diffusion processes [8, 44, 45], financial systems [46–48], and seismological predictions [49, 50, 12].

## **1.3 Detailed overview of the thesis**

This is a cumulative thesis based on the published and working papers during my PhD. The papers described below are the joint works with Prof. Didier Sornette. Chapter 2 is based on the working paper titled "Synchronized bursts of productivity and success in individual career". This is a project with my collaborators. Dr. Shyam Nandan, Sami Boulebnane and Prof. Didier Sornette. Chapter 3 is based on the preprint "Uncovering predictability of individual and team success: Significant Hot Hand Effect in International Cricket". This work is with my collaborators Dr. Shyam Nandan, and Prof. Didier Sornette. Chapter 4 is based on the published work titled "Is Seismicity Operating at a Critical Point?". This is in collaboration with Dr. Shyam Nandan, Dr. Guy Ouillon and Prof. Didier Sornette. Chapter 5 is based on the project with Prof. Didier Sornette titled "Impact of Governmental interventions on epidemic progression and workplace activity during the COVID-19 outbreak". All of the above chapters except Chapter 4 are based on my first authored projects. In Chapter 4, I am the second author with Dr. Shyam Nandan as the first author. In this project, I have participated in designing of the empirical analysis and writing of the manuscript.

### **1.3.1 Synchronized bursts of productivity and success in individual career**

This chapter investigates the contagion processes present in large datasets documenting individual productivity and success. During the last decades, there has been significant investigation of epidemic processes in biological, social, financial, and geophysical systems, however, little is known about the contagion behavior in individual productivity and success. This chapter uncover synchronized bursts in individual productivity and success. With the introduction of an epidemic model, we study the contagion of scholarly productivity and YouTube success. Our analysis reveals the strong influence of network externalities in individual careers. It is possible that endogenously fueled influence within the network forces the individual productivity and success to synchronize by fostering a persistent flow of information.

### **1.3.2 Uncovering predictability of individual and team success: Significant Hot Hand Effect in International Cricket**

This chapter investigates the presence of hot-hand effect or the contagion of success in large datasets documenting performance of teams and individuals in a Team Game. Here, we investigate the predictability and persistence (hot-hand effect) of individual and team perfor-

mance by analyzing the complete recorded history of international cricket. We introduce an original temporal representation of performance streaks, which is suitable to be modelled as a self-exciting point process. We confirm the presence of predictability and hot-hands across the individual performance and the absence of the same in team performance and game outcome. Our study contributes to recent historiographical debates concerning the presence of persistence in individual and collective productivity and success. The introduction of several metrics and methods can be useful to test and exploit clustering of performance in the study of human behavior and design of algorithms for predicting success.

### 1.3.3 Is Seismicity Operating at a Critical Point?

Seismicity and faulting within the Earth's crust are characterized by many scaling laws that are usually interpreted as qualifying the existence of underlying physical mechanisms associated with some kind of criticality in the sense of phase transitions. Using an augmented epidemic-type aftershock sequence (ETAS) model that accounts for the spatial variability of the background rates  $\mu(x,y)$ , this chapter presents a direct quantitative test of criticality. We calibrate the model to the ANSS catalog of the entire globe, the region around California, and the Geonet catalog for the region around New Zealand using an extended expectation-maximization (EM) algorithm including the determination of  $\mu(x,y)$ . We demonstrate that the criticality reported in previous studies is spurious and can be attributed to a systematic upward bias in the calibration of the branching ratio of the ETAS model, when not accounting correctly for spatial variability. We validate the version of the ETAS model that possesses a space varying background rate  $\mu(x,y)$  by performing pseudo-prospective forecasting tests. The non-criticality of seismicity has major implications for the prediction of large events.

### 1.3.4 Impact of Governmental interventions on epidemic progression and workplace activity during the COVID-19 outbreak

This chapter discusses the importance of timing and implementation of strategic policies by different countries in controlling the COVID-19 pandemic. We model the epidemic process with the help of a probabilistic contagion model with inhomogeneous source terms. We use endogenous (human to human) and exogenous (contaminated surfaces) modes of transmission of the virus to estimate the time evolution of the effective reproduction number (branching ratio). We use a number of Bayesian tools to effectively estimate the impact of governmental interventions on epidemic progression during this period. Using large scale human mobility and fine grained epidemic incidence data, we develop a framework to understand and quantify the effectiveness of the interventions implemented by various

countries to control epidemic growth. Through our analysis, we also unearth significant spatial diffusion of the epidemic before and during the first lock-down measures in several countries, casting doubt on the effectiveness or on the implementation quality of the proposed Governmental policies.

# Chapter 2

## Synchronized bursts of productivity and success in individual career

### 2.1 Introduction

Human lives are driven by their networks [51], whose structure determines what they do and achieve [52]. Thus, a better understanding of these networks is likely to help predict performance, productivity, and success of individuals [52–56].

While individuals are capable of driving their own productivity, their success is often dependent on others within their networks [52]. The stochastic flow of information propagating through the networks could influence individual productivity and success trajectories. Here, we uncover synchronized bursts of productivity and success within individual careers. Our results support the presence of strong social influence and of stochastic flows of information within social networks fueling an epidemic process, which could further lead to exuberant production and consumption of creative content.

### 2.2 Data

We develop a temporal topic-based extraction to mine two networks. First, we consider a network of scholars working and publishing in the same scientific discipline, here Signal Processing, within a given timeframe. Second, we consider a network of content creators on YouTube creating content on a particular topic: Cryptocurrency. We use state-of-the-art NLP and data aggregation techniques to prepare the datasets for our study. We analyze and extract data from the Microsoft academic graph dataset containing  $\sim 213$  million scholarly articles to prepare our dataset that contains 140 scholarly careers who have published on the topic

of Signal Processing. Similarly, we analyze 100 million YouTube videos to construct 399 careers of YouTube content creators who have created videos on the topic of Cryptocurrency (refer sections A.1 and A.2 in appendix for data preparation).

## 2.3 Productivity in Scientific Career

Understanding individual scientific productivity can help develop policies that improve each scientist's ability to succeed and enhance the prospects of Science as a whole [54]. Several factors can influence scientific productivity, like age [53], collaboration network [54], scientific discipline [55]. Knowledge creation is a complex process [54], and scholars from the same discipline often imitate, collaborate, and compete to develop understanding. This creates herding between the scholars who tend to follow the same research direction. Here we quantify the influence of the herding behavior on individual productivity. We count the total number of publications by a scholar on a particular scientific discipline in a given year and call it annual scientific productivity. For this study, we evaluate the yearly scientific productivity of the authors publishing on the topic of signal processing.

## 2.4 Success in YouTube

Effort, skill, or inherent excellence, as well as luck [57], can drive success. While it is difficult to quantify the effort and its role [57], the attention flowing through social networks is one of the drivers of success [52–55]. The position of the performer in social networks [52] and the perception of the artistic work within peers [52, 55] can influence and facilitate success. Thus, the performers constantly imitate, collaborate, and compete with each other to attract the attention of the spectators, which leads to similar content creation. Unable to quantify the quality of the content, spectators start to imitate other's choices and start herding. We consider the YouTube platform [8] to quantify the evolution of success that results from the herding behavior of spectators. The key success element in YouTube is the total number of views of the artist's content, as this decides his/her earning. We consider the total number of weekly views from all the Cryptocurrency contents as the success metric for a YouTube channel.



## 2.5 Scaling in individual productivity and success

The annual productivity of an author is the sum of all publications on Signal Processing in a given year. The peak of productivity ( $t_c$ ) is the year of the highest number of publications. To avoid the domination of highly productive careers in our statistical analysis, we normalize the productivity path of each scholar by its maximum productivity value. We then perform a peak-centered average of the annual productivity across all the authors. In the inset of fig. 2.1A, one can observe that the precursory and post-peak dynamics are governed by similar growth and decay patterns. For the YouTube data, the peak of success ( $t_c$ ) is the week of the highest number of YouTube views. As done for authors' publications, we normalize the maximum weekly views to 1 in each career and find the peak-centered average weekly views. As for scholarly productivity, we observe and quantify the burst of success in YouTube. The inset figure in fig. 2.1C shows identical growth patterns in pre-peak and post-peak dynamics. The identical scaling behavior is further validated in the main fig. 2.1. We plot the cumulative sum of the renormalized average weekly view counts for pre-peak (foreshock) and post-peak (aftershock) dynamics. For both cases, the exponent quantifying the scaling behavior is  $\sim 0.5$ .

We randomly shuffle both datasets 100 times to prepare the null hypothesis test to check that the results are not spurious and would not be present by chance. We thus determine the significance of the obtained exponent values [46]. We perform the Wilcoxon signed-rank test to determine the statistical significance. We observe that both the pre-peak and post-peak exponent obtained from data is significantly different than the null ( $p < 10^{-6}$ ).

Using an iterated least-squares fit technique [8], we calibrate the evolution of daily views. We observe unimodal distributions of relaxation exponents for both cases centered around 0.5 (see fig. A.2 B, C). The joint distribution of pre-peak and post-peak exponents cluster around ( $\sim 0.5, \sim 0.5$ ). Further, these exponents are significantly different from those obtained from randomly shuffled null datasets.

## 2.6 Synchronization of productivity and success

We note down the time when the number of publications and number of views were maximum within each career. fig. 2.1B shows the distribution of these times ( $t_c$ ) (in year), when the scholars published their maximum number of papers. Surprisingly, we witness a peak within the distribution around the year 1991. We evaluate the significance of the observed peak by comparing it with a null distribution. We randomly redistribute the publications within each career, determine the year of maximum publications, and obtain the corresponding

distribution in the gray line. A two-sample Kolmogorov-Smirnov (KS) test shows that the peak of the distribution of the non-shuffled data is significant. We repeat the above analysis with the YouTube data to better understand the bursts of success. Figure 1D shows the distribution of dates ( $t_c$ ) of the highest YouTube views within each individual career. Similar to fig. 2.1B, we observe a peculiar peak around January 2018. The small p-value from the KS test confirms the significance of the peak. The rejection of the null hypothesis for both datasets supports the hypothesis that endogenous growth of productivity and success can synchronize, likely due to herding and synchronized exuberant behaviors. Thus, in addition to the evidence for endogenous success (corroborated by the symmetric nature of the productivity peaks), there is an exogenous component at the level of individuals, which is endogenous to their global network via feedback herding loops.

## 2.7 Discussion

The endogenously fueled spontaneous interplay between a continuous stochastic flow of small external impulses and the amplifying impact of the epidemic cascade of influences results in bursts of productivity and success [58]. Thus, the bursts of productivity and success in individual careers can be effectively modeled with a simple epidemic branching process [8]. According to this model, the productivity or success at a time  $t_i$  triggers the productivity or success at a future time  $t_i$ . The influence of past on future productivity or success decays with a memory kernel  $\phi(t, t_i)$ . With the assumption that the distribution of interevent time of human activity follows a power-law behavior [58], this memory kernel  $\phi(t, t_i)$  can be expressed by eq. (2.1) [8, 6, 9, 59].

$$\phi(t, t_i) \sim \frac{1}{t - t_i^{1+\theta}}, \text{ with } 0 < \theta < 1 \quad (2.1)$$

The intensity  $\lambda(t)$  of the point process describing the arrival of papers or videos results from a combination of exogenous ( $\mu(t)$ ) and endogenous sources. Because of endogeneity, each past success or productivity can become the cause of future success or productivity according to the term  $\sum_{t_i \leq t} \phi(t, t_i)$  in the expression in eq. (2.2).

$$\lambda(t) \sim \mu(t) + \sum_{t_i \leq t} \phi(t, t_i) \quad (2.2)$$

In the absence of major external influence, the bare kernel  $\phi(t, t_i)$  of this epidemic process in equation 1 is renormalized into the dressed kernel  $A(t)$  according to eq. (2.3) [8, 6, 9, 59]. The renormalized activities in  $A(t)$  decay symmetrically around the peak  $t_c$  according to

$$A(t) = \frac{1}{|t - t_c|^{1-2\theta}} \quad (2.3)$$

fig. 2.1A, and fig. 2.1C present the time dependence of  $A(t)$  in eq. (2.3) of the epidemic process in eq. (2.2). The endogenous propagation of influence is confirmed by the similar exponent values of the cumulative pre-peak (foreshock) and post-peak (aftershock) temporal growth patterns (see fig. 2.1A). The results further provide an estimate of the exponent for the power-law influence kernel in an individual's productivity and success patterns. We estimate that, on average, there is an influence of productivity on future productivity decays as  $\phi(t) \propto \frac{1}{t^{1.15}}$ , i.e.,  $\theta = 0.2$  as defined in eq. (2.1) and the past views trigger the cascades of future views with a memory kernel  $\phi(t) \propto \frac{1}{t^{1.25}}$ , i.e.,  $\theta = 0.2$  (using eq. (2.1) and eq. (2.3)). Additionally, we show that the model accurately predicts the post-peak relaxation rate based on pre-peak dynamics. By studying the pre-peak and post-peak relaxation exponents of all the YouTube channels, we show that the joint distribution of pre-peak and post-peak exponents cluster around  $(0.5, 0.5)$  (see fig. A.2A), as predicted by the theory (eq. (2.3)) of endogenous peaks [8, 6, 9, 59].

## 2.8 Conclusion

This study shows the emergence of synchronized bursts of individual productivity and success, which reveal the existence of herding between performers and spectators. By stacking the peaks (after normalization), we show that the productivity and success trajectories are symmetric before and after the peak. Furthermore, they can be represented by a power-law with exponent  $1 - 2\theta$ , where  $1 - 2\theta$  takes approximately the same value before and after the peak. We find  $1 - 2\theta = 0.7$  for the scholarly productivity and  $0.55$  for the YouTube success. These results translate to an endogenously fueled epidemic branching process within each career, with a slowly decaying influence kernel ( $\phi(t) \propto \frac{1}{t^{1+\theta}}$ ), with  $\theta = 0.15 - 0.25$ . This value of  $\theta$  is smaller than those  $\theta = 0.3 - 0.4$ , reported in previous studies [8, 6, 9, 59].

Interestingly, we observe the peak of productivity around 1991 in most scholarly careers concerned with signal processing and the peak of success around Jan 2018 in most of the YouTube channels. We can trace the synchronized herding in scholarly productivity back to a wave of interest in wavelets during the 1990s ("Wavelet revolution" during the 1990s). The herding in YouTube success occurs one month after the peak in December 2017 of the cryptocurrency bubble (2017 Cryptocurrency mania). Thus, we demonstrate endogenously fueled herding effects in each case, where the scholars and the YouTubers, and the respective communities were most active during the two times of focus. The presence

of exuberance within the performers (in the case of scholarly productivity) and exuberance within the spectators (in the case of YouTube success) led to a synchronization of the bursts of productivity and success to synchronize.

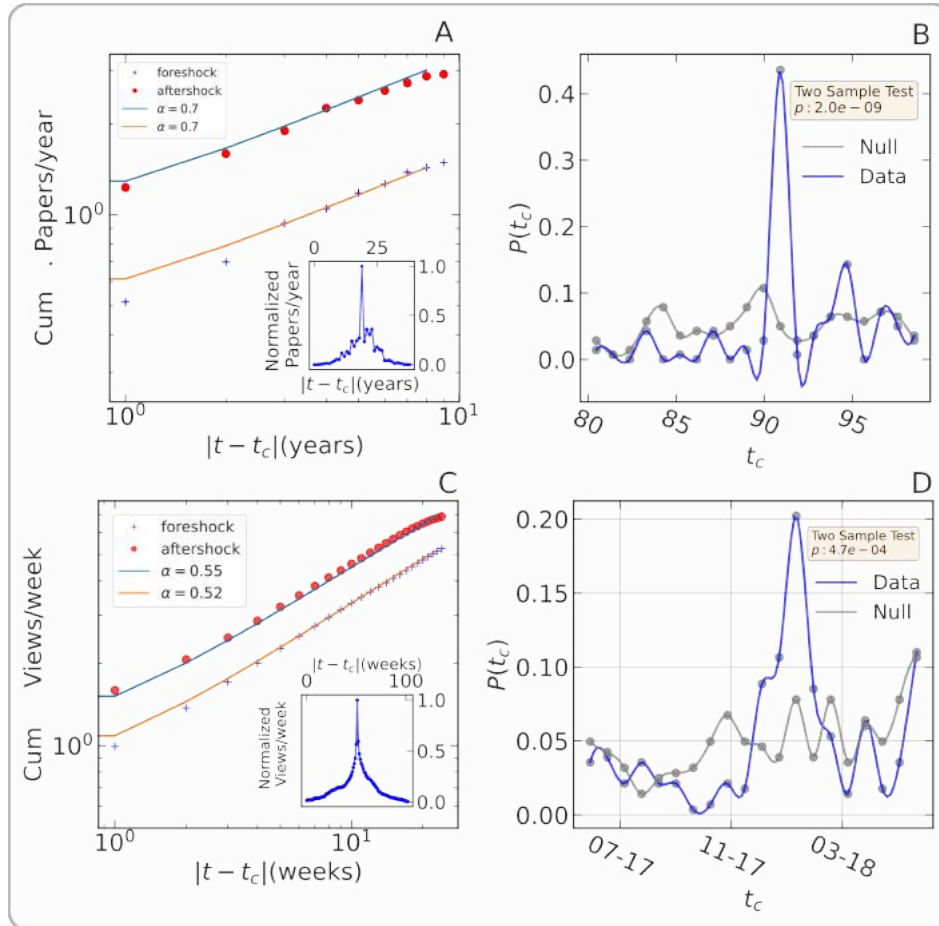


Figure 2.1 Synchronized endogeneity in scientific productivity and YouTube success: A) Cumulative precursory (“foreshock”) and relaxation (“aftershock”) of renormalized scholarly annual productivity. The exponents of scaling laws are 0.7. The inset figure shows the peak-centered average annual scholarly productivity. B) The probability distribution of time of maximum annual productivity in scholarly careers (in blue). The distribution of time of maximum productivity in randomly distributed publications within the careers (in gray). Two-sample Kolmogorov-Smirnov test result is in yellow box. C) Cumulative precursory (“foreshock”) and relaxation (“aftershock”) of weekly view counts of the YouTube (see text). The scaling laws are with exponent 0.5. The inset figure shows the peak-centered average weekly views. D) The probability distribution of the time of the highest weekly views for the YouTube channels. The blue line is obtained from the data and the gray line is for the null distribution constructed by randomly shuffling the daily weekly counts. Two-sample Kolmogorov-Smirnov test result is in yellow box.



## Chapter 3

# Uncovering predictability of individual and team success: Significant Hot Hand Effect in International Cricket

### 3.1 Introduction

The study of ‘science of success’ has become an increasingly popular field as a result of the growing availability of large datasets describing individual’s careers from which much can be learned and importantly predicted [52, 53, 60–63, 56, 64, 65]. The increasing shift towards the collaborative and team-based effort (performance) in recent years has made it ever important to quantify and predict teamwork [66–70]. However, the translation of the predictability in individual performance to team performance is still missing. In this study, we develop novel statistical tools to uncover the temporal features inside a set of performances. We explore the complete history of International cricket [71, 72] to quantify individual and team performances. We study the sequence of consecutive performances of each player and team. By investigating the scores of individual players against the index of the games, we note that success breeds success in individual career (also supported by ARIMA model in section B.3.1). We further document that the best performances in a given player’s career are clustered in time (see figure 3.3), contrary to previous findings [73, 74]. However, we cannot say the same for the teams. We uncover the presence of hot hands in individual careers in both formats of the game but the absence of the same in team performances. Our proposed Hawkes model applied to the performance time not only outperforms the traditional techniques like ARIMA (see section B.3.1) but is successful in capturing the ingredients of self-excitation in the patterns of consecutive superior performances. These findings raise

intriguing questions regarding the nature and extent of predictability of one's success and team success in a team game. This is particularly interesting, since these findings not only refute the well-established narratives of the absence of hot hands in team games [73–76] where performances are usually driven by stochastic events. Our findings suggest that the hot hand effect is not just a psychological bias [75, 73]. A part of results corroborate previous works on hot-hands [77–79, 64, 80]. To the best of our knowledge, this is the first time a detailed quantitative analysis has been done to quantify the well-known concept of 'in form' or 'out of form' present in the cricketing vocabulary.

One of the possible explanations for the observation of such a peculiar behavior in the game of cricket may be the relatively larger importance of skill in the outcomes of a player's game and luck in the outcomes of a teams' game [65, 57]. The rest of the article is structured as follows. In section 3.2, we present a short literature review to motivate our study and put it in the right context. In section 3.3, we describe the methods implemented in our study. Section 3.3.1 describes the dataset that has been used in the study and the data acquisition methodology. Section 3.3.2 and 3.3.3 presents our proposed clustering point process representation in the form of a self-excited point process model to quantify and predict the hot hands in the sequences of performances. Section 3.4 presents our main results. We conclude the results of the study in section 3.5.

## 3.2 Background

A much-debated question is whether or not a string of successes of an individual or entity is more likely to cause continued success. This is called The Hot Hand effect. The belief in it is called the Gambler's Fallacy [81]. This particular question is important, as its answer has far-reaching consequences in several research fields, including finance and econometrics [82, 83, 78, 65, 80, 84], psychology [74, 85, 73, 86] and sociology [87–90, 64]. The debate on the "Hot Hand" vs. the "Gambler's Fallacy" revolves around the deeper question: 'to what extent, human beings are capable of dealing with inherent systemic stochasticity' [65, 57].

In their seminal paper, Gilovich et al. refute the validity of "the hot hand" and "streak shooting" in the game of basketball [73]. Their analyses of the shooting records of the Philadelphia 76ers, Boston Celtics, and a controlled shooting experiment with the men and women of Cornell's varsity teams provided no evidence for a positive correlation between the outcomes of successive shots. They further showed that the belief in the hot hand and the detection of streaks in random sequences is nothing but an expression of the general misconception of chance [75], according to which even short random sequences are thought to be highly representative of their generating process. There has been very strong



support for this reasoning in the literature, especially in the field of finance and economics [91, 83, 85, 76, 84]. These studies support the existence of gambler’s fallacy, stating that the hot hand does not exist and the gambler fallacy is nothing but a psychological bias based on the “law of small numbers”, at the cost of often leading to risky decision making.

On the other side of the debate, Miller and Sanjurjo [78] have recently challenged the original findings in [73], with contrasting conclusions revealing significant evidence for streak shooting. Miller and Sanjurjo showed that the method used in [73] introduced a sampling bias because they start counting after a series of hits/misses. They further showed that the method of [73] is biased towards more misses, thus claiming that an equal rate of hits to misses after a streak presented in [73] is, in fact, a sign of a hot hand. The debate about successful streaks has gained fresh prominence in many other fields, with many arguing the presence of such streaks in large scale data sets of scientific careers, artistic career and acting careers [77, 64, 86, 92, 93].

The above debates revolve around the investigation of presence or absence of hot-hand effect in individual performances. However, they fail to show how these effects can be exploited for better prediction or how the aggregated individual performances drive the evolution of team performance. In this study we present a novel methodology to better understand and predict individual and team performances. We derive our methodology from the celebrated Hawkes process [39] – which falls under the class of self-exciting point processes –, which has garnered much attention in recent years as a consequence of its application within a variety of fields particularly the description of social diffusion processes [8, 44, 45], financial systems [46–48], and seismological predictions [49, 50, 12]. To the best of our knowledge, this is the first use of Hawkes process in the area of ‘science of success’. We apply our methodology for studying the presence (or absence) of the hot hand effect within the performance sequences in individual performance in the game of cricket. Our methodology would be useful in predicting and quantifying hot-hand effect in performance sequences in a range of other domains.

## 3.3 Methods

### 3.3.1 Data

The dataset includes 4,178 One Day International (*ODI*) matches starting from January 5, 1971 till July 1, 2019 (48 years) and 2,351 international *Test* matches spanning March, 1877 to March 2019 (142 years). We record 51,699 batting performances of 2959 *Test* batsmen and 51,088 bowling performances of 2,874 *Test* bowlers, 90,166 batting performances of 2,500

*ODI* batsmen and 90,754 bowling performances of 2,505 *ODI* bowlers (in total 283,707 records) (see figure 3.1). In order to have meaningful calibration results, we only analyse the performances of those batsmen who have played at least 30 matches (see goodness of fit in section B.1 ).

### 3.3.2 Clustering point process representation

We call  $S_j(t)$  the performance (please refer section B.2 for more details about the game of cricket) of the player  $j$  at his  $t$ -th game. We define the subordinate time process  $H_j(t)$  of the stochastic process  $S_j(t)$  [94] as

$$H_j(t) = \sum_{i=1}^t \frac{1}{S_j(t_i)} . \quad (3.1)$$

The  $t \rightarrow H_j(t)$  map represents a nonlinear transformation from the calendar time  $t$  onto an effective “performance time” of player  $j$ .  $H_j(t)$  denotes a transformed time-stamp at which the  $t$ -th event takes place for player  $j$ . This defines a point process along “performance time” with the time stamps  $\{H_j(t_1), H_j(t_2), \dots, H_j(t_n), \dots\}$ . The intuition behind definition (3.1) is that a series of strong performance values  $\{S_j(t_i), S_j(t_{i+1}), \dots\}$  are transformed into closely clustered points in “performance time”. This allows us to analyse the relationship between performances in time using simple one-dimensional techniques. In other words, by transforming  $S_j(t)$  into  $H_j(t)$ , we project the stochastic process described by the sequence  $\{S_j(t), t = 1, \dots\}$  onto an one-dimensional point process with time stamps  $\{H_j(t_1), H_j(t_2), \dots, H_j(t_n), \dots\}$ . By construction, the  $t \rightarrow H_j(t)$  transformation preserves the self-excited component of performance scores described by the stochastic process  $\{S_j(t)\}$  and amplifies it by the magnitude of the performance values.

Figure 3.2 presents the example of the career of Sachin Tendulkar, who is the highest run-scorer in both formats of the game. Top panels show the performance time  $H(t)$  as a function of  $t$ , which is the index of the  $t$ -th game, as defined in equation (3.1), for two International cricketing formats, *ODI* and *Test*. Bottom panels show the scores  $S_j(t)$  as a function of  $t$ , which indexed the  $t$ -th game, for the same two International cricketing formats, *ODI* and *Test*. The presence of local temporal clustering around the high and low performances is clearly visible in both representations of  $H(t)$  and  $S(t)$  for this player.

### 3.3.3 Hawkes point process along the “performance time”

The performance time  $H_j(t)$  of player  $j$  defined by expression (3.1) allows us to introduce a point process by the performance times  $\{H_j(t_1), H_j(t_2), \dots, H_j(t_n), \dots\}$  along the  $H$  axis. In other words, we consider the “performance time” axis  $H_j(t)$  and, along this new time axis, we identify “points” at the locations  $\{H_j(t_1), H_j(t_2), \dots, H_j(t_n), \dots\}$ . When player  $j$  has a series of large scores, this is expressed as a cluster of closely spaced points along the  $H$  axis as shown in figure 3.2.

Inspired by the analyses of [95, 11, 47] using generalized non-homogeneous Poisson processes, we propose to model the clustering of the points along the  $H$  axis of each player by using the self-excited stochastic Hawkes point process model [48, 39], augmented by some necessary ingredients for constructing a prediction model [74]. In other words, we visualise the points for a given player  $j$  along the performance time axis  $H_j(t)$  as being generated by an Hawkes model with intensity  $\lambda(t)$  given by

$$\lambda(t) = \mu + \sum_{t_i < t} \phi(t - t_i) . \quad (3.2)$$

In expression (3.2), the first term  $\mu$  in the right-hand-side is the background intensity, which quantifies the “intrinsic” performance level of a player, uninfluenced by his/her past performances. The second term describes how past points can trigger future points along the  $H$  axis. This is a convenient and elegant way to account for the possibility of a hot-hand effect, since each next point is function of the whole history, with a weight quantified by the memory or kernel function  $\phi(t - t_i) > 0$ , which is decaying as a function of its argument (points further in the past have a weaker influence). Thus, the sum  $\sum_{t_i < t} \phi(t - t_i)$  quantifies the influence of the history of past performances on a player’s present performance.

Depending on the problem, previous researchers have used different parametric forms for  $\phi$ , e.g. [95, 50, 12] use a power law kernel, whereas [11] use an exponential kernel. In the present case, as there is no reason to favour any parametric form, we decide to use a non-parametric kernel function for  $\phi$  [96, 48]. Thus, shortly after a large performance amplitude, model (3.2) describes the possibility that the excess intensity of observing a similar performance is boosted and then decays to the baseline average performance level  $\mu$  at long times.

The self-excited Hawkes conditional point process is one of the simplest model to account for how the past can influence the future, while keeping a very convenient linear dependence of the past onto the future. The most important parameter of the Hawkes model is its

branching ratio defined by

$$n = \int_0^{\infty} \phi(t) dt . \quad (3.3)$$

The branching ratio  $n$  is the average number of points (or events) triggered by a given point. It is also the fraction of points (events) that have been triggered by past events [97]. A value of  $n$  close to the critical value 1 thus qualifies a large level of triggering (strong hot hand effect) and endogeneity.

We use the expectation maximization algorithm as described [48] to calibrate the model.

## 3.4 Results

### 3.4.1 Distributions of temporal locations of best performances

To study the self-excited nature of the scores in an individual's career, we investigate the relative positions of the best three performances in each player's career. We denote  $t_j^*$  the index of the best performance in player  $j$ 's career, i.e.,

$$t_j^* = \operatorname{argmax}_t S_j(t) \quad (3.4)$$

We also define  $t_j^{**}, t_j^{***}$  as the indices of the second, third best score, and  $\tau_j$  as the length of an individual's career. We then calculate the relative difference of indices of the three best performances as

$$\Delta_j^{1,2} = (t_j^* - t_j^{**})/\tau_j, \quad \Delta_j^{1,3} = (t_j^* - t_j^{***})/\tau_j, \quad \Delta_j^{2,3} = (t_j^{**} - t_j^{***})/\tau_j \quad (3.5)$$

for all the players in our dataset and define the marginal probability density functions  $P(\Delta_j^{1,2}), P(\Delta_j^{1,3}), P(\Delta_j^{2,3})$  and the joint probability distribution  $Q(\Delta_j^{1,2}, \Delta_j^{1,3})$ . As a control, we shuffle the indices of the performances within the individual's career and reevaluate the quantities. The primed quantities correspond to the shuffled career, i.e.,  $t_j'^*$  corresponds to the index of the best performance within the random reshuffled player  $j$ 's career, and so on. We define the marginal probability density functions  $P(\Delta_j'^{1,2}), P(\Delta_j'^{1,3}), P(\Delta_j'^{2,3})$ , which are the distributions of the shuffle versions  $\Delta_j'^{1,2}$  of  $\Delta_j^{1,2}$ ,  $\Delta_j'^{1,3}$  of  $\Delta_j^{1,3}$  and  $\Delta_j'^{2,3}$  or  $\Delta_j^{2,3}$ . We define the ratios of these marginal probabilities  $R(\Delta t)$  to quantify the temporal collocation of the best performances in an individual career

$$R(\Delta t) = \frac{P(\Delta t)}{P'(\Delta t')}, \quad \text{where } \Delta t = \Delta_j^{1,2}, \Delta_j^{1,3} \text{ or } \Delta_j^{2,3}. \quad (3.6)$$

Figure 3.3 presents the joint probability distribution of relative difference of indices of best and second-best against the best and third-best ( $Q(\Delta_j^{1,2}, \Delta_j 1, 3)$ ) (top panels) defined by eq. (3.5), for ODI and Test formats over all individuals' careers. We observe a concentration of high probability around the origin (0,0) in both formats of the game. This correlation is interesting since this characteristic is a feature of the self-excited process and is not expected in a pure memoryless Poissonian process. This finding constitutes a first line of evidence for the existence of temporal clustering in the performances across players' careers.

The bottom panels of figure 3.3 shows the ratio  $R(\Delta t)$  (eq. (3.6)), which compares the marginal probability distribution of the relative difference of the indices in the real careers against the indices obtained from shuffled careers. The distinctive peak around 0 in the plots provides additional support for clustering of performance within careers.  $R\Delta t$  is approximately symmetric around the origin, indicating that the highest performances are equally likely to arrive before or after the second-highest and third-highest scores. This pattern is expected from a self-excited process with approximately equal propensity for performance persistence among the best performance streaks <sup>1</sup> [98, 99].

### 3.4.2 Hot individual hands

We partition the career of a player  $j$  into training set and validation set. We take the first 80% of the performances as the training set and the next 20% as the validation set. We transform the performance sequence in training and validation set to performance time representation as discussed in methods section. We calibrate the performance time in training set 100 times to determine background intensity  $\mu$  and the memory kernel  $\varphi$ . We then use the calibrated  $\mu$  and  $\varphi$  to evaluate the prediction performance in validation set using the log-likelihood score and call the median value  $\mathcal{L}_j^{model}$ .

Similarly, we prepare a controlled set of log-likelihood estimation for the same player. Keeping the validation set unaltered, we shuffle the sequence of the performance in the training set 100 times and use this to train the model. We evaluate the trained model on unaltered validation set to determine the median log-likelihood estimation  $\mathcal{L}_j^{control}$ . With the above constructions, we define the relative differences  $\delta \left( \mathcal{L}_j^{model}, \mathcal{L}_j^{control} \right)$  by equation 3.7.

---

<sup>1</sup>This was shown in the context of earthquake time and space clustering. Here, we can think of the highest performance as equivalent to the main shock in a seismic sequence. Then, the main shock can be shown to be triggered by large events that occur before it ("foreshocks") and the main shock itself triggers large events ("aftershocks")

$$\delta \left( \mathcal{L}_j^{\text{model}}, \mathcal{L}_j^{\text{control}} \right) = \frac{\mathcal{L}_j^{\text{model}} - \mathcal{L}_j^{\text{control}}}{\mathcal{L}_j^{\text{control}}} \quad (3.7)$$

Additionally, we estimate the branching ratios (see equation eq. (3.3)) [11, 47, 96] of the performance time for all players over the duration of their entire career. For comparison, we construct null estimations by randomly shuffling the performance time 100 times and reevaluating the 100 null branching ratios for each of the players.

The relative difference of log-likelihood prediction scores in eq. (3.7) is shown in the bottom panels of figure 3.4, for both formats of the games. The insets present the fraction of times the control set performs better than the model experiment. The results show a significant improvement in prediction score in model experiments compared to the control experiments. We plot the distribution of the branching ratios obtained from the data and the null branching ratios and compare them in the top panels of figure 3.4. In the plots, the shaded region marks the fraction of players' branching ratios that cannot be explained by the branching ratios estimated from the shuffled careers.

We then evaluate the statistical significance of having a better log-likelihood score in the model experiments compared to the control experiments. We perform a Shapiro-Wilk test on the log-likelihood scores obtained from both the experiments and found that the paired differences in each career are not normally distributed. Hence, we perform the Wilcoxon signed-rank test to determine the statistical significance. Considering a confidence level of 0.05, we observe that in 49.6% of Test careers and in 46.8% of the ODI careers, the log-likelihood prediction score in original sequence is significantly higher than the median log-likelihood prediction score in control experiments. This leads us to conclude that the probability of falsely accepting all the null hypothesis – the control experiments perform equally good – is  $< 10^{-6}$  (using a binomial probability distribution with success rate 0.05 of false test result) for both the cases. This result is sufficient to support the predictive power of our model. Furthermore, our model performs better than the standard techniques like ARIMA (please refer section B.3.1).

We thus compare the branching ratios (see eq. (3.3)) of the performance time obtained from data and null shuffling for each player to quantify the Hot-Hand effect. We perform the Wilcoxon signed-rank test to determine the statistical significance. We observe that in 56.8% of Test careers and in 53.7% of the ODI careers, the branching ratio of original performance time is significantly higher than the median branching ratio in null performance time (confidence level = 0.05). These results suggest a significant presence of Hot Hands in the players career, as the probability for the absence of Hot Hands is  $< 10^{-6}$  (using a binomial probability distribution with success rate 0.05 of false test result)

### 3.4.3 Hot team hands

We repeat the above analysis to predict and quantify the team performances (sum of all individual performances in a game) (please ref section B.2 for more details). We take the first 80% of the team performances as the training set and validate the model on the next 20%. Using the Wilcoxon signed-rank test with confidence level 0.05, we observe that only in 30% and 20% of ODI and Test teams, the log likelihood scores in model experiments is significantly better than the control experiments. These results suggest the significant reduction in prediction ( 50% reduction) compared to predictability of individual performances (please refer B.4 for more details). Further the probability of falsely accepting all the null hypothesis – the control experiments perform better – increases to  $10^{-2}$  and  $10^{-1}$  respectively (using a binomial probability distribution with success rate 0.05 of false test result). The absence of reliable prediction in the above results suggest the absence of exploitable self-excited patterns in team performance.

### 3.4.4 Hot winning hands

We investigate the presence of hot hands in the team performances by going through the complete history of games played by each team and analyze the winning streaks (i.e., the number of continuous wins without losing a single game in between). We note down the length of winning streaks and the corresponding frequencies of occurrences of such streaks in each team playing history.

Then, we construct a statistical ensemble of possible performance trajectories. We randomly shuffle the original performance sequences to generate 1000 synthetic performance trajectories. Using this statistical ensemble, we evaluate the null probability distribution for the joint occurrence of streaks of length  $n$  and of corresponding frequency  $f$ . We use this probability distribution for estimating the p-values for the observed events. we define the p-values  $p(n)$  and  $p(n_f)$  according to eq. (3.8), which respectively represent the p-value for observation of streaks with length  $n$  and streaks with length  $n$  conditional on frequency  $f$ .

$$p(n) = P(n_i \geq n), \quad p(n_f) = P(n_i \geq n|f) \quad (3.8)$$

To avoid the problem of multiple hypothesis testing [100], because of simultaneous consideration of the multiple individual tests, we correct the error rates of individual tests using multiple hypothesis testing methods [101–105]. We note down the results from the methods [101–105] and identify the extreme events (ref supporting tables for multiple hypothesis testing in section B.5.1).

Figure 3.5 presents the position of the realized winning streaks, along with the null distribution of the winning streaks for the 10 teams in the ODI format (top panel) and in the Test format (bottom panel). The red stars in figure reveal several highly improbable i.e., one or both of  $p(n)$  and  $p(n_f)$  is significant with confidence level 0.05, after multiple testing. A large number of white stars indicate the probable events i.e., none of  $p(n)$  and  $p(n_f)$  is significant. We present the  $p(n)$  and  $p(n_f)$  values for the events which pass the multiple hypothesis tests in figure.

We observe 5 out of 98 (5.1%) streaks in ODI cricket are significantly long, considering both their length ( $n$ ) and frequency ( $n_f$ ). In Test cricket, 6 out of 73 (8%) considering the length ( $n$ ) and 5 out of 73 (7%) considering the frequency  $p(n_f)$  are statistically significant. Because of the considered significance level, we expect an error rate of 0.05 in individual verification. In total we verified 98 possible streaks in ODI cricket and 73 streaks in Test cricket. The binomial probability for the observation of 5 hot hands in ODI cricket is 0.18 and more than 5 hot hands is 0.36. However, for the Test format, the probability of observing 5 and 6 hot hands are 0.14 and 0.08 and more than 5 and 6 are 0.15 and 0.07 respectively. This allows to conclude that we don't observe any Hot Hand effect in winning streaks of teams both in ODI and Test cricket.

### 3.5 Discussion

While many self-organizing systems aggregate information from its constituents to create global order [106], there is a class of systems like the stock market that digest the information from constituent entities to remove all possible global order in the system[107]. Team games are such a system aggregate information from the individuals constantly coordinating and performing in a stochastic environment, to generate the team performance and game outcome. Through this study, we quantify the predictability and persistence of individual and collective performances of the teams in a team game. We introduce a number of novel statistical tools to study the hot hand effect in a new dataset on game of Cricket. We quantify and exploit the self-excited patterns in individual and team performances to better predict the future compared to traditional methods like ARIMA.

Our investigation confirm the presence of significant hot-hands in individual performance. This is supported by the fact that the three highest performances in individual career cluster in time, particularly when players partake in hundreds of games.

Further, the shaded branching ratios in figure 4A. and 4B., are never found in simulated null data confirms the strength of the self-excitation supporting the hot-hand effect. The major finding is that these self-excitation patterns can indeed be exploited for predicting



future performances. The findings of this investigation complement those of earlier studies supporting the presence of hot hands in individual careers, while raising questions about the validity of those refuting the same.

Additionally, we show a significant reduction in prediction of team performances, suggesting the dominance of systemic stochasticity. While this is still predictable to certain extent, the outcome of the game can't be predicted nor they cluster in time. This leads to an interesting conclusion 'Cricket is a game of skill for individuals and a game of luck for the teams.'

Our study shows while an individual can consistently deal with the environmental systemic stochasticity, it is difficult for the team as a whole to perform equally well. Thus, these results open door for future research in the direction of the impact of group size in predictability and consistency of the performance. Furthermore, the present study establishes a quantitative framework for detecting and predicting the performances in individual careers. This approach will prove useful in expanding our understanding of the predictability of success in individual careers. This paper contributes to recent historiographical debates concerning the presence of hot hands in the sequence of successes in individual performances. Further work needs to be done to establish whether the presented methodology for predicting the performances can be improved for commercial usage and for financial gains exploiting the presence of self-excited patterns in individual careers. The findings of this study have a number of important implications for future research in the field of quantifying self-excited performance patterns involved in the study of human behavior and design of algorithms for predicting success.

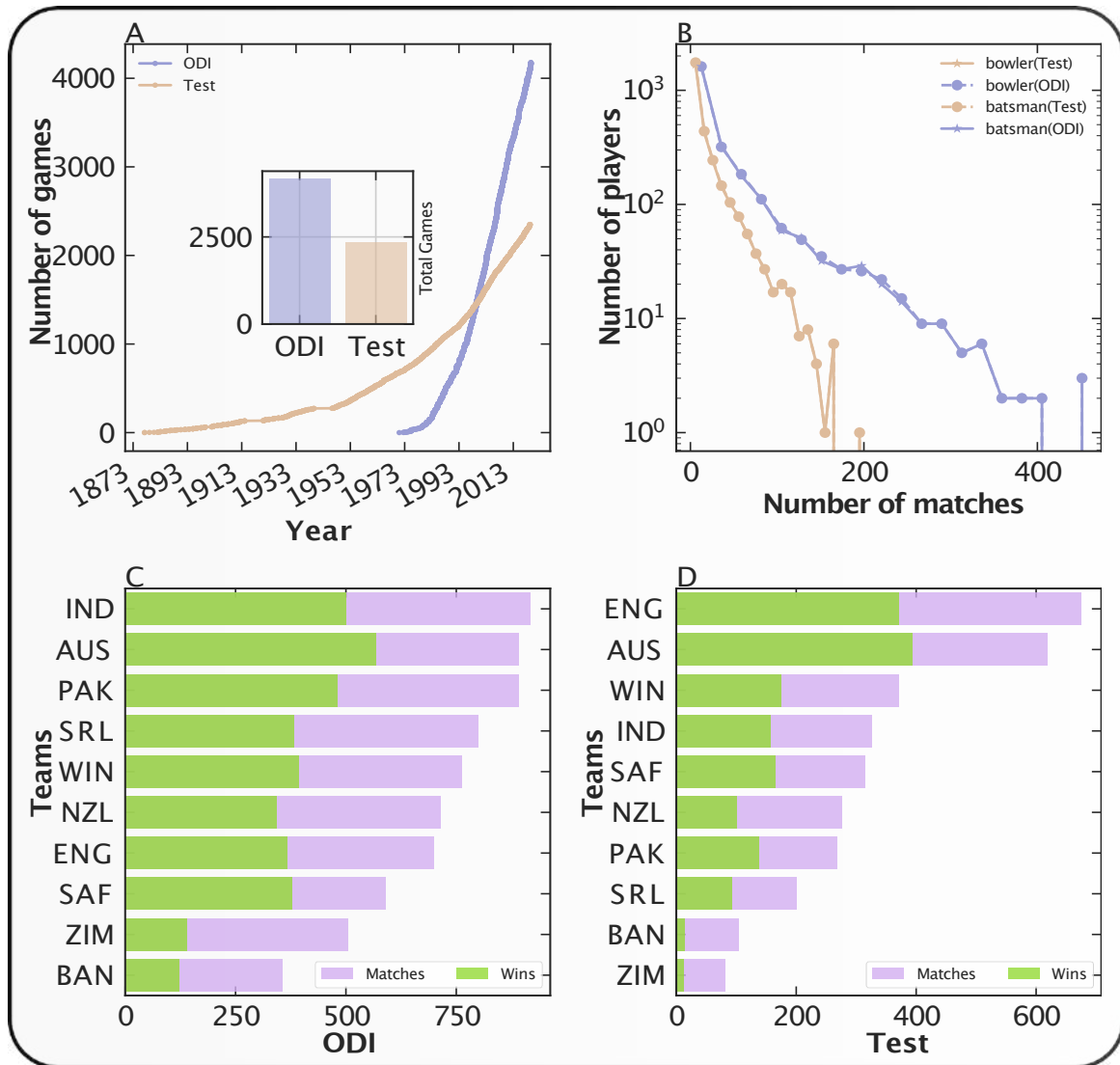


Figure 3.1 **Visual representation of the database.** . (A) The time evolution of the total number of games in *ODI* and *Test* format. The inset figure represents the total number of games in each format. The brown color represents the *Test* cricket and the purple color represents the *ODI* cricket. (B) The joint distribution of the total number of players against the total number of games played by each player. The four symbols shown in the inset can be paired since the number of batsmen and of bowlers coincide by construction of the game. The total number of *ODI* games played and the total number of games won by each country is presented in (C). The same statistics for *Test* cricket is presented in (D). The purple color represents the total games and the green color represents the total wins.

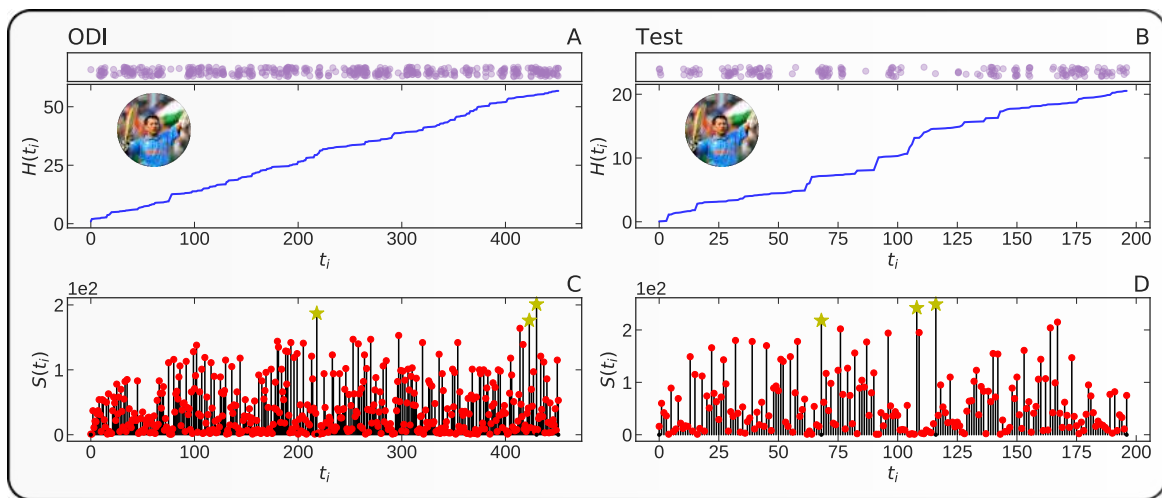


Figure 3.2 **Sequence of performances in the individual career of Sachin Tendulkar.** A) Performance time  $H(t)$  as a function of  $t$ , as defined in equation (3.1), for the highest performer in ODI cricket. (B) Performance time (3.1) for the highest run scorer in Test cricket. (C) The performance score  $S(t)$  of the player in ODI corresponding to panel (A). (D) The performance score  $S(t)$  of the player in Test cricket corresponding to panel (B). The large yellow stars represent the top 3 performances. The top insets in (A) and (B) give the point process representation of  $H(t)$ , in which each dot corresponds to an instant of time along the  $H(t)$  time axis. We have added noise along the y-direction for better visualization.

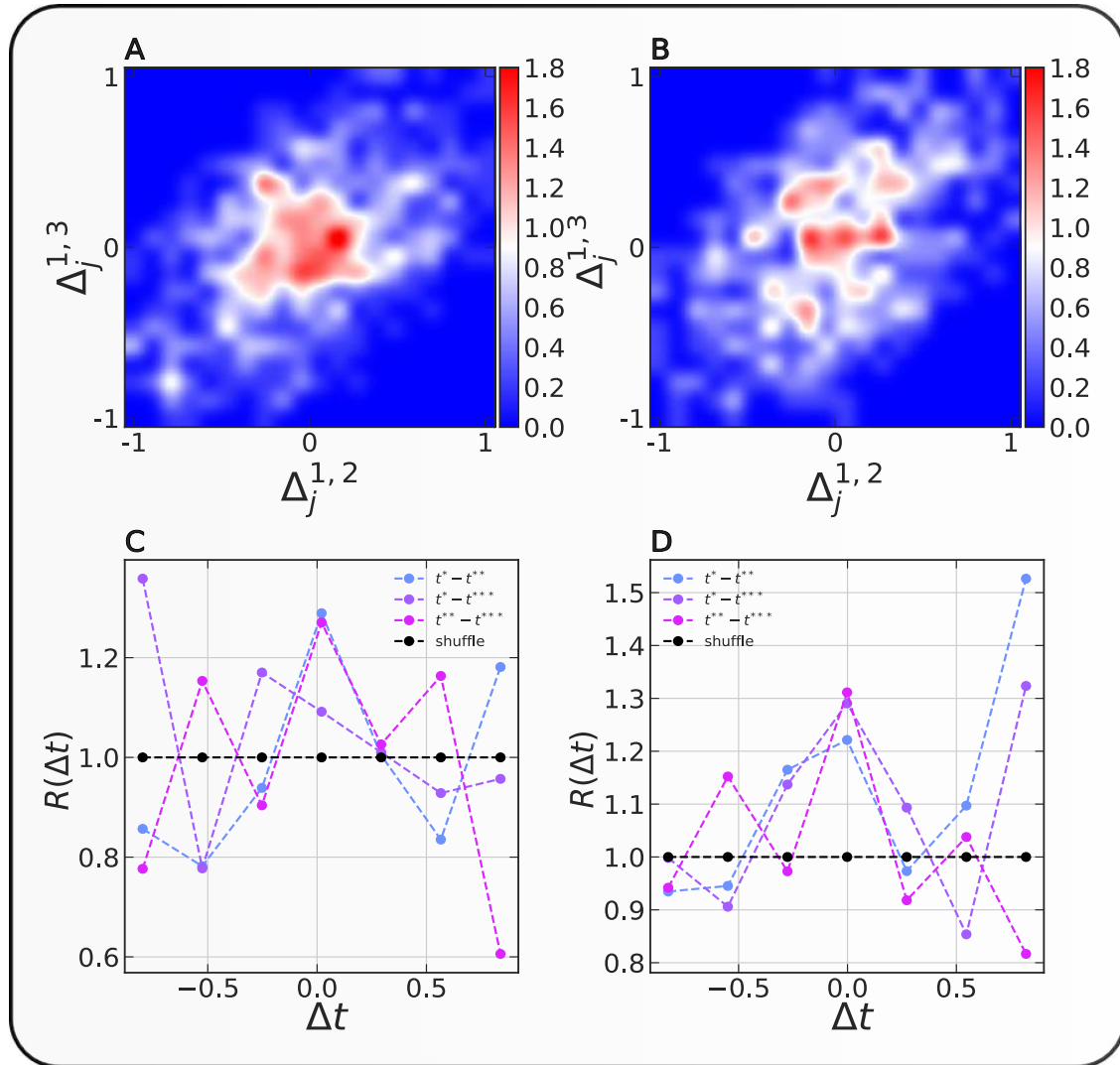


Figure 3.3 **Joint probability distribution**  $Q(\Delta_j^{1,2}, \Delta_j^{1,3})$  and  $R(\Delta t)$ . (A) and (B) show the joint distribution of the relative difference of the indices of second best from the best (defined by equation (3.5) in subsection 3.4.1), plotted against the third best from the best performances. (C) and (D) show  $R(\Delta t)$  defined by equation (3.6) for  $\Delta t = \Delta_j^{1,2}, \Delta_j^{1,3}, \Delta_j^{2,3}$ . (A) and (C) correspond to the batting performances in *ODI* cricket and (B) and (D) correspond to the performances in *Test* cricket.

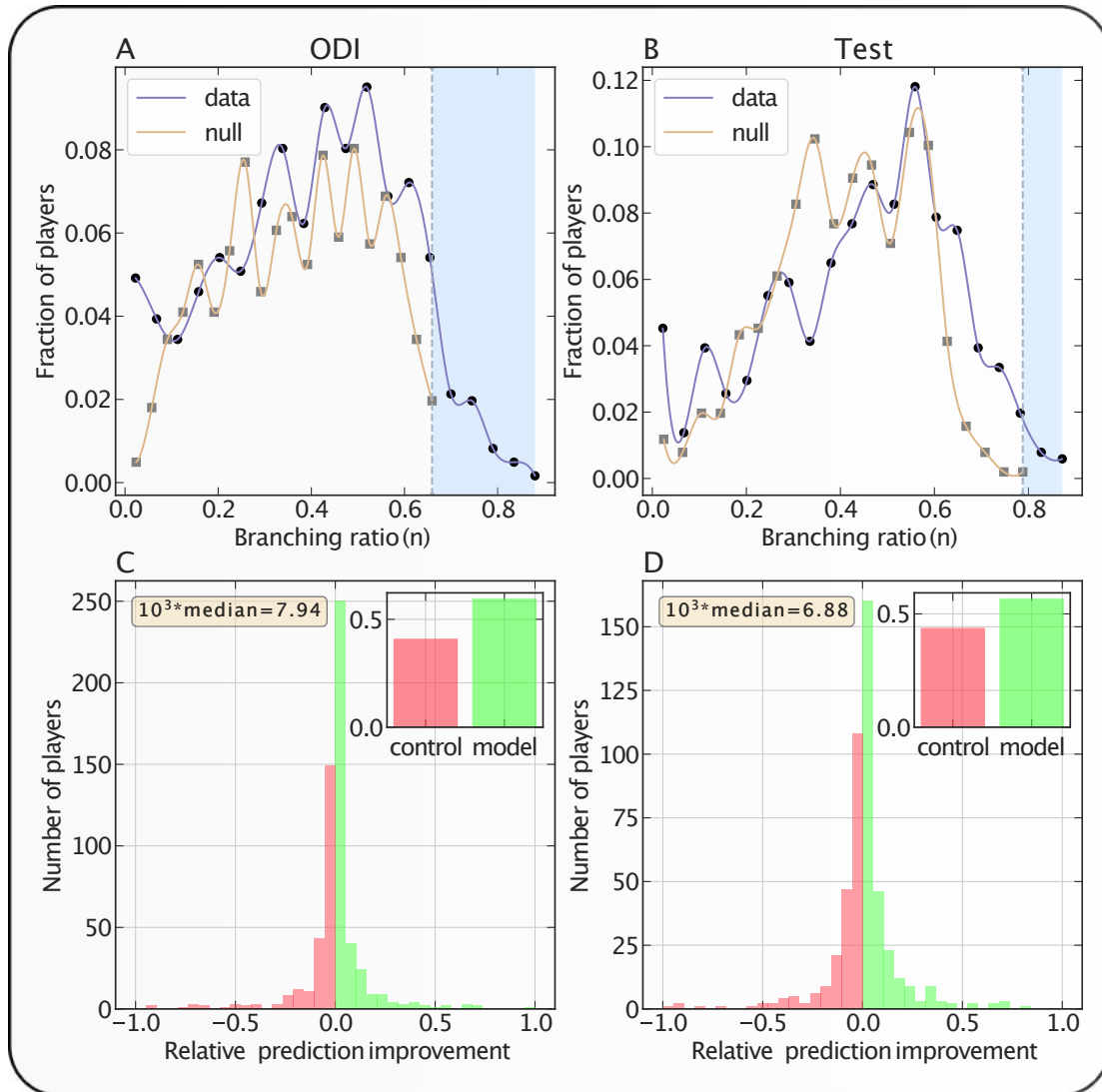


Figure 3.4 **Analysis of clustering in the time series of performance time with the self-excited point process model.** (A) and (B) represent the distribution of branching ratios over the set of players and of the branching ratios obtained from synthetic shuffled careers. (A) represents the distribution for *ODI* cricket and (B) is for Test cricket. Shaded regions in the plot represent the domain of branching ratios obtained from the real data that cannot be explained by the null models. (C) and (D) show the distribution of  $\delta(\mathcal{L}_j^{model}, \mathcal{L}_j^{control})$  (see equation (3.7)). (C) represents the distribution for *ODI* cricket, and (D) is for the Test format. The fraction of the times model experiments achieve a better log-likelihood score compared to the control experiments is colored green, otherwise the color is red. The insets show the fraction of control and model outperforming their counterparts. In *ODI* games, the fraction of times model experiment performs better than the control experiment is: 0.62; for Test cricket, this fraction is 0.60.

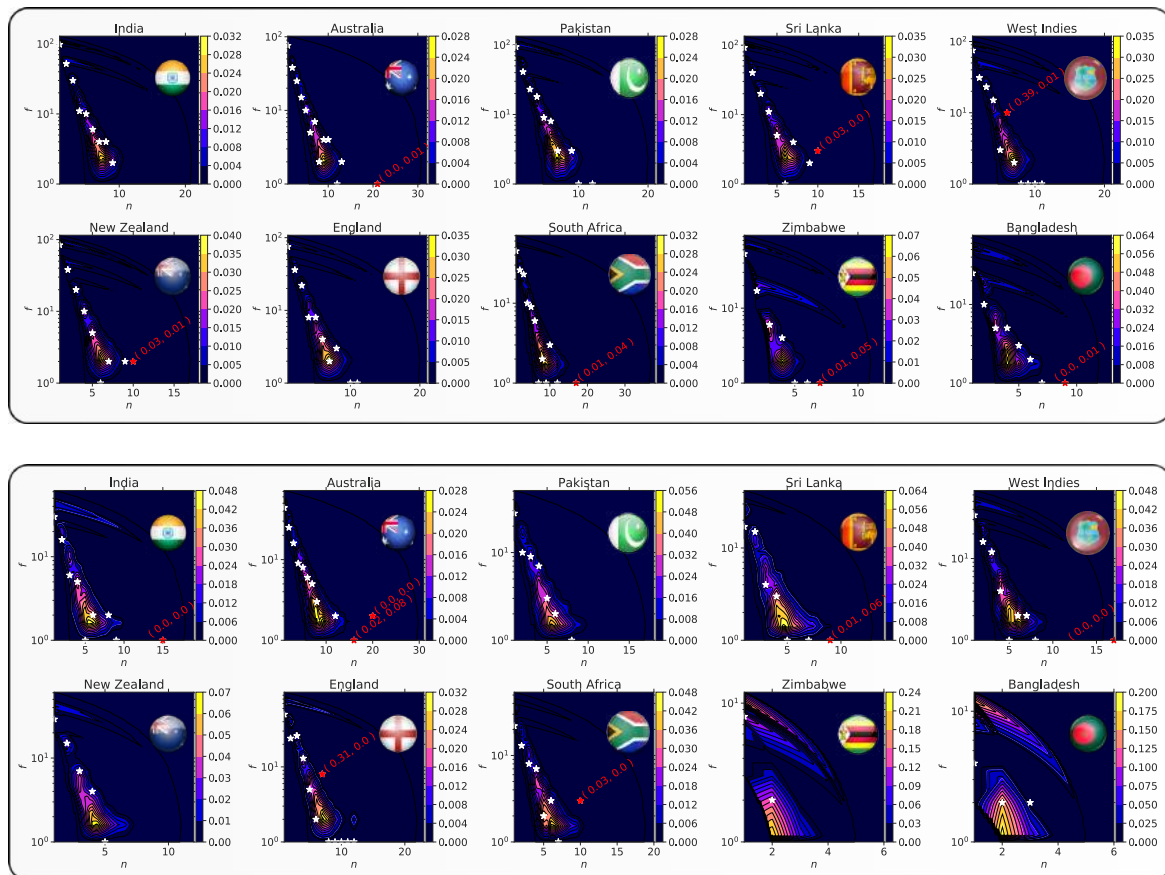


Figure 3.5 **Hot hands in cricket teams. (top panel) ODI:** ODI: Each of the 10 subplots in the figure shows the null distribution (obtained through randomly shuffling the performance sequence) of joint occurrence of winning streak length and of the corresponding frequency of occurrence. The title of each subplot provides the country of the team. Marked points on the plots represent the realized events. The white points represent the probable events and the red points represent the extreme/unlikely events (determined through multiple testing methods). The p-values ( $p(n)$  and  $p(n_f)$ ) (see equation (3.8) for definitions) for the unlikely events are provided along with the points. (Bottom panel) Test: Same as top figure for the performances in the Test format.

# Chapter 4

## Is seismicity operating at a critical point?

### 4.1 Introduction

Earthquakes have fascinated and continue to capture the imagination and interest of physicists, as they express how the huge, unbridled forces of Nature can be organized according to remarkable regular statistical laws obeying power-law statistics. The late Per Bak, one of the fathers of the concept of “self-organized criticality,” was fond of exclaiming in his talks: “I love this law,” when referring to (i) the Gutenberg-Richter (GR) distribution of earthquake seismic moments because it is valid over several decades larger than most known power laws in physical and social sciences. Several other scaling laws further characterize seismicity: (ii) the Omori law (the rate of aftershocks decays as  $\frac{1}{(t-t_m)^p}$  after a main shock that occurred at a time  $t_m$ ), (iii) a spatial Green function quantifying the power-law decay of the influence of the main shock as a function of the distance to its aftershocks, (iv) a power fertility law of the average number of aftershocks triggered as a function of the magnitude of the main shock, (v) power-law distributions of the lengths of the faults on which earthquakes occur, (vi) fractal, multifractal or hierarchical scaling of the set of earthquake epicenters as well as fault networks, and so on. For physicists, these laws suggest the existence of underlying physical mechanisms associated with some kind of criticality in the sense of phase transitions and field theory with zero mass. Indeed, many proposals in this spirit have been put forward to rationalize these power laws: self-organized criticality [108, 109] critical point behavior before large earthquakes [110–115], based on accelerating seismic release [116–119], combinations of the two [120–123].

The evidence for the criticality of the Earth’s crust is thus generally inferred from the presence of scale invariance and power-law scaling. But it is well-known that many other mechanisms can be at the origin of power-law scaling without the need for invoking criticality [124, 2]. In this context, the class of Epidemic-Type Aftershock Sequence (ETAS) models

offer a direct path to calibrate and quantify the distance to the criticality of earthquake catalogs. ETAS models combine the scaling laws (i)-(iv) to formulate a statistical framework that reproduces many of the statistical features observed in seismicity catalogs and is often used as the benchmark to assess the skills of competing earthquake forecasting models [125]. The ETAS models have an exact branching process representation [40], in which background earthquakes supposed to be driven by the forces of plate tectonics (immigrants in the language of epidemic branching processes) can trigger cohorts of earthquakes (first-generation “daughters”), themselves triggering second-generation events and so on [34]. The ETAS models exhibit a transition determined by the control parameter  $n$ , called the branching ratio, which is both the average number of triggered events of first-generation per immigrant and the fraction of all triggered events of any generation in the catalog [35]. The transition occurring at the critical value  $n = 1$  separates the sub-critical regime ( $n < 1$ ), where the sequence of events is stationary, from the super-critical regime ( $n > 1$ ) for which the number of earthquakes explodes exponentially with time with a finite probability [34]. The transition at  $n = 1$  is characterized by the standard signatures of criticality, such as diverging rate  $\mu/(1 - n)$  of events where  $\mu$  is the background rate, nonlinear “susceptibility” at  $n = 1$  [126] and various power-law statistics [127–130]. A lot of attention has been devoted to determining the empirical value of  $n$ , with most studies suggesting that  $n$  is very close to 1 [131, 50]. In contrast, some other studies averaging over broad tectonic areas find lower numbers in the range 0.35 to 0.65 [132]. If  $n \cong 1$ , a significant portion of the Earth’s crust would then function close to or even exactly at criticality, so that previous earthquakes trigger most observed earthquakes. If confirmed, this has far-reaching consequences for modeling, predicting and managing seismic risks.

Within the representation offered by ETAS models, determining whether the Earth crust is at criticality or not is crucially dependent on the ability to partition the observed seismicity clustering in space and time between the background rates and the triggered rates resulting from previous earthquakes. For instance, a misclassification of earthquakes as background events due to incomplete catalogs biases the calibrated branching ratio  $n$  downward [133, 134]. On the other hand, failing to account for spatial and temporal variations of the background rates leads to upward biases for  $n$ , as the variability may be incorrectly attributed to triggering. This effect has been recently demonstrated in the time domain for financial time series [135, 136]. Here, we revisit this question of the criticality of the Earth’s crust by offering an augmented ETAS model that improves on state of the art by accounting self-consistently for the possible spatial dependence of the key parameters.

Before presenting our method and results, we briefly review previous related studies [137–140] that have modeled the spatial dependence of the background rate  $\mu(x, y)$  using a kernel



density estimation. Generally, these studies use gaussian kernels with adaptive bandwidth. These kernel density estimates are obtained iteratively using the algorithm proposed by [137]. However, for all practical purposes, these studies fix the parameter on which the adaptive bandwidth of kernels depends, stating that the choice of this crucial parameter, which dictates the smoothing of background intensity in space, is relatively unimportant. Other studies [132, 141] pre-delineate regions in which all the parameters are assumed to be uniform and the parameters are then individually or jointly inferred in each of these regions. In some other studies, such as [142], the spatial probability density function (PDF) of background earthquakes is pre-estimated and then fed into the ETAS inversion machinery. Furthermore, [143, 50, 23] infer the optimal space variation of all ETAS parameters jointly, while more recently, [144] have proposed an alternative approach to non-parametrically model the space variation of background rate using a Gaussian process prior.

Here, we extend the iterative algorithm of [137] to jointly estimate the ETAS parameters and the kernel parameters used for obtaining the spatially variable background rate using the Expectation-Maximization (EM) algorithm of [141]. This extended algorithm leads to an augmented ETAS model that accounts specifically for the space variation of  $\mu(x, y)$ . Using this model, we show that the value of the branching ratio and other triggering parameters depend crucially on how the background rate is modeled. We demonstrate, using synthetic tests, that ignoring the spatial variation of the background rate leads to estimated parameters that are highly biased. However, when the spatial variation of the background rate is accounted for, the calibrated parameters are close to their true values and unbiased. We also document this bias effect in the case of three real catalogs. We then perform rigorous pseudo-prospective experiments and show that the ETAS model with spatially varying background rates significantly overperforms the ETAS model with uniform background rates. To the best of our knowledge, such direct comparisons have never been made between these two models. This gives us confidence that the extended ETAS model is superior to its spatially invariant counterpart, and thus reveals genuine characteristics of seismicity.

## 4.2 Methods

### 4.2.1 Self-consistent estimation of spatially variable background rate using an extended *EM* algorithm

Our implementation of the ETAS model expresses the seismicity rate  $\lambda(t, x, y | H_t)$ , at any time  $t$  and location  $(x, y)$  and conditional upon the history  $H_t$  of seismicity up to  $t$ , as eq. (4.1)

$$\lambda(t, x, y | H_t) = \mu(x, y) + \sum_{i: t_i < t} g(t - t_i, x - x_i, y - y_i, m_i) \quad (4.1)$$

$\lambda(t, x, y | H_t)$  receives contributions from the background intensity function  $\mu$  and from the sum over all past earthquakes that can trigger future earthquakes according to the triggering function given by eq. (4.2)

$$g(t - t_i, x - x_i, y - y_i, m_i) = \frac{K \exp[a(m_i - M_c)] \{t - t_i + c\}^{-1-\omega} e^{-\frac{t-t_i}{\tau}} * T_{\text{norm}} * S_{\text{norm}}}{\{(x-x_i)^2 + (y-y_i)^2 + d \exp[\gamma(m_i - M_0)]\}^{1+\rho}} \quad (4.2)$$

This expression combines the fertility law  $P(m) = K \exp[a(m_i - M_c)]$  that quantifies the expected number of first-generation aftershocks ( $\geq M_c$ ) triggered by an earthquake with magnitude  $m$ , the Omori-Utsu law  $\{t - t_i + c\}^{-1-\omega} e^{-\frac{t-t_i}{\tau}}$  and the spatial Green function, leading to the set  $\phi = \mu, K, a, c, \omega, \tau, d, \gamma, \rho$  of parameters that characterize the ETAS model.  $T_{\text{norm}}$  and  $S_{\text{norm}}$  are normalization constants for the time and space kernels, ensuring that they are proper probability density functions (PDF). With this formulation, the branching ratio is

$$n = \int_{M_c}^{M_{\text{max}}} P(m) \times f(m) dm \quad (4.3)$$

defined as the expected number of aftershocks of the first generation triggered by an earthquake, averaged over all magnitudes. The averaging over magnitude is thus performed using the GR distribution  $f(m) = b \ln(10) \frac{10^{-bm}}{10^{-bM_c} - 10^{-bM_{\text{max}}}}$ ,  $\forall M_c \leq m \leq M_{\text{max}}$ . Denoting  $\alpha = a/\ln 10$ , this yields eq. (4.4)

$$n = \begin{cases} \frac{Kb(1-10^{-(b-\alpha)(M_{\text{max}}-M_c)})}{(b-\alpha)(1-10^{-b(M_{\text{max}}-M_c)})}, & \forall \alpha \neq b \\ \frac{Kb \ln(10)(M_{\text{max}}-M_c)}{(1-10^{-b(M_{\text{max}}-M_c)})}, & \text{if } \alpha = b \end{cases} \quad (4.4)$$

We consider two variants of the ETAS model:  $ETAS_{\mu}$ , which features a spatially uniform background rate  $\mu$ , and  $ETAS_{\mu(x,y)}$ , which possesses a space varying background rate  $\mu(x, y)$ . In  $ETAS_{\mu(x,y)}$ ,  $\mu(x, y)$  is informed by the spatial positions of previous earthquakes. The proposed parameterization given below in eq. (4.5) should not be confused with the triggering part of the ETAS model given by the second term in the r.h.s. of eq. (4.1), which also involves a summation over previous earthquakes. Here, the guiding idea is that observed earthquakes occur more frequently where the background intensity is larger because the background events are, by definition, the sources of all observed seismicity. Even if many earthquakes are triggered by previous earthquakes, their locations are related to that of their background sources [121, 35].

$$\mu(x,y) = T^{-1} \sum_{i=1}^N IP_i \pi^{-1} Q D^{2Q} \left( (x-x_i)^2 + (y-y_i)^2 + D^2 \right)^{-1-Q} \quad (4.5)$$

To estimate  $\mu(x,y)$ , we extend the EM algorithm proposed by [141] as described in details in section C.1 of the appendix. In this algorithm,  $\mu(x,y)$  is inferred by smoothing the location of background earthquakes, which are dynamically obtained during the inversion using a power-law kernel [145] with a bandwidth  $D$  and exponent  $Q$ , which are estimated along with  $\phi$ .

### 4.2.2 Dataset

We apply the extended EM algorithm (section C.1 of appendix) to catalogs obtained from two sources: the Advanced National Seismic System (ANSS) and GeoNet. The ANSS catalog is used for two study regions: the entire globe and the region around California. For the area around New Zealand, we use the GeoNet catalog. The location of 600,000 earthquakes ( $M \geq 3$ ) for the entire globe between 1975-2020, as reported in the ANSS catalog, and of 1.2 million and 600,000 earthquakes ( $M \geq 1$ ) between 1975-2020 in the study regions surrounding California and New Zealand, as reported by the ANSS and Geonet catalogs respectively, are shown in figs. C.1 to C.3 of the appendix. section C.2 in appendix also presents the method to select the magnitude of completeness  $M_c$  for each catalog.

### 4.2.3 Parameter calibration

The results of the calibration of the two ETAS models,  $ETAS_\mu$  and  $ETAS_{\mu(x,y)}$ , on the global, Californian, and the New Zealand catalogs are presented in table 4.1. When going from the  $ETAS_\mu$  model to the  $ETAS_{\mu(x,y)}$  model, the most remarkable changes are that (i) the overall background rate increases by nearly 29, 3, and 16 times for the three catalogs, respectively, while consequently (ii) the branching ratio  $n$  is substantially smaller for the  $ETAS_{\mu(x,y)}$  model compared to the  $ETAS_\mu$  model. The other parameters also show substantial changes. More specifically, the branching ratio is remarkably close to 1 for the three catalogs when calibrated with the  $ETAS_\mu$  model, which would lead to the erroneous conclusion that the Earth crust is critical, as discussed in the introduction. In contrast, the  $ETAS_{\mu(x,y)}$  model gives  $n \cong 0.45$ , 0.79, and 0.61 for the global, Californian, and the New Zealand catalogs, respectively, clearly excluding criticality and qualifying the Earth crust in the sub-critical regime. The difference between the two spatial intensity of background earthquakes inferred from the calibrations of the two models is vividly illustrated in fig. 4.1 (a and b) in the case of the global catalog (and figs. C.4 and C.5 in the appendix for the Californian and New Zealand catalogs). We

find that not only is the overall background rate is different between the two models, but also that the spatial patterns of the density of background earthquakes differ substantially between the two modeling choices, as one can observe the much more refined localization of the background rates along plate boundaries in  $ETAS_{\mu(x,y)}$ .

## 4.3 Results

### 4.3.1 Synthetic tests of the bias in branching ratio $n$ due to uniform background rate

We now demonstrate that a realistic synthetic catalog with a relatively smaller branching ratio as found with the full  $ETAS_{\mu(x,y)}$  model calibrated on the real catalogs yields a spuriously large branching ratio when calibrated with the  $ETAS_{\mu}$  model, and recovers the correct value when calibrated with the full  $ETAS_{\mu(x,y)}$  model. For this, using the full  $ETAS_{\mu(x,y)}$  model, we simulate a 50-year long synthetic catalog [23] with earthquakes of magnitudes  $M \geq 5$  for the entire globe using the parameters that are listed in table 4.1 corresponding to  $ETAS_{\mu(x,y)}$ . Using the first ten years of the catalog as the auxiliary period and the remaining as the primary period [146], we calibrate the  $ETAS_{\mu}$  model on this synthetic catalog. The obtained parameters are:  $N_{bkg} = 22.27 \text{ year}^{-1}$ ,  $\log_{10} K = -0.18$ ,  $a = 1.07$ ,  $\log_{10} d = 2.17$ ,  $\rho = 0.73$ ,  $\gamma = 0.78$ ,  $\log_{10} c = -3.09$ ,  $1 + \omega = 0.80$ ,  $\log_{10} \tau = 3.9$ ,  $n = 1.17$  (see table 4.1). This should be compared with the true input parameters  $N_{bkg} = 1,013.2 \text{ year}^{-1}$  for the background rate and  $n = 0.45$  for the branching ratio for the global catalog (table 4.1). The background rate inverted with the  $ETAS_{\mu}$  model is too low, and the inferred branching ratio is too large, being remarkably close to the value inferred by inverting this model with uniform background rate  $\mu$  on the real catalog, on nearly all the parameters. For instance, the Omori exponents inferred for the synthetic catalog and the real catalog are 0.80 and 0.79, respectively. This provides an excellent self-consistent test and further supports the validity of our hypothesis that the background rate  $\mu(x,y)$  is strongly varying in space. This suggests that it is absolutely essential to account for the non-uniform background rate to obtain unbiased parameter estimates. These synthetic tests also show that the biases are predictable.

### 4.3.2 Pseudo-prospective forecasting experiments as a validation step for $ETAS_{\mu(x,y)}$

We now proceed to show that  $ETAS_{\mu(x,y)}$  leads to operationally better forecasts of future seismic activity by setting up several 30 day long pseudo-prospective forecasting experiments

at different spatial resolutions and magnitude thresholds  $M_t$  of the validation catalog. For details on these pseudo-prospective experiments, we refer the readers to section C.3 in the appendix. For more general discussions on the importance of these experiments and their design, we refer the readers to [23, 146]. At a given spatial resolution and magnitude threshold ( $M_t$ ), the log-likelihood score of a model during a given testing period is defined as  $LL = \sum_{i=1}^N \ln P(n_i)$ , where  $P(n_i)$  is the probability of observing  $n_i$  earthquakes in the  $i^{th}$  pixel during the testing period, and  $N$  is the total number of equal area pixels that tile the study region. In any pixel, the probability is constructed using the number of earthquakes observed in different simulated catalogs as described in section C.3 in the appendix.

Once the likelihoods for two models are calculated, the information gain of one over the other is simply the difference of their log-likelihoods. fig. 4.2 (a-b) shows the time-series of cumulative information gain (CIG) that the  $ETAS_{\mu(x,y)}$  model obtains over the  $ETAS_{\mu}$  model in the 368 experiments that we perform with the global catalog (see similar figs. C.6 to C.8 of the appendix for the time series of CIG at different spatial resolutions and  $M_t$  for Global, Californian and New Zealand catalogs). At all spatial resolutions and magnitude thresholds, the  $ETAS_{\mu(x,y)}$  substantially outperforms the  $ETAS_{\mu}$  model for all the three study regions.

To quantify if this over-performance of the  $ETAS_{\mu(x,y)}$  is statistically significant, we define the mean information gain as the average information that  $ETAS_{\mu(x,y)}$  obtains over the 368 testing periods. We then test the null hypothesis that this mean information gain is significantly larger than 0 against the alternative that it is not, using the student's t-test. fig. 4.2 (c-d) (and similar figs. C.9 to C.11 for all spatial resolutions and  $M_t$  for the Global, Californian, and New Zealand catalogs) confirm that the mean information gain of the  $ETAS_{\mu(x,y)}$  over  $ETAS_{\mu}$  is significantly larger than 0, thus rejecting the null hypothesis at significance levels much smaller than 0.01 in all cases.

## 4.4 Discussion and Conclusions

Our results demonstrate that, when the spatial variation of background rate is appropriately accounted for, we get a superior forecasting model and a branching ratio that is much smaller than 1.

At a fundamental level, the non-critical value of the branching ratio  $n$  invites a reexamination of the physical picture we have of the brittle rupture process in the Earth's crust. Until now, large values of  $n$  close to unity have been reported, suggesting that the loaded fault network is in a permanent critical state, compatible with the popular concept of self-organized criticality. The much lower value of  $n$  that we estimate using more appropriate assumptions and a superior algorithm suggests that fault networks mainly evolve far from a

critical point. This has major implications for the prediction of large events. Indeed, in the self-organized critical scenario, each event is no different from all others from a generating process viewpoint, making prediction impossible [147]. In contrast, if the fault network remains most of the time far from criticality, more sporadic singularities may appear via various possible mechanisms and announce upcoming catastrophic events. This suggests, for instance, the need for a reevaluation of the Accelerated Moment Release hypothesis [118], benefitting from the prior use of the  $ETAS_{\mu(x,y)}$  fitting model to better eliminate the contribution of the uncorrelated part of seismicity to the total seismicity rate.

Finally, our finding that seismicity operates in a sub-critical regime should not be confused with the phenomenon of sub-critical rupture of a single fault [148], where the stress is below the elastic stress threshold and time-dependent plastic deformations and micro-cracking occur slowly at and around the crack tip and at asperities. The finite-time singularity ending sub-critical crack growth [10] is a one-rupture problem. In contrast, the criticality (and its absence) of a large collection of interacting earthquakes is a many-body problem.

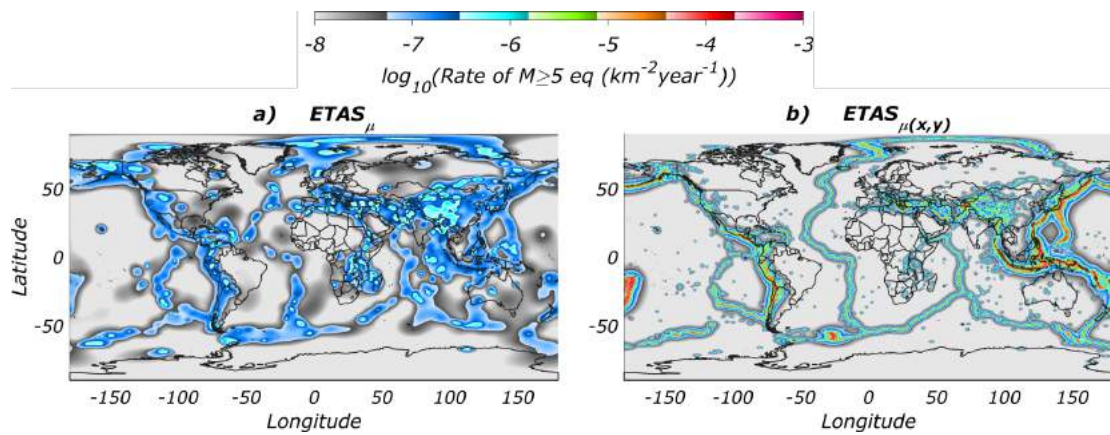


Figure 4.1 Spatial density of the earthquakes identified as background events by the  $ETAS_{\mu}$  and  $ETAS_{\mu(x,y)}$  models for the global catalog

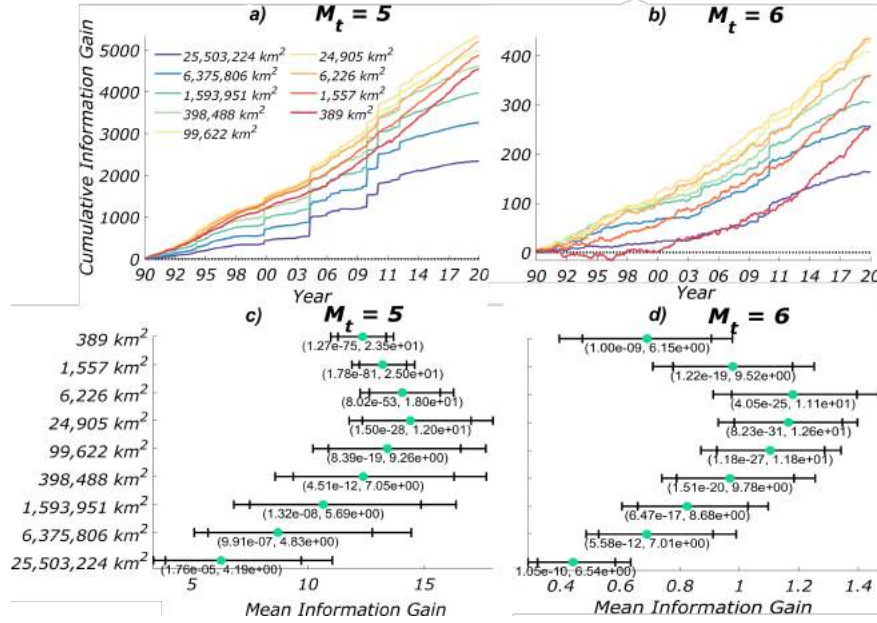


Figure 4.2 (a,b) Time series of cumulative information gain of  $ETAS_{\mu(x,y)}$  over the  $ETAS_{\mu}$  model in 368 pseudo prospective experiments with the global catalog, at nine spatial resolutions (different colors whose meaning is given in the inset) and two magnitude thresholds (different panels) of the testing catalog; (c-d) Mean information gain (MIG) of  $ETAS_{\mu(x,y)}$  over the  $ETAS_{\mu}$  model in 368 pseudo prospective experiments for the Global catalog, at nine spatial resolutions and two magnitude thresholds (different panels) of the testing catalog. The error bars indicate the 99 and 95% confidence interval of the mean information gain (MIG). The two numbers indicate the p-value and t-statistic resulting from the student's t-test, in which we test the null hypothesis that the MIG is equal to 0 against the alternative that it is larger than 0. When the p-value is smaller than 0.05, the null hypothesis is rejected.

Catalog	Model type	$N_{bkg}(\text{year}^{-1})$	$\log_{10}K$	$a$	$\log_{10}d$ (km <sup>2</sup> )	$\rho$	$\gamma$	$\log_{10}c$ (days)	$1 + \omega$	$\log_{10} \tau$	Branching Ratio $n$	$D$ (km)	$Q$
Globe	$ETAS_{\mu}$	35.35	-0.12	0.84	2.20	0.71	0.56	-3.16	0.79	3.81	1.15	NA	NA
Globe	$ETAS_{\mu(x,y)}$	1,013.2	-0.72	1.39	1.93	1.03	0.99	-2.08	1.06	3.41	0.45	13.46	0.74
California	$ETAS_{\mu}$	60.73	-0.22	0.95	-0.59	0.52	1.11	-2.94	0.94	3.87	1	NA	NA
California	$ETAS_{\mu(x,y)}$	182.42	-0.37	1.11	-0.73	0.59	1.25	-2.64	1.01	3.67	0.79	4.39	1.07
New Zealand	$ETAS_{\mu}$	9.08	-0.17	1.06	1.6	0.75	0	-3.01	0.89	3.76	1.14	NA	NA
New Zealand	$ETAS_{\mu(x,y)}$	149	-0.58	1.5	1.17	0.76	0.44	-2.28	1.09	3.37	0.61	11.79	1.49

Table 4.1  $ETAS$  parameters inverted for three catalogs (first column) using the two models  $ETAS_{\mu}$  and  $ETAS_{\mu(x,y)}$ . The branching ratio  $n$  is inferred using Eq.(4) with  $M_{max} = 10, 8.5$  and 9 for the Global, Californian, and New Zealand catalog, respectively. Parameters  $c$  and  $t$  are given in days.  $N_{bkg}$  is the total number of background events per unit time.





# Chapter 5

## Impact of Governmental interventions on epidemic progression and workplace activity during the COVID-19 outbreak

### 5.1 Introduction

The pandemic due to the SARS-CoV-2 virus [149] has been impacting the world population, health care system and economies over the first half of 2020 [150]. Since its identification in December 2019 in Wuhan, China, this novel coronavirus disease (COVID-19) has been spreading in China in Jan.-Feb. 2020. The epidemic development was detected in Italy in the second half of February and progressively diffused in the rest of the world. It was declared a global pandemic on March 11, 2020 by the World Health Organization (WHO) [151]. We have all witnessed and experienced a series of interventions with various levels of confinement measures in different countries and regions, aimed at decreasing the effective reproduction number  $R_t$  and controlling the epidemics.

Our starting hypothesis is that the development of infected cases, of the various forms of illnesses resulting from infection and of the death rates reflect the interplay between the biological and epidemiological properties of this new SARS-CoV-2 virus and the political, cultural, sociological and governance characteristics of different nations and human communities. By combining mobility data, epidemiological data and clinical data across ten countries, we develop a modelling framework to quantify the effectiveness of the interventions implemented by various countries to control the epidemic growth. This shock provides a real-life natural experiment to falsify the effectiveness of different organisations

and interventions and policies, with the goal of informing and guiding future plans against potential second and third waves of the epidemics as well as future outbreaks in general.

First, we quantify precisely how interventions in the form of mobility restrictions and social distancing have had significant impacts in controlling the development of the epidemics in many countries, as measured by the decrease in reproduction number  $R_t$  and the extend of curbing the increase of infected cases. We also document a surprisingly large heterogeneity in the reduction of  $R_t$  across regions within a given country and also across countries. Further, we observe non monotonous and fluctuating time dependence of  $R_t$  in different regions.

Our most surprising result is that, in many regions where the epidemic was not visible, the interventions led to a significant transient increase in reproduction number  $R_t$ , in contradiction with the short-term objective of lockdown measures. For instance, in several regions of France and Italy, lockdown led to an obvious strong collapse of mobility, however accompanied by an increase of  $R_t$ . We note that this is not due to the inherent delay in case detection due to incubation period, symptoms' onset and testing. We interpret this phenomenon as a result of preemptive large movements of people to relocate before the strict lockdown implementation, hence promoting new contagions and epicenters for the epidemics to mature for a while after the lockdown. Another plausible mechanism, which underlies the Japanese policy, is the effect of closed spaces and close-contact settings within confined households.

We also quantify how the timing of lockdown determined the trajectory of the epidemic by estimating the spatial correlation across regions within a country of the total increase of infected cases after the lockdown. In a number of countries, the epidemic started much before the lockdown date and was developing silently as revealed by the strong spatial diffusion of the epidemic after the lockdown. In other countries, the epidemic was better contained by intervention measures.

Our analysis overall suggests that the interventions may have not been optimal and that there are probably better alternatives to complete lockdowns. Our study reveals the importance of timing and targeting of interventions as a likely better strategy compared with undifferentiated lockdown. We observe belated intervention at the regional and local levels and hasty global lockdowns, which, in a number of regions, result in disappointing reduction of the reproduction number and curbing of the increase of new cases compared to other comparable countries.

Country	China	Italy	Spain	Switzerland	France	Canada	United States	Germany	India	Japan
Intervention date	2020-01-23	2020-03-09	2020-03-14	2020-03-16	2020-03-17	2020-03-18	2020-03-19	2020-03-20	2020-03-25	2020-04-07

Table 5.1 Starting dates of Governmental interventions (includes the lockdown) for controlling the outbreak. (source: [https://en.wikipedia.org/wiki/COVID-19\\_pandemic\\_lockdowns](https://en.wikipedia.org/wiki/COVID-19_pandemic_lockdowns))

## 5.2 Materials and Methods

### 5.2.1 Data

#### Clinical Data

**Incubation period and confirmation period:** To calculate various epidemiological parameters, we use the individual-level data [152, 153] on COVID-19 epidemic which are geo-coded and include symptoms, key dates (date of onset, admission, and confirmation), and travel history. The dataset is curated from different sources, including official government sources (official websites of Ministries of Health or Provincial Public Health Commissions), peer-reviewed scientific articles, online reports, and news websites. From this database, we select 216 individuals, who got a clinical confirmation for covid-19 and were symptomatic as well as traveled to suspected places of exposure. We have the exact dates of travel, onset of symptoms, and of clinical confirmation for these 216 individuals.

**Serial interval:** We use the infector-infectee pairs dataset that has been collected from publicly available information published in research articles and quoted from official reports of outbreak investigations [154]. We use 28 probable pairs for the estimation of generation time. For these 28 pairs of individuals, the date of illness onset for the pairs of individuals are defined as the date on which a symptom relevant to COVID-19 infection appeared and are determined by the reporting governmental body.

#### Epidemiological Data

We collect the daily reported cases at the first administrative level divisions for a list of countries from various sources. The primary sources of information for the datasets are often local news agencies, government reports, WHO reports, various medical communities. Table 5.1 gives a list of countries with their starting date of the Governmental interventions. There is no lockdown in Japan like other countries, however, prime minister Abe on 7 April, proclaimed a state of emergency. This was the first emergency declaration in Japan and we consider this date as the starting date of intervention.

## Mobility Data

We use the aggregated anonymised community mobility dataset that has been collected through the Google Maps app [155], to help understand what has changed in response to the Govt. policies aimed at flattening the curve of the COVID-19 pandemic. The dataset is anonymised to ensure that no personal data, including an individual's location, movement, or contacts, can be derived from the resulting metrics. The anonymization process for the data includes differential privacy [156], with intentionally added random noise to metrics in a way that maintains both users' privacy and the overall accuracy of the aggregated data [157]. The dataset contains the percentage changes of the anonymized mobility metrics of Google users from a baseline based on the historical part.

## Map Data

We collect the geojson files for the first-level administrative divisions of the list of countries containing the names of the regions and the geometry from a number of openly accessible github repositories.

## 5.2.2 Models and methodology

For our analysis, we select a time span that starts 10 days before the implementation of Governmental interventions – for controlling the outbreak, often through mass lockdowns – and ends 30 days following the starting date of Governmental interventions.

### Epidemiological estimations

The time between exposure and onset of symptoms is defined as the incubation period. We use a Bayesian framework to model the day of onset of symptoms following the date of exposure to the virus and the day of clinical confirmation of the virus following the date of exposure. Using the key dates for the selected 216 individuals [152, 153], we estimate the important epidemiological parameters like incubation period, confirmation period (the number of days it take to get the clinical confirmation about the virus following the date of exposure). We use the logistic model

$$P(\text{conf./sympt.}|t) = \frac{1}{1 + e^{\beta t + \alpha}} \quad (5.1)$$

to estimate the cumulative probability distribution of individuals developing symptoms or getting the clinical confirmation on the  $day = t$  that they are infected, conditional on

the fact that the individual was exposed to the virus on  $day = 0$  and will be eventually symptomatic and will get a clinical *+ve* confirmation result. With the help of Metropolis-Hastings algorithm [158], – a Markov Chain Monte Carlo (MCMC) method [159]– we sample the model parameters  $(\alpha, \beta)$ , from a standard normal prior and train our model on the dataset to estimate the posterior probability distribution of developing symptoms or getting the clinical confirmation on the  $day = t$ , following the day of exposure on  $day = 0$ .

### **Estimation of the time-dependent effective reproduction number $R_t$**

The principal epidemiological variable characterizing a disease’s transmission potential is the basic reproduction number,  $R_0$ , which is characterized as the estimated number of secondary cases caused by a typical primary case, in an entirely susceptible population. When infection is spreading across a population, working with the effectively reproductive number  $R_t$ , which is defined as the actual average number of secondary cases per primary case, is often more convenient.  $R_t$  is normally smaller than  $R_0$ , which reflects the impact of epidemic controls and the decline of susceptible individuals. The value of  $R_t$  is comparable to the branching ratio in the Hawkes process [50, 11], where  $R_t > 1$  can lead to an explosive growth of the epidemic and an ever-increasing number of new cases, while  $R_t < 1$  leads to the eventual demise of the growth process. Provided the branching structure describing who infected whom is determined, it becomes trivial to estimate  $R_t$ . However, this information is not generally available and the estimation of  $R_t$  then becomes tricky. Nevertheless, recent advances in epidemiology has made it possible to estimate the time evolution of the effective reproduction number ( $R_t$ ) through the observed epidemic curve in a geographical region [160, 161]. We use the method introduced by [161], which uses the daily number of confirmed cases and a model described below for the generation time (which is one of the key parameters dictating the severity of epidemic growth) to estimate the temporal evolution of effective reproduction number ( $R_t$ ) with the help of a Sequential Bayesian estimation approach.

The generation time is defined for source-recipient transmission pairs as the time between the infection of the source and the infection of the recipient. Because time of infection is generally not known, the generation time is often approximated by the serial interval, which is defined as the time between the onset of symptoms of the source and the onset of symptoms of the recipient. For the present case, we use the data for the serial intervals from [154], which has been constructed using the publicly available data. We calibrate the data against three models, i.e., Weibull, Log-normal, Gamma distributions to find the best possible approximation model from which the generation time is determined.

We use a probabilistic contagion model with in-homogeneous source terms to explain the progression of the COVID-19 epidemic [161]. We consider both the human to human

transmission and infections from the reservoir (contaminated surfaces) to explain the epidemic growth. Denoting  $S(t)$  and  $N(t)$  as the average number of susceptible and total population at time  $t$  and  $\beta$  and  $\gamma^{-1}$  as the contact rate and the infectious period, then

$$R_0 = \frac{\beta}{\gamma} \quad \text{and} \quad R_t = \frac{S(t)}{N(t)} \times R_0 . \quad (5.2)$$

According to this model, if we denote the number of new infections from the reservoir between  $t$  and  $t + \tau$  by  $\Delta B(t)$  and the number of new cases within this period by  $\Delta T(t + \tau)$ , then the stochastic discrete variable  $\Delta T(t + \tau)$  is generated by a probability distribution with average number of cases given by

$$\Delta T(t + \tau) \sim P\{\lambda\} \quad (5.3)$$

with

$$\lambda = \Delta B(t + \tau) + b(R_t) (\Delta T(t) - \Delta B(t) + \tau\gamma R_t \Delta B(t)) \quad (5.4)$$

where  $P\{\lambda\}$  denotes a discrete probability distribution with mean  $\lambda$ . In eq. (5.4),  $b(R_t)$  can be expressed as

$$b(R_t) = \exp[\tau\gamma(R_t - 1)] , \quad (5.5)$$

which is the slope of the tangent at the origin of case trajectories from the epidemic time delay plot ( $\Delta(t)$  vs  $\Delta(t - \tau)$ ) of surveillance data. More details about the derivation of the model can be found supplementary material or in [161].

We use a Bayesian framework to estimate the full probability distribution for the effective reproduction number  $R_t$ , conditional on the time series for new cases. The probability distribution of  $R_t$ , compatible with the observed temporal data stream, is given by

$$P[R_t | \Delta T(t + \tau) \leftarrow \Delta T(t)] = \frac{P[\Delta T(t + \tau) \leftarrow \Delta T(t) | R_t] P[R_t]}{P[\Delta T(t + \tau) \leftarrow \Delta T(t)]} , \quad (5.6)$$

where  $P[R_t]$  is the prior distribution and  $P[\Delta T(t + \tau) \leftarrow \Delta T(t)]$  is independent of  $R_t$ , and is a normalization parameter. From successive applications of Bayes' theorem, a sequential estimation scheme, that uses streaming epidemiological observations performed in real time, can be constructed using the posterior distribution for  $R_t$ , at time  $t$  as the prior in the next estimation step at time  $t + \tau$ , leading to an update scheme via iteration of eq. (5.6). The resulting probability distribution for  $R_t$  includes information on all observations up to time  $t$ , and thus is a robust estimator of the effective reproduction number compared to the estimation

by only considering the cases between  $t$  and  $t + \tau$ . Any changes in  $R_t$  over time result from the assimilation of each new data point, leading to an updated estimate of  $R_t$ . This in turn allows us the use of the estimation procedure as an anomaly detection tool.

### Estimation of the causal impact of Governmental intervention

We use the estimated time evolving effective reproduction number and the time evolving mobility metric to study the causal impact of various Governmental interventions to contain the spread of COVID-19. We use the framework proposed by [162] to infer the causal impact on the epidemic progression as well as the human mobility because of various Governmental interventions. The method uses a diffusion-regression state-space model that predicts the counterfactual evolution of effective reproduction number  $R_t$ , as well as the mobility metric in a synthetic control that would have occurred, had no intervention taken place. The model is successful in inferring the temporal evolution of attributable impact, and flexibly accommodates multiple sources of variation, including local trends, seasonality and the time-varying influence of contemporaneous covariates by incorporating empirical priors on the parameters in a fully Bayesian framework. Using the MCMC for posterior inference, we estimate the most likely counter-factual evolution of effective reproduction number and the mobility metric during the first 30 days of the interventions.

According to the model, the generalized Bayesian Structural Time Series – state-space models for time-series data – can be expressed by

$$y_t = Z_t^T \alpha_t + \beta X_t + G_t \varepsilon_t, \quad (5.7)$$

$$\alpha_{t+1} = T_t \alpha_t + H_t \eta_t, \quad (5.8)$$

where  $\varepsilon_t \sim \mathcal{N}(0, \sigma_t^2)$  and  $\eta_t \sim \mathcal{N}(0, Q_t)$  are independent of all other unknowns,  $\alpha_t$  is referred to as a “state” of the series and  $y_t$  is a linear combination of the states plus a linear regression with the *covariates*  $X$ . Eq. 5.8 is the state equation governing the evolution of the state vector  $\alpha_t$  through time, whereas  $y_t$  is a scalar observation. It is possible to model several distinct behaviors for the time series (including *ARMA* or *ARIMA*) by varying the matrices  $Z_t^T$ ,  $T_t$ ,  $G_t$ ,  $Q_t$  and  $H_t$ .

Here, we simplify (5.8) by taking  $T_t = H_t = 1$ , so that

$$\alpha_{t+1} = \alpha_t + \eta_{\mu,t} \quad (5.9)$$

is a simple random walk, also referred to as the “local level” component. This random walk component embodies the increasing uncertainty of observations as time passes. In (5.7),

we take  $Z_t^T = G_t = 1$  but augment the equation by accounting for the possible presence of seasonal components embodied into  $\gamma_t$  described below. This allows us to reduce expression (5.7) into

$$y_t = \alpha_t + \gamma_t + \beta X_t + \varepsilon_t . \quad (5.10)$$

The seasonal components in eq. (5.10) can be expressed as

$$\begin{aligned} \gamma_t &= \sum_{j=1}^h \gamma_{j,t} \\ \gamma_{j,t+1} &= \gamma_{j,t} \cos(\lambda_j) + \gamma_{j,t}^* \sin(\lambda_j) + \omega_{j,t} \\ \gamma_{j,t+1}^* &= -\gamma_{j,t}^{(1)} \sin(\lambda_j) + \gamma_{j,t}^* \cos(\lambda_j) + \omega_{j,t}^*, \\ \omega_{j,t}^*, \omega_{j,t} &\sim N(0, \sigma_{\omega^2}) \\ \lambda_j &= \frac{2\pi j}{s} \end{aligned} \quad (5.11)$$

The linear dependence on the covariates  $\beta X_t$  in eq. (5.10) further helps to explain observed data. The better this component contributes to the prediction task, the lower the local level component  $\mu_t$  should be. Finally,  $\varepsilon_t \sim \mathcal{N}(0, \sigma_t^2)$  models the noise associated with measuring  $y_t$ .

In order to estimate the impact of Governmental interventions, we follow the following methodology. During the training period, we sample the model parameters and the state vector using the Gibbs sampler – a MCMC framework – against the observed data. For the estimation of the impact on the mobility metric, the training period extends over the first 10 days in our selected dataset, i.e, from 10 days prior to the start of Governmental intervention to the start of the intervention. For the estimation of the impact on the effective reproduction number, the training period extends over the first 20 days in our selected dataset, i.e, from 10 days prior and 10 days following the start of Governmental intervention. We select this span of time since, for almost 50% of the people, it takes at least 9.92 days following the date of exposure to get a clinical confirmation of the viral infection (see fig. 5.10 for the Hessen state in Germany). Hence, on average, there is a delay of around 10 days between the date of exposure and the date of confirmation, and there will not be any immediate effect on the the daily confirmed cases as a result of the intervention. Thus, we define the estimated *effective intervention date* as being equal to the exact intervention date plus  $9.92 \approx 10$  days. We then use the posterior simulations to simulate from the posterior predictive distribution over the counterfactual time series given the observed pre-intervention activity during the training. We use the following 30 days for mobility metric and following 20 days for the effective reproduction number for this purpose.



Finally, we use the posterior predictive samples to compute the posterior distribution of the point wise impact, i.e.  $\phi_t = y_t - \tilde{y}_t$ , where  $y_t$  is the observed quantity and  $\tilde{y}_t$  is the counterfactual prediction assuming no intervention. We define the absolute impact of the intervention as the expected value of the point wise impact i.e.  $I = \langle \phi_t \rangle_t$ . We use the 30 days and 20 days of posterior predictive samples for mobility metric and  $R_t$  to estimate the point wise impact as well as absolute impact because of the intervention. We also estimate the p-value of the observe absolute impact, which measures the probability of obtaining the impact by chance under the null of no intervention.

### 5.3 Results

We use a probabilistic contagion model with inhomogeneous source terms to explain the temporal evolution of epidemic because of COVID-19 [161]. Using the daily number of confirmed cases and the generation time model (see figure (D.1)), we estimate the time evolution of effective reproduction number  $R_t$  – which is defined as the actual average number of secondary cases per primary case – with a Sequential Bayesian estimation. As a typical result of our analysis, figure 5.1 shows the time evolution of the effective reproduction number in the *Hessen state* in *Germany*.

With the help of a diffusion-regression state-space model [162] and MCMC posterior inference, we estimate the counterfactual evolution – that would have occurred had no intervention taken place – of post intervention effective reproduction number. The comparison between the effective reproduction number and the predicted  $R_t$ , had no intervention taken place, allows us to quantify the effectiveness of intervention for the *Hessen state* in *Germany* shown in fig. 5.1. We find an average reduction of  $R_t$  of 0.87, and reject the null hypothesis with a p-value of 0.04 that this reduction could result from chance under the counterfactual evolution without lockdown.

Figure 5.2 shows the map of the absolute impact on  $R_t$  as a result of intervention for different states in Germany. The color for each state represents the average impact of the intervention on  $R_t$ , i.e., the magnitude increment or decrement of  $R_t$  from the counterfactual predicted value without lockdown over the thirty days following lockdown. The radial wedges represent the temporal evolution of  $R_t$  in the corresponding state and color of the strips represent  $R_t$  on a particular day over the thirty days following lockdown. This figure illustrates the significant heterogeneity in the results, as well as an important non-monotonicity in the dynamics of  $R_t$ .

We then analyse the impact of Governmental interventions on human mobility, illustrating the results for Germany. We are able to breakdown the impact of intervention in different

mobility dimensions and quantify the level of reduction of mobility (see fig. 5.3 for details). Figure 5.4 shows the absolute impact on workplace activity (in % change from baseline) in different states in Germany, resulting from the intervention. The radial wedges represent the temporal evolution of the workplace activity in the corresponding state, and color of the strips represent activity on a particular day. Contrary to the map of the absolute impact on  $R_t$  (figure 5.2), the absolute impact of travel restrictions on mobility is much more homogenous across German states, and also consistent along the time axis.

We then evaluate the impact of Governmental interventions on  $R_t$  as well as on works place activity for the top level administrative divisions in a number of countries. Figure 5.5 and D.2 illustrate the impact of Governmental interventions on workplace activity in a number of countries and figure 5.6 and D.3 show the impact on  $R_t$ . Each subplot of figure 5.5 shows the time evolution of mobility and the impact of travel restrictions on mobility of different top level administrative divisions of a country. Each ring map represents a country and provides a visual representation of the time evolution of work place activity as well as the impact of Governmental interventions on the workplace activity. Each subplot of figure 5.6 shows the time evolution of the effective reproduction number  $R_t$  and the absolute impact of travel restrictions on  $R_t$  in ten countries. Each ring map represents a country and shows the time evolution of the effective reproduction number  $R_t$  as well as the impact of Governmental interventions on  $R_t$ . Figure 5.5 and D.2 show a rather homogeneous impact on workplace activities across the different regions of each of the nine analysed countries. However, there is a large heterogeneity across different countries, e.g. Spain being the most effected country while Japan is the least affected country in terms of workplace activity. In contrast, the impact on  $R_t$  is quite heterogeneous across different regions within a country, notwithstanding similar levels of restriction, as illustrated in figure 5.6 by Kerala (KL) and Maharastra (MH) in India. Surprisingly, across a number of regions, a significant increment of  $R_t$  is observed following the implementation of very strict lockdowns (e.g. Maharastra (MH) in India and Luzern (LU) in Switzerland.).

Figure 5.7 and D.4 show the joint distribution (obtained by kernel density estimation) of the absolute impact resulting from intervention on workplace activity and on the  $R_t$  in the administrative divisions of nine different countries. In other words, figure 5.7 compares the strictness of the governmental interventions, measured in terms of the impact on the workplace activity, against the corresponding reduction/increment in effective reproduction number. There is no significant correlation between these two variables, suggesting that other variables are controlling the reduction of reproduction rates.

Figure 5.8 and D.5 represent the spatial correlation analysis of the total increase  $\Delta S$  in the number of confirmed cases during the first 30 days of interventions. We define *Moran's I*,

which is a measure of spatial auto-correlation, as follows:

$$I = \frac{N \sum_j \sum_j w_{ij} (\Delta S_i - \Delta \bar{S}) (\Delta S_j - \Delta \bar{S})}{W \sum_i (\Delta S_i - \Delta \bar{S})^2} \quad (5.12)$$

where  $N$  is the number of spatial units indexed by  $i$  and  $j$  in a given country;  $\Delta S$  is the increase in the number of confirmed cases;  $\Delta \bar{S}$  is the mean of  $\Delta S$  over all the spatial units of the country;  $w_{ij}$  is a matrix of spatial weights with zeroes on the diagonal (i.e.,  $w_{ii} = 0$ ); and  $W$  is the sum of all  $w_{ij}$ . We define  $W$  by giving a weight  $1$  if two regions are neighbors, and  $0$  otherwise. The “spatial lag” of  $\Delta S$  for a region is defined as the weighted sum of its neighbors’  $\Delta S$ . The scatter plot in each panel of figure 5.8 shows the spatial lag as a function of its corresponding Delta S for different regions in each country. The inset gives the kernel density estimation of the simulated Moran’s  $I$  from the null model of no spatial correlation. The slope of the scatter plot of  $\Delta S$  against the spatial lag is known to converge to the *Moran’s I* [163]. We test the significance of the *Moran’s I* under the null hypothesis of no spatial auto-correlation and simulate  $1000$  realizations, by randomly shuffling the locations of the  $\Delta S$ . The spatial correlation analysis presented in figure 5.8 reveals significant spatial correlation of  $\Delta S$  in Italy, Switzerland, Japan and United States.

In order to understand the effectiveness of lockdowns, we compare the number of confirmed cases against the impact on  $R_t$  during the first 30 days of interventions. Figure 5.9 and D.6 show the kernel estimation of the bivariate distribution of the total increase in number of cases and of the total impact on  $R_t$  during the first 30 days of intervention in different regions of nine countries. The inset in each panel represents the impact on  $R_t$  during this period, (or  $\Delta R_t$ ), against the average  $R_t$  before the intervention (or  $\langle R_t \rangle_{init}$ ) over the regions in each country. The kernel density estimation of the bivariate distribution in figure 5.9 reveals a negative correlation between the above variables, showing a slowdown of the epidemic growth notwithstanding the increase of  $R_t$ . We also compare the average  $R_t$  before the intervention (or  $\langle R_t \rangle_{init}$ ) and impact on  $R_t$  (or  $\Delta R_t$ ) from the intervention to understand the effectiveness of intervention in reducing/increasing  $R_t$  from its initial values. The inset panels of figure 5.9 unveil a negative correlation between these two variables, indicating a significantly large decrease of  $R_t$  in the regions of larger initial  $R_t$ . Surprisingly, we also see the extension of the distribution to the second quadrant, revealing the fact that, in many places, the epidemic started after the lockdown.

The tables D.1 to D.10 in the supplementary materials provide the detailed values of  $R_t$  with positive and negative impact along with the corresponding growth in total number of confirmed cases as a result of intervention in the different regions for each country. The regions that are characterised by an increase of  $R_t$  after lockdown are indicated in bold face.

## 5.4 Discussion

We have found important to quantify the incubation and confirmation periods in order to assess the impact of policy interventions to contain the epidemics in different regions. This quantification shown in figures 5.10 and D.1 allows us to define a credible time interval over which to quantify the impact of interventions. With this, our analysis provides a framework to understand the effectiveness of the policies and interventions implemented by various countries to control the epidemic growth. The wealth of results presented in figures 5.1-5.9 leads to the following insights.

Overall, it is clear, and unsurprising, that interventions in the form of mobility restrictions and social distancing are found to have significant impacts in controlling the development of the epidemics in many countries. In particular, we have quantified the decrease in reproduction number  $R_t$  resulting from the intervention. As an illustration shown in fig. 5.1, for the Hessen state in Germany, we are able to quantify that intervention in this state reduced  $R_t$  by about 1 unit compared with the counterfactual scenario of no intervention.

More surprising is the large heterogeneity in the reduction of  $R_t$  across different states in Germany, as shown in figure 5.2. For this country, intervention measures had led systematically to a decrease in  $R_t$ , with quite strong differences from state to state. Even more surprising is the time dependence which exhibits a non monotonous and fluctuating behaviour of  $R_t$  in different regions.

For most of regions in different countries, because of the interventions,  $R_t$  decreased, however there are some places where it increased. Because this increase is transient and constrained by the lockdown, it does not lead to a very strong explosion of new cases. A tentative interpretation is that, as lockdown was considered and being implemented, in a number of regions, it triggered large movements of people to relocate, hence promoting new contagions and epicenters for the epidemics to mature after the lockdown. An additional mechanism, which underlies the Japanese policy, is the effect of closed spaces and close-contact settings within confined households, which has been shown to lead to increased infections within households for instance. But because the lockdown only allows the new nuclei of contagion to develop locally, the number of cases did not explode. Examples of this effect can be found in Luzern and Solothurn of Switzerland, in Bremen of Germany, in Saga Ken of Japan and in Odisha of India. This effect that has previously been described qualitatively is given here a quantitative support by our systematic analysis.

The observation that lockdown led first to an increase in  $R_t$  in a significant number of regions and countries is confirmed by the correlation analysis presented in fig. 5.9 relating increase of cases to changes of  $R_t$  in different countries. There are many regions in each country where there was no epidemics before the lockdown. The lockdown triggered a pre-

emptive movement of people to relocate, increasing  $R_t$  after the lockdown. But the increase of  $R_t$  did not increase  $\Delta S$  too much due to the effect of confinement. The numerous local infectious cases could only infect their immediate relatives. In regions where the epidemic was at more advanced stages, the lockdown had the effect of decreasing  $R_t$ , as expected. The insets in fig. 5.9 provide further support for this conclusion. The average correlations shown by the plots indicate that, where there was no epidemic, the epidemic started after the lockdown, and where there was an on-going epidemic, it came under control.

We also quantify that there is not much advantage resulting from strict lockdowns. For example, Japan and Switzerland did rather well in spite of weaker lockdowns, whereas Italy and Spain performed much more poorly with stricter confinements. This is related to the effectiveness of early stage contact tracing as well as effective awareness of the role of protection measures. Figure 5.7 exemplifies this point by showing the paradoxical results that lockdown led to an obvious strong collapse of mobility accompanied by an increase of  $R_t$  in several regions in France or Italy, for instance. For the other countries, the effect of lockdown is more as expected.

Moreover, the timing of lockdown is very important in determining the trajectory of the epidemic. Figure 5.8 shows the spatial correlation of the total growth of epidemic after the lockdown. The larger is the spatial diffusion (possible during the advanced stage of the epidemic), the larger is the spatial correlation. Our analysis shows that, for Italy, Switzerland, the United States, and Japan, the spatial correlation of  $\Delta S$  is significantly positive. This means that, in these countries, the epidemic started much before the lockdown date and was developing silently as revealed by the strong spatial diffusion of the epidemic. While Switzerland and Japan contained the epidemic with effective containment policy, Italy and United States failed to do so because the intervention was ill-adapted to the spatial developments.

When comparing the German, Swiss and US lockdowns via their mobility data, we find very similar severity levels of the confinements. However, the effectiveness of the lockdown to control the epidemic in the USA is quite low, while very significant in Germany[164] and Switzerland. While Switzerland and USA both imposed the lockdown at a rather late stage of their unfolding epidemic, the Swiss containment and awareness policy was significantly superior to that in the USA. We quantify that the epidemics has diffused to many states in the USA, as revealed by the spatial correlation in figure 5.8, even after the lockdown was implemented. This explains the failure to reduce the transmission to a large degree. For Italy and Spain, because there was a significant numbers of confirmed cases across different regions of Italy and Spain before the lockdown, it is difficult to determine the effectiveness of cross-region transmission control.

In most of the Indian states, we do not observe any significant impact of the lockdown on reducing the effective reproduction number. A possible explanation is that, in most of the places, there is no real epidemic and the majority of the infection cases are found in a small number of states. The complete lockdown of the entire country might have been ignorant to this very strong heterogeneity.

We also unearth some outliers, e.g, Maharastra state in India, Saarland in Germany where, despite a strict lockdown (quantified in our analysis by a strong reduction of workplace activity), both the number of cases and  $R_t$  exploded. This poses the questions of unobserved contagion paths, likely associated with specific events, perhaps the existence of super-spreaders and so on.

Our analysis overall suggests that the interventions may have not been optimal in many countries and that there are probably better alternatives to complete lockdowns. One alternative is a sequential and selective lockdown approach, putting in selective quarantine based on a threshold value for the number of confirmed cases while leaving the other places more open with social distancing but not complete lockdown. One should however stress that this alternative intervention requires very strong testing support in order to determine with sufficient reliability the positive cases. The case of India supports further the idea that the policy was too early in implementing a complete lockdown of the country for such a long time but too late implementing effective quarantine of people coming from effected places. The cases of Italy, France and other regions where confinements led to a transient increase of  $R_t$  over the following 30 days also underlies the plausible importance of controlling close contacts and confined places: with the aim of doing good, confining might have worsen the transmission of the disease in a number of cases.

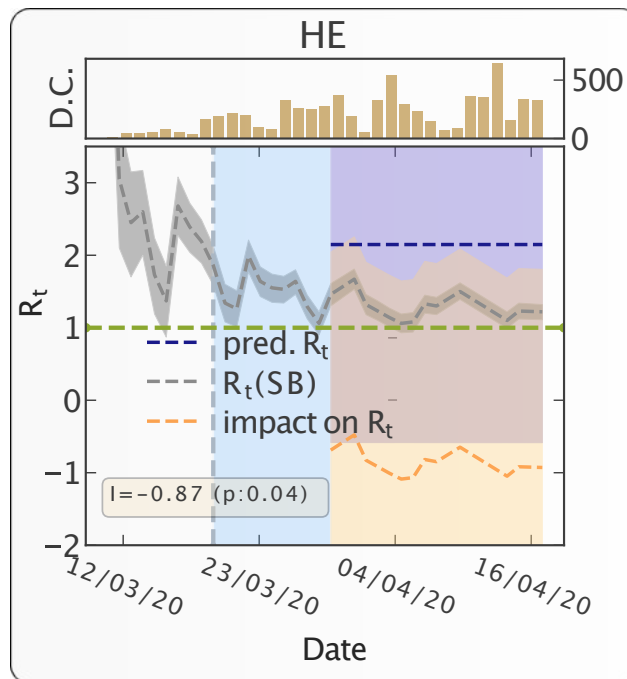


Figure 5.1 **Progression of epidemic and impact of travel restrictions on epidemic growth in the Hessen state in Germany.** The grey dashed line (with 95% CI band) represents the most likely estimation of  $R_t$ . The dotted vertical line represents the start date of intervention. The region between intervention date and effective intervention date – 10 days following the intervention date – is marked with light blue color. It indicates the period during which people exposed to the virus prior to the intervention date would keep on appearing as the new confirmed cases. The counterfactual predicted  $R_t$  is presented by the dashed blue line (with 95% CI band). The pointwise impact of the intervention is presented as the dashed orange line (with 95% CI band). The horizontal green line represents the critical value  $R_t = 1$ . The inset figure shows the time evolution of number of daily confirmed cases (D.C.). We note down the absolute impact – average value of point wise impact – of the intervention along with the p-value in the yellow box. The p-value measures the probability of observing the impact by random chance.

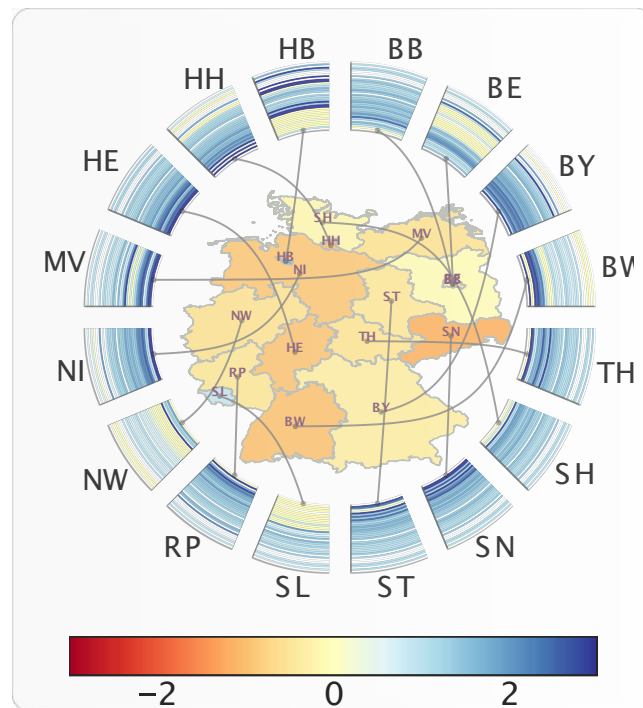


Figure 5.2 Time evolution of  $R_t$  and absolute impact of travel restrictions on  $R_t$  for different states in Germany. The color for each states (name in ISO 3166-2 code notation) represents the absolute impact (increase or decrease of  $R_t$ ) due to travel restriction in that state. The radial wedges represent the time evolution of  $R_t$  in the corresponding state and color of the strips represent  $R_t$  on a particular day.



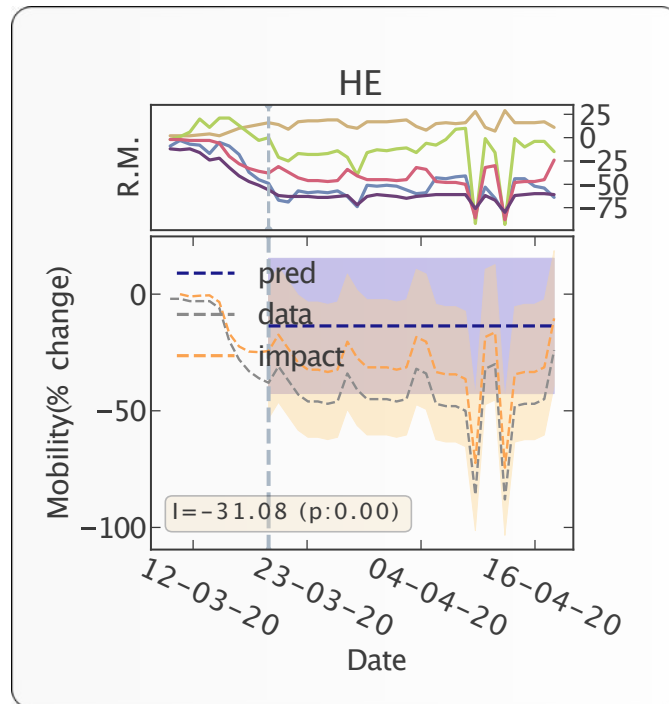


Figure 5.3 **Countrywide time evolution of human mobility and impact of travel restrictions on human mobility for Germany.** The top panel represents the time evolution of human mobility (% change from the baseline activity) in different mobility dimensions (yellow: home, red: work, blue: retail, green: grocery, violet: transit). Transit is a proxy for long-distance travel (it corresponds to petrol pumps/filling stations etc.). In the bottom panel, The y-axis represents % increase or % decrease of the average individual's activity. the grey dashed line represents the % change of activity from the baseline (baseline is set to 0) in the workplace resulting from the intervention. The dotted vertical line represents the intervention date. The counterfactual predicted evolution of workplace activity, had no intervention taken place, is presented by the dashed blue horizontal line. The point-wise impact of the intervention on mobility is presented by the orange dashed line. The yellow box indicates the absolute impact of the intervention along with its p-value.

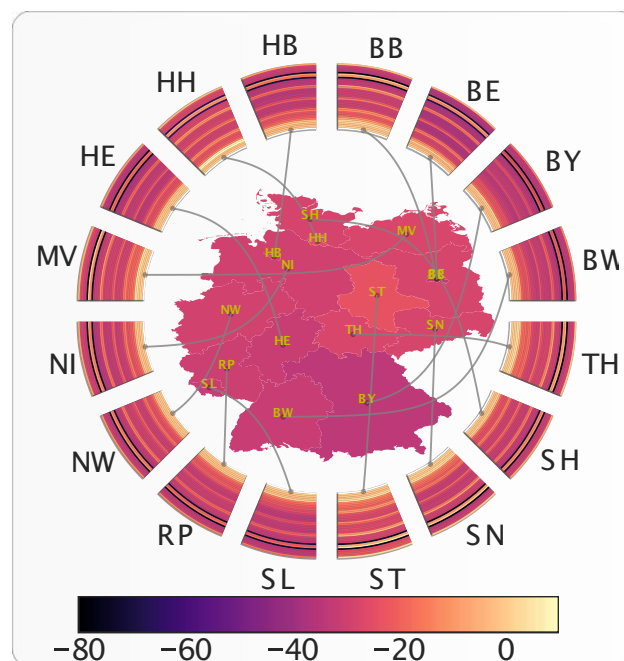


Figure 5.4 **State-by-state time evolution of mobility and absolute impact of travel restrictions on mobility.** The color of each state (name in ISO 3166-2 code notation) represents the absolute impact (increase or decrease of mobility) of interventions on workplace activity in that state. The radial wedges represent the time evolution of mobility in the corresponding state and color of the strips represent the mobility on a particular day. SL is Saarland, ST is Sachsen-Anhalt.

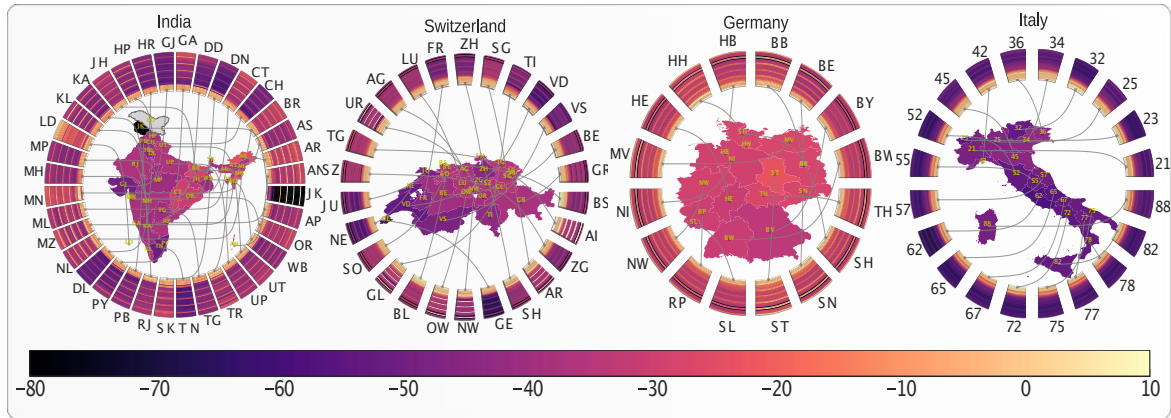


Figure 5.5 **Time evolution of mobility and impact of travel restriction on mobility in eight countries.** Each subplot shows the time evolution of mobility and impact of travel restrictions on mobility of different top level administrative divisions of a country. The countries are, for left to right: India, Switzerland, Germany, Italy. The color of the regions on the map denoted by their ISO 3166-2 code represent the impact (increase or decrease of mobility) of travel restriction on that region. The radial connected wedges represent the time evolution of the mobility in the corresponding region or state for each country. The color of the strips in the wedges represent the mobility on a particular day.

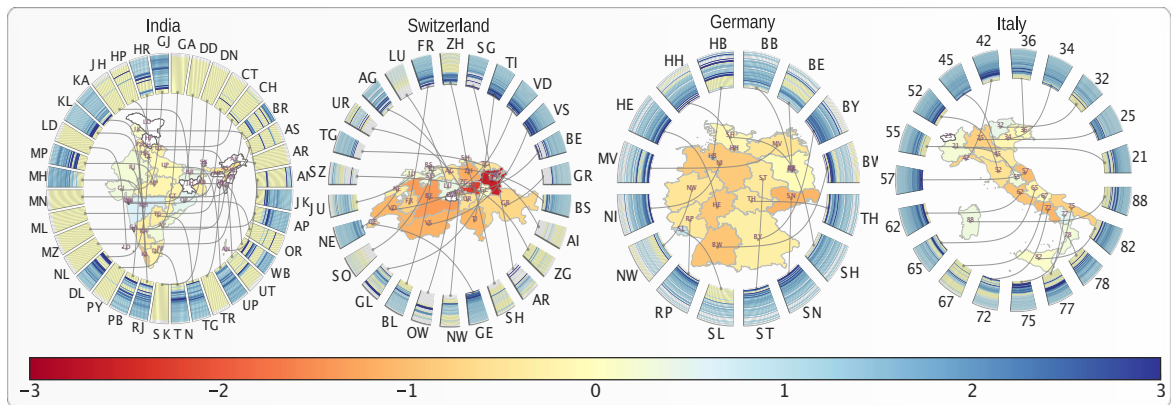


Figure 5.6 **Time evolution of effective reproduction number ( $R_t$ ) and absolute impact of travel restrictions on  $R_t$  in ten countries.** Each subplot presents time evolution of  $R_t$  and impact of interventions on  $R_t$  for different top level administrative divisions of the country. The color of the regions (name in ISO 3166-2 code notation) represents the impact (increase or decrease of  $R_t$ ) of interventions on that state. The radial wedges represent the time evolution of  $R_t$  in the corresponding state, and color of the strips represent  $R_t$  on a particular day. The countries are, from top left to bottom right: India, Switzerland, Germany, and Italy.

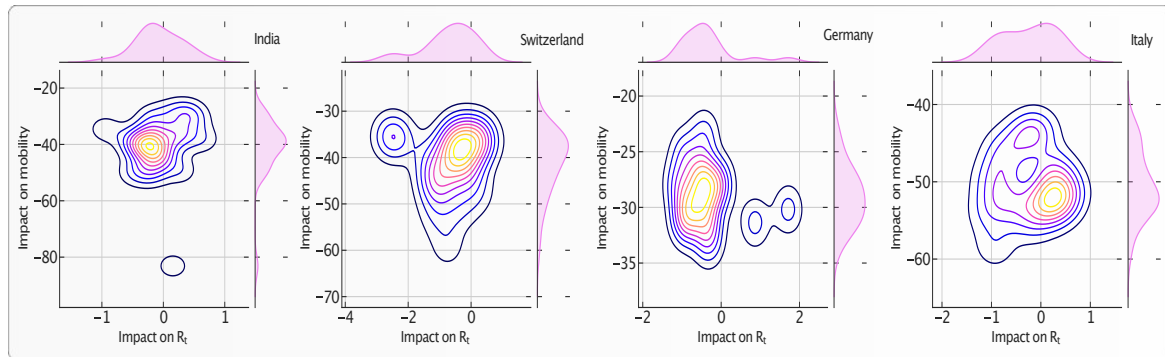


Figure 5.7 **Impact of travel restrictions on workplace activity and on epidemic progression ( $R_t$ ) in nine countries.** Each panel represents the bivariate kernel density estimation as a function of the absolute impact on workplace activity and impact on  $R_t$  in the administrative divisions of each country. The bivariate distribution is constructed over the set of regions within each country. The top and right inset of each of the nine plots represent the marginal distribution of the respective variables for each country. The countries are, from top left to bottom right: India, Switzerland, Germany, and Italy.

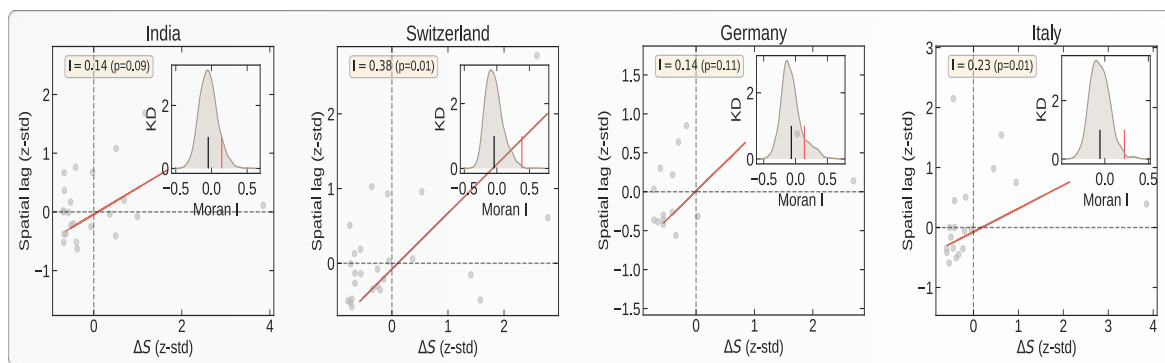


Figure 5.8 **Spatial auto-correlation of the total increase  $\Delta S$  in number of confirmed cases during the first 30 days of intervention in nine countries.** In each panel, the x-axis corresponds to the value of  $\Delta S$  in a given region in a given country; the y-axis gives the average  $\Delta S$  over the neighboring regions, called “spatial lag” in the caption along the y-axis. These two variables are *z-standardised* for better comparison. The inset in each panel represents the Kernel Density estimator for the distribution of the simulated *Moran's I*. The black vertical line in the inset represents the expected *Moran's I* from simulations with the null hypothesis of no spatial correlations. The red vertical line represents the value obtained from empirical data. *Moran's I* along with its p-value is given in yellow box. The countries are, from top left to bottom right: India, Switzerland, Germany, and Italy.

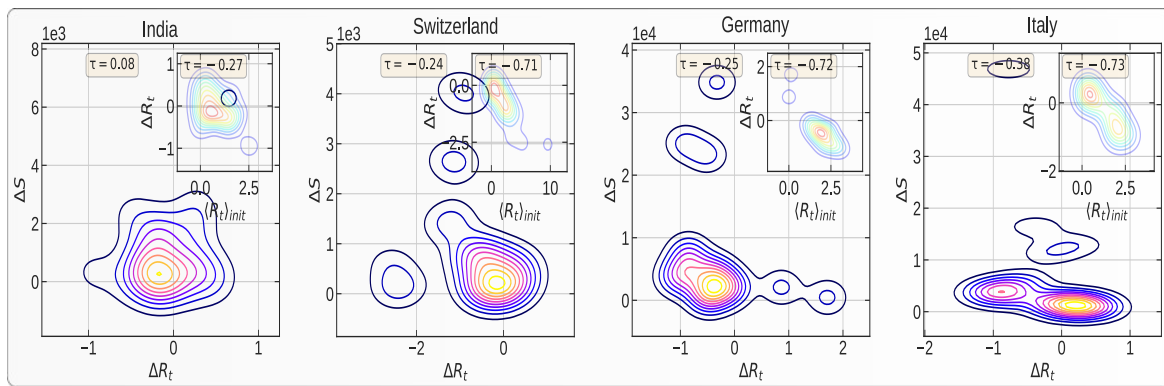
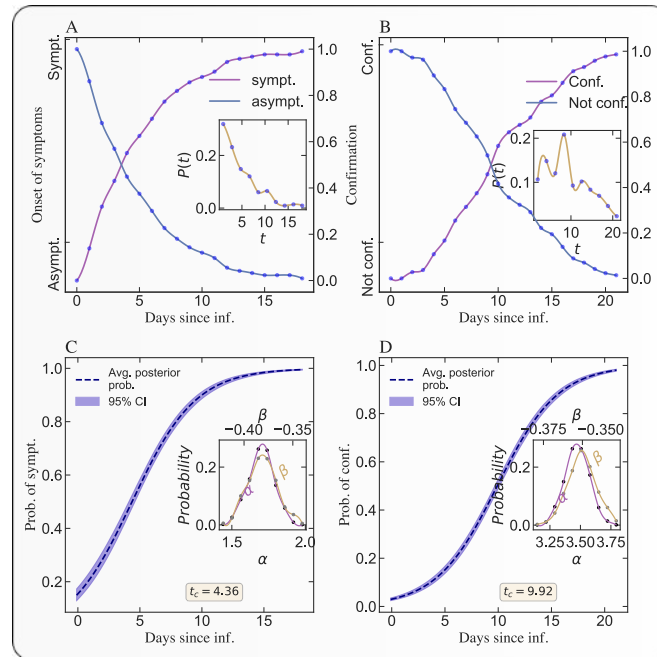


Figure 5.9 **Joint distribution of the total number  $\Delta S$  of confirmed cases within the first 30 days of intervention and of the absolute impact of intervention on the effective reproduction number ( $\Delta R_t$ ) within this period.** Each panel represents the Kernel Density Estimation for the total number of confirmed cases within the first 30 days of the intervention against the impact of interventions on effective reproduction number ( $\Delta R_t$ ) in different administrative divisions of a country. The bivariate distribution is constructed over the set of regions within each country. The yellow box contains the Kendall  $\tau$  correlation value for this joint distribution. The inset in each panel represents the variation of  $\Delta R_t$  during the intervention against the average initial  $R_t$  before the intervention. The yellow box gives the corresponding Kendalls  $\tau$ 's. The countries are, from top left to bottom right: India, Switzerland, Germany, and Italy.



**Figure 5.10 Bayesian inference of the incubation period and confirmation period.** A) Empirical cumulative probability of developing a symptom on a particular day following the day of exposure to the virus. We use individual level clinical data [152, 153] to conduct these Bayesian estimations. The lighter color of the square represents low probability and the deep color represents high probability. The solid lines represent the fraction of people remaining asymptomatic (decreasing curve) or becoming symptomatic (increasing curve) on a given day, following exposure. The inset figure represents the empirical probability distribution of duration between the date of exposure and the date of onset of symptom or the clinical confirmation. B) Same as A) but for the probability of being confirmed. C) Estimation of the posterior probability distribution (with 95% confidence interval) of the incubation period (i.e., of getting the clinical confirmation on  $day = t$ , provided the individual is exposed to the virus on  $day = 0$ ). The dashed solid line represents the most likely posterior probability estimation and the light band represent the 95% confidence interval. The inset figure presents the distributions of the two estimated parameters in the logistic model (5.1). D) Same as C) for the probability distribution of confirmation period. The median value (and also mode) of the incubation period is 4.38 days. The median value (and also mode)  $t_c$  of the confirmation period is 9.94 days.

# Chapter 6

## Conclusion

In this thesis, we studied the impact of endogenous and exogenous shocks on epidemic processes in out-of-equilibrium self-organizing systems. We addressed the questions using massive datasets from social, biological, and physical systems and quantified the endogenous and exogenous origin of contagion. We further extended our discussion on how to build robust data aggregation systems to better understand the out of equilibrium systems.

In second chapter, we studied the herding in scholarly productivity and success. We showed that the productivity and success trajectories are symmetric before and after the peak and they could be represented by a power-law with exponent  $1 - 2\theta$ , where  $1 - 2\theta$  takes approximately the same value before and after the peak. We found  $1 - 2\theta = 0.7$  for the scholarly productivity and 0.55 for the YouTube success. These results translated to an endogenously fueled epidemic branching process within each career, with a slowly decaying influence kernel ( $\phi(t) = \frac{1}{t-t_i^{1+\theta}}$ ), with  $\theta = 0.15 - 0.25$ . This value of  $\theta$  is smaller than those  $\theta = 0.3 - 0.4$ , reported in previous studies. Interestingly, we observed the peak of productivity around 1991 in most scholarly careers concerned with signal processing and the peak of success around Jan 2018 in most of the YouTube channels. We could trace the synchronized herding in scholarly productivity back to a wave of interest in wavelets during the 1990s (“Wavelet revolution” during the 1990s). The herding in YouTube success occurred one month after the peak in December 2017 of the cryptocurrency bubble (2017 Cryptocurrency mania). Thus, we demonstrated endogenously fueled herding effects in each case, where the scholars and the YouTubers, and the respective communities were most active during the two times of focus making the growth of productivity and success within the respective fields evolve endogenously in synchrony.

In chapter 3, we evaluated the predictability and persistence of individual and collective team performance in a team game. We introduced a number of new statistical methods for studying the hot hand effect in game of cricket. In comparison to established methods such

as ARIMA, we quantified and used self-excited patterns in individual and team performance to better predict the future. Our findings supported the existence of considerable hot-hands in individual performance. Furthermore, we found a considerable decline in team performance prediction, indicating the dominance of systemic stochasticity. While this was still somewhat predictable, the outcome of the game could not be predicted, nor did they cluster in time. This led to an intriguing conclusion: 'Cricket is a game of skill for individuals and luck for teams.' While an individual can consistently deal with ambient systemic stochasticity, it is difficult for a team to function comparably effectively. As a result, these findings paved the way for further research into the effect of group size on predictability and consistency of performance. Furthermore, the current study developed a quantitative framework for detecting and predicting individual career achievement. This technique would help us better understand the predictability of success in individual careers.

In chapter 4, we showed that the spatial variation of background seismicity rate is a key parameter in the determination of the criticality of the fault network. We showed when this spatial variation is properly accounted for, we get a superior forecasting model and a branching ratio that is much smaller than 1. This suggests that fault networks mainly evolve far from a critical point. This has major implications for the prediction of large events. Indeed, in the self-organized critical scenario, each event is no different from all others from a generating process viewpoint, making prediction impossible. In contrast, if the fault network remains most of the time far from criticality, more sporadic singularities may appear via various possible mechanisms and announce upcoming catastrophic events. This suggests, for instance, the need for a reevaluation of the Accelerated Moment Release hypothesis, benefitting from the prior use of the ETAS model to better eliminate the contribution of the uncorrelated part of seismicity to the total seismicity rate.

Chapter 5 presented a framework to understand and quantify the effectiveness of various lockdown and distantiating measures in controlling the spread of the SARS-CoV-2 epidemic. We showed that most regions obtained positive results in reducing the reproduction number after implementing lockdown. However, one of the surprising results of our study was that there was a transient increase in the contagion after the lockdown in some regions. We hypothesize that this resulted in part from large geographic movements and relocation of people performed in reaction to the announced lockdown measures, hence creating new contagion epicenters for the epidemics to mature after the lockdown. However, other factors, such as the dynamics of hospital/ICU admissions, would need to be considered for a better and more conclusive interpretation. The methodology used in our study is based on the probabilistic contagion model developed by Bettencourt and Ribeiro. There exist different methods to estimate the reproduction number  $R_t$  and they differ in their accuracy with pros



and cons that go beyond the scope of our study. A future direction for further improvement would be to perform systematic comparisons of how different estimation methods for  $R_t$  would impact our main conclusions. More important would be to improve the calibration of  $R_t$ , both in consistency, accuracy, and time resolution. Further, our results were based on estimating counterfactual evolutions of the effective reproduction number. Thus the presented results are conditional to the limitations of the above method. Another limitation of our study is that, for the sake of simplicity, we only considered lockdown and distantiation measures at national levels, thus neglecting regional heterogeneities. While most local lockdown measures were close to their national-level policies, more accurate dates for the regional lockdown measures would improve the results of this study.

A general framework for predicting how out-of-equilibrium systems will respond to endogenous and exogenous shocks is still an open problem. Through this thesis, we developed number of novel approaches and tried to characterize the properties of these systems, by studying the corresponding responses against both kinds of shocks. We applied different variations of self-excited conditional point process models to study the epidemic responses exhibited by these systems in spatial, temporal, and magnitude dimensions. Our studies led to many interesting findings that develop insights into better understanding about the underlying systems. Our findings show that the social network, that we are part of, acts like a critical system that further drives individuals' productivity and success. In team games as a out-of-equilibrium system, our study revealed while an individual can consistently deal with the systemic stochasticity of the environment, it is difficult for the team as a whole to perform equally predictably. In contrary to traditional belief behind the generating processes of earthquakes, our study suggests the fault network remains most of the time far from criticality. The epidemic due to SARS-CoV-2 virus is an example of epidemic processes in a out-of-equilibrium system. The development of infected cases, of the various forms of illnesses resulting from infection and of the death rates reflect the interplay between the biological and epidemiological properties of the virus. Our study revealed the importance of early intervention measures in preventing the outbreak in this complex interconnected system.

In the future, we plan to study how contagion phenomena evolve in scholarly networks. We are interested in investigating how the citation of articles evolves in a self-excited manner, and in exploring the different driving forces that influence this evolution. Some possible factors we might look at include the emergence of new research topics on scholarly networks. Additionally, we plan to investigate the impact of information flowing through social networks on financial systems, as well as any feedback effects that may be present.



# Bibliography

- [1] D. Sornette, “Endogenous versus exogenous origins of crises,” in *Extreme events in nature and society*, pp. 95–119, Springer, 2006.
- [2] D. Sornette, *Critical phenomena in natural sciences: chaos, fractals, selforganization and disorder: concepts and tools*. Springer Science & Business Media, 2006.
- [3] D. Sornette, *Why stock markets crash: critical events in complex financial systems*, vol. 49. Princeton University Press, 2017.
- [4] D. Sornette, “Critical market crashes,” *Physics Reports*, vol. 378, no. 1, pp. 1–98, 2003.
- [5] D. Sornette, “Predictability of catastrophic events: Material rupture, earthquakes, turbulence, financial crashes, and human birth,” *Proceedings of the National Academy of Sciences*, vol. 99, no. suppl 1, pp. 2522–2529, 2002.
- [6] D. Sornette, F. Deschâtres, T. Gilbert, and Y. Ageon, “Endogenous versus exogenous shocks in complex networks: An empirical test using book sale rankings,” *Physical Review Letters*, vol. 93, p. 228701, nov 2004.
- [7] D. Sornette, “Discrete-scale invariance and complex dimensions,” *Physics Reports*, vol. 297, no. 5, pp. 239–270, 1998.
- [8] R. Crane and D. Sornette, “Robust dynamic classes revealed by measuring the response function of a social system,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, pp. 15649–15653, oct 2008.
- [9] D. Sornette and A. Helmstetter, “Endogenous versus exogenous shocks in systems with memory,” *Physica A: Statistical Mechanics and its Applications*, vol. 318, pp. 577–591, feb 2003.
- [10] C. G. Sammis and D. Sornette, “Positive feedback, memory, and the predictability of earthquakes,” *Proceedings of the National Academy of Sciences*, vol. 99, no. suppl 1, pp. 2501–2508, 2002.
- [11] V. Filimonov and D. Sornette, “Quantifying reflexivity in financial markets: Toward a prediction of flash crashes,” *Physical Review E*, vol. 85, no. 5, p. 056108, 2012.
- [12] S. Nandan, G. Ouillon, D. Sornette, and S. Wiemer, “Forecasting the rates of future aftershocks of all generations is essential to develop better earthquake forecast models,” *Journal of Geophysical Research: Solid Earth*, vol. 124, no. 8, pp. 8404–8425, 2019.

- [13] N. T. Bailey *et al.*, *The mathematical theory of infectious diseases and its applications*. Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE., 1975.
- [14] P. S. Dodds and D. J. Watts, “Universal behavior in a generalized model of contagion,” *Physical review letters*, vol. 92, no. 21, p. 218701, 2004.
- [15] J. K. Shultis and R. E. Faw, *Fundamentals of Nuclear Science and Engineering Third Edition*. CRC press, 2016.
- [16] D. J. Watts and P. S. Dodds, “Influentials, networks, and public opinion formation,” *Journal of consumer research*, vol. 34, no. 4, pp. 441–458, 2007.
- [17] D. Centola, J. Becker, D. Brackbill, and A. Baronchelli, “Experimental evidence for tipping points in social convention,” *Science*, vol. 360, no. 6393, pp. 1116–1119, 2018.
- [18] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, *et al.*, “The science of fake news,” *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.
- [19] A. Madan, M. Cebrian, D. Lazer, and A. Pentland, “Social sensing for epidemiological behavior change,” in *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pp. 291–300, 2010.
- [20] F. Allen and D. Gale, “Financial contagion,” *Journal of political economy*, vol. 108, no. 1, pp. 1–33, 2000.
- [21] M. Billio, M. Getmansky, A. W. Lo, and L. Pelizzon, “Econometric measures of connectedness and systemic risk in the finance and insurance sectors,” *Journal of financial economics*, vol. 104, no. 3, pp. 535–559, 2012.
- [22] D. Acemoglu, V. M. Carvalho, A. Ozdaglar, and A. Tahbaz-Salehi, “The network origins of aggregate fluctuations,” *Econometrica*, vol. 80, no. 5, pp. 1977–2016, 2012.
- [23] S. Nandan, G. Ouillon, D. Sornette, and S. Wiemer, “Forecasting the rates of future aftershocks of all generations is essential to develop better earthquake forecast models,” *Journal of Geophysical Research: Solid Earth*, vol. 124, no. 8, pp. 8404–8425, 2019.
- [24] T. S. Robertson, “The process of innovation and the diffusion of innovation,” *Journal of marketing*, vol. 31, no. 1, pp. 14–19, 1967.
- [25] J. Kaminski, “Diffusion of innovation theory,” *Canadian Journal of Nursing Informatics*, vol. 6, no. 2, pp. 1–6, 2011.
- [26] P. M. Hirsch, “Processing fads and fashions: An organization-set analysis of cultural industry systems,” *American journal of sociology*, vol. 77, no. 4, pp. 639–659, 1972.
- [27] S. Bikhchandani, D. Hirshleifer, and I. Welch, “A theory of fads, fashion, custom, and cultural change as informational cascades,” *Journal of political Economy*, vol. 100, no. 5, pp. 992–1026, 1992.
- [28] S. Blackmore and S. J. Blackmore, *The meme machine*, vol. 25. Oxford Paperbacks, 2000.

- [29] T. Besley and T. Persson, “The logic of political violence,” *The quarterly journal of economics*, vol. 126, no. 3, pp. 1411–1445, 2011.
- [30] M. Pelling and K. Dill, “Disaster politics: tipping points for change in the adaptation of sociopolitical regimes,” *Progress in human geography*, vol. 34, no. 1, pp. 21–37, 2010.
- [31] A. Wehrli, D. Sornette, *et al.*, “Classification of flash crashes using the Hawkes (p, q) framework,” tech. rep., Swiss Finance Institute, 2020.
- [32] D. Sornette, R. Woodard, and W.-X. Zhou, “The 2006–2008 oil bubble: Evidence of speculation, and prediction,” *Physica A: Statistical Mechanics and its Applications*, vol. 388, no. 8, pp. 1571–1576, 2009.
- [33] W.-X. Zhou and D. Sornette, “Is there a real-estate bubble in the us?,” *Physica A: Statistical Mechanics and its Applications*, vol. 361, no. 1, pp. 297–308, 2006.
- [34] A. Helmstetter and D. Sornette, “Subcritical and supercritical regimes in epidemic models of earthquake aftershocks,” *Journal of Geophysical Research: Solid Earth*, vol. 107, no. B10, pp. ESE–10, 2002.
- [35] A. Helmstetter, G. Ouillon, and D. Sornette, “Are aftershocks of large Californian earthquakes diffusing?,” *Journal of Geophysical Research: Solid Earth*, vol. 108, no. B10, 2003.
- [36] A. Helmstetter and D. Sornette, “Diffusion of epicenters of earthquake aftershocks, Omori’s law, and generalized continuous-time random walk models,” *Physical Review E*, vol. 66, no. 6, p. 061104, 2002.
- [37] S. Altizer, A. Dobson, P. Hosseini, P. Hudson, M. Pascual, and P. Rohani, “Seasonality and the dynamics of infectious diseases,” *Ecology letters*, vol. 9, no. 4, pp. 467–484, 2006.
- [38] M. J. Keeling and P. Rohani, *Modeling infectious diseases in humans and animals*. Princeton university press, 2011.
- [39] A. G. Hawkes, “Spectra of some self-exciting and mutually exciting point processes,” *Biometrika*, vol. 58, no. 1, pp. 83–90, 1971.
- [40] A. G. Hawkes and D. Oakes, “A cluster process representation of a self-exciting process,” *Journal of Applied Probability*, pp. 493–503, 1974.
- [41] R. F. Engle and A. Lunde, “Trades and quotes: a bivariate point process,” *Journal of Financial Econometrics*, vol. 1, no. 2, pp. 159–188, 2003.
- [42] A.-L. Barabasi, “The origin of bursts and heavy tails in human dynamics,” *Nature*, vol. 435, no. 7039, pp. 207–211, 2005.
- [43] A. Saichev, T. Maillart, and D. Sornette, “Hierarchy of temporal responses of multivariate self-excited epidemic processes,” *The European Physical Journal B*, vol. 86, no. 4, pp. 1–19, 2013.

- [44] J. D. O'Brien, A. Aleta, Y. Moreno, and J. P. Gleeson, "Quantifying uncertainty in a predictive model for popularity dynamics," *Physical Review E*, vol. 101, no. 6, p. 062311, 2020.
- [45] A. N. Medvedev, J.-C. Delvenne, and R. Lambiotte, "Modelling structure and predicting dynamics of discussion threads in online boards," *Journal of Complex Networks*, vol. 7, no. 1, pp. 67–82, 2019.
- [46] V. Filimonov and D. Sornette, "Spurious trend switching phenomena in financial markets," *European Physical Journal B*, vol. 85, p. 155, may 2012.
- [47] V. Filimonov and D. Sornette, "Apparent criticality and calibration issues in the Hawkes self-excited point process model: application to high-frequency financial data," *Quantitative Finance*, vol. 15, no. 8, pp. 1293–1314, 2015.
- [48] E. Lewis and G. Mohler, "A nonparametric em algorithm for multiscale hawkes processes," *Journal of Nonparametric Statistics*, vol. 1, no. 1, pp. 1–20, 2011.
- [49] R. Shcherbakov, J. Zhuang, G. Zöller, and Y. Ogata, "Forecasting the magnitude of the largest expected earthquake," *Nature communications*, vol. 10, no. 1, pp. 1–11, 2019.
- [50] S. Nandan, G. Ouillon, S. Wiemer, and D. Sornette, "Objective estimation of spatially variable parameters of epidemic type aftershock sequence model: Application to california," *Journal of Geophysical Research: Solid Earth*, vol. 122, no. 7, pp. 5118–5143, 2017.
- [51] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, pp. 814–818, jun 2005.
- [52] S. P. Fraiberger, R. Sinatra, M. Resch, C. Riedl, and A. L. Barabási, "Quantifying reputation and success in art," *Science*, vol. 362, pp. 825–829, nov 2018.
- [53] R. Sinatra, D. Wang, P. Deville, C. Song, and A. L. Barabási, "Quantifying the evolution of individual scientific impact," *Science*, vol. 354, nov 2016.
- [54] S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, A. Vespignani, L. Waltman, D. Wang, and A. L. Barabási, "Science of science," mar 2018.
- [55] D. Wang, C. Song, and A. L. Barabási, "Quantifying long-term scientific impact," *Science*, vol. 342, pp. 127–132, oct 2013.
- [56] L. Liu, Y. Wang, R. Sinatra, C. L. Giles, C. Song, and D. Wang, "Hot streaks in artistic, cultural, and scientific careers," *Nature*, vol. 559, pp. 396–399, jul 2018.
- [57] D. Sornette, S. Wheatley, and P. Cauwels, "The fair reward problem: the illusion of success and how to solve it," *Advances in Complex Systems*, vol. 22, no. 3, p. 1950005 (52 pages), 2019.
- [58] A. L. Barabási, "The origin of bursts and heavy tails in human dynamics," *Nature*, vol. 435, pp. 207–211, may 2005.

- [59] F. Deschatres and D. Sornette, “Dynamics of book sales: Endogenous versus exogenous shocks in complex networks,” *Physical Review E*, vol. 72, no. 1, p. 016112, 2005.
- [60] P. Deville, D. Wang, R. Sinatra, C. Song, V. D. Blondel, and A.-L. Barabási, “Career on the move: Geography, stratification, and scientific impact,” *Scientific reports*, vol. 4, p. 4770, 2014.
- [61] J. Berger and D. Pope, “Can losing lead to winning?,” *Management Science*, vol. 57, no. 5, pp. 817–827, 2011.
- [62] S. F. Way, A. C. Morgan, A. Clauset, and D. B. Larremore, “The misleading narrative of the canonical faculty productivity trajectory,” *Proceedings of the National Academy of Sciences*, vol. 114, no. 44, pp. E9216–E9223, 2017.
- [63] A. Clauset, S. Arbesman, and D. B. Larremore, “Systematic inequality and hierarchy in faculty hiring networks,” *Science advances*, vol. 1, no. 1, p. e1400005, 2015.
- [64] O. E. Williams, L. Lacasa, and V. Latora, “Quantifying and predicting success in show business,” *Nature communications*, vol. 10, no. 1, p. 2256, 2019.
- [65] M. J. Mauboussin, *The success equation: Untangling skill and luck in business, sports, and investing*. Harvard Business Press, 2012.
- [66] A. V. Carron, S. R. Bray, and M. A. Eys, “Team cohesion and team success in sport,” *Journal of sports sciences*, vol. 20, no. 2, pp. 119–126, 2002.
- [67] S. Wuchty, B. F. Jones, and B. Uzzi, “The increasing dominance of teams in production of knowledge,” *Science*, vol. 316, no. 5827, pp. 1036–1039, 2007.
- [68] N. J. Cooke, M. L. Hilton, *et al.*, *Enhancing the effectiveness of team science*. National Academies Press Washington, DC, 2015.
- [69] L. Wu, D. Wang, and J. A. Evans, “Large teams develop and small teams disrupt science and technology,” *Nature*, vol. 566, no. 7744, pp. 378–382, 2019.
- [70] V. Larivière, Y. Gingras, C. R. Sugimoto, and A. Tsou, “Team size matters: Collaboration and scientific impact since 1900,” *Journal of the Association for Information Science and Technology*, vol. 66, no. 7, pp. 1323–1332, 2015.
- [71] S. Mukherjee, “Identifying the greatest team and captain—a complex network approach to cricket matches,” *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 23, pp. 6066–6076, 2012.
- [72] S. Mukherjee, “Quantifying individual performance in cricket—a network analysis of batsmen and bowlers,” *Physica A: Statistical Mechanics and its Applications*, vol. 393, pp. 624–637, 2014.
- [73] T. Gilovich, R. Vallone, and A. Tversky, “The hot hand in basketball: On the misperception of random sequences,” *Cognitive psychology*, vol. 17, no. 3, pp. 295–314, 1985.

- [74] D. Kahneman and A. Tversky, "On the psychology of prediction," *Psychological review*, vol. 80, no. 4, p. 237, 1973.
- [75] A. Tversky, D. Kahneman, S. Kahneman, and Tversky, "Belief in the law of small numbers," *A Handbook for Data Analysis in the Behavioral Sciences*, vol. 1, p. 341, 2014.
- [76] D. Kahneman and M. W. Riepe, "Aspects of investor psychology," *Journal of portfolio management*, vol. 24, no. 4, pp. 52–+, 1998.
- [77] L. Liu, Y. Wang, R. Sinatra, C. L. Giles, C. Song, and D. Wang, "Hot streaks in artistic, cultural, and scientific careers," *Nature*, vol. 559, no. 7714, p. 396, 2018.
- [78] J. B. Miller and A. Sanjurjo, "Surprised by the hot hand fallacy? a truth in the law of small numbers," *Econometrica*, vol. 86, no. 6, pp. 2019–2047, 2018.
- [79] J. J. Koehler and C. A. Conley, "The "hot hand" myth in professional basketball," *Journal of sport and exercise psychology*, vol. 25, no. 2, pp. 253–259, 2003.
- [80] D. Hendricks, J. Patel, and R. Zeckhauser, "Hot hands in mutual funds: Short-run persistence of relative performance, 1974–1988," *The Journal of finance*, vol. 48, no. 1, pp. 93–130, 1993.
- [81] C. J. Roney and L. M. Trick, "Sympathetic magic and perceptions of randomness: The hot hand versus the gambler's fallacy," *Thinking & reasoning*, vol. 15, no. 2, pp. 197–210, 2009.
- [82] E. F. Fama and K. R. French, "Luck versus skill in the cross-section of mutual fund returns," *The journal of finance*, vol. 65, no. 5, pp. 1915–1947, 2010.
- [83] D. Hirshleifer, "Investor psychology and asset pricing," *The Journal of Finance*, vol. 56, no. 4, pp. 1533–1597, 2001.
- [84] M. M. Carhart, "On persistence in mutual fund performance," *The Journal of finance*, vol. 52, no. 1, pp. 57–82, 1997.
- [85] G. Gigerenzer and H. Brighton, "Homo heuristicus: Why biased minds make better inferences," *Topics in cognitive science*, vol. 1, no. 1, pp. 107–143, 2009.
- [86] R. K. Merton, "The matthew effect in science: The reward and communication systems of science are considered," *Science*, vol. 159, no. 3810, pp. 56–63, 1968.
- [87] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, *et al.*, "Computational social science," *Science*, vol. 323, no. 5915, pp. 721–723, 2009.
- [88] R. Sinatra, D. Wang, P. Deville, C. Song, and A.-L. Barabási, "Quantifying the evolution of individual scientific impact," *Science*, vol. 354, no. 6312, p. aaf5239, 2016.
- [89] S. P. Fraiberger, R. Sinatra, M. Resch, C. Riedl, and A.-L. Barabási, "Quantifying reputation and success in art," *Science*, vol. 362, no. 6416, pp. 825–829, 2018.



- [90] I. Iacopini, S. Milojević, and V. Latora, “Network dynamics of innovation processes,” *Physical review letters*, vol. 120, no. 4, p. 048301, 2018.
- [91] T. Heatherton, D. M. Tice, *et al.*, *Losing control: How and why people fail at self-regulation*. San Diego, CA: Academic Press, Inc, 1994.
- [92] T. Bol, M. de Vaan, and A. van de Rijt, “The matthew effect in science funding,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 19, pp. 4887–4890, 2018.
- [93] A. Clauset, D. B. Larremore, and R. Sinatra, “Data-driven predictions in the science of science,” *Science*, vol. 355, no. 6324, pp. 477–480, 2017.
- [94] M. Jagielski, R. Kutner, and D. Sornette, “Theory of earthquakes interevent times applied to financial markets,” *Physica A: Statistical Mechanics and its Applications*, vol. 483, pp. 68–73, 2017.
- [95] R. Crane and D. Sornette, “Robust dynamic classes revealed by measuring the response function of a social system,” *Proceedings of the National Academy of Sciences*, vol. 105, no. 41, pp. 15649–15653, 2008.
- [96] D. Sornette and S. Utkin, “Limits of declustering methods for disentangling exogenous from endogenous events in time series with foreshocks, main shocks, and aftershocks,” *Physical Review E*, vol. 79, no. 6, p. 061110, 2009.
- [97] A. Helmstetter and D. Sornette, “Importance of direct and indirect triggered seismicity in the etas model of seismicity,” *Geophys. Res. Lett.*, vol. 30, no. 11, p. doi:10.1029/2003GL017670, 2003.
- [98] A. Helmstetter, D. Sornette, and J.-R. Grasso, “Mainshocks are aftershocks of conditional foreshocks: How do foreshock statistical properties emerge from aftershock laws,” *J. Geophys. Res. (Solid Earth)*, vol. 108, no. B10, pp. 2046, doi:10.1029/2002JB001991, 2003.
- [99] A. Helmstetter and D. Sornette, “Foreshocks explained by cascades of triggered seismicity,” *J. Geophys. Res. (Solid Earth)*, vol. 108, no. B10, p. 2457 10.1029/2003JB002409 01, 2003.
- [100] L. Fiévet and D. Sornette, “Decision trees unearth return sign predictability in the s&p 500,” *Quantitative Finance*, vol. 18, no. 11, pp. 1797–1814, 2018.
- [101] 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)*, vol. 57, no. 1, pp. 289–300, 1995.
- [102] Y. Hochberg, “A sharper bonferroni procedure for multiple tests of significance,” *Biometrika*, vol. 75, no. 4, pp. 800–802, 1988.
- [103] S. Holm, “A simple sequentially rejective multiple test procedure,” *Scandinavian journal of statistics*, pp. 65–70, 1979.

- [104] Z. Šidák, “Rectangular confidence regions for the means of multivariate normal distributions,” *Journal of the American Statistical Association*, vol. 62, no. 318, pp. 626–633, 1967.
- [105] J. D. Storey and R. Tibshirani, “Statistical significance for genomewide studies,” *Proceedings of the National Academy of Sciences*, vol. 100, no. 16, pp. 9440–9445, 2003.
- [106] T. Vicsek and A. Zafeiris, “Collective motion,” *Physics reports*, vol. 517, no. 3-4, pp. 71–140, 2012.
- [107] W. B. Arthur, J. H. Holland, B. LeBaron, R. Palmer, and P. Tayler, “Asset pricing under endogenous expectations in an artificial stock market,” *The economy as an evolving complex system II*, vol. 27, 1996.
- [108] P. Bak and C. Tang, “Earthquakes as a self-organized critical phenomenon,” *Journal of Geophysical Research: Solid Earth*, vol. 94, no. B11, pp. 15635–15637, 1989.
- [109] A. Sornette and D. Sornette, “Self-organized criticality and earthquakes,” *EPL (Europhysics Letters)*, vol. 9, no. 3, p. 197, 1989.
- [110] D. Sornette and C. G. Sammis, “Complex critical exponents from renormalization group theory of earthquakes: Implications for earthquake predictions,” *Journal de Physique I*, vol. 5, no. 5, pp. 607–619, 1995.
- [111] V. Keilis-Borok, “What comes next in the dynamics of lithosphere and earthquake prediction?,” *PEPI*, vol. 111, no. 3-4, pp. 179–185, 1999.
- [112] D. S. Fisher, K. Dahmen, S. Ramanathan, and Y. Ben-Zion, “Statistics of earthquakes in simple models of heterogeneous faults,” *Physical review letters*, vol. 78, no. 25, p. 4885, 1997.
- [113] G. Zöller, S. Hainzl, and J. Kurths, “Observation of growing correlation length as an indicator for critical point behavior prior to large earthquakes,” *Journal of Geophysical Research: Solid Earth*, vol. 106, no. B2, pp. 2167–2175, 2001.
- [114] K. Tiampo and M. Anghel, “Critical point theory and space–time pattern formation in precursory seismicity,” *Tectonophysics*, vol. 1, no. 413, pp. 1–3, 2006.
- [115] M. F. Xia, Y. J. Wei, F. J. Ke, and Y. L. Bai, “Critical sensitivity and trans-scale fluctuations in catastrophic rupture,” in *Earthquake Processes: Physical Modelling, Numerical Simulation and Data Analysis Part II*, pp. 2491–2509, Springer, 2002.
- [116] C. G. Bufe and D. J. Varnes, “Predictive modeling of the seismic cycle of the greater san francisco bay region,” *Journal of Geophysical Research: Solid Earth*, vol. 98, no. B6, pp. 9871–9883, 1993.
- [117] D. Bowman, G. Ouillon, C. Sammis, A. Sornette, and D. Sornette, “An observational test of the critical earthquake concept,” *Journal of Geophysical Research: Solid Earth*, vol. 103, no. B10, pp. 24359–24372, 1998.

- [118] A. Mignan, “Retrospective on the accelerating seismic release (asr) hypothesis: Controversy and new horizons,” *Tectonophysics*, vol. 505, no. 1-4, pp. 1–16, 2011.
- [119] M. Acosta, F. X. Passelègue, A. Schubnel, R. Madariaga, and M. Violay, “Can precursory moment release scale with earthquake magnitude? a view from the laboratory,” *Geophysical Research Letters*, vol. 46, no. 22, pp. 12927–12937, 2019.
- [120] Y. Huang, H. Saleur, C. Sammis, and D. Sornette, “Precursors, aftershocks, criticality and self-organized criticality,” *EPL (Europhysics Letters)*, vol. 41, no. 1, p. 43, 1998.
- [121] A. Helmstetter, S. Hergarten, and D. Sornette, “Properties of foreshocks and aftershocks of the nonconservative self-organized critical olami-feder-christensen model,” *Physical Review E*, vol. 70, no. 4, p. 046120, 2004.
- [122] Y. Ben-Zion, “Collective behavior of earthquakes and faults: Continuum-discrete transitions, progressive evolutionary changes, and different dynamic regimes,” *Reviews of Geophysics*, vol. 46, no. 4, 2008.
- [123] J. Kazemian, K. Tiampo, W. Klein, and R. Dominguez, “Foreshock and aftershocks in simple earthquake models,” *Physical review letters*, vol. 114, no. 8, p. 088501, 2015.
- [124] M. Mitzenmacher, “A brief history of generative models for power law and lognormal distributions,” *Internet mathematics*, vol. 1, no. 2, pp. 226–251, 2004.
- [125] Y. Ogata, “Statistics of earthquake activity: Models and methods for earthquake predictability studies,” *Annual Review of Earth and Planetary Sciences*, vol. 45, pp. 497–527, 2017.
- [126] A. Saichev and D. Sornette, “Superlinear scaling of offspring at criticality in branching processes,” *Physical Review E*, vol. 89, no. 1, p. 012104, 2014.
- [127] A. Saichev and D. Sornette, “Anomalous power law distribution of total lifetimes of branching processes: Application to earthquake aftershock sequences,” *Physical Review E*, vol. 70, no. 4, p. 046123, 2004.
- [128] A. Saichev and D. Sornette, “Power law distribution of seismic rates: theory and data analysis,” *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 49, no. 3, pp. 377–401, 2006.
- [129] A. Saichev and D. Sornette, ““universal” distribution of interearthquake times explained,” *Physical review letters*, vol. 97, no. 7, p. 078501, 2006.
- [130] A. Saichev, A. Helmstetter, and D. Sornette, “Power-law distributions of offspring and generation numbers in branching models of earthquake triggering,” *Pure and applied geophysics*, vol. 162, no. 6-7, pp. 1113–1134, 2005.
- [131] S. Seif, A. Mignan, J. D. Zechar, M. J. Werner, and S. Wiemer, “Estimating etas: The effects of truncation, missing data, and model assumptions,” *Journal of Geophysical Research: Solid Earth*, vol. 122, no. 1, pp. 449–469, 2017.
- [132] A. Chu, F. P. Schoenberg, P. Bird, D. D. Jackson, and Y. Y. Kagan, “Comparison of etas parameter estimates across different global tectonic zones,” *Bulletin of the Seismological Society of America*, vol. 101, no. 5, pp. 2323–2339, 2011.

- [133] D. Sornette and M. J. Werner, “Apparent clustering and apparent background earthquakes biased by undetected seismicity,” *Journal of Geophysical Research: Solid Earth*, vol. 110, no. B9, 2005.
- [134] A. Saichev and D. Sornette, “Renormalization of branching models of triggered seismicity from total to observable seismicity,” *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 51, no. 3, p. 443, 2006.
- [135] A. Wehrli, S. Wheatley, and D. Sornette, “Scale-, time- and asset-dependence of Hawkes process estimates on high frequency price changes,” *Swiss Finance Institute Research Paper*, no. 20-39, 2020.
- [136] S. Wheatley, A. Wehrli, and D. Sornette, “The endo–exo problem in high frequency financial price fluctuations and rejecting criticality,” *Quantitative Finance*, vol. 19, no. 7, pp. 1165–1178, 2019.
- [137] J. Zhuang, Y. Ogata, and D. Vere-Jones, “Stochastic declustering of space-time earthquake occurrences,” *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 369–380, 2002.
- [138] A. M. Lombardi, “Estimation of the parameters of etas models by simulated annealing,” *Scientific reports*, vol. 5, p. 8417, 2015.
- [139] J. Zhuang, Y. Ogata, and D. Vere-Jones, “Analyzing earthquake clustering features by using stochastic reconstruction,” *Journal of Geophysical Research: Solid Earth*, vol. 109, no. B5, 2004.
- [140] J. Zhuang, “Next-day earthquake forecasts for the Japan region generated by the etas model,” *Earth, planets and space*, vol. 63, no. 3, pp. 207–216, 2011.
- [141] A. Veen and F. P. Schoenberg, “Estimation of space–time branching process models in seismology using an em–type algorithm,” *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 614–624, 2008.
- [142] A. Helmstetter, Y. Y. Kagan, and D. D. Jackson, “Comparison of short-term and time-independent earthquake forecast models for southern California,” *Bulletin of the Seismological Society of America*, vol. 96, no. 1, pp. 90–106, 2006.
- [143] Y. Ogata, “Significant improvements of the space-time etas model for forecasting of accurate baseline seismicity,” *Earth, planets and space*, vol. 63, no. 3, p. 6, 2011.
- [144] C. Molkenhain, C. Donner, S. Reich, G. Zöller, S. Hainzl, M. Holschneider, and M. Opper, “Gp-etAs: Semiparametric Bayesian inference for the spatio-temporal epidemic type aftershock sequence model,” *arXiv preprint arXiv:2005.12857*, 2020.
- [145] A. Helmstetter, Y. Y. Kagan, and D. D. Jackson, “High-resolution time-independent grid-based forecast for m 5 earthquakes in California,” *Seismological Research Letters*, vol. 78, no. 1, pp. 78–86, 2007.
- [146] S. Nandan, Y. Kamer, G. Ouillon, S. Hiemer, and D. Sornette, “Global models for short-term earthquake forecasting and predictive skill assessment,” *Eur. Phys. J. Special Topics*, 2020.

- [147] R. J. Geller, D. D. Jackson, Y. Y. Kagan, and F. Mulargia, "Earthquakes cannot be predicted," *Science*, vol. 275, no. 5306, pp. 1616–1616, 1997.
- [148] B. K. Atkinson, "Subcritical crack growth in geological materials," *Journal of Geophysical Research: Solid Earth*, vol. 89, no. B6, pp. 4077–4114, 1984.
- [149] W.-j. Guan, Z.-y. Ni, Y. Hu, W.-h. Liang, C.-q. Ou, J.-x. He, L. Liu, H. Shan, C.-l. Lei, D. S. Hui, *et al.*, "Clinical characteristics of coronavirus disease 2019 in china," *New England journal of medicine*, vol. 382, no. 18, pp. 1708–1720, 2020.
- [150] W. H. Organization *et al.*, "Coronavirus disease 2019 (covid-19): situation report, 73," 2020.
- [151] J. Bedford, D. Enria, J. Giesecke, D. L. Heymann, C. Ihekweazu, G. Kobinger, H. C. Lane, Z. Memish, M.-d. Oh, A. Schuchat, *et al.*, "Covid-19: towards controlling of a pandemic," *The Lancet*, vol. 395, no. 10229, pp. 1015–1018, 2020.
- [152] B. Xu, M. U. Kraemer, and D. C. Group, "Open access epidemiological data from the covid-19 outbreak," *The Lancet. Infectious Diseases*, 2020.
- [153] B. Xu, B. Gutierrez, S. Mekaru, K. Sewalk, L. Goodwin, A. Loskill, E. L. Cohn, Y. Hswen, S. C. Hill, M. M. Cobo, *et al.*, "Epidemiological data from the covid-19 outbreak, real-time case information," *Scientific data*, vol. 7, no. 1, pp. 1–6, 2020.
- [154] H. Nishiura, N. M. Linton, and A. R. Akhmetzhanov, "Serial interval of novel coronavirus (covid-19) infections," *International journal of infectious diseases*, 2020.
- [155] A. Aktay, S. Bavadekar, G. Cossoul, J. Davis, D. Desfontaines, A. Fabrikant, E. Gabrilovich, K. Gadepalli, B. Gipson, M. Guevara, *et al.*, "Google covid-19 community mobility reports: Anonymization process description (version 1.0)," *arXiv preprint arXiv:2004.04145*, 2020.
- [156] R. J. Wilson, C. Y. Zhang, W. Lam, D. Desfontaines, D. Simmons-Marengo, and B. Gipson, "Differentially private sql with bounded user contribution," *arXiv preprint arXiv:1909.01917*, 2019.
- [157] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography conference*, pp. 265–284, Springer, 2006.
- [158] W. K. Hastings, "Monte carlo sampling methods using markov chains and their applications," 1970.
- [159] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan, "An introduction to mcmc for machine learning," *Machine learning*, vol. 50, no. 1-2, pp. 5–43, 2003.
- [160] J. Wallinga and P. Teunis, "Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures," *American Journal of epidemiology*, vol. 160, no. 6, pp. 509–516, 2004.
- [161] L. M. Bettencourt and R. M. Ribeiro, "Real time bayesian estimation of the epidemic potential of emerging infectious diseases," *PLoS One*, vol. 3, no. 5, p. e2185.doi:10.1371/journal.pone.0002185, 2008.

- 
- [162] K. H. Brodersen, F. Gallusser, J. Koehler, N. Remy, S. L. Scott, *et al.*, “Inferring causal impact using bayesian structural time-series models,” *The Annals of Applied Statistics*, vol. 9, no. 1, pp. 247–274, 2015.
- [163] R. L. Munasinghe and R. D. Morris, “Localization of disease clusters using regional measures of spatial autocorrelation,” *Statistics in Medicine*, vol. 15, no. 7-9, pp. 893–905, 1996.
- [164] J. Dehning, J. Zierenberg, F. P. Spitzner, M. Wibral, J. P. Neto, M. Wilczek, and V. Priesemann, “Inferring change points in the spread of covid-19 reveals the effectiveness of interventions,” *Science*, 2020.
- [165] T. Puoliväli, S. Palva, and J. M. Palva, “Influence of multiple hypothesis testing on reproducibility in neuroimaging research: A simulation study and python-based software,” *Journal of Neuroscience Methods*, p. 108654, 2020.
- [166] B. W. Silverman, *Density estimation for statistics and data analysis*, vol. 26. CRC press, 1986.
- [167] S. Tinti and F. Mulargia, “Confidence intervals of b values for grouped magnitudes,” *Bulletin of the Seismological Society of America*, vol. 77, no. 6, pp. 2125–2134, 1987.
- [168] D. Schorlemmer and M. Gerstenberger, “Relm testing center,” *Seismological Research Letters*, vol. 78, no. 1, pp. 30–36, 2007.

# **Appendix A**

## **Supplementary Materials to Synchronized Bursts of Productivity and Success in individual career**

### **A.1 Preparation of citation data**

For our study, we use a database containing 256 million scholars, who have published 213 million articles on 53 thousand research topics. We mine the database to extract all the articles on the topic of Signal Processing which has been published between 1971-01-01 and 2000-01-01. We find a total of 36 thousand publications within this duration. We further match these publications with their authors to reconstruct the careers. To identify the productivity peak in each career accurately, we consider the authors with at least 20 scientific articles.

### **A.2 Preparation of YouTube data**

We develop technique to classify channels who create content on a particular topic on YouTube. We analyze the voice inside the videos with state-of-the-art Natural Language Processing (NLP) techniques. We categorize the videos and subsequently the channels. We extracted the information about the evolution of view counts, likes, shares, subscribes for the channel creating similar content at a particular time point.

### **A.2.1 Quasi-Monte Carlo search for finding related YouTube channels**

YouTube provides a list of suggested channels that might be interesting for the viewers. We use the above feature to sample the list of channels, which might be creating similar content. We manually select few YouTube channels that create content related to cryptocurrency. We use them as initial seeds to query YouTube and sample the list of channels using a Quasi-Monte Carlo search algorithm. First, we find the list of suggested channels for a given channel. Further, we proceed with the search by finding the list of suggested channels for the suggested channels and so on. We redo the query with many initial conditions for sampling. We sample a list of 1 million YouTube channels who has created 100 million videos. However, this doesn't ensure that the contents in the channels are on the same topic, hence we move to the next step for further processing.

### **A.2.2 Analyzing the contents of the videos**

We systematically analyze the content of each video from the YouTube channels. We use Google speech to text synthesizer to extract the content of the videos in text format. We then use this text for analyzing and matching the voices to classify videos on the basis of the content. We analyze the texts using a customized NLP algorithm that we develop to categorize the videos. For this we first construct a vocabulary containing thousands of words that can be used to recognize the conversations related to cryptocurrency. With the help of this vocabulary and the Gestalt matching, we classified the cryptocurrency videos.

### **A.2.3 Reconstructing the YouTube channel's career**

After classifying the videos, we map back the videos with the channels and categorize the channels. We extract the evolution of daily view count for the videos. We sum the daily view counts from all the videos to find the daily view counts for the YouTube channel.

## **A.3 Sample Data**

In panel A of fig. A.1, we plot a sample of the daily view count for one of the YouTube channels. We observe a distinctive peak as well as significant precursory growth followed by an almost symmetric relaxation as predicted by our model. We further, observe the similar scaling behavior for pre-peak and post-peak view count dynamics as presented in the inset figure. This observation validates the presence of peculiar growth patterns predicted by our model in an individual channel's dynamics. In panel B, we show a typical scholarly



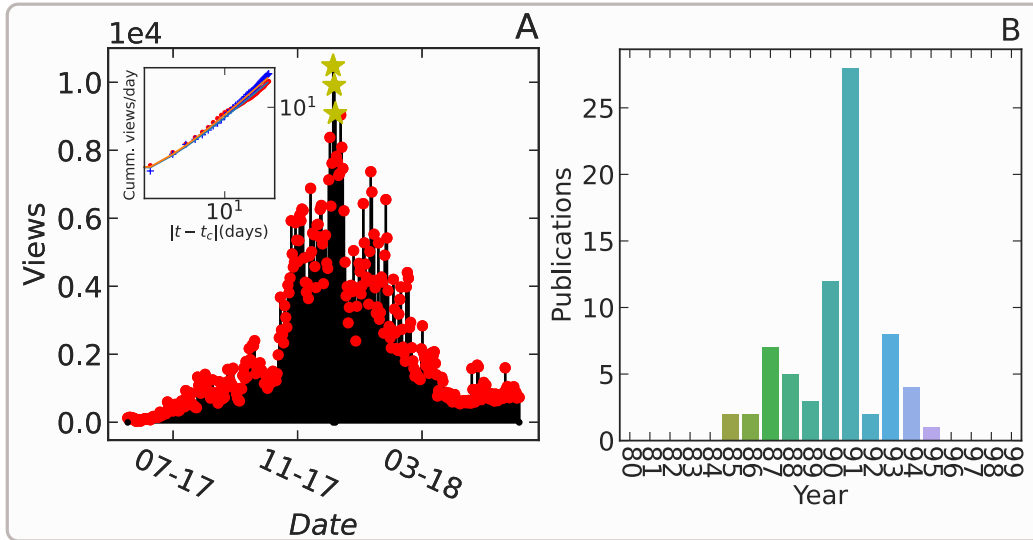


Figure A.1 Sample Scholarly productivity and YouTube success: A) An illustration of the typical response found in selected channel’s daily view time series on YouTube. (Inset) The cumulative precursory (“foreshock”) and relaxation (“aftershock”) on a log-log scale, revealing the power-law behavior that lasts over months. B) Annual productivity (no. of publications/year) within a sample scholarly career.

productivity trajectory. We observe a distinctive peak of productivity around 1991, which is present in a large number of scholars in our dataset. Further, we observe a symmetric decay of productivity around the peak.

## A.4 Validation of model prediction

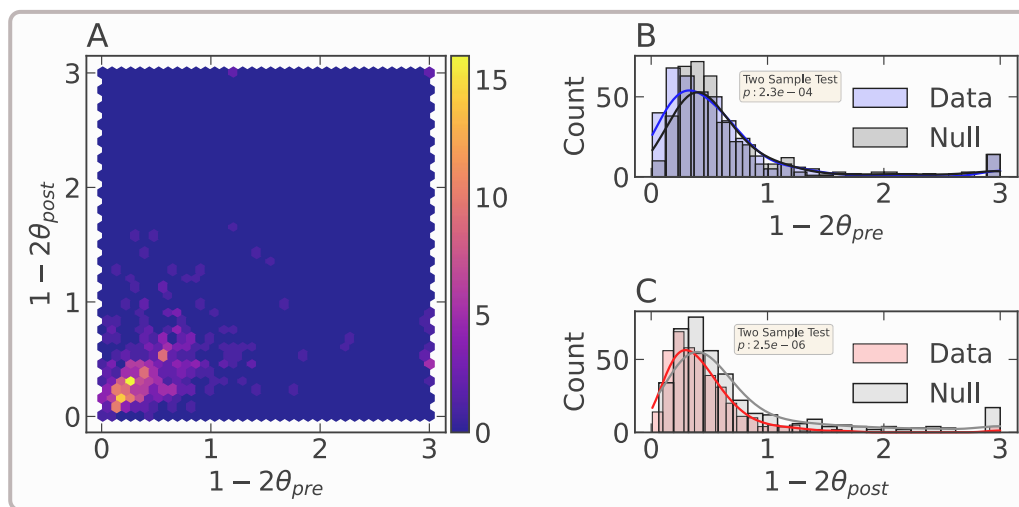


Figure A.2 Joint and marginal distribution of pre-peak and post-peak relaxation exponents in YouTube success: We calibrate the individual daily channel view time series to find the dependence of pre- and post-peak exponents. A) Joint distribution of pre-peak and post-peak exponents. The highest density of points cluster near 0.5. The fuzzy C-mean clustering gives the centroid of the distribution to be  $\sim (0.5, 0.5)$ . B) Marginal distribution of pre-peak exponents obtained from the data and the pre-peak exponents obtained from randomly shuffled null data. The two sample Kolmogorov-Smirnov test suggests that the distribution of the exponents obtained from data is significantly different than that of the exponents in null. C) same as B, except the exponent values represent the post-peak exponents.

# **Appendix B**

## **Supplementary Materials to uncovering predictability of individual and team success: Significant Hot Hand Effect in International Cricket**

### **B.1 Data Preparation**

We acquire the player and team performance dataset from <https://www.espnricinfo.com> and <http://howstat.com/> websites.

The <https://www.espnricinfo.com> website provides the complete list of cricket players, who have played at least one international cricket match, through its sub-URL <https://stats.espnricinfo.com/ci/engine/stats/index.html>. Furthermore, the website provides a graphical user interface for manually downloading and storing the list of players in HTML format. In addition to that, another sub-URL <http://search.espnricinfo.com/ci/content/player/search.html> provides a graphical user interface for accessing the information (performance in each match) about a given player and download it in HTML format. Even though the above dataset is available and accessible for public usage, doing a systematic quantitative analysis of the complete database is difficult, because of the involved accessibility problem. To solve this problem, we constructed an automated web-navigation system in python. The system sends requests to the website to obtain the dataset in HTML format with sufficient delay between consecutive requests in order to mimic a human's navigation pattern, without harming the website. We then parsed the HTML source to obtain all the useful information for our analysis

and stored the structured data in a MongoDB database for the analysis. We also stored the structured data for the Team Performance, which we obtained from <http://www.howstat.com/>.

We used the web-navigation framework to browse through <http://howstat.com/> and collected the team performance dataset. We listed all the international games played between various teams, along with the dates and the outcome. Furthermore, we navigated through the scoreboards of each game to collect the total scores for each team in each game.

## B.2 About game of cricket

- In *ODI* cricket, each team gets a chance to perform (bat) once. The sequence of performance within the game is decided with a coin toss. The individuals within the teams get the chance to perform (bat) one after another to maximize the team performance. Hence the goal of each team in the game is to maximize the team performance within a limited time frame (50 overs) and limited number of dismissals (10 individuals).
- Similarly in *Test* cricket each team gets a chance to perform (bat) a maximum of two times (two innings) within a game period of maximum 5 days. The individuals within the teams get the chance to perform one after another to maximize the team performance. Hence the goal of each team in the game is to maximize the team performance within a limited number of dismissals (10 individuals) with no over restrictions.
- In both formats the team performance in can be considered as the aggregation of participating individual performances.
- **Individual performance**
  1. We look at the individual batting performances for our study.
  2. An innings is one of the divisions of a cricket match during which one team takes its turn to bat.
  3. We call the total scored runs by an individual as the performance. While doing this, we add a water level of 1 with the runs, i.e,  $S_j(t) = Run_j(t) + 1$ . By doing this, we set the smallest score to 1, which provides us with a well-defined performance fingerprint value for all the performance values. In other words, adding 1 removes the singularity associated with zero run scored in our performance fingerprint (1).
- **Team performance**

1. For both *ODI* and *Test* cricket, the performance of the first batting team determines the trajectory for rest of the game. Hence, we only consider the batting performances of the first batting team for the quantification.
2. For the analysis, we only take into account the games that had a definite outcome, i.e, Win or Loss for either of the teams. We don't study the games where there was no winning team for the match.

### B.3 Supporting Results

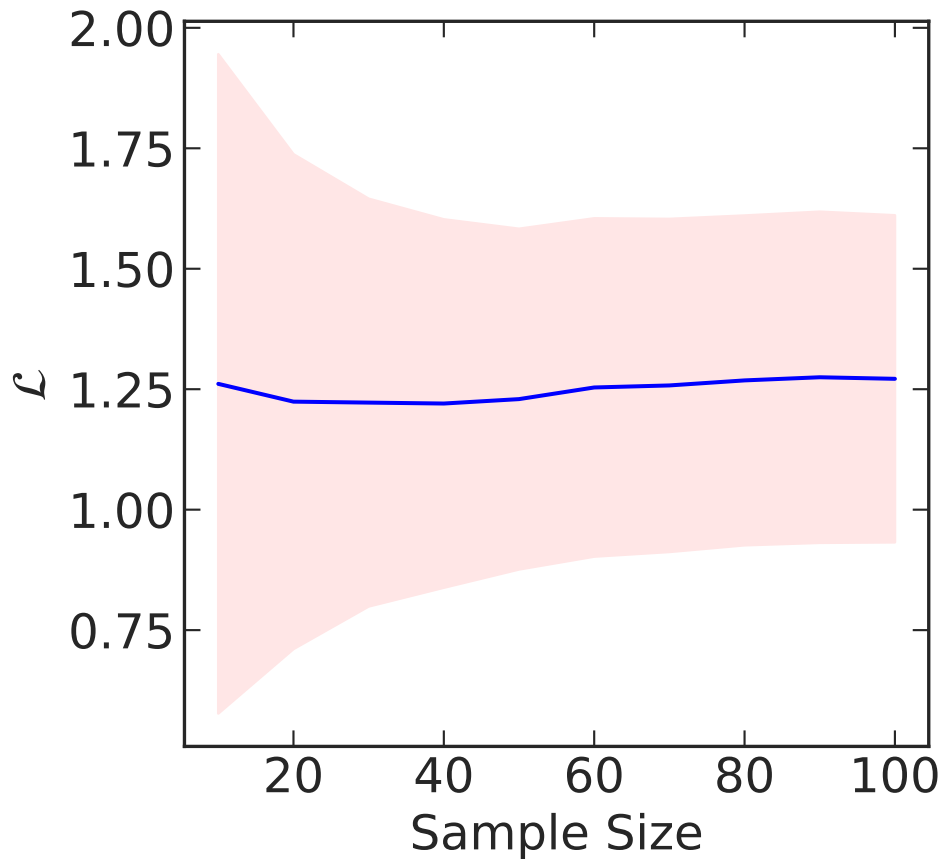


Figure B.1 **Goodness of fit:** The mean (blue line) and standard deviation (red band) of log-likelihood scores is plotted against the sample size. We observe that the estimation becomes reliable close to sample size of 30.

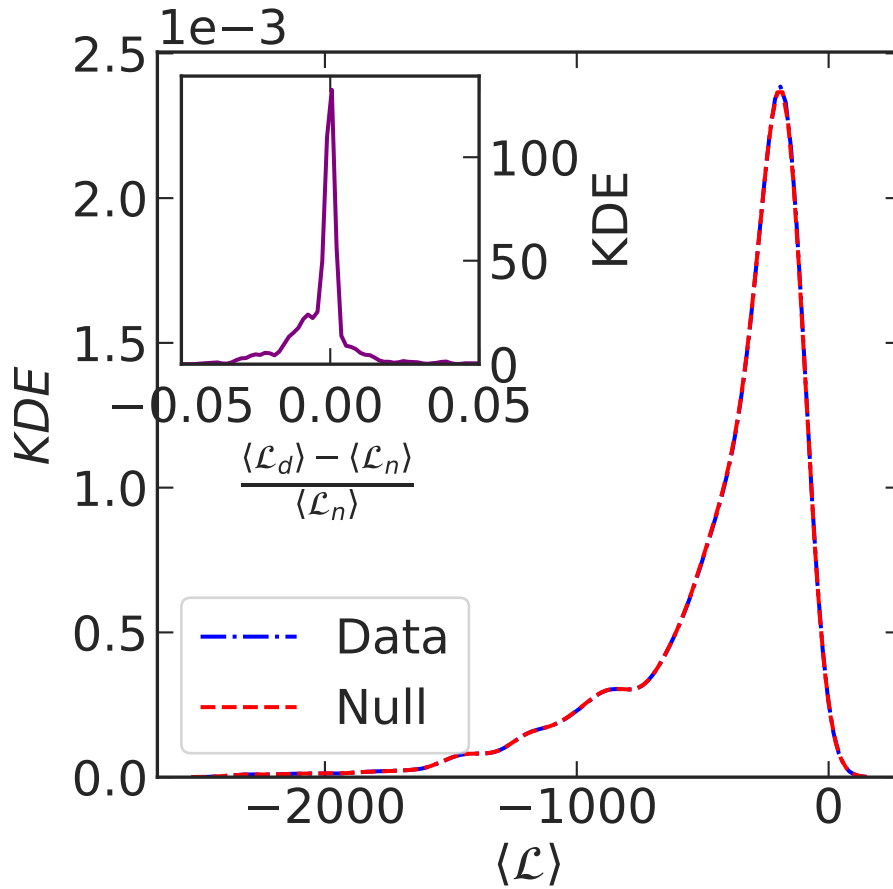


Figure B.2 **ARIMA Model:** We calibrate all the individual performances in ODI cricket using ARIMA model. We conduct differencing tests (Kwiatkowski-Phillips-Schmidt-Shin, Augmented Dickey-Fuller and Phillips-Perron) to determine the order of differencing, ‘d’. We analyse the auto-correlation functions (ACF), partial auto-correlation functions (PACF) and use Canova-Hansen method to determine the optimal value of ‘p’ and ‘q’ for each model. We perform the ARIMA calibration 100 times on each performance sequence and randomly shuffled performance sequence (null), and note down the log-likelihood scores. The main figure represents the Kernel Density Estimation (KDE) of median log-likelihood scores obtained from the original performance sequence and the shuffled performance sequence. The inset figure shows the KDE for relative difference of median log-likelihood scores obtained from the original performance sequence and the shuffled null performance sequence.

### B.3.1 Model comparison

- We perform Shapiro-Wilk test on the the original null log-likelihood distributions for each career to find that the paired differences are not normally distributed.
- We thus perform the Wilcoxon signed-rank test to determine the significance of the observed results.

- 
- We found 223 out of 610 players (or 36% players) have a significantly better (with confidence level 5%) median log-likelihood score in original performance sequence compared to the shuffled sequence. This makes us to believe that the probability of falsely rejecting the null hypothesis is ( $< 10^{-6}$ ). And we conclude that there is some predictable pattern hidden inside the performance sequences.
  - While we compare these results with the model proposed in the main text, we find that our model produces better prediction results on 46.8% of the player compared to 36% players better prediction results with the ARIMA model. We thus validate the superiority of our model compared to tradition forecasting models.

## B.4 Hot team hands

Team	$n$				$\mathcal{L}$			
	$\langle Data \rangle$	$\langle Null \rangle$	$\langle Data \rangle > \langle Null \rangle$	p	$\langle Model \rangle$	$\langle Control \rangle$	$\langle Model \rangle > \langle Control \rangle$	p
Australia	1.30E-04	5.48E-05	True	$< 10^{-6}$	4.53E+00	4.53E+00	False	$< 10^{-6}$
Bangladesh	5.58E-10	1.82E-10	True	$< 10^{-6}$	4.44E+00	4.44E+00	False	$< 10^{-6}$
England	8.55E-07	8.54E-07	True	$< 10^{-6}$	4.56E+00	4.56E+00	False	$< 10^{-6}$
India	7.63E-05	1.90E-05	True	$< 10^{-6}$	4.54E+00	4.54E+00	False	$< 10^{-6}$
New Zealand	8.73E-06	4.47E-06	True	$< 10^{-6}$	4.53E+00	4.53E+00	False	$3 \cdot 10^{-1}$
Pakistan	5.71E-05	5.13E-05	True	$6 \cdot 10^{-1}$	4.49E+00	4.49E+00	True	$< 10^{-6}$
South Africa	4.71E-08	2.56E-08	True	$< 10^{-6}$	4.58E+00	4.58E+00	False	$< 10^{-6}$
Sri Lanka	9.00E-05	1.24E-05	True	$< 10^{-6}$	4.43E+00	4.43E+00	False	$< 10^{-6}$
West Indies	6.88E-07	1.81E-06	False	$< 10^{-6}$	4.41E+00	4.41E+00	True	$< 10^{-6}$
Zimbabwe	2.56E-08	6.94E-08	False	$< 10^{-6}$	4.25E+00	4.25E+00	True	$< 10^{-6}$

**Table B.1 Prediction of team performance in ODI format.** We partition the performances into training and validation set to quantify ( $\mathcal{L}$ ) the Control and Model forecasts 100 times and note down the median values. Similarly, we estimate the branching ratio ( $n$ ) on the original data and shuffled data (Null), 100 times and note down the median values. We perform the Wilcoxon signed-rank test to determine the significance values for the observations.

Team	$n$				$\mathcal{L}$			
	$\langle Data \rangle$	$\langle Null \rangle$	$\langle Data \rangle > \langle Null \rangle$	p	$\langle Model \rangle$	$\langle Control \rangle$	$\langle Model \rangle > \langle Control \rangle$	p
Australia	4.44E-05	1.22E-06	True	$< 10^{-4}$	4.72E+00	4.72E+00	False	$< 10^{-4}$
Bangladesh	1.40E-31	1.62E-31	False	$< 10^{-4}$	4.26E+00	4.26E+00	False	$5 \cdot 10^{-2}$
England	8.97E-04	6.66E-05	True	$< 10^{-4}$	4.38E+00	4.38E+00	True	$< 10^{-4}$
India	2.26E-10	9.77E-11	True	$3 \cdot 10^{-1}$	4.70E+00	4.70E+00	False	$< 10^{-4}$
New Zealand	2.30E-10	4.99E-12	True	$< 10^{-4}$	4.58E+00	4.58E+00	False	$< 10^{-4}$
Pakistan	5.06E-12	9.84E-13	True	$< 10^{-4}$	4.45E+00	4.45E+00	False	$< 10^{-4}$
South Africa	1.44E-10	1.94E-13	True	$< 10^{-4}$	4.78E+00	4.78E+00	False	$< 10^{-4}$
Sri Lanka	1.90E-18	2.70E-19	True	$1 \cdot 10^{-1}$	4.63E+00	4.63E+00	False	$1 \cdot 10^{-1}$
West Indies	4.51E-11	2.00E-11	True	$3 \cdot 10^{-3}$	4.51E+00	4.51E+00	True	$< 10^{-4}$
Zimbabwe	6.75E-42	7.92E-42	False	$6 \cdot 10^{-2}$	4.55E+00	4.55E+00	False	NA

**Table B.2 Prediction of team performance in Test format.** We partition the performances into training and validation set to quantify ( $\mathcal{L}$ ) the Control and Model forecasts 100 times and note down the median values. Similarly, we estimate the branching ratio ( $n$ ) on the original data and shuffled data (Null), 100 times and note down the median values. We perform the Wilcoxon signed-rank test to determine the significance values for the observations.



## B.5 Hot winning hands

### B.5.1 Multiple Hypothesis Testing

We face the problem of multiple hypothesis testing, while simultaneously considering multiple individual tests on the same dataset or dependent datasets. In order to resolve the above problem, a number of methods have been proposed, which tries to solve the issue by correcting the error rates of individual tests according to the number of simultaneously considered hypotheses as well as  $p$ -values for the individual tests. Below is a list of methods that we have implemented to test our hypotheses [165].

#### Sidak's Test

Let  $p_1, \dots, p_m$  be the  $p$ -values for the family of  $m$  null hypotheses  $H_1, \dots, H_m$ . If we set the family-wise alpha level to  $\alpha$ , according to the test, we reject all null hypotheses that have a  $p$ -value lower than  $\alpha_{SID} = 1 - (1 - \alpha)^{\frac{1}{m}}$ . This test produces a family-wise Type I error rate of exactly  $\alpha$  when the tests are independent from each other and all null hypotheses are true.[104]

#### Holm's Test

Let  $p_1, \dots, p_m$  be the  $p$ -values for the family of  $m$  null hypotheses  $H_1, \dots, H_m$ , and let us introduce the sorted values (from lowest to highest) denoted  $p_{(1)} \dots p_{(m)}$  with the associated hypotheses be  $H_{(1)}, \dots, H_{(m)}$ . Then for a given significance level  $\alpha$ , if  $k$  is the minimal index such that  $p_{(k)} > \frac{\alpha}{m+1-k}$ , we reject the null hypotheses  $H_{(1)}, \dots, H_{(k-1)}$  and do not reject  $H_{(k)}, \dots, H_{(m)}$ . If  $k = 1$  then we do not reject any of the null hypotheses. If no such  $k$  exists, then we reject all of the null hypotheses. This method ensures that  $\text{FWER} \leq \alpha$ , where FWER is the family-wise error rate.[103]

#### Hochberg's Test

Let  $p_1, \dots, p_m$  be the  $p$ -values for the family of  $m$  null hypotheses  $H_1, \dots, H_m$ , and let us introduce the sorted values (from lowest to highest) denoted  $p_{(1)} \dots p_{(m)}$  with the associated hypotheses be  $H_{(1)}, \dots, H_{(m)}$ . For a given  $\alpha$ , let  $R$  be the largest  $k$  such that  $P_{(k)} \leq \frac{\alpha}{m-k+1}$ . Then, we reject the null hypotheses  $H_{(1)} \dots H_{(R)}$ . [102]

### Bonferroni's Test

Let  $p_1, \dots, p_m$  be the  $p$ -values for the family of  $m$  null hypotheses  $H_1, \dots, H_m$ , with  $m_0$  number of true null hypotheses. Let the family-wise error rate ( $FWER$ ) be the probability of rejecting at least one true  $H_i$ , that is, of making at least one type I error. Then the Bonferroni correction rejects the null hypothesis for each  $p_i \leq \frac{\alpha}{m}$ , thereby controlling the  $FWER$  at  $\leq \alpha$ . [104]

### Classic FDR Test

Let  $p_1, \dots, p_m$  be the  $p$ -values for the family of  $m$  null hypotheses  $H_1, \dots, H_m$ , and let us define the sorted values (from lowest to highest) denoted  $p_{(1)} \dots p_{(m)}$  with the associated hypotheses be  $H_{(1)}, \dots, H_{(m)}$ . The Benjamini-Hochberg test controls the FDR (False Discovery Rate) at level  $\alpha$ . For a given  $\alpha$ , if  $k$  is the largest value such that  $P_{(k)} \leq \frac{k}{m}\alpha$ , we reject the null hypothesis (i.e., declare discoveries) for all  $H_{(i)}$  for  $i = 1, \dots, k$ . [101, 102]

### Storey's Test

Let  $T_1, \dots, T_m$  be i.i.d. random variables representing the test statistics associated with  $m$  tests of the null hypothesis  $H_0$  versus an alternative hypothesis  $H_i$ , such that

$$T_i | H_i \sim (1 - H_i) \cdot F_0 + H_i \cdot F_1$$

In other words,  $T_i$  follows the null distribution  $F_0$  if  $H_i = 0$ , i.e,  $H_0$  is true for test  $i$ , else  $T_i$  follows the alternative distribution  $F_1$  if  $H_i = 1$ , i.e,  $H_1$  is true. Suppose  $H_i \sim \text{Bernoulli}(\pi_1)$ , with probability of success for  $H_i$  and  $H_0$  respectively  $\pi_1$  and  $\pi_0 = 1 - \pi_1$ . We denote the critical region at significance level  $\alpha$  by  $\Gamma_\alpha$ , corresponding to the values of  $T_i$  for which  $H_0$  is rejected beyond  $\Gamma_\alpha$ . Let an experiment yield a value  $t$  for the test statistic. The  $q$ -value of  $t$  is formally defined as

$$\inf_{\{\Gamma_\alpha: t \in \Gamma_\alpha\}} \text{pFDR}(\Gamma_\alpha)$$

That is, the  $q$ -value is the infimum of the pFDR if  $H_0$  is rejected for test statistics with values  $\geq t$ . Equivalently, the  $q$ -value equals

$$\inf_{\{\Gamma_\alpha: t \in \Gamma_\alpha\}} \Pr(H = 0 | T \in \Gamma_\alpha)$$

which is the infimum of the probability that  $H_0$  is true given that  $H_0$  is rejected (the false discovery rate). We use the  $q$ -value to reject or accept the null hypothesis.[105]

## B.5.2 Winning Streaks in ODI Teams

Country	$n$	$f$	$p(n)$	$p(n_f)$	Sidak		Holm		Hochberg		Bonferroni		Classic FDR		Storey FDR		Storey's Q	
					$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$
AUS	1	76	1.00	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
AUS	2	38	0.91	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
AUS	3	25	0.82	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
AUS	4	15	0.73	0.98	False	False	False	False	False	False	False	False	False	False	False	False	0.98	0.98
AUS	5	10	0.65	0.96	False	False	False	False	False	False	False	False	False	False	False	False	0.96	0.96
AUS	6	5	0.56	0.97	False	False	False	False	False	False	False	False	False	False	False	False	0.97	0.97
AUS	7	7	0.47	0.40	False	False	False	False	False	False	False	False	False	False	False	False	0.31	0.31
AUS	8	2	0.38	0.92	False	False	False	False	False	False	False	False	False	False	False	False	0.76	0.92
AUS	9	4	0.30	0.20	False	False	False	False	False	False	False	False	False	False	False	False	0.07	0.07
AUS	10	4	0.22	0.07	False	False	False	False	False	False	False	False	False	False	False	False	0.22	0.15
AUS	12	1	0.11	0.47	False	False	False	False	False	False	False	False	False	False	True	False	0.03	0.06
AUS	13	2	0.07	0.05	False	False	False	False	False	False	False	False	False	False	False	False	0.07	0.07
AUS	21	1	0.00	0.01	True	True	True	True	True	True	True	True	True	True	True	True	0.00	0.01

Table B.3 Winning streak statistics for team Australia in *ODI* format. In the table  $n$  denotes the length of the winning streak,  $f$  is the corresponding frequency of occurrence.  $p(n)$  and  $p(n_f)$  are the  $p$ -values for observation of streaks with length  $n$  and streaks with length  $n$  conditional on frequency  $f$ . With the help of multiple hypothesis tests mentioned in the previous section, we check the significance of the  $p$ -values and note down the results. We write result to be True, for the positive result, implying the significant  $p$ -value, otherwise it is False.

Country	$n$	$f$	$p(n)$	$p(n_f)$	Sidak		Holm		Hochberg		Bonferroni		Classic FDR		Storey FDR		Storey's Q	
					$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$
BAN	1	25	1.00	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
BAN	2	10	0.79	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
BAN	3	5	0.58	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
BAN	4	5	0.37	0.18	False	False	False	False	False	False	False	False	False	False	False	False	0.11	0.11
BAN	5	3	0.18	0.12	False	False	False	False	False	False	False	False	False	False	False	False	0.18	0.18
BAN	6	2	0.07	0.05	False	False	False	False	False	False	False	False	False	False	False	False	0.07	0.07
BAN	7	1	0.02	0.12	<b>True</b>	False	False	False	<b>True</b>	False	<b>True</b>	False	<b>True</b>	False	<b>True</b>	False	0.05	0.12
BAN	9	1	0.00	0.01	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	0.01	0.01

Table B.4 Winning streak statistics for team Bangladesh in *ODI* format. In the table  $n$  denotes the length of the winning streak,  $f$  is the corresponding frequency of occurrence.  $p(n)$  and  $p(n_f)$  are the  $p$ -values for observation of streaks with length  $n$  and streaks with length  $n$  conditional on frequency  $f$ . With the help of multiple hypothesis tests mentioned in the previous section, we check the significance of the  $p$ -values and note down the results. We write result to be True, for the positive result, implying the significant  $p$ -value, otherwise it is False.

Country	$n$	$f$	$p(n)$	$p(n_f)$	Sidak		Holm		Hochberg		Bonferroni		Classic FDR		Storey FDR		Storey's Q	
					$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$
ENG	1	75	1.00	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
ENG	2	36	0.88	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
ENG	3	22	0.76	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
ENG	4	8	0.64	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
ENG	5	8	0.52	0.68	False	False	False	False	False	False	False	False	False	False	False	False	0.68	0.68
ENG	6	4	0.40	0.70	False	False	False	False	False	False	False	False	False	False	False	False	0.70	0.70
ENG	7	2	0.28	0.71	False	False	False	False	False	False	False	False	False	False	False	False	0.57	0.71
ENG	8	3	0.19	0.12	False	False	False	False	False	False	False	False	False	False	False	False	0.19	0.19
ENG	10	1	0.06	0.29	False	False	False	False	False	False	False	False	False	False	<b>True</b>	<b>True</b>	0.01	0.02
ENG	11	1	0.03	0.17	False	False	False	False	False	False	False	False	False	False	False	False	0.07	0.17

Table B.5 Winning streak statistics for team England in *ODI* format. In the table  $n$  denotes the length of the winning streak,  $f$  is the corresponding frequency of occurrence.  $p(n)$  and  $p(n_f)$  are the  $p$ -values for observation of streaks with length  $n$  and streaks with length  $n$  conditional on frequency  $f$ . With the help of multiple hypothesis tests mentioned in the previous section, we check the significance of the  $p$ -values and note down the results. We write result to be True, for the positive result, implying the significant  $p$ -value, otherwise it is False.

Country	$n$	$f$	$p(n)$	$p(n_f)$	Sidak		Holm		Hochberg		Bonferroni		Classic FDR		Storey FDR		Storey's Q	
					$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$
IND	1	98	1.00	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
IND	2	52	0.89	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
IND	3	30	0.78	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
IND	4	11	0.67	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
IND	5	10	0.56	0.90	False	False	False	False	False	False	False	False	False	False	False	False	0.90	0.90
IND	6	6	0.45	0.70	False	False	False	False	False	False	False	False	False	False	False	False	0.70	0.70
IND	7	4	0.34	0.47	False	False	False	False	False	False	False	False	False	False	False	False	0.28	0.28
IND	8	4	0.24	0.10	False	False	False	False	False	False	False	False	False	False	False	False	0.24	0.20
IND	9	2	0.15	0.30	False	False	False	False	False	False	False	False	False	False	<b>True</b>	<b>True</b>	0.04	0.04

Table B.6 Winning streak statistics for team India in *ODI* format. In the table  $n$  denotes the length of the winning streak,  $f$  is the corresponding frequency of occurrence.  $p(n)$  and  $p(n_f)$  are the  $p$ -values for observation of streaks with length  $n$  and streaks with length  $n$  conditional on frequency  $f$ . With the help of multiple hypothesis tests mentioned in the previous section, we check the significance of the  $p$ -values and note down the results. We write result to be True, for the positive result, implying the significant  $p$ -value, otherwise it is False.

Country	$n$	$f$	$p(n)$	$p(n_f)$	Sidak		Holm		Hochberg		Bonferroni		Classic FDR		Storey FDR		Storey's Q	
					$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$
NZL	1	84	1.00	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
NZL	2	38	0.87	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
NZL	3	20	0.73	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
NZL	4	10	0.60	0.99	False	False	False	False	False	False	False	False	False	False	False	False	0.99	0.99
NZL	5	5	0.46	0.91	False	False	False	False	False	False	False	False	False	False	False	False	0.91	0.91
NZL	6	1	0.33	0.98	False	False	False	False	False	False	False	False	False	False	False	False	0.66	0.98
NZL	7	2	0.21	0.44	False	False	False	False	False	False	False	False	False	False	False	False	0.15	0.16
NZL	9	2	0.06	0.03	False	False	False	False	False	False	False	False	False	False	False	False	0.06	0.06
NZL	10	2	0.03	0.01	False	<b>True</b>	False	<b>True</b>	<b>True</b>	<b>True</b>	False	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	0.03	0.02

Table B.7 Winning streak statistics for team New Zealand in *ODI* format. In the table  $n$  denotes the length of the winning streak,  $f$  is the corresponding frequency of occurrence.  $p(n)$  and  $p(n_f)$  are the  $p$ -values for observation of streaks with length  $n$  and streaks with length  $n$  conditional on frequency  $f$ . With the help of multiple hypothesis tests mentioned in the previous section, we check the significance of the  $p$ -values and note down the results. We write result to be True, for the positive result, implying the significant  $p$ -value, otherwise it is False.

Country	$n$	$f$	$p(n)$	$p(n_f)$	Sidak		Holm		Hochberg		Bonferroni		Classic FDR		Storey FDR		Storey's Q	
					$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$
PAK	1	94	1.00	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
PAK	2	41	0.89	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
PAK	3	23	0.78	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
PAK	4	18	0.66	0.95	False	False	False	False	False	False	False	False	False	False	False	False	0.95	0.95
PAK	5	9	0.55	0.89	False	False	False	False	False	False	False	False	False	False	False	False	0.89	0.89
PAK	6	8	0.44	0.22	False	False	False	False	False	False	False	False	False	False	False	False	0.17	0.17
PAK	7	3	0.33	0.65	False	False	False	False	False	False	False	False	False	False	False	False	0.65	0.65
PAK	9	3	0.14	0.05	False	False	False	False	False	False	False	False	False	False	False	False	0.14	0.10
PAK	10	1	0.09	0.41	False	False	False	False	False	False	False	False	False	False	True	False	0.04	0.08
PAK	12	1	0.03	0.14	False	False	False	False	False	False	False	False	False	False	False	False	0.05	0.14

Table B.8 Winning streak statistics for team Pakistan in *ODI* format. In the table  $n$  denotes the length of the winning streak,  $f$  is the corresponding frequency of occurrence.  $p(n)$  and  $p(n_f)$  are the  $p$ -values for observation of streaks with length  $n$  and streaks with length  $n$  conditional on frequency  $f$ . With the help of multiple hypothesis tests mentioned in the previous section, we check the significance of the  $p$ -values and note down the results. We write result to be True, for the positive result, implying the significant  $p$ -value, otherwise it is False.

Country	$n$	$f$	$p(n)$	$p(n_f)$	Sidak		Holm		Hochberg		Bonferroni		Classic FDR		Storey FDR		Storey's Q	
					$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$
SAF	1	45	1.00	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
SAF	2	26	0.90	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
SAF	3	23	0.81	0.79	False	False	False	False	False	False	False	False	False	False	False	False	0.81	0.81
SAF	4	10	0.71	0.98	False	False	False	False	False	False	False	False	False	False	False	False	0.98	0.98
SAF	5	9	0.62	0.71	False	False	False	False	False	False	False	False	False	False	False	False	0.71	0.71
SAF	6	6	0.52	0.69	False	False	False	False	False	False	False	False	False	False	False	False	0.69	0.69
SAF	7	1	0.43	0.99	False	False	False	False	False	False	False	False	False	False	False	False	0.86	0.99
SAF	8	2	0.34	0.76	False	False	False	False	False	False	False	False	False	False	False	False	0.67	0.76
SAF	9	1	0.25	0.83	False	False	False	False	False	False	False	False	False	False	False	False	0.50	0.83
SAF	10	3	0.18	0.09	False	False	False	False	False	False	False	False	False	False	False	False	0.18	0.18
SAF	12	1	0.08	0.34	False	False	False	False	False	False	False	False	False	False	True	False	0.02	0.05
SAF	17	1	0.01	0.04	True	False	True	False	True	True	True	False	True	True	True	True	0.02	0.04

Table B.9 Winning streak statistics for team South Africa in *ODI* format. In the table  $n$  denotes the length of the winning streak,  $f$  is the corresponding frequency of occurrence.  $p(n)$  and  $p(n_f)$  are the  $p$ -values for observation of streaks with length  $n$  and streaks with length  $n$  conditional on frequency  $f$ . With the help of multiple hypothesis tests mentioned in the previous section, we check the significance of the  $p$ -values and note down the results. We write result to be True, for the positive result, implying the significant  $p$ -value, otherwise it is False.

Country	$n$	$f$	$p(n)$	$p(n_f)$	Sidak		Holm		Hochberg		Bonferroni		Classic FDR		Storey FDR		Storey's Q	
					$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$
SRL	1	92	1.00	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
SRL	2	40	0.87	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
SRL	3	20	0.74	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
SRL	4	11	0.60	0.97	False	False	False	False	False	False	False	False	False	False	False	False	0.97	0.97
SRL	5	5	0.47	0.94	False	False	False	False	False	False	False	False	False	False	False	False	0.94	0.94
SRL	6	1	0.34	0.99	False	False	False	False	False	False	False	False	False	False	False	False	0.68	0.99
SRL	7	4	0.22	0.10	False	False	False	False	False	False	False	False	False	False	False	False	0.22	0.21
SRL	9	2	0.06	0.05	False	False	False	False	False	False	False	False	False	False	False	False	0.06	0.06
SRL	10	3	0.03	0.00	False	<b>True</b>	False	<b>True</b>	<b>True</b>	<b>True</b>	False	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	0.03	0.00

Table B.10 Winning streak statistics for team Sri Lanka in *ODI* format. In the table  $n$  denotes the length of the winning streak,  $f$  is the corresponding frequency of occurrence.  $p(n)$  and  $p(n_f)$  are the  $p$ -values for observation of streaks with length  $n$  and streaks with length  $n$  conditional on frequency  $f$ . With the help of multiple hypothesis tests mentioned in the previous section, we check the significance of the  $p$ -values and note down the results. We write result to be True, for the positive result, implying the significant  $p$ -value, otherwise it is False.

Country	$n$	$f$	$p(n)$	$p(n_f)$	Sidak		Holm		Hochberg		Bonferroni		Classic FDR		Storey FDR		Storey's Q	
					$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$
WIN	1	75	1.00	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
WIN	2	31	0.88	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
WIN	3	23	0.76	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
WIN	4	15	0.63	0.90	False	False	False	False	False	False	False	False	False	False	False	False	0.63	0.63
WIN	5	3	0.51	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
WIN	6	10	0.39	0.01	False	<b>True</b>	False	<b>True</b>	False	<b>True</b>	False	<b>True</b>	False	<b>True</b>	False	<b>True</b>	0.12	0.00
WIN	7	2	0.27	0.72	False	False	False	False	False	False	False	False	False	False	False	False	0.54	0.72
WIN	8	1	0.17	0.70	False	False	False	False	False	False	False	False	False	False	False	False	0.34	0.70
WIN	9	1	0.10	0.45	False	False	False	False	False	False	False	False	False	False	<b>True</b>	False	0.03	0.08
WIN	10	1	0.06	0.28	False	False	False	False	False	False	False	False	False	False	<b>True</b>	<b>True</b>	0.01	0.01
WIN	11	1	0.03	0.16	False	False	False	False	False	False	False	False	False	False	False	False	0.06	0.16

Table B.11 Winning streak statistics for team West Indies in *ODI* format. In the table  $n$  denotes the length of the winning streak,  $f$  is the corresponding frequency of occurrence.  $p(n)$  and  $p(n_f)$  are the  $p$ -values for observation of streaks with length  $n$  and streaks with length  $n$  conditional on frequency  $f$ . With the help of multiple hypothesis tests mentioned in the previous section, we check the significance of the  $p$ -values and note down the results. We write result to be True, for the positive result, implying the significant  $p$ -value, otherwise it is False.

Country	$n$	$f$	$p(n)$	$p(n_f)$	Sidak		Holm		Hochberg		Bonferroni		Classic FDR		Storey FDR		Storey's Q		
					$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$
ZIM	1	53	1.00	1.00	False	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
ZIM	2	17	0.77	1.00	False	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
ZIM	3	6	0.53	1.00	False	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
ZIM	4	4	0.30	0.23	False	False	False	False	False	False	False	False	False	False	False	False	False	0.10	0.10
ZIM	5	1	0.12	0.55	False	False	False	False	False	False	False	False	False	False	False	False	False	0.23	0.55
ZIM	6	1	0.03	0.19	False	False	False	False	False	False	False	False	False	False	False	False	False	0.07	0.19
ZIM	7	1	0.01	0.05	<b>True</b>	False	<b>True</b>	False	<b>True</b>	False	<b>True</b>	False	<b>True</b>	False	<b>True</b>	False	<b>True</b>	0.02	0.05

Table B.12 Winning streak statistics for team Zimbabwe in *ODI* format. In the table  $n$  denotes the length of the winning streak,  $f$  is the corresponding frequency of occurrence.  $p(n)$  and  $p(n_f)$  are the  $p$ -values for observation of streaks with length  $n$  and streaks with length  $n$  conditional on frequency  $f$ . With the help of multiple hypothesis tests mentioned in the previous section, we check the significance of the  $p$ -values and note down the results. We write result to be True, for the positive result, implying the significant  $p$ -value, otherwise it is False.



### B.5.3 Winning Streaks in Test Teams

Country	$n$	$f$	$p(n)$	$p(n_f)$	Sidak		Holm		Hochberg		Bonferroni		Classic FDR		Storey FDR		Storey's Q	
					$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$
AUS	1	44	1.00	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
AUS	2	25	0.91	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
AUS	3	16	0.81	0.99	False	False	False	False	False	False	False	False	False	False	False	False	0.99	0.99
AUS	4	9	0.72	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
AUS	5	8	0.62	0.85	False	False	False	False	False	False	False	False	False	False	False	False	0.85	0.85
AUS	6	6	0.53	0.72	False	False	False	False	False	False	False	False	False	False	False	False	0.72	0.72
AUS	7	5	0.43	0.40	False	False	False	False	False	False	False	False	False	False	False	False	0.30	0.30
AUS	8	3	0.34	0.52	False	False	False	False	False	False	False	False	False	False	False	False	0.23	0.23
AUS	12	2	0.08	0.07	False	False	False	False	False	False	False	False	False	False	False	False	0.08	0.08
AUS	16	1	0.02	0.08	True	False	True	False	True	False	True	False	True	False	True	False	0.03	0.08
AUS	20	2	0.00	0.00	True	True	True	True	True	True	True	True	True	True	True	True	0.00	0.00

Table B.13 Winning streak statistics for team Australia in *Test* format. In the table  $n$  denotes the length of the winning streak,  $f$  is the corresponding frequency of occurrence.  $p(n)$  and  $p(n_f)$  are the  $p$ -values for observation of streaks with length  $n$  and streaks with length  $n$  conditional on frequency  $f$ . With the help of multiple hypothesis tests mentioned in the previous section, we check the significance of the  $p$ -values and note down the results. We write result to be True, for the positive result, implying the significant  $p$ -value, otherwise it is False.

Country	$n$	$f$	$p(n)$	$p(n_f)$	Sidak		Holm		Hochberg		Bonferroni		Classic FDR		Storey FDR		Storey's Q	
					$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$
BAN	1	4	1.0	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.0	1.00
BAN	2	2	0.5	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.0	1.00
BAN	3	2	0.1	0.04	False	False	False	False	False	False	False	False	False	False	False	False	0.1	0.08

Table B.14 Winning streak statistics for team Bangladesh in *Test* format. In the table  $n$  denotes the length of the winning streak,  $f$  is the corresponding frequency of occurrence.  $p(n)$  and  $p(n_f)$  are the  $p$ -values for observation of streaks with length  $n$  and streaks with length  $n$  conditional on frequency  $f$ . With the help of multiple hypothesis tests mentioned in the previous section, we check the significance of the  $p$ -values and note down the results. We write result to be True, for the positive result, implying the significant  $p$ -value, otherwise it is False.

Country	$n$	$f$	$p(n)$	$p(n_f)$	Sidak		Holm		Hochberg		Bonferroni		Classic FDR		Storey FDR		Storey's Q			
					$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$
ENG	1	50	1.00	1.00	False	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00	
ENG	2	24	0.89	1.00	False	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00	
ENG	3	26	0.77	0.99	False	False	False	False	False	False	False	False	False	False	False	False	False	0.99	0.99	
ENG	4	13	0.66	0.97	False	False	False	False	False	False	False	False	False	False	False	False	False	0.97	0.97	
ENG	5	5	0.54	0.98	False	False	False	False	False	False	False	False	False	False	False	False	False	0.98	0.98	
ENG	6	2	0.43	0.97	False	False	False	False	False	False	False	False	False	False	False	False	False	0.85	0.97	
ENG	7	8	0.31	0.00	False	<b>True</b>	False	<b>True</b>	False	<b>True</b>	False	<b>True</b>	False	<b>True</b>	False	<b>True</b>	False	<b>True</b>	0.09	0.00
ENG	8	1	0.21	0.78	False	False	False	False	False	False	False	False	False	False	False	False	False	0.42	0.78	
ENG	9	1	0.13	0.56	False	False	False	False	False	False	False	False	False	False	False	False	False	0.26	0.56	
ENG	10	1	0.08	0.38	False	False	False	False	False	False	False	False	False	False	False	<b>True</b>	False	0.03	0.07	
ENG	11	1	0.05	0.22	False	False	False	False	False	False	False	False	False	False	False	False	False	0.09	0.22	
ENG	12	1	0.02	0.12	<b>True</b>	False	False	False	<b>True</b>	False	<b>True</b>	False	<b>True</b>	False	<b>True</b>	False	<b>True</b>	False	0.05	0.12

Table B.15 Winning streak statistics for team England in *Test* format. In the table  $n$  denotes the length of the winning streak,  $f$  is the corresponding frequency of occurrence.  $p(n)$  and  $p(n_f)$  are the  $p$ -values for observation of streaks with length  $n$  and streaks with length  $n$  conditional on frequency  $f$ . With the help of multiple hypothesis tests mentioned in the previous section, we check the significance of the  $p$ -values and note down the results. We write result to be True, for the positive result, implying the significant  $p$ -value, otherwise it is False.

Country	$n$	$f$	$p(n)$	$p(n_f)$	Sidak		Holm		Hochberg		Bonferroni		Classic FDR		Storey FDR		Storey's Q		
					$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$
IND	1	30	1.00	1.00	False	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
IND	2	16	0.84	1.00	False	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
IND	3	6	0.69	1.00	False	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
IND	4	5	0.53	0.87	False	False	False	False	False	False	False	False	False	False	False	False	False	0.87	0.87
IND	5	1	0.38	0.98	False	False	False	False	False	False	False	False	False	False	False	False	False	0.76	0.98
IND	6	2	0.24	0.48	False	False	False	False	False	False	False	False	False	False	False	False	False	0.17	0.17
IND	8	2	0.07	0.04	False	False	False	False	False	False	False	False	False	False	False	False	False	0.07	0.07
IND	9	1	0.03	0.15	False	False	False	False	False	False	False	False	False	False	False	False	False	0.07	0.15
IND	15	1	0.00	0.00	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	0.00	0.00

Table B.16 Winning streak statistics for team India in *Test* format. In the table  $n$  denotes the length of the winning streak,  $f$  is the corresponding frequency of occurrence.  $p(n)$  and  $p(n_f)$  are the  $p$ -values for observation of streaks with length  $n$  and streaks with length  $n$  conditional on frequency  $f$ . With the help of multiple hypothesis tests mentioned in the previous section, we check the significance of the  $p$ -values and note down the results. We write result to be True, for the positive result, implying the significant  $p$ -value, otherwise it is False.

Country	$n$	$f$	$p(n)$	$p(n_f)$	Sidak		Holm		Hochberg		Bonferroni		Classic FDR		Storey FDR		Storey's Q	
					$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$
NZL	1	29	1.00	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
NZL	2	15	0.79	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
NZL	3	7	0.58	0.95	False	False	False	False	False	False	False	False	False	False	False	False	0.95	0.95
NZL	4	4	0.38	0.34	False	False	False	False	False	False	False	False	False	False	False	False	0.25	0.25
NZL	5	1	0.19	0.69	False	False	False	False	False	False	False	False	False	False	False	False	0.38	0.69

Table B.17 Winning streak statistics for team New Zealand in *Test* format. In the table  $n$  denotes the length of the winning streak,  $f$  is the corresponding frequency of occurrence.  $p(n)$  and  $p(n_f)$  are the  $p$ -values for observation of streaks with length  $n$  and streaks with length  $n$  conditional on frequency  $f$ . With the help of multiple hypothesis tests mentioned in the previous section, we check the significance of the  $p$ -values and note down the results. We write result to be True, for the positive result, implying the significant  $p$ -value, otherwise it is False.

Country	$n$	$f$	$p(n)$	$p(n_f)$	Sidak		Holm		Hochberg		Bonferroni		Classic FDR		Storey FDR		Storey's Q	
					$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$
PAK	1	28	1.00	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
PAK	2	10	0.85	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
PAK	3	9	0.69	0.95	False	False	False	False	False	False	False	False	False	False	False	False	0.95	0.95
PAK	4	7	0.54	0.38	False	False	False	False	False	False	False	False	False	False	False	False	0.21	0.21
PAK	5	3	0.39	0.60	False	False	False	False	False	False	False	False	False	False	False	False	0.08	0.08
PAK	6	2	0.25	0.49	False	False	False	False	False	False	False	False	False	False	False	False	0.20	0.20
PAK	8	1	0.08	0.33	False	False	False	False	False	False	False	False	False	False	<b>True</b>	<b>True</b>	0.02	0.05

Table B.18 Winning streak statistics for team Pakistan in *Test* format. In the table  $n$  denotes the length of the winning streak,  $f$  is the corresponding frequency of occurrence.  $p(n)$  and  $p(n_f)$  are the  $p$ -values for observation of streaks with length  $n$  and streaks with length  $n$  conditional on frequency  $f$ . With the help of multiple hypothesis tests mentioned in the previous section, we check the significance of the  $p$ -values and note down the results. We write result to be True, for the positive result, implying the significant  $p$ -value, otherwise it is False.

Country	$n$	$f$	$p(n)$	$p(n_f)$	Sidak		Holm		Hochberg		Bonferroni		Classic FDR		Storey FDR		Storey's Q	
					$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$
SAF	1	22	1.00	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
SAF	2	13	0.86	1.00	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
SAF	3	8	0.71	0.99	False	False	False	False	False	False	False	False	False	False	False	False	0.99	0.99
SAF	4	7	0.57	0.66	False	False	False	False	False	False	False	False	False	False	False	False	0.66	0.66
SAF	5	2	0.43	0.92	False	False	False	False	False	False	False	False	False	False	False	False	0.85	0.92
SAF	6	3	0.29	0.33	False	False	False	False	False	False	False	False	False	False	False	False	0.17	0.17
SAF	7	1	0.18	0.64	False	False	False	False	False	False	False	False	False	False	False	False	0.36	0.64
SAF	10	3	0.03	0.00	False	<b>True</b>	False	<b>True</b>	<b>True</b>	<b>True</b>	False	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	0.03	0.00

Table B.19 Winning streak statistics for team South Africa in *Test* format. In the table  $n$  denotes the length of the winning streak,  $f$  is the corresponding frequency of occurrence.  $p(n)$  and  $p(n_f)$  are the  $p$ -values for observation of streaks with length  $n$  and streaks with length  $n$  conditional on frequency  $f$ . With the help of multiple hypothesis tests mentioned in the previous section, we check the significance of the  $p$ -values and note down the results. We write result to be True, for the positive result, implying the significant  $p$ -value, otherwise it is False.

Country	$n$	$f$	$p(n)$	$p(n_f)$	Sidak		Holm		Hochberg		Bonferroni		Classic FDR		Storey FDR		Storey's Q		
					$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$
SRL	1	17	1.00	1.00	False	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
SRL	2	15	0.82	1.00	False	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
SRL	3	4	0.63	1.00	False	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
SRL	4	3	0.45	0.79	False	False	False	False	False	False	False	False	False	False	False	False	False	0.79	0.79
SRL	5	1	0.28	0.84	False	False	False	False	False	False	False	False	False	False	False	False	False	0.55	0.84
SRL	7	1	0.07	0.29	False	False	False	False	False	False	False	False	False	False	False	<b>True</b>	<b>True</b>	0.01	0.02
SRL	9	1	0.01	0.06	<b>True</b>	False	<b>True</b>	False	<b>True</b>	False	<b>True</b>	False	<b>True</b>	False	<b>True</b>	False	<b>True</b>	0.03	0.06

Table B.20 Winning streak statistics for team Sri Lanka in *Test* format. In the table  $n$  denotes the length of the winning streak,  $f$  is the corresponding frequency of occurrence.  $p(n)$  and  $p(n_f)$  are the  $p$ -values for observation of streaks with length  $n$  and streaks with length  $n$  conditional on frequency  $f$ . With the help of multiple hypothesis tests mentioned in the previous section, we check the significance of the  $p$ -values and note down the results. We write result to be True, for the positive result, implying the significant  $p$ -value, otherwise it is False.

Country	$n$	$f$	$p(n)$	$p(n_f)$	Sidak		Holm		Hochberg		Bonferroni		Classic FDR		Storey FDR		Storey's Q		
					$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	
WIN	1	34	1.00	1.00	False	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
WIN	2	16	0.85	1.00	False	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
WIN	3	12	0.69	1.00	False	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.00
WIN	4	4	0.54	0.98	False	False	False	False	False	False	False	False	False	False	False	False	False	0.98	0.98
WIN	5	1	0.38	0.99	False	False	False	False	False	False	False	False	False	False	False	False	False	0.76	0.99
WIN	6	2	0.24	0.49	False	False	False	False	False	False	False	False	False	False	False	False	False	0.17	0.17
WIN	7	2	0.13	0.15	False	False	False	False	False	False	False	False	False	False	False	False	False	0.15	0.15
WIN	8	1	0.07	0.29	False	False	False	False	False	False	False	False	False	False	False	<b>True</b>	<b>True</b>	0.01	0.02
WIN	17	1	0.00	0.00	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	<b>True</b>	0.00	0.00

Table B.21 Winning streak statistics for team West Indies in *Test* format. In the table  $n$  denotes the length of the winning streak,  $f$  is the corresponding frequency of occurrence.  $p(n)$  and  $p(n_f)$  are the  $p$ -values for observation of streaks with length  $n$  and streaks with length  $n$  conditional on frequency  $f$ . With the help of multiple hypothesis tests mentioned in the previous section, we check the significance of the  $p$ -values and note down the results. We write result to be True, for the positive result, implying the significant  $p$ -value, otherwise it is False.

Country	$n$	$f$	$p(n)$	$p(n_f)$	Sidak		Holm		Hochberg		Bonferroni		Classic FDR		Storey FDR		Storey's Q		
					$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	$p(n)$	$p(n_f)$	
ZIM	1	8	1.00	1.0	False	False	False	False	False	False	False	False	False	False	False	False	False	1.00	1.0
ZIM	2	2	0.49	1.0	False	False	False	False	False	False	False	False	False	False	False	False	False	0.99	1.0

Table B.22 Winning streak statistics for team Zimbabwe in *Test* format. In the table  $n$  denotes the length of the winning streak,  $f$  is the corresponding frequency of occurrence.  $p(n)$  and  $p(n_f)$  are the  $p$ -values for observation of streaks with length  $n$  and streaks with length  $n$  conditional on frequency  $f$ . With the help of multiple hypothesis tests mentioned in the previous section, we check the significance of the  $p$ -values and note down the results. We write result to be True, for the positive result, implying the significant  $p$ -value, otherwise it is False.

# Appendix C

## Supplementary material for “Is the Earth crust operating at a critical point?”

### C.1 Extended Expectation-Maximization Algorithm for joint estimation of ETAS parameters and space varying background rate

We start with an initial guess of the independence probability,  $IP_i$ , for all the earthquakes in the primary catalog. This initial guess can be created by simply drawing a uniform random number between 0 and 1 corresponding to each earthquake. We also make initial guesses of the triggering parameters  $\{K, a, c, \omega, \tau, d, \gamma, \rho\}$  as well as of the smoothing parameters  $(D, Q)$  for the background kernel. E step: Using the current value of the parameters  $(K, a, c, \omega, \tau, d, \gamma, \rho, \eta, D, Q)$  and the IP's, we define the probability that the  $i^{th}$  earthquake triggered the  $j^{th}$  earthquake as:

$$P_{ij} = \frac{g(t_j - t_i, x_j - x_i, y_j - y_i, m_i)}{\mu(x_j, y_j) + \sum_{i: t_i < t_j} g(t_j - t_i, x_j - x_i, y_j - y_i, m_i)} \quad (C.1)$$

where  $\mu(x_j, y_j)$  is given by a modification of expression eq. (4.5) obtained by removing the term  $i=j$  in the sum in order to avoid over-influencing the determination of the background rate by the specific location of the earthquakes in the catalog (in a way similar to the leave-one-out validation scheme of [147] p. 127). The new estimates of the independence probabilities can be obtained as  $IP_j = 1 - \sum_i P_{ij}$ . M step 1: Using the current estimate of the  $IP_i$ , we obtain

the value of the PDF of the locations of background earthquakes at the location of the  $j^{\text{th}}$  background earthquake as:

$$\mu_{PDF}(x_j, y_j) = \frac{\sum_{i \neq j} IP_i \pi^{-1} Q D^{2Q} \left( (x_j - x_i)^2 + (y_j - y_i)^2 + D^2 \right)^{-1-Q}}{\sum_{i \neq j} IP_i} \quad (\text{C.2})$$

Using  $\mu_{PDF}(x_j, y_j)$  and  $IP_j$ , we can define the complete data loglikelihood for the spatial distribution of the background earthquakes as:

$$LL_{bkg} = \sum_{j=1}^N IP_j \times \ln \mu_{PDF}(x_j, y_j) \quad (\text{C.3})$$

$LL_{bkg}$  can be maximized with respect to the parameters  $D$  and  $Q$  to obtain their new estimates. To ensure consistency with the data quality, the minimum value of  $D$  is set to the average location error. We also remark that the expression of  $\mu_{PDF}(x_j, y_j)$  omits the term  $i = j$  in the sum in the r.h.s of eq. (4.5). Otherwise, the optimization of  $LL_{bkg}$  leads to  $D$  being very close to 0 and to very large  $Q$  values, creating a Dirac function at the location of the  $j^{\text{th}}$  earthquake, thus leading to the maximal possible value of  $LL_{bkg}$ . This singularity is avoided by using all earthquakes except the  $j^{\text{th}}$  earthquake to explain the background rate at its location [166]. M step 2: Maximize the loglikelihood function  $LL_{trig}$  for the triggering part of the seismicity rate, given by the sum in the r.h.s. of eq. (4.1), with respect to the parameters  $\{K, a, c, \omega, \tau, d, \gamma, \rho\}$ . Update the current estimates of  $(K, a, c, \omega, \tau, d, \gamma, \rho, D, Q)$  to the new estimates obtained in M step 1 and M step 2. Repeat steps E step, M step 1, and M step 2 until convergence. The latter requires that the loglikelihood increase is smaller than 1 (a sufficient criterion as we shall only consider datasets with thousands of samples in the primary catalog) and that the relative variation of the sum of inverted kernel parameters is less than 0.01.

## C.2 The three data sets and their magnitude of completeness

Locations of 600,000 earthquakes ( $M \geq 3$ ) for the entire globe between 1975-2020, as reported in the ANSS catalog, are shown in a of figs. C.1 to C.3 show the location of 1.2 million and 600,000 earthquakes ( $M \geq 1$ ) between 1975-2020 in the study region surrounding California and New Zealand, as reported by the ANSS and Geonet catalogs respectively. Some important metadata related to these catalogs is reported in table C.1.

figs. C.1 to C.3(b) show the time series of the cumulative number of earthquakes (black curve) corresponding to earthquakes shown in panel (a). All these curves show significant convexity, which expresses the increase in reporting rate of earthquakes at the chosen magnitude threshold. In essence, this convexity indicates that the global, Californian, and New Zealand catalogs are not complete above  $M \geq 3$ ,  $M \geq 1$ , and  $M \geq 1$ , respectively. To decide an appropriate completeness level for each of these catalogs, we estimate how the b-value changes with different choices of  $M_c$ . Before this, we bin the magnitude reported in the catalogs at 0.1 units. With these binned magnitudes, we then compute the b-values for different choices of  $M_c$  using the formula proposed by [167]. We find that the estimated b-value first increases and then stabilizes. The stability of the b-value with change in chosen  $M_c$  has been proposed as a proxy for the magnitude above which the catalog is complete. Based on the stability of the b-value for each catalog, we conservatively choose a magnitude of completeness of 5, 3, and 4, respectively, for the global, Californian, and New Zealand catalogs. In panel d of each of the three figs. C.1 to C.3, we show the fit of the theoretical Gutenberg-Richter (GR) law, with parameters estimated using magnitudes above corresponding  $M_c$ , to the empirical magnitude distribution. Due to an appropriate choice of  $M_c$ , the theoretical GR law seems to describe the observed magnitude distribution above  $M_c$  quite well. As a last check of the appropriateness of the estimated  $M_c$ , we again plot the time series of the cumulative number of earthquakes above  $M_c$  in panel (b) in figs. C.1 to C.3 using the orange curve. We see that the convexity present in the black curve now disappears in the orange curve, further indicating that our choice of  $M_c$  for the three catalogs is appropriate.

### C.3 Setup of pseudo-prospective experiments conducted on the three regions

Starting on January 1, 1990, we perform 368, 368, and 371 pseudo prospective experiments for the three catalogs. In these experiments, the duration of the testing period is always 30 days long, and all the testing periods are non-overlapping. Only the data before the beginning of the testing periods are used to calibrate the models. Each model simulates 400,000 catalogs for each of the testing periods. The forecasts are made as stochastic event sets as this allows us to express the full PDF of the forecasts and not just the mean (see [23] for a discussion on the importance of the PDF). Since the parametric form of this dressed PDF is not known, it is expressed by sampling many times from the underlying ETAS model, justifying the need for many stochastic catalogs. The stochastic catalogs are then used to construct the forecast of the models at any spatial resolution and magnitude threshold during the testing

Catalog Source	Region	Region Details	Auxiliary Period	Primary Period	$N_{evt} (Primary) \geq M_c$	$N_{evt} (Auxiliary) \geq M_c$	$M_c (Primary)$	b-value
ANSS	Globe	Entire Globe	1975-1980	1981- 24 March 2020	65,200	9,609	5	1.06
ANSS	California	Collection Polygon [168]	1975-1980	1981- 24 March 2020	27,413	5,610	3	1.04
Geonet	New Zealand	Centre: [174.3, -41] Radius: 278 km	1975-1980	1981- 23 May 2020	12,875	1,041	4	1.14

Table C.1 Main characteristics of the catalog used in this study.

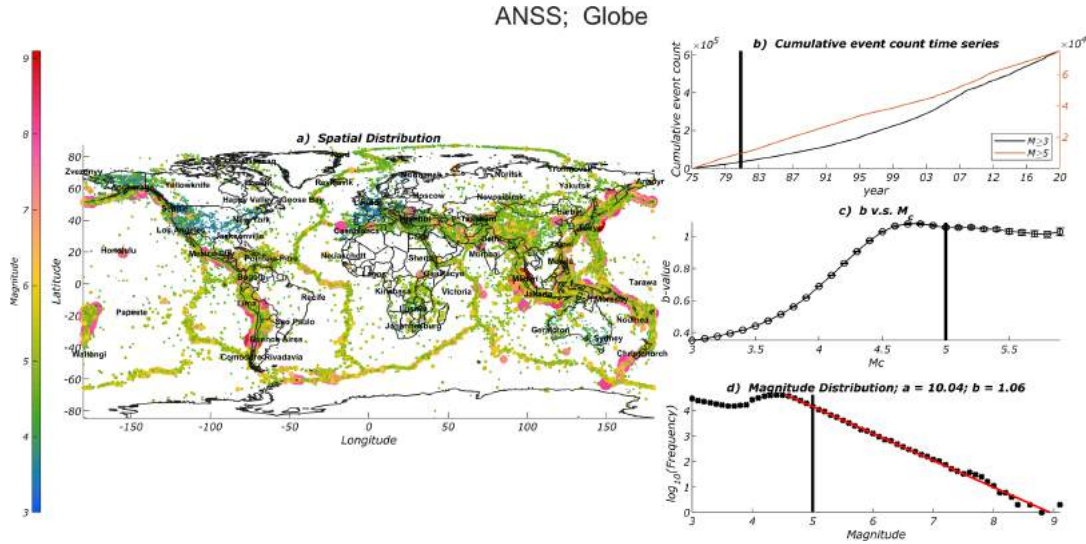


Figure C.1 a) Spatial distribution of 600,000,  $M \geq 3$  earthquakes, distributed over the entire globe and reported in the ANSS catalog; size and colors of the earthquakes scales with their magnitude according to the color code on the left; (b) time series of the cumulative number of  $M \geq 3$  (black) and  $M \geq 5$  (orange) earthquakes since 1975; (c) b-value as a function of the assumed magnitude of completeness  $M_c$ ; solid black line indicates the chosen magnitude of completeness ( $M_c = 5$ ) (d) Empirical magnitude distribution (black circles) and best fit to a Gutenberg-Richter law (red line) obtained using  $M \geq M_c$ .

periods. Given the size of the region and the computational complexity that follows, the global catalog is evaluated at nine spatial resolutions:  $389 \text{ km}^2$ ,  $4 * 389 \text{ km}^2$ ,  $4^2 * 389 \text{ km}^2 \dots$ ,  $4^8 * 389 \text{ km}^2$  and six magnitude thresholds: 5, 5.2, 5.4, ..., 6. For the forecast experiments, the Californian catalog is evaluated at five spatial resolutions:  $4^{-1} * 389 \text{ km}^2$ ,  $389 \text{ km}^2$ ,  $4 * 389 \text{ km}^2$ ,  $4^2 * 389 \text{ km}^2$ ,  $4^3 * 389 \text{ km}^2$  and six magnitude thresholds: 3, 3.2, 3.4, ..., 4. The New Zealand catalog is evaluated at five spatial resolutions:  $4^{-1} * 389 \text{ km}^2$ ,  $389 \text{ km}^2$ ,  $4 * 389 \text{ km}^2$ ,  $4^2 * 389 \text{ km}^2$ ,  $4^3 * 389 \text{ km}^2$  and six magnitude thresholds: 4, 4.2, 4.4, ..., 5. For more details on the testing setup, we refer the readers to [146].



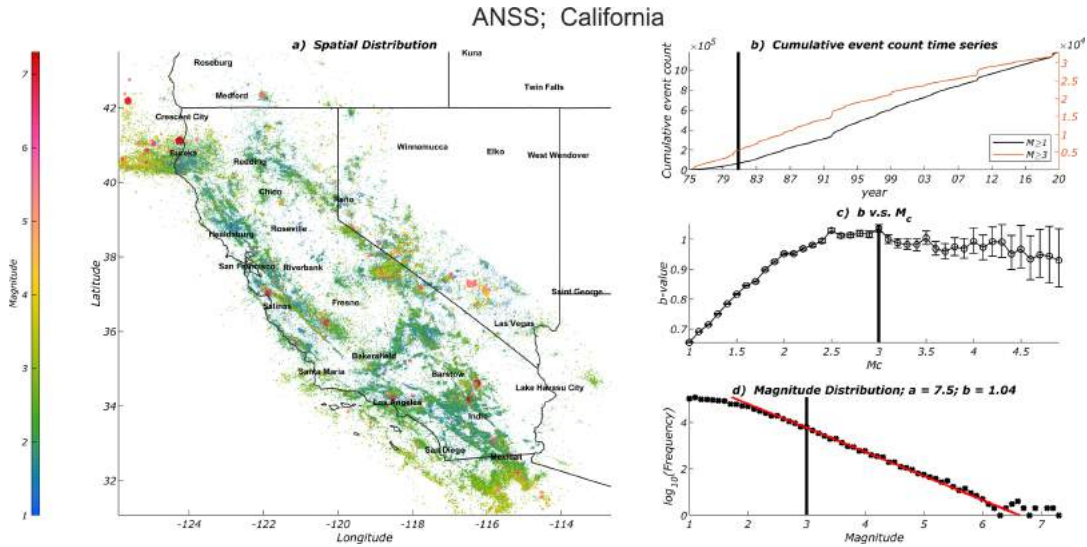


Figure C.2 (a) Spatial distribution of 1.1 million,  $M \geq 1$  earthquakes, distributed around California and reported in the ANSS catalog; size and colors of the earthquakes scales with their magnitude according to the color code on the left; (b) time series of the cumulative number of  $M \geq 1$  (black) and  $M \geq 3$  (orange) earthquakes since 1975; (c) b-value as a function of the assumed magnitude of completeness  $M_c$ ; solid black line indicates the chosen magnitude of completeness ( $M_c = 3$ ) (d) Empirical magnitude distribution (black circles) and best fit to a Gutenberg-Richter law (red line) obtained using  $M \geq M_c$ .

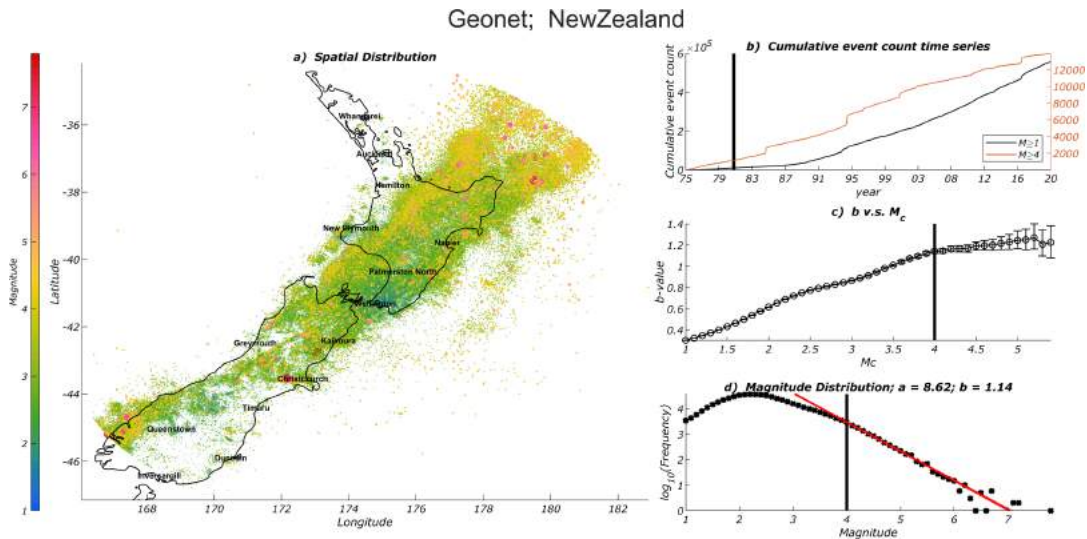


Figure C.3 a) Spatial distribution of 600,000,  $M \geq 1$  earthquakes, distributed around New Zealand and reported in the GeoNet catalog; size and colors of the earthquakes scales with their magnitude according to the color code on the left; (b) time series of the cumulative number of  $M \geq 1$  (black) and  $M \geq 4$  (orange) earthquakes since 1975; (c) b-value as a function of the assumed magnitude of completeness  $M_c$ ; solid black line indicates the chosen magnitude of completeness ( $M_c = 4$ ) (d) Empirical magnitude distribution (black circles) and best fit to a Gutenberg-Richter law (red line) obtained using  $M \geq M_c$ .

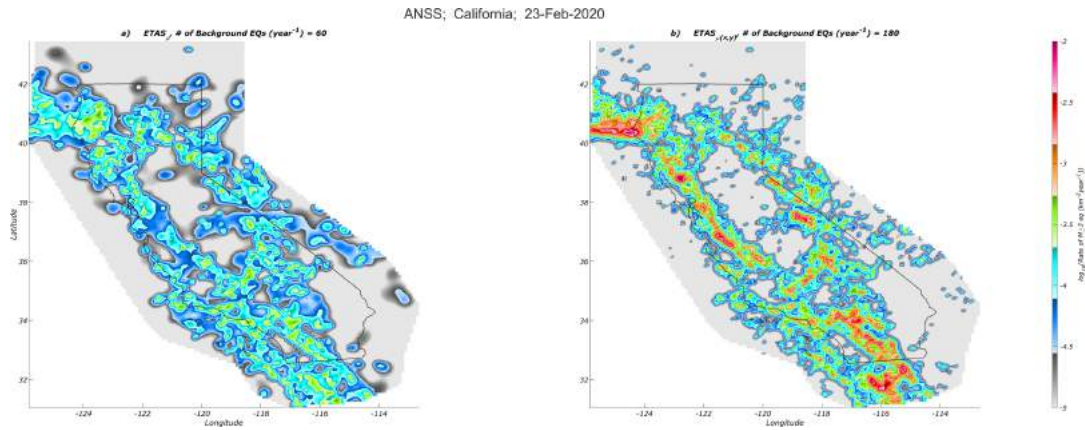


Figure C.4 Spatial density of the earthquakes identified as background by the  $ETAS_{\mu}$  and  $ETAS_{\mu(x,y)}$  models for the Californian catalog. The title indicates the date up to which earthquakes in the catalog were used to calibrate the two models and obtain these maps.

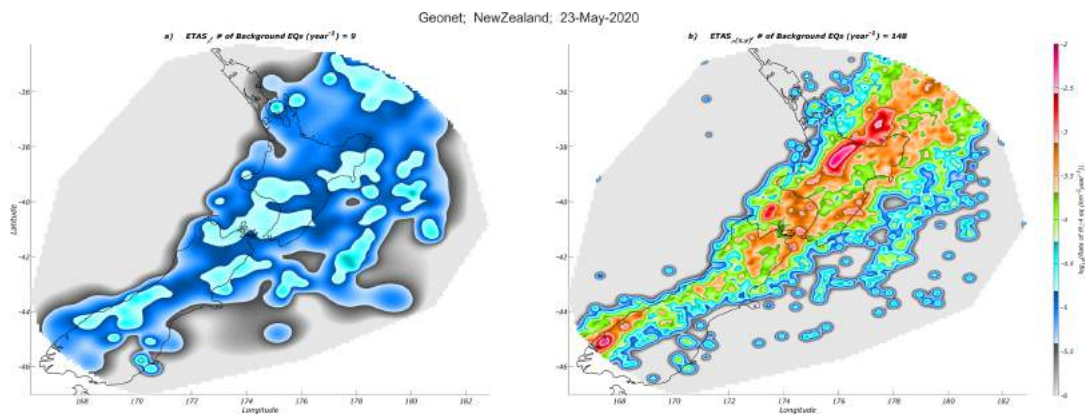


Figure C.5 Spatial density of the earthquakes identified as background by the  $ETAS_{\mu}$  and  $ETAS_{\mu(x,y)}$  models for the New Zealand catalog. The title indicates the date up to which earthquakes in the catalog were used to calibrate the two models and obtain these maps.

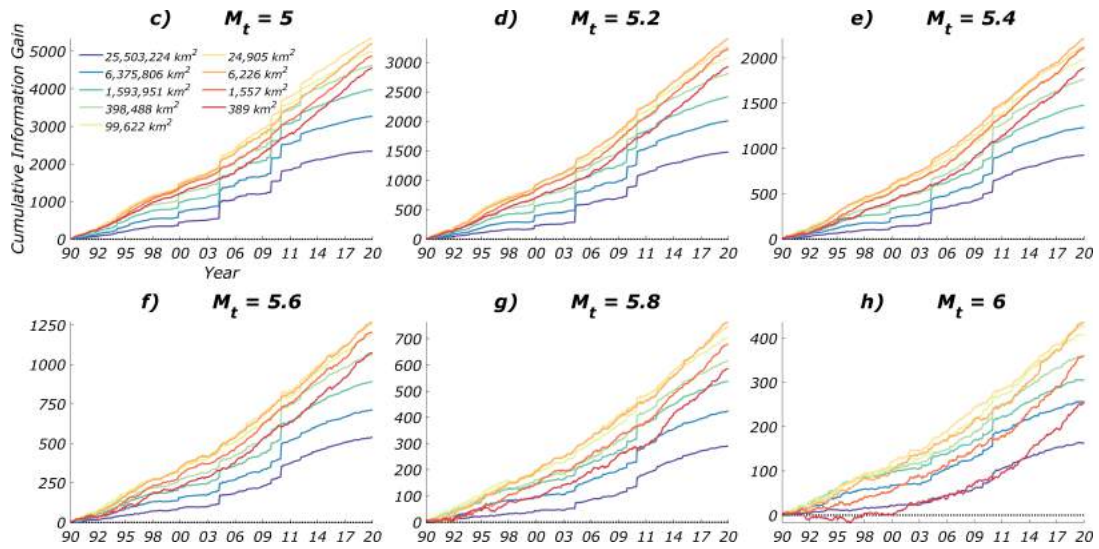


Figure C.6 : Time series of cumulative information gain of  $ETAS_{\mu(x,y)}$  over the  $ETAS_{\mu}$  model in 368 pseudo prospective experiments for the Global ANSS catalog, at different spatial resolutions (different colors whose meaning is given in the inset) and magnitude threshold (different panels) of the testing catalog

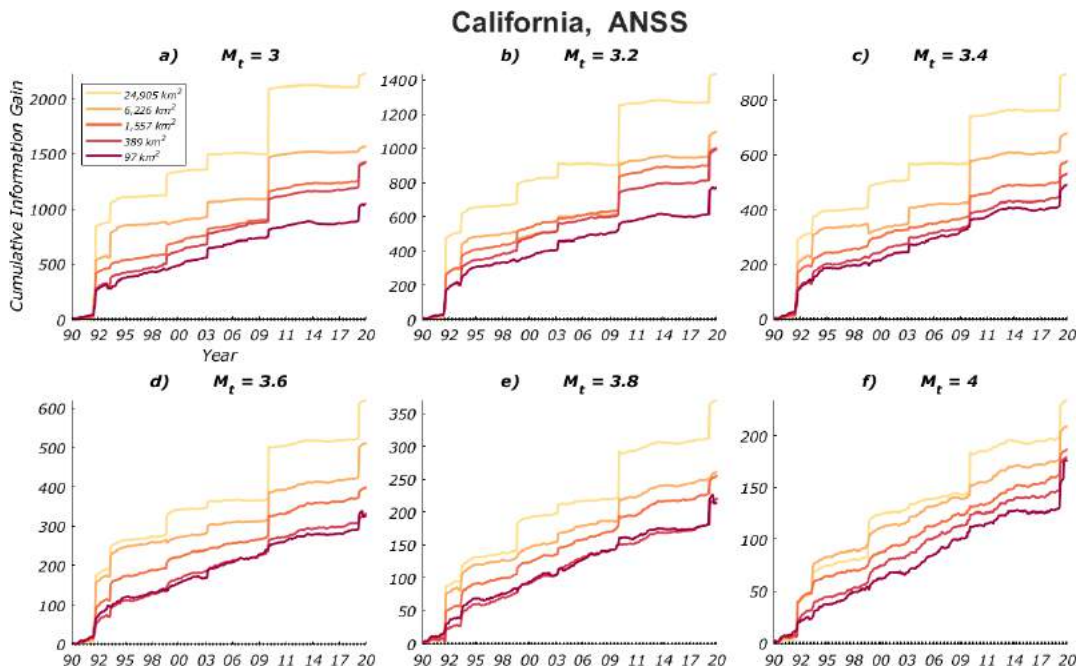


Figure C.7 Time series of cumulative information gain of  $ETAS_{\mu(x,y)}$  over the  $ETAS_{\mu}$  model in 368 pseudo prospective experiments for the Californian ANSS catalog, at different spatial resolutions (different colors whose meaning is given in the inset) and magnitude threshold (different panels) of the testing catalog

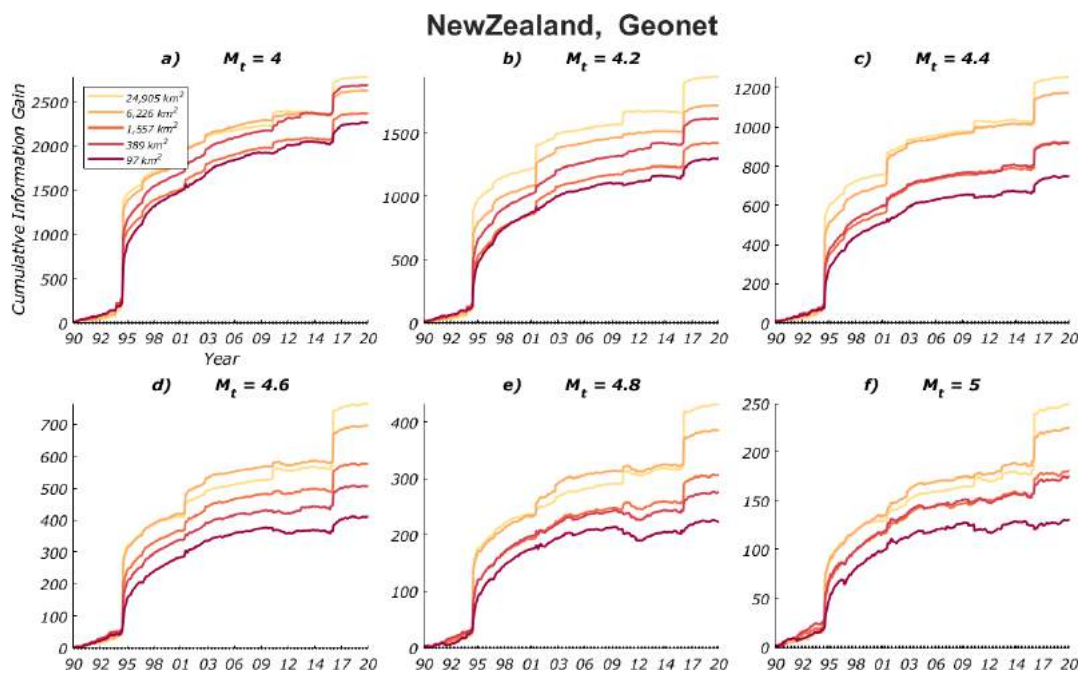


Figure C.8 Time series of cumulative information gain of  $ETAS_{\mu(x,y)}$  over the  $ETAS_{\mu}$  model in 371 pseudo prospective experiments for the New Zealand catalog, at different spatial resolutions (different colors whose meaning is given in the inset) and magnitude threshold (different panels) of the testing catalog.



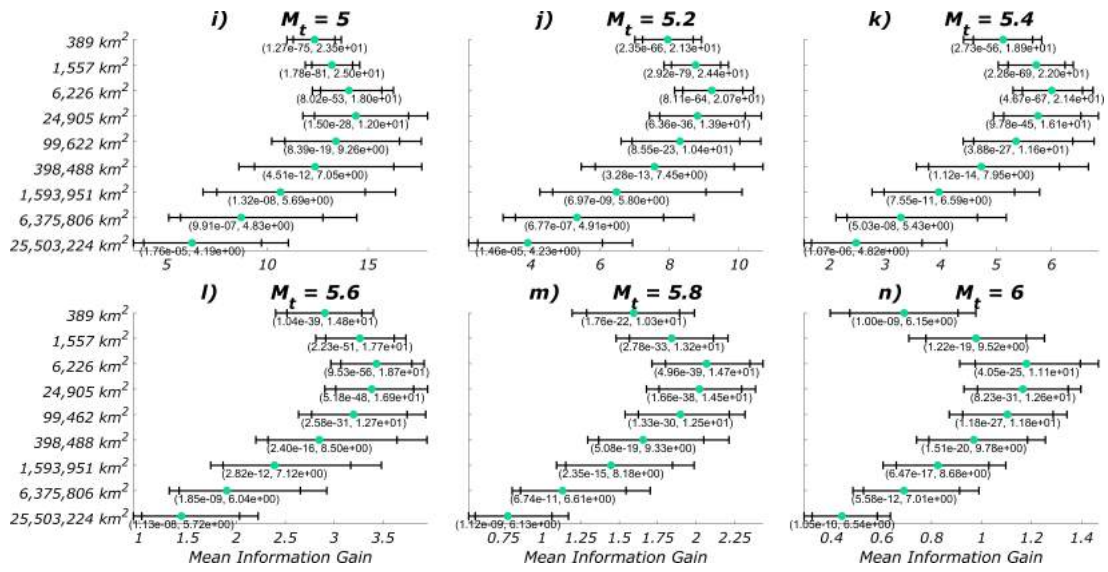


Figure C.9 Mean information gain (MIG) of  $ETAS_{\mu}(x,y)$  over the  $ETAS_{\mu}$  model in 368 pseudo prospective experiments for the Global ANSS catalog, at different spatial resolutions and magnitude threshold (different panels) of the testing catalog. The error bars indicate the 99 and 95% confidence interval of the mean information gain (MIG). The two numbers indicate the p-value and t-statistic resulting from the student's t-test, in which we test the null hypothesis that the MIG is equal to 0 against the alternative that it is larger than 0. When the p-value is smaller than 0.05, the null hypothesis is rejected, and the circles are green. The notations  $e - xx$  means  $10^{-xx}$ .

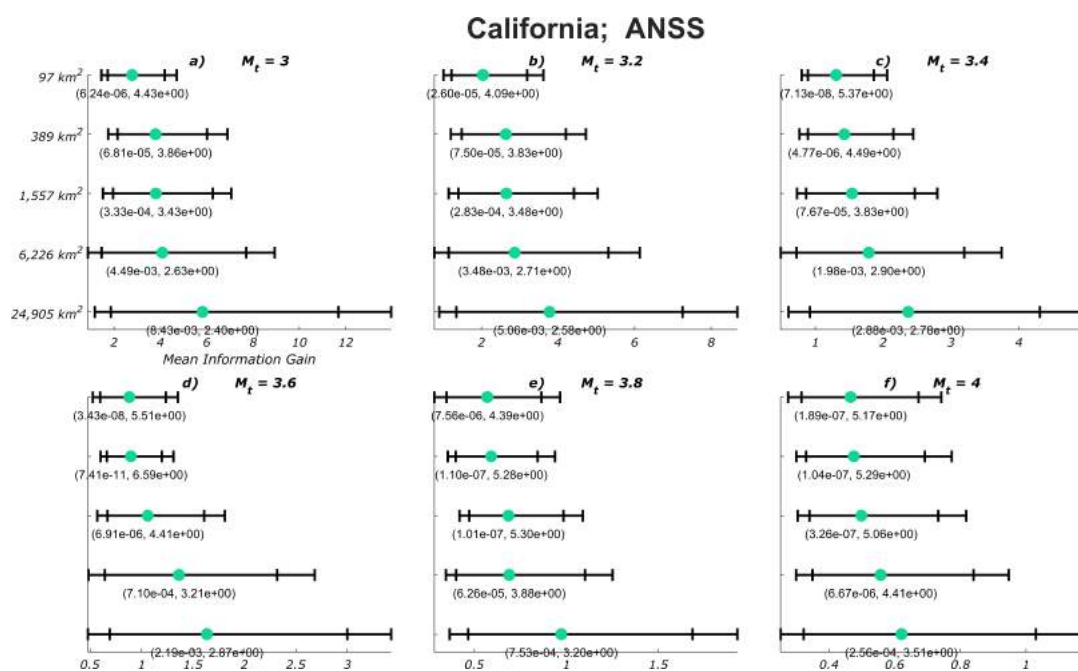


Figure C.10 Mean information gain (MIG) of  $ETAS_{\mu(x,y)}$  over the  $ETAS_{\mu}$  model in 368 pseudo prospective experiments for the Californian catalog, at different spatial resolutions and magnitude threshold (different panels) of the testing catalog. The error bars indicate the 99 and 95% confidence interval of the mean information gain (MIG). The two numbers indicate the p-value and t-statistic resulting from the student's t-test, in which we test the null hypothesis that the MIG is equal to 0 against the alternative that it is larger than 0. When the p-value is smaller than 0.05, the null hypothesis is rejected, and the circles are green. The notations  $e - xx$  means  $10^{-xx}$ .

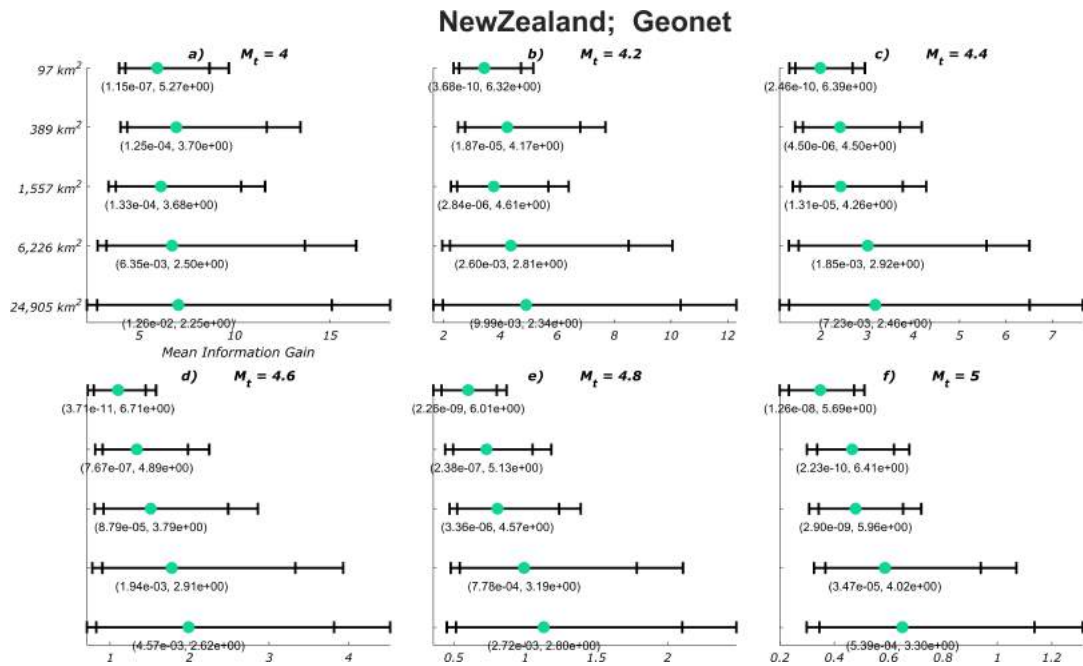


Figure C.11 Mean information gain (MIG) of  $ETAS_{\mu(x,y)}$  over the  $ETAS_{\mu}$  model in 371 pseudo prospective experiments with New Zealand catalog, at different spatial resolutions and magnitude threshold (different panels) of the testing catalog. The error bars indicate the 99 and 95 % confidence interval of the mean information gain (MIG). The two numbers indicate the p-value and t-statistic resulting from the student's t-test, in which we test the null hypothesis that the MIG is equal to 0 against the alternative that it is larger than 0. When the p-value is smaller than 0.05, the null hypothesis is rejected, and the circles are green. The notations  $e - xx$  means  $10^{-xx}$ .





# **Appendix D**

## **Appendix 1**

### **D.1 Supporting figures**

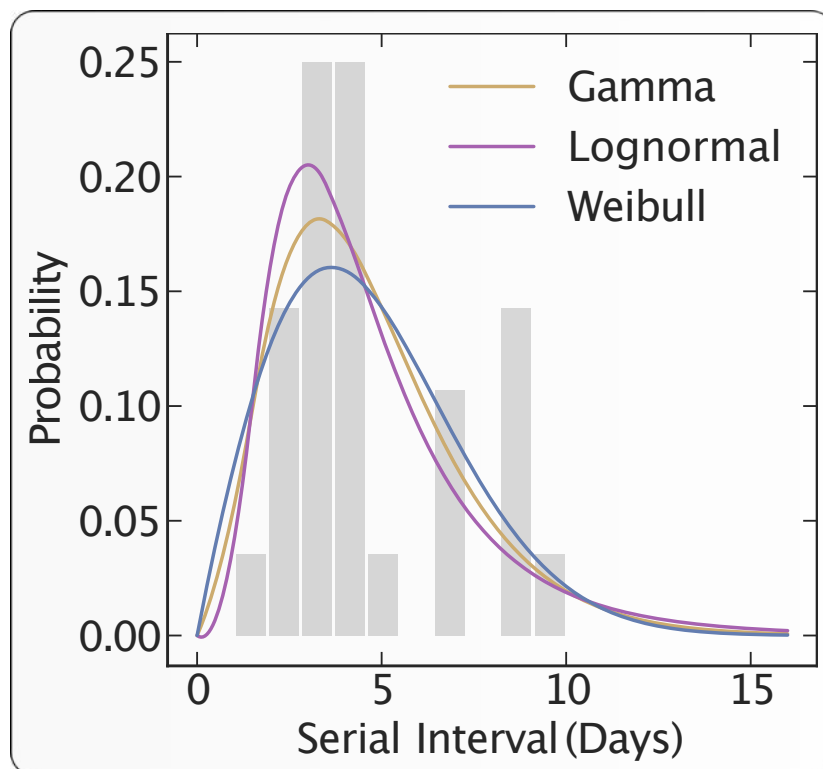


Figure D.1 **Generation time distribution:** The distribution of the serial intervals are calibrated with 3 different models, i.e, Gamma, Log-normal and Weibull and using maximum likelihood estimation. We find the most suitable model for the generation time distribution by comparing the log-likelihood scores of the above models. We observe that the log-normal distribution gives the best fitting with  $\mu = 4.6$  and  $\sigma = 2.8$ . The grey bars represents the empirical observations and the three fitted models are presented as solid lines.

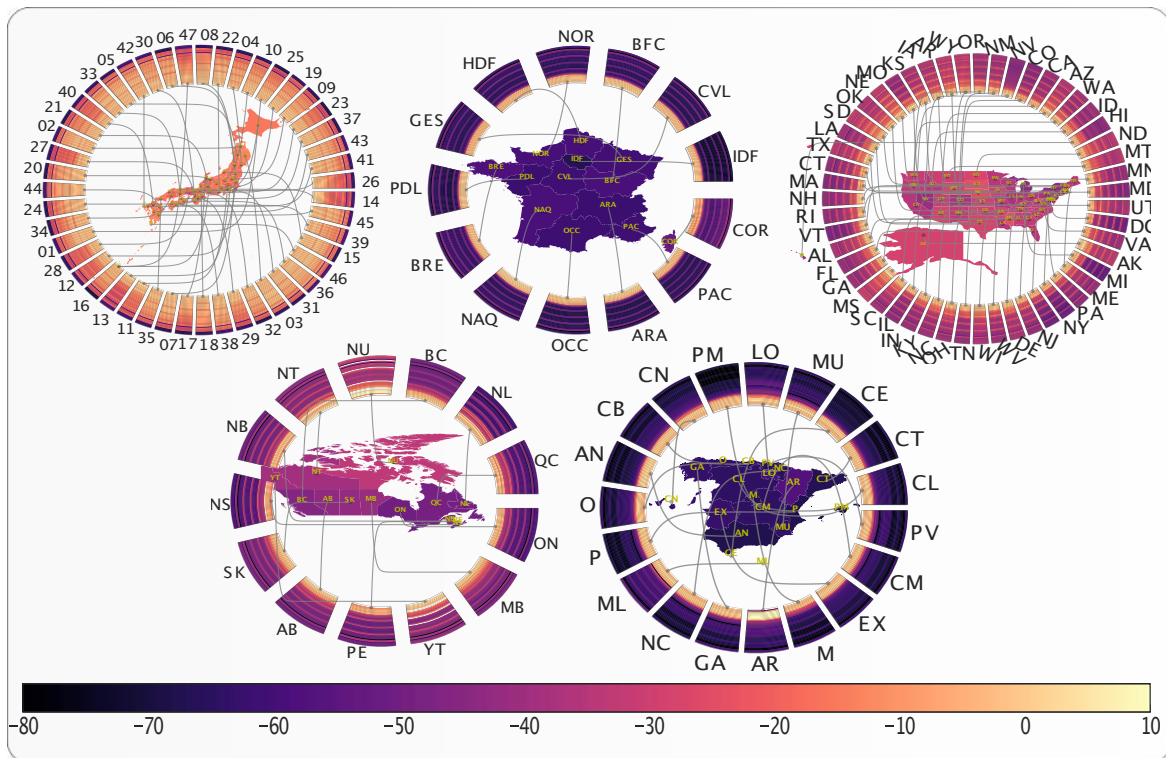
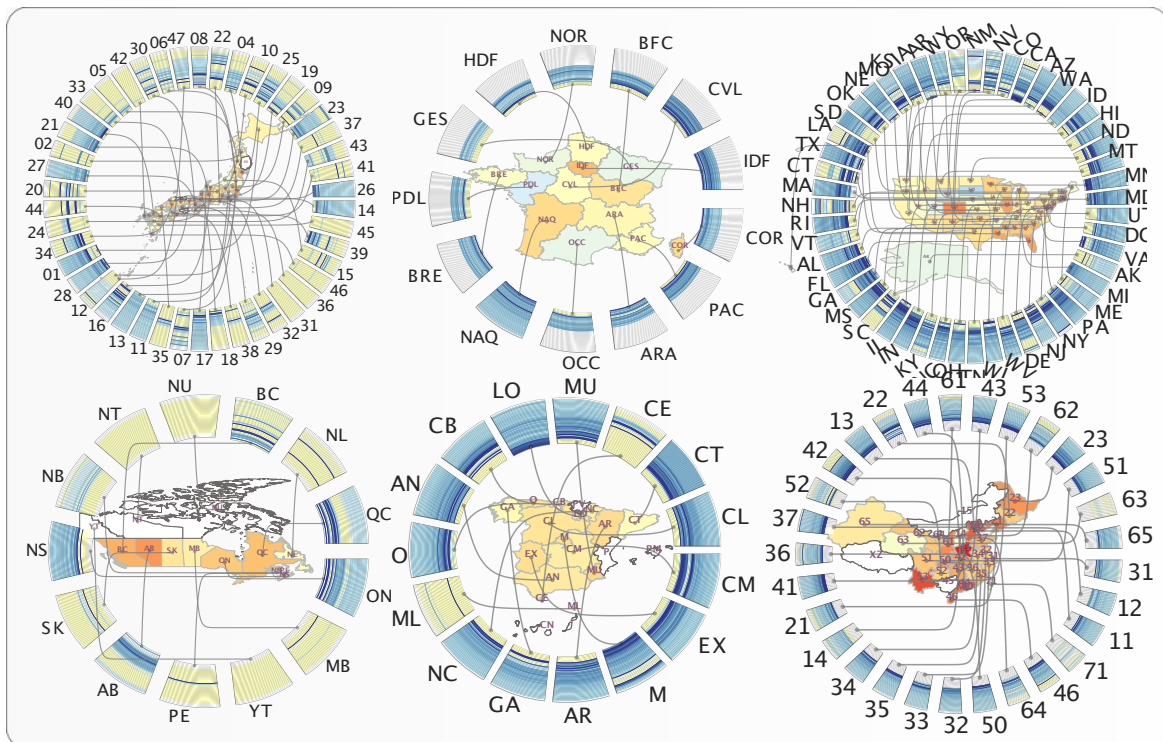
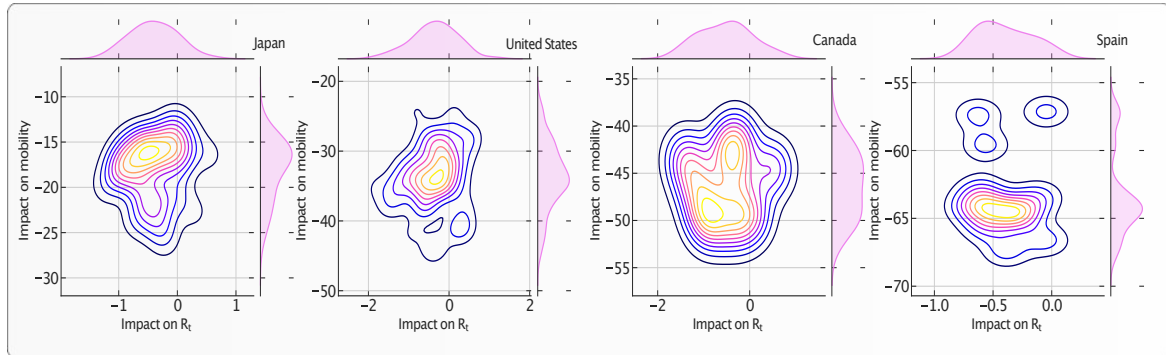


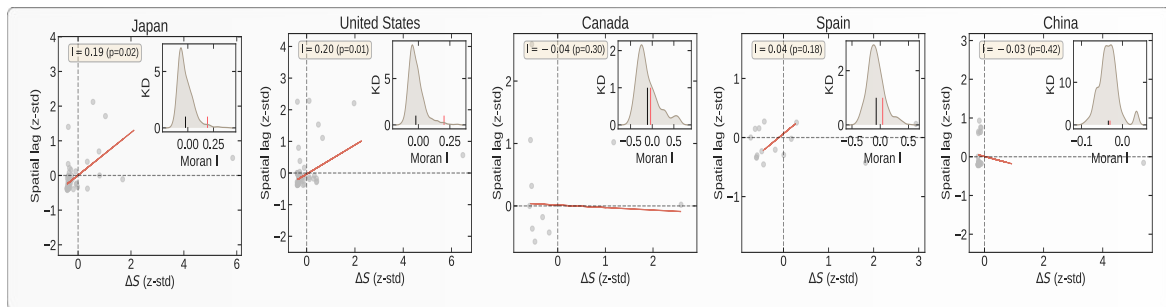
Figure D.2 **Time evolution of mobility and impact of travel restriction on mobility in eight countries.** Each subplot shows the time evolution of mobility and impact of travel restrictions on mobility of different top level administrative divisions of a country. The countries are, for top left to bottom right: Japan, France, USA, Canada, and Spain. The color of the regions on the map denoted by their ISO 3166-2 code represent the impact (increase or decrease of mobility) of travel restriction on that region. The radial connected wedges represent the time evolution of the mobility in the corresponding region or state for each country. The color of the strips in the wedges represent the mobility on a particular day.



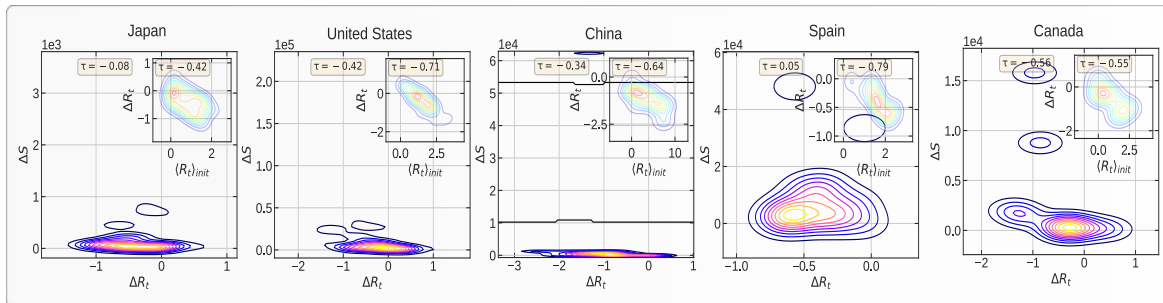
**Figure D.3 Time evolution of effective reproduction number ( $R_t$ ) and absolute impact of travel restrictions on  $R_t$  in ten countries.** Each subplot presents time evolution of  $R_t$  and impact of interventions on  $R_t$  for different top level administrative divisions of the country. The color of the regions (name in ISO 3166-2 code notation) represents the impact (increase or decrease of  $R_t$ ) of interventions on that state. The radial wedges represent the time evolution of  $R_t$  in the corresponding state, and color of the strips represent  $R_t$  on a particular day. The countries are, from top left to bottom right: Japan, France, USA, Canada, Spain and China.



**Figure D.4 Impact of travel restrictions on workplace activity and on epidemic progression ( $R_t$ ) in nine countries.** Each panel represents the bivariate kernel density estimation as a function of the absolute impact on workplace activity and impact on  $R_t$  in the administrative divisions of each country. The bivariate distribution is constructed over the set of regions within each country. The top and right inset of each of the nine plots represent the marginal distribution of the respective variables for each country. The countries are, from left to right: Japan, USA, Canada, and Spain.



**Figure D.5 Spatial auto-correlation of the total increase  $\Delta S$  in number of confirmed cases during the first 30 days of intervention in nine countries.** In each panel, the x-axis corresponds to the value of  $\Delta S$  in a given region in a given country; the y-axis gives the average  $\Delta S$  over the neighboring regions, called “spatial lag” in the caption along the y-axis. These two variables are *z-standardised* for better comparison. The inset in each panel represents the Kernel Density estimator for the distribution of the simulated *Moran's I*. The black vertical line in the inset represents the expected *Moran's I* from simulations with the null hypothesis of no spatial correlations. The red vertical line represents the value obtained from empirical data. *Moran's I* along with its p-value is given in yellow box. The countries are, from left to right: Japan, USA, Canada, Spain and China.



**Figure D.6 Joint distribution of the total number  $\Delta S$  of confirmed cases within the first 30 days of intervention and of the absolute impact of intervention on the effective reproduction number ( $\Delta R_t$ ) within this period.** Each panel represents the Kernel Density Estimation for the total number of confirmed cases within the first 30 days of the intervention against the impact of interventions on effective reproduction number ( $\Delta R_t$ ) in different administrative divisions of a country. The bivariate distribution is constructed over the set of regions within each country. The yellow box contains the Kendall  $\tau$  correlation value for this joint distribution. The inset in each panel represents the variation of  $\Delta R_t$  during the intervention against the average initial  $R_t$  before the intervention. The yellow box gives the corresponding Kendalls  $\tau$ 's. The countries are, from top left to bottom right: Japan, USA, China, Spain, and Canada.

## D.2 Supporting Tables

### D.2.1 Impact of interventions on effective reproduction number and growth of epidemic

Country	Administrative Division	p-value	$\Delta R_t$	$\Delta S$
Canada	Alberta	0	-1.39626	2061.0
Canada	British Columbia	0.027972	-1.16284	1375.0
Canada	<b>New Brunswick</b>	0.000999001	0.491501	115.0
Canada	Ontario	0.011988	-0.855158	8772.0
Canada	Quebec	0.024975	-0.967118	15783.0
Canada	Saskatchewan	0.0589411	-0.4651	303.0

Table D.1 Impact of intervention on  $R_t$  and the total growth of confirmed cases in Canada.

Country	Administrative Division	p-value	$\Delta R_t$	$\Delta S$
China	Anhui Province	0.03996	-0.802842	979.0
China	Beijing	0.032967	-1.26934	382.0
China	Chongqing City	0.021978	-1.56294	558.0
China	Fujian Province	0.000999001	-1.10573	283.0
China	Gansu Province	0.004995	-0.904039	89.0
China	Guangdong Province	0	-1.67239	1301.0
China	Jiangxi Province	0.00599401	-0.729367	241.0
China	Hainan Province	0.011988	-1.50345	163.0
China	Hebei Province	0.021978	-1.91644	306.0
China	Heilongjiang Province	0.00699301	-1.41766	478.0
China	Henan Province	0.013986	-2.48088	1262.0
China	Hubei Province	0	-1.33345	62218.0
China	Hunan Province	0	-1.24172	1002.0
China	Jiangsu Province	0	-0.782538	622.0
China	Jiangxi Province	0.001998	-2.23217	931.0
China	Liaoning Province	0.046953	-1.58059	118.0
China	Ningxia Hui Autonomous Region	0.002997	-0.84003	70.0
China	Shaanxi Province	0.028971	-1.48649	242.0
China	Shandong Province	0	-1.04356	743.0
China	Shanghai	0.02997	-0.619683	318.0
China	Sichuan Province	0	-0.906403	517.0
China	Yunnan Province	0.000999001	-2.06287	172.0
China	Zhejiang Province	0.002997	-0.787457	1176.0

Table D.2 Impact of intervention on  $R_t$  and the total growth of confirmed cases in China.

Country	Administrative Division	p-value	$\Delta R_t$	$\Delta S$
France	Nouvelle-Aquitaine	0.002997	-0.648405	3298.0

Table D.3 Impact of intervention on  $R_t$  and the total growth of confirmed cases in France.



Country	Administrative Division	p-value	$\Delta R_t$	$\Delta S$
Germany	<b>Bremen</b>	0	1.71413	457.0
Germany	Hamburg	0.000999001	-0.942453	3633.0
Germany	Hessen	0.038961	-0.86765	6053.0
Germany	Niedersachsen	0.003996	-0.847555	8027.0
Germany	<b>Saarland</b>	0	0.862704	2090.0
Germany	Sachsen	0.045954	-1.10722	3683.0

Table D.4 Impact of intervention on  $R_t$  and the total growth of confirmed cases in Germany.

Country	Administrative Division	p-value	$\Delta R_t$	$\Delta S$
India	<b>Maharashtra</b>	0.002997	0.650857	6305.0
India	<b>Odisha</b>	0	0.494264	87.0

Table D.5 Impact of intervention on  $R_t$  and the total growth of confirmed cases in India.

Country	Administrative Division	p-value	$\Delta R_t$	$\Delta S$
Italy	Puglia	0.026973	-0.778201	2464.0
Italy	Emilia-Romagna	0	-0.381702	16439.0
Italy	Liguria	0.037962	-0.734625	4648.0
Italy	Marche	0	-1.00811	4387.0

Table D.6 Impact of intervention on  $R_t$  and the total growth of confirmed cases in Italy.

Country	Administrative Division	p-value	$\Delta R_t$	$\Delta S$
Japan	Aichi Ken	0.004995	-0.553495	238.0
Japan	Akita Ken	0.03996	-0.362313	5.0
Japan	<b>Chiba Ken</b>	0.002997	0.297088	572.0
Japan	Fukui Ken	0.000999001	-0.988835	57.0
Japan	Fukuoka Ken	0.034965	-0.573188	451.0
Japan	Gifu Ken	0	-1.02953	81.0
Japan	Hyogo Ken	0.00799201	-0.721471	447.0
Japan	Ibaraki Ken	0.048951	-0.935754	90.0
Japan	Kochi Ken	0.017982	-0.676786	36.0
Japan	Kyoto Fu	0.00799201	-0.416234	196.0
Japan	Miyagi Ken	0	-1.02827	56.0
Japan	Okayama Ken	0.00899101	-0.567645	11.0
Japan	Osaka Fu	0.016983	-0.359852	1217.0
Japan	<b>Saga Ken</b>	0	0.564563	34.0
Japan	Tochigi Ken	0.0579421	-0.567938	34.0
Japan	Tokyo To	0.036963	-0.201202	3611.0
Japan	Yamagata Ken	0.01998	-0.612601	50.0
Japan	Yamaguchi Ken	0.037962	-0.4729	21.0
Japan	Yamanashi Ken	0.00999001	-0.855139	34.0

Table D.7 Impact of intervention on  $R_t$  and the total growth of confirmed cases in Japan.

Country	Administrative Division	p-value	$\Delta R_t$	$\Delta S$
Spain	Asturias	0.048951	-0.571728	1958.0
Spain	La Rioja	0.00899101	-0.562582	3304.0
Spain	Madrid	0.000999001	-0.568361	48108.0

Table D.8 Impact of intervention on  $R_t$  and the total growth of confirmed cases in Spain.

Country	Administrative Division	p-value	$\Delta R_t$	<i>DeltaS</i>
Switzerland	Appenzell Ausserrhoden	0.04995	-2.53811	68.0
Switzerland	Bern	0.016983	-1.3457	1347.0
Switzerland	Fribourg	0	-0.764614	834.0
Switzerland	Genève	0	-0.749412	3902.0
Switzerland	Graubünden	0.0599401	-0.690593	625.0
Switzerland	<b>Luzern</b>	0	0.384942	524.0
Switzerland	Neuchâtel	0.001998	-0.370057	506.0
Switzerland	<b>Solothurn</b>	0	0.292663	282.0
Switzerland	St. Gallen	0.0569431	-2.59314	617.0
Switzerland	Ticino	0.00899101	-1.08246	2544.0
Switzerland	Valais	0.001998	-1.13367	1535.0
Switzerland	Vaud	0.00599401	-1.08239	4125.0
Switzerland	Zürich	0	-1.19565	2739.0

Table D.9 Impact of intervention on  $R_t$  and the total growth of confirmed cases in Switzerland.

Country	Administrative Division	p-value	$\Delta R_t$	$\Delta S$
United States	<b>Alaska</b>	0.0559441	0.44278	303.0
United States	Colorado	0.00599401	-1.53705	8459.0
United States	Connecticut	0.004995	-1.7927	16713.0
United States	Florida	0.004995	-1.27691	23729.0
United States	Georgia	0	-0.887848	16907.0
United States	Louisiana	0.000999001	-1.28319	22771.0
United States	Maryland	0.002997	-0.52522	11465.0
United States	Massachusetts	0	-0.700434	34896.0
United States	Michigan	0	-0.460859	26631.0
United States	Minnesota	0	-0.768463	1982.0
United States	Mississippi	0.035964	-0.80443	3743.0
United States	New Jersey	0.040959	-0.934117	77725.0
United States	Pennsylvania	0.011988	-0.584873	29256.0
United States	South Carolina	0.03996	-1.14461	3871.0
United States	<b>South Dakota</b>	0	1.08317	1400.0
United States	<b>Wyoming</b>	0.020979	0.536284	278.0

Table D.10 Impact of intervention on  $R_t$  and the total growth of confirmed cases in United States.

## D.3 Supplemental figures

### D.3.1 Mobility Impact

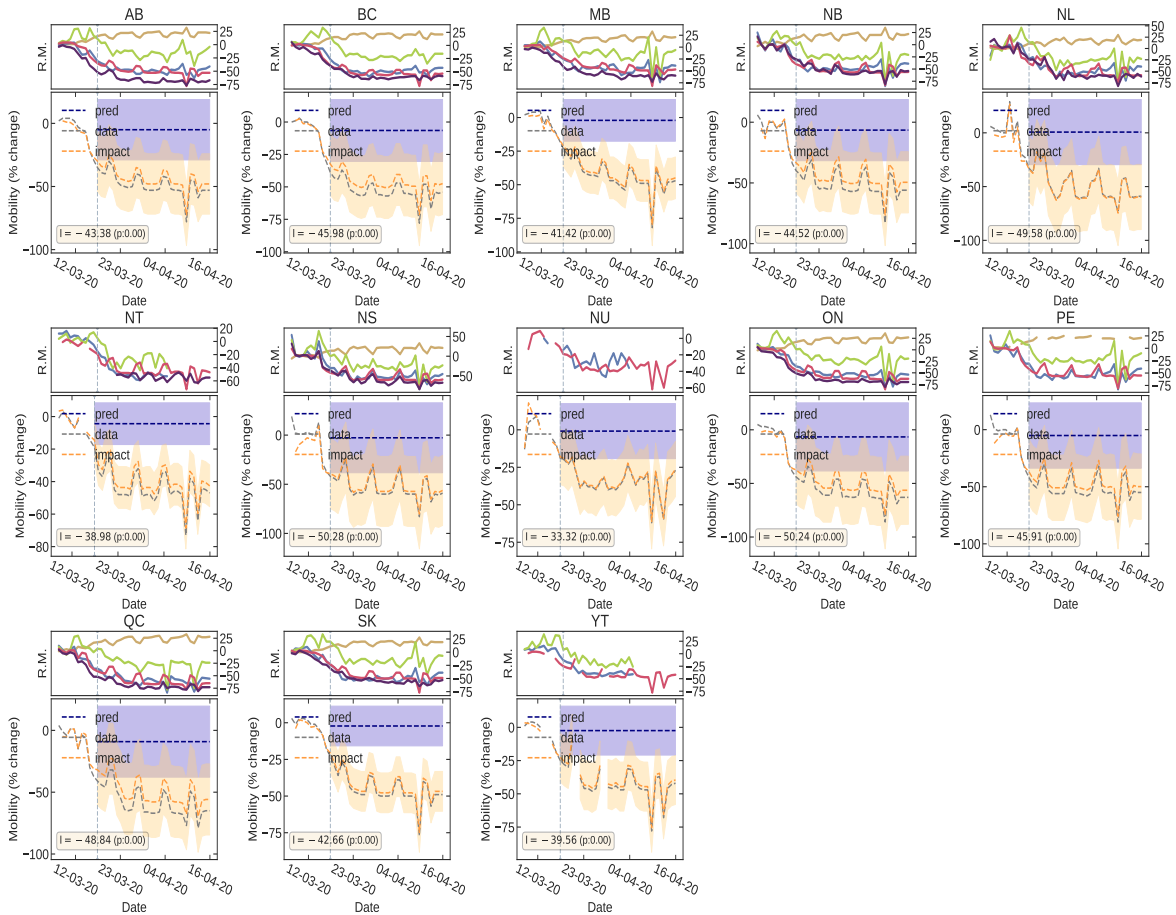


Figure D.7 Impact of Governmental interventions on mobility in different administrative divisions of Canada.

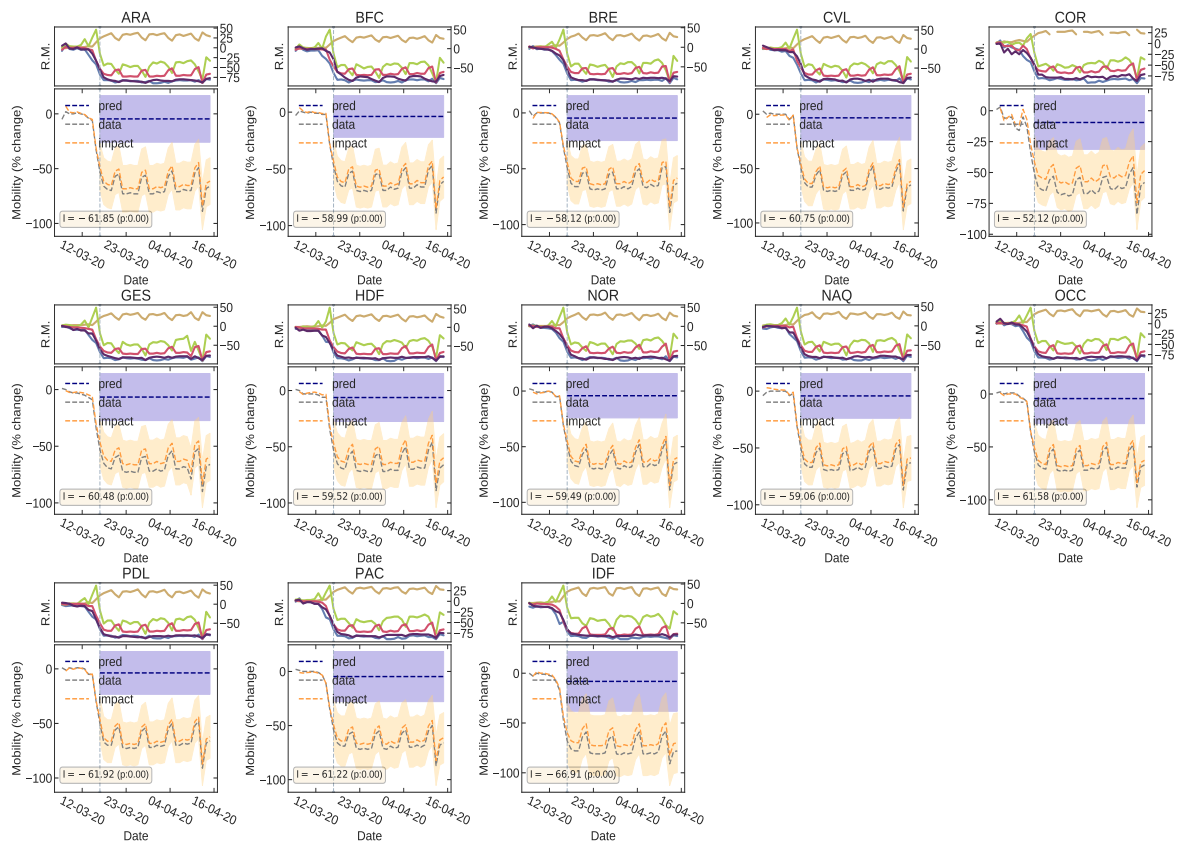


Figure D.8 Impact of Governmental interventions on mobility in different administrative divisions of France.

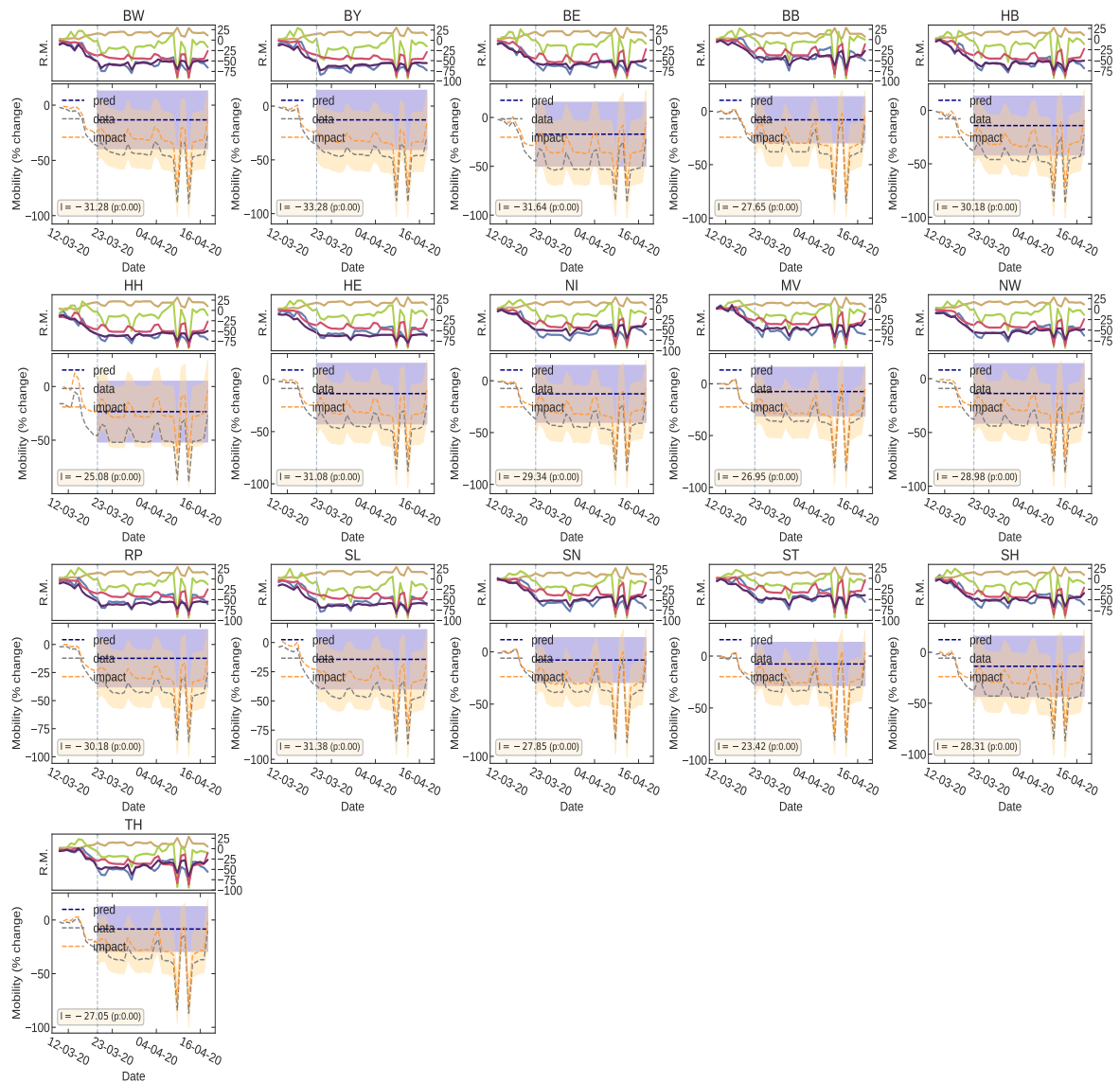


Figure D.9 Impact of Governmental interventions on mobility in different administrative divisions of Germany.

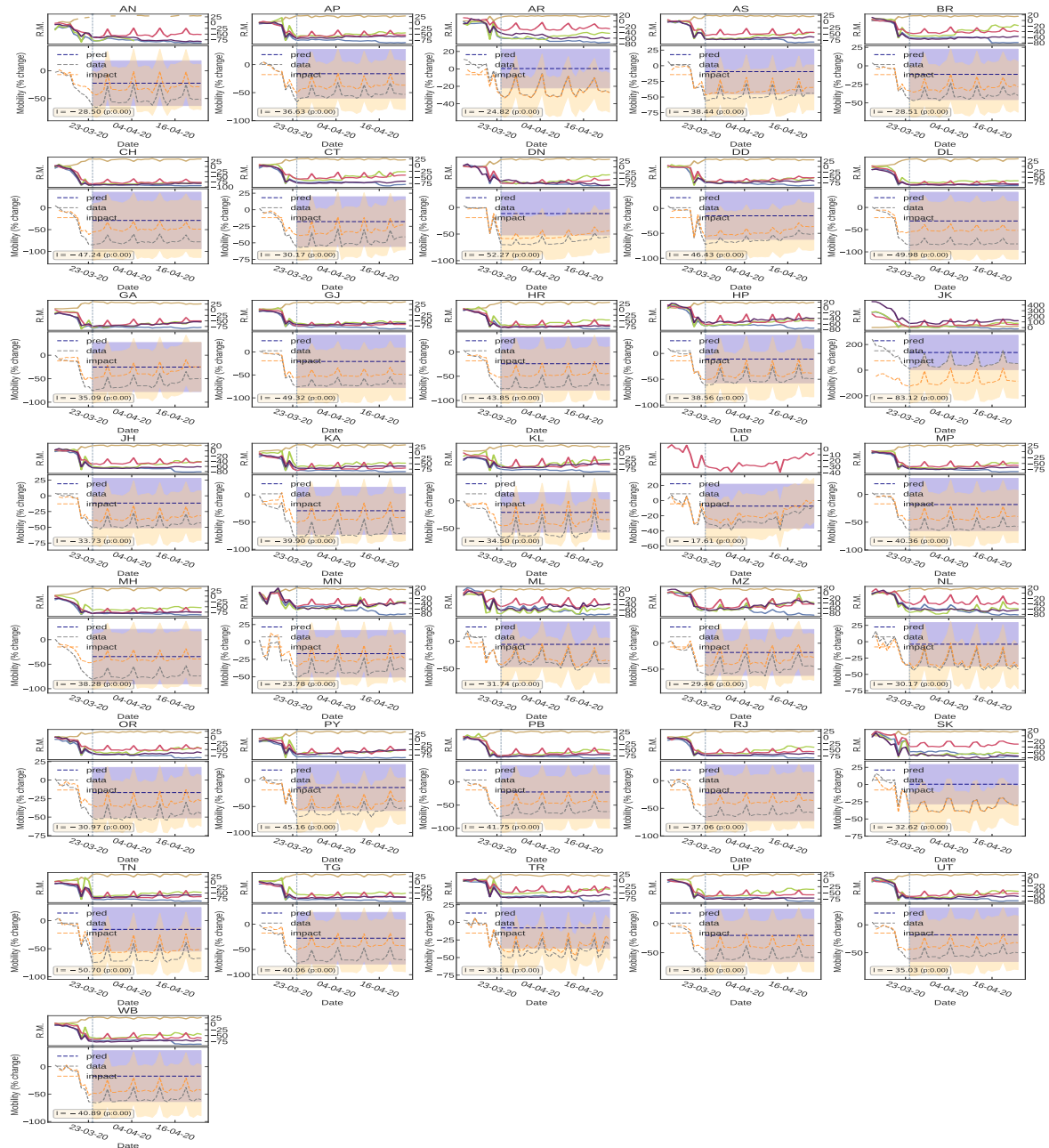


Figure D.10 Impact of Governmental interventions on mobility in different administrative divisions of India.



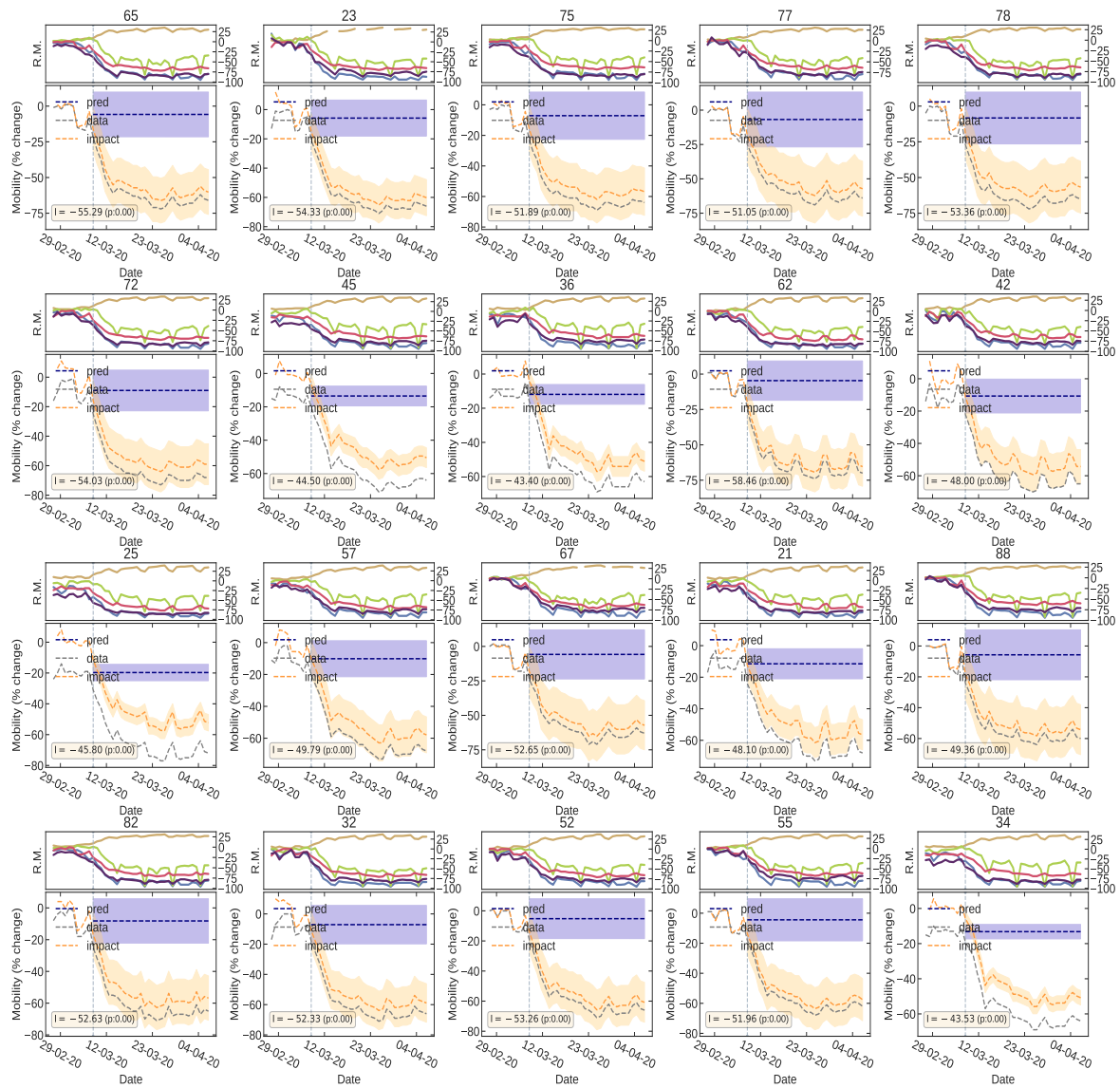


Figure D.11 Impact of Governmental interventions on mobility in different administrative divisions of Italy.

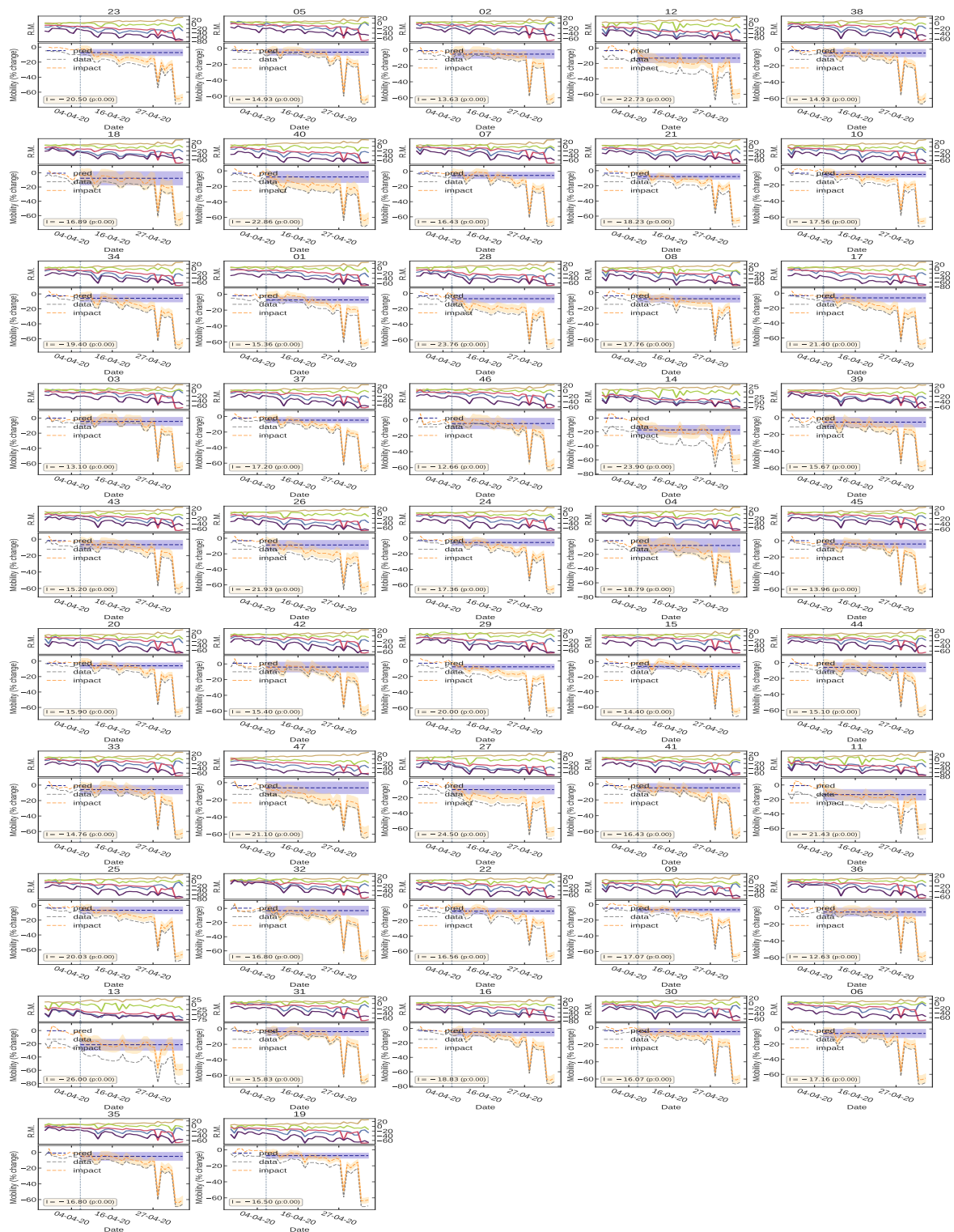


Figure D.12 Impact of Governmental interventions on mobility in different administrative divisions of Japan.

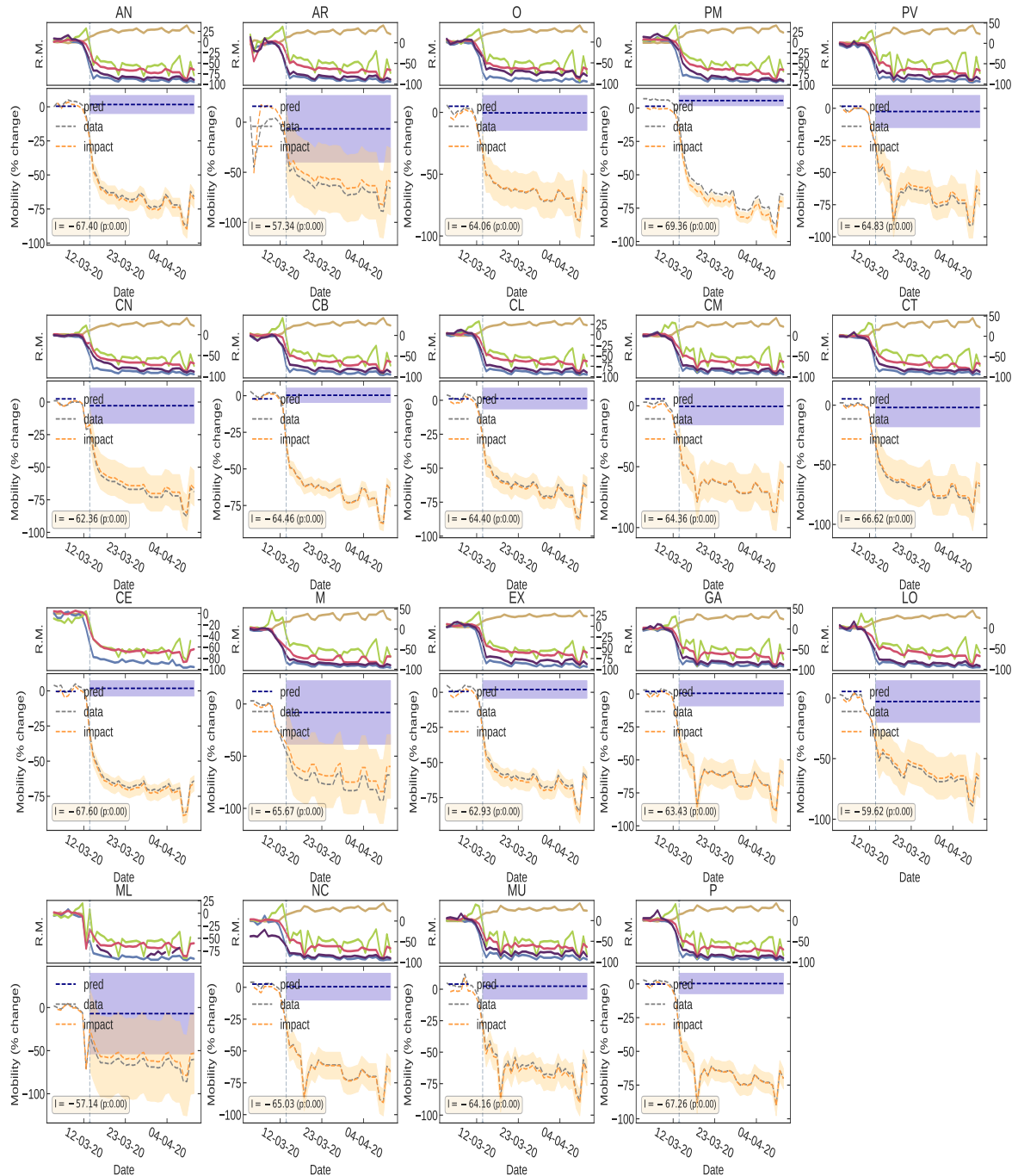


Figure D.13 Impact of Governmental interventions on mobility in different administrative divisions of Spain.

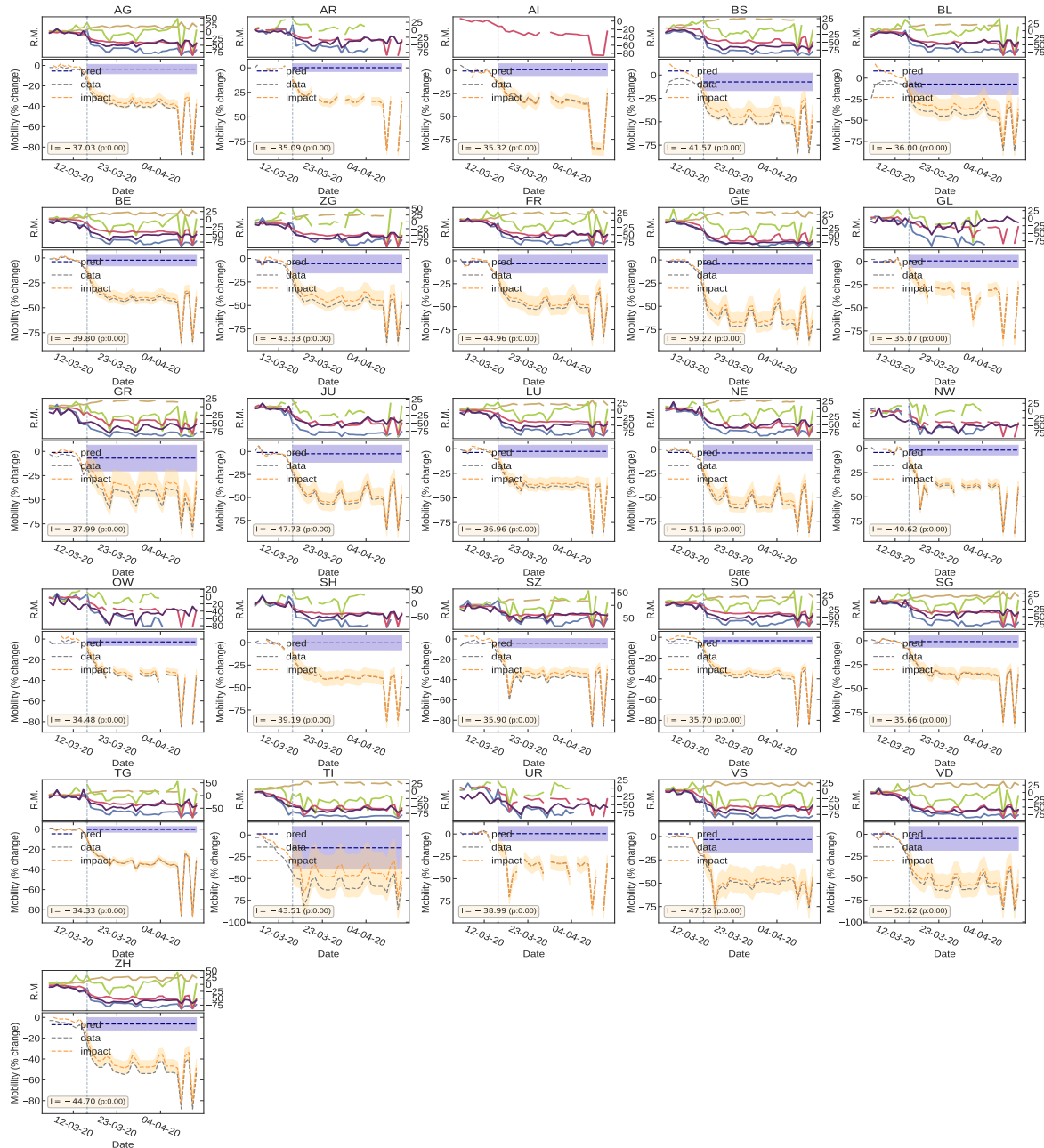


Figure D.14 Impact of Governmental interventions on mobility in different administrative divisions of Switzerland.



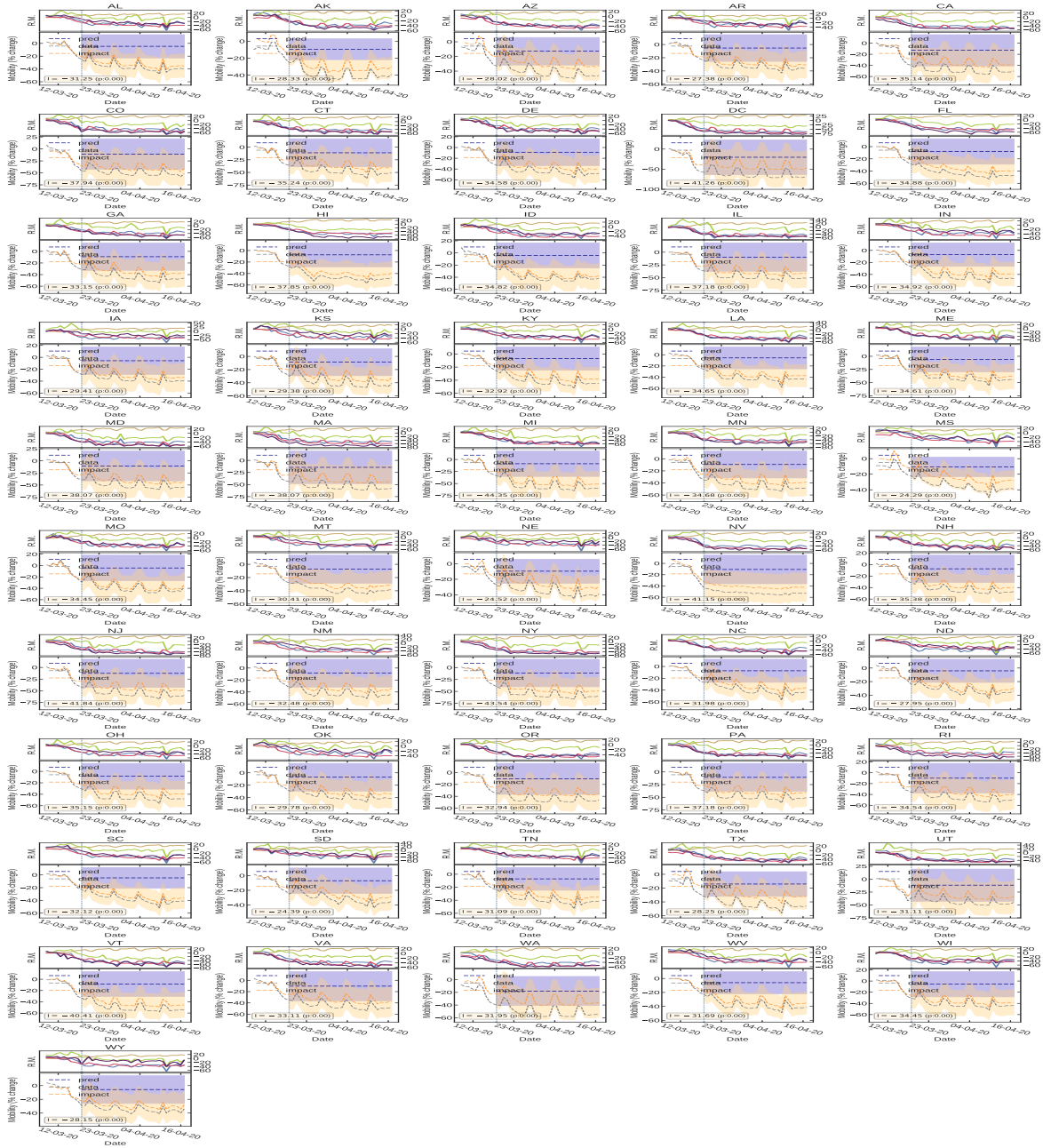


Figure D.15 Impact of Governmental interventions on mobility in different administrative divisions of United States.

### D.3.2 Epidemic Impact

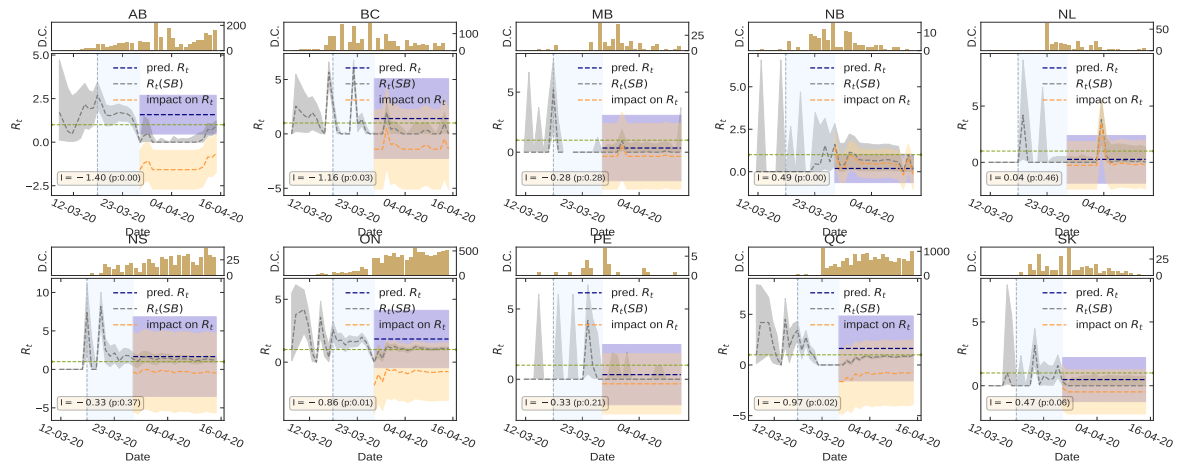


Figure D.16 Impact of Governmental interventions on  $R_t$  in different administrative divisions of Canada.

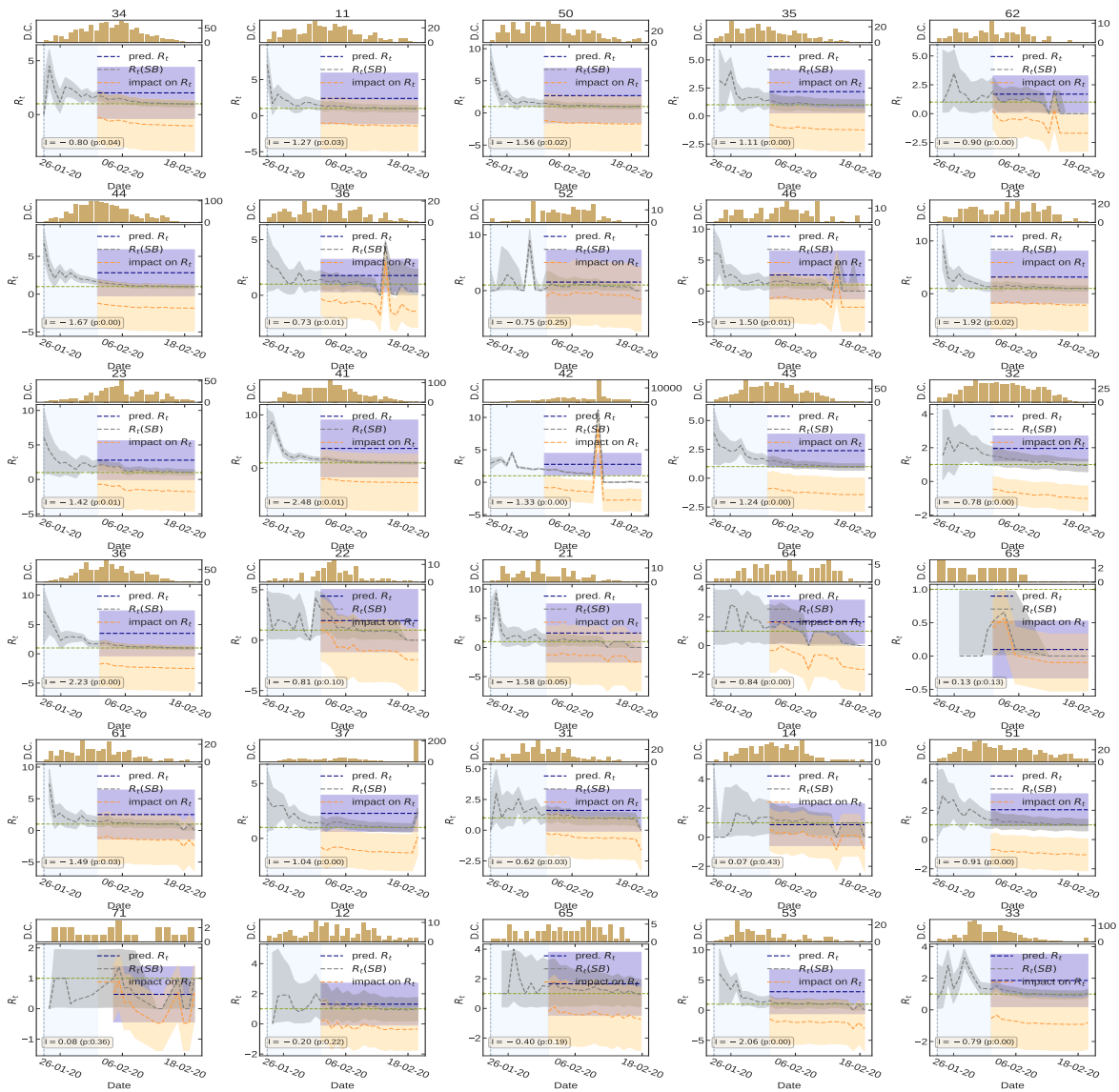


Figure D.17 Impact of Governmental interventions on  $R_t$  in different administrative divisions of China.

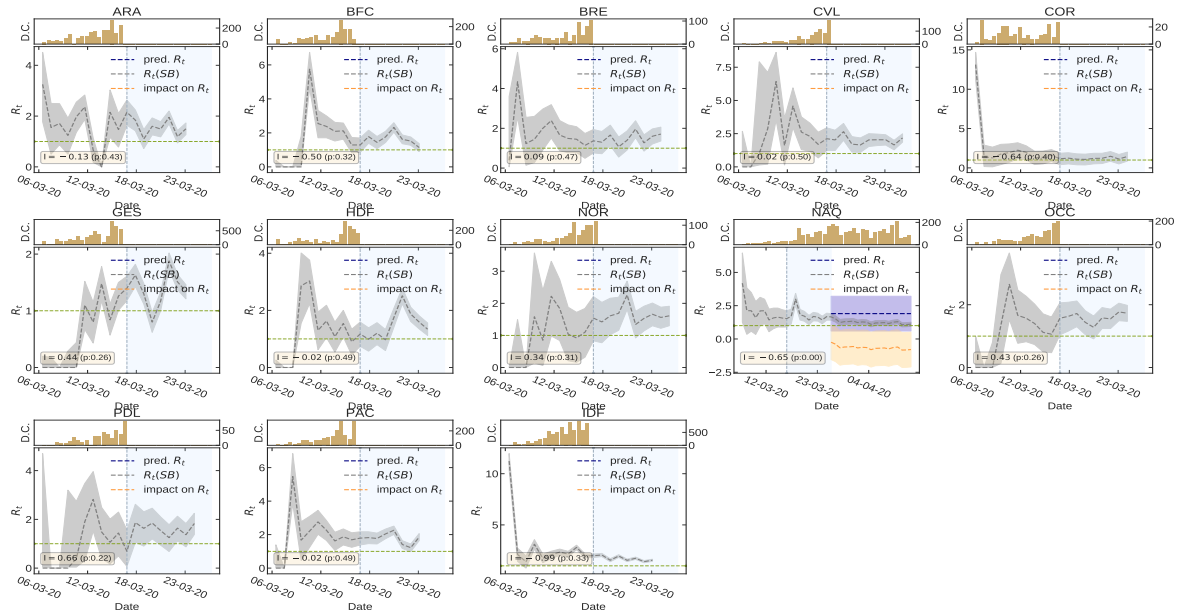


Figure D.18 Impact of Governmental interventions on  $R_t$  in different administrative divisions of France.

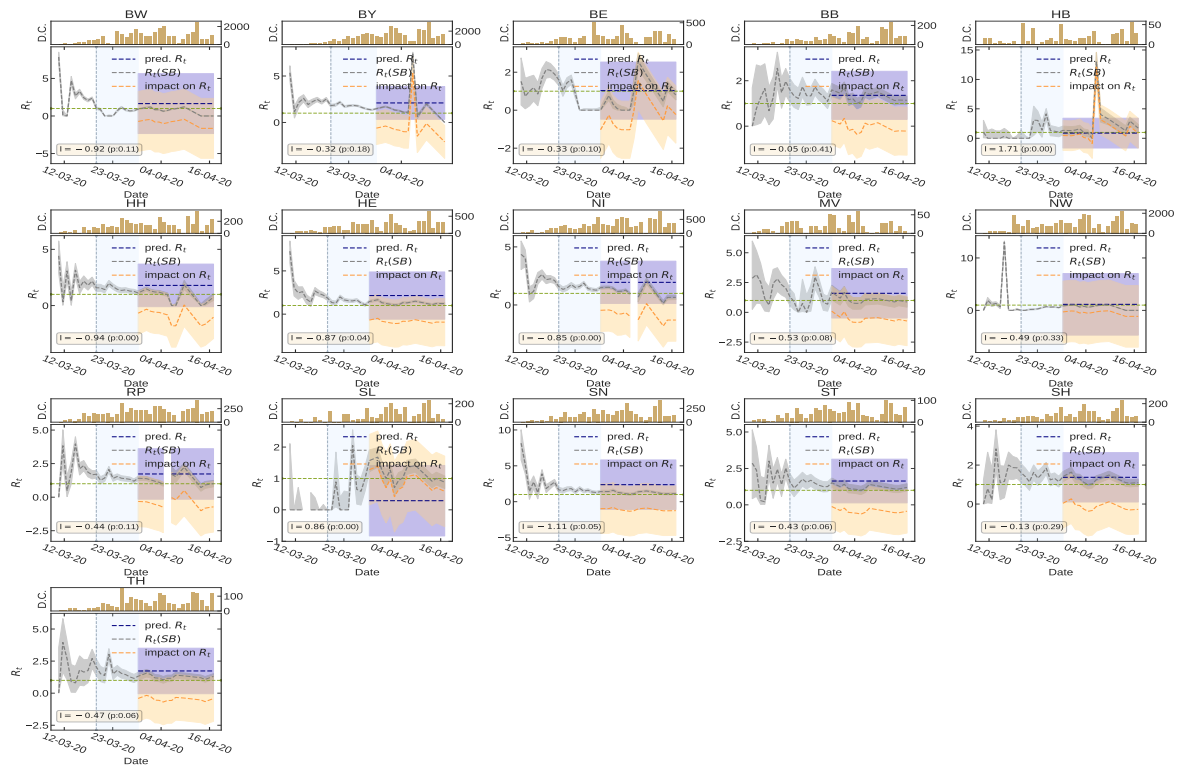


Figure D.19 Impact of Governmental interventions on  $R_t$  in different administrative divisions of Germany.



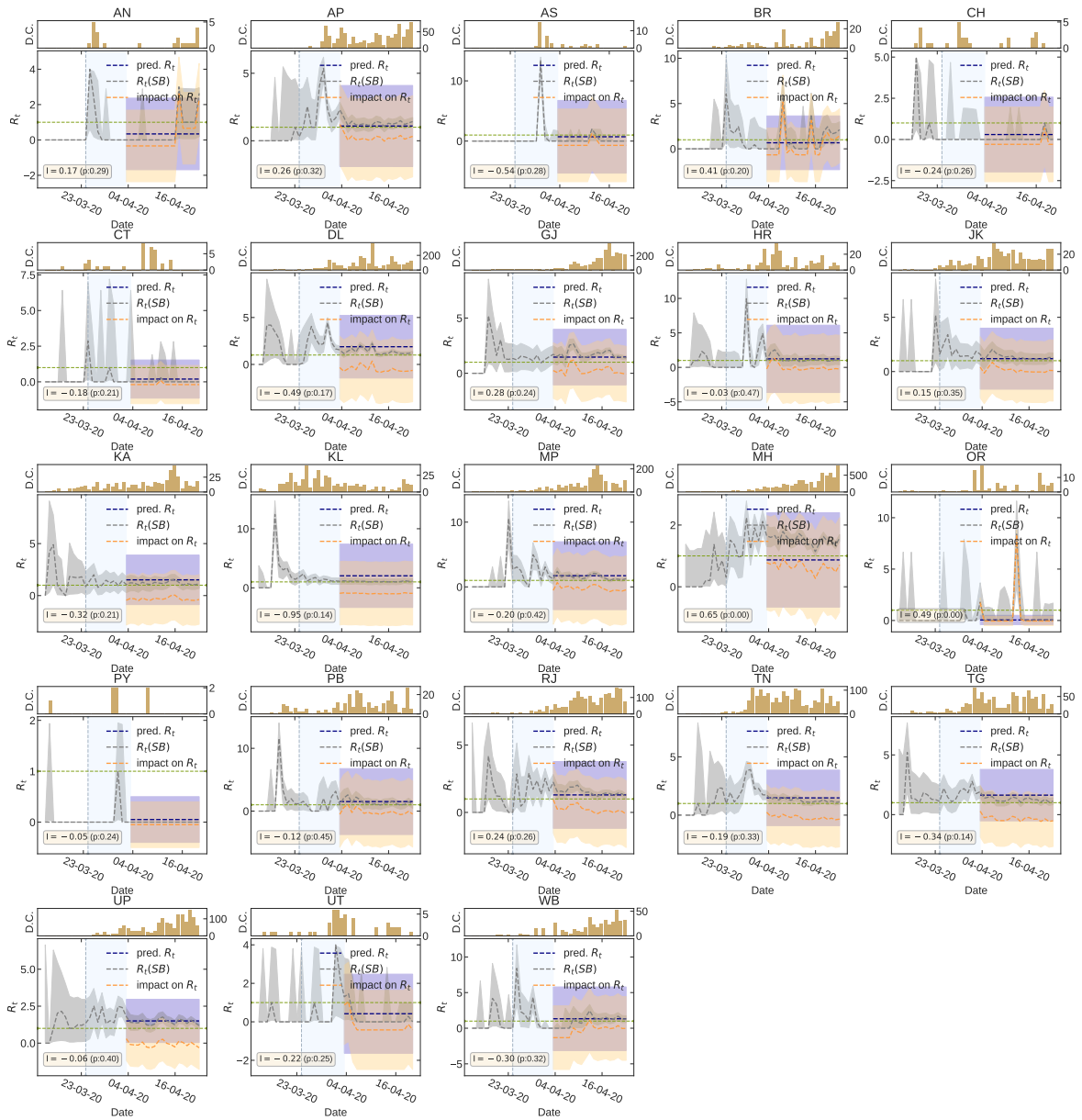


Figure D.20 Impact of Governmental interventions on  $R_t$  in different administrative divisions of India.

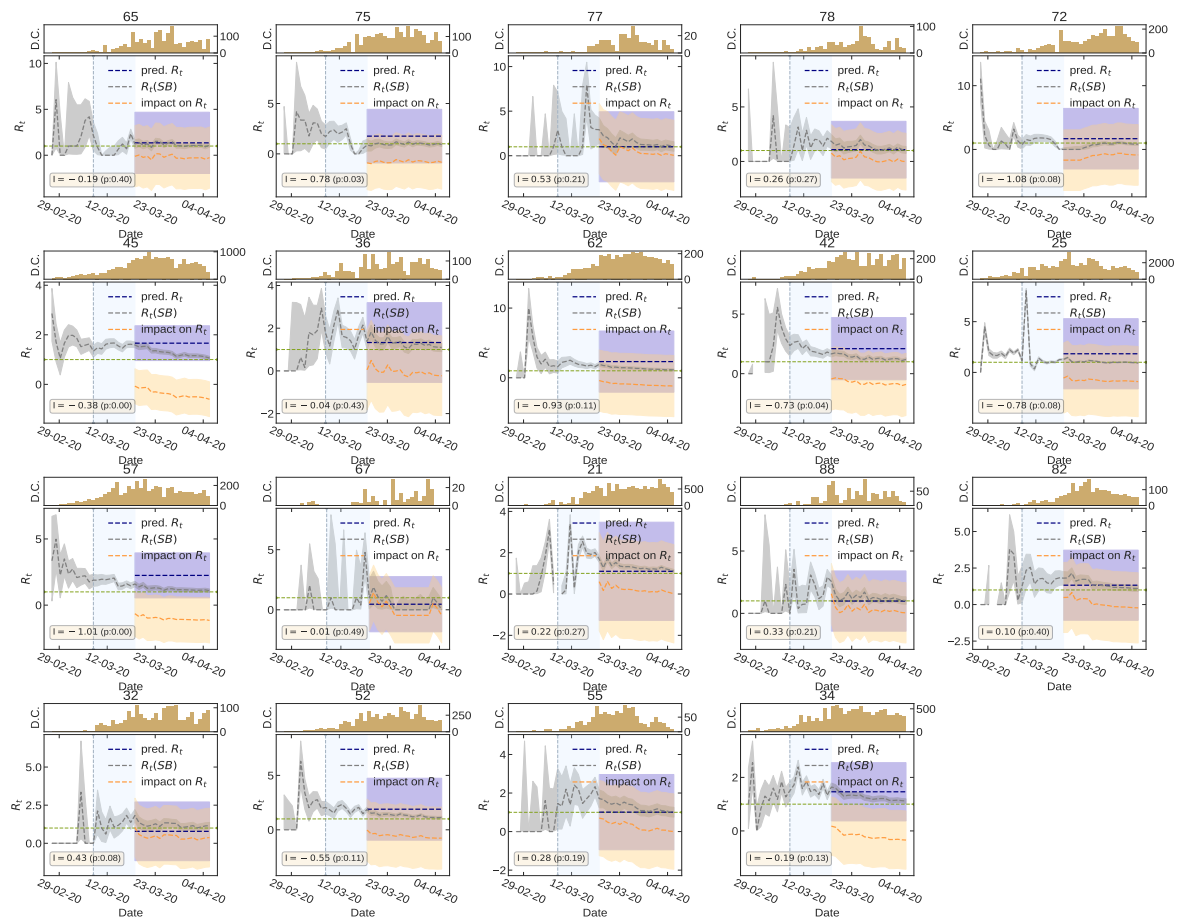


Figure D.21 Impact of Governmental interventions on  $R_t$  in different administrative divisions of Italy.

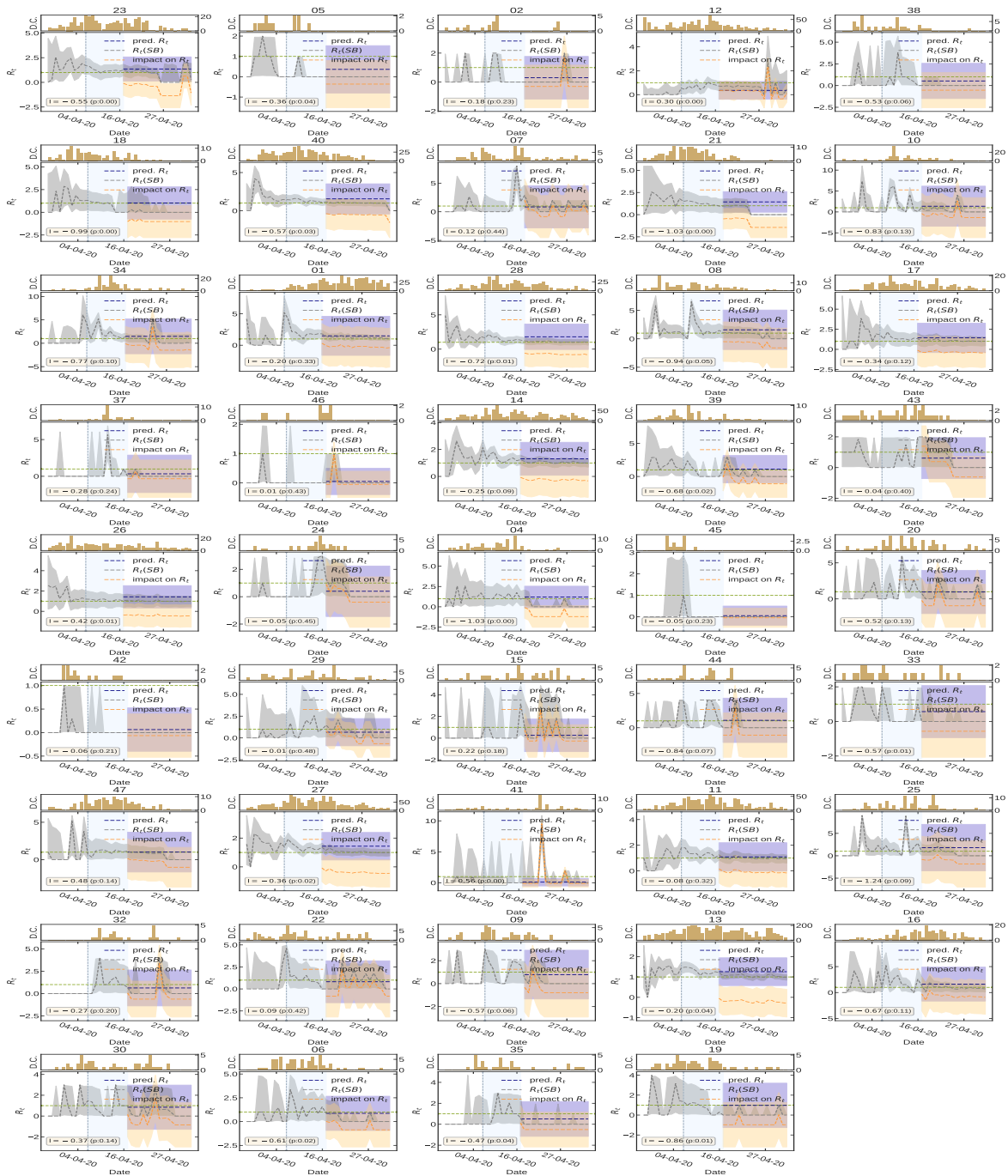


Figure D.22 Impact of Governmental interventions on  $R_t$  in different administrative divisions of Japan.

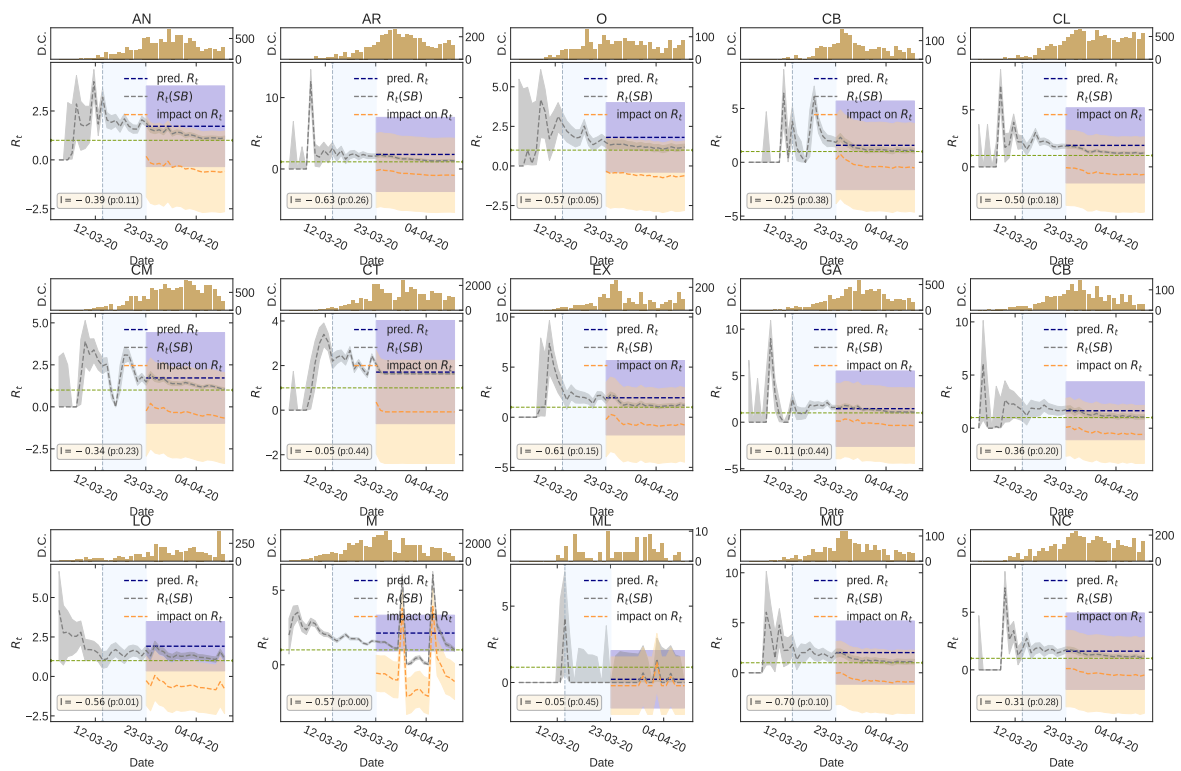


Figure D.23 Impact of Governmental interventions on  $R_t$  in different administrative divisions of Spain.

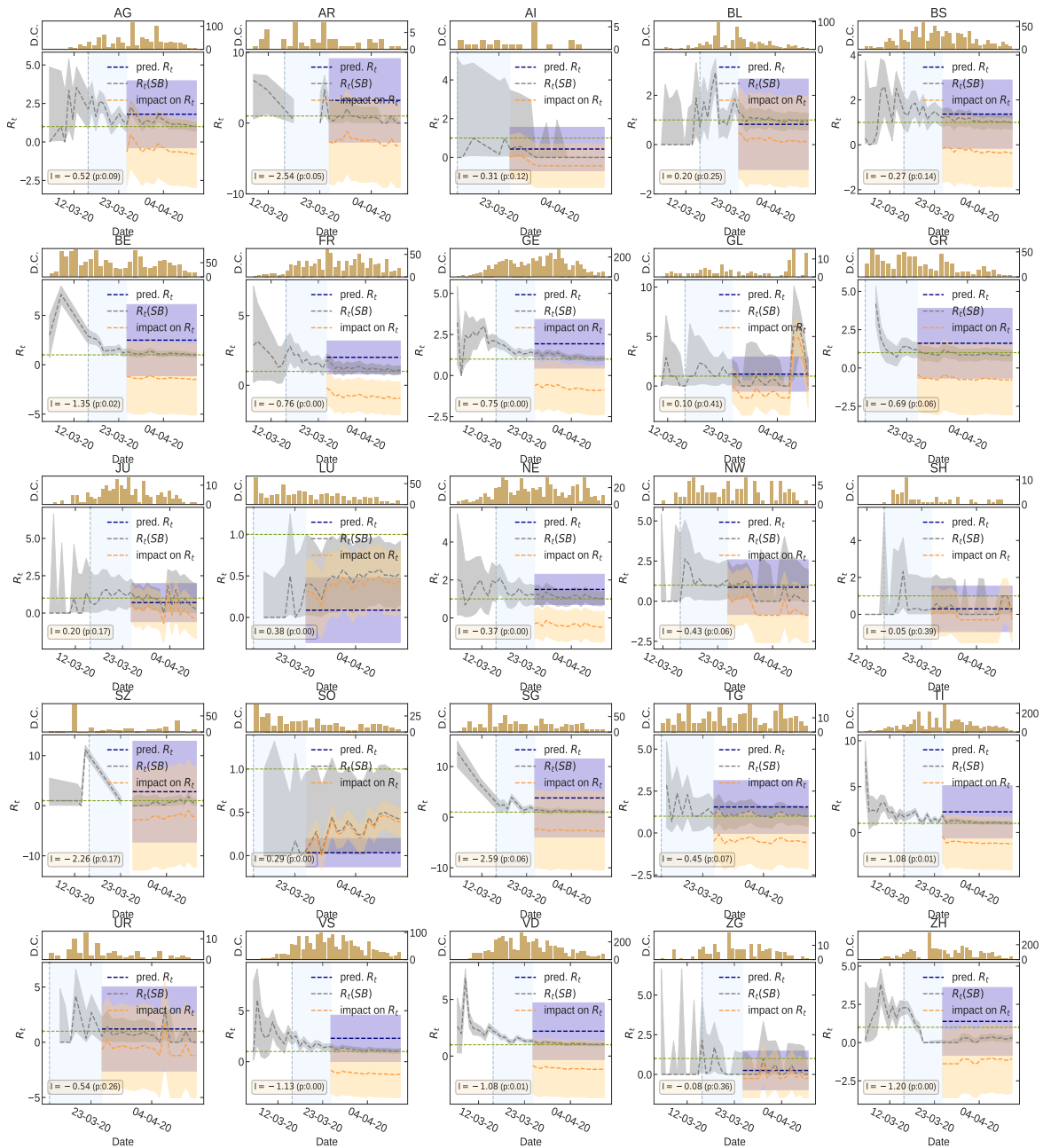


Figure D.24 Impact of Governmental interventions on  $R_t$  in different administrative divisions of Switzerland.

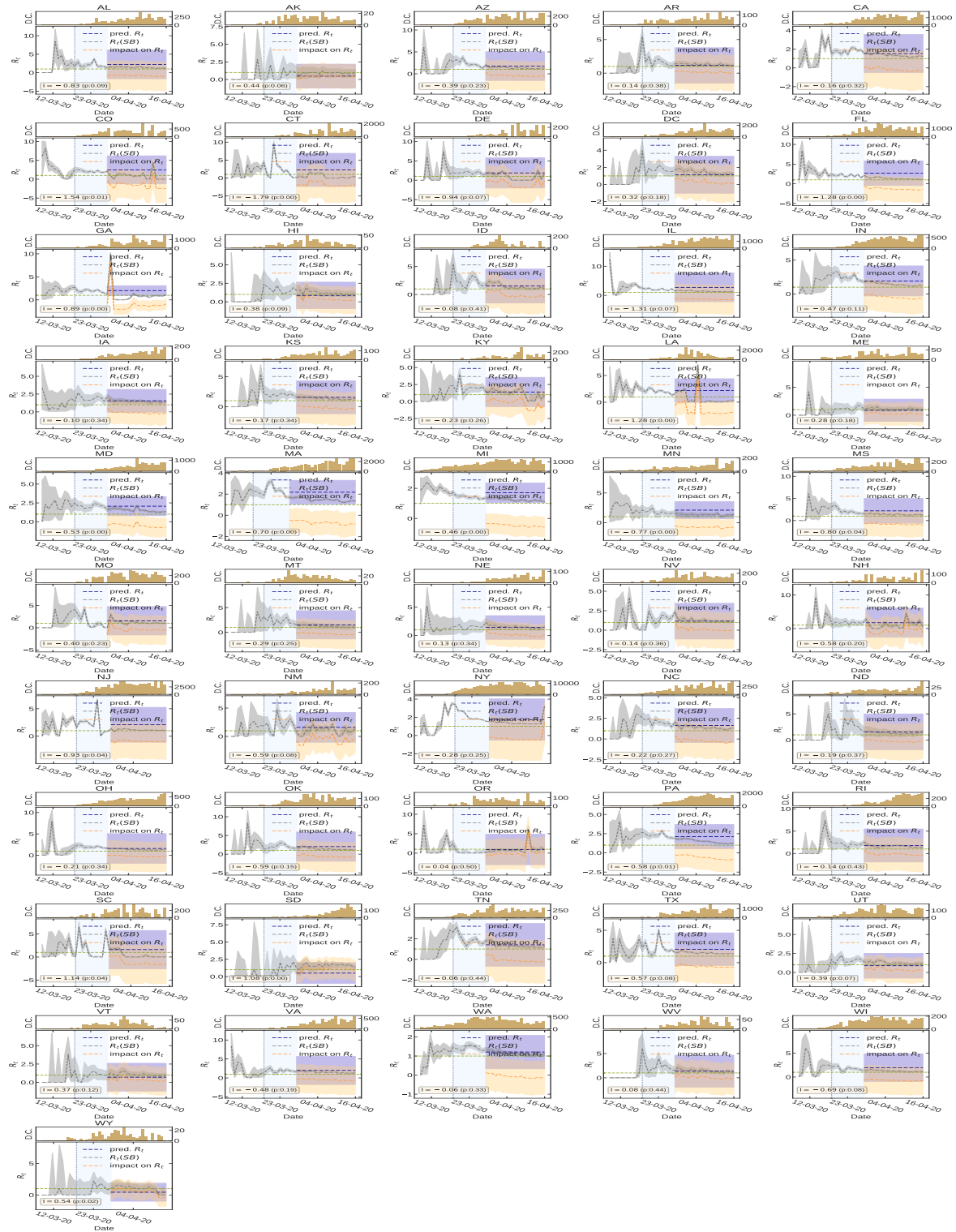


Figure D.25 Impact of Governmental interventions on  $R_t$  in different administrative divisions of United States.

# Sumit Kumar RAM

## PERSONAL DATA

---

DATE OF BIRTH: 24 Feb 1995 (age 26)  
CONTACT: [sram@ethz.ch](mailto:sram@ethz.ch)

## EDUCATION

---

OCT 2017-Jan 2021 | *PhD* in Quant Finance, **ETH Zurich**  
Supervisor: **Prof. Didier Sornette**  
AUG 2012-Jun 2017 | *BS-MS Dual degree* in Physical Sciences, **IISER Kolkata**

## QUANTITATIVE SKILLS

---

Optimization: Evolutionary Algorithms, Swarm Intelligence, Simulated Annealing, MCMC.  
Signal processing: Financial, Spatio-Temporal, Adaptive.  
Agent Based Models: Financial Markets, Evolutionary games, Networked societies, Epidemic processes.  
Bayesian Inferences: Stochastic processes, Simulation-based Inference, Time Series, Graphical Models.

## COMPUTER SKILLS

---

Languages: Python, R, Bash, MATLAB, Fortran, Mathematica.  
Databases: NoSQL (MongoDB), SQL (MySQL, SQLite). GraphDB (Neo4j)  
Front: HTML, CSS, Bootstrap, Markdown, Adobe Illustrator, Affinity Designer, JavaScript.  
Automation: Selenium.  
Software: Eikon, Datastream, Bloomberg, Eview, Stata, Tableau, LabView, ORIGIN,  $\LaTeX$ , MS Office.  
OS: MacOS, Linux (Ubuntu), Windows.

## SCHOLARSHIPS

---

JUL-AUG 2017 | Max Planck Institute for Physics of Complex Systems, Germany  
JUL-AUG 2016 | Eötvös Loránd University, Hungary  
MAY-JUL 2016 | Aalto University, Finland  
DEC-JAN 2016 | NTNU, Norway  
MAY-AUG 2015 | NTNU, Norway  
AUG-MAY 2012-17 | Dept. of Sci. and Tech., Govt. of India

## TEACHING AND PUBLICATIONS

---

TEACHING | 3 Post-Graduate courses  
SUPERVISION | 5 Master Thesis supervision  
PUBLICATIONS | 4 Peer Reviewed Journal Publications

## MEDIA COVERAGE

---

- [NRK](#)
- [GEMINI](#)
- [TU.no](#)

## PAST EMPLOYMENTS

---

RESEARCHER | Max Planck Institute, Germany  
RESEARCHER | Aalto, Finland  
RESEARCHER | NTNU, Norway

## REFERENCES

---

[PROF. DIDIER SORNETTE](#) | Professor, Department of Management Technology, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland. [dsornette@ethz.ch](mailto:dsornette@ethz.ch)